Selective Bayes: Attentional load and crowding

Peter Dayan [a,*], Joshua A. Solomon [b]

[a] Gatsby Computational Neuroscience Unit, UCL, 17 Queen Square, London WC1N 3AR, United Kingdom
[b] Department of Optometry and Visual Science, City University, Northampton Square, London EC1V 0HB, United Kingdom

ARTICLE INFO

ABSTRACT

The simple neural observation that the receptive fields of visual neurons are spatially extended lies at the heart of accounts of psychophysical phenomena to do with a sometimes unrequited need for spatial selection. In this paper, we consider its role in three anomalies associated with selective attention: the apparently undue influence of distractor stimuli when decisions in the Eriksen flanker task have to be made under time pressure; the phenomenon associated with attentional load that distractors distal to a target exert more effect when the demands on selective attention are smaller rather than larger; and the observation that crowding, a breakdown in peripheral discriminability in the presence of flankers, can under some circumstances be asymmetrical with respect to the relative proximity to the fovea of target and flanker. We show how these seeming anomalies can arise from normative Bayesian inference in the face of spatially confounded input.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

One of the ostensively less contentious forms of visual attention concerns spatial selection, i.e., suppressing or eliminating some parts of the input, whilst possibly boosting other parts (Desimone & Duncan, 1995; Driver, 2001; Pashler, 1998). A huge range of paradigms probes aspects of spatial selection, most of which quantify the influence of relatively nearer or more distant distractor stimuli on the discrimination of a target stimulus. In such tasks, distractors must be excluded from consideration, a basic requirement for selection.

Although there are many influential algorithmic models of bottom-up and top-down attention (Boynton, 2009; Grossberg, 2001; Itti & Koch, 2000; Koch & Ullman, 1985; Li, 2001, 2002; Lu & Dosher, 1998; Navalpakkam & Itti, 2005; Reynolds, Chelazzi, & Desimone, 1999; Rolls & Deco, 2002; Treisman & Gelade, 1980; Tsotsos, 1990; Wolfe, Cave, & Franzel, 1989; Yen & Finkel, 1998; Zhaoping, 2006), there is a comparative dearth (Dayan & Zemel, 1999; Yu, Dayan, & Cohen, 2009; and for discussion and review Eckstein, Peterson, Pham, & Droll, 2009; Whiteley, 2008) of the sort of accounts that are of increasing importance in many other aspects of psychology and cognitive science involving statistically normative Bayesian ideal observation, possibly in the face of confounded or corrupted input (Chater, Tenenbaum, & Yuille, 2006; Oaksford & Chater, 2007). Indeed, from an optimising viewpoint, selection appears dangerous,

since once information is discarded over the course of processing, it can never subsequently be recovered.

Here, we start from the evident fact that visual receptive fields (RFs) are extended in space, and build an extremely simplified normative model of discrimination in which spatial selection emerges as a bottom-up computational principle. The foundation of our model is a treatment (Yu et al., 2009) of a perplexing finding in the Eriksen flanker task (Eriksen & Eriksen, 1974). In a standard version of this task (shown in Fig. 1A), subjects have to make a speeded report as to whether a relatively foveal target (the central letter) is an 'S' or an 'H', ignoring flanking letters that can either be the same as (compatible) or opposite to (incompatible) the target. For consistency with later sections, we call these flanking letters distractors, since they can impact the report of the target. The odd finding is that in the incompatible condition, subjects make more than 50% errors for a range of reaction times (RTs; Gratton, Coles, Sirevaag, Eriksen, & Donchin, 1988). In the spatial uncertainty model of this (Yu et al., 2009), Bayesian inference is performed about the target based on accumulating information from inputs that integrate over regions of space. Spatial selection, which is required to separate target from distractors, emerges directly over the course of this inference process, in a computational rather than an algorithmic form. For short RTs, it is optimal to include confounded information about the distractors, leading to the large error rates.

In this paper, we consider two major extensions to this model, in each case coupled to an aspect of selective attention that has attracted substantial separate, but not always normative, study, namely attentional load and crowding. In Section 3, we consider

* Corresponding author.
E-mail addresses: dayan@gatsby.ucl.ac.uk (P. Dayan), j.a.solomon@city.ac.uk (J.A. Solomon).
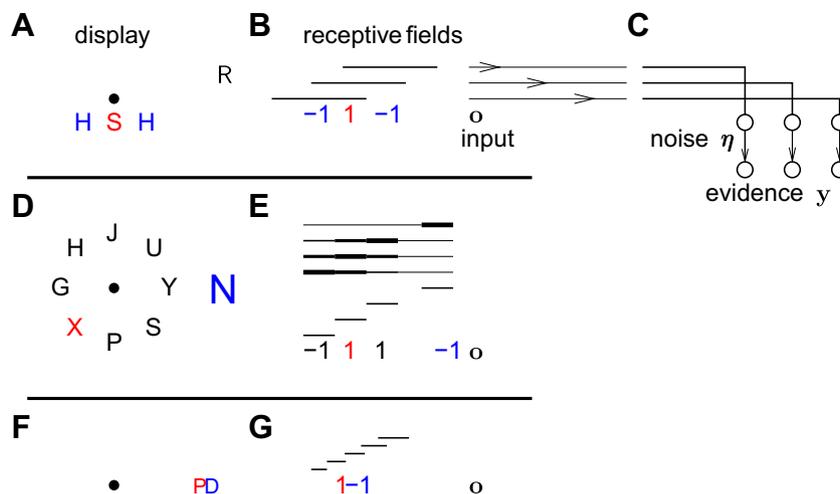
Fig. 1. Tasks and model for target discrimination. (A) Eriksen flanker task; subjects have to make a speeded judgment of the identity of the central stimulus ('S'; shown here in red; in the actual task, all stimuli are the same color), ignoring the distractor stimuli ('H's, blue) that can be the same (or require the same response) or opposite. The black dot is the fixation point (also in D and F). (B) Abstraction of the Eriksen stimuli with the input $\mathbf{o} = [-1, 1, -1]^T$ comprising a central target (here, the 1) whose sign has to be discriminated, and distractors (here, the -1's, with strength $\gamma = 1$). The stimuli are coded by spatially extended receptive fields (RFs), R, shown by the three horizontal black bars; their topographic arrangement cartooning their locations in input space. The RFs combine target and distractor stimuli linearly. (C) The generative model is completed by specifying the source of late noise $\boldsymbol{\eta}$, leading to evidence $\mathbf{y}$. Recognition based on $\mathbf{y}$ is exact as in an ideal observer (or very nearly so). (D) An example attentional load task (Lavie & de Fockert, 2003). Subjects have to indicate whether the central ring of letters contains an 'X' or an 'N', ignoring the large, salient distractor (the 'N') outside the ring. This configuration is an example of a high load task. Under conditions of low load, the ring is empty apart from the target. The non-target characters in the ring are flankers, and are relatively easily distinguished from the potential targets. In one condition, the target letter is not only smaller than the distractor, but it can also have lower contrast. (E) In the model of this task, the flankers are balanced, so they do not favour either 1 or -1. There are now two classes of RF; small ones (the short bars), and large ones (the longer bars). The thicknesses of the parts of the long bars indicate their relative preferences for parts of the visual field. The rest of the architecture is as in (C). (F) A crowding display. If presented by themselves, 'P' or 'D' could be easily discriminated; together, though, they are confused. We consider the relative influence of proximal (to the fixation point, 'P') and distal ('D') inputs on each other. (G) In the model of this, we consider RFs of increasing size (and decreasing density) progressing away from the fovea.

the effects in the model of multiple scales of receptive fields (Dayan, 2009). These turn out to offer an account of phenomena associated with attentional load (Lavie, 2005; Lavie & Tsal, 1994; Lavie, Hirst, de Fockert, & Viding, 2004). The key finding in these paradigms is that when subjects are faced with an attentionally challenging discrimination task (discriminating between a single isolated 'X' or 'N' in a ring of readily discriminable flanking letters (the 'G, H, J, U, Y, S, P') as in Fig. 1C), distractors (the large 'N' outside the ring) can easily be excluded. However, when the requirements on attention are more lax, when the target is not flanked by any letter, the same distractors can exert a significantly more deleterious effect. One interpretation of this is that there is a mandatory bandwidth for attentional selection, and so distractors will be 'let through' if nothing else occupies it. We offer an alternative, more normative, account of these findings.

Finally, spatially extended RFs also lie at the heart of some (rather more qualitative) accounts of crowding (Levi, 2008; Pelli, Cavanagh, Desimone, Tjan, & Treisman, 2007a), the phenomenon that parafoveally- and peripherally-presented letters that can be perfectly well discriminated in isolation, are hard or impossible to discriminate when presented simultaneously with distractors at appropriately small spacings, even when their existence can be properly detected (the effect of the 'D' on the 'P', or vice-versa, in Fig. 1E). Central to crowding is the scaling of this critical spacing with eccentricity – we therefore extend the original model to consider the effects of eccentricity. In particular, we focus on the anomalous asymmetry of crowding – that under some, but not all, circumstances, targets are more affected by a distal distractor than by a foveal one. Along with Motter and Simoni (2007) and van den Berg et al. (2010), we argue that this arises from the way that RF sizes increase with eccentricity, i.e. the functional geometry of the cortical representation of visual space (Cowey & Rolls, 1974; Daniel & Whitteridge, 1961; Dow, Snyder, Vautin, & Bauer, 1981; Duncan & Boynton, 2003; Gattass, Sousa, & Rosa, 1987; Johnston, 1989; Levi, Klein, & Aitsebaomo, 1985; Paradiso, 1988; Rolls & Cowey, 1970; Schwartz,

1980; Virsu & Rovamo, 1979). Finally, in Section 5, we discuss the shortcomings associated with various oversimplifications in the model, and possible extensions.

2. The model

We first describe the model in the context of the Eriksen task, since it is a very close relative of one of the two models in Yu et al. (2009) (the 'spatial uncertainty model'). In later sections we show how the other tasks can be captured through modest changes in the model. We adopt an ideal observer framework (Green & Swets, 1966), specifying the generative model that leads from the experimenter-determined stimulus to the net evidence (here called $\mathbf{y}$, which reflects the signal, any distractors, and all sources of noise), and then the recognition model that performs inference to discriminate the target.

Fig. 1A shows an example caricature of the Eriksen task. Here, the dot is the fixation point, the target stimulus is the central member of the array (the 'S'), and the distractors are the two elements in neighbouring locations (the 'H's). Subjects have to make a speeded report of the identity of the target in the face of identical or (as here) opposite distractors. As mentioned above, they show anomalous behaviour in the case of incongruency between the two.

The Eriksen task uses rather complex stimuli (such as letters) that involve conjunctions of multiple features. To analyse the basic effect of the distractors without having to model the complexities of conjunctions, we simplify the stimuli in the generative model by considering them as spartan, binary, options (such as a left or right tilted line), which we label with the numbers -1 and 1 (reserving 0 for the absence of a stimulus at a location, and allowing our units to have both positive and negative activities). Fig. 1B shows the representation of this input stimulus in the form of the (column) vector $\mathbf{o} = [-1, 1, -1]^T$, with the central 1 being the target. This vector describes the content of the input stimulus. In addition, we scale the components associated with the distractors by a factor γ to allow

them to have arbitrary strengths. For example, to model the effects of high-contrast distractors on a low-contrast target, we can set $\gamma > 1$; to model performance in the absence of distractors, we set $\gamma = 0$. These capture changes in effective contrast, since as we will see, the input signal has ultimately to compete with late noise.

These inputs are processed by RFs (shown in Fig. 1B) to provide the net evidence on which inference about the target is based. The most critical aspect of the RFs is their spatial location and extent (cartooned by the horizontal lines). That they can cover more than one input location implies that they are potentially influenced by both target and distractors. In this paper, we show how the separate tasks reveal three successively more complex aspects of the receptive fields.

Finally, the outputs of the RFs are subjected to late noise $\boldsymbol{\eta}$ (we discuss the issue of early noise later) to provide the net evidence $\mathbf{y}$. From this evidence, the recognition model has to infer the sign of the target, i.e., whether it was ± 1.

We formalise the stimulus to allow simple linear algebra. First, we write the target as $\alpha^c = \pm 1$ and the identity of the distractors as $\alpha^d = \pm 1$. Allowing for the scaling factor γ for the distractors, we can therefore write the overall input as

$$\mathbf{o} = \alpha^c \mathbf{o}^c + \alpha^d \gamma \mathbf{o}^d. \tag{1}$$

where $\mathbf{o}^c = [0,1,0]^T$ is a vector capturing the support of (i.e., the input locations covered by) the target, and $\mathbf{o}^d = [1,0,1]^T$ is a vector capturing the support of the distractors. That is to say that the second element o_2 of input vector $\mathbf{o}$ is the target; the first and third, $\{o_1, o_3\}$ are the distractors.

Receptive fields sample the input array – just three are shown in Fig. 1B and C. Here, we assume instantaneous sampling with late noise, where the neural output associated with RF R_i at timepoint τ is given by

$$y_i(\tau) = \sum_j R_{ij} o_j(\tau) + \eta_i(\tau) \tag{2}$$

where the rows of R quantify the spatial structure of the RFs (the black lines in Fig. 1B), and $\eta_i(\tau)$ is late additive noise (which is assumed to be independent between different RFs and timepoints). We assume that increases in the density or number of different sorts of RFs translate into effective reductions in the strength of the added noise η_i (since, except in the case of extreme correlations, the more neurons that report noisily on an underlying quantity, the more accurately that quantity can be determined). This simple, additive, scheme in Eqs. (1) and (2) suffices for our purposes. Most types of non-linear interaction (Fukushima, 1980; Lampl, Ferster, Poggio, & Riesenhuber, 2004; Miller, Gochin, & Gross, 1993; Reynolds et al., 1999; Rolls & Tovee, 1995; Riesenhuber & Poggio, 1999; Zoccolan, Cox, & DiCarlo, 2005) can be summarised by changing the statistics of the noise $\eta_i(\tau)$; indeed, we will see later that one main effect of the presence of nearby stimuli is increasing this late noise.

All of the putatively attentional effects we model involve changes in either the nearby distractors or, for the model of attentional load, a more remote distractor. The RFs never change.

As mentioned, the task always involves discriminating whether α^c (which is just the second element o_2 of $\mathbf{o}$) is positive or negative. Even if one RF were perfectly matched to the target, other RFs could still contain useful information about the target's identity, and it would be optimal to integrate all such sources.

Eq. (2) specifies the momentary output from the RFs. This information accumulates over time, being automatically weighted via Bayesian inference to provide a current log posterior odds ratio that $\alpha^c > 0$.[1] In the Eriksen task, subjects are often asked to provide

speeded responses. We translated this into an evidence-independent, noisy, timing process that could force early responses even for weak log likelihood ratios, overlaid on integration-to-bound inference (Smith & Ratcliff, 2004) that reported a choice when the log odds reached a threshold.

More concretely, $\mathbf{y}(\tau)$ given $\mathbf{o}$ is drawn from a multivariate Gaussian distribution, with mean $R \cdot \mathbf{o}$ (using $\cdot$ to denote matrix–vector multiplication) which only depends on the input stimulus $\mathbf{o}$ mapped through the RFs R, and covariance matrix H, which specifies the late or output noise $\eta(\tau)$. We write the density as:

$$p(\mathbf{y}(\tau)|\mathbf{o}) \sim \mathcal{N}[R \cdot \mathbf{o}, H] \propto e^{-\frac{1}{2}(\mathbf{y}(\tau) - R \cdot \mathbf{o}) \cdot H^{-1} \cdot (\mathbf{y}(\tau) - R \cdot \mathbf{o})} \tag{3}$$

We also assume a flat prior probability distribution over the target, with $p(\alpha^c = 1) = p(\alpha^c = -1) = 0.5$.

In the special case that the distractors are known to be the same (or indeed absent) on every trial, and the covariance matrix H does not depend on the input, the recognition model takes a sequence of samples $\mathcal{Y}(T) = \{\mathbf{y}(1) \cdots \mathbf{y}(T)\}$ with mean

$$\bar{\mathbf{y}}(T) = \frac{1}{T} \sum_{\tau=1}^{T} \mathbf{y}(\tau) \tag{4}$$

and evaluates the net evidence in favour of $\alpha^c = \pm 1$. The net evidence can be summarised by the log posterior odds ratio associated with the two options,

$$l_+(T) = \log \frac{P(\alpha^c = +1 | \mathcal{Y}(T))}{P(\alpha^c = -1 | \mathcal{Y}(T))}$$

which, for simple Gaussian input stimuli, as in Eq. (3), is additive in the samples, and so depends only on their mean (see Appendix B.1):

$$l_+(T) = 2T(\bar{\mathbf{y}}(T) - \alpha^d \gamma R \cdot \mathbf{o}^d) \cdot H^{-1} \cdot R \cdot \mathbf{o}^c \tag{5}$$

As expected, if the subjects can predict the nature of the distractors, then the latter do not affect inference (by virtue of having their mean values subtracted from $\mathbf{y}(\tau)$). Consider, therefore, the case that $\gamma = 0$. Then the distribution of $\bar{\mathbf{y}}(T)$ is Gaussian, with mean $\alpha^c R \mathbf{o}^c$ and covariance H/T. Thus, $\bar{\mathbf{y}}(T) \cdot H^{-1} \cdot R \cdot \mathbf{o}^c$, which, by direct calculation, has distribution

$$\mathcal{N}[\alpha^c \mathbf{o}^c \cdot R \cdot H^{-1} \cdot R \cdot \mathbf{o}^c, \mathbf{o}^c \cdot R \cdot H^{-1} \cdot R \cdot \mathbf{o}^c / T]$$

is a sufficient statistic for (i.e., contains all the relevant information in the input about) the sign of α^c, with signal-to-noise ratio $\sqrt{T \mathbf{o}^c \cdot R \cdot H^{-1} \cdot R \cdot \mathbf{o}^c}$. This signal-to-noise ratio determines the overall quality of inference, and $H^{-1} \cdot R \cdot \mathbf{o}^c$ the weightings of the various RF contributions. This depends on both their relative noise (captured by H) and their signal (captured by $\mathbf{o}^c$). The weighting shows the potentially rather subtle operation of selective attention.

For the normal version of the Eriksen task, there are two possible values of the distractors $\alpha^d = \pm 1$. Since the likelihood depends on which of these is true, assessing the inferential contribution of $\mathcal{Y}(T)$ requires summing over both possibilities, weighted by their prior probabilities, which are here equal $P(\alpha^d = 1) = P(\alpha^d = -1) = 0.5$. This is called marginalizing out the effect of α^d. The log posterior odds then become (see Appendix B.2):

$$l_+(T) = 2T\bar{\mathbf{y}}(\tau) \cdot H^{-1} \cdot R \cdot \mathbf{o}^c$$
$$+ \log \frac{\cosh \left[\gamma T (\bar{\mathbf{y}}(T) - R \cdot \mathbf{o}^c) \cdot H^{-1} \cdot R \cdot \mathbf{o}^d \right]}{\cosh \left[\gamma T (\bar{\mathbf{y}}(T) + R \cdot \mathbf{o}^c) \cdot H^{-1} \cdot R \cdot \mathbf{o}^d \right]} \tag{6}$$

For small T, the effect reported by Gratton et al. (1988) arises, with early responses in the incompatible condition ($\alpha^c = -\alpha^d$) being biased to favour $\alpha^c = \alpha^d$ slightly. This is because the spatial merging inherent in Eq. (2) would only be incompletely resolved after a short time; and the dominance of α^d from the multiple distractors would cause the error. More formally, the first, linear term on the right hand side of Eq. (6) favours $\alpha^c = \alpha^d$ (because of the expected

---

[1] It has been shown that the exact Bayesian computations can be approximated by a form of absorbing drift-diffusion process, which is itself related to the sequential probability ratio test (Liu, Yu, & Holmes, 2009).

value of $\bar{y}(T)$). This is incompletely balanced by the second, non-linear, term (involving the $\log \cosh$ expressions) for small values of T. The apparent paradox is that it would seem to be better for the subjects to shut their eyes and guess ±1 at random. This is of course not true – in fact, on trials on which the distractors are compatible with the target ($\alpha^c = \alpha^d$), the performance is better than 50% by *more* than it is worse than 50% on the incompatible trials.

Fig. 2 shows these effects in a didactic example in which the RFs substantially mix the target and the distractors (here the distractors are of intermediate strength, $\gamma = 1$; other parameters are provided in Appendix A.1). Fig. 2A shows the average value of $l_+(T)$ when the true value was $\alpha^c = 1$ across multiple samples, as a function of T for compatible ($\alpha^c = \alpha^d$; solid) and incompatible ($\alpha^c = -\alpha^d$; dashed) cases. The curves overlay the results of a large number of samples and a closed-form approximation (based on substituting $\log \cosh(\xi) \simeq |\xi| - \log 2 + \phi \exp(-\psi|\xi|)$ for suitable values of α and β; the fit is so close that the curves are indistinguishable). The fact that the dashed curve drops below 0 for small numbers of steps underlies the Gratton et al. (1988) effect, since it shows that incompatible distractors can bias choices the wrong way. The inset plot shows the standard deviation of the log odds.

Fig. 2B depicts the sensitivity of inference to the distractors in a richer way. This shows the derivative of the average log odds at each time step with respect to the factor γ that scales $\mathbf{o}^d$, again for compatible (solid) and incompatible (dashed) cases. For small T, this is positive for the compatible case, implying that making the distractors stronger would aid inference; but asymptotes at 0, since the subjects will ultimately infer the identity of the distractor stimuli and ignore them. However for the incompatible case, for

small T, boosting the strength of the distractors would have an even more devastating effect on inference; but for large T, boosting the distractors would actually help inference, since it would make them easier to to explain away. In this case, in incompatible trials, the influence of the distractors does not vanish asymptotically. Rather, they continually weigh on the assessment of the target.

Fig. 2C shows the effect of substantially boosting the distractors, setting $\gamma = 3$. In this case, the distractors, because their signature in $\mathbf{y}$ is so different from that of target, are easier to eliminate, and so only have a very fleeting negative impact on incompatible trials. Fig. 2D shows the sensitivities to scaling up the strength of the distractors. Now, since it is straightforward to infer the true value of the distractors, inference about the target is asymptotically the same for compatible and incompatible trials. Nevertheless, the distractors still exert an effect in early timesteps because of the substantial amount of noise.

We can understand the effect at large T by noting that $\log \cosh(\alpha T) \simeq |\alpha|T$ and $\bar{y}(T) \simeq R \cdot (\alpha^d \gamma \mathbf{o}^d + \alpha^c \mathbf{o}^c)$. Thus, depending on the signs of $\gamma(2\mathbf{o}^c + \alpha^d \gamma \mathbf{o}^d) \cdot R \cdot H^{-1} \cdot R \cdot \mathbf{o}^d$ and $\gamma(-2\mathbf{o}^c + \alpha^d \gamma \mathbf{o}^d) \cdot R \cdot H^{-1} \cdot R \cdot \mathbf{o}^d$ (which are the asymptotic forms of the factors inside the cosh terms), Eqs. (5) and (6) may or may not have the same asymptotic behaviour. They do for the case of Fig. 2C and D, but do not for Fig. 2A and B.

Fig. 3 shows the joint effect of the factor γ scaling $\mathbf{o}^d$ and the number of samples T on the log odds in a different way. For the compatible case (Fig. 3A), the contours are nearly vertical, i.e., inference is almost independent of the strength of the distractors. The only difference is that for values of γ near 1 for small T and near 0 for larger T, the compatible distractors speeds up inference about α^c.
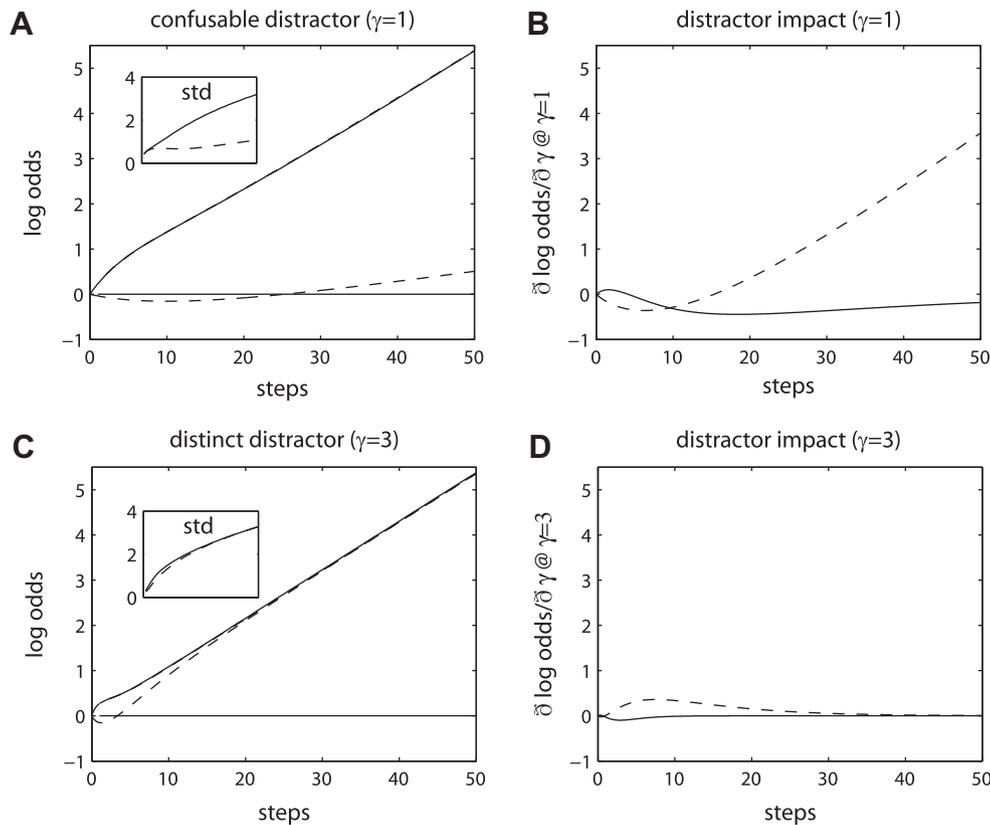

Fig. 2. Inference in the Eriksen task. (A and C) Average log odds (standard deviations in the insets) in favour of the true target α^c for compatible ($\alpha^d = \alpha^c$; solid) and incompatible ($\alpha^d = -\alpha^c$; dashed) distractor stimuli. Approximate and sampled curves for the means completely overlap. When the log odds drop below 0, the distractors are causing errors in inference. Here $\gamma = 1$. In (C), the distractors are $\gamma = 3$ times stronger than in (A; parameters in Appendix A.1); this makes the distractors easy to explain away in a bottom-up manner. (B and D) derivative of the log odds with respect to γ at $\gamma = 1$ (B) and $\gamma = 3$ (D). Here, inference in the incompatible case is improved when the distractors get *stronger* (or, of course, get set to 0). This non-monotonicity comes from the possibility of explaining away.
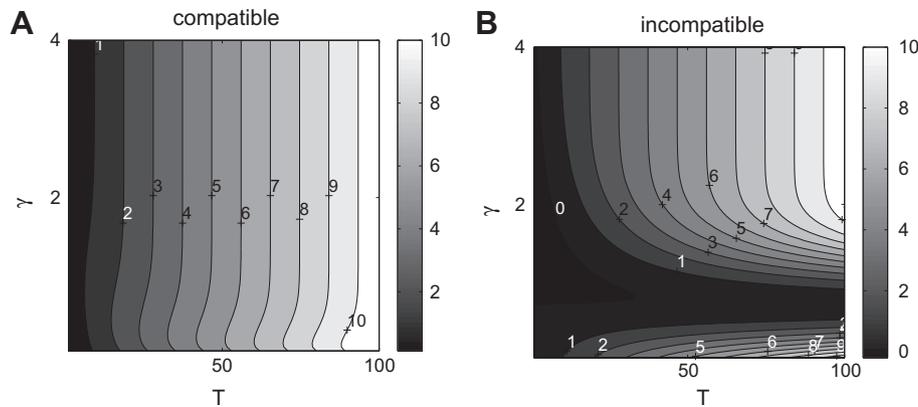
Fig. 3. Contour plot of log odds in favour of the true value of α^c as a function of the factor γ scaling $\mathbf{o}^d$ and the number T of samples for compatible (A; left) and incompatible (B; right) cases. When the contours bend left (in A), inference is faster and/or more accurate; when they bend right (in B), inference is slower and/or less accurate.

For the incompatible case (Fig. 3B), the effects of the distractors are more complex. Most obvious is the ridge of negative impact of the distractors that for small T is maximal for values of γ near 1 (the case shown in Fig. 2A), and as T increases, tends towards the value of $\gamma = 0.73$ for which the first, linear, contribution to Eq. (6) is 0. Comparing Fig. 3A and B, the devastating impact of the distractors is apparent. By contrast weaker distractors ($\gamma \rightarrow 0$) or stronger distractors (larger values of γ) exert much less malign an effect for moderate numbers of samples T. The negative impact of the distractors for small T is somewhat obscured by the contours.

In sum, this model (Yu et al., 2009) shows that human performance is consistent with a Bayesian consideration of all the available evidence for target identity. Attention need not attenuate input from the distractors. In the next two sections, we consider two extensions of the model – to RFs at multiple scales and to RFs whose sizes scale with their radial eccentricity – and study their effects in popular attentional paradigms.

3. Multiple scales and attentional load

A key limitation of the model of the Eriksen flanker task is that all the RFs had the same scale. The problem of attentional selection becomes even more stark in the case that there are multiple scales, one so fine that it includes just a single input; and one or more that are coarser, and so include extra information.

We follow Dayan's (2009) model of Lavie and de Fockert's (2003) task which assesses attentional load effects on selection (Lavie, 2005; Lavie & Tsal, 1994). Fig. 1D shows an example display from this task. There is a ring of possible locations for a target, which can be either an 'X' or an 'N'. The other positions in the ring associated with the target can either be filled with flankers (letters other than 'X' or 'N' that can readily be distinguished from the targets), or are left blank. For consistency with the other tasks, the black circle is a fixation point – in the original experiment, each possible location in the ring had a positional marker before the onset of the letters, and so fixating in the centre would be a logical strategy, without knowing the ring position at which the target would appear. Outside the ring there can be a distractor stimulus, that can be the same as, or different from, the target.

A characteristic experiment assessing attentional load compares the effect of the distant distractor on the target across two conditions: with and without the flankers. The striking and paradoxical finding is that in the seemingly *easier* case, with no flankers, the distractor exerts a *greater* effect on the target than in the seemingly harder case in which the flankers are present. We will show that our model simulates this experimental behaviour.

Fig. 1E shows the basic structure of the abstraction, for comparison with that in Fig. 1B. One key difference in the input is that the stimuli flanking the target, if present, are balanced (one is 1; the other is −1). This implies that they do not bias decision-making, but only add noise, if present.[2] The other is that there is now a distal distractor (shown as −1 in the figure). In terms of the generative model, there are now more RFs; four have the finest scale, one each per input location; the other four cover all the inputs, but favouring one more than the others (shown by the thicknesses of the horizontal bars). That the distractor is relatively farther from the array of target and flanker stimuli is reflected in the structure of the RFs. The same structure as in Fig. 1C is replicated to realise the evidence.

Fig. 4A shows this case in more detail. Here, the target is shown as c; the distractor as d and the unbiased flankers as n. Hinton diagrams are shown for the eight RFs. The finer scale RFs distinguish each input location from all the others; the broader scale inputs include all the locations, but with weights reflecting their proximities (shown in greyscale). Given the (noisy) evidence associated with these RFs, the process of inference is just as in the previous section.

Fig. 4B shows the means $\bar{\mathbf{y}}$ and inverse standard deviations (the reciprocal of the square roots of the diagonal entries of H) of the simulated inputs for the cases of low load, without flankers, and so low noise, and high load, with flankers, and so high noise; details are provided in Appendix A.2. The critical points for the model are: (a) the larger-scale RFs that include the distractor do contain some valid information about the target; but (b) in the presence of flankers, the noise associated with these RFs becomes very large. Thus, in the easier case, without flankers, the large-scale RFs will exert some influence over the log odds, and thereby allow the distractor to influence the decision. Conversely, in the harder case, the excess noise associated with the large-scale RFs will force their relative exclusion from the inference about the target, and therefore eliminate an effect of the distractor.

The solid lines in Fig. 5 show these effects. Fig. 5A shows the accumulation of the average log odds as a function of timestep within a trials in the easy, no-distractor, case for compatible (downwards triangles) and incompatible (upwards triangles) distractors. Information about the target accumulates quickly (i.e., the slopes are high), but with a big difference between compatible and incompatible cases. This implies that the distractor exerts a large effect on the processing of the target. The solid line in

---

[2] In fact, Eriksen and Eriksen (1974) used a case a bit like this in their original presentation of their flanker task, showing that it produced an intermediate impairment compared with complete incongruence.
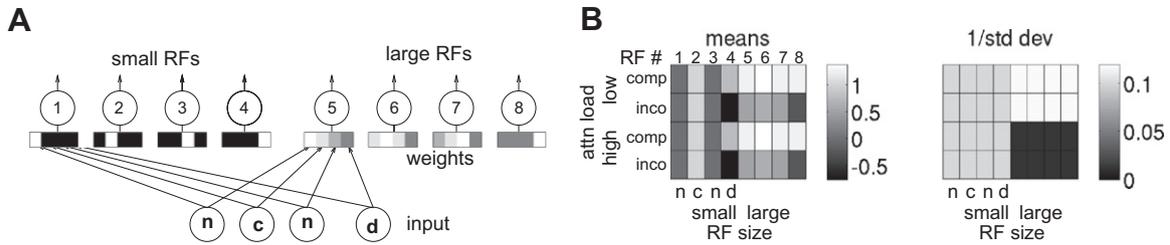
Fig. 4. Attentional load. (A) In the model, the four input units, representing flankers (n), the target (c) and the distractor (d), feed eight RFs which fall into two groups, small and large. Input weights are illustrated using Hinton diagrams. Large weights are indicated with bright boxes; small weights are indicated with dark boxes. The darkest boxes indicate a weight of zero. Each small RF has only one non-zero weight. The large RFs respond to all input elements. (B) The means (left) and the reciprocals of the standard deviations (right) for all eight RFs in four different conditions: low load/compatible distractor (top rows); low load/incompatible distractor (second rows); high load/compatible distractor (third rows); and high load/incompatible distractor (bottom rows). Here, high load is defined as the presence of flankers. Means are simply the weighted sum of flankers, target and distractor. Standard deviations have been adjusted to produce the curves in Fig. 5. Note that "attentional load" affects only the standard deviation of output from large RFs.
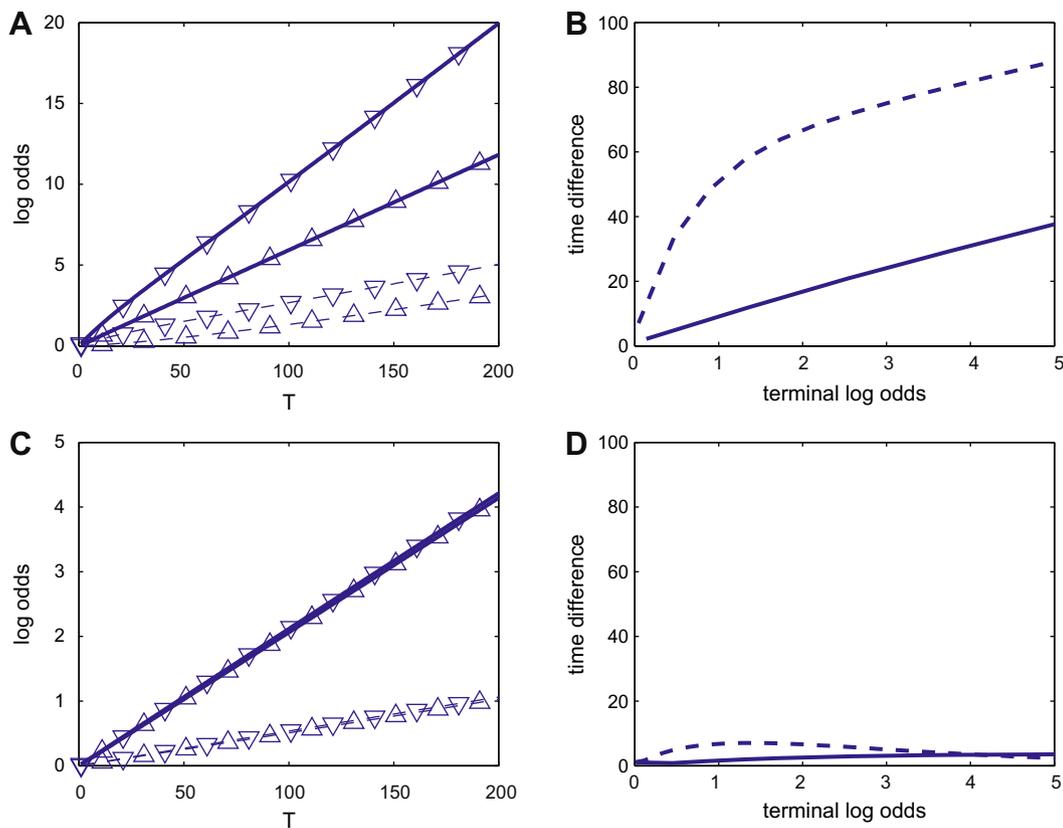

Fig. 5. Attentional load effects. (A) Average log odds as a function of timestep T within a trial in favour of the target for the case that there are no flankers. Solid lines, strong target; dashed lines, weak target; downwards triangles, compatible case (distractor and target are the same); upwards triangles, incompatible case (distractor and target are opposite). The differences between the two solid and two dashed lines show the effects being exerted by the distractor. (B) Difference in the times at which the average log odds reaches a given terminal value between incompatible and compatible cases for strong (solid) and weak (dashed) targets. Inference is slowed for the incompatible case (showing the effect of the distractor), by more for weak than strong targets. (C and D) The same as (A and B) but for the case with flankers that increase the noise associated with the large-scale RFs. Although inference is slowed, there is now barely any influence of the distractor.

Fig. 5B shows this result in a different way, reporting how much extra time it takes for the average log odds to reach a given terminal value for incompatible versus compatible distractors. The larger this is, the greater the effect of the distractors.

The solid lines in Fig. 5C and D show the same information, but now for the case that there are flankers, greatly increasing the noise associated with the large-scale RFs. In this case, inference as a whole is substantially slower (note the change in the scale of the axis reporting the average log odds in Fig. 5C); however that the curves for compatible and incompatible distractors lie directly on top of one another indicates the lack of effect of the distractor.

Fig. 5D shows the same fact, indicating that inference would take about the same time independent of the distractor.

One concern about the effects of attentional load is that the elimination of the distractor in the harder case might simply arise from the extra demands on, and thus slowing down of, inference, irrespective of the stimuli. Lavie and de Fockert (2003) tested this directly by making the task more difficult by weakening the strength of the signal. They showed that, even though this slows inference down, the distractor actually exerts a stronger rather than a weaker influence. The dashed lines in Fig. 5 show this effect. They come from a case in which the strength of the target is

halved. The slopes of the accumulation of evidence are all decreased (Fig. 5A and C) but, in the low load case, Fig. 5B, inference is actually harmed (i.e., slowed) more by an incompatible distractor, than for the case with the stronger target.

4. Asymmetries and crowding

The final effect we sought to consider in this modelling framework is associated with positional asymmetries in crowding (Levi, 2008). Crowding (Bouma, 1970, 1973; Korte, 1923) is the phenomenon that discriminating stimuli such as letters can be very much harder when they are flanked by nearby stimuli (which we again call distractors, for consistency) than when they are presented in isolation. Fig. 1F shows an example; fixating at the black dot, it is very hard to identify the two letters in the right visual field.

Crowding is important, for instance playing a key role in determining the maximum legible density of letters in print and reading rate (Pelli & Tillman, 2008; Pelli et al., 2007b). It is also extremely complex and contentious, with a large range of theoretical and experimental disputes about such things as its relationship with various forms of surround suppression and masking (Levi, Hariharan, & Klein, 2002a; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Pelli, Palomares, & Majaj, 2004; Petrov, Popple, & McKee, 2007) and whether or not crowding in the fovea is the same as crowding in the parafovea and periphery (Danilova & Bondarko, 2007; Hess, Dakin, & Kapoor, 2000; Levi, Klein, & Hariharan, 2002b; Strasburger, Harvey, & Rentschler, 1991). There are two key empirical results associated with crowding. One is that there is a critical spacing between the stimuli at which crowding is effective, which grows linearly with eccentricity, at least outside the fovea. The other, which was present already in some of the earliest reports of the phenomenon (Bouma, 1970, 1973; Shaw, 1969), and is further confirmed by more directed studies (such as Chastain, 1982, 1988; Krumhansl & Thomas, 1977; Toet & Levi, 1992), is that more peripheral distractors exert a *greater* impact on more foveal targets than vice-versa. Whether this is only true for complex stimuli such as letters or also simpler ones such as Gabor patches (Petrov & Popple, 2007) is open to question (see the supplement to van den Berg, Roerdink, & Cornelissen, 2007). The asymmetry may be somewhat unexpected, since the more foveal targets are themselves more readily detected and discriminated in the absence of a distractor. It is therefore our focus.

The models in Sections 2 and 3 already exhibit forms of crowding, that is, they exhibit interference between stimuli that are neighbouring in space. This interference in our model arises from the spatial uncertainty inherent in larger RFs. Just as suggested by van den Berg, Roerdink, and Cornelissen (2010), Motter and Simoni (2007), for instance, we might then expect that the way that the RFs scale with eccentricity should be responsible for both the empirical relationships mentioned above. Motter and Simoni (2007) explain the asymmetry by nothing that the geometry of the retino-cortical map is warped such that given a target in one location, a distractor nearer the fovea than the target is effectively a greater distance away in cortical space than a flank that is further away from the fovea than the target. This is certainly true; however results such as those by Chastain (1988), that the magnitude of the effect on the same two visual objects depending on which is the target and which is not, are not directly explained by this fact itself. Indeed, Petrov and Popple (2007) discuss a number of theories of an empirical asymmetry related to this (albeit one that is controversial; van den Berg et al., 2007), and put forward an additional one of their own. Classical models (Krumhansl, 1977; Wolford, 1975) suggested that features associated with the objects could be mislocalised nearer the fovea (possibly being duplicated along the way), and then be pooled with features actually present at those locations. This would obviously be highly non-normative.

Petrov and Popple (2007) suggested that it is only differences or contrast in features that would affect inference, but again in a manner whose inferential approximations and even basic anisotropy lack explanations.

We therefore considered whether there are circumstances under which anisotropy might arise through normative inference with overlapping spatial RFs of the form arising as in Motter and Simoni (2007). The obvious basis for anisotropy is that receptive fields in general get larger (cartooned by the RFs in Fig. 1G), the further are their centers from the fovea. What is more tricky is understanding circumstances under which this leads to the combination of *decreasing* signal-to-noise ratio for unperturbed targets as a function of eccentricity, and yet the *greater* effect of the more peripheral distractors on the more foveal target than vice-versa when one might expect the latter to enjoy *smaller,* and hence less confounded RFs.

Fig. 6A shows a very simple analysis of this. There are two possible target locations: 'inner', at eccentricity e_{in}, which is closer to the fovea, and 'outer', at eccentricity $e_{out} = e_{in} + \partial e$, which is farther away. Given spatially distributed RFs, there are three classes of RFs which we represent by three units: y_{in}, which is influenced by the inner and not the outer target location, y_{out}, influenced by the outer and not the inner target location, and y_{both}, influenced by both. In turn, the coding of the two locations is controlled by the three factors in the model – (i) the strength of the connection (the entries in R), which decrease with eccentricity:

$$R = \begin{pmatrix} R_{in} & 0 \\ R_{both}^{in} & R_{both}^{out} \\ 0 & R_{out} \end{pmatrix} \tag{7}$$

(ii) the density of the RFs, which we consider to be modelled by the added noise η_i, whose variance increases with eccentricity

$$H^{no\ flank} = \text{diag}(H_{in}, H_{both}, H_{out}) \tag{8}$$

and finally (iii) the effect of having a distractor in one of the locations and the target in the other

$$H^{flank} = \text{diag}(H_{in}, H_{both} + H_{flank}, H_{out}), \tag{9}$$

which is the same in both cases.

Fig. 6A shows a cartoon of the first two factors (the lower plot shows R; the upper H). Crudely, the phenomenology of the effect suggests that the representation of the inner location depends more on y_{both} than does that of the outer location (one requirement for which is that $R_{both}^{in} > R_{both}^{out}$). Then, if the simultaneous presence of target and flank makes the noise $H_{both} + H_{flank}$ sufficiently large to render this information significantly less useful, then the inner location can ultimately suffer more from an outer distractor than vice-versa. In the limit that y_{both} is useless in the presence of a distractor, the key constraint concerns the swapped signal-to-noise ratio for the remaining inner and outer targets. This is satisfied if

$$\frac{R_{out}}{R_{in}} > \sqrt{\frac{H_{out}}{H_{in}}}.$$

This places a constraint on the shape of the RFs, such that their outward tails are subject to less overlap than their inward tails relative to their densities.

Quantifying this effect precisely is difficult, since it depends on how the sizes and shapes of RFs depend on eccentricity (which itself depends on the distribution of their spatial scales), and on the cortical magnification factor. One crude way to proceed is to start from the function $\rho(e)$, reporting the half-width of RFs centered on eccentricity e, and quantify how many would be represented by each of the three categories y_{in}, y_{out} and y_{both}. Fig. 6B shows how
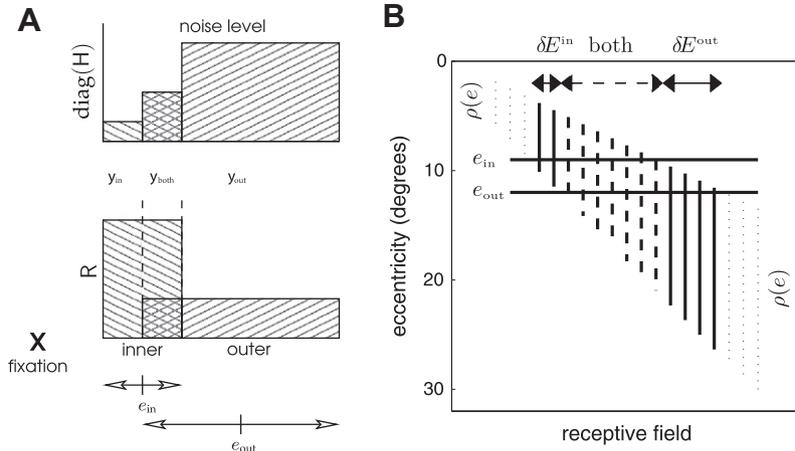
Fig. 6. Asymmetry in crowding. (A) Exaggerated three-unit cartoon of the basis of asymmetry in crowding. Inner and outer targets are covered by successive RFs. Some RFs (leading to y_{in}) are exclusive to the inner target, and so are unaffected by an outer distractor. Some (leading to y_{out}) are exclusive to the outer target, and so are unaffected by an inner distractor. The remainder are covered by both (leading to y_{both}). Two factors control the effects of the RFs – the weights R, which we consider come from coverage of RFs, determined by their sizes $\rho(e)$, and the noise H, controlled by the cortical magnification factor $M(e)$. (B) Illustration of how a linear form of $\rho(e)$ leads to the excess size of δE^{out} compared with δE^{in}. The long horizontal lines show e_{in} and e_{out}; the vertical lines show sample RFs centered on increasing eccentricities with widths growing according to $\rho(e)$. The dotted RFs are not influenced by inner or outer targets; the dashed RFs are associated with y_{both} (delineated by the dashed arrow on top of the RFs); the solid RFs with y_{in} (the solid arrow labelled as δE^{in}) and y_{out} (the solid arrow labelled as δE^{out}). If function $\rho(e)$ grows fast enough, then the outer tails will be significantly larger than the inner tails.

to do this – for instance RFs centered at eccentricity e will fall under y_{in} if

$$e - \rho(e) \leqslant e_{in} \leqslant e + \rho(e) \quad \text{and} \quad e + \rho(e) < e_{out} \tag{10}$$

and under y_{out} if

$$e_{in} < e - \rho(e) \quad \text{and} \quad e - \rho(e) \leqslant e_{out} \leqslant e + \rho(e) \tag{11}$$

There are data on $\rho(e)$ in various species of monkeys from V1 from a number of authors, including Dow et al. (1981), Gattass et al. (1987), Van Essen, Newsome, and Maunsell (1984), and in V4 from Motter, 2009. Although the fovea is clearly special (a fact that we would associate also with the complexities of foveal crowding; Danilova & Bondarko, 2007; Hess et al., 2000; Levi et al., 2002b; Strasburger et al., 1991), outside it, $\rho(e)$ follows the roughly linear form

$$\rho(e) = \alpha + \beta e,$$

with $\alpha = 0.42$; $\beta = 0.03$ in capuchin V1 (Gattass et al., 1987) and $\alpha = 0.70$; $\beta = 0.36$ in macaque V4 (Motter, 2009). The geometry of the overlap implies that the range of eccentricities in y_{in} and y_{out} are respectively

$$E^{in} = \left[\frac{e_{in} - \alpha}{1 + \beta}, \min\left\{\frac{e_{out} - \alpha}{1 + \beta}, \frac{e_{in} + \alpha}{1 - \beta}\right\}\right] \tag{12}$$

and

$$E^{out} = \left[\max\left\{\frac{e_{in} + \alpha}{1 - \beta}, \frac{e_{out} - \alpha}{1 + \beta}\right\}, \frac{e_{out} + \alpha}{1 - \beta}\right]. \tag{13}$$

Assuming that ∂e is sufficiently small that the RFs centered at e_{in} and e_{out} cover both inner and outer targets, we have that the sizes of these ranges are $\delta E^{in} = \partial e/(1 + \beta)$ and $\delta E^{out} = \partial e/(1 - \beta)$. Since $\beta > 0$ (as RFs grow with eccentricity), the outer tail is indeed longer than the inner tail; however, whereas for the value of β associated with V1, this is only by 6%, for V4, this is by around 200%.

Completing this account requires three more specifications: (i) how the scales of δE^{in} and δE^{out} translate to R_{in} and R_{out}; (ii) the effective noise terms H_{in} and H_{out} corrupting these signals associated with the density of appropriate RFs; and (iii) the values associated with 'both'. One critical fact is that psychophysical acuity appears to be controlled by the size of the image on cortex of the representation of the input, which in turn has been argued to be

the product of the cortical magnification factor $M(e)$, which determines how the centre of the cortical representation scales with eccentricity, and $\rho(e)$ (see Cowey & Rolls, 1974; Daniel & Whitteridge, 1961; Dow et al., 1981; Duncan & Boynton, 2003; Levi et al., 1985; Gattass et al., 1987; Paradiso, 1988; Rolls & Cowey, 1970; Virsu & Rovamo, 1979). $M(e)$ is roughly inversely proportional to e in V1 (estimated as $7.8e^{-0.94}$ for the capuchin monkey by Gattass et al., 1987 and $18.5/(1 + 1.2e)$ for humans in Duncan & Boynton, 2003), and it has been argued by Motter (2009) that V4 simply inherits this scaling from V1. Indeed the general form of the relationship with eccentricity is preserved for letter discrimination in humans too (Anderson & Thibos, 1999a, 1999b).

In total, we therefore make the three crude assumptions that:

$$(1) \quad R_{in} \propto \delta E^{in}; \quad H_{in} \propto \frac{1}{(M(e_{in}))^2} \tag{14}$$

coming from the regions devoted to each location individually, with noise associated with the inverse of the density of RFs;

$$(2) \quad H_{both} = \frac{1}{(M(e_{in} + \partial e/2))^2} \tag{15}$$

approximating the density associated with y_{both} from half way between e_{in} and e_{out}; and

$$(3) \quad \frac{(R_{in})^2}{H_{in}} + \frac{(R_{both}^{in})^2}{H_{both}} \propto \frac{(\rho(e_{in}))^2}{H_{in}}; \quad \frac{(R_{out})^2}{H_{out}} + \frac{(R_{both}^{out})^2}{H_{both}}$$

$$\propto \frac{(\rho(e_{out}))^2}{H_{out}} \tag{16}$$

coming from the overall acuity associated with targets at inner and outer locations in the absence of distractors.

The dotted line in Fig. 7A shows the ratio $(\rho(e_{out})/\sqrt{H_{out}})/(\rho(e_{in})/\sqrt{H_{in}})$ between the signal-to-noise ratio for the unflanked inner and outer targets (via assumption 3, and using the parameters associated with V4). The inner target has a slightly higher signal-to-noise ratio because of the cortical magnification factor, making the ratio slightly less than 1. The solid and dashed lines in the figure show the ratios $(R_{out}/\sqrt{H_{out}})/(R_{in}/\sqrt{H_{in}})$ between the signal-to-noise ratio for flanked targets (for the parameters for V4 and V1 respectively), taking advantage of the first
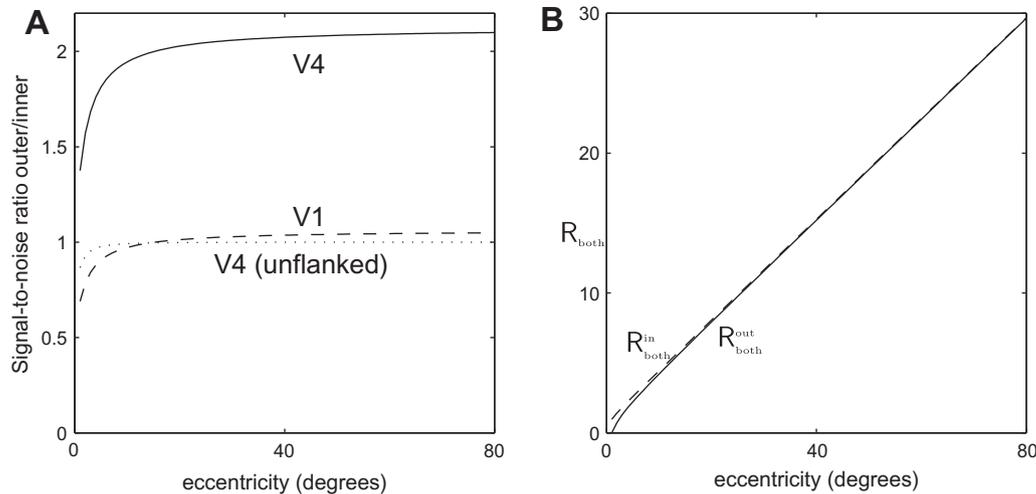
Fig. 7. Asymmetry in crowding. (A) The ratio between the signal-to-noise ratio associated with outer versus the inner targets (values greater than 1 favour outer targets). Solid: ratio for the flanked targets using $\rho(e)$ associated with V4; dashed: ratio for flanked targets using $\rho(e)$ associated with V1; dotted: ratio for unflanked targets with parameters for V4. Although there is a relatively small preference for the inner target in the absence of distractors, there is a large benefit for the outer target in their presence. (B) Values of R_{both}^{in} and R_{both}^{out} associated with the full signal-to-noise ratios for the unflanked inner and outer targets.

assumption above. The key asymmetry, that the outer target is left much better off than the inner one, is clear for the parameters associated with V4.

Fig. 7B uses the third assumption above to show the growth of R_{both}^{in} and R_{both}^{out} with eccentricity. The tiny difference between the two supports the small excess signal-to-noise ratio for the unflanked inner target. By comparison the asymptotic values as the eccentricity grows are $R_{in} = 0.74$, and $R_{out} = 1.56$.

In sum, according to this model, the asymmetry between inner and outer locations arises from the way that the sizes of receptive fields grow with eccentricity. The growth implies that the outer tails, whose RFs are not influenced by a confusing inner distractor, are larger than the inner tails, whose RFs are not influenced by a confusing outer distractor. By making a key assumption about the mapping from the RFs to the signal-to-noise ratio associated with cortical processing, we derive the basic asymmetry. However, the magnitude of the asymmetry depends on the rate of growth, which in turn depends on the cortical area concerned. For V1, and so putatively simple patterns such as oriented Gabors, the effect is minimal. For V4 and beyond, associated with more complex inputs such as letters, the effect is much greater.

Note that as ∂e gets larger, fewer RFs will be affected by both inner and outer locations, and so the interference will reduce. The size of the region of interference also scales linearly with eccentricity, consistent, as in Motter and Simoni (2007), with the common finding that the zone of crowding (independent of the asymmetry) scales in this manner (Levi, 2008).

5. Discussion

We have built an extremely simple model of spatially extended receptive fields, and showed quantitative cartoons of the way it addresses three characteristic phenomena associated with spatial attention: the Eriksen flanker task, attentional load, and anisotropic interference in crowding. The model involves only Gaussian stimuli and noise and Bayes-optimal marginalization and inference.

For clarity, we further simplified the presentation of the (already spartan) model. In particular, unlike the original accounts in Dayan (2009), Yu et al. (2009), we did not model the speeded integration-to-bound accumulation process (Smith & Ratcliff, 2004) leading to the decision, and only reported average log pos-

terior odds in favour of particular targets. This is of special consequence for Fig. 5C and D, because of the non-linearity associated with tuning log odds into a decision; however, the basic point about the influence exerted by distractors remains. We also only modelled late noise ($\eta_i(\tau)$) and explicit flankers and distractors. It would be formally straightforward to consider early or equivalent noise (corrupting **o**), which would lead to correlations in the noise in **y** coming from the extended RFs. However, we would need to consider the effect on this of normalising interactions (Carandini & Heeger, 1994; Carandini, Heeger, & Movshon, 1997; Zoccolan et al., 2005).

There are at least three further important directions for adding richness and complexity to the model. The first (a) is to use distinct visual features (e.g., oriented bars, junctions or letters) rather than positive or negative numerical values. This would allow us to consider the general effects of statistical similarity and grouping among the distractors and between targets and distractors and the complexities of competition between more or less similar subparts of letters when they are targets and distractors (Felisberti, Solomon, & Morgan, 2005; Gheri, Morgan, & Solomon, 2007; Kooi, Toet, Tripathy, & Levi, 1994; Livne & Sagi, 2007; Louie, Bressler, & Whitney, 2007). It would also allow us to investigate more fully issues such as the apparent absence of crowding for feature detection rather than discrimination (Andriessen & Bouma, 1976; Levi et al., 2002b; Pelli et al., 2004). Indeed, Chastain (1988) tested asymmetry not by comparing inner and outer targets with inner and outer distractors (as we did in Section 4), but rather distracting with by more and less confusable stimuli. One basic phenomenon is the same – that the inner target is more affected by a less confusable outer distractor than is the outer target by the same inner distractor. However, we would need to model confusability itself to capture the other phenomenon, that increasing the confusability of the distractor has a much greater effect on the outer than the inner target.

A second direction (b) would be to allow hierarchies of these features (as, for instance, in Fukushima (1980) & Riesenhuber & Poggio (1999)), with RFs in one layer feeding into RFs in the next, and with Bayesian integration happening simultaneously in all layers. A feature-based model of this sort would also allow a more straightforward notion of spatial pooling – one of the anomalies of the current model of attentional load is that strong distractors are more distinguishable from targets by virtue of being too strong;

if the subparts of RFs reported in a more all-or-nothing manner on the presence of relevant features, then this would not occur. It would also (c) be interesting to consider other forms of contextual interactions such as the divisive normalisation that we have previously (Schwartz, Sejnowski, & Dayan, 2006) argued arises from a Gaussian scale mixture model for natural scenes (Wainwright, Simoncelli, & Willsky, 2001), and used to account for effects such as the tilt illusion (Schwartz, Sejnowski, & Dayan, 2009). Indeed, masking interactions like surround suppression can also arise in statistically normative models of contextual interactions (Schwartz & Simoncelli, 2001, 2006; Schwartz et al., 2009). On the other hand, the fact that crowding scales with eccentricity rather than target size argues against these sorts of contextual interactions as an explanation for crowding, since their scale is determined by natural scene statistics rather than the way these statistics are sampled by a variable resolution visual system, and so would be dependent on target size rather than eccentricity (for a discussion of this, see Levi (2008)).

In considering the hierarchical extension of the model, we should return to a central issue for the model that it is not possible to change how the input **o** is processed by manipulating the RFs R; rather, attention operates at or beyond the RF integration that produced **y**. If attention could directly affect R, then there would be no spatial pooling or uncertainty, since it could all be removed at the first instance; that so many phenomena argue for spatial uncertainty argues that the ability to manipulate R is at best very limited. Of course, it would be statistically optimal to be able to exclude $\mathbf{o}^d$ from the outset; however, doing so would either require the simultaneous bottom-up construction of all possible RFs or the fine top-down control of the microstructure of low-level RFs. Either of these solutions imposes large wiring and processing costs. In the context of a processing hierarchy, however, the output **y** of one layer is the input **o** to the next, and so attention can have these more complex effects.

In terms of the model of attentional load – the original hypothesis (Lavie, 2005) suggested that when little attention is required to solve the set task, inputs associated with distractor stimuli *leak* through with little attenuation, and so cause disruption; when the task is difficult, attention is totally occupied with the set task, leaving nothing left over. By contrast, we have suggested that an inferential model taking advantage of all the information in the input will show exactly the same characteristic, with the key issue being whether the units with large RFs, which include the distractor, are rendered useless by the flankers that make for the high load in the first place. The advantage of this version of an attenuation theory (Treisman, 1960, 1969) of attention is that it obviates the requirement to appeal to an inexplicable inefficiency, over and above the existence of units with large RFs, and indeed relates this set of selective attentional tasks to the wide range of other accounts of probabilistically-correct sensory inference.

In terms of crowding – we adopted Motter and Simoni (2007)'s conclusion that the essential feature and the anisotropy of interference is a straightforward consequence of spatial pooling (see also van den Berg et al., 2010); however, the details of the anisotropy are less mandatory for models in this class. It arose for us because of the structure of noise induced by the sharing of RFs, sharing that differs between V1 and V4 because of the differing cortical magnification and RF size scaling. It would be particularly interesting in this case to consider the effect of correlations induced by early noise, and also to simulate the additional effect captured by van den Berg et al. (2010) that the strength of the anisotropy depends non-monotonically on the exact spacing.

One other important characteristic of crowding that we have not so far modelled is the apparent fact that when a target is competently crowded, no increase in observation time would improve the ability to localise and bind together the spatial features associ-

ated with the target. One obvious possibility for modelling this is to allow for large spatial uncertainty and leaky information accumulation, so that there is a (possibly low) asymptotic certainty about the target. However, it would be interesting to seek a more direct experimental test of this.

Perhaps the model's most important characteristic, along with those in (Dayan, 2009; Yu et al., 2009) is that it lacks an explicit attentional mechanism in inference which has the capacity to downplay some input units over others. The model does know the location of the targets, and automatically, through inference, focuses all its resources on it. However, it lacks any way of boosting or suppressing some receptive fields compared with others. In other words, the form of selection it considers is an *output* from inference rather than an *input* into it.

Acknowledgments

Appendix A. Parameters

A.1. Eriksen task

For the model of the Eriksen task (Fig. 2), we consider a slight adaptation of the parameters from Yu et al. (2009). Here, there are three input units $j = -1, 0, 1$, with $\mathbf{o}^c = [0,1,0]^T$ and $\gamma \mathbf{o}^d = \gamma[1,0,1]^T$, with $\gamma = 1$ for Fig. 2A and B and $\gamma = 3$ for Fig. 2C and D. The other key parameters are

$$\mathrm{R} = \begin{pmatrix} 1.1 & 0.9 & 0.0 \\ 0.9 & 1.1 & 0.9 \\ 0.0 & 0.9 & 1.1 \end{pmatrix} \quad \mathrm{H} = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & 0 & 0 \\ 0 & \sigma_1^2 + 2\sigma_2^2 & 0 \\ 0 & 0 & \sigma_1^2 + \sigma_2^2 \end{pmatrix}$$

where $\sigma_1 = 6$, $\sigma_2 = 3.5$, loosely reflecting the requirement that greater signals are associated with greater noise.

A.2. Attentional load

For the model of the attentional load task (Figs. 4 and 5), we use a slight adaptation of the parameters from Dayan (2009). Here, there are four input units, with $j = -1, 0, 1, 5$, with $\mathbf{o}^c = \gamma[0,1,0,0]^T$, with $\gamma = 1$ for the solid lines in Fig. 5 and $\gamma = 0.5$ for the dashed lines, $\mathbf{o}^d = [0,0,0,1]^T$, and with flankers that are modelled as just increasing the late noise. The other key parameters are

$$\mathrm{R} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1.0 & 0.9 & 0.7 & 0.5 \\ 0.9 & 1.0 & 0.9 & 0.5 \\ 0.7 & 0.9 & 1.0 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1.0 \end{pmatrix}$$

$$\mathrm{H} = \mathrm{diag}\left[\sigma_f^2, \sigma_f^2, \sigma_f^2, \sigma_f^2, \sigma_l^2, \sigma_l^2, \sigma_l^2, \sigma_l^2\right]$$

where $\sigma_f = 10$ (for the fine-scale RFs) and $\sigma_l = 8.5$ for the large-scale RFs when there are no flankers (Fig. 5A and B) and $\sigma_l = 80$ when there are.

Appendix B. Formulæ

B.1. Eq. (5)

We derive Eq. (5) from Eq. (3). From the latter equation and the knowledge of the distractors (and up to a constant which is $-\log\det(H) - n\log(2\pi)$ when the dimensionality of $\mathbf{y}$ is n):

$$
\begin{aligned}
2\log p(\mathbf{y}(\tau)|\alpha^c = +1) &= -(\mathbf{y}(\tau) - R\cdot\mathbf{o}^c - \alpha^d\gamma R\cdot\mathbf{o}^d)\cdot H^{-1}\cdot(\mathbf{y}(\tau) \\
&\quad - R\cdot\mathbf{o}^c - \alpha^d\gamma R\cdot\mathbf{o}^d) \\
&= -(\mathbf{y}(\tau) - \alpha^d\gamma R\cdot\mathbf{o}^d)\cdot H^{-1}\cdot(\mathbf{y}(\tau) - \alpha^d\gamma R\cdot\mathbf{o}^d) \\
&\quad - \mathbf{o}^c\cdot R\cdot H^{-1}\cdot R\cdot\mathbf{o}^c + 2(\mathbf{y}(\tau) \\
&\quad - \alpha^d\gamma R\cdot\mathbf{o}^d)\cdot H^{-1}\cdot R\cdot\mathbf{o}^c
\end{aligned} \tag{17}
$$

$$
\begin{aligned}
2\log p(\mathbf{y}(\tau)|\alpha^c = -1) &= -(\mathbf{y}(\tau) + R\cdot\mathbf{o}^c - \alpha^d\gamma R\cdot\mathbf{o}^d)\cdot H^{-1}\cdot(\mathbf{y}(\tau) \\
&\quad + R\cdot\mathbf{o}^c - \alpha^d\gamma R\cdot\mathbf{o}^d) \\
&= -(\mathbf{y}(\tau) - \alpha^d\gamma R\cdot\mathbf{o}^d)\cdot H^{-1}\cdot(\mathbf{y}(\tau) - \alpha^d\gamma R\cdot\mathbf{o}^d) \\
&\quad - \mathbf{o}^c\cdot R\cdot H^{-1}\cdot R\cdot\mathbf{o}^c \\
&\quad - 2(\mathbf{y}(\tau) - \alpha^d\gamma R\cdot\mathbf{o}^d)\cdot H^{-1}\cdot R\cdot\mathbf{o}^c
\end{aligned} \tag{18}
$$

and so

$$
\log\frac{p(\mathbf{y}(\tau)|\alpha^c = +1)}{p(\mathbf{y}(\tau)|\alpha^c = -1)} = \log p(\mathbf{y}(\tau)|\alpha^c = +1)
$$
$$
- \log p(\mathbf{y}(\tau)|\alpha^c = -1) \tag{19}
$$
$$
= 2(\mathbf{y}(\tau) - \alpha^d\gamma R\cdot\mathbf{o}^d)\cdot H^{-1}\cdot R\cdot\mathbf{o}^c. \tag{20}
$$

Since each time point is independent, $\sum_{\tau=1}^{T}\mathbf{y}(\tau) = T\bar{\mathbf{y}}(T)$, and Eq. (20) is linear in $\mathbf{y}(\tau)$,

$$
\log\frac{p(\mathcal{Y}(T)|\alpha^c = +1)}{p(\mathcal{Y}(T)|\alpha^c = -1)} = 2T(\bar{\mathbf{y}}(T) - \alpha^d\gamma R\cdot\mathbf{o}^d)\cdot H^{-1}\cdot R\cdot\mathbf{o}^c. \tag{21}
$$

Now, by Bayes rule

$$
\log\frac{P(\alpha^c = +1|\mathcal{Y}(T))}{P(\alpha^c = -1|\mathcal{Y}(T))} = \log\frac{p(\mathcal{Y}(T)|\alpha^c = +1)}{p(\mathcal{Y}(T)|\alpha^c = -1)} + \log\frac{P(\alpha^c = +1)}{P(\alpha^c = -1)}, \tag{22}
$$

which, since $P(\alpha^c = \pm 1) = 0.5$,

$$
= 2T(\bar{\mathbf{y}}(T) - \alpha^d\gamma R\mathbf{o}^d)\cdot H^{-1}\cdot R\cdot\mathbf{o}^c, \tag{23}
$$

which is just Eq. (5).

B.2. Eq. (6)

Eq. (6) is more complicated than Eq. (5) because of the need to marginalize out the unknown value of α^d. Formally:

$$
\begin{aligned}
p(\mathbf{y}(\tau)|\alpha^c = +1) &= p(\mathbf{y}(\tau), \alpha^d = +1|\alpha^c = +1) \\
&\quad + p(\mathbf{y}(\tau), \alpha^d = -1|\alpha^c = +1)
\end{aligned} \tag{24}
$$
$$
\begin{aligned}
&= 0.5(p(\mathbf{y}(\tau)|\alpha^c = +1, \alpha^d = +1) \\
&\quad + p(\mathbf{y}(\tau)|\alpha^c = +1, \alpha^d = -1))
\end{aligned} \tag{25}
$$

since $P(\alpha^d = \pm1|\alpha^c = 1) = 0.5$. Now, writing $\mathbf{a} = R\cdot\mathbf{o}^c$; $\mathbf{b} = \gamma R\cdot\mathbf{o}^d$

$$
p(\mathbf{y}(\tau)|\alpha^c = +1)
$$
$$
\propto e^{-(\mathbf{y}(\tau)-\mathbf{a}-\mathbf{b})\cdot H^{-1}\cdot(\mathbf{y}(\tau)-\mathbf{a}-\mathbf{b})/2} + e^{-(\mathbf{y}(\tau)-\mathbf{a}+\mathbf{b})\cdot H^{-1}\cdot(\mathbf{y}(\tau)-\mathbf{a}+\mathbf{b})/2} \tag{26}
$$

and, similarly,

$$
p(\mathbf{y}(\tau)|\alpha^c = -1)
$$
$$
\propto e^{-(\mathbf{y}(\tau)+\mathbf{a}-\mathbf{b})\cdot H^{-1}\cdot(\mathbf{y}(\tau)+\mathbf{a}-\mathbf{b})/2} + e^{-(\mathbf{y}(\tau)+\mathbf{a}+\mathbf{b})\cdot H^{-1}\cdot(\mathbf{y}(\tau)+\mathbf{a}+\mathbf{b})/2} \tag{27}
$$

collecting and cancelling terms, this implies that

$$
\log\frac{p(\mathbf{y}(\tau)|\alpha^c = +1)}{p(\mathbf{y}(\tau)|\alpha^c = -1)} = 2\mathbf{y}(\tau)\cdot H^{-1}\cdot\mathbf{a} + \log\frac{e^{(\mathbf{y}(\tau)-\mathbf{a})\cdot H^{-1}\cdot\mathbf{b}} + e^{-(\mathbf{y}(\tau)-\mathbf{a})\cdot H^{-1}\cdot\mathbf{b}}}{e^{(\mathbf{y}(\tau)+\mathbf{a})\cdot H^{-1}\cdot\mathbf{b}} + e^{-(\mathbf{y}(\tau)+\mathbf{a})\cdot H^{-1}\cdot\mathbf{b}}} \tag{28}
$$
$$
= 2\mathbf{y}(\tau)\cdot H^{-1}\cdot\mathbf{a} + \log\frac{\cosh\left[(\mathbf{y}(\tau)-\mathbf{a})\cdot H^{-1}\cdot\mathbf{b}\right]}{\cosh\left[(\mathbf{y}(\tau)+\mathbf{a})\cdot H^{-1}\cdot\mathbf{b}\right]} \tag{29}
$$

since, again, the samples are independent

$$
\log\frac{p(\mathcal{Y}(T)|\alpha^c = +1)}{p(\mathcal{Y}(T)|\alpha^c = -1)} = 2T\bar{\mathbf{y}}\cdot H^{-1}\cdot\mathbf{a}
$$
$$
+ \log\frac{\cosh\left[T(\bar{\mathbf{y}}(T) - \mathbf{a})\cdot H^{-1}\cdot\mathbf{b}\right]}{\cosh\left[T(\bar{\mathbf{y}}(T) + \mathbf{a})\cdot H^{-1}\cdot\mathbf{b}\right]} \tag{30}
$$

and, again by Bayes rule, and filling in the values of $\mathbf{a}$ and $\mathbf{b}$, we get

$$
\log\frac{P(\alpha^c = +1|\mathcal{Y}(T))}{P(\alpha^c = -1|\mathcal{Y}(T))} = 2T\bar{\mathbf{y}}(\tau)\cdot H^{-1}\cdot R\cdot\mathbf{o}^c
$$
$$
+ \log\frac{\cosh\left[\gamma T(\bar{\mathbf{y}}(T) - R\cdot\mathbf{o}^c)\cdot H^{-1}\cdot R\cdot\mathbf{o}^d\right]}{\cosh\left[\gamma T(\bar{\mathbf{y}}(T) + R\cdot\mathbf{o}^c)\cdot H^{-1}\cdot R\cdot\mathbf{o}^d\right]} \tag{31}
$$

which is Eq. (6).

For the approximate analytical calculations, we made the empirical approximation based on averaging over a large number of moderate-sized values of ξ:

$$
\log\cosh(\xi) \simeq |\xi| - \log(2) + 0.7175e^{-1.7134|\xi|}
$$

This is within 0.025 of the true value.

References

Anderson, R. S., & Thibos, L. N. (1999a). Relationship between acuity for gratings and for tumbling-E letters in peripheral vision. *Journal of the Optical Society of America A – Optics Image Science and Vision, 16*(10), 2321–2333.

Anderson, R. S., & Thibos, L. N. (1999b). Sampling limits and critical bandwidth for letter discrimination in peripheral vision. *Journal of the Optical Society of America A - Optics Image Science and Vision, 16*(10), 2334–2342.

Andriessen, J. J., & Bouma, H. (1976). Eccentric vision: Adverse interactions between line segments. *Vision Research, 16*(1), 71–78.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature, 226*(5241), 177–178.

Bouma, H. (1973). Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Research, 13*(4), 767–782.

Boynton, G. M. (2009). A framework for describing the effects of attention on visual responses. *Vision Research, 49*(10), 1129–1143.

Carandini, M., & Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science, 264*(5163), 1333–1336.

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience, 17*(21), 8621–8644.

Chastain, G. (1982). Confusability and interference between members of parafoveal letter pairs. *Perception & Psychophysics, 32*(6), 576–580.

Chastain, G. (1988). Does order of letter analysis contribute to the parafoveal identification asymmetry? *The Journal of General Psychology, 115*(4), 397–402.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10*(7), 287–291.

Cowey, A., & Rolls, E. T. (1974). Human cortical magnification factor and its relation to visual acuity. *Experimental Brain Research, 21*(5), 447–454.

Daniel, P. M., & Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *The Journal of Physiology, 159*, 203–221.

Danilova, M. V., & Bondarko, V. M. (2007). Foveal contour interactions and crowding effects at the resolution limit of the visual system. *Journal of Vision, 7*(2), 1–18.

Dayan, P. (2009). Load and attentional Bayes. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.). *Advances in neural information processing systems* (Vol. 21, pp. 369–376). NIPS.

Dayan, P., & Zemel, R. (1999). Statistical models and sensory attention. *ICANN* (pp. 1017–1022). IEEE.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.

Dow, B. M., Snyder, A. Z., Vautin, R. G., & Bauer, R. (1981). Magnification factor and receptive field size in foveal striate cortex of the monkey. *Experimental Brain Research, 44*(2), 213–228.

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology, 92*(Pt. 1), 53–78.

Duncan, R. O., & Boynton, G. M. (2003). Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron, 38*(4), 659–671.

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

Eriksen, B., & Eriksen, C. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*(1), 143–149.

Felisberti, F., Solomon, J., & Morgan, M. (2005). The role of target salience in crowding. *Perception, 34*, 823–833.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*(4), 193–202.

Gattass, R., Sousa, A. P., & Rosa, M. G. (1987). Visual topography of V1 in the Cebus monkey. *Journal of Comparative Neurology, 259*(4), 529–548.

Gheri, C., Morgan, M. J., & Solomon, J. A. (2007). The relationship between search efficiency and crowding. *Perception, 36*(12), 1779–1787.

Gratton, G., Coles, M. G., Sirevaag, E. J., Eriksen, C. W., & Donchin, E. (1988). Pre- and poststimulus activation of response channels: A psychophysiological analysis. *Journal of Experimental Psychology: Human Perception and Performance, 14*(3), 331–344.

Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Grossberg, S. (2001). Linking the laminar circuits of visual cortex to visual perception: Development, grouping, and attention. *Neuroscience & Biobehavioral Reviews, 25*(6), 513–526.

Hess, R. F., Dakin, S. C., & Kapoor, N. (2000). The foveal 'crowding' effect: Physics or physiology? *Vision Research, 40*(4), 365–370.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10-12), 1489–1506.

Johnston, A. (1989). The geometry of the topographic map in striate cortex. *Vision Research, 29*(11), 1493–1500.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*(4), 219–227.

Kooi, F. L., Toet, A., Tripathy, S. P., & Levi, D. M. (1994). The effect of similarity and duration on spatial interaction in peripheral vision. *Spatial Vision, 8*(2), 255–279.

Korte, W. (1923). Uber die gestaltauffassung im indirekten sehen. *Zeitschrift für Psychologie, 93*, 17–82.

Krumhansl, C. L. (1977). Naming and locating simultaneously and sequentially presented letters. *Perception & Psychophysics, 22*, 293–302.

Krumhansl, C. L., & Thomas, E. A. C. (1977). Effect of level of confusability on reporting letters from briefly presented visual displays. *Perception & Psychophysics, 21*, 269–279.

Lampl, I., Ferster, D., Poggio, T., & Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex. *Journal of Neurophysiology, 92*(5), 2704–2713.

Lavie, N. (2005). Distracted and confused? Selective attention under load. *Trends in Cognitive Sciences, 9*(2), 75–82.

Lavie, N., & de Fockert, J. W. (2003). Contrasting effects of sensory limits and capacity limits in visual selective attention. *Perception & Psychophysics, 65*(2), 202–212.

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology – General, 133*(3), 339–354.

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics, 56*(2), 183–197.

Levi, D. M. (2008). Crowding – An essential bottleneck for object recognition: A mini-review. *Vision Research, 48*(5), 635–654.

Levi, D. M., Hariharan, S., & Klein, S. A. (2002a). Suppressive and facilitatory spatial interactions in amblyopic vision. *Vision Research, 42*(11), 1379–1394.

Levi, D. M., Klein, S. A., & Aitsebaomo, A. P. (1985). Vernier acuity, crowding and cortical magnification. *Vision Research, 25*(7), 963–977.

Levi, D. M., Klein, S. A., & Hariharan, S. (2002b). Suppressive and facilitatory spatial interactions in foveal vision: Foveal crowding is simple contrast masking. *Journal of Vision, 2*(2), 140–166.

Li, Z. (2001). Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex. *Neural Computation, 13*(8), 1749–1780.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*(1), 9–16.

Liu, Y. S., Yu, A., & Holmes, P. (2009). Dynamical analysis of Bayesian inference models for the Eriksen task. *Neural Computation, 21*(6), 1520–1553.

Livne, T., & Sagi, D. (2007). Configuration influence on crowding. *Journal of Vision, 7*(2), 1–12.

Louie, E. G., Bressler, D. W., & Whitney, D. (2007). Holistic crowding: Selective interference between configural representations of faces in crowded scenes. *Journal of Vision, 7*(2), 1–11.

Lu, Z. L., & Dosher, B. A. (1998). External noise distinguishes attention mechanisms. *Vision Research, 38*(9), 1183–1198.

Miller, E. K., Gochin, P. M., & Gross, C. G. (1993). Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus. *Brain Research, 616*(1–2), 25–29.

Motter, B. C. (2009). Central V4 receptive fields are scaled by the V1 cortical magnification and correspond to a constant-sized sampling of the V1 surface. *Journal of Neuroscience, 29*(18), 5749–5757.

Motter, B. C., & Simoni, D. A. (2007). The roles of cortical image separation and size in active visual search performance. *Journal of Vision, 7*(2), 1–15.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45*(2), 205–231.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, England: Oxford University Press.

Paradiso, M. A. (1988). A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biological Cybernetics, 58*(1), 35–49.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739–744.

Pashler, H. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.

Pelli, D., Cavanagh, P., Desimone, R., Tjan, B., & Treisman, A. (2007a). Crowding: Including illusory conjunctions, surround suppression, and attention. *Journal of Vision, 7*(2), 1.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision, 4*(12), 1136–1169.

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience, 11*(10), 1129–1135.

Pelli, D. G., Tillman, K. A., Freeman, J., Su, M., Berger, T. D., & Majaj, N. J. (2007b). Crowding and eccentricity determine reading rate. *Journal of Vision, 7*(2), 20.1–20.36.

Petrov, Y., & Popple, A. V. (2007). Crowding is directed to the fovea and preserves only feature contrast. *Journal of Vision, 7*(2), 8.1–8.9.

Petrov, Y., Popple, A. V., & McKee, S. P. (2007). Crowding and surround suppression: Not to be confused. *Journal of Vision, 7*(2), 12.1–12.9.

Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience, 19*(5), 1736–1753.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019–1025.

Rolls, E., & Deco, G. (2002). *Computational neuroscience of vision*. Oxford, England: Oxford University Press.

Rolls, E. T., & Cowey, A. (1970). Topography of the retina and striate cortex and its relationship to visual acuity in rhesus monkeys and squirrel monkeys. *Experimental Brain Research, 10*(3), 298–310.

Rolls, E. T., & Tovee, M. J. (1995). The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. *Experimental Brain Research, 103*(3), 409–420.

Schwartz, E. L. (1980). Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding. *Vision Research, 20*(8), 645–669.

Schwartz, O., Sejnowski, T., & Dayan, P. (2009). Perceptual organization in the tilt illusion. *Journal of Vision, 9*, 1–20.

Schwartz, O., Sejnowski, T. J., & Dayan, P. (2006). Soft mixer assignment in a hierarchical generative model of natural scene statistics. *Neural Computation, 18*(11), 2680–2718.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience, 4*(8), 819–825.

Shaw, P. (1969). Processing of tachistoscopic displays with controlled order of characters and spaces. *Perception & Psychophysics, 6*, 257–266.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences, 27*(3), 161–168.

Strasburger, H., Harvey, L. O., & Rentschler, I. (1991). Contrast thresholds for identification of numeric characters in direct and eccentric view. *Perception & Psychophysics, 49*(6), 495–508.

Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research, 32*(7), 1349–1357.

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology, 12*, 242–248.

Treisman, A. M. (1969). Strategies and models of selective attention. *Psychological Review, 76*(3), 282–299.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.

Tsotsos, J. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences, 13*(3), 423–469.

van den Berg, R., Roerdink, J. B. T. M., & Cornelissen, F. W. (2007). On the generality of crowding: Visual crowding in size, saturation, and hue compared to orientation. *Journal of Vision, 7*(2), 1–11.

van den Berg, R., Roerdink, J. B. T. M., & Cornelissen, F. W. (2010). A neurophysiologically plausible population code model for feature integration explains visual crowding. *PLoS Computational Biology, 6*(1), e1000646.

Van Essen, D. C., Newsome, W. T., & Maunsell, J. H. (1984). The visual field representation in striate cortex of the macaque monkey: Asymmetries, anisotropies, and individual variability. *Vision Research, 24*(5), 429–448.

Virsu, V., & Rovamo, J. (1979). Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research, 37*(3), 475–494.

Wainwright, M., Simoncelli, E., & Willsky, A. (2001). Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Applied and Computational Harmonic Analysis, 11*(1), 89–123.

Whiteley, L. (2008). *Uncertainty, reward and attention in the Bayesian brain*. PhD thesis, Gatsby Computational Neuroscience Unit, UCL.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 419–433.

Wolford, G. (1975). Perturbation model for letter identification. *Psychological Review, 82*(3), 184–199.

Yen, S. C., & Finkel, L. H. (1998). Extraction of perceptually salient contours by striate cortical networks. *Vision Research, 38*(5), 719–741.

Yu, A. J., Dayan, P., & Cohen, J. D. (2009). Dynamics of attentional selection under conflict: Toward a rational Bayesian account. *Journal of Experimental Psychology: Human Perception and Performance, 35*(3), 700–717.

Zhaoping, L. (2006). Theoretical understanding of the early visual processes by data compression and data selection. *Network, 17*(4), 301–334.

Zoccolan, D., Cox, D. D., & DiCarlo, J. J. (2005). Multiple object response normalization in monkey inferotemporal cortex. *Journal of Neuroscience, 25*(36), 8150–8164.