In *ICANN 1999.*

Statistical Models and Sensory Attention

Peter Dayan
GCNU
17 Queen Square
London WC1N 3AR
dayan@gatsby.ucl.edu

Richard S Zemel
Department of Psychology
University of Arizona
Tucson, AZ 85721
zemel@u.arizona.edu

Abstract

Physiological investigations into the neural basis of sensory attention have led to puzzling and contradictory results. Attention can seemingly lead to increased, decreased and unchanged neural activities, according to features of attentional experiments that are not well understood. We take one particular case in which activities increase as a result of attention, model its possible statistical underpinning, and relate our model to other attentional suggestions. Increased activities in population codes are associated with increased certainty about the encoded quantities. This increased certainty has to come from somewhere – in our model it emerges from particular changes in the model's processing strategy.

1 Introduction

Although hard to define very precisely, *attention* has been a highly seductive target for experiments and models alike. The core idea, that some stimuli or stimulus features are treated differently from others (in particular, more effectively) for the purposes of representation, processing and/or learning, is common to a whole set of observable effects, a set whose neural basis may well be far from unitary. The range of experiments include cases for which features and/or locations are pitted either against nothing (*ie* background noise) or against one another.

One common starting point for understanding attentional effects is to build models in which competition is a natural consequence of the structure of a computational task. A good example comes from the field of animal conditioning, in which multiple conditioned stimuli (CSs) such as lights and tones have to be used to predict unconditioned stimuli (USs) such as rewards and punishments (see Dickinson, 1980; Mackintosh, 1983). There are rich interactions in the ways that collections of CSs learn, that have been modeled in terms of attentional competition amongst the CSs (Mackintosh, 1975; Pearce & Hall, 1980; Grossberg, 1988). Arguing from a statistical perspective, Dayan & Long (1998) separated two underlying components of attentional competition, one governing *representation,* based on how *reliable* a CS is at predicting a US, and one governing *learning,* based on how *uncertain* the quantitative prediction made of a US by a (reliable or unreliable) CS (Sutton, 1992).

In both these cases, different forms of attention, which would likely have completely different neural implementations, arise as statistically rational solutions to particular problems in learning. Our long term aim is to provide similarly rational accounts for aspects of sensory attention, albeit based on quite different statistical premises.

2 Attention in the ventral stream

Attention is multi-faceted, and an incredible wealth of quite different experiments, designed to test quite different things, bear on it. For instance, since attentive vision is partly defined by opposition to preattentive vision, the many experiments that test things like surround suppression and contour enhancement for texture segmentation, pop-out and contour integration (see

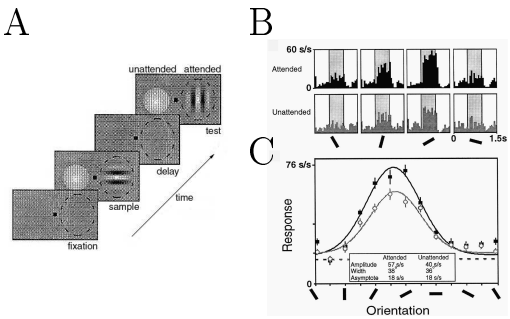

Figure 1: Attentional effects in V4. A) Delayed match to sample paradigm. The animal is instructed to pay attention either to orientation or colour (shown here as grey) in blocks of trial. Within a single trial, it has to respond if the attended feature in the test pattern is the same as that in the sample, irrespective of the unattended stimulus. The dashed line shows an example receptive field. B) Responses of a single V4 cell to different orientations in the attended (upper) and unattended (lower) conditions. C) Attended (filled squares) and unattended (open circles) tuning functions for the cell. Black (attended) and gray (unattended) dashed lines at the bottom show the undriven activity of the cell. From McAdams & Maunsell (1999).

Li, 1998 for models and references to the data) are of obvious relevance. We focus on neurophysiological findings (*eg* Moran & Desimone, 1985; Spitzer, Desimone & Moran, 1988; Motter, 1993; 1994a;b; Desimone & Duncan, 1995; Connor, Gallant, Preddie & Van Essen, 1996; Connor, Preddie, Gallant & Van Essen, 1997; Motter, 1998) in particular a core result involving a very simple visual display in which there is competition (McAdams & Maunsell, 1999).

Figure 1A, from McAdams & Maunsell (1999) shows a basic delayed match to sample paradigm. Here, two stimuli (one coloured, one oriented) are shown in circumstances under which the animal has been instructed to pay attention for a whole block of trials to just one or the other. In this case, the animal is supposed to remember the orientation of the oriented cue in the sample phase and act at the test according to whether the orientation is the same or different, ignoring any similarities or changes in the colour. The dashed line shows schematically the receptive field of the neuron being tested – it comfortably contains just one of the two stimuli. The location of the oriented stimulus during the test is always the same as that of the oriented stimulus during the sample.

Figure 1B shows the effect of the prior instruction on the activity of a single, orientation selective, neuron in V4. In the so-called attended condition, when the monkey is forced to attend to orientation, the neuron responds more strongly than in the unattended condition, when the monkey attends to colour (although see Moran & Desimone, 1985). Although the neuron is more active in the face of attention, its tuning curve has essentially the same width in both conditions, and the baseline activity also does not change. 55% of the 223 orientation-tuned cells recorded were modulated by attention, and, for most of them, activities in the attended condition were higher than in the unattended condition. Only about 9% of the cells exhibited significantly different tuning widths in the two conditions, and of these some were broader and some narrower. McAdams & Maunsell (1999) show that something similar is true for cells in V1, although the magnitude of the effect is much smaller and a smaller proportion of the cells (31%) are modulated by attention.

We later discuss various possible interpretations of this change in the activity. For the present, consider one consequence that emerges naturally from considering the activity of the cells as part of a population code for orientation that is providing the basic information required for the monkey's inferences (Paradiso, 1988; Seung & Sompolinsky, 1993; Snippe, 1996). Many population coding models turn the activity of the cells into a *posterior probability* distribution over the variable they code (in this case, orientation). The tightness of the posterior distribution is a measure of the quality of the coding scheme, and is well quantified by the Fisher information the code provides about the variable. If neuron i has tuning curve $f_i(\theta)$ for orientation θ and is corrupted by Poisson noise, then the Fisher information for θ is proportional to $\mathcal{F}(\theta) \propto \sum_i \frac{(f_i'(\theta))^2}{f_i(\theta)}$ where the individual components of the sum take the form of signal/noise ratios. Amplifying the tuning curves $f_i(\theta)$, thereby multiplicatively increasing the firing rates, pro-

portionally increases the Fisher information (at least ignoring the baseline). Something similar is true for some population coding models that consider the activity as determining a full probability distribution over θ (Anderson, 1994; Zemel, Dayan & Pouget, 1998; Hinton & Ghahramani, personal communication); for many of them, the stronger the firing, the tighter the posterior distribution over θ, and so the increased certainty.

The key point of this paper is that, from a statistical perspective, the increased certainty has to originate somewhere. In the end, we would like an algorithmic account of its provenance and its instantiation in the amplification of tuning curves; however, in this paper we merely seek a computational account. One suggestion is that the neurons always have available all the information needed to report on θ accurately, but that the metabolic cost of doing so by spiking fast is very high. In this case, attention can be seen as changing the *loss function,* weighing accuracy about a given variable more heavily and the energetic cost less heavily, and so making higher activity levels appropriate.

A (not necessarily mutually exclusive) alternative is that, through the medium of attention, the information on which neuron i is basing its activity has changed. The key question is why this information might change in a way that licenses more certain inference. In the next section we present an example, based on the rather standard (*eg* Pelli, 1985; Graham, 1989; Palmer, 1984), though not wholly complete (Downing, 1988; Kontsevick & Tyler, 1999) psychophysical idea of a model of positional uncertainty.

3 Attentional control over receptive fields

Figure 2A shows the structure of a simple inference problem. The brightness of each little patch shows the activity r of a single input unit as a function of position (x) and angle (θ). The activities are caused by a characteristic noisy, Gaussian-shaped input which can be centred *anywhere* along the x and θ coordinates (actually using circular boundary conditions in x to avoid edge effects). The inferential task for the units shown in the circle on top is to represent the angle of the characteristic stimulus, marginalising away the effect of positional uncertainty. We consider the V4 orientation selective cell whose responses are shown in figure 1B;C as an example of one of the cells in the circle reporting on the net angle of the stimulus, and each V1 cell within its receptive field corresponds to an input unit. Furthermore, in keeping with distributional interpretations of population codes, the V4 units must represent the *posterior uncertainty* about this angle given all the input $\mathbf{r}$. That is, their activity must represent the distribution

$$\mathcal{P}[\theta|\mathbf{r}] = \int_x \mathcal{P}[\theta, x|\mathbf{r}]dx \propto \int_x \mathcal{P}[\mathbf{r}|\theta, x]dx \tag{1}$$

assuming a flat prior over x and θ. Figure 2B shows an example of the true marginal posterior distribution over θ given the entire input – in this case it has a unimodal shape centered roughly on the true value of the angle ($\theta = 0$). This is designed to be one of the simplest possible cases in which a set of units is intended to have pattern selectivity (for Gaussian bump stimuli) whilst marginalising across input position.

As described in the previous section, the overall magnitude of the activities of the units in a population code is a way of encoding the overall uncertainty of the values that they are coding. Rather than build an explicit encoding model which performs the mapping in figure 2A from the input to the model V4 neurons, we study some particular general properties the encoding model must possess if it is to perform correct inference on its inputs. That is, we focus on the posterior distribution that can be inferred from the input rates $\mathbf{r}$, and assume that the firing rates in the model V4 neurons will reflect that distribution. At present, the model ignores the (small) changes in the activities of V1 neurons.

Since θ is an angle variable, a computationally convenient way to characterise its posterior distribution is to find the parameters of the best fitting circular normal distribution, whose distribution function is $f(\theta) = \exp(k\cos(\theta - \langle\theta\rangle))/2\pi I_0(k)$, where $\langle\theta\rangle$ is the angular mean, k is a form of inverse variance, and $I_0(k)$ is the modified Bessel func-

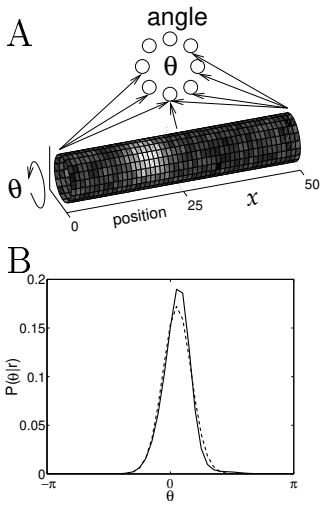

Figure 2: Simple attentional model. A) Example characteristic activity $\mathbf{r}$ across a set of input units arranged according to preferred position x and angle θ. The input is derived from a Gaussian bump on top of which noise is added. Circular boundary conditions for x are actually assumed in order to eliminate edge effects. The task for the network is to determine the posterior distribution over the angle based on the input and represent this posterior distribution in the activities of a population code. The population coding units (upper circle) are not actually simulated – we just consider the properties they would have to have. B) Posterior distribution $\mathcal{P}[\theta|\mathbf{r}]$ (solid line) over the original angle of presentation θ after marginalising away the unknown position x. In the simulations, the noise was an order of magnitude greater than in (A) and the variance in x was smaller in order for the posterior distributions to be non-trivial. The dashed line shows the circular normal distribution that best fits this activity.

tion of the first kind and order 0. The dashed line in figure 2B shows the circular normal distribution fit to this activity.

Figure 3A;B show histograms of the values of $\langle\theta\rangle$ and k for the case that the true value of $\theta = 0$ for the stimulus and for a large number of different drawings of the noise. The posterior means are correctly clustered around the value $\langle\theta\rangle = 0$, and the inverse variances are quite small, indicating that the posterior distributions are broad.

We implement a spatial form of attention in a very simple way, by restricting the positions x which are relevant for the purposes of inferring the angle presented. In this case, the added noise in the areas that are filtered out is irrelevant, and so the posterior distribution over the angle θ of the stimulus will likely be tighter. Figure 3C shows a histogram of the inverse variances of the posterior distribution in the case that the window of integration for x in equation 1 is 20% of the value used to generate figure 3B. The inverse variances are significantly greater, corresponding to the increased accuracy with which θ can be determined. Figure 3D shows how the median inverse variance (including the median values from the plots in figures 3B;C) varies with the size of the attentional window. By the arguments above, this increase in *accuracy* should correctly be paralleled by an increase in the *activities* of the units forming the population code, thereby statistically legitimising the effect shown by McAdams & Maunsell (1999).

4 Discussion

We have constructed a very simple model of a particular neurophysiological attentional effect. In the model, the enhanced activity of the cells in a population code is treated as reporting with greater accuracy the underlying variable they encode. This greater accuracy is made legitimate on account of a neural processing strategy in which the unattended part of the input to the population code is eliminated from consideration. The statistical price to pay is, of course, that if the stimulus is actually presented in one of the unattended locations, then its orientation will not be correctly determined. In a less strict model, processing in unattended areas might merely be attenuated, essentially implementing a softer prior as to where the stimulus might be found. Note that McAdams & Maunsell's (1999) paradigm does not allow us to compare the response of the cell in the total absence of any competing stimulus.

The essential underpinnings of the model are not new (see, for example, Pelli, 1985; Downing, 1988; van der Heijden, 1996), and its general philosophy is based on the notion (Allport, 1989) that ways of viewing atten-

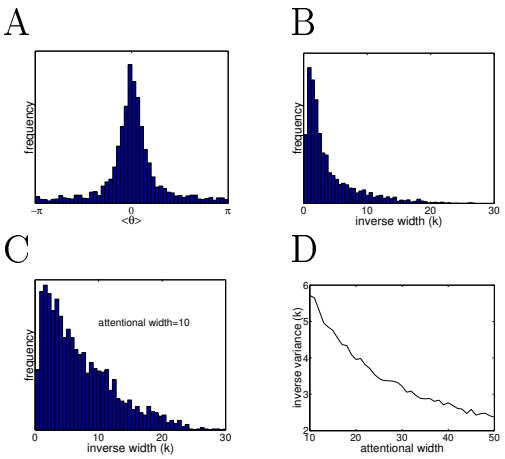

Figure 3: Histograms of the means ($\langle\theta\rangle$; A) and inverse variances (k ; B;C) of the circular normal distributions fit to the posterior distributions $\mathcal{P}[\theta|\mathbf{r}]$. (A;B) are without focal attention; (C) shows the effect of reducing the window of attention to 10 units, 20% of the width of the full input shown in figure 2. D) The median of the inverse variances of the posterior distributions as a function of the width of the attention.

tion in terms of distinctions such as early and late selection are unhelpful – it bites only where there is a deleterious or beneficial interaction between the tuning function properties of cells (including the location of their receptive fields and the features to which they are sensitive) and the locational and/or featural aspects of the current focus of attention. In our particular case, as with Moran & Desimone (1985; see also Desimone & Duncan, 1995), receptive fields are the critical processing resource that need to be managed spatially. However, this is only one very simple case, and in experiments in which attention to a particular *feature* and not to locations is important (*eg* Motter, 1994a), one might expect different attentional processing strategies to be employed in which the elements from which the feature is computed are comparatively favoured, leading once again to greater statistical accuracy. Note that the same simple strategy—restricting the range of a dimension along which a feature may be inferred—can lead to increased certainty for features just as for positions. Featural attention, and the consequences of attention in hierarchical systems with cells with multiple selectivities are the most pressing areas for future work.

As always, it is important to distinguish the computational basis of the model from any mechanistic substrate. That increased firing rates are statistically rational says nothing as to how they might arise – in particular, it is silent as to whether input cells from locations of the input that are deemed unattended should also be suppressed as a whole, or whether input cells from attended locations should be boosted strongly so that they overcome input from unattended locations, or whether only particular outputs of these input cells should be suppressed or boosted appropriately. Neurophysiological data bear on these questions – for instance, simple forms of models in which the activities of input cells are suppressed or boosted as a whole are contradicted by Moran & Desimone's (1985) results on the circumstances under which the activities of cells in V4 whose receptive fields cover or do not cover the locus of attention are unaffected by attention. However, the various results are based on different paradigms, and can be hard to reconcile with each other.

The cases for which it is hardest to account using our model come when the consequence of attending to the location of the receptive field of orientation selective cells is actually to *reduce* the activity of those cells (Motter, 1993). One possibility is that this result reflects a competitive interaction between spatial and featural aspects of attention in which the fact that those cells are tuned to other features (such as colour) which are irrelevant for the task results in their being suppressed. Motter (1993) found some weak evidence for this in that the population of cells in V2 that showed these reductions were only weakly tuned to orientation.

The main competing computational model to explain changes in activity in the face of attention suggests that it represents a basis function strategy for performing *normalisation,* (*ie* presenting images of translated and rotated objects in a canonical frame (Olshausen, Anderson & Van Essen, 1993; Salinas & Abbott, 1997; Riesenhuber & Dayan, 1997). This interpretation captures the multiplicative modulation seen in the sensitivity of visually responsive parietal cells to the position of the eyes (see Pouget & Sejnowski, 1997), only with some form of attentional focus (possibly with both spatial and featural dimensions) taking the place of

eye position (and also head position, hand position, *etc*), and has been suggested on the basis of various experiments (*eg* Conor *et al*, 1996; 1997; McAdams & Maunsell, 1999). Although these ideas are quite compelling for normalisation, they are hard to relate to some aspects of the neural data, such as the effects of changing the number and nature of competing stimuli.

Acknowledgements

We are very grateful to Jochen Braun, Zhaoping Li and Alex Pouget for discussion and comments. Funding is from the Gatsby Charitable Foundation (PD); ONR Young Investigator Award N00014-98-1-0509 (RZ).

Bibliography

Allport, A (1989). Visual attention. In MI Posner, editor *Foundations of Cognitive Science.* Cambridge, MA: MIT Press, 631-682.

Anderson, CH (1994). Basic elements of biological computational systems. *International Journal of Modern Physics C* **5**:135-137.

Connor, CE, Gallant, JL, Preddie, DC & Van Essen, DC (1996). Responses in area V4 depend on the spatial relationship between stimulus and attention. *Journal of Neurophysiology* **75**:1306-1308.

Connor, CE, Preddie, DC, Gallant, JL & Van Essen, DC (1997). Spatial attention effects in macaque area V4. *Journal of Neuroscience* **17**:3201-14.

Dayan, P & Long, T (1998). Statistical models of conditioning. In MI Jordan, M Kearns & SA Solla, editors, *Advances in Neural Information Processing Systems, 10.* Cambridge, MA: MIT Press, 117-123.

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

Dickinson, A (1980) *Contemporary Animal Learning Theory.* Cambridge: Cambridge University Press.

Downing, CJ (1988). Expectancy and visual-spatial attention: Effects on perceptual quality. *Journal of Experimental Psychology: Human Perception & Performance* **14**:188-202.

Graham, NVS (1989). *Visual pattern analyzers.* New York: Oxford University Press.

Grossberg, S, editor (1988) *Neural Networks and Natural Intelligence.* Cambridge, MA:MIT Press.

van der Heijden, AHC (1996). Selective attention as a computational function. In AF Kramer, MGH Coles & GD Logan, editors, *Converging Operations in the Study of Visual Selective Attention.* Washington, DC: APA, 459-482.

Jacobs, RA, Jordan, MI & Barto, AG (1991). Task decomposition through competition in a modular connectionist architecture: the what and where vision tasks. *Cognitive Science* **15**:219-250.

Kontsevick, LL & Tyler, CW (1999). Distraction of attention and the slope of the psychometric functions. *Journal of the Optical Society of America. A.* **16**:217-222.

McAdams, CJ, Maunsell, JHR (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience* **19**:431-441.

Mackintosh, NJ (1975) A theory of attention: Variations of the associability of stimuli with reinforcement. *Psychological Review* **82**:276-298.

Moran, J & Desimone, R (1985). Selective attention gates visual processing in the extrastriate cortex. *Science* **229**:782-784.

Motter, BC (1998). Neurophysiology of visual attention. In R Parasuraman, editor, *The Attentive Brain.* Cambridge, MA: MIT Press, 51-69.

Motter, BC (1994a). Neural correlates of attentive selection for color or luminance in extrastriate area V4. *Journal of Neuroscience* **14**:2178-2189.

Motter, BC (1994b). Neural correlates of feature selective memory and pop-out in extrastriate Area V4. *Journal of Neuroscience* **14**:2190-2199.

Motter, BC (1993) Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of Neurophysiology* **70**:909-919.

Olshausen, BA, Anderson, CH & Van Essen, DC (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience* **13**:4700-4719.

Palmer, J (1994). Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks. *Vision Research,* **34**:1703-1721.

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

Pearce, JM & Hall, G (1980) A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review* **87**:532-552.

Pelli, DG (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America. A.* **2**:1508-1532.

Pouget, A & Sejnowski, TJ (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience* **9**:222-237.

Riesenhuber, M & Dayan, P (1996). Neural models for part-whole hierarchies. In MC Mozer, MI Jordan & T Petsche, editors, *Advances in Neural Information Processing Systems, 9.* Cambridge, MA: MIT Press, 17-23.

Salinas, E & Abbott, LF (1997). Invariant visual responses from attentional gain fields. *Journal of Neurophysiology* **77**:3267-3272.

Seung, HS & Sompolinsky, H (1993) Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences of the United States of America* **90**:10749-10753.

Snippe, HP (1996) Theoretical considerations for the analysis of population coding in motor cortex. *Neural Computation* **8**:29-37.

Spitzer, H, Desimone, R & Moran, J (1988). Increased attention enhances both behavioral and neuronal performance. *Science* **240**:338-340.

Sutton, RS (1992). Adapting bias by gradient descent: an incremental version of delta-bar-delta. *Proceedings Tenth National Conference on Artificial Intelligence.* Menlo Park, CA: AAAI Press, 171-176.

Zemel, RS, Dayan, P & Pouget A (1998). Probabilistic interpretation of population codes. *Neural Computation* **10**:403-430.