Adaptive learning of Gaussian categories leads to decision bounds and response surfaces incompatible with optimal decision making

Michael Kalish
Department of Psychology
Indiana University
Bloomington, IN 47405
mkalish@ucs.indiana.edu

Abstract

Two experiments in category learning are used to examine two types of categorization models. In both a two and four choice experiment, subjects are shown to fail to learn to optimally classify two dimensional stimuli. The general recognition theory (GRT) of Ashby & Maddox (1990) predicts quadratic decision bounds. The first experiment disconfirms this. The extended GRT predicts that learners adopt a bound of complexity equivalent to the optimal one. The second experiment disconfirms this as well. Both experiments support the idea that general resources of adaptive systems can provide explanations of observed sub-optimal behavior.

Introduction

People are readily able to learn new perceptual categories, which is not surprising given the underlying importance of this ability. Determining the value of any of the myriad affordances which make up our niche is a large part of our daily existence. If not in the laboratory, than these categorical decisions are made in the grocery store, or, recreationally, while out birding or mushroom hunting.

The structure of perceptual categories is largely unknown, but the process of learning them is most often conceived as a process of optimization of attention to the relevant perceptual dimensions (Gibson, 1966). A number of theories of category learning which apply to this problem have been presented as mathematical models, and analysis of these (eg., Estes 1989) have shown many to be asymptotically similar, if not identical. In essence, most leading models of categorization are able to account for optimal segregation of multi-dimensional stimuli into two or more categories.

Optimality is, of course, a relative term. In the context of categorization, maximizing the likelihood of a correct response is one objective which immediately comes to mind. Deviations from this goal are common when categories are overlapping and have graded membership (eg., Ashby & Maddox 1990). Different models account for these deviations differently, and this paper reports attempts to distinguish a number of models in two different experimental contexts. In particular, standard back-propagation trained multi-layer perceptrons, radial basis function networks and a hybrid model combining rule and distribution information will be contrasted on data from two and four choice categorization tasks.

Categorization Models

A description of a concrete categorization problem will ease exposition of the models under consideration. Imagine two overlapping distributions lying in a two-dimensional space (figure 1).

Optimal response selection requires sensitivity to the conditional expectation of category A given the stimulus, P(class=A|stimulus=**s**), which for any stimulus s is the probability mass function p(A|s). If the prior probability of class membership, q_A, and the class-conditional densities, $f_A(\mathbf{s})$ and $f_B(\mathbf{s})$, are known then Bayes' Theorem tells us how to compute posterior probabilities:

$$p(A|\mathbf{s}) = q_A f_A(\mathbf{s})/f(\mathbf{s}) \quad (1)$$

Where $\mathbf{s} = (s_X, s_Y)$, the conjunction of features that makes up the stimulus, and the evidence f(**s**) is given by $\sum_u q_u f_u(\mathbf{s})$, where the summation is over both categories. Selecting the class with the maximum likelihood (in this case, choosing A if p(A|s) is > p(B|s)) maximizes the number of correct responses in the long run. This decision boundary corresponding to this deterministic response strategy is shown in figure 1.

The two distributions were chosen so as to encourage subjects to use all the information available about category membership. In particular, optimal categorization required sensitivity to the covariance between the dimensions within each distributions, along with the means and variances of the individual dimensions.

Since this experiment is essentially a replication of Ashby & Maddox (1990), sub-optimal empirical decision boundaries are expected. On the other hand, the gradient of the response surface about the boundary may or may not be optimal. If it is not, then both these types of non-optimality need to be explained.

Noise

One notion of what separates the observed from optimal response surfaces is that there is likely to be **perceptual**, **criterial** (Ashby & Maddox 1993) or **response** (Kalish, 1993) noise. Allow perceptual noise z to be normally distributed with mean $\mu_Z = 0$ and $\Sigma_Z = \mathbf{I}\sigma_Z$, then the perceived stimulus s' is the sum of the presented stimulus $\mathbf{s}_i$ and the noise z. Recall that $\mathbf{s}_i$ (whether i is category A or B) is normal with mean μ_i and covariance Σ_i,

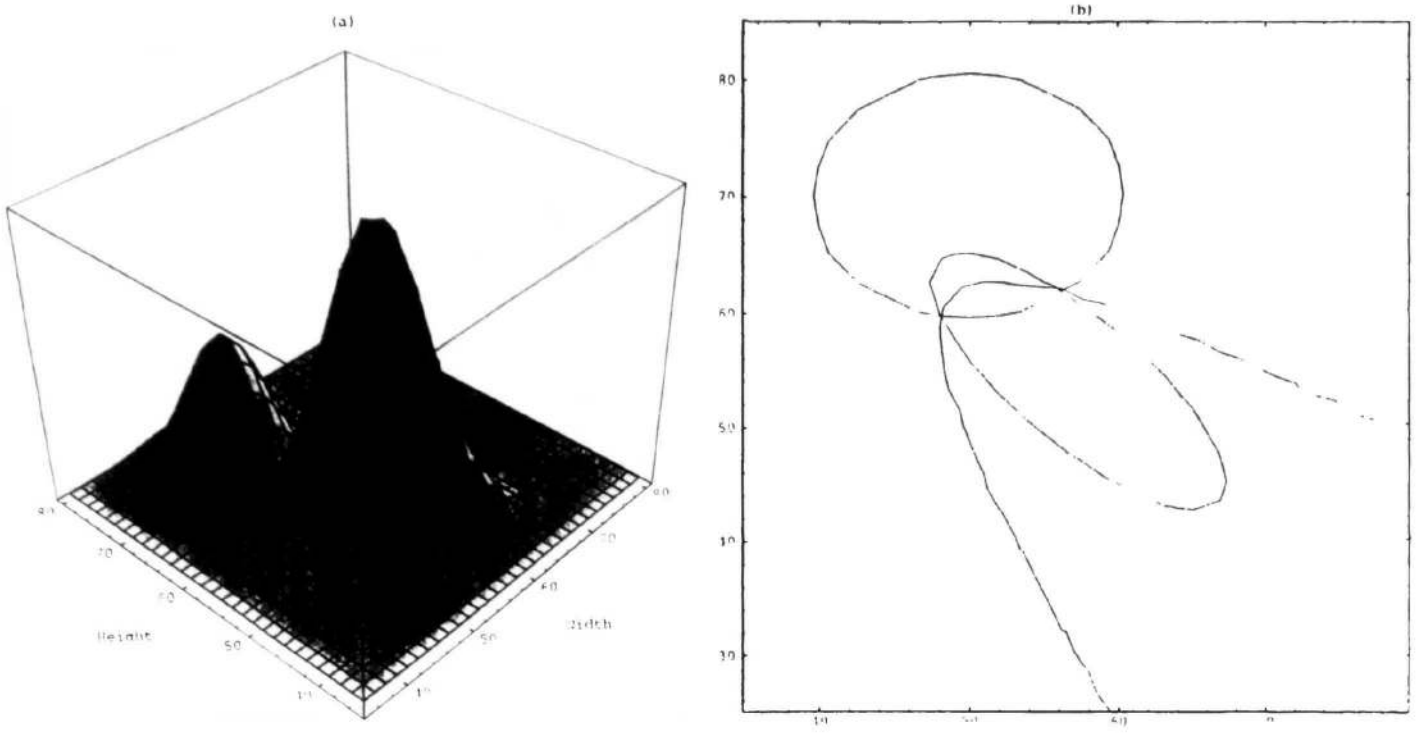

Figure 1 The two categories from experiment (1) are shown in (a) as two frequency distributions. A slice parallel to the stimulus plane (b) reveals equiprobability contours which are shown along with the optimal decision boundary.

then s' is normal with mean μ_i and covariance $\Sigma_i+\Sigma_z$. Therefore, the posterior probability of category A given perceived stimulus **s**' is just: p'(A|**s**) = p(A|**s**').

Since increased noise is equivalent to drawing stimuli from categories with larger variance, perceptual noise effectively moves the two distributions closer together (makes them less discriminable) and therefore moves the optimal decision boundary. This movement is always toward less curved boundaries, moving in this case from an ellipse to a hyperbola to a line.

Criterial noise is a process which perterbs the criterion (or the subject's ability to discriminate the posterior from the criterion). Allow criterial noise *c* such that:

$$r(A|\mathbf{s}) = \begin{cases} 1 \text{ if } p'(A|\mathbf{s}) - p'(B|\mathbf{s}) > c \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

If *c* is a Laplace random variable, then the step-function is transformed into a logistic. Proofs of the equivalence between the deterministic and probabilistic response selection models extend back to Luce (1963). For brevity, this paper will focus only on the probability matching surface, i.e., where the probability of responding A when shown stimulus s, r(A|**s**) is equal to p'(A|**s**).

Response noise is equivalent to randomly guessing about which category label to assign to a stimulus some fixed proportion of the time. So, assuming no bias in guessing:

$$r'(A|\mathbf{s}) = (1-\alpha)\,r(A|\mathbf{s}) + \alpha/2 \quad (3)$$

where α is the guessing rate. Response noise has no effect on the decision boundary, but does change the response surface gradient; the slope of the logistic is decreased with increasing α.

One might reasonably hypothesize that subjects could reduce the size of these two noise quantities during learning. If the level of σ goes down during learning, then subjects are perceiving the stimuli more accurately at the end of the experiment. If α decreases, then subjects are being more cautious about making avoidable errors.

Restricted optimization

Response noise alone cannot produce changes in decision boundaries, and additive Gaussian perceptual noise can result only in boundaries which are less sharply curved than optimal. If the boundaries are more complex than is optimal, or if they are too sharply curved, perceptual noise cannot provide an explanation. An alternative is to consider category learning as an adaptive process, and look to adaptive systems for explanations of non-optimal behavior. This is the theory behind the use of connectionist networks to understand human category learning (Kruschke 1993).

Back propagation trained networks are, asymptotically, optimal classifiers. However, they depend on a number of resources in order to converge on the optimal weights, the

effects of which can largely be determined only empirically. The resources relevant to networks as a model of human category learning are:

1) PARAMETERIZATION: A sufficiently parameterized network is one such that additional parameters cannot significantly reduce error.

2) CONVERGENCE: Network training algorithms require enough exposures to the data at appropriate learning rates to minimize error.

3) DATA SUFFICIENCY: Data are sufficient when a fully converged, sufficiently parameterized network trained on sufficient data will generalize perfectly.

4) NETWORK COMPLEXITY: The basis functions (hidden units) of a network determine what functions will be approximated most readily.

5) DATA COMPLEXITY: A network of fixed parameters is limited in the complexity of the function it can estimate. For example, category boundaries must be of bounded dimensionality.

6) COST FUNCTION: For any distribution of training data, optimal network parameters can only be guaranteed when the information-theoretically appropriate cost function is used.

A modified optimality hypothesis (Kalish 1993) holds that these restrictions are the cause of observable non-optimalities in categorization. Parameterization and convergence can be continuously varied, but network complexity is more difficult to modulate. In this paper, complexity is varied by using two types of basis functions: linear sigmoids and radial Gaussians with tunable covariance.

To recap: for any given categorization task, performance is either optimal with respect to a reasonable objective function, or it is not. Any number of models can achieve optimal categorization, but each makes different predictions about how performance changes as people learn. The optimality model uses noise to explain suboptimal performance, while the restricted adaptive system models depend on the effects of their various resources. In order to distinguish the models, behavior of subjects in two category learning experiments was compared against model predictions.

Experiment One

The GRT of Ashby & Maddox (1990) proposed that subjects learn to use quadratic decision boundaries. This experiment tested that hypothesis by providing a quadratic optimal bound. The extent to which subjects approached that bound, and how they did so serve as the evidence for distinguishing the optimal and RAS models.

Method

Subjects Seven undergraduate students from UCSD were paid to participate in single one hour sessions. They received a bonus payment which increased with the proportion of correct responses.

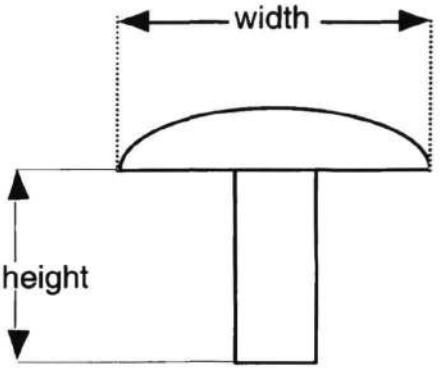

Figure 2 A sample stimulus from the experiments. The width of the mushroom's cap and height of its stem were the two dimensions along which categories were defined.

Apparatus The stimuli were schematic mushroom shapes, pictured in figure 2. Stimuli were drawn from the two categories in figure 1, with the minimum possible change in either stimulus dimension set at 1/12th of an inch. Stimuli were displayed by a Macintosh IIci computer on a color monitor. Subjects made responses on a numeric keypad was covered by a shield through which the two response keys alone extended

Procedure Each subject was read a set of instructions by the experimenter while viewing a sample stimulus. Subjects were told to do their best to classify the stimuli into two types, as indicated by a tone received when the response did not match the class from which the stimulus had been drawn. Subjects were told that perfect performance was attainable only by chance, and that the decision of class membership would be more equivocal for some stimuli than others.

Table 1: The optimal and best fitting polynomial bounds for each block

Block	Bound
1	$-3.18 = 0.527y - 8.1*10^{-4}x^2 - 0.012y^2 + 9.3*10^{-5}xy^2 - 8.62*10^{-7}xy^3 + 6.48*10^{-7}y^4$
2	$-9.58 = 0.1607x - 0.2857y - 2.0*10^{-5}x^2y + 2.475*10^{-7}xy\text{^}3$
3	$-5.74 = 0.2116x - 0.2601y - 1.0*10^{-5}x^3 + 6.58*10^{-8}y^4$
4	$-3.26 = 0.2546x - 0.2204y - 2.0*10^{-5}x^3$
Optimal	$-300.0 = -6.19x - 5.0y + 0.03x^2 + 0.056xy + 0.02y^2$

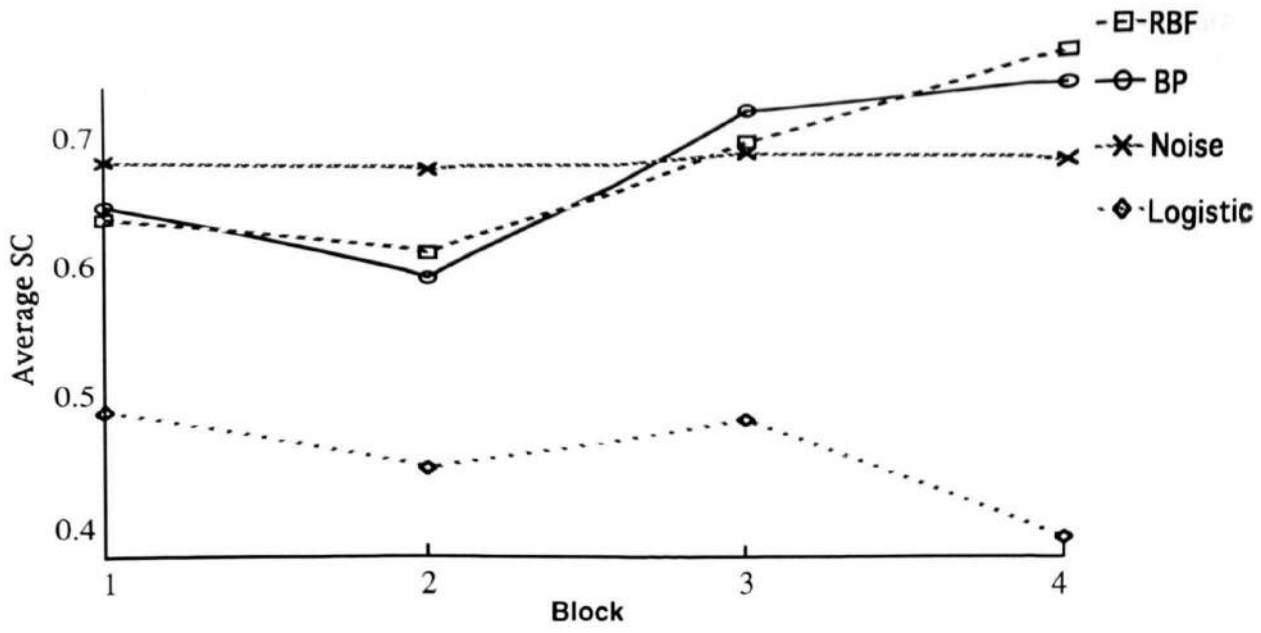

Figure 3. Schwartz criterion fits of models to data from Experiment One. The RAS models (BP=linear sigmoid, RBF=radial gaussian), which embody a learning theory, fit the data as well as the noise-affected optimal model, which was fit at each block independently. The superiority of the logistic shows that the response surfaces are not quadratic.

Each trial was generated by first randomly selecting one of the two equally likely categories. The particular stimulus displayed was selected according to the covariance matrix for the chosen category. If the subject chose the wrong response key, and error tone sounded for 100ms. Then, or 50ms after a correct response, the stimulus disappeared and the next was displayed following a 150ms intertrial interval.

All subjects saw the same 2000 experimental trials in 4 blocks of 500 trials each. At the end of each block subjects were given a self-timed rest period and informed of their proportion of correct responses, and of the value of their bonus payment.

Results

Of the seven subjects in the experiment, only one failed to do better than chance at distinguishing the categories. The data from that subject was discarded from later analysis; data from the other six subjects was collapsed after confirming that there was no significant difference in the means of the judged categories. Subjects were also analyzed individually; the findings there do nothing to contradict the analysis of the group data.

The shape of the response surface for each block was estimated using a stepwise logistic regression procedure. The four bounds are shown in table 1. Each is significantly non-linear, and also contains significant coefficients of higher-than-second order terms. The success of the polynomially bounded logistic in characterizing the data are shown in figure 3.

The ability of the optimal model to fit these data was assayed by optimizing the levels of σ and α at each block. The model fits are shown in figure 3. The RAS models were fit on a trial-by-trial basis, optimizing learning rates, biases and a scale factor which represents the relative discriminability of the two stimulus dimensions. These results are also shown in figure 3, where all fits are measured by the AIC, which equalizes for different numbers of parameters.

Discussion

These data provide a partial replication of Ashby & Maddox (1990), in that subjects are shown to have non-optimal category boundaries. However, inspection of the boundaries through polynomial regression shows them to be different from the optimal (quadratic) in form and order. RAS models fit (nearly) as well as by-block noisy optimality model, and provide more plausible interpretation of learning. The noise parameters of the optimality model changed non-monotonically across blocks, which is in contrast to the theory of learning-as-error-reduction.

Experiment Two

This experiment introduces non-quadratic boundaries in order to further test the GRT model. Four Gaussian distributions with separate covariance matrices make up the four categories. The boundary between one category and the rest is thus a difference between mixtures, and has a complex shape. The chosen distributions are shown in figure 4, along with the optimal decision boundaries.

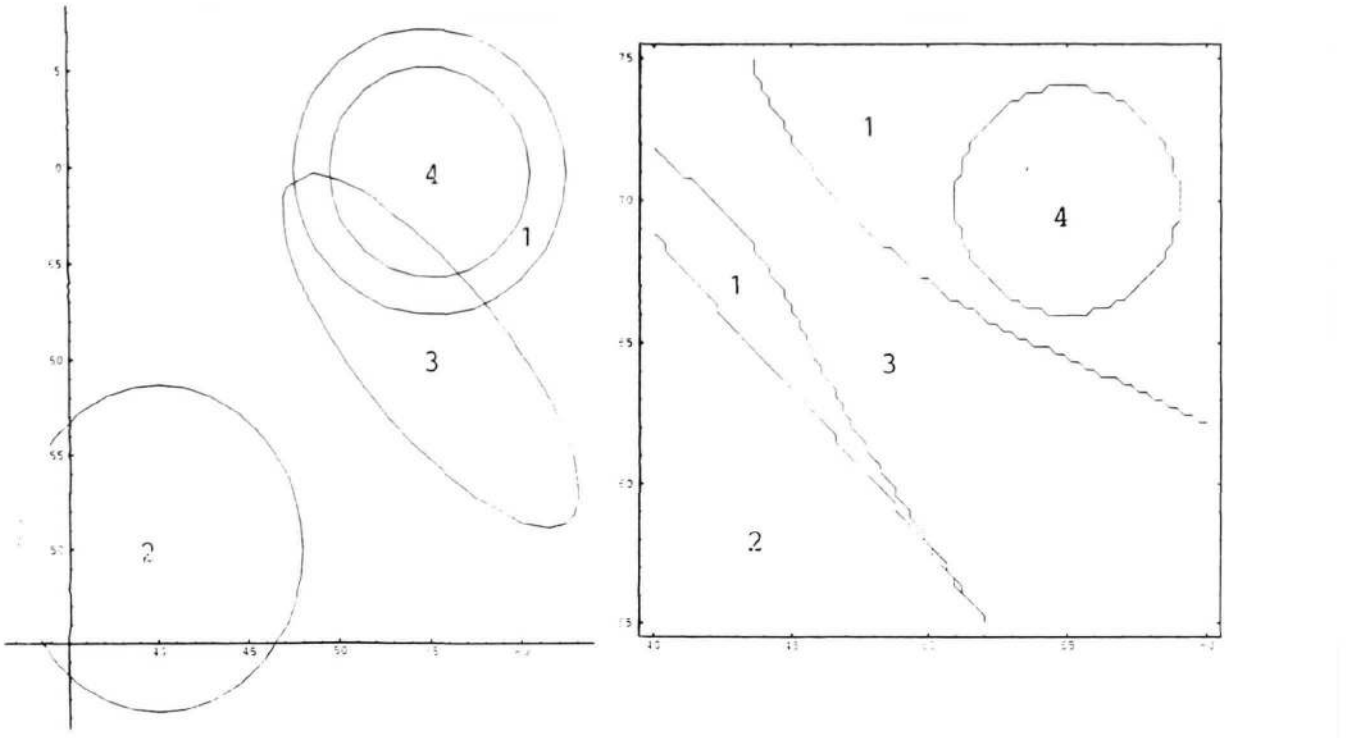

Figure 4: (a) Category equiprobability contours, and (b) Optimal decision boundaries, for Experiment Two.

As figure 4(a) shows, two of the categories have equal means and zero covariance, differing only in their variance terms. Only one of the distributions has a non-zero covariance, and one boundary is nearly linear. This configuration of categories poses difficult challenges for the learners, and so an extended training regime was used

Method

Subjects Four UCSD students participated for payment in three one hour sessions over a single week.

Apparatus The same equipment as in Experiment One was used, with the exception that four keys were constructed from the numeric keypad

Procedure Similar instructions as in Experiment One were given, but subjects were told there would be four 'different types of mushrooms' presented. The three training sessions each included the same 2000 stimuli, divided into four blocks. As before, only negative feedback was provided to the subjects.

Results

The decision regions of the four categories are not simply quadratically bounded. However, pairwise comparisons between bivariate normal categories will reveal quadratic bounds. Optimality can thus be measured by considering the order of subjects pairwise bounds.

The subjects' decision bounds were determined from their responses for each of the twelve blocks of training. The observed decision bounds contained many significant coefficients of the qubic and quartic terms, indicating that subjects were not using a quadratic boundary. The empirical optimal bounds were estimated from the training data at each block, and were well described by at most qubic polynomials. The subjects bounds, where of the same order, were of different form (eg., different qubic components) from the optimal. As shown if figure 5, the RBF and linear sigmoid networks both achieved similar levels of fit, indicating that the added complexity of the RBF nodes was not necessary to approximate the change in subjects' responses during learning.

Discussion

As in Experiment 1, subjects adopted non-optimal category boundaries. Their responses varied in proportion to the likelihood of the chosen category, as shown by the fit of the RAS models. However, the similarity between different RAS fits suggest that more general properties of the systems (eg., gradient descent technique) are shared by the subjects in the experiment. In addition, the non-monotonic changes in α and σ needed to fit the optimal model argue against noise-reduction as a vehicle of learning.

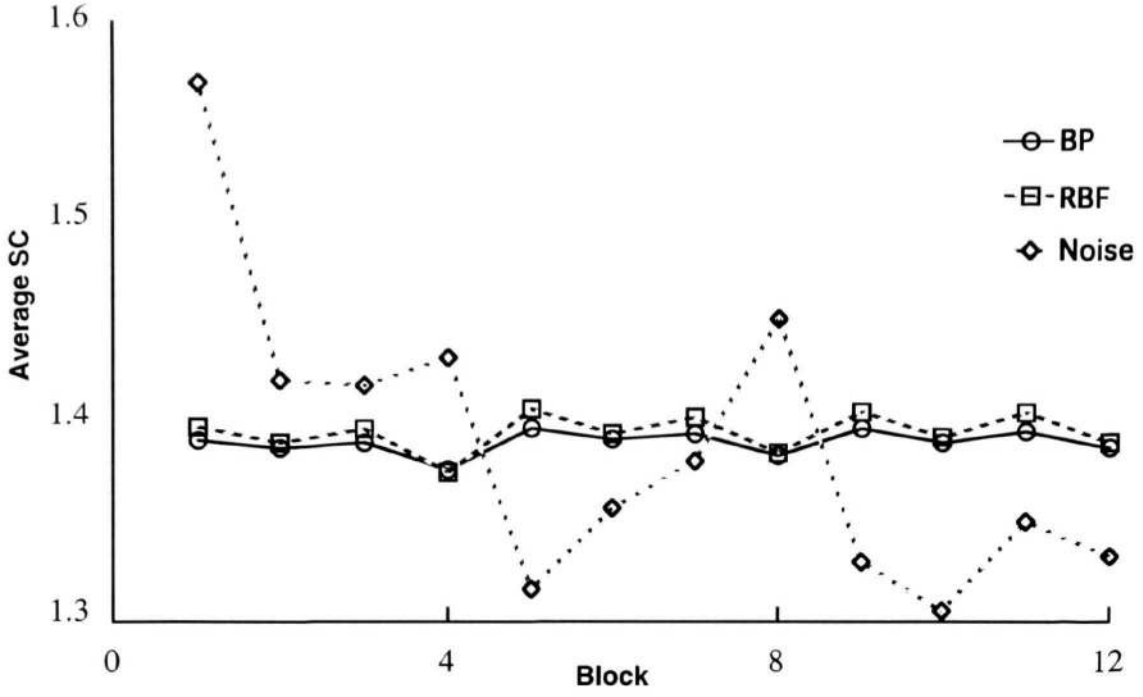

Figure 5 - Schwartz criterion fits of the three theoretical models to Experiment Two. The linear sigmoid and radial gaussian models are nearly identical, indicating that common properties of error reduction are at work. As with Experiment One, the response surfaces here are non quadratic, although the fit of the polynomial logistic is not shown.

Conclusion

Bivariate normal categories are optimally searated by quadratic bounds. The posterior probability of one category versus another is also always quadratic. Leading models of human categorization predict quadratic response surfaces whenever the training data are bivariate normal. In contrast, the Modified Optimality Hypothesis predicts that response surfaces only become quadratic assymptotically. The RAS systems provide a pricipled account of changing non-optimal behavior.

In these experiments, the complexity of the RAS basis function (composition of the network hidden layer) had less to do with this success than did the common factors of gradient descent method and objective function. Two factors might explain this emphasis. First, the network models were all constrained to match the data on a trial-by-trial basis. As has been noted (Chapman 1991), this restriction is likely too harsh. Subjects likely rehearse the stimuli, at least covertly, and so effectively resample the training data Second, the chosen basis functions may both be inadequate for describing people's adaptive learning (Kruschke 1993). Using more psychologically plausible models of categorization, it is possible to formulate and test RAS systems which have a closer correspondance to human function.

References

Ashby, F.G. & Maddox, W.T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance,* **16**, 598-612.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, **37**, 372-400.

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgement. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **17**, 837-854.

Estes, W K. (1986). Array models of category learning. *Cognitive Psychology*, **18**, 500-549.

Gibson, E. J. (1969). *Principles of perceptual learning and development.* NY: Appleton-Century-Crofts.

Kalish, M. L. (1993). *Information integration: an ecological and connectionist approach.* Doctoral dissertation, University of California, San Diego.

Kruschke, J. K. (1993). Three principles for models of category learning. In. G. V. Nakamura, R. Taraban and D. L. Medin (eds.), *Categorization by Humans and Machines: The Psychology of Learning and Motivation, Volume 29*, pp. 57-90. San Diego: Academic Press.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.) *Handbook of Mathematical Psychology* (pp. 103-189). New York: Wiley.