From Natural Kinds to Complex Categories

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia

Abstract

The rational analyses of generalization proposed by Shepard (1987) and Tenenbaum and Griffiths (2001a) are extended to category learning. Categories are assumed to possess a rich microstructure of subtypes, and people are assumed to adapt to the context implied by the learning task. Selective attention and typicality gradients are shown to emerge from Bayesian inference over the possible category microstructures and task contexts.

In outlining the universal law of generalization, Shepard (1987, p. 1322) wrote: "We generalize from one situation to another not because we cannot tell the difference between the two situations but because we judge that they are likely to belong to a set of situations having the same consequence". According to the theory, the probability of generalizing from one stimulus to another decays exponentially with distance in a suitable psychological space. The theory argues that people perceive stimuli as members of *natural kinds*, where a natural kind is assumed to occupy a *consequential region* that is connected, convex, and centrally symmetric. The probability of generalizing from x to y is treated as the probability that the two belong to the same natural kind. Naturally, these probabilities are dependent on the assumptions one makes about the distribution of natural kinds. However, Shepard demonstrates that under a wide range of choices for this distribution, the generalization probability is well-approximated by an exponential function of psychological distance. This roughly-exponential decay is well-documented, and Shepard's law is now widely adopted when modelling higher-level cognitive tasks.

The success of the exponential law has led other researchers to extend the theory, to see if these notions of kinds and consequences can deal with a wider range of cognitive phenomena. Recently, Tenenbaum and Griffiths (2001a) reformulated the law as a form of Bayesian inference. Each possible kind/region constitutes a hypothesis for how stimuli were generated. The advantage to this approach is that it naturally handles stimuli that cannot be represented spatially, and the explicitly Bayesian formulation makes it simple to extend Shepard's law to multiple examples. Moreover, by treating concepts as natural kinds, Tenenbaum (1999) used the theory to predict how people learn simple concepts from a set of positive examples.

This attempt to extend Shepard's work has not been uncontroversial. For example, the work has been criticized for not accounting for structural relations in similarity judgments (Gentner, 2001) and for not providing methods for deriving suitable hypothesis spaces (Boroditsky & Ramscar, 2001). While Tenenbaum and Griffiths (2001b) discuss how many of these problems might be addressed, the majority of traditional category learning research still remains outside the scope of the theory (Love, 2001; Heit, 2001). Yet, to some extent "the value of this model surely will be its ability to address already documented phenomena in generalization, categorization, and inductive inference ... To address this large body of existing research, the Bayesian model itself would require some further generalization" (Heit, 2001, p. 673). This paper provides an initial, though tentative, attempt to do so. As with other 'rational' accounts, the paper is concerned with statistical structures that may explain why we act the way we do, rather than the process by which we exploit this structure.

From Natural Kinds ...

The impressive achievement of Shepard's theory is that it is possible to make strong predictions using only the consequential regions that he associates with natural kinds. Yet while Shepard's work is commonly cited to justify the use of exponential functions (e.g., Kruschke, 1992; Love, Medin & Gureckis, 2004; Nosofsky, 1984), very few papers use the consequential regions upon which the exponential law relies. In that sense, the Tenenbaum and Griffiths' model (henceforth the T&G model) is a rare example of a *consequential region model*.

Consequential region models require a stimulus representation $\mathcal{X}$ and a hypothesis space $\mathcal{H}$. The stimulus representation defines the relationships between different objects x, and the hypothesis space contains all the consequential regions that the observer might encounter. A critical constraint is that the hypothesis space should respect the topology of the stimulus representation. For spatially represented stimuli, this implies that $\mathcal{H}$ should be a set of connected regions in the stimulus space, denoted $\mathcal{R}$ (Shepard, 1987). In Shepard's (1987) one-point generalization model, and the basic version of the T&G model for a spatial $\mathcal{X}$, the probability of generalizing from a set of old items $x = (x_1, \ldots, x_n)$ to a new one y is given by the probability that y belongs to the same region r that generated x. However, since r is unknown,

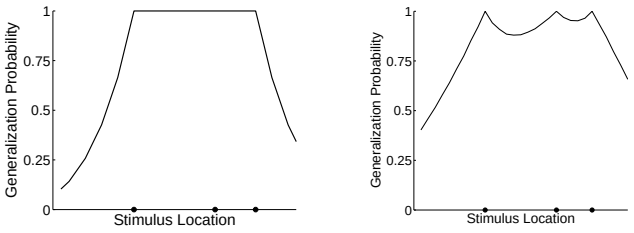

Figure 1: Generalization gradients produced by a consequential region model. The left panel shows the gradients for the original T&G model in which every x comes from the same region, while the right panel shows the gradients for a model in which every x comes from an independently sampled region.

we obtain[1]

$$p(y \in r \mid x) = \sum_{r \mid y \in r} p(r \mid x),$$

where Bayes' theorem implies that $p(r \mid x) \propto p(x \mid r)p(r)$. For objects located in a one-dimensional space, regions are defined by a mean m and a width s. When items are generated from a region, it is assumed that every element of the region is equally likely, so for items inside the region, $p(x \mid r) = 1/s_r$. In any case, because the regions respect the structure of the stimulus representation, the generalization gradient is flat over the locations spanned by the previously observed stimuli, as shown by the left panel in Figure 1.

An extension to the theory (Tenenbaum & Griffiths, 2001b) allows items to be generated from different regions. In an extreme case, the observer might assume that all old items were generated from independently chosen regions. Under this assumption, the chance that a novel item y falls inside at least one of the regions $\cup r = r_1 \cup r_2 \ldots \cup r_n$ is now given by,

$$p(y \in \cup r \mid x) = 1 - \sum_{\cup r \mid y \notin \cup r} \prod_i p(r_i \mid x_i).$$

This generalization gradient is illustrated in the right panel of Figure 1. The difference in the gradients is striking, and illustrates the importance of considering the manner in which multiple stimuli are generated from consequential regions. Naturally, if one makes different assumptions about the generative process, one observes different generalization gradients.

. . . To Complex Categories

In discussing category learning, Anderson (1990, p. 411) remarks that, "[p]eople notice that a number of objects serve similar functions and proceed to form a category to include them". The plural term "similar functions" is crucial. Most cricket bats are wooden, fairly large and covered with reddish marks that indicate they have been used for the function of playing cricket. Others

[1]If the hypothesis space is continuous, the summations should be replaced with integrals.

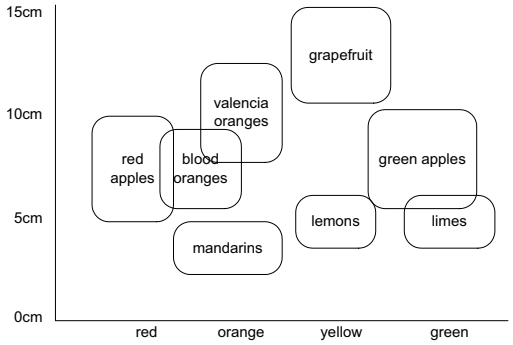

Figure 2: A highly stylized partial map of the (author's) category of fruit, built up from basic-level classes such as apple, orange and lemon. While it is simple to observe that superordinate categories consist of many subregions in the space, it is also true of basic level categories. Red apples and green apples occupy distinct regions, as do Valencia oranges and blood oranges.

lack reddish marks, but are instead covered with the signatures of famous cricketers, indicating that they are used as memorabilia. Still others are plastic, lightweight and yellow, indicating that they are used by children in primary school. As a result, the category of cricket bats is unlikely to consist of a single a consequential region. Rather, the category has a rich structure of subtypes, each of which entail their own idiosyncratic consequences. For instance, Figure 2 shows a map of various different fruit. Not only is the superordinate category 'fruit' broken into multiple regions, so too are the basic level categories of 'apples' and 'oranges'. In view of Wittgenstein's (1953) discussion of 'games' and Rosch's (1978) notion of category hierarchies, it seems likely that this phenomenon is quite general.

Motivated by this discussion, I outline a statistical model for category learning in which each category is built from an arbitrary number of subtypes, each associated with a single consequential region as illustrated in Figure 2. Category learning is primarily associated with the statistical task of inferring these consequential regions, referred to as the *microstructure* of a category. This learning problem has parallels with the cluster recruitment procedure suggested by Love et al. (2004) and the rational category learning approach proposed by Anderson (1990). In all three cases, the learner is required to infer which stimuli should be clustered together. However, while there may be an arbitrary *number* of subtypes, the learner is unlikely to assume that the locations and sizes of the regions are arbitrary. This has important implications for the learning problem, discussed next.

Consider the four building types shown in Figure 3. If a learner is exposed only to towers and skyscrapers, he or she will encounter tall objects that vary much more in height than in width. In contrast, the reverse will be true if the learner only observes airports and factories. Though it may be true that "psychological space has been shaped over evolutionary history so that consequential regions . . . are not consistently elongated or flattened in particular directions" (Shepard 1987, p. 1319), evolu-

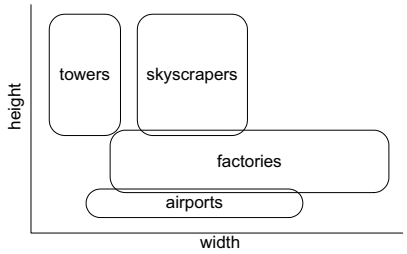

Figure 3: Heights and widths of different kinds of buildings. Observe that the sizes of airports and factories differ from towers and skyscrapers, on the horizontal and vertical dimensions, both on average and in variability.

tionary adaptation will produce an appropriate shaping only on average. In any *particular* context, the regions may be highly elongated. Moreover, these characteristics are evident to the observer: no-one expects novel types of fruit to average two meters in size, or to vary in size only by millimeters, despite the fact that these are both evolutionarily relevant scales for objects. Instead, people recognize that only some parts of the psychological space are plausible locations for consequential regions.

Given this, I assume that the category learner discovers not only the microstructure of the individual categories, but also the gross statistical characteristics of the task, such as where in psychological space the relevant regions tend to be located, and on what scale variation is manifested.

The Statistical Model

A consequential region model for category learning is now developed based on the principles discussed, and is illustrated in Figure 4. Much of the exposition here is necessarily technical. However, since the model is intended only to illustrate principles, the precise details are of less importance than the ideas discussed above.

Richly Structured Categories. Regions are assumed to be rectangular, defined by a mean m and a size s along each dimension, and when generating a stimulus x from a region, all points inside the region are assumed to be equally likely. However, some subtypes are more important than others, so we have a set of importance weights w that are applied to the regions. Formally,

$$\begin{array}{rcll} x_{ij} & | & m, s, r, z & \sim & \text{Uniform}(m^{(z)}_{jr_i}, s^{(z)}_{jr_i}) \\ r_i & | & w, z & \sim & \text{Discrete}(w^{(z)}), \end{array} \quad (1)$$

where x_{ij} denotes the location of the ith stimulus on the jth dimension, r_i denotes the region from which the stimulus is generated, and z is an index for the category.

Learning Region Characteristics. The learner is taken to postulate the existence of an unknown *consequence distribution* over the possible locations and sizes of regions. For simplicity, I assume that this consequential

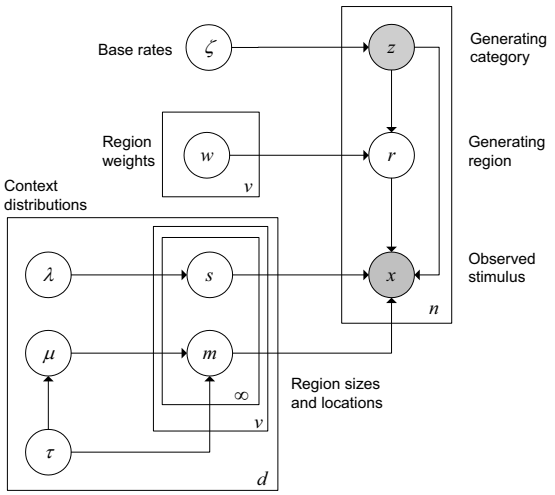

Figure 4: Graphical representation of the Bayesian categorization model. Shaded circles denote observed variables and white circles indicate latent, unobserved variables. Arrows indicate dependencies between variables, while plates enclose a set of independent replications. In this figure, n is the number of stimuli, v is the number of categories, and d is the number of dimensions. Since this paper is concerned with supervised learning, the category variables z are shaded. Naturally, it is only the previous category labels that are observed, not a novel one.

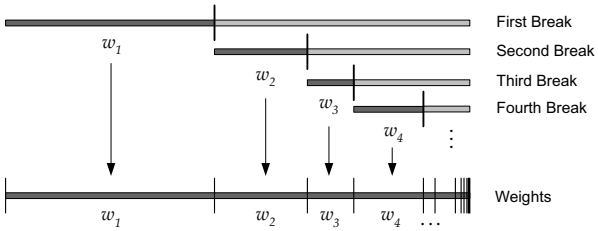

Figure 5: A graphical depiction of the stick-breaking process, showing successive breaks of a stick with starting length one. The final lengths produce an infinite set of weights.

distribution is shared by all categories,[2]

$$\begin{array}{rcll} s_{jr} & | & \lambda_j, \xi & \sim & \text{Gamma}(\xi, \lambda_j) \\ m_{jr} & | & \mu_j, \tau_j & \sim & \text{Normal}(\mu_j, 1/\tau_j). \end{array} \quad (2)$$

Since gammas and normals are both maximum entropy distributions (see Cover & Thomas, 1991), these distributions are plausible choices for a learner possessing little prior knowledge.

Learning Region Assignments. If a category contains regions of variable prominence, we need a prior for these weights. A simple choice is the *stick-breaking process* (e.g., Ishwaran & James, 2001), which provides arbitrarily rich microstructures and allows some regions to be weighted strongly enough to recur frequently during the

[2]This assumption is plausible only when the categories are highly similar to one another. A more general formulation would give each category a unique distribution, and allow the context to induce dependencies between them. However, this complication is unnecessary for the current paper.

generative process[3]. Formally,

$$w \quad | \quad \alpha \quad \sim \quad \text{Stick}(\alpha). \tag{3}$$

Stick-breaking processes are conceptually very simple: an infinite set of weights is produced by taking a 'stick' of length 1, and snapping pieces off as illustrated in Figure 5. If the proportion of the stick broken at each step follows a Beta(1,α) distribution, then the process can continue indefinitely, and the (infinite number of) fragment lengths follow a Stick(α) distribution. Stick-breaking priors are very simple to work with (see Navarro, Griffiths, Steyvers and Lee, 2006), since the weights can be integrated out: if $\mathcal{R}_+$ denotes the set of regions that have previously generated stimuli, then

$$\begin{aligned} p(r_i = k \,|\, r_1, \ldots, r_{i-1}, \alpha, k \in \mathcal{R}_+) &= \tfrac{n_k}{n+\alpha} \\ p(r_i \notin \mathcal{R}_+ \,|\, r_1, \ldots, r_{i-1}, \alpha) &= \tfrac{\alpha}{n+\alpha} \end{aligned} \tag{4}$$

where n_k denotes the number of stimuli previously generated from the kth region. If r_i is not in $\mathcal{R}_+$, then the location and size of the region are sampled from the distributions discussed in the next section.

Learning the Context. Given the inherent variety to the number of situations one might encounter, I assume the learner approaches an unknown context with a diffuse prior, but subsequently learns these gross characteristics of the consequential regions as the context becomes clear:

$$\begin{array}{rcrl} \lambda_j & | & \beta_1, \beta_2 & \sim \text{Gamma}(\beta_1, \beta_2) \\ \mu_j & | & \mu_0, \tau_0, \tau & \sim \text{Normal}(\mu_0, 1/\tau\tau_0) \\ \tau_j & | & \phi_1, \phi_2 & \sim \text{Gamma}(\phi_1, \phi_2). \end{array} \tag{5}$$

These distributions provide conjugate priors for the region distributions in Equation 2. The hyper-parameters β_1, β_2, ϕ_1 and ϕ_2 are all fixed at 1, yielding standard exponential priors, with $\mu_0 = 0$ and $\tau_0 = .001$ yielding a very diffuse Gaussian. The shape parameter ξ is held fixed. Since the space is separable, the distributions for each dimension are kept independent in order to preserve the city-block metric structure that is typical for separable stimuli (see Shepard, 1987).

Base Rates. In general, category base rates (i.e., the relative likelihood of observing items from different categories) can vary. A simple learning model for this would place some prior over the possible base rates, which is updated as category exemplars are learned. A standard model is:

$$\begin{array}{rcrl} z & | & \zeta & \sim \text{Discrete}(\zeta) \\ \zeta & | & \eta & \sim \text{Dirichlet}(\eta). \end{array} \tag{6}$$

Unless there is a reason for people to have an *a priori* bias for one category over another, the prior over ζ is uniform, so I fix $\eta = 1$. Since the Dirichlet is a conjugate prior, it is trivial to find the expected posterior base rate for a category. This is simply $(n_z + 1)/(n + v)$, where v is the number of categories, and n_z is the number of times a member of category z has been observed.

[3]In fact, the stick-breaking priors include the Dirichlet process prior that Anderson's model implicitly adopts.

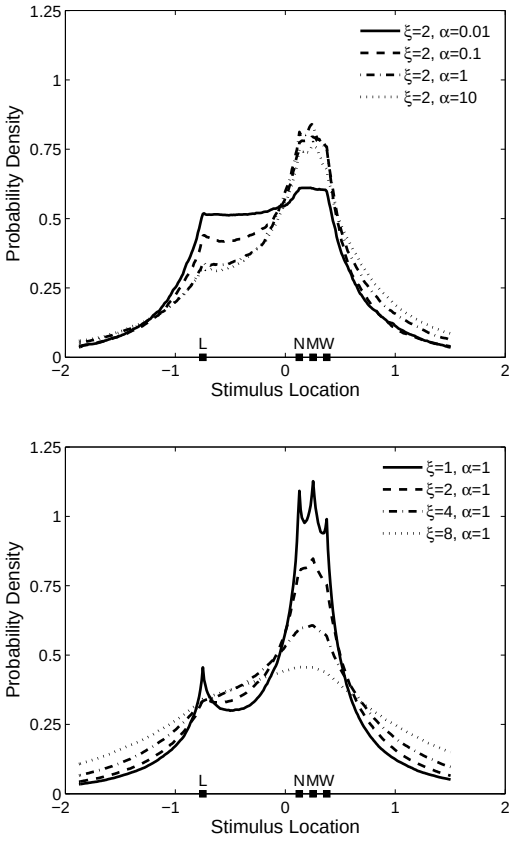

Figure 6: Category distributions for a variety of parameter values.

Statistical Inference. Learning takes place by inverting the generative model, learning the posterior distributions over the category microstructure, and the context-specific region distributions. Since analytic expressions for this posterior distribution do not exist, we can use Markov Chain Monte Carlo methods (e.g., Gilks, Richardson & Spiegelhalter, 1995) to sample from the posterior, allowing numerical approximation. A method for doing so is discussed in a technical note available from the author (Navarro, 2006).

Applications

Two brief applications of the model are presented, showing typicality effects (e.g., Mervis & Rosch, 1981) and selective attention effects (e.g., Kruschke, 1993), chosen because the primitive unit for the model (a natural kind) behaves much like a classical category and shows no graded structure, and because selective attention appears to be a necessary component to concept learning.

Typicality Effects

Typicality effects imply that category membership is graded: some exemplars are better members than others. Although consequential regions have a flat, non-graded structure, some exemplars are more likely to fall within highly-weighted regions than others, producing

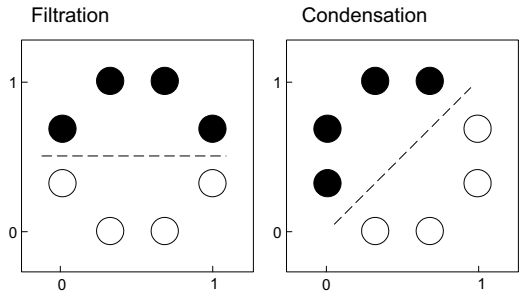

Figure 7: Category structures for a filtration task (left) and a condensation task (right).

typicality effects. To see this, consider a simple perceptual category consisting of the letters M, N, W and L. Multidimensional scaling on the pairwise similarities collected by Rothkopf (1957) shows that a single spatial dimension can account for 90% of the variance[4]. Using this dimension as the representation, we can find the category distributions implied by the Bayesian model.

Figure 6 shows these distributions across a substantial range of the parameter space. Typicality effects can be seen in two respects. Firstly, the exception item L is accorded much lower probability than M, N and W, since a region covering three items receives higher weight than a region encompassing only one. The second typicality effect shown in Figure 6 is prototype enhancement. The letter M, which has vertical sides and twin diagonals, lies between N (vertical sides) and W (twin diagonals). Since the hypothesis space respects the continuity of $\mathcal{X}$, any time that W and N share a region, so does M. Naturally, the reverse is not true. So, ignoring the influence of L, M will always be more typical than N or W.

The figure also shows the effect of varying the two key model parameters, α and ξ. As α increases, the learner is more prepared to assume that new items arise from new regions, leading to a proliferation of regions containing few members. This tends to produce more jagged distributions. In contrast, varying ξ alters the assumptions made about the size of the regions rather than their members. As ξ increases, the regions tend to be larger, which has a smoothing effect. So both α and ξ are smoothing parameters of a sort, but they smooth in different ways.

Selective Attention Effects

Human participants adapt to learning tasks by attending more heavily to dimensions that are diagnostic of the underlying category structure. Accordingly, a key test of any account of category learning is the ability to accommodate attentional effects. Kruschke's (1993) filtration-condensation task provides an initial test, involving linearly-separable categories represented in two separable dimensions (height and position). As shown in Figure 7, there are eight stimuli, arranged octagonally. In a filtration task, the decision boundary runs parallel

[4]Of course, this representation is not entirely appropriate, since continuity is probably violated.

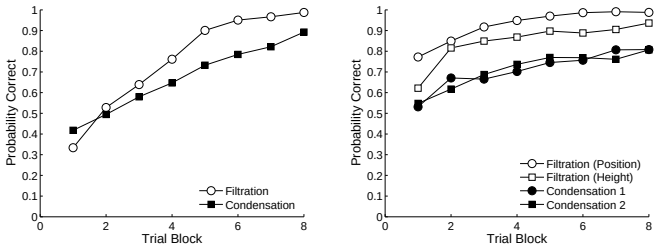

Figure 8: Learning curves for the Bayesian model (left) with $\alpha = 6$ and $\xi = 8$ on a filtration-condensation task, with two empirical data sets shown on the right (from Kruschke, 1993). The model learns faster, but reproduces the filtration advantage effect.

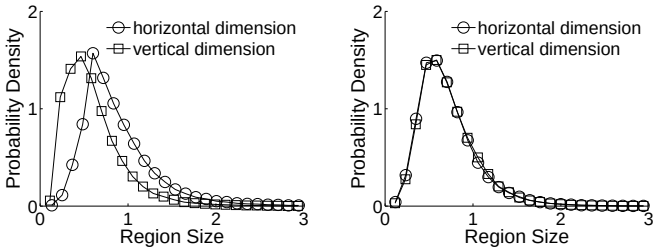

Figure 9: Distributions over region size for the two dimensions in filtration tasks (left) and condensation tasks (right), at the end of the 6th trial block.

to one of the representational axes, while in a condensation task the boundary runs at an angle. With parameter values of $\alpha = 6$ and $\xi = 8$, and injecting a small amount of trial-to-trial Gaussian noise (with $\sigma = 0.025$) to the stimulus representations as a very crude approximation to perceptual and memory errors, the Bayesian model produces the learning curves shown in Figure 8. As one might expect, the Bayesian model learns faster than humans. Still, for the purposes of an initial investigation, what matters is that the model learns to filter faster than to condense.

It is noteworthy that the model can produce this effect without having an explicit attentional mechanism. To some (perhaps limited) extent, attention can be justified as an adaptation to the consequential context implied by category exemplars. This adaptation can be observed in Figure 9, which shows the learned distributions over region sizes at the end of the sixth trial block. In the condensation task, the two dimensions have identically distributed regions, so there is no *differential* scaling of the dimensions. In contrast, the filtration task shows a clear asymmetry between the two dimensions, in that the regions are smaller along the diagnostic dimension. The resulting category distributions for the two tasks are shown in Figure 10.

Discussion

In a sense, the model used in this paper is not novel, and is probably best thought of as an attempt to examine how the consequential regions employed by Shepard (1987) and Tenenbaum and Griffiths (2001a) might be

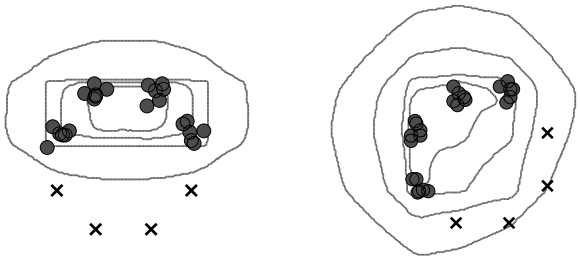

Figure 10: Category distributions after the 6th "trial block" (i.e., presentation of each stimulus once) in a condensation-filtration task, displayed as contours located at the 20th, 40th, 60th and 80th percentiles. Observed instances are shown as grey circles, and the locations of the opposing category members are indicated with crosses. The filtration category (left panel) has a compact distribution that covers the stimuli (black dots) without extending very far along the vertical dimension. In comparison, the condensation category (right) is fairly diffuse, and extends vertically just as far as it extends horizontally.

applied in a supervised learning context. The major observation is that, in order to cover disjoint categories, it is necessary to allow categories to be built from multiple regions. This development was in fact suggested by Shepard (1994), and later explored in a generalization context by Tenenbaum and Griffiths (2001b). Following Anderson's (1990) approach, a flexible prior over the number of (manifest) regions was employed. The resulting model is very similar to Anderson's model, perhaps unsurprisingly. The primary differences lie in the use of flat, sharply bounded regions, and in the inclusion of the context distributions. The context distributions themselves relate to the second aim of the paper, namely to explore the use of hierarchical structure in category learning models. The intuition here is that models of concept learning may need to address Rosch's (1978) horizontal dimension (variation in type) and vertical dimension (variation in abstraction) of categories.

While these initial results are promising, caution is advisable. The model used here is clearly incomplete, and the applications are illustrative at best. If microstructures and context distributions are to be mapped onto "category hierarchies", the same kinds of distributions should govern the relations between types at different levels. This is not currently the case, since the context distributions have a different form to the distributions over possible microstructures. Also, the current analysis is based on the assumption that observations are exchangeable, which seems implausible in changeable environments. To account for trial-order effects such as highlighting (e.g., Kruschke, 2003), this assumption may need to be altered.

Acknowledgements

This research was supported by Australian Research Council grant DP-0451793. I thank Simon Dennis and Michael Lee for helpful comments and discussions, as well as Matt Jones and the other reviewers for their very thorough and useful reviews that have greatly improved this paper.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Hillsdale, NJ: Lawrence Erlbaum.

Boroditsky, L. & Ramscar, M. (2001). "First, we assume a spherical cow ...". *Behavioral and Brain Sciences, 24*, 656-657.

Cover, T. & Thomas, J. (1991). *Elements of Information Theory* New York, NY: Wiley Interscience.

Gilks, W. R. , Richardson, S., & Spiegelhalter, D. J. (1995). *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

Heit, E. (2001). What is the probability of the Bayesian model, given the data? *Behavioral and Brain Sciences, 24*, 672-673.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association, 96*, 161-173.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learing. *Psychological Review, 99*, 22-44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science 5*, 336.

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science, 12*, 171-175.

Love, B. C. (2001). Three deadly sins of category learning modelers. *Behavioral and Brain Sciences, 24*, 687-688.

Love, B. C., Medin, D. L. & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309-332.

Mervis, C. B. & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*, 89-115.

Navarro, D. J. (2006). Statistical inference in a Bayesian model for category learning. *Unpublished manuscript.*

Navarro, D. J., Griffiths, T. L., Steyvers, M, & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology, 50*, 167-179.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 104-114.

Rosch, E. (1978). Principles of categorization. In E. Rosch and B. B. Lloyd (Eds), *Cognition and Categorization* (pp. 27-77). Hillsdale, NJ: Erlbaum.

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks.*Journal of Experimental Psychology, 53*, 94-101.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science, *Science, 237*, 1317-1323.

Shepard, R. N. (1994. Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin and Review, 1*, 2–28.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems, 11* (pp. 59-65). Cambridge, MA: MIT Press.

Tenenbaum, J. B. & Griffiths, T. L. (2001a). Generalization, similarity, and Bayesian Inference. *Behavioral and Brain Sciences, 24*, 629-641.

Tenenbaum, J. B. & Griffiths, T. L. (2001b). Some specifics about generalization. *Behavioral and Brain Sciences, 24*, 762-778.

Wittgenstein, L. (1953). *Philosophical Investigations.* New York: Macmillan.