Learning to learn categories

Amy F. Perfors (amy.perfors@adelaide.edu.au)
School of Psychology, University of Adelaide
Adelaide, SA 5005 Australia

Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain & Cognitive Science, Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Abstract

Learning to categorize objects in the world is more than just learning the specific facts that characterize individual categories. We can also learn more abstract knowledge about how categories in a domain tend to be organized – extending even to categories that we've never seen examples of. These abstractions allow us to learn and generalize examples of new categories much more quickly than if we had to start from scratch with each category encountered. We present a model for "learning to learn" to categorize in this way, and demonstrate that it predicts human behavior in a novel experimental task. Both human and model performance suggest that higher-order and lower-order generalizations can be equally as easy to acquire. In addition, although both people and the model show impaired generalization when categories have to be inferred compared to when they don't, human performance is more strongly affected. We discuss the implications of these findings. **Keywords:** overhypotheses; word learning; Bayesian modelling; shape bias

Introduction

Learning is often thought of as acquiring knowledge, as if it simply consists of gathering facts like pebbles scattered on the ground. Very often, however, effective learning also requires learning *how* to learn: forming abstract inferences about how those pebbles are scattered – how that knowledge is organized – and using those inferences to guide one's future behavior. Indeed, most learning operates on many levels at once. We do gather facts about specific objects and actions, and we also learn about categories of objects and actions. But an even more powerful form of human learning, evident throughout development, extends to even higher levels of abstraction: learning about kinds of categories and making inferences about what categories are like in general. This knowledge enables us to learn entirely new categories quickly and effectively, because it guides the generalizations we can make about even small amounts of input.

Consider, for instance, a learner acquiring knowledge about different kinds of animals. He might realize that CATS have four legs and a tail, SPIDERS have eight legs and no tail, MONKEYS have two legs and a tail, FISH have no legs and a tail, and so on. The knowledge supports what we call a *first-order* generalization: given a new animal that has eight legs and no tail, it is more likely to be some kind of spider rather than a cat or a monkey. However, the learner may also have realized something more abstract: that while the number of legs or the presence of a tail varies a lot *between* categories, these features tend to be homogenous *within* categories. By contrast, surface colorings might vary significantly both between and within categories. This more abstract knowledge, or overhypothesis[1], supports *second-order* generalizations about categories one has never seen: upon seeing a new animal with six legs and a tail, it is reasonable to conclude that other examples of that animal will also have six legs and no tail, but not necessarily the same superficial markings. This higher-order overhypothesis is what allows the learner to form a reasonable prototype of an entirely new kind of animal from only one instance, as well as how to generalize to new instances.

Children as young as 24 months are able to form abstract inferences about how categories are organized, realizing that categories corresponding to count nouns tend to have a common shape, but not a common texture or color (Landau, Smith, & Jones, 1988; Soja, Carey, & Spelke, 1991), whereas categories corresponding to foods often have a common color but not shape (e.g., Macario, 1991; Booth & Waxman, 2002). The advantages of acquiring this overhypothesis, or "shape bias", is clear: teaching children a few novel categories strongly organized by shape results in early acquisition of the shape bias as well as faster learning even of other, non-taught words (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). This is a noteworthy result because it demonstrates that overhypotheses can rapidly be acquired on the basis of little input, but it raises questions about what enables such rapid acquisition. The work in this paper is motivated by these questions about how knowledge is acquired on higher levels of abstraction, and how that kind of learning interacts with lower-level learning about specific items.

In a broader sense, acquiring knowledge on a higher, more abstract level – learning to learn – is important in many contexts besides categorization. In the causal domain, people must draw general conclusions about different novel causal *types* and their characteristic interactions as well as the causal roles fulfilled by specific objects (Kemp, Goodman, & Tenenbaum, 2007). Children learning language must simultaneously acquire knowledge about specific verbs and which arguments they take, as well as higher-order knowledge about entire classes of verbs, some of which may take a certain kind of argument (e.g., a direct object) and others of which cannot. It is this higher-order knowledge that enables people to make intelligent second-order generalizations about verbs they have never seen before (Pinker, 1989).

[1] This terminology is borrowed from Goodman (1955).

For computational theories of learning, the ability to learn on multiple levels at once poses something of a chicken-and-egg problem: the learner cannot acquire overhypotheses without having attained some specific item-level knowledge first, but acquiring specific item-level knowledge would be greatly facilitated by already having a correct overhypothesis about how that knowledge might be structured. Often it is simply presumed that acquiring knowledge on the higher (overhypothesis) level must always follow the acquisition of more specific knowledge.

Recently, a computational framework called hierarchical Bayesian modelling has emerged which can help to explain how learning on multiple levels might be possible. This framework has been applied to domains as disparate as causal reasoning (Kemp, Goodman, & Tenenbaum, 2007), the acquisition of abstract syntactic principles (Perfors, Tenenbaum, & Regier, 2006), and learning about feature variability (Kemp, Perfors, & Tenenbaum, 2007). In the hierarchical Bayesian framework, inferences about data are made on multiple levels: the lower level, corresponding to specific item-based information, and the overhypothesis level, corresponding to abstract inferences about the lower-level knowledge.

In this paper we present a model of category learning which acquires knowledge about how specific items should be categorized as well as higher-order overhypotheses about how categories in general are organized. It is an extension of an earlier model by Kemp, Perfors, and Tenenbaum (2007), which was capable of making inferences at the overhypothesis level but required specific items to be grouped into basic-level categories as part of the input. Our new model can discover how to cluster items at the category level on the basis of their featural similarity, at the same time that it makes inferences about higher-level parameters (or overhypotheses) indicating which features are most important for organizing items into basic-level categories. We show that both first- and second-order generalizations can emerge in tandem, even when category information is not given; it is not necessary for the lower-level knowledge to be acquired first. We compare model predictions with human performance on a novel categorization task with second-order generalization, and demonstrate that human learners follow the same pattern.

Our model is also capable of performing both supervised and unsupervised category learning, which enables us to address the question of how useful category labels are to an ideal learner that can form generalizations on multiple levels. This is a topic of some debate in the infant word learning literature (Xu, 2002; Smith, Jones, Yoshida, & Colunga, 2003). We demonstrate that both human and ideal learners benefit from receiving category information, but human learners benefit more; this may suggest that humans differ from the ideal in their ability to infer the correct category assignments when no category information is given. Both types of learners make stronger generalizations on the basis of highly coherent categories. We discuss the implications and limitations of these findings.

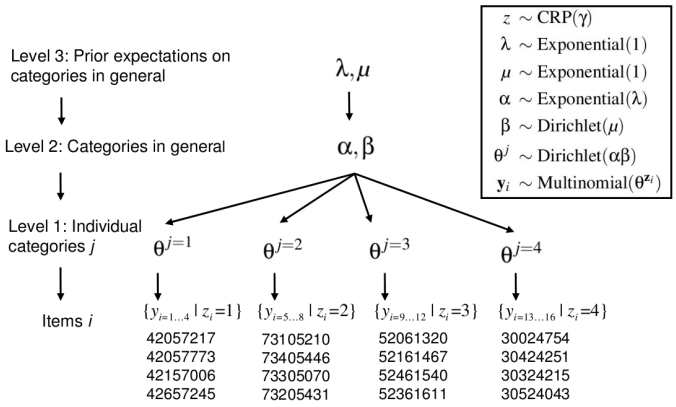

Figure 1: Our hierarchical Bayesian model. Each setting of (α, β) is an overhypothesis: β represents the distribution of features across items within categories, and α represents the variability/uniformity of features within categories (i.e., the degree to which each category tends to be coherently organized with respect to a given feature, or not). The model is given data consisting of the features $\mathbf{y}_i$ corresponding to individual items i, depicted here as a sequence of digits (although representing features as digits implies that order matters; this is not the case for the actual data). Learning categories corresponds to identifying the correct assignment $\mathbf{z}$ of items to categories.

Model

Computational details

Our hierarchical Bayesian model supports the acquisition of two kinds of knowledge: the ability to put uncatgorized items into sensible categories on the basis of their featural similarity, and the ability to acquire more abstract knowledge about the formation of categories in general. An example of the former would be the realization that two entities that share many features (e.g., eat bananas, have two legs, have long tails) are examples of the same category (say, MONKEYS); an example of the latter would be the realization that categories in general tend to be coherent with respect to some features (like number of legs) and not others. The former ability is realized in our model by performing Bayesian inference over possible category assignments; the latter by performing inference over the hyperparameters governing the overhypotheses.

We depict this type of learning graphically in Figure 1 and formalize it more precisely as follows. Each item i is associated with a vector of feature counts $\mathbf{y}_i$, which are drawn from category $j = z_i$. Giving the model category information consists of presenting the model with a partition of items into possible categories, represented by a vector $\mathbf{z}$; if the model is not given category information, it tries to find the best possible $\mathbf{z}$.[2] The prior distribution on $\mathbf{z}$ is induced by the Chinese Restaurant Process, which can be defined recursively by extending a partition over items 1 through $k-1$ to a new item k:

$$P(z_k = c | z_1, \ldots, z_{k-1}) = \begin{cases} \frac{n^j}{k-1+\gamma} & n^j > 0 \\ \frac{\gamma}{k-1+\gamma} & k \text{ is a new category} \end{cases}$$

[2]Throughout the paper boldfaced $\mathbf{z}$ and $\mathbf{y}$ refer to the entire dataset – the full set of $\mathbf{y}_i$ and $\mathbf{z}_i$ for every item i.

Here n^j is the number of items previously assigned to category j and γ is a hyperparameter which captures the degree to which the process favors simpler category assignments (we set $\gamma = 1$, consistent with previous work with this model). The Chinese Restaurant Process prefers to assign items to categories that already have many members, and therefore tends to prefer partitions with fewer categories.

At the same time that the model is attempting to identify the best category assignments, it is also performing inference about the nature of those categories and the overhypotheses that govern them. Level 1 knowledge about the features and items associated with a specific category j is represented by θ^j, which can be understood as the parameters of multinomials that govern how the features $\mathbf{y}_i$ of items i in that category are distributed. This knowledge is acquired with respect to a more abstract both of knowledge, Level 2 knowledge, which in this case is knowledge about the distribution of features across categories in general. It is represented in our model by two parameters, α and β: roughly speaking, α captures the extent to which each individual category is organized by a given feature (or not), and β captures the average distribution of features across all categories in the world.[3]

Level 2 knowledge depends on knowledge at a higher level, Level 3, which is represented in our model by two hyperparameters λ and μ. They capture prior knowledge about α and β, respectively: the range of values expected about the uniformity of features within a category(λ), and the range of values of the expected distribution of features in the world (μ). Our model learns λ and μ in addition to α and β, and assumes that knowledge at the next highest level is given.[4] Inferences about λ, μ, α, and β – in conjunction with inferences about the category assignments $\mathbf{z}$ – can be made by drawing a sample from $P(\alpha,\beta,\lambda,\mu,\mathbf{z}|\mathbf{y})$, which is given by:

$$P(\alpha,\beta,\lambda,\mu,\mathbf{z}|\mathbf{y}) \propto P(\mathbf{y}|\alpha,\beta,\mathbf{z})P(\alpha|\lambda)P(\beta|\mu)P(\lambda)P(\mu)P(\mathbf{z})$$

Inferences about the category-specific distributions θ^j are computed by integrating out α,β,λ,μ, and $\mathbf{z}$:

$$P(\theta^j|\mathbf{y}) = \int_{\alpha,\beta,\lambda,\mu} \sum_{\mathbf{z}} P(\theta^j|\alpha,\beta,\lambda,\mu,\mathbf{z})P(\alpha,\beta,\lambda,\mu,\mathbf{z}|\mathbf{y})d\alpha d\beta d\lambda d\mu$$

Inference is performed by performing a standard numerical stochastic integration technique known as Markov Chain Monte Carlo (Gilks, Richardson, & Spiegelhalter, 1996). When $\mathbf{z}$ is not given, the process of inference alternates between fixing the category assignments $\mathbf{z}$ and sampling the space of hyperparameters α,β,λ, and μ, vs. fixing the hyperparameters and sampling from category assignments. Learning in an HBM thus corresponds to making inferences about category assignments $\mathbf{z}$, as well as the parameters and hyperparameters, based on the input data. First- and second-order generalization are calculated by computing $p(z_k = z_i|\mathbf{y})$, which is the likelihood of a new item k being in the same category as some item i, given their observed feature vectors $\mathbf{y}_k, \mathbf{y}_i$, and all the other observed data in $\mathbf{y}$. This can be calculated[5] by integrating over all of the hyperparameters and all possible category assignments $\mathbf{z}$:

$$P(z_k = z_i|\mathbf{y}) = \int_{\alpha,\beta,\lambda,\mu} \sum_{\mathbf{z}} P(\alpha,\beta,\lambda,\mu,\mathbf{z}|\mathbf{y})\delta_{z_k=z_i} d\alpha d\beta d\lambda d\mu$$

The difference between first and second order generalization is whether item i is already represented in the training set $\mathbf{y}$, or is a new item altogether. All results represent averages across 4 runs of the model.

Datasets

As the category-learning experiments of Smith et al. (2002) demonstrated, it is possible for children to acquire an overhypothesis about the role of shape in categorization after being taught only a few novel nouns; however, it is not clear precisely what aspects of the input enabled such rapid acquisition. Was it the fact that the categories were organized on the basis of highly coherent features, or because the individual items were consistently labelled, effectively providing strong evidence about category assignments? Was it because a certain number of items or categories is required to effectively form overhypotheses, and the children were at the precise critical point in development? Or perhaps people are biased to form overhypotheses about salient features, such as shape, implying that it would be more difficult to acquire overhypotheses about less salient features.

To address these questions we design datasets that vary systematically in terms of (a) coherence of category features; (b) the number of items and categories to be learned; and (c) whether category information is given (the SUPERVISED condition) or must be inferred (the UNSUPERVISED condition). How do these factors affect first-order and second-order generalization? Our goal is to obtain predictions from our model about what an ideal Bayesian learner would do when presented with this sort of input, and then to present human learners with datasets with precisely the same characteristics.

In all datasets, items are associated with eight independent features, four of which have values that are randomly assigned (these are denoted f_R), and four of which are coherent with respect to category membership (f_C). A coherence level of c means that a feature value has a $(100-c)$% chance of being random. By systematically varying the factors of interest, we obtain datasets that correspond to a particular factorial

[3] One way of thinking about the relationship between θ, α, and β is that α captures how close, on average, each individual θ is to β (i.e., how close each individual category's feature distribution is to the overall distribution across all categories). Low α would indicate that each item in a category tends to share a certain feature value, but does not say anything about *what* value that might be: if a category had low α for the shape feature, one would know that it was organized by shape, but not know precisely what shape it was.

[4] We also evaluated performance of a model that assumed that knowledge about λ and μ is given ($\lambda = \mu = 1$, as in Kemp, Perfors, and Tenenbaum (2007); results were qualitatively similar in all cases, but learning at Level 3 as well as Level 2 resulted in a quantitatively better match to human data.

[5] δ is the Kronecker delta function, equal to 1 if $z_k = z_i$, 0 if not.

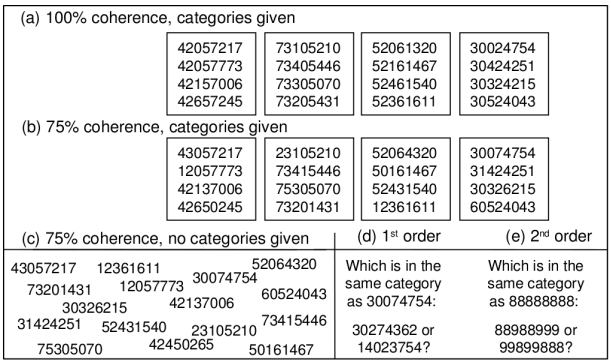

Figure 2: A schematic depiction of the nature of different datasets presented to both humans and our model. Items are associated with four coherent features (f_C) and four random ones (f_R); here we depict each feature as a digit, and its value as the digit value. (a) An example dataset in the SUPERVISED condition with 16 items four of whose f_C features are 100% coherent (all items in the category share the same feature value). (b) As an illustration, we show an example dataset whose four f_C features are 75% coherent: for each feature and item, there is 25% probability that its value will differ from the value shared by most members in the category. (c) The same dataset as in (b), but in the UNSUPERVISED condition. Here the model must learn both the proper categorization as well as the higher-order inference about which features are coherent. (d) A sample first-order generalization task: given an item seen already, which of the test items are in the same category, the one sharing features f_C or the one sharing features f_R? (e) Second-order generalization, which is the same except that the model is presented with entirely new items and feature values.

experimental design: 2 (SUPERVISED or UNSUPERVISED) x 3 (coherence level of 60%, 80%, or 100%) x 2 (containing 8 or 16 items total) x 3 (categories made of 2, 4, or 8 items), slightly complicated by the constraint that each category must have at least two items. As a result of this constraint, the last two factors, when crossed, lead to 5 possible category structures at each coherence level, once in the SUPERVISED and once in the UNSUPERVISED condition.

We assess model performance by examining first-order and second-order generalization. First-order generalization corresponds to presenting the model with an item that occurs in the dataset and querying whether it is more likely to be in the same category as an item that shares coherent features f_C (a "correct" generalization) or random features f_R? Second-order generalization is identical, except the model is presented with an item and features that have not occurred before. Figure 2 contains further details.

Results

Figure 3 shows the model's probability of correct generalization as a function of three factors – whether categories were given for the training data or had to be inferred (SUPERVISED versus UNSUPERVISED), whether the generalization was first-order or second-order, and the coherence level of the training dataset – averaged across all trials with the same levels of these factors.[6] Interestingly, there is no difference between first-order and second-order generalization in either condition ($p > 0.05$, n.s., two-tailed). This result may seem counterintuitive, but further reflection suggests that it is sensible: second-order generalization occurs on the basis of inferences about the overhypothesis, and these inferences effectively have more data bearing on them (all datapoints, not just the specific ones).

Generalization is better in the SUPERVISED condition than in the UNSUPERVISED condition ($p = 0.0006$, two-tailed), although the size of the effect is not large: though category information helps somewhat, especially when the features are less coherent, the fairly high performance of the model in the UNSUPERVISED condition suggests that to the extent that the features of a category are coherent enough to support generalization, they also support categorization, and an ideal learner can take advantage of this. Since there will always be uncertainty about which categories are most appropriate there is some benefit to being given category information, but it is not huge. The affect of coherence on generalization in the model is significant[7], which is sensible: if categories are more incoherent, less generalization is appropriate.

To what extent do humans look like our learner? Do people also find first-order and second-order generalization equivalently easy? Is category information useful? Do they too show differential performance based on how coherent the categories are? We address these questions in the next section.

Experiment

Our experiment is designed to present participants with the exact task and dataset presented to our model, in order to most closely compare performance between the two.

Items. Because the model was presented with items that each had eight independently-generated features, four random (f_R) and four more coherent (f_C), we designed items with the same characteristics for the experiment. They consisted of a square with four characters (one in each quadrant) surrounded by circles at the corner, each containing a character of its own. The characters corresponded to the features of the items in the model datasets, and were designed to ensure that they were salient and discrete, as in the model. Which of the four features varied coherently changed from trial to trial and participant to participant, to eliminate order or saliency effects of any particular feature or feature combination.

Trial structure. Each trial had several phases. In the first phase, participants were shown a set of novel objects on a computer screen and either asked to sort them by moving them around the screen with a mouse and drawing boxes around the ones they thought would be in the same category (in the UNSUPERVISED trials) or were shown the objects already sorted with boxes drawn around them (in the SUPERVISED trials.

After the first phase, each participant was asked two generalization questions, presented in random order. In the first-

[6] We also examined effects of the number of categories and number of items per category, but for space reasons these analyses will be deferred to a longer report.

[7] One-way ANOVA, $p < 0.0001, F = 7.19$.

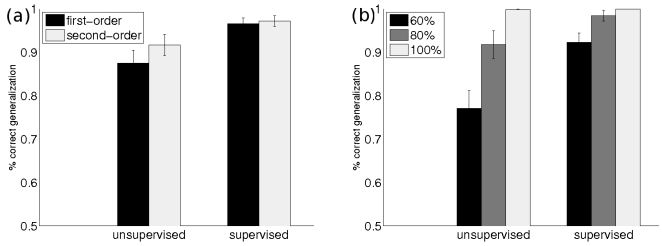

Figure 3: (a) Model generalization averaged across all datasets based on the nature of the category information given. There is no significant difference between first- and second-order generalization. Although category information aids in generalization, the effect is small. (b) Coherence affects generalization, especially in the UNSUPERVISED condition.

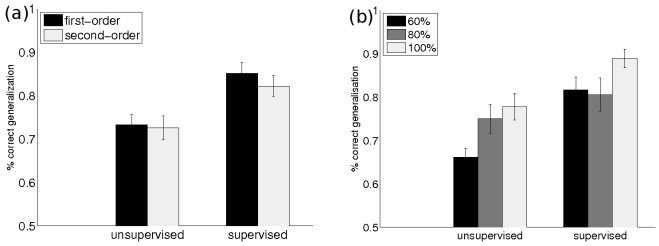

Figure 4: (a) Subject performance on categorization task by condition. Like the model, participants performed equally well for both first- and second-order generalization (SUPERVISED condition, $p = 0.1176$, n.s.; UNSUPERVISED condition, $p = 0.7551$, n.s., both two-tailed). However, they did worse without category information than with it ($p = 0.0001$, one-tailed). (b) Subject generalization, like in the model, was affected by coherence (one-way repeated measures (within-subject) ANOVA: UNSUPERVISED condition, $p = 0.0081, F = 5.16$; SUPERVISED condition, $p = 0.0446, F = 3.25$).

order generalization questions, they were shown an item corresponding to one of the items they had already seen, and asked which of two other novel items were most likely to belong in the same category as that one. The second-order generalization questions were identical except that the participants were presented with items and feature values they had not seen before. All of the sorted items were visible to participants throughout the task. To maintain interest in the task, after completing both questions participants were told how many of the two they got correct, but not which ones.

Procedure. Each participant was shown 30 trials, half SUPERVISED and half UNSUPERVISED, in random order. The factorial design of the experiment corresponded precisely to the design of the datasets presented to the model.

Participants. 18 subjects were recruited from a paid participant pool largely consisting of undergraduate psychology students and their acquaintances. The experiment took 1 hour to complete and participants were paid $12 for their time.

Results

Figure 4(a) demonstrates that, as predicted by the model, first-order and second-order generalization do not significantly differ for human learners. This may be somewhat contrary to intuition, but the fact that this is evident for both human learners as well as the model lends further support to the notion that higher-order generalization need not be more difficult than lower-order. Learning to learn is not only useful, but apparently not too difficult either.

Figure 4(b) shows that people's generalizations depend on coherence, although this result is far noisier than shown by the model. We also see that humans, like the model, were aided by being given category information; however, people's generalizations deteriorated substantially more in the UNSUPERVISED condition. Why do humans have poorer generalization when the categories were not given? One possibility is that they simply fail to identify the correct categories, and in these cases generalize incorrectly. Another possibility is that they succeed in identifying the correct categories most of the time, but are less confident in those categories or less able to make generalizations on the basis of them.

To decide between these hypotheses, we evaluate the correctness of category assignments using the adjusted rand index *adjR* (Hubert & Arabie, 1985), a measure of similarity between two clusterings (in this case, the correct categories vs. the category assignments made by the participants). Most trials (67%) in the UNSUPERVISED condition had high *adjR* values (over 0.5), indicating substantial agreement between the correct categories and the category assignments made; a full 92% were better than chance. Figure 5(a) suggests that people's relatively poorer performance in the UNSUPERVISED condition is carried by the minority of situations in which they were unable to find the correct categories, since when they found the correct ones their generalization performance was quite high. As Figure 5(b) shows, the effect of coherence disappears when considering only those trials in which people found the correct categories; they look more like the model in the SUPERVISED condition.

Discussion

One interesting finding of our work is that both the model and our participants show that first-order and second-order learning – learning to learn – can occur at the same time as each other; it need not be harder to perform second-order generalization than it is to perform first-order generalizations. Our model predicted this result, and we confirmed it empirically in human performance as well. The fact that higher-order generalization may at times be easier (or at least equivalently easy) to lower-order generalization has interesting implications for questions of innateness: although we generally infer that higher-order generalizations must be innate if they are observed early in development, this result implies that such an inference may not always be valid.

Another interesting aspect of this work is the comparison of model and human performance when given category information and when not. For both humans and the model, generalization worsened when not given the category information, but human performance worsened substantially more. This is probably because people had a harder time identifying the correct categories than the model, perhaps due to capacity limitations. It may be possible to model such limitation

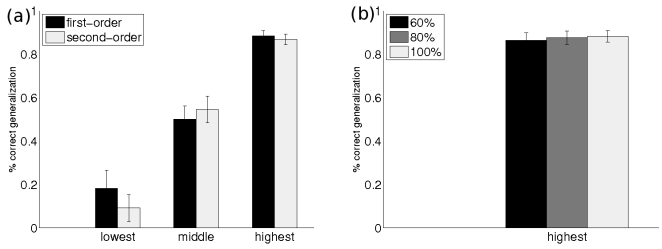

Figure 5: (a) Subject performance on categorization task based on categorization success. The HIGHEST group succeeded in finding the correct categories (had *adjR* scores above 0.5); the MIDDLE group had *adjR* scores above chance, but not substantially; and the LOWEST group were below chance performance in sorting items into categories. Participants who succeeded in finding the correct categories had high generalization performance, indicating that people's relatively poorer performance in the UNSUPERVISED condition was probably due to a difficulty in identifying the correct categories. (b) Among trials in which the correct categories were found (i.e., the HIGHEST *adjR* group), overall generalization (collapsed across first- and second-order) was uniformly high, regardless of coherence.

through the use of particle filters, a limited MCMC process, or memory constraints captured by dropping data. In future work we aim to explore this in more detail.

Although hierarchical Bayesian models have been applied in many other domains, the essential insight – that learning proceeds on multiple levels of inference at once – is rarely explicitly incorporated into models of category learning. Our model can be seen as a version of the hierarchical Dirichlet Process framework introduced by Griffiths, Canini, Sanborn, and Navarro (2007), but with one crucial difference. Our model infers the higher-level parameters describing overhypotheses directly from the data subjects observe, as part of modeling their learning process[8]; in contrast, Griffiths et al. (2007) fit these parameters directly to subjects' behavioral data, without modeling how subjects might infer them. It is the inference of higher-level parameters that supports "learning to learn." In this sense, our model is perhaps most similar to the hierarchical model proposed by Navarro (2006). However, like the original overhypothesis model of which this work is an extension, Navarro's model does not learn to categorize specific items in addition to performing more abstract inferences.

More abstractly, the notion that part of category learning consists of making inferences about which features "matter" is widespread, but is typically framed as the learning of attentional weights rather than as an inference about the abstract principles underlying categorization in a domain (see Kruschke (2008) for an overview). Most models of categorization with learned attentional weights adjust those weights through a process of supervised learning (Kruschke, 2008), and thus do not explain how people learn what features matter in an unsupervised situation as in our experiment. An unsupervised version of SUSTAIN would perhaps be closest to our models' ability to simultaneously discover a system of categories as well as the inductive biases that constrain those categories (Love, Medin, & Gureckis, 2004). Although SUSTAIN and other unsupervised category learning models have not (to our knowledge) been applied to problems of "learning to learn", they could be. Our framework would still offer distinctive insights stemming from its rational basis and the few free parameters.

In the real world, unlike in our experiments or our models, knowledge about which features matter for categorizing is usually restricted to just a certain domain of categories. For instance, children's strong shape bias applies only to categories of solid artifacts, not to living kinds or non-solid substances. An extension of our model can simultaneously discover categories and multiple overhypotheses, as well as which overhypotheses are applicable to which subsets of categories. It incorporates a higher-level nonparametric clustering of categories into 'ontological types' (Kemp, Perfors, & Tenenbaum, 2007), in addition to clustering objects into categories; overhypotheses about categories are shared only within these ontological types. Testing the predictions of this extended model is an important avenue for future work.

This paper presents a computational framework capturing "learning to learn" in categorization and shows that it predicts human performance. The ability to learn on multiple levels at once is a fundamental aspect of human cognition, and our results serve as a step toward understanding that ability.

[8]In HDP terms, it infers the parameters of the base distribution.

Acknowledgments

We thank Dan Navarro, Charles Kemp, and Fei Xu for useful discussions, and Wai Yee Li for her help in running the experiments. JBT was supported by AFOSR grant FA9550-07-1-0075.

References

Booth, A., & Waxman, S. (2002). Word learning is 'smart': Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition, 84*, B11-B22.
Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall.
Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard Univ Press.
Griffiths, T., Canini, K., Sanborn, A., & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical Dirichlet Process. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*.
Hubert, L., & Arabie, P. (1985). Comparing partitions. *Jn of Classification*, 193–218.
Kemp, C., Goodman, N., & Tenenbaum, J. (2007). Learning causal schemata. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*.
Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10*(3), 307–321.
Kruschke, J. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 267–301). New York: Cambridge University Press.
Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*, 299–321.
Love, B., Medin, D., & Gureckis, T. (2004). Sustain: A network model of category learning. *Psychological Review, 111*(2), 309–332.
Macario, J. F. (1991). Young children's use of color in classification: Foods as canonically colored objects. *Cognitive Development, 6*, 17–46.
Navarro, D. (2006). From natural kinds to complex categories. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.
Perfors, A., Tenenbaum, J., & Regier, T. (2006). Poverty of the stimulus? A rational approach. *28th Annual Conference of the Cognitive Science Society*.
Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.
Smith, L., Jones, S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science, 13*(1), 13–19.
Smith, L., Jones, S., Yoshida, H., & Colunga, E. (2003). Whose DAM account? Attentional learning explains Booth and Waxman. *Cognition, 87*, 209–213.
Soja, N., Carey, S., & Spelke, E. (1991). Ontological categories guide young children's inductions of word meaning. *Cognition, 38*, 179–211.
Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition, 85*, 223–250.