A Hierarchical Probabilistic Language-of-Thought Model of Human Visual Concept Learning

Matthew C. Overlan (moverlan@bcs.rochester.edu)
Robert A. Jacobs (robbie@bcs.rochester.edu)
Steven T. Piantadosi (spiantadosi@bcs.rochester.edu)
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY USA

Abstract

How do people rapidly learn rich, structured concepts from sparse input? Recent approaches to concept learning have found success by integrating rules and statistics. We describe a hierarchical model in this spirit in which the rules are stochastic, generative processes, and the rules themselves arise from a higher-level stochastic, generative process. We evaluate this *probabilistic language-of-thought* model with data from an abstract rule learning experiment carried out with adults. In this experiment, we find novel generalization effects, and we show that the model gives a qualitatively good account of the experimental data. We then discuss the role of this kind of model in the larger context of concept learning.

Keywords: Probabilistic language of thought, Bayesian inference, abstract rule learning, computational model, induction, generalization, behavioral experiment

Introduction

A foundational question about human cognition is how we learn as much as we do from input that is often extremely limited. From a few or even just one example, we can make powerful and accurate generalizations. In language learning, for example, there have long been debates about how we learn the complex grammatical structures that we do, given input that is relatively sparse. The ability to learn complex mental representations on the basis of small data sets is also at work in other cognitive domains such as visual perception (Marr & Nishihara, 1978) and causal reasoning (Gopnik & Sobel, 2000) . These mental representations are important because they abstract and summarize the regularities in our environment. By generalizing knowledge gained in specific settings, they allow us to act in novel settings. For the field of cognitive science, a key question is, since there are infinitely many generalizations that are consistent with a finite input, why and how do we generalize in the ways that we do?

To understand people's generalizations, cognitive scientists often debate the merits of statistical versus rule-based approaches. Statistical approaches are appealing because the statistical regularities of data can often be learned relatively easily. Advocates of rule-based approaches argue however, that statistical regularities are inadequate and that, without explicit rules, it is difficult to explain the full richness and complexities of people's generalizations (Marcus, 1999).

As pointed out by others, statistical and rule-based approaches are not mutually exclusive, but rather can be profitably combined (Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Aslin & Newport, 2012). For example, in the field of artificial intelligence, the natural language processing community has long used models that use statistics to infer structured, arguably rule-based representations of syntax (Manning & Schütze, 1999). Within the cognitive sciences, the Rational Rules model (Goodman, Tenenbaum, Feldman, & Griffiths, 2008) showed how we can account for human performance by considering rule learning as Bayesian statistical inference over a structured rule space. Hybrid statistical/rule-based models are sometimes referred to as "probabilistic language of thought" models.

Work in developmental psychology has strongly suggested that even infants generalize in ways that go beyond simple statistical co-occurrences. Marcus, Vijayan, Bandi Rao, & Vishton (1999) showed that seven-month-old infants can learn to recognize sequences of syllables that follow an *ABA* pattern like "ga ti ga" and "wo fe wo", where the first and third syllables are the same but differ from the middle syllable. Gerken (2006) later showed that when infants are exposed to stimuli that are consistent with both a broader (e.g., *ABA*) and a narrower generalization (e.g., *AxA* where x is a specific syllable, not a class or set of syllables), the infants tend to prefer the narrower generalization. This ability to seemingly learn abstract rules from small data sets has even been glimpsed in non-human animals (van Heijningen, Chen, van Laatum, van der Hulst, & ten Cate, 2013).

To account for these types of experimental findings, Frank & Tenenbaum (2011) modeled rule-like patterns with strings that have a symbol for each token in the pattern. The symbols indicate whether the token in a given position is a particular token x, (is_x), the same as another token at position n ($=_n$), or a wildcard (i.e., the symbol could be any token). They modeled learning as Bayesian inference over these structures: $P(h|D) \propto P(D|h)\ P(h)$, where h is a hypothesis representing a specific choice of symbols and D is the observed data. The likelihood $P(D|h)$ is the probability of obtaining the exemplars in D via independent random draws from the set of all strings consistent with h. Importantly, this likelihood follows the "size principle" (Tenenbaum & Griffiths, 2001). If h is a broad hypothesis that is consistent with many possible sets of strings, then obtaining the specific set D from h is small (i.e., $P(D|h)$ is small). In contrast, if h is a focused hypothesis that is consistent with relatively few possible sets of strings, then obtaining D from h is large (i.e., $P(D|h)$ is large). Consequently, likelihood functions that follow the size principle fa-

vor focused hypotheses over broad hypotheses. In Frank and Tenenbaum's model, hypotheses were *a priori* equally likely (i.e., $P(h)$ was a uniform distribution).

Their model gives an impressive account of findings in the literature for abstract rule learning across several domains. Although this work is an important early step in developing a probabilistic language-of-thought account of human generalization, it leaves open many important questions. Their model is limited in the sense that it only includes the bare machinery necessary to account for the specific findings that they consider. To what extent can their model, or rather their general theoretical framework, serve as a foundation for a richer and more broadly-applicable model providing a more comprehensive account of generalization? What is the full range of rules that people might learn and that cognitive models will need to account for? As noted by Frank and Tenenbaum, people have built-in biases for certain hypotheses over others. What are those biases, and how can they be included in a cognitive model?

In light of these outstanding questions, we developed a new probabilistic language-of-thought model for rule learning. This model uses a two-level generative process for explaining items in a data set. At the top level is a stochastic generative process for generating rules. As explained below, the generative process in our model is a probabilistic context-free grammar, and this grammar generates rules. At the bottom level, each rule is a stochastic generative process for generating data items. An innovative aspect of our model is that rules are themselves stochastic generative processes. Because data items are generated stochastically from rules, and because rules are generated stochastically from a probabilistic grammar, the overall generative process forms a hierarchy.

To our knowledge, only one other computational model of concept learning employs such a structure. Lake, Salakhutdinov, & Tenenbaum (2015) created a model of handwritten character recognition that employed a two-level generative model. The top level defined a distribution over the abstract, symbolic representation of a character (the *type*), and then given that specification, the bottom layer defined a distribution over concrete instances of that character as visual strokes (the *token*). Our work relates to and extends this work by casting a hierarchical model in a more general context. The Lake et al. model is highly customized for its domain, so it is unclear how to apply insights from that work to other domains, except at the broadest conceptual level. Our model, however, is built upon the more general Language of Thought framework. Since this framework has already been successfully applied to other domains, and since models in this framework only require the specification of very general primitives, our work is much more readily adaptable to other domains. This also makes for a much more plausible cognitive explanation, as the learning system requires far less manual engineering.

The model is a natural evolution of prior work on rule learning. Previously, theoretical progress was made by incorporating stochasticity into the rule-learning process, and here we incorporate stochasticity into the rules themselves. The remainder of the paper is an exploration of this idea in an abstract rule learning task in a visual domain.

Experiment

We conducted an experiment with adults to test their ability to learn rule-based visual concepts from a small number of examples. Our visual stimuli were part-based 3D objects where the parts act as tokens in an abstract rule (see Figure 1). There were two groups of subjects (30 subjects per group). For one group of subjects, the experiment used the rule *ABA* (as in the experiment by Marcus et al. discussed above). For the other group, the experiment used the rule *xBB* (as in the experiment by Gerken discussed above; *x* is a specific token that is identical in all exemplars).

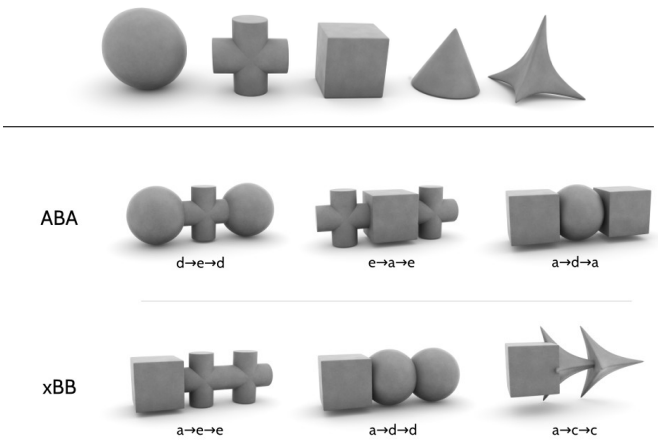

Figure 1: On top, the five parts used in the experiment. Participants viewed these during the instructions phase. On the bottom are the training exemplars for *ABA* and *xBB* conditions.

The experiment was web-based, carried out on Amazon Mechanical Turk. All subjects were US residents over the age of 18. To eliminate the possibility of order and experience effects, each subject participated in only a single condition. The experiment consisted of three stages: an instruction stage, a training stage, and a testing stage. As part of the instruction stage, participants were shown all five possible part shapes. Following the instruction stage, subjects participated in a training stage where they were shown three exemplars from a category. Each subject was allowed to view the exemplars for as long as he or she wished. Training was followed by testing. During testing, subjects were shown an array of 24 test items. Test items had the same general structure as training exemplars (three parts arranged linearly), but differed in which parts occupied each position in an item. Participants chose 'yes' or 'no' for each test item to indicate whether it belonged to the same category as the training exemplars. The exemplars remained available for viewing at the top of the

web page for the duration of the test stage. At least one exemplar was present in the test items. If a participant answered 'no' to that item, his or her results were excluded from the analysis (1 and 2 subjects were excluded in the *ABA* and *xBB* conditions, respectively).

Cognitive Model

In our implementation, concepts or hypotheses begin as lambda calculus expressions. Lambda calculus is a form of logic that is a universal model for computation (i.e., it is equivalent in power to a Turing machine). It characterizes computation using function abstraction and application via variable binding and substitution. For ease of readability, we present these expressions here as procedural "computer programs". These programs construct objects by sampling parts from a fixed alphabet and then combining those parts in a spatial order. They do so using simple set operations, such as removing an element from a set. They can also abstract over parts by assigning them to variables which can be reused. An example program is the following:

```
let x1 = sample(A)
let x2 = sample(A − x1)
output x1 → x2 → x1
```

This program generates the set of objects following an *ABA* pattern. The object is constructed by first randomly sampling part x_1 from the full alphabet, then randomly sampling x_2 from the set consisting of all parts except x_1, and finally combining those parts in the order $x_1 \rightarrow x_2 \rightarrow x_1$. The arrows specify the spatial relationship between parts. Although repetition detection is not explicitly built in, it arises as a natural consequence of the ability to form variable abstractions.

We model learning by assuming that people select the most probable rule or hypothesis h given the set of training exemplars D: $\text{argmax}_h P(h|D)$. The posterior distribution over h can be calculated using Bayes' Rule: $P(h|D) \propto P(D|h)P(h)$. This expression has a natural interpretation in our framework, with the two probabilities corresponding to the two levels of the hierarchy.

The likelihood $P(D|h)$ is the probability that hypothesis h generated the training exemplars in data set D. Assuming the exemplars are drawn independently with replacement (known as the "strong sampling hypothesis" (Tenenbaum & Griffiths, 2001)),

$$P(D|h) = \prod_{d \in D} P(d|h)$$

where d is an individual exemplar. Note that each h, as in the example above, is itself a stochastic generative model. Therefore it naturally defines a distribution over its outputs. Consider "running" the example program repeatedly. Each run produces an independent output that depends on the randomly sampled tokens. $P(d|h)$ is the limiting distribution over those outputs.

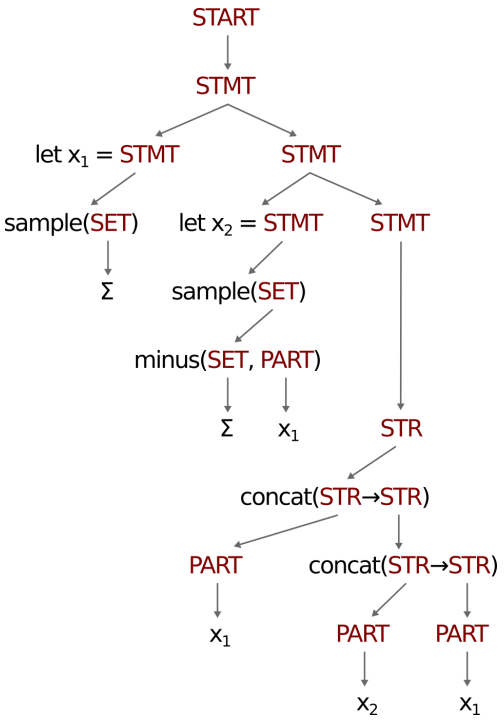

Figure 2: Parse tree for the example program discussed in the main text.

The prior distribution $P(h)$ is the prior probability of hypothesis h. By employing the language-of-thought framework, it too has a natural interpretation. Consistent with earlier language-of-thought models, our model implements the idea that hypotheses are language-like in that they are compositional. Just as sentences are structures built out of words, our model's hypotheses are structures built out of primitives.

This structure is specified by a probabilistic context-free grammar G which defines the syntax for how primitives can be combined. For this experiment, we provided primitive functions *sample(SET)* which samples uniformly from a set, *set_minus(SET, PART)* which removes a part from a set, and *concatenate(STR, STR, ...)* which concatenates strings (which are in turn made up of parts). These primitives are in addition to variable abstraction, which is an inherent property of hypotheses by virtue of their lambda calculus core.

Because the grammar is probabilistic, it defines a distribution over the structures it generates. Each non-terminal in the grammar has an associated distribution that specifies the probability that a production rule will be used to expand that non-terminal. The prior probability of hypothesis h is the product of the probabilities for each of the production rules used in constructing h. In other words, if $\mathcal{T}$ is the parse tree for h and r_i is a rule in this tree, then

$$P(h) = \prod_{r_i \in \mathcal{T}} P(r_i|G).$$

Note that this prior distribution implements a form of Occam's Razor. Since each probability $P(r_i|G)$ is less than one,

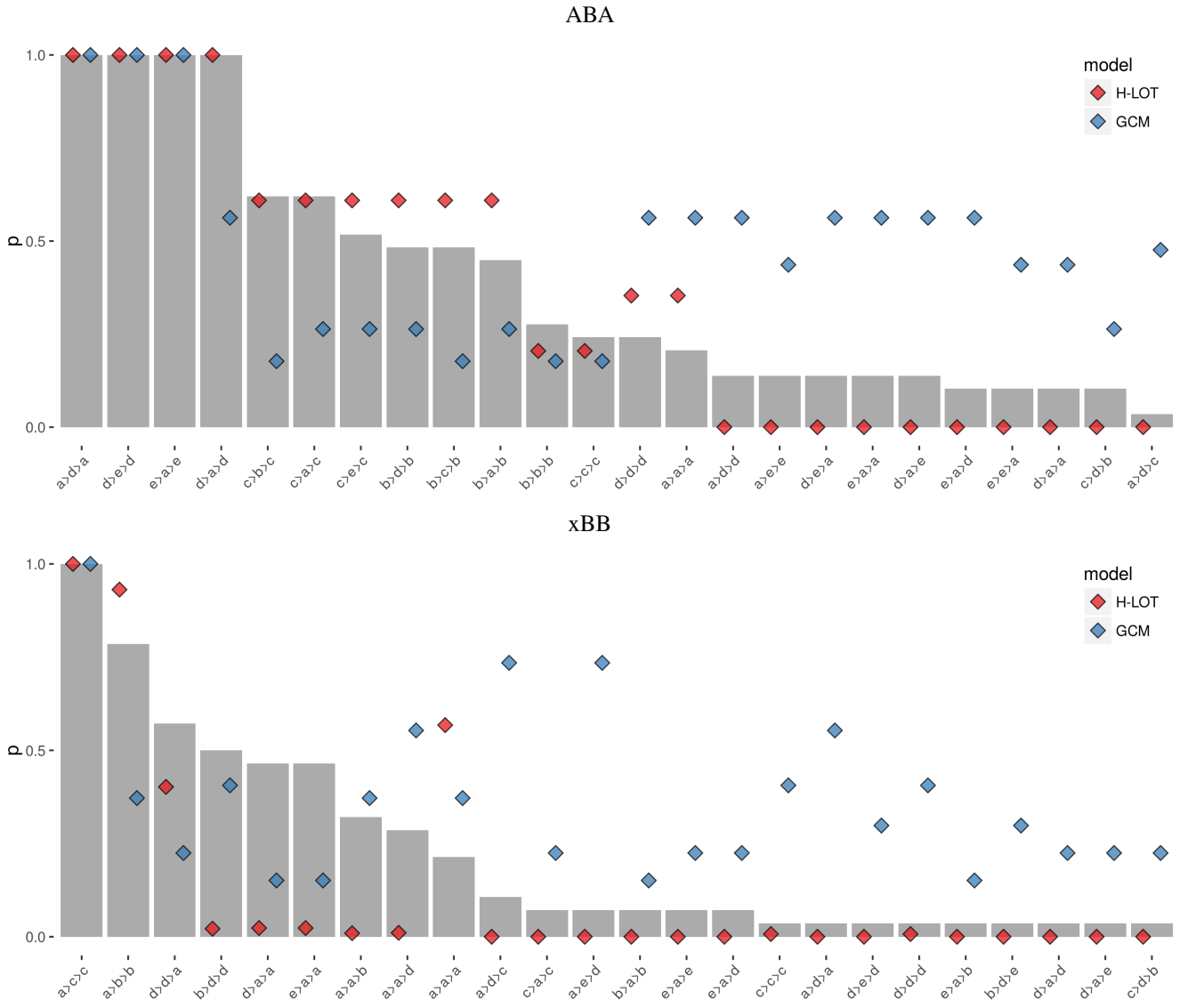

Figure 3: For each test item, the graphs show the probability that subjects judged the item to be from the same category as the training exemplars (plotted by the gray bars in each graph) compared to the probability predicted by our model (labeled H-LOT) and the GCM model. The top and bottom graphs are for the *ABA* and *xBB* conditions, respectively.

hypotheses with short derivations tend to have higher probability than those with long derivations. The parse tree for the example program discussed above is shown in Figure 2.

Finally, each hypothesis has an associated vector of variables θ which allow us to model additional factors that influence generalization. For this experiment, we used two variables. One indicated that the hypothesis should be *orientation invariant*, so that an expression that produced the object $a \rightarrow d \rightarrow d$ would also produce $d \rightarrow d \rightarrow a$. The other indicated that the alphabet should only contain parts that have been seen in the training exemplars rather than the full alphabet of parts. Each of these parameters has an associated prior probability, making the full posterior

$$P(h, \theta | D) \propto P(D|h, \theta) P(h) P(\theta).$$

We estimated this distribution in two steps. First we fixed $P(\theta)$ and sampled the discrete variables h and θ using a Metropolis-Hastings sampling algorithm (a type of Markov chain Monte Carlo algorithm). Because we are sampling in a discrete space, we can approximate the full distribution by saving unique samples and then normalizing. In our Metropolis-Hastings algorithm, we used a slightly modified version of the standard tree regeneration proposal distribution by Goodman et al. (2008). Next, since we have no *a priori* information or theory to indicate how strong the prior tendency to generalize to novel parts or to show invariance to orientation should be, we fit $P(\theta)$ via gradient descent. Since

our search space is tractable enough for the initial sampling step to obtain an approximately complete set of samples, we do not need to do any more sampling after fitting θ.

Results

The results are shown in Figure 3. For each test item, these graphs show the probability that subjects judged the item to be from the same category as the training exemplars (plotted by the gray bars in each graph) as well as the probabilities predicted by two different models. For our model, its prediction for a given test item was calculated by summing the posterior probabilities for all hypotheses that produce that item:

$$P(t|D) = \sum_{h,\theta} P(h,\theta|D) I_{ext(h_\theta)}(t)$$

where I is the indicator function and $ext(h_\theta)$ is the extension of (the set of objects generated by) hypothesis h given parameters θ.

We also show results for the Generalized Context Model (GCM) (Nosofsky, 1986), an exemplar model of category learning that is commonplace in the cognitive science literature. The GCM is a similarity-based model; it determines the category membership of a test item based on its similarity (or inverse of distance) to the exemplars. For this domain, a natural distance function would be one that assigns low distances (and thus high similarities) to pairs of objects that have many parts in common, and high distances to those that have parts that differ. This distance function would serve as a useful comparison because, unlike our model, its representation is relatively unstructured, and it does not have internal variables. We chose the Levenshtein string edit distance, which gives the minimum number of insertions, deletions, or substitutions needed to transform one string into another. We gave this distance function a string representation of the objects. For test item t, the predicted proportion of responses is given by

$$P(t|D) = \sum_{d \in D} e^{-c \cdot Lev(t,d)}$$

where c is a scaling parameter that we fit via gradient descent.

Subjects' responses in our experiment showed large variability, as illustrated by the fact that many subject probabilities (see gray bars in Figure 3) are not near 0 or 1. Despite this variability, our model provides a reasonably good account of subjects' responses, particularly in the *ABA* condition. The GCM model performs poorly; people's generalizations in this task reflect the latent structure present in the objects, thereby going beyond simple similarities.

The tables in Figure 4 show the three hypotheses with the highest probabilities according to our model. These results suggest that the model correctly infers the target rules (*ABA* in the top table of Figure 4 and *xBB* in the bottom table). However, people often deviate from the exact patterns given by these rules, sometimes in interesting ways. For example, the model and subjects may or may not generalize to test items containing parts beyond those used by the training exemplars.

ABA

p	hypothesis	extension set size
.4	let x_1 = sample(A) let x_2 = sample($A - x_1$) output $x_1 \rightarrow x_2 \rightarrow x_1$	20
.24	let x_1 = sample(A_R) let x_2 = sample($A_R - x_1$) output $x_1 \rightarrow x_2 \rightarrow x_1$	6
.15	let x_1 = sample(A) let x_2 = sample(A) output $x_1 \rightarrow x_2 \rightarrow x_1$	25

xBB

p	hypothesis	extension set size
.39	let x_1 = "a" let x_2 = sample(A) output $x_1 \rightarrow x_2 \rightarrow x_2$	5
.21	let x_1 = "a" let x_2 = sample($A - x_1$) output $x_1 \rightarrow x_2 \rightarrow x_2$	4
.18	let x_1 = "a" let x_2 = sample(A) output $x_1 \rightarrow x_2 \rightarrow x_2$ or $x_2 \rightarrow x_2 \rightarrow x_1$	5

Figure 4: The three top scoring hypotheses for each condition as given by the model, along with their posterior probability (p) and the number of objects each hypothesis generates. A is the set of all parts, and A_R is the set of parts that are present in any of the exemplar objects. The model predicts that learned hypotheses should not strictly follow the size principle, as shown by the non-monotonic set sizes.

In addition, the model and subjects may or may not generalize to the linear reversal of patterns (e.g., *BBx* instead of *xBB*). These creative generalizations suggest that people's concept space may be rich in ways that have only rarely been explored by computational models in the cognitive science literature.

An interesting finding is that people often generalize in ways that violate the "size principle" discussed above. For instance, subjects in the *xBB* condition seem to often infer the rule *ABB*. That is, despite the fact that the same token appeared in the leftmost position of all training exemplars, subjects seemed to infer that the leftmost token can be any part so long as it differs from the tokens appearing in the other po-

sitions. In general, the set sizes that people learned are highly variable. This tells us that people are being strongly influenced by the biases and structure of their hypothesis space, which is captured reasonably well by the biases and structure of the language-of-thought prior.

Discussion

We have shown how our theory of rule-based concepts as structured, generative models provides a framework which can profitably model concept learning in the domain of abstract rules. Our approach has several features that make it attractive as a general paradigm for theorizing in this domain. First, it operates as an *ideal observer* (Geisler, 2003)—that is, it defines optimal behavior under a set of assumptions. The language-of-thought framework has the attractive property that the modeler is forced to make those assumptions explicit; they must be encoded directly in the rules, choices of primitives, and probabilities of the grammar. For instance, the model in this paper assumed two computational primitives: a uniform sample operation and an operation to remove an element from a set. Furthermore, these assumptions are psychologically interpretable. The modeling choices afforded by the framework, such as the production probability of a rule, can typically be mapped directly onto psychological phenomena. For instance, we saw that in our experiment, subjects only mildly penalized the complexity added by the use of set operations. Because the framework allows us to decompose the structure of concepts in these ways, we can identify the relevant dimensions along which to aim further work.

As an example of how the framework gives us a lens through which we can frame analyses, we identify several avenues for further investigation both in the domain of abstract rule learning and in wider concept learning. The data hinted that people may be incorporating primitives other than those that we included in our model, perhaps ones that arbitrarily permute or shuffle tokens, or ones that invert them, swapping parts A and B. Would the incorporation of such primitives improve model performance in this domain, and would those primitives be relevant in other domains? There are several dimensions of variation that may influence generalization—the number of training exemplars, the number of tokens in an object, the number of unique tokens across all exemplars, etc. The probabilistic basis of the model allows it to make predictions along all of these dimensions, but further empirical data is needed to test those predictions. More broadly, as the framework allows us to identify potential representational biases, we can then ask why people have those biases? Are they the result of some deeper computational principle? And do they need to be innate or can those biases be learned?

The work presented here is a proof of concept that a two-level hierarchy of generative models can be a powerful framework for modeling and interpreting human rule-learning phenomena. Thinking of concepts as structured, stochastic rules has promising potential to be a general theoretical tool for investigating concept learning in many contexts and domains.

Acknowledgments

This work was supported by AFOSR (FA9550-12-1-0303) and NSF (BCS-1400784) research grants.

References

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current directions in psychological science*, *21*(3), 170–176.

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*, 360–371.

Geisler, W. S. (2003). Ideal Observer Analysis. *The visual neurosciences*, *10*(7), 12–12.

Gerken, L. (2006). Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition*, *98*(3), B67–B74.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: how young children use information about novel causal powers in categorization and induction. *Child development*, *71*(5), 1205–1222.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). MIT Press.

Marcus, G. F. (1999). Do Infants Learn Grammar with Algebra or Statistics? *Science*, *284*(5413), 436–437.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule Learning by Seven-Month-Old Infants. *Science*, *283*(5398), 77–80.

Marr, D., & Nishihara, H. K. (1978). Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes. *Proceedings of the Royal Society B: Biological Sciences*, *200*(1140), 269–294.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of experimental psychology. General*, *115*(1), 39–61.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *The Behavioral and brain sciences*, *24*, 629–640; discussion 652–791.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science (New York, N.Y.)*, *331*(6022), 1279–85.

van Heijningen, C. a. a., Chen, J., van Laatum, I., van der Hulst, B., & ten Cate, C. (2013). Rule learning by zebra finches in an artificial grammar learning task: which rule? *Animal cognition*, *16*(2), 165–75.