Revealing Priors on Category Structures Through Iterated Learning

Thomas L. Griffiths (Thomas_Griffiths@brown.edu)
Brian R. Christian (Brian_Christian@brown.edu)
Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912

Michael L. Kalish (kalish@louisiana.edu)
Institute of Cognitive Science, University of Louisiana at Lafayette, Lafayette, LA 70504

Abstract

We present a novel experimental method for identifying the inductive biases of human learners. The key idea behind this method is simple: we use participants' responses on one trial to generate the stimuli they see on the next. A theoretical analysis of this "iterated learning" procedure, based on the assumption that learners are Bayesian agents, predicts that it should reveal the inductive biases of the learners, as expressed in a prior probability distribution. We test this prediction through two experiments in iterated category learning.

Many of the cognitive challenges faced by human beings can be framed as inductive problems, in which observed data are used to evaluate underdetermined hypotheses. To take two common examples, in language acquisition the hypotheses are languages and the data are the utterances to which the learner is exposed, while in category learning the hypotheses are category structures and the data are the observed members of a category. Analyses of inductive problems in both philosophy (Goodman, 1955) and learning theory (Geman, Bienenstock, & Doursat, 1992; Kearns & Vazirani, 1994; Vapnik, 1995) stress the importance of combining the evidence provided by the data with *a priori* biases about the plausibility of hypotheses. These biases prevent learners from jumping to outlandish conclusions that might be consistent with the data, and can produce successful inductive inferences so long as they approximately capture the nature of the learner's environment.

If we want to understand how people solve inductive problems, we need to understand the biases that constrain their inferences. However, identifying these biases can be a challenge. Inductive biases can result from biological constraints on learning, general-purpose principles such as a preference for simplicity, or previous domain-specific experience, and in many cases will be a mixture of all three. Not all of these factors are available to introspection, and as a consequence assessment of the biases of learners has tended to be indirect. In the past, people's inductive biases have been evaluated using experiments that examine whether, for example, certain category structures are easier or harder to learn (e.g., Shepard, Hovland, & Jenkins, 1961), or by assessing how well models that embody particular biases correspond to human judgments (e.g., Tenenbaum, 1999).

In this paper, we explore a novel experimental method that makes it possible to directly determine the biases of learners. The basic idea behind this method is simple: having people solve a series of inductive problems where the hypothesis selected on one trial is used to generate the data observed on the next. We call this method "iterated learning", due to its close correspondence to a class of models that have been used to study language evolution (Kirby, 2001). Our use of iterated learning is motivated by a theoretical analysis that shows that, in the case where the learners are Bayesian agents, the probability that a learner chooses a particular hypothesis will ultimately be determined by their inductive biases, as expressed in a prior probability distribution over hypotheses (Griffiths and Kalish, 2005). We tested this prediction in two experiments with stimuli for which people's inductive biases are well understood, examining whether the outcome of iterated learning is consistent with previous work on the difficulty of learning different category structures (Shepard et al., 1961; Feldman, 2000).

The plan of the paper is as follows. First, we outline the theoretical background behind our approach, laying out the formal framework that justifies the use of iterated learning as a method for determining the biases of learners. We then provide a more detailed analysis of the specific case of inferring category structures from observed members, presenting a Bayesian model of this task. The predictions of this model, and of our more general theoretical framework, are tested through two experiments. We close by discussing the implications of these experiments for iterated learning as a method for revealing inductive biases, and some future directions.

Iterated learning reveals inductive biases

Iterated learning has been discussed most extensively in the context of language evolution, where it is seen as a potential explanation for the structure of human languages. Language, like many other aspects of human culture, can only be learned from other people, who were once learners themselves. The consequences of this fact have been studied using what Kirby (2001) termed the *iterated learning model*, in which several generations of one or more learners each learn from data produced by the previous generation. For example, with one learner per generation, the first learner is exposed to some initial data, forms a hypothesis about the language it represents, and generates new data from that language. This new data are passed to the second learner, who infers a hypothesis and generates data from it that are provided

to the third learner, and so forth. Through simulations, Kirby and his colleagues have shown that languages with properties similar to those of human languages can emerge from iterated learning with simple learning algorithms (Kirby, 2001; Smith, Kirby, & Brighton, 2003).

Griffiths and Kalish (2005) provided a formal analysis of the consequences of iterated learning for the case where learners are Bayesian agents. Assume that a learner has a set of hypotheses, $\mathcal{H}$, and that their biases are encoded through a *prior* probability distribution, $P(h)$, specifying the probability a learner assigns to the truth of each hypothesis $h \in \mathcal{H}$ before seeing some data d. Bayesian agents evaluate hypotheses using a principle of probability theory called Bayes' rule. This principle states that the *posterior* probability $P(h|d)$ that should be assigned to each hypothesis h after seeing d is

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')} \quad (1)$$

where $P(d|h)$, the *likelihood*, indicates the probability of the data d under hypothesis h.

We can now formally analyze the consequences of iterated learning with Bayesian learners. Each learner uses Bayes' rule to compute a posterior distribution over the hypothesis of the previous learner, samples a hypothesis from this distribution, and generates the data provided to the next learner using this hypothesis. The probability that the nth learner chooses hypothesis h_n given that the previous learner chose hypothesis h_{n-1} is

$$P(h_n|h_{n-1}) = \sum_d P(h_n|d)P(d|h_{n-1}) \quad (2)$$

where $P(h_n|d)$ is the posterior probability obtained from Equation 1. This specifies the transition matrix of a Markov chain, since the hypothesis chosen by each learner depends only on that chosen by the previous learner. Griffiths and Kalish (2005) showed that when the learners share a common prior, $P(h)$, the stationary distribution of this Markov chain is simply the prior assumed by the learners. The Markov chain will converge to this distribution under fairly general conditions (e.g., Norris, 1997). This means that the probability that the last in a long line of learners chooses a particular hypothesis is equal to the prior probability of that hypothesis, regardless of the data provided to the first learner.

Testing convergence to the prior

The theoretical results summarized in the previous section raise a tantalizing possibility: if iterated learning converges to the prior, perhaps we can reproduce it in the laboratory as a means of determining people's inductive biases. However, these results are based on the assumption that the learners are Bayesian agents. Whether the predictions of this account will be borne out with human learners is an empirical question.

To test whether iterated learning with human learners will converge to an equilibrium reflecting people's inductive biases, we need to use a set of stimuli for which these biases are well understood. One such set of stimuli comes from the literature on category learning. Shepard et al. (1961) conducted an experiment exploring the relative difficulty of learning different kinds of category structures defined on objects that vary along three binary dimensions, such as shape, color, and size. Categories are defined in terms of which subsets of the eight possible objects they contain. In principle, there are 256 different category structures, but if we restrict ourselves to categories with four members, this number is reduced to 70. If we collapse together structures that are identical up to rotation and negation, this number is reduced still further, giving us a total of six different types of category structures. Examples of categories belonging to these six types are shown in Figure 1(a).

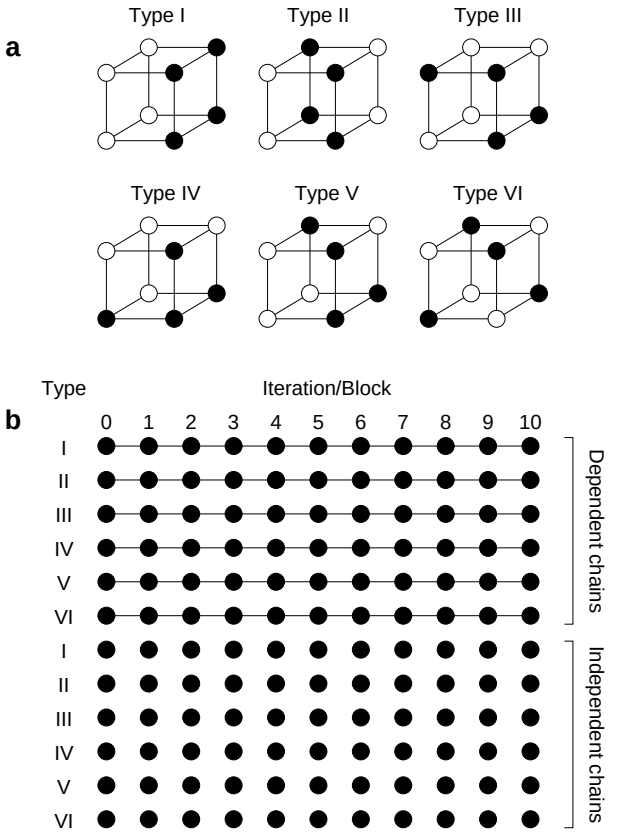

Figure 1: (a) Types of category structures for stimuli defined on three binary dimensions. Vertices are objects, with color indicating category membership. (b) Design of iterated category learning experiments (see Method).

Shepard et al. (1961) found that there is great variation in the ease with which people learn different types of category structures. Type I, in which membership is defined along a single dimension, is easiest to learn, followed by Type II, in which two dimensions are sufficient to identify members. Next come Types III, IV, and V, which all correspond to a one-dimensional rule plus an exception, and are about equally difficult to learn. Type VI, in which no two members share a value along more than one dimension, is hardest to learn. Similar results have been obtained by Nosofsky, Gluck, Palmeri, McKinley, and Glauthier (1994) and Feldman (2000).

Since difficulty in learning a hypothesis is an indication that it may be inconsistent with the inductive biases of the learner, these stimuli provide a way to test the predictions of our theoretical account of iterated learning: we can examine whether iterated category learning using these stimuli converges to a distribution over hypotheses consistent with the results of Shepard et al. (1961). However, in order to do this efficiently, we need to introduce one more innovation. Our discussion so far has focused on cases where iterated learning occurs "between subjects", with each learner seeing data generated by a previous learner. Iterated learning experiments using such a design can be cumbersome, requiring a large number of participants in order to have chains of learners of any appreciable length. Fortunately, the same analysis applies to iterated learning with a "within subjects" design, where a single learner responds to stimuli that are based on his or her own previous responses. In the remainder of the paper, we discuss two experiments in within-subjects iterated category learning. However, before we present these experiments, we will describe a formal model that we will use to make quantitative predictions about the dynamics of iterated category learning.

Modeling iterated category learning

Our account of category learning is based on a model developed by Tenenbaum (1999) and Tenenbaum and Griffiths (2001). The data, d, that people observe will consist of m positive examples – objects that belong to a category. If we assume that objects are drawn by sampling without replacement, then the probability of a particular set of m positive examples is

$$P(d|h) = \begin{cases} (|h|-m)!/|h|! & d \subset h \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $|h|$ denotes the number of objects in the category associated with hypothesis h and $d \subset h$ indicates that all m objects in d are members of h (and $m < |h|$). We can now use Bayes' rule to evaluate the posterior probability of any hypothesis h given some set of objects d. Combining the likelihood given by Equation 3 with a prior on hypotheses, $P(h)$, we obtain

$$P(h|d) = \frac{P(h)\,(|h|-m)!/|h|!}{\sum_{h' \supset d} P(h')\,(|h'|-m)!/|h'|!} \quad (4)$$

for all hypotheses h such that $d \subset h$, and 0 otherwise. All categories are of the same size ($|h| = 4$), so we have

$$P(h|d) = \frac{P(h)}{\sum_{h' \supset d} P(h')} \quad (5)$$

which is simply the prior, normalized over all hypotheses consistent with d.

The likelihood given by Equation 3 and posterior distribution from Equation 5 can be substituted into Equation 2 to find the transition matrix of the Markov chain on hypotheses induced by iterated learning. The result is a square matrix where the number of rows and columns is equal to the number of hypotheses. Given an initial distribution over hypotheses, represented as a column vector, the distribution over hypotheses at each subsequent iteration can be computed by multiplying this vector by the transition matrix. This can be used to make predictions not just about the asymptotic distribution over hypotheses, which we know to be the prior, $P(h)$, but also the dynamics of iterated learning. The asymptotic distribution will not be affected by the amount of data seen by the learners, but the dynamics will change significantly depending on the degree to which the data constrain the choices of the learners.

Our two experiments examine iterated category learning in two regimes: with two positive examples ($m = 2$), and with three positive examples ($m = 3$). The goals of these experiments are twofold. First, to determine whether iterated learning converges to a distribution over hypotheses consistent with people's inductive biases, and second, to establish whether the fine-grained dynamics of this process are consistent with the Bayesian framework presented above.

Experiment 1: Two positive examples

Method

Participants Participants were 20 members of the Brown University community, paid $8 per hour for their participation, and 97 University of Louisiana at Lafayette undergraduates participating for course credit.

Stimuli Following Feldman (2000), stimuli were "amoebae" with a wavy cell wall and an internal nucleus. Nuclei varied along three binary dimensions: shape (round/square), color (black/white), and size (large/small). Categories were "species" of amoebae.

Procedure The design of the experiment is shown in Figure 1(b). Each participant completed 120 trials of category learning, being presented with two positive examples from a category and selecting one of the fifteen category structures that were consistent with those examples. To remove memory demands, examples and hypotheses were presented simultaneously on a computer screen, as shown in Figure 2 (a), and participants selected category structures using a mouse. The 120 trials were divided into 10 blocks of 12, corresponding to 10 iterations of learning. Within each block, six trials belonged to "dependent" chains, with the objects being

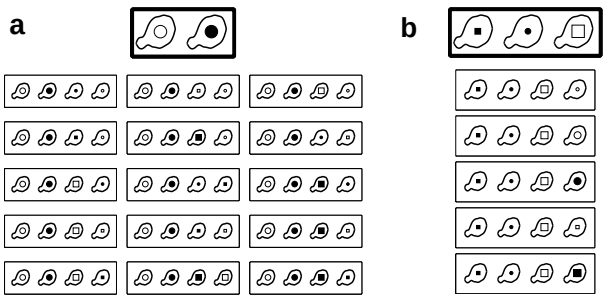

Figure 2: Sample displays showing stimuli and possible responses for (a) Experiment 1 and (b) Experiment 2.

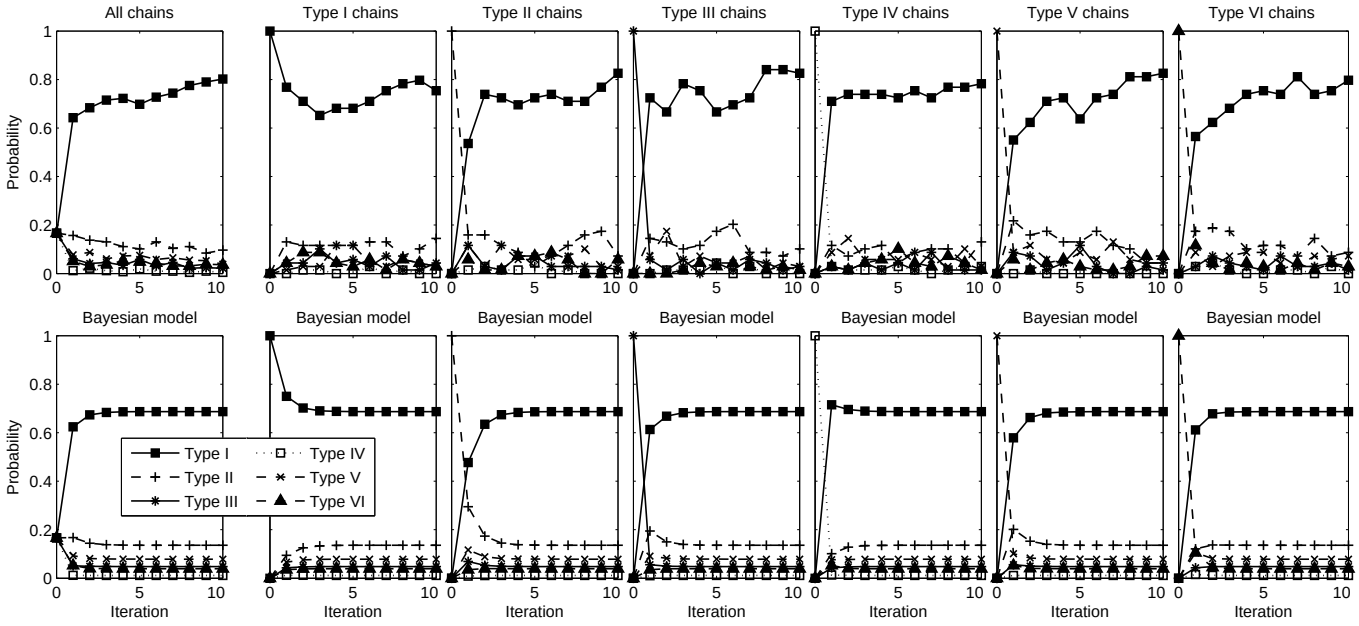

Figure 3: Results of Experiment 1. The leftmost panel shows data aggregated over all chains, while the remaining panels break this down by the type of category structure used to initialize each chain. Each point represents the contribution of 69 subjects, so the maximum standard error is 0.025 for the aggregate data, and 0.060 otherwise.

generated at random from the hypothesis selected on the previous trial in that chain. Each dependent chain was initialized with a hypothesis corresponding to a different type of category structure (these hypotheses are referred to as iteration 0). The other six trials within each block were part of "independent" chains, with the objects being generated from a randomly selected hypothesis corresponding to one of the six types. Trials were randomized within blocks.

Results and Discussion

While previous work makes qualitative predictions about the relative prior probabilities of the six different types of category structure, the model presented above provides the opportunity to estimate these quantities directly, and use them to make quantitative predictions about the dynamics of iterated learning. We specified a prior over hypotheses, $P(h)$, by assuming that the prior probability was affected only by the type of category structure to which a hypothesis corresponds, being uniform within types. Since there are six such types, the prior can be completely specified by five parameters, giving the probabilities of Types I-V (the probability of Type VI follows from the fact that probabilities sum to one).

The parameters of the prior were estimated from the frequencies with which participants selected hypotheses given different sets of examples, aggregated across dependent and independent chains, rather than by optimizing the fit of the model to the dynamics of the data. According to the Bayesian model, people's choices should follow the distribution given by Equation 5. Parameters were found using maximum-likelihood estimation. Preliminary analyses indicated that a subset of the participants were responding at random, so a variant of the EM algorithm (Dempster et al., 1977) was used to simultaneously estimate the prior and probabilistically classify participants as either responding at random or in accord with the model. The resulting parameter estimates gave Types I-VI prior probabilities of 0.687, 0.136, 0.048, 0.012, 0.079, and 0.039, respectively. Computing the actual prior probability of a category structure of each type requires dividing by the number of categories of each type, being 6, 6, 24, 8, 24, and 2, respectively. These probabilities are consistent with previous findings concerning the relative difficulty of learning different types of category structures, with the only possible exception being the relatively high probability of Type VI structures. Using these parameters, we computed the probability that each participant was responding at random. The remainder of our analyses use only the 69 participants for whom this probability was less than 0.5.

The leftmost panel of Figure 3 shows how the proportion of participants selecting a hypothesis of each type varies as a function of the number of iterations, aggregating over all six dependent chains. To evaluate whether iterated learning was having an effect on responses, we ran a χ^2 test comparing the proportions of the six types across the independent and dependent chains at each iteration. The results of these tests were statistically significant for all iterations after the third, with $p < .01$.

Figure 3 also shows the predictions of the Bayesian model outlined in the previous section when applied to this task. As can be seen from the figure, there is a remarkably close correspondence between the predictions of the model and the human data, with a linear correlation of $r(58) = 0.997$. In particular, both the model and the human data converge to an asymptotic distribution over hypotheses consistent with the prior. This close cor-

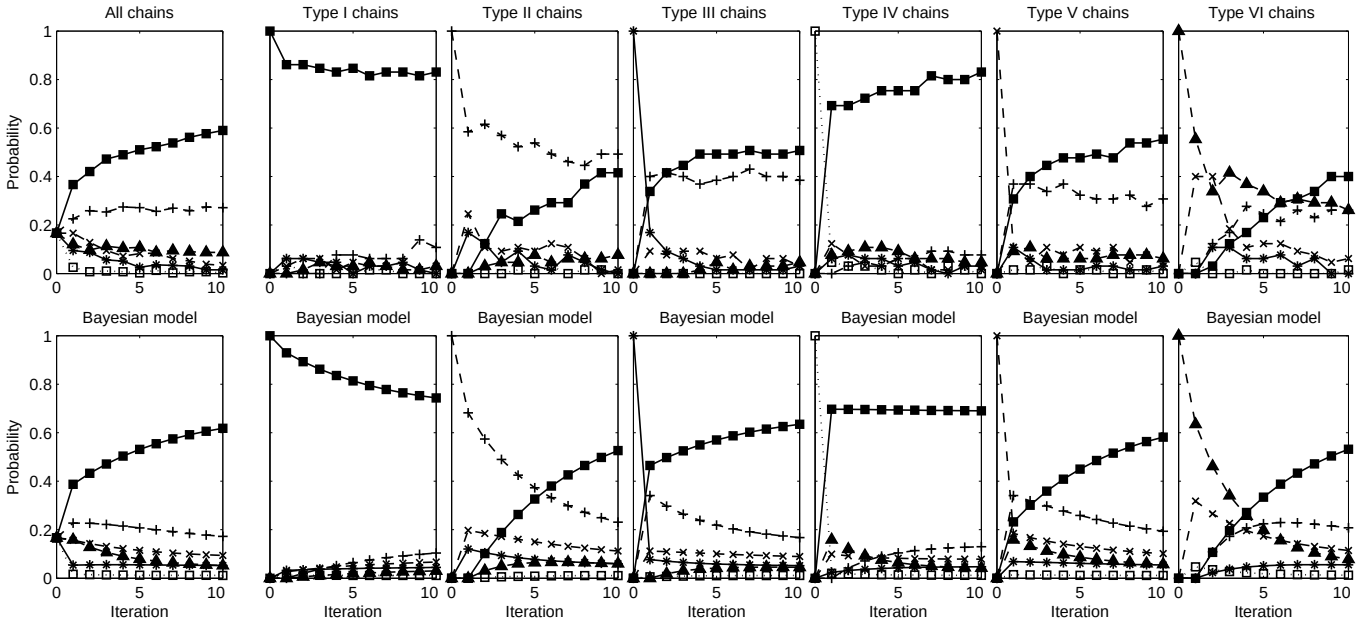

Figure 4: Results of Experiment 2. The leftmost panel shows data aggregated over all chains, while the remaining panels break this down by the type of category structure used to initialize each chain. Each point represents the contribution of 64 subjects, so the maximum standard error is 0.026 for the aggregate data, and 0.062 otherwise.

respondence prevails despite the fact that the model parameters were estimated from the hypotheses that people selected across all trials, rather than explicitly attempting to capture the dynamics of iterated learning. The remaining panels of Figure 3 show a more fine-grained analysis of the correspondence between model and data, breaking the aggregate data shown in the leftmost panels up based on the type of category structure with which the chains were initialized. The model and data still exhibit a strong correlation, with $r(358) = 0.990$, and both converge to a distribution consistent with the prior regardless of initial conditions.

The results of this first experiment bear out the predictions of our theoretical framework, with human learners converging to a distribution over hypotheses consistent with their inductive biases. Furthermore, the dynamics of this process correspond well with the quantitative predictions of our Bayesian model, with convergence occurring extremely rapidly regardless of initial conditions. However, the speed of convergence prevents a detailed analysis of the dynamics of iterated learning, with very little variation in behavior following the second or third iteration. To address this problem, our second experiment examined iterated category learning in a context where the data provided stronger constraints on hypotheses, reducing the rate of convergence.

Experiment 2: Three positive examples

Method

Participants Participants were 20 members of the Brown University community, paid $8 per hour for their participation, and 53 University of Louisiana at Lafayette undergraduates participating for course credit.

Stimuli Stimuli were those used in Experiment 1.

Procedure The procedure was that of Experiment 1, but three positive examples of each category were presented on each trial. Since this meant only one member of the category was unknown, participants had to choose a response from just five consistent hypotheses. Figure 2 (b) shows a sample display from the experiment.

Results and Discussion

The procedure developed for Experiment 1 was used to estimate the parameters of the prior, resulting in probabilities of 0.651, 0.195, 0.040, 0.008, 0.062, and 0.044 for Types I-VI respectively. These parameters were consistent with both previous research and the estimates from Experiment 1. Using these parameters, 64 participants were classified as responding non-randomly, and were used in the remainder of our analyses. The leftmost panel of Figure 4 shows how the proportion of participants selecting a hypothesis of each type varies as a function of the number of iterations. χ^2 tests found a significant difference between dependent and independent chains for every iteration after the third, with $p < .01$.

Figure 4 also shows the predictions of the Bayesian model, which again correlated extremely well with the human data, $r(58) = 0.992$. The stronger constraints on choices imposed by using more examples resulted in much slower convergence towards the prior for both the Bayesian model, and the human data. In particular, it seems that iterated learning has not fully converged after 10 iterations, with the prevalence of Type I still increasing, and the prevalence of other types still decreasing. One small difference from the predictions of the Bayesian model appears for the Type II structures, which should

still be decreasing at the end of the experiment, but appear to have stabilized at a slightly higher probability than predicted by the model.

The remaining panels in Figure 4 show the data broken down across the six dependent chains. The slower convergence results in some interesting differences in the distribution over hypotheses across chains. For example, in the data for the Type VI chains, the dominance of Type I categories only emerges after a period in which Type V increases in popularity. As can be seen from the figure, the Bayesian model does a good job of capturing these dynamics, with a correlation of $r(358) = 0.990$. The slight over-prevalence of Type II category structures relative to the model predictions is more pronounced for the Type II and III chains, and seems to be complemented by under-prevalence in Type I and IV chains. With sufficiently many iterations, the probabilities across all chains should converge. The fact that they remain quite different at the end of the experiment suggests that these discrepancies may be the result of noise rather than a systematic failure of the model.

General Discussion

The results of our experiments bear out the predictions of both our theoretical framework, and our Bayesian model of iterated category learning. In both experiments, the distribution over category structures converged towards an equilibrium consistent with previous research on learning difficulty (Shepard et al., 1961; Feldman, 2000; Nosofsky et al., 1994), with Type I structures being most prevalent, followed by Type II, and then the other four types. The dynamics of this convergence, as represented by the distribution over category structures at each iteration, were also strongly in accord with our Bayesian model: the greater constraints on hypotheses provided by three positive examples resulted in slower convergence to the prior, and the Markov chains initialized with different types of category structures showed fine-grained dynamics that closely matched the predictions of the Bayesian model. These results suggest that iterated learning may provide a viable method for determining the inductive biases of learners.

One interesting aspect of our data is the persistently high probability of Type VI category structures. The previous research mentioned above suggested that Type VI structures are hardest to learn, but the prior that seemed to characterize people's inferences in our experiments gave these structures higher probability than Types III-V. One possible explanation for this difference is the lack of memory demands in our task. The experiments that suggest Type VI structures are hard to learn required participants to remember a set of examples from the category, while in our experiments participants could see both the examples and the full set of possible category structures. Type VI structures actually have far greater symmetry and simplicity than Types III-V, being describable as the structures for which every two members have the same value on exactly one dimension. Our presentation format could have made this property more apparent, resulting in a stronger preference for Type VI.

There are a number of directions in which the experiments presented in this paper could be extended. First, a more complete test of the predictions of our framework, and of the Bayesian model outlined above, could be conducted by considering a wider range of category learning tasks, potentially producing a deeper picture of the dynamics of iterated category learning. Second, given the original proposal of iterated learning as a mode of intergenerational knowledge transmission, exploration of whether similar dynamics are observed when iterated learning occurs "between subjects" could provide insight into questions relating to the consequences of cultural evolution. However, perhaps the most exciting direction for future research is the investigation of people's inductive biases in contexts where they remain unknown. By reproducing iterated learning in the laboratory, we may be able to map out the implicit biases that are at the heart of the remarkable human ability to solve problems requiring inductive inference.

Acknowledgements We thank Anu Asnaani, Rebecca Cremona, Alana Firl, and Vikash Mansinghka for discussions about this project and assistance in running experiments.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, 407:630–633.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias-variance dilemma. *Neural Computation*, 4:1–58.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Harvard University Press, Cambridge.

Griffiths, T. L. and Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In Bara, B. G., Barsalou, L., and Bucciarelli, M., editors, *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*, pages 827–832. Erlbaum, Mahwah, NJ.

Kearns, M. and Vazirani, U. (1994). *An introduction to computational learning theory*. MIT Press, Cambridge, MA.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, 5:102–110.

Norris, J. R. (1997). *Markov Chains*. Cambridge University Press, Cambridge, UK.

Nosofsky, R. M., Gluck, M., Palmeri, T. J., McKinley, S. C., and Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition*, 22:352–369.

Shepard, R. N., Hovland, C. I., and Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, 75. 13, Whole No. 517.

Smith, K., Kirby, S., and Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, 9:371–386.

Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning*. PhD thesis, Massachussets Institute of Technology, Cambridge, MA.

Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24:629–641.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer, New York.