

Running head: LEARNING OVERHYPOTHESES

Learning overhypotheses with hierarchical Bayesian models

Charles Kemp, Amy Perfors & Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences

Massachusetts Institute of Technology

Address for correspondence:

Charles Kemp

77 Massachusetts Avenue Building 46-4053

Cambridge, MA 02139

e-mail: ckemp@mit.edu

phone: 617 324 2894

Abstract

Inductive learning is impossible without overhypotheses, or constraints on the hypotheses considered by the learner. Some of these overhypotheses must be innate, but we suggest that hierarchical Bayesian models help explain how the rest can be acquired. To illustrate this claim, we develop models that acquire two kinds of overhypotheses — overhypotheses about feature variability (e.g. the shape bias in word learning) and overhypotheses about the grouping of categories into ontological kinds like objects and substances.

Learning overhypotheses with hierarchical Bayesian models

Compared to our best formal models, children are remarkable for learning so much from so little. A single labelled example is enough for children to learn the meanings of some words (Carey & Bartlett, 1978), and children develop grammatical constructions that are rarely found in the sentences that they hear (Chomsky, 1980). These inductive leaps appear even more impressive when we consider the many interpretations of the data that are logically possible but apparently never entertained by children (Goodman, 1955; Quine, 1960).

Learning is impossible without constraints of some sort, but the apparent ease of children’s learning may rely on relatively strong inductive constraints. Researchers have suggested, for example, that the M-constraint (Keil, 1979) and the shape bias (Heibeck & Markman, 1987) help explain concept learning, that universal grammar guides the acquisition of linguistic knowledge (Chomsky, 1980), and that constraints on the properties of physical objects (Spelke, 1990) support inferences about visual scenes. Constraints like these may be called theories or schemata, but we will borrow a term of Goodman’s and refer to them as overhypotheses.¹

Although overhypotheses play a prominent role in nativist approaches to development (Keil, 1979; Chomsky, 1980; Spelke, 1990), some overhypotheses are probably learned (Goldstone & Johansen, 2003). One such overhypothesis is the shape bias — the expectation that all of the objects in a given category tend to have the same shape, even if they differ along other dimensions, such as color and texture. Smith, Jones, Landau, Gershkoff-Stowe, and Samuelson (2002) provide strong evidence that the shape bias is learned by showing that laboratory training allows children to demonstrate this bias at an age before it normally emerges. Other overhypotheses that appear to be learned include constraints on the rhythmic pattern of a child’s native language (Jusczyk, 2003),

and constraints on the kinds of feature correlations that are worth tracking when learning about artifacts or other objects (Madole & Cohen, 1995).

The acquisition of overhypotheses raises some difficult challenges for formal models. It is difficult at first to understand how something as abstract as an overhypothesis might be learned, and the threat of an infinite regress must also be confronted — what are the inductive constraints that allow inductive constraints to be learned? Connectionist models have been able to overcome some of these challenges: Samuelson (2002) and Colunga and Smith (2005) have developed models that acquire the shape bias, and Rogers and McClelland (2004) suggest that connectionist models can acquire many other kinds of abstract knowledge. The connectionist approach, however, aims to provide a mechanistic account of learning, and is not ideal for explaining the computational principles that support the acquisition of overhypotheses. Connectionist models, for instance, do not clearly distinguish between knowledge at different levels of abstraction, and it is difficult to analyze a successful model and decide which overhypotheses are responsible for its success, and how they might have been acquired.

This paper suggests suggest that hierarchical Bayesian models (Good, 1980; Gelman, Carlin, Stern, & Rubin, 2003) can help to explain the computational principles which allow overhypotheses to be learned. Hierarchical Bayesian models (HBMs) include representations at multiple levels of abstraction, and show how knowledge can be acquired at levels quite remote from the data given by experience. To illustrate these points, we describe one of the simplest possible HBMs and use it to suggest how overhypotheses about feature-variability (e.g. the shape bias) are acquired and used to support categorization. We also present an extension of the basic model that groups categories into ontological kinds (e.g. objects and substances) and discovers the features and the patterns of feature variability that are characteristic of each kind.

The hierarchical Bayesian approach shows how knowledge can be simultaneously

acquired at multiple levels of abstraction, and may help to reconcile two competing approaches to cognitive development. The bottom-up approach suggests that concrete knowledge is available before abstract knowledge, and that abstract knowledge is acquired by generalizing over instances of concrete knowledge (Piaget & Inhelder, 1969; Smith et al., 2002; Karmiloff-Smith, 1992). One instance of this approach is the claim that perceptual categories are acquired before more abstract categories (Cohen, 1998; French, Mareschal, Mermillod, & Quinn, 2004). An alternative approach — the top-down approach — suggests that abstract knowledge is sometimes available before more concrete knowledge is securely in place (Keil, 1998; Mandler, 2003). Mandler and McDonough (1993), for example, argue that infants form global categories like “animal” and “vehicle” before basic-level categories like “rabbit” and “car.” The hierarchical Bayesian approach provides a unifying framework that accommodates both top-down and bottom-up learning. More precisely, HBMs support the simultaneous acquisition of abstract and concrete knowledge, and allow for three possibilities: depending on the task and the data set, learning may be significantly faster for concrete knowledge than for abstract knowledge, about equally rapid for both kinds of knowledge, or significantly faster for abstract knowledge than for concrete knowledge. We use the simple model we develop to provide examples of all three cases.

Overhypotheses and HBMs

Goodman introduces the notion of an overhypothesis with an example based on bags of colored marbles (Goodman, 1955). Suppose that S is a stack containing many bags of marbles. We empty several bags and discover that some bags contain black marbles, others contain white marbles, but that the marbles in each bag are uniform in color. We now choose a new bag — bag n — and draw a single black marble from the bag. On its own, a single draw would provide little information about the contents of the new bag, but

experience with previous bags may lead us to endorse the following hypothesis:

H: All marbles in bag n are black.

If asked to justify the hypothesis, we might invoke the following overhypothesis:

O: Each bag in stack S contains marbles that are uniform in color.

Goodman gives a precise definition of “overhypothesis” but we use the term more generally to refer to any form of abstract knowledge that sets up a hypothesis space at a less abstract level. By this criterion, *O* is an overhypothesis since it sets up a space of hypotheses about the marbles in bag n : they could be uniformly black, uniformly white, uniformly green, and so on.

Insert Figure 1 about here

Hierarchical Bayesian models (Gelman et al., 2003) capture this notion of overhypothesis by allowing hypothesis spaces at several levels of abstraction. We give an informal introduction to this modelling approach, leaving all technical details for the next section. Suppose that we wish to explain how a certain kind of inference can be drawn from a given body of data. In Goodman’s case, the data are observations of several bags (\mathbf{y}^i indicates the observations for bag i) and we are interested in the ability to predict the color of the next marble to be drawn from bag n (Figure 1a). The first step is to identify a kind of knowledge (level 1 knowledge) that explains the data and that supports the ability of interest. In Goodman’s case, level 1 knowledge is knowledge about the color distribution of each bag (θ^i indicates the color distribution for the i th bag).

We then ask how the level 1 knowledge might be acquired, and the answer will make reference to a more abstract body of knowledge (level 2 knowledge). For the marbles

scenario, level 2 knowledge is knowledge about the distribution of the θ variables. As described in the next section, this knowledge can be represented using two parameters, α and β (Figure 1a). Roughly speaking, α captures the extent to which the marbles in each individual bag are uniform in color, and β captures the average color distribution across the entire stack of bags. If we now go on to ask how the level 2 knowledge might be acquired, the answer will rely on a body of knowledge at an even higher level, level 3. In Figure 1a, this knowledge is represented by λ , which captures prior knowledge about the values of α and β . The parameter λ and the pair (α, β) are both overhypotheses, since each sets up a hypothesis space at the next level down. We will assume that the level 3 knowledge is specified in advance, and show how an overhypothesis can be learned at level 2.

Within cognitive science, linguists have provided the most familiar example of this style of model building. Language comprehension can be explained using parse trees for individual sentences (level 1 knowledge). Parse trees, in turn, can be explained with reference to a grammar (level 2 knowledge), and the acquisition of this grammar can be explained with reference to Universal Grammar (level 3 knowledge). There are few settings where cognitive scientists have discussed more than three levels, but there is no principled reason to stop at level 3. Ideally, we should continue adding levels until the knowledge at the highest level is simple enough or general enough that it can be plausibly assumed to be innate.

As the grammar-learning example suggests, it has long been known that hierarchical models are capable in principle of explaining the acquisition of overhypotheses. The value of hierarchical *Bayesian* models (HBMs) is that they explain how overhypotheses can be acquired by rational statistical inference. Given observations at the lowest level of a HBM, statistical inference can be used to compute posterior distributions over entities at the higher levels. In the model of Figure 1a, for instance, acquiring an overhypothesis is a

matter of acquiring knowledge at level 2. The posterior distribution $p(\alpha, \beta | \mathbf{y})$ represents a normative belief about level 2 knowledge — the belief, given the data \mathbf{y} , that the marbles in each bag are close to uniform in color.

At the level of computational theory, the problem of overhypothesis acquisition can be reduced to a search problem. Given a space of possible overhypotheses, the learning problem can be solved by searching for the candidate (e.g. the pair (α, β)) with maximum posterior probability. The claim that the set of candidates is known in advance may seem inconsistent with the intuition that the repertoire of a learner can grow over time. Some formal models appear to capture this intuition: for example, the hypothesis space of a constructivist neural network (Fahlman & Lebiere, 1990; Shultz, 2003) appears to grow when new units are added. The apparent inconsistency here rests on a mismatch between two possible definitions of “hypothesis space.” At the level of computational theory, a hypothesis space represents the abstract potential of a learning system: if one imagines all possible streams of input that a learning system could receive, the hypothesis space includes all states of knowledge which the system could possibly reach. For a constructivist neural network, this notion of a hypothesis space includes all configurations that could be reached by adding new units. The second possible definition is often used when describing a process model. In this context, “hypothesis space” often refers to the set of hypotheses that are currently entertained by the model, and this set will usually change over time. Since our goal is to develop computational theories, we will work with the abstract, static definition of “hypothesis space,” but it is important to note that our theories have many possible implementations, some of which appear just as dynamic and as flexible as constructivist neural networks. For example, one of our theories (Figure 1b) has implementations that can “grow” by introducing new ontological kinds when new data are observed.

A computational theory of feature variability

We now describe one formal instantiation of the model in Figure 1a. There may be other ways to formalize overhypotheses about feature variability, but ours is perhaps the simplest account of how these overhypotheses can be acquired and simultaneously used to guide learning at lower levels. Suppose we are working with a set of k colors. Initially we set $k = 2$ and use white and black as the colors. Let θ^i indicate the true color distribution for the i th bag in the stack: if 60% of the marbles in bag 7 are black, then $\theta^7 = [0.4, 0.6]$. Let y^i indicate a set of observations of the marbles in bag i . If we have drawn 5 marbles from bag 7 and all but one are black, then $y^7 = [1, 4]$.

Insert Figure 2 about here

We assume that y^i is drawn from a multinomial distribution with parameter θ^i : in other words, the marbles responsible for the observations in y^i are drawn independently at random from the i th bag, and the color of each depends on the color distribution θ^i for that bag. The vectors θ^i are drawn from a Dirichlet distribution parameterized by a scalar α and a vector β . The parameter α determines the extent to which the colors in each bag tend to be uniform, and β represents the distribution of colors across the entire collection of bags (Figure 2). Each possible setting of (α, β) is an overhypothesis, and to discover which of these settings is best we need to formalize our *a priori* expectations about the values of these variables. We use a uniform distribution on β and an exponential distribution on α , which captures a weak prior expectation that the marbles in any bag will tend to be uniform in color (Figure 3a.i). The mean of the exponential distribution is λ , and each possible setting of λ is an over-overhypothesis (Figure 1a), or an overhypothesis one level higher than the level of (α, β) . The qualitative predictions of

our model are relatively insensitive to changes in λ , and all simulations described in this paper use $\lambda = 1$.

Insert Figure 3 about here

So far, we have assumed we are working with a single dimension — for Goodman, marble color. Suppose, however, that some marbles are made from metal and others are made from glass, and we are interested in material as well as color. A simple way to deal with multiple dimensions is to assume that each dimension is independently generated, and to introduce separate values of α and β for each dimension. When working with multiple features, we often use α to refer to the collection of α values along all dimensions, β for the set of all β vectors, and \mathbf{y} for the set of counts along all dimensions.

To fit the model to data we assume that counts \mathbf{y} are observed for one or more bags. Our goal is to compute the posterior distribution $p(\alpha, \beta, \{\theta^i\} | \mathbf{y})$: in other words, we wish to simultaneously discover level 2 knowledge about α and β and level 1 knowledge about the color distribution θ^i of each individual bag i . As described in Appendix A, inferences about α and β can be made by drawing a sample from $p(\alpha, \beta | \mathbf{y})$, the posterior distribution on (α, β) given the observed data. Figures 3a.ii and 3a.iii show posterior distributions on $\log(\alpha)$ and β for two sets of counts. Inferences about θ^i , the color distribution of bag i , can be computed by calculating the mean prediction made by all pairs (α, β) in the sample. Note that this inference scheme is merely a convenient way of computing the predictions of our computational theory. Any computational theory can be implemented in many ways, and the particular implementation we have chosen is not intended as a process model.

Modelling inductive reasoning

Since Goodman, psychologists have confirmed that children (Macario, Shipley, & Billman, 1990) and adults (Nisbett, Krantz, Jepson, & Kunda, 1983) have overhypotheses about feature variability, and use them to make inductive leaps given very sparse data. We provide an initial demonstration of our model using data inspired by one of the tasks of Nisbett et al. (1983). As part of this task, participants were asked to imagine that they were exploring an island in the Southeastern Pacific, that that they had encountered a single member of the Barratos tribe, and that this tribesman was brown and obese. Based on this single example, participants concluded that most Barratos were brown, but gave a much lower estimate of the proportion of obese Barratos (Figure 4). When asked to justify their responses, participants often said that tribespeople were homogeneous with respect to color but heterogeneous with respect to body weight (Nisbett et al., 1983).

Insert Figure 4 about here

To apply our model to this task, we replace bags of marbles with tribes. Suppose we have observed 20 members from each of 20 tribes. Half the tribes are brown and the other half are white, but all of the individuals in a given tribe have the same skin color. Given these data, the posterior distribution on α indicates that skin color tends to be homogenous within tribes (i.e. α is probably small) (Figure 3a.ii). Learning that α is small allows the model to make strong predictions about a sparsely observed new tribe: having observed a single, brown-skinned member of a new tribe, the posterior distribution on θ^{new} indicates that most members of the tribe are likely to be brown (Figures 3a.ii and 4). Note that the posterior distribution on θ^{new} is almost as sharply peaked as the posterior distribution on θ^{11} : the model has realized that observing one member of a new tribe is almost as informative as observing 20 members of that tribe.

Suppose now that obesity is a feature that varies within tribes: a quarter of the 20 tribes observed have an obesity rate of 10%, and the remaining quarters have rates of 20%, 30%, and 40%. Obesity is represented in our model as a second binary feature, and the posterior distributions on α and β (Figure 3a.iii) indicate that obesity varies within tribes (α is high), and that the base rate of obesity is around 25% (β_2 is around 0.25). Again, we can use these posterior distributions to make predictions about a new tribe, and now the model requires many observations before it concludes that most members of the new tribe are obese (Figure 4). Unlike the case in Figure 3a.ii, the model has learned that a single observation of a new tribe is not very informative, and the distribution on θ^{new} is now similar to the average of the θ distributions for all previously observed tribes.

Accurate predictions about a new tribe depend critically on learning at both level 1 and level 2 (Figure 1a). Learning at level 1 is needed to incorporate the observation that the new tribe has at least one obese, brown-skinned member. Learning at level 2 is needed to discover that skin color is homogeneous within tribes but that obesity is not, and to discover the average rate of obesity across many tribes. Figure 3b shows inferences drawn by an alternative model that is unable to discover overhypotheses — instead, we fix α and β to their expected values under the prior distributions used by our model. Since it cannot learn at level 2, this alternative model cannot incorporate any information about the 20 previous tribes when reasoning about a new tribe. As a result, it makes identical inferences about skin color and obesity — note that the distribution on θ^{new} is the same in Figures 3b.ii and 3b.iii. Note also that the mean of this distribution (0.75) is lower than the mean of the distribution in Figure 3a.ii (0.99) — both models predict that most members of the new tribe have brown skin, but our model alone accounts for the human judgment that almost all members of the new tribe have brown skin (Figure 4).

Learning the shape bias

The Barratos task does not address an important kind of reasoning that overhypotheses support: reasoning about *new* feature values along known dimensions. Based on the data in Figure 1a, a learner could acquire at least two different overhypotheses: the first states that the marbles in each bag are uniform along the dimension of color, and the second states that the marbles in each bag are either all white or all black. One way to distinguish between these possibilities is to show the learner that a single green marble is drawn from the new bag. A learner with the first overhypothesis will predict that all marbles in the new bag are green, but a learner with the second overhypothesis will be lost.

There are many real-world problems that involve inferences about novel features. Children know, for example, that animals of the same species tend to make the same sound. Observing one horse neigh is enough to conclude that most horses neigh, even though a child may never have heard an animal neigh before (Shipley, 1993). Similarly, by the age of $2\frac{1}{2}$ children show a “shape bias:” they know that shape tends to be homogeneous within object categories. Given a single exemplar of a novel object category, children extend the category label to similarly shaped objects ahead of objects that share the same texture or color as the exemplar (Heibeck & Markman, 1987; Landau, Smith, & Jones, 1988).

Insert Figure 5 about here

The model in Figure 1a deals naturally with inferences like these. We illustrate using stimuli inspired by the work of Smith et al. (2002). In their first experiment, these authors trained 17-month olds on two exemplars from each of four novel categories. Novel

names (e.g. “zup”) were provided for each category, and the experimenter used phrases like “this is a zup — let’s put the zups in the wagon.” Within each category, the two exemplars had the same shape but differed in size, texture and color (Figure 5a). After eight weeks of training, the authors tested *first-order* generalization by presenting T_1 , an exemplar from one of the training categories, and asking children to choose another object from the same category as T_1 . Three choice objects were provided, each of which matched T_1 in exactly one feature (shape, color or size) (Figure 5b). Children preferred the shape match, showing that they were sensitive to feature distributions within a known category. Smith et al. (2002) also tested *second-order* generalization by presenting children with T_2 , an exemplar from a novel category (Figure 5c). Again, children preferred the shape match, revealing knowledge that shape in general is a reliable indicator of category membership. Note that this result depends critically on the training summarized by Figure 5a:

19-month olds do not normally reveal a shape bias on tests of second-order generalization.

We supplied our model with counts \mathbf{y}^i computed from the feature vectors in Figure 5a. For example, \mathbf{y}^1 indicates that the data for category 1 include two observations of shape value 1, one observation of texture value 1, one observation of texture value 2, and so on. The key modelling step is to allow for more values along each dimension than appear in the training set. This policy allows the model to handle shapes, colors and textures it has never seen during training, but assumes that the model is able to recognize a novel shape as a kind of shape, a novel color as a kind of color, and so on. We allowed for ten shapes, ten colors, ten textures and two sizes: for example, the shape component of \mathbf{y}^1 indicates that the observed exemplars of category 1 include two objects with shape value 1 and no objects with shape values 2 through 10.

Figure 5d shows the patterns of generalization predicted by the model. Smith et al. (2002) report that the shape match was chosen 88% (66%) of the time in the test of first-order generalization, and 70% (65%) of the time in the second-order task (percentages

in parentheses represent results when the task was replicated as part of Experiment 2). Our model reproduces this general pattern: shape matches are preferred in both cases, and are preferred slightly more strongly in the test of first-order generalization.

Insert Figure 6 about here

Smith et al. (2002) also measured real-world generalization by tracking vocabulary growth over an eight week period. They report that experience with the eight exemplars in Figure 5a led to a significant increase in the number of object names used by children. Our model helps to explain this striking result. Even though the training set includes only four categories, the results in Figure 5b show that it contains enough statistical information to establish or reinforce the shape bias, which can then support word learning in the real world. Similarly, our model explains why providing only two exemplars per category is sufficient. In fact, if the total number of exemplars is fixed, our model predicts that the best way to teach the shape bias is to provide just two exemplars per category. We illustrate by returning to the marbles scenario.

Each point in Figure 6a represents a simulation where 64 observations of marbles are evenly distributed over some number of bags. The marbles drawn from any given bag are uniform in color — black for half of the bags and white for the others. When 32 observations are provided for each of two bags (Figure 6b.i), the model is relatively certain about the color distributions of those bags, but cannot draw strong conclusions about the homogeneity of bags in general. When two observations are provided for each of 32 bags, (Figure 6b.ii), the evidence about the composition of any single bag is weaker, but taken together, these observations provide strong support for the idea that α is low and most bags are homogeneous. When just one observation is provided for each of 64 bags, the model has no information about color variability within bags, and the posterior

distribution on α is identical to the prior on α , which has a mean value of 1. If the total number of observations is fixed, Figure 6a suggests that the best way to teach a learner that bags are homogeneous in general is to provide two observations for as many bags as possible. The U-shaped curve in Figure 6a is a novel prediction of our model, and could be tested in developmental experiments.

At least three outcomes are possible when learning proceeds in parallel at levels 1 and 2. Figure 6b.i is a case where the learner is more confident about level 1 knowledge than level 2 knowledge: note that the distributions for the two individual bags (θ^1 and θ^2) are more tightly peaked than the distributions on α and β , which capture knowledge about bags in general. Figure 6b.ii is a case where the learner is relatively confident about the values of the variables at both levels. Figure 6b.iii is a case where the learner is more confident about level 2 than level 1. In this case, two observations are provided for each of 32 bags: 22 of the observed pairs are mixed, and there are 5 white pairs and 5 black pairs. The model is now relatively uncertain about the color distribution of any individual bag, but relatively certain about the values of α and β . A second prediction of our model, then, is that learning can sometimes be faster at level 2 than at level 1. This prediction distinguishes the hierarchical Bayesian approach from bottom-up approaches to learning overhypotheses (e.g. the four-step method of Smith et al. (2002)), which suggest that some of the variables at level 1 must be securely known before learning can take place at level 2.

Although our model provides some insight into the findings of Smith et al. (2002), it does not account for all of their results. Their second experiment includes a *no-name* condition where children received the same training as before (Figure 5a) except that category labels were not provided. Instead of naming the training objects, the experimenter used phrases like “here is one, here is another — let’s put them both in the wagon.” Children in this condition showed first-order but not second-order generalization, which supports the view that the shape bias reflects attention to shape in the context of

naming (Smith, Jones, & Landau, 1996). An alternative view is that the shape bias is not specifically linguistic: shape is important not because it is linked to naming in particular, but because it is a reliable cue to category membership (Ward, Becker, Hass, & Vela, 1991; Bloom, 2000). Our model is consistent with this second view, and predicts that learning in the no-name condition should not have been impaired provided that children clearly understood which training objects belonged to the same category. This discrepancy between model predictions and empirical results calls for further work on both sides. On the modelling side, it is important to develop hierarchical models that allow an explicit and privileged role for linguistic information. On the empirical side, it seems possible that children in the no-name condition did not achieve second-order generalization because they did not realize that each pair of identically-shaped objects was supposed to represent a coherent category.² Observing associations between similarly-shaped objects may have led them only to conclude that shape was a salient feature of each of these objects, which would have been enough for them to pass the test of first-order generalization.

Insert Figure 7 about here

Discovering ontological kinds

The model in Figure 1a is a simple hierarchical model that acquires something like the shape bias, but to match the capacities of a child it is necessary to apply the shape bias selectively — to object categories, for example, but not to substance categories. Selective application of the shape bias appears to demand knowledge that categories are grouped into ontological kinds and that there are different patterns of feature variability within each kind. Before the age of three, for instance, children appear to know that shape tends to be homogeneous within object categories but heterogeneous within

substance categories (Soja, Carey, & Spelke, 1991; Imai, Gentner, & Uchida, 1994; Samuelson & Smith, 1999), that color tends to be homogeneous within substance categories but heterogeneous within object categories (Landau et al., 1988; Soja et al., 1991), and that both shape and texture tend to be homogeneous within animate categories (Jones, Smith, & Landau, 1991).

Figure 1b shows how we can give our model the ability to discover ontological kinds. The model assumes that categories *may* be grouped into several ontological kinds, and that there is a separate α^k and β^k for each ontological kind k . The model, however, is not told which categories belong to the same kind, and is not even told how many different kinds it should look for. Instead, we give it a prior distribution on the partition of categories into kinds (see Appendix A). This prior assigns some probability to all possible category partitions, but favors the simpler partitions — those that use a small number of kinds. To fit the model to data, we again assume that feature counts \mathbf{y} are observed for one or more categories. Our goal is to simultaneously infer the partition of categories into kinds, along with the α^k and β^k for each kind k and the feature distribution θ^i for each category. Appendix A describes how these inferences can be carried out.

Jones and Smith (2002) have shown that training young children on a handful of suitably structured categories can promote the acquisition of ontological knowledge. We gave our model a data set of comparable size. During training, the model saw two exemplars from each of four categories: two object categories and two substance categories (Figure 7a). Exemplars of each object category were solid, matched in shape, and differed in material and size. Exemplars of each substance category were non-solid, matched in material, and differed in shape and size. Second-order generalization was tested using exemplars from novel categories — one test exemplar (S) was solid and the other (N) was not (Figure 7b). Figure 7c shows that the model chooses a shape match for the solid exemplar and a material match for the non-solid exemplar.

Figure 7d confirms that the model correctly groups the stimuli into two ontological kinds: object categories and substance categories. This discovery is based on the characteristic features of ontological kinds (β) as well as patterns of feature variability within each kind (α). If the object categories are grouped into kind k , α^k indicates that shape is homogeneous within categories of that kind, and β^k indicates that categories of that kind tend to be solid. The β parameter, then, is responsible for the inference that the category including S should be grouped with the two object categories, since all three categories contain solid objects.

The results in Figure 7 predict that a training regime with a small number of categories and exemplars should allow children to simultaneously acquire a shape bias for solids and a material bias for substances. Samuelson (2002) ran a related study where she attempted to teach one group of children a precocious shape bias and another a precocious material bias. Only the shape bias was learned, suggesting that the shape bias is easier to teach than the material bias, but leaving open the possibility that the material bias could have been acquired with more training. Simultaneously teaching a shape bias for solids and a material bias for substances may raise some difficult practical challenges, but Jones and Smith (2002) have shown that children can simultaneously learn two kind-specific biases. By the end of their training study, children had learned that names for animate exemplars (exemplars with eyes) should be generalized according to shape and texture, and that names for objects (exemplars without eyes) should be generalized only according to shape. Our model accounts for these results: given the data provided to the children in these experiments, it discovers that there are two ontological kinds, and makes selective generalizations depending on whether or not a novel exemplar has eyes.

Related models

Our models address tasks that have been previously modelled by Colunga and Smith (2005). These authors develop a connectionist network that acquires a shape bias for solid objects and a material bias for non-solid objects. The network uses a set of hidden nodes to capture high-order correlations between nodes representing the shape, material, and solidity of a collection of training objects, and generates results similar to Figure 7c when asked to make predictions about novel objects. Our model is similar to this connectionist model in several respects: both models show that abstract knowledge can be acquired, and both models are statistical, which allows them to deal with noise and uncertainty and to make graded generalizations. These models, however, differ in at least two important respects.

First, the two models aim to provide different kinds of explanations. Our contribution is entirely at the level of computational theory (Marr, 1982), and we have not attempted to specify the psychological mechanisms by which our model might be implemented. Colunga and Smith (2005) describe a process model that uses a biologically-inspired learning algorithm, but provide no formal description of the problem to be solved. Their network can probably be viewed as an approximate implementation of some computational theory,³ but the underlying computational theory may not be ideal for the problem of word learning. For instance, it is not clear that the network adequately captures the notion of a category. In tests of second-order generalization (e.g. Figure 7c), our model is able to compute the probability that a choice object belongs to the same category as the test exemplar. Colunga and Smith (2005) compute model predictions by comparing the similarity between hidden-layer activations for the choice object and the test exemplar. Objects in the same category may often turn out to have similar representations, but there are some well-known cases where similarity and categorization diverge (Keil, 1989; Rips, 1989).

A second limitation of the connectionist approach is that it does not extend naturally to contexts where structured representations are required. We defined models that generate scalars (α) and vectors ($\beta, \theta, \mathbf{y}$), but hierarchical probabilistic models can generate many other kinds of representations, including taxonomies (Kemp, Perfors, & Tenenbaum, 2004), ontologies (Schmidt, Kemp, & Tenenbaum, in press), causal networks (Mansinghka, Kemp, Tenenbaum, & Griffiths, in press), parse trees (Perfors, Tenenbaum, & Regier, in press), and logical theories (Milch et al., 2005). Many overhypotheses correspond to constraints on structured representations: for example, the M-constraint states that ontological knowledge is better described by a tree structure than by a set of arbitrarily overlapping clusters (Keil, 1979), and Universal Grammar may include many overhypotheses that constrain the structure of possible grammars. Hierarchical Bayesian models may eventually explain how overhypotheses like these might be acquired, and Schmidt et al. (in press) and Perfors et al. (in press) describe some initial steps towards this goal.

Previous researchers have developed Bayesian models of categorization (Anderson, 1991) and word learning (Tenenbaum & Xu, 2000), and our work continues in this tradition. The hierarchical approach, however, attempts to address a problem raised by most Bayesian models of cognition. In terms of our hierarchical framework, a conventional Bayesian model incorporates two levels of knowledge: the elements in its hypothesis space represent level 1 knowledge, and the prior (generally fixed) represents knowledge at level 2. One common reservation about Bayesian models is that different priors account for different patterns of data, and the success of any given Bayesian model depends critically on the modeller's ability to choose the right prior. Hierarchical models disarm this objection by showing that knowledge at level 2 need not be specified in advance, but can be learned from raw data.

Hierarchical Bayesian models (HBMs) still rely on some prior knowledge, since the

prior at the highest level must be specified in advance. The ultimate goal, however, is to design models where this prior is simple enough to be unobjectionable. Our models demonstrate that HBMs can sometimes rely on much simpler priors than conventional Bayesian models. If we were only interested in inferences about level 1 knowledge (inferences about the θ^i for each bag i), α and β (Figure 1a) would not be essential: in other words, a conventional Bayesian model could mimic the predictions of our model if it used the right prior distribution on the set $\{\theta^i\}$. If specified directly, however, this prior would look extremely complicated — much more complicated, for example, than the prior $\{\theta^i\}$ used by the conventional model in Figure 3b, which assumes that all the θ^i are independent. We avoided this problem by specifying the prior on $\{\theta^i\}$ *indirectly*. We introduced an extra layer of abstraction — the layer including α and β — and placed simple priors on these variables. These simple distributions on α and β induce a complicated prior distribution on $\{\theta^i\}$ — the same distribution that a conventional Bayesian model would have to specify directly.

Beyond feature variability

In order to establish the generality of the hierarchical Bayesian approach to development, it will be necessary to develop HBMs that account for the acquisition of overhypotheses in many different domains. Much work remains to be done, but HBMs have recently been applied to several different problems. Perfors et al. (in press) present a hierarchical model that discovers whether a corpus of child-directed speech is better described by a regular grammar or a context-free grammar. Discovering abstract properties of the underlying grammar may help language learners zero in on a specific grammar that accounts well for the data they have observed. The M-constraint (Keil, 1979) may help children learn which entities and predicates can be sensibly paired, and Schmidt et al. (in press) have argued that this constraint may be learnable from raw data.

Finally, knowledge about causal types (e.g. diseases and symptoms) and relationships between these types (diseases cause symptoms) places useful constraints on causal learning: for instance, a learner need not consider causal networks that state that lung cancer causes smoking. Mansinghka et al. (in press) describe a HBM that uses raw co-occurrence data to discover abstract knowledge about causal types.

Future work should also explore the ability of HBMs to learn simultaneously at several levels of abstraction. Figure 6b illustrates three possible patterns of learning: the third makes the surprising prediction that abstract knowledge can sometimes be acquired faster than knowledge at lower levels of abstraction. When abstract knowledge is available very early in development, a natural conclusion is that the knowledge is innate. HBMs suggest a possible alternative: in some cases, abstract knowledge may appear to be innate only because it is acquired much faster than knowledge at lower levels of abstraction. This possibility may apply in situations where a child has access to a large number of sparse or noisy observations — any individual observation may be difficult to interpret, but taken together they may provide strong support for a general conclusion. For example, a hierarchical Bayesian model of grammar induction may be able to explain how a child becomes confident about some property of a grammar even though most of the individual sentences that support this conclusion are poorly understood. Similarly, a hierarchical approach may explain how a child can learn that visual objects are cohesive, bounded and rigid (cf. Spelke (1990)) before developing a detailed understanding of the appearance and motion of any individual object.

Conclusion

The hierarchical Bayesian approach is familiar to statisticians (Good, 1980), but is just beginning to be explored as a framework for modelling human learning and reasoning (Tenenbaum, Griffiths, & Kemp, in press; Lee, 2006). Ultimately, hierarchical

Bayesian models may help to explain the acquisition of overhypotheses across a broad range of domains. We focused on overhypotheses about feature variability, and presented HBMs that help explain the acquisition of the shape bias, and the acquisition of overhypotheses about feature variability within ontological kinds.

Although we suggested that overhypotheses can be learned by HBMs, we do not claim that overhypotheses can be generated out of thin air. Any HBM will assume that the process by which each level is generated from the level above is known, and that the prior at the topmost level is provided. Any account of induction must rely on *some* initial knowledge: the real question for a learning framework is whether it allows us to build models that require no initial assumptions beyond those we are willing to make. Whether the hierarchical Bayesian approach will meet this challenge is not yet clear, but it deserves to be put to the test.

Appendix A

The model in Figure 1a is known to statisticians as a Dirichlet-multinomial model (Gelman et al., 2003). Using statistical notation, it can be written as:

$$\alpha \sim \text{Exponential}(1)$$

$$\beta \sim \text{Dirichlet}(\mathbf{1})$$

$$\theta^i \sim \text{Dirichlet}(\alpha, \beta)$$

$$\mathbf{y}^i | n^i \sim \text{Multinomial}(\theta^i)$$

where n^i is the number of observations for bag i .

To compute the predictions of this model, we estimate $p(\alpha, \beta | \mathbf{y})$ using numerical integration or a Markov chain Monte Carlo (MCMC) scheme. Inferences about the θ^i are computed by integrating out α and β :

$$p(\theta^i | \mathbf{y}) = \int_{\alpha, \beta} p(\theta^i | \alpha, \beta, \mathbf{y}) p(\alpha, \beta | \mathbf{y}) d\alpha d\beta$$

Our MCMC sampler uses Gaussian proposals on $\log(\alpha)$, and proposals for β are drawn from a Dirichlet distribution with the current β as its mean. The results in Figure 5 represent averages across 30 Markov chains, each of which was run for 50,000 iterations (1000 were discarded as burn-in). The model predictions in Figures 3, 4 and 6 were computed using numerical integration.

The model in Figure 1b partitions the categories into one or more ontological kinds. Each possible partition can be represented by a vector \mathbf{z} : the partition in 1b is represented by the vector $[1, 1, 1, 2, 2, 2]$ which indicates that the first three categories belong to one ontological kind, and the remaining three belong to a second kind. Our prior distribution on \mathbf{z} is induced by the Chinese Restaurant Process (CRP, Aldous, 1985):

$$p(z_i = a | z_1, \dots, z_{i-1}) = \begin{cases} \frac{n_a}{i-1+\gamma} & n_a > 0 \\ \frac{\gamma}{i-1+\gamma} & a \text{ is a new kind} \end{cases}$$

where z_i is the kind assignment for category i , n_a is the number of categories previously assigned to kind a , and γ is a hyperparameter (we set $\gamma = 0.5$). This process prefers to assign new categories to kinds which already have many members, and therefore favors partitions that use a small number of kinds.

The entire model in Figure 1b can be written as follows:

$$\begin{aligned} \mathbf{z} &\sim \text{CRP}(\gamma) \\ \alpha^k &\sim \text{Exponential}(\lambda) \\ \beta^k &\sim \text{Dirichlet}(\mathbf{1}) \\ \theta^i &\sim \text{Dirichlet}(\alpha^{z_i} \beta^{z_i}) \\ \mathbf{y}^i | n^i &\sim \text{Multinomial}(\theta^i) \end{aligned}$$

If \mathbf{z} is known, this model reduces to several independent copies of the model in Figure 1a, and model predictions (including $p(\theta^i | \mathbf{z}, \mathbf{y})$) can be computed using the

techniques already described. Since \mathbf{z} is unknown, we integrate over this quantity:

$$p(\boldsymbol{\theta}^i|\mathbf{y}) = \sum_{\mathbf{z}} p(\boldsymbol{\theta}^i|\mathbf{z}, \mathbf{y})P(\mathbf{z}|\mathbf{y}).$$

To compute this sum we use $P(\mathbf{z}|\mathbf{y}) \propto P(\mathbf{y}|\mathbf{z})P(\mathbf{z})$, where $P(\mathbf{z})$ is the CRP prior on \mathbf{z} .

Computing $P(\mathbf{y}|\mathbf{z})$ reduces to the problem of computing several marginal likelihoods

$$P(\mathbf{y}^l) = \int_{\alpha, \boldsymbol{\beta}} P(\mathbf{y}^l|\alpha, \boldsymbol{\beta})p(\alpha, \boldsymbol{\beta})d\alpha d\boldsymbol{\beta}$$

for the model in Figure 1a. We estimate each of these integrals by drawing 10,000 samples from the prior $p(\alpha, \boldsymbol{\beta})$.

References

- Ackley, D., Hinton, G., & Sejnowski, T. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*, 147–169.
- Aldous, D. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983* (pp. 1–198). Berlin: Springer.
- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.
- Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.
- Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and reports on child language development*, *15*, 17–29.
- Chomsky, N. (1980). *Rules and representations*. Oxford: Basil Blackwell.
- Cohen, L. B. (1998). An information-processing approach to infant perception and cognition. In F. Simion & G. Butterworth (Eds.), *The development of sensory, motor and cognitive capacities in early infancy*. Hove, UK: Psychology Press.
- Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychological Review*, *112*(2).
- Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems 2* (pp. 524–532). San Francisco, CA: Morgan Kaufmann.
- French, R. M., Mareschal, D., Mermillod, M., & Quinn, P. C. (2004). The role of bottom-up processing in perceptual categorization by 3 to 4-month old infants: simulations and data. *Journal of Experimental Psychology: General*, *133*(3), 382–397.

- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall.
- Goldstone, R. L., & Johansen, M. K. (2003). Conceptual development from origins to asymptotes. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development* (pp. 403–418). New York: Oxford University Press.
- Good, I. J. (1980). Some history of the hierarchical Bayesian methodology. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 489–519). Valencia: Valencia University Press.
- Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge: Harvard University Press.
- Heibeck, T., & Markman, E. (1987). Word learning in children: an examination of fast mapping. *Child Development*, *58*, 1021–1024.
- Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: the role of shape similarity in early acquisition. *Cognitive Development*, *9*, 45–76.
- Jones, S. S., & Smith, L. B. (2002). How children know the relevant properties for generalizing object names. *Developmental Science*, *5*(2), 219–232.
- Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development*, *62*, 499–516.
- Jusczyk, P. W. (2003). Chunking language input to find patterns. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development* (pp. 17–49). New York: Oxford University Press.
- Karmiloff-Smith, A. (1992). *Beyond modularity*. Cambridge, MA: MIT Press.
- Keil, F. C. (1979). *Semantic and conceptual development*. Cambridge, MA: Harvard University Press.

- Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.
- Keil, F. C. (1998). Cognitive science and the origins of thought and knowledge. In R. M. Lerner (Ed.), *Theoretical models of human development* (Vol. I). New York: Wiley.
- Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual conference of the cognitive science society* (pp. 672–678). Lawrence Erlbaum Associates.
- Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*, 299–321.
- Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science, 30*, 555–580.
- Macario, J. F., Shipley, E. F., & Billman, D. O. (1990). Induction from a single instance: formation of a novel category. *Journal of Experimental Child Psychology, 50*, 179–199.
- Madole, K. L., & Cohen, L. B. (1995). The role of object parts in infants' attention to form-function correlations. *Developmental Psychology, 31* (4), 637–648.
- Mandler, J. M. (2003). Conceptual categorization. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development* (pp. 103–131). New York: Oxford University Press.
- Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. *Cognitive Development, 8*, 291–318.
- Mansinghka, V. K., Kemp, C., Tenenbaum, J. B., & Griffiths, T. L. (in press). Structured priors for structure learning. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*.

- Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.
- Medin, D., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge: Cambridge University Press.
- Milch, B., Marthi, B., Russell, S., Sontag, D., Ong, D. L., & Kolobov, A. (2005). BLOG: Probabilistic models with unknown objects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (pp. 1352–1359).
- Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339–363.
- Perfors, A., Tenenbaum, J. B., & Regier, T. (in press). Poverty of the stimulus? A rational approach. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.
- Piaget, J., & Inhelder, B. (1969). *The psychology of the child*. New York: Basic Books.
- Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.
- Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (p. 21–59). Cambridge: Cambridge University Press.
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: a Parallel Distributed Processing approach*. MIT Press.
- Samuelson, L. (2002). Statistical regularities in vocabulary guide language acquisition in connectionist models and 15–20 month olds. *Developmental Psychology*, *38*(6), 1016–1037.
- Samuelson, L., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition*, *73*, 1–33.

- Schmidt, L. A., Kemp, C., & Tenenbaum, J. B. (in press). Nonsense and sensibility: Discovering unseen possibilities. In *Proceedings of the 28th Annual Conference of the Cognitive Science society*.
- Shipley, E. F. (1993). Categories, hierarchies and induction. *Psychology of Learning and Motivation, 30*, 265–301.
- Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT Press.
- Smith, L. B., Jones, S. S., & Landau, B. (1996). Naming in young children: a dumb attentional mechanism? *Cognition, 60*, 143–171.
- Smith, L. B., Jones, S. 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.
- Soja, N., Carey, S., & Spelke, E. (1991). Ontological categories guide young children's inductions of word meanings. *Cognition, 38*, 179–211.
- Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14*, 29–56.
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (in press). Theory-based Bayesian models for inductive learning and reasoning. *Trends in Cognitive Science*.
- Tenenbaum, J. B., & Xu, F. (2000). Word learning as Bayesian inference. In *Proceedings of the 22nd Annual Conference of the Cognitive Science society* (pp. 517–522). Hillsdale, NJ: Erlbaum.
- Ward, T. B., Becker, A. H., Hass, S. D., & Vela, E. (1991). Attribute availability and the shape bias in children's category generalization. *Cognitive Development, 6*, 143–167.

Author Note

We thank two anonymous reviewers for helpful suggestions. This work was supported by the William Asbjornsen Albert Memorial fellowship (CK), a NDSEG graduate fellowship (AP), and the Paul E. Newton Career Development Chair (JBT).

Footnotes

¹Other authors distinguish between theories, schemata, scripts, and overhypotheses. There are important differences between these varieties of abstract knowledge, but it is useful to have a single term (for us, overhypothesis) that includes them all.

²For those who support an essentialist view of categories (Medin & Ortony, 1989; Bloom, 2000), the issue at stake is whether the identically-shaped objects were believed to have the same essence. A shared name is one indication that two objects have the same essence, but other indications are possible — for example, children might be told “Here’s one and here’s another. Look, they are both the same kind of thing. I wonder what they’re called.”

³The network used by Colunga and Smith (2005) is related to a Boltzmann machine (Ackley, Hinton, & Sejnowski, 1985), which is an exact implementation of a known computational theory.

Figure Captions

Figure 1. (a) A hierarchical Bayesian model. Each setting of (α, β) is an overhypothesis: β represents the color distribution across all categories, and α represents the variability in color within each category. (b) A model with separate overhypotheses for two ontological kinds meant to correspond loosely to objects and substances. α^1 represents knowledge about feature variability within the first ontological kind (object categories are homogeneous in shape but not in material), and β^1 captures the characteristic features of the entities belonging to the first kind (objects tend to be solid).

Figure 2. The Dirichlet distribution serves as a prior on θ , the color distribution of a bag of marbles. Assume that there are two possible colors — white and black — and let θ_2 be the proportion of black marbles within the bag. Shown here are distributions on θ_2 when the parameters of the Dirichlet distribution (α and β) are systematically varied. When α is small, the marbles in each individual bag are near-uniform in color (θ_2 is close to 0 or close to 1), and β determines the relative proportions of bags that are mostly white and bags that are mostly black. When α is large, the color distribution for any individual bag is expected to be close to the color distribution across the entire population of bags (θ_2 is close to β_2).

Figure 3. (a) Generalizations made by the model in Figure 1a. (i) Prior distributions on $\log(\alpha)$, β and $\{\theta^i\}$ indicate the model’s expectations before any data have been observed (ii) posterior distributions after observing 10 all-white bags and 10 all-black bags; (iii) posterior distributions after observing 20 mixed bags inspired by the obesity condition of the Barratos task. After observing 20 bags that are either all white or all black (ii), the model realizes that most bags are near-uniform in color (α is small), and that about half of these bags are black (β_2 is around 0.5). These posterior distributions allow the model to predict that the proportion of black marbles in the new, sparsely observed bag (θ_2^{new})

is very close to 1. After observing 20 mixed bags, the model realizes that around 25% of marbles are black (β_2 is around 0.25), and that roughly 25% of the marbles in each individual bag are black (α is high). These posterior distributions allow the model to predict that the new, sparsely observed bag is likely to contain more white marbles than black marbles (θ_2^{new} is not close to 1). (b) Generalizations of a conventional Bayesian model that learns only at the level of θ (α and β are fixed). The model does not generalize correctly to new, sparsely observed bags: since α and β are fixed, observing 20 previous bags provides no information about a new bag, and the posterior distributions on θ_2^{new} are identical for cases (ii) and (iii).

Figure 4. Generalizations about a new tribe after observing 1, 3, or 20 obese, brown-skinned individuals from that tribe. Human generalizations are replotted from Nisbett et al. (1983). For each set of observations, our model learns a distribution over the feature proportions θ^{new} for a new tribe (Figure 3a). Plotted here are the means of those distributions. A single observation allows the model to predict that most individuals in the new tribe have brown skin, but many more observations are needed before the model concludes that most tribe members are obese.

Figure 5. Learning the shape bias. (a) Training data based on Smith et al. (2002). Each column represents an object, and there are 10 possible colors, textures, and shapes, and 2 possible sizes. (b) First-order generalization was tested by presenting the model with exemplar T_1 , and asking it to choose which of three objects (a shape match, a texture match and a color match) was most likely to belong to the same category as T_1 . (c) Second-order generalization was tested using T_2 , an exemplar of a category that was not seen during training. (d) Model predictions for both generalization tasks. Each bar represents the probability that a choice object belongs to the same category as the test exemplar (probabilities have been normalized so that they sum to 1 across each set of

choice objects). The model makes exact predictions about these probabilities: we computed 30 estimates of these predictions, and the error bars represent the standard error of the mean.

Figure 6. (a) Mean α values after observing 32 white marbles and 32 black marbles divided evenly across some number of homogenous bags. The model is most confident that bags in general are homogeneous (i.e. α is low) when given 2 samples from each of 32 bags. (b) Three possible outcomes when learning occurs simultaneously at level 1 and level 2. (i) After observing 2 homogeneous bags, the model is more certain about the variables at level 1 than the variables at level 2. (ii) After observing pairs of marbles from 32 homogeneous bags, the model is fairly certain about both levels. (iii) After observing pairs of marbles from 32 bags (5 white pairs, 22 mixed pairs, and 5 black pairs), the model is more certain about level 2 than level 1.

Figure 7. Learning a shape bias for solids and a material bias for non-solids. (a) Training data. (b) Second-order generalization was tested using solid and non-solid exemplars (S , N). In each case, two choice objects were provided — a shape match and a material match. (c) The model chooses the shape match given the solid exemplar and the material match given the non-solid exemplar. The model makes exact predictions about the probabilities plotted, and the error bars represent standard error across 8 estimates of these probabilities. (d) The model groups the categories into two kinds: objects (categories 1, 2 and 5) and substances (categories 3, 4 and 6). Entry (i, j) in the matrix is the posterior probability that categories i and j belong to the same ontological kind (light colors indicate high probabilities).

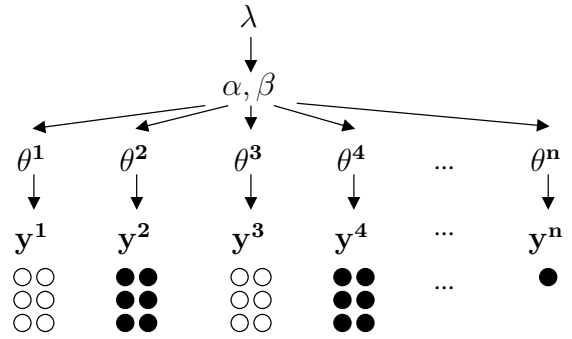
Learning overhypotheses, Figure 1

a) Level 3: Over-overhypotheses

Level 2: Overhypotheses

Level 1: Category means

Data

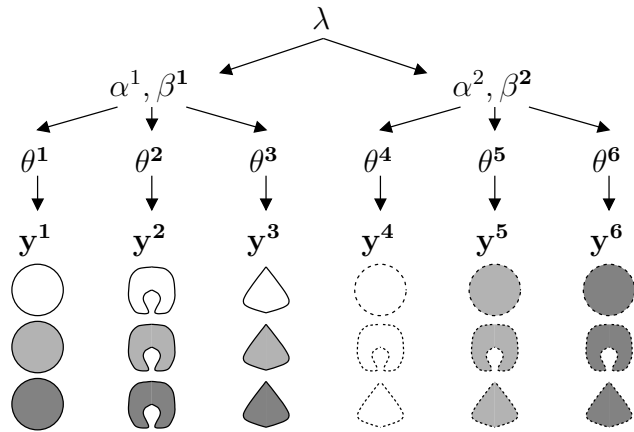


b) Level 3: Over-overhypotheses

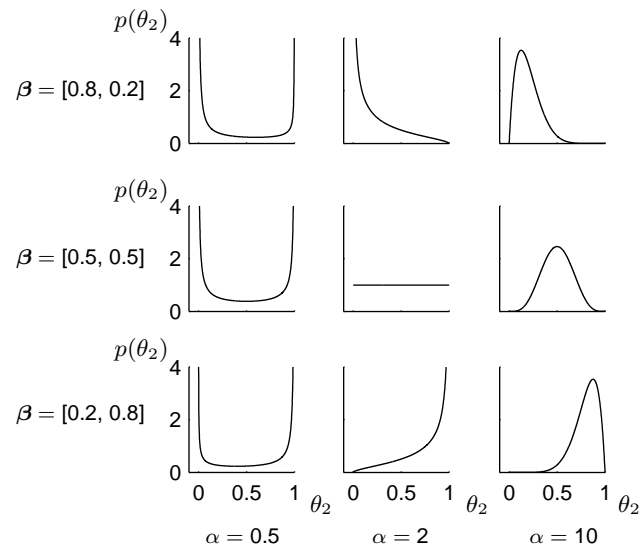
Level 2: Overhypotheses

Level 1: Category means

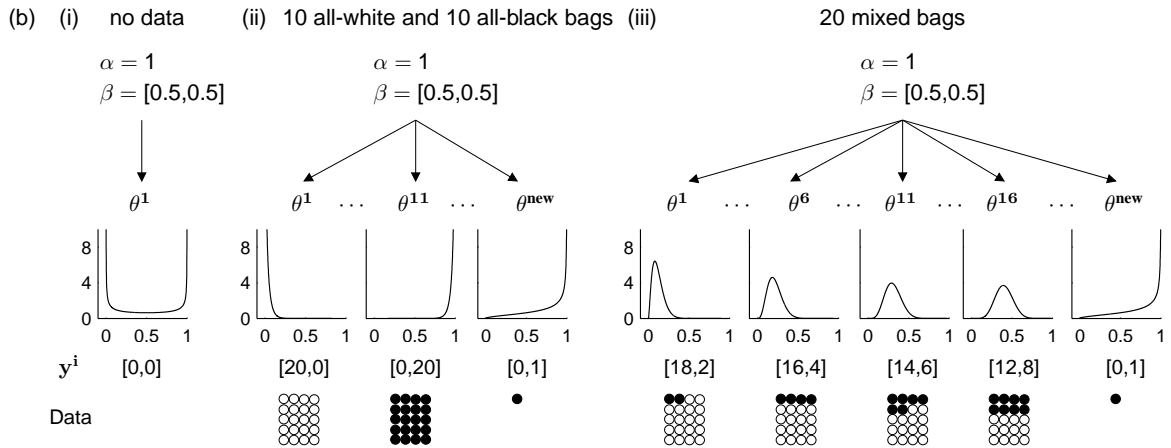
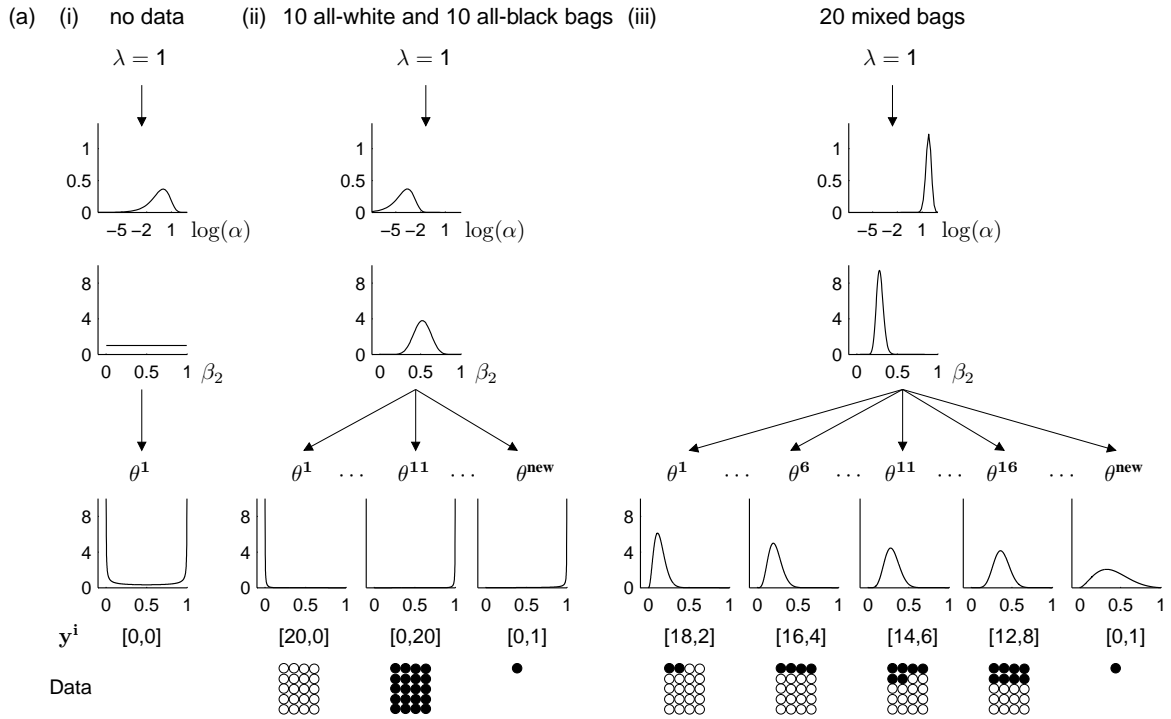
Data



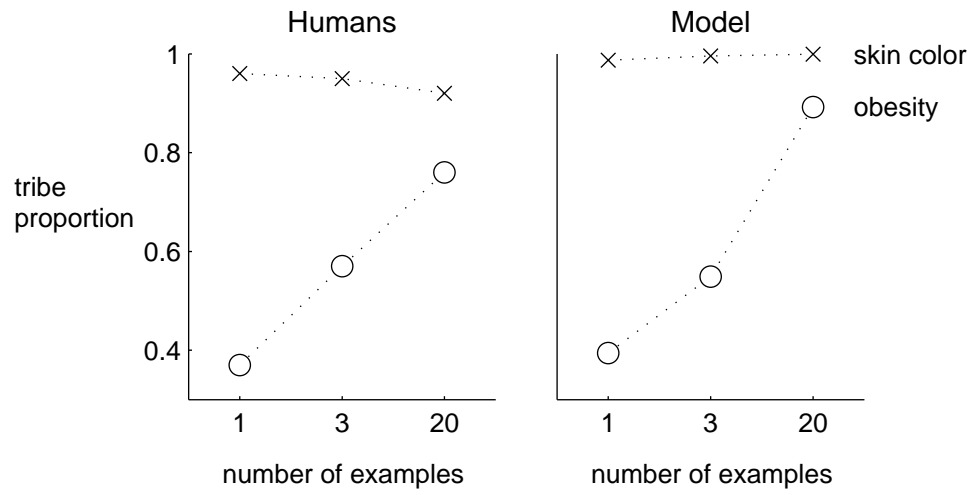
Learning overhypotheses, Figure 2



Learning overhypotheses, Figure 3

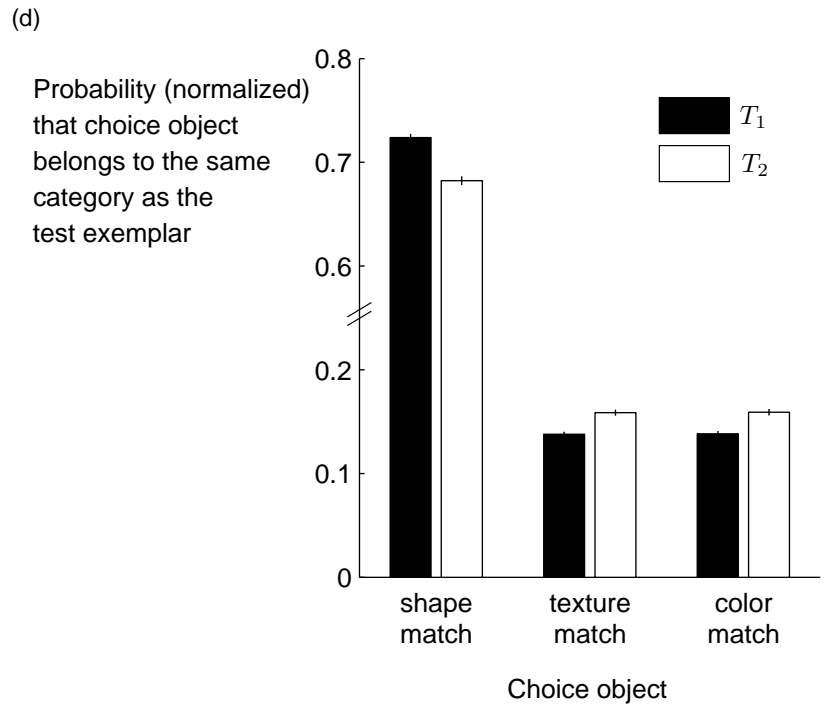


Learning overhypotheses, Figure 4

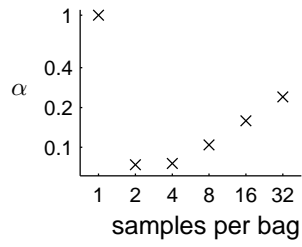


Learning overhypotheses, Figure 5

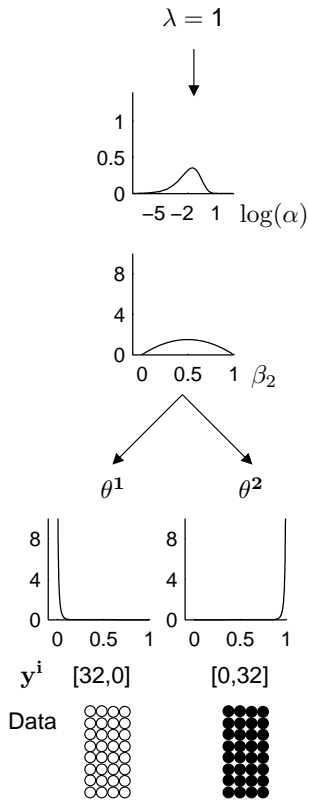
(a) Training	(b) First-order generalization	(c) Second-order generalization
	T_1	T_2
Category	1 ? ? ?	5 ? ? ?
Shape	1 1 6 6	5 5 6 6
Texture	1 9 1 9	9 10 9 10
Color	1 9 9 1	9 10 10 9
Size	1 1 1 1	1 1 1 1



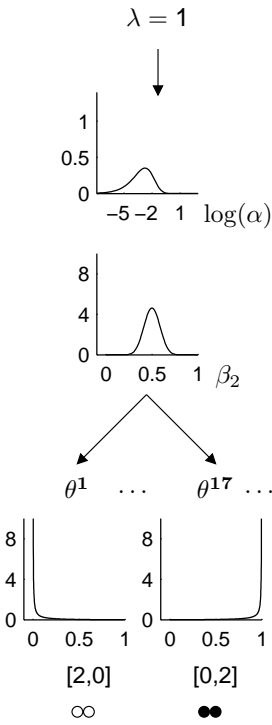
(a)



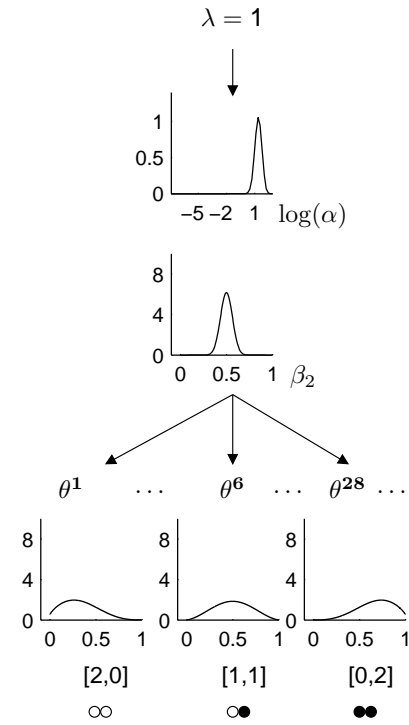
(b) (i) More certain about level 1



(ii) Fairly certain about both levels



(iii) More certain about level 2



Learning overhypotheses, Figure 7

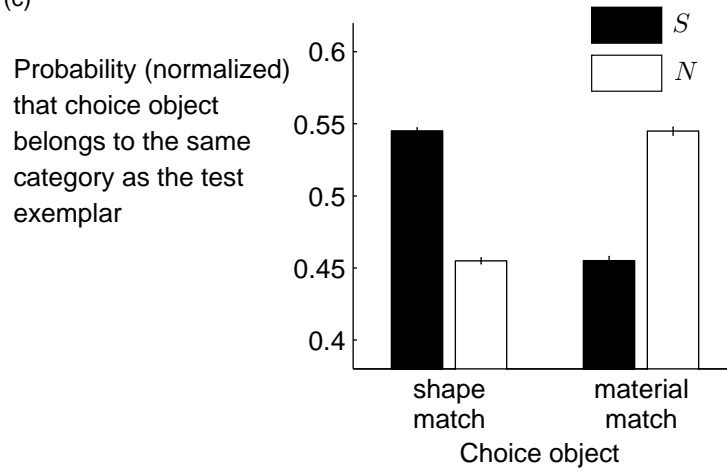
(a) Training

Category	1	1	2	2	3	3	4	4
Shape	1	1	2	2	3	4	5	6
Material	1	2	3	4	5	5	6	6
Size	1	2	1	2	1	2	1	2
Solidity	1	1	1	1	2	2	2	2

(b) Second-order generalization

<i>S</i>		<i>N</i>	
5	? ?	6	? ?
7	7 8	8	8 9
7	8 7	8	9 8
1	1 1	1	1 1
1	1 1	2	2 2

(c)



(d)

