Learning Overhypotheses

Charles Kemp, Amy Perfors & Joshua B. Tenenbaum
{ckemp, perfors, jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

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. The hierarchical approach also addresses a common question about Bayesian models of cognition: where do the priors come from? To illustrate our claims, we consider two specific kinds of overhypotheses — overhypotheses about feature variability (eg the shape bias in word learning) and overhypotheses about the grouping of categories into ontological kinds like objects and substances. The models we develop account for several existing datasets.

Compared to machine-learning algorithms, humans are remarkable for doing so much with so little. A single labelled example is enough for children to learn the meanings of some words (Heibeck and Markman, 1987), 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 human learners (Goodman, 1955).

Several authors have proposed that the apparent ease of human learning depends on constraints that guide induction. This view has been applied to many cognitive problems: the M-constraint and the shape bias help explain concept acquisition, universal grammar guides the acquisition of linguistic knowledge (Chomsky, 1980), and the development of folk biology is guided by the idea that living kinds can be organized hierarchically (Atran, 1998). Constraints like these may be called framework theories or schemata, but we will use a term of Goodman’s and refer to them as overhypotheses.1

Some overhypotheses must be innate, but others are probably learned. For at least two reasons, though, the acquisition of overhypotheses has received less attention than it deserves. First, the authors who have argued most convincingly for the importance of overhypotheses often suggest that these overhypotheses are innate, and are therefore not deeply motivated to grapple with the question of how they might be acquired. Second, the study of overhypothesis acquisition raises some formidable methodological challenges. Designing adult experiments to address the problem is difficult, since adults bring a lifetime of learning experience to any experiment and have already distilled overhypotheses that help them deal with most novel tasks. Infant experiments are challenging for different reasons, but can address the acquisition of some of the most fundamental overhypotheses. For instance, Smith and colleagues have explored the development of the shape bias (Colunga and Smith, 2005), and Piaget and colleagues considered how abstract kinds of knowledge (such as the concrete operations) can be acquired and used to support many different learning tasks (Piaget, 1970).

The problem of overhypothesis acquisition is closely related to a problem raised by Bayesian models of cognition. These models usually rely on a prior distribution chosen by the modeller, and a natural response is to wonder where the prior comes from. Hierarchical Bayesian models address this question: in the framework we adopt, learning an overhypothesis amounts to learning a prior distribution. The hierarchical approach shows that priors (or overhypotheses) can be learned given knowledge at a higher level of abstraction.

Hierarchical Bayesian modelling can be used to explore the acquisition of overhypotheses in many different domains. Here we introduce one of the simplest possible HBMs and use it to model how overhypotheses about feature-variability are acquired and used to support categorization. One such overhypothesis is the shape bias, the expectation that shape features are homogeneous within object categories. We also present an extension of the basic model that groups categories into ontological kinds (eg objects and substances) and discovers the features and the patterns of feature variability that are characteristic of each kind.

Overhypotheses and HBMs

Goodman introduces the idea of overhypotheses using bags of colored marbles (Goodman, 1955). Suppose we are faced with a stack $S$ containing many bags of marbles. We empty several bags and discover that some bags contain black marbles, others contain white marbles, but each bag is uniform in color. We now choose a new bag — bag $n$ — and draw a single black marble from the bag. This observation may lead us to endorse the following

---

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

\( H \): All marbles in bag \( n \) are black.

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

\( O \): All bags in \( S \) 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 bag \( n \): it could be uniformly black, uniformly white, uniformly green, and so on. \( O \), however, is just one kind of overhypothesis. To give a very different example, Universal Grammar is an overhypothesis that sets up a space of hypotheses (ie a space of possible grammars) for language learning.

A hierarchical Bayesian model (Gelman et al., 1995) includes hypothesis spaces at several levels of abstraction. Suppose that we are given a body of data, and we wish to account for a given cognitive ability. In Goodman’s case, the data are observations of several bags (\( y_i \) indicates the observations for bag \( i \)) and we are interested in the ability to predict the next marble to be drawn from bag \( n \). 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 bags (\( \theta_i \) indicates the color distribution for the \( i \)th bag). We then ask how the level 1 knowledge is acquired, and the answer will make reference to a body of even more abstract knowledge (level 2 knowledge). In Goodman’s case, level 2 represents knowledge about the distribution of the \( \theta \) variables — knowledge that each bag tends to be uniform in color. As described below, we can formalize this knowledge using two parameters, \( \alpha \) and \( \beta \) (Figure 1a). If we now 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 \( \lambda \). The parameter \( \lambda \) and the pair \( (\alpha, \beta) \) 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 structural descriptions of sentences (Level 1 knowledge). Structural descriptions, 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 modellers have gone beyond 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) in particular is that they provide a formal account of the acquisition process. The acquisition process at the heart of a HBM is statistical inference, which suggests that overhypothesis learning can be approximated as a rational response to the data available to a learner. In the model of Figure 1a, for example, the posterior \( p(\alpha, \beta | y) \) represents a normative belief about level 2 knowledge — the belief, given the data \( y \), that most bags are close to uniform in color.

We have argued that HBMs go beyond previous hierarchical models proposed by cognitive scientists, but they also represent an advance over the standard Bayesian models used in cognitive science. A standard Bayesian model has two levels of knowledge: the elements in its hypothesis space represent level 1 knowledge, and the prior (generally fixed) represents knowledge at level 2. A common objection to Bayesian modelling is that a prior can chosen to approximate almost any pattern of data, and that the success of a Bayesian model owes more to the modeller’s ability to choose the right prior than to any interesting learning on the part of the model. HBMs disarm the objection by showing that knowledge at level 2 need not be specified in advance, but can be learned using the knowledge represented at higher levels. Of course, the prior at the highest level must still be specified in ad-
vance, but the ultimate hope is to design models where this prior is simple enough to be unobjectionable.

**Computational Theory**

We use the marbles example to describe a formal instantiation of the model in Figure 1a. Suppose we are working with a set of \( k \) colors. Initially we set \( k = 2 \) and use black and white as our colors. Let \( \theta_1 \) indicate the true color distribution for the \( i \)th bag in the stack: if 60% of the marbles in bag 7 are black and the remainder are white, then \( \theta_7 = [0.6, 0.4] \). 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 \( y_7 = [4, 1] \).

We assume that \( y_i \) is drawn from a multinomial distribution with parameter \( \theta_i \); 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 \( \theta_i \) of that bag. The vectors \( \theta_i \) are drawn from a Dirichlet distribution parameterized by a scalar \( \alpha \) and a vector \( \beta \). Here \( \beta \) represents the expected distribution of colors across the stack and \( \alpha \) captures the notion of feature variability. The larger the value of \( \alpha \), the more likely that the color distribution for any given bag will be close to the vector \( \beta \). When \( \alpha \) is small, however, each individual bag is likely to be near-uniform in color, and \( \beta \) will determine the relative proportions of ‘mostly black’ and ‘mostly white’ bags.

Each possible setting of \((\alpha, \beta)\) is an overhypothesis. In order to discover values for these variables, we need prior distributions on \( \beta \) and \( \alpha \). We use a uniform distribution (Dirichlet(1)) on \( \beta \) and an exponential distribution with mean \( \lambda \) on \( \alpha \). For all simulations in this paper we set \( \lambda = 1 \).

This model is familiar to statisticians as a Dirichlet-multinomial model Gelman et al. (1995). It can be expressed as follows:

\[
\begin{align*}
\alpha & \sim \text{Exponential}(1) \\
\beta & \sim \text{Dirichlet}(1) \\
\theta_i & \sim \text{Dirichlet}(\alpha, \beta) \\
y_i | n_i & \sim \text{Multinomial}(\theta_i)
\end{align*}
\]

where \( n_i \) is the number of observations for bag \( i \). As written, the model assumes we are working with a single dimension — for Goodman, marble color. Perhaps, however, some marbles are made from metal and others are made from glass. We deal with multiple dimensions by assuming that each dimension is independently generated according to the model, and introducing separate values of \( \alpha \) and \( \beta \) for each dimension. When working with multiple features, we will often use \( \alpha \) to refer to the collection of \( \alpha \) values along all dimensions, and \( \beta \) for the set of all \( \beta \) vectors.

**Modelling behavioral data**

Since Goodman, psychologists have confirmed that adults (Nisbett et al., 1983) and children (Macario et al., 1990) have overhypotheses about feature variability, and use them to support inductive leaps on the basis of very sparse data. Nisbett et al. (1983) asked subjects to imagine that they were exploring a little known island in the Southeastern Pacific. As part of the task, subjects were told that they had encountered a single member of the “Barratos” tribe, and that the tribesman was brown and obese. Based on this single example, subjects concluded that most Barratos were brown, but gave a much lower estimate of the proportion of obese Barratos (Figure 2a). When asked to justify their responses, subjects often said that tribespeople were “homogeneous with respect to color” but “heterogeneous with respect to body weight.”

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 share the same skin color. Given these data, the model learns a posterior distribution on \( \alpha \) indicating that \( \alpha \) is probably small, which means that skin color tends to be homogeneous within tribes. We can also make predictions about a sparsely observed new tribe: having observed a single, brown-skinned member of tribe number 21, the posterior distribution on \( \theta_{21} \) indicates that most members of the tribe are likely to be brown (Figure 2a). Suppose now that obesity is a feature that is variable across 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 model now requires much more evidence to decide that most members of tribe 21 share the feature (Figure 2).

**Novel feature values**

The Barratos task does not address an important kind of inference that overhypotheses support: reasoning about new features. We have seen that experience with bags of marbles that are all white and others that are all black can lead a subject to conclude that a new bag is black after seeing a single black marble drawn from the bag. Perhaps, however, the subject does not have Goodman’s overhypothesis \( O \)

\[ O: \text{All bags in } S \text{ are uniform in color} \]

but instead has overhypothesis \( O' \):

\[ O': \text{All bags in } S \text{ are all black or all white} \]
First and second-order generalization were tested using exemplars and textures, 5 possible shapes, and 2 possible sizes. For example, the count vector $y_1$ says that the observed exemplars of category 1 include 2 objects with shape value 1 and no objects with shape value 10. This policy allows the model to handle shapes, colors and textures it has never seen during training, but assumes, of course, 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. Smith (1984) provides experimental evidence that three-year olds have this capacity.

Generalizations of the model are shown in Figure 3b. Smith et al. (2002) report that shape matches are chosen 88% of the time given exemplar $T_1$, and 70% of the time given exemplar $T_2$. The model reproduces this general pattern: shape matches are preferred in both cases, and are preferred more strongly when the exemplar belongs to a familiar category.

### Discovering ontological kinds

The model in Figure 1a is a simple HBM 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, that color tends to be homogeneous within object categories but heterogeneous within object categories, and that both shape and texture tend to be homogeneous within animate categories.

Figure 1b shows how we can give our model the ability to discover ontological kinds. The model assumes that categories may be grouped into ontological kinds, and that each kind is associated with a different $\alpha$ and $\beta$. The model, however, is not told which categories belong together, 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. The prior is induced by a Chinese Restaurant process (CRP): it assigns some probability to all possible category partitions, although it favors partitions with a small number of kinds. The same prior lies at the heart of Anderson’s (1991) rational model of categorization, although Anderson works with partitions of objects into categories rather than partitions of categories into ontological kinds.

The new model can be written as:

$$z \mid n_{\text{cat}} \sim \text{CRP}(\gamma)$$

$$\alpha_k \sim \text{Exponential}(\lambda)$$

$$\beta_k \sim \text{Dirichlet}(1)$$

$$\theta_i \sim \text{Dirichlet}(\alpha_{z_i}, \beta_{z_i})$$

$$y_i \mid n_i \sim \text{Multinomial}(\theta_i)$$

### Table 3a: Counts of Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Training</th>
<th>$T_1$</th>
<th>$T_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape</td>
<td>11 2 2 3 4 4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Texture</td>
<td>1 2 3 4 5 6 7 8</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Color</td>
<td>1 2 3 4 5 6 7 8</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Size</td>
<td>1 2 1 2 1 2 1 2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
where \( n_{\text{cat}} \) is the number of categories, \( \gamma \) is the concentration parameter for the CRP, \( z_i \) is the kind label for category \( i \) and there is a separate \( \alpha_k \) and \( \beta_k \) for each ontological kind \( k \).

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 dataset of comparable size (Figure 4a). During training, the model saw two exemplars from each of four categories: two object categories and two substance categories. 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 4b shows that the model chooses a shape match for the solid exemplar and a material match for the non-solid exemplar.

Figure 4c 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 (\( \beta \)) as well as patterns of feature variability within each kind (\( \alpha \)). If the object categories are grouped into kind \( k \), \( \alpha_k \) indicates that shape is homogeneous within categories of that kind, and \( \beta_k \) indicates that categories of that kind tend to be solid. The \( \beta \) parameter, then, is responsible for the inference that the test exemplar \( S \) should be grouped with the two object categories, since all three categories contain solid objects.

**Learning at multiple levels**

Although the model in 1a is relatively simple, this very simplicity makes it a useful source of metaphors that help to illuminate more difficult problems. Here we suggest that the model helps to address the chicken-and-egg problem of learning at multiple levels of abstraction, even when those levels appear to depend on each other.

The problem seems most challenging in the context of difficult tasks like grammar acquisition. The evidence that supports grammar learning presumably includes structural descriptions of sentences, yet these structural descriptions are provided by the grammar itself. The tasks we considered have been much simpler than grammar acquisition, but the same apparent circularity still arises. Knowledge about feature variability helps learners discover the distribution of features within categories (as in the Barratos example), but knowledge of feature variability is driven in turn by distributional evidence. Bayesian inference overcomes the apparent circularity: given a pattern of data we can compute the posterior distribution on both the feature-variability parameter \( \alpha \) and the feature distribution \( \theta \) within any category \( i \).

Figure 5 shows that the model in 1a can become relatively certain about the value of the \( \alpha \) parameter even though it has very sparse evidence about any given category. Each point 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, the model has strong evidence about the color distribution of each bag: one is mostly black and the other is mostly white. No matter how well those two bags are understood, knowledge about only two bags does not provide much evidence about feature variability within bags in general. The case where two observations are provided for each of 32 bags represents the other extreme. The evidence about the composition of any single bag is weak, but taken together, these observations provide strong support for the idea that most bags are homogeneous. Similar inferences may be found in more natural settings. For example, learning that most people voted consistently across the last two presidential elections supports the idea that most of the voters are committed Republicans or committed Democrats, and will vote similarly in future elections.

**Discussion**

We presented hierarchical Bayesian models that help explain the acquisition of the shape bias, and of over-hypotheses about feature variability within ontological kinds. The model in Figure 1a, however, addresses several other psychological phenomena. When shown a circle with a diameter of three inches, subjects report that the circle is more likely to be a pizza than a quarter, even though the circle is closer in size to the average quarter than the average pizza (Rips, 1989). Our model suggests that this decision is driven by knowledge about the variability of the size feature. The model also accounts for some of Harlow’s experiments on “learning to
learn” (Harlow, 1949). Harlow gave monkeys a blocked forced-choice decision task, where the same object was rewarded within each block regardless of whether it appeared on the left or the right. After many blocks, Harlow found that his monkeys were almost always choosing correctly after second trial in each block. They had evidently learned that the rewarded objects in each block were homogeneous in shape and color, but heterogeneous in position.

We know of no previous attempts to provide rational computational theories (Marr, 1982) of the acquisition of the shape bias, or of other overhypotheses about feature variability. Colunga and Smith (2005) present a connectionist model that acquires knowledge of this sort, but our approach is different in emphasis and explanatory effect. We provided a computational theory but have not attempted to specify the psychological mechanisms by which it might be implemented, and Colunga and Smith (2005) provide a process model but do not give a computational theory which formally specifies the task to be solved by the model. It is possible that our two approaches are complementary, and that a system like Colunga and Smith (2005) describe may be able to implement the computations required by our theory.

We suggested earlier that our models provides metaphors that help explain the acquisition of overhypotheses about objects with representations more complex than lists of features. One reason why these metaphors seem useful is that our probabilistic approach extends naturally to contexts where structured representations are required: computational linguists, for example, work with probabilistic grammars that generate parse trees. The connectionist model of Colunga and Smith (2005) does not share this property, since it is unclear how it could be extended to deal with structured objects like parse trees.

Like other rational analyses of categorization, our model follows the general approach described by Anderson (1991). Anderson suggests that HBMs are necessary to account for data like those shown in Figure 2a, but we know of no attempts to follow up on this idea. Unlike the formal model presented by Anderson (1991), both of our models assume that category labels are observed for each object. This assumption can be avoided: for example, we can use the nested CRP (Blei et al., 2003) to develop an alternative to the model in 1b that simultaneously discovers ontological kinds and categories within each kind.

Although we have argued that overhypotheses can be acquired 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 need no initial assumptions beyond those we are willing to make. Whether the hierarchical Bayesian approach will meet this challenge is far from clear, but it deserves to be put to the test.

### References


---

**Figure 5:** (a) Mean $\alpha$ values after seeing 32 white marbles and 32 black marbles. Low values of $\alpha$ indicate that bags are expected to be homogeneous in color. The model is most confident that bags are homogeneous when given the data in (c) (b) Dataset with 32 samples per bag (c) Dataset with 2 samples per bag.