MIT Libraries | DSpace@MIT

MIT Open Access Articles

A tutorial introduction to Bayesian models of cognitive development

Massachusetts Institute of Technology

A tutorial introduction to Bayesian models of cognitive development

Amy Perfors

School of Psychology, University of Adelaide

Joshua B. Tenenbaum

Brain and Cognitive Sciences, Massachusetts Institute of Technology

Thomas L. Griffiths

Fei Xu

Department of Psychology, University of California, Berkeley

Abstract

We present an introduction to Bayesian inference as it is used in probabilistic models of cognitive development. Our goal is to provide an intuitive and accessible guide to the *what*, the *how*, and the *why* of the Bayesian approach: what sorts of problems and data the framework is most relevant for, and how and why it may be useful for developmentalists. We emphasize a qualitative understanding of Bayesian inference, but also include information about additional resources for those interested in the cognitive science applications, mathematical foundations, or machine learning details in more depth. In addition, we discuss some important interpretation issues that often arise when evaluating Bayesian models in cognitive science.

Keywords: Bayesian models; cognitive development

1 Introduction

One of the central questions of cognitive development is how we learn so much from such apparently limited evidence. In learning about causal relations, reasoning about object categories or their properties, acquiring language, or constructing intuitive theories, children routinely draw inferences that go beyond the data they observe. Probabilistic models provide a general-purpose computational framework for exploring how a learner might make these inductive leaps, explaining them as forms of Bayesian inference.

This paper presents a tutorial overview of the Bayesian framework for studying cognitive development. Our goal is to provide an intuitive and accessible guide to the *what*, the *how*, and the *why* of the Bayesian approach: what sorts of problems and data the framework is most relevant for, and how and why it may be useful for developmentalists. We consider three general inductive problems that learners face, each grounded in specific developmental challenges:

1. Inductive generalization from examples, with a focus on learning the referents of words for object categories.

2. Acquiring inductive constraints, tuning and shaping prior knowledge from experience, with a focus on learning to learn categories.

3. Learning inductive frameworks, constructing or selecting appropriate hypothesis spaces for inductive generalization, with applications to acquiring intuitive theories of mind and inferring hierarchical phrase structure in language.

We also discuss several general issues as they bear on the use of Bayesian models: assumptions about optimality, biological plausibility, and what idealized models can tell us about actual human minds. The paper ends with an appendix containing a glossary and a collection of useful resources for those interested in learning more.

2 Bayesian Basics: Inductive generalization from examples

The most basic question the Bayesian framework addresses is how to update beliefs and make inferences in light of observed data. In the spirit of Marr's (1982) computational-level of analysis, it begins with understanding the logic of the inference made when generalizing from examples, rather than the algorithmic steps or specific cognitive processes involved. A central assumption is that degrees of belief can be represented as probabilities: that our conviction in some hypothesis h can be expressed as a real number ranging from 0 to 1, where 0 means something like "h is completely false" and 1 that "h is completely true." The framework also assumes that learners represent probability distributions and that they use these probabilities to represent uncertainty in inference. These assumptions turn the mathematics of probability theory into an engine of inference, a means of weighing each of a set of mutually exclusive and exhaustive hypotheses $\mathcal{H}$ to determine which best explain the observed data. Probability theory tells us how to compute the degree of belief in some hypothesis h_i, given some data d.

Computing degrees of belief as probabilities depends on two components. One, called the *prior probability* and denoted $P(h_i)$, captures how much we believe in h_i prior to observing the data d. The other, called the *likelihood* and denoted $P(d|h_i)$, captures the probability with which we would expect to observe the data d if h_i were true. These combine to yield the *posterior probability* of h_i, given via Bayes' Rule:

$$P(h_i|d) = \frac{P(d|h_i)P(h_i)}{\sum_{h_j \in \mathcal{H}} P(d|h_j)P(h_j)}. \tag{1}$$

As we will see, the product of priors and likelihoods often has an intuitive interpretation. It balances between a sense of plausibility based on background knowledge on one hand and the data-driven sense of a "suspicious coincidence" on the other. In the spirit of Ockham's Razor, it expresses the tradeoff between the intrinsic complexity of an explanation and how well it fits the observed data.

The denominator in Equation 1 provides a normalizing term which is the sum of the probability of each of the possible hypotheses under consideration; this ensures that Bayes' Rule will reflect the proportion of all of the probability that is assigned to any single hypothesis h_i, and (relatedly) that the posterior probabilities of all hypotheses sum to one. This captures what we might call the "law of conservation of belief": a rational learner has a fixed "mass" of belief to allocate over different hypotheses, and the act of observing data just pushes this mass around to different regions of the hypothesis space. If the data lead us to strongly believe one hypothesis, we must decrease our degree of belief in all other hypotheses. By contrast, if the data strongly disfavor all but one hypothesis, then (to paraphrase Sherlock Holmes) whichever remains, however implausible *a priori*, is very likely to be the truth.

To illustrate how Bayes' Rule works in practice, let us consider a simple application with three hypotheses. Imagine you see your friend Sally coughing. What could explain this? One possibility (call it h_{cold}) is that Sally has a cold; another (call it h_{cancer} is that she has lung cancer; and yet another (call it $h_{\text{heartburn}}$) is that she has heartburn. Intuitively, in most contexts, h_{cold} seems by far the most probable, and may even be the only one that comes to mind consciously. Why? The likelihood favors h_{cold} and h_{cancer} over $h_{\text{heartburn}}$, since colds and lung cancer cause coughing, while heartburn does not. The prior, however, favors h_{cold} and $h_{\text{heartburn}}$ over h_{cancer}: lung cancer is thankfully rare, while colds and heartburn are common. Thus the posterior probability – the product of these two terms – is high only for h_{cold}.

The intuitions here should be fairly clear, but to illustrate precisely how Bayes' Rule can be used to back them up, it can be helpful to assign numbers.[1] Let us set the priors as follows: $P(h_{\text{cold}}) = 0.5$, $P(h_{\text{heartburn}}) = 0.4$, and $P(h_{\text{cancer}}) = 0.1$. This captures the intuition that colds are slightly more common than heartburn, but both are

---

[1]Note that we have assumed that these are the only possible hypotheses, and that exactly one applies. That is why the priors are much higher than the base rates of these diseases. In a real setting, there would be many more diseases under consideration, and each would have much lower prior probability. They would also not be mutually exclusive. Adding such details would make the math more complex but not change anything else, so for clarity of exposition we consider only the simplified version.

significantly more common than cancer. We can set our likelihoods to be the following: $P(d|h_{\text{cold}}) = 0.8$, $P(d|h_{\text{cancer}}) = 0.9$, and $P(d|h_{\text{heartburn}}) = 0.1$. This captures the intuition that both colds and cancer tend to lead to coughing, and heartburn generally does not. Plugging this into Bayes' Rule gives:

$$P(h_{\text{cold}}|d) = \frac{P(d|h_{\text{cold}})P(h_{\text{cold}})}{P(d|h_{\text{cold}})P(h_{\text{cold}}) + P(d|h_{\text{cancer}})P(h_{\text{cancer}}) + P(d|h_{\text{heartburn}})P(h_{\text{heartburn}})}$$
$$= \frac{(0.8)(0.5)}{(0.8)(0.5) + (0.9)(0.1) + (0.1)(0.4)}$$
$$= \frac{0.4}{0.4 + 0.09 + 0.04} = 0.7547.$$

Thus, the probability that Sally is coughing because she has a cold is much higher than the probability of either of the other two hypotheses we considered. Of course, these inferences could change with different data or in a different context. For instance, if the data also included coughing up blood, chest pain, and shortness of breath, you might start to consider lung cancer as a real possibility: the likelihood now explains that data better than a cold would, which begins to balance the low prior probability of cancer in the first place. On the other hand, if you had other information about Sally – e.g., that she had been smoking two packs of cigarettes per day for 40 years – then it might raise the prior probability of lung cancer in her case. Bayes' Rule will respond to these changes in the likelihood or the prior in a way that accords with our intuitive reasoning.

The Bayesian framework is generative, meaning that observed data are assumed to be generated by some underlying process or mechanism responsible for creating the data. In the example above, data (symptoms) are generated by an underlying illness. More cognitively, words in a language may be generated by a grammar of some sort, in combination with social and pragmatic factors. In a physical system, observed events may be generated by some underlying network of causal relations. The job of the learner is to evaluate different hypotheses about the underlying nature of the generative process, and to make predictions based on the most likely ones. A probabilistic model is simply a specification of the generative processes at work, identifying the steps (and associated

probabilities) involved in generating data. Both priors and likelihoods are typically describable in generative terms.

To illustrate how the nature of the generative process can affect a learner's inference, consider another example, also involving illness. Suppose you observe that 80% of the people around you are coughing. Is this a sign that a new virus is going around? Your inference will depend on how those data were generated – in this case, whether it is a random sample (composed, say, of people that you saw on public transport) or a non-random one (composed of people you see sitting in the waiting room at the doctor's office). The data are the same – 80% of people are coughing – regardless of how it was generated, but the inferences are very different: you are more likely to conclude that a new virus is going around if you see 80% of people on the bus coughing. A doctor's office full of coughing people means little about whether a new virus is going around, since doctor's offices are never full of healthy people.

How can the logic of Bayesian inference, illustrated here with these medical examples, apply to problems like word and concept learning, the acquisition of language, or learning about causality or intuitive theories? In these cases, there is often a huge space of hypotheses (possibly an infinite one). It may not be clear how the models in question should be interpreted generatively, since they seem to delineate sets (e.g., the set of instances in a concept, the set of grammatical sentences, or the set of phenomena explained by a theory). Here we illustrate how Bayesian inference works more generally in the context of a simple schematic example. We will build on this example throughout the paper, and see how it applies and reflects problems of cognitive interest.

Our simple example, shown graphically in Figure 1, uses dots to represent individual data points (e.g., words or events) generated independently from some unknown process (e.g., a language or a causal network) that we depict in terms of a region or subset of space: the process generates data points randomly within its region, never outside. Just as each of the hypotheses in the medical example above (i.e., cold, heartburn, or cancer) are associated with different data (i.e., symptoms), each hypothesis here

Figure 1: (i) Example data and hypothesis. Graphical representation of data and one possible hypothesis about how those data were generated. There are three hypotheses here, each corresponding to a single rectangle. The black data points can only be generated by the solid or the dashed rectangle. A new data point in position a might be generated if the dashed rectangle is correct, but not the solid or dotted one. (ii) Some hypotheses in the hypothesis space for this example. Hypotheses consist of rectangles; some are well-supported by the data and some are not.

encodes a different idea about which subset of space the data are drawn from. Figure 1(i) depicts three possible hypotheses, each consisting of a single rectangle in the space: h_{solid} corresponds to the solid line, h_{dashed} to the dashed line, and h_{dotted} to the dotted line. Before seeing data, a learner might have certain beliefs about which hypothesis is most likely; perhaps they believe that all are equally likely, or they have a bias to prefer smaller or larger rectangles. These prior beliefs, whatever they are, would be captured in the prior probability of each hypothesis: $P(h_{\text{solid}})$, $P(h_{\text{dashed}})$, and $P(h_{\text{dotted}})$. The different hypotheses also yield different predictions about what data one would expect to see; in Figure 1(i), the data are consistent with h_{solid} and h_{dashed}, but not h_{dotted}, since some of the points are not within the dotted rectangle. This would be reflected in their likelihoods; $P(d|h_{\text{solid}})$ and $P(d|h_{\text{dashed}})$ would both be non-zero, but $P(d|h_{\text{dotted}})$ would be zero. Bayesian inference can also yield predictions about about unobserved data. For instance, one would only observe new data at position a if h_{dashed} is correct, since $P(a|h_{\text{solid}}) = 0$, but $P(a|h_{\text{dashed}}) > 0$. In this sense, inferring the hypotheses most likely to have generated the observed data guides the learner in generalizing beyond the data to new situations.

The hypothesis space $\mathcal{H}$ can be thought of as the set of all possible hypotheses,

defined by the structure of the problem that the learner can entertain. Figure 1(ii) shows a possible hypothesis space for our example, consisting of all possible rectangles in this space. Note that this hypothesis space is infinite in size, although just a few representative hypotheses are shown.

The hypothesis space is defined by the nature of the learning problem, and thus provided to the learner *a priori*. For instance, in our example, the hypothesis space would be constrained by the range of possible values for the lower corner (x and y), length (l), and width (w) of rectangular regions. Such constraints need not be very strong or very limiting: for instance, one might simply specify that the range of possible values for x, y, l, and w lies between 0 and some extremely large number like 10^9, or be drawn from a probability distribution with a very long tail. In this sense, the prior probability of a hypothesis $P(h_i)$ is also given by a probabilistic generative process – a process operating "one level up" from the process indexed by each hypothesis that generates the observed data points. We will see below how these hypothesis spaces and priors need not be built in, but can be constructed or modified from experience.

In our example the hypothesis space has a very simple structure, but because a Bayesian model can be defined for any well-specified generative framework, inference can operate over any representation that can be specified by a generative process. This includes, among other possibilities, probability distributions in a space (appropriate for phonemes as clusters in phonetic space); directed graphical models (appropriate for causal reasoning); abstract structures including taxonomies (appropriate for some aspects of conceptual structure); objects as sets of features (appropriate for categorization and object understanding); word frequency counts (convenient for some types of semantic representation); grammars (appropriate for syntax); argument structure frames (appropriate for verb knowledge); Markov models (appropriate for action planning or part-of-speech tagging); and even logical rules (appropriate for some aspects of conceptual knowledge). The appendix contains a detailed list of papers that use these and other representations.

The representational flexibility of Bayesian models allows us to move beyond some of the traditional dichotomies that have shaped decades of research in cognitive development: structured knowledge vs. probabilistic learning (but not both), or innate structured knowledge vs. learned unstructured knowledge (but not the possibility of knowledge that is both learned and structured). As a result of this flexibility, traditional critiques of connectionism that focus on their inability to adequately capture compositionality and systematicity (e.g., Fodor & Pylyshyn, 1988) do not apply to Bayesian models. In fact, there are several recent examples of Bayesian models that embrace language-like or compositional representations in domains ranging from causal induction (Griffiths & Tenenbaum, 2009) to grammar learning (Perfors, Tenenbaum, & Regier, submitted) to theory acquisition (Kemp, Tenenbaum, Niyogi, & Griffiths, 2010).

2.1   A case study: learning names for object categories

To illustrate more concretely how this basic Bayesian analysis of inductive generalization applies in cognitive development, consider the task a child faces in learning names for object categories. This is a classic instance of the problem of induction in cognitive development, as many authors have observed. Even an apparently simple word like "dog" can refer to a potentially infinite number of hypotheses, including all dogs, all Labradors, all mammals, all animals, all pets, all four-legged creatures, all dogs except Chihuahuas, all things with fur, all running things, etc. Despite the sheer number of possible extensions of the word, young children are surprisingly adept at acquiring the meanings of words – even when there are only a few examples, and even when there is no systematic negative evidence (Markman, 1989; Bloom, 2000).

How do children learn word meanings so well, so quickly? One suggestion is that infants are born equipped with strong prior knowledge about what sort of word meanings are natural (Carey, 1978; Markman, 1989), which constrains the possible hypotheses considered. For instance, even if a child is able to rule out part-objects as possible
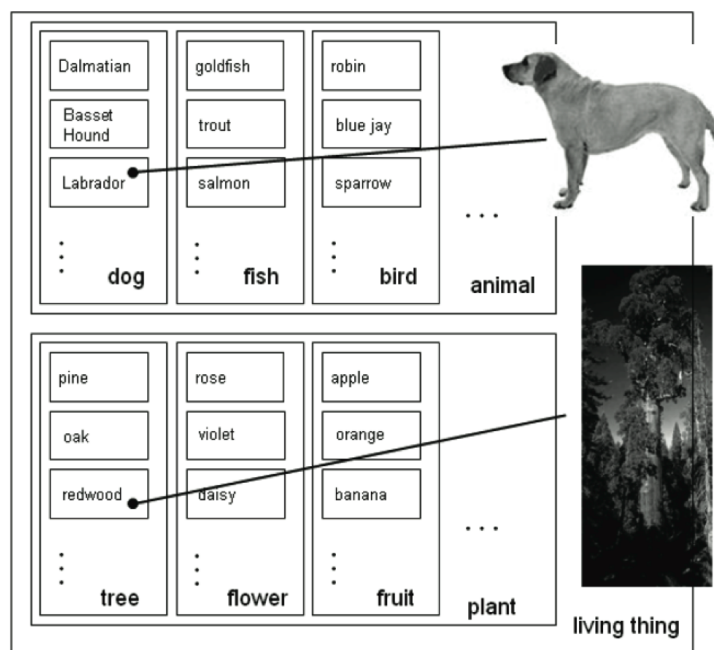
Figure 2: Schematic view of hypotheses about possible extensions considered by the learner in Xu & Tenenbaum (2007); because the taxonomy is hierarchical, the hypotheses are nested within each other. Figure reproduced from Xu & Tenenbaum (2007).

extensions, she cannot know what level of a taxonomy the word applies: whether "dog" actually refers to dogs, mammals, Labradors, canines, or living beings. One solution would be to add another constraint – the presumption that count nouns map preferentially to the basic level in a taxonomy (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). This preference would allow children to learn names for basic-level categories, but would be counterproductive for every other kind of word.

Xu and Tenenbaum (2007b) present a Bayesian model of word learning that offers a precise account of how learners could make meaningful generalizations from one or a few examples of a novel word. This problem can be schematically depicted as in Figure 2: for concepts that are organized in a hierarchical taxonomy, labelled examples are consistent with multiple different extensions. For instance, a single label "Labrador" could pick out only Labradors, but it could also pick out dogs, animals, or living things. This problem is faced by a child who, shown one or many objects with a given label, must decide which hypothesis about possible extensions of the label is best. Intuitively,
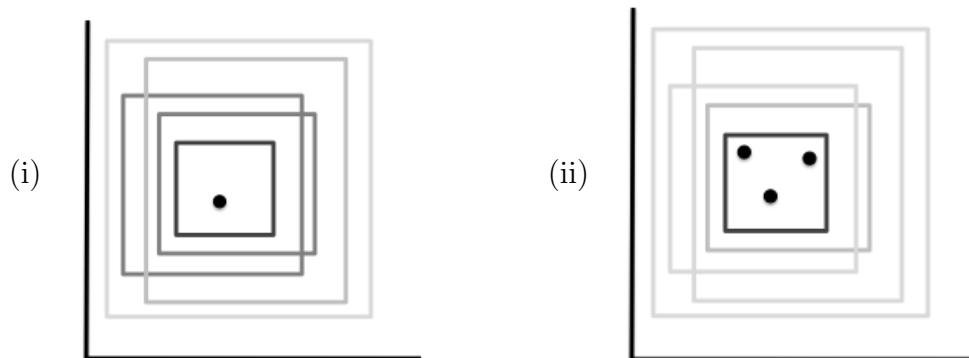
Figure 3: Learning object words. (i) Hypothesis space that is conceptually similar to that in Figure 2, now depicted as a two-dimensional dot diagram; hypotheses with higher probability are darker rectangles. With one data point, many hypotheses have some support. (ii) With three examples, the most restrictive hypothesis is much more strongly favored.

we would expect that when given one object, a reasonable learner should not strongly prefer any of the hypotheses that include it, though the more restricted ones might be slightly favored. If the learner were shown *three* examples, we would expect the most closely-fitting hypothesis to be much more strongly preferred. For instance, given one Labrador as an example of a "fep", it is unclear whether "fep" refers to Labradors, dogs, mammals, or animals. But if given three Labradors as the first three examples of "fep", it would be quite surprising – a highly suspicious coincidence – if "fep" in fact referred to a much more general class such as all dogs.

The same problem is depicted more abstractly in the dot diagram in Figure 3. Superordinate hypotheses (e.g., "animal") are represented as larger rectangles. Sometimes they fully enclose smaller rectangles (corresponding to more subordinate hypotheses), just as the extension of "animals" includes all Labradors. Sometimes they can also cross-cut each other, just as the extension of "pets" includes many (but not all) Labradors. The smaller rectangles represent hypotheses with smaller extensions, and we can use this to understand how Bayesian reasoning captures the notion of a suspicious coincidence, explaining the tendency to increasingly favor the smallest hypothesis that is consistent with the data as the number of data points increases.

This ability emerges due to the likelihood $p(d|h)$, the probability of observing the

data d assuming hypothesis h is true. In general, more restrictive hypotheses, corresponding to smaller regions in the data space, receive more likelihood for a given piece of data. If a small hypothesis is the correct extension of a word, then it is not too surprising that the examples occur where they do; a larger hypothesis could be consistent with the same data points, but explains less well exactly why the data fall where they do. The more data points we observe falling in the same small region, the more of a suspicious coincidence it would be if in fact the word's extension corresponded to a much larger region.

More formally, if we assume that data are sampled uniformly at random from all cases consistent with the concept, then the probability of any single data point d consistent with h is inversely proportional to the size of the region h picks out – call this the "size of h." This is why when there is one data point, as in Figure 3(i), there is a slight preference for the most restrictive (smallest) hypothesis; however, the preference is only slight, because it could still easily have been generated by any of the hypotheses that include it. But if multiple data points are generated independently from the concept, as in Figure 3(ii), the likelihood of h with n consistent examples is inversely proportional to the size of h, raised to the nth power. Thus the preference for smaller consistent hypotheses over larger hypotheses increases exponentially with the number of examples, and the most restrictive consistent hypothesis is strongly favored. This assumption is often referred to as the size principle (Tenenbaum & Griffiths, 2001).

The math behind the size principle is best understood concretely if we think about the hypotheses as discrete subsets of possible objects we might observe, such as bags of colored marbles, rather than as continuous regions such as rectangular regions in a two-dimensional space. Suppose bag A contains two marbles (a red and a green) and bag B contains three (a red, a green, and a yellow). The probability of pulling the red marble out of bag A is $\frac{1}{2} = 0.5$, since there are two possible marbles to choose from. The probability of pulling the red marble out of bag B is $\frac{1}{3} = 0.33$ for similar reasons. Thus, if you know only that a red marble has been pulled out of a bag (but not which

bag it is), you might have a weak bias to think that it was pulled out of bag A, which is $\frac{0.5}{0.33} = 1.67$ times as likely as bag B.

Now suppose that someone draws out the following series of marbles, shaking the bag fully between each draw: red, green, red, green. At this point most people would be more certain that the bag is A. The size principle explains why. If the probability of pulling one red (or green) marble from bag A is $\frac{1}{2}$, the probability of pulling that specific series of marbles is $\frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = \frac{1}{2^4} = \frac{1}{16} = 0.0625$, since each draw is independent. By contrast, the probability of drawing those marbles from bag B is $\frac{1}{3} * \frac{1}{3} * \frac{1}{3} * \frac{1}{3} = \frac{1}{3^4} = \frac{1}{91} = 0.0109$. This means that bag A is now $\frac{0.0625}{0.0109} = 5.73$ times as likely as B. In essence, the slight preference for the smaller bag is magnified over many draws, since it becomes an increasingly unlikely coincidence for only red or green marbles to be drawn if there is also a yellow one in there. This can be magnified if the number of observations increases still further (e.g., consider observing a sequence of red, green, red, green, green, green, red, green, red, red, green) or the relative size of the bags changes (e.g., suppose the observations are still red, green, red, green, but that the larger bag contains six marbles, each of a different color, rather than three). In either case bag A is now preferred to bag B by over a factor of 80, and there is little doubt that the marbles were drawn from bag A. In a similar way, a small hypothesis makes more precise predictions; thus, if the data are consistent with those predictions, then the smaller hypothesis is favored.

The size principle explains how it is possible to make strong inferences based on very few examples. It also captures the notion of a suspicious coincidence: as the number of examples increases, hypotheses that make specific predictions – those with more explanatory power – tend to be favored over those that are more vague. This provides a natural solution to the "no negative evidence" problem: deciding among hypotheses given positive-only examples. As the size of the data set approaches infinity, a Bayesian learner rejects larger or more overgeneral hypotheses in favor of more precise ones. With limited amounts of data, the Bayesian approach can make more subtle predictions, as

the graded size-based likelihood trades off against the preference for simplicity in the prior. The likelihood in Bayesian learning can thus be seen as a principled quantitative measure of the weight of implicit negative evidence – one that explains both how and when overgeneralization should occur.

The results of Xu and Tenenbaum (2007b) reflect this idea. Adults and 3- and 4-year-old children were presented with 45 objects distributed across three different superordinate categories (animals, vegetables, and vehicles), including many basic-level and subordinate-level categories within those. Subjects were then shown either one or three labelled examples of a novel word such as "fep", and were asked to pick out the other "feps" from the set of objects. Both children and adults responded differently depending on how many examples they were given. Just as in Figure 3, with one example, people and the model both showed graded generalization from subordinate to superordinate matches. By contrast, when given three examples, generalizations became much sharper and were usually limited to the most restrictive level.

This also illustrates how assumptions about the nature of the generative process affect the types of inferences that can be made. We have seen that people tend show restricted generalizations on the basis of three examples; however, this *only* if they think the experimenter was choosing those examples sensibly (i.e., as examples of the concept). If people think the data were generated in some other way – for instance, another learner was asking about those particular pictures – then their inferences change (Xu & Tenenbaum, 2007a). In this case, the lack of non-Labradors no longer reflects something the experimenter can control; though it is a coincidence, it is not a *suspicious* one. The data are the same, but the inference changes as the generative process underlying the data changes. In other words, the size principle applies in just those cases where the generative process is such that data are generated from the concept (or, more generally, hypothesis) itself.

So far we have illustrated how Bayesian inference can capture generalization from just a few examples, the simultaneous learning of overlapping extensions, and the use

of implicit negative evidence. All of these are important, but it is also true that we have built in a great deal, including a restricted and well-specified hypothesis space. Very often, human learners must not make reasonable specific generalizations within a set hypothesis space, they also much be able to make generalizations *about* what sort of generalizations are reasonable. We see an example of this in the next section.

3   Acquiring inductive constraints

One of the implications of classic problems of induction is the need for generalizations about generalizations, or inductive constraints, of some sort. The core problem is how induction is justified based on a finite sample of any kind of data, and the inevitable conclusion is that there must be some kind of constraint that enables learning to occur. Nearly every domain studied by cognitive science yields evidence that children rely on higher-level inductive constraints. Children learning words prefer to apply them to whole objects rather than parts (Markman, 1990). Babies believe that agents are distinct from objects in that they can move without contact (Spelke, Phillips, & Woodward, 1995) and act in certain ways in response to goals (Woodward, 1998; Gergely & Csibra, 2003). Confronted with evidence that children's behavior is restricted in predictable ways, the natural response is to hypothesize the existence of innate constraints, including the whole object constraint (Markman, 1990) core systems of object representation, psychology, physics, and biology (Carey & Spelke, 1996; Spelke & Kinzler, 2007; Carey, 2009), and so on. Given that they appear so early in development, it seems sensible to postulate that these constraints are innate rather than learned.

However, it may be possible for inductive constraints to be learned, at least in some cases. For instance, consider the problem of learning that some features "matter" for categorizing new objects while others should be ignored (e.g., Nosofsky, 1986). Acquiring higher-level abstract knowledge would enable one to make correct generalizations about an object from a completely novel category, even after seeing only one example. A wealth of research indicates that people are capable of acquiring this sort of knowl-

edge, both rapidly in the lab (Nosofsky, 1986; Perfors & Tenenbaum, 2009) and over the course of development (Landau, Smith, & Jones, 1988; L. Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Children also acquire other sorts of inductive constraints over the course of development, including the realization that categories may be organized taxonomically (Rosch, 1978), that some verbs occur in alternating patterns and others don't (e.g., Pinker, 1989) or that comparative orderings should be transitive (Shultz & Vogel, 2004).

How can an inductive constraint be learned, and how might a Bayesian framework explain this? Is it possible to acquire an inductive constraint faster than the specific hypotheses it is meant to constrain? If not, how can we explain people's learning in some situations? If so, what principles explain this acquisition?

A familiar example of the learning of inductive constraints was provided by Goodman (1955). Suppose we have many bags of colored marbles and discover by drawing samples that some bags seem to have black marbles, others have white marbles, and still others have red or green marbles. Every bag is uniform in color; no bag contains marbles of more than one color. If we draw a single marble from a new bag in this population and observe a color never seen before – say, purple – it seems reasonable to expect that other draws from this same bag will also be purple. Before we started drawing from any of these bags, we had much less reason to expect that such a generalization would hold. The assumption that color is uniform within bags is a learned *overhypothesis*, an acquired inductive constraint. The ability to infer such a constraint is not in itself a solution to the ultimate challenges of induction; it rests on other, arguably deeper assumptions – that the new bag is like the previous bags we have seen in relevant ways. Yet it is still a very useful piece of abstract knowledge that guides subsequent generalizations and can itself be induced from experience.

We can illustrate a similar idea in the rectangle world by imagining a learner who is shown the schematic data in Figure 4(i). Having seen point a only, the learner has no way to decide whether b or c is more likely to be in the same category or region as
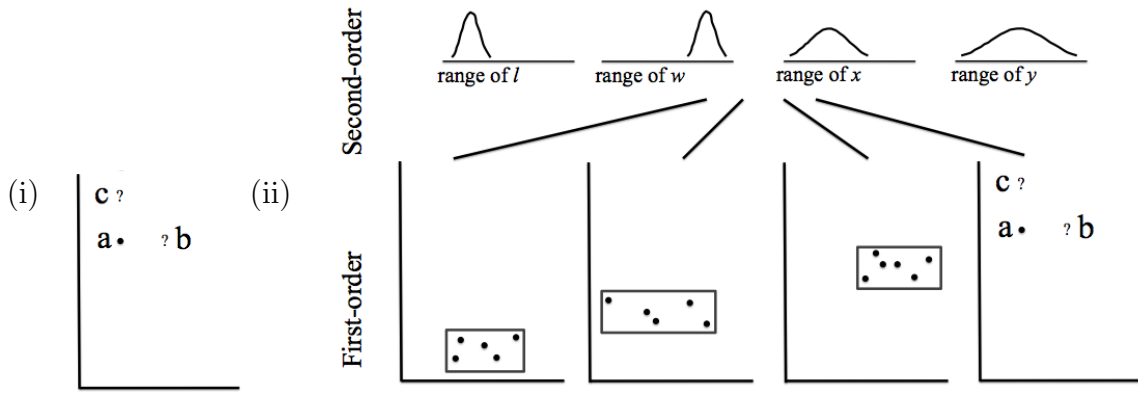
Figure 4: Learning higher-order information. (i) Given point a, one cannot identify whether b or c is more likely. (ii) Given additional data, a model that could learn higher-order information about hypotheses might favor regions that tend to be long, thin rectangles oriented along the y axis (i.e., regions for which the length l tends to be short, the width w tends to be long, and the location (x and y coordinates) can be nearly anywhere). If this is the case, points a and b are probably within the same region, but a and c are not.

a. However, if the learner has also seen the data in Figure 4(ii), they might infer both first-order and second-order knowledge about the data set. First-order learning refers to the realization that the specific rectangular regions constitute the best explanation for the data points seen so far; second-order (overhypothesis) learning would involve realizing that the regions tend to be long, thin, and oriented along the y-axis. Just as learning the how categories are organized helps children generalize from new items, this type of higher-order inference helps with the interpretation of novel data, leading to the realization that point b is probably in the same region as a but point c is not, even though b and c are equidistant from a.

A certain kind of Bayesian model, known as a hierarchical Bayesian model (HBM), can learn overhypotheses by not only choosing among specific hypotheses, but by also making higher-order generalizations about those hypotheses. As we've already seen, in a non-hierarchical model, the modeler sets the range of the parameters that define the hypotheses. In a hierarchical model, the modeler instead specifies *hyperparameters* – parameters defining the parameters – and the model learns the range of the parameters themselves. So rather than being given that the range of each of the the lower corner

(x, y), length l, and width w values lies between 0 and 10^9, a hierarchical model might instead learn the typical range of each (e.g., that l tends to be short while w tends to be long, as depicted in Figure 4(ii)) while the modeler specifies the range of the ranges.

The idea of wiring in abstract knowledge at higher levels of hierarchical Bayesian models may seem reminiscent of nativist approaches, but several key features fit well with empiricist intuitions about learning. The top level of knowledge in an HBM is prespecified, but every level beneath that can be learned. As one moves up the hierarchy, knowledge becomes increasingly abstract and imposes increasingly weak constraints on the learners specific beliefs at the lower levels. Thus, a version of the model that learns at higher levels builds in weaker constraints than a version that learns only at lower levels. By adding further levels of abstraction to an HBM while keeping pre-specified parameters to a minimum, at the highest levels of the model, we can come increasingly close to the classical empiricist proposal for the bottom-up, data-driven origins of abstract knowledge.

Although the precise mathematical details of any HBM are too complex to go into detail here, we can give a simplified example designed to motivate how it might be possible to learn on multiple levels simultaneously. Imagine you are faced with the marble example described earlier. We can capture this problem by saying that for each bag b, you have to learn the distribution of colors in the bag: call this distribution θ_b. At the same time, you want to make two inferences about the sets of bags as a whole: how uniform colors tend to be within bags (call this α) and what sorts of colors exist overall (call this β). Here, α and β are the hyperparameters of each of the θ_b values, since how likely any particular bag is will depend on the higher-level assumptions about bags: if you think, for instance, that colors tend to be uniform within bags, then a bag with lots of marbles of different colors in it will be low probability relative to a bag with only one. We can use this fact to learn on the higher level as well. A Bayesian model that sees three bags, all uniform in color, will search for the setting of α, β, and θ that make the observed data most probable; this will correspond to α values that

tend to favor uniformity within bags, and β and θ values that capture the observed color distributions. The Bayesian model learns these things simultaneously in the sense that it seeks to maximize the joint probability of all of the parameters, not just the lower-level ones.

This example is a simplified description of one of the existing hierarchical Bayesian models for category learning (Kemp, Perfors, & Tenenbaum, 2007); there are several other HBMs for the same underlying problem (Navarro, 2006; Griffiths, Sanborn, Canini, & Navarro, 2008; Heller, Sanborn, & Chater, 2009). Though they differ in many particulars, what all of these models have in common is that they can perform inference on multiple levels of abstraction. When presented with data consisting of specific objects and features, these models are capable of making generalizations about the specific objects as well as the appropriate generalizations about categorization in general. For instance, children in an experiment by L. Smith et al. (2002) were presented with four novel concepts and labels and acquired a bias to assume not only that individual categories like chairs tend to be organized by shape, but also that categories of solid artifacts *in general* are as well. A hierarchical Bayesian model can make the same generalization on the basis of the same data (Kemp, Perfors, & Tenenbaum, 2007).

A surprising effect of learning in hierarchical models is that, quite often, the higher-order abstractions are acquired before all of the specific lower-level details: just as children acquire some categorization biases even before they have learned all categories, the model might infer parameter values such that l tends to be short and w tends to be long, significantly before the size and location of each rectangular region is learned with precision. This effect, which we might call the "blessing of abstraction"[2], is somewhat counterintuitive. Why are higher-order generalizations like this sometimes easier for a Bayesian learner to acquire?

One reason is that the higher-level hypothesis space is often smaller than the lower-level ones. As a result, the model has to choose between fewer options at the higher level, which may require less evidence. In our rectangle example, the higher-level knowledge

---

[2]We thank Noah Goodman for this coinage.

may consist of only three options: l and w are approximately equal, l is smaller than w, or w is smaller than l. Even if a learner doesn't know whether l is 10 units or 11 units long and w is 20 or 22, it might be fairly obvious that l is smaller than w.

More generally, the higher-level inference concerns the lower-level hypothesis space (and is therefore based on the data set as a whole), whereas the lower-level inference is only relevant for specific data points. A single data point is informative only about the precise size and location of a single region. However, it – and every other single data point – is informative about all of the higher-level hypotheses. There is, in effect, more evidence available to the higher levels than the lower ones, and they can therefore be learned quite quickly.

Is there empirical evidence that people acquire higher-level abstract knowledge at the same time as, or before, lower-level specific knowledge? Adult laboratory category learning studies indicate that generalizations on the basis of abstract knowledge occurs at least as rapidly as lower-level generalizations (Perfors & Tenenbaum, 2009). There is also some indication that children show an abstract to concrete shift in both biological knowledge (Simons & Keil, 1995) and categorization, tending to differentiate global, super-ordinate categories before basic level kinds (Mandler & McDonough, 1993). Even infants have been shown to have the capacity to form overhypotheses given a small amount of data, providing initial evidence that the mechanisms needed for rapidly acquired inductive constraints exist early in development (Dewar & Xu, in press).

There is also a great deal of research that demonstrates the existence of abstract knowledge before any concrete knowledge has been acquired. For instance, the core knowledge research program suggests that before children learn about many specific physical objects or mental states, they have abstract knowledge about physical objects and intentional agents in general (e.g., Spelke & Kinzler, 2007). The core knowledge view suggests that the presence of this abstract knowledge so early in development, and before the existence of specific knowledge, implies that the abstract knowledge must be innate in some meaningful sense. More broadly, the basic motivation for positing innate

constraints on cognitive development is that without these constraints, children would be unable to infer the specific knowledge that they seem to acquire from the limited data available to them. What is critical to the argument is that some constraints are present prior to learning some of the specific data, not that those constraints must be innate. Approaches to cognitive development that emphasize learning from data typically view the course of development as a progressive layering of increasingly abstract knowledge on top of more concrete representations; under such a view, learned abstract knowledge would tend to come in after more specific concrete knowledge is learned, so the former could not usefully constrain the latter.

This view is sensible in the absence of explanations that can capture how abstract constraints could be learned together with (or before) the more specific knowledge they are needed to constrain. However, the hierarchical Bayesian framework provides such an explanation (or, at minimum, evidence that such a thing is possible). A model with the capability of acquiring abstract knowledge of a certain form[3] can identify *what* abstract knowledge is best supported by the data by learning which values of hyper-parameters (like α and β) are the most probable given the data seen so far. If an abstract generalization like this can be acquired very early and can function as a constraint on later acquisition of specific data, it may function effectively as if it were an innate domain-specific constraint, even if it is in fact not innate and instead is acquired by domain-general induction from data.

In sum, then, hierarchical Bayesian models offer a valuable tool for exploring questions of innateness due to the ability to limit built-in knowledge to increasingly abstract levels and thereby learn inductive constraints at other levels. As we will see in the next section, the Bayesian framework is also a useful way of approaching these questions for another reason – their ability to evaluate the rational tradeoff between the simplicity of a hypothesis and its goodness-of-fit to the evidence in the world. Because of this, Bayesian learners can make inferences that otherwise appear to go beyond the amount

---

[3]See Section for a more thorough discussion of how this degree of supervision is consistent with a non-innatist perspective, and is in fact impossible to avoid in any model of learning.

of evidence available.

4 Developing inductive frameworks

The hierarchical Bayesian models described above explain the origins of inductive biases and constraints by tuning priors in response to data observed from multiple settings or contexts. But the acquisition of abstract knowledge often appears more discrete or qualitative – more like constructing an appropriate hypothesis space, or selecting an appropriate hypothesis space from a higher level "hypothesis space of hypothesis spaces". Consider the "theory theory" view of cognitive development. Children's knowledge about the world is organized into intuitive theories with a structure and function analogous to scientific theories (Carey, 1985; Gopnik & Meltzoff, 1997; Karmiloff-Smith, 1988; Keil, 1989). The theory serves as an abstract framework that guides inductive generalization at more concrete levels of knowledge, by generating a space of hypotheses. Intuitive theories have been posited to underlie real-world categorization (Murphy & Medin, 1985), causal induction (Waldmann, 1996; Griffiths & Tenenbaum, 2009), biological reasoning (Atran, 1995; Inagaki & Hatano, 2002; Medin & Atran, 1999), physical reasoning (McCloskey, 1983) and social interaction (Nichols & Stich, 2003; Wellman, 1990). For instance, an intuitive theory of mind generates hypotheses about how a specific agent's behavior might be explained in particular situations – candidate explanations framed in terms of mental states such as goals, beliefs, or preferences. Under this view, cognitive development requires recognizing that a current theory of a domain is inadequate, and revising it in favor of a new theory that makes qualitative conceptual distinctions not made in the earlier theory (Carey, 1985; Gopnik, 1996). Probabilistic models provide a way to understand how such a process of theory change might take place, and in particular how a learner might weigh the explanatory power of alternative theories against each other. In the Bayesian framework, developmental change is a result of model selection; as data accumulate, eventually one theory becomes more likely than another, and the learner prefers a different one than before. In the

next section we describe how and why this transition occurs.

4.1   Trading off parsimony and goodness-of-fit

One of the most basic challenges in choosing between theories (or grammars, or other kinds of inductive frameworks) is trading off the parsimony, or simplicity, of a theory with how well it fits the observed data. To take a developmental example inspired by one of the papers that appears in this special issue (Lucas et al., submitted), we can imagine a child choosing between two theories of human choice behavior. Under one theory, everybody shares essentially the same assumptions about what kinds of things are desirable, such as having the same preferences for different kinds of food (and hence everybody has the same preferences as the child). Under the other theory, different people can possess different preferences. The developmental data suggest that a transition between these two theories occurs when children are between 14 and 18 months of age (Repacholi & Gopnik, 1997). However, the second theory is significantly more complex than the first, with the information required to specify the preferences of everybody the child knows increasing with the number of people. This extra complexity makes the theory more flexible, and thus better able to explain the pattern of choices a group of people might make. However, even if it were the case that everybody shared the same preferences, any random variation in people's choices could be explained by the more complex theory in terms of different people having different preferences. So, how can the child know when a particular pattern of choices should lead to the adoption of this more complex theory?

Developing intuitive theories requires trading off parsimony with goodness-of-fit. A more complex theory will always fit the observed data better, and thus needs to be penalized for its additional flexibility. While our example focuses on the development of theories of preference, the same problem arises whenever theories, grammars or other inductive frameworks that differ in complexity need to be compared. Just as a higher-order polynomial is more complicated but can fit a data set more precisely, so too can a

highly expressive theory or grammar, with more internal degrees of freedom, fit a body of data more exactly. How does a scientist or a child recognize when to stop positing ever more complex epicycles, and instead adopt a qualitatively different theoretical framework? Bayesian inference provides a general-purpose way to formalize a rational tradeoff between parsimony and fit.

As we saw earlier, goodness-of-fit for a hypothesis h is captured by the likelihood term in Bayes' Rule, or $P(d|h)$, while the prior $P(h)$ reflects other sources of a learner's beliefs. Priors can take various forms, but in general, a preference for simpler or more parsimonious hypotheses will emerge naturally without having to be engineered deliberately. This preference derives from the generative assumptions underlying the Bayesian framework, in which hypotheses are themselves generated by a stochastic process that produces a space of candidate hypotheses and $P(h)$ reflects the probability of generating h under that process.

To illustrate, consider the three hypotheses shown in Figure 5. We expand on our previous example by now stipulating that individual hypotheses may include more than one rectangular subregion. As a result, hypotheses are generated by first choosing a number of rectangular subregions and then choosing l, w, x, and y for each subregion. The first choice of how many subregions could be biased towards smaller numbers, but it need not be. Simpler hypotheses, corresponding to those with fewer subregions, would still receive higher prior probability because they require fewer choice points in total to generate. The simplest hypothesis A, with one subregion, can be fully specified by making only four choices: l, w, x, and y. Hypothesis C, at the other extreme, contains sixteen distinct rectangular subregions, and therefore requires 64 separate choices to specify, four for each subregion. Intuitively, the more complicated a pattern is, the more "accidental" it is likely to appear; the more choices a hypothesis requires, the more likely it is that those choices could have been made in a different way, resulting in an entirely different hypothesis. More formally, because the prior probability of a hypothesis is the product of the probabilities for all choices needed to generate it, and
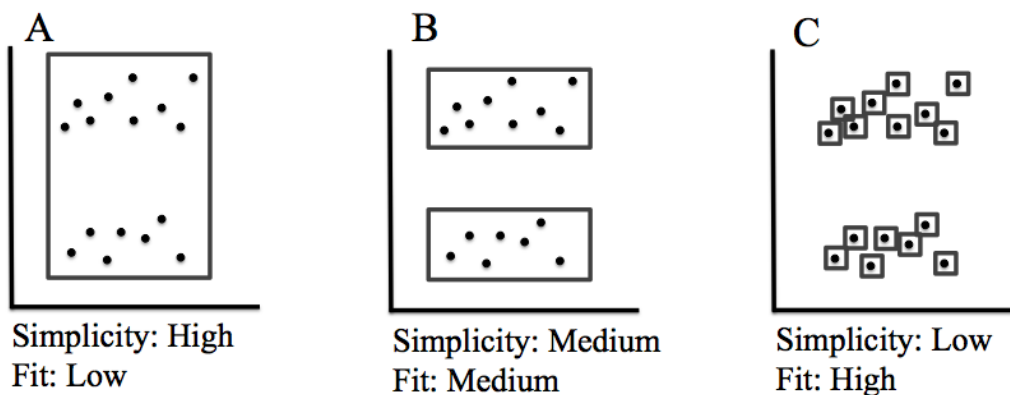
Figure 5: Hypothesis A is too simple, fitting the observed data poorly; C fits closely but is too complex; while B is "just right." A Bayesian analysis naturally ensures that the best explanation of the data optimizes a tradeoff between complexity and fit, as in B.

the probability of making any of these choices in a particular way must be less than one, a hypothesis specified by strictly more choices will in general receive strictly lower prior probability.

There are other ways of generating the hypotheses shown in Figure 5 – for instance, we could choose the upper-right and lower-left corners of each rectangular subregion, rather than choosing one corner, a height and a width. These might generate quantitatively different prior probabilities but would still give a qualitatively similar tradeoff between complexity and fit. The "Bayesian Ockham's razor" (MacKay, 2003) thus removes much of the subjectivity inherent in assessing simplicity of an explanation.[4] Note that in any of these generative accounts where hypotheses are generated by a sequence of choices, earlier or higher-up choices tend to play more important roles because they can affect the number and the nature of choices made later on or lower down. The

---

[4]That said, it is always possible to imagine bizarre theories, generating hypotheses from very different primitives than we typically consider, in which hypotheses that are intuitively more complex receive higher (not lower) prior probabilities. For instance, suppose that the hypotheses shown in Figure 5 were generated not by choosing the dimensions of one or more rectangles from some generic distribution, but by starting with just the twenty-one small rectangles in Figure 5C, and then making choices about whether to add or remove rectangles to or from this set. In that case, hypothesis C would have higher prior probability than A or B. Because the simplicity of a hypothesis is only meaningful relative to the primitives out of which hypotheses are generated, the decision of which primitives to include in a probabilistic model of cognition is a crucial choice, which we consider in more depth later. For now, we simply note that this is a key concern for any cognitive modeler, Bayesian or otherwise inclined. It can be seen as a virtue of the Bayesian framework that it forces us to make these choices and their consequences explicit, and that it provides a tool to evaluate the primitives we choose.

initial choice of how many rectangular subregions to generate determines how many choices about positions and side lengths are made later on. Perhaps we could have also chosen initially to generate circular subregions instead of rectangles; then each subregion would involve only three choices rather than four.

The same general logic applies to cognitively interesting hypothesis spaces, not just circles and rectangles: for instance, more complex grammars incorporate more rules and non-terminals (and therefore more choices are involved in specifying each one), and more complex causal theories involve more hidden causes or a greater degree of specification about the form that the causal relationship takes. These higher-level "choices that control choices" characterize the learner's "hypothesis space of hypothesis spaces"; they embody a more discrete, qualitative version of the hierarchical Bayesian ideas introduced in the previous section. They capture the role that intuitive theories or grammars play in providing frameworks for inductive inference in cognition, or the analogous role that higher-level frameworks or paradigms play in scientific theory building (Henderson, Goodman, Tenenbaum, & Woodward, 2010).

The logic outlined in the preceding paragraphs has been used to analyze developmental theory transitions in several settings. Elsewhere in this issue, Lucas et al. (submitted) show that the change from believing everybody shares the same preferences (analogous to hypothesis A in Figure 5) to believing everybody has different preferences (analogous to hypothesis C in Figure 5) can be produced simply by providing more data, a mechanism that we discuss in more detail in the next section. Goodman et al. (2006) show that the same approach can be used to explain the development of understanding of false beliefs, with a theory in which the beliefs that people maintain are influenced by their access to information being more complex but providing a better fit to the data than a theory without this principle. Schmidt, Kemp, and Tenenbaum (2006) demonstrated that a high-level theory about the properties of semantic predicates known as the M-constraint (essentially the constraint that predicates respect the structure of an ontological hierarchy; Sommers, 1971; Keil, 1979) can be induced from

linguistic data consistent with that theory, providing an alternative to the idea that this constraint is innate. Perfors, Tenenbaum, and Regier (2006) and Perfors et al. (submitted) reanalyze one version of a famous "poverty of stimulus" argument, and demonstrates that highly abstract and universal features of language – in particular, the principle that grammars incorporate hierarchical phrase structure – need not be built in as a language-specific bias but instead can be inferred on the basis of only a few hours of child-directed speech, given certain reasonable assumptions. This is because hierarchical grammars offer a more parsimonious explanation of the observed sentences: the grammars are shorter, with fewer non-terminals and fewer rules – that is, fewer choice points.

4.2 Adapting Ockham's Razor to the data

A key advantage of Bayesian approaches over earlier approaches to selecting grammars or theories based on data can be seen in how they adapt the preference for simpler hypotheses as the amount of observed data increases. In language acquisition, a traditional solution to the problem of constraining generalizing in the absence of negative evidence is the Subset Principle (Wexler & Culicover, 1980; Berwick, 1986): learners should choose the most specific grammar consistent with the observed data. In scientific theorizing, the classical form of Ockham's Razor speaks similarly: entities should not be multiplied beyond necessity. The difficulty with these approaches is that because their inferential power is too weak, they require additional constraints in order to work – and those constraints often apply only in a way we can recognize *post hoc*. In Figure 5, for instance, the preference for hypothesis B over A can be explained by the Subset Principle, but to explain why B is better than C (a subset of B), we must posit that C is ruled out *a priori* by some innate constraint; it is just not a natural hypothesis and should never be learnable, regardless of the data observed.

A Bayesian version of Ockham's Razor, in contrast, will naturally modulate the tradeoff between simplicity and goodness-of-fit based on the available weight of data,

even if the data are always generated by the same underlying process. This adaptiveness is intuitively sensible and critical for human learning. Consider Figure 6, which shows three data sets generated from the same underlying process but varying in the amount of data observed. The best hypothesis fits the five data points in data set 1 quite loosely, but because there are so few points this does not impose a substantial penalty relative to the high prior probability of the hypothesis. Analogously, early on in development children's categories, generalizations, and intuitive theories are likely to be coarser than those of adults, blurring distinctions that adults consider highly relevant and therefore being more likely to over-generalize.[5] As data accumulate, the relative penalty imposed for poor fit is greater, since it applies to each data point that is not predicted accurately by the hypothesis. More complex hypotheses become more plausible, and even hypothesis C that looked absurd on the data in Figure 5 could become compelling given a large enough data set, containing many data points all clustered into the sixteen tiny regions exactly as the theory predicts. The Subset Principle is not flexible in the same way. Being able to explain the process of development, with different theories being adopted by the child at different stages, requires being able to adapt to the data. This property makes it possible for the gradual accumulation of data to be the driving force in theory change, as in the examples discussed above.

Looking at Figure 6, one might guess that as the data increase, the most complex

---

[5]Adopting a sequence of ever more complex theories as the relevant data come to light seems like a plausible account of cognitive development, but it appears to be at odds with the familiar phenomenon of U-shaped learning curves (e.g., Marcus et al. (1992); see also Siegler (2004) for an overview). A U-shaped learning pattern occurs when a learner initially appears to have correctly acquired some piece of knowledge, producing it without error, but then follows this by an interval of incorrect performance marked by overgeneralization before eventually self-correcting. It may be possible to understand U-shaped acquisition patterns by considering a learner who can simply memorize individual data points in addition to choosing among hypotheses about them. In our example, memorizing a data point would require two choices to specify – its x and y coordinates – but even the simplest hypothesis would require at least four (x, y, l, and w). Moreover, a single data point also has the highest possible likelihood, since it predicts the data (itself) exactly. A data set with only one or a few data points, therefore, would be preferred in both the prior *and* the likelihood. Only as the number of data points increases would the penalty in the prior become high enough to preclude simply memorizing each data point individually: this is when overgeneral, highly simple hypotheses begin to be preferred. Thus, whether a U-shaped pattern occurs depends on the tradeoff in complexity that it takes to represent individual data points as opposed to entire hypotheses: if it is cheaper to memorize a few data points, then that would have both a higher prior and likelihood than would an extremely vague, overly general hypothesis.
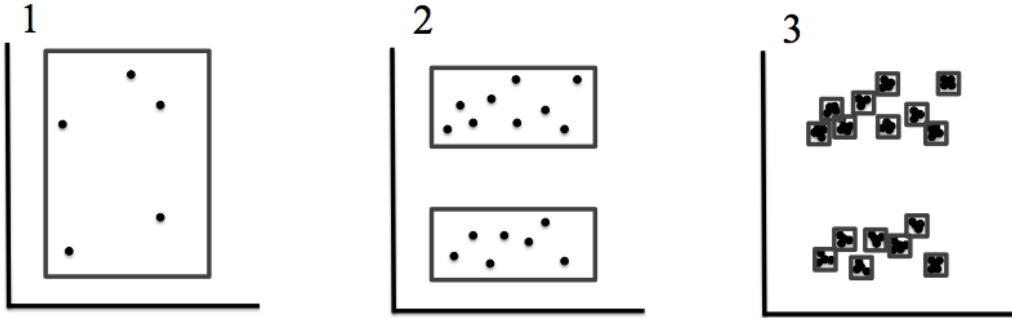
Figure 6: Role of data set size. Three datasets with increasing numbers of data points and their corresponding best hypotheses. For dataset 1, there are so few data points that the simplicity of the hypothesis is the primary consideration; by dataset 3, the preferred hypothesis is one that fits the clustered data points quite tightly.

hypotheses will eventually always be preferred. This is not true in general, although as data accumulate a Bayesian learner will tend to consider more complex hypotheses. Yet the preferred hypothesis will be that which best trades off between simplicity and goodness-of-fit, and ultimately, this will be the hypothesis that is closest to the true generative process (MacKay, 2003).[6] In other words, if the data are truly generated by a process corresponding to twenty-one different rectangular regions, then the points will increasingly clump into clusters in those regions, and hypothesis C will eventually be preferred. But if the data are truly generated by a process inhabiting two larger regions, then hypothesis B would still have a higher posterior probability as more data accumulate.

---

[6]Technically, this result has been proven for information-theoretic models in which probabilities of data or hypotheses are replaced by the lengths (in bits) of messages that communicate them to a receiver. The result is known as the "MDL Principle" (Rissanen, 1978), and is related to Kolmogorov complexity (Solomonoff, 1964; Kolmogorov, 1965). The Bayesian version applies given certain assumptions about the randomness of the data relative to the hypotheses and the hypotheses relative to the prior (Vitànyi & Li, 2000). Both versions apply only to the hypotheses in the hypothesis space: if no hypothesis corresponding to the true data generating process exists in the space, then it will never be considered, much less ultimately preferred. Thus, the hypothesis that is preferred by the model in the limit of infinite data is the "best" hypothesis only in the sense that it is closest to the true data generating process out of all of the hypotheses considered.

5 Discussion

Several issues are typically raised when evaluating Bayesian modelling as a serious computational tool for cognitive science. Bayesian reasoning characterizes "optimal" inference: what does this mean? How biologically plausible are these models, and how much does this matter? And finally, where does it all come from – the hypothesis space, the parameters, the representations? The answers to each of these questions affect what conclusions about actual human cognition we can draw on the basis of Bayesian models; we therefore consider each in turn.

5.1 Optimality: What does it mean?

Bayesian probability theory[7] is not simply a set of *ad hoc* rules useful for manipulating and evaluating statistical information: it is also the set of unique, consistent rules for conducting plausible inference (Jaynes, 2003). In essence, it is a extension of deductive logic to the case where propositions have degrees of truth or falsity – that is, it is identical to deductive logic if we know all the propositions with 100% certainty. Just as formal logic describes a deductively correct way of thinking, Bayesian probability theory describes an inductively correct way of thinking. As Laplace (1816) said, "probability theory is nothing but common sense reduced to calculation."

What does this mean? If we were to try to come up with a set of desiderata that a system of "proper reasoning" should meet, they might include things like consistency and qualitative correspondence with common sense – if you see some data supporting a new proposition A, you should conclude that A is more plausible rather than less; the more you think A is true, the less you should think it is false; if a conclusion can be

---

[7]Bayesian methods are often contrasted to so-called "frequentist" approaches, which are the basis for many of the standard statistical tests used in the social sciences, such as t-tests. Although frequentist methods often correspond to special cases of Bayesian probability theory, Bayesian methods have historically been relatively neglected, and often attacked, in part because they are viewed as unnecessarily subjective. This perception is untrue – Bayesian methods are simply more explicit about the prior information they take into account. Regardless, the issue of subjectivity seems particularly irrelevant for those interested in modelling human cognition, where accurately capturing "subjective belief" is part of the point.

reasoned multiple ways, its probability should be the same regardless of how you got there; etc. The basic axioms and theorems of probability theory, including Bayes' Rule, emerge when these desiderata are formalized mathematically (Cox, 1946, 1961), and correspond to common-sense reasoning and the scientific method. Put another way, Bayesian probability theory is "optimal inference" in the sense that a non-Bayesian reasoner attempting to predict the future will always be out-predicted by a Bayesian reasoner in the long run (de Finetti, 1937).

Even if the Bayesian framework captures optimal inductive inference, does that mean it is an appropriate tool for modelling human cognition? After all, people's everyday reasoning can be said to be many things, but few would aver that it is always optimal, subject as it is to emotions, heuristics, and biases of many different sorts (e.g., Tversky & Kahneman, 1974). However, even if humans are non-optimal thinkers in many ways – and there is no reason to think they are in *every* way – it is impossible to know this without being able to precisely specify what optimal thinking would amount to. Understanding how humans *do* think is often made easier if one can identify the ways in which people depart from the ideal: this is approximately the methodology by which Kahneman and Tversky derived many of their famous heuristics and biases, and the flexibility of the Bayesian approach makes it relatively easy to incorporate constraints based on memory, attention, or perception directly into one's model.

Many applications of Bayesian modelling operate on the level of computational theory (Marr, 1982), which seeks to understand cognition based on what its goal is, why that goal would be appropriate, and the constraints on achieving that goal, rather than precisely how it is implemented algorithmically. Understanding at this level is important because the nature of the reasoning may often depend more on the learner's goals and constraints than it does on the particular implementation. It can also enhance understanding at the other levels: for instance, analyzing connectionist networks as an implementation of a computational-level theory can elucidate what sort of computations they perform, and often explain why they produce the results they do (Hertz, Krogh,

& Palmer, 1991; MacKay, 2003).

Being able to precisely specify and understand optimal reasoning is also useful for performing ideal learnability analysis, which is especially important in the area of cognitive development. What must be "built into" the newborn mind in order to explain how infants eventually grow to be adult reasoners, with adult knowledge? One way to address this question is to establish the bounds of the possible: if some knowledge couldn't possibly be learned by an optimal learner presented with the type of data children receive, it is probably safe to conclude either that actual children couldn't learn it either, or that some of the assumptions underlying the model are inaccurate. The tools of Bayesian inference are well-matched to this sort of problem, both because they force modelers to make all of these assumptions explicit, and also because of their representational flexibility and ability to calculate optimal inference.

That said, not all Bayesian models operate on the computational level, and not all Bayesian models strive to capture optimal inference. Rational process models (see, e.g., Doucet, Freitas, & Gordon, 2001; Sanborn, Griffiths, & Navarro, 2010) are Bayesian models that focus on providing *approximations* to optimal reasoning. As such, they span the algorithmic and computational level, and can provide insight into how a resource-limited learner might reason. Likewise, much work in computational neuroscience focuses on the implementational level, but is Bayesian in character (e.g., Pouget, Dayan, & Zemel, 2003; T. Lee & Mumford, 2003; Zemel, Huys, Natarajan, & Dayan, 2005; Ma, Beck, Latham, & Pouget, 2006; Doya, Ishii, Pouget, & Rao, 2007; Rao, 2007). We discuss the implications of this work in the next section.

5.2 Biological plausibility

Because cognitive scientists are ultimately interested in understanding human cognition, and human cognition is ultimately implemented in the brain, it is important that our computational-level explanations be realizable on the neurological level, at least potentially. This is one reason for the popularity of the Parallel Distributed Process-

ing, or connectionist, approach, which was developed as a neurally inspired model of the cognitive process (Rumelhart & McClelland, 1986). Connectionist networks, like the brain, contain many highly interconnected, active processing units (like neurons) that communicate with each other by sending activation or inhibition through their connections. As in the brain, learning appears to involve modifying connections, and knowledge is represented in a distributed fashion over the connections. As a result, representations degrade gracefully with neural damage, and reasoning is probabilistic and "fuzzy" rather than all-or-none.

In contrast, Bayesian models may appear implausible from the neurological perspective. One of the major virtues of Bayesian inference – the transparency of its computations and the explicitness of its representation – is, in this light, potentially a major flaw: the brain is many wonderful things, but it is neither transparent nor explicit. How could structured symbolic representations like grammars or logics be instantiated in our neural hardware? How could our cortex encode hypotheses and compare them based on a tradeoff between their simplicity and goodness-of-fit? Perhaps most problematically, how could the brain approximate anything like optimal inference in a biologically realistic timeframe, when conventional algorithms for Bayesian inference running on conventional computing hardware take days or weeks to tackle problems that are vastly smaller than those the brain solves?

These are good questions, but there is growing evidence for the relevance of Bayesian approaches on the neural level (e.g., Doya et al., 2007). Probability distributions can in fact be represented by neurons, and they can be combined according to a close approximation of Bayes' Rule; posterior probability distributions may be encoded in populations of neurons in such a way that Bayesian inference is achieved simply by summing up firing rates (Pouget et al., 2003; Ma et al., 2006). Spiking neurons can be modelled as Bayesian integrators accumulating evidence over time (Deneve, 2004; Zemel et al., 2005). Recurrent neural circuits are capable of performing both hierarchical and sequential Bayesian inference (Deneve, 2004; Rao, 2004, 2007). Even specific brain

areas have been studied: for instance, there is evidence that the recurrent loops in the visual cortex integrate top-down priors and bottom-up data in such a way as to implement hierarchical Bayesian inference (T. Lee & Mumford, 2003).

This work, though still in its infancy, suggests that concerns about biological plausibility may not, in the end, prove to be particularly problematic. It may seem to us, used to working with serial computers, that searching these enormous hypothesis spaces quickly enough to perform anything approximating Bayesian inference is impossible; but the brain is a parallel computing machine made up of billions of highly interconnected neurons. The sorts of calculations that take a long time on a serial computer, like a sequential search of a hypothesis space, might be very easily performed in parallel. They also might not; but whatever the future holds, the indications so far serve as a reminder of the danger of advancing from the "argument from incredulity" to any conclusions about biological plausibility.

It is also important to note that, for all of their apparent biological plausibility, neural networks are unrealistic in important ways, as many modelers acknowledge. Units in neural networks are assumed to have both excitatory and inhibitory connections, which is not neurally plausible. This is a problem because the primary learning mechanism, backpropagation, relies on the existence of such connections (Rumelhart & McClelland, 1986; Hertz et al., 1991). There is also no analogue of neurotransmitters and other chemical transmission, which play an important role in brain processes (Gazzaniga, Ivry, & Mangun, 2002). These issues are being overcome as the state of the art advances (see Rao, Olshausen, and Lewicki (2002) for some examples), but for the models most commonly used in cognitive science – perceptrons, multilayered recurrent networks, and Boltzmann machines – they remain a relevant concern.

Different techniques are therefore biologically plausible in some ways and perhaps less so in others. Knowing so little about the neurological mechanisms within the brain, it is difficult to characterize how plausible either approach is or how much the ways they fall short impact their utility. In addition, biological plausibility is somewhat ir-

relevant on the computational level of analysis. It is entirely possible for a system to be emergently or functionally Bayesian, even if none of its step-by-step computations map onto anything like the algorithms used by current Bayesian models. Just as optimal decision-making can be approximated under certain conditions by simple heuristics (Goldstein & Gigerenzer, 2002), it may be possible that the optimal reasoning described by Bayesian models can be approximated by simple algorithms that look nothing like Bayesian reasoning in their mechanics. If so – in fact, even if the brain couldn't implement anything even heuristically approximating Bayesian inference – Bayesian models would still be useful for comprehending the goals and constraints faced by the cognitive system and comparing actual human performance to optimal reasoning. To the extent that neural networks are relevant to the computational level, the same is true for them.

5.3   Where does it all come from?

For many, a more important critique is that, in some sense, Bayesian models do not appear to be *learning* at all. The entire hypothesis space, as well as the evaluation mechanism for comparing hypotheses, has been given by the modeler; all the model does is choose among hypotheses that already exist. Isn't learning, particularly the sort of learning that children perform over the first years of their life, something more than this? Our intuitive notion of learning certainly encompasses a spirit of discovery that does not appear at first glance to be captured by a model that simply does hypothesis testing within an already-specified hypothesis space.

The same intuition lies at the core of Fodor's famous puzzle of concept acquisition (Fodor, 1975, 1981). His essential point is that one cannot learn anything via hypothesis testing because one must possess it in order to test it in the first place. Therefore, except for those concepts that can be created by composing them from others (which Fodor believes to be in the minority), all concepts (including CARBURETOR and GRANDMOTHER) must be innate.

To understand how this intuition can be misleading, it is helpful to make a distinc-

tion between two separate notions of what it means to build in a hypothesis space. A trivial sense is to equip the model with the representational capacity to represent any of the hypotheses in the space: if a model has this capacity, even if it is not currently evaluating or considering any given hypothesis, that hypothesis is in some sense latent in that space. Thus, if people have the capacity to represent some given hypothesis, we say it can be found in their *latent hypothesis space*. The ability to represent possible hypotheses in a latent hypothesis space is necessary for learning of any sort, in any model or being. We can contrast this with hypotheses that may be explicitly considered or evaluated – the hypotheses that can be actively represented and manipulated by the conceptual system – which we refer to as the *explicit hypothesis space*.

As an analogy, consider a standard English typewriter with an infinite amount of paper. There is a space of documents that it is capable of producing, which includes things like *The Tempest* and does not include, say, a Vermeer painting or a poem written in Russian. This typewriter represents a means of generating the hypothesis space for a Bayesian learner: each possible document that can be typed on it is a hypothesis, the infinite set of documents producible by the typewriter is the latent hypothesis space[8], and the documents that have actually been typed out so far make up the explicit hypothesis space. Is there a difference between documents that have been created by the typewriter and documents that exist only in the latent hypothesis space? Of course there is: documents that have been created can be manipulated in all sorts of ways (reading, burning, discussing, editing) that documents latent in the space cannot. In the same way, there may be a profound difference between hypotheses that have been considered by the learner and hypotheses that are simply latent in the space: the former can be manipulated by the cognitive system – evaluated, used in inference, compared to other hypotheses – but the latter cannot. Hypothesis generation would describe the process by which hypotheses move from the latent space to the explicit

---

[8]Note that the latent hypothesis space does not need to be completely enumerated in order to exist; it must simply be defined by some sort of process or procedure. Indeed, in practice, exhaustive hypothesis enumeration is intractable for all but the simplest models; most perform inference via guided search, and only a subset of the hypotheses within the space are actually evaluated.

space – the process by which our typist decides what documents to produce. Hypothesis testing would describe the process of deciding which of the documents produced should be preferred (by whatever standard). Learning, then, would correspond to the entire process of hypothesis generation and testing – and hence would never involve new hypotheses being added to the latent hypothesis space. This is what some critics object to: it doesn't "feel" like learning, since in some sense everything is already "built in."

However, this intuitive feeling is misleading. If we take "learning" to mean "learning in the Fodorian sense" or, equivalently, "not built into the latent hypothesis space", then there are only two conclusions possible. Either the hypotheses appear in the latent hypothesis space completely arbitrarily, or *nothing* can ever be learned. In other words, there is no interpretation of learning "in the Fodorian sense" that allows for an interesting computational model or theory of learning to emerge.

How is this so? Imagine that we could explain how a new hypothesis could be added to a latent hypothesis space; such an explanation would have to make reference to some rules or some kind of process for adding things. That process and those rules, however, would implicitly define a meta latent space of their own. And because this meta-space is pre-specified (implicitly, by that process or set of rules) in the exact same way the original hypothesis space was pre-specified (implicitly, by the original generative process), the hypotheses within it are "innate" in precisely the same way that the original hypotheses were. In general, the only way for something to be learned in the Fodorian sense – the sense that underlies this critique – is for them to be able to spring into a hypothesis space in such a way that is essentially random (i.e., unexplainable via some process or rule). If this is truly what learning is, it seems to preclude the possibility of studying it scientifically; but luckily, this is not what most of us generally mean by learning.

One implication is that *every* computational learning system – any model we build, and the brain if it can be understood as a kind of computer – must come equipped with a latent hypothesis space that consists of everything that it can possibly represent

and compute; all learning must happen within this space. This is not a novel or controversial point – all cognitive scientists accept that *something* must be built in – but it is often forgotten; the fact that hypothesis spaces are clearly defined within the Bayesian framework makes them appear more "innate" than if they were simply implicit in the model. But even neural networks – which are often believed to presume very little in the way of innate knowledge – implicitly define hypotheses and hypothesis spaces via their architecture, functional form, learning rule, etc. In fact, neural networks can be viewed as implementations of Bayesian inference (e.g., Funahashi, 1998; McClelland, 1998; MacKay, 2003), corresponding to a computational-level model whose hypothesis space is a set of continuous functions (e.g., Funahashi, 1989; Stinchcombe & White, 1989). This is a large space, but Bayesian inference can entertain hypothesis spaces that are equivalently large.

Does this mean that there is no difference between Bayesian models and neural networks? In one way, the answer is yes: because neural networks are universal approximators, it is always possible to construct one that approximates the input-output functionality of a specific Bayesian model. In practice, however, the answer is usually no: the two methods have very different strengths and weaknesses, and therefore their value as modelling tools varies depending on the questions being asked (see Griffiths, Chater, Kemp, Perfors, and Tenenbaum (2010) and McClelland et al. (2010) for a more thorough discussion of these issues). One difference is that connectionist models make certain commitments about representation that make it difficult to capture explicit symbolic knowledge, of the sort that is commonly incorporated into cognitive theories. Another difference relates to how the models trade off between simplicity and goodness-of-fit; in most Bayesian models, that tradeoff is (or approximates) optimality. By contrast, neural network models perform a similar tradeoff, but generally non-optimally and in a more *ad hoc* manner, avoiding overfitting by limiting the length of training and choosing appropriate weights, learning rules, and network architecture.[9]

---

[9]There is an interesting subfield called Bayesian neural networks studying how to construct models that make these choices for themselves, pruning connections in a Bayes-optimal way (e.g., MacKay,

In the Bayesian framework, what is built in is the generative process, which implicitly defines the assignment of prior probabilities, the representation, and the size of the hypothesis space; in the PDP framework, these things are built in through choices about the architecture, weights, learning rule, training procedure, etc.

It is therefore incorrect to say one framework assumes more innate knowledge than another: *specific models* within each may assume more or less, but it can be quite difficult to compare them precisely, in part because neural networks incorporate it implicitly. Which model assumes more innate knowledge is often not even the interesting question. A more appropriate one might be: *what* innate knowledge does it assume? Instead of asking whether one representation is a stronger assumption than another, it is often more productive to ask which predicts human behavior better. The answer will probably depend on the problem and the domain, but the great advantage of computational modelling is that it allows us to explore this dependence precisely.

5.4   Limitations of Bayesian models

Because of their combination of representational flexibility and powerful domain-general statistical learning mechanisms, Bayesian models are a useful tool for modeling in cognitive science and language acquisition. However, no approach can be all things to all people. What are some of the limitations of the Bayesian paradigm?

One of the most important is that Bayesian modeling is not an appropriate tool for every question. Bayesian models address inductive problems, which cover a large range of the problems in cognitive science. However, there are many important problems in cognitive science that are not obviously cast as inductive problems. For instance, many scientists are concerned with understanding how different cognitive characteristics are related to each other (for instance, IQ and attention), and how that changes over the lifespan. Bayesian models have also had little to say about emotional regulation or psychopathology. This is not to say that Bayesian models could not be applicable to

---

1995; Neal, 1994, 1996).

these problems, but to the extent that induction is not the central concern here, they are unlikely to be illuminating.

Another set of limitations stems from a general factor that afflicts any model or theory: if the assumptions behind that model or theory are wrong, then it will not describe human behavior. Broadly speaking, we see two key ways in which the assumptions underlying Bayesian modelling might be wrong. One would occur if it turns out that human behavior can only be explained by appealing to some hardware (or implementation-level, or biological) characteristics of the cognitive system. For instance, if some behavior emerges only because of the particular architecture of the brain or the way in which action potentials are propagated – and there is no computational-level explanation for why those aspects of the system should be the way they are – then Bayesian models would not be able to explain that behavior. Rational process models (Sanborn et al., 2010), which explore ways in which to approximate optimal inference, might explain some types of deviation from optimality, but not all.

The second major way that Bayesian modelling might be wrong is that it might make the wrong *computational*-level assumptions about the human mind. For instance, Bayesian models assume that a computational-level description of human inference should follow the mathematics of probability theory. Although doing so is rational for all of the reasons described earlier, it is still possible that human reasoners nevertheless do not do it (or even approximate it). If this is the case, Bayesian models would fail to match human behavior for reasons that could not be attributable the sorts of computational-level factors that are typically explored within the Bayesian modelling framework, like different specifications of the problem or the goal of the learner.

If there are places where Bayesian models err in either of these two ways, the most straightforward way to identify these places is to do exactly what the field is currently doing: pairing good experimental work with theoretical explorations of the capabilities of a broad array of Bayesian models. Acquiring solid empirical data about how humans behave is vital when evaluating the models, and systematically exploring the space

of models is vital for determining whether some behavior cannot be accounted for by such models. Thus, even if it is ultimately the wrong explanation for some behavior of interest, a Bayesian model may still be useful for identifying when that behavior departs from optimality, and clarifying how it departs, as a cue to its algorithmic basis.

A final limitation exists more in practice than in principle. As Bayesian models get increasingly complex, their computational tractability decreases dramatically. Currently, no Bayesian model exists that can deal with a quantity of input that is within orders of magnitude as much as a developing child sees over the course of a few years: the search space is simply too large to be tractable. Improvements in computer hardware (Moore's Law) and machine learning technologies will reduce this limitation over time; however, for now, it does mean that generating precise predictions on the basis of large amounts of data, especially when the domain is highly complex, is difficult. In fact, even effectively searching through extremely high-dimensional hypothesis spaces with multimodal posteriors (such as grammars) is currently effectively intractable. The problem of computational intractability on large problems is one that affects all computational models of learning, because the problems are intrinsically hard. We expect developments in computer hardware and machine learning technology over the coming years to offer dramatically new possibilities for Bayesian models of cognition and other approaches as well.

6 Conclusion

Bayesian models offer explanatory insights into many aspects of human cognition and development. The framework is valuable for defining optimal standards of inference, and for exploring tradeoffs between simplicity and goodness-of-fit that must guide any learner's generalizations from observed data. Its representational flexibility makes it applicable to a wide variety of learning problems, and its transparency makes it easy to be clear about what assumptions are being made, what is being learned, and why learning works.

7 Acknowledgements

References

Alishahi, A., & Stevenson, S. (2008). A computational model for early argument structure acquisition. *Cognitive Science*, *32*(5), 789-834.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychology Review*, *98*(3), 409–429.

Andrews, M., & Vigliocco, G. (2009). Learning semantic representations with Hidden Markov Topics Models. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *31st Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463-498.

Ashby, F., & Alfonso-Reese, L. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216-233.

Atran, S. (1995). Classifying nature across cultures. In E. Smith & D. Osherson (Eds.), *Thinking: An invitation to cognitive science* (Vol. 3, p. 131-174). Cambridge, MA: MIT Press.

Baker, C., Tenenbaum, J. B., & Saxe, R. (2007). Goal inference as inverse planning. In D. McNamara & J. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical

knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284-17289.

Berger, J. (1993). *Statistical decision theory and Bayesian analysis*. New York: Springer.

Berwick, R. (1986). Learning from positive-only examples: The subset principle and three case studies. *Machine Learning*, *2*, 625–645.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Box, G., & Tiao, G. (1992). *Bayesian inference in statistical analysis*. New York: Wiley.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

Carey, S. (1985). *Conceptual change in childhood*. Bradford Books, MIT Press.

Carey, S. (2009). *Origin of concepts*. Oxford University Press.

Carey, S., & Spelke, E. (1996). Science and core knowledge. *Philosophy of Science*, *63*, 515–533.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.

Cox, R. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics*, *14*, 1–13.

Cox, R. (1961). *The algebra of productive inference*. Baltimore, MD: Johns Hopkins University Press.

de Finetti, B. (1937). Prevision, its logical laws, its subjective sources. In H. Kyburg & H. Smokler (Eds.), *In studies in subjective probability* (2nd ed.). New York: J. Wiley and Sons.

de Finetti, B. (1974). *Theory of probability* (2nd ed.). New York: J. Wiley and Sons.

Deneve, S. (2004). Bayesian inference in spiking neurons. *Advances in Neural Infor-*

mation Processing Systems, *17*.

Dewar, K., & Xu, F. (in press). Induction, overhypothesis, and the origin of abstract knowledge: evidence from 9-month-old infants. *Psychological Science*.

Doucet, A., Freitas, N. d., & Gordon, N. (2001). *Sequential Monte Carlo in practice.* Springer-Verlag.

Dowman, M. (2000). Addressing the learnability of verb subcategorizations with Bayesian inference. In L. Gleitman & A. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Doya, K., Ishii, S., Pouget, A., & Rao, R. (2007). *Bayesian brain: Probabilistic approaches to neural coding.* Cambridge, MA: MIT Press.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*, 630-633.

Feldman, J., & Tremoulet, P. (2008). *The attribution of mental architecture from motion: Towards a computational theory* (Tech. Rep. No. RuCCS TR-87). Rutgers University.

Feldman, N., & Griffiths, T. L. (2009). Learning phonetic categories by learning a lexicon. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *31st Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Feldman, N., Morgan, J., & Griffiths, T. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782.

Fisher, R. (1933). Probability, likelihood, and quantity of information in the logic of uncertain inference. *Proceedings of the Royal Society*, *146*, 1–8.

Fodor, J. (1975). *The language of thought.* New York, NY: Thomas Y. Crowell Company.

Fodor, J. (1981). *Representations: Philosophical essays on the foundations of cognitive*

science. Cambridge, MA: MIT Press.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *Cognition*(28), 3-71.

Frank, M., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2007). Modeling human performance in statistical word segmentation. In D. McNamara & J. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Frank, M., Goodman, N., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578-585.

Frank, M., Ichinco, D., & Tenenbaum, J. B. (2008). Principles of generalization for learning sequential structure in language. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, *2*, 183–192.

Funahashi, K. (1998). Multilayer neural networks and bayes decision theory. *Neural Networks*, *11*(2), 209–213.

Gazzaniga, M., Ivry, R., & Mangun, G. (2002). *Cognitive neuroscience: The biology of the mind* (2nd ed.). New York, NY: W.W. Norton & Company.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis* (2nd ed.). Chapman & Hall.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naive theory of rational action. *Trends in Cognitive Science*, *7*(7), 287–292.

Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice.* Chapman & Hall.

Goldstein, D., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75-90.

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Interpolating between types and tokens by estimating power law generators. *Advances in Neural Information Processing Systems*, *18*.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2007). Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21-54.

Goodman, N. (1955). *Fact, fiction, and forecast.* Cambridge, MA: Harvard University Press.

Goodman, N., Baker, C., Bonawitz, E., Manisinghka, V., Gopnik, A., Wellman, H., et al. (2006). Intuitive theories of mind: A rational approach to false belief. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Goodman, N., Griffiths, T. L., Feldman, J., & Tenenbaum, J. B. (2007). A rational analysis of rule-based concept learning. In D. McNamara & J. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Gopnik, A. (1996). The scientist as child. *Philosophy of Science*, *63*, 485-514.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*(1), 1–30.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* Cambridge, MA: MIT Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357-364.

Griffiths, T. L., & Kalish, M. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*, 441-480.

Griffiths, T. L., Sanborn, A., Canini, K., & Navarro, D. (2008). Categorization as nonparameteric Bayesian density estimation. In M. Oaksford & N. Chater (Eds.),

The probabilistic mind: Prospects for Bayesian cognitive science. Oxford: Oxford University Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661-716.

Grünwald, P., Myung, J., & Pitt, M. (2005). *Advances in minimum description length: Theory and applications.* Cambridge, MA: MIT Press.

Heller, K., Sanborn, A., & Chater, N. (2009). Hierarchical learning of dimensional biases in human categorization. In *Advances in neural information processing systems* (Vol. 22). Cambridge, MA: MIT Press.

Henderson, L., Goodman, N., Tenenbaum, J. B., & Woodward, J. (2010). The structure and dynamics of scientific theories: A hierarchical Bayesian perspective. *Philosophy of Science*, *77*(2).

Hertz, J., Krogh, A., & Palmer, R. (1991). *Introduction to the theory of neural computation: Santa Fe Institute studies in the science of complexity* (Vol. 1). Reading, MA: Perseus Books.

Hsu, A., & Griffiths, T. L. (2009). Differential use of implicit negative evidence in generative and discriminative language learning. In *Advances in Neural Information Processing Systems* (Vol. 22). Cambridge, MA: MIT Press.

Inagaki, K., & Hatano, G. (2002). *Young children's thinking about biological world.* New York: Psychology press.

Jaynes, E. (2003). *Probability theory: The logic of science.* Cambridge: Cambridge University Press.

Jeffreys, H. (1939). *Theory of probability.* Oxford: Clarendon Press.

Karmiloff-Smith, A. (1988). The child is a theoretician, not an inductivist. *Mind and*

Language, *3*, 183-195.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective.* Cambridge, MA: Harvard University Press.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kemp, C., Goodman, N., & Tenenbaum, J. (2007). Learning causal schemata. In D. McNamara & J. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

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* (p. 672-677). Austin, TX: Cognitive Science Society.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, *114*(2), 165-196.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241-5245.

Kolmogorov, A. (1965). Three approaches to the quantitative definition of information. *Problems of Information and Transmission*, *1*(1), 1–7.

Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299–321.

Laplace, P. S. (1816). *A philosophical essay on probabilities.* Dover Publications.

Lee, M. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*(3), 555-580.

Lee, M., Fuss, I., & Navarro, D. (2007). A Bayesian approach to diffusion models of decision-making and response time. In B. Scholkopf, J. Platt, & T. Hoffman

(Eds.), *Advances in Neural Information Processing Systems* (Vol. 19, p. 809-815). Cambridge, MA: MIT Press.

Lee, P. (1997). *Bayesian statistics.* New York: Wiley.

Lee, T., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*, 1434–1448.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P., & Holyoak, K. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*, 955–984.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., et al. (submitted). The child as econometrician: A rational model of preference understanding in children. *Cognition*.

Ma, W., Beck, J., Latham, P., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

MacKay, D. (1995). Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, *6*, 469–505.

MacKay, D. (1998). Introduction to Monte Carlo methods. In M. Jordan (Ed.), *Learning in graphical models* (pp. 175–204). Cambridge, MA: MIT Press.

MacKay, D. (2003). *Information theory, inference, and learning algorithms.* Cambridge, UK: Cambridge University Press.

Mandler, J., & McDonough, L. (1993). Concept formation in infancy. *Cognitive Development*, *9*, 291–318.

Marcus, G., Pinker, S., Ullman, M., Hollander, M., Rosen, T., Xu, F., et al. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*, 1–178.

Markman, E. (1989). *Categorization and naming in children.* MIT Press.

Markman, E. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*, 57–77.

Marr, D. (1982). *Vision: A computational investigation into the human representation*

and processing of visual information. Henry Holt & Company.

McClelland, J. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 21–53). Oxford: Oxford University Press.

McClelland, J., Votvinick, M., Noelle, D., Plaut, D., Rogers, T., Seidenberg, M., et al. (2010). Letting structure emerge: Connectionist and dynamic systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348-356.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *284*, 114-123.

Medin, D. L., & Atran, S. (Eds.). (1999). *Folkbiology.* Cambridge, MA: Bradford Books.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289-316.

Navarro, D. (2006). From natural kinds to complex categories. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (p. 621-626). Austin, TX: Cognitive Science Society.

Navarro, D., & Griffiths, T. L. (2007). A nonparametric Bayesian method for inferring features from similarity judgments. *Advances in Neural Information Processing Systems*, *19*.

Navarro, D., Griffiths, T. L., Steyvers, M., & Lee, M. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101-122.

Neal, R. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). University of Toronto.

Neal, R. (1994). *Priors for infinite networks* (Tech. Rep. No. CRG-TR-94-1). University of Toronto.

Neal, R. (1996). *Bayesian learning for neural networks.* Springer.

Nichols, S., & Stich, S. (2003). *Mindreading: An integrated account of pretense, self-awareness and understanding other minds.* Oxford: Oxford University Press.

Nosofsky, J. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39-57.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge University Press.

Perfors, A., & Tenenbaum, J. B. (2009). Learning to learn categories. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (p. 136-141). Austin, TX: Cognitive Science Society.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2006). Poverty of the stimulus? A rational approach. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (p. 663-668). Austin, TX: Cognitive Science Society.

Perfors, A., Tenenbaum, J. B., & Regier, T. (submitted). The learnability of abstract syntactic principles. *Cognition*.

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*, 607-642.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure.* Cambridge, MA: MIT Press.

Pouget, A., Dayan, P., & Zemel, R. (2003). Inference and computation with population codes. *Annual Reviews in Neuroscience*, *26*, 381–410.

Rao, R. (2004). Bayesian computation in recurrent neural circuits. *Neural Computation*, *16*, 1–38.

Rao, R. (2007). Neural models of Bayesian belief propagation. In K. Doya, S. Ishii, A. Pouget, & R. Rao (Eds.), *Bayesian brain: Probabilistic approaches to neural coding* (pp. 239–267). Cambridge, MA: MIT Press.

Rao, R., Olshausen, B., & Lewicki, M. (2002). *Probabilistic models of the brain: Perception and neural function.* Cambridge, MA: MIT Press.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from

14- and 18-month-olds. *Developmental Psychology*, *33*(1), 12-21.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Rosch, E. (1978). Principles of categorization. In E. Rosch & L. B. (Eds.), *Cognition and categorization* (p. 27-48). Lawrence Erlbaum.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Roy, D., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2006). Learning annotated hierarchies from relational data. *Advances in Neural Information Processing Systems*, *19*.

Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

Sanborn, A., Griffiths, T. L., & Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*.

Savage, L. (1954). *Foundations of statistics.* New York, NY: J. Wiley & Sons.

Schmidt, L., Kemp, C., & Tenenbaum, J. B. (2006). Nonsense and sensibility: Inferring unseen possibilities. In R. Sun & N. Miyake (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., & Tenenbaum, J. B. (2006). Learning cross-cutting systems of categories. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Shultz, T., & Vogel, A. (2004). A connectionist model of the development of transitivity. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (p. 1243-1248).

Siegler, R. (2004). U-shaped interest in U-shaped development – and what it means. *Journal of Cognition and Development*, *5*(1), 1–10.

Simons, D., & Keil, F. (1995). An abstract to concrete shift in the development of biological thought: The *insides* story. *Cognition*, *56*, 129–163.

Sivia, D. (1996). *Data analysis: A Bayesian tutorial*. Oxford: Oxford University Press.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (p. 697-702). Austin, TX: Cognitive Science Society.

Smith, L., Jones, S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Sobel, D., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*, 303-333.

Solomonoff, R. (1964). A formal theory of inductive inference, parts 1 and 2. *Information and Control*, *7*(1–22), 224-254.

Sommers, F. (1971). Structural ontology. *Philosophia*, *1*, 21-42.

Spelke, E., & Kinzler, K. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Spelke, E., Phillips, A., & Woodward, A. (1995). Infants' knowledge of object motion and human action. In *Causal cognition: A multidisciplinary debate* (pp. 44–78). Oxford: Oxford University Press.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453–489.

Stinchcombe, M., & White, H. (1989). Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. *Proceedings of the International Joint Conference on Neural Networks*, 607–611.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian

inference. *Behavioral and Brain Sciences*, *24*, 629–641.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *135*, 1124–1131.

Vallabha, G., McClelland, J., Pons, F., Werker, J., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273–13278.

van Dantzig, D. (1957). Statistical priesthood (Savage on personal probabilities). *Statistica Neerlandica*, *2*, 1–16.

Verma, D., & Rao, R. (2006). Goal-based imitation as probabilistic inference over graphical models. *Advances in Neural Information Processing Systems*, *18*.

Vitànyi, P., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, *46*(2), 446–464.

Waldmann, M. R. (1996). Knowledge-based causal induction. In *The psychology of learning and motivation* (Vol. 34, p. 47-88). San Diego: Academic Press.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wexler, K., & Culicover, P. (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

Woodward, A. (1998). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior and Development*, *22*(2), 145–160.

Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*, 288-297.

Xu, F., & Tenenbaum, J. B. (2007b). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245-272.

Zemel, R., Huys, Q., Natarajan, R., & Dayan, P. (2005). Probabilistic computation in spiking populations. *Advances in Neural Information Processing Systems*, *17*.

A Appendix

This appendix contains additional references that may be useful to those interested in learning more about different aspects of Bayesian learning.

A.1 Glossary

This is a brief glossary of some of the terms that may be encountered when learning about Bayesian models.

Bayesian Ockham's Razor : Describes how a preference for "simpler" models emerges in a Bayesian framework.

Blessing of abstraction : The phenomenon whereby higher-level, more abstract knowledge, may be easier or faster to acquire than specific, lower-level knowledge.

Conditional distribution : The probability of one variable (e.g., a) given another (e.g., b), denoted $p(a|b)$.

Graphical model : A probabilistic model for which a graph denotes the conditional independence structure between random variables. A directed graphical model identifies which of the nodes are the parents, and thus enables the joint distribution to be factored into conditional distributions. A directed graphical model is also known as a Bayesian network.

Hierarchical Bayesian model (HBM) : A type of Bayesian model capable of learning at multiple levels of abstraction.

Hyperparameters : The higher-level parameters learned in a hierarchical Bayesian model. These parameters capture the *overhypothesis* knowledge and govern the choice of lower-level parameters.

Hypothesis space : The set of all hypotheses a learner could entertain. This is divided into the *latent hypothesis space*, which consists of all logically possible

hypothesis spaces and is defined by the structure of the learning problem, and the *explicit hypothesis space*, which contains the hypotheses a learner has explicitly considered or enumerated.

Joint distribution : The probability of multiple variables (e.g., a and b) occurring jointly, denoted $p(a, b)$.

Likelihood : The probability of having observed some data d if some hypothesis h is correct, denoted $p(d|h)$.

Marginal distribution : The probability distribution of a subset of variables, having averaged over information about another. For instance, given two random variables a and b whose joint distribution is known, the marginal distribution of a is the probability distribution of a averaging over information about b, generally by summing or integrating over the joint probability distribution $p(a, b)$ with respect to b.

Markov chain Monte Carlo (MCMC) : A class of algorithms for sampling probability distributions. It is generally used when the probability distributions are too complex to be calculated analytically, and involves a series of sampling steps. Metropolis-Hastings and Gibbs sampling are two common types of MCMC methods.

Markov model : A model which captures a discrete random process in which the current state of the system depends only on the previous state of the system, rather than on states before that.

Overhypothesis : A higher-level inductive constraint that guides second-order generalization (or above). The term originates from Goodman (1955).

Posterior probability : The degree of belief assigned to some hypothesis h after having seen some data d (combines the *likelihood* and the *prior*, denoted $p(h|d)$.

Prior probability : The degree of belief assigned to some hypothesis h before having seen the data, denoted $p(h)$.

Probability distribution : Defines either the probability of a random variable (if the variable is discrete) or the probability of the value of the variable falling in a particular interval (when the variable is continuous).

Size principle : The preference for smaller hypotheses over larger ones, all else being equal, naturally instantiated by the likelihood term.

Stochastic : Random.

A.2   Applications

Recent years have seen a surge of interest in applying Bayesian techniques to many different problems in cognitive science. Although an exhaustive overview of this research is beyond the scope of this paper, we list here some example references, loosely organized by topic, intended to give the interested reader a place to begin, and also to illustrate the flexibility and scope of this framework. In addition, *Trends in Cognitive Sciences* (2007) published a special issue (Volume 10, Issue 7) focused on probabilistic models in cognition.

1. **Learning and using phonetic categories**: Vallabha, McClelland, Pons, Werker, and Amano (2007); N. Feldman, Morgan, and Griffiths (2009); N. Feldman and Griffiths (2009)

2. **Acquisition and nature of causal reasoning**: Cheng (1997); Pearl (2000); Steyvers, Tenenbaum, Wagenmakers, and Blum (2003); Sobel, Tenenbaum, and Gopnik (2004); Gopnik et al. (2004); Griffiths and Tenenbaum (2005); Lu, Yuille, Liljeholm, Cheng, and Holyoak (2008); Griffiths and Tenenbaum (2009)

3. **Abstract reasoning and representation based on graphical structures**: Kemp, Perfors, and Tenenbaum (2004); Roy, Kemp, Mansinghka, and Tenenbaum

(2006); Schmidt et al. (2006); Xu and Tenenbaum (2007b)

4. **Abstract semantic representations**: Navarro and Griffiths (2007); Griffiths, Steyvers, and Tenenbaum (2007); Andrews and Vigliocco (2009)

5. **Category learning and categorization**: Anderson (1991); Ashby and Alfonso-Reese (1995); Navarro (2006); Kemp, Perfors, and Tenenbaum (2007); Shafto, Kemp, Mansinghka, Gordon, and Tenenbaum (2006); Griffiths et al. (2008); Perfors and Tenenbaum (2009); Heller et al. (2009); Sanborn et al. (2010)

6. **Decision making**: M. Lee (2006); M. Lee, Fuss, and Navarro (2007)

7. **Grammar learning and representation**: Dowman (2000); Perfors et al. (2006, submitted); Bannard, Lieven, and Tomasello (2009)

8. **Individual differences**: Navarro, Griffiths, Steyvers, and Lee (2006)

9. **Language evolution**: Griffiths and Kalish (2007); Kirby, Dowman, and Griffiths (2007); K. Smith (2009)

10. **Morphological acquisition**: Goldwater, Griffiths, and Johnson (2006); Frank, Ichinco, and Tenenbaum (2008)

11. **Planning and inferences about agents**: Verma and Rao (2006); Baker, Tenenbaum, and Saxe (2007); J. Feldman and Tremoulet (2008); Lucas et al. (submitted)

12. **Learning logical rules**: J. Feldman (2000); Goodman, Griffiths, Feldman, and Tenenbaum (2007)

13. **Theory learning**: Kemp, Goodman, and Tenenbaum (2007); Kemp et al. (2010)

14. **Verb learning**: Alishahi and Stevenson (2008); Hsu and Griffiths (2009); Perfors, Tenenbaum, and Wonnacott (2010)

15. **Word learning**: Xu and Tenenbaum (2007b); Andrews, Vigliocco, and Vinson (2009); Frank, Goodman, and Tenenbaum (2009)

16. **Word segmentation**: Goldwater, Griffiths, and Johnson (2007); Frank, Goldwater, Griffiths, and Tenenbaum (2007)

A.3   Further reading

The mathematical foundations of Bayesian inference extend back decades if not centuries. Sivia (1996) and P. Lee (1997) are good introductory textbooks; more advanced texts include Berger (1993) and Jaynes (2003).

As discussed briefly within the paper, Bayesian probability theory brings up several issues related to the subjectivity of the prior probability, relation to frequentist statistical approaches, and the interpretation and nature of probability in the first place. Classic work from a frequentist perspective includes Fisher (1933) and van Dantzig (1957), and from a Bayesian perspective Jeffreys (1939), Cox (1946), Savage (1954), and de Finetti (1974). Box and Tiao (1992) explores how the frequentist approach may be interpreted from a Bayesian perspective, and Jaynes (2003) provides a nice overview, bringing the threads of many of these arguments together.

There is a great deal of work exploring the relationship between Bayesian learning and information-theoretic or minimum description length (MDL) approaches. Vitànyi and Li (2000), Jaynes (2003), MacKay (2003) and Grünwald, Myung, and Pitt (2005) provide excellent discussions and overview of some of the issues that arise. More classic texts include Rissanen (1978), Solomonoff (1964), and Kolmogorov (1965).

One of the largest areas of research in machine learning is focused on developing more effective techniques for searching the (sometimes quite large) hypothesis spaces defined by Bayesian models. Bayesian methods in artificial intelligence and machine learning are described generally in Russell and Norvig (2010) and MacKay (2003). One of the standard approaches includes Markov chain Monte Carlo (MCMC), which is introduced and explained in Neal (1993); MacKay (1998); Gilks, Richardson, and

Spiegelhalter (1996) and Gelman, Carlin, Stern, and Rubin (2004) provide examples of how to incorporate these methods into Bayesian models. In addition, sequential Monte Carlo methods (e.g., Doucet et al., 2001; Sanborn et al., 2010) provide a means to explore capacity limitations and a more "on-line" processing approach.