Theory Acquisition as Stochastic Search

Tomer D. Ullman, Noah D. Goodman, Joshua B. Tenenbaum
{tomeru, ndg, jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

We present an algorithmic model for the development of children's intuitive theories within a hierarchical Bayesian framework, where theories are described as sets of logical laws generated by a probabilistic context-free grammar. Our algorithm performs stochastic search at two levels of abstraction – an outer loop in the space of theories, and an inner loop in the space of explanations or models generated by each theory given a particular dataset – in order to discover the theory that best explains the observed data. We show that this model is capable of learning correct theories in several everyday domains, and discuss the dynamics of learning in the context of children's cognitive development.

Introduction

As children learn about the world, they learn more than just a large stock of specific facts. They organize their knowledge into abstract coherent frameworks, or *intuitive theories*, that guide inference and learning within particular domains (Carey, 1985; Wellman & Gelman, 1992). Much recent work in computational cognitive modeling has attempted to formalize how intuitive theories are structured, used and acquired from experience (Tenenbaum, Griffiths, & Kemp, 2006), working broadly within a hierarchical Bayesian framework shown in Figure 1 (and explained in more detail below). While this program has made progress in certain respects, it has treated the problem of theory acquisition only in a very ideal sense. The child is assumed to have a hypothesis space of possible theories constrained by some "Universal Theory", and to be able to consider all possible theories in that space, in light of a given body of evidence. Given sufficient evidence, and a suitably constrained hypothesis space of theories, it has been shown that an ideal Bayesian learner can identify the correct theory underlying core domains of knowledge such as causality (Goodman, Ullman, & Tenenbaum, 2009), kinship and other social structures (Kemp, Goodman, & Tenenbaum, 2008). These Bayesian computational analyses have not to date been complemented by working algorithmic models of the search process by which a child can build up an abstract theory, piece by piece, generalizing from experience. Here we describe such an algorithmic model for Bayesian theory acquisition. We show that our algorithm is capable of constructing correct if highly simplified theories for several everyday domains, and we explore the dynamics of its behavior – how theories can change as the learner's search process unfolds as well as in response to the quantity and quality of the learner's observations.

At first glance, the dynamics of theory acquisition in childhood look nothing like the ideal learning analsyes of hierarchical Bayesian models – and may not even look particularly rational or algorithmic. Different children see different random fragments of evidence and make their way to adult-like intuitive theories at different paces and along different paths. It seems unlikely that children can simultaneously evaluate many candidate theories at once; on the contrary, they appear to hold just one theory in mind at any time. Transitions between theories appear to be local, myopic, and semi-random, rather than systematic explorations of the hypothesis space. They are prone to backtracking or "two steps forward, one step back". We suggest that these dynamics are indicative of a stochastic search process, much like the Markov chain Monte Carlo (MCMC) methods that have been proposed for performing approximate probabilistic inference in complex generative models. We show how a search-based learning algorithm can begin with little or no knowledge of a domain, and discover the underlying structure that best organizes it by generating new hypotheses and checking them against its current conceptions of the world using a hiearchical Bayesian framework. New hypotheses are accepted probabilistically if they can better account for the observed data, or if they compress it in some way. Such a search-based learning algorithm is capable of exploring a potentially infinite space of theories, but given enough time and sufficient data it tends to converge on the correct theory – or at least some approximation thereof, corresponding to a small set of abstract predicates and laws.

The plan of the paper is as follows. We first introduce our framework for representing and evaluating theories, based on first-order logic and Bayesian inference in a hierarchical probabilistic model that specifies how the theory's logical structure constrains the data observed by a learner. We then describe our algorithmic approach to theory learning based on MCMC search, using simulated annealing to aid convergence. Finally we study the search algorithm's behavior on two case studies of theory learning in everyday cognitive domains: the taxonomic organization of object categories and properties, and a simplified version of magnetism.

Formal framework

We work with the hierarchical probabilistic model shown in Figure 1, based on those in (Katz, Goodman, Kersting, Kemp, & Tenenbaum, 2008; Kemp et al., 2008). We assume that a domain of cognition is given, comprised of one or more systems, each of which gives rise to some observed data. The learner's task is to build a theory of the domain: a set of abstract concepts and explanatory laws that together generate a hypothesis space and prior probability distribution over candidate models for systems in that domain. The laws and

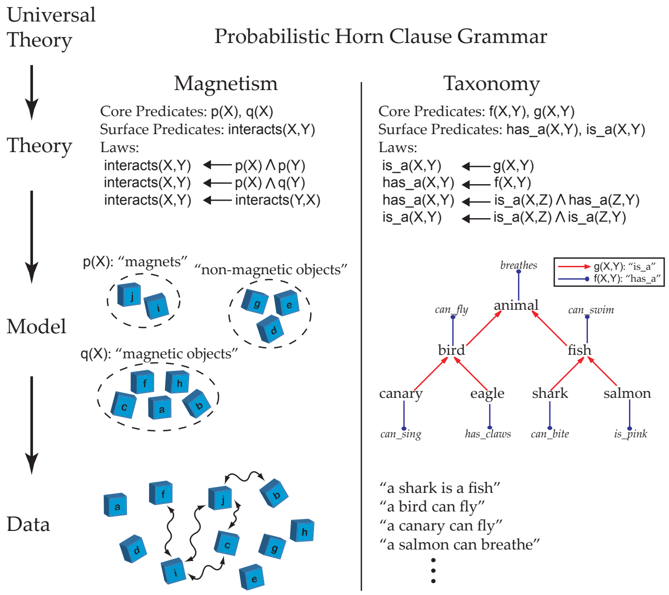

Figure 1: A hierarchical Bayesian framework for theory acquisition

concepts are written in logical form, a "language of thought", typically a subset of first-order logic. The learner's model of a system specifies what is true of that system, and thereby generates a probability distribution over possible observations that can be made for that system.

For example, consider a child learning about the domain of magnetism. She might begin by playing with a few pieces of metal and notice that some of the objects interact, exerting strange pulling or pushing forces on each other. She could describe the data directly, as "Object i interacts with object f", "Object i interacts with object j", and so on. Or she could form a simple theory, in terms of abstract concepts such as magnet, magnetic object and non-magnetic object, and laws such as "Magnets interact with other magnets", "Magnets interact with magnetic objects", and "Interactions are symmetric" (but no other interactions take place). Systems in this domain correspond to specific subsets of objects, such as the set of objects $\{a, \ldots, i\}$ in Figure 1. A model of a system specifies the minimal facts needed to apply the abstract theory to the system, in this case which objects are magnetic, which are magnets, and which are non-magnetic. From these core facts the laws of the theory determine all other true facts – in our example, this means all the pairwise interactions between the objects: e.g., objects i and j, being magnets, should interact, but i and e should not, because e is non-magnetic. Finally, the true facts generate the actual data observed by the learner via a noisy observation process.

While the abstract concepts in this simplified magnetism theory are attributes of objects, more complex relations are possible. Consider for example a domain of taxonomy, as in Collins and Quillian's classic model of semantic memory as an inheritance hierarchy (Collins & Quillian, 1969). Here the abstract concepts are is_a relations between categories and has_a relations between categories and properties. The theory underlying taxonomy has two basic laws: "The has_a relation inherits down is_a relations" and "The is_a relation is transitive" (laws 3 and 4 on the right side of Figure 1). A system consists of a specific set of categories and properties, such as salmon, eagle, breathes, can fly, and so on. A model specifies the minimal is_a and has_a relations, typically corresponding to a tree of is_a relations between categories with properties attached by has_a relations at the broadest category they hold for: e.g., "A canary is a bird", "A bird is an animal", "An animal can breathe", and so on. The laws then determine that properties inherit down chains of is_a relations to generate many other true facts that can potentially be observed, e.g., "A canary can breathe".

Equipped with this hierarchical generative model, a learner can work backwards from observed data to multiple levels of latent structure. Given the correct theory, the learner can infer the most likely model underlying a set of noisy, sparse observations and predict facts that have not been directly observed (Katz et al., 2008; Kemp et al., 2008). If the true theory is unknown, the learner can consider a hypothesis space of candidate theories, generated by higher-level "Universal Theory (UT)" knowledge. UT defines a distribution over the space of possible theories, $P(T|UT)$, which can then be used by a learner to infer the correct theory describing a domain, according to the standard Bayesian formulation:

$$P(T|D,UT) \propto P(D|T)P(T|UT) \qquad (1)$$

Bayes' rule here captures the intuition of Occam's razor. The theory that best explains the data, or has highest posterior probability $P(T|D,UT)$, should be based on two considerations: how well the theory fits the data, as measured by the likelihood $P(D|T)$, and how simple or short is the theory, as measured by the prior $P(T|UT)$. We now define these hypothesis spaces and probabilities more formally, and then describe a learning algorithm that searches the space of theories by proposing small random changes to the current theory and accepting changes stochastically based on whether they are likely to lead to higher overall probability.

A language for theories. Following (Katz et al., 2008) we represent the laws in a theory as Horn clauses: logical expressions of the form $t \leftarrow (p \wedge q \wedge ... \wedge r)$. Horn clauses express logical implications – a set of conjunctive conditions under which t holds – but can also capture intuitive causal relations under the assumption that any propositions not generated by the theory are assumed to be false. In our formulation, the clauses contain two kinds of predicates: "core" and "surface". Core predicates are a minimal set of predicates that determine all other predicates when combined with the theory's laws. Surface predicates are derived from other predicates, either surface or core, via the laws. Predicates may or may not be directly observable in the data. The core predicates can be seen as compressing the full model into just the minimal bits necessary to specify all true facts. In the magnetism example above, the core could be expressed in terms of two predicates p(X) and q(X). Based on an assignment of truth values to these core predicates, the

Top level theory			
(S1)	S	$\Rightarrow$	(Law) $\wedge$ S
(S2)	S	$\Rightarrow$	(Tem) $\wedge$ S
(S3)	S	$\Rightarrow$	Stop
Random law generation			
(Law)	Law	$\Rightarrow$	($P_{left} \leftarrow P_{right} \wedge$ Add)
(Add1)	A	$\Rightarrow$	P $\wedge$ Add
(Add2)	A	$\Rightarrow$	Stop
Predicate generation			
($P_{left}1$)	P_{left}	$\Rightarrow$	$surface1()$
$\vdots$			
($P_{left}\ \alpha$)	P_{left}	$\Rightarrow$	$surface\alpha()$
($P_{right}1$)	P_{right}	$\Rightarrow$	$surface1()$
$\vdots$			
($P_{right}\ \alpha$)	P_{right}	$\Rightarrow$	$surface\alpha()$
($P_{right}(\alpha+1)$)	P_{right}	$\Rightarrow$	$core1()$
$\vdots$			
($P_{right}(\alpha+\beta)$)	P_{right}	$\Rightarrow$	$core\beta()$
Law templates			
(Tem1)	Tem	$\Rightarrow$	template1()
$\vdots$			
(Temγ)	Tem	$\Rightarrow$	templateγ()

Figure 2: Production rules of the Probabilistic Horn Clause Grammar. S is the start symbol and Law, Add, P and Tem are non-terminals. α, β, and γ are the numbers of surface predicates, core predicates, and law templates, respectively.

learner can use the theory's laws such as $\mathsf{interacts(X,Y)} \leftarrow \mathsf{p(X)} \wedge \mathsf{q(Y)}$ to derive values for the observable surface predicate $\mathsf{interacts(X,Y)}$. Notice that $\mathsf{p(X)}$ and $\mathsf{q(X)}$ are abstract predicates, which acquire their meaning as concepts picking out magnets or magnetic objects respectively in virtue of the role they play in the theory's laws. In constructing such a theory the learner essentially creates new concepts (Carey, 1985). Entities may be typed and predicates restricted based on type constraints: e.g., in taxonomy, $\mathsf{has\_a(X,Y)}$ requires that X be a category and Y be a property, while $\mathsf{is\_a(X,Y)}$ requires that X and Y both be categories. Forcing candidate models and theories to respect these type constraints provides the learner with a valuable and cognitively natural inductive bias.

The theory prior $P(T|UT)$. We posit UT knowledge in the form of a probabilistic context-free Horn clause grammar (PHCG) that generates the hypothesis space of possible Horn-clause theories, and a prior $P(T|UT)$ over this space (Figure 2). This grammar and the Monte Carlo algorithms we use to sample or search over the theory posterior $P(T|D,UT)$ are based heavily on Goodman, Tenenbaum, Feldman, and Griffiths (2008), who introduced the approach for learning single rule-based concepts rather than the larger theory structures we consider here. We refer readers to Goodman et al. (2008) for many technical details. Given a set of possible predicates in the domain, the PHCG draws laws from a random construction process (Law) or from law templates (Tem; explained in detail below) until the Stop symbol is reached, and then grounds out these laws as horn clauses. The prior $p(T|UT)$ is the product of the probabilities of choices made at each point in this derivation. All these probabilities are less than one, so overall the prior favors simpler theories with shorter derivations. The precise probabilities of different rules in the grammar are treated as latent variables and integrated out, favoring re-use of the same predicates and law components within a theory (Goodman et al., 2008).

P(X,Y)	$\leftarrow$	P(X,Z)$\wedge$P(Z,Y)	P(X,Y)	$\leftarrow$	P(X)$\wedge$P(Y)
P(X,Y)	$\leftarrow$	P(Z,X)$\wedge$P(Z,Y)	P(X,Y)	$\leftarrow$	P(Y,X)
P(X,Y)	$\leftarrow$	P(X,Z)$\wedge$P(Y,Z)	P(X,Y)	$\leftarrow$	P(X,Y)
P(X,Y)	$\leftarrow$	P(Z,X)$\wedge$P(Y,Z)	P(X)	$\leftarrow$	P(X)
P(X,Y)	$\leftarrow$	P(X,Y)$\wedge$P(X)	P(X)	$\leftarrow$	P(X,Y)$\wedge$P(X)
P(X,Y)	$\leftarrow$	P(Y,X)$\wedge$P(X)	P(X)	$\leftarrow$	P(Y,X)$\wedge$P(X)
P(X,Y)	$\leftarrow$	P(X,Y)$\wedge$P(Y)	P(X)	$\leftarrow$	P(X,Y)$\wedge$P(Y)
P(X,Y)	$\leftarrow$	P(Y,X)$\wedge$P(Y)	P(X)	$\leftarrow$	P(Y,X)$\wedge$P(Y)

Figure 3: The list of templates available to in the PHCG.

Law templates. We make the grammar more likely to generate useful laws by equipping it with templates, or canonical forms of laws that capture structure likely to be shared across many domains. While it is possible for the PHCG to reach each of these law forms without the use of templates, their inclusion allows the most useful laws to be invented more readily. They can also serve as the basis for transfer learning across domains. For instance, instead of having to reinvent transitivity anew in every domain with some specific transitive predicates, a learner could recognize that the same transitivity template applies in several domains. It may be costly to invent transitivity for the first time, but once found – and appreciated! – its abstract form can be readily reused. The specific law templates used are described in Figure 3. Each "$\mathsf{P}(\cdot)$" symbol stands for a non-terminal representing a predicate of a certain -arity. This non-terminal is later instantiated by a specific predicate. For example, the template $\mathsf{P(X,Y)} \leftarrow \mathsf{P(X,Z)} \wedge \mathsf{P(Z,Y)}$ might be instantiated as $\mathsf{is\_a(X,Y)} \leftarrow \mathsf{is\_a(X,Z)} \wedge \mathsf{is\_a(Z,Y)}$ (a familiar transitive law) or as $\mathsf{has\_a(X,Y)} \leftarrow \mathsf{is\_a(X,Z)} \wedge \mathsf{has\_a(Z,Y)}$ (the other key law of taxonomy, stating that "has_a is transitive over is_a").

The theory likelihood $P(D|T)$. An abstract theory makes predictions about the observed data in a domain only indirectly, via the models it generates. A theory typically generates many possible models: even if a child has the correct theory and abstract concepts of magnetism, she could categorize a specific set of metal bars in many different ways, each of which would predict different interactions that could be observed as data. Expanding the theory likelihood,

$$P(D|T) = \sum_{M} P(D|M)P(M|T) \tag{2}$$

we see that theory T predicts data D well if it assigns high prior $P(M|T)$ to models M that make the data probable under the observation process $P(D|M)$.

The model prior $P(M|T)$ reflects the intuition that a theory T explains some data well if T requires few additional degrees of freedom beyond its abstract concepts and laws to make its predictions. That is, few specific and contingent facts about the system under observation are required in addition to the theory's general prescriptions. This intuition is captured by a prior that encourages the core predicates to be as sparse as possible, penalizing theories that can only fit well by "overfitting" with many extra degrees of freedom. Formally, following (Katz et al., 2008), we model all values of the core predicates as independent Bernoulli random variables with conjugate beta priors encouraging most variables to have the same value (on or off). We assume that any proposition potentially in the model M is false unless it is a core predicate turned on by this Bernoulli process or is derived from the core predicates through the theory's laws (the *minimal model* assumption of logic programming).

Finally, the model likelihood $P(D|M,T)$ comes from assuming that we are observing randomly sampled true facts (sampled with replacement, so the same fact could observed on multiple occasions), which also encourages the model extension to be as small as possible.

Stochastic search in theory space: a grammar-based Monte-Carlo algorithm. Following Goodman et al. (2008), we use a grammar-based Metropolis-Hastings (MH) algorithm to sample theories from the posterior distribution over theories conditioned on data, $P(T|D,UT)$. This algorithm is applicable to any grammatically structured theory space, such as the one generated by our PHCG. The MH algorithm proceeds by randomly proposing changes to the current theory, and accepting or rejecting these changes. Each proposed change to the current theory corresponds to choosing a grammatical constituent of the theory then regenerating it from the PHCG. For example, if our theory of magnetism includes the law $\mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{p}(\mathsf{X}) \wedge \mathsf{q}(\mathsf{Y})$, the MH procedure might propose to add or delete a predicate (e.g., $\mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{p}(\mathsf{X}) \wedge \mathsf{q}(\mathsf{Y}) \wedge \mathsf{p}(\mathsf{Y})$ or $\mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{p}(\mathsf{X})$), to change one predicate to an alternative of the same form (e.g., $\mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{p}(\mathsf{X}) \wedge \mathsf{p}(\mathsf{Y})$) or a different form if available (e.g., $\mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{p}(\mathsf{X}) \wedge \mathsf{r}(\mathsf{X},\mathsf{Y})$); to resample the law from a template (e.g., $\mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{r}(\mathsf{X},\mathsf{Z}) \wedge \mathsf{r}(\mathsf{Z},\mathsf{Y})$); or to add or delete a whole law.

These proposals are accepted with probability equal to the minimum of 1 and the MH acceptance ratio,

$$\frac{P(T'|D,UT)}{P(T|D,UT)} \cdot \frac{Q(T|T')}{Q(T'|T)} \tag{3}$$

where T is the current theory, T' is the new proposed theory, and $Q(\cdot|\cdot)$ is the transition probability from one theory to the other, derived from the PHCG (Goodman et al., 2008). To aid convergence we raise the posterior ratio to a power greater than 1, which we increase very slightly after each MH step in a form of simulated annealing. The learner initially explores alternative theories freely, but with time becomes increasingly likely to reject theory changes unless they lead to an improved posterior probability.

While this MH algorithm could be viewed merely as a way to approximate the calculations necessary for a hierarchical Bayesian analysis, we suggest that it could also capture in a schematic form the dynamic processes of theory acquisition and change in young children. Stochastic proposals to add a new law or change a predicate within an existing law are consistent with some previous characterizations of children's theory learning dynamics (Siegler & Chen, 1998). These dynamics were previously proposed on purely descriptive grounds, but here they emerge as a consequence of a rational learning algorithm for effectively searching an infinite space of logical theories.

Approximating the theory score. Computing the theory likelihood $P(D|T)$, necessary to compare alternative theories in Equation 3, requires a summation over all possible models consistent with the current theory (Equation 2). Because this sum is typically very hard to evaluate exactly, we approximate $P(D|T)$ with $P(D|M^*)P(M^*|T)$, where M^* is an estimate of the maximum a-posteriori (MAP) model inferred from the data: the most likely values of the core predicates. The MAP estimate M* is obtained by running a Gibbs sampler over the values of the core predicates, as in (Katz et al., 2008), annealing slightly on each Gibbs sweep to speed convergence and lock in the best solution. The Gibbs sampler over models generated by a given theory is thus an "inner loop" of sampling in our learning algorithm, operating within each step of an "outer loop" sampling at a higher level of abstract knowledge, the MH sampler over theories generated by UT knowledge.

Case Studies

We now explore the performance of this stochastic approach to theory learning in two case studies, using simulated data from the domains of taxonomy and magnetism introduced above. We examine the learning dynamics in each domain and make more explicit the possible parallels with human theory acquisition.

Taxonomy

Katz et al. (2008) defined a similar hierarchical Bayesian framework and showed that a theory of taxonomic reasoning about properties and categories in an inheritance hierarchy could be correctly selected from among several alternatives, on the basis of data. However, they did not address the harder challenge of constructing the theory from the ground up, or selecting it from an effectively infinite hypothesis space of theories (which could be used to describe many other domains). That is our goal here. Following Katz et al. (2008), we take the correct theory to have two unobservable core predicates, $\mathsf{g}(\mathsf{X},\mathsf{Y})$ and $\mathsf{f}(\mathsf{X},\mathsf{Y})$, and two observable surface predicates, $\mathsf{is\_a}(\mathsf{X},\mathsf{Y})$ and $\mathsf{has\_a}(\mathsf{X},\mathsf{Y})$. There are four laws:

$$\begin{aligned}
&\text{Law 1:} && \mathsf{has\_a}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{f}(\mathsf{X},\mathsf{Y})\\
&\text{Law 2:} && \mathsf{is\_a}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{g}(\mathsf{X},\mathsf{Y})\\
&\text{Law 3:} && \mathsf{has\_a}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{is\_a}(\mathsf{X},\mathsf{Z}) \wedge \mathsf{has\_a}(\mathsf{Z},\mathsf{Y})\\
&\text{Law 4:} && \mathsf{is\_a}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{is\_a}(\mathsf{X},\mathsf{Z}) \wedge \mathsf{is\_a}(\mathsf{Z},\mathsf{Y})
\end{aligned}$$

Laws 1 and 2 set up the core predicates to represent the mini-

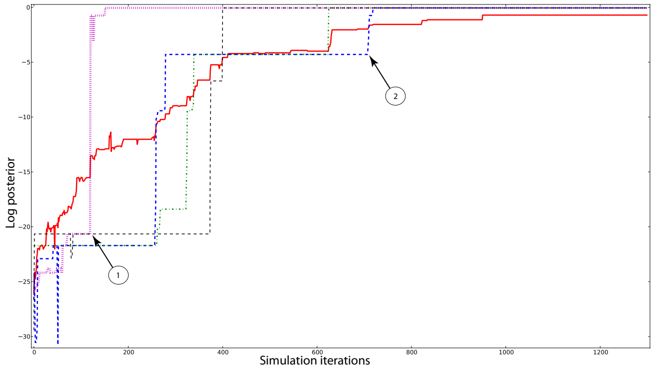

Figure 4: Log posterior score for representative runs of theory learning in Taxonomy. Dashed lines show different runs. Solid line is the average across all runs. Node 1 marks the acquisition of law 3, node 2 marks the acquisition of law 4.

mal is_a and has_a links, on top of which are defined the laws of property inheritance (Law 3) and transitive category membership (Law 4). We take laws 1 and 2 as given, assuming the structure and meaning of the core predicates as Katz et al. did, and ask whether a learner can successfully construct laws 3 and 4. Following Katz et al., we consider a concrete domain with 7 categories and 7 properties in a balanced taxonomy, as shown in Figure 1. Observations include all positive facts asserting that a property is true of a category, as in "An eagle has claws". (The data used for this section and the following case study can be found at http://web.mit.edu/tomeru/www/tlss.) We ran 10 simulations for 1300 iterations of the outer MH loop. Learning curves for representative runs as well as the average over all runs are shown in Figure 4. Out of 10 simulations, 8 found the correct theory within the given number of iterations, and 2 discovered a partial theory which included only law 3 (property inheritance). Several observations are worth noting.

Abstract learning is possible. Using only stochastic local search moves, a learner can navigate the space of potential theories to discover the laws underlying the domain. Even a relatively small dataset (with 7 categories and 7 properties) is sufficient to learn the correct abstract domain theory.

Individual learning curves show sudden changes and high variability in what is learned when, while on average learning is smooth and follows a characteristic timecourse. The learning algorithm's local dynamics are highly stochastic and variable across runs, because of the randomness in what theory changes are proposed when, and the fact that a small theory change can make a big difference in predictive power. Yet there is still a meaningful sense in which we can talk about "typical" learning behavior, even though any one learner may not look much like this average. If stochastic local search is a key component in children's theory construction, it could explain why cognitive development shows this same dual nature: systematic and graded progression at the population level, despite random, discontinuous and highly variable learning rates in any one child.

Although proposals are random, there is a systematic and rational order to learning. While there are many routes through theory space to a given endpoint, a sequence of random MH proposals may still prefer some orders of knowledge acquisition over others. Here, when law 4 is discovered (on 8/10 runs), it is always acquired after law 3. This is because law 4 (transitivity of category membership) provides much more explanatory power – and hence is more stable under our stochastic theory-learning dynamics – given law 3 (property inheritance) and a reasonable domain model specifying which properties hold for which categories. This order is also consistent with the order of acquisition in human cognitive development (Wellman & Gelman, 1992): children learn to generalize properties of biological categories to instances well before they learn that categories can be arranged in a multilevel hierarchy supporting transitive inferences of category membership.

Magnetism

After showing that stochastic search can learn the correct laws in a domain theory, we now consider a second case study in which the acquisition of new laws corresponds to a shift in the meaning of the core predicates, and new (i.e., previously unassigned) core predicates are introduced during learning – akin to some of the conceptual changes described by Carey (1985). Our domain here is the simple version of magnetism described above, with two unobservable core predicates: $\mathsf{p(X)}$ and $\mathsf{q(X)}$, and one observable surface predicate: $\mathsf{interacts(X,Y)}$. There are three laws:

$$\begin{aligned} &\text{Law 1:} && \mathsf{interacts(X,Y)} \leftarrow \mathsf{p(X)} \wedge \mathsf{p(Y)} \\ &\text{Law 2:} && \mathsf{interacts(X,Y)} \leftarrow \mathsf{p(X)} \wedge \mathsf{q(Y)} \\ &\text{Law 3:} && \mathsf{interacts(X,Y)} \leftarrow \mathsf{interacts(Y,X)} \end{aligned}$$

We consider a concrete system with 3 magnets, 5 magnetic objects and 2 non-magnetic objects. These concepts are initially unknown to the learner. The core predicates $\mathsf{p(X)}$ and $\mathsf{q(X)}$ are completely abstract and initially uninterpreted. They will acquire their meaning as concepts picking out magnets and magnetic objects respectively in virtue of the role they play in the theory's laws, specifying that objects in one subset (the p's) interact with each other and with objects in a second set (the q's), but q's do not interact with each other. In constructing a theory, the learner introduces these abstract predicates via new laws, or new roles in existing laws, and thereby essentially creates these concepts where she did not have them before (Carey, 1985).

We ran 10 simulations for 1600 iterations of the outer MH loop. Representative runs are displayed in Figure 5, as well as the average over all the runs. The results were similar to the taxonomy case study in several respects, which we also expect to hold for a variety of other domains. The correct theory was usually learned, with some variation: 9/10 simulations found the correct theory or a variant of it, and one discovered a partial theory containing only law 1. Only some runs learned the exact form of law 3, asserting that interactions are symmetric. Others found variants that were extensionally equivalent to symmetry in this domain, but slightly more

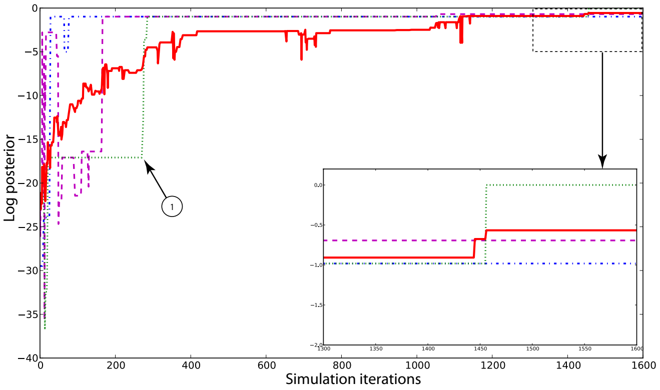

Figure 5: Representative runs of theory learning in Magnetism. Dashed lines show different runs. Solid line is the average across all runs. Node 1 marks the acquisition of law 1 and the confounding of magnets with magnetic objects. Lower right panel zooms into the end of the simulation, showing acquisition of the final correct theory.

complex in their logical form. Individual runs of learning showed discrete jumps with high variability, while average-case behavior was smooth, with systematic order effects. Law 3 is never learned first, because alone it has no explanatory power. Either law 1 or the combination of laws 2 and 3 tend to be learned first, followed by the other, although sometimes laws 1 and 2 are learned first, followed by law 3. Law 1 tends to be learned first overall because it is most likely under the prior (which is also the proposal distribution for local search moves), and also because, as explained below, it represents a reasonable first approximation to the domain's structure.

The algorithm's learning dynamics in this case study are particularly interesting for how they parallel key transitions in childrens' cognitive development: restructuring or construction of new concepts, as when one concept differentiates into two (Carey, 1985). When our simulations of learning about magnetism construct law 1 first, without laws 2 and 3, they find a simpler theory capturing many of the observed facts at the cost of over-generalizing. That is, under law 1 alone, the optimal setting of the core predicates – the most probable model – equates magnets and magnetic objects, making $\mathsf{p}(\mathsf{X})$ true for both. This is a good first approximation, even as it collapses two categories of objects with fundamentally different causal properties: the generators of magnetic force (the "magnets") and the objects on which that force acts (the "magnetic objects"). Only once all three laws have been constructed does the learner come to distinguish between magnets and magnetic objects, reflected in the difference between the roles played by the two core predicates $\mathsf{p}(\mathsf{X})$ and $\mathsf{q}(\mathsf{X})$. Only once law 2 is available does the learner have reason to restrict the extension of $\mathsf{p}(\mathsf{X})$ to just magnets, excluding other magnetic objects.

Conclusion and Future Directions

We have presented an algorithmic model of theory acquisition as stochastic search in a hierarchical Bayesian framework and explored its dynamics in two case studies. We were encouraged by the general pattern of successes on these examples and by several qualitative parallels with phenomena of human cognitive development. These results suggest that previous ideal learning analyses of Bayesian theory acquisition can be realized approximately by algorithms that are cognitively plausible for child learners, and indeed potentially descriptive of the dynamics of development.

Previous hierarchical Bayesian analyses of learning abstract knowledge have focused on the role of accumulating data in driving changes to the learner's hypotheses (Kemp & Tenenbaum, 2008). In contrast, here we have focused on how changes to the learner's theories and abstract concepts are driven by a different source, the stochastic dynamics of the learning algorithm. Data-driven and algorithm-driven theory change can have a similar character, first discovering simpler, rougher approximations to reality and then refining those to more complex, accurate representations; sometimes changing by adjusting small details, but other times by making large qualitative transitions or discoveries. In future work we plan to explore further the similarities, differences and interactions between these two drivers of learning dynamics, both in computational analyses and experimental work. We hope to establish tighter quantitative correspondences with human learning curves in development, as well as with controlled laboratory studies of theory learning in adults, where some of the same mechanisms might be at work. We will also consider a broader range of algorithmic approaches, stochastic as well as deterministic, evaluating them both as behavioral models and as effective computational approximations to the theory search problem for larger domains.

Acknowledgments. We thank Yarden Katz for valuable discussions. This work was funded in part by grants from the McDonnell Causal Learning Collaborative, AFOSR (FA9550-07-1-0075), and ONR (N00014-09-0124).

References

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

Collins, A., & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247.

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

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2009). Learning a theory of causality. In *Proceedings of the 31st annual conference of the cognitive science society.*

Katz, Y., Goodman, N. D., Kersting, K., Kemp, C., & Tenenbaum, J. B. (2008). Modeling semantic cognition as logical dimensionality reduction. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory acquisition and the language of thought. In *Proceedings of thirtieth annual meeting of the cognitive science society.*

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Siegler, R. S., & Chen, Z. (1998). Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology*, *36*, 273-310.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337-375.