Probabilistic Models of Cognitive Development

Fei Xu (University of California, Berkeley),
Tom Griffiths (University of California, Berkeley)

May 24–29, 2009

This workshop focused on some of the major issues in the study of cognitive development, especially from the computational modeling point of view. Forty participants from developmental psychology, computational cognitive science, philosophy, and education engaged in five days of talks, poster presentations, and many discussion sessions. It was the first time that these researchers were brought together in a forum, and the workshop was a huge success. Currently the organizers are putting together a special issue of the journal *Cognition*, one of the most prestigious journals in cognitive science, based on contributions from the workshop participants.

1 Overview of the Field

This workshop aimed to capitalize on a major new direction in research on formal models of human cognition. The question of how people come to know so much about the world on the basis of their limited experience has been at the center of the study of the mind since it was first asked by Plato. This question takes a modern form in the nature-nurture debate, which has guided the study of cognitive development from infants to middle childhood over the last few decades. Nativists, favoring strong innate constraints provided by nature, have emphasized competences found in young infants (e.g., constraints on word learning, early understanding of the physical world) whereas empiricists, who focus on the role of experience and nurture, have emphasized learning mechanisms (e.g., keeping track of frequencies and correlations). However, this debate has been hard to resolve without formal tools for evaluating what might plausibly be learned from experience, and what kind of constraints are necessary to support the inferences that children make.

In recent years, several researchers in the cognitive sciences have argued that the nature-nurture framework may have set up a false dichotomy. A more fruitful and productive research strategy may be to find principled ways of combining prior constraints with statistical information in the input. In particular, a number of researchers have begun to use the principles of Bayesian statistics to establish a formal framework for investigating empirical phenomena in development and building computational models of developmental processes. The technical advances that have been made in the use of probabilistic models over the last twenty years in statistics, computer science, and machine learning have made this research enterprise possible, providing psychologists with a set of mathematical and computational tools that can be used to build explicit models of psychological phenomena. By indicating the conclusions that a rational learner might draw from the data provided by experience, Bayesian models can be used to investigate how nature and nurture contribute to human knowledge.

2 Recent Developments and Open Problems

The goal of this workshop was to explore a new approach for studying cognitive development: analyzing childrens learning from the perspective of rational statistical inference. Bayesian statistics and probability theory provide the formal tools that allow us to investigate what prior constraints and what input data are needed in order to justify a particular conclusion. Most importantly, this approach indicates how the kind of prior constraints that might be provided by nature should be combined with the data provided by experience when a learner evaluates a set of alternative hypotheses. More formally, imagine that a learner has a set of hypotheses $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$ about the structure of her environment, and has degrees of belief in those hypotheses that can be expressed through a "prior" probability distribution $p(h)$, where $p(h_i)$ indicates her degree of belief in h_i. She is then provided with some data d, and needs to revise these beliefs in light of evidence. Bayes' rule indicates that the resulting distribution over hypotheses, $p(h|d)$, known as the "posterior" distribution, is given by

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

where $p(d|h)$, the "likelihood," indicates the probability of observing the data d if the hypothesis h were true.

Under Bayes' rule, the posterior probability of a hypothesis is proportional to the product of its prior probability and its likelihood, with the ultimate beliefs of the learner being the result of combining her prior dispositions with the evidence provided by the data. This has direct implications for understanding cognitive development, where innate constraints can be viewed as influencing the prior probability of hypotheses, or even which hypotheses are considered. This approach thus provides a natural compromise between the nativist position, in which strong innate constraints are the key to learning, and the empiricist position, where these constraints are taken to be extremely weak. By exploring the consequences of using different prior distributions, we can determine what kind of constraints are necessary in order to explain the state achieved by adults from the data available to children. This probabilistic approach has already been shown to be productive in studying cognition in human adults, providing accounts of how people make predictions, generalize from examples, form categories, and learn causal relationships [2, 1, 6, 9]. Explaining the inferences that people make requires going beyond the simple formal ideas expressed in the last two paragraphs. Probabilistic models of cognition use cutting-edge tools from statistics and computer science tools that have largely been developed over the last two decades. Understanding human causal learning requires a formal language for representing and reasoning about causal relationships, which is provided by causal graphical models [8, 7]. The properties of with recursive generative systems, such as linguistic syntax, can be captured using probabilistic context-free grammars and other structured statistical models from computational linguistics [4]. Performing probabilistic inference in large, structured models requires using modern Monte Carlo methods, such as Markov chain Monte Carlo [5]. Finally, capturing the flexibility of human mental representations, and the capacity for these representations to increase in complexity when warranted by the data leads to the use of ideas from nonparametric Bayesian statistics, such as the Dirichlet process [3]. Applying these statistical tools in novel contexts can often lead to new insights, and we hope that new formal methods will result from tackling some of the most difficult problems in cognitive development.

3 Presentation Highlights

At our workshop, a number of formal problems were presented and discussed extensively, with the aim to develop further the mathematical/computational tools for modeling cognitive development.

1. *Iterated learning.* A basic question for the cognitive sciences is how information is changed when it is transmitted from person to person. In real cases of cultural evolution, children are often the agents of transmission, meaning that this question has relevance to understanding how cognitive development links to culture. Recent accounts have emphasized the effects of innate constraints on learning on cultural transmission, arguing that universals in languages, religious concepts, and social conventions can be explained in terms of the structure of the human mind. However, these claims have not been supported through detailed mathematical analysis.

Griffiths presented a formal account of cultural evolution by iterated learning that makes it clear in what sense the expectations of learners influence the outcome of cultural evolution. Formally, imagine a sequence of learners, each of whom observes data d_i, forms a hypothesis h_i about the process that generated those data, and then generates new data d_{i+1} that is presented to the next learner. This process can be shown to define a Markov chain, and if the hypotheses are selected by sampling from a Bayesian posterior distribution, the stationary distribution of this Markov chain is the prior of the learners. This means that we should expect that the distribution over hypotheses that are selected will converge to the prior distribution as the process of cultural transmission continues.

This mathematical result is interesting in providing a connection between constraints on learning and the outcome of cultural evolution, but also suggests a way that we can explore the question of what constraints guide human learning. By simulating this process of cultural transmission in the laboratory and seeing what hypotheses emerge in the minds of the participants, we can estimate the priors used by human learners. This process is analogous to estimating a complex probability distribution via Markov chain Monte Carlo, a deep connection that opens the possibility of other randomized algorithms being relevant to exploring the subjective probability distributions maintained by people.

2. *Probabilistic models for the diagnosis of learner knowledge.* Michael Lee gave a presentation illustrating how probabilistic models can be used to infer the knowledge that young children have about numbers. As preschool children begin to understand the relationship between the size of a set of objects (e.g. two mice) and its numerical label ("two"), they go through several stages. These stages correspond roughly to understanding the referents of the words "one", "two", and "three", at each point not understanding the meaning of any of the terms for larger sets, and then transitioning to a complete understanding of the mapping between the terms used for numbers and the size of the sets they describe. Lee considered the problem of how a researcher or clinician could solve the problem of diagnosing the knowledge that a given child has from the way they perform on tasks that require producing sets of objects of different sizes in response to verbal requests.

   The basic idea behind the method that was used to solve this problem was to estimate the distribution over responses produced by children identified as being at each of these stages of knowledge. This information could then be used to apply Bayesian inference, providing a probability distribution over the level of knowledge that a learner seems to possess. Formally, each level of knowledge is a hypothesis h about the learner, and the behavior that the learner produces is data d. The probability of each hypothesis based on the data is given by Bayes' rule (Equation 1). The likelihood, $p(d|h)$, encodes the probability of the learners behavior if the hypothesis is true, and is estimated from the responses of children whose level of knowledge has been diagnosed by an expert. The prior, $p(h)$, reflects expectations about the relative probability of those knowledge states, and can be left uniform if the goal is simply to identify the amount of evidence in favor of each knowledge state.

   A similar strategy for assessing the knowledge of learners can be used in other cases where there are common patterns of understanding or errors that need to be diagnosed. If those patterns of understanding can be identified with a probability distribution over responses, Bayesian inference can be used to work backwards from responses to a picture of the knowledge that the learner has. All that is required is specification of a set of hypotheses about the knowledge of the learner, and estimation of the likelihood function $p(d|h)$ that characterizes the behavior associated with each knowledge state.

3. *Enriching our view of learning to include multiple levels of abstraction.* Many of the probabilistic models that were used to explain aspects of cognitive development shared the use of hierarchical Bayesian inference. In its standard form, Bayesian inference represents a way for a learner to optimally update his or her beliefs about a set of hypotheses in light of data. The basic computations are exactly like those described above: probabilistic models of cognition identify the hypotheses that a learner might entertain, and describe the beliefs of the learner in terms of a probability distribution over those hypotheses, with Bayes rule acting as a learning algorithm for updating those distributions. Hierarchical Bayesian inference assumes a richer representation on the part of the learner, with knowledge at multiple levels of abstraction.

   Many learning problems require making inferences not just about the current problem, but about more general principles that organize a domain. For example, many of the hierarchical Bayesian models presented at the workshop assumed that learners considered not just hypotheses that could be used to explain the most recently observed data, but higher-level theories that captured regularities linking the current hypothesis with hypotheses from past learning opportunities. To take some examples from the presentations: a learner could be forming hypotheses about the meaning of particular words based on labels provided by a parent, but simultaneously developing a theory about the kinds of objects that tend to share a label (such as shape being an important cue about whether two objects can be labeled with the same word); alternatively, a learner could be forming hypotheses about the causal relationships that exist in a particular physical system, while simultaneously forming a theory about how causal relationships operate in that system (such as causes deterministically producing their effects, or combining additively).

The hierarchical Bayesian approach provides a richer picture of learning than that assumed in many computational approaches, with the learner considering not just the solution to a particular problem but also forming generalizations about what solutions to these problems look like. In this way, a learner can form "overhypotheses" that guide future inferences. The ability to make inferences at multiple levels of abstraction provides a way to understand how a child can "learn to learn": as the theory of the domain becomes more accurate, it provides information that reduces the amount of data required to evaluate a particular hypothesis. Formally, we assume that we have random variables at three levels – the data d, the hypotheses h, and a higher level "theory" t. Different learning situations will involve different observed data (say d_1 and d_2), and inferring different hypotheses to explain those data (say h_1 and h_2), but the same theory t can apply across those situations. As a consequence, the prior that is used in one situation is informed by the data observed in the other.

4. *Formalizing pedagogical reasoning.* By providing a way to describe learning, Bayesian inference also provides a way to formalize optimal teaching. The presentation by Patrick Shafto focused on a framework for formalizing pedagogical reasoning. This framework can be built up in three steps. First, we consider the problem of a teacher selecting what data d to provide a student in order to best support learning of a hypothesis h. This problem is solved by finding the d that maximizes $p(h|d)$. Next, we assume that the learner knows that he or she is being taught, and takes this into account in assessing which data d the teacher would produce if h were the intended hypothesis. Intuitively, the learner should expect data that are diagnostic of the hypothesis, sharpening the original likelihood function. Finally, we allow the teacher to take into account this modification of the likelihood function by the learner, potentially changing the data provided to better discriminate between hypotheses. This set of assumptions results in the definition of a system of equations that characterize optimal pedagogical reasoning on the part of teachers and learners. Shafto presented results suggesting that adults produce behavior consistent with a solution to this system of equations when learning about simple categories and causal relationships.

   The basic idea behind this approach is that the teacher should choose data that are maximally informative about the target hypothesis h that the learner should acquire. This gives

   $$p_{\text{teacher}}(d|h) \propto p_{\text{learner}}(h|d)^{\alpha} \tag{2}$$

   where the exponent reflects the degree of noise in the teacher's production of data. Accordingly, the learner should use this distribution when inferring hypotheses, with

   $$p_{\text{learner}}(h|d) \propto p_{\text{teacher}}(d|h)p(h) \tag{3}$$

   which is just Bayes' rule substituting the teacher's behavior for the likelihood. This defines a system of equations that can be solved by iteration, with the iterative solution predicting a pattern of behavior by both teachers and learners.

   In its current form, this research on optimal pedagogical reasoning could potentially guide the development of more efficient automated tutoring systems. However, it seems that this approach has significant potential for further impact on education, provided similar results hold with children and in more complex learning situations. One of the other presentations at the workshop, by Elizabeth Bonawitz, provided preliminary evidence that this pedagogical framework can be used to explain inferences that children make in relatively complex causal reasoning situations.

5. Probabilistic inference on logical reprentations. Goodman and colleagues have recently developed new formal tools for understanding how probabilistic inference can be done on logical representations. Statistical and logical methods have both been influential in cognitive science. The compositionality of logical representations is key for capturing productivity and systematicity of human thought, while probabilistic or statistical inference is crucial for capturing flexible reasoning under uncertainty. Yet compositionality and statistics have rarely been combined in a meaningful way.

   This line of research explore this unification through the probabilistic language of thought hypothesis (PLoT): that mental representations which subserve higher-level cognition are compositional, their meaning is probabilistic, and their function follows from the probabilistic inferences they support. Further, these representations describe generative processes—causal models of the world that may be used to make many different predictions, explanations, and actions.

   To formalize the PLoT hypothesis, Goodman and colleagues have turned to the stochastic λ-calculus. The stochastic λ-calculus is a formal system that extends untyped λ-calculus with stochastic primitive operations and a primitive conditioning operator. The evaluation of an expression in stochastic λ-calculus (which can be understood in terms of a definitional interpreter, or in terms of reduction rules) results in a randomly sampled value. Intuitively this induces a distribution on return values. Indeed, it can be shown that any expression in stochastic λ-calculus that halts almost-always induces a computable distribution on return values, and that any computable distribution can be thus represented. Thus stochastic λ-calculus is universal for representing probability distributions, and for probabilistic reasoning, yet it is a compositional representation language. This compositionality is particularly important when we attempt to understand conceptual development in terms of the PLoT: we view concept learning as induction of expressions in stochastic λ-calculus.

4 Scientific Progress Made

The primary scientific progress resulted from direct contact between empirically-focused developmental psychologists and researchers pursuing mathematical models of human cognition. This interaction has led to several new collaborations, as well as a greater understanding of these models by the broader developmental psychology community.

5 Outcome of the Meeting

In addition to less tangible outcomes resulting from the interaction between these researchers, the meeting has led to the production of a special issue of the journal *Cognition* focusing on probabilistic models of cognitive development. *Cognition* is a leading journal in the cognitive sciences, and the ideal venue for this kind of work. The special issue is edited by Dr. Xu and Dr. Griffiths, and will include a tutorial introduction as well as approximately ten short empirical papers presenting these ideas, authored by the attendees of the meeting.

References

[1] M. Oaksford, and N. Chater, *Bayesian rationality*, Oxford University Press, Oxford, 2007.

[2] N. Chater and M. Oaksford, *Bayesian cognitive science*, Oxford University Press, Oxford, 2008.

[3] T. Ferguson, A Bayesian analysis of some nonparametric problems, *Annals of Statistics*, **1** (1973), 209–230.

[4] S. Geman and M. Johnson, M., A survey of probabilistic grammars. In M. Johnson, S. Khudanpur, M. Ostendorf and R. Rosenfeld, (Eds.), *Mathematical foundations of speech and language processing*, Springer-Verlag, New York, 2004.

[5] W. R. Gilks, S. Richardson and D.J. Spiegelhalter, *Markov Chain Monte Carlo in practice*, Chapman & Hall/CRC, London, 1996.

[6] T. L. Griffiths and J.B. Tenenbaum, Statistics and the Bayesian mind, *Significance*, **3** (2006), 130–133.

[7] J. Pearl, *Causality: models, reasoning, and inference*, Cambridge University Press, Cambridge, 2000.

[8] P. Spirtes, C. Glymour, and R. Scheines, R., *Causation, prediction, and search*, Springer-Verlag, New York, 1993.

[9] J.B. Tenenbaum, T.L. Griffiths, and C. Kemp. Theory-based Bayesian models of inductive learning and reasoning, *Trends in Cognitive Science*, **10** (2006), 309–316.