Connecting human and machine learning via probabilistic models of cognition

Thomas L. Griffiths

Department of Psychology, University of California, Berkeley, USA
tom_griffiths@berkeley.edu

Abstract

Human performance defines the standard that machine learning systems aspire to in many areas, including learning language. This suggests that studying human cognition may be a good way to develop better learning algorithms, as well as providing basic insights into how the human mind works. However, in order for ideas to flow easily from cognitive science to computer science and vice versa, we need a common framework for describing human and machine learning. I will summarize recent work exploring the hypothesis that probabilistic models of cognition, which view learning as a form of statistical inference, provide such a framework, including results that illustrate how novel ideas from statistics can inform cognitive science. Specifically, I will talk about how probabilistic models can be used to identify the assumptions of learners, learn at different levels of abstraction, and link the inductive biases of individuals to cultural universals.

Index Terms: human learning, machine learning, probabilistic models

1. Introduction

Despite the significant advances that have been made in artificial intelligence and machine learning over the last 50 years, it is easy to think of things that people do better than computers. One of the most compelling cases is the ease with which people learn from limited data. The gap between human and machine learning is apparent at many scales. Almost every human child succeeds in learning language purely from linguistic input, while no computer can solve this problem. However, even learning a single new word or concept poses a significant challenge for computers, often requiring hundreds of examples where a person might need only a handful. This gap represents an opportunity both to improve automated systems and to develop a deeper understanding of the formal principles that underlie human cognition.

For computer science and cognitive science to interact profitably, we need a common language for talking about human and machine learning. In the past, these two disciplines have been brought together by formal approaches that provide a way to develop both intelligent automated systems and computational models of human cognition. Logic-based approaches to artificial intelligence grew out of and supported symbolic models of cognition [1, 2, 3]. Artificial neural networks demonstrated how gradient descent and distributed representations could be combined to produce both effective learning algorithms and innovative models of human cognition [4]. Recently, however, machine learning research has begun to draw on probability and statistics as a source of tools for solving challenging learning problems (e.g., [5, 6]). This move has been complemented by work in cognitive science that explores the potential of probabilistic models of human cognition [7, 8, 9].

Probabilistic models of cognition typically focus on the question of how a rational learner should solve an inductive problem, in which the learner considers a variety of hypotheses that might account for observed data. If we assume that the learner represents his or her degree of belief in each hypothesis by assigning a probability to that hypothesis, solving an inductive problem becomes a matter of appropriately modifying these degrees of belief in light of the data. The solution is provided by Bayes' rule: using $p(h)$ to denote the degree of belief in the hypothesis h before seeing data d (known as the *prior* probability) and $p(h|d)$ to denote the degree of belief in h after seeing d (the *posterior* probability), we have

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

where $p(d|h)$ indicates the probability of d if h is true (the *likelihood*), and $\mathcal{H}$ is the set of all hypotheses under consideration (the *hypothesis space*). Bayes' rule can thus be interpreted as a model of learning, with the choice of prior, likelihood, and hypothesis space giving a transparent characterization of the assumptions of the learner. In particular, the prior summarizes the *inductive biases* of a learner – those factors that lead him or her to favor one hypothesis over another when those hypotheses are equally consistent with the data.

Learning and using a language requires solving a variety of inductive problems, and thus provides a natural setting in which to apply probabilistic models of cognition. In this paper, I will discuss three ways in which probabilistic models can be used to shed light on language learning and language evolution: identifying the assumptions necessary for learning to take place, indicating how learning can simultaneously take place at different levels of abstraction, and linking the inductive biases of individual learners to linguistic universals. These three uses of probabilistic models highlight some of the properties that have made them useful in the context of machine learning, and illustrate how they can contribute to a deeper understanding of some of the central issues in cognitive science.

2. Identifying the assumptions of learners

Arguments from philosophy [10] and formal analyses of learning [11, 12, 13] indicate that the key to solving inductive problems is constraining the set of hypotheses under consideration, but the nature and origins of the constraints on human inferences are controversial. For example, much of the language acquisition literature focuses on two extreme positions: that the relevant constraints are strong, innate, and specific to language (e.g., [14]), or that the constraints are weak, and the result of general-purpose learning mechanisms (e.g., [15]). Probabilistic models provide a way to formalize the inductive biases that guide human learning, expressing these constraints through the choice of hypothesis space and the prior distribution over hypotheses.

The simplest way to use probabilistic models to reveal human inductive biases is to construct a set of models that assume different priors, and examine which of these models best characterizes human performance. This approach is consistent with a long tradition of computational modeling in cognitive science, in which parameterized models are fit to human data and compared in order to evaluate claims about the processes behind behavior. Recently, this approach has been used to explore the inductive biases that seem to characterize aspects of human language learning. For example, learning the meaning of a novel word can be modeled as Bayesian inference, with data consisting of a pairing of a word and object and hypotheses corresponding to sets of objects that could be possible referents of the word. Using a model of this kind makes it possible to analyze the prior distribution over word meanings assumed by human learners [16]. A similar approach can be used to analyze how people learn about probabilistic variation in a language: in some circumstances, people tend to "regularize" inconsistent input, making their language more deterministic [17]. This tendency can be captured through a prior that favors extreme probability distributions, where the probability of each variant of a linguistic structure is close to zero or one [18].

A second way to explore inductive biases relevant to language learning is to examine what kind of information simulated learners with different priors can extract from corpora of the speech that adults produce when interacting with children. This strategy can be used to address questions such as whether children receive sufficient information to identify particular kinds of syntactic structures (e.g., [19]). It can also be used to determine what kinds of assumptions a learner needs to make in order to reach adult linguistic competence. For example, an analysis of this kind for the problem of word segmentation – learning the words that appear in continuous speech – suggests that assumptions about the nature of the interaction between words can have a significant effect on how well a simulated learner recovers the correct words from an unsegmented corpus [20].

These two approaches to exploring human inductive biases are quite compatible: corpus based analyses can lead to new hypotheses that are tested through laboratory experiments. In the case of word segmentation, the probabilistic model mentioned in the previous paragraph inspired a series of experiments examining how human word segmentation is affected by various manipulations of the statistical properties of the input [21]. However, analyses that combine these two approaches remain rare, and more research of this kind will be important in order to gather clear evidence about the constraints that guide human language learning, facilitating the development of machine learning systems that exploit these constraints.

3. Learning at different levels of abstraction

Learning a language requires solving inductive problems at many different levels: identifying the phonetic categories into which continuous acoustic signals are divided, grouping these sounds into words, learning the morphological rules characterizing the structure of these words, and inferring how words can be combined together to form sentences. These inductive problems each build on one another, so a natural approach might be to imagine that each problem is solved in turn, with the learner using the solution to one problem as input to the next. However, the results of learning at a higher level (such as current guesses about the words in the language) can be relevant to learning at a lower level (such as identifying phonetic categories), suggesting that a learner might benefit by simultaneously making inferences at both of these levels. Such inferences can be captured by *hierarchical Bayesian models* [9].

A hierarchical Bayesian model is a probabilistic model in which the hypotheses to be inferred from data are expressed at multiple levels of abstraction. The basic idea behind these models is that the knowledge we draw upon in solving inductive problems is represented at many levels, and that Bayesian inference can be applied at any of these levels. To the extent that there are principles that are relevant to many inductive inferences within a domain – such as the sequences of sounds that are likely to comprise words, or the properties of objects that are typically picked out by a word – these principles can be abstracted from experience and used to constrain future inferences. Mathematical analyses show that hierarchical Bayesian models require less and less data to make accurate inductive inferences [22], providing an account of the process of "learning to learn". These formal ideas can be applied to aspects of human cognition, providing insight into how learning at one level can provide useful constraints at another.

The main insight that has resulted from applying hierarchical Bayesian models to language learning is that introducing more unknown variables can often – paradoxically – make learning easier. For example, consider the problem of learning phonetic categories from acoustic signals. This problem can be formalized in terms of estimating a probability distribution over acoustic signals associated with each category, and recent work has shown that this approach results in reasonably good estimation of well-separated phonetic categories [23]. However, infants learning phonetic categories are not solving this problem in isolation from other aspects of language learning: at the same time, they are beginning to pick out the words that comprise their language. Those words consist of a sequence of phonetic categories, and result in regularities in the tendency for sounds to appear in close temporal proximity. Introducing another level of abstraction to the learning problem, in which the learner seeks to simultaneously identify these words as well as the phonetic categories, improves learning by making it possible to exploit this regularity [24].

Similar constraints hold across levels of abstraction in other language learning problems, and can also be captured by hierarchical Bayesian models. For example, such models can be used to explain how children learn that the labels used for objects typically depend on their shape [25]. In other domains, such as causal learning, machine learning systems that make inferences at multiple levels of abstraction outperform more traditional methods that focus on a single level [26]. These results suggest that the ability to operate at multiple levels of abstraction is a property of human learning that might profitably be used in other machine learning systems.

4. Linking inductive biases to universals

Human languages form a subset of all logically possible communication schemes, with some univeral properties being common across languages [27, 28]. These linguistic universals have been taken as providing evidence for strong constraints on language learning, but connecting the inductive biases of individuals to the languages that are spoken in different communities requires considering the role that language learning plays in language evolution. In particular, it requires considering how linguistic universals might emerge from the process of languages

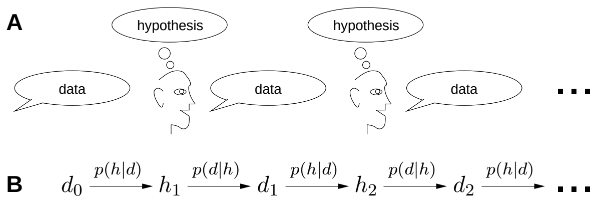
A
hypothesis
hypothesis
data
data
data
B
$d_0 \xrightarrow{p(h|d)} h_1 \xrightarrow{p(d|h)} d_1 \xrightarrow{p(h|d)} h_2 \xrightarrow{p(d|h)} d_2 \xrightarrow{p(h|d)} \ldots$

Figure 1: Iterated learning. (A) Each learner sees data produced by the previous generation, forms a hypothesis about the process by which those data were produced, and uses this hypothesis to produce the data that will be supplied to the next generation. (B) In iterated learning with Bayesian agents, each learner sees data d, and uses Bayes' rule to compute the posterior probability of each hypothesis h, $p(h|d)$. The learner samples a hypothesis from this distribution, and then generates data from the distribution $p(d|h)$.

being learned anew by each generation (e.g., [29]).

The connection between language learning and language evolution can be explored using the *iterated learning model*, a simplified model of the process by which languages are transmitted from one generation to the next [30]. In this model, each generation consists of one or more learners. Each learner sees some data, forms a hypothesis about the process that produced that data, and then produces the data which will be supplied to the next generation of learners. A schematic illustration of this process appears in Figure 1A. By making assumptions about the way in which learners select hypotheses – such as using Bayesian inference – we can begin to explore how individual inductive biases influence the languages that are ultimately produced by iterated learning.

In iterated learning with Bayesian agents, each learner uses Bayes' rule (Equation 1) to infer the language spoken by the previous learner, and generates the data provided to the next learner using the results of this inference (see Figure 1B). The first learner sees data d_0, computes a posterior probability distribution over hypotheses according to Equation 1, chooses a hypothesis h_1 according to this distribution, and generates new data d_1 by sampling from the likelihood function associated with that hypothesis. These data are provided to the second learner, and the process continues, with the nth learner seeing data d_{n-1}, inferring a hypothesis h_n, and generating new data d_n. We assume that all learners share the same prior probability distribution.

As a first step in analyzing this process, we can examine how the hypotheses chosen by the learners change as a result of iterated learning. The probability that the nth learner chooses hypothesis i given that the previous learner chose hypothesis j is

$$p(h_n = i|h_{n-1} = j) = \sum_d p(h_n = i|d)p(d|h_{n-1} = j) \quad (2)$$

where $p(h_n = i|d)$ is the posterior probability obtained from Equation 1. This specifies the transition matrix of a Markov chain, with the hypothesis chosen by each learner depending only on that chosen by the previous learner.

The stationary distribution of the Markov chain defined by Bayesian iterated learning is $p(h)$, the prior assumed by the learners [31]. The Markov chain will converge to this distribution, provided it satisfies the conditions for ergodicity – roughly, that there is no state that cannot be reached in a finite amount of time from any other state (e.g., [32]). This means that the probability that the last in a long line of learners chooses a particular hypothesis is simply the prior probability of that hypothesis. A similar result can be obtained if we consider how the data generated by the learners change: after many generations, the probability that a learner generates data d will be $p(d) = \sum_h p(d|h)p(h)$, the probability of d under the prior predictive distribution.

The convergence of iterated Bayesian learning to a distribution determined by the prior of the learners has important implications for explaining linguistic universals. It indicates that the asymptotic probability with which a language is spoken depends only upon its prior probability, and is not affected by any of the properties of the language. Explaining linguistic universals thus requires explaining why learners assign high prior probability to particular properties of languages. It is tempting to attempt to explain these prior probabilities in terms of an innate language faculty [14], but the biases reflected in the prior probabilities of the learners need not be innate, or language-specific. Every learning algorithm assumes some kind of inductive bias, and this bias is essential to the success of the algorithm [33, 34]. The biases exhibited by human learners could reflect general-purpose information-processing constraints, or result from knowledge acquired in other domains.

Going beyond language evolution, these theoretical results suggest another method for exploring human inductive biases: implement iterated learning in the laboratory with human learners, and examine which hypotheses survive. To test this method, we need to simulate iterated learning using stimuli for which human inductive biases are already well understood. Experiments using one-dimensional functions [35], binary concepts [36], and predictions of everyday quantities [37] have confirmed that iterated learning seems to converge to a distribution over hypotheses consistent with inductive biases documented using other tasks.

Having established that the basic prediction of convergence to the prior seems to hold, we can begin to explore how laboratory simulations of iterated learnimg can be used to explore questions more directly related to language acquisition. As mentioned above, one basic question in language acquisition is how learners deal with probabilistic variation resulting from inconsistent input: whether they regularize the language to a more deterministic form, or simply reproduce the probabilities in the input. Iterated learning provides a sensitive tool for exploring this question, and experiments in the transmission of a simple language suggest that people have a bias towards regularization that emerges over several generations [18].

When combined with probabilistic models of cognition, iterated learning provides a powerful tool for exploring inductive biases. It also provides an interesting link between human and machine learning that goes beyond simply analyzing human learning as statistical inference. The stochastic process that iterated learning with Bayesian agents defines on (h, d) pairs is formally equivalent to Gibbs sampling, a form of Markov chain Monte Carlo that is widely used in Bayesian statistics and statistical physics [38, 39]. This connects iterated learning to an extensive literature examining the properties of Gibbs sampling, including results on convergence rates (e.g., [40]). This connection makes it possible to explore how assumptions about the nature of the hypothesis space considered by language learners influence the rate at which languages converge to the prior, allowing us to begin to explore whether the properties that we see across human languages might simply be residues of a common origin [41].

5. Conclusion

Probabilistic models of cognition provide a new set of tools for characterizing human learning. These tools allow us to analyze the assumptions that underlie human learning, understand how learning can take place at multiple levels of abstraction, and explore the link between learning and cultural evolution. However, they also provide the foundation for developing new machine learning methods that might come closer to human performance: by identifying the nature of human inductive biases, we can begin to develop automated systems that possess the same biases, and hopefully the same ability to rapidly and accurately solve inductive problems.

6. References

[1] A. Newell and H. Simon, "The logic theory machine: A complex information processing system," *IRE Transactions on Information Theory*, vol. IT-2, pp. 61–79, 1956.

[2] J. Laird, P. Rosenbloom, and A. Newell, "Soar: An architecture for general intelligence," *Artificial Intelligence*, vol. 33, pp. 1–64, 1987.

[3] J. R. Anderson, *Rules of the mind.* Hillsdale, NJ: Erlbaum, 1993.

[4] J. McClelland and D. Rumelhart, Eds., *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press, 1986.

[5] S. J. Russell and P. Norvig, *Artificial intelligence: A modern approach*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 2002.

[6] C. M. Bishop, *Pattern recognition and machine learning.* New York: Springer, 2006.

[7] J. R. Anderson, *The adaptive character of thought.* Hillsdale, NJ: Erlbaum, 1990.

[8] M. Oaksford and N. Chater, Eds., *Rational models of cognition.* Oxford: Oxford University Press, 1998.

[9] J. B. Tenenbaum, T. L. Griffiths, and C. Kemp, "Theory-based Bayesian models of inductive learning and reasoning," *Trends in Cognitive Science*, vol. 10, pp. 309–318, 2006.

[10] N. Goodman, *Fact, fiction, and forecast.* Cambridge: Harvard University Press, 1955.

[11] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias-variance dilemma," *Neural Computation*, vol. 4, pp. 1–58, 1992.

[12] M. Kearns and U. Vazirani, *An introduction to computational learning theory.* Cambridge, MA: MIT Press, 1994.

[13] V. N. Vapnik, *The nature of statistical learning theory.* New York: Springer, 1995.

[14] N. Chomsky, *Aspects of the theory of syntax.* Cambridge, MA: MIT Press, 1965.

[15] J. L. Elman, E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett, *Rethinking innateness: A connectionist perspective.* Cambridge, MA: MIT Press, 1996.

[16] F. Xu and J. B. Tenenbaum, "Word learning as Bayesian inference," *Psychological Review*, vol. 114, pp. 245–272, 2007.

[17] C. L. Hudson-Kam and E. L. Newport, "Regularizing unpredictable variation: The roles of adult and child learners in language formation and change," *Language Learning and Development*, vol. 1, pp. 151–195, 2005.

[18] F. Reali and T. L. Griffiths, "The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning," *Cognition*, vol. 111, pp. 317–328, 2009.

[19] S. Foraker, T. Regier, N. Khetarpal, A. Perfors, and J. Tenenbaum, "Indirect evidence and the poverty of the stimulus: The case of anaphoric "one"," *Cognitive Science*, vol. 33, pp. 287–300, 2009.

[20] S. Goldwater, T. L. Griffiths, and M. Johnson, "Contextual dependencies in unsupervised word segmentation," in *Proceedings of Coling/ACL 2006*, 2006.

[21] M. C. Frank, S. Goldwater, V. Mansinghka, T. Griffiths, and J. Tenenbaum, "Modeling human performance in statistical word segmentation," in *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*, 2007.

[22] J. Baxter, "A Bayesian/information theoretic model of learning to learn via multiple task sampling," *Machine Learning*, vol. 28, pp. 7–39, 1997.

[23] G. K. Vallabha, J. L. McClelland, F. Pons, J. Werker, and S. Amano, "Unsupervised learning of vowel categories from infant-directed speech," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 13 273–13 278, 2007.

[24] N. H. Feldman, T. L. Griffiths, and J. L. Morgan, "Learning phonetic categories by learning a lexicon," in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2009.

[25] C. Kemp, A. Perfors, and J. B. Tenenbaum, "Learning overhypotheses with hierarchical Bayesian models," *Developmental Science*, vol. 10, pp. 307–321, 2007.

[26] V. K. Mansinghka, C. Kemp, J. B. Tenenbaum, and T. L. Griffiths, "Structured priors for structure learning," in *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

[27] B. Comrie, *Language universals and linguistic typology.* Chicago: University of Chicago Press, 1981.

[28] J. Greenberg, Ed., *Universals of language.* Cambridge, MA: MIT Press, 1963.

[29] E. Briscoe, Ed., *Linguistic evolution through language acquisition: Formal and computational models.* Cambridge, UK: Cambridge University Press, 2002.

[30] S. Kirby, "Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity," *IEEE Journal of Evolutionary Computation*, vol. 5, pp. 102–110, 2001.

[31] T. L. Griffiths and M. L. Kalish, "A Bayesian view of language evolution by iterated learning," *Cognitive Science*, vol. 31, pp. 441–480, 2007.

[32] J. R. Norris, *Markov Chains.* Cambridge, UK: Cambridge University Press, 1997.

[33] T. M. Mitchell, *Machine learning.* New York: McGraw Hill, 1997.

[34] D. J. C. Mackay, *Information theory, inference, and learning algorithms.* Cambridge: Cambridge University Press, 2003.

[35] M. L. Kalish, T. L. Griffiths, and S. Lewandowsky, "Iterated learning: Intergenerational knowledge transmission reveals inductive biases," *Psychonomic Bulletin and Review*, vol. 14, pp. 288–294, 2007.

[36] T. L. Griffiths, B. R. Christian, and M. L. Kalish, "Using category structures to test iterated learning as a method for identifying inductive biases," *Cognitive Science*, vol. 32, pp. 68–107, 2008.

[37] S. Lewandowsky, T. L. Griffiths, and M. L. Kalish, "The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning," *Cognitive Science*, in press.

[38] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[39] W. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds., *Markov Chain Monte Carlo in Practice.* Suffolk, UK: Chapman and Hall, 1996.

[40] J. S. Liu, W. H. Wong, and A. Kong, "Covariance structure and convergence rate of the Gibbs sampler with various scans," *Journal of the Royal Statistical Society B*, vol. 57, pp. 157–169, 1995.

[41] A. Rafferty, T. L. Griffiths, and D. Klein, "Convergence bounds for language evolution by iterated learning," in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2009.