Advanced Review

Learnability theory

Sean A. Fulop[1]* and Nick Chater[2]

Learnability theory is a body of mathematical and computational results concerning questions such as: when is learning possible? What prior information is required to support learning? What computational or other resources are required for learning to be possible? It is therefore complementary both to the computational project of building machine learning systems and to the scientific project of understanding learning in people and animals through observation and experiment. Learnability theory includes work within a variety of theoretical frameworks, including, for example, identification in the limit, and Bayesian learning, which idealize learning in different ways. Learnability theory addresses one of the foundational questions in cognitive science: to what extent can knowledge be derived from experience?

How to cite this article:
WIREs Cogn Sci 2013, 4:299–306. doi: 10.1002/wcs.1228

INTRODUCTION

Learnability theory is the formal theory of learning, which complements the more empirically motivated concerns of *learning theory* broadly conceived. In learnability theory, one is primarily concerned with theoretical results about whether learning is possible in a given scenario; the details of specific learning algorithms are usually secondary. This article provides a survey of learnability theory with a view to cognitive science applications.

Haussler and Warmuth[1] note that it would be dangerous to try to formalize anything as complex and varied as learning so that it can be subjected to rigorous mathematical analysis. Inevitably, most people will feel that something important about learning has been left out of the idealization—and they will be right. But, as in any area of inquiry, idealization is required to make progress; different idealizations will be appropriate in modeling different aspects of learning.

One fundamental dichotomy is the distinction between *supervised* and *unsupervised* learning. Supervised learning works with *labeled* examples of the target concept, and possibly also labeled nonexamples. It is this type of learning that is the traditional province of learnability theory. Unsupervised learning refers to a wide range of techniques which attempt to extract 'hidden structure' from unlabeled data. Because this approach to learning does not specify precisely what the goal of the learning is, 'effectiveness is a matter of opinion and cannot be verified directly'.[2] As a result, there is not a significant learnability theory for unsupervised learning. There are useful hybrid approaches, however, in which, for example the data is partially labeled, or the learning goal somehow extends beyond the labeling.

*Correspondence to: sfulop@csufresno.edu

[1]Department of Linguistics, California State University Fresno, Fresno, CA, USA

[2]Behavioural Science Group, Warwick Business School, University of Warwick, Coventry, UK

FORMAL MODELS OF LEARNING

A standard way to model learning supposes that the learner receives a data sequence D—this is the *example set* or *learning sample* or *training set*—one item at a time. The data sequence consists of examples. The learner then proposes a hypothesis to characterize what it has learned after each example. Common types of learning problems including learning functions from set X to set Y, or binary-valued assignments to a set of Boolean attributes, or learning sets of formal expressions known as *languages*. One natural goal for the learner is to recover (or perhaps to approximate) the 'true' function, concept, or language from which the data D has been generated.

For example, consider some 'true' function f, such that $y = f(x)$. The learner may then be given pairs

of values (x, y), together with the information that it is a positive example of the function; or perhaps that it is *not* within the function, thus making it a negative example. A learning sample in this setting would be a sequence of such positive/negative pairs. Similarly for language learning, an example might consist of a sequence of symbols, plus the information that this expression is (or is not) within the target language. A learning sample in this setting would then be a sequence of such positive/negative expressions.

Learnability theory is traditionally concerned chiefly with how to set up a problem so whether the 'true' function, concept, or language is learnable can be assessed by mathematical analysis. By contrast, studies of learning within machine learning, cognitive science and psychology have more often focused on the learning algorithm itself—how it could be computationally feasible, cognitively plausible, practically useful, or capture aspects of human learning. Learnability theory also usually focuses on learning as a process in which the learner's hypothesis approaches the target concept as more data is analyzed. Key differences among theoretical frameworks center on the specific way to model the notion of 'approaching the target.'

IDENTIFICATION IN THE LIMIT

While not the first formal analysis of the general problem of learning[3]—surely it was Gold[4] who initiated the modern debates about learnability in linguistics and cognitive science. We thus begin with a brief account of Gold's learning framework, known as *identification in the limit* (i.i.l.). In this framework, the learner receives an unending sequence of examples (either all positive; or labeled as either positive or negative) and the task is to correctly infer a concept (or function, language, grammar, etc.) from a given class of hypotheses. The learner typically will not recognize its own success (for all the learner knows, the very next example might be a counterexample); the solution of the learning problem is the convergence of the learner's hypothesis to the correct one in the limit (i.e., after some finite number of steps, the learner will have 'locked in' on the correct hypothesis and never deviate from it), irrespective of whether we know that the convergence has occurred. If this convergence can occur, in theory, for every concept (or function, or language) within some class, that entire class is said to be learnable in the limit; learnability is thus a property that is relative to a class.

Gold proved results in the specific settings of language learning and function learning, but his most well-known (though not always correctly interpreted) result was the following: given the hypothesis class $\boldsymbol{H}$ encompassing all finite languages (based on the fixed vocabulary), the learning problem on $\boldsymbol{H}$ is solvable in the limit from an unending sequence of positive examples; the problem becomes unsolvable, however, if just a single infinite language is included with the finite languages in $\boldsymbol{H}$. This result resonated with the linguistics community, which was grappling at the time with the best ways to model apparently 'infinite' natural languages using finite representations such as grammars. Gold's result was widely seen[5] as proving that no infinite language, and therefore no human language, was learnable. Given that children do appear to learn (apparently infinite) natural languages readily enough, this result is reminiscent of the folklore that a bumblebee cannot fly according to the laws of aerodynamics. In reality, of course, Gold's theorem does not lead to a catastrophic conclusion. For one thing, no reasonable hypothesis class of human languages should include 'all finite languages,' and should perhaps include *no* finite languages, given that human languages are generally infinite. Nonetheless, i.i.l. from positive examples is also not possible under many other circumstances—and researchers in linguistics and cognitive science focused on learning from positive evidence, because it has widely been assumed that children learn language from experience with the language (i.e., positive examples), and do not receive or utilize information about negative examples (e.g., from having their own mistaken sentences corrected)—this was part of the Input Conditions of Pinker.[6]

It has been possible to make some important mathematical progress in analyzing learning from positive data. Key theoretical results in the Gold framework include the 'locking sequence lemma',[7] showing that if a language is identifiable in the limit from positive examples, there must be a sequence of examples such that, if it begins the learning sample, locks the learner onto the correct conjecture *no matter how the rest of the learning sample continues*. An interesting result about learnable *classes* is the 'tell-tale subset theorem',[8] which states that in any class of languages learnable from positive examples, each language L must have a finite 'tell-tale' subset with the property that, if it is also a subset of another language L' in the class, then L' cannot be a subset of L. This result has an interesting interpretation in the realm of language learning, wherein it explicates the intuitive notion that each natural language must have a finite subset from which the learner can conjecture the entire language without the danger of overgeneralizing (overgeneralization is often viewed as creating a fundamental problem for language acquisition[9]).

Other work in the Gold paradigm has studied the properties of languages which are learnable from a specified finite set of 'good' examples, rather than an unending sequence.[10] This is not identification in the limit, but a variation that concentrates on learning from finite samples; as such it seems especially relevant for cognitive learning. Many other specific variations of identifiability in the limit have been studied, which modify, for example, the constraints on convergence or the constraints on the use of information. A prominent example of the latter is that of a *set-driven* learner,[11] which is required to make its successive conjectures based entirely on the set of distinct examples derived from the learning sequence at each stage, thus ignoring all matters of organization or statistical structure within the sequence.

Another variation of the Gold paradigm that has been suggested on natural cognitive grounds is the model of a *memory-limited* learner. This includes learning algorithms which conjecture hypotheses based only on the latest example received, together with the previous conjecture,[11] and cannot recall previous input. It also includes extensions wherein some bounded amount of additional memory is also permitted.[12] This type of learning has also been developed into a formal definition of a *U-shaped learner;* this is a learning algorithm which goes from correct conjectures to incorrect ones and then back to the correct conjecture as the example sequence is received,[12] a pattern often observed in child language acquisition.

BAYESIAN LEARNING

A very different approach to learning focuses on statistical properties of the input—and attempts to infer the underlying structure that generated that input. Here, the fundamental aim is not identification in the limit, but rather assigning a suitably high probability to the true hypothesis (or a hypothesis that is, in some sense, close to the true hypothesis). This approach applies Bayes' celebrated rule of probability:

$$\Pr(H|D) = \frac{\Pr(D|H)\Pr(H)}{\sum \Pr(D|H_i)\Pr(H_i)}, \qquad (1)$$

where the various H_i are the individual (mutually exclusive and exhaustive) hypotheses in hypothesis class **H**, and D are 'data:' sets of observed sentences, instances of concepts, and so on. The formula states that the probability (the 'posterior' on the left side) of a particular hypothesis H given data D, is proportional to the product of the learner's prior probability for H multiplied by the probability of D on the assumption that H is true (this term, $\Pr(D|H)$, is known as the *likelihood*).

Bayes' rule is an elementary theorem of probability theory, directly derivable from the definition of the conditional probability $\Pr(H|D)$. From the point of view of machine learning and cognitive science, what is distinctive in the Bayesian approach is viewing probabilities as measures of 'degrees of belief' (rather than, e.g., limiting frequencies of repeated events)—this is known as the *subjective* interpretation of probability. The idea is that the learner has some initial prior beliefs, and revises these in the light of observed data, to create posterior beliefs; these posteriors then become the right priors on the arrival of new data, creating new posterior beliefs and so on.

The large literature on Bayesian models in cognitive science[13] views learning and other cognitive problems as involving uncertain reasoning, which can be modeled using probability theory. This raises the controversial question of how far, and in what sense, thought itself should be viewed as probabilistic inference, and whether Bayesian updating, or some approximation, can be empirically observed during learning. For example, Xu and Tenenbaum[14] empirically check the Bayesian updating model against human performance, thereby seeking to validate it as a model of mental processes, rather than simply as a tool to understand the nature of the learning problem. Moreover, in theoretical neuroscience there has been considerable interest in the idea that neural circuits may be carrying out Bayesian calculations, to some approximation.[15]

In most of this research, little attention is paid to learnability. Bayesian methods are often just assumed to work. A natural question, of *convergence*, from the point of view of learnability, is whether, as the number of examples the learner receives increases, the posterior distribution gradually converges on the 'true' hypothesis, that is, the one that actually generated the data.

One result which validates the asymptotic success of Bayesian inference where we consider the posterior distribution (or, more narrowly, we select the *maximum a posteriori* (MAP) hypothesis from that distribution) is known as the Bernstein–von Mises Theorem. It was first proven by Doob,[16] and finally in a general form by Freedman.[17] We begin with the scenario in which the space of possible outcomes is discrete and finite (e.g., rolling dice, flipping coins). Freedman's result states roughly that if we have independent identically distributed (i.i.d.) discrete observations that are limited to a finite set of possible outcomes, 'the posterior probability

converges to point mass at the true parameter value (or hypothesis) among almost all sample sequences.' This result—known as *consistency* of the inference process—holds for any initial prior probabilities which satisfy 'Cromwell's rule,' meaning that none are set equal to 0. The result generalizes to the case of a countable infinity of possible observations, but only provided the initial priors are 'good.' Importantly, there are many 'bad' priors, and therefore many setups for Bayesian inference in which the posterior probability was shown to 'almost never' converge or to converge on the wrong answer.[18,19] In most interesting settings in machine learning or cognitive science, there are at least countably many possible hypotheses (e.g., there is a countable infinity of languages constructed from a finite set of symbols). So, in these settings we cannot in general be sure that Bayesian inference may not accurately converge. There are many subtleties lurking here, and the consistency of Bayesian learning under various conditions is still the subject of much current research.

A second question, of *prediction*, concerns the number of mistakes the learner makes, as learning proceeds—we might hope that this number is as few as possible—this question focuses on the quality of predictions throughout the learning process, rather than picking out the true hypothesis. Three broad approaches to Bayesian prediction can be distinguished. First, given a posterior probability distribution over languages, concepts, or functions, we can predict using the single most probable hypothesis. A second approach is less accurate, but often computationally easier: to sample a single hypothesis from the a posteriori distribution over **H**, for example Ref 20. The third approach is most accurate: to make predictions by averaging over the entire posterior distribution over the hypothesis space,[21] since this leverages all the information that has been gained. This approach is known as *marginalization* or *hypothesis averaging*,[22] and is very frequently approximated in machine learning and cognitive science.[23] It turns out that, with remarkably mild restrictions (e.g., no requirement that the data is sampled independently) bounds can be derived for each of the three approaches (e.g., Refs 24 and 22, extending work by Solomonoff[25] and others). Most strikingly, the cumulative prediction error in predicting a sequence using marginalization is, roughly, bounded by the minimum description length of the true hypothesis.[25]

Note, though, that in many learning contexts, the 'true' hypothesis cannot even be represented by the learner (i.e., it is not contained in **H**)—the hypothesis class is 'mis-specified,' to use statistical terminology. When **H** is mis-specified, convergence to the true hypothesis is out of the question. Indeed, this is presumably the typical case in machine learning, cognition, and natural science: we are only able to create approximate models of reality. What can we say about convergence and prediction in this situation? Does the learner converge on the hypothesis in **H** which is, in some sense, nearest to the truth? And even if the learner definitely has the wrong hypothesis, can its predictions still be, to some extent, reliable? These questions are not well-understood, although positive results are available under quite narrow conditions (e.g., each item of data is independently sampled from the same distribution).[26]

Model Selection

One approach to the mis-specification is simply to reduce our ambitions, from finding the 'true' hypothesis, to finding what is, in some sense, the 'best' hypothesis in **H**.[27] This is the starting point for 'model selection' in statistics. But what defines the best? One approach is to focus on *generalization*; given a fixed set of *training data,* find the model that best predicts unseen data. This type of scenario still invokes Bayes' rule, but now we view it as a formula for the posterior probability of each model, based on the training data. This is a useful procedure for comparing models of unequal complexity (e.g., with different numbers of parameters). There are many possible ways of comparing hypotheses, but few of these are truly 'Bayesian.' One such is to check the relative merits of pairs of models by using the *posterior odds:*

$$\frac{\Pr(H_m|D)}{\Pr(H_l|D)} = \frac{\Pr(H_m)}{\Pr(H_l)} \times \frac{\Pr(D|H_m)}{\Pr(D|H_l)}, \qquad (2)$$

where D is the training data as before. Comparison of the posterior odds is equivalent to evaluating models using the *Bayesian evidence*,[28] the normalizing factor (denominator) in Bayes' formula. As this procedure has massive computational complexity in practice (because the likelihood term can only be calculated by marginalizing over what are usually vast numbers of other unknown parameters), it can sometimes be well approximated by choosing the model with the smallest Bayesian Information Criterion (BIC):[29]

$$\mathrm{BIC} \equiv -2 \ln \mathcal{L}_{\max} + k \ln N \qquad (3)$$

in which $\mathcal{L}_{\max}$ is the maximum likelihood achievable by the model and N is the number of data points provided (size of the learning sample). In practice, $\mathcal{L}_{\max}$ must typically itself be estimated. This criterion has been shown to be asymptotically consistent,

meaning that given a family of hypotheses including the 'true' target, the probability that the BIC will give us the correct hypothesis approaches one as $N \to \infty$ so long as the assumptions underlying the approximation are met.[2] Nonetheless, Bayesian evidence and BIC can diverge dramatically: for example, the inclusion in the models of unconstrained parameters (irrelevant to the learning data) is penalized by the BIC but not by the Bayesian evidence.[30]

Computability and the Universal Prior Distribution

It is sometimes, although not generally, assumed that the function, concept or language generating the data that the learner receives is, in some sense, *computable* (e.g., that the function can be implemented, or the language generated, by some computer program). Focusing on computability turns out to have a surprising side-effect—it allows the formulation of a powerful and general prior distribution.

Suppose that we consider an infinitely long sequence of data (which might have all manner of complex sequential structure, as in music or speech), and suppose that our learner considers only computable probability distributions over these sequences (strictly, these are lower semicomputable semimeasures, a more general class than probability distributions, but we ignore these complications for lack of space here; see Ref 31 for formal discussion). Suppose we wish to choose a prior over these distributions.

It turns out that there is a prior such that, whatever the 'true' distribution (if it is computable), this prior assigns 'almost as much' probability to each sequence as the true distribution (i.e., the ratio between the true probability of a sequence and the probability according to this prior is bounded by a multiplicative constant, however, much data we consider). This prior is known as the *universal prior distribution*.[3,31–33] In essence, the universal prior can be viewed as a mixture of probability distributions over sequences of data, assigning greater weights to the 'simplest' distributions.[31] The simplicity of a probability distribution (or grammar, concept, or function) can be measured by the length of the shortest binary computer program that generates it, a quantity known as *Kolmogorov complexity*.[31] This aspect is recognizable as a version of Occam's razor.

Bayesian prediction using the universal prior is surprisingly powerful. A useful review of the technical results about Bayesian prediction from the universal prior is provided by Hutter,[34] and many are also provided in Ref 31. Chater and Vitányi[35] show that these results have direct applications to learnability in the context of language. It turns out that, using Bayesian inference over the universal prior, it is possible to learn to predict language, assess sentence grammaticality, and produce language arbitrarily well, from exposure to sufficiently large amounts of linguistic input (assuming that input is generated by some computable probabilistic process). There is, though, a crucial limitation, which prevents these results being turned into practical machine learning technology or cognitive scientific computer models: the universal prior is not computable—and hence any algorithm for Bayesian prediction according to this prior can only be an approximation. Moreover, for such approximations, learnability results are typically not available.

PROBABLY APPROXIMATELY CORRECT

The problem of computability is only the tip of a large computational iceberg, which we have so far ignored. A problem afflicting both the i.i.l. approach and Bayesian learning is that, even where computable, the computational complexity of the typical methods are too high, making them cognitively infeasible, for example Ref 36. Indeed, Clark and Lappin[37] argue that computational complexity is a much more powerful constraint on learning than the richness of the available data in domains such as language acquisition. The *probably approximately correct* (PAC) learning framework was introduced by Valiant[38] in order to require the learning to approximate, and equally crucially, be computationally feasible. This is accomplished by allowing approximation of the target, rather than insisting on convergence to the target. We can define PAC-learnability of a concept class or language (possibly comprising functions, languages, etc.) as being fulfilled when we have a learning algorithm such that for each element f in the class and probability ε, the algorithm converges with probability at least $1 - \varepsilon$ to some element h which disagrees with f only on a set of examples having total probability less than ε.[31] In the PAC literature it is typically also required that the learning algorithm should run in polynomial time depending on the size of the provided example sequence.

Is the restriction to tractable algorithms compatible with a focus on elegant, but generally intractable, approaches to learning, such as Bayesian methods? Not necessarily—indeed finding tractable approximations to Bayesian methods is a major

research area. Nonetheless, tractability does, in general, place very substantial constraints on theories of learning.

APPLICATIONS OF LEARNABILITY THEORY

One of the early applications of learnability theory in linguistics was the analysis by Hamburger and Wexler,[39] which showed that the transformational component (mapping from deep structure to surface structure) of a transformational grammar (a Universal Base grammar was assumed to be innate) is learnable from positive examples of deep structure–surface structure pairs (these are not, of course, available to the child). Their learning algorithm was later simplified[11] so that it would be more computationally tractable, which resulted in part from restricting the representational complexity of the input examples from which it would be possible to extract the relevant information. This strategy resulted from the authors' position that 'restricting the power of possible grammars has been, and continues to be, the most important method in our attempt to move toward learnability'.[11] More recently, language learnability research has focused on alternative ways of finding frameworks within which learnability is possible.

Constraining the Hypothesis Space

Later attacks on the learnability problem attempted to constrain the hypothesis space so that it would be learnable. There is a subtle difference between this sort of proposal and Wexler and Culicover's idea, since it is possible to constrain a hypothesis class without necessarily limiting the complexity of the languages or concepts in it. The idea of directly constraining the hypothesis class as providing the answer to learnability has underlain a number of influential theories in cognitive science, particularly in linguistics. Indeed, this seems to be the main argument in favor of Universal Grammar. Perhaps the most highly developed linguistic theory incorporating a heavily constrained hypothesis class is the theory of Principles and Parameters. According to this theory, all natural languages conform to a set of principles of Universal Grammar, and are then learned by a process of setting a finite number of parameters which govern the systematic differences observed between different languages.

The theory of Principles and Parameters was supposed to virtually trivialize the learning problem[40]: as so much of grammar is supposed to be innate, the learner need only set a finite number of binary parameters in response to linguistic input. A simple, intuitive algorithm called Triggering was developed to explain this 'learning by parameter setting.'[41] An amazing analysis by Niyogi,[42] however, proved that learning parameter settings from linguistic data actually ranges from computationally hard to impossible, and the Triggering algorithm is not guaranteed to succeed even asymptotically. This result appears to undercut one of the primary motivations for the Principles and Parameters theory.[a]

The role of the hypothesis space in learnability questions can be very subtle. On the one hand, the hypothesis space is generally thought of as a proper subset of the entire concept space, and this can play some role in constraining the learning problem. On the other hand, it is possible for a learner to conjecture 'worthless' hypotheses in response to some input, while nonetheless being able to converge to targets in some learnable class. This means that the learnable class of concepts can actually be a proper subset of the hypothesis class, for a functioning successful learner in a certain setting.[37] It is not necessary for the learner to always formulate conjectures which are themselves learnable; since the process is asymptotic, only convergence to (or toward) learnable concepts is necessary.

Distributional Learning from Good Examples

One strategy for language learning that has been suggested more than once requires rich learning data to induce a complete grammar. The strategy involves 'semantic bootstrapping'[43]—extracting information from sentence examples annotated with information about their meanings and syntactic structures—together with 'distributional learning,'[6,43] the induction of word usage information from the distribution of words evident in the learning sample. A version of this strategy was described in detail by Fulop,[44,45] who also proved that it could learn a grammar powerful enough for natural language (even a context-sensitive grammar) from a finite set of *good examples*,[10] while also learning the system of parts of speech (grammatical categories) needed for the target language—thus making the procedure a sort of hybrid supervised/unsupervised algorithm. The success of this strategy results in part from the richly annotated sentence data, but also results crucially from strong assumptions about the nature of natural languages—the hypothesis class is essentially restricted to languages in which expressions of the same type are intersubstitutable in a structurally defined way.

CONCLUSION

The study of learnability considers questions of fundamental importance to cognitive science: when is learning possible? What prior information is required? What computational resources are needed? The aim is to prove general mathematical results, which abstract from specific representations and algorithms—and this requires an abstract representation of the learning problem. We have seen that various abstract conceptions of learning are possible, from identification in the limit, to Bayesian approaches, and the PAC framework. While initial results were sometimes interpreted as suggesting very strong 'innate' constraints on the learner, the picture is more nuanced—under some idealizations, it is possible to prove surprisingly strong positive learnability results, under fairly general assumptions. In addition, learnability results can provide a much-needed 'reality check' on cavalier assumptions within statistical frameworks like Bayesian inference. Given that almost all areas of cognition, from perception to motor control to language, appear to involve huge learning challenges, it is likely that the theory of learnability, alongside concrete computational models of learning, will be influential in the future development of cognitive science.

NOTE

[a]Niyogi's proof of the nonviability of the 'learnability results' pertaining to Principles and Parameters went almost completely unnoticed in linguistics.

ACKNOWLEDGMENT

N.C. was supported by ERC Advanced Research Grant 'Cognitive and Social Foundations of Rationality.'

REFERENCES

1. Haussler D, Warmuth M. The Probably Approximately Correct (PAC) and other learning models. In: Wolpert DH, ed. *The Mathematics of Generalization: Proceedings of the SFI/CNLS Workshop on Formal Approaches to Supervised Learning*. Santa Fe, NM: Westview Press; 1995, 17–36.
2. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York: Springer; 2001.
3. Solomonoff RJ. A formal theory of inductive inference Part I. *Inform Cont* 1964, 7:1–22.
4. Gold EM. Language identification in the limit. *Inform Cont* 1967, 10:447–474.
5. Clark A. Unsupervised Language Acquisition: Theory and Practice. D.Phil. Thesis, University of Sussex, 2001.
6. Pinker S. Formal models of language learning. *Cognition* 1979, 7:217–283.
7. Blum L, Blum M. Toward a mathematical theory of inductive inference. *Inform Cont* 1975, 28:125–155.
8. Angluin D. Inductive inference of formal languages from positive data. *Inform Cont* 1980, 45:117–135.
9. Baker CL, McCarthy JJ. *The Logical Problem of Language Acquisition*. Cambridge, MA: MIT Press; 1981.
10. Lange S, Nessel J, Wiehagen R. Language learning from good examples. In: Arikawa S, Jantke KP, eds. *Algorithmic Learning Theory*. Berlin: Springer; 1994, 423–437.
11. Wexler K, Culicover P. *Formal Principles of Language Acquisition*. Cambridge, MA: The MIT Press; 1980.
12. Carlucci L, Case J, Jain S, Stephan F. Results on memory-limited U-shaped learning. *Inform Comput* 2007, 205:1551–1573.
13. Perfors A, Tenenbaum JB, Griffiths TL, Xu F. A tutorial introduction to Bayesian models of cognitive development. *Cognition* 2011, 120:302–321.
14. Xu F, Tenenbaum JB. Word learning as Bayesian inference. *Psychol Rev* 2007, 114:245–272.
15. Ma WJ, Beck J, Latham P, Pouget A. Bayesian inference with probabilistic population codes. *Nature Neurosci* 2006, 9:1432–1438.
16. Doob JL. Applications of the theory of martingales. *Le Calcul des Probabilités et ses Applications*. Paris, 1948, 22–28.
17. Freedman DA. On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann Math Stat* 1963, 34:1386–1403.
18. Freedman DA. On the asymptotic behavior of Bayes estimates in the discrete case II. *Ann Math Stat* 1965, 36:454–456.
19. Diaconis P, Freedman DA. On inconsistent Bayes estimates of location. *Ann Stat* 1986, 14:68–87.
20. Kirby S, Dowman M, Griffiths TL. Innateness and culture in the evolution of language. *Proc Natl Acad Sci U S A* 2007, 104:5241–5245.
21. Wolpert DH. The relationship between PAC, the Statistical Physics framework, the Bayesian framework, and the VC framework. In: Wolpert DH, ed. *The Mathematics of Generalization: Proceedings of the SFI/CNLS*

Workshop on Formal Approaches to Supervised Learning. Santa Fe, NM: Westview Press; 1995, 117–214.

22. Poland J. Consistency of discrete Bayesian learning. *Theoret Comp Sci* 2008, 405:256–273.
23. Tenenbaum JB, Griffiths TL. Generalization, similarity, and Bayesian inference. *Behav Brain Sci* 2001, 24:629–640.
24. Haussler D, Kearns M, Schapire RE. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Mach Learn* 1994, 14:83–113.
25. Solomonoff RJ. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans Inform Theory* 1978, 24:422–432.
26. Vapnik VN. *The Nature of Statistical Learning Theory*. New York: Springer; 1995.
27. Bretthorst GL. Bayesian model selection: examples relevant to NMR. In: Skilling J, ed. *Maximum Entropy and Bayesian Methods*. Dordrecht: Kluwer; 1989, 377–388.
28. Lasenby A, Hobson M. Methods and tools for statistical analyses of CMB data. *CMB and Physics of the Early Universe*, 2006.
29. Schwartz G. Estimating the dimension of a model. *Ann Stat* 1978, 6:461–464.
30. Liddle AR. Information criteria for astrophysical model selection. *Month Notice Roy Astron Soc Lett* 2007, 377:L74–L78.
31. Li M, Vitányi PMB. *An Introduction to Kolmogorov Complexity and Its Applications*. 3rd ed. New York: Springer; 2008.
32. Solomonoff RJ. A formal theory of inductive inference Part II. *Inform Cont* 1964, 7:224–254.
33. Li M, Vitányi PMB. *An Introduction to Kolmogorov Complexity and Its Applications*. New York: Springer-Verlag; 1993.
34. Hutter M. On universal prediction and Bayesian confirmation. *Theoret Comp Sci* 2007, 384:33–48.
35. Chater N, Vitányi PMB. 'Ideal learning' of natural language: Positive results about learning from positive evidence. *J Math Psychol* 2007, 51:135–163.
36. Kwisthout J, Wareham T, van Rooij I. Bayesian intractability is not an ailment that approximation can cure. *Cogn Sci* 2011, 35:779–784.
37. Clark A, Lappin S. Complexity in language acquisition. *Topics Cogn Sci* 2013, 5:89–110.
38. Valiant LG. A theory of the learnable. *Commun ACM* 1984, 27:1134–1142.
39. Hamburger H, Wexler K. A mathematical theory of learning transformational grammar. *J Math Psychol* 1975, 12:137–177.
40. Chomsky N. *Lectures on Government and Binding*. Dordrecht: Foris; 1981.
41. Gibson E, Wexler K. Triggers. *Ling Inq* 1994, 25:407–454.
42. Niyogi P. *The Informational Complexity of Learning: Perspectives on Neural Networks and Generative Grammar*. Boston: Kluwer Academic Publishers; 1998.
43. Pinker S. *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press; 1984.
44. Fulop SA. Grammar induction by unification of type-logical lexicons. *J Logic Lang Inform* 2010, 19:353–381.
45. Fulop SA. Erratum to: grammar induction by unification of type-logical lexicons. *J Logic Lang Inform* 2011, 20:135–136.

