Convergence Bounds for Language Evolution by Iterated Learning

Anna N. Rafferty (rafferty@cs.berkeley.edu)
Computer Science Division, University of California, Berkeley, CA 94720 USA

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Dan Klein (klein@cs.berkeley.edu)
Computer Science Division, University of California, Berkeley, CA 94720 USA

Abstract

Similarities between human languages are often taken as evidence of constraints on language learning. However, such similarities could also be the result of descent from a common ancestor. In the framework of iterated learning, language evolution converges to an equilibrium that is independent of its starting point, with the effect of shared ancestry decaying over time. Therefore, the central question is the *rate* of this convergence, which we formally analyze here. We show that convergence occurs in a number of generations that is $O(n \log n)$ for Bayesian learning of the ranking of n constraints or the values of n binary parameters. We also present simulations confirming this result and indicating how convergence is affected by the entropy of the prior distribution over languages.

Introduction

Human languages share a surprising number of properties, ranging from high level characteristics like compositional mapping between sound and meaning to relatively low-level syntactic regularities (Comrie, 1981; Greenberg, 1963; Hawkins, 1988). One explanation for these universal properties is that they reflect constraints on human language learning, with the mechanisms by which we acquire language being restricted to languages with these properties (e.g., Chomsky, 1965). However, if all modern languages are descended from a common ancestor, these similarities could just reflect the properties of that ancestor. Evaluating these different possibilities requires establishing how constraints on learning influence the properties of languages, and how long it takes for this process to remove the influence of a common ancestor. In this paper, we explore these questions using a simple model of language evolution.

We model language evolution as a process of *iterated learning* (Kirby, 2001). This model assumes that each generation of people learns language from utterances generated by the previous generation. While this model makes certain simplifying assumptions, such as a lack of interaction between learners in the same generation, it has the advantage that it can be analyzed mathematically. Previous research has shown that after some number of generations, the distribution over languages produced by learners converges to an equilibrium that reflects the constraints that guide learning (Griffiths & Kalish, 2007). After convergence, the behavior of learners is independent of the language spoken by the first generation.

These results provide a way to relate constraints on learning to linguistic universals. However, convergence to the equilibrium has to occur in order for these constraints to be the sole factor influencing the languages learners acquire. Our key contribution is providing bounds on the number of generations required for convergence, known as the *convergence time*, which we obtain by analyzing Markov chains associated with iterated learning. Bounding the convergence time is a step towards understanding the source of linguistic universals: If convergence occurs in relatively few generations, it suggests constraints on learning are more likely than common descent to be responsible for linguistic universals.

To bound the number of generations required for iterated learning to converge, we need to make some assumptions about the algorithms and representations used by learners. Following previous analyses (Griffiths & Kalish, 2007), we assume that learners update their beliefs about the plausibility of a set of linguistic hypotheses using Bayesian inference. We outline how this approach can be applied using two kinds of hypothesis spaces that appear in prominent formal linguistic theories: constraint rankings, as used in Optimality Theory (Prince & Smolensky, 2004), and vectors of binary parameter values, consistent with a simple Principles and Parameters model (Chomsky & Lasnik, 1993). In each case, we show that iterated learning with a uniform prior reaches equilibrium after $O(n \log n)$ generations, where n is the number of constraints or parameters.

Analyzing Iterated Learning

Iterated learning has been used to model a variety of aspects of language evolution, providing a simple way to explore the effects of cultural transmission on the structure of languages (Kirby, 2001; Smith, Kirby, & Brighton, 2003). The basic assumption behind the model – that each learner learns from somebody who was themselves a learner – captures a phenomenon we see in nature: Parents pass on language to their children, and these children in turn pass on language to their own children. The sounds that the children hear are the input, and the child produces language (creates output) based on this input, as well as prior constraints on the form of the language.

Formally, we conceptualize iterated learning as follows (see Figure 1). A first learner receives data, forms a hypothesis about the process that generated these data, and then produces output based on this hypothesis. A second learner receives the output of the first learner as data and produces a new output that is in turn provided as data to a third learner.

This process may continue indefinitely, with the t^{th} learner receiving the output of the $(t-1)^{th}$ learner. The iterated learning models we analyze make the simplifying assumptions that language evolution occurs in only one direction (previous generations do not change their hypotheses based on the data produced by future generations) and that each learner receives input from only one previous learner. We first characterize how learning occurs, independent of specific representation, and then give a more detailed description of the form of these hypotheses and data.

Our models assume that learners represent (or act as if they represent) the degree to which constraints predispose them to certain hypotheses about language through a probability distribution over hypotheses, and that they combine these predispositions with information from the data using Bayesian inference. Starting with a *prior* distribution over hypotheses $p(h)$ for all hypotheses h in a hypothesis space H, the *posterior* distribution over hypotheses given data d is given by Bayes' rule,

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

where the *likelihood* $p(d|h)$ indicates the probability of seeing d under hypothesis h. The learners thus shape the language they are learning through their own biases in the form of the prior probabilities: the prior $p(h)$ incorporates the human learning constraints. These probabilities might, for example, tend to favor lword forms with alternating consonant-vowel phonemes. We assume that learners' expectations about the distribution of the data given the hypothesis are consistent with the actual distribution (i.e. that the probability of the previous learner generating data d from hypothesis h matches the likelihood function $p(d|h)$). Finally, we assume that learners choose a hypothesis by sampling from the posterior distribution (although we consider other ways of selecting hypotheses in the Discussion section).[1]

The analyses we present in this paper are based on the observation that iterated learning defines a Markov chain. A Markov chain is a sequence of random variables X_t such that each X_t is independent of all preceding variables when conditioned on the immediately preceding variable, X_{t-1}. Thus, $p(x_t|x_1,\ldots,x_{t-1}) = p(x_t|x_{t-1})$. There are several ways of reducing iterated learning to a Markov chain (Griffiths & Kalish, 2007). We will focus on the Markov chain on hypotheses, where transitions from one state to another occur each generation: the t^{th} learner assumes the data were generated by h_t, where these data are dependent only on the hypothesis h_{t-1} chosen by the previous learner. The transition probabilities for this Markov chain are obtained by summing over the data from the previous time step d_{i-1}, with $p(h_t|h_{t-1}) = \sum_{d_{i-1}} p(h_t|d_{i-1})p(d_{i-1}|h_{t-1})$ (see Figure 1).

Identifying iterated learning as a Markov chain allows us to draw on mathematical results concerning the convergence of

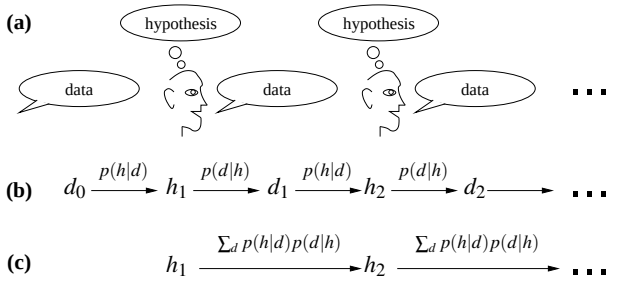

Figure 1: Language evolution by iterated learning. (a) Each learner sees data, forms a hypothesis, and generates the data provided to the next learner. (b) The underlying stochastic process, with d_t and h_t being the data generated by the tth learner and the hypothesis selected by that learner respectively. (c) We consider the Markov chain over hypotheses formed by summing over the data variables. All learners share the same prior $p(h)$, and each learner assumes the input data were created using the same $p(d|h)$.

Markov chains. In particular, Markov chains can converge to a *stationary distribution*, meaning that after some number of generations t, the marginal probability that a variable X_t takes value x_t becomes fixed and independent of the value of the first variable in the chain (Norris, 1997). Intuitively, the stationary distribution is a distribution over states in which the probability of each state is not affected by further iterations of the Markov chain; in our case, the probability that a learner learns a specific grammar at time t is equal to the probability of any future learner learning that grammar. The stationary distribution is thus an equilibrium state that iterated learning will eventually reach, regardless of the hypothesis of the first ancestral learner, provided simple technical conditions are satisfied (see Griffiths & Kalish, 2007, for details).

Previous work has shown that the stationary distribution of the Markov chain defined by Bayesian learners sampling from the posterior is the learners' prior distribution over hypotheses, $p(h)$ (Griffiths & Kalish, 2007). These results illustrate how constraints on learning can influence the languages that people come to speak, indicating that it is possible for iterated learning to converge to an equilibrium that is determined by these constraints and independent of the language spoken by the first learner in the chain.

However, characterizing the stationary distribution of iterated learning still leaves open the question of whether enough generations of learning have occurred for convergence to this distribution to have taken place in human languages. To understand the degree to which linguistic universals reflect constraints on learning rather than descent from a common ancestor, it is necessary to establish bounds on convergence time. Previous work has identified factors influencing the rate of convergence in very simple settings (e.g., Griffiths & Kalish, 2007). Our contribution is to provide analytic upper bounds on the convergence time of iterated learning with relatively complex representations of the structure of a language that are consistent with linguistic theories.

[1]Note that these various probabilities form our model of the learners. Learners need not actually hold them explicitly, nor perform the exact computations, provided that they act as if they do.

Bayesian Language Learning

Defining a Bayesian model of language learning requires choosing a representation of the structure of a language. In this section, we outline Bayesian models of language learning compatible with two formal theories of linguistic representation: Principles and Parameters (Chomsky & Lasnik, 1993), and Optimality Theory (Prince & Smolensky, 2004).

Principles and Parameters

The Principles and Parameters framework postulates that all languages obey a finite set of principles, with specific languages defined by setting the values of a finite set of parameters (Chomsky & Lasnik, 1993). For example, one parameter might encode the head directionality of the language (with the values indicating left- or right-headedness), while another might encode whether covert subjects are permitted. For simplicity, we will assume that parameters are binary. Learning a language is learning the settings for these parameters. In reality, learning is not an instantaneous process. Learners are presented with a series of examples from the target language and may change their parameters after each example. The exact model of learning varies based on assumptions about the learners' behavior (e.g., Gibson & Wexler, 1994). However, in this work, we do not model this fine-grained process, but rather lump acquisition into a single computation, wherein a single hypothesis h is selected on the basis of a single data representation d.

To specify a Bayesian learner for this setting, we define a hypothesis space H, a data representation space D, a prior distribution $p(h)$, and a likelihood $p(d|h)$. Our hypothesis space is composed of all binary vectors of length n: $H = \{0,1\}^n$. We represent the data space as strings in $\{0,1,?\}^n$ in which 0 and 1 indicate the values of parameters that are fully determined by the evidence and question marks indicate underdetermined parameters. For now, we assume a uniform prior, with $p(h) = 1/2^n$ for all $h \in H$. To define the likelihood, we assume the data given to each generation fully specify all but m of the n parameters, with the m unknown parameters chosen uniformly at random without replacement. Then, $p(d|h)$ is zero for all strings d with a 0 or 1 not matching the binary vector h or that do not have exactly m question marks (i.e. those consistent with h). Moreover, we assume that $p(d|h)$ is equal for all strings consistent with h. There are $\binom{n}{m}$ strings consistent with any hypothesis, so $p(d|h) = \frac{m!(n-m)!}{n!}$ for all d consistent with h (see Figure 2).

Applying Bayes' rule (Equation 1) with this hypothesis space and likelihood, the posterior distribution is

$$p(h|d) = \begin{cases} \frac{p(h)}{\sum_{h':h' \vdash d} p(h')} & h \vdash d \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $h \vdash d$ indicates that h is consistent with d. This follows from the fact that $p(d|h)$ is constant for all h such that $h \vdash d$, meaning that the likelihood cancels from the numerator and denominator and the posterior is the prior renormalized over the set of consistent hypotheses. For a uniform prior, the posterior probability of a consistent hypothesis is simply the reciprocal of the number of consistent hypotheses. In the uniform case, 2^m of our hypotheses are valid, so $p(h|d) = \frac{1}{2^m}$.

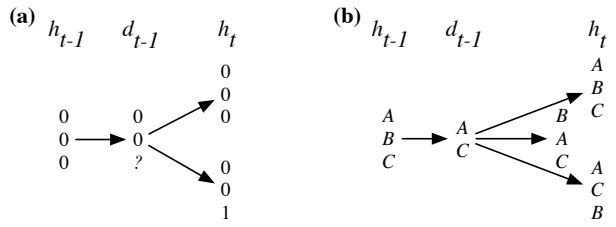

Figure 2: Bayesian language learning. (a) Representation of a hypothesis and data item for Principles and Parameters. On the left is a possible hypothesis for $n = 3$; the center shows a possible data output derived from this hypothesis (with $m = 1$), and the right shows all hypotheses consistent with this data output. (b) A similar representation for OT. The relative ordering of A and C is preserved in the data, but not the ordering of B.

Optimality Theory

In Optimality Theory (OT), learning a language is learning the rankings of various constraints (Prince & Smolensky, 2004; McCarthy, 2004). These constraints are universal across languages and encode linguistic properties. For example, one constraint might encode that high vowels follow high vowels. Whether a construction is well-formed in a language is based on the ranking of the constraints that the construction violates. Specifically, well-formed constructions are those that violate the lowest-ranked constraints. Producing well-formed constructions thus requires determining how constraints are ranked in the target language.

To specify a Bayesian learner, we again need to identify the hypothesis space H, data space D, prior $p(h)$, and likelihood $p(d|h)$. In the OT case, each hypothesis is an ordered list of n constraints, with the order of constraints representing rank. The hypothesis space H is thus the symmetric group of permutations of rank n, S_n, and is of size $n!$. We assume learners see sufficient data to specify the relative ordering of all but m constraints. The data space is then strings over $\{1,2,\ldots n\}$ of length $n-m$, with no repeated elements, ordered from left-to-right in order of precedence (see Figure 2). The relative ordering of the $n-m$ specified constraints is maintained exactly from the generating hypothesis. We again see that the likelihood, $p(d|h)$, is 0 for all orderings not consistent with our hypothesis and equal for all consistent orderings. Analogously to the previous case, we select m constraints to remove from the ranking randomly, giving $\binom{n}{m}$ possible data strings for each hypothesis. This gives $p(d|h) = \frac{m!(n-m)!}{n!}$. Thus, the posterior is the same as that in Equation 2. Since we can freely permute m of our parameters, we have $\frac{n!}{(n-m)!}$ consis-

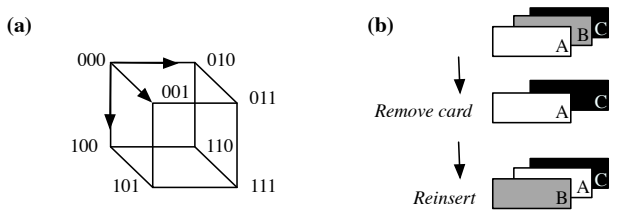

Figure 3: (a) The Principles and Parameters case is analogous to a walk on the hypercube when $m = 1$. Above, the corners (hypotheses) that could be reached after one step (iteration) beginning at 000 are shown. (b) The OT case is analogous to a shuffle in which a random card (in this case, the grey card) is removed and reinserted into a random spot.

tent hypotheses for any data string d. If our prior is uniform, then $p(h|d) = \frac{(n-m)!}{n!}$ for all consistent h and 0 otherwise.

Convergence of Iterated Learning

We now seek to bound the time to convergence of the Markov chain formed by IL. Bounds on the time to convergence are often expressed based on total variation distance. This is a distance measure between two probability distributions μ and ν on some space Ω that is defined as $\|\mu - \nu\| \equiv \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| = \max_{A \subseteq \Omega} |\mu(A) - \nu(A)|$ (Diaconis & Saloff-Coste, 1996). We seek to bound the rate of convergence of the marginal distributions of the h_i to the stationary distribution, expressed via the number of iterations for the total variation distance to fall below a small number ε. This allows us to analytically determine how many iterations the Markov chain must be run to conclude that the current distribution is within ε of the stationary distribution.

To establish these bounds on the convergence rate, we show that the Markov chains associated with iterated learning are analogous to Markov chains for which there are known bounds. First, we consider the Principles and Parameters approach. As described above, we assume each learner has received sufficient data to set all but m of the n parameters in the hypothesis. We first consider the case where there is only one unknown parameter ($m = 1$). Then at each generation of iterated learning, one parameter may be changed at a time. This is equivalent to a random walk on a hypercube, where the hypercube has vertices with binary labels and each vertex is connected by an edge to only those vertices that differ in exactly one digit (see Figure 3). In this Markov chain, vertices are states and edges indicate the possible transitions out of each state. We also assume that there is a transition from each state to itself; this accounts for the case where a learner chooses the same parameter values as the previous generation. Previous analyses show that this Markov chain converges at the rate of $O(n \log n)$ (i.e. at a rate upper-bounded by some constant multiplied by $n \log n$) (Diaconis & Saloff-Coste, 1996). The multiplicative constant absorbs the value of ε indicating the desired distance to convergence.

An intuition for this result comes from the following argument. A sufficient condition for convergence using the previously defined prior is that all parameters have been sampled at least once. This is true because each sample changes the value of the parameter in a way that is insensitive to its current value, making the result equivalent to drawing a vector of values uniformly at random. The time to convergence is thus upper-bounded by the time required to sample all parameters at least once. This is a version of the *coupon-collector* problem, being equivalent to asking how many boxes of cereal one must purchase to collect n distinct coupons, assuming coupons are distributed uniformly over boxes. The first box provides one coupon, but then the chance of getting a new coupon in the next box is $(n-1)/n$. In general, the chance of getting a new coupon in the next box after obtaining i coupons is $(n-i)/n$. The expected time to find the next coupon is thus $n/(n-i)$, and the expected time to find all coupons is $n \sum_{i=1}^{n} \frac{1}{i}$, or n times the nth harmonic number. The bound of $n \log n$ results from an asymptotic analysis of the harmonic numbers, showing that the largest term in the asymptotic approximation grows as $n \log n$ as n becomes large.

Now, we incorporate the fact that each learner does not know m parameters, and can thus change up to m parameters at each iteration. We assume this constant m is fixed and constant across learners. Changing m parameters at each step is equivalent to redefining an iteration as a collection of m successive steps, each of which changes one parameter. Consider choosing which parameter to change at each step independently; this means that we might change a single parameter multiple times in one iteration. This process must converge in $O(\frac{n}{m} \log n)$ iterations, since each iteration in which we can change up to m parameters is equivalent to m steps in our original Markov chain. In our situation, however, we choose the m parameters without replacement, so no parameter changes more than once per iteration. Since the net effect of changing the same parameter twice in one iteration is equivalent to changing it once (from the original value to the final value), changing m different parameters brings us at least as close to convergence as changing fewer than m different parameters. Thus, the Markov chain corresponding to the Principles and Parameters approach to grammar learning is $O(\frac{n}{m} \log n)$ in time to convergence.

We next consider bounding the convergence rate under the assumption that learners learn a ranking over constraints as in OT. We now assume that, at each generation, the learner has sufficient data to rank all but m of n constraints. Again, we first consider the case where $m = 1$. The process of changing the ordering of one item in a permutation while leaving the relative ordering of the other items unchanged has been studied previously in the context of a random-to-random shuffle (see Figure 3). The best bound for the random-to-random shuffle is $O(n \log n)$ (Diaconis & Saloff-Coste, 1993), with the intuitive argument being similar to that given above. As before, we view each iteration as m successive steps, making time to convergence $O(\frac{n}{m} \log n)$.

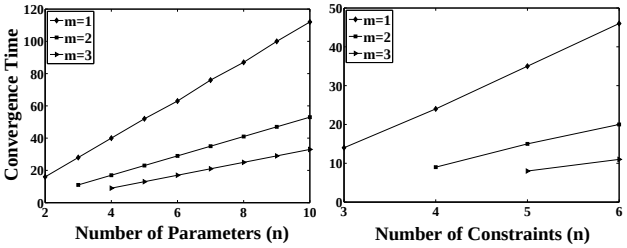

Figure 4: Rate of convergence using a uniform prior with the Principles and Parameters model (left) and the OT model (right). In both cases, time to convergence is proportional to $\frac{n}{m}\log n$.

Empirical Simulations

The preceding sections provide a mathematical analysis of convergence rates for iterated learning; we now turn to several simulations to show convergence behavior and to confirm and extend our mathematical results. We first demonstrate convergence time for various choices of the size of the hypothesis space and the amount of information contained in the data, fixing a uniform prior, and then examine whether the entropy of the prior has an effect on convergence time.

To demonstrate the dependence of convergence time on n and m, we show in Figure 4 the effect of varying these quantities for the uniform prior. For all simulations, we show expected iterations to convergence (number of iterations for $\|p(h_t) - p(h)\| < 0.0001$) given that the observed starting point is distributed according to the stationary distribution. As expected, as n increases, convergence time increases slightly more than linearly, and as m increases, convergence time decreases proportionally.

One variable that could affect convergence time that was not previously considered is the entropy of the prior distribution: i.e., whether the prior made all hypotheses equally likely or put almost all weight on only a few hypotheses. Our previous analyses show how convergence time varies with the size of the hypothesis space, but non-uniform priors provide another way to model constraints on learning that might influence convergence. The uniform prior is the unique maximum entropy distribution for any hypothesis space. However, there is no unique solution for achieving a given entropy for a distribution with k values. Thus, we altered entropy in the following, non-unique way. We define one hypothesis h_p as the prototypical hypothesis. Then, we calculate the distance between h_p and h for each hypothesis h using an appropriate distance measure Δ. For Principles and Parameters, we used the Hamming distance. For OT, we used Kendall's tau, a distance measure for permutations (Diaconis, 1998). Then, for all h, $p(h) \propto \exp(-\beta\Delta(h,h_p))$. Changing β changes the entropy of the distribution. Changing entropy in this manner gives our priors a characteristic shape: h_p has maximum probability, and the probability of other hypotheses decreases with distance from h_p.

In the simulations involving variable entropy, we fixed n and m for the two linguistic representations ($n = 7, m = 2$ for Principles and Parameters; $n = 5, m = 2$ for OT) and varied β to adjust entropy. Our results showed that entropy and expected time to convergence were positively correlated for both representations: as entropy increased, expected time to convergence also increased (Figure 5). This suggests that the constraints on learning provided by a non-uniform prior behave similarly to a reduction in the size of the hypothesis space in their effects on convergence time.

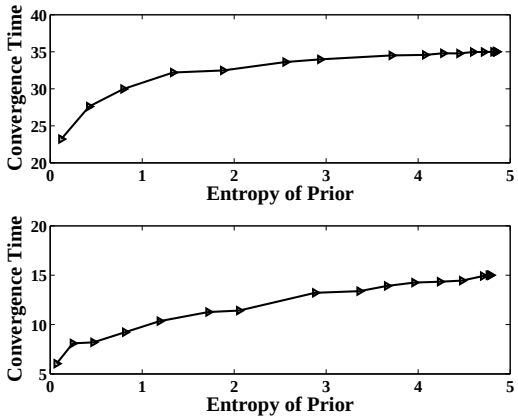

Figure 5: Relationship of the entropy of the prior and convergence behavior. For both Principles and Parameters (top) and OT (bottom), entropy and time to convergence are positively correlated.

Discussion

We began with the question of whether sufficiently many generations of language learning have occurred for similarities across languages to be the result of biases in human learning rather than a common origin. Using iterated learning to explore this question, we showed that the convergence time of Markov chains associated with iterated learning can be bounded, and our simulations confirm the relationship between the complexity of the hypothesis space (n), the degree to which incoming data (language) limits the choice of language (m), and number of generations to convergence. The key result is that the time to convergence for two plausible linguistic representations is $O(\frac{n}{m}\log n)$.

This result has two interesting implications. First, the fidelity with which languages are transmitted between generations – reflected in the value of m – has a direct effect on the rate of convergence. While we have only considered integer values of m, our results also apply to fractional values. For example, if a parameter changes value on average every ten generations, then the convergence time is bounded by taking m to be 0.1; m is thus the expected number of parameters changed per generation. Consequently, it may be possible to estimate m from historical records for different languages. Second, when we vary n, the time to convergence increases a little more than linearly but the size of the hypothesis space increases exponentially. Thus, relatively rapid convergence should occur even with very large hypothesis spaces.

These results provide constraints on the size of the hypoth-

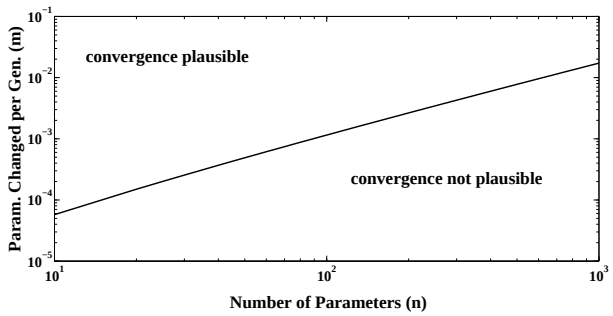

Figure 6: Values of m and n for which human language could have converged for the Principles and Parameters model, under our simplifying assumptions.

esis space and the fidelity of learning necessary for the distribution of human languages to have reached an equilibrium. In the case of the Principles and Parameters model, the exact (rather than asymptotic) bound is known, and the bound itself very tightly tracks the point at which convergence occurs. Consequently, we can identify exactly which values of m and n would make convergence plausible, given some assumptions about the number of generations over which languages have been evolving. Figure 6 shows the plausible values of m and n given the approximately 100,000 years that anatomically modern humans have existed and assuming 25 years per generation. Several authors have estimated the number of parameters that might be required for a Principles and Parameters model; these estimates range from as low as 20 to 50-100 parameters (Kayne, 2000; Lightfoot, 1999). Thus, given this range of values for n, we can see that convergence is plausible for a variety of values of m. While this graph is only a rough guide given the strong simplifying assumptions of our model, it allows some understanding of how our analysis applies to actual human language.

In conducting our analyses, we assumed that learners sample from the posterior distribution over hypotheses. Alternative methods of selecting a hypothesis, such as selecting the hypothesis with the maximum posterior probability (MAP) and exponentiated sampling, have been considered in previous work (Griffiths & Kalish, 2007; Kirby, Dowman, & Griffiths, 2007). In the case of a uniform prior, both methods are equivalent to the sampling method we considered since all hypotheses with non-zero probability have the same probability in the uniform case; thus, our analyses of convergence time hold. In the non-uniform case, raising the posterior to the power of γ before sampling is equivalent to multiplying the β parameter in the model we used to construct our non-uniform priors by γ. Thus, predictions concerning convergence time can also be made for exponentiated sampling in the non-uniform case. For MAP in the non-uniform case, convergence to the prototype hypothesis (that with the highest probability in the non-uniform prior) will occur. The time in which this occurs is still $O(\frac{n}{m}\log n)$: at every step, the learner changes unknown parameters to match the prototype, producing another coupon collector problem with the worst case being that where all n parameters differ from the prototype.

Our key result is thus that language evolution by iterated learning can converge remarkably quickly to the prior – in time that increases linearly as the hypothesis space increases exponentially in size. This result is suggestive about the nature of linguistic universals, although we are hesitant to draw strong conclusions yet. Several restrictive assumptions went into our analysis that could affect convergence time. For example, the lack of interaction between learners means that there is no pressure to adopt a shared communication scheme, and learning from just one other learner removes the opportunity for errors in transmission to be corrected. We also make no assertions about the source or nature of the constraints that limit the size or entropy of the prior over hypotheses. However, our analyses are a step towards a more complete understanding of the origin of universals, with future models exploring the impact of these assumptions and developing a more complete account of how languages change over time.

Acknowledgements We thank the anonymous reviewers for their helpful comments. This work is supported by an NSF Graduate Fellowship (ANR) and by NSF grant BCS-0704034 (TLG).

References

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N., & Lasnik, H. (1993). The theory of principles and parameters. In J. Jacobs, A. von Stechow, W. Sternefeld, & T. Vannemann (Eds.), *Syntax: An international handbook of contemporary research* (pp. 506–569). Berlin: Walter de Gruyter.

Comrie, B. (1981). *Language universals and linguistic typology*. Chicago: University of Chicago Press.

Diaconis, P. (1998). *Group representations in probability and statistics*. Institute of Mathematical Statistics.

Diaconis, P., & Saloff-Coste, L. (1993). Comparison techniques for random walk on finite groups. *The Annals of Probability*, *21*(4), 2131–2156.

Diaconis, P., & Saloff-Coste, L. (1996). Random walks on finite groups: A survey of analytic techniques. *Probability measures on groups and related structures*, *XI*.

Greenberg, J. (Ed.). (1963). *Universals of language*. Cambridge, MA: MIT Press.

Griffiths, T. L., & Kalish, M. L. (2007). A Bayesian view of language evolution by iterated learning. *Cognitive Science*, *31*, 441-480.

Hawkins, J. (Ed.). (1988). *Explaining language universals*. Oxford: Blackwell.

Kayne, R. S. (2000). *Parameters and universals*. New York: Oxford University Press.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, *5*, 102-110.

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

Lightfoot, D. (1999). *The development of language: Acquisition, change and evolution*. Blackwell: Oxford.

McCarthy, J. J. (2004). *Optimality theory in phonology: A reader*. Malden: Wiley-Blackwell.

Norris, J. R. (1997). *Markov chains*. Cambridge, UK: Cambridge University Press.

Prince, A., & Smolensky, P. (2004). *Optimality theory: Constraint interaction in generative grammar*. Blackwell Publishing.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: a framework for the emergence of language. *Artificial Life*, *9*, 371–386.