Cognitive Science 31 (2007) 441–480

Language Evolution by Iterated Learning With Bayesian Agents

Thomas L. Griffiths[a], Michael L. Kalish[b]

[a]*University of California, Berkeley*
[b]*University of Louisiana, Lafayette*

Received 15 September 2005; received in revised form 4 May 2006; accepted 26 July 2006

Abstract

Languages are transmitted from person to person and generation to generation via a process of iterated learning: people learn a language from other people who once learned that language themselves. We analyze the consequences of iterated learning for learning algorithms based on the principles of Bayesian inference, assuming that learners compute a posterior distribution over languages by combining a prior (representing their inductive biases) with the evidence provided by linguistic data. We show that when learners sample languages from this posterior distribution, iterated learning converges to a distribution over languages that is determined entirely by the prior. Under these conditions, iterated learning is a form of Gibbs sampling, a widely-used Markov chain Monte Carlo algorithm. The consequences of iterated learning are more complicated when learners choose the language with maximum posterior probability, being affected by both the prior of the learners and the amount of information transmitted between generations. We show that in this case, iterated learning corresponds to another statistical inference algorithm, a variant of the expectation-maximization (EM) algorithm. These results clarify the role of iterated learning in explanations of linguistic universals and provide a formal connection between constraints on language acquisition and the languages that come to be spoken, suggesting that information transmitted via iterated learning will ultimately come to mirror the minds of the learners.

Keywords: Bayesian models; Language evolution; Iterated learning; Cultural transmission

1. Introduction

Languages change as they are passed from person to person, and from generation to generation. A variety of explanations have been proposed for different aspects of language change, as part of a growing literature on language evolution (e.g., Hurford, Studdert-Kennedy, & Knight, 1998; Briscoe, 2002; Christiansen & Kirby, 2003). The key idea motivating much of this work is that language change can be understood as a process of cultural evolution, with

Correspondence should be addressed to T. L. Griffiths, University of California, Berkeley, Department of Psychology, 3210 Tolman Hall #1650, Berkeley, CA 94720-1650. E-mail: tom_griffiths@berkeley.edu

languages themselves being subject to evolutionary forces. Accounts of language evolution differ in the kind of forces they see as fundamental, appealing to analogues of the forces of selection, mutation, and genetic drift that appear in biological evolution. For example, accounts focusing on selection consider the consequences of a language for the "fitness" of its speakers – that is, their tendency to produce further speakers of that language (e.g., Komarova et al., 2001; Nowak et al., 2001, 2002).

While explanations of the properties of languages based on the fitness of their speakers are intuitively appealing, it is important to take into account the possibility that those properties could be produced by other processes. In biology, the prominent role of selection in early evolutionary theory has more recently been supplemented by the suggestion that much of the variation in the genome is the consequence of mutation and genetic drift, the forces that govern the fidelity with which genetic information is transmitted from one generation to the next (Kimura, 1983). In the case of language evolution, biological transmission is replaced by cultural transmission, and in particular, learning. Each person learns a language from the utterances produced by other people who were once language learners themselves. The variation introduced by learning is the analogue of mutation and genetic drift in language evolution, and it has been suggested that many of the properties of human languages might simply arise from this process of *iterated learning* (Kirby, 1999, 2001; Brighton, 2002; Briscoe, 2002; Kirby & Hurford, 2002; Smith, Kirby, & Brighton, 2003; Kirby et al., 2004).

In order to study the consequences of learners learning from other learners, Kirby (2001) introduced the *iterated learning model*. In this model, the process of language evolution is idealized as the transmission of languages over a sequence of discrete generations, each consisting of one or more learners. The first learner sees some linguistic data such as a set of utterances, and forms a hypothesis about the language that might have produced it. Using this hypothesis, the learner generates a new set of utterances, which are provided to the next learner as data. This process continues, with each learner seeing data, forming a hypothesis, and generating data for the next learner, as illustrated schematically in Fig. 1(a). Formalizing the process of iterated learning in this way makes it possible to analyze its predictions for language change (e.g., Kirby, 2001; Brighton, 2002; Smith et al., 2003).

Accounts of language evolution have the potential to shed light not just on the dynamics of language change, but on the kinds of languages we might ultimately expect to be produced by such a process. One of the main applications of the iterated learning model has been attempting to explain the origins of linguistic universals (e.g., Kirby, 2001; Brighton, 2002; Smith et al., 2003). Human languages form a subset of all logically possible communication schemes, with certain universal properties being shared by all languages. Some of these properties concern the fundamental structure of language. For example, all human languages are *compositional*: the elements of utterances correspond to the elements of the events they describe (Krifka, 2001). Other linguistic universals govern subtle and surprisingly specific aspects of the grammar of human languages (Greenberg, 1963; Comrie, 1981; Hawkins, 1988).

Linguistic universals have traditionally been explained by appealing to innate constraints imposed by a system specific to the acquisition of language that is part of the human genetic endowment (e.g., Chomsky, 1965). Universals are viewed as a manifestation of these innate constraints. Iterated learning potentially provides an alternative explanation, suggesting that properties such as compositionality can emerge as a consequence of many generations of

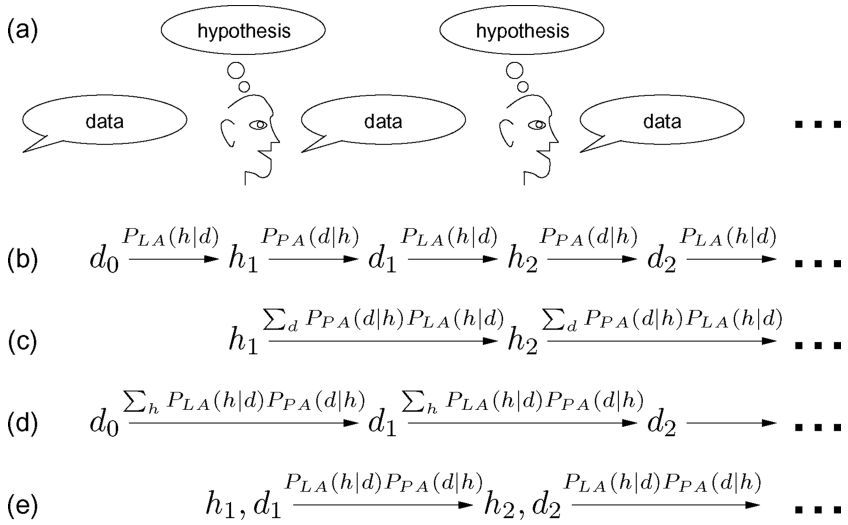

Fig. 1. (a) Language is learned anew by each generation, a process which has been proposed as an explanation for the existence of linguistic universals. Each learner sees data—a set of utterances—produced by the previous generation, forms a hypothesis about the language from which those utterances were produced, and uses this hypothesis to produce the data that will be supplied to the next generation. (b) Dependencies among variables in the stochastic process induced by iterated learning. (c) Reduction to a Markov chain on hypotheses. (d) Reduction to a Markov chain on data. (e) Reduction to a Markov chain on hypothesis-data pairs.

learners applying general-purpose learning algorithms. In particular, explanations of linguistic universals based upon iterated learning have tended to focus on the idea that the finite amount of information that can be communicated from one speaker to another imposes a "bottleneck" on the transmission of language between generations. If particular properties of languages make it easier to pass through that bottleneck, then many generations of iterated learning might allow those properties to become universal. For example, it has been argued that the regular structure of compositional languages means that they can be learned from less data, and are thus more likely to pass through the information bottleneck (Kirby, 2001; Brighton, 2002; Smith et al., 2003).

The central role of iterated learning in language evolution suggests that we should attempt to develop a deeper understanding of its implications. In particular, the possibility that iterated learning with general-purpose learning algorithms might explain linguistic universals requires determining the consequences of iterated learning for learners with different kinds of inductive biases. Previous work has explored the influence of language acquisition on language evolution via a number of avenues (Briscoe, 2002), including simulation of language learners (Kirby, 2001; Brighton, 2002; Smith et al., 2003) and use of formal models of population dynamics (Niyogi & Berwick, 1995, 1997a,b; Nowak, Plotkin, & Jansen, 2000; Komarova et al., 2001; Nowak et al., 2001, 2002). These approaches indicate that languages with specific properties, such as compositionality, can be produced by iterated learning with specific learning algorithms. However, there are no general results indicating the consequences of iterated learning for arbitrary properties of languages or broad classes of learning algorithms.

In this paper, we present a detailed analysis of iterated learning for the case where the learners are rational Bayesian agents. Assuming that our learners use Bayes' rule allows us to characterize their biases through a prior probability distribution over hypotheses. We consider two learning algorithms based on Bayesian inference: sampling from the posterior distribution over hypotheses, and choosing the hypothesis that has maximum posterior probability. In both cases, the consequences of iterated learning are strongly influenced by the prior of the learners. When learners sample, iterated learning results in convergence of the probability that a learner speaks a particular language to the prior probability the learner assigns to that language. This convergence occurs regardless of the nature of the languages or the amount of data available to each learner, indicating that iterated learning can produce systematic results in the absence of an information bottleneck effect. The consequences of iterated learning when learners maximize are more complicated, with the amount and accuracy of the data passed between learners playing an important role, but are still governed by the prior distribution assumed by the learners. We also show that iterated learning with sampling and maximizing each correspond to algorithms that are widely used in statistics, providing a rich source of further formal results.

The plan of the paper is as follows. Section 2 formally defines iterated learning, and presents some general results characterizing its consequences. Section 3 introduces the basic ideas behind Bayesian inference, and how these ideas apply to modeling language acquisition. Section 4 presents our results on the consequences of iterated learning when learners sample languages from their posterior distributions. Section 5 considers the case where learners choose the hypothesis with greatest posterior probability. Section 6 illustrates the predictions of this account by discussing an example of iterated learning in detail: the emergence of compositionality. Section 7 shows how this framework can be extended to characterize the consequences of iterated learning in an unbounded population of Bayesian learners. Section 8 concludes the paper, considering the implications of these results for understanding language evolution and processes of cultural transmission more generally.

2. Analyzing iterated learning

Let $\mathcal{D}$ denote the set of data, d, that a learner might observe, and $\mathcal{H}$ denote the set of hypotheses, h, that the learner might entertain about the origins of those data.[1] In the case of language learning, each hypothesis $h \in \mathcal{H}$ is a language, and the data $d \in \mathcal{D}$ are a set of utterances. Each learner has a *learning algorithm* that specifies a procedure for choosing a hypothesis h upon observing data d, and a *production algorithm* that specifies a procedure for choosing new data d given a hypothesis h. Each learning algorithm, LA, defines a probability distribution over hypotheses given data, $P_{LA}(h|d)$, and each production algorithm, PA, defines a probability distribution over data given hypotheses, $P_{PA}(d|h)$. In the following analyses, we will assume that all learners use the same learning and production algorithms.

Simulations of iterated learning are typically conducted in a setting where each generation consists of a single learner who receives data produced by the learner in the previous generation (Kirby, 2001; Brighton, 2002; Smith et al., 2003). Thus, the first learner sees data d_0, samples a hypothesis h_1 from $P_{LA}(h_1|d_0)$, and generates new data d_1 from $P_{PA}(d_1|h_1)$. These data

are provided to the second learner, and the process continues, with the nth learner sampling a hypothesis from $P_{LA}(h_n|d_{n-1})$, and generating new data from $P_{PA}(d_n|h_n)$. This defines a stochastic process on the variables $d_0, d_1, \ldots$ and $h_1, h_2, \ldots$, as illustrated in Fig. 1 (b).

Using this formal framework, we can be precise about the questions we aim to answer by analyzing iterated learning. To model language change, we need to understand the *dynamics* of iterated learning: how the distribution over h_n and d_n changes as a function of time. To evaluate the predictions that iterated learning makes about linguistic universals, we need to understand its *asymptotic behavior*: what the distribution on h_n and d_n will be after many generations (that is, as n becomes large). We can answer both of these questions by analyzing the stochastic process defined by iterated learning using mathematical results characterizing the behavior of Markov chains. Before outlining the relevance of these results to iterated learning, we will briefly summarize the properties of Markov chains. More detailed introductions are provided by Rosenthal (1995), Norris (1997), and Kemeny & Snell (1983).

2.1. A brief introduction to Markov chains

A Markov chain is a sequence of random variables $v_0, v_1, \ldots$ such that

$$P(v_n|v_0, v_1, \ldots, v_{n-1}) = P(v_n|v_{n-1}) \tag{1}$$

meaning that v_n is independent of all of its predecessors, given v_{n-1}. We will restrict our attention to finite Markov chains, where each state of the Markov chain v_n is an element of a discrete set $\mathcal{V}$, with k members. We will index the elements of this set using $i, j \in \{1, \ldots, k\}$. A Markov chain is *homogeneous* if $P(v_n|v_{n-1})$ is constant for all values of n. In this case, we can fully describe the Markov chain with a *transition matrix* $\mathbf{T} = (t_{ij})$, such that

$$t_{ij} = P(v_n = i|v_{n-1} = j) \tag{2}$$

where i and j are states of the Markov chain. Since $P(v_n = i|v_{n-1} = j)$ is a probability distribution, $\sum_{i=1}^{k} t_{ij} = 1$.

The dynamics of a finite Markov chain can be characterized using linear algebra. Let $\mathbf{p}_0 = (p_{0j})$ be a vector encoding our knowledge about the initial state of the Markov chain, with $p_{0j} = P(v_0 = j)$. Then we can write

$$P(v_1 = i|\mathbf{p}_0) = \sum_j P(v_1 = i|v_0 = j)P(v_0 = j|\mathbf{p}_0) \tag{3}$$

$$= \sum_{j=1}^{k} t_{ij} p_{0j} \tag{4}$$

where $P(v_1 = i|\mathbf{p}_0)$ is shorthand for the probability that $v_1 = i$ given that $\mathbf{p}_0$ encodes the distribution of v_0. If we use $\mathbf{p}_1$ to denote the distribution $P(v_1 = i|\mathbf{p}_0)$, we can write the result in matrix form, with

$$\mathbf{p}_1 = \mathbf{T}\mathbf{p}_0 \tag{5}$$

being the product of the matrix $\mathbf{T}$ and the vector $\mathbf{p}_0$. Similarly, we can write

$$\mathbf{p}_n = \mathbf{T}^n \mathbf{p}_0 \tag{6}$$

multiplying by $\mathbf{T}$ each time we add a new variable.

One way of understanding the asymptotic behavior of a Markov chain is to look for the equivalent of "fixed points." The *stationary distribution* of a Markov chain with transition matrix $\mathbf{T}$ is a distribution $\boldsymbol{\pi}$ such that

$$\boldsymbol{\pi} = \mathbf{T}\boldsymbol{\pi} \tag{7}$$

meaning that the probability distribution over states at point n is the same as the distribution over states at point $n - 1$. This distribution is "stationary" because once it has been reached, the probability of a variable being in a particular state will remain constant. Drawing on linear algebra once again, Eq. 7 identifies $\boldsymbol{\pi}$ as an *eigenvector* of the matrix $\mathbf{T}$ with an *eigenvalue* of 1. The requirement that $\sum_{i=1}^{k} t_{ij} = 1$ means that $\mathbf{T}$ is a stochastic matrix, so its largest (or "first") eigenvalue, λ_1, is 1, making $\boldsymbol{\pi}$ the first eigenvector of $\mathbf{T}$. A standard result in the theory of Markov chains indicates that the asymptotic distribution over states of the chain will approach the stationary distribution as n becomes large, regardless of the initial state of the chain. More formally,

$$\lim_{n\to\infty} \mathbf{T}^n \mathbf{p}_0 = \boldsymbol{\pi} \tag{8}$$

for any $\mathbf{p}_0$, implying that $\lim_{n\to\infty} P(v_n|v_0) = \pi(v_n)$ for any v_0. The rate of convergence depends on the magnitude of the second eigenvalue, λ_2, decreasing as $|\lambda_2|$ increases towards 1 (a simple proof is provided by Rosenthal, 1995).

Equation 8 makes it straightforward to determine the asymptotic behavior of a Markov chain: we need only find the first eigenvector of the transition matrix, which can be done using a variety of analytic and numerical methods (e.g., Stewart, 1994). However, there is one caveat on this convergence result: the Markov chain needs to be *ergodic*. Markov chains with finite state spaces are ergodic if they satisfy two conditions, being *irreducible* and *aperiodic*. A Markov chain is irreducible if every state has a non-zero probability of ever reaching every other state after some finite number of iterations. It is aperiodic if the greatest common divisor of the times at which it is possible for a state to return to itself is 1 for all states.

The most common way in which ergodicity is violated is through the existence of "sinks"—states (or sets of states) that the chain enters but never leaves (in violation of irreducibility). A chain with multiple sinks will eventually become stuck in one of them. The probability of entering a particular sink will depend on the initial state, so the asymptotic independence implied by Eq. 8 does not hold. However, it is easy to check whether any finite Markov chain is ergodic by finding the eigenvalues of the transition matrix: a Markov chain is ergodic if and only if it has just one eigenvalue of unit magnitude (see Rosenthal, 1995). Even if the underlying chains are not ergodic, predictions can still be made about the asymptotic probability of different states, although they require using more sophisticated tools for the analysis of Markov chains (e.g., Kemeny & Snell, 1983).

2.2. Markov chains on hypotheses and data

If we can reduce a stochastic process to a Markov chain, we have gone a long way towards understanding its properties: its dynamics are specified by Eq. 6, and its asymptotic behavior is characterized by Eq. 8. We now return to the stochastic process defined by iterated learning, showing that we can obtain answers to our questions about the consequences of iterated learning by reducing it to two Markov chains: a Markov chain on hypotheses, and a Markov chain on data.

A standard way to analyze probabilistic models is to consider the consequences of summing out a subset of the random variables in the model. The iterated learning model outlined above defines a joint probability distribution on both the data seen by learners and the hypotheses they infer, $P(d_0, h_1, d_1, h_2, \ldots)$. Summing over all possible values for the data seen by each learner defines a probability distribution on hypotheses alone, $P(h_1, h_2, \ldots)$. The dependencies among these variables are shown in Fig. 1(c), taking the form of a Markov chain. The state space of this Markov chain is $\mathcal{H}$, and its transition matrix is $\mathbf{Q} = (q_{ij})$, whose elements give the probability that a learner chooses hypothesis i after seeing data generated from hypothesis j,

$$q_{ij} = P_{LA,PA}(h_n = i|h_{n-1} = j) = \sum_{d \in \mathcal{D}} P_{LA}(h_n = i|d) P_{PA}(d|h_{n-1} = j), \tag{9}$$

where $P_{LA}(h_n = i|d)$ and $P_{PA}(d|h_{n-1} = j)$ are determined by the learning and production algorithms respectively, as specified above. This Markov chain has a stationary distribution $\boldsymbol{\theta} = (\theta_i)$ satisfying

$$\boldsymbol{\theta} = \mathbf{Q}\boldsymbol{\theta} \tag{10}$$

or, departing from vector notation for a moment, $\theta_i = \sum_j q_{ij}\theta_j$.

We can use a similar approach to derive a Markov chain on the data generated by the learners. Summing over the hypotheses entertained by each learner, we obtain a probability distribution $P(d_0, d_1, \ldots)$ with the dependency structure shown in Fig. 1 (d), being a Markov chain on $d_0, d_1, \ldots$. The state space of this Markov chain is $\mathcal{D}$, and the transition matrix $\mathbf{R}$ has elements r_{ij} indicating the probability that a learner produces data $d_n = i$ given that they saw data $d_{n-1} = j$,

$$r_{ij} = P_{LA,PA}(d_n = i|d_{n-1} = j) = \sum_{h \in \mathcal{H}} P_{PA}(d_n = i|h) P_{LA}(h|d_{n-1} = j). \tag{11}$$

The stationary distribution of this Markov chain is $\boldsymbol{\rho} = (\rho_i)$ such that $\boldsymbol{\rho} = \mathbf{R}\boldsymbol{\rho}$.

The Markov chains on hypotheses and data induced by iterated learning can be used to answer questions about its dynamics and asymptotic behavior. The process of language change can be predicted using the transition matrices $\mathbf{Q}$ and $\mathbf{R}$. The properties of the languages produced by iterated learning are indicated in the stationary distributions $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$, providing the underlying Markov chains are ergodic. The conditions for ergodicity have intuitive interpretations in the context of iterated learning. For example, the Markov chain on hypotheses would not be ergodic if there existed more than one language such that once a learner had chosen that language every subsequent learner would also select that language (i.e., the language acts as a sink).

2.3. An example: Two languages

To provide a concrete illustration of the ideas introduced in the previous section, we will work through a simple example of iterated learning. Following a strategy adopted by Niyogi and Berwick (1995, 1997a, 1997b), we will consider the case where learners are faced with a choice between two languages, L_1 and L_2. In this case, the transition matrix $\mathbf{Q}$ has four elements

$$\mathbf{Q} = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix} \tag{12}$$

where q_{ij} is defined in Eq. 9, with $h = i$ being the hypothesis that the language is L_i. q_{11} and q_{22} represent the probability that each language is faithfully transmitted from one learner to the next, while q_{21} and q_{12} indicate failures of transmission, being the probability that a learner acquires L_2 from data produced from L_1 and the probability that a learner acquires L_1 from data produced from L_2 respectively.

With just two languages, it is easy to find the stationary distribution of the Markov chain. $\boldsymbol{\theta}$ will be a distribution over just two languages, with θ_1 being the probability that $h = 1$ and θ_2 being the probability that $h = 2$. By the definition of the stationary distribution, we have

$$\theta_1 = q_{11}\theta_1 + q_{12}\theta_2 \tag{13}$$

from which we obtain

$$\theta_1 = \frac{q_{12}}{q_{12} + q_{21}} \tag{14}$$

by exploiting the fact that $q_{21} = 1 - q_{11}$, $q_{12} = 1 - q_{22}$, and $\theta_2 = 1 - \theta_1$. Thus, the stationary probability of each of the two languages is determined by the relative fidelity with which those languages are transmitted.

Finally, we can compute the eigenvalues of $\mathbf{Q}$. Since $\mathbf{Q}$ is a 2×2 matrix, it has two (not necessarily unique) eigenvalues. The first eigenvalue, λ_1, is 1, since $\mathbf{Q}$ is the transition matrix of a Markov chain, and corresponds to the eigenvector defined by Eq. 14. The second eigenvalue is given by

$$\lambda_2 = 1 - q_{12} - q_{21} \tag{15}$$

and corresponds to the eigenvector $(1, -1)^T$. The second eigenvalue thus gets closer to 1 as q_{12} and q_{21} get closer to zero, slowing convergence of the Markov chain to its stationary distribution, because it becomes difficult to move between states. When $q_{12} = q_{21} = 0$ there is no movement between languages across generations, and both languages act as a sink. The language spoken by the nth generation is thus completely determined by the language spoken by the first learner. The Markov chain is thus not ergodic, and this is reflected in the fact that $\lambda_2 = 1$. This simple example can also be used to illustrate the other way in which ergodicity can be violated. If $q_{12} = q_{21} = 1$, then every generation speaks a different language from that which preceded it. In this case, the language spoken by the nth learner is completely determined by the language spoken by the first learner and the parity of n. Applying Eq. 15

gives $\lambda_2 = -1$, so the magnitude of the second eigenvalue is $|\lambda_2| = 1$, consistent with the fact that the Markov chain is not ergodic.

2.4. Summary

Formalizing iterated learning makes analyzing its consequences straightforward. The reduction of iterated learning to a Markov chain allows us to determine its dynamics and asymptotic behavior by computing a transition matrix and finding its first eigenvector. For example, in the case of the Markov chain on hypotheses, we want the transition matrix $\mathbf{Q}$ and the probability distribution $\boldsymbol{\theta}$ that satisfies Eq. 7. The first eigenvector $\boldsymbol{\theta}$ can be computed numerically for any choice of learning and production algorithms, provided the state space, $\mathcal{H}$, is small or the transition matrix, $\mathbf{Q}$, is sparse. Stronger assumptions about the learning and production algorithms make it possible to go beyond these general results and obtain analytic expressions for $\boldsymbol{\theta}$. In previous work, this has been done for two simple learning algorithms: one that memorizes the data, and one that has no memory at all (Komarova et al., 2001; Nowak et al., 2001, 2002; Komarova & Nowak, 2003). However, these previous analyses assumed that all languages are equally similar to one another, and only allowed a very coarse characterization of the biases of learners via the size of the set of languages being considered. In the remainder of this paper, we derive analytic results that apply to a broad class of learning algorithms which can incorporate a variety of biases for arbitrarily related sets of languages. Our analysis is based upon the assumption that the learners are Bayesian agents.

3. Iterated learning with Bayesian agents

Bayesian agents use a principle of probability theory, called Bayes' rule, to infer the process that was responsible for generating some observed data. Assume that a learner has a *prior* probability distribution, $P(h)$, that encodes that learner's biases by specifying the probability the learner assigns to the truth of each hypothesis $h \in \mathcal{H}$ before seeing d. Bayes' rule states that the probability that an agent should assign to each hypothesis after seeing d—known as the *posterior* probability, $P(h|d)$—is

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)} \tag{16}$$

where $P(d|h)$—the *likelihood*—indicates how likely d is under hypothesis h, and $P(d)$ is the probability of d averaged over all hypotheses,

$$P(d) = \sum_{h \in \mathcal{H}} P(d|h)P(h) \tag{17}$$

which is sometimes called the *prior predictive distribution*.

There are several arguments for exploring iterated learning with Bayesian agents. First, Bayes' rule is a fundamental principle of rational action in statistics and economics (e.g., Savage, 1954; Robert, 1994; Jaynes, 2003), and is used in a variety of models of human cognition (e.g., Anderson, 1990; Oaksford & Chater, 1998; Chater & Oaksford, 1999; Tenenbaum & Griffiths, 2001). Consequently, our analyses will have a direct connection to formal models of

learning and decision-making that are already used to explain human behavior. Second, algorithms based on Bayesian inference are widely used for learning different aspects of language in computational linguistics (e.g., Manning & Schütze, 1999), and previous work on iterated learning has examined algorithms which have a direct Bayesian interpretation, such as minimum description length (Brighton, 2002; Smith et al., 2003). Finally, Bayes' rule makes the biases of learners explicit, encoding those biases in a prior probability distribution over hypotheses. Using Bayesian agents thus provides us with a direct way to explore the influence of the biases of learners on the consequences of iterated learning.

We can use Bayes' rule to model language acquisition by assuming that each hypothesis h is a language, and the data d are a set of utterances sampled from the target language.[2] The likelihood, $P(d|h)$, indicates the probability of observing a particular set of utterances d if the language h were the target. If we assume that the learners have accurate knowledge of the production algorithm in use, the probability they should associate with d if the target language is h is simply the probability of d under that production algorithm, $P_{PA}(d|h)$. Applying Bayes' rule, we have

$$P(h|d) = \frac{P_{PA}(d|h)P(h)}{P_{PA}(d)} \tag{18}$$

where

$$P_{PA}(d) = \sum_{h \in \mathcal{H}} P_{PA}(d|h)P(h). \tag{19}$$

The specific values of $P(h|d)$ will be determined by the prior, $P(h)$. The assumption that all learners share the same learning algorithm (made above, to guarantee that our Markov chain is homogeneous) requires that all learners share the same prior.

3.1. Interpreting priors, hypotheses, and data

The standard interpretation of the prior, $P(h)$, as representing the extent to which the learner believes in a hypothesis before seeing any data is perhaps not the best way to understand the role that it plays under this view of language acquisition. The prior is better seen as determining the amount of evidence that a learner would need to see in order to adopt a particular language. Thinking of the prior as expressing the amount of evidence a learner would need in order to choose a particular language makes it clear how it can encode the biases of learners: only hypotheses with positive prior probability will enter into consideration, and hypotheses with higher prior probabilities are easier to learn (requiring less evidence, and ultimately less data).

Our formal analyses will not make a commitment to the nature of the prior, the hypotheses, or the data. Consequently, they are consistent with many different approaches to modeling language acquisition, from artificial neural network models (Rumelhart & McClelland, 1986), in which the hypotheses are continuous functions represented by the weights of a network (MacKay, 1995; Neal, 1992), to parameter-setting models, in which hypotheses are configurations of a small number of discrete parameters (Gibson & Wexler, 1994; Niyogi & Berwick, 1996). The Bayesian framework is not supposed to be interpreted as a statement of the mechanistic process by which language acquisition takes place, with learners maintaining

a hypothesis space in their heads and updating a distribution over those hypotheses. Rather, it is a *computational level* analysis (Marr, 1982), as is generally emphasized in rational models of cognition (Anderson, 1990; Oaksford & Chater, 1998; Chater & Oaksford, 1999), focusing on the abstract computational problem and a method for solving that problem. So long as the actual process underlying language acquisition approximates this solution, our results will have implications for understanding human behavior.

Finally, it is important to note that the prior distribution assumed by a learner should not be interpreted as reflecting innate constraints specific to language acquisition. The prior simply collects together all of the factors affecting how easily a learner will come to entertain a particular hypothesis. There are many such factors other than language-specific innate constraints: data from other domains that is independent of the observed linguistic data given a hypothesis about the structure of language, but nonetheless affects the beliefs that the learner entertains about that structure; information-processing constraints, such as limitations on working memory; or the inductive bias associated with some kind of general-purpose learning algorithm. Every learning algorithm assumes some kind of inductive bias, and this bias is essential to the success of the algorithm (Geman, Bienenstock, & Doursat, 1992; Kearns & Vazirani, 1994; Vapnik, 1995).

3.2. Applying Bayes' rule with just two languages

We can illustrate how Bayes' rule can be used to model language acquisition by returning to the case where learners are faced with a choice between just two languages. For the purpose of illustration, and in accord with previous work on iterated learning (Kirby, 2001; Brighton, 2002; Smith et al., 2003), we will assume that each language is a mapping from meanings to utterances. The data d consist of a set of a single input, x, and corresponding output, y, and hypotheses h are probability distributions over y for each x. The nth learner sees data, $d_{n-1} = (x_{n-1}, y_{n-1})$, and then generates an output y_n in response to a new input x_n.

To keep the example as simple as possible, we will assume that outputs y can only take two values, which we will denote 0 or 1. To slightly abuse an example introduced by Quine (1960), we might imagine that both languages contain only a single utterance—"Gavagai!"—which is made in response to the presence of some objects ($y = 1$), but not for others ($y = 0$). We will use $\mathcal{X}$ to denote the set of objects, and $\mathcal{G}_1$ and $\mathcal{G}_2$ to denote the sets of values of $x \in \mathcal{X}$ for which $y = 1$ for the languages L_1 and L_2, respectively. Using $\mathcal{S}$ to denote the subset of the values of x on which the two languages agree, we have $\mathcal{S} = (\mathcal{G}_1 \cap \mathcal{G}_2) \cup (\overline{\mathcal{G}_1} \cap \overline{\mathcal{G}_2})$. The relationship among these sets is illustrated in Fig. 2.

Applying Bayes' rule (Eq. 18) requires us to specify the production algorithm and prior. We assume that the production of an (x, y) pair involves two stages: an object x is sampled from the world according to some distribution $P(x)$ that is constant across the two languages, and a learner then produces the appropriate value of y with probability $1 - \epsilon$, where ϵ is small (and definitely less than 0.5).[3] Consequently, we have

$$P_{PA}(d = (x, y)|h = i) = \begin{cases} P(x)(1 - \epsilon) & \text{if } y = I(x \in \mathcal{G}_i) \\ P(x)\epsilon & \text{otherwise} \end{cases} \tag{20}$$

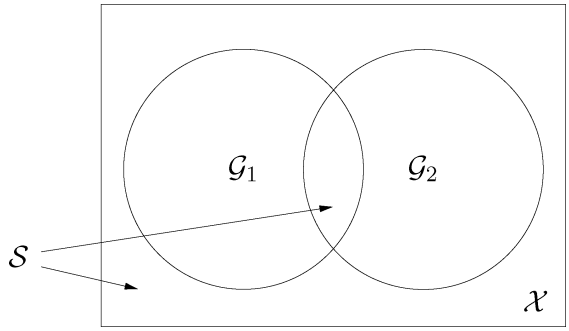

Fig. 2. Sets involved in the two language example. $\mathcal{X}$ is the set of all objects in the domain, with each $x \in \mathcal{X}$ being an object. $\mathcal{G}_1$ and $\mathcal{G}_2$ are the sets of objects that are likely to elicit a spontaneous utterance of "Gavagai!" from a speaker of the languages L_1 and L_2, modeled as the output y associated with the input x taking value 1. $y = 0$ for all objects outside the appropriate set. $\mathcal{S}$ is the set of objects on which the two languages agree, with speakers of either language being equally likely to produce the same response.

where $I(\cdot)$ takes the value 1 when its argument is true, and 0 otherwise. As a prior, we will assume that $P(h = 1) = \alpha$ and $P(h = 2) = 1 - \alpha$, where $0.5 < \alpha < 1$ (i.e., both hypotheses have positive prior probability, but L_1 is favored by the prior).

The posterior distribution over hypotheses given an observed (x, y) pair breaks down into three cases. The first case is where $x \in \mathcal{S}$, and both languages make the same prediction about the value of y. Consequently, $P_{PA}(d|h)$ is constant, and the posterior probability that $h = 1$ is simply the prior probability, α. The second case is where $x \in (\mathcal{G}_1 - \mathcal{G}_2)$ and $y = 1$, or $x \in (\mathcal{G}_2 - \mathcal{G}_1)$ and $y = 0$. The posterior probability that $h = 1$ is then

$$P(h = 1 \mid d) = \frac{(1 - \epsilon)\alpha}{(1 - \epsilon)\alpha + \epsilon(1 - \alpha)} \tag{21}$$

which will favor L_1 over L_2. The third case is when the values of y are reversed (that is, if $x \in (\mathcal{G}_1 - \mathcal{G}_2)$ and $y = 0$, or $x \in (\mathcal{G}_2 - \mathcal{G}_1)$ and $y = 1$). The posterior probability that $h = 1$ is then

$$P(h = 1|d) = \frac{\epsilon\alpha}{\epsilon\alpha + (1 - \epsilon)(1 - \alpha)} \tag{22}$$

which will favor L_1 only if $\epsilon > 1 - \alpha$.

This example illustrates how the prior probability of a language can be interpreted as the amount of evidence that needs to be seen for a learner to choose that language. According to Bayes' rule (Eq. 16), the posterior probability of a language is simply the normalized product of the likelihood and prior. If we see a set of utterances d such that $P(d|h = 2) > P(d|h = 1)$, then d provides evidence for L_2 over L_1. This is exactly the third case mentioned in the previous paragraph, where the value of y for the observed x is consistent with L_2 but not L_1. However, the posterior distribution will only favor L_2 over L_1 if $P(d|h = 2)P(h = 2) > P(d|h = 1)P(h = 1)$. Thus, if the prior strongly favors L_1 over L_2, we need to see evidence that is strongly in favor of L_2 to even consider it a possibility. Hence the condition on ϵ: the observed data have to be sufficiently unlikely under the hypothesis that the target language is L_1 to overwhelm the prior.

3.3. Summary

Bayesian inference allows us to characterize how learners should update their beliefs about hypotheses in the light of data, and makes the inductive biases of learners explicit through the use of a prior distribution. However, the idea that learners use Bayes' rule does not directly identify a learning algorithm of the form required for our analysis of iterated learning. By our definition, a learning algorithm has to specify the probability with which a learner selects a hypothesis after observing data d. In the remainder of the paper, we will analyze iterated learning using two learning algorithms based on Bayesian inference: sampling from the posterior distribution over hypotheses, and choosing the hypothesis with greatest posterior probability.

4. Sampling from the posterior distribution

The simplest way to translate a posterior distribution over hypotheses into a learning algorithm is to assume that learners sample hypotheses according to their posterior probability. That is,

$$P_{\text{samp}}(h \mid d) = P(h \mid d) \tag{23}$$

where $P(h \mid d)$ is defined in Eq. 18. This approach postulates that learners engage in a form of "probability matching," with their choices directly reflecting their posterior distribution. Probability matching is a robust phenomenon observed in human learning (reviews are provided by Myers, 1976 and Vulkan, 2000), and is commonly assumed in cognitive modeling. Response probabilities are often taken as directly corresponding to posterior probabilities in Bayesian models of cognition (e.g., Anderson, 1990; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003), and a similar assumption appears in a variety of models of categorization and choice behavior in the guise of Luce's (1959) choice rule (Nosofsky, 1986, 1987; Ashby, 1992; Kruschke, 1992; Ashby & Maddox, 1993; Ashby & Alfonso-Reese, 1995).

Analyzing sampling from the posterior distribution can also be motivated from the perspective of Bayesian statistics. The posterior distribution encodes a great deal of information, which can be lost by selecting only a single hypothesis. The best way to make accurate predictions is to use all of this information, averaging over hypotheses (e.g., Hoeting, Madigan, Raftery, & Volinsky, 1999). Thus, a learner who has seen data d_{n-1} and wants to predict data d_n should compute

$$P(d_n \mid d_{n-1}) = \sum_{h \in \mathcal{H}} P(d_n \mid h) P(h \mid d_{n-1}) \tag{24}$$

where d_n and d_{n-1} are taken to be independent conditioned on h. This strategy of hypothesis averaging is widely used in Bayesian models of cognition (e.g., Shepard, 1987; Anderson, 1990; Tenenbaum & Griffiths, 2001). While we generally assume that each learner selects a single hypothesis, our analysis of the Markov chain on data that results from the assumption that learners sample from the posterior distribution also applies to the case where learners average over hypotheses, since the definition of the transition matrix $\mathbf{R}$ in Eq. 11 remains the same whether h is sampled or summed out analytically.

4.1. The evolution of hypotheses and data

We can now formally analyze the consequences of iterated learning, calculating the transition matrices and stationary distributions for the Markov chains on hypotheses and data. We will consider these two cases in turn.

For the Markov chain on hypotheses, with transition matrix $\mathbf{Q}$ defined in Eq. 9, taking $\theta_i = P(h = i)$ satisfies the definition of the stationary distribution given in Eq. 10, with $\boldsymbol{\theta} = \mathbf{Q}\boldsymbol{\theta}$. Specifically, we have

$$P(h_n = i) = \sum_j P_{\text{samp},PA}(h_n = i \mid h_{n-1} = j)P(h_{n-1} = j) \tag{25}$$

$$= \sum_j \sum_{d \in \mathcal{D}} P_{\text{samp}}(h_n = i \mid d)P_{PA}(d|h_{n-1} = j)P(h_{n-1} = j) \tag{26}$$

$$= \sum_{d \in \mathcal{D}} P_{\text{samp}}(h_n = i \mid d) \sum_j P_{PA}(d|h_{n-1} = j)P(h_{n-1} = j) \tag{27}$$

$$= \sum_{d \in \mathcal{D}} P_{\text{samp}}(h_n = i \mid d)P_{PA}(d) \tag{28}$$

$$= \sum_{d \in \mathcal{D}} \frac{P_{PA}(d \mid h_n = i)P(h_n = i)}{P_{PA}(d)} P_{PA}(d) \tag{29}$$

$$= P(h_n = i) \sum_{d \in \mathcal{D}} P_{PA}(d \mid h_n = i), \tag{30}$$

where Eq. 28 uses Eq. 19 and Eq. 29 uses Eq. 23 and 18. Since $P_{PA}(d \mid h_n = i)$ is a probability distribution over d for any production algorithm PA, the sum in the last line evaluates to 1, providing our result.

The stationary distribution of the Markov chain on hypotheses is thus the prior distribution. Using the results on the convergence of Markov chains summarized in Section 2.1, this gives a simple characterization of the asymptotic behavior of iterated learning: the probability that a learner entertains a particular hypothesis h will converge to the prior probability of that hypothesis, $P(h)$, as the number of learners in the chain increases. Thus, in the case of language evolution, the probability that a learner speaks a particular language will converge to the prior probability of that language, and the distribution over languages will directly reflect the inductive biases of the learners.

An analogous result can be obtained for the Markov chain on data. For the transition matrix $\mathbf{R}$ defined in Eq. 11, taking $\rho_i = P_{PA}(d = i)$, the prior predictive distribution specified in Eq. 19, satisfies $\boldsymbol{\rho} = \mathbf{R}\boldsymbol{\rho}$. Specifically, we have

$$P_{PA}(d_n = i) = \sum_j P_{\text{samp},PA}(d_n = i|d_{n-1} = j)P_{PA}(d_{n-1} = j) \tag{31}$$

$$= \sum_j \sum_{h \in \mathcal{H}} P_{PA}(d_n = i|h)P_{\text{samp}}(h \mid d_{n-1} = j)P_{PA}(d_{n-1} = j) \tag{32}$$

$$= \sum_j \sum_{h \in \mathcal{H}} P_{PA}(d_n = i \mid h) \frac{P_{PA}(d_{n-1} = j \mid h) P(h)}{P_{PA}(d_{n-1} = j)} P_{PA}(d_{n-1} = j) \quad (33)$$

$$= \sum_j \sum_{h \in \mathcal{H}} P_{PA}(d_n = i \mid h) P_{PA}(d_{n-1} = j \mid h) P(h) \quad (34)$$

$$= \sum_{h \in \mathcal{H}} P_{PA}(d_n = i \mid h) P(h) \sum_j P_{PA}(d_{n-1} = j \mid h) \quad (35)$$

$$= \sum_{h \in \mathcal{H}} P_{PA}(d_n = i \mid h) P(h), \quad (36)$$

where Eq. 33 uses Eq. 23 and 18 and Eq. 35 uses the fact that $P_{PA}(d_{n-1} = j \mid h)$ sums to 1 over all d_{n-1}. The result is the definition of $P(d_n = i)$ from Eq. 19.

This analysis of the behavior of the Markov chain on data complements our analysis of the Markov chain on hypotheses, indicating that the stationary distribution is the prior predictive distribution. Consequently, the probability that a learner produces data d converges to the probability that d would be produced by sampling a hypothesis from the prior and then sampling data according to that hypothesis. In the case of language evolution, this means that the distribution over utterances produced by a learner will ultimately be the distribution we would expect if learners were simply sampling languages according to their prior.

The stationary distributions of these two Markov chains indicate that iterated learning converges to the prior when learners sample from their posterior distributions. An intuitive explanation for this result is that the inference made by each learner provides another opportunity for the prior to affect the distribution over hypotheses. The data seen by the first learner or the hypothesis they entertain will affect the conclusions drawn by the next few learners, but is ultimately only a single piece of information, while the prior asserts its effect on each iteration. Thus, the distribution over hypotheses should move closer to the prior on each iteration. Since the only distribution over hypotheses that is invariant under this influence is the prior itself (demonstrated formally through the fact that the prior is the stationary distribution for the Markov chain on hypotheses), we should expect that iterated learning will remain at that distribution once it has been reached.

4.2. Iterated learning by sampling and the Gibbs sampler

A deeper formal explanation for the consequences of iterated learning can be obtained by noting a correspondence between iterated learning and a class of inference algorithms used in Bayesian statistics. This requires considering another way of reducing the stochastic process that iterated learning defines on both hypotheses and data to a Markov chain. If we choose to group the hypothesis inferred by a learner and the data generated by that learner into a single variable, we obtain the dependency structure shown in Fig. 1(e), which is a Markov chain on hypothesis-data pairs. The state space of this Markov chain is the Cartesian product of $\mathcal{H}$ and $\mathcal{D}$, and the transition matrix has elements corresponding to $P(h_n, d_n \mid h_{n-1}, d_{n-1}) = P_{LA}(h_n \mid d_{n-1}) P_{PA}(d_n \mid h_n)$. Again, this Markov chain will converge to a stationary distribution, provided it satisfies the conditions for ergodicity. Determining

the stationary distribution is straightforward, as this Markov chain takes a form commonly encountered in Markov chain Monte Carlo.

Bayesian statistics often requires working with complex probability distributions, which can be hard to compute analytically. A standard solution to this problem is to apply the Monte Carlo principle, drawing a set of samples from a distribution and performing calculations with those samples rather than the distribution itself (excellent tutorials on Monte Carlo methods are provided by Neal, 1993 and Mackay, 2003). However, sometimes even sampling from a distribution can be difficult, particularly if the distribution is defined over a large state space. This has led to the development of a variety of algorithms for generating samples from probability distributions using Markov chains, which are known as Markov chain Monte Carlo (MCMC) algorithms (an introduction to MCMC is given in Gilks, Richardson, & Spiegelhalter, 1996).

The basic idea behind MCMC is to construct a Markov chain that has the distribution from which one wants to sample as its stationary distribution. Thus, for a target distribution on a variable $\mathbf{v}$ with k components, $\mathbf{v} = (v_1, v_2, \ldots, v_k)$, we would define a Markov chain with a state space corresponding to the different values of $\mathbf{v}$ and a transition matrix designed to produce the target distibution, $P(\mathbf{v})$ as its stationary distribution. The MCMC algorithm then samples a succession of states from this Markov chain, each being a set of values for $\mathbf{v}$. Once this has been done sufficiently many times for the Markov chain to converge to its stationary distribution, the subsequent samples can be treated like samples from $P(\mathbf{v})$ (although the fact that these samples are drawn from a Markov chain means that they will be correlated, making the effective sample size smaller than the total number of samples).

One of the most common methods that is used to construct Markov chains that converge to a particular stationary distribution is *Gibbs sampling* (Geman & Geman, 1984). The Gibbs sampler for a target distribution $P(\mathbf{v})$ is the Markov chain defined by drawing each component of $\mathbf{v}$ from its distribution conditioned on the current values of all other variables, with v_i being drawn from $P(v_i \mid v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k)$. There are a number of variants on this procedure, but one standard method is to cycle through the variables in turn (this is called a "systematic scan" Gibbs sampler). Thus, we would initialize the Markov chain by setting $v_1, v_2, \ldots, v_k$ to some arbitrary initial values, and then draw v_1 from its distribution conditioned on the current values of $v_2, v_3, \ldots, v_k$, then v_2 conditioned on the current values of $v_1, v_3, \ldots, v_k$ (including the newly assigned value of v_1), and so forth. Each complete sweep through the variables constitutes one iteration of the Markov chain, and this procedure is repeated until the Markov chain has converged to its stationary distribution and the desired number of samples have been drawn.

To return to iterated learning, consider the Gibbs sampler for the target distribution $P(d, h) = P_{PA}(d \mid h)P(h)$. The variable of interest is the hypothesis-data pair (d, h), having components d and h, and the sampler would alternate between drawing h conditioned on the current value of d and d conditioned on the current value of h. The corresponding conditional distributions are $P_{\text{samp}}(h|d)$ and $P_{PA}(d \mid h)$, respectively. Alternating between drawing from these conditional distributions is exactly the procedure followed in iterated learning, with the Markov chain defined by Gibbs sampling being that shown in Fig. 1 (e). Consequently, the convergence results for the Gibbs sampler apply to iterated learning, with the distribution over hypothesis-data pairs converging to $P_{PA}(d, h)$. This relationship can also be used to derive the

results for the stationary distribution of the Markov chains on hypotheses and data presented in Section 4.1.

This demonstration that iterated learning is a Gibbs sampler is, to our knowledge, the first instance of a connection between Markov chain Monte Carlo and human cognition. In addition to offering insight into why iterated learning converges to the prior, it provides a source of further formal results about the dynamics of iterated learning, thus helping to characterize the process of language change. The rate of convergence of Markov chains induced by Gibbs sampling has been extensively analyzed by statisticians (Geman & Geman, 1984; Tanner & Wong, 1987; Schervish & Carlin, 1992; Liu, Wong, & Kong, 1995). These analyses indicate that the distance between the distribution over the state space after n iterations and the target distribution decreases geometrically with n, using some standard measures of the distance between probability distributions. The same result carries over to iterated learning, indicating that each learner will bring us ever closer to the prior. The analysis of MCMC algorithms is still an ongoing project in statistics, and any further results characterizing the properties of Gibbs sampling will likewise have implications for iterated learning.

4.3. Sampling from the posterior with two languages

We can illustrate some of the results outlined above by returning to our example with just two languages. In this case, the probability with which a learner chooses a particular hypothesis, defined in Eq. 23, is given by the posterior distribution derived in Section 3.2. This distribution was specified for three cases, corresponding to different kinds of data a learner can see. We can also compute the probability of generating data that match each of these three cases. For either language, the probability of generating an (x, y) pair such that $x \in \mathcal{S}$ is simply $s = \sum_{x \in \mathcal{S}} P(x)$. The probability that we generate an (x, y) pair that corresponds to the second case is $(1 - s)(1 - \epsilon)$ for $h_{n-1} = 1$ and $(1 - s)\epsilon$ for $h_{n-1} = 2$. Finally, the probability that we generate an (x, y) pair that corresponds to the third case is $(1 - s)\epsilon$ for $h_{n-1} = 1$ and $(1 - s)(1 - \epsilon)$ for $h_{n-1} = 2$.

Putting the results in the previous paragraph together with those from Section 3.2 gives us the transition probabilities for the Markov chain on hypotheses, as specified in Eq. 9. Summing over the three cases gives

$$\begin{aligned} q_{12} &= \alpha s + \frac{(1-\epsilon)\alpha}{(1-\epsilon)\alpha + \epsilon(1-\alpha)}(1-s)\epsilon + \frac{\epsilon\alpha}{\epsilon\alpha + (1-\epsilon)(1-\alpha)}(1-s)(1-\epsilon) \\ &= \alpha\left[s + (1-s)(1-\epsilon)\epsilon\left(\frac{1}{\alpha+\epsilon-2\alpha\epsilon} + \frac{1}{1-\alpha-\epsilon+2\alpha\epsilon}\right)\right] \end{aligned} \tag{37}$$

and

$$\begin{aligned} q_{21} &= (1-\alpha)s + \frac{(1-\epsilon)(1-\alpha)}{(1-\epsilon)(1-\alpha) + \epsilon p}(1-s)\epsilon + \frac{\epsilon(1-\alpha)}{\epsilon(1-\alpha) + (1-\epsilon)\alpha}(1-s)(1-\epsilon) \\ &= (1-\alpha)\left[s + (1-s)(1-\epsilon)\epsilon\left(\frac{1}{\alpha+\epsilon-2\alpha\epsilon} + \frac{1}{1-\alpha-\epsilon+2\alpha\epsilon}\right)\right]. \end{aligned} \tag{38}$$

These two values specify the transition matrix $\mathbf{Q}$, since $q_{22} = 1 - q_{12}$ and $q_{11} = 1 - q_{21}$.

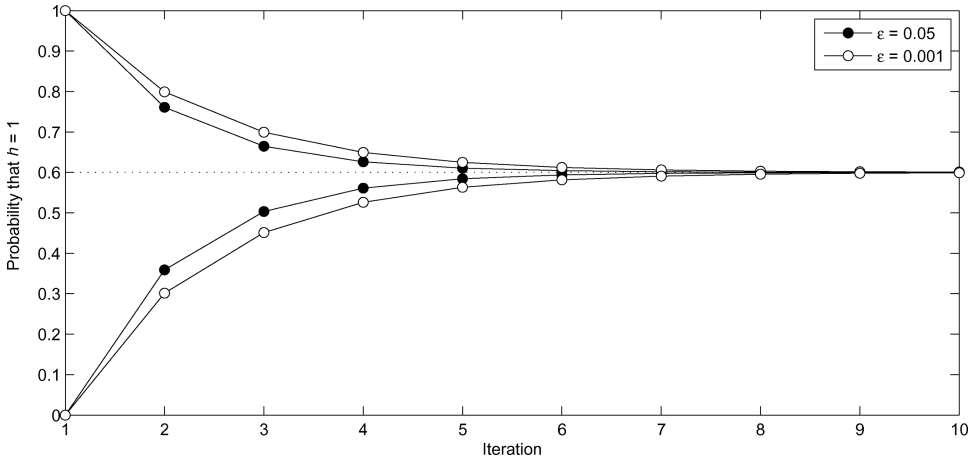

Fig. 3. Dynamics of iterated learning by sampling from the posterior. Solid lines show the probability that a learner chooses $h = 1$ as a function of the number of iterations, starting from a state where the first learner speaks either L_1 or L_2, with $\epsilon = 0.05$ and $\epsilon = 0.001$. In all cases, $\alpha = 0.6$ and $s = 0.5$. The probability of choosing $h = 1$ rapidly converges to the stationary probability, α, indicated by the dotted line.

Knowing the transition matrix allows us to characterize the dynamics and asymptotic behavior of iterated learning. We can use Eq. 6 to compute the probability that a learner acquires L_1 at each iteration given that the first learner spoke a particular language, producing the results shown in Fig. 3. We can compute the stationary distribution of the Markov chain using Eq. 14. The bracketed term in Eq. 37 and 38 is constant across the two expressions, so the stationary distribution has $\theta_1 = \alpha$ (unless $s = 0$ and $\epsilon = 0$). Thus, as indicated by the results outlined above and illustrated in Fig. 3, the probability that a learner chooses $h = 1$ converges to its prior probability, α. From Eq. 15, we find that the second eigenvalue is

$$\lambda_2 = 1 - \left[s + (1-s)(1-\epsilon)\epsilon \left(\frac{1}{\alpha + \epsilon - 2\alpha\epsilon} + \frac{1}{1 - \alpha - \epsilon + 2\alpha\epsilon} \right) \right] \tag{39}$$

which is less than 1 unless $s = 0$ and $\epsilon = 0$. Consequently, the Markov chain will be ergodic if there is any agreement between the languages or any noise in production, since this makes it possible to move between languages. When $s = 0$ and $\epsilon = 0$ there is no agreement and no noise, so it is impossible for a speaker of one language to generate data that are consistent with the other language. In this case, both languages act as sinks and the Markov chain is not ergodic. Fig. 3 shows that the rate of convergence increases as ϵ increases, since this makes it easier to move between languages, and this is reflected in the effect of ϵ on λ_2.

4.4. Summary

When learners sample from their posterior distribution over hypotheses, the consequences of iterated learning are determined entirely by the biases of the learners. The probability that a learner entertains a particular hypothesis will converge to the prior probability of that hypothesis, and the probability that a learner produces particular data will converge to the

probability distribution over data produced by choosing a hypothesis from the prior, and then generating data from that hypothesis. The asymptotic outcome of iterated learning thus does not depend on the amount or structure of the data seen by the learners, or the properties of the hypotheses those learners consider, except insofar as those factors influence the prior probabilities assumed by the learners. Before considering further conclusions we might draw from this case, we will examine the consequences of what seems like a small change to our assumptions, considering what happens when learners choose the hypothesis that has maximum posterior probability rather than sampling.

5. Choosing hypotheses with maximum posterior probability

While it may be the simplest, sampling from the posterior distribution is not the only way to define a learning algorithm based on Bayesian inference. An alternative is to assume that learners select the hypothesis with the maximum posterior probability (known as *maximum a posteriori* or MAP estimation). If we let $\mathcal{H}^*(d)$ denote the set of hypotheses h^* such that

$$h^* = \arg\max_h P(h \mid d) = \arg\max_h P_{PA}(d|h)P(h) \tag{40}$$

for some dataset d, this learning algorithm is associated with the probability distribution

$$P_{\text{MAP}}(h \mid d) = \begin{cases} 1/|\mathcal{H}^*(d)| & h \in \mathcal{H}^*(d) \\ 0 & \text{otherwise} \end{cases} \tag{41}$$

where $|\mathcal{H}^*(d)|$ is the size of the set of hypotheses with maximum posterior probability.

The strategy of selecting the hypothesis with the greatest posterior probability can be justified from the perspective of Bayesian decision theory, as it maximizes the probability of selecting the hypothesis from which the data were generated (e.g., Robert, 1994). Thus, MAP estimation is the scheme with the greatest fidelity of transmission of hypotheses across generations. It is also consistent with the approach taken in previous work on iterated learning. Some of this work has explicitly approached language acquisition as a problem of Bayesian inference (Kirby et al., 2004), while other work has considered algorithms based on minimum description length (Brighton, 2002; Smith et al., 2003). The principle of maximizing the posterior probability is equivalent to minimizing the length of the representation of a language in a particular encoding scheme (this idea is explored in detail in Chater, 1996, and Li & Vitanyi, 1997). Other learning algorithms, such as gradient descent algorithms for artificial neural networks, can also be interpreted as a form of MAP estimation (Neal, 1992; MacKay, 1995).

Unfortunately, iterated learning by MAP estimation is more difficult to analyze than iterated leaning by sampling from the posterior. However, as with sampling, it can be reduced to an inference algorithm that is used in statistics, and for which some analytic results exist. We will establish this correspondence and summarize the relevant results, and then turn to a more detailed analysis of the case of iterated learning with just two languages that we have used throughout the paper. This example helps to highlight the differences between MAP estimation

and sampling, and provides some intuitions about the consequences of iterated learning by MAP estimation.

5.1. Iterated learning by MAP estimation and the stochastic EM algorithm

The correspondence between iterated learning by sampling and the Gibbs sampler suggests that we might be able to analyze other cases of iterated learning by identifying corresponding algorithms used in statistics. This strategy provides us with a way to analyze iterated learning by MAP estimation, which corresponds to a variant on the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977), which is widely used in modern statistics and machine learning. An introduction to the EM algorithm is given by Bilmes (1997), and a more detailed treatment appears in McLachlan & Krishnan (1997).

The EM algorithm is typically used to obtain the maximum-likelihood estimate of the parameters of a model that contains *latent variables*—variables that are involved in generating data, but are not themselves observed. A classic example is a clustering problem, where we observe the locations of a set of points, but do not know the clusters from which those points were generated or the parameters that characterize each cluster. The EM algorithm makes it possible to estimate the cluster parameters even though we do not know the cluster assignments. It alternates between two steps: an expectation (E) step, in which the probability distribution over cluster assignments is computed, and a maximization (M) step in which the parameters of the clusters are updated based on the probabilities with which the different points are assigned to those clusters.

More formally, assume we have observed data $\boldsymbol{x}$, latent variables z, and we use h to represent a hypothesis about the parameters of the model.[4] In the clustering problem mentioned in the previous paragraph, $\boldsymbol{x}$ is the location of the points, z the cluster assignments, and h the cluster parameters. Our goal is to obtain a maximum-likelihood estimate of h, choosing the hypothesis that maximizes $P(\boldsymbol{x}|h)$ (or, equivalently, the log-likelihood $\log P(\mathbf{x}|h)$). However, this is made complicated by the involvement of the latent variables, since typically it is far easier to find the parameters that maximize $P(\boldsymbol{x}, z|h)$ than $P(\boldsymbol{x}|h)$. If we knew the values of the latent variables, we could find h easily, and if we knew h, we could work out the distribution on the latent variables. The EM algorithm exploits this by alternating between adjusting the distribution on the latent variables and the values of the parameters. On iteration n of the algorithm, the E step involves computing the posterior distribution over z given $\boldsymbol{x}$ and the previous choice of h, $P(z|\boldsymbol{x}, h_{n-1})$, and then taking the expectation of $\log P(\boldsymbol{x}, z|h_n)$ with respect to this distribution for each hypothesis h_n, to give

$$E_{P(z|\boldsymbol{x},h_{n-1})}[\log P(\boldsymbol{x}, z|h_n)] = \sum_z P(z|\boldsymbol{x}, h_{n-1}) \log P(\boldsymbol{x}, z|h_n). \tag{42}$$

In the M step, we choose the value of h_n that maximizes this expectation. This procedure is guaranteed to produce a series of estimates of h_n for which $P(\boldsymbol{x}|h_n)$ is non-decreasing (Dempster, Laird, & Rubin, 1977; Neal & Hinton, 1998). If the set of hypotheses under consideration is continuous, the EM algorithm converges to a local maximum (or saddle-point)

of $P(\boldsymbol{x}|h)$. With discrete hypothesis spaces, the maximum-likelihood solution is one fixed point, but the algorithm can also converge on other, suboptimal, hypotheses (Friedman, 1998).

Performing the expectation in Eq. 42 can be difficult, leading to the development of approximate EM algorithms. In Monte Carlo EM (Wei & Tanner, 1990), the expectation is approximated using samples from $P(\mathbf{z}|\boldsymbol{x}, h_{n-1})$, with

$$E_{P(z|\boldsymbol{x},h_{n-1})}[\log P(\boldsymbol{x}, z|h_n)] \approx \frac{1}{m_n}\sum_{\ell=1}^{m_n} \log P\left(\boldsymbol{x}, \mathbf{z}^{(\ell)}|h_n\right) \tag{43}$$

where m_n is the number of samples on iteration n, and $\mathbf{z}^{(\ell)}$ is the ℓth sample of the latent variables $\boldsymbol{z}$. The number of samples, m_n, typically increases with n, producing similar convergence guarantees to those of the standard EM algorithm (Sherman, Ho, & Dalal, 1999; Fort & Moulines, 2003). However, some authors have also advocated algorithms in which m_n remains constant. The case where $m_n = 1$ for all n is called *stochastic EM* (Celeux & Diebolt, 1985; Diebolt & Ip, 1996).

The additional variability introduced by sampling in stochastic EM means that it is less likely to get stuck in sub-optimal solutions, although the output of the algorithm is a distribution over hypotheses rather than a single hypothesis. The sequence of hypotheses produced by stochastic EM form a homogeneous Markov chain, and conditions for the ergodicity of this chain have been established (Ip, 1994, 2002; Diebolt & Ip, 1996; Nielsen, 2000). Empirical and theoretical results indicate that the stationary distribution over hypotheses produced by this Markov chain is approximately centered on the maximum-likelihood solution, with a variance that increases as a function of the rate at which the hypotheses change across iterations (Celeux & Diebolt, 1985, 1988; Celeux, Chauveau, & Diebolt, 1995; Diebolt & Ip, 1996; Ip, 1994; Nielsen, 2000). A more precise characterization of the consequences of stochastic EM can be given in special cases, such as estimating parameters for the kind of clustering problem introduced above (Diebolt & Celeux, 1993; Nielsen, 2000), but there are no explicit results characterizing the asymptotic behavior of stochastic EM when the set of hypotheses is discrete.

The EM algorithm and its variants can also be used to perform MAP estimation simply by replacing $\log P(\boldsymbol{x}, z|h_n)$ with $\log P(\boldsymbol{x}, z|h_n)P(h_n)$ in Eq. 42 or 43. The resulting algorithm converges to a local maximum of the joint probability $P(\boldsymbol{x}|h)P(h)$ rather than the likelihood $P(\mathbf{x}|h)$. Since the posterior distribution $P(h|\boldsymbol{x})$ is directly proportional to $P(\boldsymbol{x}|h)P(h)$, this is a MAP solution. The stochastic EM algorithm for MAP estimation with $m_n = 1$ would thus take h_n to be the value of h that maximizes $P(\boldsymbol{x}, z|h)P(h)$ when $\boldsymbol{z}$ is drawn from the distribution $P(z|\boldsymbol{x}, h_{n-1})$, and converge to a stationary distribution approximately centered at the maximum of $P(h|\boldsymbol{x})$.

With this background in place, the correspondence between iterated learning by MAP estimation and stochastic EM can be stated: iterated learning corresponds to stochastic EM in a model where the data passed from one generation to the next, d, plays the role of the latent variable, $\boldsymbol{z}$, and there are no observations, $\boldsymbol{x}$. In this case, the stochastic EM algorithm reduces to taking h_n to be the hypothesis that maximizes $P_{PA}(d|h_n)P(h_n)$ when d is drawn from the distribution $P_{PA}(d|h_{n-1})$. This is exactly the procedure followed in iterated learning using MAP estimation. Consequently, results characterizing the behavior of stochastic EM apply to this form of iterated learning. Since this correspondence applies to the case where there are

no observations, $\boldsymbol{x}$, the MAP solution is simply the maximum of the prior (with no $\boldsymbol{x}$, the joint distribution of $\mathbf{x}$ and h is just $P(h)$). Thus, the stationary distribution over hypotheses produced by iterated learning by MAP estimation should be approximately centered on the maximum of the prior, with a variance that increases as a function of the rate at which hypotheses change across generations.

The relationship between iterated learning by MAP estimation and stochastic EM provides a rough characterization of the consequences of iterated learning: the probability that a learner acquires a particular language will converge to a distribution that emphasizes languages with high prior probability. Unlike sampling, this distribution will be affected by the properties of the languages involved and the amount of information transmitted between learners, with these factors determining the amount of variation around the high probability languages exhibited by the stationary distribution. As with our previous observation about the relationship between iterated learning by sampling and the Gibbs sampler, this correspondence establishes a route by which results in statistics can be used to gain a deeper understanding of the processes of language evolution.

5.2. MAP estimation with two languages

Identifying iterated learning by MAP estimation with the stochastic EM algorithm provides an abstract characterization of its behavior. To obtain a deeper understanding of this process, we will return to the example of iterated learning with two languages introduced in Section 2.3 and embellished on in Sections 3.2 and 4.3. In this simple case, we can determine the consequences of iterated learning by MAP estimation directly, and give analytic results for the stationary distribution and the rate of convergence that provide some valuable intuitions.

We established the fundamentals of Bayesian iterated learning with just two languages in Section 3.2. We need only change the learning algorithm that is used to translate the posterior distribution into the choice of a hypothesis, replacing sampling with maximizing. The same three cases are relevant. When $x \in \mathcal{S}$ the posterior probability that $h = 1$ is simply the prior probability, α. By assumption, $\alpha > 0.5$, so the MAP hypothesis is $h = 1$. Thus, we should select $h = 1$ with probability 1. In the second case, $h = 1$ is likewise dominant, as $\alpha > 0.5$ and $\epsilon < 0.5$. However, in the third case, which hypothesis has highest posterior probability depends on the relative values of α and ϵ. We can write the transition probabilities as

$$q_{12} = s + (1-s)\epsilon + R\left(\frac{\epsilon\alpha}{\epsilon\alpha + (1-\epsilon)(1-\alpha)}\right)(1-s)(1-\epsilon)$$

and

$$q_{21} = R\left(\frac{\epsilon(1-\alpha)}{\epsilon(1-\alpha) + (1-\epsilon)\alpha}\right)(1-s)(1-\epsilon)$$

where $R(\cdot)$ is a rounding function, taking the value 1 when its argument is greater than 0.5, 0 when its argument is less than 0.5, and 0.5 when its argument is exactly 0.5.

Table 1 gives the values of q_{12}, q_{21}, θ_1, and λ_2 under different conditions on the relationship between α and ϵ. L_1 is almost always favored over L_2, with the stationary probability that $h = 1$, θ_1, being greater than 0.5 for any $s > 0$ because L_1 is always chosen when $x \in \mathcal{S}$.

Table 1
Properties of the Markov chain on hypotheses for iterated learning with MAP estimation

Condition	q_{12}	q_{21}	θ_1	λ_2
$\epsilon < 1-\alpha$	$s+(1-s)\epsilon$	$(1-s)\epsilon$	$\frac{s+(1-s)\epsilon}{s+2(1-s)\epsilon}$	$(1-s)(1-2\epsilon)$
$\epsilon = 1-\alpha$	$s+(1-s)(1+\epsilon)/2$	$(1-s)\epsilon/2$	$\frac{s+(1-s)(1+\epsilon)/2}{s+(1-s)(1+2\epsilon)/2}$	$(1-s)(1-2\epsilon)/2$
$\epsilon > 1-\alpha$	1	0	1	0

When $\epsilon > 1-\alpha$, there is so much uncertainty that L_1 is also able to dominate, being selected on every iteration after the first (hence $\lambda_2 = 0$, indicating the fastest rate of convergence possible). More interesting results are obtained when $\epsilon < 1-\alpha$. In this case, the preference for L_1 gradually decreases as ϵ increases, as higher values of ϵ make it more likely that we will observe data that are consistent with L_2 being generated by speakers of L_1. In fact, θ_1 is completely independent of α so long as $\epsilon < 1-\alpha$, implying that a broad range of inductive biases will result in exactly the same stationary distribution. The relationship between s, ϵ, and θ_1 is shown in Fig. 4, under the assumption that $\epsilon < 1-\alpha$. The second eigenvalue, λ_2, also decreases as ϵ increases, indicating that convergence to the stationary distribution becomes faster in the presence of more noise.

The differences between the results for MAP estimation outlined in this section and the results for sampling presented in Section 4.3 are instructive. When learners sample from their

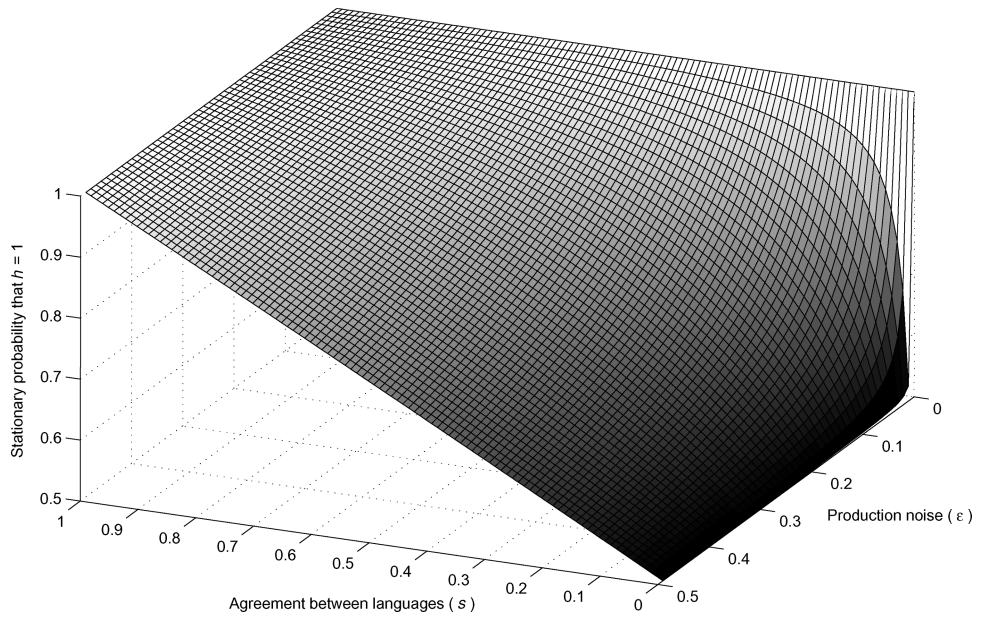

Fig. 4. Stationary probability that $h = 1$ as a function of agreement betwen languages, s, and level of production noise, ϵ, in iterated learning with MAP learners. The stationary probability is unaffected by the prior probability that $h = 1$, α, provided $\epsilon < 1-\alpha$.

posterior distributions, the stationary probability that $h = 1$ is simply the prior probability, α, for almost any amount of overlap between languages, s, and production noise, ϵ. The consequences of iterated learning are thus determined entirely by the prior. When learners use MAP estimation, the stationary probability that $h = 1$ is stable over a large range of values of α (for $0.5 < \alpha < 1 - \epsilon$), but depends directly on s and ϵ. The consequences of iterated learning are thus the same for many priors, being determined by the overlap between languages and the amount of production noise.

5.3. Summary

While iterated learning by MAP estimation is harder to analyze than sampling from the prior, we can obtain some insight into its consequences by observing a correspondence with the stochastic EM algorithm and by analyzing specific cases. The results of these analyses indicate that iterated learning by MAP estimation still favors languages with higher prior probability, but the stationary distribution depends on the nature of the hypotheses and the amount of noise. The correspondence with stochastic EM provides a rough characterization of the stationary distribution of iterated learning by MAP estimation, indicating that this distribution should be approximately centered on the hypotheses with greatest prior probability, but will exhibit variance around these hypotheses. Analyzing the case of just two languages provides an illustration of these general trends: the stationary distribution favors the language with greatest prior probability, L_1, but the extent to which this is the case depends on the amount of noise in the data, as represented by ϵ. When ϵ is higher, it is easier to move between hypotheses, and it is easier to generate data that result in the acquisition of L_2.

6. An example: The emergence of compositionality

The results presented in the last two sections provide a characterization of the consequences of iterated learning when Bayesian learners either sample from their posterior distribution or select the hypothesis with the greatest posterior probability. In order to explore the implications of these results for language evolution, we chose to examine their predictions in a setting that is closer to that used in previous work on iterated learning. To this end, we simulated iterated learning in a simplified version of a scenario that has been used in several papers exploring the emergence of compositionality (Kirby, 2001; Brighton, 2002; Smith et al., 2003).

As in our two-language example, we model language acquisition as learning a mapping between meanings and utterances. The data d consist of a set of m inputs, $\boldsymbol{x} = \{x_1, \ldots, x_m\}$, and m corresponding outputs, $\boldsymbol{y} = \{y_1, \ldots, y_m\}$, and hypotheses h are probability distributions over y for each x. The nth learner sees data, $(\boldsymbol{x}_{n-1}, \boldsymbol{y}_{n-1})$, and then generates outputs $\mathbf{y}_n$ in response to new inputs $\boldsymbol{x}_n$. Meanings and utterances each vary along two binary dimensions. This yields a total of four meanings and four utterances, each corresponding to the set $\{00, 01, 10, 11\}$.

In a *compositional* language, the mapping between meanings and utterances depends upon their parts: the two dimensions of meanings are mapped onto the two dimensions of utterances (for simplicity, we assumed that the order is preserved), and the only uncertainty is in which

values map to one another. There are $2^2 = 4$ such languages. In a *holistic* language, the mapping between meanings and utterances is arbitrary, and a single word is chosen to represent each meaning without any constraints. There are $4^4 = 256$ such languages. $\mathcal{H}$ thus contains 260 hypotheses, each a mapping between meanings and utterances. For each $h \in \mathcal{H}$, we define the probability distribution over outputs y given the input x to be

$$P(y \mid x, h) = \begin{cases} 1 - \epsilon & x \text{ maps to } y \text{ in } h \\ \frac{\epsilon}{3} & \text{otherwise} \end{cases} \tag{44}$$

where ϵ is the error rate of production, as in our example with just two languages. The prior probability of each hypothesis is

$$P(h) = \begin{cases} \frac{\alpha}{4} & h \text{ refers to a compositional language} \\ \frac{1-\alpha}{256} & h \text{ refers to a holistic language} \end{cases}. \tag{45}$$

This is a *hierarchical prior*, allocating a probability of α to the set of compositional languages and $1 - \alpha$ to the set of holistic languages, and then spreading this probability uniformly over the hypotheses within those sets.

Since every language is simply a mapping from meanings to utterances, $\mathcal{H}$ includes four holistic languages that each give the same mapping as one of the four compositional languages. These languages make the same predictions about inputs and outputs, as determined by Eq. 44, and thus cannot be discriminated by any data. Any advantage of the compositional languages over their holistic counterparts results from the prior defined in Eq. 45. If compositional and holistic languages are equally probable a priori ($\alpha = 0.5$), then the relatively small number of compositional languages means that any particular compositional language is more probable than any particular holistic language. Consequently, it would be very unlikely to see a holistic language that just happened to produce a compositional mapping. As α becomes smaller, it becomes less likely that one would see a compositional language at all, and a holistic language that just happened to produce a compositional mapping becomes more plausible.

The transition matrix for the Markov chain on hypotheses, $\mathbf{Q}$, for each algorithm can be obtained by summing over all $(\boldsymbol{x}, \boldsymbol{y})$ pairs. Since there are $(2^2 2^2)^m$ such pairs, this is intractable for large m. Consequently, matrices for $m > 4$ were computed approximately using Monte Carlo, with 1000 samples for each hypothesis. We computed transition matrices for $\alpha \in \{0.01, 0.5\}$, $\epsilon \in \{0.01, 0.05\}$, and $m \in \{1, 2, \ldots, 10\}$ for both sampling and MAP. We will discuss the results for the two algorithms in turn.

6.1. Results for sampling from the posterior

The first column of Fig. 5 shows a portion of some of the transition matrices produced by sampling from the posterior. The second column shows 1000 iterations for four Markov chains (initialized by choosing h_1 uniformly at random), while the third and fourth columns (labeled "Chain" and "Prior") show the distribution over hypotheses from a single sample of 10000 iterations from those chains and the prior $P(h)$, respectively. Our theoretical results predict that the asymptotic probability that a learner infers a particular language depends only on its prior probability, and not on other properties of the language. This was confirmed by

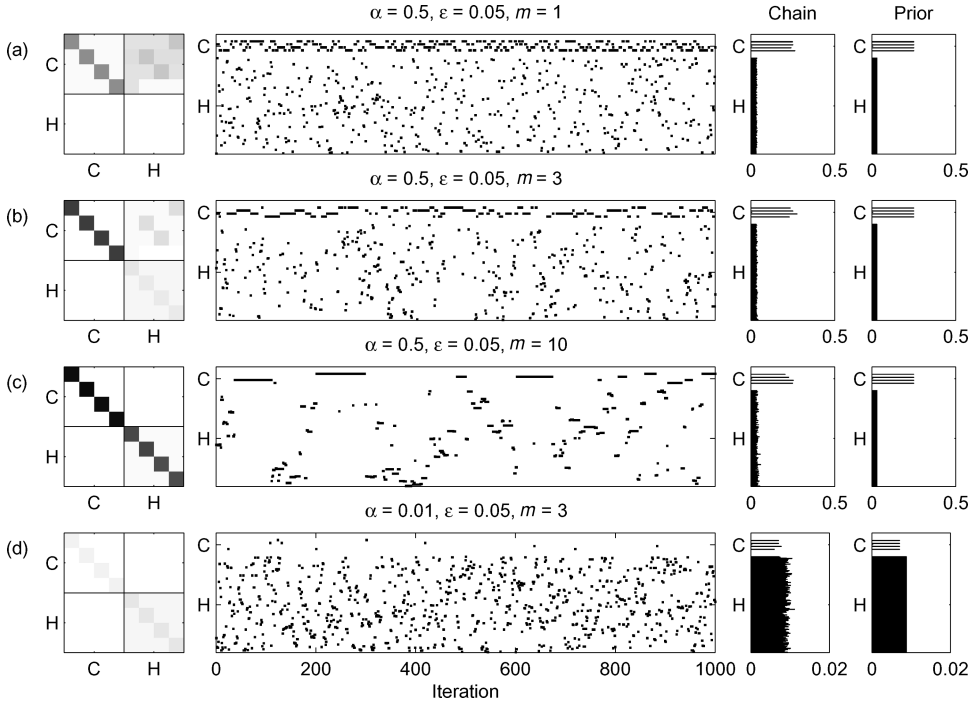

Fig. 5. Markov chains on hypotheses for the evolution of compositionality when learners sample from their posterior distribution. Different rows correspond to different parameter values. For each set of parameters, the first column shows a portion of the transition matrix, **Q**, with four compositional languages (labeled C) and four holistic languages (labeled H). Columns are h_{n-1}, rows are h_n, and darker grey indicates a higher value of $q_{ij} = P(h_n = i|h_{n-1} = j)$. The second column shows a sample of 1000 iterations from this matrix, the third shows the relative frequency of hypotheses across 10000 iterations, and the fourth shows the prior, $P(h)$. The quantities in the third and fourth columns are subjected to a square root transformation in order to make the full range of variation apparent.

our simulations, as can be seen by comparing Fig. 5(b) and (d). The Markov chain shown in Fig. 5(b) used $\alpha = 0.5$, and compositional languages appear with high frequency. The Markov chain shown in Fig. 5 (d) used $\alpha = 0.01$, and compositional languages appear only infrequently. As shown in the third and fourth columns of the figure, the relative frequencies of the different languages correspond closely to their prior probabilities. Thus, compositional languages are favored by iterated learning only if they have high prior probability.

The results shown in Fig. 5(a)–(c) illustrate that the asymptotic probability that a language is spoken is not affected by the amount of data seen by the learners. While the Markov chain develops a greater tendency to remain in the same state as m increases, indicated by the strong diagonal in the transition matrices and the length of the streaks in the samples, the relative frequencies of the different languages remain the same. These relative frequencies match the corresponding prior probabilities, consistent with our mathematical analysis. Thus, the emergence of languages with particular properties does not require a bottleneck on the amount of information passed from one generation to the next.

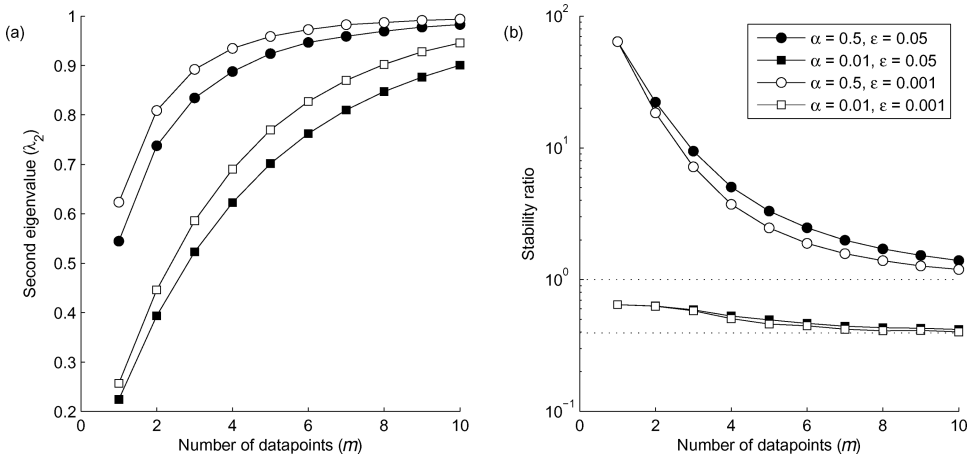

Fig. 6. Quantities derived from Markov chains on hypotheses for learners sampling from the posterior, as a function of number of datapoints, m, prior on composite languages, α, and error rate, ϵ. (a) Second eigenvalue of transition matrix, λ_2. (b) Stability ratio. The dotted line shows the stability ratio as $m \to \infty$.

While the amount of data seen by the learners does not influence the asymptotic consequences of iterated learning by sampling from the posterior, it does affect other properties of the underlying Markov chain. Figure 6(a) shows how λ_2 is affected by α, ϵ, and m. As α brings $P(h)$ away from uniformity, it increases the probability that successive learners will share the same hypothesis. This increase in the fidelity of transmission means that it will take longer to move from an initial hypothesis to a hypothesis with higher probability under the stationary distribution. As a consequence, convergence to the stationary distribution will be slower, and λ_2 increases as α increases.

Changing ϵ and m also decreases the rate of convergence (and increases λ_2), as the data received by each learner become more informative. Decreasing ϵ increases the probability that a learner produces a set of utterances consistent with their current hypothesis, and thus the probability that the next learner will infer the same hypothesis. As a consequence, movement between hypotheses is slower, convergence takes longer, and λ_2 increases. Increasing m increases the amount of information available to learners, reducing the chance that a misleading set of utterances will be generated. Consequently, the probability that successive learners choose the same hypothesis increases, slowing movement between hypotheses, decreasing the rate of convergence, and increasing λ_2. With large m, it is likely that a single hypothesis will be maintained across several generations, as can be seen in Fig. 5(c).

The parameters α, m, and ϵ also influence the relative stability of compositional and holistic languages, assessed via the ratio of the mean probability that a particular compositional language would appear as both h_{n-1} and h_n (i.e., the mean of q_{ii} where i is a compositional language) to the corresponding mean probability for holistic languages. The effects of m, α, and ϵ on this ratio are shown in Fig. 6(b). The stability ratio is strongly affected by α: if the prior probability of a compositional language is high, it is more likely that a learner will acquire that language, and consequently that language is more stable. The magnitude of this

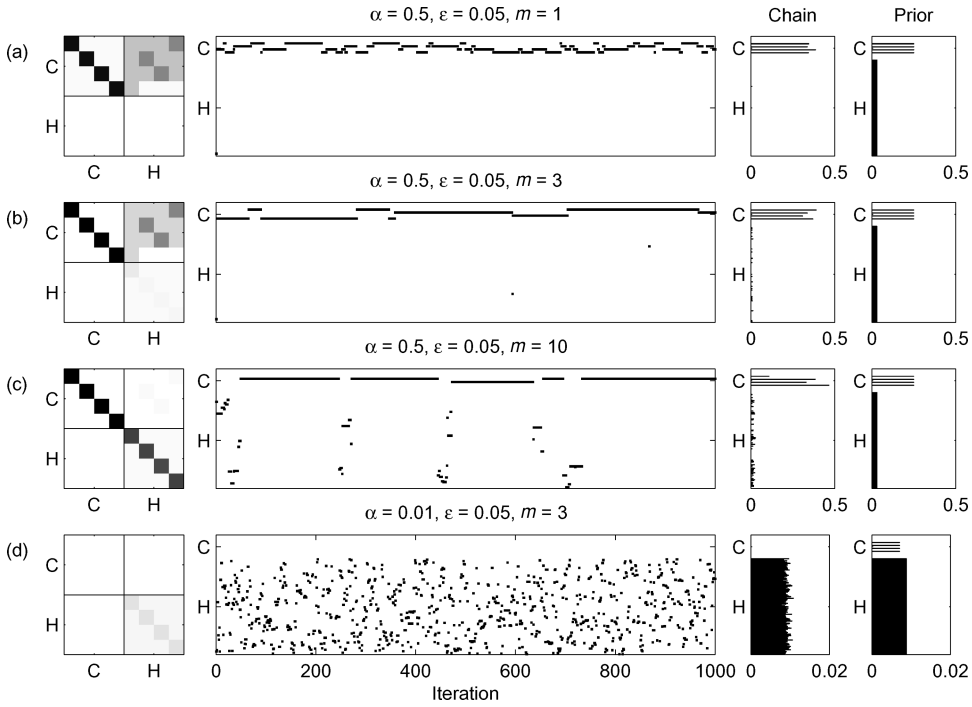

Fig. 7. Markov chains on hypotheses for the evolution of compositionality with MAP estimation. Different rows correspond to different parameter values. For each set of parameters, the first column shows a portion of the transition matrix, **Q**, with four compositional languages (labeled C) and four holistic languages (labeled H). Columns are h_{n-1}, rows are h_n, and darker grey indicates a higher value of $q_{ij} = P(h_n = i \mid h_{n-1} = j)$. The second column shows a sample of 1000 iterations from this matrix, the third shows the relative frequency of hypotheses across 10000 iterations, and the fourth shows the prior, $P(h)$. The quantities in the third and fourth columns are subjected to a square root transformation in order to make the full range of variation apparent.

effect is modulated by the number of datapoints, m, with α having the greatest effect when m is small. As m increases, the data begin to overcome the influence of the prior.[5]

6.2. Results for MAP estimation

Fig. 7 shows the same quantities as Fig. 5 for learners choosing the MAP hypothesis. The results are quite different from the corresponding cases for learners sampling from the posterior: the influence of the prior on the behavior of the Markov chains is significantly increased. Since every compositional language is also contained in the set of holistic languages, learners never choose a compositional language when holistic languages have greater prior probability (i.e., when $\alpha = 0.01$). The effect is less marked when compositional languages have higher prior probability (i.e., with $\alpha = 0.5$), since there are many holistic languages which are not dominated by compositional languages, but the holistic languages are still far less common than in the chains produced by sampling with the same values of m and ϵ.

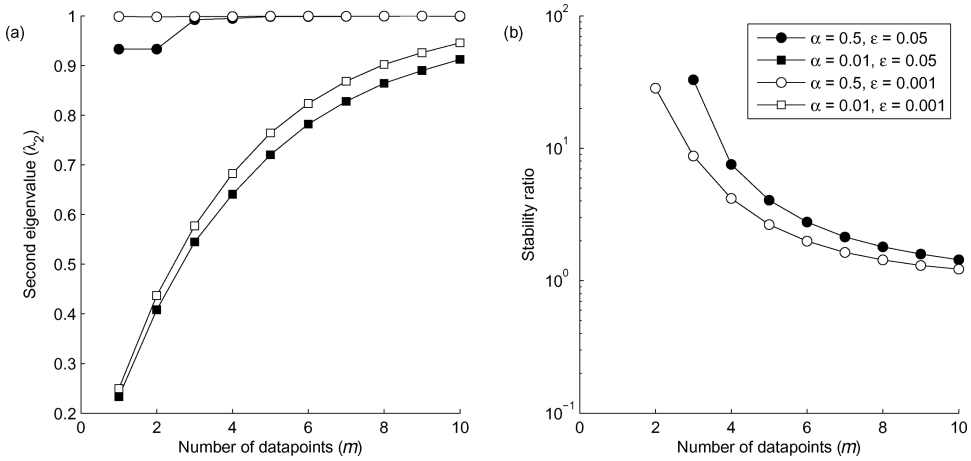

Fig. 8. Quantities derived from Markov chains on hypotheses for learners using MAP estimation, as a function of number of datapoints, m, prior on composite languages, α, and error rate, ϵ. (a) Second eigenvalue of transition matrix, λ_2. (b) Stability ratio. The stability ratios for $\alpha = 0.01$ are constant at zero, since compositional hypotheses are never chosen, and infinite for $m < 3$ and $m < 2$ when $\epsilon = 0.05$ and 0.01 respectively for $\alpha = 0.5$, since in these cases holistic languages are never chosen.

A second qualitative difference from the results for sampling is that there are fewer transitions between languages in the chains favoring compositional languages (i.e., with $\alpha = 0.5$). This occurs because if the compositional languages are the only hypotheses under consideration, moving from one language to another requires generating data that are consistent with that language. Because the compositional languages do not overlap, this requires errors in production. The probability of such errors is set by ϵ, and the chance of a compositional languge producing data that are more consistent with another compositional language than itself decreases as m increases. Because of this, the second eigenvalue of the transition matrix, λ_2, for the chains with $\alpha = 0.5$ is consistently close to 1, as shown in Fig. 8(a). The values of λ_2 when $\alpha = 0.01$ are more comparable to those produced by sampling from the posterior, since it is easier to move between holistic languages.

Finally, unlike sampling from the posterior, m affects the asymptotic consequences of iterated learning by MAP estimation. Looking just at the cases with $\alpha = 0.5$, when $m = 1$ only compositional languages appear in the chain. This is because the evidence provided by a single utterance is insufficient to overwhelm the higher prior probability of compositional languages. The probability of holistic languages under the stationary distribution is thus zero. While it is not shown here, the same phenomenon occurs with $m = 2$. When $m = 3$, it finally becomes possible to generate data that result in selection of holistic languages, and these languages have non-zero probabilities under the stationary distribution. This trend is more pronounced when $m = 10$, with holistic languages appearing more often, albeit for short intervals before returning to a compositional language. As with sampling from the posterior, increasing m also reduces the ratio of the stability of compositional to holistic languages, following the trend shown in Fig. 8(b).

6.3. Summary

The simulations summarized in this section allowed us to analyze the consequences of iterated learning by sampling from the posterior and by MAP estimation in a setting more comparable to previous research. The results of these simulations bear out the predictions produced by our theoretical analyses: both forms of Bayesian learning result in a distribution over languages that reflects the inductive biases of learners, with the influence of the prior being emphasized by MAP estimation. This setting also makes it possible for us to examine the effect of the number of utterances, m, seen by each learner. As expected, this has no effect on the distribution over languages ultimately produced by sampling from the posterior, although it does affect the rate of convergence to this distribution. In contrast, when learners use MAP estimation the dominance of languages with high prior probability is attenuated with larger values of m, as it becomes possible to generate sets of utterances that provide strong evidence for a language with low prior probability.

7. Population dynamics and iterated learning

The results we have discussed so far characterize the consequences of iterated learning in a setting where each generation consists of a single learner. However, several prominent analyses of language evolution have focused a different setting, examining how the proportion of an unbounded population that speaks a particular language changes in continuous time (Komarova et al., 2001; Nowak et al., 2001, 2002). Our results can be extended into this setting, indicating how iterated learning will affect the asymptotic proportion of a population that learns a particular language.

Let p_i denote the proportion of a population of learners entertaining hypothesis i at a given moment t, and q_{ij} denote the probability that a learner chooses hypothesis i after seeing data generated from hypothesis j, as defined in Eq. 9. If we assume that each learner learns from a random member of the population at the previous instant, then the population proportions evolve as

$$\frac{dp_i}{dt} = \sum_j f_j q_{ij} p_j - \phi\, p_i, \tag{46}$$

where f_j is the *fitness* of speakers of language j in a population with proportions $\boldsymbol{p} = (p_i)$, $\phi = \sum_k f_k p_k$, is the mean fitness, and the second term on the right hand side ensures that $\sum_i p_i = 1$. This is the "language dynamical equation" explored by Nowak et al. (2001, 2002). In such models, fitness is typically assumed to be a function of how well speakers of a particular language can communicate with the population at large, implementing a selection pressure for communication. If we assume that all speakers have equal fitness, $f_j = 1$, Eq. 46 simplifies to

$$\frac{dp_i}{dt} = \sum_j q_{ij} p_j - p_i, \tag{47}$$

which is a linear dynamical system. A special case of this model was analyzed by Komarova and Nowak (2003).

The asymptotic behavior of this linear dynamical system is straightforward to analyze. By setting $\frac{dp_i}{dt}$ equal to zero, we find that the population proportions will be in equilibrium for $\mathbf{p}$ such that $p_i = \sum_j q_{ij} p_j$, which is the same as the condition used to define the stationary distribution $\boldsymbol{\theta}$ in Eq. 7. Consequently, this system has an equilibrium at $\mathbf{p} = \boldsymbol{\theta}$. The properties of this equilibrium depend on the eigenvalues of the matrix $\mathbf{Q} - \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. As mentioned above, the largest eigenvalue of $\mathbf{Q}$ will be 1, since $\mathbf{Q}$ is a stochastic matrix. Under conditions analogous to those for the ergodicity of the Markov chain on hypotheses, this largest eigenvalue will be unique. In this case, $\mathbf{Q} - \mathbf{I}$ will have a single eigenvalue of 0, and the real components of the remaining eigenvalues will be negative. It follows that the equilibrium $\boldsymbol{\theta}$ defined above is a sink, a unique asymptotically stable equilibrium to which the population proportions will converge (Hirsch & Smale, 1974).[6]

Iterated learning with infinite populations evolving in continuous time thus displays similar asymptotic behavior to iterated learning with discrete generations of single learners. The key difference is in the nature of the quantities that converge: with discrete generations of single learners, the probability that a particular learner entertains hypothesis i converges to θ_i; with an infinite population evolving in continuous time, it is the proportion of the population that entertains hypothesis i that converges to θ_i. Consequently, the results from the previous sections characterize the consequences of iterated learning not just for individuals, but for populations. This provides an additional justification for the use of the iterated learning model in studying language evolution: if the stationary probability corresponds to the proportion of the population that learn a particular language, estimates of the stationary distribution produced by running simulations with discrete generations consisting of a single learner can be generalized to the level of populations.

8. Discussion

Studying iterated learning with Bayesian agents provides the opportunity to determine the influence of the inductive biases of learners on language evolution. Our results indicate that these inductive biases have a strong effect on the consequences of iterated learning. When learners sample from their posterior distributions, the probability that a learner acquires a particular language converges to the prior probability assigned to that language as iterated learning proceeds. When learners choose the hypothesis with greatest posterior probability, iterated learning produces in convergence to a distribution in which hypotheses with high prior probability dominate, although the exact distribution depends on the amount of information transmitted between learners. Furthermore, these results apply not just to the probability that individual learners acquire a language, but to the proportion of the members of a population who will acquire that language.

Our analyses provide simple conditions for determining when a particular property of languages will emerge from iterated learning with Bayesian agents: that property will emerge if it is favored by the prior. In the remainder of the paper, we explore some questions raised by these results. First, we summarize the assumptions behind our analysis, and consider their appropriateness. We then highlight some further results suggested by our analyses, and consider how our results relate to previous work on iterated learning. Finally, we return to

some of the larger issues mentioned at the start of the paper, discussing their implications for explaining linguistic universals and language change.

8.1. Summary of assumptions behind analyses

Since the assumptions behind our analyses were introduced incrementally, it is worth taking a moment to summarize those assumptions, and to consider their appropriateness in modeling language evolution. Our formulation of iterated learning as a homogeneous Markov chain relied upon two assumptions:

Assumption 1 All learners use the same learning algorithm, LA, and production algorithm, PA.

Assumption 2 We have a sequence of discrete generations of learners, with one learner per generation who receives data produced by the previous learner.

These assumptions are standard in applications of the iterated learning model (e.g., Kirby, 2001; Brighton, 2002; Smith et al., 2003). In Section 7, we showed that Assumption 2 can be replaced with

Assumption 2 We have an infinite (or large) population of learners evolving in continuous time, where each learner receives data from a randomly selected member of the population at the previous instant.

which is consistent with models of population dynamics that have been applied to language evolution (Komarova et al., 2001; Nowak et al., 2001, 2002).

When the learning algorithm is specified by Bayesian inference, Assumption 1 requires that all learners share the same set of hypotheses $\mathcal{H}$ and the same prior distribution over those hypotheses, $P(h)$. We also assumed

Assumption 3 All learners have knowledge of the probability distribution over utterances d produced by use of the production algorithm PA for the language h.

This assumption is used in specifying how Bayes' rule is applied in language acquisition in Eq. 18. Essentially, it requires consistency between learning and production: that learners produce utterances from the same probability distribution that they use in evaluating languages. Again, assumptions of this kind are common in formulations of iterated learning in terms of minimum descripton length (Brighton, 2002; Smith et al., 2003) or Bayesian inference (Kirby et al., 2004).

Finally, our characterization of the dynamics and asymptotic behavior of iterated learning required specification of the learning algorithm (as either sampling, specified by Eq. 23, or MAP estimation, specified by Eq. 41), and the ergodicity of the underlying Markov chain. As discussed in Section 2.1, ergodicity is straightforward to check for a finite Markov chain, and most of the examples we have discussed are ergodic. The possibility remains that non-ergodic Markov chains play an important role in language evolution, but our results indicate that many of the properties emphasized in previous work on iterated learning are consistent with the asymptotic behavior of ergodic Markov chains.

8.2. Language evolution and algorithms for statistical inference

A key feature of our analysis of iterated learning by both sampling and MAP estimation was its correspondence to an algorithm used for statistical inference. In Section 4.2, we showed that iterated learning by sampling from the posterior is a form of Gibbs sampling, a procedure that is used for estimating the form of complex probability distributions. In Section 5.2, we showed that iterated learning by MAP estimation corresponds to a variant of the EM algorithm, a procedure that is used for estimating the parameters of a model that contains latent variables. The fact that this correspondence exists in both cases suggests that there may be a more general relationship between iterative estimation procedures and cultural evolution.

In some ways, the existence of a relationship between estimation algorithms and cultural evolution should come as no surprise, since such relationships are well known in the context of biological evolution. For example, the standard model of population dynamics by selection in a single locus model with no mutation and constant fitness for different alleles corresponds to gradient ascent on the mean fitness of the population (e.g., Rice, 2004). Biological evolution may provide a good cautionary tale in terms of reading too much into such relationships, since the conditions identified in the previous sentence are relatively restrictive and not particularly biologically plausible: with multiple loci, mutation, and a fitness function that depends on the composition of the population the simple story of optimization that is often associated with biological evolution is no longer true. Similarly, allowing for the effects of selection (removing the assumption of uniform fitness) could easily disrupt the connections between cultural evolution and statistical inference that we have identified in this paper.

Despite this need for caution, it seems that there is a great deal of further potential for the relationship between iterated learning and algorithms for statistical inference to yield insight into processes of language evolution. Statisticians have explored a great many variants on the EM algorithm, some of which have natural interpretations in the context of iterated learning. In particular, Monte Carlo EM algorithms where $m_n > 1$ (as opposed to stochastic EM, where $m_n = 1$) are directly applicable to cases of iterated learning, and have been studied extensively (Sherman, et al., 1999; Fort & Moulines, 203). Statisticians have also investigated a version of the stochastic EM algorithm in which the samples of latent variables from previous iterations are also incorporated, providing a natural way of modeling language evolution when learners are exposed to linguistic data produced by more than one previous generation (Celeux & Diebolt, 1992; Celeux et al., 1995; Delyon, Lavielle, & Moulines, 1999).

8.3. Relationship to previous work on iterated learning

Previous work on iterated learning has emphasized the emergence of structured languages from general-purpose learning algorithms, and argued that this is partly due to the "information bottleneck" imposed by the finite nature of the data that is used to transmit language between generations (Kirby, 1999, 2001; Brighton, 2002; Smith et al., 2003; Kirby et al., 2004). Our results indicate that these effects depend to a surprising extent on the choice of learning algorithm. If learners sample from the posterior distribution, their inductive biases need to strongly favor structured languages in order for these languages to emerge, and the amount of data communicated between generations has no effect on the asymptotic probability

that learners will speak a particular langauge. However, if learners use a learning algorithm equivalent to selecting the hypothesis with greatest posterior probability, their biases can be emphasized, and the asymptotic distribution over languages is affected by the amount of information transmitted between generations.

Our analysis of the consequences of iterated learning with learners who sample from their posterior distributions initially seems to provide a counter-example to the idea that the information bottleneck is involved in the emergence of structured languages. Specifically, it shows that sensible learning algorithms exist that can produce structured languages, without an information bottleneck effect. However, in this case, structured languages will emerge only if they are strongly favored by the inductive biases of the learning algorithms being used. Thus, while these results show that iterated learning need not always exhibit an information bottleneck effect, they do not undermine the argument that such effects might play a central role in the emergence of linguistic structure when learners use general-purpose learning algorithms that assert only a weak preference for structured languages.

The case where learners select hypotheses with greatest posterior probability provides closer parallels with previous work on iterated learning. Indeed, as pointed out in Section 5, many of the algorithms that have been examined in previous work can be construed as a form of MAP estimation. This analysis potentially indicates how structured languages could emerge using general-purpose learning algorithms that embody only a weak preference for such structure. The relationship between this form of iterated learning and the stochastic EM algorithm implies that the stationary distribution over languages will center around those languages with highest prior probability. Since the maximum of the prior is a relative notion, the prior need not strongly favor languages for them to dominate the stationary distribution. For example, in the case of iterated learning with two languages explored in Section 5.2, the prior probability that $h = 1, \alpha$, could be manipulated over a wide range for a fixed value of ϵ, and have no effect on the stationary probability that $h = 1, \theta_1$. Thus, learners having a weak preference for L_1, with α only slightly greater than 0.5, would result in emergence of L_1 with exactly the same probability as having a strong preference, with α only slightly less than $1 - \epsilon$. This example illustrates how MAP learning can emphasize the weak biases of learners.

While these results provide some suggestive connections, there is still much work to be done in understanding the effects identified in previous work on iterated learning. In particular, the correspondence between iterated learning and stochastic EM leaves some unknowns about the factors influencing the stationary distribution over languages. In the example explored in Section 6, we obtained results consistent with an information bottleneck effect arising from MAP estimation, with the preference for compositional languages being emphasized when m was small. Obtaining systematic results connecting the stationary distribution with the size of the bottleneck is an important topic for future research. Some preliminary work exploring bottleneck effects with MAP learners appears in Dowman, Kirby, and Griffiths (2006).

8.4. Implications for explaining linguistic universals

In Section 1, we outlined two possible explanations for linguistic universals: the traditional idea that these universals are the result of strong constraints that are specific to language learning, and the alternative hypothesis that iterated learning might be able to produce languages

with the structural properties of human languages even if learners use general-purpose learning algorithms. Unfortunately, our formal results do not indicate which of these explanations is more plausible. However, they do help to fill in some gaps in the traditional argument, and provide some insight into the kind of questions that need to be asked to resolve the debate.

Advocates of the traditional explanation of linguistic universals in terms of constraints arising from an innate language faculty can seize upon our results for learners who sample from the posterior distribution, which dictate a one-to-one correspondence between the inductive biases of learners and the languages that ultimately manifest in a population. These results provide an important missing piece of the traditional story, indicating how soft constraints on individual learning can influence the languages spoken by a population. If the prior distribution is taken as reflecting innate constraints specific to language learning (a view which we do not encourage, as mentioned in Section 3), then our results show that iterated learning can act as the engine by which these constraints result in universals. Those who prefer the idea that linguistic universals can be produced by iterated learning exaggerating the weak biases expressed in general-purpose learning algorithms can take heart from our results for learners who use MAP estimation, which illustrate that iterated learning can potentially emphasize weak biases. The information bottleneck also provides a further factor that could contribute to this effect.

Even though our present results have something to offer for both sides of the debate, they do place some important constraints on possible explanations. In particular, our analysis suggests two questions that we might pursue in order to evaluate the plausibility of different explanations for linguistic universals. The first question is whether human language learners are better approximated as sampling from the posterior distribution or selecting the hypothesis with greater posterior probability. We provided some empirical arguments for the appropriateness of sampling in Section 4, but this is an issue that has not been explored in depth in the case of language learning (although see Hudson-Kam & Newport, 2005). If learners are closer to sampling than maximizing, strong constraints are necessary. If they are closer to maximizing, then the weak biases of general-purpose learning mechanisms might be sufficient. The second question is whether we can identify general-purpose learning algorithms that have inductive biases (albeit weak ones) consistent with the properties of human languages. Our results indicate that only languages favored by the prior will be produced by iterated learning, regardless of whether learners sample or maximize. In previous work, the inductive biases of different algorithms have largely been investigated by running iterated learning simulations (Kirby, 2001; Brighton, 2002; Smith et al., 2003). The connection between biases and universals indicated by our results suggests a more productive mode of analysis might be to investigate these biases directly, determining which kinds of languages are easier to learn with different prospective general-purpose learning algorithms.

8.5. Conclusion

Iterated learning is one of the basic mechanisms by which languages are passed from person to person, and generation to generation. Understanding the consequences of iterated learning is thus a crucial step towards understanding the process of language change. In this paper, we

have laid out a framework for analyzing iterated learning, making it possible to characterize both the dynamics and asymptotic behavior of processes in which learners learn from other learners. The formulation of iterated learning in terms of Markov chains on hypotheses and data makes it straightforward to determine the distribution over languages that should be expected to arise at any iteration (and the change in this distribution across iterations), as well as the asymptotic probability that any particular language will be chosen by a learner. The connection between the stationary distribution of the Markov chain on hypotheses and the equilibrium of population dynamics based on iterated learning also makes it possible to generalize the conclusions reached with sequences of individual learners to the level of the population. These results strengthen the contributions that use of the iterated learning model can make to our understanding of language change.

We have used this framework to provide a detailed account of the consequences of iterated learning with Bayesian learners. The use of Bayesian inference makes it possible to explicitly identify the inductive biases of our learners, and to demonstrate that these biases strongly affect the outcome of iterated learning. Formally establishing the basic predictions that result from iterated learning paves the way towards being able to develop more complete models of language evolution, exploring the interaction between iterated learning and forces of cultural and biological selection. It also provides insight into the plausibility of different explanations of linguistic universals, providing new terms in which to formulate the debate about the relationship between linguistic universals and the inductive biases of learners.

Going beyond language evolution, our analysis of iterated learning with Bayesian learners provide conditions under which information transmitted via iterated learning will ultimately come to mirror the structure of the mind of the learners. This information will be influenced by the inductive biases of learners, regardless of the source of those biases and whether the learners sample from the posterior or use MAP estimation. This correspondence suggests that we should look closely at the universal properties of human languages, since, under this account, they should reflect the biases behind human language learning. More generally, our results suggest that any information transmitted by a process of iterated learning—not just languages, but also legends, religious concepts, and social norms—will ultimately come to be tailored to match people's inductive biases, providing a formal justification for treating these phenomena as a source of clues about the assumptions that guide human thought.

Notes

1. We will assume that $\mathcal{H}$ and $\mathcal{D}$ are both finite sets. This is not a necessary assumption, but simplifies the statement of the results we present in this paper.
2. A similar use of Bayes' rule to characterize the role of inductive biases in language acquisition in the context of iterated learning appears in Kirby, Smith, and Brighton (2004).
3. This definition of ϵ treats it as the error rate of production, but it could equivalently be considered the error rate of perception. Under either of these interpretations, it characterizes the variation in the perceptual data available to the learner.

4. We use $\boldsymbol{x}$ and $\boldsymbol{z}$ simply to denote vector-valued random variables in this section. The choice of these variable names matches much of the literature on the EM algorithm, and should not be connected with the use of $\boldsymbol{x}$ and $\boldsymbol{z}$ in other sections.
5. A small influence of the prior remains even at asymptote due to the presence of holistic hypotheses that are equivalent to compositional hypotheses. These hypotheses cannot be separated by any amount of data, so the stability ratio approaches $\frac{64}{62+1/\alpha}$ as $m \rightarrow \infty$. If $\mathcal{H}$ did not include hypotheses that make equivalent predictions about the data, the stability ratio would approach 1 as $m \rightarrow \infty$. Consequently, the decrease in the stability ratio as a function of m for the cases where $\alpha = 0.01$ is due to the specific structure of $\mathcal{H}$, rather than being a general trend.
6. The fact that $\mathbf{Q} - \mathbf{I}$ has an eigenvector with eigenvalue 0 indicates that this system has a one-dimensional linear subspace of equilibrium solutions, corresponding to multiples of the corresponding eigenvector. The uniqueness of $\boldsymbol{\theta}$ as an equilibrium follows from the fact that $\boldsymbol{\theta}$ is the only multiple of this eigenvector that is a probability distribution, with $\sum_i \theta_i = 1$.

Acknowledgments

We thank Henry Brighton, Nick Chater, Mike Dowman, Mark Johnson, Simon Kirby, Stephan Lewandowsky, Tania Lombrozo, Tim O'Donnell, Steven Sloman, and an anonymous reviewer for comments and conversations about this work. A preliminary version of the results for learners sampling from their posterior distributions was presented at the 27th Annual Conference of the Cognitive Science Society.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Ashby, F. G. (1992). *Multidimensional models of perception and cognition*. Hillsdale, NJ: Erlbaum.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216–233.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372–400.

Bilmes, J. A. (1997). *A gentle tutorial of the EM algorithm and its applications to parameter estimation for Gaussian mixture and hidden Markov models* (Tech. Rep. No. TR-97-021). Berkeley, CA: International Computer Science Institute.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, 25–54.

Briscoe, E. (2002). *Linguistic evolution through language acquisition: Formal and computational models*. Cambridge, UK: Cambridge University Press.

Celeux, G., Chauveau, D., & Diebolt, J. (1995). *On stochastic versions of the EM algorithm* (Tech. Rep. No. 2514). Montbonnot, France: Institut National de Recherche en Informatique et en Automatique.

Celeux, G., & Diebolt, J. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, *2*, 73–82.

Celeux, G., & Diebolt, J. (1988). A probabilistic teacher algorithm for iterative maximum likelihood estimation. In H. H. Bock (Ed.), *Classification and related methods of data analysis* (pp. 617–623). North-Holland: Elsevier.

Celeux, G., & Diebolt, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics Reports*, *41*, 119–134.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*, 566–581.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science*, *3*, 57–65.

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

Christiansen, M. H., & Kirby, S. (Eds.). (2003). *Language evolution*. Oxford: Oxford University Press.

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

Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, *27*, 94–128.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, *39*.

Diebolt, J., & Celeux, G. (1993). Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. *Communications in Statistics—Stochastic models*, *9*, 599–613.

Diebolt, J., & Ip, E. H. S. (1996). Stochastic EM: method and application. In W. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 259–273). Suffolk, UK: Chapman and Hall.

Dowman, M., Kirby, S., & Griffiths, T. L. (2006). Innateness and culture in the evolution of language. In A. Cangelosi, A. D. M. Smith, & K. Smith (Eds.), *The evolution of language: Proceedings of the 6th international conference*. Hackensack, NJ: World Scientific.

Fort, G., & Moulines, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *The Annals of Statistics*, *31*, 1220–1259.

Friedman, N. (1998). The Bayesian structural EM algorithm. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI 14)* (pp. 129–138).

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias-variance dilemma. *Neural Computation*, *4*, 1–58.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*, 355–407.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Suffolk, UK: Chapman and Hall.

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

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

Hirsch, M., & Smale, S. (1974). *Differential equations, dynamical systems, and linear algebra*. New York: Academic Press.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, *14*, 382–417.

Hudson-Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*, 151–195.

Hurford, J., Studdert-Kennedy, M., & Knight, C. (Eds.). (1998). *Approaches to the evolution of language: Social and cognitive bases*. Cambridge: Cambridge University Press.

Ip, E. H. (2002). On single versus multiple imputation for a class of stochastic algorithms estimating maximum likelihood. *Computational Statistics*, *17*, 517–524.

Ip, E. H. S. (1994). *A stochastic EM estimator in the presence of missing data—Theory and applications* (Tech. Rep.). Stanford, CA: Department of Statistics, Stanford University.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Kearns, M., & Vazirani, U. (1994). *An introduction to computational learning theory*. Cambridge, MA: MIT Press.

Kemeny, J. G., & Snell, J. L. (1983). *Finite Markov chains*. New York: Springer-Verlag.

Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.

Kirby, S. (1999). *Function, selection and innateness: The emergence of language universals*. Oxford: 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., & Hurford, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 121–148). London: Springer Verlag.

Kirby, S., Smith, K., & Brighton, H. (2004). From UG to universals: linguistic adaptation through iterated learning. *Studies in Language*, *28*, 587–607.

Komarova, N. L., Niyogi, P., & Nowak, M. A. (2001). The evolutionary dynamics of grammar acquisition. *Journal of Theoretical Biology*, *209*, 43–59.

Komarova, N. L., & Nowak, M. A. (2003). Language dynamics in finite populations. *Journal of Theoretical Biology*, *221*, 445–457.

Krifka, M. (2001). Compositionality. In R. A. Wilson & F. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences*. Cambridge, MA: MIT Press.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Li, M., & Vitanyi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. London: Springer Verlag.

Liu, J. S., Wong, W. H., & Kong, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society B*, *57*, 157–169.

Luce, R. D. (1959). *Individual choice behavior.* New York: John Wiley.

MacKay, D. (1995). Probable networks and plausible predictions—A review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, *6*, 469–505.

Mackay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

McLachlan, G., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.

Myers, J. L. (1976). Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Approaches to human learning and motivation* (pp. 171–205). Hillsdale, NJ: Erlbaum.

Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial Intelligence*, *56*, 71–113.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). University of Toronto.

Neal, R. M., & Hinton, G. E. (1998). A view EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.), *Learning in graphical models*. Cambridge, MA: MIT Press.

Nielsen, S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli*, *6*, 457–489.

Niyogi, P., & Berwick, R. C. (1995). *The logical problem of language change* (Tech. Rep.). AI Lab, MIT. (AI Memo-1516)

Niyogi, P., & Berwick, R. C. (1996). A language learning model for finite parameter spaces. *Cognition*, *61*, 161–193.

Niyogi, P., & Berwick, R. C. (1997a). A dynamical systems model for language change. *Complex Systems*, *11*, 161–204.

Niyogi, P., & Berwick, R. C. (1997b). Evolutionary consequences of language learning. *Linguistics and Philosophy*, *20*, 697–719.

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–108.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2001). Evolution of universal grammar. *Science*, *291*, 114–118.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature*, *417*, 611–617.

Nowak, M. A., Plotkin, J. B., & Jansen, V. A. A. (2000). The evolution of syntactic communication. *Nature*, *404*, 495–498.

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition.* Oxford: Oxford University Press.

Quine, W. V. O. (1960). *Word and object.* Cambridge, MA: MIT Press.

Rice, S. (2004). *Evolutionary theory: Mathematical and conceptual foundations.* Sunderland, MA: Sinauer.

Robert, C. P. (1994). *The Bayesian choice: A decision-theoretic motivation.* New York: Springer.

Rosenthal, J. S. (1995). Convergence rates of Markov chains. *SIAM Review*, *37*, 387–405.

Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: MIT Press.

Savage, L. J. (1954). *Foundations of statistics.* New York: John Wiley & Sons.

Schervish, M. J., & Carlin, B. P. (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical Statistics*, *1*, 111–127.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Sherman, R. P., Ho, Y.-Y. K., & Dalal, S. R. (1999). Conditions for convergence of Monte-Carlo EM sequences with an application to product diffusion modeling. *The Econometrics Journal*, *2*, 248–267.

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

Stewart, W. J. (1994). *Introduction to the numerical solution of Markov chains.* Princeton, NJ: Princeton University Press.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453–489.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, *82*, 528–550.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–641.

Vapnik, V. N. (1995). *The nature of statistical learning theory.* New York: Springer.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*, 101–118.

Wei, G. C. G., & Tanner, M. A. (1990). A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Assocation*, *85*, 699–704.