# ITERATED LEARNING OF MULTIPLE LANGUAGES FROM MULTIPLE TEACHERS

DAVID BURKETT

*Computer Science Division, University of California, Berkeley*
*Berkeley, CA, 94720, USA*
dburkett@cs.berkeley.edu

THOMAS L. GRIFFITHS

*Psychology Department, University of California, Berkeley*
*Berkeley, CA, 94720, USA*
tom_griffiths@berkeley.edu

Language learning is an iterative process, with each learner learning from other learners. Analysis of this process of iterated learning with chains of Bayesian agents, each of whom learns from one agent and teaches the next, shows that it converges to a distribution over languages that reflects the inductive biases of the learners. However, if agents are taught by multiple members of the previous generation, who potentially speak different languages, then a single language quickly dominates the population. In this work, we consider a setting where agents learn from multiple teachers, but are allowed to learn multiple languages. We show that if agents have a sufficiently strong expectation that multiple languages are being spoken, we reproduce the effects of inductive biases on the outcome of iterated learning seen with chains of agents.

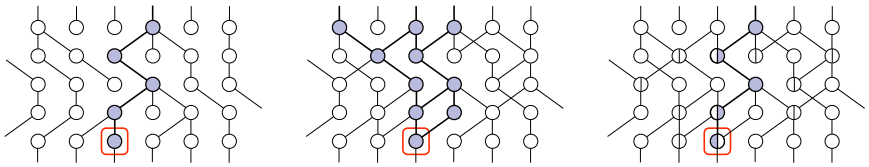
## 1. Introduction

Natural languages change as they are passed from person to person and from generation to generation. Although many explanations have been proposed for these changes, one way to analyze the process is to focus on the iterative nature of language learning: people learning a language are being taught by other people who themselves previously learned that language. Formal analysis of this "iterated learning" process has yielded some important insights into how learners' biases affect the languages likely to be used by a population (Kirby, 2001). In particular, it has been shown that if we assume the learners are Bayesian agents who compute posterior distributions over languages based on their prior beliefs and evidence from the previous generation of learners, then the iterated learning procedure will converge on a population whose preferences reflect the learners' prior beliefs (Kirby, Dowman, & Griffiths, 2007; Griffiths & Kalish, 2007). This relationship between the prior beliefs of a population and the stationary distribution

over linguistic characteristics is important for understanding the evolution of human language use. For example, it suggests that we should more closely examine universal properties of natural languages, since they are likely to reflect the biases underlying human language learning.

Most analyses of iterated learning with Bayesian agents have assumed that each learner receives linguistic data from exactly one member of the previous generation. This has led to the criticism that such learning dynamics are unrealistic and do not adequately model the full range of linguistic evolutionary processes (Niyogi & Berwick, 2009). However, a recent study begins to address this critique, showing that if learners consider evidence from the entire previous generation of speakers, then the results of iterated learning with Bayesian agents do not depend entirely on the learners' prior beliefs (Smith, 2009). Instead, the population comes to almost entirely speak one language, and the initial composition of the linguistic community is important. In particular, learners can consistently overcome their prior beliefs if they are learning from a multi-speaker population whose distribution of languages conflicts strongly enough with this prior.

Although the results of iterated learning with multiple teachers seems to conflict with earlier findings, there is a sense in which this model has an unusual dynamics. Each learner is attempting to decide on a single language despite the fact that the population that the learner is using as a source of evidence is, in the aggregate, multilingual. The learners are violating the principle of Bayesian rational analysis (Ferdinand & Zuidena, 2009) by making an unjustified assumption that only one language is being spoken. An alternative is that the learner not only assumes that the evidence is coming from multiple speakers, but also assumes that different speakers may be speaking different languages (see Figure 1).



(a) Single teacher (b) Multiple teachers (c) One language per teacher

Figure 1. (a) If each learner learns from a single teacher in the previous generation, then the learning dynamics for infinitely sized discrete generations are equivalent to those for chains of individual learners, as each learner's ancestry is still a single chain. (b) If learners still learn a single language, but get data from multiple teachers, then we see different learning dynamics. (c) However, as learners consider multilingual hypotheses, in the limiting case where each teacher is assumed to speak a different language, we recover the learning dynamics from a single chain of learners.

One way to allow hypotheses consistent with data received from multiple speakers is to directly estimate the probabilities of words. We show that if learners adopt this approach, the iterated learning procedure will converge to reflect their

prior beliefs, as in the simpler single teacher formulation. However, this learning model makes the most sense in a context where it is reasonable to interpolate between the languages spoken by one's predecessors: say, when learning from teachers who speak different dialects of the same basic language. In order for learners to deal appropriately with truly divergent inputs, we need a more complex learning process that explicitly models the possibility that they need to learn multiple complete linguistic systems. Intuitively, we expect that if learners are able to appropriately separate their input into distinct languages, then the learning dynamics will resemble those from the single teacher setting, as shown in Figure 1 (c). We test this intuition by modeling learners of this sort, and showing that in simulations, iterated learning with learners who believe that they are receiving data from multiple languages once more converges on the learners' prior beliefs.

## 2. Iterated Learning with Bayesian Agents

For Bayesian agents, learning is modeled as a statistical inference about the hypothesis $h$ that generated data $d$. The agent computes the posterior distribution:

$$p(h|d) \propto p(d|h)p(h) \tag{1}$$

where $p(h)$ is the agent's prior distribution over hypotheses, and $p(d|h)$ is a likelihood expressing the probability of data $d$ being generated for hypothesis $h$.

In iterated learning, the data that each agent learns from is generated from the previous generation of learners. If we let $p_t(h)$ represent the proportion of agents speaking language $h$ at time $t$ and assume that each learner receives data generated by one teacher, the probability of an agent receiving data $d$ at time $t$, $p_t(d)$, is:

$$p_t(d) = \sum_h p_t(h)p(d|h) \tag{2}$$

Similarly, if each learner samples a hypothesis from $p(h|d)$, the next generation's distribution over hypotheses is: $p_{t+1}(h) = \sum_d p_t(d)p(h|d)$, with $p(h|d)$ as given in Equation 1. This is the model considered in previous analyses of iterated learning by Bayesian agents, and results in convergence of $p_t(h)$ to the prior $p(h)$ as $t \to \infty$ (Griffiths & Kalish, 2007).

This model can be extended to allow multiple teachers. Assume that the data $d$ consists of a collection of independently produced words, $w$, with $p(d|h) = \prod_{w \in d} p(w|h)$. The case where all the words are generated by a single teacher is given by expanding Equation 2 as $p_t(d) = \sum_h p_t(h) \left(\prod_{w \in d} p(w|h)\right)$. If we allow each word to be potentially generated by a different teacher, as in the alternate model used in Smith (2009), we have to select a new hypothesis for each word, resulting in the modified distribution:

$$p_t(d) = \prod_{w \in d} \left( \sum_h p_t(h)p(w|h) \right) \tag{3}$$

Smith (2009) showed that when this model was used with a hypothesis space consisting of two hypotheses, each having one highly diagnostic word, $p_t(h)$ converged to a distribution that was dominated by one hypothesis, with the specific hypothesis resulting from an interaction between the prior and the initial proportion of the population who used that hypothesis.

The Bayesian inference described in Equation 1 is intended to identify which single hypothesis generated a collection of words, $d$. Therefore, the learners are making an estimate that is consistent with data generated according to Equation 2, but *not* with data generated according to Equation 3. Therefore, to properly model a Bayesian learner who receives data according to Equation 3, we need to consider a different hypothesis space: one that takes into account the fact that the learner is receiving data from multiple underlying distributions. This will be the focus of the remainder of this paper.

## 3. Learning Distributions over Words

One way to allow hypotheses consistent with data received from multiple speakers is to directly estimate the probabilities of words from the vocabulary. This results in a continuous hypothesis space: for a vocabulary of size $V$, a hypothesis $h$ is a member of the $V$-dimensional simplex. For the simple two-word vocabulary used by Smith (2009), $h$ can be summarized by a single parameter $\theta \in [0, 1]$, and the production probabilities can be written as $p(w|\theta) = \theta^{\delta(w,w_0)}(1-\theta)^{\delta(w,w_1)}$, where $\delta$ is the Kroenecker's delta function. Note that with this production probability, we can rewrite Equation 3 as:

$$p_t(d) = \prod_{w \in d} \left( \int_\theta p_t(\theta) p(w|\theta) d\theta \right) = E_t(\theta)^{n_0} \left(1 - E_t(\theta)\right)^{n_1}$$

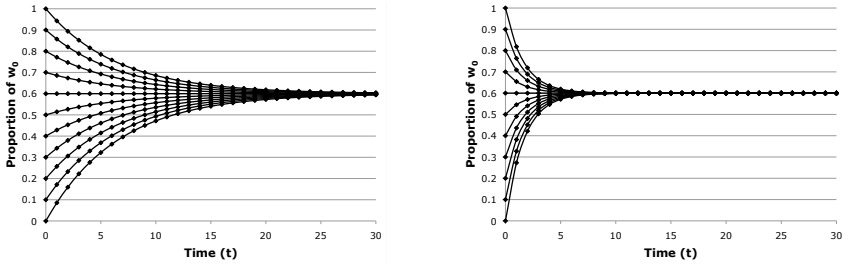
where $E_t(\theta)$ is the mean of the current generation's distribution over hypotheses $p_t(\theta)$, and $n_i = \sum_{w \in d} \delta(w, w_i)$ for $i \in \{0, 1\}$. Because our hypothesis space is equivalent to the set of Bernoulli distributions, the conjugate prior is the beta distribution.[a] Therefore, we define $p(\theta)$ to be a beta prior, with hyperparameters $\alpha$ and $\beta$, which results in an iterated Bayesian learning process equivalent to the Wright-Fisher model of genetic drift (Reali & Griffiths, in press).

Given the equivalence to the Wright-Fisher model, we expect that the iterated learning procedure will converge to a distribution over $\theta$ that is closely related to the prior (Reali & Griffiths, in press). To verify this, we ran a simple computational simulation. Recall that under our transmission model, the behavior of an entire generation of learners can be summarized by a single value: $E_t(\theta)$. We assume learners select a hypothesis by sampling from their posterior distribution over hypotheses, so linearity of expectation makes this straightforward to compute

[a] More generally, for a vocabulary of size $V$, the conjugate prior is the $V$-dimensional Dirichlet.

from $E(\theta|d)$, the posterior mean. $E_{t+1}(\theta) = \sum_d p_t(d)E(\theta|d)$. Fortunately, due to the conjugacy of the beta prior, $E(\theta|d)$ has a simple form: $\frac{\alpha+n_0}{\alpha+\beta+n_0+n_1}$.

Following Smith (2009), we ran simulations for various proportions of $w_0$ spoken in the initial population, assuming an infinite population of learners. We fixed $|d| = 3$. Varying this parameter resulted in slower convergence as $|d|$ increased, but did not affect the qualitative results. We also experimented with different values of $\alpha$, but fixed $\beta = \frac{2}{3}\alpha$ so that the expected value of $\theta$ under the prior was always 0.6, slightly favoring $w_0$. The results of our simulations are in Figure 2. The main finding is that under all the settings under consideration, the proportion of $w_0$ spoken in the population converged to 0.6. In other words, this model exhibited the expected convergence to a proportion favored by the prior.




(a) $\alpha = 0.3$ (b) $\alpha = 1.5$

Figure 2. Simulation results for the two hypothesis model. Here, the number of words seen by each learner $|d| = 3$, and the prior is parameterized with $\beta = \frac{2}{3}\alpha$, favoring proportions of $w_0$ close to 0.6.

## 4. Learning Distributions over Languages

A more general way to deal appropriately with multiple linguistically divergent teachers is for learners to explicitly consider arbitrary sets of languages, but to shift the hypothesis space from individual languages to *distributions* over languages. Thus, a learner is simultaneously inducing from data *which* languages are being spoken by the previous generation and their relative frequencies. We use $h$ to refer to a full hypothesis: a distribution over languages, $l$, where each language $l$ is a distribution over words $w$. For example, in the two-word setting, there might be two languages: $l_0$, where $w_0$ is spoken with probability 0.95 and $w_1$ with probability 0.05, and $l_1$, where those probabilities are reversed. A bilingual agent with a slight preference for $l_0$ can be represented as having hypothesis $h_{0.6}$, where $p(l_0|h_{0.6}) = 0.6$ and $p(l_1|h_{0.6}) = 0.4$.

The Dirichlet process (DP) (Ferguson, 1973) provides a suitable family of priors for this hypothesis space. A DP prior has two parameters: $\alpha$, which affects the learner's prior belief in the number of languages being spoken (a learner with a higher value of $\alpha$ will tend to predict a larger number of distinct languages), and

a base distribution $G_0$, which is a distribution over languages, specifying which languages are preferred.

### 4.1. *Inference*

Exact inference in the space of distributions over languages under a DP prior is intractable. However, we can approximate the dynamics of the Markov chain by running Monte Carlo simulations with collections of artificial agents separated into discrete generations. Procedurally, these simulations are straightforward. We start out with an initial collection of agents, $A_0$. Each of these agents receives some data according to the starting conditions of that simulation (see Section 4.2 for details). Then, based on that data, the agent picks a specific hypothesis. We then create a new generation of agents, $A_1$. Each agent in $A_1$ receives data generated by the agents in $A_0$, and chooses its own hypothesis accordingly. This procedure is iterated for some fixed number of generations, with each agent in $A_t$ receiving data collectively from the agents in $A_{t-1}$.

There are thus two steps that we have to perform repeatedly. First, given some data $d$, we need to be able to draw a sample from the posterior distribution $p(h|d)$. Though we omit the mathematical details here, a sample can be obtained efficiently by using a Gibbs sampler based on the Chinese Restaurant Process (Aldous, 1985). The second step is to sample some data, $d$, from a collection of agents, $A$. We sample each word independently according to a multistep procedure. First, an agent $a$ is selected uniformly from $A$. Then, a language is sampled according to that agent's hypothesis. Finally, the word is sampled from the selected language. This procedure is repeated for each word in the data. The number of words in the data, $|d|$, is fixed as before. This amounts to drawing each word according to: $p(w) = \sum_{a \in A} \frac{1}{|A|} \sum_l p(l|h_a) \sum_w p(w|l)$.

### 4.2. *Simulations*

We ran simulations of this learning procedure using the 260-language compositional vs. holistic setting from Griffiths and Kalish (2007). In this setting, each word, $w$, represents a form-meaning pair, $(x, y)$, where $x$ and $y$ each have a two-bit representation. Each language corresponds to a mapping between forms and meanings on which its production probabilities depend. The holistic languages range over all $4^4 = 256$ possible mappings between the 4 forms and 4 meanings, whereas the compositional languages map each bit individually, and thus range over only $2^2 = 4$ possible mappings. The actual production probabilities are:

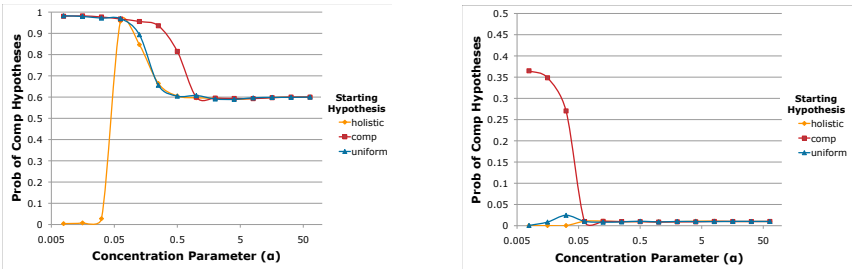
$$p(x, y|l) = p(y)p(x|y, l) = \begin{cases} \frac{1}{4}(1-\epsilon) & y \text{ maps to } x \text{ in } l \\ \frac{1}{4}\frac{\epsilon}{3} & \text{otherwise} \end{cases}$$

with $\epsilon$ a free parameter. The base distribution, $G_0$, is parameterized by $p_0$, determining the probability of a compositional language. The distribution selects

uniformly given the class of language:

$$G_0(l) = \begin{cases} \frac{p_0}{4} & l \text{ is compositional} \\ \frac{1-p_0}{256} & l \text{ is holistic} \end{cases}$$

For these simulations, we fixed the number of agents in each generation: $|A| = 100$, and each agent learned from a data set of fixed size: $|d| = 20$. We set $\epsilon = 0.05$ and ran the simulations for 50 generations each. There are three different types of starting conditions: in "holistic," 90% of the starting data was generated by a particular holistic hypothesis (one with minimal overlap with the compositional hypotheses) and the remaining 10% was drawn uniformly from the set of possible words. In "compositional," 90% of the starting data was generated by a particular compositional hypothesis. In "uniform," all the starting data was generated uniformly. We report values for various settings of $\alpha$ and $p_0$. Here, the values reported are the total final probabilities of compositional hypotheses, averaged over 50 runs.[b] The results are in Figure 3.




(a) $p_0 = 0.6$ (b) $p_0 = 0.01$

Figure 3. Simulation results with a richer pool of languages and multilingual learners. As the concentration parameter $\alpha$ increases, so does the extent to which the learner expects teachers to speak different languages. $p_0$ controls the strength of the prior in favor of compositional hypotheses.

The general trend of these results is that for low values of $\alpha$, the population of learners tends to converge to the hypothesis most favored by the initial conditions, whereas for high values of $\alpha$, we see convergence to the prior. This is consistent with previous work, which can be viewed as limiting conditions of this framework: as $\alpha \to 0$, we obtain a prior assumption by the learner that only one language is being spoken in the general population, as in Smith (2009), while as $\alpha \to \infty$, we obtain an assumption that each individual word is generated from a separate

[b]For smaller values of $\alpha$, the individual hypotheses most consistent with the starting data were favored, whereas for larger values of $\alpha$, the probabilities of individual hypotheses were generally uniform over each hypothesis class.

hypothesis, which is equivalent to the learning dynamics from Griffiths and Kalish (2007) (see Figure 1). Thus, the concentration parameter of the DP prior provides a natural way to interpolate between these two patterns of results.

## 5. Conclusion

The simulations we have presented show that when an agent's hypothesis space explicitly takes into account the possibility of receiving input from multiple teachers with possibly divergent hypotheses, then iterated Bayesian learning generally converges to reflect learners' inductive biases, as when agents learn from only a single teacher. However, if we explicitly encode a bias in the agent towards believing that the teachers all share a single hypothesis, then we may observe results that more closely align with initial data conditions. These results provide a way to understand how learners might learn from multiple teachers, but nonetheless show significant effects of inductive biases in the languages that they come to speak.

## References


Aldous, D. (1985). Exchangeability and related topics. In *Lecture notes in mathematics* (Vol. 1117, p. 1-198). Springer, Berlin.

Ferdinand, V., & Zuidena, W. (2009). Thomas' theorem meets Bayes' rule: a model of the iterated learning of language. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*, 209-230.

Griffiths, T. L., & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*, 441-480.

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

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

Niyogi, P., & Berwick, R. C. (2009). The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences*, *106*, 10124-10129.

Reali, F., & Griffiths, T. L. (in press). Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of the Royal Society, Series B*.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*