Self-Sustaining Iterated Learning*

Bernard Chazelle[1] and Chu Wang[2]

1 **Princeton University, Princeton, USA**
chazelle@cs.princeton.edu
2 **Nokia Bell Labs, Murray Hill, USA**
chu.wang@nokia-bell-labs.com

Abstract

An important result from psycholinguistics (Griffiths & Kalish, 2005) states that no language can be learned iteratively by rational agents in a self-sustaining manner. We show how to modify the learning process slightly in order to achieve self-sustainability. Our work is in two parts. First, we characterize iterated learnability in geometric terms and show how a slight, steady increase in the lengths of the training sessions ensures self-sustainability for any discrete language class. In the second part, we tackle the nondiscrete case and investigate self-sustainability for iterated linear regression. We discuss the implications of our findings to issues of non-equilibrium dynamics in natural algorithms.

1998 ACM Subject Classification I.2.6 Learning

Keywords and phrases Iterated learning, language evolution, iterated Bayesian linear regression, non-equilibrium dynamics

Digital Object Identifier 10.4230/LIPIcs.ITCS.2017.17

1 Introduction

Consider this hypothetical scenario: A native speaker of Quenya[1] sets out to teach the language to an English speaker; after a year of teaching, the learner considers herself fluent enough to teach Quenya to some other English speaker, who a year later does the same. In this form of *iterated learning*, agents teach each other in sequence: X teaches Y, who then teaches Z, who then teaches...[2, 8, 7, 15, 12, 9, 14, 16, 18, 11]. By a classic result of Griffiths and Kalish [7], Quenya will vanish after a finite number of iterations, at which point the agents, assumed to be rational, will be "teaching" each other plain English. In other words, after a while, learners will be taught nothing they don't already know: iterated learning is not self-sustaining.

Such findings are hard to validate empirically but variants of it are within the reach of experimental psychology. As early as 1932, in fact, the English psychologist Frederic Bartlett used iterated learning to expose hidden biases among humans. He presented a picture of an owl to a person for given period of time and then asked her to draw it from memory. Her picture was then shown to the next learner for the same amount of time, who then proceeded to draw it back from memory. After 20 iterations of this process, to Bartlett's surprise, what was being drawn was no longer an owl but, quite clearly, a cat! The challenge was to explain why humans would exhibit a pro-feline bias without falling into the trap of just-so stories.

* This work was supported in part by NSF grant CCF-1420112.
[1] Quenya is one of J.R.R. Tolkien's fictional languages.

Griffiths et al. [11] repeated the same experiment ten years ago, this time trading owls for lines. The goal was to see if linear regression could be iterated: the answer was a resounding No. Skipping over logistical details, the experiment presents the first learner with a cloud of 20 points drawn randomly, with noise, from the line $Y = 1 - X$. The cloud vanishes and the learner is then asked to reconstruct it from memory. She then becomes the teacher by passing on her own cloud to the next learner, who likewise, looks at it for a while, and then tries to reconstruct it from memory, etc. Surprisingly, iterating this process a mere nine times leads the last learner in the sequence to draw a cloud that regresses to the line $Y = X$; in other words, teaching about descending lines iteratively has precisely the opposite effect! In fact the initial picture is essentially is irrelevant. A random cloud of points will also lead to $Y = X$.

Unlike the Quenya scenario, where the bias toward English is not unexpected, the cat and line experiments both reveal a hidden prior among the participants. Humans seem to love cats and possess a strong positive correlation bias; it is easy to speculate why.[2] It is noteworthy that the prior should prevail even in the absence of any sort of priming. Indeed, this experiment fails miserably if you try it yourself by playing the role of all the agents in sequence. The use of different learners ensures that the training does not acquire long-term memory. Similar laboratory experiments with human subjects (well, undergraduates) have confirmed the unstainability of iterated learning [11, 2, 19, 1, 9].

In our first example, Quenya gets “washed out" by English in a way reminiscent of the fixation of an allele through genetic drift. Indeed, the original impetus for studying iterated learning in psycholinguistics was to look for a parallel to Kimura’s neutral theory of molecular evolution in the area of cultural transmission. People learn their native tongue from speakers who themselves learned it from others. This process introduces variation along the way, some of which is retained durably. The selectionist view seeks to explain this process by fitness considerations at the population level. Iterated learning suggests a different explanation. Language acquisition suffers from a well-documented information bottleneck (the notorious “poverty of stimulus"), so one might expect languages to evolve so as to be easy to learn: could complexity theory be the key? This push for simplicit would then trigger the emergence of linguistic universals (eg, compositionality) that one finds present in all languages [8]. This view complements – some will argue, contradict – Chomsky’s interpretation of universals as the product of constraints imposed by an innate genetic endowment.

Following Chomsky and Lasnik’s theory of “Principles and Parameters," Rafferty et al. [15] model languages by means of a handful of parameters: think of a few knobs whose settings specify any given language. Language evolution thus entails the trans-generational update of a probability distribution over that parameter space. Assuming that the learners are rational Bayesian agents, iterated learning acts as a Gibbs sampler for a joint probability distribution over languages and their sentences. By converging to a stationary distribution, iterated learning proves incapable of sustaining itself past the mixing time. In that model, languages evolve to reflect the priors of the learners while losing all trace of the ancestor language. While this phenomenon is of central relevance in the study of universal grammars, it leaves open the possibility that changes in the sampling algorithm might make iterated learning self-sustaining. Of course, it is easy to think of situations where this feature would be highly desirable (eg, school teaching, social transmission of norms, legends, jokes, etc.) We show how keeping the length of the training sessions growing slightly allows iterated

[2] Our favorite piece of anecdotal evidence in support of the positive slope bias is that no road sign in the US features an aircraft on a descending path.

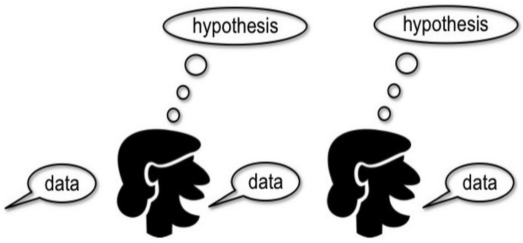

Figure 1 Chained iterated learning.

learning to be sustained in perpetuity.

In the first part of the paper, we characterize iterated learnability in geometric terms and show how a slight, steady increase in the lengths of the learning sessions ensures self-sustainability for any discrete language class. In the second part, we tackle the nondiscrete case and investigate self-sustainability for iterated Bayesian linear regression. In all cases, self-sustainability requires making the underlying Markov process time-inhomogeneous in order to stay out of equilibrium. This gives us an opportunity to offer a few thoughts on the growing importance of non-equilibrium in natural algorithms.

Background

Following [2, 8, 7, 15, 12, 9, 14, 16, 18, 11], we begin with *chained iterated learning*: a learner's prior is modeled by a distribution over a hypothesis space $\mathcal{H}$, which is itself equipped with a likelihood function: $\mathbb{P}[d|h]$ indicates the probability of generating data $d \in \mathcal{D}$ given the hypothesis $h \in \mathcal{H}$. The first learner samples m_1 items *iid* from the initial hypothesis $\boldsymbol{h}_{\texttt{init}}$: these items provide the training data $\mathbf{d}_1 = (d_{1,1}, \ldots, d_{1,m_1})$ with which the first learner Bayes-updates its prior. Its posterior is given by setting $t = 1$ in this formula:

$$\mathbb{P}[h|\mathbf{d}_t] = \mathbb{P}[\mathbf{d}_t|h]\,\mathbb{P}[h]/\mathbb{P}[\mathbf{d}_t], \quad \text{with } \mathbb{P}[\mathbf{d}_t] = \sum_{h\in\mathcal{H}} \mathbb{P}[\mathbf{d}_t|h]\,\mathbb{P}[h]. \tag{1}$$

From that point on, each successive learner updates its prior from their predecessor. For any $t > 1$, learner t receives m_t items sampled from the posterior of agent $t-1$ to form the training set $\mathbf{d}_t$. To do that, she picks a random hypothesis h from $\mathcal{H}$ with probability $\mathbb{P}[h|\mathbf{d}_{t-1}]$ (the posterior of learner $t-1$) and then samples m_t items *iid* from h to form $\mathbf{d}_t \in \mathcal{D}^{m_t}$. The posterior $\mathbb{P}[h|\mathbf{d}_t]$ is derived according to (1). Note that learner t has no direct access to the posterior of learner $t-1$ but only to data drawn from a hypothesis sampled from the posterior. Our formulation assumes a discrete space $\mathcal{H}$ but extends to continuous settings, as we show in §3.

In the case of linguistic transmission, each hypothesis $h \in \mathcal{H}$ is a "knob" whose setting is given by a number between 0 and 1, specifically the prior probability $\mathbb{P}[h]$. All learners share the same prior. Picking some h from that prior specifies a *language* (also denoted h for convenience). In this case, a language is defined as a probability distribution over $\mathcal{D}$, interpreted here as a set of *sentences*. In this way, the prior can be viewed as a mixture over $\mathcal{H}$: by abuse of terminology, we call it a *mixed* hypothesis, which we distinguish from a *pure* hypothesis of the form $h \in \mathcal{H}$ (corresponding to a single-point distribution). Access to language h is achieved by random sampling: the sentence $d \in \mathcal{D}$ is picked with probability $\mathbb{P}[d|h]$.

Iterated learning proceeds as follows. After selecting language h with probability $\mathbb{P}[h|\mathbf{d}_{t-1}]$, learner t collects m_t independent samples from h. Thus, given a tuple $\mathbf{d}_t = (d_1, \ldots, d_{m_t})$ of sentences from $\mathcal{D}$, the likelihood $\mathbb{P}[\mathbf{d}_t|h]$ is equal to $\prod_{1\le k\le m_t} \mathbb{P}[d_k|h]$. The learner is now

ready to Bayes-update its prior. Of course, the first one ($t = 1$) samples directly from the language $\boldsymbol{h}_{\texttt{init}}$ chosen for iterated learning. The notation is boldfaced to indicate that $\boldsymbol{h}_{\texttt{init}}$ may be a mixed hypothesis or, in other words, a distribution over hypotheses.

Suppose that $\mathcal{D} = \{d_1, \ldots, d_s\}$ and $\mathcal{H} = \{h_1, \ldots, h_n\}$ are both finite. While sampling from the posterior of learner $t-1$, if learner t winds up choosing h_i then, by Bayesian updating, the probability $P_{ij}^{|t}$ that its posterior picks h_j is given by:

$$P_{ij}^{|t} = \sum_{\mathbf{d} \in \mathcal{D}^{m_t}} \mathbb{P}[h_j|\mathbf{d}]\, \mathbb{P}[\mathbf{d}|h_i] = \sum_{\mathbf{d} \in \mathcal{D}^{m_t}} \frac{\mathbb{P}[\mathbf{d}|h_i]\, \mathbb{P}[\mathbf{d}|h_j]\, \mathbb{P}[h_j]}{\sum_{k=1}^{n} \mathbb{P}[\mathbf{d}|h_k]\, \mathbb{P}[h_k]}\,. \tag{2}$$

To our knowledge, the entire literature on the topic assumes a common, fixed sample size for all the learners: $m_t = m$. Equation (2) can be then interpreted as marginalizing a Gibbs sampler over the data space, which creates a Markov chain over the hypothesis space $\mathcal{H}$: if $\boldsymbol{h}^t$ denotes the row vector formed by the n probabilities $\mathbb{P}[h_k \,|\, \mathbf{d}_t]$, then $\boldsymbol{h}^t = \boldsymbol{h}^{t-1} P^t$, where $\boldsymbol{h}^0 = \boldsymbol{h}_{\texttt{init}}$. Assuming ergodicity (in this case, a fairly inconsequential technical assumption), the chain can be shown to converge to a unique stationary distribution $\boldsymbol{h}$. It can be easily checked that it coincides with the prior: $\boldsymbol{h} = (\mathbb{P}[h_1], \ldots, \mathbb{P}[h_n])$ [7, 13]; see [15, 16] for an analysis of the mixing time in specific linguistic scenarios. This convergence reveals the long-term unsustainability of iterated learning. We show how diversifying the sample sizes m_t, hence making the Markov chain time-inhomogeneous, can overcome this weakness.

Our results

In §2, we show how to achieve self-sustainability in the discrete setting [8, 7], using only a logarithmically increasing sample size; specifically, the new hypothesis to be learned is acquired by all the (infinitely many) learners with probability at least $1 - \varepsilon$ using a sample size of $O(\log \frac{t}{\varepsilon})$ for the t-th learner. The constant factor depends on the geometry of the hypothesis space. By relaxing the objective and allowing learners to settle on an arbitrarily close approximation of the hypothesis to be learned, we can remove all dependency on the geometry of the hypothesis space.

In §3, we extend the iterated learning model to a Gaussian setting for an infinite hypothesis space and show that a sample size of $O(t)^{1+o(1)}$ is sufficient to ensure self-sustainability. We also show that allowing learners to pick their teachers at random cuts down the sample size to $O(\log t)^{1+o(1)}$. The arguments used for the discrete case bump into singularities so we use a different approach, which allows us to exploit various "stability" properties of the Gaussian setting.

In §4, we turn our attention to the iterated version of Bayesian linear regression and prove a high-probability statement about self-sustainability. This requires spectral arguments from random matrix theory and, in particular, bounds on the lowest singular value of Wishart matrices.

Discussion

Before moving to the technical part of this work, we add a few thoughts about its larger context and relevance. For a dynamicist, the loss of Quenya is a byproduct of the memory-erasing ergodicity implied by mixing. For a physicist, the loss is due to the Second Law of thermodynamics and the bounded supply of free energy available to each agent: together these two constraints make it impossible to keep the system out of equilibrium. For a biologist, this entropic pull toward equilibrium is the hallmark of a dying system. Evolution is nature's attempt to optimize the absorption of free energy into work while maximizing

the production of entropy. The first requirement is keeping the system out of equilibrium over timescales well in excess of the metabolic rate (here, the teaching rate). From that perspective, our work can be seen as an effort to find out the minimum conditions necessary to keep a target dynamics active in perpetuity. There are several approaches to this question and the two we follow are among the simplest: (i) increasing the supply of free energy (eg, lengthening the training sessions) and (ii) mixing timescales (eg, rewiring the communication network).

Most of the work on Markov chains in theoretical computer science regards mixing as a blessing: large spectral gaps are good while small ones are to be avoided. In biology, however, mixing often means death. In fact, much of life can be seen as nature's attempt to keep mixing at bay. This paper explores what can be done to prevent a Markov chain from reaching equilibrium. We expect this theme to gain prominence in future work on natural algorithms.

2 Self-Sustainability

We show how to make iterated learning self-sustaining in the presence of a finite hypothesis space $\mathcal{H} = \{h_1, \ldots, h_n\}$. This involves specifying a sequence of training session lengths $m_1, m_2, \ldots$ so that the posterior of any learner ends up differing from $\boldsymbol{h}_{\texttt{init}}$ by an arbitrarily small amount. Formally, given any $\delta, \varepsilon \geq 0$, we say that iterated learning is (δ, ε)-*self-sustaining* if, with probability at least $1 - \varepsilon$, a random $h \in \mathcal{H}$ picked from any learner's posterior distribution differs from $\boldsymbol{h}_{\texttt{init}}$ in total variation by at most δ. We recall a few facts: the hypothesis h denotes a language modeled as a probability distribution over $\mathcal{D}$; the total variation distance is half the ℓ_1-norm; and the posterior of learner t after the t-th iteration is defined by marginalizing $\mathbb{P}[h|\mathbf{d}_t]$ over all samples $\mathbf{d}_t$ drawn from a random h picked from the posterior of learner $t-1$ (or $\boldsymbol{h}_{\texttt{init}}$ if $t = 1$). As a shorthand, we speak of ε-self-sustainability to refer to the case $\delta = 0$.

The parameters δ and ε allow us to distinguish between two metrics: the distance between two languages over $\mathcal{D}$ and the distance between two mixtures over $\mathcal{H}$. The two notions could differ widely. For example, if all of $\mathcal{H}$ corresponds to languages very close to $\boldsymbol{h}_{\texttt{init}}$, to achieve (δ, ε)-self-sustainability might be easy for a tiny $\delta > 0$ but hopelessly difficult for $\delta = 0$. The complexity of iterated learning depends on the geometry of the languages formed by the pure hypotheses. This is best captured by introducing a metric that, though more specialized than the total variation (it works only on the simplex of probability vectors) brings all sorts of technical benefits: the *root-sine distance* between two probability distributions $\boldsymbol{a} = (a_1, \ldots, a_s)$ and $\boldsymbol{b} = (b_1, \ldots, b_s)$ over $\mathcal{D}$ is defined as

$$d_{RS}(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\frac{1}{2} \sum_{i,j=1}^{s} \left(\sqrt{a_i b_j} - \sqrt{a_j b_i} \right)^2} = \sqrt{1 - \left(\sum_{i=1}^{s} \sqrt{a_i b_i} \right)^2}. \tag{3}$$

It would be surprising if this distance had not been used before, but we could not find a reference. We prove that it is indeed a metric in the Appendix and also explain its name. We show that it is related to the Hellinger, Bhattacharyya and total variation distances, d_H, d_B, d_{TV} by the following relations:

$$\begin{cases} d_H = \sqrt{1 - \sqrt{1 - d_{RS}^2}}\,; \\ d_B = -\frac{1}{2} \ln(1 - d_{RS}^2)\,; \\ d_{TV} \leq \sqrt{2s}\, d_{RS}. \end{cases} \tag{4}$$

2.1 The results

We focus on the "pure" case $\boldsymbol{h}_{\texttt{init}} \in \mathcal{H}$, and later briefly discuss how to generalize the method to mixed hypotheses. Using the shorthand $\mathtt{d}_{ij}$ for $d_{RS}(\mathbb{P}[\cdot|h_i], \mathbb{P}[\cdot|h_j])$, we define $\mathtt{d}_i := \min_{j:j\neq i} \mathtt{d}_{ij}$. Let $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ be the prior distribution over $\mathcal{H}$, where $p_i := \mathbb{P}[h_i]$. We can obviously assume that each p_i is positive and that all the pure hypotheses are distinct, hence $\mathtt{d}_i > 0$. The two theorems below assume that $\boldsymbol{h}_{\texttt{init}} = h_1$.

▶ **Theorem 1.**. *For any positive $\varepsilon < 1$, the following sample size sequence makes iterated learning ε-self-sustaining:*

$$m_t = \frac{4}{\mathtt{d}_1^2} \ln \frac{nt}{\varepsilon\, p_1} = \frac{4}{\mathtt{d}_1^2}\Big(\log \frac{t}{\varepsilon} + C\Big),$$

for some $C > 0$ independent of $t, \varepsilon, \mathtt{d}_1$.

The factor 4 can be reduced to $2^{1+o(1)}$ if we adjust the constant C. It is to be expected that the lengths of the training sessions should grow to infinity as p_1 tends to zero, as the vanishing prior makes it increasingly difficult for the posteriors to "attach" to h_1. The session lengths are sensitive to the minimum distance between the languages specified by $\mathcal{H}$ and the target language h_1. Settling for (δ, ε)-self-sustainability allows us to remove this dependency.

▶ **Theorem 2.**. *For any positive $\delta, \varepsilon < 1$, the following sample size sequence makes iterated learning (δ, ε)-self-sustaining:*

$$m_t = \frac{8sn^2}{\delta^2}\Big(\ln \frac{t}{\varepsilon} + C\Big).$$

for some $C > 0$ independent of t, δ, ε.

2.2 The proofs

To establish Theorem 1, we estimate the probability P^* that each leaner ends up picking h_1. Recall that $\boldsymbol{h}^t$ is the posterior distribution of learner t, by the Markovian property of the system,

$$P^* = \mathbb{P}[\boldsymbol{h}^0 = h_1] \prod_{t\geq 0} \mathbb{P}[\boldsymbol{h}^{t+1} = h_1 | \boldsymbol{h}^t = h_1] = \prod_{t\geq 1} P_{11}^{|t}. \tag{5}$$

Since the matrix $P^{|t}$ is the transition matrix of a Markov chain, we proceed by bounding its off-diagonal elements $P_{ij}^{|t}$ for $i \neq j$. By (2) and Young's inequality,

$$\begin{aligned}
P_{ij}^{|t} &\leq \sum_{\mathbf{d}\in\mathcal{D}^{m_t}} \frac{\mathbb{P}[\mathbf{d}|h_i]\,\mathbb{P}[\mathbf{d}|h_j]p_j}{\mathbb{P}[\mathbf{d}|h_i]p_i + \mathbb{P}[\mathbf{d}|h_j]p_j} = \frac{p_j}{p_i} \sum_{\mathbf{d}\in\mathcal{D}^{m_t}} \frac{\big(\frac{p_i}{p_j}\big)\mathbb{P}[\mathbf{d}|h_i]\,\mathbb{P}[\mathbf{d}|h_j]}{\big(\frac{p_i}{p_j}\big)\mathbb{P}[\mathbf{d}|h_i] + \mathbb{P}[\mathbf{d}|h_j]} \\
&\leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}} \sum_{\mathbf{d}\in\mathcal{D}^{m_t}} \sqrt{\mathbb{P}[\mathbf{d}|h_i]\,\mathbb{P}[\mathbf{d}|h_j]} = \frac{1}{2}\sqrt{\frac{p_j}{p_i}}\Big(\sum_{d\in\mathcal{D}} \sqrt{\mathbb{P}[d|h_i]\,\mathbb{P}[d|h_j]}\Big)^{m_t} \\
&\leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}} \exp\left\{\frac{m_t}{2}\Big(\Big(\sum_{d\in\mathcal{D}} \sqrt{\mathbb{P}[d|h_i]\,\mathbb{P}[d|h_j]}\Big)^2 - 1\Big)\right\}.
\end{aligned}$$

By definition of the root-sine distance, we have

$$P_{ij}^{|t} \leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}} e^{-\frac{1}{2}\mathtt{d}_{ij}^2 m_t} \quad (i \neq j). \tag{6}$$

Setting $i = 1$ in (6) and summing over $2 \le j \le n$, it follows by Cauchy-Schwarz that

$$\sum_{j=2}^{n} P_{1j}^{|t} \le \frac{1}{2}\sqrt{\frac{n(1-p_1)}{p_1}}\, e^{-\frac{1}{2}\mathtt{d}_1^2 m_t}. \tag{7}$$

Combining (5) and (7) yields

$$P^* \ge \prod_{t\ge 1}\left(1 - \frac{1}{2}\sqrt{\frac{n(1-p_1)}{p_1}}\, e^{-\frac{1}{2}\mathtt{d}_1^2 m_t}\right) \ge 1 - \frac{1}{2}\sqrt{\frac{n(1-p_1)}{p_1}}\sum_{t\ge 1} e^{-\frac{1}{2}\mathtt{d}_1^2 m_t}. \tag{8}$$

Given $0 < \varepsilon < 1$, we constrain the sequence (m_t) to satisfy:

$$\sum_{t\ge 1} e^{-\frac{1}{2}\mathtt{d}_1^2 m_t} < \varepsilon\sqrt{\frac{4p_1}{n(1-p_1)}}. \tag{9}$$

For example, we can pick the sequence

$$m_t = \frac{1}{\mathtt{d}_1^2}\ln\frac{n(1-p_1)t^4}{\varepsilon^2 p_1},$$

which completes the proof. A closer look at the calculation shows that the factor t^4 can be reduced to $C_\alpha t^{2+\alpha}$ for any small $\alpha > 0$ and a suitable constant $C_\alpha > 0$, which makes the dependency on t arbitrarily close to $(2/\mathtt{d}_1^2)\ln t$. ◀

To prove Theorem 2, we set a target distance $\rho := \delta/(n\sqrt{2s})$ and find a subset $A \subseteq \mathcal{H}$ such that (i) $\mathtt{d}_{1j} \le \rho n$ for $j \in A$ and (ii) $\mathtt{d}_{ij} \ge \rho$ for $i \in A$ and $j \notin A$. To see why such a subset must exist, consider spheres centered at $\boldsymbol{h}_{\mathtt{init}} = h_1$ of radius $k\rho$, for $k = 1, \ldots, n+1$ (with respect to d_{RS}). These define $n+1$ disjoint (open) regions and, by the pigeonhole principle, at least one of them must be empty. We set A to include all the points in the regions preceding the empty one; note that $h_1 \in A$. The claim follows from the triangular inequality. We begin with a straightforward generalization of (7): for any $i \in A$,

$$\sum_{j\notin A} P_{ij}^{|t} \le \frac{1}{2}\sqrt{\frac{n(1-p_A)}{p_A}}\, e^{-\frac{1}{2}\rho^2 m_t}, \tag{10}$$

where $p_A := \min_{i\in A} p_i$. Now let P^* be the probability that $\boldsymbol{h}^t \in A$ for each t, then (5) and (8) are generalized to

$$P^* \ge \prod_{t\ge 1}\left(1 - \max_{i\in A}\sum_{j\notin A} P_{ij}^{|t}\right) \ge 1 - \frac{1}{2}\sqrt{\frac{n(1-p_A)}{p_A}}\sum_{t\ge 1} e^{-\frac{1}{2}\rho^2 m_t}. \tag{11}$$

Setting

$$m_t = \frac{1}{\rho^2}\ln\frac{n(1-p_A)t^4}{\varepsilon^2 p_A} \tag{12}$$

ensures that $P^* > 1-\varepsilon$. The root-sine distance between the languages denoted by h_1 and any $h \in A$ is at most ρn, so that, by (4), the total variation distance is bounded by $\sqrt{2s}\rho n = \delta$, which concludes the proof of Theorem 2. ◀

So far, we have analyzed only the "pure" case $\boldsymbol{h}_{\mathtt{init}} \in \mathcal{H}$. The idea of the training is to prevent the prior to "drag" the posterior mixture all across $\mathcal{H}$. It should be clear that a

similar result obtains if $\boldsymbol{h}_{\texttt{init}} \in \Delta\mathcal{H}$ is concentrated on a subset A of $\mathcal{H}$. The proof follows the path charted in Theorem 2 and need not be repeated here. It is crucial to note, however, that this result is to be understood in a coarse-graining sense: iterated learning cannot ensure that the original weights in the mixture $\boldsymbol{h}_{\texttt{init}}$ are retained but only that A contributes most of the mass in the posteriors. To retain the weights would require changing the stationary distribution to conform with $\boldsymbol{h}_{\texttt{init}}$, as the process unfolds, something that straightforward Bayesian learning seems unable to do. Learning pure hypotheses bypasses that difficulty.

2.3 Applications

We briefly discuss a direct application of our results to a well-known model of language acquisition via iterated learning and we mention some natural extensions of the techniques.

Language evolution

Rafferty et al. [15] show how iterated learning fails rapidly in a simple model of language evolution. Given n hypotheses, iterated learning with fixed-length training sessions ceases to learn anything new after only $O(\log n \log\log n)$ rounds. The previous theorems show how to turn this around and achieve self-sustainality. In the model, $\mathcal{H} = \{h_1, \ldots, h_n\}$, where $n = 2^k$ and h_i denotes the language whose sentences are words in $\{0, 1, ?\}^k$ with exactly m question marks and $0, 1$ matching the binary decomposition of $i-1$ outside the question marks. For example, if $k = 4$ and $m = 2$, then h_3 denotes the language

$$\{\, 00??, 0?1?, ?01?, 0??0, ?0?0, ??10 \,\}.$$

We can assume that m is much smaller than k. Each language has the same length $\binom{k}{m}$ and the total number of sentences is $s = \binom{k}{m}2^{k-m}$. The prior is given by $\mathbb{P}[h_i] = p_i = 1/n$. Given a hypothesis h_i, $\mathbb{P}[d|h_i] = 1/\binom{k}{m}$ if d has m question marks and match the bits of $i-1$ elsewhere; else it is 0 (and d, h are called incompatible). Given $h \in \mathcal{H}$,

$$\begin{cases} \mathbb{P}[d] = \sum_{h\in\mathcal{H}} \mathbb{P}[d|h]\mathbb{P}[h] = 2^{m-k}/\binom{k}{m}\,; \\ \mathbb{P}[h|d] = \mathbb{P}[d|h]\mathbb{P}[h]/\mathbb{P}[d] = 2^{-m} \quad \text{(or 0 if } d, h \text{ are incompatible).} \end{cases}$$

We easily check that $\mathtt{d}_1^2 = 1 - \left(\sum_{i=1}^s \sqrt{a_i b_i}\right)^2 \geq 1 - \left(\frac{m}{k}\right)^2 > \frac{1}{2}$; hence, by Theorem 1, session lengths m_t no larger than $O(\log \frac{t}{\varepsilon})$ are sufficient to maintain ε-self-sustainability.

Meanings and utterances

In the use of iterated learning for studying language evolution [7, 14], it is common to model the data $\mathbf{d}$ as a joint distribution $(\boldsymbol{x}, \boldsymbol{y})$ over a product space $\mathcal{X}^{m_t} \times \mathcal{Y}^{m_t}$. The idea is to distinguish between "meanings" $\boldsymbol{x}$ and "utterances" $\boldsymbol{y}$. In this setting, $\mathbb{P}[\mathbf{d}|h] = \mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h]\mu(\boldsymbol{x})$, where $\mu(\boldsymbol{x})$ is the probability of generating $\boldsymbol{x}$. The transition matrix of the Markov chain thus becomes

$$\begin{aligned} P_{ij}^{|t} &= \sum_{\boldsymbol{x}\in\mathcal{X}^{m_t}} \sum_{\boldsymbol{y}\in\mathcal{Y}^{m_t}} \mathbb{P}[h_j|\boldsymbol{x}, \boldsymbol{y}]\, \mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_i]\mu(\boldsymbol{x}) \\ &= \sum_{\boldsymbol{x}\in\mathcal{X}^{m_t}} \sum_{\boldsymbol{y}\in\mathcal{Y}^{m_t}} \frac{\mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_i]\, \mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_j]\, \mathbb{P}[h_j]}{\sum_{k=1}^m \mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_k]\, \mathbb{P}[h_k]} \mu(\boldsymbol{x})\,. \end{aligned} \tag{13}$$

Since the output $\boldsymbol{y}$ now depends on both the hypothesis and the input data, we redefine $\mathtt{d}_{ij}$ as the root-sine distance between the two distributions $\mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_i]\mu(\boldsymbol{x})$ and $\mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_j]\mu(\boldsymbol{x})$:

$$\mathtt{d}'_{ij} := 1 - \left(\sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{\boldsymbol{y}\in\mathcal{Y}} \sqrt{\mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_i]\mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_j]}\,\mu(\boldsymbol{x}) \right)^2 \tag{14}$$

and we define $\mathtt{d}'_i := \min_{j:j\neq i} \mathtt{d}'_{ij}$. Given any $i \neq j$,

$$\begin{aligned}
P_{ij}^{|t} &\leq \sum_{\boldsymbol{x}\in\mathcal{X}^{m_t}} \sum_{\boldsymbol{y}\in\mathcal{Y}^{m_t}} \frac{\mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_i]\,\mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_j]\,p_j}{\mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_i]\,p_i + \mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_j]\,p_j}\mu(\boldsymbol{x}) \\
&\leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}} \sum_{\boldsymbol{x}\in\mathcal{X}^{m_t}} \sum_{\boldsymbol{y}\in\mathcal{Y}^{m_t}} \sqrt{\mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_i]\,\mathbb{P}[\mathbf{y}|\boldsymbol{x}, h_j]}\,\mu(\boldsymbol{x}) \\
&\leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}} \Big(\sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{\boldsymbol{y}\in\mathcal{Y}} \sqrt{\mathbb{P}[\boldsymbol{y}|h_i]\,\mathbb{P}[\boldsymbol{y}|h_j]}\,\mu(\boldsymbol{x}) \Big)^{m_t} \\
&\leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}} \exp\left\{ \frac{m_t}{2} \Big(\Big(\sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{\boldsymbol{y}\in\mathcal{Y}} \sqrt{\mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_i]\,\mathbb{P}[\boldsymbol{y}|\boldsymbol{x}, h_j]}\,\mu(\boldsymbol{x}) \Big)^2 - 1 \Big) \right\}.
\end{aligned}$$

This gives us this new version of inequality (6), which we can use as the basis for a repeat of the argument of the previous section:

$$P_{ij}^{|t} \leq \frac{1}{2}\sqrt{\frac{p_j}{p_i}} e^{-\frac{1}{2}\mathtt{d}'^2_{ij} m_t} \qquad (i \neq j). \tag{15}$$

3 Iterated Learning in Continuous Spaces

When iterated learning operates over a hypothesis space $\mathcal{H}$ parametrized continuously, say, in $\mathbb{R}$, the minimum root-sine distance usually vanishes and the previous arguments run into singularities and collapse. A new approach is needed. To make our discussion concrete, we assume that the prior distribution of each learner is a Gaussian $\mathbb{P}[h] \sim N(\bar{\mu}, \bar{\sigma}^2)$ and that the likelihood of producing data d given hypothesis h is also normal: $\mathbb{P}[d|h] = N(h, \sigma^2)$. The likelihood can also be understood as a noisy measurement of h: $d = h + \phi$, where the noise $\phi \sim N(0, \sigma^2)$. We assume that the data received by the first learner comes from $N(\mu_0, \sigma_0^2)$. This is the simplest instance of a continuous setting in which the root-sine distance argument fails. We discuss it in some detail, considering both chained learning and its generalizations; and then we use the results to treat the case of iterated Bayesian linear regression.

During its training session, the t-th learner receives data $\mathbf{d}_t = (d_{t,1}, \ldots, d_{t,m_t})$ from its predecessor: it is obtained by first picking a random hypothesis h from the posterior of learner $t-1$ and then collecting m_t independent random samples from $N(h, \sigma^2)$. For the case $t = 1$, we can treat the original teacher as learner 0 with its posterior equal to $N(\mu_0, \sigma_0^2)$. Learner t Bayes-updates its posterior as follows:

$$\mathbb{P}[h|\mathbf{d}_t] \propto \mathbb{P}[\mathbf{d}_t|h]\mathbb{P}[h] \propto \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{m_t}(d_{t,i} - h)^2\Big) \exp\Big(-\frac{1}{2\bar{\sigma}^2}(h - \bar{\mu})^2\Big),$$

which is still Gaussian, with mean and variance denoted by μ_t and σ_t^2, respectively. Carrying out the usual square completion gives up these update rules: for $t > 0$,

$$\begin{cases} \mu_t = \frac{1}{\bar{\tau}+m_t\tau}\left(\bar{\tau}\bar{\mu} + \tau(d_{t,1} + d_{t,2} + \cdots + d_{t,m_t})\right) \\ \tau_t \;= \bar{\tau} + m_t\tau, \end{cases} \tag{16}$$

where we define the precisions $\tau = 1/\sigma^2$, $\bar{\tau} = 1/\bar{\sigma}^2$, and $\tau_t = 1/\sigma_t^2$. We say that iterated learning is *ε-self-sustaining* if $|\mathbb{E}\,\mu_t - \mu_0| \le \varepsilon$ and $\sigma_t^2 + \operatorname{var}\mu_t$ remains bounded for all t. If $\sigma_t^2 + \operatorname{var}\mu_t \to 0$ as $t \to \infty$, we say that iterated learning is *strongly ε-self-sustaining*. We consider successively the case of chained iterated learning and the more challenging "hopping" scenario in which a new learner picks a random teacher from the past (instead of the previous one).

3.1 Chained learning

In chained iterated learning, the data $d_{t,i}$ is a noisy message drawn from the posterior of the $(t-1)$-th learner; hence $d_{t,i} \sim N(\mu_{t-1}, \sigma_{t-1}^2 + \sigma^2)$. In view of (16), μ_t is itself Gaussian. By taking the expectation and variance of equation (16), we find the following recursive relations for $\mathbb{E}\,\mu_t$ and $\operatorname{var}\mu_t$: for $t > 0$,

$$\begin{cases} \mathbb{E}\,\mu_t = \frac{1}{\bar{\tau}+m_t\tau}\big(\bar{\tau}\bar{\mu} + m_t\tau\,\mathbb{E}\,\mu_{t-1}\big); \\ \operatorname{var}\mu_t = \frac{m_t\tau^2}{(\bar{\tau}+m_t\tau)^2}\big(\operatorname{var}\,\mu_{t-1} + \sigma_{t-1}^2 + \sigma^2\big). \end{cases} \tag{17}$$

If we define $\beta_t := m_t\tau/(\bar{\tau} + m_t\tau)$, then (17) becomes $\mathbb{E}\,\mu_t = \beta_t\,\mathbb{E}\,\mu_{t-1} + (1-\beta_t)\bar{\mu}$. If $m_t = m$ is a constant, then so is β_t, and the recursive relation (17) becomes

$$\mathbb{E}\,\mu_t - \bar{\mu} = \beta_1^t(\mu_0 - \bar{\mu}),$$

which shows that $\mathbb{E}\,\mu_t$ converges to $\bar{\mu}$ exponentially fast. As in the discrete case, iterated learning is not self-sustainable with constant-length training sessions. By letting m_t increase as $O(t^{1+o(1)})$ order, however, we can achieve self-sustainability:

▶ **Theorem 3.** *For any $0 < \varepsilon < 1$, the following sample size sequence makes chained iterated learning strongly ε-self-sustaining:*

$$m_t = \frac{|\mu_0 - \bar{\mu}|}{\varepsilon}\Big(1 + \frac{1}{c}\Big)\Big(\frac{\sigma}{\bar{\sigma}}\Big)^2 t^{1+c},$$

for an arbitrarily small constant $c > 0$.

Proof. We observe that $\mathbb{E}\,\mu_t$ is a convex combination of $\bar{\mu}$ and $\mathbb{E}\,\mu_s$ $(s < t)$; specifically,

$$\mathbb{E}\,\mu_t = \prod_{s=1}^{t}\beta_s\mu_0 + \Big(1 - \prod_{s=1}^{t}\beta_s\Big)\bar{\mu}. \tag{18}$$

Because $\sum_{s>0}(1/s)^{1+c} < 1 + \int_1^\infty x^{-1-c}\,dx = 1 + 1/c$, we have

$$\begin{aligned} 1 \;&\ge\; \prod_{s=1}^{t}\beta_s = \prod_{s=1}^{t}\Big(1 - \frac{\bar{\tau}}{m_s\tau + \bar{\tau}}\Big) \ge 1 - \sum_{s=1}^{t}\frac{\bar{\tau}}{m_s\tau + \bar{\tau}} \\ &\ge\; 1 - \frac{\varepsilon}{|\mu_0 - \bar{\mu}|}\Big(\frac{c}{c+1}\Big)\sum_{s=1}^{\infty}\frac{1}{s^{1+c}} > 1 - \frac{\varepsilon}{|\mu_0 - \bar{\mu}|}\,. \end{aligned}$$

This shows that

$$|\mathbb{E}\,\mu_t - \mu_0| = \Big(1 - \prod_{s=1}^{t}\beta_s\Big)|\bar{\mu} - \mu_0| \le \varepsilon.$$

By (16), $\sigma_t^2 = 1/\tau_t < 1/m_t\tau \to 0$. Since $\sigma_{t-1}^2 \le \bar{\sigma}^2$ for $t > 1$, it follows from (17) that $\operatorname{var}\mu_t \le (\operatorname{var}\mu_{t-1} + \sigma^2 + \bar{\sigma}^2)/m_t$ for $t > 1$, and $\operatorname{var}\mu_1 \le (\sigma_0^2 + \sigma^2)/m_1$. Writing $M_t := m_t m_{t-1} \dots m_1$, we have

$$\begin{aligned} M_t \operatorname{var}\mu_t &\le M_{t-1}\operatorname{var}\mu_{t-1} + M_{t-1}(\sigma^2 + \bar{\sigma}^2) \\ &\le tM_{t-1}(\sigma_0^2 + \sigma^2 + \bar{\sigma}^2), \end{aligned}$$

and thus $\operatorname{var}\mu_t \le (\sigma_0^2 + \sigma^2 + \bar{\sigma}^2)t/m_t \to 0$ since $m_t = \Omega(t^{1+c})$. ◀

3.2 Hopped learning

We consider the "hopped learning" scenario in which learner t hops back to pick a teacher from $\{0, 1, \dots, t-1\}$ at random, and then samples m_t bits of data from her posterior. The recursive relation for μ_t becomes

$$\mu_t = \frac{\beta_t}{m_t} \sum_{s=0}^{t-1} \chi_{t,s} \sum_{i=1}^{m_t} d_{t,s,i} + (1 - \beta_t)\bar{\mu}, \tag{19}$$

where, given t, the random variable $\chi_{t,s}$ is 1 for a value of s picked at random between 0 and $s-1$, and is zero elsewhere; recall that $\beta_t := m_t\tau/(\bar{\tau} + m_t\tau)$. Hopped iterated learning provides access to earlier data, so one would expect the lengths of the training sessions to grow more slowly than in chained learning. The change is indeed quite dramatic:

▶ **Theorem 4.** *For any positive $\varepsilon < |\mu_0 - \bar{\mu}|$, the following sample size sequence makes hopped iterating learning ε-self-sustaining:*

$$m_t = B_c \frac{|\mu_0 - \bar{\mu}|}{\varepsilon} \left(\frac{\sigma}{\bar{\sigma}}\right)^2 (1 + \log t)^{1+c},$$

for an arbitrarily small $c > 0$ and a constant B_c that depends only on c.

Proof. By taking expectation on both sides of (19), for any $t > 0$,

$$\mathbb{E}\,\mu_t = \frac{\beta_t}{t} \sum_{s=0}^{t-1} \mathbb{E}\,\mu_s + (1 - \beta_t)\bar{\mu},$$

We define $\gamma_1 = \beta_1$ and, for $t > 1$,

$$\gamma_t := (1 + \beta_1)\left(1 + \frac{\beta_2}{2}\right) \cdots \left(1 + \frac{\beta_{t-1}}{t-1}\right) \frac{\beta_t}{t}.$$

We verify easily that $\mathbb{E}\,\mu_t = \gamma_t\mu_0 + (1 - \gamma_t)\bar{\mu}$, for $t > 0$; therefore, the first part in establishing ε-self-sustainability consists of proving that

$$1 \ge \gamma_t \ge 1 - \frac{\varepsilon}{|\mu_0 - \bar{\mu}|}, \tag{20}$$

which will show that $|\mathbb{E}\,\mu_t - \mu_0| \le \varepsilon$. Note that

$$\gamma_t \le \frac{1}{t} \prod_{s=1}^{t-1} \left(1 + \frac{1}{s}\right) = 1.$$

Now define

$$\alpha_s = \frac{\varepsilon}{B_c |\mu_0 - \bar{\mu}| s (1 + \log s)^{1+c}}.$$

for $s > 0$. We pick a constant B_c large enough so that α_s is small enough to carry out first-order Taylor approximations around $1 + \alpha_s$. We find that

$$1 + \frac{\beta_s}{s} = 1 + \frac{1}{s}\Big(1 - \frac{1}{1 + m_t\tau/\bar\tau}\Big) \ge \Big(1 + \frac{1}{s}\Big)\Big(1 - \frac{1}{(s+1)m_t\tau/\bar\tau}\Big)$$
$$\ge \Big(1 + \frac{1}{s}\Big)\Big(1 - \frac{s\alpha_s}{s+1}\Big) \ge \Big(1 + \frac{1}{s}\Big)(1 - \alpha_s) \ge \Big(1 + \frac{1}{s}\Big)e^{-2\alpha_s}.$$

Thus,

$$\gamma_t \ge \frac{\beta_t}{t}\prod_{s=1}^{t-1}\Big(1 + \frac{1}{s}\Big)e^{-2\sum_{s=1}^{t-1}\alpha_s} = \beta_t e^{-2\sum_{s=1}^{t-1}\alpha_s} \ge 1 - \frac{\varepsilon}{|\mu_0 - \bar\mu|},$$

which establishes (20). Our derivation relies on the fact that

$$\beta_t \ge 1 - \frac{\varepsilon}{B_c|\mu_0 - \bar\mu|(1 + \log t)^{1+c}} \ge 1 - \frac{\varepsilon}{2|\mu_0 - \bar\mu|}$$

and

$$\sum_{s=1}^{t-1}\frac{1}{s(1+\log s)^{1+c}} \le 1 + \frac{1}{(\log e)^{1+c}}\int_2^{t-1}\frac{1}{x(\ln x)^{1+c}}\,dx = O\Big(\frac{1}{c}\Big);$$

hence,

$$e^{-2\sum_{s=1}^{t-1}\alpha_s} \ge e^{-O(\varepsilon/(cB_c|\mu_0-\bar\mu|))} \ge 1 - \frac{\varepsilon}{2|\mu_0 - \bar\mu|}.$$

Having shown that $|\mathbb{E}\,\mu_t - \mu_0| \le \varepsilon$ for all t, it now suffices to prove that $\sigma_t^2 + \operatorname{var}\mu_t$ remains bounded. We note that $\tau_t > m_t\tau \to \infty$, hence $\sigma_t^2 = 1/\tau_t \to 0$, so the remainder of the proof needs to establish that the variance of μ_t stays bounded. Writing $D_{t,s} := d_{t,s,1} + \cdots + d_{t,s,m_t}$, we have $\operatorname{var} D_{t,s} = m_t \operatorname{var} d_{t,s,1} = m_t(\sigma_s^2 + \sigma^2 + \operatorname{var}\mu_s)$; hence

$$\mathbb{E}\,D_{t,s}^2 = \operatorname{var} D_{t,s} + (\mathbb{E}\,D_{t,s})^2 = m_t(\sigma_s^2 + \sigma^2 + \operatorname{var}\mu_s) + m_t^2(\mathbb{E}\,\mu_s)^2.$$

In (19), the variables $\chi_{t,s}$ and $D_{t,s}$ are independent, for $0 \le s \le t-1$; furthermore, $\mathbb{E}\,\chi_{t,s} = \mathbb{E}\,\chi_{t,s}^2 = 1/t$, and $\mathbb{E}\,\chi_{t,s_1}\chi_{t,s_2} = 0$ if $s_1 \ne s_2$; therefore,

$$\operatorname{var}[\chi_{t,s}D_{t,s}] = \mathbb{E}\,\chi_{t,s}^2\,\mathbb{E}\,D_{t,s}^2 - (\mathbb{E}\,\chi_{t,s})^2(\mathbb{E}\,D_{t,s})^2 = \frac{\mathbb{E}\,D_{t,s}^2}{t} - \frac{(\mathbb{E}\,D_{t,s})^2}{t^2}$$
$$= \Big(\frac{m_t}{t}\Big)\big(\sigma_s^2 + \sigma^2 + \operatorname{var}\mu_s + m_t(\mathbb{E}\,\mu_s)^2\big) - \Big(\frac{m_t}{t}\Big)^2(\mathbb{E}\,\mu_s)^2 \tag{21}$$

and, for $s_1 \ne s_2$,

$$\begin{aligned}\operatorname{cov}[\chi_{t,s_1}D_{t,s_1}, \chi_{t,s_2}D_{t,s_2}] &= \mathbb{E}[\chi_{t,s_1}\chi_{t,s_2}D_{t,s_1}, D_{t,s_2}] - \mathbb{E}[\chi_{t,s_1}D_{t,s_1}]\mathbb{E}[\chi_{t,s_2}D_{t,s_2}]\\ &= \mathbb{E}[\chi_{t,s_1}\chi_{t,s_2}]\mathbb{E}[D_{t,s_1}D_{t,s_2}] - \mathbb{E}\,\chi_{t,s_1}\mathbb{E}\,D_{t,s_1}\mathbb{E}\,\chi_{t,s_2}\mathbb{E}\,D_{t,s_2}\\ &= -\frac{1}{t^2}\mathbb{E}\,D_{t,s_1}\mathbb{E}\,D_{t,s_2} = -\Big(\frac{m_t}{t}\Big)^2\mathbb{E}\,\mu_{s_1}\mathbb{E}\,\mu_{s_2}.\end{aligned} \tag{22}$$

Then, by taking the variance on both sides of (19), we have

$$\begin{aligned}
\operatorname{var}\mu_t &= \Big(\frac{\beta_t}{m_t}\Big)^2 \operatorname{var}\sum_{s=0}^{t-1}\chi_{t,s}D_{t,s}\\
&= \Big(\frac{\beta_t}{m_t}\Big)^2\Big(\sum_{s=0}^{t-1}\operatorname{var}\left[\chi_{t,s}D_{t,s}\right] + \sum_{0\le s_1\ne s_2\le t-1}\operatorname{cov}\left[\chi_{t,s_1}D_{t,s_1},\chi_{t,s_2}D_{t,s_2}\right]\Big)\\
&= \Big(\frac{\beta_t}{m_t}\Big)^2\Big(\sum_{s=0}^{t-1}\Big(\frac{m_t}{t}\Big)\big(\sigma_s^2+\sigma^2+\operatorname{var}\mu_s+m_t(\mathbb{E}\,\mu_s)^2\big) - \Big(\frac{m_t}{t}\Big)^2\Big(\sum_{s=0}^{t-1}\mathbb{E}\,\mu_s\Big)^2\Big)\\
&\le \frac{1}{tm_t}\sum_{s=0}^{t-1}\big(\sigma_s^2+\sigma^2+\operatorname{var}\mu_s+m_t(\mathbb{E}\,\mu_s)^2\big).
\end{aligned}$$

Notice that $\sigma_s^2\to 0$ and $(\mathbb{E}\,\mu_s)^2$ is bounded since $|\mathbb{E}\,\mu_t-\mu_0|\le\varepsilon$. We conclude that $\sigma_t^2+\operatorname{var}\mu_t$ remains bounded for all t. ◀

4 Iterated Bayesian Linear Regression

The iterated version of Bayesian linear regression has been the subject of extensive study in the field of psychology [11, 2, 19, 1, 9]. The work has involved experimentation with human subjects but little in the way of theoretical analysis. This section is a first step toward filling this void. The task at hand is to estimate a hypothesis $h\in\mathcal{H}:=\mathbb{R}^d$ given a noisy measurements on the hyperplane $y=h^Tx$, where $x\in\mathbb{R}^d$. In the Bayesian setting, we assume a Gaussian prior on the hypothesis space: $\mathbb{P}[h]\sim N(\bar\mu,\bar\sigma^2 I_d)$. The data is given by (x,y), where $x\sim N(0,I_d)$ and $y=h^Tx+\phi$, for $\phi\sim N(0,\sigma^2)$ (with x,ϕ independent). Since we typically make several measurements, we write this (likelihood) relation in matrix form: $y=Xh+\phi$, where $y\in\mathbb{R}^m$ (with m the number of measurements); $\phi\sim N(0,\sigma^2 I_m)$; and X is an m-by-d matrix each of whose rows denotes a random vector $x\sim N(0,I_d)$. This means that the matrix X is random (a fact of key importance in our discussion below). We have:

$$\begin{cases}
\mathbb{P}[\phi]\sim \exp\{-\frac{1}{2\sigma^2}\|\phi\|_2^2\} & \texttt{(noise)}\\
\mathbb{P}[h]\sim \exp\{-\frac{1}{2\bar\sigma^2}\|h-\bar\mu\|_2^2\} & \texttt{(prior)}\\
\mathbb{P}[y|X,h]\sim \exp\{-\frac{1}{2\sigma^2}\|y-Xh\|_2^2\} & \texttt{(likelihood)}
\end{cases}$$

In iterated Bayesian linear regression, the t-th learner receives her data from learner $t-1$. Here, learner 0 is treated just like any other agent, except that his prior $\mathbb{P}[h]\sim N(\mu_0,\bar\sigma^2 I_d)$ is the distribution to be learned iteratively. Since sampling from the prior is independent of X, Bayesian updating gives the posterior $N(\mu_t,\Sigma_t)$, where

$$\mathbb{P}[h|X,y]=\mathbb{P}[h]\,\mathbb{P}[y|X,h]/\mathbb{P}[y|X]\sim\exp\Big\{-\frac{1}{2\bar\sigma^2}\|h-\bar\mu\|_2^2-\frac{1}{2\sigma^2}\|y-Xh\|_2^2\Big\}.$$

Completing the square in the usual fashion shows that the posterior of learner t is given by:

$$\begin{cases}
\Sigma_t=\left(\bar\sigma^{-2}I_d+\sigma^{-2}X_t^TX_t\right)^{-1};\\
\mu_t=\Sigma_t\left(\bar\sigma^{-2}\bar\mu+\sigma^{-2}X_t^Ty_t\right),
\end{cases}\tag{23}$$

where (X_t,y_t) is the data gathered by learner t from her predecessor: specifically, $y_t=X_th+\phi_t$, where h is collected from the $(t-1)$-th learner by sampling his posterior distribution $N(\mu_{t-1},\Sigma_{t-1})$.

▶ **Theorem 5.** *Given any small enough $\delta, \varepsilon > 0$, the following sample size sequence for iterated Bayesian linear regression ensures that $\|\mathbb{E}\mu_t - \mu_0\|_2 \leq \delta$ with probability greater than $1 - \varepsilon$:*

$$m_t = D_c \frac{\|\mu_0 - \bar{\mu}\|_2}{\delta} \left(\frac{\sigma}{\bar{\sigma}}\right)^2 t^{1+c} + D_c\, d \log \frac{t+1}{\varepsilon},$$

for an arbitrarily small $c > 0$ and a constant D_c that depends only on c.

Proof. We proceed in two steps: first, we show that to keep $\mathbb{E}\mu_t$ arbitrarily close to μ_0 for all t hinges on spectral properties of certain random matrices; second, we call on known facts about the singular values of random Gaussian matrices to translate the spectral condition into a high-probability event. The proof unfolds as a series of simple relations, which we state first and then demonstrate. The first one follows directly from (23):

$$\mathbb{E}\mu_t = (I_d + M_t)^{-1} (\bar{\mu} + M_t\, \mathbb{E}\mu_{t-1}), \quad \text{where} \quad M_t := \left(\frac{\bar{\sigma}}{\sigma}\right)^2 X_t^T X_t. \tag{24}$$

Note that (24) is a randomized recursive relation since the data points $X_1, X_2, \ldots$ are themselves random. We note that all the matrices whose inverses are taken are positive definite, hence nonsingular. To move on to our second relation, we define the matrix

$$Q_t := (I_d + M_t)^{-1} M_t (I_d + M_{t-1})^{-1} M_{t-1} \cdots (I_d + M_1)^{-1} M_1,$$

for $t > 0$, with $Q_0 = I_d$, and prove by induction that

$$\mathbb{E}\,\mu_t = Q_t \mu_0 + (I_d - Q_t)\bar{\mu}. \tag{25}$$

The base case is obvious so we assume that $t > 0$: by (24),

$$\begin{aligned}
\mathbb{E}\,\mu_t &= (I_d + M_t)^{-1}(\bar{\mu} + M_t\, \mathbb{E}\,\mu_{t-1}) \\
&= (I_d + M_t)^{-1}(\bar{\mu} + M_t Q_{t-1}\mu_0 + M_t(I_d - Q_{t-1})\bar{\mu}) \\
&= (I_d + M_t)^{-1} M_t Q_{t-1}\mu_0 + (I_d + M_t)^{-1}(I_d + M_t(I_d - Q_{t-1}))\bar{\mu} \\
&= Q_t\, \mu_0 + (I_d - (I_d + M_t)^{-1} M_t Q_{t-1})\bar{\mu},
\end{aligned}$$

which proves (25). Our next goal is to bound the information decay $\|\mathbb{E}\,\mu_t - \mu_0\|_2$. To do that, we investigate the spectral norm of the matrix $I_d - Q_t$, which leads to our third relation. We prove by induction that, for $t > 0$,

$$\|I_d - Q_t\|_2 \leq \sum_{s=1}^{t} \|A_s\|_2, \tag{26}$$

where $A_s := (I_d + M_s)^{-1}$. For $t = 1$, $Q_1 = (I_d + M_1)^{-1} M_1 = I_d - (I_d + M_1)^{-1}$ and the claim follows. If $t > 1$, then

$$\begin{aligned}
\|I_d - Q_t\|_2 &= \|(I_d - Q_{t-1}) + (Q_{t-1} - Q_t)\|_2 \\
&\leq \|I_d - Q_{t-1}\|_2 + \|Q_t - Q_{t-1}\|_2 \leq \sum_{s=1}^{t-1} \|A_s\|_2 + \|\Psi\|_2,
\end{aligned}$$

where $\Psi := (A_t M_t - I_d) Q_{t-1}$. Since $A_t(I_d + M_t) = I_d$, we have $\Psi = -A_t Q_{t-1}$. Each matrix M_s is positive semidefinite, so the eigenvalues of $(I_d + M_s)^{-1} M_s$ are of the form $\lambda/(1+\lambda)$, where $\lambda \geq 0$. This shows that all the eigenvalues of Q_s are between 0 and 1;

therefore $\|Q_s\|_2 \le 1$. The eigenvalues of $I_d - A_t M_t$ are the same as those of A_t; hence, by submultiplicativity, $\|\Psi\|_2 \le \|A_t\|_2 \|Q_{t-1}\|_2 \le \|A_t\|_2$, which establishes (26).

We are now ready to express the information decay in spectral terms. Pick an arbitrarily small constant $c > 0$ and assume that

$$\|A_s\|_2 \le \frac{\delta}{\|\bar{\mu} - \mu_0\|_2} \Big(\frac{c}{1+c}\Big) \Big(\frac{1}{s}\Big)^{1+c}. \tag{27}$$

By (25), $\mathbb{E}\,\mu_t - \mu_0 = (I_d - Q_t)(\bar{\mu} - \mu_0)$; therefore, by (26),

$$\begin{aligned} \|\mathbb{E}\,\mu_t - \mu_0\|_2 &\le \|\bar{\mu} - \mu_0\|_2 \sum_{s=1}^{t} \|A_s\|_2 \le \frac{\delta c}{1+c} \sum_{s=1}^{t} s^{-1-c} \\ &\le \frac{\delta c}{1+c} \Big(1 + \int_1^{\infty} x^{-1-c}\, dx\Big) = \delta, \end{aligned} \tag{28}$$

The relation says that, on average, the means of any of the agents' posteriors can be brought as close to the original mean to be learned as we want. We can turn this into a high-probability event by using some basic random matrix theory. Recall that $\mathbb{E}\,\mu_t$ is itself a random variable whose stochasticity comes from the matrices X_s, which are all drawn from Gaussians. Because M_s is positive semidefinite,

$$\|A_s\|_2 \le \|M_s^{-1}\|_2 \le \frac{(\sigma/\bar{\sigma})^2}{\lambda_{\min}(X_t^T X_t)} \le \Big(\frac{\sigma/\bar{\sigma}}{\sigma_1(X_t)}\Big)^2, \tag{29}$$

which gives us a relation between the spectral norm of $(I_s + M_s)^{-1}$ and the smallest singular value $\sigma_1(X_t)$ of an m_t-by-d matrix X_t whose elements are drawn *iid* from $N(0,1)$. The asymptotic behavior of $\sigma_1(X_t)$ for large values of m_t has been extensively studied within the field of random matrix theory [5, 6, 17]. Following Theorem II.13 in (Davidson & Szarek [5]), for any $\gamma_t > 0$,

$$\mathbb{P}[\sigma_1(X_t) < \sqrt{m_t} - \sqrt{d} - \gamma_t] \le e^{-\gamma_t^2/2}.$$

We use C below as a generic constant large enough to satisfy the inequalities where it appears. Setting $\gamma_t = C\sqrt{\log((t+1)/\varepsilon)}$ ensures that $\sum_{t>0} e^{-\gamma_t^2/2} < \varepsilon$, hence that $\sigma_1(X_t) < \sqrt{m_t} - \sqrt{d} - \gamma_t$ holds for all t with probability less than ε. With our setting of m_t, this means that, for all $t > 0$,

$$\mathbb{P}\Big[\,\sigma_1(X_t) \ge \frac{\sqrt{m_t}}{2}\,\Big] > 1 - \varepsilon. \tag{30}$$

Assuming the event in (30), it follows from (29) and our setting of m_t that

$$\|A_t\|_2 \le \frac{4}{m_t} \Big(\frac{\sigma}{\bar{\sigma}}\Big)^2 \le \frac{\delta}{\|\bar{\mu} - \mu_0\|_2} \Big(\frac{4}{D_c}\Big) \Big(\frac{1}{t}\Big)^{1+c};$$

hence (27) for D_c large enough. By (28, 30), this proves that, with probability greater than $1 - \varepsilon$, $\|\mathbb{E}\,\mu_t - \mu_0\|_2 \le \delta$ for all $t > 0$, which completes the proof. ◀

References

1 FC Bartlett. Remembering: a study in experimental and social psychology.(1932). 317 pp.

2 Aaron Beppu and Thomas L Griffiths. Iterated learning and the cultural ratchet. In *Proceedings of the 31st annual conference of the cognitive science society*, pages 2089–2094. Citeseer, 2009.

3 A Bhattachayya. On a measure of divergence between two statistical population defined by their population distributions. *Bulletin Calcutta Mathematical Society*, 35:99–109, 1943.

4 Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142–149. IEEE, 2000.

5 Kenneth R Davidson and Stanislaw J Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

6 Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, 9(4):543–560, 1988.

7 Thomas L Griffiths and Michael L Kalish. A bayesian view of language evolution by iterated learning. In *Proceedings of the 27th annual conference of the cognitive science society*, pages 827–832, 2005.

8 Thomas L Griffiths and Michael L Kalish. Language evolution by iterated learning with bayesian agents. *Cognitive Science*, 31(3):441–480, 2007.

9 Thomas L Griffiths, Michael L Kalish, and Stephan Lewandowsky. Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1509):3503–3514, 2008.

10 Michiel Hazewinkel. *Encyclopaedia of Mathematics*. Springer Science & Business Media, 2013.

11 Michael L Kalish, Thomas L Griffiths, and Stephan Lewandowsky. Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14(2):288–294, 2007.

12 Simon Kirby, Tom Griffiths, and Kenny Smith. Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28:108–114, 2014.

13 James R Norris. *Markov chains*. Cambridge university press, 1998.

14 Amy Perfors and Daniel Navarro. Language evolution is shaped by the structure of the world: An iterated learning analysis. In *Annual Conference*, 2011.

15 Anna N Rafferty, Thomas L Griffiths, and Dan Klein. Convergence bounds for language evolution by iterated learning. In *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*, 2009.

16 Anna N Rafferty, Thomas L Griffiths, and Dan Klein. Analyzing the rate at which languages lose the influence of a common ancestor. *Cognitive science*, 38(7):1406–1431, 2014.

17 Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009.

18 Kenny Smith. Iterated learning in populations of bayesian agents. In *Proceedings of the 31st annual conference of the cognitive science society*, pages 697–702. Citeseer, 2009.

19 Mónica Tamariz and Simon Kirby. Culture: copying, compression, and conventionality. *Cognitive science*, 39(1):171–183, 2015.

A Appendix

The two forms of the function d_{RS} in (3) make it clear that $0 \leq d_{RS}(\boldsymbol{a}, \boldsymbol{b}) \leq 1$ and $d_{RS}(\boldsymbol{a}, \boldsymbol{b}) = 0$ if and only if $\boldsymbol{a}$ and $\boldsymbol{b}$ are identical. We easily check that d_{RS} makes the simplex $\mathcal{S}$ of distributions over $\mathcal{D}$ into a metric space. Indeed, $d_{RS}(\cdot, \cdot)$ is obviously symmetric, and $d_{RS}(\boldsymbol{a}, \boldsymbol{b}) = 0$ implies that $\boldsymbol{a} = \boldsymbol{b}$. To check the triangular inequality, notice that

$$d_{RS}(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{1 - \Big(\sum_{i=1}^{s} \sqrt{a_i b_i}\Big)^2} = \sin\langle \sqrt{\boldsymbol{a}}, \sqrt{\boldsymbol{b}} \rangle, \tag{31}$$

where $\langle\sqrt{\boldsymbol{a}},\sqrt{\boldsymbol{b}}\,\rangle$ is the angle between the unit vectors $\sqrt{\boldsymbol{a}}$ and $\sqrt{\boldsymbol{b}}$, using the notation $\sqrt{\boldsymbol{v}} = (\sqrt{v_1},\ldots,\sqrt{v_s})$. To prove that $d_{RS}(\boldsymbol{a},\boldsymbol{b}) + d_{RS}(\boldsymbol{b},\boldsymbol{c}) \geq d_{RS}(\boldsymbol{a},\boldsymbol{c})$ for any $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c} \in \mathcal{S}$, we denote by α,β,γ the corresponding angles in that order, ie, $\alpha = \langle\sqrt{\boldsymbol{a}},\sqrt{\boldsymbol{b}}\,\rangle$, etc. The coordinates in $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}$ are nonnegative; therefore $0 \leq \alpha,\beta,\gamma \leq \pi/2$. These form the three angles at the origin of a tetrahedron with a vertex at the origin; therefore, by the triangular inequality in spherical geometry, $\alpha+\beta \geq \gamma$. If $\alpha+\beta \leq \frac{\pi}{2}$, then $\sin\alpha+\sin\beta \geq \sin\alpha\cos\beta+\cos\alpha\sin\beta = \sin(\alpha+\beta) \geq \sin\gamma$. On the other hand, if $\alpha+\beta > \pi/2$, then $\sin\alpha+\sin\beta = 2\sin\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2} \geq 2\sin\frac{\pi}{4}\cos\frac{\pi}{4} = 1 \geq \sin\gamma$, which establishes the triangular inequality.

Relation to the Euclidean distance

Shrinking the simplex $\mathcal{S}$ by a tiny amount, we define $\mathcal{S}_\varepsilon := \{\boldsymbol{a} \in \mathcal{S} : \varepsilon \leq a_i \leq 1-\varepsilon\}$ and note that

$$d_E(\boldsymbol{a},\boldsymbol{b}) := \|\boldsymbol{a}-\boldsymbol{b}\|_2 = \sqrt{\sum_{i=1}^{s}(\sqrt{a_i}-\sqrt{b_i})^2(\sqrt{a_i}+\sqrt{b_i})^2}.$$

It follows that, for $\boldsymbol{a},\boldsymbol{b} \in \mathcal{S}_\varepsilon$,

$$\frac{1}{2}d_E(\boldsymbol{a},\boldsymbol{b}) \leq d_E(\sqrt{\boldsymbol{a}},\sqrt{\boldsymbol{b}}\,) \leq \frac{1}{2\sqrt{\varepsilon}}\,d_E(\boldsymbol{a},\boldsymbol{b}). \tag{32}$$

On the other hand, $\|\sqrt{\boldsymbol{a}}\|_2 = \|\sqrt{\boldsymbol{b}}\|_2 = 1$, so the vectors $\sqrt{\boldsymbol{a}}$ and $\sqrt{\boldsymbol{b}}$ form an isosceles triangle; hence

$$d_E(\sqrt{\boldsymbol{a}},\sqrt{\boldsymbol{b}}\,) = 2\sin\frac{1}{2}\langle\sqrt{\boldsymbol{a}},\sqrt{\boldsymbol{b}}\rangle = \frac{\sin\langle\sqrt{\boldsymbol{a}},\sqrt{\boldsymbol{b}}\rangle}{\cos\frac{1}{2}\langle\sqrt{\boldsymbol{a}},\sqrt{\boldsymbol{b}}\rangle} = \frac{d_{RS}(\boldsymbol{a},\boldsymbol{b})}{\cos\frac{1}{2}\langle\sqrt{\boldsymbol{a}},\sqrt{\boldsymbol{b}}\rangle}\,.$$

Since $0 \leq \langle\sqrt{\boldsymbol{a}},\sqrt{\boldsymbol{b}}\,\rangle \leq \frac{\pi}{2}$,

$$d_{RS}(\boldsymbol{a},\boldsymbol{b}) \leq d_E(\sqrt{\boldsymbol{a}},\sqrt{\boldsymbol{b}}\,) \leq \sqrt{2}\,d_{RS}(\boldsymbol{a},\boldsymbol{b}).$$

Together with (32) this shows that, for any $\boldsymbol{a},\boldsymbol{b} \in \mathcal{S}_\varepsilon$,

$$\frac{1}{2\sqrt{2}}\,d_E(\boldsymbol{a},\boldsymbol{b}) \leq d_{RS}(\boldsymbol{a},\boldsymbol{b}) \leq \frac{1}{2\sqrt{\varepsilon}}\,d_E(\boldsymbol{a},\boldsymbol{b}), \tag{33}$$

which shows that the Euclidean distance and the metric d_{RS} are equivalent in $\mathcal{S}_\varepsilon$.

Relation to other distances

The metric d_{RS} is related to the Hellinger and Bhattacharyya distances. Writing $C(\boldsymbol{a},\boldsymbol{b}) = \sum_{i=1}^{s}\sqrt{a_i b_i}$ [4], then $d_{RS}(\boldsymbol{a},\boldsymbol{b}) = \sqrt{1-C(\boldsymbol{a},\boldsymbol{b})^2}$. The Hellinger distance is defined as $d_H(\boldsymbol{a},\boldsymbol{b}) = \sqrt{1-C(\boldsymbol{a},\boldsymbol{b})}$ [10], while the Bhattacharyya distance is defined as $d_B(\boldsymbol{a},\boldsymbol{b}) = -\ln C(\boldsymbol{a},\boldsymbol{b})$ [3]. The total variation distance d_{TV} is half the ℓ_1-norm; therefore $d_{TV}(\boldsymbol{a},\boldsymbol{b}) \leq \frac{1}{2}\sqrt{s}\,d_E(\boldsymbol{a},\boldsymbol{b})$. Combining these observations with (33) establishes (4):

$$\begin{cases} d_H = \sqrt{1-\sqrt{1-d_{RS}^2}}\,; \\ d_B = -\frac{1}{2}\ln(1-d_{RS}^2)\,; \\ d_{TV} \leq \sqrt{2s}\,d_{RS}. \end{cases}$$