UNIVERSITY OF CALIFORNIA,
IRVINE

A Nonlinear Approach To Learning From An Inconsistent Source
(with some applications)

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematics

by

Timmy Ma

Dissertation Committee:
Professor Natalia L. Komarova, Chair
Associate Professor German Enciso
Professor Patrick Guidotti

2018

Version 1.0.

Dedication

To my grandma,
Tinh Duong
who is proud of me,
always.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my appreciation to my thesis committee chair, Professor Natalia Komarova, who helped me investigate a beautiful area of mathematics I never knew I could.

I would also like to thank my committee members, Professor German Enciso and Professor Patrick Guidotti, for their interest in my research.

I would like to thank Springer Nature Terms and Conditions for RightsLink Permissions Springer Customer Service Centre GmbH for allowing the inclusion of marked portions of this paper. The text of chapter 1 is a reprint of the material as it appears in *Bulletin of Mathematical Biology*. The co-author listed in this publication directed and supervised research which forms the basis for the thesis/dissertation.

Finally, I would like to acknowledge the support of my family and the wonderful communities that I am proud to be a part of: the Mathletes and UCI Pokemon GO.

Curriculum Vitæ

Timmy Ma

Department of Mathematics, University of California, Irvine *sites.uci.edu/timmym/*
532 Rowland Hall *timmym@math.uci.edu*

Research Interests

- Applied Mathematics - stochastic processes, Markov chains, statistical learning, probability
- Complex Social Phenomena - linguistics, language learning

Education

- PhD Mathematics - University of California, Irvine May 2018
  - Thesis Advisor: Natalia Komarova, PhD
  - Dissertation: A Nonlinear Approach to Learning From an Inconsistent Source (with some applications)
- M.S. in Mathematics - University of California, Irvine Sept 2014
- B.A. in Mathematics - University of California, Berkeley May 2011
- A.S. in Mathematics - El Camino Community College May 2009

Papers and Preprints

- Ma, T. and Komarova, N.L., 2017. "Object-Label-Order in a noisy learning environment." Submitted to Cognition.
- Ma, T., Wood, K., Xu, D., Guidotti, P., Pantano, A., Komarova, N.L., 2017. "Diversity and Admission Predictors for Mathematics PhD Success." To be submitted to Notices of the AMS.
- Ma, T. and Komarova, N.L., 2017. "Mathematical Modeling of Learning from an Inconsistent Source: A Nonlinear Approach." Bulletin of mathematical biology, 79(3), pp.635-661.

Abstract of the Dissertation

A Nonlinear Approach To Learning From An Inconsistent Source

(with some applications)

By

Timmy Ma

Doctor of Philosophy in Mathematics

University of California, Irvine, 2018

Professor Natalia Komarova, Chair

Learning in natural environments is often characterized by a degree of inconsistency from an input. These inconsistencies occur e.g. when learning from more than one source, or when the presence of environmental noise distorts incoming information; as a result, the task faced by the learner becomes ambiguous. In this study we present a new interpretation of existing algorithms to model and investigate the process of a learner learning from an inconsistent source. Our model allows us to analyze and present a theoretical explanation of a frequency boosting property, whereby the learner surpasses the fluency of the source by increasing the frequency of the most common input. We then focus on two applications of our model. One is using our model to describe the "Object-Label-Order" effect. The other is to describe the evolution of the first word.

Introduction

> "Education's purpose is to replace an empty mind with an open one."
>
> *Malcolm Forbes*

0.1 Regularization

From the development of the Nicaraguan Sign Language [67], whereby younger cohorts of children in schools of Western Nicaragua in the 1970s and 1980s learned and improved upon the "home-signing" system of their older cohorts, to the creolization of pidgin languages [1], where languages that have basic and inconsistent foundation evolve to acquire a complex and consistent system, to the discussion of the plural allomorphs in English [8], the examples of language regularization by learners are abundant.

The ability of humans to improve on and modify the linguistic input they receive is a natural occurrence of language learning. A prime example of such a phenomenon is reported by Elissa Newport and colleagues [27]. They analyzed the language of a deaf boy (named Simon) who received all of his linguistic input from his parents, who were not fluent in American Sign Language (ASL) since they learned it after the age of 15. Due to the parents' late introduction to ASL, they served as an "inconsistent source" that Simon received as an ASL learner. Amazingly, comparing Simon to his parents and children who were native speakers of ASL, Newport reported that Simon greatly outperformed his parents when being tested on ASL, and in many aspects did as well as native speakers' children

of similar age.

Newport and her colleagues used the term *frequency boosting* to describe the margin of regularization of learning that is exhibited by learners upon the language of the source. Suppose that the source of linguistic input, such as a teacher (or parents in the case of Simon) is inconsistent. By "inconsistent", we mean that there is a probabilistic element of using a particular form of a certain rule. Such as two ways to pronounce "po-tay-toe" or "po-ta-toe", or two variations of gesturing the letter "J" in ASL. Frequency boosting is the ability of a language learner to increase the frequency usage of a particular form compared to the source. *Frequency matching* happens when the learner reproduces the same frequency of usage as the source. In [27] it is reported that (i) children are more prone to frequency boosting than adults, and (ii) adults can also frequency boost, depending on the structure of input.

0.2 Learning algorithms

Previously [34, 58] we proposed a class of simple reinforcement learning algorithms (see e.g. [42, 45, 75] for background on reinforcement learning) that can exhibit frequency boosting behavior. These algorithms are based on the classic Bush-Mosteller algorithm [11], which we explain here to motivate the subsequent development. Let us suppose that a learner is receiving input on the usage of two possible, alternative forms of a rule. In this process, the learner develops a certain propensity of using form 1 or form 2. We will denote that learner's probability to use form 1 as x, where $0 \leq x \leq 1$. The learner's state is thus uniquely characterized by the value of x, such that $x = 1$ means that form 1 has been completely

learned, and $x = 0$ corresponds to the learner never using form 1. Bush-Mosteller algorithm postulates that the state of the learner changes in a sequence of updates, each update following one bit of the source's input. If the source uses form 1, this will increase the propensity of the learner to use form 1: $x \to x + \Delta x$. It is further postulated that the size of the increment, Δx, is a linear function of the "novelty" or "surprise" effect: the farther the current state, x, is from 1, the larger the increment. Mathematically this is expressed as $\Delta x = a(1-x)$, where a is some constant.

Since its introduction, Bush-Mosteller type learning algorithms have been used in the literature to describe a wide variety of learning behaviors, both in biological and behavioral sciences [9, 10, 43]. Some of the context include modeling the response of phenological traits, such as timing of breeding, to climatic conditions [48]; voting behavior of individuals in a population [6, 23], conflict resolution in humans [22], and the role of innovations and change in human language [41]. In the context of experiments with human subjects, several authors have used simple reinforcement learning models to successfully explain and predict behavior in a wide range of experiments [18, 38, 39, 60].

A central assumption of the Bush-Mosteller algorithm, and subsequently, the influential Rescorla-Wagner algorithm [53, 54, 55], is the notion that learning happens more quickly if an element of surprise is present. This principle has been studied in the neurophysiological literature [21, 57, 63]). For example, [62] discovered that neurons "show reward activations only when the reward occurs unpredictably and fail to respond to well-predicted rewards, and their activity is depressed when the predicted reward fails to occur". Further empirical background for models of the Bush-Mosteller type is provided by [5, 13, 17, 26].

The class of algorithms presented in [34, 58] is consistent with the central assumption of the Bush-Mosteller and Rescorla-Wagner models, that the speed of learning is positively correlated with "novelty", but instead of assuming that the increment of learning is a linear function of the "novelty", it uses more general, and possibly nonlinear, non-decreasing functions. In [34, 58] we demonstrate the existence of the boosting property of the algorithm for several interesting cases. It however remains unclear under what circumstances can one expect to see the boosting property.

In chapter 1, we introduce an alternative and more elegant mathematical framework that allows us to model language regularization in the context of a teacher-learner pair. The advantage of the novel approach is that it allows us to develop and provide a mathematical proof of sufficient conditions in which frequency boosting can occur.

0.3 Learning with noise

One aspect of learning concerns inconsistencies, or "noise", in the input. Typically, input consists of exemplars that become a basis for our internal associations and concept formation. In natural situations, these exemplars are often contradictory. For instance, if learning from more than one teacher, such contradictions are inevitable. Another reason for inconsistencies can simply be the noise inherent in any information sharing exercise. How do we deal with internal contradictions in the source? To explore this issue experimentally and theoretically, in this paper we restrict ourselves with a comparatively simplistic learning environment, where a student is exposed to a series of images, each combined with a specific utterance. The input is intentionally contradictory, such that different utterances

may accompany the same image and different images can co-occur with the same utterance. The objective is to create a mental map of image-utterance associations.

0.4 Object-Label-Ordering effect

In the following study, we explore the possibility that the learner may have an easier time dealing with the source's inconsistency, depending on the way the input is presented. Namely, we vary the order in which the image and corresponding utterance are received (that is, whether an utterance is followed by an object or the other way around).

The order of presentation has been studied by other groups in the context of the so called Feature-Label-Order (FLO) effect, see e.g. [3, 30, 49]. For example, Ramscar and colleagues studied the difference between two types of symbolic learning procedures: the Feature-Label (FL) and the Label-Feature (LF) process [49]. Under the FL process, when learning images and utterances, images ("features") are shown first, and then the label is heard. Under the LF process, the order is reversed. To investigate whether FLO made a difference, a series of experiments were performed where each exemplar had one non-discriminating and one discriminating feature. A non-discriminating feature occurred in multiple categories, thus making the objects difficult to correlate to one particular category. Each discriminating feature only occurred with one category but at different frequencies. Using human subjects as well as computer simulations based on the Rescorla-Wagner type model, the authors found that FL learners were able to identify and discriminate both high and low frequency categories at a higher

accuracy than LF learners.

Inspired by these studies, we hypothesize that the order in which images and utterances are received by the learner may make a difference in how efficiently the learners handle inconsistencies of the source. There are two important differences between our concept and the FLO effect described above. In FLO studies, objects that appeared in the learning tasks possessed different "features". Some features were shared among two or more objects, thus making the task of label assignment difficult. In our study, the objects are "holistic" and do not share features. Instead, what makes the learners' task difficult/ambiguous in our study is the presence of noise, or inconsistencies, of the source. In order to avoid confusion between the two experimental and conceptual settings (the setting with shared features by [49], and the present setting with inconsistencies in the source), we refer to our framework as "Object-Label-Order" (OLO) effect.

In chapter 2, we investigate the interplay between the Object-Label-Order effect on the one hand, and regularization as a way of processing noisy, inconsistent source of learning, on the other. In [33, 34, 58], we developed a new mathematical approach to describe learning from an inconsistent source and, in particular, to explain the regularization phenomenon. Here, we extend our framework to study how learners process inconsistencies of the source in the OL and the LO learning environments. Does OLO make a difference in how the inconsistency is processed? What is a better way of setting up the learning task? Chapter 2 explores these questions both by theory and experiment.

Apart from regularization and noisy language learning, there are several other broad areas of application for this work. First is symbolic learning, where one ex-

amines how symbols are represented and used as tools of learning [49]. Another important area is categorization and concept learning, which is regarded as a process of determining how and which objects are grouped together or separately [4, 20, 77]. More details are presented in section 2.4.

0.5 Language evolution: Zipf's Law of Abbreviation

Language is a fundamental feature of human behavior. Its evolution has puzzled and caused controversy across many disciplines [15, 16, 29, 64]. Both real-life behavioral data and theoretical modeling have thus far provided the scientific community tools to approach language evolution. The emergence of the first word, however, has remained unexplored. In hindsight, the "first word" may have never occurred in one particular event but instead could be a result of a process that occurred over generations. How does the first word emerge in a population of social individuals? With the complexity of the new word, is there a more likelihood that the word will be forgotten? Can we say then that, from the evolution of more complex words, the "simpler" words would be used less frequently?

In chapter 3 we explore our theoretical modeling from chapter 1 and use regularization to describe the formation of the first word. In addition, we use language evolution to explore Zipf's Law of Abbreviation. This law states that frequently used words tend to be shorter [78, 79]. In [7], they studied a sample of 1263 texts written in 986 different languages of 80 different families show a negative correlation between word frequency and word length. However, with our theoretical model and the concepts of learning with noise, regularization, and symbolic learning, we explore a potential environment which will be in violation of Zipf's Law of Abbreviation.

Student-Learner-Pair

"Intelligence is the ability to adapt to change."

Stephen Hawking

In this chapter we introduce two algorithms, which exhibit frequency boosting presented e.g. by [27]. We first describe the algorithm that was developed in [58]. We discuss some of its basic properties and the role it takes in describing the frequency boosting property. The second algorithm is new; we describe the formulation and discuss the advantages it provides.

1.1 The Variable Increment algorithm

This section is to highlight the basic algorithm discussed in [58]. Let us suppose that a certain rule has 2 forms, or variants. A learner is characterized by 2 positive numbers, (x_1, x_2) such that $x_1 + x_2 = 1$ (a generalization to n forms is described in [58]). We will use the same terminology as in [58] for consistency. Note that x_1 uniquely defines the state of the learner, and the quantity x_1 (x_2) represents the probability of the learner to use form 1 (form 2).

The learning process is modeled as a sequence of steps, which are responses of the learner to the source's input. The changes in x_1 at each step are defined by (i) the input from the source, which will be described as a probability ν to utter form 1, and (ii) the state of the learner, through a given response function. If the source's input is form 1, then the learner will update its frequencies according to

the following update rules:

$$x_1 \to \begin{cases} x_1 + F^+(x_1), & \text{if form 1 is received,} \\ x_1 - F^-(x_1), & \text{if form 2 is received,} \end{cases} \tag{1.1}$$

where F^+ and F^- are some functions of one variable on $[0, 1]$. The rules above can be rewritten for the probability to utter form 2, $x_2 = 1 - x_1$:

$$x_2 \to \begin{cases} x_2 - F^+(x_2), & \text{if form 1 is received,} \\ x_2 + F^-(x_2), & \text{if form 2 is received.} \end{cases} \tag{1.2}$$

Because of symmetry considerations, we can argue that $F^-(x_1) = F^+(1 - x_1) \equiv F(x_1)$, such that the superscripts can be omitted. We will refer to the function F as the *update function*. Note that because $x_1, x_2 \in [0, 1]$, we must require that

$$F(x) \leq x. \tag{1.3}$$

Fig. 1.1(a) provides a schematic of this simple algorithm.

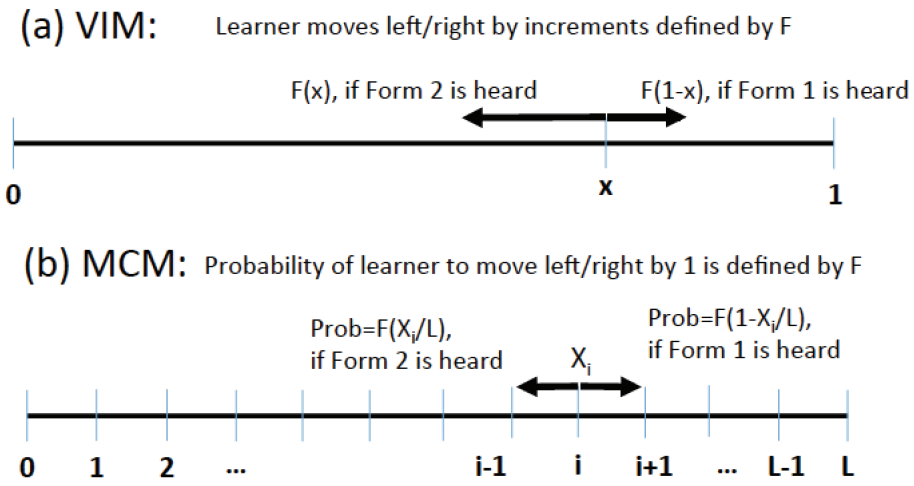

Figure 1.1: A schematic illustration of the learning algorithms in the case of $n = 2$ forms. (a) The VIM. (b) The MCM.

We will call this model the Variable Increment Method (VIM). While it captures the phenomena described in [27, 58], it has some shortcomings. (i) The

model has a non-constant increment. We can see in Fig. 1.1 that as the learner's propensity to use form 1 increases (x moves to the right), that the size of positive updates must become smaller (see requirement (1.3)). This puts a restriction on the types of functions that one can use in conjunction with this algorithm. Practically, for a given function $f(z)$, one can simply set

$$F(z) = \begin{cases} f(z), & f(z) \leq z, \\ z, & \text{otherwise.} \end{cases}$$

(ii) The second shortcoming of the VIM is its analytical intractability. In the next section we introduce a different algorithm where the theory of Markov Chains can be used to study its properties.

1.2 The Markov Chain algorithm

For this method, which we call the Markov Chain Method (MCM), we assume that the state of the learner is defined by positive integers, X_1 and X_2, that takes values $0, 1, \ldots, L$, with $X_1 + X_2 = L$. We will refer to parameter L as the *discretization parameter*. Then $X_i/L = x_i$ is the probability that the learner uses form i. At each time step, the learner's value of X_i may move up or down by 1 with the probability defined by the source's input and the learner's current position. This gives rise to a Markov chain with the following transition matrix, $P = \{P_{i,j}\}$:

$$P_{i,i+1} = \nu F(1 - i/L), \; P_{i,i-1} = (1 - \nu)F(i/L), \tag{1.4}$$

$$P_{i,i} = 1 - P_{i,i+1} - P_{i,i-1}, \; 0 < i < L, \tag{1.5}$$

$$P_{0,1} = \nu F(1), \quad P_{0,0} = 1 - \nu F(1), \tag{1.6}$$

$$P_{L,L-1} = (1 - \nu)F(1), \quad P_{L,L} = 1 - (1 - \nu)F(1), \tag{1.7}$$

with the rest of the elements being zero. Here, the function F is defined on $[0, 1]$ and has the meaning of probability. Although the value $F(0)$ does not appear

in the transition matrix, we formally assume that $F(0) = 0$, because the factor $F(0)$ multiplies the probability to reduce X upon receiving form 2 from the input, when $X = 0$, and this quantity should be zero. A schematic of this method is presented in Fig. 1.1(b).

The MCM formulation does not require function F to satisfy condition (1.3). Further, as will be demonstrated below, this Markov chain formulation allows for easy proofs of some of the algorithm's important properties.

In order to satisfy the requirements of the psychologically and physiologically based Bush-Mosteller and Rescorla-Wagner models, the function F used in our formulation(s) must satisfy certain requirements. We will require that (i) $F(0) = 0$, and (ii) $F(x)$ is nondecreasing.

To compare the two algorithms, suppose that function F satisfies inequality (1.3), and calculate the expected increments in the state of the learner, given that the frequency of the source is ν. Suppose further that the current state is given by x_1 for VIM and $X_1 = Lx_1$ for MCM. It is easy to check that the expected increment both in VIM and MCM is given by

$$\Delta = \nu F(1 - x_1) - (1 - \nu)F(x_1).$$

Note however that while in VIM, $x_1 \to x_1 + \Delta$, and this increment translates directly to the increment in the learner's probability to utter form 1, in MCM we have $X_1 \to X_1 + \Delta$, and thus the corresponding increment in the learner's probability to utter form 1 is given by Δ/L. To make the two methods commensurate, one has to scale the function F used in MCM by a factor of L.

In section A.1 we present a detailed numerical investigation of the two methods' behavior, including convergence and applications to the setting described in [58]. To reproduce the results in [27], we had to introduce two parameters in VIM to capture an asymmetric aspect in the model as dicussed in [58]. We do the same for MCM to complete our comparison of the two methods. We call the asymmetric versions of VIM (and MCM) the AVIM (and the AMCM). A detailed comparison of AVIM and AMCM is presented in section A.1.2. In summary, the two algorithms coincide as the discretization parameter, L, of MCM (AMCM) increases to infinity.

1.2.1 Generalization of the MCM algorithm to N forms

For the general N form update, suppose that the state of the learner is given by

$$(X_1, X_2, \ldots, X_N),$$

where X_i is an integer from 0 to L for $i = 1, \ldots, N$ and that $\sum_i^N X_i = L$ for an integer L. Suppose that the source utters form i, then for each $j \neq i$, the following update occurs: with probability $F(X_j/L)$,

$$X_j \to X_j - 1, \quad X_i \to X_i + 1, \tag{1.8}$$

For example, when $N = 3$, suppose that the source utters form 1, then each X_2 and X_3 have a chance to update (negatively), matched with positive update in X_1. If X_2 decreases (with probability $F(X_2/L)$), then X_1 increases. Similarly, if X_3 decreases (which happens with probability $F(X_3/L)$), then X_1 increases.

1.3 The Frequency Boosting Effect

In [34], an observation was made about "frequency boosting" for power functions. Here, we will discuss this effect in more generality. In particular, we will describe

a class of functions that exhibit boosting behavior. Let us remind ourselves of the conditions that must be met to satisfy the Bush-Mostellar and Rescorla-Wagner models: (i) $F(0) = 0$, and (ii) $F(x)$ is nondecreasing. Examples of functions that satisfy these conditions are linear functions $F(x) = ax$ or power functions such as $F(x) = ax^{\alpha}$, where α is any real number. The theory presented below is however more general and includes a very large family of functions.

1.3.1 The Frequency Boosting Theorem

Frequency boosting is the event where the frequency of the learner to use the most frequent form surpasses that of the source. If we denote by ν the frequency of the source to use its most frequent from, and by ν_{learn} the frequency of the learner to use the same form, frequency boosting is simply $\nu_{learn} > \nu$.

Fig. 1.2 gives an example of this behavior from MCM with $F(X/L) = (X/L)^{0.4}$ and $L = 100$. The source in this example will use form 1 60% of the time ($\nu = 0.6$) and form 2 40% of the time. As the learner is being exposed to the source's usage of forms 1 and 2, over time, the learner's frequency of usage of form 1 increases from its initial value and even surpasses 60%. Eventually, the learner will reach a quasi-steady state and oscillate close to 70%. This is the frequency boosting behavior, where the learner decreases the source's inconsistency by increasing the usage of the more frequent form.

Formally, for the two-form example, if ν and ν_{learn} denote the frequency of form 1 for the source and the learner respectively, then the frequency boosting

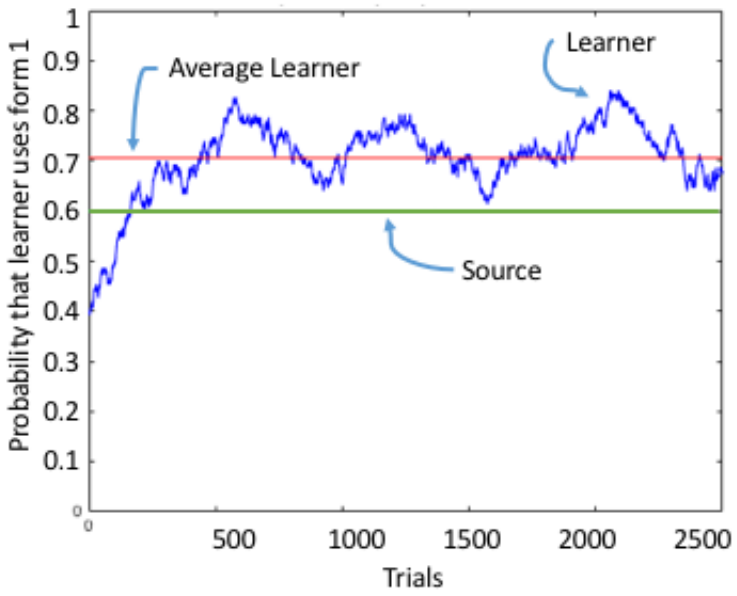

Figure 1.2: Stochastic simulation with MCM illustrating the Frequency Boosting Property. Two forms are used by the source, with $\nu = 0.6$. The frequency of the learner is plotted as a function of time (updates). The resulting quasi-steady state is characterized by the average frequency that exceeds the frequency of the source for the more frequent form.

property is equivalent to

$$\nu_{learn} \begin{cases} > \nu, & \text{if } \nu > \frac{1}{2}, \\ < \nu, & \text{if } \nu < \frac{1}{2}, \\ = \nu, & \text{if } \nu = \frac{1}{2}. \end{cases}$$

Fig. 1.3 shows two examples, one of the boosting property and the other as a non-example. Fig. 1.3(a) is the plot of MCM with $F(X/L) = (X/L)^{0.2}$, and it shows that as long as $\nu > 0.5$, the resulting frequency of the learner, ν_{learn}, exceeds, ν. Thus we can see that the given function exhibits boosting property. In Fig. 1.3(b), we used a different function, $F(X/L) = (X/L)^{1.5}$. It can be seen that in this case, there is no frequency boosting property, and an opposite trend

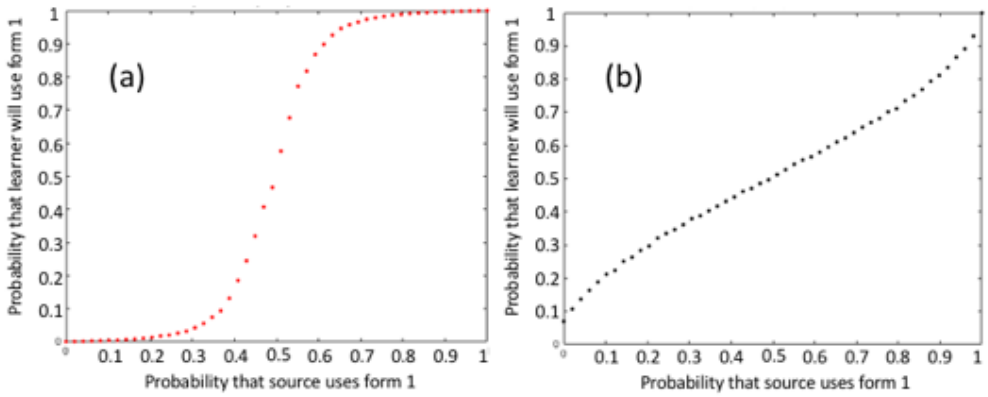

Figure 1.3: Plot of MCM with functions (a) $F(X/L) = (X/L)^{0.2}$ and (b) $F(X/L) = (X/L)^{1.5}$, with $L = 100$. Average probability of the learner to use form 1 is plotted against the probability of the source to use form 1.

seems to take place (we refer to this effect as *frequency demotion* or *under-matching*).

An example of *frequency matching* is when the learner's frequency of using form 1 matches that of the source. A function that produces this result is $F(X/L) = aX/L$, a linear function, which is shown in Fig. 1.4.

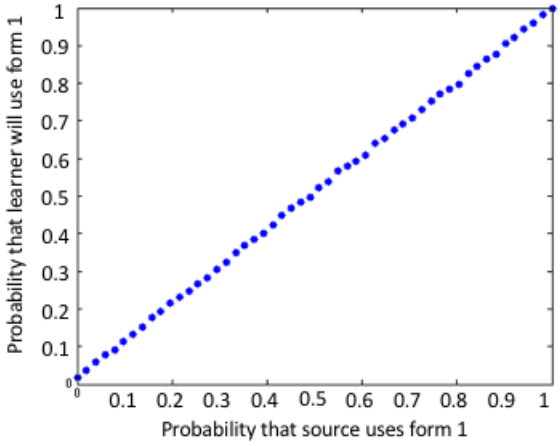

Figure 1.4: Algorithm MCM with function $F(X/L) = X/L$; $L = 100$.

After hypothesizing various nonlinear functions, we conjectured that a sufficient types of function that exhibits frequency boosting are nonlinear functions

that are concave down, that is, at least some of secant lines are below the graph of F. In other words, F is concave down on interval I if for all $x, y, z \in I$ such that $x < z < y$,

$$\frac{F(z) - F(x)}{z - x} \geq \frac{F(y) - F(x)}{y - x},$$

and the inequality is strict for at least some points.

We now formalize this in the following theorem:

Frequency Boosting Theorem. *Suppose that the source uses two forms, and* $\nu > \frac{1}{2}$ *is the frequency of the source to use form* 1. *Further, suppose that the learner operates according to MCM with an update function* F *and discretization parameter* L. *Denote by* ν_{learn} *the limiting average frequency of the learner to use form* 1 *(corresponding to the number of updates increasing to infinity). Then the learner will exhibit the boosting property, that is,* $\nu_{learn} > \nu$, *if the function* F *has the following properties on a set of discrete points* $\{0,\ 1/L,\ 2/L, \ldots,\ 1\}$ *for some integer* L*:*

1. $F(0) = 0$,
2. *F is non-decreasing,*
3. $F \leq 1$,
4. *F is concave down in the sense explained above.*

The proof of the theorem is presented in A.2.

1.3.2 The frequency of the learner – an analytical expression

At each time step, the MCM learner can move up or down the Markov walk by 1 with a probability defined by the source's input and the learner's current position, according to the transition matrix P defined by (1.4-1.7). Stochastic matrix

P has eigenvalues $1 = \lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_N$. The steady state of the learner is characterized by the left eigenvector, $Y = (Y_0, \ldots, Y_L)$, of the transition matrix P that corresponds to the unit eigenvalue, $\lambda_0 = 1$:

$$YP = Y.$$

The eigenvector Y can be found explicitly, for a general function F. Let us define the notation

$$C_n^k = \frac{\prod_{i=1}^{n} F(i/L)}{\prod_{i=1}^{k} F(i/L) \prod_{i=1}^{n-k} F(i/L)}. \tag{1.9}$$

We also formally define $\prod_{i=1}^{0} F(i/L) = 1$ (in analogy with the definition of 0!). Note that in the case where $F(j/L) = aj/L$, we simply have

$$C_n^k = \left(\begin{array}{c} n \\ k \end{array} \right),$$

so one can think of the quantities C_n^k as a generalization of the binomial coefficients to the nonlinear functions F. The entries of the eigenvector Y are given by

$$Y_i = C_L^i \nu^i (1-\nu)^{L-i}, \quad 0 \leq i \leq L,$$

and the expected frequency of the learner is simply

$$\nu_{learn} = \frac{\sum_{i=0}^{L} Y_i i}{L \sum_{i=0}^{L} Y_i}. \tag{1.10}$$

The speed of convergence to the steady state is defined by the quantity $1 - \lambda_1$, where λ_1 is the second largest eigenvalue of P.

1.3.3 Power functions and other examples

There are several important examples that we will use to illustrate the above theory.

Simon model. This example comes from [34, 58] and corresponds to $F(i/L) = a$, a constant. In this case,

$$C_n^k = 1, \quad Y_i = \nu^i (1-\nu)^{L-i},$$

and

$$\nu_{learn} = \frac{\nu}{L} \frac{(1-\nu)^{L+1} + \nu^L (2\nu L - L + \nu - 1)}{(2\nu - 1)(\nu^{L+1} - (1-\nu)^{L+1})}, \tag{1.11}$$

which is the same as obtained from the VIM in [34]. Certainly $\lambda_0 = 1$ is an eigenvalue for the transition matrix P. It is possible to show that the next largest eigenvalue, λ_1, has the property

$$\lim_{L\to\infty} (1 - \lambda_1) = a(1 - 2\sqrt{\nu(1-\nu)}).$$

This quantity is related to the speed of convergence of the learning algorithm. It follows that for a given value ν, the speed of convergence is defined by the quantity a, the increment of the learner. This result coincides with the one reported in [34] for the VIM.

It is also instructive to find the limit of expression (1.11) as $L \to \infty$. If we denote

$$\kappa = \frac{1-\nu}{\nu}, \tag{1.12}$$

the expression in (1.11) can be rewritten as

$$\nu_{learn} = \frac{\frac{\nu\kappa^{L+1}}{L} + (2\nu - 1) - \frac{1-\nu}{L}}{(2\nu - 1)(1 - \kappa^{L+1})}.$$

If $\nu < \frac{1}{2}$, then $\kappa > 1$ and $\lim_{L\to\infty} \kappa^{L+1} = \infty$. In this case, the expression above behaves as $\nu/L/(1-2\nu) \to 0$ as $L \to \infty$. On the other hand, if $\nu > \frac{1}{2}$, then $\kappa < 1$ and $\lim_{L\to\infty} \kappa^{L+1} = 0$, such that the expression for ν_{learn} tends to 1. Therefore we have

$$\lim_{L\to\infty} \nu_{learn} = \begin{cases} 0, & \nu < \frac{1}{2}, \\ 1, & \nu > \frac{1}{2}. \end{cases} \tag{1.13}$$

The power function model. Let us suppose that $F(x) = ax^\alpha$, with $\alpha \geq 0$. In this case, we have

$$C_n^k = \left(\frac{n!}{k!(n-k)!} \right)^\alpha = \left(\begin{array}{c} n \\ k \end{array} \right)^\alpha .$$

Implementing formula (1.10), we obtain

$$\nu_{learn} = L^{\alpha-1} \frac{\nu}{1-\nu} \frac{{}_\alpha F_{\alpha-1}(1-L, \ldots, 1-L; 2, \ldots, 2; -\frac{\nu}{1-\nu})}{{}_\alpha F_{\alpha-1}(-L, \ldots, -L; 1, \ldots, 1; -\frac{\nu}{1-\nu})}. \qquad (1.14)$$

In the particular case where $\alpha = 1$, the expressions simplify and we obtain from (1.10),

$$\nu_{learn} = \nu.$$

In other words, the linear function F leads to frequency matching behavior, as was already illustrated by example of Fig. 1.4. The speed of convergence can be found for the linear functions. The eigenvalues of the stochastic matrix are given by

$$\lambda_i = 1 - ai/L, \quad 0 \leq i \leq L,$$

and the speed of convergence is then given by $1 - \lambda_1 = a/L$.

Frequency matching solutions. We would like to ask the question: apart from the linear function, what functions, $F(i/L)$, $0 \leq i \leq L$, lead to the frequency matching behavior? Using expression (1.10), we can consider the equation

$$\nu_{learn} - \nu = 0.$$

This equation has to be satisfied for all values of $\nu \in [0, 1]$, and is equivalent to an algebraic system of L equations for $F(1/L), F(2/L), \ldots, F(L/L)$. These equations are obtained by equating the coefficients of the powers $1, 2, \ldots, L$ of ν to zero.

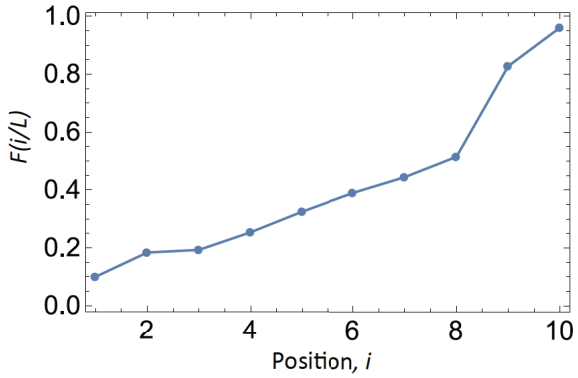

Figure 1.5: An example of function $F(i/L)$ that satisfies equation (1.16) and thus yields $\nu_{learn} = \nu$ for all ν. This function was created by taking $F(i/L) = F((i-1)/L) + \beta_i \xi_i$, for $0 \le i \le L/2$, and applying equation (1.15). Here $L = 10$, $\beta_i = F\left(\frac{i-1}{L}\right)\left(\frac{i(L+2-i)}{(L-i+1)(i-1)} - 1\right)$, and ξ_i is a uniformly distributed random number in $[0, 1]$. The choice of β_i guarantees that the function F is nondecreasing.

They can be solved iteratively, starting from the equation for the first power. The solution reads:

$$F\left(\frac{L+1-k}{L}\right) = F\left(\frac{k}{L}\right)\frac{L+1-k}{k}, \quad k \in \mathbb{Z}, \quad 1 \le k \le \frac{L+1}{2}. \tag{1.15}$$

If L is odd, the equation for $F\left(\frac{L+1}{2L}\right)$ turns into an identity. The above equation can be rewritten as

$$kF\left(\frac{L+1-k}{L}\right) = (L+1-k)F\left(\frac{k}{L}\right). \tag{1.16}$$

This equation defines an infinite family of sequences $\{F(k/L)\}$ that has the frequency matching property. One possible solution is, of course, the linear function, $F(X) = aX$. Fig. 1.5 shows a different example of a function F that yields frequency matching behavior.

Let us denote $X = k/L$. Then, in the limit $L \to \infty$, equation (1.16) becomes

$$XF(1-X) = (1-X)F(X). \tag{1.17}$$

We will encounter this limiting case in the next section, when talking about alternative methods of finding the frequency of the learner.

1.3.4 The mean approximation

While for the examples presented above it was possible to use formula (1.10) directly, the expressions do not simplify easily for a wider class of functions $F(X)$. In this case we can use an alternative method. At the equilibrium state, X, we assume that the probability of a positive increment exactly balances the probability of the negative increment. This leads to the following simple equation:

$$\nu F(1-X) = (1-\nu)F(X). \tag{1.18}$$

In particular, for frequency matching, where $X = \nu$, the function $F(X)$ has to satisfy

$$\nu F(1-\nu) = (1-\nu)F(\nu), \tag{1.19}$$

which is the same as equation (1.17) obtained as a limit of the discrete Markov process as $L \to \infty$. It is clear that the function $F(X) = aX$ satisfies this equation. As was mentioned above, this is only one of many functions that yield frequency matching.

Suppose that $F(X) = aX^\alpha$. We will focus on values $\alpha \geq 0$ because this function has to be nondecreasing and bounded on $[0, 1]$. First of all we note that equation (1.19) restricted to power functions becomes

$$\left(\frac{1-\nu}{\nu}\right)^{\alpha-1} = 1,$$

which holds for all $0 < \nu < 1-\nu$ only if $\alpha = 1$. Therefore, among the power functions, the linear function is the only one that leads to frequency matching.

Further, we can find the equilibrium value $\nu_{learn} = X$ by solving equation (1.18). We obtain

$$\nu_{learn} = \frac{1}{1 + \left(\frac{1-\nu}{\nu}\right)^{1/\alpha}}. \tag{1.20}$$

Manipulating this expression we can see that

$$\frac{1-\nu_{learn}}{\nu_{learn}} = \left(\frac{1-\nu}{\nu}\right)^{1/\alpha}, \text{ or } \kappa_{learn} = \kappa^{1/\alpha},$$

see definition (1.12). Let us first assume that $\nu > \frac{1}{2}$, $\kappa < 1$. In this case, $\kappa^{1/\alpha} > \kappa$ if $\alpha > 1$ and $\kappa^{1/\alpha} < \kappa$ if $\alpha < 1$. This is equivalent to the statement that $\nu_{learn} < \nu$ if $\alpha > 1$ and $\nu_{learn} > \nu$ if $\alpha < 1$. Repeating this argument for the case of $\nu < \frac{1}{2}$, we obtain the following result. For the power functions $F(X) = aX^{\alpha}$,

- $0 < \alpha < 1$ corresponds to frequency boosting,
- $\alpha = 1$ to frequency matching, and
- $\alpha > 1$ to frequency "demotion".

Of course, the first of these results is a direct consequence of the Frequency Boosting Theorem.

Finally, we would like to investigate the case $\alpha = 0$. Taking the limit $\alpha \to 0$ in expression (1.20), we obtain exactly the same result as predicted from the exact model with $L + 1$ discrete states, formula (1.13).

1.4 Discussion

In this chapter we presented an alternate and simpler version of the mathematical model that was able to capture several findings of Hudson Kam & Newport in their work with adult and children's learning [27]. From their discussion on frequency boosting, it was possible for us to examine the properties of the phenomena in [58] with an algorithm, which in the present chapter is called VIM. With the formulation of the novel algorithm, MCM, we can also provide a rigorous and general mathematical framework of frequency boosting to supplement

this discussion.

The main results obtained from this chapter can be formulated as follows:

- The novel stochastic nonlinear reinforcement learning algorithm is formulated. Its advantage is analytical tractability, and the fact that it coincides with the old algorithm in the limit where the discretization of the grid becomes more refined.

- With the new algorithm, the frequency of the learner is calculated analytically. The expressions are quite simple in several important cases, such as the linear (Bush-Mosteller) algorithm, "Simon" algorithm [34, 58], and power law algorithm.

- The new formulation also allows for a straightforward evaluation of the speed of convergence.

- A set of sufficient conditions was formulated, which defines update functions that exhibit frequency boosting property: the function is required to be non-decreasing and concave down.

Situations for which our theory is applicable provide a certain idealization of the real problem of language learning. More precisely, we only concentrate on a relatively small sub-task of language learning, while ignoring a set of other aspects of this grandiose task. For example, we assume that a learner's input is a string, and the learner is able to extract (segment), from all the utterances received, the correct mutually exclusive forms of the rule under investigation. This in itself is a challenge studied in the literature, see e.g. [37, 61, 66]. We will not address this pre-processing step in the present manuscript.

By following this reductionist approach, we try to shed light on one aspect of language learning; a similar philosophy was used by others in the literature, see e.g. [24, 32, 44, 46, 72]). It is therefore important to realize that our modeling approach may be constructive in certain circumstances, while being insufficient in others. For example, our modeling approach has been useful in describing the experiments of [19] and [27] (see [58]). In the next chapter, we can further apply our model to the ideas of [49], providing a possible explanation of the feature-label order effect in the given type of setting.

On the other hand, other circumstances require more complex models compared to the ones studied here. In those cases, our models have to be amended and expanded, to handle more complex experimental settings. In particular, here we will mention the phenomenon of implicit negative evidence that was studied in particular in [49]. In this paper it was demonstrated that negative learning and cue competition are necessary to explain the learning process under a range of realistic conditions. To this end, two novel experiments were described. One was presenting a group of adult learners with a set of objects to test how the frequency of objects can affect the discrimination of features. The other study was to consider the feature-label phenomena associating with children's word learning, particularly with learning color words. Further study of the feature-label order effect and way to model it with MCM is subject of the next chapter.

Object-Label-Order Effect

> "It doesn't matter how beautiful your theory is…If it doesn't agree with experiment, it's wrong."
>
> *Ricard P. Feynman*

In this chapter, we will briefly begin with the Markov Chain Method (MCM) algorithm and the frequency boosting effect as it was formulated in the last chapter. Then we will describe the concept of object-object and label-label interactions and the related phenomenon of negative evidence, which will allow us to adapt the MCM model to describe learning from an inconsistent source.

As mentioned before, in section 0.4 , in contrast to [49] and other papers by Ramscar's group, we will not be distinguishing between individual features of the objects of learning. In our setup, the objects are treated "holistically" and they do not share features. Therefore, to avoid confusion, we call our learners "object-label" and "label-object" learners.

2.1 Theory

This section will review the MCM model from section 1.2 and describe a new novel approach to explain negative feedback.

2.1.1 MCM framework

The MCM model gives a mathematical framework to describe the interaction of a teacher (a source) and a learner. It is a nonlinear generalization of the Mosteller-

Bush and Rescorla-Wagner models [12, 56], and in particular, allows one to describe the frequency boosting effect.

If the source of information uses rule variant 1, the learner updates $X \to X+1$ with probability $1 - F(X)$, where we use the updating function

$$F(X/L) = c(X/L)^r,$$

where $c > 0$ and $r \geq 0$ are some constants. If the source uses rule variant 2, then the learner updates $X \to X - 1$ with probability $F(X/L)$. After a number of updates, the learner converges to a certain statistical equilibrium, where the mean learner frequency characterizes the learned behavior. It has been proven in section 1.3.4 that this model exhibits different properties depending on the parameter r:

- In the special case where $r = 1$, we have the usual Bush Mosteller algorithm, and the limiting mean frequency of the learner is exactly equal to the frequency of the source. This is the phenomenon of frequency matching.
- If $r < 1$, the phenomenon of frequency boosting (or regularization) takes place where the limiting frequency of the learner is larger than ν in the case where $\nu > \frac{1}{2}$, and it is smaller than ν if $\nu < \frac{1}{2}$. In other words, the learner uses the preferred form of the source *m*ore frequently than the source.
- Finally, if $r > 1$, we have the phenomenon of "undermatching", where the frequency of usage of both forms becomes closer to $\frac{1}{2}$ (in other words, differences in usage between more frequent and less frequent variants diminish).

There is a key difference between our MCM model and, for example, the Mosteller-Bush model. The MCM can be viewed as a generalization of the Mosteller-Bush, which includes a (possibly nonlinear) update function, F. We hypothesize

that different types of learners (such as OL and LO learners) are characterized by differences in their update function. If we can identify the differences in function F using our model, then we can explain the contrasting behaviors of the two learners. These key parameters responsible for the differences in behavior can be identified by fitting our model to the empirical data that we collect.

2.1.2 Negative evidence: the matrix model

In the very simple model described above, only one rule is being learned at a time. It is not unreasonable to assume however that when several rules are learned simultaneously, there may be a certain degree of interaction among them. For example, consider two labels, Lab1 and Lab2, "competing" to be the primary label for a given object (say, Obj1). This is an example of the one-to-many learning process [20]. Suppose that Lab1 is used exclusively with Obj1 (X instances of usage), and Lab2 is used with the same object but also with another object, Obj2, Y_1 and Y_2 times respectively. Let us further assume that $Y_1 > X$. The input thus described is summarized in the following table:

	Obj1	Obj2
Lab1	X	0
Lab2	Y_1	Y_2

If we only focus on choosing the label for Obj1 and ignore associations with Obj2, then it would seem that Lab2 is a better candidate for Obj1's label. If however we consider possible interactions between Obj1 and Obj2, we will immediately see that Lab2 is shared between the objects, and Lab1 is reserved only for Obj1. The fact that Lab2 co-occurs with Obj2 will weaken its association with Obj1, which could actually make Lab1 a better choice as the label of Obj1. Note that label-label interactions can be set up in a similar manner, when learning for

the best object described by a given label.

This type of object-object and label-label interaction brings us to the concept of implicit negative evidence in learning . In [51] the importance of implicit negative evidence is illustrated by using the example of plural noun learning. A similar idea is explored in [50], where the concept of "informativity" in learning was studied. In this section we present a very simple method to incorporate implicit negative evidence in a learning model. Later on (in section 2.1.3), we create a more sophisticated, stochastic model of learning which uses similar ideas.

In the experiment described in [50], three different objects (A, B, and C) were used. Objects A and B appeared simultaneously, accompanied by the word "dax", then objects B and C appeared simultaneously, accompanied by the word "pid". These are examples of the many-to-one learning process [20]. This was repeated a number of times (N times). This table describes the input:

	Objects		
	A	B	C
dax	N	N	0
pid	0	N	N

The subjects were asked to find the object that was described by words "dax" and "pid", along with a new word, "wug". The experiments were performed separately with a group of adults and with a group of children. While both children and adults mostly agreed that "dax" was A and that "pid" was C, the two groups showed a different response for "wug" : the majority of adults picked object "B", but about half of the children picked A and the other half, C. To explain this, the

authors evoked the concept of informativity.

Indeed, object B was present every time "dax" or "pid" were uttered. This suggests that B is some background object that is not informative for "dax" or "pid". On the other hand, A and C are informative about "dax" and "pid" respectively (these objects only appeared together with those words). Therefore, most subjects paired A with "dax" and C with "pid". Then, the children extrapolated the non-informativity of B and therefore chose one of the two informative objects, A or C, to describe a new word, "wug".[1] In what follows, we describe one simple way to implement the concept of negative evidence.

First we will model the simpler part of the experiment, where the participants were asked to identify objects that are called "dax" and "pid". In principle, "dax" could be either A or B, and "pid" could be either B or C. We need to make the objects "compete" based on their informativity. Therefore, we normalize the input matrix column by column (that is, we divide each element by the sum of the elements in its column):

	Objects		
	A	B	C
dax	1	$\frac{1}{2}$	0
pid	0	$\frac{1}{2}$	1

Since object B appears more often than objects A and C, the entries corresponding to B in the above matrix are lowered as a result of the normalization. We can see that if A and B compete to be the object for "dax", A should have an advantage, because it has exclusively appeared with "dax", while B appeared both

[1] Interestingly, adults used a different algorithm. They used the fact that objects A and C were already taken by the two words, and therefore they assigned B to mean "wug", which was interpreted as a "logical" component of adult learning.

with "dax" and with "pid". Therefore, to choose the best candidate for "dax", we simply find the largest element in the appropriate row, which corresponds to A. Similar procedure is performed for "pid". We obtain the following result, which coincides with the subject's responses:

	Objects		
	A	B	C
dax	*		
pid			*

Next, we extend the model to predict the response to the question about an unknown word "wug". We expand the association matrix to include this word, and insert small nonzero elements instead of zeros:

	Objects		
	A	B	C
dax	N	N	ϵ_1
pid	ϵ_2	N	N
wug	ϵ_3	ϵ_4	ϵ_5

The values ϵ_i stand for "noise". A subject may have a shade of doubt whether a certain word has been associated with an object. Now, we follow the same procedure as before, that is, normalize column by column, and then pick the largest entries in each row:

	A	B	C
dax	N	N	ϵ_1
pid	ϵ_2	N	N
wug	ϵ_3	ϵ_4	ϵ_5

$\Rightarrow$

	A	B	C
dax	1	$\frac{1}{2}$	$\frac{\epsilon_1}{N}$
pid	$\frac{\epsilon_2}{N}$	$\frac{1}{2}$	1
wug	$\frac{\epsilon_3}{N}$	$\frac{\epsilon_4}{2N}$	$\frac{\epsilon_5}{N}$

$\Rightarrow$

	A	B	C
dax	$*$		
pid			$*$
wug	$*$		$*$

Note that in the middle matrix obtained by normalizing the columns, we assumed that $\epsilon_i \ll 1$ (and $\epsilon_i/\epsilon_j \sim 1$, that is, the values ϵ_i are of the same order of

magnitude), and only kept the highest order terms in all the expressions. In the rightmost matrix, if the values $\epsilon_3 \approx \epsilon_4 \approx \epsilon_5$, then the first or the last entry of the third row could be the winner. Therefore, the "wug" row contains two stars. We can see that the result is consistent with the response of children. More examples demonstrating the versatility of the matrix method are presented in section A.3.

The simple method demonstrated here is an example of how interaction between different "rules" can be quantified. The procedure used here is a way to implement implicit negative evidence in learning. A drawback of this approach is that it belongs to the well studied, but cognitively unrealistic "batch learner" models, in that it operates on the whole body of the input information received by the learner (and thus theoretically, it possesses an infinite memory capacity) [46]. In a more realistic scenario explored in the next section, information is received gradually, and the learner's internal representation of reality is modified accordingly.

2.1.3 Stochastic model of learning with negative evidence

To create a model that studies the OLO effect in learning from an inconsistent source, we combine the concept of stochastic learning with the mechanisms of implicit negative evidence. To describe the algorithm, we will use a three-object, three-label system. We assume that a learner stores object-label associations in the form of two 3×3 matrices with nonnegative integer entries, that we call P and Q; as before, the rows correspond to labels and columns to objects. The matrix P

is the "object matrix",

$$P = \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{pmatrix}, \tag{2.1}$$

where we assume that the columns sum up to L. This matrix characterizes the learner's answer to "what is object?" questions. For example, the probability for the learner to answer "label 1" to "what is object 1?" question is given by α_1/L. The fact that the entries in each column sum up to a fixed number represents the competition of different labels to represent a given object. The matrix Q is the "label matrix",

$$Q = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix}, \tag{2.2}$$

and its rows sum up to L, which represents the competition of different objects to associate with a given label. This matrix characterizes the learner's answer to "what is label?" questions, for instance, the probability for the learner to answer "object 1" to "what is label 1?" question is given by x_1/L. In our framework the two matrices are kept and developed separately; this resembles an earlier model used by [28] in the evolutionary theory of mutual indelibility of linguistic agents.

The learning procedure is set up as follows. The source produces object-label pairs, and the learner updates entries of its P and Q matrices. In the absence of negative evidence, the evolution of the columns of matrix P and the rows of matrix Q is completely described by the MCM algorithm. For example, after hearing an "Obj1 – Lab2" pair, the learner will be updating the 1st column of the P matrix and the 2nd row of the Q matrix.

In the presence of negative evidence, interactions among columns of matrix P and rows of matrix Q must be taken into account. We will illustrate the procedure of negative update by using the columns of matrix P; the procedure is similar with the rows of matrix Q. In the case where a positive update took place, we assume that negative updates occur with probability μ, where $0 \leq \mu \leq 1$. Once the learner determines that the update will occur, then the learner will decide which of the remaining columns of matrix P will be updated negatively. Since the learner updated positively on β_1, then the learner will negatively update column i with probability:

$$P_i = \frac{\beta_i^{-1}}{\beta_2^{-1} + \beta_3^{-1}}, \quad i \in \{2, 3\},$$

that is, the probability of negative update is larger for weaker entries corresponding to the label in question. Note that if either β_2 or β_3 are zero, then the learner will choose the non-zero β_i to update. If they are both zero, then no negative update occurs. Suppose that the learner chose to negatively update column 2, that is $[\alpha_2, \beta_2, \gamma_2]$. The learner will then perform the following update:

$$\beta_2 \to \beta_2 - 1 \quad \text{with probability } F(\beta_2/L). \tag{2.3}$$

Finally, to keep the column elements on the simplex, the learner performs

$$\begin{aligned} \alpha_2 &\to \alpha_2 + 1 \quad \text{with probability } \tfrac{\alpha_2}{\alpha_2+\gamma_2}, \\ \gamma_2 &\to \gamma_2 + 1 \qquad\qquad \text{otherwise.} \end{aligned}$$

Note that both positive and negative updates tend to strengthen stronger associations and to weaken weaker associations.

2.2 Experiments

In the experiment described below we studied OL and LO learning in the presence of inconsistencies of the source, in a simple setting of two label/two object

system plus a control. We explored the differences in the learning results of OL and LO learners in the context of both "what is label?" and "what is object?" type questions.

2.2.1 Participants

Two-hundred-fifty undergraduate students from a Southern California public university participated. After filtering the results by using the criterium of correctly identifying the control object/label pair (see below), we were left with one-hundred-fifty-three learners.

2.2.2 Set up and procedures

Three fictional "animals" or objects were constructed for training, see Fig. 2.1. Each of these objects had a corresponding fictional name ("yosh", "wug", and "niz"). None of these objects shared any body shape or coloring and thus became unique in order for participants to identify them.

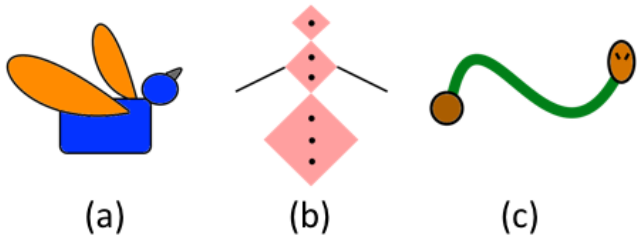

Figure 2.1: The three objects used for experimentation. (a) Obj1 was given the name "yosh". (b) Obj2 was given he name "wug". (c) Obj3 was given the name "niz". In this experiment, niz acted as control.

In order to train the participants, we created a training set that consisted of 10 yoshs, 10 wugs, and 10 nizzes. In addition, we also introduced 10 exemplars where the objects were incorrectly named. Table 2.1 summarizes our training set. Note that Obj3 and niz is the only consistent pair, and they served as control in

		Objects		
		Obj1	Obj2	Obj3
Labels	yosh	10	10	0
	wug	0	10	0
	niz	0	0	10

Table 2.1: Summary table of the training set. Participants were trained on learning 3 fictional objects (pictures are in Fig. 2.1) and three fictional labels.

our experiments. We expect participants to always correctly identify Obj3 and niz and not confuse them with any of the other objects or labels.

Participants were randomly split into two groups, one for OL (85 participants before filtering, 76 participants after filtering) and the other for LO (94 participants before filtering, 77 after filtering). Each participant was trained through a series of slides that showed the fictional objects followed by a correct or incorrect label (OL) or the label followed by a correct or incorrect object (LO) as outlined in table 2.1. The labels were both presented in writing and pronounced by a recording of a native English speaking female. The training set consisted of the 40 exemplars shown in a semi-randomized order. In particular, we made sure that the order of the exemplars was the same in the OL and LO group. Further, we never presented an incorrect pair following a correct pair for the same object or for the same label. We always started the training with a control pair and ended the training with a control pair; this set of restrictions prevented "first impression" bias.

Fig. 2.2 demonstrates how objects and labels were presented to both groups. Objects were presented for 175 ms to limit participants' ability to strategize. Labels were presented with the sentence "this is a (label)" for the LO group or "that was a (label)" for the OL group. Blank slides were presented for 150 ms between

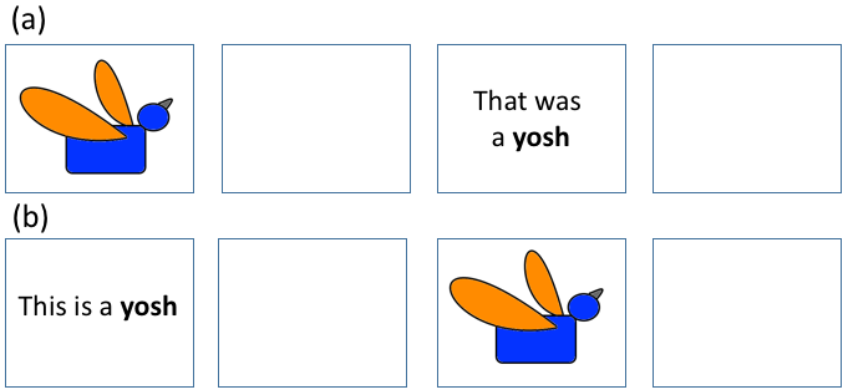

Figure 2.2: Diagram of the order of presentation. (a) is the OL scenario. (b) is the LO scenario. The presentation's slide duration is 150 ms for blank slides, 175 ms for slides with objects, and 1000 ms for slides with labels.

each object/label pair. For the analysis, we filtered out the participants that failed to identify correctly the non-ambiguous control pair. This left us with the total of 153 participants, of which 76 were in the OL group and 77 in the LO group.

After training on all 40 examples, participants were asked six questions to identify the objects (three "what is object?" type questions) and labels (three "what is label?" type questions). These were multiple choice questions and participants were instructed to select exactly one answer per question. Fig. 2.3 is a diagram of the test task. To enforce the OL and LO behavior, we minimized participants' opportunities to collaborate or strategize by asking participants to not speak among each other and to answer the questions to the best of their ability within the several seconds allocated to each question.

2.2.3 Experimental results

For the analysis below, we define ambiguous and non-ambiguous questions. Non-ambiguous questions are the questions where only one type of answer occurs in the presentation. The non-ambiguous questions are "what is wug?" and "what is

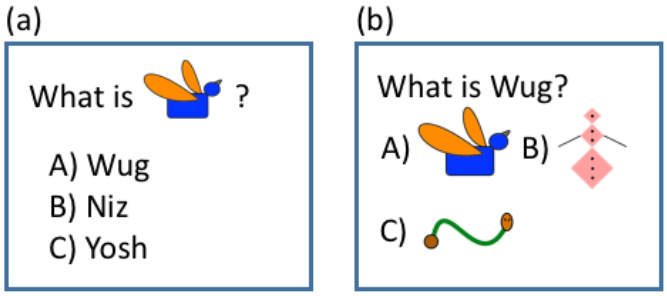

Figure 2.3: Diagram of the test task for the participants. (a) is the "what is object" type question. (b) is the "what is label" type question.

Obj1?". Ambiguous questions are characterized by multiple answers that are consistent with (part of) the training data. Ambiguous questions are "what is yosh?" and "what is Obj2?".

From table 2.1 we note that in the learning procedures employed here, the underlying source is extremely inconsistent. In fact, it is not immediately clear what are the "correct" association pairs. We can see that while Obj1 is always yosh, yosh can be both Obj1 and Obj2 with the same frequency. Similarly, while wug is always Obj2, Obj2 can be yosh or wug with the same frequency. We can say that Obj1 is a non-ambiguous object and wug is a non-ambiguous label. Seeing that Obj1 is always yosh and wug is always Obj2, we can tentatively say that Obj1-yosh and Obj2-wug are the "correct" associations and Obj2-yosh is "noise". The theoretical justification of this assignment (which does not alter our results) was put forward by [50].

The extremely inconsistent source used here is motivated by our experience gained from earlier attempts: (i) Experimentally, we determined that the length of the training session should not be much longer than what is used here, which limits the number of associations used, and (ii) Results of the experiments tended

to have a high degree of inter-response variability. To minimize the impact of participants who did not pay attention, we filtered out the responses of those who failed to correctly identify the control pair. Even after this filtering, we still wanted to maximize the experimental difference between ambiguous and non-ambiguous objects and labels. This was achieved by the setting of table 2.1. Effects of a subtler presence of source inconsistency could be simply masked by the noisiness of the experimental setup.

Results of the experiments are summarized in Fig. 2.4.

(a) What is object? OL

		Objects		
		Obj1	Obj2	Obj3
Labels	yosh	83%	32%	0%
	wug	17%	64%	0%
	niz	0%	4%	100%

(b) What is object? LO

		Objects		
		Obj1	Obj2	Obj3
Labels	yosh	69%	42%	0%
	wug	29%	56%	0%
	niz	2%	2%	100%

(c) What is label? OL

		Objects		
		Obj1	Obj2	Obj3
Labels	yosh	79%	21%	0%
	wug	12%	83%	5%
	niz	0%	0%	100%

(d) What is label? LO

		Objects		
		Obj1	Obj2	Obj3
Labels	yosh	57%	40%	3%
	wug	27%	66%	6%
	niz	0%	0%	100%

Figure 2.4: Experimental results. The percentages of participants giving the different answers to the multiple choice questions are presented for (a,b) "what is object?" questions and (c,d) "what is label?" questions. In (a) and (c), participants are from the OL group ($n = 76$). In (b) and (d), participants are from the LO group ($n = 77$). The arrows in the upper left corners of the tables indicate the direction in which the table entries sum up to 100%.

We notice the following trends.

- The "correct" answers (as predicted by the informativity theory of [50]) correspond to the largest percent of responses for all questions, for both learner types. In other words, the diagonal elements in the tables of Fig. 2.4 are larger than non-diagonal elements.

- OL learners provided a higher proportion of correct answers compared with the LO learners. Using the Fisher test, we can calculate the p-value between all the questions of comparing the number of OL learners versus the number of LO learners who answered the questions correctly.

  - For the non-ambiguous object ("what is Obj1") question, we have that $p = 0.058$. Thus since the p-value is relatively small, then we can conclude that OL learners perform better than LO learners.

  - For the ambiguous object ("what is Obj2") question, $p = 0.3229$; this is the only question for which we cannot conclude here that OL learners perform better. More testing is needed.

  - For the non-ambiguous label ("what is wug") question ($p = 0.0254$) and the ambiguous label ("what is yosh") question ($p = 0.005$), we can see that OL learners significantly answer these questions at a higher accuracy than LO learners and thus the OLO effect is observed.

- For the "what is object" questions:

  - For the OL learners 83% of participants correctly identified Obj1 as yosh. 64% of participants correctly identified wug when answering "what is Obj2" question. We can see that OL participants had difficulty with identifying ambiguous objects' labels. We used the Fisher

test to see if the number of OL participants correctly answering the ambiguous question was significantly lower than that for non-ambiguous questions.The p-value is 0.016, suggesting that the difference between the OL participants in identifying the two different types of objects is statistically significant.

  - For the LO learners, 69% answered yosh when asked "what is Obj1", 56% answered wug when asked "what is Obj2". When calculating the p-value for the participants for these two questions, we have that $p = 0.1341$.

- For the "what is label" questions:

  - Among the OL participants, 79% correctly answered Obj1 when asked "what is yosh", and 83% correctly answered Obj2 when asked "what is wug". The p-value here is 0.68 using the Fisher exact test on the number of OL participants.

  - With the LO learners, 57% correctly answered Obj1 when asked "what is yosh" and 66% correctly answered Obj2 when asked "what is wug" ($p = 0.32$).

  Even though we can see that there is a higher percentage of learners correctly answering non-ambiguous "what is label" questions, the p-values are not small enough to determine that there is a statistical significance to this trend; more testing will be required.

It is interesting to make a connection between the setting of our experiments and categorization/ concept learning literature, see e.g. [20]. The ambiguous "what is label?" question that involves one label and two objects is an example of the many-to-one category learning. The ambiguous "what is object?" question

that involves one object and two labels is an example of the one-to-many category learning. It has been argued that the many-to-one learning process is a stronger mechanism of learning than the one-to-many learning process [20]. This is evident from our experimental results. We see that both learners perform better with correctly answering ambiguous "what is label" questions (many-to-one) than the ambiguous "what is object" questions (one-to-many).

In short, we observe that both OL and LO learners can identify non-ambiguous associations better than ambiguous ones. In addition, both learners are able to identify labels correlating to two or more objects better than objects correlating to two or more labels. OL learners are also significantly better at identifying both non-ambiguous and ambiguous associations compared to LO learners.

2.2.4 Comparison with the matrix method

Let us start with the matrix that records the input:

		Objects		
		Obj1	Obj2	Obj3
Labels	yosh	10	10	0
	wug	0	10	0
	niz	0	0	10

We would like to assess to what extent the theoretical predictions of the matrix method are consistent with the data presented in Fig. 2.4.

"What is object" questions. In the absence of negative evidence, we can obtain the expected percentages of different answers to the questions by simply normalizing the input matrix over columns:

No negative evidence:

	Objects		
	Obj1	Obj2	Obj3
yosh	100%	50%	0
wug	0	50%	0
niz	0	0	100%

In order to account for possible negative evidence, recall from section 2.1.2 that we can apply the matrix method to predict the responses of the participants. For the "what is object" questions, we normalize over rows; in section 2.1.2, we then simply picked the largest element in each column to predict the most typical answer. Here we could take a step further and actually calculate the percentage of responders giving different answers to the "what is object" questions. To do this, we simply normalize over the columns:

With negative evidence:

	Objects		
	Obj1	Obj2	Obj3
yosh	$\frac{1}{2}$	$\frac{1}{2}$	0
wug	0	1	0
niz	0	0	1

$\Rightarrow$

	Objects		
	Obj1	Obj2	Obj3
yosh	100%	33%	0
wug	0	67%	0
niz	0	0	100%

"What is label" questions. For "what is label" questions, in the absence of negative evidence, we simply normalize the input matrix row by row:

No negative evidence:

	Objects		
	Obj1	Obj2	Obj3
yosh	50%	50%	0
wug	0	100%	0
niz	0	0	100%

To incorporate the influence of negative evidence, we normalize over columns and then over rows, to obtain the predicted percentage of people giving different answers to the "what is label" questions:

With negative evidence:

	Objects		
	Obj1	Obj2	Obj3
yosh	1	$\frac{1}{2}$	0
wug	0	$\frac{1}{2}$	0
niz	0	0	1

$\Rightarrow$

	Objects		
	Obj1	Obj2	Obj3
yosh	67%	33%	0
wug	0	100%	0
niz	0	0	100%

From the two resulting matrices, non-ambiguous questions ("what is Obj1?" and "what is wug?"), 100% of the participants are expected to give correct answers. For the ambiguous questions "what is Obj2?" and "what is yosh?", the learners should split in proportion $2 : 1$. This is however not exactly what we see when we compare these theoretical results with the experimental outcome in Fig. 2.4.

The first difference is that the correct answers to non-ambiguous questions are given by fewer than 100% of the people (except for the control questions; there, the correct answer is given 100% of the time because we discarded the responses of the participants that failed to answer the control questions correctly). This result is expected because of the presence of inconsistencies and imperfections of learning.

The second difference between the theoretical prediction of the matrix method and reality becomes obvious when we compare the performance of the OL and

LO learners in their answers to ambiguous questions. We can see that LO learners produce the "correct" answers fewer than 67% of the time (namely, they answer the "what is Obj2?" and "what is yosh?" questions correctly 56% and 57% of the times, respectively). In comparison, OL learners do a lot better, and in the case of the "what is wug?" question, they actually get it right 83% of the times, compared to the predicted 67%.

These trends remind us of the phenomena of frequency boosting in the case of OL learners, and frequency undermatching in the case of LO learners. We will explore these by means of mathematical modeling in the next section.

2.3 Computational results

We implement a computation model of the experiments in order to identify the key differences between the OL and LO learners that are responsible for the different experimental outcomes. The objective is to come up with a minimal model capable of capturing the patterns observed experimentally.

Using the stochastic learning model with negative evidence, we will focus on two parameters. One comes from the update function $F(x) = x^r$ used in MCM. The parameter r is responsible for the frequency boosting (and undermatching) of the learner, as described in [33], see also Section 1.2. In particular, recall that $r = 1$ corresponds to frequency matching, $r < 1$ to frequency boosting, and $r > 1$ to undermatching. We hypothesize that the two types of learners may differ in quality and degree of frequency manipulation.

Further, we assume that the value of r may also depend on the type of question that is being addressed. For example, an OL learner in the process of learning uses objects to predict the labels, and thus, objects are competing with one another. On the other hand, as LO learners are presented the material, they are predicting objects given a label. As such, an LO learner is correlating the associations of objects to labels when answering a "what is label?" question. This simple argument suggests that for an OL learner, a "what is object?" question may be processed differently from a "what is label?" question, but in some sense more similar to a "what is label?" question for an LO learner. We therefore conclude that there could by up to four distinct r parameters describing the learners. We also hypothesize that the magnitude ordering for the two pairs (r for OL for "what is object?", r for OL for "what is label?") and (r for LO for "what is label?", r for LO for "what is object?") is the same.

The second parameter that can affect learning is the strength of negative evidence, μ. This parameter defines how often the learner processes negative associations. To exhibit a difference between OL and LO learners, as well as the two types of questions, we would need varying negative evidence values, giving another 4 independent parameter values.

In order to understand how these two parameters affect the learning outcome, we implement the stochastic learning model with negative evidence on the training system (table 2.1), and analyze the behavior of the association of yosh-Obj1 and wug-Obj2, as parameters r and μ change. These two associations are of the main interest throughout the experiments. Let us refer to the matrices P and Q from (2.1) and (2.2). The values of α_1 and β_2 determine the association of yosh and wug to Obj1 and Obj2 respectively; similarly, the values of x_1 and y_2 determine

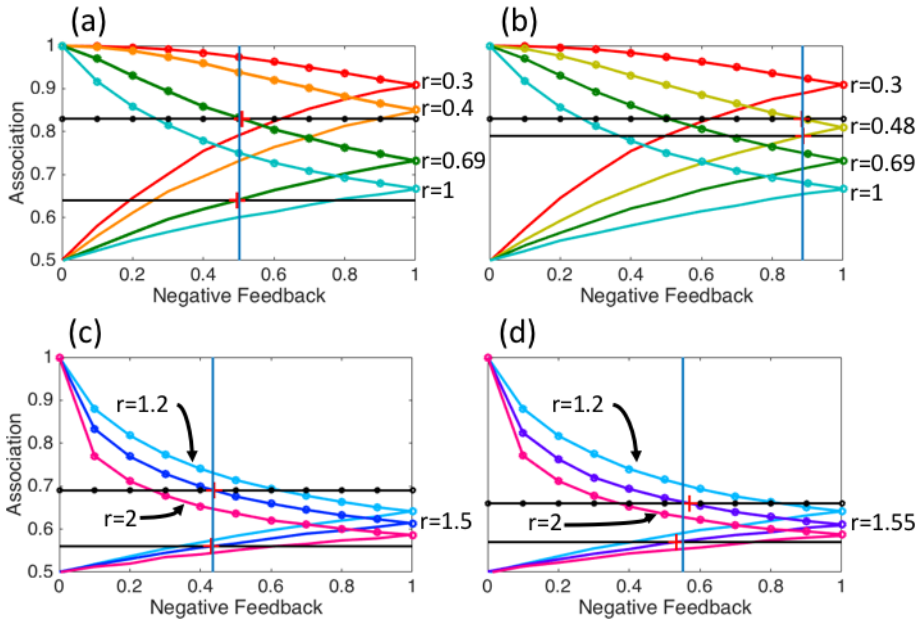

Figure 2.5: Stochastic simulation results showing the associations of learners as functions of parameters: (a) OL learners answering "what is object" questions; (b) OL learners answering "what is label" questions; (c) LO learners answering "what is object" questions and (d) LO learners answering "what is label" questions. The ambiguous (lines) and non-ambiguous (circles) associations are plotted as functions of parameter μ (the strength of the negative evidence) for several values of r (the frequency boosting/demotion parameter). Horizontal lines correspond to the experimentally observed values. Vertical lines mark the strength of negative feedback, where the stochastic learning model with negative evidence and experimental values coincide for the same r.

the association of Obj1 and Obj2 to yosh and wug respectively. Further note that α_1 represents the value of a non-ambiguous label for Obj1, whereas β_2 represents the value of an ambiguous label for Obj2. Similarly, x_1 represents the value of an ambiguous object for yosh, whereas y_2 represents the value of a non-ambiguous object for wug.

Fig. 2.5 summarizes the results of the stochastic learning model with negative evidence. Four graphs are presented, which differ by the learning type (OL in the top row, LO in the bottom row) and the types of questions ("what is object" in the

left column and "what is label" in the right column). The ambiguous association is plotted by lines and the non-ambiguous association by connected circles. These values are presented as functions of the strength of negative evidence ($0 \leq \mu \leq 1$) and are plotted for several different values of r. Note that we present results for $r < 1$ for OL learners and $r > 1$ for LO learners. It was hypothesized in section 2.2.4 that OL learners exhibited features of frequency boosting, because the experimentally observed values for the correct associations were higher than those obtained from the matrix method, indicating that $r < 1$. On the other hand, LO learners had values smaller than those predicted by the matrix method, hinting at the presence of "undermatching", $r > 1$. The results of the fitting procedure confirm that these are indeed the correct ranges for the two learning types.

We have the following observations of the graphs Fig. 2.5:

- Non-ambiguous associations are always higher (or equal) to the ambiguous associations.
- In the absence of negative evidence ($\mu = 0$), we have the ambiguous and nonambiguous values at $\frac{1}{2}$ and 1 respectively, as predicted by the matrix theory, see section 2.1.2.
- As μ (the strength of negative updates) increases, ambiguous associations increase and the non-ambiguous association decreases.
- When $\mu = 1$, then values of the ambiguous and non-ambiguous associations coincide, as observed in section A.4.
- For a fixed μ value, as r decreases, the association of both ambiguous and non-ambiguous values increases. This is due to the frequency boosting

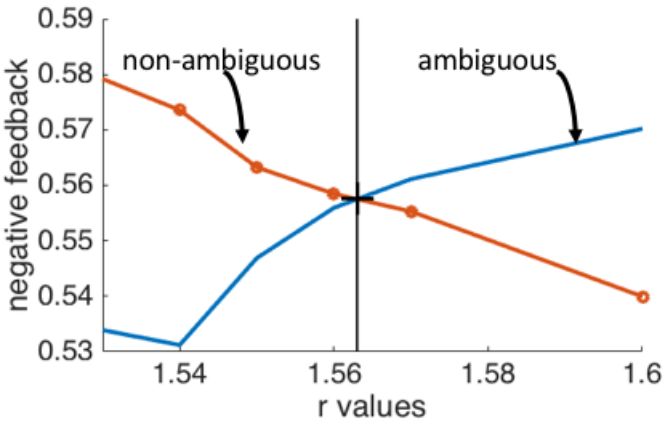

Figure 2.6: Plot of the (r, μ) pair where the non-ambiguous and ambiguous association values of the stochastic simulations (for LO learners in the "what is label" question) intersect the experimental values as depicted by Fig. 2.5. Red is the intersections of the non-ambiguous associations. Blue is the intersections of the ambiguous associations. The (r, μ) pair where both of these plots coincide is the parameter pair that best describes the experimental values from the simulations.

property. That is, as r decreases, then the update function of MCM exhibits a stronger frequency boosting property and thus associations will increase.

In order to find the parameters r and μ that best describe the experimental results, for each of the four situations of Fig. 2.5, we need to find a pair (r, μ) such that both the ambiguous and the non-ambiguous association values coincides with the ones obtained in the experiments. For example, in the case of "what is label" questions for LO learners, the intersection of the ambiguous and non-ambiguous associations with their observed values for $r = 1.55$ both occur at $\mu = 0.56$ and $\mu = 0.58$. To refine our search to find one μ value, Fig. 2.6 shows the intersection for non-ambiguous and ambiguous association values of the simulation with the experiments. We can see that $r = 1.563$ and $\mu = 0.556$ is a pair where the intersection will exist. Using this procedure, we find the best fitting values for all case, see table 2.2.

		r	μ
OL	what is object	0.6976	0.5038
	what is label	0.4765	0.8885
LO	what is object	1.5177	0.4333
	what is label	1.5630	0.5575

Table 2.2: Parameters to fit the experimental data with the stochastic learning model with negative evidence. OL learners have update functions, x^r, with $r < 1$ while LO learners have $r > 1$.

Fig. 2.7-2.8 are bar graphs of the simulation with the parameters listed in table 2.2 and the experimental data. Fig. 2.7 represents the percentage of correct answers for the "what is object?" questions grouped into OL and LO answers. Fig. 2.8 represents the percentage of the correct answers for the "what is label?" questions, again grouped into OL and LO answers. We can see that the parameters for the simulation match the experimental data.

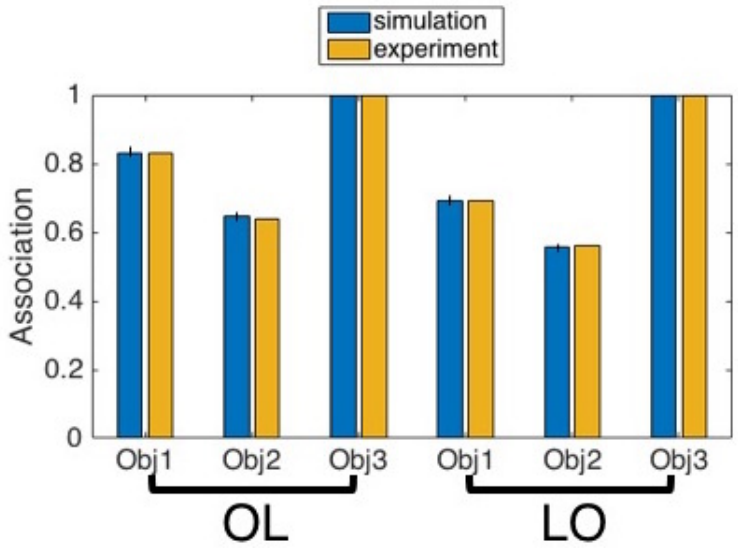

Figure 2.7: Simulation versus experimental data for "what is object" questions. We can see that the parameters of rows 1 and 3 of table 2.2 match the data.

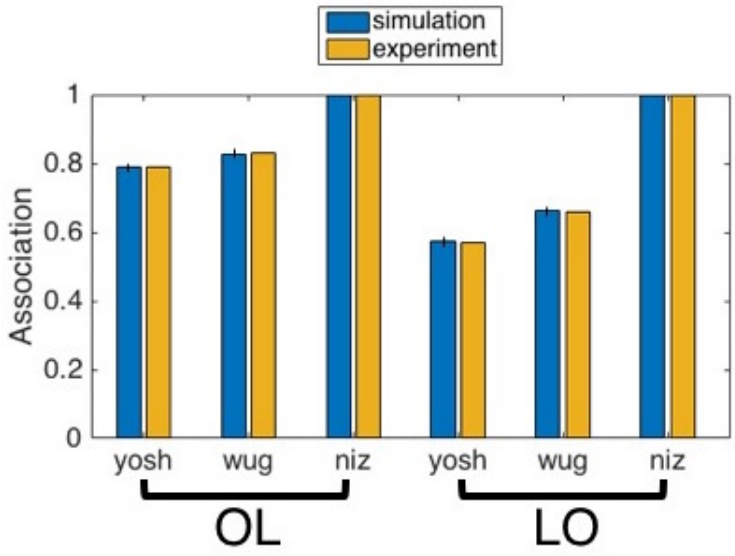

Figure 2.8: Simulation versus experimental data for "what is label" questions. We can see that the parameters of rows 2 and 4 of table 2.2 match the data.

2.4 Discussion

In this chapter we studied the object-label-order effect in the context of noisy learning, or learning from an inconsistent source. In our setting, during the learning phase, the same object can be paired with more than one label, and the same label can be used to name multiple objects. The learners are facing the difficult task of making sense of such inconsistent input, and creating associations that can serve as a representation of reality that potentially can be useful in communication.

We combined novel experimental data with a theoretical model of learning. The model's most important components are (i) its ability to regularize the input (and also to do the opposite!) and (ii) its ability to take account of interactions (negative evidence) between objects (and labels) that are learned simultaneously.

It turned out that these two components were necessary and sufficient to describe the experimental data obtained. It is important however to point out that this work is not an exercise in parameter fitting. It is not hard to match 8 experimental points with 8 parameters, and this was by far not the goal of this study. The point was to deduce differences between the two learning mechanisms, OL and LO. It appears that OL learners are performing frequency boosting (overmatching) while LO learners are performing an undermatching operation. It further transpires that the negative evidence is necessary to account for the observed patterns, and moreover, the strength of negative evidence appears to be somewhat different in different learning scenarios, namely, "what is label?" questions seem to evoke a stronger reaction to negative evidence compared to the "what is object?" questions.

The value of mathematical modeling here is to test various hypotheses. By looking at the learning data (such as those presented in Fig. 2.4), one may create various scenarios that may intuitively explain the observations. It is however impossible to be sure that the assumptions can lead all the way to the conclusions without using a systematic way of testing it. In this context, mathematical modeling provides a tool that enables us to follow the hypotheses all the way through to their logical outcomes. In the context of the questions posed in this paper, we can say that OL and LO learners are characterized by different degrees of regularization power, and they also utilize negative evidence to different degrees.

Regularization in language learning has been extensively studied in the literature because, among other factors, of its implications for language evolution. Evidence of regularization in children is vast, and is considered at the core of language dynamics, exemplified e.g. by Nicaraguan sign language creation [67, 68,

69]. Although it has been argued that children have a higher tendency to regularize compared to adults [25], there is strong evidence of adult regularization behavior in language learning [27]. In fact, it has been argued that under some circumstances, adults regularized as often or even more than children [40, 73, 76]. It has been proposed that creolization occurs due to cohorts of adult learners [2, 35]. Regularization is at the basis of the iterated learning paradigm, in which one learner's output is given as input to the next learner, in a chain-like fashion. It has been found that initial inconsistencies of language are slowly phased out from the language [52, 71]. [47] suggests that in adults, "probability matching is surprising and apparently irrational behavior." She argues that regularization allows for more efficient communication, as is established in learning theory and decision making. The cases of frequency matching (which have also been documented extensively in adults) can be explained by rational, pragmatic reasoning on the part of learning adults.

The second important aspect of our conceptual model is negative evidence and how learners can incorporate it into their internal algorithm. First of all, it is very clear from our experimental results that some sort of negative evidence plays a significant role in the participants' responses. For example, from the summary of the training set, table 2.1, we can see that label "yosh" was used 10 times with Obj1 and also 10 times with Obj1. In the absence of any information about the the other labels, it would not be possible to determine whether "yosh" would be chosen as the leading label for Obj1 or Obj2. So we would expect that about the same percentage of people would think that "yosh" is Obj1 and Obj2. Examining the experimental data in Fig. 2.4(c,d) we can see that 79% of OL learners and 57% of LO learners identifies "yosh" as Obj1. It appears that in a sense, the presence of "wug"-Obj2 associations serves as negative evidence to weaken the "yosh"-Obj2

associations and thus to strengthen the complimentary, "yosh"-Obj1 associations.

A very informative discussion of the concept of negative evidence and its controversies in language learning can be found in [49, 51]. In our work, by "negative evidence" we always mean "implicit negative evidence". This is in contrast to "explicit negative evidence", such as direct statements obtained by the learner that a certain construct (or association) is incorrect.[2] Implicit negative evidence is generally information that can be extracted from statistical properties of the stochastic source [59, 65]. As was observed in [14], distributional information can provide "a kind of 'negative evidence'", when expectations are formed which can be subsequently violated. As pointed out in [31], "at least some cases, the so-called 'logical problems' associated with the no negative evidence hypothesis may be solved by admitting the stochastic information." In our models, we account for this type of negative evidence of statistical kind. It is interesting that according to our model, it is utilized to a different extent by the two types of learners, OL and LO.

This work has implications to the general area of concept formation. We are studying the learning process, by which information is acquired about the probabilistic relationships between objects, and the process in which people predict and categorize these objects. Several types of concept learning have been described in the literature, such as relational, perceptual, associative [77]. In this paper we are exploring associative concept learning, which is where objects share no physical features among each other, but still share functional properties [70]. This type of learning is used in cognitive sciences to study symbolic processes of both

[2]Explicit negative evidence could also include the phenomenon of subtle or covert explicit evidence, which can be observed in children learning a language, when a parent or a teacher has a greater tendency to rephrase ungrammatical compared with grammatical utterances. Such types of negative evidence are not part of our constructs.

human and nonhuman cognition. Focusing on human cognition, we illustrate that this type of learning occurs when matching objects to categories or labels, creating conditional correlations involving unrelated features. These features can be grouped by association to their categories in two ways. Many-to-one learning process associates the one categorical response from two or more exemplars. This type of association has been studied in depth [36, 49]. There is also the reverse, one-to-many learning process, in which a common exemplar is associated with two or more categories. It has been suggested that the one-to-many learning process affects generalized categorization learning in a negative way or that it is not as strong of a mechanism of learning as its counterpart, many-to-one correlation [20]. In the context of our experiments, these two types of learning correspond to the two types of ambiguous questions ("what is label?" and "what is object?") that appear in our experimental setting.

This study can be extended in a number of ways. For example, in the present setting, the objects used in the training process did not share features (i.e. each object presented a single cue dimension). It would be interesting to include feature variance/overlap in the training exemplars of the objects, and to study the resulting cue competition, in addition to the source inconsistency effects, in the two different types of learners. We further emphasize that the present study does not give all the answers even for the simplified setting that it explores, but instead proposes several questions that can be addressed by further research in cognitive sciences, both in the experimental and computational terms. In particular, we suggest the possibility that the two types of learners, OL and LO, regularize the input and utilize implicit negative evidence to different extents. In the present paper we show that this is consistent with the observed learning patterns. Future experimentation will be able to test this hypotheses further.

Evolution of the first word

> "One voice can change a room…your voice can change the world."
>
> *Barack H. Obama*

In this chapter, we address how the first word emerges in a population. In our environment, we assume that there is a qualitative difference between pre-existing, innate vocalizations and the first word. In regards to our model, the new word is more prone to be forgotten by individuals in our population since it is a larger, complex combination of innate vocalizations (consonant-vowel syllable). We assume that the emergence of this "language" is stored in a set of word-event associations for each individual. These associations quantify how a given individual will use the new word in each context, as well as how likely they will interpret an event or the new word. Let us consider two vocalizations: one that is innate and simple such as "a" and the newly learned, syllable-like word "ba". In addition to "a" and "ba", we will model their association to two events, E_1 and E_2, which occur at different frequencies. Suppose that E_1 is the more frequent of the two and furthermore suppose that the majority of the population associates the innate vocalization "a" with event E_1. We explore whether there is a possibility that the new, complex word can be learned by the population and can associate itself with the most frequent event.

3.1 Set-up and patterns

To explore this scenario, we introduce an "inventor" of the new word, "ba", into a population of people who have only established using "a" and nothing else.

The "inventor" uses both "ba" and "a" with a certain frequency in association with event E_1. Then the training scenario begins: Suppose that E_1 occurs, then a random individual in the population is chosen to utter something in association with E_1. The remaining individuals then update their association probabilistically, strengthening their association of E_1 with the uttered vocalization and weakening their association of E_1 to utterances that were expected but not heard.

The strengthening of associations is based on (possibly nonlinear versions of) the Bush-Mosteller and Resocrla-Wagner models as discussed in chapter 1. We will use the MCM model for our learning algorithm. Let us briefly remind ourselves of the parameters that will characterizs our learning. First, we have the update function,

$$F(X/L) = (X/L)^r,$$

where r is the regularization parameter. In section 1.3.1 we discussed the different types of regularization. When $r = 1$, we exhibit frequency matching; when $r < 1$, we exhibit frequency boosting. Parameter L is the discretization parameter, which characterizes the Markov walk in MCM. It will be noted in section 3.4 that the size of L plays a role in describing the learning behavior. We introduce a new parameter, p, which will measure the probability of forgetting; $p = 0$ means that the word will not be forgotten.

We have performed extensive simulations of the learning system. The following two patterns were observed for a wide variety of initial conditions and parameters:

- *Shared, steady solution.* After some time, the population converged to a steady state characterized by a common, shared usage of the words, that

is, the associations of the individuals oscillated around the same means (see Fig. 3.1).

- *Inherent stochasticity.* The steady shared solution reached by the population may be different from simulation to simulation. Therefore, one can measure the frequency (or likelihood) of each outcome (see section 3.5).

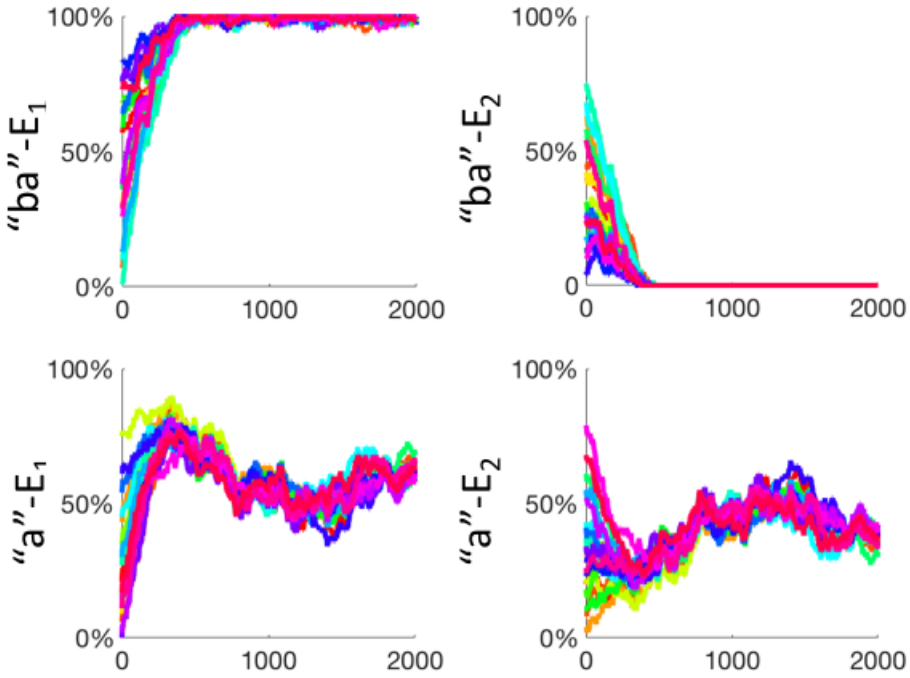

Figure 3.1: Starting from random initial conditions for all learners in the population, the individual trajectories for the "ba"-E_1 association strength are shown for the whole population. In this simulation, this association rose to dominance. Convergence for "ba"-E_2, "a"-E_1, and "a"-E_2 association strength is also observed.

3.2 Population learning algorithm

Suppose that the two events, E_1 and E_2, occur with frequencies Q_1 and Q_2 respectively ($Q_2 = 1 - Q_1$). We assume that the first event is more frequent, that is $Q_1 > \frac{1}{2}$. Further, let us suppose that initially the majority of the population uses "a" to describe both of the events, and introduce "ba" at a low frequency. We

assume that "ba" is more complex, and if it is not heard by individuals during training, it has a larger probability to be forgotten compared to "a".

In a population of individuals, every person is equipped with a matrix of associations:

	E_1	E_2	
"ba"	α	β	ϵ_1
"a"	γ	δ	ϵ_2

where α is the association strength between "ba" and event E_1. Associations are stochastic variables that take integer values between 0 and L (here L is some integer number), and change according to the Markov chain described by MCM. Note that we assume

$$\alpha + \beta + \epsilon_1 = L, \quad \gamma + \delta + \epsilon_2 = L.$$

The values ϵ_1 and ϵ_2 stand for the "empty" meanings of "ba" and "a" respectively. For example, if an individual does not use "ba" at all and uses "a" equally for events E_1 and E_2, their association matrix is given by

	E_1	E_2	
"ba"	0	0	L
"a"	$L/2$	$L/2$	0

We can describe the training scenario with the association matrix in the following way. At each time-step an event occurs (E_1 with probability Q_1 and E_2 with probability $1 - Q_1$). Then a random individual is drawn from the population. This individual will then utter "ba" or "a" from corresponding to the event. For example, if E_1 occurred, then the individual selected will use their association matrix and utter "ba" with probability $\frac{\alpha}{\alpha+\gamma}$ and utter "a" with probability $\frac{\gamma}{\alpha+\gamma}$.

Then everyone else in the population will update their association according to this input using the MCM algorithm.

At each step, only one (out of two) utterances is heard. We assume that the word that was not heard may have a chance to be forgotten to a certain degree, that is, its associations with events E_1 and E_2 are weakened. For example, if the word "a" is heard, this means that "ba" is not heard, and everyone has a chance p_1 to weaken their "ba"-E_1 and "ba"-E_2 associations. Similarly, in the case that "ba" was heard, the associations of "a" are weakened with probability p_2. In the analysis presented here, we will assume that $p_1 = p > 0$ and $p_2 = 0$, that is, the word "ba" is forgotten with a finite probability, and the word "a" is not forgotten. The forgetting routine is implemented in the following way. If word "ba" was not heard, then if $\alpha > 0$, we perform

$$\alpha \to \alpha - 1, \quad \epsilon_1 \to \epsilon_1 + 1,$$

with probability p. Similarly, if $\beta > 0$, we perform

$$\beta \to \beta - 1, \quad \epsilon_1 \to \epsilon_1 + 1,$$

with probability p.

3.3 Rise of the first word

It can be shown that even if the new word, "ba", is initially used infrequently, the population may establish a shared solution where the associations "ba"-E_1 and "a"-E_2 dominate. Specifically, the new word "takes off," and adopts the meaning of the more frequent event E_1, pushing the simple innate vocalization, "a" to (the less frequent meaning) E_2. Fig. 3.2 shows typical outcomes of multiple simulations that show the mean population dynamics of different associations. Here

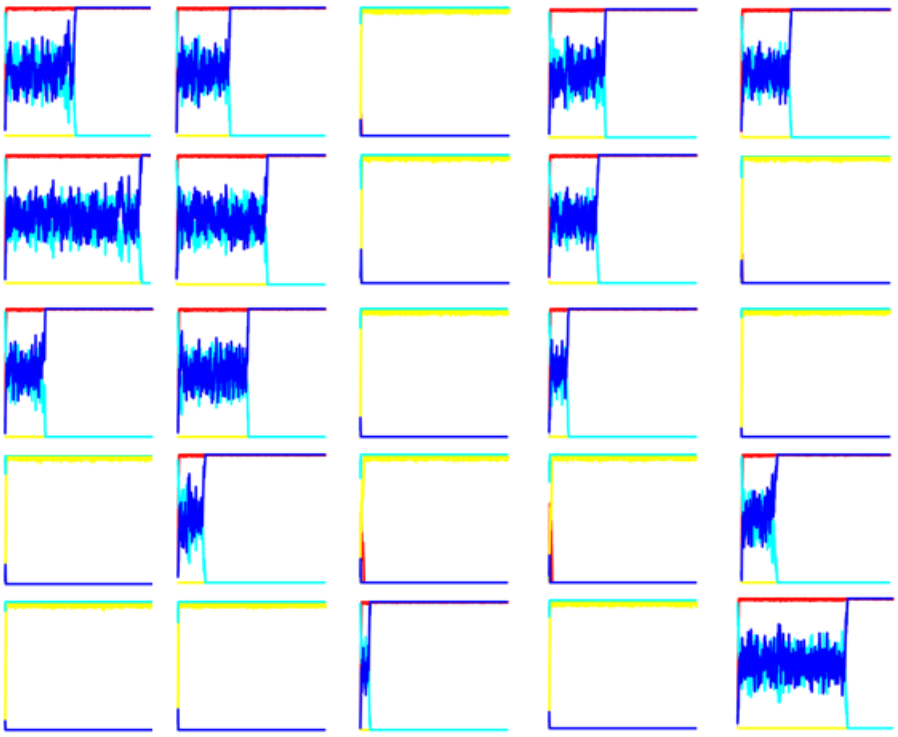

Figure 3.2: Shown are numerical simulations of language dynamics, starting from the initial condition where the usage of the composite word "'ba" was uniformly low. Plotted are the population mean association strengths for "ba"-E_1 (red), "a"-E_2 (dark blue), "a"-E_1 (cyan), and 'ba"-E_2 (yellow). The horizontal axes have the meaning of time (iterations). Out of 25 runs presented here, 14 resulted in the dominance of "ba"-E_1 and "a"-E_2. The parameters were: $r = 0.55$, $p = 0.04$.

associations "ba"-E_1 and "a"-E_2 are shown in red and dark blue respectively. We can see that after initial dynamics, in 14 out of 25 cases, these two associations come to dominate the language.

Examining the typical time-series of the word-context associations (Fig. 3.2) we observe the existence of at least two different time scales. While the rise of the "ba"-E_1 association, if it happens, occurs relatively quickly (that is, "ba" starts being used for E_1 only, and never for E_2), the innate "a" continues to be used, with similar probabilities for both E_1 and E_2. In other words, the association matrix at this time contains a weaker synonym for E_1 ("a"), and at this time "a" is also used

as a homonym, to be associated both with E_1 and E_2. Only after a relatively long time-span, the "a"-E_2 association finally comes to dominance, displacing "a"-E_1. At this point, we can say that with respect to these two words, the population has developed an unambiguous communication system. In the context of our model, we observe the disambiguation of the meaning of "a" happens only for the values of the regularization parameter, r, smaller than a threshold. In other words, a certain degree of nonlinearity of the learning dynamics is required to see the transition to efficient communication.

We further explored the behavior of a population where initially, most people do not use the word "ba," and only a minority of the population had the word "ba" in their vocabulary (see Fig. 3.3(a), the association matrices on the left). For simulations presented in Fig. 3.3, one out of the 20 individuals used the innate "a" in the same way as everyone else, but also used the new word "ba" to denote mostly event E_1. Starting with this configuration, the learning process was simulated many times, until a steady state was reached by the population. The frequency of different outcomes was recorded. We are particularly interested in the frequency of the outcome where association "ba"-E_1 is dominant. Fig. 3.3 shows a typical simulation where the new word catches on. Trajectories of individual association strength for "ba"-E_1 are plotted.

We observe that for the parameters chosen for this simulation, the new word's association with the more frequent event spreads through the population, reaching dominancy. Fig 3.3(b) shows the dependence on the forgetting parameter, p, to have such an outcome (a more detailed study is presented in section 3.5). Interestingly, the likelihood of the "ba"-E_1 association dominating reaches a maximum for a nonzero forgetting parameter. In other words, it is beneficial for the

system to have a certain probability of forgetting the learned, composite word, in order for it to become linked with the more frequent event. In Fig. 3.4, we observed that there is a forgetting p value for each regularizing parameter r where the frequency of observed "ba"-E_1 dominance reaches a maximum.

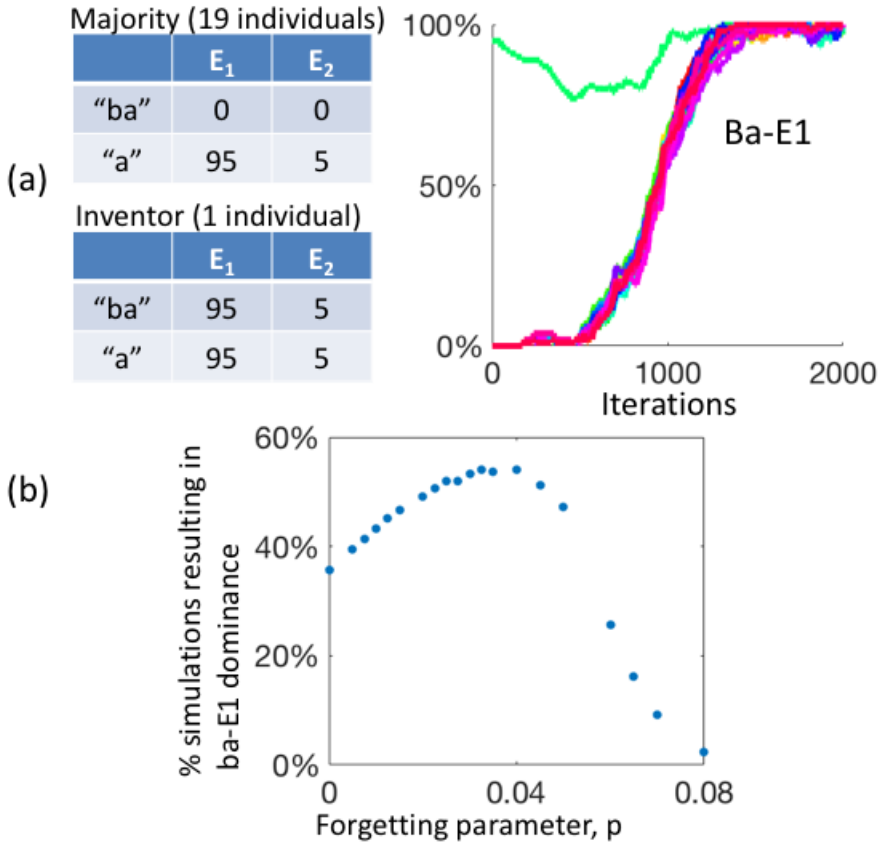

Figure 3.3: (a) Starting from the initial condition where only one individual (inventor) uses the new word (left), the individual trajectories for the "ba"-E_1 association strength are shown for the whole population. In this simulation, this association rose to dominance. The parameters are $r = 0.55$ and $p = 0.04$ (b) The probability for the "ba"-E_1 association rises (for $r = 0.55$) to dominance as a function of the forgetting parameter.

Intuitively, it is clear that if the probability to forget the word "ba" is too high, this can destroy the word, resulting in the population only using the shorter word "a". On the other hand, zero forgetting can be inferior to a certain degree of for-

getting as far as the "ba"-E_1 association is concerned. Indeed, in the presence of forgetting, association "ba"-E_2 is unstable because E_2 is an infrequent event, and if the population only hears "ba" when E_2 occurs, there are too many opportunities for a negative, "forgetting" update for this word. On the other hand, the association "ba"-E_1 appears more stable because of the frequent "reminders" about the longer word, which help maintain in it in the population.

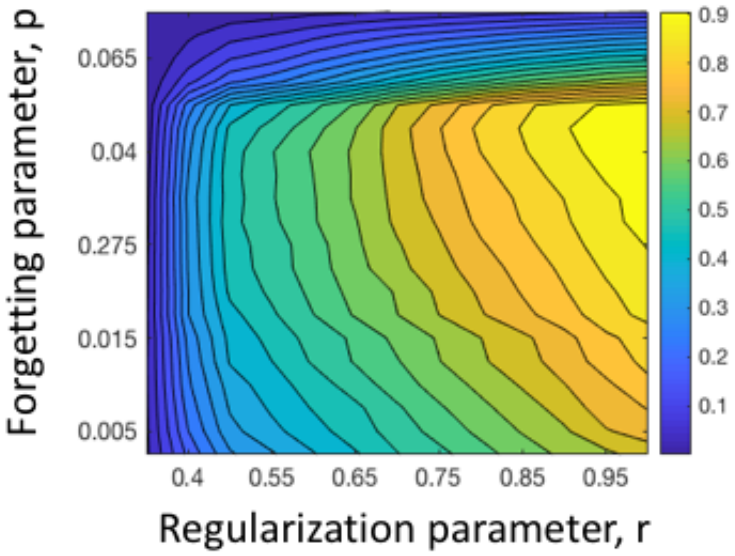

Figure 3.4: The frequency of the shared solution with a dominant "ba"-E_1 association, as a function of the regularization and forgetting parameter, out of 100 independent runs with the initial condition described in Fig. 3.3. The lighter color corresponds to higher frequency of the association.

3.4 Talkative versus quiet speakers

We further explored the difference between quiet and talkative species. Which of these two species is more readily associated with the rise of a new, learned, utterance? In order to investigate this aspect of word dynamics we assumed that in a "talkative" species frequent utterances inevitably result in more frequent, but smaller, updates of the association values, while for the "quiet" species, the updates are less frequent but are relatively large in value. The model suggests that for

the "quiet" species, the dynamics are noisier (see also related study in [74]). At the same time, the rise of the new, learned word occurs (i) at a higher probability, see Fig. 3.5, and (ii) faster compared to that in the "talkative" species (given that the rest of the parameters are equal between the two cases and the initial conditions are equivalent), see Fig. 3.6. This holds in particular for small and large forgetting rates. In Fig. 3.7, we can see that the intermediate forgetting rates that are optimal for "talkative" species seem to be detrimental resulting in a lower probability of adopting the compositional utterance, see Fig. 3.7.

It should be noted that since L, discretization parameter, is different for "talkative" ($L = 100$) and "quiet" ($L = 20$) species that the initial conditions for each should be scaled appropriately for comparison. For Fig. 3.5 - 3.7 the initial conditions for $L = 100$ are the same as the initial conditions for Fig. 3.3. The initial conditions for $L = 20$ are: 19 individuals with

	E_1	E_2	
"ba"	0	0	20
"a"	19	1	0

and 1 individual with

	E_1	E_2	
"ba"	19	1	0
"a"	19	1	0

Fig. 3.6 presents the histogram of simulated time that it took for a population to have a stable, shared, dominant association strength for "ba"-E_1. The time for the new word dominance is significantly (p-value is 10^{-6} with T-test) larger for the "talkative" species. Note that in these simulations, the talkative species were assumed to produce utterances 5 times more frequently compared to the quiet

species, and the physical time of learning was scaled accordingly. The final comparison between the "talkative" and "quiet" species was performed in the context of disambiguation of the meaning of "a" after the new composite word has taken root. We observe that transition to efficient communication occurs more readily for quieter species. For example, for the parameters for Fig. 3.6, out of 50 populations of "talkative" learners simulated for $1,000,000$ time-steps, none exhibited the transition to unambiguous communication system. On the other hand, out of 50 populations of "quiet" learners, 17 were able to disambiguate the usage of "a".

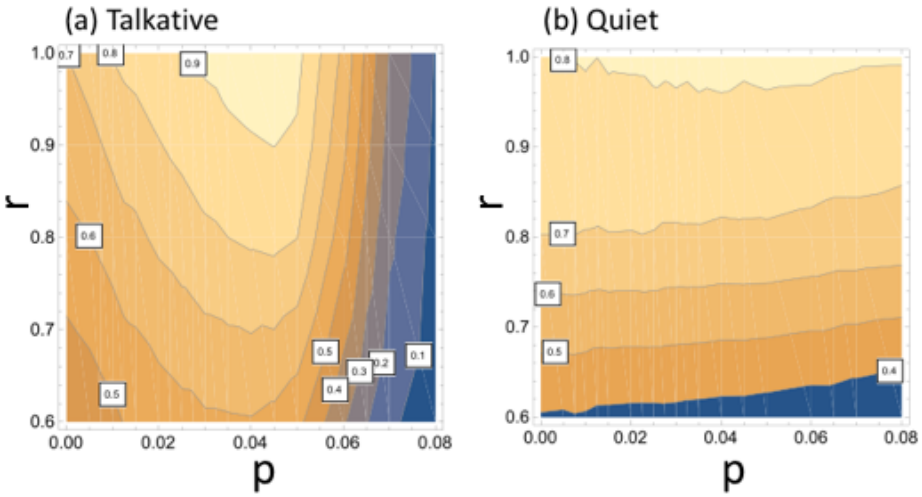

Figure 3.5: Frequency of shared solution with a dominant "ba"-E_1 association as a function of the regularization and forgetting parameters for (a) "talkative" species ($L = 100$) and (b) "quiet" species ($L = 20$) with the initial conditions described in Fig 3.3.

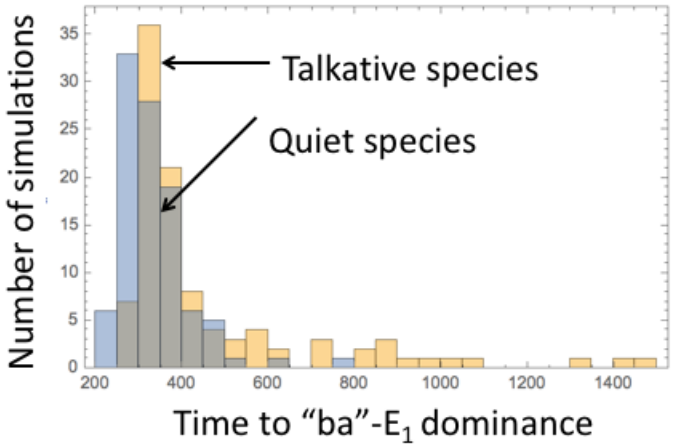

Figure 3.6: The time to "ba"-E_1 dominance (conditioned on the occurrence of this event) for many independent simulations is shown in the form of histograms. The yellow histogram is for "talkative" species ($L = 100$) while the blue histogram is for the "quiet" species ($L = 20$). Other parameters are: $r = 0.9$, $p = 0$.

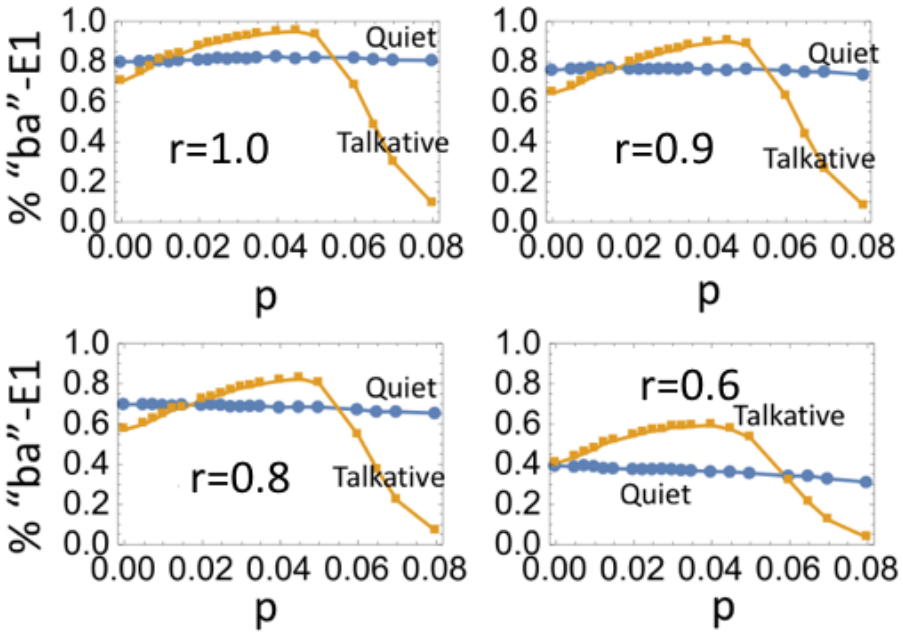

Figure 3.7: Frequency of shared solution with a dominant "ba"-E_1 association for $r = 1$, $r = 0.9$, $r = 0.8$, and $r = 0.6$ as a function of forgetting parameters. "Talkative" species ($L = 100$) are in yellow and "quiet" ($L = 20$) are in blue.

3.5 ODE approach

Consider the matrix of associations that represents the averaged associations of the whole population:

	E_1	E_2	
"ba"	α	β	ϵ_1
"a"	γ	δ	ϵ_2

where

$$\alpha + \beta + \epsilon_1 = 1, \quad \gamma + \delta + \epsilon_2 = 1,$$

(note that instead of using parameter L, we normalized the sum of associations to 1). Suppose that the frequencies of events E_1 and E_2 are Q_1 and Q_2 respectively, with $Q_1 + Q_2 = 1$. The following ODEs describe the average population dynamics of learning:

$$\dot{\alpha} = Q_1 \frac{\alpha}{\alpha+\gamma}\left[\beta^r + (1-\alpha-\beta)^r\right] - Q_2 \frac{\beta}{\beta+\delta}\alpha^r - pH(\alpha)G, \tag{3.1}$$

$$\dot{\beta} = Q_2 \frac{\beta}{\beta+\delta}\left[\alpha^r + (1-\alpha-\beta)^r\right] - Q_1 \frac{\alpha}{\alpha+\gamma}\beta^r - pH(\beta)G, \tag{3.2}$$

$$\dot{\gamma} = Q_1 \frac{\gamma}{\alpha+\gamma}\left[\delta^r + (1-\gamma-\delta)^r\right] - Q_2 \frac{\delta}{\beta+\delta}\gamma^r, \tag{3.3}$$

$$\dot{\delta} = Q_2 \frac{\delta}{\beta+\delta}\left[\gamma^r + (1-\gamma-\delta)^r\right] - Q_1 \frac{\gamma}{\alpha+\gamma}\delta^r, \tag{3.4}$$

where $H(.)$ is the Heaviside function, and $G = \left(Q_1 \frac{\gamma}{\gamma+\alpha} + Q_2 \frac{\delta}{\delta+\beta}\right)$. The right hand side of these equations are the expected increments of the association strengths, obtained by considering all the events that can influence them, with their respective probabilities. For example, the first equation describes changes in the association between "ba" and E_1. This association can change under the following circumstances:

- Event E_1 happens (probability Q_1) and word "ba" is uttered (probability $\frac{\alpha}{\alpha+\gamma}$), and in a two-step update, association "ba"-E_2 is reduced (probability

β^r) and/or association "ba"-"nothing" is reduced (probability $\epsilon_1^r = (1-\alpha-\beta)^r$). In these cases, α is strengthened.

- Event E_2 happens (probability Q_2), word "ba" is uttered (probability $\frac{\beta}{\beta+\delta}$), and association "ba"-$E_1$ is reduced (probability α^r). In this case, α is weakened.

- Forgetting of word "ba". This process reduces association "ba"-E_1 if either event E_1 happens and "a" is used (probability $Q_1\frac{\gamma}{\gamma+\alpha}$) or E_2 happens and "a" is used (probability $Q_2\frac{\delta}{\delta+\beta}$). Forgetting happens with probability p, only if association α is nonzero (factor $pH(\alpha)$).

The rest of the equations are constructed in a similar way.

If we use

$$F(z) = z,$$

that is, $r = 1$ as the regularization parameter, the ODEs that describe the learning process are somewhat simplified:

$$\dot{\alpha} = Q_1\frac{\alpha}{\alpha+\gamma}(1-\alpha) - Q_2\frac{\beta}{\beta+\delta}\alpha - pH(\alpha)G, \tag{3.5}$$

$$\dot{\beta} = Q_2\frac{\beta}{\beta+\delta}(1-\beta) - Q_1\frac{\alpha}{\alpha+\gamma}\beta - pH(\beta)G, \tag{3.6}$$

$$\dot{\gamma} = Q_1\frac{\gamma}{\alpha+\gamma}(1-\gamma) - Q_2\frac{\delta}{\beta+\delta}\gamma, \tag{3.7}$$

$$\dot{\delta} = Q_2\frac{\delta}{\beta+\delta}(1-\delta) - Q_1\frac{\gamma}{\alpha+\gamma}\delta. \tag{3.8}$$

3.5.1 No forgetting

In the absence of forgetting ($p = 0$) we have the following steady states:

$$S_0^{(1)}: \quad \alpha = Q_1, \quad \beta = 1 - Q_1, \quad \gamma = \delta = 0, \tag{3.9}$$

$$S_0^{(2)}: \quad \alpha = \beta = 0, \quad \gamma = Q_1, \quad \delta = 1 - Q_1, \tag{3.10}$$

$$S_1: \quad \alpha = 1, \beta = 0, \gamma = 0, \delta = 1, \tag{3.11}$$

$$S_2: \quad \alpha = 0, \beta = 1, \gamma = 1, \delta = 0, \tag{3.12}$$

$$S_3: \quad \beta = 1 - \alpha, \quad \gamma = 2Q_1 - \alpha, \quad \delta = 1 - 2Q_1 + \alpha. \tag{3.13}$$

Solutions $S_0^{(1)}$ and $S_0^{(2)}$ are unstable. Solutions S_1 and S_2 are also unstable except when $Q_1 = \frac{1}{2}$, when they are neutrally stable. S_3 is a neutral family of solutions which is meaningful for

$$2Q_1 - 1 < \alpha < 2Q_1,$$

since γ and δ cannot be negative. It has a zero eigenvalue, eigenvalue $\frac{1}{2}$ of multiplicity 2, and eigenvalue

$$\lambda = \frac{(\alpha - Q_1)^2 - Q_1(1 - Q_1)}{2Q_1(1 - Q_1)}.$$

Solution S_3 is neutrally stable for

$$Q_1 - \sqrt{Q_1(1 - Q_1)} < \alpha < Q_1 + \sqrt{Q_1(1 - Q_1)}. \tag{3.14}$$

We will assume that $\frac{1}{2} \leq Q_1 \leq 1$. In this case condition (3.14) becomes

$$\alpha > Q_1 - \sqrt{Q_1(1 - Q_1)},$$

that is, the "ba"-E_1 association has to be strong enough. Values of Q_1 close to 1 correspond to solutions of type

$$\begin{matrix} * & 0 \\ * & 0 \end{matrix}.$$

Values of Q_1 close to $\frac{1}{2}$ correspond to solutions of type

$$\begin{matrix} \alpha & 1-\alpha \\ 1-\alpha & \alpha \end{matrix}.$$

3.5.2 With forgetting

In the presence of forgetting $p > 0$, let us first assume that at the steady state, $\alpha > 0$ and $\beta > 0$. Then in system (3.5-3.8) the first two equations are modified to

$$\dot{\alpha} = Q_1 \frac{\alpha}{\alpha+\gamma}(1-\alpha) - Q_2 \frac{\beta}{\beta+\delta}\alpha - pG, \tag{3.15}$$

$$\dot{\beta} = Q_2 \frac{\beta}{\beta+\delta}(1-\beta) - Q_1 \frac{\alpha}{\alpha+\gamma}\beta - pG. \tag{3.16}$$

We have the following solution types:

- $S_0^{(1)} : \alpha = Q_1, \beta = 1-Q_1, \gamma = \delta = 0$. This solution is unstable as before.
- Two solutions corresponding to $\gamma = 0$, $\delta = 1$. One of them becomes S_1 for $p = 0$ but has a negative component for $p > 0, Q_1 > \frac{1}{2}$. The other one corresponds to S_3 with $\alpha = 2Q_1$ and is meaningless for $Q_1 > \frac{1}{2}$.
- Two solutions corresponding to $\gamma = 1$, $\delta = 0$. One of them is given by (to the first order of expansion in p)

  $$S_A = \begin{pmatrix} \frac{Q_1 p}{2Q_1-1} & 1 - \frac{Q_1(3Q_1-1)p}{(1-Q_1)(2Q_1-1)} \\ 1 & 0 \end{pmatrix}.$$

  It becomes S_2 for $p = 0$, and can be stable for a subset of $p \in [0,1]$, $Q_1 \in [\frac{1}{2}, 1]$. The other corresponds to S_3 with $\alpha = 2Q_1 - 1$ and can also be stable for a subset of $p \in [0,1]$, $Q_1 \in [\frac{1}{2}, 1]$. Its first order expansion in p is given by

  $$S_B = \begin{pmatrix} 2Q_1 - 1 - \frac{Q_1 p}{2Q_1-1} & 2 - 2Q_1 + \frac{(3Q_1-2)p}{2Q_1-1} \\ 1 & 0 \end{pmatrix}.$$

- Two interior solutions. One of them corresponds to $S_0^{(2)}$ as $p \to 0$ and is unstable. The other one corresponds to S_3 with $\alpha = \frac{1}{2}$ as $p \to 0$ and can be stable for a portion of the relevant domain. Its first order expansion in p is given by

$$S_C = \begin{pmatrix} \frac{1}{2} - p & \frac{1}{2} - p \\ 2Q_1 - \frac{1}{2} - (2Q_1 - 1)p & \frac{3}{2} - 2Q_1 + (2Q_1 - 1)p \end{pmatrix}.$$

To systematically obtain all the equilibria, we need to consider three more cases:

1. $\alpha = 0$, $\beta > 0$. In this case we have system (3.5-3.8) with the first equation replaced with equation (3.15). We have the following equilibria (we denoted $S = \sqrt{1 - 4p}$):

$$S_{D1} = \begin{pmatrix} 0 & \frac{1}{2}(1 - S) \\ \frac{Q_1}{2}(3 - S) & \frac{1}{2}(2 + (-3 + S)Q_1) \end{pmatrix},$$

$$S_{D2} = \begin{pmatrix} 0 & \frac{1}{2}(1 + S) \\ \frac{Q_1}{2}(3 + S) & \frac{1}{2}(2 - (3 + S)Q_1) \end{pmatrix},$$

$$S_{D3} = \begin{pmatrix} 0 & 1 - \frac{pQ_1}{1-Q_1} \\ 1 & 0 \end{pmatrix}.$$

Let us perturb solution S_{D1} (denoted by $\alpha_*, \beta_*, \gamma_*, \delta_*$),

$$\alpha = \alpha_* + \epsilon z_1, \ldots, \delta = \delta_* + \epsilon z_4,$$

where $\epsilon \ll 1$, expand equations (3.5-3.8) into Taylor series in ϵ, and in each equation, only keep the terms up to the highest order in terms of the perturbation. The equation for z_1 becomes

$$\dot{z}_1 = -\frac{2p}{\epsilon(3 - S)} + O(\epsilon^0);$$

the order ϵ^{-1} term on the right comes from $H(\alpha) = H(\epsilon z_1) \sim \epsilon^0$. This equation suggests that any positive perturbation of α will rapidly decay to zero. Once it is zero, the term $H(\alpha) = 0$, and the association strength will no longer decay. Negative perturbations of association strength $\alpha = 0$ are not relevant. Because of these rapid dynamics of the first component, the other three association strengths can be analyzed separately. Performing the standard linear stability analysis (and setting $\alpha = 0$), we obtain the characteristic polynomial of the 3×3 Jacobian to be

$$P(\lambda) \propto \left(\lambda + \frac{2}{3-S}\right)(\lambda^2 + c_1\lambda + c_0),$$

where coefficients c_0 and c_1 satisfy

$$c_0 = \hat{c}_0(p, Q_1)\left(Q_1 - \frac{2}{3-S}\right), \quad c_1 = \hat{c}_1(p, Q_1)\left(\frac{5+S}{6+p} - Q_1\right),$$

and functions $\hat{c}_0(p, Q_1)$ and $\hat{c}_1(p, Q_1)$ are positive for $0 \leq p < 1/4$ and $1/2 < Q_1 < 1$. For stability, we require $c_0 > 0$ and $c_1 > 0$, which results in the condition

$$\frac{2}{3-S} < Q_1 < \frac{5+S}{6+p}, \tag{3.17}$$

see figure 3.8. On the other hand, checking the positivity of the δ component of solution S_{D1}, we obtain condition

$$Q_1 < \frac{2}{3-S},$$

which is incompatible with condition (3.17).

Similar analysis shows that solution S_{D2} is stable and positive if

$$Q_1 < \frac{2}{3+S}.$$

Finally, solution Q_{D3} is stable and positive in region

$$\frac{2}{3+S} < Q_1 < \frac{2}{3-S}.$$

The regions of stability for S_{D1}, S_{D2}, and S_{D3} are shown in figure 3.8.

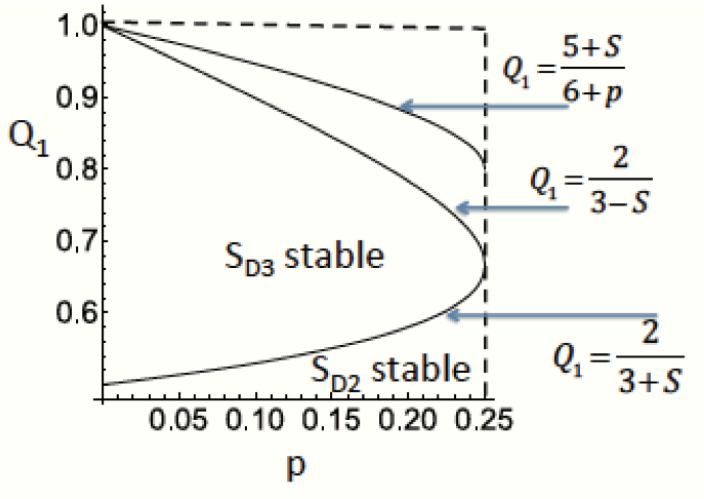

Figure 3.8: Stability regions for solutions S_{D1}, S_{D2}, S_{D3}.

2. $\alpha > 0, \beta = 0$. In this case we have system (3.5-3.8) with the second equation replaced with equation (3.16). We have the following equilibria:

$$S_{E1} = \begin{pmatrix} \frac{1}{2}(1-S) & 0 \\ \frac{1}{2}(-1+S+(3-S)Q_1) & \frac{1}{2}(3-S)(1-Q_1) \end{pmatrix},$$

$$S_{E2} = \begin{pmatrix} \frac{1}{2}(1+S) & 0 \\ \frac{1}{2}(-1-S+(3+S)Q_1) & \frac{1}{2}(3+S)(1-Q_1) \end{pmatrix},$$

$$S_{E3} = \begin{pmatrix} 1-\frac{p(1-Q_1)}{Q_1} & 0 \\ 0 & 1 \end{pmatrix}.$$

Similar to the $\alpha = 0, \beta > 0$ case, a positive perturbation of one of the components (this time, the β association) has a negative ϵ^{-1} term in the right hand side of the corresponding ODE, and rapidly decays to zero. The rest

of the components can be analyzed separately by methods of standard linear analysis. We obtain that solutions S_{E1} and S_{E3} are unstable for $Q_1 > 1/2$ and solution S_{E2} is stable for $Q_1 > 1/2$.

In particular, solution S_{E3} has eigenvalues $-(1-Q_1)$, $-Q_1$, and

$$\frac{p(1-Q_1)^2 + Q_1(2Q_1-1)}{Q_1 - p(1-Q_1)},$$

where the latter quantity is positive as long as $Q_1 > 1/2$ and S_{E3} is positive ($Q_1 > p(1-Q_1)$). Furthermore, the characteristic polynomial for the Jacobian corresponding to solution S_{E1} is given by

$$P(\lambda) \propto \left(\lambda + \frac{2}{3-S}\right)(\lambda^2 + c_1\lambda + c_0),$$

where

$$c_0 = \hat{c}_0(p,Q_1)\left(\frac{1+2p-S}{4+2p} - Q_1\right),$$

and function $\hat{c}_0(p,Q_1)$ is positive for $0 \leq p < 1/4$ and $1/2 < Q_1 < 1$. For stability, we require $c_0 > 0$ and $c_1 > 0$, which results in a condition

$$Q_1 < \frac{1+2p-S}{4+2p},$$

which is violated for $Q_1 > 1/2$. Finally, for S_{E2}, we have

$$P(\lambda) \propto \left(\lambda + \frac{2}{3-S}\right)(\lambda^2 + c_1\lambda + c_0),$$

where

$$c_0 = \hat{c}_0(p,Q_1)\left(Q_1 - \frac{1+2p+S}{4+2p}\right),$$
$$c_1 = \hat{c}_1(p,Q_1)\left(Q_1 - \frac{1+p+S}{6+p}\right),$$

and functions $\hat{c}_0(p,Q_1)$ and $\hat{c}_1(p,Q_1)$ are positive for $0 \leq p < 1/4$ and $1/2 < Q_1 < 1$. For stability, we require $c_0 > 0$ and $c_1 > 0$, which results in

conditions

$$Q_1 > \frac{1+2p+S}{4+2p}, \quad Q_1 > \frac{1+p+S}{6+p},$$

both of which are satisfied for $Q_1 > 1/2$.

3. Finally, setting $\alpha = \beta = 0$, we use system (3.15,3.16,3.7,3.8), to obtain solution $S_0^{(2)}$:

$$S_0^{(2)}: \quad \begin{pmatrix} 0 & 0 \\ Q_1 & 1-Q_1 \end{pmatrix}.$$

Stability analysis can be performed as follows. The first two equations are (to the highest order):

$$\dot{z}_1 = -p/\epsilon + O(\epsilon^0), \quad \dot{z}_2 = -p/\epsilon + O(\epsilon^0),$$

and the positive perturbations decay to zero. The remaining two components satisfy a liner system with a double eigenvalue of -1. Therefore, solution $S_0^{(2)}$ is stable.

It appears that depending on the value of parameter p, different solutions are reached, see Fig. 3.9. For $p = 0$, solution S_3 is observed (with the value α which depends on the initial condition). As p increases, the symmetry is broken and solution S_{E2} becomes stable. As p increases further, this solution loses stability and we observe $S_0^{(2)}$. These dynamics are also shown in Fig. 3.10, where we simulated the ODEs up to a larger time t_0, and plotted the resulting values of the functions α (denoted as "ba"-E_1) and δ (denoted as "a"-E_2) as functions of the parameter p. We can see that the "ba"-E_1 association becomes dominant for an intermediate region of p values, which corresponds to the S_{E2} stationary point.

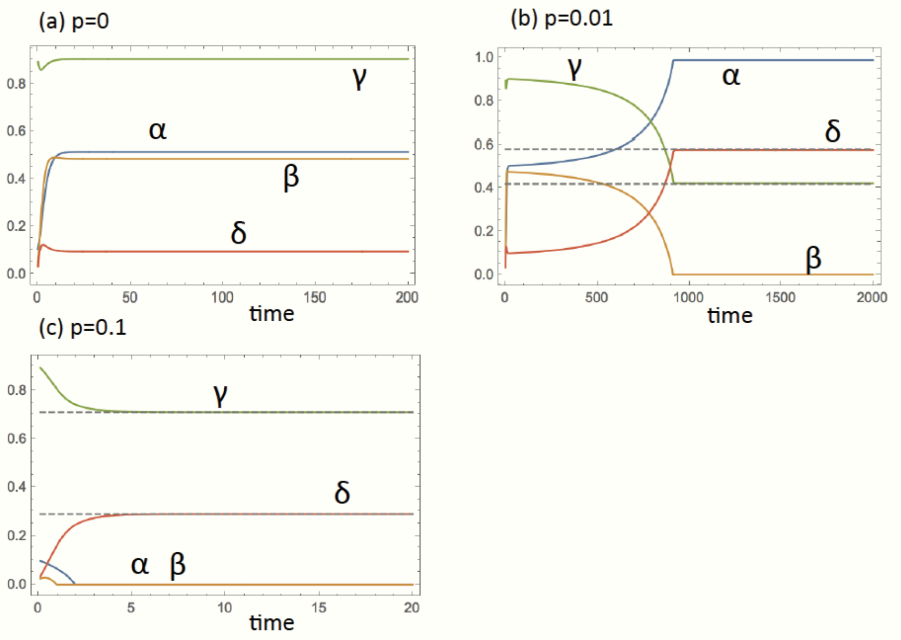

Figure 3.9: ODE simulations of the learning dynamics for different values of p: (a) $p = 0$, (b) $p = 0.01$, (c) $p = 0.1$. In panels (b) and (c), horizontal dashed lines show the values for components γ and δ of steady state solutions S_{E2} and $S_0^{(2)}$, respectively. Parameter $Q_1 = 0.71$ and $t_0 = 2 \times 10^5$.

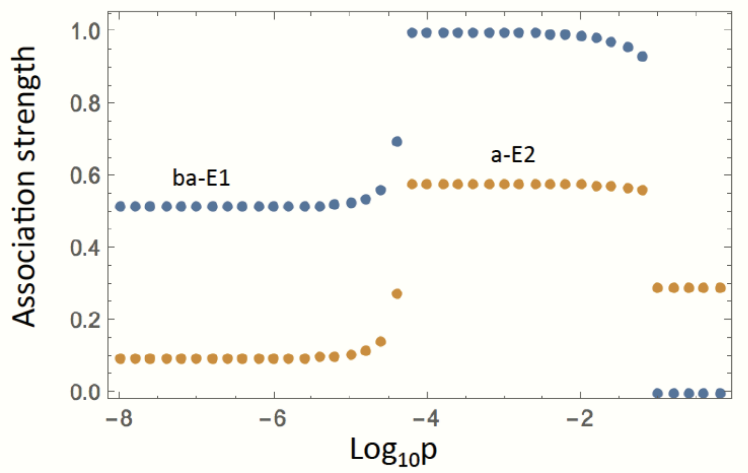

Figure 3.10: Long-term outcomes of the learning dynamics. The values $\alpha(t_0)$ (denoted as "ba"-E_1) and $\delta(t_0)$ (denoted as "a"-E_2), plotted as functions of the parameter p. Parameter $t_0 = 2 \times 10^5$, and the rest of the parameters are as in Fig. 3.9.

3.6 Discussion

In this chapter, we have used mathematical modeling to hypothesize the dynamics of early communication. We established that under very modest assumptions on the interactions of communicating individuals, a new, composite word can rise from low numbers and displace a current innate vocalization for the usage in a frequent context. Interestingly, a degree of increased forgetting of the composite word (compared to the innate) can be beneficial in shaping the communication system where the more complex word gets paired up with the more common event.

It is worth nothing that the model employed in this study is intentionally simplistic. The goal of this study was to identify the minimal number of assumptions (the simplest model) compatible with the possibility of a composite word rising from low usage and displacing the old, innate vocalization. The model used here does demonstrate this behavior, indicating that such scenario can perhaps occur in the natural environment. This reductionist model however does not take account of other numerous factors that may not be necessary for this particular phenomenon, but shape other aspects of learning and communication dynamics. For example, the process of disambiguation of the innate word(after the new word has taken root) takes a relatively long time in our simulations. Such dynamics are a consequence of the specific modeling choice made here. Many other models used in the field contain penalty for inefficient communication imposed through various mechanisms. If we included such features in our model, the long coexistence of two meanings for "a" (and thus a long stretch of inefficient communication) could be significantly shortened. Therefore, we do not propose that the present model realistically describes the timing of change. Instead, it can

be used as a proof of principle, showing that under very minimal assumptions, a new word can arise, and invade in the absence of any externally imposed "fitness" advantage compared to the commonly used innate word.

The following summarizes the findings of this chapter:

- A simplistic model was developed using MCM to describe the evolution of the first word.
  - There are three parameters to describe the behavior of our learning environment: the regularization parameter, r, the forgetting parameter, p, and the discretization parameter, L.
- "ba"-E_1 (the association of the new word with the more frequent event) occurs relatively quickly (if at all).
  - Disambiguation (where "ba" is associated with with E_1 and "a" is associated with E_2) occurs only for the values of the regularization parameter smaller than a threshold.
- The likelihood of "ba"-E_1 association dominating reaches a maximum for a nonzero forgetting parameter, p.
- Rise of the new word occurs faster and at a higher probability for "quiet" species than "talkative" species.

In regards to Zipf's Law of Abbreviation it is shown that the most frequent word that occurs is the shortest word in terms of length [7, 78, 79]. However, from our simplistic model and communication system described in section 3.3, we can see that a population starts with an innate simple vocalization and then inherits a complex consonant-vowel word. The population even associates the new,

complex word with the more frequent event. The likelihood of such a behavior is dependent on a regularization factor as well as a forgetting characteristic. These observations (although in a simplistic model) gives an example of a system that violates Zipf's law. More observation, experimentation, and analysis for a more complex model is subject for future work.

Appendix

"Whatever it is that you do in this life, it's not legendary unless your friends are there to see it."

Barney Stinson, How I Met Your Mother

A.1 Comparing the VIM and the MCM

This section will discuss the comparison of the two methods in two respects: symmetric and asymmetric.

A.1.1 Symmetric Algorithm

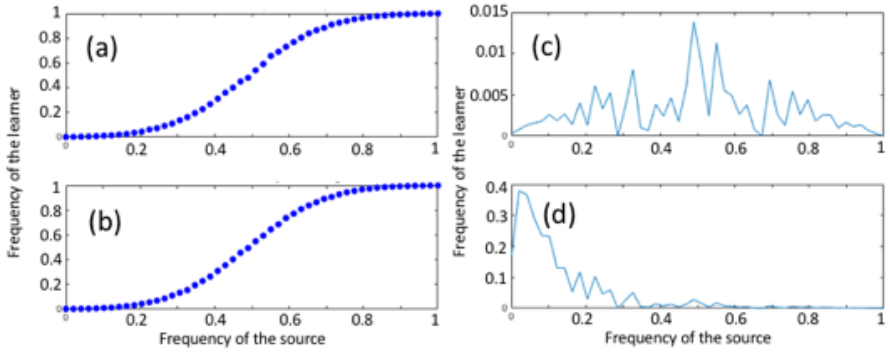

Figure A.1: Comparison of a) VIM and b) MCM where we ran them for $500,000$ time steps over various frequency of the source, ν, from 0 to 1, with $F(X/L) = (X/L)^{0.4}$. And $L = 100$. Plot of (c) relative error and (d) absolute error between VIM and MCM.

To compare the symmetric version methods, we plot the average values of the learner for various ν values from 0 to 1 in Fig. A.1. The average value is noted as approximately the steady state of the learner. Power functions will be used for

the remainder of the discussion.

In Fig. A.1, both methods produce similar shaped plots. The absolute error for each ν between the methods are also plotted in Fig. A.1; note that the absolute error is no more than 0.012 and that the relative error decreases as ν increases.

A.1.2 Asymmetric Algorithm

We now introduce an asymmetric version of VIM and MCM, where the increments and decrements are different depending on whether the source's input value matches the highest-propensity, or preferred form, of the learner.

Let F be defined as above in the formulation of VIM. Let us define the preferred form of the learner as form m, that is $x_m = \max_{i=1,2}\{x_i\}$. Suppose that $x_m = x_1$, if the source emits form 1, then we have the following updates for the VIM version of the algorithm:

$$x_1 \to \begin{cases} x_1 + sF^+(x_1), & \text{if form 1 is received,} \\ x_1 - pF^-(x_1), & \text{if form 2 is received,} \end{cases} \tag{A.1}$$

where F^+ and F^- are some functions of one variable on $[0, 1]$. The rules above can be rewritten for the probability to utter form 2, $x_2 = 1 - x_1$:

$$x_2 \to \begin{cases} x_2 - sF^+(x_2), & \text{if form 1 is received,} \\ x_2 + pF^-(x_2), & \text{if form 2 is received,} \end{cases} \tag{A.2}$$

where s and p are values between 0 and 1. The MCM asymmetric algorithm is just the translation of the VIM asymmetric algorithm into a Markov walk as described in section 1.2 with the transition matrix, P, and function F.

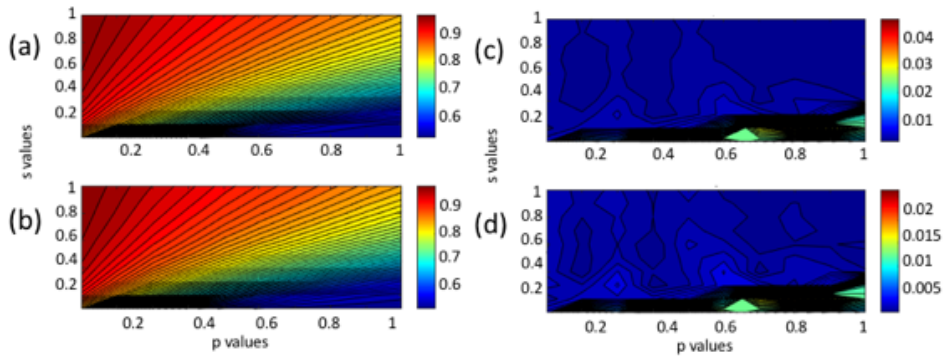

Figure A.2: Plot of (a) AVIM and (b) AMCM, with $\nu = 0.8$, and initial $X = [0.6, 0.4]$. $s, p \in [0.1, 0.2, \ldots, 1]$. Along with the plots of the (c) relative and (d) absolute error between AVIM and AMCM.

For simplicity, we will refer to the asymmetric version of VIM and MCM as AVIM and AMCM respectively. To compare AVIM and AMCM, we executed each algorithm with a fixed s and p value, given function $(x^{0.4})$ and calculated the average learner values as we did with the symmetric case. Fig. A.2-A.3 are the contour plots comparing both methods over various s and p values.

A.1.2.1 Observation: Differences in initial learner values

An important observation here is that updates occur more frequently for certain values of s and p when the learner is reassured, otherwise updates occur less frequently. Fig. A.2 is the plot of AVIM and AMCM where the learner and the source both prefers form 1. We can see that when $s \gg p$, the learner will experience frequency boosting to nearly 1. This is due to the fact that in the algorithm, we use the "preferred" update when the source utters a form that coincides with the preferred form of the learner; the learner is reassured and takes a much bigger increment of update for its frequency of usage than it would otherwise.

Fig. A.3 is the plot of AVIM and AMCM when the learner and the source prefers different forms. Again, we can see that when $s \gg p$, updates are per-

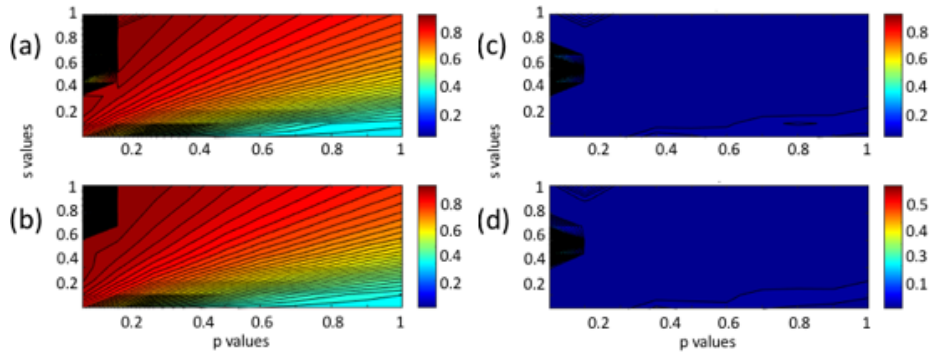

Figure A.3: Plot of (a) AVIM and (b) AMCM, with $\nu = 0.8$, and initial $X = [0.3, 0.7]$. $s, p \in [0.1, 0.2, \ldots, 1]$. Along with plots of (c) relative and (d) absolute errors of AVIM and AMCM.

formed when the learner is reassured that its preferred form is used by the source; otherwise, the frequencies are updated by a significantly smaller amount.

A.1.2.2 Observation: Varying the value of L

The discretization parameter, L, has a role in the stochastic behavior of the updates and ultimately determines when both methods are the same. Note that VIM (and AVIM) algorithm is updating with $\Delta x = F/L$; which is L times slower than MCM (and AMCM). Thus the updates for VIM and AVIM are relatively small increments as opposed to the increments of MCM and AMCM. In Fig. A.4(a), the plots of AVIM has small increments for updates of x_1 and will oscillate around 0.38. The values for X_1 however may oscillate around 0.4 in the early stages but will eventually reach 0.5 given some random probability which then means that the learner's preferred form will be the same as the source and experience the observations in section A.1.2.1. This is also the cause of the discrepancy in the relative and absolute error seen in Fig. A.3.

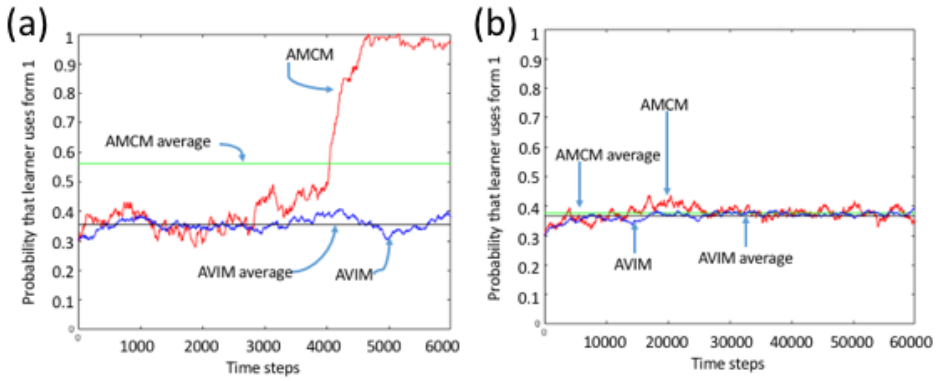

Figure A.4: Plot of AVIM and AMCM for $s = 0.6$, $p = 0.1$ over (a) 6000 time steps with $L = 100$, and (b) 60000 time steps with $L = 500$.

Fig. A.4(b) is AVIM and AMCM with a bigger L value. Since L was increased then X_1 now updates with constant increments of a much smaller expectancy than before, and thus the frequency stabilizes around the same value as what is exhibited in AVIM.

Fig. A.5 is a plot of the max relative and absolute error of the results for AVIM and AMCM when we vary L over all $s, p \in [0, 1]$. We see that as L increases, then we have that the relative error between AVIM and AMCM decreases. Thus as $L \to \infty$, the resulting learner values from the two algorithms coincide.

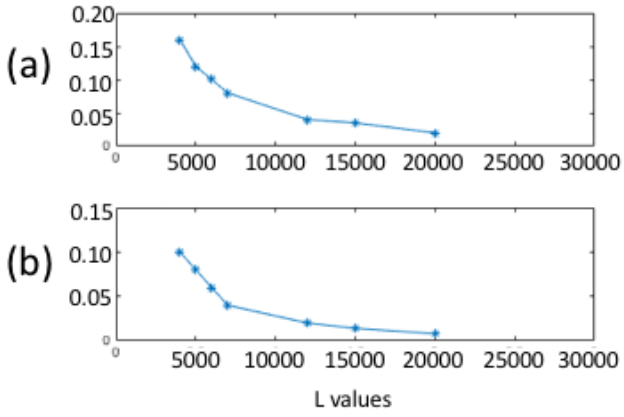

Figure A.5: Plot of L versus (a) relative and (b) absolute error of AVIM and AMCM.

A.2 Proof of Frequency Boosting Theorem

A.2.1 Proof of the theorem for a particular example

Here we set $L = 2$ and prove the Theorem in this simple case. The general proof is provided in section A.2.2. We will use the convenient notation

$$\lambda = \frac{\nu}{1-\nu}.$$

We want to show that $\nu_{learn} > \nu$ when $\nu > \frac{1}{2}$. Thus with the notation above, it suffices to show that $\nu_{learn} > \frac{\lambda}{1+\lambda}$, when $\lambda > 1$. That is

$$\frac{\sum_{i=0}^{L} i\lambda^i C_L^i}{L\sum_{i=0}^{L}\lambda^i C_L^i} > \frac{\lambda}{1+\lambda}. \tag{A.3}$$

For $L = 2$, the LHS of (A.3) becomes:

$$\frac{\lambda\frac{F(1)}{F(\frac{1}{2})} + 2\lambda^2}{2(1+\lambda\frac{F(1)}{F(\frac{1}{2})} + \lambda^2)}.$$

So we want to show that

$$\frac{\lambda\frac{F(1)}{F(\frac{1}{2})} + 2\lambda^2}{2(1+\lambda\frac{F(1)}{F(\frac{1}{2})} + \lambda^2)} > \frac{\lambda}{1+\lambda}.$$

Which is true if and only if

$$\frac{\frac{F(1)}{2F(\frac{1}{2})} + \lambda}{(1+\lambda\frac{F(1)}{F(\frac{1}{2})} + \lambda^2)} > \frac{1}{1+\lambda}.$$

Which is equivalent to

$$(1+\lambda)\Big(\frac{F(1)}{2F(\frac{1}{2})} + \lambda\Big) > (1+\lambda\frac{F(2)}{F(1)} + \lambda^2).$$

After some algebra, we can simplify the expression to

$$\lambda\Big(1 - \frac{F(1)}{2F(\frac{1}{2})}\Big) > 1 - \frac{F(1)}{2F(\frac{1}{2})}. \tag{A.4}$$

Note that $\lambda > 1$, so it suffices to show that

$$1 - \frac{F(1)}{2F(\frac{1}{2})} > 0. \tag{A.5}$$

That is we want to show that

$$\frac{F(\frac{1}{2})}{\frac{1}{2}} > \frac{F(1)}{1}. \tag{A.6}$$

From the definition of a function that is concave down as was described in section 1.3.1, we can conclude that we have (A.6) and thus (A.3) is true when $L = 2$.

A.2.2 The general case

Now to the proof of the Theorem for a general integer L.

Proof. Expanding out (A.3) we have that the frequency boosting property now becomes

$$\frac{\lambda C_L^1 + 2\lambda^2 C_L^2 + \cdots + L\lambda^L}{L(1 + \lambda C_L^1 + \lambda^2 C_L^2 + \cdots + \lambda^L)} > \frac{\lambda}{1+\lambda}, \tag{A.7}$$

which is equivalent to

$$(1+\lambda)\Big(\frac{1}{L}C_L^1 + \frac{2}{L}\lambda C_L^2 + \cdots + \lambda^{L-1}\Big) > 1 + \lambda C_L^1 + \lambda^2 C_L^2 + \cdots + \lambda^L.$$

Simplifying we get,

$$\frac{1}{L}C_L^1 + \frac{2}{L}\lambda C_L^2 + \cdots + \lambda^{L-1} + \frac{1}{L}\lambda C_L^1 + \frac{2}{L}\lambda^2 C_L^2 + \cdots + \lambda^L > 1 + \lambda C_L^1 + \lambda^2 C_L^2 + \cdots + \lambda^L.$$

Combining like terms, we get,

$$\sum_{k=1}^{L-1} \lambda^k \Big(\frac{k+1}{L} C_L^{k+1} - C_L^k(1 - \frac{k}{L})\Big) > 1 - \frac{C_L^1}{L}. \tag{A.8}$$

For notation, let

$$\alpha_{k+1} = \frac{k+1}{L} C_L^{k+1} - C_L^k\Big(1 - \frac{k}{L}\Big) = \frac{k+1}{L} C_L^{k+1} - \frac{L-k}{L} C_L^k. \tag{A.9}$$

Note that $1 - \frac{C_L^1}{L} = \frac{(L-1)+1}{L} C_L^{(L-1)+1} - \frac{L-(L-1)}{L} C_L^{L-1} = \alpha_{(L-1)+1} = \alpha_L$. So then (A.8) becomes

$$\sum_{k=1}^{L-1} \lambda^k \alpha_{k+1} > \alpha_L.$$

Since $\alpha_L = 1 - \frac{C_L^1}{L}$, then by similar reasoning as in the $L = 2$ case, with $x = 1$ and $y = L$, in (A.6), we have that $\frac{C_L^1}{L} < 1$, that is $\alpha_L > 0$.

Since $\lambda > 1$, and $\alpha_L > 0$, we have that $\lambda^{L-1}\alpha_L > \alpha_L$, thus to show (A.8), it suffices to show

$$\lambda\alpha_2 + \lambda^2\alpha_3 + \cdots + \lambda^{L-2}\alpha_{L-1} > 0. \tag{A.10}$$

We can make more simplifications to (A.10) with the notation of (1.9). Note that $C_L^{L-k} = C_L^k$, thus we have that (A.9) becomes

$$\alpha_{k+1} = \frac{k+1}{L} C_L^{L-(k+1)} - \frac{L-k}{L} C_L^{L-k}, \tag{A.11}$$

which can also be written as

$$-\alpha_{k+1} = \frac{L-k}{L} C_L^{L-k} - \frac{L-(L-k-1)}{L} C_L^{L-k-1}. \tag{A.12}$$

From (A.9), by replacing the $k+1$ with $L-k$, we have that

$$\alpha_{L-k} = \frac{L-k}{L} C_L^{L-k} - \frac{L-(L-k-1)}{L} C_L^{L-k-1}. \tag{A.13}$$

Thus

$$\alpha_{L-k} = -\alpha_{k+1}. \tag{A.14}$$

Therefore, (A.10) simplifies into two cases, when L is even or odd.

Suppose first that $L = 2j$ for some integer j. Then using (A.14), we have that (A.10) simplifies to showing that

$$\alpha_2(\lambda - \lambda^{2j-2}) + \alpha_3(\lambda^2 - \lambda^{2j-3}) + \cdots + \alpha_j(\lambda^{j-1} - \lambda^{2j-j}) > 0.$$

Which is equivalent to showing

$$\alpha_2\lambda(1 - \lambda^{2j-3}) + \alpha_3\lambda^2(1 - \lambda^{2j-5}) + \cdots + \alpha_j\lambda^{j-1}(1 - \lambda) > 0. \tag{A.15}$$

Similarly when $L = 2j+1$ for some integer j. Then (A.10) simplifies to showing that

$$\alpha_2(\lambda - \lambda^{2j-1}) + \alpha_3(\lambda^2 - \lambda^{2j-2}) + \cdots + \alpha_j(\lambda^{j-1} - \lambda^{2j-(j-1)}) + \alpha_{j+1}\lambda^j > 0.$$

Note that since $L = 2j + 1$, we have that $\alpha_{L-j} = \alpha_{j+1}$ and that $\alpha_{L-j} = -\alpha_{j+1}$ from (A.14), thus $\alpha_{j+1} = 0$. Thus for the odd case, (A.10) simplifies to showing that

$$\alpha_2\lambda(1 - \lambda^{2j-2}) + \alpha_3\lambda^2(1 - \lambda^{2j-4}) + \cdots + \alpha_j\lambda^{j-1}(1 - \lambda^2) > 0. \tag{A.16}$$

Thus for either L is even or odd, from (A.15) and (A.16), it suffices to show that

$$\alpha_i < 0,$$

for $i = 2, 3, \ldots, j$, since $1 - \lambda^m < 0$ for any m, because $1 < \lambda$.

From (1.9), we can also write

$$C_L^i = C_L^{i-1}\frac{F(\frac{L-i+1}{L})}{F(\frac{i}{L})}.$$

Thus after simplification we have that (A.9) becomes

$$\alpha_i = C_L^{i-1}\Big(\frac{iF(\frac{L-i+1}{L})}{LF(\frac{i}{L})} - \frac{L-(i-1)}{L}\Big).$$

Because $F(x) > 0$ for all $x > 0$, then $C_L^{i-1} > 0$, then it suffices to show that

$$\frac{iF(\frac{L-i+1}{L})}{LF(\frac{i}{L})} - \frac{L-(i-1)}{L} < 0.$$

That is, it suffices to show that

$$\frac{F(\frac{L-i+1}{L})}{\frac{L-i+1}{L}} < \frac{F(\frac{i}{L})}{\frac{i}{L}}, \tag{A.17}$$

for $i = 2, 3, \ldots, j$, where $L = 2j$ or $L = 2j + 1$.

Using the same argument made in the $L = 2$ case, we have (A.17), and therefore can conclude that $\alpha_i < 0$, for $i = 2, 3, \ldots, j$ where $L = 2j$ or $L = 2j + 1$, and therefore can conclude (A.15)-(A.16), and thus have (A.10). Hence (A.3) is true for any integer L and for a function F that satisfies the properties listed in section 1.3.1. □

A.3 The matrix method: more examples

A.3.1 Frequency of usage

Now let us modify the experiment and assume that the A-B and B-C pairs appear different numbers of times. Let us assume that A-B comes up N times and B-C comes up M times, and $M > N$. The question again is to identify what is "dax", "pid", and "wug". We have the following algorithm:

	A	B	C
dax	N	N	ϵ_1
pid	ϵ_2	M	M
wug	ϵ_3	ϵ_3	ϵ_3

$\Rightarrow$

	A	B	C
dax	1	$\frac{N}{N+M}$	$\frac{\epsilon_1}{M}$
pid	$\frac{\epsilon_2}{N}$	$\frac{M}{N+M}$	1
wug	$\frac{\epsilon_3}{N}$	$\frac{\epsilon_3}{N+M}$	$\frac{\epsilon_3}{M}$

$\Rightarrow$

	A	B	C
dax	$*$		
pid			$*$
wug	$*$		

We can see now that when it comes to identifying "wug", object A wins because it appears less often and thus is more informative than object C.

A.3.2 More objects

Let us keep the rules the same as before, but among the responders' choices, include an extra object, D, than has not appeared during the training sessions. Now the result is slightly different:

	A	B	C	D
dax	N	N	ϵ_1	ϵ_1
pid	ϵ_2	M	M	ϵ_2
wug	ϵ_3	ϵ_3	ϵ_3	ϵ_3

$\Rightarrow$

	A	B	C	D
dax	1	$\frac{N}{N+M}$	$\frac{\epsilon_1}{M}$	$\frac{\epsilon_1}{\sum \epsilon_i}$
pid	$\frac{\epsilon_2}{N}$	$\frac{M}{N+M}$	1	$\frac{\epsilon_2}{\sum \epsilon_i}$
wug	$\frac{\epsilon_3}{N}$	$\frac{\epsilon_3}{N+M}$	$\frac{\epsilon_3}{M}$	$\frac{\epsilon_3}{\sum \epsilon_i}$

$\Rightarrow$

	A	B	C	D
dax	$*$			
pid			$*$	
wug				$*$

Note that in the third row of the middle matrix, the last element is always the largest because it is ~ 1, while the other elements are $\sim \epsilon_i$. We can see that given a new object, the new word is naturally associated with that object.

A.3.3 Finding a label for an object

Finally, we change the task and assume that the responders are asked to produce a word that describes a given object. Now, we have a different procedure because labels are now competing for associations with objects. Therefore, the first step is to normalize the matrix by rows, and then to pick the largest element in each column.

	A	B	C	D
dax	N	N	ϵ_3	ϵ_4
pid	ϵ_1	M	M	ϵ_4
wug	ϵ_1	ϵ_2	ϵ_3	ϵ_4

$\Rightarrow$

	A	B	C	D
dax	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{\epsilon_3}{2N}$	$\frac{\epsilon_4}{2N}$
pid	$\frac{\epsilon_1}{2M}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{\epsilon_4}{2M}$
wug	$\frac{\epsilon_1}{\sum \epsilon_i}$	$\frac{\epsilon_2}{\sum \epsilon_i}$	$\frac{\epsilon_3}{\sum \epsilon_i}$	$\frac{\epsilon_4}{\sum \epsilon_i}$

$\Rightarrow$

	A	B	C	D
dax	$*$	$*$		
pid		$*$	$*$	
wug				$*$

Again, in the rightmost column of the middle matrix, the largest element will correspond to "wug", because it is ~ 1. We can see that when asked to match an object to a given word, the subjects are expected to pick either "dax" or "pid" for B. For the novel object D, the subjects are expected to choose the novel word "wug". If however this word is not among the possible choices, then the word "dax" is expected to be chosen, because it has been heard fewer times ($N < M$). If, as in the original experiment, $N = M$, then the words "dax" and "pid" will be chosen to describe D an equal number of times.

A.4 Stochastic algorithm: additional properties

Let us consider the stochastic algorithm and the behavior of the matrices, for different values of μ. As we have noted from our model in explaining the experimental results, OL and LO learners utilize negative evidence in different ways when trying to associate objects to labels versus labels to objects. According to our model, varying μ values means varying occurrence of negative update. When $\mu = 0$, then the negative update will not occur and we are only performing the basic MCM algorithm. When $\mu = 1$, then negative update will always occur. To describe what happens at different μ values, suppose the learner is acquiring information from the following training matrix:

$$\begin{pmatrix} N & N \\ 0 & N \end{pmatrix}. \tag{A.18}$$

This matrix details that the training set will contain N pairings of Obj1 with the label A, N pairings of Obj2 with label A, and N pairings of Obj2 with label B.

If $\mu = 1$, then the resulting normed P matrix after implementing the algorithm will be:

$$\begin{pmatrix} \delta & 1-\delta \\ 1-\delta & \delta \end{pmatrix}, \tag{A.19}$$

where δ is determined by the update function F of the stochastic learning model with negative evidence.

We can see that for a 2×2 matrix P:

$$\begin{pmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{pmatrix}.$$

according to the algorithm described in section 2.1.3, if the learner sees object 2 and then hears label 1 then they will perform the following MCM update on column 2,

$$\begin{aligned} \alpha_2 &\to \alpha_2 + 1, \\ \beta_2 &\to \beta_2 - 1. \end{aligned} \tag{A.20}$$

Then, with the negative update on column 1, they will perform

$$\begin{aligned} \alpha_1 &\to \alpha_1 - 1, \\ \beta_1 &\to \beta_1 + 1. \end{aligned} \tag{A.21}$$

That is, the diagonal entries of P (α_1 and β_2) update simultaneously, at the same rate and magnitude. Similar examples have the same observation.

We can summarize the observations of this matrix:

- the diagonal entries of (A.19) are the same.

- the resulting normed Q matrix will be the same as (A.19).
- Learners trying to answer "what is object" questions will give the same corresponding answer to "what is label" questions.

Note that as μ varies from 0 to 1, the first component of the resulting normed P matrix will decrease from 1 to δ. This is due to the nature of the negative evidence which will allow learners to classify more accurately objects to their labels and vice versa.

Bibliography

[1] Roger W Andersen. *Pidginization and Creolization as Language Acquisition.* ERIC, 1983.

[2] Jacques Arends. Towards a gradualist model of creolization. In *Atlantic meets Pacific: a global view of pidginization and creolization*, pages 371–380, Amsterdam, 1993. John Benjamins.

[3] F Gregory Ashby, W Todd Maddox, and Corey J Bohil. Observational versus feedback training in rule-based and information-integration category learning. *Memory & cognition*, 30(5):666–677, 2002.

[4] Lawrence W Barsalou. Deriving categories to achieve goals. *Psychology of Learning and Motivation*, 27:1–64, 1991.

[5] Jonathan Bendor, Dilip Mookherjee, and Debraj Ray. Aspiration-based reinforcement learning in repeated interaction games: An overview. *International Game Theory Review*, 3(02n03):159–174, 2001.

[6] Jonathan Bendor, Daniel Diermeier, and Michael Ting. A behavioral model of turnout. *American Political Science Review*, 97(02):261–280, 2003.

[7] Christian Bentz and Ramon Ferrer i Cancho. *Zipf's Law of Abbreviation as a Language Universal*. Universitätsbibliothek Tübingen, 2016.

[8] Jean Berko. The child's learning of english morphology. *Word*, 14(2-3):150–177, 1958.

[9] Matthew M Botvinick, Yael Niv, and Andrew C Barto. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262–280, 2009.

[10] Jerome R Busemeyer and Timothy J Pleskac. Theoretical tools for understanding and aiding dynamic decision making. *Journal of Mathematical Psychology*, 53(3):126–138, 2009.

[11] Robert R Bush and Frederick Mosteller. *Stochastic models for learning*. John Wiley & Sons, Inc., 1955.

[12] Robert R Bush and Frederick Mosteller. *A mathematical model for simple learning*. Springer, 2006.

[13] Colin Camerer. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press, 2003.

[14] Noam Chomsky. Lectures on government and binding. *Dordrecht: Foris*, 1981.

[15] Morten H Christiansen and Simon Kirby. Language evolution: Consensus and controversies. *Trends in cognitive sciences*, 7(7):300–307, 2003.

[16] Morten H Christiansen and Simon Kirby. Language evolution: the hardest problem in science? *Studies in the Evolution of Language*, 3:1–15, 2003.

[17] John Duffy. Agent-based models and human subject experiments. *Handbook of computational economics*, 2:949–1011, 2006.

[18] Ido Erev and Alvin E Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, pages 848–881, 1998.

[19] Maryia Fedzechkina, T Florian Jaeger, and Elissa L Newport. Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44):17897–17902, 2012.

[20] Lanny Fields, Kenneth F Reeve, Priya Matneja, Antonios Varelas, James Belanich, Adrienne Fitzer, and Kim Shamoun. The formation of a generalized categorization repertoire: Effect of training with multiple domains, samples, and comparisons. *Journal of the Experimental Analysis of Behavior*, 78(3):291–313, 2002.

[21] Christopher D Fiorillo, Philippe N Tobler, and Wolfram Schultz. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614): 1898–1902, 2003.

[22] Andreas Flache and Michael W Macy. Stochastic collusion and the power law of learning a general reinforcement learning model of cooperation. *Journal of Conflict Resolution*, 46(5):629–653, 2002.

[23] James H Fowler. Habitual voting and behavioral turnout. *Journal of Politics*, 68(2): 335–344, 2006.

[24] Anne S Hsu, Nick Chater, and Paul Vitányi. Language learning from positive evidence, reconsidered: A simplicity-based approach. *Topics in Cognitive Science*, 5(1): 35–55, 2013.

[25] Carla L Hudson Kam and Elissa L Newport. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2):151–195, 2005.

[26] Luis R Izquierdo, Segismundo S Izquierdo, Nicholas M Gotts, and J Gary Polhill. Transient and asymptotic dynamics of reinforcement learning in games. *Games and Economic Behavior*, 61(2):259–276, 2007.

[27] Carla L Hudson Kam and Elissa L Newport. Getting it right by getting it wrong: When learners change languages. *Cognitive psychology*, 59(1):30–66, 2009.

[28] Natalia Komarova and Partha Niyogi. Optimizing the mutual intelligibility of linguistic agents in a shared world. *Artificial Intelligence*, 154(1-2):1–42, 2004.

[29] Adriano R Lameira, Ian Maddieson, and Klaus Zuberbühler. Primate feedstock for the evolution of consonants. *Trends in cognitive sciences*, 18(2):60–62, 2014.

[30] Kimery R Levering and Kenneth J Kurtz. Observation versus classification in supervised category learning. *Memory & cognition*, 43(2):266–282, 2015.

[31] John D Lewis and Jeffrey L Elman. Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th Annual Conference on Language Development*, 2001.

[32] Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A Nowak. Quantifying the evolutionary dynamics of language. *Nature*, 449(7163): 713–716, 2007.

[33] Timmy Ma and Natalia L Komarova. Mathematical modeling of learning from an inconsistent source: A nonlinear approach. *Bulletin of mathematical biology*, 79(3): 635–661, 2017.

[34] Yelena Mandelshtam and Natalia L Komarova. When learners surpass their models: mathematical modeling of learning from an inconsistent source. *Bulletin of mathematical biology*, 76(9):2198–2216, 2014.

[35] Patrick-André Mather. Second language acquisition and creolization: Same (i-) processes, different (e-) results. *Journal of Pidgin and Creole Languages*, 21(2):231–274, 2006.

[36] Mark A McDaniel, Katherine Hannah Nuefeld, and Sandra Damico-Nettleton. Many-to-one and one-to-many associative learning in a naturalistic task. *Journal of Experimental Psychology: Applied*, 7(3):182, 2001.

[37] Padraic Monaghan, Laurence White, and Marjolein M Merkx. Disambiguating durational cues for speech segmentation. *The Journal of the Acoustical Society of America*, 134(1):EL45–EL51, 2013.

[38] Dilip Mookherjee and Barry Sopher. Learning behavior in an experimental matching pennies game. *Games and Economic Behavior*, 7(1):62–91, 1994.

[39] Dilip Mookherjee and Barry Sopher. Learning and decision costs in experimental constant sum games. *Games and Economic Behavior*, 19(1):97–132, 1997.

[40] James D Moran and John C McCullers. Reward and number of choices in children's probability learning: An attempt to reconcile conflicting findings. *Journal of Experimental Child Psychology*, 27(3):527–532, 1979.

[41] Roland Mühlenbernd and Jonas David Nick. Language change and the force of innovation. In *Pristine Perspectives on Logic, Language, and Computation*, pages 194–213. Springer, 2014.

[42] Kumpati S Narendra and Mandayam AL Thathachar. *Learning automata: an introduction*. Courier Dover Publications, 2012.

[43] Yael Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154, 2009.

[44] Partha Niyogi. *The computational nature of language learning and evolution*. MIT press Cambridge, MA:, 2006.

[45] M.F. Norman. *Markov Processes and Learning Models*. Academic Press, New York, 1972.

[46] Martin A Nowak, Natalia L Komarova, and Partha Niyogi. Evolution of universal grammar. *Science*, 291(5501):114–118, 2001.

[47] Amy Perfors. Adult regularization of inconsistent input depends on pragmatic factors. *Language Learning and Development*, 12(2):138–155, 2016.

[48] Martijn Van de Pol and Andrew Cockburn. Identifying the critical climatic time window that affects trait expression. *The American Naturalist*, 177(5):698–707, 2011.

[49] Michael Ramscar, Daniel Yarlett, Melody Dye, Katie Denny, and Kirsten Thorpe. The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34(6):909–957, 2010.

[50] Michael Ramscar, Melody Dye, and Joseph Klein. Children value informativity over logic in word learning. *Psychological science*, 24(6):1017–1023, 2013.

[51] Michael Ramscar, Melody Dye, and Stewart M McCauley. Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, 89(4): 760–793, 2013.

[52] Florencia Reali and Thomas L Griffiths. The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, 111 (3):317–328, 2009.

[53] RA Rescorla and A Wagner. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black and W.F. Prokasy, editors, *Classical Conditioning II: Current Research and Theory*. Appleton-Century-Crofts, New York, 1972.

[54] Robert A Rescorla. Probability of shock in the presence and absence of cs in fear conditioning. *Journal of comparative and physiological psychology*, 66(1):1, 1968.

[55] Robert A Rescorla. Pavlovian conditioning: It's not what you think it is. *American Psychologist*, 43(3):151, 1988.

[56] Robert A Rescorla, Allan R Wagner, et al. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 2:64–99, 1972.

[57] Jose JF Ribas-Fernandes, Alec Solway, Carlos Diuk, Joseph T McGuire, Andrew G Barto, Yael Niv, and Matthew M Botvinick. A neural signature of hierarchical reinforcement learning. *Neuron*, 71(2):370–379, 2011.

[58] Jacquelyn L Rische and Natalia L Komarova. Regularization of languages by adults and children: A mathematical framework. *Cognitive Psychology*, 84:1–30, 2016.

[59] Douglas LT Rohde and David C Plaut. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72(1):67–109, 1999.

[60] Alvin E Roth and Ido Erev. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and economic behavior*, 8(1):164–212, 1995.

[61] Deb K Roy and Alex P Pentland. Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146, 2002.

[62] Wolfram Schultz. Behavioral theories and the neurophysiology of reward. *Annu. Rev. Psychol.*, 57:87–115, 2006.

[63] Wolfram Schultz. Behavioral dopamine signals. *Trends in neurosciences*, 30(5):203–210, 2007.

[64] Thomas C Scott-Phillips. Evolutionary psychology and the origins of language: (editorial for the special issue of journal of evolutionary psychology on the evolution of language). *Journal of Evolutionary Psychology*, 8(4):289–307, 2010.

[65] Mark S Seidenberg. Language acquisition and use: Learning and applying probabilistic constraints. *Science*, 275(5306):1599–1603, 1997.

[66] Amanda Seidl and Elizabeth K Johnson. Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science*, 9(6):565–573, 2006.

[67] Ann Senghas. The development of nicaraguan sign language via the language acquisition process. In *Proceedings of the 19th annual Boston University conference on language development*, pages 543–552. Boston: Cascadilla Press, 1995.

[68] Ann Senghas and Marie Coppola. Children creating language: How nicaraguan sign language acquired a spatial grammar. *Psychological Science*, 12(4):323–328, 2001.

[69] Ann Senghas, Marie Coppola, Elissa L Newport, and Ted Supalla. Argument structure in nicaraguan sign language: The emergence of grammatical devices. In *Proceedings of the 21st Annual Boston University Conference on Language Development*, volume 2, pages 550–561, 1997.

[70] Murray Sidman. *Equivalence relations and behavior: A research story*. Authors Cooperative Boston, 1994.

[71] Kenny Smith and Elizabeth Wonnacott. Eliminating unpredictable variation through iterated learning. *Cognition*, 116(3):444–449, 2010.

[72] Luc Steels. Language as a complex adaptive system. In *Parallel Problem Solving from Nature PPSN VI*, pages 17–26. Springer, 2000.

[73] Harold W Stevenson and Kenneth L Hoving. Probability learning as a function of age and incentive. *Journal of Experimental Child Psychology*, 1(1):64–70, 1964.

[74] Richard L Street Jr, Nancy James Street, and Anne Van Kleek. Speech convergence among talkative and reticent three year-olds. *Language Sciences*, 5(1):79–96, 1983.

[75] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. Cambridge Univ Press, 1998.

[76] Morton W Weir. Developmental changes in problem-solving strategies. *Psychological review*, 71(6):473, 1964.

[77] Thomas R Zentall, Mark Galizio, and Thomas S Critchfield. Categorization, concept learning, and behavior analysis: An introduction. *Journal of the experimental analysis of behavior*, 78(3):237–248, 2002.

[78] George Kingsley Zipf. The psycho-biology of language. 1935.

[79] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.