Stability in Stochastic Language Change Models

by

Benjamin Brandon Goodman

M.Sc., McMaster University, 2013
B.Sc., McMaster University, 2012

Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Mathematics
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2017

Approval

Name:	**Benjamin Brandon Goodman**
Degree:	**Doctor of Philosophy (Mathematics)**
Title:	***Stability in Stochastic Language Change Models***
Examining Committee:	**Chair:** Steven Ruuth Professor

Paul Tupper
Senior Supervisor
Professor _____

John Stockie
Committee Member
Professor _____

Ben Adcock
Internal Examiner
Assistant Professor _____

David F. Anderson
External Examiner
Associate Professor
Department of Mathematics
University of Wisconsin _____

Date Defended: 24 July 2017

Abstract

Exemplar models are a popular class of models used to describe language change. Exemplars are detailed memories of stimuli people are exposed to, and when modelling language change are represented as vectors where each component is a phonetic variable. Each exemplar is given a category label, representing what that sound is identified as. New sounds are categorized based on how close they are to the exemplars in each category. Newly categorized exemplars become a part of the system and affect how the future sounds are produced and perceived.

It is possible in certain situations in language for a category of sound to become extinct, such as a pronunciation of a word. One of the successes of exemplar models has been to model extinction of sound categories. The focus of this dissertation will be to determine whether categories become extinct in certain exemplar models and why.

The first model we look at is an exemplar model which is an altered version of a k-means clustering algorithm by MacQueen. It models how the category regions in phonetic space vary over time among a population of language users. For this particular model, we show that the categories of sound will not become extinct: all categories will be maintained in the system for all time. Furthermore, we show that the boundaries between category regions fluctuate and we quantitatively study the fluctuations in a simple instance of the model.

The second model we study is a simple exemplar model which can be used to model direct competition between categories of sound. Our aim in investigating this model is to determine how limiting the memory capacity of an individual in exemplar models affects whether categories become extinct. We will prove for this model that all the sound categories but one will always become extinct, whether memory storage is limited or not.

Lastly, we create a new model that implements a bias which helps align all the categories in the phonetic space, using the framework of an earlier exemplar model. We make an argument that this exemplar model does not have category extinction.

Keywords: Exemplar Models; Exemplar Dynamics; Stochastics; Language Change; Linguistics; Extinction

Dedication

To my parents, Don and Jamie, for their love and support, and without whom I would not be the person I am today. To my sister, Megan, for her love and support.

Acknowledgements

I would like to thank my supervisor Paul Tupper. He has been an amazing mentor, role model, and coworker over the last 4 years. I cannot imagine having had a better supervisor for my PhD. He has been fantastic to work for, and I consider myself lucky to have had the opportunity to learn from him and work with him.

I would like to thank my committee member John Stockie. Taking part in his research group meetings has been useful and insightful. He has given me sound career advice going into the workplace.

I would like to thank Marni Mishna for helping me with my writing. Her course on communicating research has been a great benefit to my work.

Last but not least, I would like to thank all of my family and friends for their support. You know everything I've been through to get here and have been there for me all the way. Thank you for your support, patience, advice, comic relief, and for believing in me even when I doubt myself.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Exemplar Models and Extinction

In linguistics, exemplar models have been shown to effectively model how humans perceive and produce sounds in speech, as well as the evolution of speech. In exemplar models each individual stores memories of sounds/words they have heard, and uses those memories to speak sounds/words and identify new ones. Each memory is given a category label which is its identifier; for example, if classifying vowel sounds, the category labels will be the different vowels. As an example, in an exemplar model, each time the words "bet" and "bit" are heard by someone, the occurrences of their vowels are stored as memories. Newly stored sounds affect the production and perception processes of categories. The production of exemplars is affected because the speaker draws from the exemplars they have stored to produce new sounds. The listener classifies sounds based on their stored exemplars. These models evolve via a production-perception loop [24]. Many of these exemplar models have not been studied with mathematical rigour.

In this thesis, we work towards determining the mathematical properties for some of these models. We have been particularly interested in whether categories become extinct in these models and why. In our research, we have determined how changing certain features of these models alter whether categories of sound become extinct.

Category extinction is a major feature of language change. For some English speakers, the pronunciation of the words "caught" and "cot" have merged, and are indistinguishable when spoken [33]. Exemplar models have been used to successfully model sound merger (a type of extinction) in [24, 31].

We can also model the evolution of alternative spellings of words in the same way we do pronunciations (see Section 3.1). Figure 1.1, which was generated by Google Ngram Viewer [19], shows in the written lexicon a comparison between the usage of the word "entrenchment" and its archaic spelling "intrenchment". For many years it seems both spellings were used about the same amount. Around 1900 though, "entrenchment" began

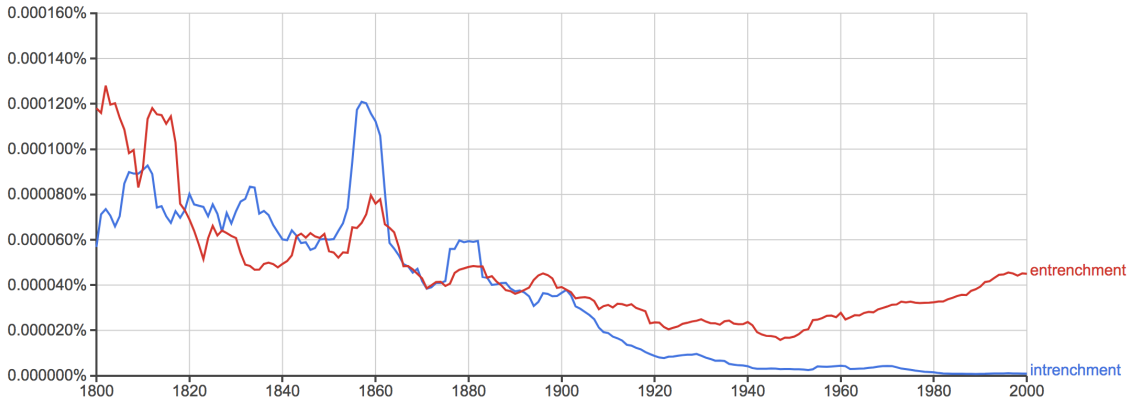

Figure 1.1: Comparison of the usage of "entrenchment" (red) and its archaic spelling "intrenchment' (blue) within a corpus of books between the years 1800 and 2000. The y-axis represents the percentage of the usages of the words in the entire database. This image was generated by Google Ngram Viewer [19].

to become the prominent spelling and "intrenchment" has had a slow decline in usage ever since. By the year 1985, the spelling "intrenchment" became practically extinct. This kind of competition between two pronunciations or spellings of words will be modelled in Chapter 3.

1.2 A Brief History of Exemplar Models in Linguistics

Exemplar models were first introduced by Nosofsky [20, 22]. Nosofsky hypothesized that people store detailed memories of stimuli they are exposed to which he called exemplars [31]. The first successful application of exemplar theory to model speech perception was done by Johnson [11]. Johnson wanted to replicate human perception of vowel sounds using an exemplar model from Nosofsky [21]. In [11], utterances of vowel sounds spoken by a collection of English speakers were treated as exemplars. The exemplars in [11] were vectors where each component represented an acoustic parameter. He was able to alter the parameters in the model so that it classified sounds with an accuracy comparable with humans identifying synthesized sounds as found in [13, 29].

Exemplar dynamics, a special type of exemplar modelling, was introduced in Pierrehumbert [24]. The work done in [24] built on exemplar theory by creating an exemplar model which implemented speech production and not just perception. As such, the model had a production-perception loop between two individuals with their own stored exemplars, allowing us to study the evolutionary dynamics of language over generations [34]. In [24], a target exemplar is picked from the list of catalogued exemplars of a chosen label. People do not perfectly replicate sounds they have from memory, as such the target exemplar is subject to noise. The produced sound also undergoes a lenition, tending towards a preferred

value in the phonetic space, and is biased towards the center of the cloud of exemplars of the chosen category label (entrenchment). Pierrehumbert's results gave a possible explanation for historical shifts (on the order of decades) seen in languages, and provided an example showing exemplar dynamics could be used to model sound mergers [24]. Ever since, many have used exemplar dynamics to model spoken and written language such as [9, 32, 31, 6, 35] to list a few.

Exemplar models being studied to model language change follow the basic layout of the model in [24]. Here we describe that layout. There are usually two people speaking to one another. Each individual has a store of labeled exemplars. At every time step a new sound is produced by a speaker, which is then perceived by the listener and classified based on the listener's stored exemplars. Instead of this, sometimes a population of speakers is represented using one store of exemplars which both produces and perceives sounds as done in [31]. The population could be any group of 2 or more people; the population size that these models emulate best is yet to be studied. The way the sound is produced varies depending on the model. Usually a new sound is produced randomly by adding noise and bias to a pre-existing exemplar like in [24]. The listener usually categorizes sounds based on their 'closeness' to the cloud of exemplars stored for each category. Nosofsky introduced the notion of each exemplar having a weight (or activation) associated with it, representing how predominant or recent the memory of the sound is [20]. In [24], new exemplars start with weights at some fixed value but they decay over time. This represents how older memories will become less influential in the system as time passes. Newly categorized sounds become a part of the perception process, continually evolving the system [24].

Exemplar models are most often studied with one or two stores of exemplars. Exemplar models for language change are meant to study the evolution of language within a population. While studying exemplar models we are making certain assumptions. First, we assume all people within the population have approximately the same distribution of exemplars. Second, we assume that behaviours seen in models with fewer stores of exemplars will still exist in models with a large population. Neither of these assumptions have been justified, and as such it is a topic for future research to investigate.

1.3 Chapter Summaries

In this dissertation three exemplar models will be described, all of which are based on previous exemplar models. Chapters 2 and 3 focus on two different models for which we have determined whether the sound categories eventually become extinct or not. In Chapter 4, we develop a model by combining features from the models in [32] and [31].

The first model, studied in Chapter 2, is an altered version of a k-means clustering algorithm created by MacQueen in [15]. We have taken the model in [15] and introduced memory decay into it. We are able to prove for this model that category extinction cannot

occur. The long term behaviour of the perceptual boundary between classification regions is then studied via a probabilistic model.

Chapter 3 focuses on a simple model which helped us understand the results found in [32]. This simple model only depends on the weights of exemplars which have no phonetic variables associated with them. In this exemplar model, the extinction of categories depends on one category being able to outnumber or outweigh another. The work done here helps us understand the behaviour of the model studied in [32]. This model is unrelated to the work done in Chapter 2.

In Chapter 4, we develop a model with segmental bias. Segmental bias in exemplar models results in the breaking up of vocalizations into distinct units which are re-used [23]. For example this happens with phonemes, which are the different units of sound put together to distinguish words [12]. The result of this phenomenon is that categories end up segmented in the phonetic space. Segmental bias was first introduced into exemplar models by Wedel in [32]. The results found in [31] and in Chapter 4 indicate that the model in [32] is not stable. Combining the work done in [32] and [31], we develop a model incorporating segmental bias which we believe is stable. In the conclusion, we discuss what features of the exemplar models we have studied cause category extinction and why.

1.4 Future Publications and Contributions of the Authors

This dissertation is mostly composed of manuscripts currently in the editing process of publication, with the exception of Chapter 4.

Chapter 2: Based on a paper titled *Stability and Fluctuations in a Simple Model of Phonetic Category Change*, coauthored by Benjamin Goodman and Paul F. Tupper. BG and PFT discussed the development of the model, the methods used in the proofs, and how the behaviour of the perceptual boundary was analyzed. BG worked on and wrote out the proofs, and ran simulations. PFT and BG wrote the manuscript.

Chapter 3: Based on a paper titled *Effects of Limiting Memory Capacity on the Behaviour of Exemplar Dynamics*, coauthored by Benjamin Goodman and Paul F. Tupper. BG and PFT discussed the development of the model, and the methods used in proofs. BG worked on and wrote out the proofs, ran simulations, and wrote the manuscript.

The manuscript in Chapter 2 has for the most part not been changed from how it was originally written, although it is missing its abstract (of which pieces have been used in the second paragraph of the abstract for this dissertation). Chapter 3 has had portions of its introduction removed from it, and some other pieces changed to connect with Chapter 4.

Chapter 2

Stability and Fluctuations in a Simple Model of Phonetic Category Change

2.1 Introduction

Exemplar models are used in linguistics to describe how language users store linguistic categories [34]. Examples of the type of linguistic categories we have in mind are vowel sounds like ɑ, ɜ, ɪ (corresponding to the vowel sounds in 'bat', 'bet', 'bit', respectively). When a person hears a vowel sound within a word, they have to classify it as belonging to one of the categories of vowels based on the sound's acoustic properties. This classification will determine what word the person understands is being uttered, e.g. 'tack', 'tech', or 'tick'. An important issue in linguistics is how language users perform this classification.

Exemplar models provide one answer to this question [24]. According to exemplar models, every member of a linguistic community stores a multitude of detailed memories of every sound that they hear. These memories are called exemplars. Exemplars consist of detailed acoustic information about the sound, as well as a category label: information about what the sound is classified as. For example, with the case of vowels, according to exemplar models, every person holds a detailed memory of every vowel they have ever heard, labeled with the corresponding vowel category, ɑ, ɜ, ɪ, etc [9].

Exemplar models provide a theory of both perception and production. When the language user hears a new vowel sound, the sound is compared to other exemplars already in memory and is classified according to the labels of exemplars that it is close to. When the language user needs to produce a new instance of a vowel, they select an exemplar from the set of all exemplars with the appropriate label and utter a copy of it, usually with either noise or bias added.

An important feature of many exemplar models is that exemplars do not remain in memory unchanged forever. A popular choice is for each exemplar to have a weight that decays with time [24, 32]. These exemplar weights enter into both perception and production, indicating that certain exemplars are more important for the relevant process than others. New exemplars are created (every time a new instance is perceived) with some large weight which then decays exponentially with time. This allows old exemplars to be forgotten and the general population of exemplars in a language user's mind to change with time.

There has been relatively little mathematical analysis of exemplar models. In [31] Tupper studies a fairly elaborate exemplar model that is able to account for the phenomenon of sound merger. He was able to obtain analytical results by looking at a certain limiting case of the model, a limit in which there are no stochastic fluctuations. In [31] perceptual boundaries, where language users switch from classifying a stimulus as one sound versus another, approach a stable configuration in perceptual space. Our intent here is to study the fluctuations of perceptual boundaries within an exemplar model.

Our starting point is an exemplar model that was studied by MacQueen in 1967, originally as an algorithm for k-means clustering [15]. We can put MacQueen's work in the context of exemplar models as follows. Suppose an individual has a phonetic space (that is, a space of possible sounds) with k labeled exemplars, one for each of k categories. Suppose the individual receives an independent identically distributed (i.i.d.) sequence of acoustic inputs that they have to classify into these k categories. Rather than the criteria for categories being pre-given, the classification is performed on the fly using the exemplars that are already stored. Thus as more exemplars are stored the criteria for classification changes. If we assume that (i) the weights of the exemplars are all equal and do not change with time, (ii) for each category the mean acoustic value of all exemplars in that category is stored, (iii) new exemplars are classified according to which category mean they are closest to, we obtain the MacQueen model.

To explain MacQueen's [15] results, we recall the definition of a centroidal Voronoi tessellation [8]. Given a set of *generators*, which are just a finite set of points in the space, the Voronoi tessellation is a partitioning of the space where each point is assigned to a cell based on which generator it is closest to. A centroidal Voronoi tessellation of a region is a Voronoi tessellation in which each generator is the centroid (i.e. the center of mass) of its cell. Centroidal Voronoi tessellations have already been established as being fundamental in some game theoretic models of language [10]. MacQueen's result is that in his model the distance between the category means and generators of centroidal Voronoi tessellations converge to 0. This implies in turn that the perceptual boundaries of the language-users align with the boundaries of centroidal Voronoi tessellations.

Taken as an exemplar model, the MacQueen model deviates from more realistic models of language use and development in several ways. The main difference is that in MacQueen's

model weights of exemplars do not decay over time. The idea of weight decay in exemplar models for language use was introduced by Pierrehumbert in [24] and has been used since. In order to move in the direction of analyzing more realistic exemplar models, our contribution in this chapter is to introduce weight decay into MacQueen's model. This alteration makes the model more realistic because weight decay simulates how memories fade over time. In our model every exemplar starts with weight 1 and the weight then exponentially decays with time. As we will show, this causes the Voronoi regions to no longer settle down into a stable configuration, but instead continue to move randomly for all time; this is the main result of Section 2.3. We then consider a simple special case of our model for which we perform a quantitative analysis of the motion of the perceptual boundary between two categories.

In Section 2.2 we formally specify the exemplar model model we study here. In Section 2.3 we go on to investigate a number of properties of the model, and most significantly, prove that when we have decay of exemplar weights then the exemplar means do not converge. In Section 2.4 we go on to study our model in a special one-dimensional case with two categories. We provide a probabilistic model for the motion of the perceptual boundary.

2.2 The k-Means Exemplar Model

We imagine a language user who hears a sequence of sounds and classifies each of the sounds into one of k categories, $k \geq 2$. The acoustic properties of the sound heard at the nth time step are denoted by $z_n \in \mathbb{R}^N$. We assume that all z_n lie in a set $E \subseteq \mathbb{R}^N$ that is bounded, convex, closed, and has a non-empty interior. E corresponds to the space of all phonetically possible sounds. The sounds $z_1, z_2, \ldots$ are generated in E independently according to a fixed probability measure $\mathbf{P}$. We assume the probability measure $\mathbf{P}$ can be written as

$$\mathbf{P}(A) = \int_A f(x)dx + \sum_{m \in \Omega} p_m \delta_m(A) \tag{2.1}$$

for all measurable $A \subseteq E$, where $f(x) > 0$, for all $x \in E$ is a Borel-measurable function, $\Omega \subset E$ is a countable set, δ_m is the point mass distribution located at m, and p_m is the corresponding probability of each point mass. Based on the definition of a probability measure, these parameters must be under the condition that $\mathbf{P}(E) = 1$. (Our results in Section 2.3 hold for measures of this generality, but in practice we can just imagine a measure without atoms.)

At the start of the model, we imagine that our language user already has a number of exemplars (each with a corresponding value in E) in each of the k categories. At time n, in category $j = 1, \ldots, k$, the language user has exemplars with phonetic values $a_j^i \in E$ and weights v_j^i for $i = 1, \ldots, n_j$, where n_j is the number of exemplars in category j. At time step n, a new sound with phonetic parameters $z_n \in E$ is heard. This sound is stored as an

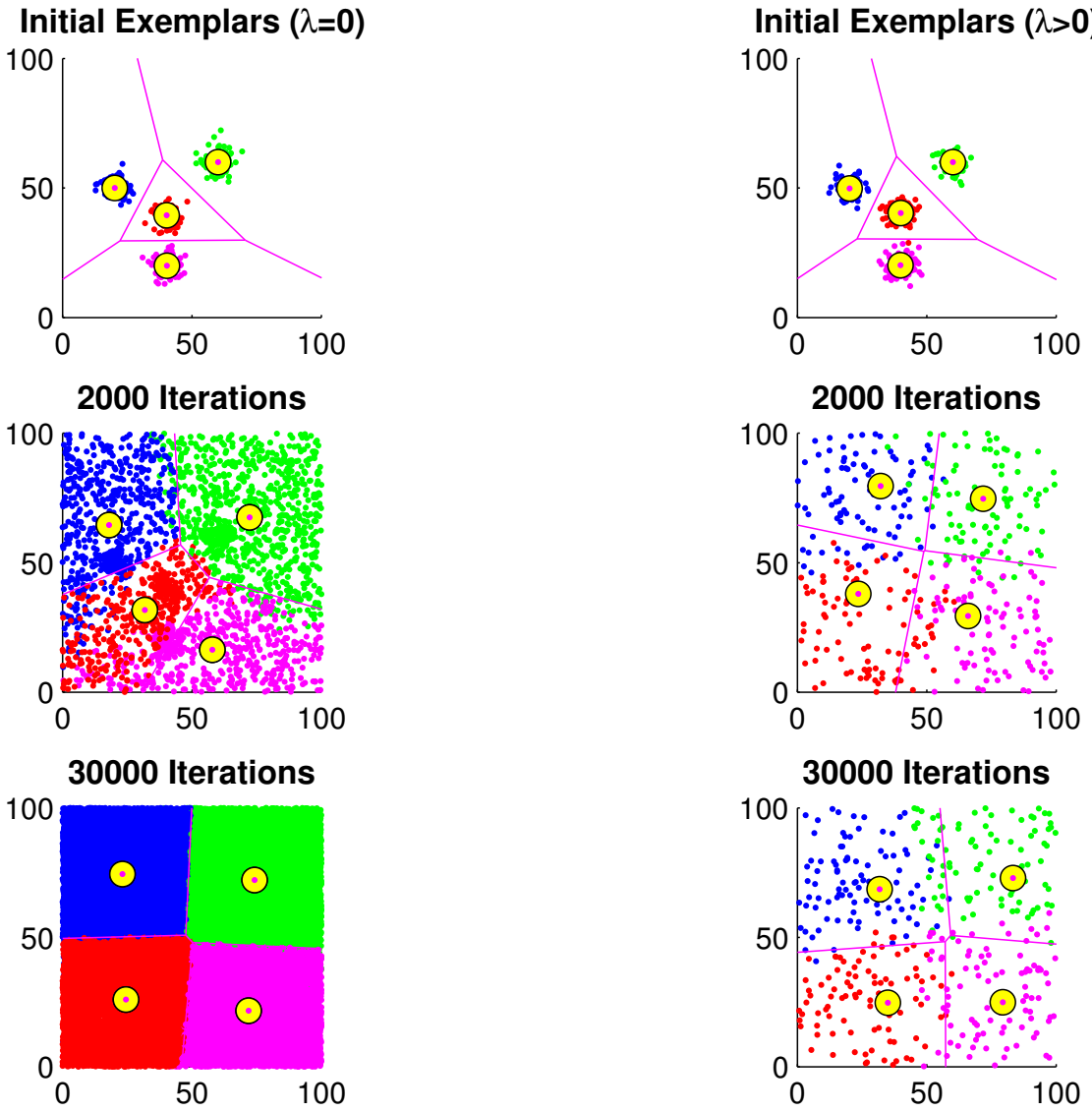

Figure 2.1: The exemplar dynamics model for $\lambda = 0$ (left) and $\lambda = 0.05$ (right). Small dots indicate individual exemplars, color-coded to indicate which category they were classified into. All dots with weight larger than 0.01 are plotted. The larger yellow dots indicate category means. Magenta lines show the boundaries between the Voronoi regions defined by the category means.

exemplar in one of the language user's categories. Which category the new sound is stored in depends on the category means of all the categories. We define the category means at time step n, x_j^n for $j = 1, \ldots, k$ to be

$$x_j^n = \frac{\sum_{i=1}^{n_j} v_j^i a_i^i}{\sum_{i=1}^{n_j} v_j^i}$$

and the total category weights by

$$w_j^n = \sum_{i=1}^{n_j} v_j^i.$$

We assign the new phonetic value z_n to be an exemplar of category j if for all ℓ

$$|z_n - x_j^n| \leq |z_n - x_\ell^n|.$$

(In the case of a tie, the exemplar is assigned to the category with the lower category index.) In this chapter $|\cdot|$ represents the Euclidean norm.

We can specify the update procedure in another way as follows: New exemplars entering the system are classified according to which Voronoi cell they are in according to the Voronoi tesselation generated by the category means. In other words, exemplars are assigned to the category mean they are closest to. More formally, if the category means are $x_1^n, \ldots, x_k^n$ we define the Voronoi cell $S_i(x^n)$ as the set

$$\begin{aligned} S_i(x^n) = \{\xi : &\xi \in E, |\xi - x_i^n| \leq |\xi - x_j^n| \text{ for } j = 1, 2, \ldots, k, \\ &\text{and if } i > 1,\ \xi \notin S_m(x^n), \text{ for all } m < i.\} \end{aligned} \tag{2.2}$$

$S_i(x^n)$ contains points in E closest to x_i^n with tied points being assigned to the lower index. We will let $S_i^n = S_i(x^n)$ for convenience of notation.

Our system will evolve as an iterative process in the following way. At each step all weights in the system decay, which we model by multiplying weights by $e^{-\lambda}$, where λ is a positive parameter. In addition, if $z_n \in S_i^n$ then we update the value of the average x_i^n by including z_n in it, and also by increasing the total weight of the category by 1. So if $z_n \in S_i^n$ we set

$$x_i^{n+1} = \frac{x_i^n w_i^n e^{-\lambda} + z_n}{w_i^n e^{-\lambda} + 1}, \qquad w_i^{n+1} = w_i^n e^{-\lambda} + 1,$$

and for all $j \neq i$ we set

$$x_j^{n+1} = x_j^n, \qquad w_j^{n+1} = w_j^n e^{-\lambda}.$$

An important feature of our model is that, as can be seen above, information about the individual's exemplars, a_j^i and v_j^i, are not needed to update the values of x_j and w_j. Given z_n, only x_j^n and w_j^n are needed to compute x_j^{n+1} and w_j^{n+1}. Accordingly, we let $x^n = (x_1^n, x_2^n, \ldots, x_k^n)$, where $x_j^n \in E$ represents the weighted mean of category j at time

step n with associated weights $w^n = (w_1^n, w_2^n, \ldots, w_k^n)$. These weights are the sum of all weights of the exemplars in each category. We will refer to x_j^n as the exemplar mean for category j. The parameter $\lambda > 0$ is our decay rate.

Initial conditions can be prescribed by giving the locations and weights of each exemplar in each category. However, since the individual locations and weights of exemplars do not enter in to the dynamics of the Voronoi cells, we only need to set the category means and category weights. For each $j = 1, 2, \ldots, k$, we define category means $x_j^0 \in E$ such that $x_j^0 \neq x_i^0$ whenever $j \neq i$, and let $w_j^0 > 0$.

Figure 2.1 shows what typical runs of the model look like with $\lambda = 0$ and $\lambda > 0$. In each case the model is initialized with 4 categories. The 4 categories were generated by selecting 4 arbitrary points in the phonetic space (in this case a 100 by 100 square) and distributing 100 exemplars with weight 1 about the points according to a Gaussian distribution with standard deviation 3. In both cases, at every time step a new exemplar was randomly generated uniformly on the square ($\mathbf{P}(A) = \int_A 1/|E| dx$, where $|E|$ is the area of E), and assigned to the category of the category mean it is closest to. In the $\lambda = 0$ case exemplars remain at their initial weight for all time, and so the number of exemplars in the system and visible in the plot increases continually. The division of the square into a Voronoi tesselation converges to a centroidal Voronoi tesselation, as shown in [15]. In the $\lambda > 0$ case on the right, although added exemplars remain in the system for all time, the weights of the exemplars decrease over time. To help visualize this, we only plot exemplar with weight greater than 0.01. Accordingly, the number of exemplars in the plot converges to a steady state. However, the Voronoi cells now continue to move for all time, no matter our choice of parameters as described above. This will be proved in the next section.

The condition $f(x) > 0$ is necessary for our result. Consider the following counterexample. Let $E_1 = [0, 1]$ and set our probability distribution to be $\mathbf{P}(x) = \delta_{1/2}(x)$. If $k = 2$ we can choose $x_1^0 = 1/2$ and $x_2^0 = 1/4$. The probability distribution P can only generate the value $1/2$, and as such $z^n = 1/2$, for all n. As such, $x_1^n = 1/2$ and $x_2^n = 1/4$, for all n, no matter the choice of the initial weights. As such, both category centers in this case will converge to points in E. We have thus given an example where $f(x)$ is not strictly greater than 0 where the exemplar means converge, which is the reason why we require $f(x)$ to be strictly greater than 0.

2.3 General Results for k-Means Exemplar Model

In this section, we prove for the model described in Section 2.2 that none of the exemplar means converge, and that the categories do not collapse when E is bounded. In particular, our main result is the following.

Theorem 1. *Let $E \subseteq \mathbb{R}^n$ be a bounded, convex, closed subset having non-empty interior. Let $\lambda > 0$, $f(x) > 0$ for all $x \in E$, s.t. f is Borel-measurable, $w_1^0, \ldots, w_k^0 > 0$, $x_1^0, \ldots x_k^0 \in E$,*

where $x_i^0 \neq x_j^0$ for $i \neq j$, and $z_n, n \geq 1$ be an independent sequence of random variables each with distribution given by $\mathbf{P}$ *as defined in Equation 2.1. Define x_j^n, w_j^n for $n > 0$ by: if i is the minimal index such that $|x_i^n - z_n|$ is minimized, then*

$$x_i^{n+1} = \frac{x_i^n w_i^n e^{-\lambda} + z_n}{w_i^n e^{-\lambda} + 1}, \qquad w_i^{n+1} = w_i^n e^{-\lambda} + 1, \tag{2.3}$$

and for all $j \neq i$ we set

$$x_j^{n+1} = x_j^n, \qquad w_j^{n+1} = w_j^n e^{-\lambda}. \tag{2.4}$$

Then, for each j,

1. *x_j^n does not converge as $n \to \infty$,*
2. *the volume of S_j^n (defined in Equation 2.2) does not converge to zero as $n \to \infty$,*
3. *$z_n \in S_j^n$ for infinitely many n.*

Proof. Result 1 is implied by Corollary 1 and Result 2 is implied by Theorem 4, both of which are proved below. Result 3 follows from Result 1, since the only way x_j^n can move is if $z_n \in S_j^n$. □

Informally, the main result says that (1) none of the exemplar means converge, (2) the volumes of none of the Voronoi regions converge to zero, and (3) every category has exemplars classified in it infinitely often. The key to proving all three results is showing that there is an $\varepsilon > 0$ such that for each category j there infinitely many times n such that a ball of radius ε is contained in S_j^n. This fact is the content of Theorem 4. To prove Theorem 4 we will require several preliminary lemmas.

Lemma 1. *There exists a $\gamma \in \mathbb{R}$ depending only on λ and the initial vector of weights w^0, such that $w_i^n \leq \gamma$, for $i = 1, 2, \ldots, k$, and for all $n \geq 0$.*

Proof. If we let W^n represent the total weight of our system at time step n, it is straightforward to show $W^n = \sum_{i=1}^k w_i^n = W^{n-1}e^{-\lambda} + 1$ for all realizations. Since $e^{-\lambda} < 1$ we know that W^n converges to $W := (1 - e^{-\lambda})^{-1}$. Since W^n converges monotonically to W,

$$W^n \leq \max\left\{W^0, \frac{1}{1 - e^{-\lambda}}\right\} = \gamma, \tag{2.5}$$

for all n. This in turn implies the result. □

Lemma 1 proves the total weight of all exemplars in the system is uniformly bounded above. This is in contrast to the MacQueen model where the total weight of the system diverges, and new exemplars have less influence every iteration. The value W (that W^n converges to) will come up later when we investigate the long term behaviour of the perceptual boundary.

The following lemma shows that for a fixed r the probability of a new exemplar z_n landing in a ball of radius r centred at a point $x \in E$ is bounded away from 0, uniformly with respect to x in bounded sets.

Lemma 2. *If $r > 0$ is fixed, and $F \subseteq E$ is closed then*

$$\inf_{x \in F} \mathbf{P}(B(x, r) \cap E) > 0.$$

Proof. We first want to show if $r > 0$ is fixed, there exists an $r' > 0$ such that for all $x \in F$ one can find a $x' \in E$ such that $B(x', r') \subseteq B(x, r) \cap E$.

Let $B_0 = B(x_0, r_0)$ be a subset of E, where $r_0 > 0$. We know B_0 exists because E has a non-empty interior. Fix an x in F. For any $\alpha \in [0, 1)$, the set $B_\alpha := (1 - \alpha)B_0 + \alpha x$ is an open ball which is contained in E because E is convex. Furthermore, B_α is centred at point $x_0 + \alpha(x - x_0)$ and has radius $(1 - \alpha)r_0$.

We want to find all such α so that B_α is completely within $B(x, r)$. This containment will hold for $\alpha \in [0, 1)$ if and only if

$$(1 - \alpha)|x - x_0| + (1 - \alpha)r_0 \leq r$$

or, rearranging,

$$\alpha \geq 1 - \frac{r}{|x - x_0| + r_0}.$$

Let $\alpha' := \max\left(0, \sup_{x \in F} 1 - \frac{r}{|x - x_0| + r_0}\right)$. Since F is bounded, $\alpha' < 1$. Then $B_{\alpha'} = (1 - \alpha')B_0 + \alpha' x$ is contained in $B(x, r)$ for all $x \in F$. But $B_{\alpha'}$ is a ball of radius $r' = (1 - \alpha')r_0$ and so we proved the existence of such an r'.

We have now shown that

$$\mathbf{P}(B(x, r) \cap E) \geq \int_{B(x,r) \cap E} f(y)dy \geq \int_{B(x'(x), r')} f(y)dy,$$

where $x'(x) = x_0 + \alpha'(x - x_0)$. To establish our result, we just need to show that

$$\inf_{x \in F} \int_{B(x'(x), r')} f(y)dy > 0.$$

Suppose for contradiction that $\inf_{x \in F} \int_{B(x'(x), r')} f(y)dy = 0$. There must exist a sequence $\{z_n\}_{n>0}$ in F such that $\lim_{n \to \infty} \int_{B(x'(z_n), r')} f(y)dy = 0$. Because F is bounded, there exists a subsequence such that $z_{n_i} \to z \in E$, as $i \to \infty$. Furthermore, since $B(x'(z_{n_i}), r') \subseteq E$, and E is closed, $B(x'(z), r')$ is also a subset of E, where we have used the fact that x' is a continuous function of x. We know that for all $\varepsilon > 0$, there exists an I, such that for all $i \geq I$, $|z_{n_i} - z| < \varepsilon$. Let $\varepsilon = r'/2$, implying that $B(x'(z), \varepsilon) \subseteq B(x'(z_{n_i}), r')$

for all $i \geq I$. So for all $i \geq I$ we have

$$\int_{B(x'(z_{n_i}),r')} f(x)dx \geq \int_{B(x'(z),\varepsilon)} f(x)dx > 0$$

since f is non-zero inside E, and $B(x'(z),\varepsilon)$ is contained in E. This gives us our contradiction, since we know the left hand side converges to zero as $i \to \infty$. □

The following lemma shows that if a long enough sequence of new exemplars arrive within ε of a point, then some category mean will arrive within 2ε of the point.

Lemma 3. *Let $z \in E$ and $\varepsilon > 0$ be given. There exists a $p > 0$, such that if z_q is in $B(z,\varepsilon)$ for $n \leq q \leq n+p-1$, then for some $m \in \{1,\ldots,k\}$, the exemplar mean $x_m^{n+p} \in B(z,2\varepsilon) \cap E$. Parameter p only depends on k,* diam(E) *(the diameter of set E) and ε.*

Proof. Let's assume that z_q is in $B(z,\varepsilon)$ for $q \in \{n,\ldots,n+kp'-1\}$, for some p' we will determine later. We know that one category $m \in \{1,\ldots,k\}$ will be classified at least p' times over that time interval. So there exists a subsequence $\{q_i\}_{i\geq 1} \subseteq \{n,\ldots,n+kp'-1\}$ such that $z_{q_i} \in S_m^{q_i}$ for all i, and $|\{q_i\}_{i\geq 1}| \geq p'$. Note that in this time interval the exemplar means either stay fixed or move closer to z, if they are not already in the ball $B(z,\varepsilon)$. We want to show there is a p large enough such that we are guaranteed the mth exemplar mean will be inside $B(z,2\varepsilon)$ by the time $n+kp'$.

Let $\eta := \min_i |x_i^n - z|$ be the distance of the closest exemplar mean to z at time n. Note that since new exemplars only arrive at locations in $B(z,\varepsilon)$, only exemplars within $\eta+\varepsilon$ of z will move. In particular, $|x_m^n - z| \leq \eta + \varepsilon$.

Without loss of generality, let us assume z is at 0, so that $|z_q| \leq \varepsilon$ for all q. Let $y^q := |x_m^q|$.

If $z_q \in S_m^q$, then

$$x_m^{q+1} = \frac{x_m^q w_m^q e^{-\lambda} + z_q}{w_m^q e^{-\lambda} + 1}.$$

Let $\rho = w_m^q e^{-\lambda}/(w_m^q e^{-\lambda}+1)$. Note that $\rho \in (0,1)$ and is bounded away from 1, because of the bound on the total weights provided by Lemma 1. So

$$x_m^{q+1} = \rho x_m^q + (1-\rho) z_q$$

which implies

$$\begin{aligned} y^{q+1} &\leq \rho y^q + (1-\rho)|z_q| \\ &\leq \rho y^q + (1-\rho)\varepsilon. \end{aligned}$$

If $z_q \notin S_m^q$ then $y^{q+1} = y^q$.

We know that new exemplars will be classified in category m at least p' times. So, using the inequality version of the identity for geometric series:

$$y^{n+kp'} \leq \rho^{p'} y^n + \frac{1-\rho^{p'}}{1-\rho}(1-\rho)\varepsilon \leq \rho^{p'}(\eta+\varepsilon) + (1-\rho^{p'})\varepsilon \leq \rho^{p'}\mathrm{diam}(E) + \varepsilon.$$

The limit of the right-hand side as $p' \to \infty$ is ε, so there is a large enough p' so that $y^{n+kp} < 2\varepsilon$, and hence $x_m^{n+kp'} \in B(z, 2\varepsilon)$. Let $p = kp'$ gives the required result. □

Lemma 3 will be used to prove that the exemplar means don't converge towards one another, and as such, there is no collapse in the system. How it will be utilized will become apparent in the following lemma. Here we show if a collection of one or more exemplar means are close to a point z in the interior of E, with positive probability one of the exemplar means will be moved away from z in a bounded number of steps. Meanwhile, all the exemplar means that are far away from z will not be moved.

In what follows, we will use ∂E to denote the boundary of E, and for any subset F of $\mathbb{R}^N$, $d(z, F)$ to denote the distance between point $z \in \mathbb{R}^n$ and F:

$$d(z, F) = \inf_{x \in F} |z - x|.$$

Lemma 4. *Let $\delta > 0$, $z \in E$ with $\delta \leq d(z, \partial E)$, ε be a constant such that $\varepsilon \leq \delta/10$, and*

$$\begin{aligned} A_1 &= \{i,\ s.t.\ |x_i^{n_0} - z| \geq \delta\}, \\ A_2 &= \{i,\ s.t.\ |x_i^{n_0} - z| < \delta/2 - 4\varepsilon,\ and\ x_i^{n_0} \neq z\}, \\ A_3 &= \{i,\ s.t.\ x_i^{n_0} = z\}. \end{aligned}$$

If $|A_1| + |A_2| + |A_3| = k$ (so there are no exemplars between distances $\delta/2 - 4\varepsilon$ and δ from point z), then there exists a $y \in B(z, \delta/2)$, and a $p > 0$, such that if $z_n \in B(y, \varepsilon)$ for $n_0 \leq n < n_0 + p$, then $\max_{i \in A_2} |x_i^{n_0+p} - z| \geq \delta/2 - 4\varepsilon$, and $x_i^{n_0+p} = x_i^{n_0}$ for all $i \in A_1 \cup A_3$.

Proof. Let $q = \arg\max_{i \in A_2} |x_i^{n_0} - z|$, and

$$y = z + \frac{x_q^{n_0} - z}{|x_q^{n_0} - z|}(\delta/2 - 2\varepsilon),$$

so that y is a distance $\delta/2 - 2\varepsilon$ away from z in the direction of point $x_q^{n_0}$. We know because $\delta \leq d(z, \partial E)$, that $B(y, \varepsilon) \subseteq E$. If $z_n \in B(y, \varepsilon)$, consecutively, then the nth sound will always be categorized as a category in A_2 for the following reasons:

1. x_q^n is closer to any point in $B(y, \varepsilon)$ than z is. So new exemplars are never classified in categories $i \in A_3$.

2. x_q^n will always be closer to $B(y,\varepsilon)$ than any x_i^n such that $i \in A_1$. So new examplars are never classified in categories $i \in A_1$.

Let constant $p > 0$ be as determined by Lemma 3, which will depend on k, ε and $\text{diam}(E)$. If $z_n \in B(y,\varepsilon)$ for $n_0 \le n \le n_0 + p$, we know $z_n \in S_i^n$ where $i \in A_2$ for all n such that $n_0 \le n \le n_0 + p$. By Lemma 3, we know there exists a category m such that $x_m^{n_0+p} \in B(y, 2\varepsilon)$. The category m must be in A_2 because the categories in $A_1 \cup A_3$ do not move in the time interval. This implies there exists an $i \in A_2$, such that $x_i^{n_0+p} \in B(y, 2\varepsilon)$, and because $|y - z| = \delta/2 - 2\varepsilon$, we know that $|x_i^{n_0+p} - z| \ge \delta/2 - 4\varepsilon$, giving the result. □

Using Lemma 4 we will establish Theorem 2 which states the following: for a given j, as long as all exemplar means are away from the boundary of E, there is a probability bounded away from zero that at some time later, all other exemplar means will be moved away from the jth one and the jth one will be moved away from the boundary.

Theorem 2. *For any $\delta > 0$, $j \in 1, \ldots, k$, and time n_0, there exists an $\varepsilon > 0$, $M > 0$, and $H > 0$ such that*

$$\mathbf{P}\left(\min_{i \ne j} |x_i^{n_0+M} - x_j^{n_0+M}| \ge \varepsilon, d(x_j^{n_0+M}, \partial E) \ge \varepsilon \middle| d(x_j^{n_0}, \partial E) \ge \delta\right) \ge H$$

In particular, ε, M, and H depend only on δ, E, and k.

Proof. To prove this theorem, we will begin by showing there is an event which pulls all the exemplar means (except category j's) away from category j's exemplar mean x_j^n, under the condition $d(x_j^{n_0}, \partial E) \ge \delta$. Futhermore, x_j^n will also be away from the boundary. We then will show the event has a positive probability of happening.

Before describing the event, we will define some parameters. We let $\varepsilon = \delta 2^{-k}/9$. Let $\{d_i\}_{i=1}^{k-1}$ be defined as $d_0 = \delta$ and for all $i > 1$, $d_{i+1} = d_i/2 - 4\varepsilon$. One can calculate d_i explicitly as

$$d_i = \frac{\delta}{2^i} - 8\varepsilon\left(1 - \frac{1}{2^i}\right)$$

for $i \in \{1, \ldots, k-1\}$. This is a decreasing sequence, such that $B(x_j^{n_0}, d_i + 4\varepsilon) \subseteq E$ for $i \ge 1$. Our choice of ε guarantees that $d_i \ge \varepsilon$ for all i.

An event will now be described which pulls every exemplar mean (except for category j's) at least distance ε away from x_j^n. The event will be described in an algorithmic manner, as a sequence of steps that must occur as the index i runs from 1 to $k-1$. For each step, what needs to happen depends on whether or not there are already i exemplar means distance d_i or greater from x_j^n. If there are not, we use Lemma 4 to pull one of the exemplars within distance d_i from x_j^n away from it. If there already are, we select new exemplars from a ball of radius ε far away from x_j^n so as not to disturb exemplars close to x_j^n. At the end of step i, we will be guaranteed at least i exemplar means will be distance d_i or farther from x_j^n.

Let $z = x_j^{n_0}$ and $n = n_0$. We will start our process with $i = 1$ and proceed through to $i = k - 1$.

For $i = 1$ to $k - 1$, let $s_i^n = |\{q \text{ s.t. } |x_q^n - z| \geq d_i\}|$, the number of category means that are distance d_i or greater from z. There are two possibilities.

Case 1: $s_i^n = i - 1$.

We need to move at least one exemplar mean inside $B(z, d_i)$ outside of $B(z, d_i)$, without moving x_j^n or any of the exemplar means that are outside of $B(z, d_{i-1})$.

By the inductive hypothesis, we know that there are $i - 1$ exemplars farther than d_{i-1} from z. So this means that there are no exemplars in $B(z, d_{i-1}) \backslash B(z, d_i)$.

We also know z is at least distance δ from the boundary, $d_i = d_{i-1}/2 - 4\varepsilon \geq \varepsilon$, and $\delta > d_i$, which implies that we can implement Lemma 4.

By Lemma 4, there exists a $y \in B(z, d_i + 4\varepsilon)$, and a $p > 0$, such that if $z_m \in B(y, \varepsilon) \subseteq E$, for $n \leq m \leq n + p - 1$, then $\max_{i \in A_2} |x_i^{n+p} - z| \geq d_i$, and $x_i^{n+p} = x_i^n$ for all exemplar means outside of $B(z, d_{i-1})$. In other words, if new sounds are generated consecutively in $B(y, \varepsilon)$, one exemplar mean will be moved outside of $B(z, d_i)$, but neither x_j^n nor the exemplar means outside of $B(z, d_{i-1})$ will be moved. We can take the p determined by the lemma, which only depends on k, $\text{diam}(E)$, and ε, and not on the specific configuration of category means.

Case 2: $s_i^n \geq i$. There are already enough exemplars far enough away from z at this step. Let ℓ be such that the exemplar x_ℓ^n is furthest from z. Let y be distance $\delta - \varepsilon$ away from z in the direction of x_ℓ^n. Allow the next p exemplars introduced to the system to fall within $B(y, \varepsilon)$, where p is the same value that would have been chosen in Case 1. The effect of this will only be to move around exemplar means of distance farther than d_i from z, and it cannot move them within d_i of z.

Now observe that whichever of these cases occur, the exemplar mean x_j^n is not moved. Since it started out at least distance δ from the boundary (and $\varepsilon < \delta$), we complete this step with x_j^n at least ε away from the boundary.

After one of these two cases has been performed for $i = 1$ to $k - 1$, the sequence of $p(k - 1)$ events guarantees that all exemplar means except x_j^n are distance ε away from x_j^n at time $n = n_0 + p(k - 1)$, and x_j^n is also at least distance ε from the boundary. We just need to show that the probability of this event is bounded away from 0. By Lemma 2, letting $F = E$, each event $(z_n \in B(y, \varepsilon))$ has a probability greater than $Q > 0$ uniformly with respect to y. So the total event has probability at least $Q^{p(k-1)}$ as required. Letting $H = Q^{p(k-1)}$ and $M = p(k - 1)$ gives the result. □

The probability in Theorem 2 is conditioned on all the category means $x_j^{n_0}$ being a distance δ away from ∂E. We have to establish that this event happens with a non-zero probability. We will show there exists a sequence of events which bring all the exemplar

means at least a distance δ^* away from ∂E. The probability of the event occurring is bounded below like in Theorem 2. Two lemmas are required to prove it.

Lemma 5. *Let z^* be in the interior of E, $\Gamma \in (0,1)$, and define the function $f(x) = z^* + \Gamma(x - z^*)$. There exists a $\delta > 0$ such that $d(f(x), \partial E) \geq \delta$, for all $x \in E$.*

Proof. We first observe that for all $x \in E$, $f(x)$ is in the interior of E. To see this, let x be in the interior of E. Note that $B(z^*, \rho)$ must be contained in E for some $\rho > 0$ because z^* is in the interior of E. By convexity, $B(z^* + \Gamma(x - z^*), (1 - \Gamma)\rho)$ must also be contained in E. So $f(x)$ is in the interior of E.

Since E is compact, f is continuous, and continuous images of compact sets are compact, $f(E)$ is a compact set. Two compact sets must have respective points whose distance is the same as the distance between the sets. Since nowhere do $f(E)$ and ∂E intersect, this distance must be positive. So, there exists a $\delta > 0$ such that $\text{dist}(f(E), \partial E) \geq \delta$. □

Before proving the next necessary lemma, we must define a sequence. Let $\Gamma = 1/2$. For $\varepsilon > 0$, let $\{f_i\}_{i=1}^k$ be a sequence such that $f_1 = 0$ and

$$f_{i+1} = \frac{\Gamma + (f_i + 2\varepsilon)}{2} + 2\varepsilon,$$

for all $i > 1$. f_i can be written explicitly as $f_i = (\Gamma + 6\varepsilon)(1 - 2^{1-i})$. We will need to choose ε small enough so that $f_i < \Gamma$ for $i = 1, \ldots, k$. This is accomplished by letting $\varepsilon \leq 2^{-k}/6$. The sequence $\{f_i\}_{i=1}^k$ will play a similar role to the sequence $\{d_i\}_{i=1}^{k-1}$ in Theorem 2, though, $\{f_i\}_{i=1}^k$ does not depend on the set E, ranges between 0 and 1, and increases instead of decreases.

Lemma 6. *Fix a time n_0. Let z^* be in the interior of E and let $\alpha > 0$ be such that $B(z^*, \alpha) \subseteq E$. Let $\{c_q\}_{q=1}^k$ be a reordering of $\{1, \ldots, k\}$ such that*

$$|x_{c_1}^{n_0} - z^*| \leq |x_{c_2}^{n_0} - z^*| \leq \ldots \leq |x_{c_k}^{n_0} - z^*|.$$

Let $q \in \{1, \ldots, k\}$, and let $s = \max\{\alpha, |x_{c_q}^{n_0} - z^*|\}$. Let $\Gamma = 1/2$ and let

$$\varepsilon = \min\left\{\frac{\delta}{3\text{diam}(E)}, \frac{2^{-k}}{6}\right\} > 0.$$

Let $f_i = (1 - 2^{1-i})(\Gamma + 6\varepsilon)$ for $i = 1, \ldots, k$, and let $\delta > 0$ be the constant determined by Lemma 5, where we use z^* as the interior point in E.

If $|x_{c_q}^{n_0} - z^*| > (f_q + 2\varepsilon)s$, and (if $q > 1$) $|x_{c_{q-1}}^{n_0} - z^*| \leq (f_{q-1} + 2\varepsilon)s$, then there exists a y in the interior of E, and a $p > 0$, such that if $z_n \in B(y, \varepsilon\alpha)$, for $n_0 \leq n < n_0 + p$, there will exist an $i \geq q$ such that

$$|x_{c_i}^{n_0+p} - z^*| \leq (f_q + 2\varepsilon)s, \text{ and } d(x_{c_i}^{n_0+p}, \partial E) \geq \delta/3.$$

Additionally $x_{c_r}^n = x_{c_r}^{n_0}$*, for all* $r < q$*, and* n *such that* $n_0 \leq n \leq n_0 + p$*.*

Proof. Let

$$y = z^* + \frac{(x_{c_q}^{n_0} - z^*)}{|x_{c_q}^{n_0} - z^*|} s f_q,$$

so y is distance sf_q away from z^* in the direction of point $x_{c_q}^{n_0}$. Note that if we let $w = z^* + \frac{sf_q}{\Gamma} \frac{(x_{c_q}^{n_0} - z^*)}{|x_{c_q}^{n_0} - z^*|}$, then

$$y = z^* + \Gamma(w - z^*), \tag{2.6}$$

where $w \in E$, since it is a convex combination of z^* and $x_{c_q}^{n_0}$. By Lemma 5, we therefore have that y is at least distance δ from ∂E.

We will show that exemplars z_n falling in $B(y, \varepsilon\alpha)$ will always be classified in categories c_i for $i \geq q$. First note that this is immediate if $k = 1$ so assume $k \geq 2$. For $z_n \in B(y, \varepsilon\alpha)$ we have

$$\begin{aligned} |z_n - x_{c_q}^{n_0}| &\leq |y - x_{c_q}^{n_0}| + |y - z_n| \\ &\leq s - f_q s + \varepsilon\alpha \\ &\leq s[1 - f_q + \varepsilon], \end{aligned}$$

and, for $i < q$,

$$\begin{aligned} |z_n - x_{c_i}^{n_0}| &\geq |y - z^*| - |z^* - x_{c_i}^{n_0}| - |z_n - y| \\ &\geq f_q s - (f_{q-1} + 2\varepsilon)s - \varepsilon\alpha \\ &\geq s\left[f_q - f_{q-1} - 3\varepsilon\right]. \end{aligned}$$

The definition of f_q in terms of f_{q-1} then allows us to show $|z_n - x_{c_q}^{n_0}| < |z_n, x_{c_i}^{n_0}|$, and so exemplars falling in $B(y, \varepsilon\alpha)$ will always be classified in categories c_i for $i \geq q$.

Let p be the constant determined by Lemma 3 which will depend on $\text{diam}(E)$ and $\varepsilon\alpha$ (in place of ε in the lemma). If $z_n \in B(y, \varepsilon\alpha)$ for all n such that $n_0 \leq n < n_0 + p$, there must exist an m such that $x_m^{n_0+p} \in B(y, 2\varepsilon\alpha)$. We know $m \geq q$, because the other exemplar means cannot move. As such there exists an $m \geq q$, such that $x_{c_m}^{n_0+p} \in B(y, 2\varepsilon s)$, since $s \geq \alpha$. This gives

$$|x_{c_m}^{n_0+p} - z^*| \leq |x_{c_m}^{n_0+p} - y| + |y - z^*| \leq 2\varepsilon s + f_q s = (f_q + 2\varepsilon)s.$$

Additionally $x_{c_r}^{n_0+p} = x_{c_r}^{n_0}$, for all $r < q$.

We know $\varepsilon \leq \delta / (3\text{diam}(E))$, which ensures that $\varepsilon s \leq \delta/3$. Since y is at least distance δ from ∂E and the radius of the ball $B(y, 2\varepsilon s)$ is at most $2\delta/3$, we have that $d(x_i^{n_0+p}, \partial E) \geq \delta/3$. □

Lemma 5 will be used in the proof of the following theorem. The proof of the theorem will be similar to the proof of Theorem 2; we will describe an event in which all exemplar means are pulled away from the boundary, and which has a positive probability of occurring.

Theorem 3. *There exists a $\delta^* > 0$, an $M > 0$, and an $H > 0$ such that, for every time n_0*

$$\mathbf{P}\left(\min_i d(x_i^{n_0+M}, \partial E) \geq \delta^*\right) \geq H.$$

δ^*, M, and H only depend on E.

Proof. To prove this lemma, we are going to show there is an event which can bring all the exemplar means away from ∂E. We will prove this event has a positive probability of happening. Before describing the event, we will define some variables.

Let z^* be a point in the interior of E. Let $\alpha = d(z^*, \partial E) > 0$.

We use the sequence $\{f_i\}_{i=1}^k$, which was defined before Lemma 6. To repeat: let $\Gamma = 1/2$, and use Lemma 5 to give us a $\delta > 0$ so that $d(z^* + \Gamma(x - z^*), \partial E) \geq \delta$ for all $x \in E$. Let $\varepsilon = \min\left\{\delta/(3\text{diam}(E)), 2^{-k}/6\right\}$. Then let $f_i = (1 - 2^{1-i})(\Gamma + 6\varepsilon)$ for $i = 1, \ldots, k$.

As in Theorem 2 we will describe a sequence of steps that pulls the ith exemplar to within distance $(f_i + 2\varepsilon)\text{diam}(E)$ of z^* and where all the exemplars are at least distance $\delta^* = \delta/3$ away from ∂E. We start with $i = 1$ and then increase i up to k.

Let $n = n_0$. We repeat the following steps for $i = 1, \ldots, k$. First, we let the indices $c_\ell, \ell = 1, \ldots, k$ be such that

$$|x_{c_1}^n - z^*| \leq |x_{c_2}^n - z^*| \leq \cdots \leq |x_{c_k}^n - z^*|$$

Let $s = \max\{\alpha, |x_{c_i}^n - z^*|\}$. Let $s_i^n = |\{q \text{ s.t.} |x_q^n - z^*| < (f_i + 2\varepsilon)s\}|$ be the number of category means that are within $(f_i + 2\varepsilon)s$ of z^*. There are two possibilities.

Case 1: $s_i^n = i - 1$.

We need to move at least one exemplar outside of $B(z^*, (f_i + 2\varepsilon)s)$ to within distance $(f_i + 2\varepsilon)s$ of z^*. Lemma 6 shows precisely this: there is a p such that if z_m falls in $B(y, \varepsilon\alpha)$ for $n \leq m < n + p$ then at least one exemplar mean will be moved in to $B(z^*, (f_i + 2\varepsilon)s)$. Furthermore, none of the exemplar means that are already closer to z^* will be moved, and the exemplar mean that is moved will be farther than distance $\delta/3$ of ∂E.

Case 2: $s_i^n \geq i$.

In this case, none of the exemplar means need to be moved for the condition to be satisfied. In this case we allow p exemplars in a row to fall within $B(z^*, \varepsilon\alpha)$. Since $\varepsilon\alpha \leq (f_i + 2\varepsilon)s$ for all i, we have $B(z^*, \varepsilon\alpha) \subset B(z^*, (f_i + 2\varepsilon)s)$ for all i, and this does not change any of the necessary containments.

Finally, we update n to $n + p$.

Repeating the procedure for $i = 1, \ldots, k$ gives the required conditions. What was necessary was that event of the form $z_n \in B(y, \varepsilon\alpha)$ for some $y \in E$, kp times in a row.

Lemma 2 shows that the probability of $z_n \in B(y, \varepsilon\alpha)$ is bounded below uniformly in $y \in E$ by some $h > 0$, and so the probability of it happening kp times in a row for varying y is bounded below by $H = h^M$, where $M = kp$. □

The inequalities in Theorems 2 and 3 do not change no matter how you condition the event inside the probability on the events prior to the initial condition. With these theorems, we now have enough to prove there is no categorical collapse for the system described in Section 2.2.

Theorem 4. *There exists an $\varepsilon > 0$ such that the event $\{B(x_j^n, \varepsilon) \subseteq S_j^n\}$ occurs infinitely often for every $j \in \{1, \ldots, k\}$.*

Proof. We will show the statement is true by proving that infinitely often the exemplar means will be at least a certain distance away from one another and from the boundary.

Let time $n_0 \geq 0$ and $j \in \{1, \ldots, k\}$ be given. By Theorem 3, we know there exists a $\delta^* > 0$, $M_1 > 0$, and $H_1 > 0$, such that

$$\mathbf{P}\left(d(x_j^{n_0+M_1}, \partial E) \geq \delta^*\right) \geq H_1.$$

By Theorem 2 there exists an $\varepsilon_1 > 0$, $M_2 > 0$, and an $H_2 > 0$, such that

$$\mathbf{P}\left(\min_{i \neq j} |x_i^{n_0+M_1+M_2} - x_j^{n_0+M_1+M_2}| \geq \varepsilon_1, d(x_j^{n_0+M_1+M_2}, \partial E) \geq \varepsilon_1 \middle| d(x_j^{n_0+M_1}, \partial E) \geq \delta^*\right) \geq H_2.$$

Combining these two bounds we get that

$$\mathbf{P}\left(\min_{i \neq j} |x_i^{n_0+M_1+M_2} - x_j^{n_0+M_1+M_2}| \geq \varepsilon_1, d(x_j^{n_0+M_1+M_2}, \partial E) \geq \varepsilon_1\right) \geq H_1 H_2.$$

Since H_1, H_2 and M_1, M_2 do not depend on n_0, this is enough to show that

$$\min_{i \neq j} |x_i^n - x_j^n| \geq \varepsilon_1, \quad d(x_j^n, \partial E) \geq \varepsilon_1,$$

for infinitely many n. The exemplar means will thus be separated from each other at least a distance ε_1 infinitely often. This implies $B(x_j^n, \varepsilon_1/2) \subseteq S_j^n$ infinitely often for every $j \in \{1, \ldots, k\}$. Letting $\varepsilon = \varepsilon_1/2$ gives the result. □

Theorem 4 implies that the areas of the categorical regions S_j^n cannot approach zero for any $j \in \{1, \ldots, k\}$. This in turn implies Result 3 of Theorem 1.

Corollary 1. *There exists a $\delta > 0$ such that for every $z \in E$, the event $\{|x_j^n - z| \geq \delta\}$ occurs infinitely often for every $j \in \{1, \ldots, k\}$.*

Proof. Let $j \in \{1, \ldots, k\}$, and assume for a contradiction that there exists a $z \in E$ such that $|x_j^n - z| \to 0$ as $n \to \infty$.

By Theorem 4, we know there exists an $\varepsilon > 0$ such that $B(x_j^n, \varepsilon) \subseteq S_j^n$ infinitely often. This implies there must exist a subsequence $\{x_j^{n_p}\}_{p=1}^{\infty}$, such that $B(x_j^{n_p}, \varepsilon) \subseteq S_j^{n_p}$ for all p.

If $z \in \partial E$, then we get a contradiction because when $B(x_j^{n_p}, \varepsilon) \subseteq S_j^n$, it implies that $|x_j^n - z| > \varepsilon$, which must occur infinitely often. So z must be in the interior of E.

Let $F_p = B(x_i^{n_p}, \varepsilon) \backslash B(x_i^{n_p}, \varepsilon/2)$. We know there always exists a ball of radius $\varepsilon/4$ which is a subset of F_p. This implies by Lemma 2 that the probability of the event $\{z_{n_p} \in F_p\}$ is bounded below by some constant $Q > 0$, and will thus occur infinitely often.

By the definition of a limit, we know there must exist an N such that $|x_i^n - z| \leq \varepsilon/8\gamma$, for all $n \geq N$, where γ is the bound on the weights given by Lemma 1. Choose P large enough so that $n_p \geq N$ for all $p \geq P$.

If $z_{n_p} \in F_p$, then $|z_{n_p} - x_i^n| \geq \varepsilon/2$, implying

$$|x_j^{n_p+1} - z| + |x_j^{n_p} - z| \geq |x_j^{n_p+1} - x_j^{n_p}| = \frac{|z_{n_p} - x_j^{n_p}|}{w_i^{n_p+1}} \geq \frac{\varepsilon}{2\gamma}.$$

By this inequality, either $|x_j^{n_p} - z| \geq \varepsilon/4\gamma$ or $|x_j^{n_p+1} - z| \geq \varepsilon/4\gamma$. Because $z_{n_p} \in F_p$ infinitely often, we know it must occur for some $p \geq P$, and thus gives us a contradiction. □

2.4 A Simple Model for the Motion of the Perceptual Boundary

In the remainder of this chapter, we will study a simple special case of our model. We consider a system with just two categories. We let our domain be $E = [0, 1]$ and we stipulate that new exemplars arrive in the system with uniform probability density on E. These choices correspond to $k = 2$ and $f(x) = 1$ for all $x \in E$. Figure 2.2 shows a state of our model for these choices. We perform a detailed study of the dynamics of the perceptual boundary in this case, providing a simple probabilistic model of its motion.

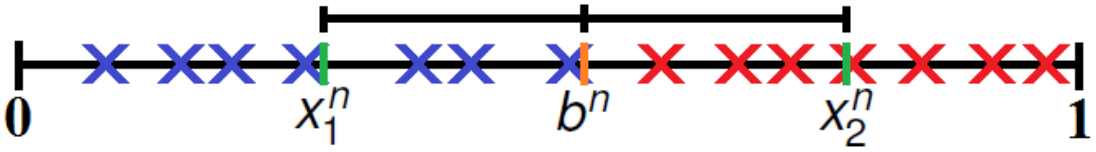

Figure 2.2: A representative state of our model in the special case of $E = [0, 1]$, uniform probability distribution on $[0, 1]$ and two categories (blue and red). x_1^n and x_2^n are the corresponding category means, and b^n is the perceptual boundary between the two categories.

Our goal in this section is to characterize the behaviour of b^n, the location of the perceptual boundary as function of n. The sequence $b^n, n \geq 0$ is a discrete-time stochastic

process, where the randomness in the evolution of b^n enters through the i.i.d. random variables $z_n, n \geq 0$.

To motivate the study of the perceptual boundary b^n and its fluctuations, see Figure 2.3 in which we show time series for x_1^n, x_2^n and b^n. The left graph shows the evolution of the system where $\lambda > 0$, and the right where $\lambda = 0$ (MacQueen's model [15]). The red lines represent the two weighted exemplar means x_1^n and x_2^n. The blue line is the perceptual boundary given by $b^n = (x_1^n + x_2^n)/2$, which represents the boundary between the Voronoi cells (S_1^n and S_2^n) of the two categories. In the left plot ($\lambda > 0$) the category means fluctuate with roughly the same amplitude for the whole interval, whereas on the right they appear to converge. Even in this simplest case of our model, we are not able to analyze the system completely, so we instead derive an even simpler approximation to our model.

We derive an autoregressive model as an approximation to the behaviour of b^n [16]. An autoregressive-moving average model of order (p, ℓ) ($ARMA(p, \ell)$ model), for a real time series $y_n \in \mathbb{R}$, $n = 0, 1, 2, \ldots$ is given by

$$y_{n+1} = \alpha_1 y_n + \alpha_2 y_{n-1} + \ldots + \alpha_p y_{n-p+1} + \eta_n + \beta_1 \eta_{n-1} + \ldots + \sigma_\ell \eta_{n-\ell},$$

where $\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_\ell \in \mathbb{R}$, and η_n is an error sequence which is frequently assumed to be Gaussian [7, 18]. The model (described in Section 2.2) is a Markov process, so we can approximate it as an $ARMA(1, 0)$ model. An $ARMA(1, 0)$ model, is better known as an autoregressive first-order ($AR(1)$) model. Using this result, we will determine the best approximation to the dynamics of b^n among $AR(1)$ models, deriving coefficients in terms of λ.

An $AR(1)$ model for a real time series $y_n \in \mathbb{R}$, $n = 0, 1, 2, \ldots$ is given by

$$y_{n+1} = \alpha y_n + \sigma \eta_n, \tag{2.7}$$

where $\alpha \in \mathbb{R}$, η_n is a sequence of i.i.d. standard Gaussian random variables, and $\sigma \geq 0$ is a scalar. When $|\alpha| < 1$ the sequence y_n converges to a stationary stochastic process with stationary distribution $N(0, \sigma^2/(1 - \alpha^2))$.

We assume the initial exemplar means are ordered such that $x_1^0 < x_2^0$, implying $x_1^n < x_2^n$, for all $n \geq 0$. Let the perceptual boundary between the 2 exemplar means be b^n. It is straightforward to show $b^n = (x_1^n + x_2^n)/2$.

Recall from Section 2.2 that the evolution of the category means of the system can be completely expressed using only the category means and the category weights: the positions of individual stored exemplars do not enter the dynamics. Accordingly, we begin deriving our approximate model by defining the sequence of random vectors

$$\mathbf{Z}^n = \begin{bmatrix} x_1^n, & x_2^n, & w_1^n, & w_2^n \end{bmatrix}^T,$$

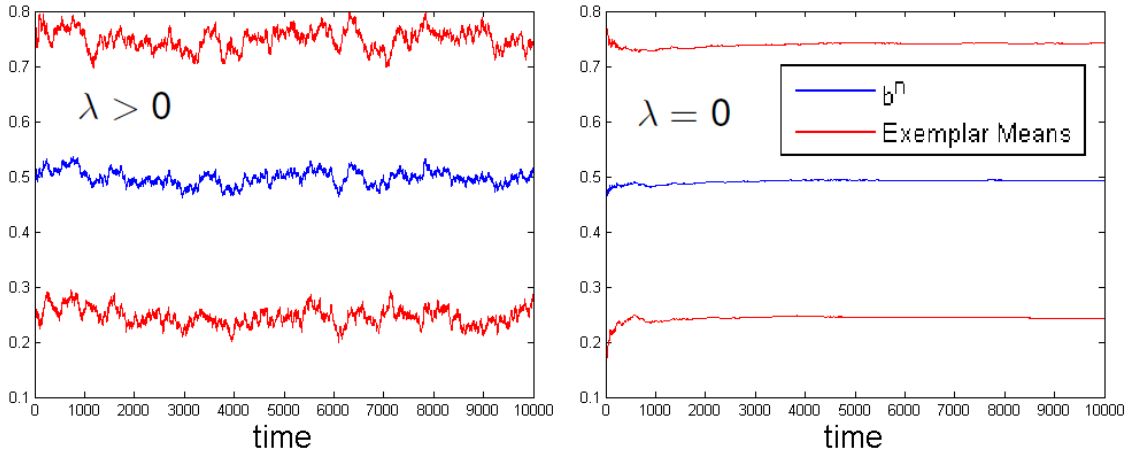

Figure 2.3: A comparison of the evolution of simulations of the two-category uniform distribution system 1-D case. The left graph exhibits what the system looks like when $\lambda = 0.01$ and initial values $w_1^0 = w_2^0 = W/2 \cong 50$. On the right we set $\lambda = 0$ and $w_1^0 = w_2^0 = 10$. For both graphs we set $x_1^0 = 1/6$, and $x_2^0 = 2/3$.

for $n \geq 0$. We can exactly express the evolution of $\mathbf{Z}^n$ as a random dynamical system

$$\mathbf{Z}^{n+1} = \Phi(\mathbf{Z}^n, z_n)$$

where we see that each $\mathbf{Z}^{n+1}$ is determined as a function of the previous value $\mathbf{Z}^n$ and the random variable z_n. Hence the randomness of $\mathbf{Z}^{n+1}$ only enters through z_n once $\mathbf{Z}^n$ is known. The expression for Φ is

$$\Phi(\mathbf{Z}^n, z_n) = \begin{cases} \left[\left(\dfrac{x_1^n w_1^n e^{-\lambda} + z_n}{w_1^n e^{-\lambda} + 1}\right), \quad x_2^n, \quad \left(w_1^n e^{-\lambda} + 1\right), \quad w_2^n e^{-\lambda}\right]^T & \text{if } z_n \leq \dfrac{x_1^n + x_2^n}{2} \\ \left[x_1^n, \quad \left(\dfrac{x_2^n w_2^n e^{-\lambda} + z_n}{w_2^n e^{-\lambda} + 1}\right), \quad w_1^n e^{-\lambda}, \quad \left(w_2^n e^{-\lambda} + 1\right)\right]^T & \text{if } z_n > \dfrac{x_1^n + x_2^n}{2} \end{cases}.$$

This expression is exact but unwieldy, so we derive an approximate model by linearizing the system about a point where we expect the invariant measure to be densest. Define $F(\mathbf{x}) = \mathbb{E}(\Phi(\mathbf{x}, z))$, where z is uniform on $[0, 1]$. Let $\mathbf{Z}^*$ be the vector such that $\mathbf{Z}^* = \mathbb{E}(\Phi(\mathbf{Z}^*, z))$ [1]. This can be thought of as an analogue of the fixed point of a dynamical system for our random dynamical system. We linearize the random dynamical system about this point. We can expect our system to be well approximated by the linearized system if fluctuations about $\mathbf{Z}^*$ are not too large.

First, one can determine that $\mathbf{Z}^*$ is

$$\mathbf{Z}^* = \left[1/4, \quad 3/4, \quad W/2, \quad W/2\right]^T, \tag{2.8}$$

where $W = (1 - e^{-\lambda})^{-1}$ as defined in the proof for Lemma 1.

Define another random variable $\mathbf{y}^n = \mathbf{Z}^n - \mathbf{Z}^*$. The Jacobian of F at $\mathbf{Z}^*$ is $J := \partial F(\mathbf{Z}^*)$, and the covariance matrix of the random perturbation at $\mathbf{Z}^*$ is $H := \mathbb{E}(G(\mathbf{Z}^*)G(\mathbf{Z}^*)^T)$, where $G(\mathbf{x}, z) = \Phi(\mathbf{x}, z) - F(\mathbf{x})$. Let $\mathbf{d}^n$, for $n \geq 0$, be an i.i.d. sequence of random variables each distributed as $N(0, H)$. The $AR(1)$ model of $\mathbf{y}^n$ is written

$$\mathbf{y}^{n+1} = J\mathbf{y}^n + \mathbf{d}^n, \tag{2.9}$$

and is an approximation for the dynamics of $\mathbf{Z}^n - \mathbf{Z}^*$ [17, 36]. This dynamical system can also be expressed as $\mathbf{y}^{n+1} = J\mathbf{y}^n + H^{1/2}N(0, I)$.

The matrices J, H, and $H^{1/2}$, can be found through some tedious calculations to be

$$J = \begin{bmatrix} \frac{5 - e^{-\lambda}}{4(2 - e^{-\lambda})} & \frac{1 - e^{-\lambda}}{4(2 - e^{-\lambda})} & 0 & 0 \\ \frac{1 - e^{-\lambda}}{4(2 - e^{-\lambda})} & \frac{5 - e^{-\lambda}}{4(2 - e^{-\lambda})} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & e^{-\lambda} & 0 \\ -\frac{1}{2} & -\frac{1}{2} & 0 & e^{-\lambda} \end{bmatrix},$$

$$H = \begin{bmatrix} \frac{(1 - e^{-\lambda})^2}{24(2 - e^{-\lambda})^2} & 0 & 0 & 0 \\ 0 & \frac{(1 - e^{-\lambda})^2}{24(2 - e^{-\lambda})^2} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & -\frac{1}{4} \\ 0 & 0 & -\frac{1}{4} & \frac{1}{4} \end{bmatrix},$$

and

$$H^{1/2} = \frac{1}{2\sqrt{2}} \begin{bmatrix} \frac{(1-e^{-\lambda})}{\sqrt{3}(2-e^{-\lambda})} & 0 & 0 & 0 \\ 0 & \frac{(1-e^{-\lambda})}{\sqrt{3}(2-e^{-\lambda})} & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

An $AR(1)$ model for the perceptual boundary b^n can be derived using the $AR(1)$ process we have found for random variable $\mathbf{y}^n$. First, we calculate the first 2 components of $\mathbf{y}^n$ using Equation 2.9,

$$\begin{bmatrix} x_1^{n+1} - \frac{1}{4} \\ x_2^{n+1} - \frac{3}{4} \end{bmatrix} = \begin{bmatrix} \frac{5-e^{-\lambda}}{4(2-e^{-\lambda})}\left(x_1^n - \frac{1}{4}\right) + \frac{1-e^{-\lambda}}{4(2-e^{-\lambda})}\left(x_2^n - \frac{3}{4}\right) + \frac{1-e^{-\lambda}}{2\sqrt{6}(2-e^{-\lambda})}N(0,1) \\ \frac{5-e^{-\lambda}}{4(2-e^{-\lambda})}\left(x_2^n - \frac{3}{4}\right) + \frac{1-e^{-\lambda}}{4(2-e^{-\lambda})}\left(x_1^n - \frac{1}{4}\right) + \frac{1-e^{-\lambda}}{2\sqrt{6}(2-e^{-\lambda})}N(0,1) \end{bmatrix}.$$

Noting that $b^n - 1/2 = \left[(x_1^n - 1/4) + (x_2^n - 3/4)\right]/2$, we add the two components of this vector together and divide by 2 to get

$$b^{n+1} - \frac{1}{2} = \frac{1}{2}\left(\frac{3-e^{-\lambda}}{2-e^{-\lambda}}\right)\left(b^n - \frac{1}{2}\right) + \frac{1}{4\sqrt{3}}\left(\frac{1-e^{-\lambda}}{2-e^{-\lambda}}\right)N(0,1).$$

This is an $AR(1)$ model for the perceptual boundary, it can be expressed as

$$Y^{n+1} = \alpha Y^n + \sigma\eta_n, \tag{2.10}$$

where $Y^n = b^n - 1/2$, the variable η_n is a noise term drawn from the standard normal distribution $N(0,1)$, and

$$\alpha = (3-e^{-\lambda})\left[2(2-e^{-\lambda})\right]^{-1}, \qquad \sigma = \left(1-e^{-\lambda}\right)\left[4\sqrt{3}(2-e^{-\lambda})\right]^{-1}.$$

Equation 2.10 is our simplified probabilistic model for the motion of the boundary $b^n - 1/2$.

As $\lambda \to 0$, the variables α and σ approach 1 and 0 respectively, meaning the stochastic process approaches the case where $b^{n+1} = b^n$. This fits with the fact that in when $\lambda = 0$, our original model reverts to the MacQueen model, in which there are no fluctuations as $n \to \infty$.

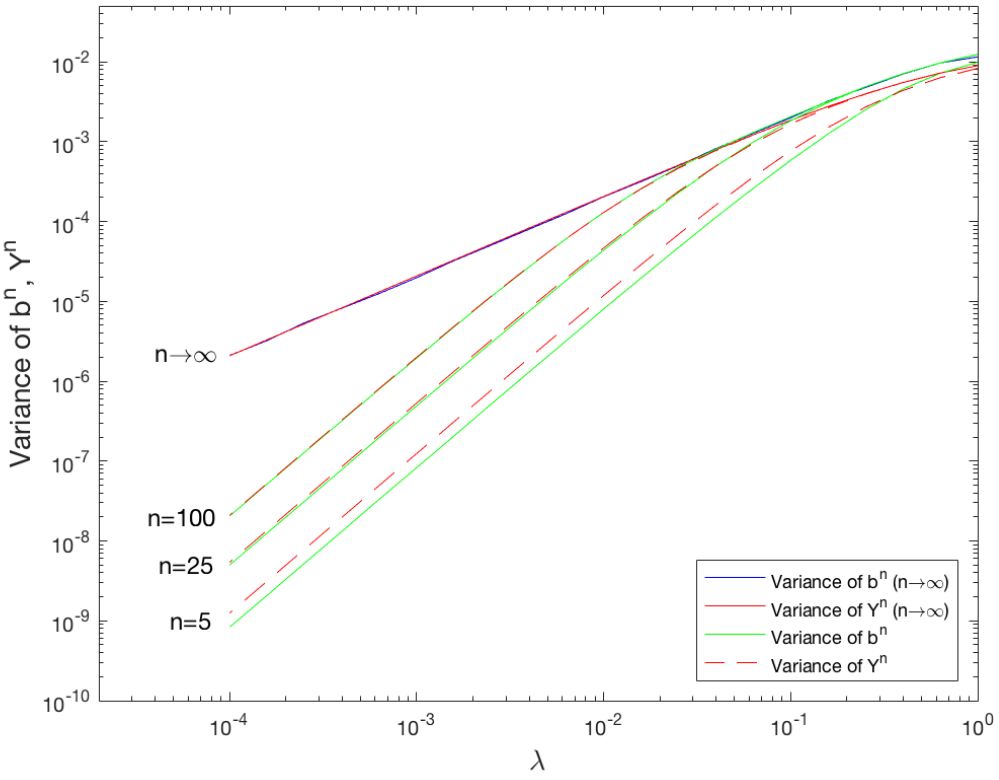

Figure 2.4: The variances of b^n and Y^n versus λ for various times n. We set $x_1^0 = 1/4$, $x_2^0 = 3/4$, and $w_1^0 = w_2^0 = W/2$, so that we start at the fixed point of the system as given in Equation 2.8. The $n \to \infty$ case for b^n is approximated by setting $n = \lceil 400/\lambda \rceil$.

In order to assess the quality of the AR(1) process Y^n as a model for b^n, we compare the variance of Y^n with the variance of b^n while varying n and λ. For the process Y^n, [28] provides an exact formula for the variance as a function of time when $Y^0 = 0$:

$$\begin{aligned}
\mathrm{Var}[Y^n] = \mathbb{E}(Y^n)^2 &= \sum_{j=0}^{n-1} \alpha^{2j}\sigma^2 \\
&= \frac{1}{192}\left(\frac{We^{-\lambda}}{2} + 1\right)^{-2} \sum_{j=0}^{n-1}\left(1 - \frac{1}{2We^{-\lambda}}\right)^{2j} \\
&= \frac{1}{48}\left(\frac{1-e^{-\lambda}}{2-e^{-\lambda}}\right)^{2} \sum_{j=0}^{n-1}\left(\frac{3e^{-\lambda}-1}{2e^{-\lambda}}\right)^{2j}.
\end{aligned}$$

One can show using algebra that when taking the limit as n approaches infinity of the above equation that the sum converges under the condition that $0 < \lambda < \ln(5) \cong 1.6$,

$$\lim_{n\to\infty} \mathrm{Var}[Y^n] = \frac{1 - e^{-\lambda}}{12e^{2\lambda}(2 - e^{-\lambda})^2(5e^{-\lambda} - 1)}.$$

For the original process b^n, we estimate the variance using Monte Carlo simulation, starting from an initial condition where $\mathbf{Z}^0 = \mathbf{Z}^*$, meaning that $\mathbb{E}b^n = b^0$ for all n. Figure 2.4 shows a comparison between the variance of b^n (calculated from simulations of the model) in green and blue, and the variance of the AR(1) approximation Y^n in red.

As expected, for all λ the variances of b^n and Y^n increase up to the equilibrium value as $n \to \infty$. The equilibrium values of the variances of b^n and Y^n are indistinguishable to within the accuracy we compute them here. However, for lower values of n the variance Y^n is a significantly poorer approximation to that of b^n indicating that the AR(1) model is not as good at catching non-equilibrium behaviour.

Chapter 3

Effects of Limiting Memory Capacity on the Behaviour of Exemplar Dynamics

3.1 Introduction

Figure 3.1: Comparison of the usage of "cider" (blue) and its archaic spelling "cyder" (red) within a corpus of books between the years 1800 and 2000. The y-axis represents the percentage of the usages of the words in the entire database. This image was generated by Google Ngram Viewer [19].

In spoken and written language, there are instances where there are two or more variants of a word, each of which is equivalent from the point of view of communication. We can think of instances of the word as belonging to one of two or more categories. For example, a population might pronounce the word "either" as both "ee-ther" and "eye-ther". Another example is when you have different spellings of a word. Figure 3.1, which was generated by

Google Ngram Viewer [19], shows in the written lexicon a comparison between the usage of the word "cider" and its archaic spelling "cyder". In the year 1800 "cyder" seems to have been the more popular spelling but it has become practically extinct since then. As we see in this example, it is possible for a category to become extinct, passing out of usage. Here we study a model for just this kind of category extinction.

The extinction of a category occurs when the weights of all the exemplars labelled in that category approach zero. This represents the listener no longer remembering the category. The listener will cease to produce tokens from that category. A necessary condition for the extinction of a category is that the probability of classifying a sound as that category must approach zero. In this chapter we will be particularly interested in when there is extinction of all but one category.

This chapter is motivated by research in exemplar dynamics done by Tupper [31] and Wedel [32] (these models are described in detail in Chapter 4, Sections 4.2 and 4.3 respectively). They both studied similar exemplar models, with the following differences:

1. [32] used anti-ambiguity (see Section 4.2) to keep categories from becoming extinct, whereas [31] used rejection (see Section 4.3).

2. In Wedel's simulations of the model in [32], the number of stored exemplars per category was limited to a maximum of 100. In [31], there was no limitation on the number of stored exemplars per category.

3. A segmental bias was implemented in [32] to keep the categories aligned, whereas [31] did not.

4. The word and preferred value biases were implemented differently in the two papers.

We will discuss the difference between using anti-ambiguity and rejection, as well as segmental bias later in this thesis (see Chapter 4). For now it will suffice to say that rejection is better than anti-ambiguity at keeping categories from becoming extinct. Based on simulations we ran, as well as the results of this thesis, we do not believe that the segmental bias or the difference in how the word and preferred value biases were implemented affect the stability of the system. The subtle difference between how in [32] the number of exemplars were limited to 100, and in [31] they were unlimited, will be of particular interest to us in this chapter. Although it might not seem like it, limiting the number of exemplars per category in an exemplar model has an effect on the long term evolution of the system.

It was demonstrated in [31] that when the number of exemplars stored per category is unlimited, then there is extinction of all but one category. In [32], it was observed that there is no category extinction in simulations for a certain choice of parameters when the exemplars stored per category is limited to 100. However, in [32], Wedel only did numerical simulations of his model up to 4000 iterations.

This begs the question, when you limit the number of exemplars to be stored per category, will categories eventually become extinct? In this chapter we seek the answer to this question. The models of [31] and [32] are too complicated to investigate rigorously, so we study a simpler model which captures some of their essential features.

In Section 3.2, we describe our simple exemplar model. Our model only depends on three parameters: the number of categories k, the decay rate λ, and the number of exemplars stored per category N. Two particular cases of this general model will be studied: one where we limit the number of exemplars ($N < \infty$, as in [32]) in Section 3.3, and another where we do not ($N = \infty$, as in [31]) in Section 3.4. We prove in both cases that all categories but one will become extinct. In Section 3.5 we discuss computational results, which demonstrates how limiting the number of exemplars affects the system's evolution. The numerical simulations in this section will help us explain the effect N and λ have on the expected time to extinction.

3.2 Simple Exemplar Weight Model

In this section we describe a simplified exemplar model. The parameters for the system are the number of categories, k, the number of exemplars stored per category, N, and the decay rate, λ. The listener starts with some exemplars with associated weights in each category, and then receives a stream of new inputs (sounds). The listener in this model will decide how to classify new sounds only using the total weights of the exemplars in each category. The phonetic information stored in exemplars will not be utilized in the categorization process.

At time n, let $w_{j,m}^n$ be the weight of the mth exemplar where $m \in \mathbb{N}$, for category j. At time n, these k infinite sequences of real numbers comprise the state of the system. Let N be the maximum number of exemplars per category the listener is permitted to store. Let $\lambda > 0$ be the decay rate of the weights, so that at each time step n, the weights of old memories will decay by a factor of $\beta = e^{-\lambda}$. New exemplars are given a weight $W_0 = 1$. Additionally, when $N < \infty$, if there are $N+1$ exemplars in a category with non-zero weight upon adding a new exemplar, then the exemplar with the lowest weight is discarded.

We assume that exemplars are ordered by weight at all times, so that $0 \leq w_{j,m+1}^n \leq w_{j,m}^n \leq 1$, for all $n \in \mathbb{N}_0$, $j \in \{1, \ldots, k\}$, and $m < N$. The initial conditions of the weights are non-random, and can be anything such that $0 \leq w_{j,m}^0 \leq 1$, for all $j \in \{1, \ldots, k\}$, and $m \leq N$, at least one of the weights in one category must be non-zero, and if $N < \infty$, then $w_{j,m}^0 = 0$ for all $j \in \{1, \ldots, k\}$, and $m > N$.

Let $W_j^n := \sum_{m=1}^N w_{j,m}^n$ be the total weight of exemplars in category $j \in \{1, \ldots, k\}$, and $W_{tot}^n := \sum_{j=1}^k W_j^n$ be the total weight of all exemplars.

Let x_n be the category we classify the nth sound as at time n. For example, $x_n = j$ means we classified the nth sound as category j. We let the probability of classifying the

nth sound as category j ($x_n = j$), given the state of the system in the previous time step be W_j^n/W_{tot}^n. This classification procedure is the Luce choice rule [14]. As such, the categorization of sounds only depends on the weights of the exemplars, unlike other models where the phonetic information stored in exemplars is used to classify sounds.

Before discussing a σ-field, first we look at the definition of a probability triple as written in [26].

Definition 1. *A σ-field (or equivalently σ-algebra) $\mathcal{F}$ is a collection of subsets of Ω (the sample space), containing Ω itself and the empty set $\emptyset$, and closed under the formation of complements and countable unions and countable intersections.*

We let

$$\mathcal{F}_n = \sigma(w_{j,m}^q,\ 0 < q \leq n,\ j \in \{1, \ldots, k\},\ m \in \mathbb{N}), \tag{3.1}$$

so $\mathcal{F}_n$ is the smallest σ-field with respect to which each $w_{j,m}^q$, such that $0 < q \leq n$, $j \in \{1, \ldots, k\}$, and $m \in \mathbb{N}$, is measurable [3, pg.64]. This σ-field will be important in the next two sections. One thing to note is that this sequence of σ-fields forms a filtration $\{\mathcal{F}_n\}_{n\geq 1}$. A filtration $\{\mathcal{F}_n\}_{n\geq 1}$, is an indexed family of σ-fields such that $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$, for all $n \geq 1$ [5, pg.491].

Another way to describe the process is the following: At time step n, $\mathbf{P}(x_n = j|\mathcal{F}_n) = W_j^n/W_{tot}^n$, for each $j \in \{1, \ldots, k\}$. If $x_n = j$, then:

1. Let $w_{j,m+1}^{n+1} = \beta w_{j,m}^n$, for all $m < N$, and $w_{j,1}^{n+1} = W_0 = 1$. If $N < \infty$, the exemplar corresponding to the Nth position of category j from the previous time step will be discarded: Thus, if $N < \infty$, we let $w_{j,N+1}^{n+1} = 0$.

2. For all $i \neq j$, and $m \in \{1, \ldots, N\}$, let $w_{i,m}^{n+1} = \beta w_{i,m}^n$.

The next couple of sections are devoted to proving the model just described always results in the extinction of all but one category. Sections 3.3 and 3.4 will respectively look at the cases where $N < \infty$ and $N = \infty$.

3.3 Finite Stored Exemplars Model

In this section we will show that when $N < \infty$, all but one category will become extinct with a probability of 1. That is, it will be proved that with a probability of 1, there exists an M and a j, such that $x_n = j$, for all $n \geq M$.

The following lemma proves if one classifies p consecutive sounds as category j, then it only increases the probability of the next sound being classified as category j. Before we state the lemma, recall that a conditional probability can only be defined a.s. (on sets of non-zero measure) as in [26].

Lemma 7. *If $N < \infty$, and $\mathcal{F}_n$ is the σ-field defined by Equation 3.1, then*

$$\mathbf{P}(x_{n+p} = j | x_{n+p-1} = j, \ldots, x_n = j, \mathcal{F}_n) \geq \mathbf{P}(x_n = j | \mathcal{F}_n),$$

a.s., for all $p \in \mathbb{N}_0$, and $j \in \{1, \ldots, k\}$.

Proof. We will prove this lemma using induction. Let $S(p)$ be the statement that

$$\mathbf{P}(x_{n+p} = j | x_{n+p-1} = j, \ldots, x_n = j, \mathcal{F}_n) \geq \mathbf{P}(x_n = j | \mathcal{F}_n),$$

a.s. We want to prove $S(p)$ is true for all $p \in \mathbb{N}_0$.

The initial statement $S(0)$, which is $\mathbf{P}(x_n = j|\mathcal{F}_n) \geq \mathbf{P}(x_n = j|\mathcal{F}_n)$, is true because the two sides are equal.

Now we assume the inductive hypothesis $S(p)$ is true; $\mathbf{P}(x_{n+p} = j|x_{n+p-1} = j, \ldots, x_n = j, \mathcal{F}_n) \geq \mathbf{P}(x_n = j|\mathcal{F}_n)$. We want to show $S(p+1)$ is true. If $\{x_{n+p} = j, \ldots, x_n = j\}$, then $W_j^{n+p+1} = \beta W_j^{n+p} + 1 - \beta w_{j,N}^{n+p}$, and $W_{tot}^{n+p+1} = \beta W_{tot}^{n+p} + 1 - \beta w_{j,N}^{n+p}$. One can show via simple algebra, using the facts that $\beta^{-1}(1 - \beta w_{j,N}^{n+p}) > 0$, and $W_j^{n+p} \leq W_{tot}^{n+p}$, that

$$\begin{aligned} \mathbf{P}(x_{n+p+1} = j | x_{n+p} = j, \ldots, x_n = j, \mathcal{F}_n) &= \frac{W_j^{n+p} + \beta^{-1}(1 - \beta w_{j,N}^{n+p})}{W_{tot}^{n+p} + \beta^{-1}(1 - \beta w_{j,N}^{n+p})} \\ &\geq \frac{W_j^{n+p}}{W_{tot}^{n+p}} \end{aligned} \tag{3.2}$$

a.s. The right hand side of Equation 3.2 is equal to $\mathbf{P}(x_{n+p} = j|x_{n+p-1} = j, \ldots, x_n = j, \mathcal{F}_n)$, which implies by the induction hypothesis that statement $S(p+1)$ is true. □

In order to prove the next lemma, we require Theorem 15.5 in [27], which states the following:

Theorem 5. *Suppose $0 \leq u_n < 1$. Then*

$$\prod_{n=1}^{\infty}(1 - u_n) > 0 \text{ if and only if } \sum_{n=1}^{\infty} u_n < \infty.$$

Lemma 8. *Let $A_n = \{\exists j, x_m = j, \forall m \geq n\}$, be the event that we only classify input sounds as a single category from time step n onwards. If $N < \infty$, and $\mathcal{F}_n$ is the σ-field defined by Equation 3.1, then there exists a $Q > 0$, such that $\mathbf{P}(A_n|\mathcal{F}_n) \geq Q$, for all n.*

Proof. At time step n, there must exist a category $c \in \{1, \ldots, k\}$, such that $\mathbf{P}(x_n = c|\mathcal{F}_n) \geq k^{-1}$. By Lemma 7, we get the following inequality,

$$\begin{aligned} &\mathbf{P}(x_{n+N-1} = c, \ldots, x_{n+1} = c, x_n = c|\mathcal{F}_n) \\ &= \prod_{p=0}^{N-1} \mathbf{P}(x_{n+p} = c|x_{n+p-1} = c, \ldots, x_n = c, \mathcal{F}_n) \geq k^{-N}. \end{aligned} \tag{3.3}$$

If $x_q = c$, for $n \leq q \leq n + N - 1$, then $W_c^{n+N} = \sum_{q=0}^{N-1} \beta^q$, and if we continue to categorize $x_q = c$, for $q > n + N - 1$, the weight for category c will stay constant. Let $\Omega = \sum_{q=0}^{N-1} \beta^q$. Upon inspection, it is apparent that $W_i^n \leq \Omega$, for all i and $n > N - 1$, because it is the maximum total weight a category can have after there have been at least N time steps.

If $x_p = c$, for $n \leq p \leq n + q$, where $q \geq N - 1$, then

$$W_{tot}^{n+q} = \sum_{p=1}^{k} W_p^{n+q} = W_c^{n+q} + \sum_{p \neq c} W_p^{n+q} \leq \Omega + \sum_{p \neq c} \beta^q \Omega < \Omega(1 + k\beta^q).$$

Let $G_{n,N} = \{x_{n+N-1} = c, \ldots, x_{n+1} = c, x_n = c\}$. The probability of categorizing the next sound as c, given that we have only categorized as c since time step n, and have at least done so N times in a row can be bounded below,

$$\begin{aligned} \mathbf{P}(x_{n+N-1+q} = c|x_{n+N-2+q} = c, \ldots, x_{n+N} = c, G_{n,N}, \mathcal{F}_n) &= \frac{\Omega}{W_{tot}^{n+N-1+q}} \\ &\geq \frac{\Omega}{\Omega(1 + k\beta^q)} \\ &= 1 - \frac{k\beta^q}{1 + k\beta^q}, \end{aligned} \tag{3.4}$$

for all $n, q, N > 0$. Note we used the fact that $W_j^{n+N-1+q} = \Omega$, because the event $G_{n,N}$ had already occurred.

Utilizing Equations 3.3 and 3.4,

$$\begin{aligned} \mathbf{P}(x_m = c, \forall m \geq n|\mathcal{F}_n) &= \mathbf{P}\left(\bigcap_{q=0}^{\infty} \{x_{n+q} = c\} \,\middle|\, \mathcal{F}_n\right) \\ &= \mathbf{P}\left(\bigcap_{q=1}^{\infty} \{x_{n+N-1+q} = c\} \,\middle|\, G_{n,N}, \mathcal{F}_n\right) \mathbf{P}(G_{n,N}|\mathcal{F}_n) \\ \geq k^{-N} \prod_{q=1}^{\infty} \mathbf{P}(x_{n+N-1+q} = c|x_{n+N-2+q} = c, \ldots, x_{n+N} = c, G_{n,N}, \mathcal{F}_n) \\ &\geq k^{-N} \prod_{q=1}^{\infty} \left(1 - \frac{k\beta^q}{1 + k\beta^q}\right). \end{aligned} \tag{3.5}$$

By Theorem 5 [27], the product in Equation 3.5 is strictly greater than 0 if and only if

$$\sum_{q=1}^{\infty} \frac{k\beta^q}{1+k\beta^q} < \infty.$$

By the ratio test we know this series is convergent. Therefore, there is a $Q > 0$, such that $\mathbf{P}(x_m = c, \forall m \geq n | \mathcal{F}_n) \geq Q > 0$.

Since

$$\mathbf{P}\left(\exists j, x_m = j, \forall m \geq n | \mathcal{F}_n\right) \geq \mathbf{P}\left(x_m = c, \forall m \geq n | \mathcal{F}_n\right) \geq Q > 0,$$

we get the final result. □

Lemma 8 states that the probability of A_n occurring, given any event which only depends on the events up to time step $n-1$, can be bounded below by a constant $Q > 0$. In other words, the probability of x_m being classified as the same category for all $m \geq n$, always has at least a certain probability of happening no matter what occurs before it.

Note, if A_n is true for any value of n, the rest of the categories $i \neq j$ will become extinct. If we prove that $\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = 1$, then we have proved there is almost surely extinction of all but one category when $N < \infty$.

Lemma 9. *Let $\mathcal{G}$ be a σ-field, and $G \in \mathcal{G}$, such that $\mathbf{P}(G) > 0$. If there exists a $Q > 0$, s.t. $\mathbf{P}(X|\mathcal{G}) \geq Q$, a.s., then $\mathbf{P}(X|G) \geq Q$.*

Proof. By the definition of the probability of an event conditioned on a σ-field [26, p.155] which states,

$$\mathbb{E}(\mathbf{P}(X|\mathcal{G})\mathbb{1}_G) = \mathbf{P}(X \cap G)$$

we get the result

$$\mathbf{P}(X|G) = \frac{\mathbf{P}(X \cap G)}{\mathbf{P}(G)} = \frac{\mathbb{E}\left[\mathbf{P}(X|\mathcal{G})\mathbb{1}_G\right]}{\mathbf{P}(G)} \geq \frac{\mathbb{E}\left[Q\mathbb{1}_G\right]}{\mathbf{P}(G)} \geq Q.$$

□

Theorem 6. *When $N < \infty$, all categories but one will become extinct with a probability of 1: that is, $\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = 1$, where $A_n = \{\exists j, x_m = j, \forall m \geq n\}$.*

Proof. The proof utilizes Murphy's Law, a general statement proven in [30]. Murphy's Law states the following: Let $(G_n, n \geq 1)$ be any sequence of events satisfying the condition $G_n \subseteq G_{n+1}$, for all $n \geq 1$, and let $G = \bigcup_{n=1}^{\infty} G_n$. If $\mathbf{P}(G|G_n^c) \geq \varepsilon > 0$, for all $n \geq 1$, then $\mathbf{P}(G) = 1$.

We know $A_n \subseteq A_{n+1}$, for all $n \geq 1$. Let $A = \bigcup_{n=1}^{\infty} A_n$. By Murphy's law, if we can show $\mathbf{P}(A|A_n^c) \geq \varepsilon > 0$, for all n, then $\mathbf{P}(A) = 1$, proving the theorem.

Let $Y_n = \min\{m \in \mathbb{N} : x_{n+m} \neq x_n\}$, and if it is not defined then let $Y_n = \infty$. The event $\{Y_n = m\}$ is a subset of $A_n^c = \{\exists j > n, \text{ s.t. } x_j \neq x_n\}$, for all n, and $A_n^c = \bigcup_{m>0}\{Y_n = m\}$. Using the fact that the events $\{Y_n = i\}$ and $\{Y_n = j\}$ are disjoint when $i \neq j$, we obtain the following,

$$\begin{aligned}
\mathbf{P}(A|A_n^c) &= \frac{\mathbf{P}(A \cap A_n^c)}{\mathbf{P}(A_n^c)} = \frac{\mathbf{P}\left(A \cap \left(\bigcup_{m>0}\{Y_n = m\}\right)\right)}{\mathbf{P}(A_n^c)} \\
&= \frac{\mathbf{P}\left(\bigcup_{m>0}\{A \cap \{Y_n = m\}\}\right)}{\mathbf{P}(A_n^c)} \\
&= \frac{\sum_{m>0} \mathbf{P}\left(A \cap \{Y_n = m\}\right)}{\mathbf{P}(A_n^c)} \\
&\geq \frac{\sum_{m>0} \mathbf{P}\left(A_{n+m+1} \cap \{Y_n = m\}\right)}{\mathbf{P}(A_n^c)}
\end{aligned}$$

since $A_{n+m+1} \subseteq A$. Using Lemma 8 with Lemma 9 (noting $\{Y_n = m\}$ is in $\mathcal{F}_{n+m}$), and that $\bigcup_{m>0}\{Y_n = m\} = A_n^c$, we obtain

$$\begin{aligned}
\frac{\sum_{m>0} \mathbf{P}\left(A \cap \{Y_n = m\}\right)}{\mathbf{P}(A_n^c)} &= \sum_{m>0} \mathbf{P}\left(A_{n+m+1}|\{Y_n = m\}\right) \frac{\mathbf{P}\left(Y_n = m\right)}{\mathbf{P}(A_n^c)} \\
&\geq Q \sum_{m>0} \mathbf{P}\left(Y_n = m|A_n^c\right) \\
&= Q \cdot \mathbf{P}\left(\bigcup_{m>0} \{Y_n = m\} \middle| A_n^c\right) = Q > 0.
\end{aligned}$$

□

3.4 Infinite Stored Exemplars Weight Model

This section will be devoted to studying the special case of the model where the listener stores an infinite number of exemplars, so $N = \infty$. The proof for showing there is almost surely extinction of all but one category in this special case will be different from the previous section.

Let

$$Z_j^n = \mathbf{P}(x_n = j|\mathcal{F}_n) = W_j^n / W_{tot}^n, \tag{3.6}$$

where $\mathcal{F}_n$ is as defined in Equation 3.1. Note the combined weights of all categories which are not j is equal to $W_{tot}^n - W_j^n$. We will first re-describe the model's evolutionary process in terms of W_j^n and W_{tot}^n, in order to simplify the proof. The evolutionary process evolves as follows:

- If $x_n = j$, then the total weight of category j becomes $W_j^{n+1} = 1 + W_j^n\beta$, and the total weight of all other categories besides j becomes $W_{tot}^{n+1} - W_j^{n+1} = (W_{tot}^n - W_j^n)\beta$.
- If $x_n \neq j$, then the total weight of all categories besides j is $W_{tot}^{n+1} - W_j^{n+1} = 1 + (W_{tot}^n - W_j^n)\beta$, and the total weight of category j is $W_j^{n+1} = W_j^n\beta$.

We want to prove there exists a category j, such that $Z_j^n \to 1$, a.s., as $n \to \infty$, and for the rest of the categories $q \neq j$, that $Z_q^n \to 0$, a.s. We will then show that if $Z_j^n \to 0$, a.s., then $W_j^n \to 0$, a.s. As such we would prove all categories but one become extinct. In order to prove this result, we require a few lemmas.

Lemma 10. *Let $Z_j^n = W_j^n/W_{tot}^n$. If the number of exemplars per category stored is $N = \infty$, then the random variable Z_j^n is a martingale with respect to the filtration $\{\mathcal{F}_n\}_{n\geq 1}$.*

Proof. We know that Z_j^n is $\mathcal{F}_n$-measurable [4, p.68], and $\mathbb{E}(|Z_j^n|) \leq 1$. Due to the fact that Z_j^{n+1} conditioned on $\mathcal{F}_n$ only depends on the values of W_j^n, and W_{tot}^n, we obtain:

$$\begin{aligned}\mathbb{E}(Z_j^{n+1}|\mathcal{F}_n) &= \mathbb{E}(Z_j^{n+1}|W_j^n, W_{tot}^n) \\ &= \frac{W_j^n}{W_{tot}^n}\left(\frac{W_j^n\beta + 1}{W_{tot}^n\beta + 1}\right) + \frac{W_{tot}^n - W_j^n}{W_{tot}^n}\left(\frac{W_j^n\beta}{W_{tot}^n\beta + 1}\right) \\ &= \frac{W_j^n}{W_{tot}^n} = Z_j^n,\end{aligned}$$

implying that Z_j^n is a martingale with respect to the filtration $\{\mathcal{F}_n\}_{n\geq 1}$ [4, p.458]. □

We will also require Lemma 1, from Chapter 2, Section 2.3. It is proven in the same manner as before, and proves that the total weight of all exemplars W_{tot}^n, is uniformly bounded above by a value γ. We are able to prove Theorem 8 by using Lemmas 1 and 10, and the Martingale Convergence Theorem [3, p.490]. For the purpose of this thesis, we will first state the Martingale Convergence Theorem as written in [3].

Theorem 7. *(The Martingale Convergence Theorem) Let $X_1, X_2, \ldots$ be a submartingale. If $K = sup_n\, \mathbb{E}(|X_n|) < \infty$, then $X_n \to X$ with probability 1, where X is a random variable satisfying $\mathbb{E}(|X|) \leq K$.*

Recall by their definitions, that if $X_1, X_2, \ldots$ is a martingale, then $X_1, X_2, \ldots$ is a submartingale [26].

Theorem 8. *If $N = \infty$, then for all $j \in \{1, \ldots, k\}$, Z_j^n converges a.s. to a random variable Z_j, which can only equal 0 or 1, a.s.*

Proof. To prove this theorem, we will require an expression for $\mathrm{Var}(Z_j^{n+1}|\mathcal{F}_n)$, where $\mathcal{F}_n = \sigma(w_j^m,\ m \le n,\ j \in \{1,\dots,k\})$, as in Lemma 10. First we determine $\mathbb{E}((Z_j^{n+1})^2|\mathcal{F}_n)$,

$$\begin{aligned}\mathbb{E}((Z_j^{n+1})^2|\mathcal{F}_n) &= \frac{W_j^n}{W_{tot}^n}\left(\frac{W_j^n\beta+1}{W_{tot}^n\beta+1}\right)^2 + \frac{W_{tot}^n - W_j^n}{W_{tot}^n}\left(\frac{W_j^n\beta}{W_{tot}^n\beta+1}\right)^2 \\ &= \frac{(W_j^n)^2 W_{tot}^n\beta^2 + 2(W_j^n)^2\beta + W_j^n}{W_{tot}^n\left(W_{tot}^n\beta+1\right)^2}.\end{aligned}$$

This allows us to calculate the conditional variance,

$$\begin{aligned}\mathrm{Var}(Z_j^{n+1}|\mathcal{F}_n) &:= \mathbb{E}((Z_j^{n+1})^2|\mathcal{F}_n) - \mathbb{E}(Z_j^{n+1}|\mathcal{F}_n)^2 \\ &= \frac{(W_j^n)^2 W_{tot}^n\beta^2 + 2(W_j^n)^2\beta + W_j^n}{W_{tot}^n\left(W_{tot}^n\beta+1\right)^2} - \left(\frac{W_j^n}{W_{tot}^n}\right)^2 \\ &= W_j^n(W_{tot}^n - W_j^n)(W_{tot}^n)^{-2}\left(W_{tot}^n\beta+1\right)^{-2} \\ &= Z_j^n(1-Z_j^n)\left(W_{tot}^n\beta+1\right)^{-2}. \qquad (3.7)\end{aligned}$$

By the Martingale Convergence Theorem, because Z_j^n is a submartingale and $\sup_n \mathbb{E}\,|Z_j^n| \le 1$, we know there is a random variable Z_j, such that $Z_j^n \to Z_j$ a.s. This implies $Z_j^{n+1} - Z_j^n \to 0$ a.s., and we know that $|Z_j^{n+1} - Z_j^n| \le 2$ for all n. By the Dominated Convergence Theorem [26], this implies that $\mathbb{E}(|Z_j^{n+1} - Z_j^n|^2) \to 0$, as $n \to \infty$.

By Lemma 1, $W_{tot}^n \le \gamma$, for all n. Because Z_j^n is $\mathcal{F}_n$-measurable, and $\mathbb{E}(Z_j^{n+1}|\mathcal{F}_n) = Z_j^n$,

$$\begin{aligned}\mathbb{E}(|Z_j^{n+1} - Z_j^n|^2) &= \mathbb{E}((Z_j^{n+1})^2 - 2Z_j^{n+1}Z_j^n + (Z_j^n)^2) \\ &= \mathbb{E}\left[\mathbb{E}((Z_j^{n+1})^2 - 2Z_j^{n+1}Z_j^n + (Z_j^n)^2|\mathcal{F}_n)\right] \\ &= \mathbb{E}\left[\mathrm{Var}(Z_j^{n+1}|\mathcal{F}_n)\right]. \qquad (3.8)\end{aligned}$$

Using Equations 3.7 and 3.8, as well as Lemma 1, we get the following

$$\begin{aligned}\mathbb{E}(|Z_j^{n+1} - Z_j^n|^2) &= \mathbb{E}\left[Z_j^n(1-Z_j^n)\left(W_{tot}^n\beta+1\right)^{-2}\right] \\ &\ge (\gamma\beta+1)^{-2}\,\mathbb{E}\left[Z_j^n(1-Z_j^n)\right].\end{aligned}$$

Taking the limit as $n \to \infty$ on both sides, we obtain $\mathbb{E}\left[Z_j^n(1-Z_j^n)\right] \to 0$, as $n \to \infty$. Because convergence in L_1 implies convergence in probability [25, p.85], we know $\mathbf{P}(Z_j^n(1-Z_j^n) < \varepsilon) \to 1$, for all $\varepsilon > 0$. This implies there exists a subsequence such that $Z_j^{n_i}(1-Z_j^{n_i}) \to 0$ a.s. [2, p.7]. As such Z_j can only equal 0 or 1, since we know there must exist a Z_j such that $Z_j^n \to Z_j$, a.s. □

Which brings us to our final result.

Theorem 9. *When $N = \infty$, in the model described in Section 3.2, all categories but one will become extinct with a probability of* 1.

Proof. We know by Lemma 1, and Equation 3.6, that $Z_j^n = W_j^n / W_{tot}^n \geq W_j^n \gamma^{-1} \geq 0$, implying if $Z_j^n \to 0$, then $W_j^n \to 0$ as well. By Theorem 8, for every category $j \in \{1 \dots k\}$, $Z_j^n \to Z_j$, a.s., where Z_j can only be 0 or 1, and we know $\sum_j Z_j = 1$. As such $Z_j^n \to 0$, a.s, for every category j, but one. This implies all but one category will become extinct with a probability of 1. □

3.5 Simulations and Time to Extinction

In the last two sections, we proved extinction of all but one category occurs for our model regardless of the value of N. In this section we will discuss some of the results obtained by computer simulations of the simplified weight model. These simulations will demonstrate how changing the variables N and λ affects how long it takes until there is only one non-extinct category left in the system.

Before discussing the results of our computer simulations, we will explain weight thresholds. Analytically a category j becomes extinct when $W_j^n \to 0$, as $n \to \infty$. Extinction of all but one category means there exists a category j such that $W_i^n \to 0$, as $n \to \infty$, for all $i \neq j$. When running computer simulations, we cannot possibly know for certain if a category's weight approaches zero, but we do something else to detect if it is most likely going to. In simulations, once a category's weight goes below a value we call a weight threshold, we assume that the category becomes extinct. The time it takes for all but one of the category's weights to go below the weight threshold will be referred to as the extinction time. For Figures 3.2 and 3.3, the number of categories $k = 2$, so the extinction time is how soon one of the two categories goes extinct.

Figure 3.2 plots three simulations each for three separate values of N, where the number of categories $k = 2$. We show the evolution of the random variable Z_1^n (defined by Equation 3.6 in Section 3.4) for the values $N = 1$, 10 and ∞. When Z_1^n hits either 0 or 1 the simulation ends, representing that either category 1 or 2 has become extinct respectively. Upon inspection we see that the larger N is, the faster categories become extinct.

Figure 3.3 plots how the expected extinction time changes based on the values of our decay rate λ, and the limitation on the number of exemplars N, when the number of categories $k = 2$. The expected value for the extinction time is found by averaging over 1000 simulations for each value of N and λ. As N decreases, we observe as we did for Figure 3.2 that the extinction time increases. Likewise as λ decreases, the extinction time increases as well.

It is straightforward to explain how λ affects the extinction time, but the explanation for the effect N has is more subtle. To help understand the effect N has on the extinction time, we will consider two examples. For both examples, let $\beta = 0.5$, $k = 2$ (two categories),

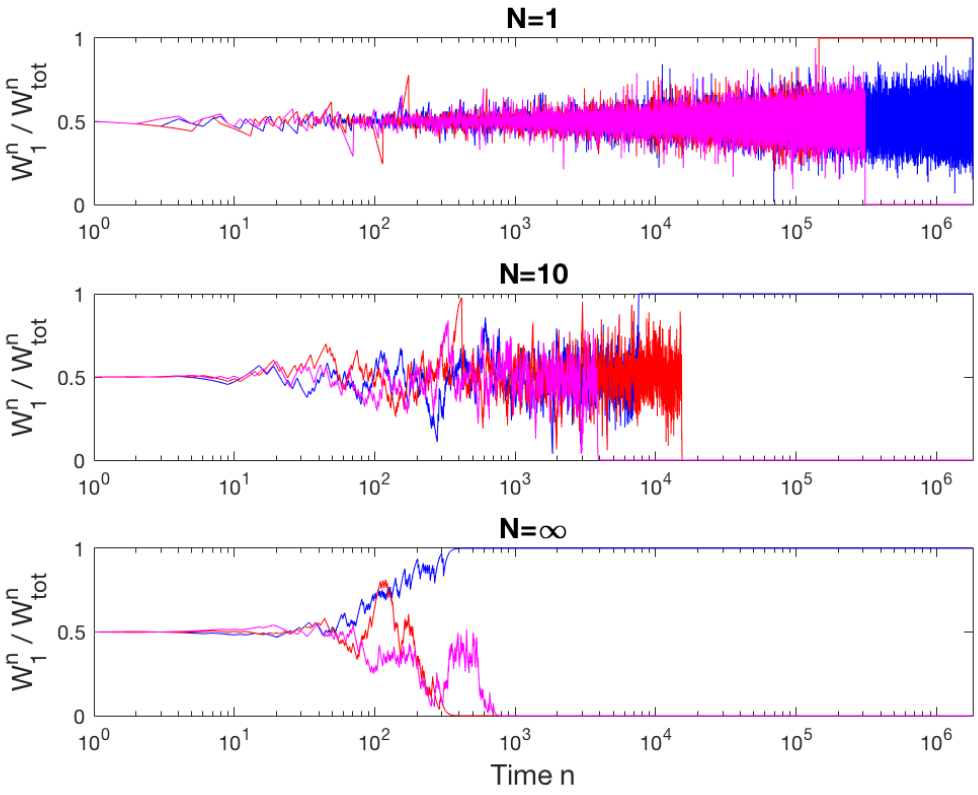

Figure 3.2: Plots of $Z_1^n = W_1^n/W_{tot}^n$ for single simulations when $k = 2$, $\lambda = 0.06$, and the weight threshold is $10^{-4}W_0$. For each value of N we have plotted Z_1^n against time step n for three simulations.

and the initial weight of the first two exemplars in each list is $W_0 = 1$, while the rest of the exemplar weights are zero.

1. First consider the case where $N = 2$. If $x_0 = 2$, the weights of category 2 will be $w_{2,1}^2 = 1$, and $w_{2,2}^2 = 0.5$, and the weights of category 1 will be $w_{1,1}^2 = w_{1,2}^2 = \beta = 0.5$. This implies the probability that $x_1 = 2$, given that $x_0 = 2$, is 60%.

2. Now consider the case where $N = \infty$. If $x_0 = 2$, then the total weight of categories 1 and 2 respectively will be $W_1^n = 2\beta = 1$, and $W_2^n = 2\beta + 1 = 2$. This implies the probability that $x_1 = 2$, given that $x_0 = 2$, is approximately 66.7%.

It is more probable when $N = \infty$, for a category to be consecutively categorized. When $N = 2$, it is rarer for the exemplar weights to decay close to zero than when $N = \infty$. This demonstrates why limiting the number of exemplars makes extinction take longer. When $N = \infty$, exemplars getting stored in a category consecutively adds comparatively more

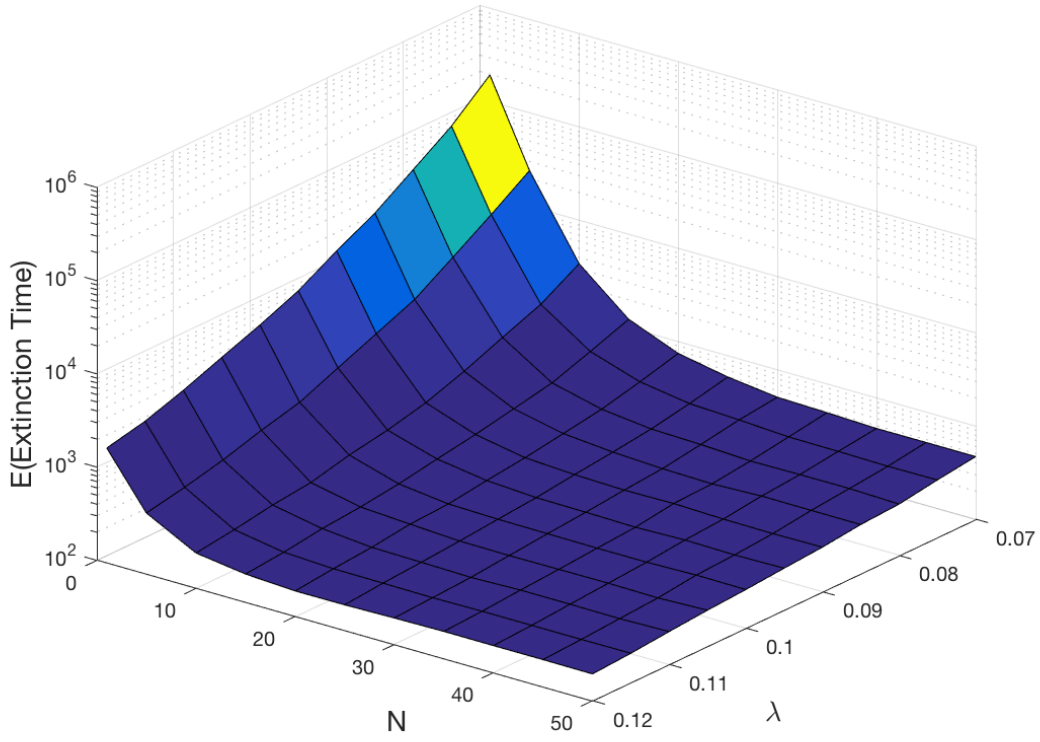

Figure 3.3: Plotting the expected extinction time as we change variables N and λ. We use a weight threshold equal to $10^{-4}W_0$.

weight to the category. This explains the effect of N on the extinction time, as seen in Figures 3.2 and 3.3.

The behaviour of the expected extinction time increasing as λ decreases is much easier to explain. The weights are decaying slower, so it will take longer for the weights to approach zero. If $\lambda = 0$, then there would be no decay and thus no category extinction. Because of this, as $\lambda \to 0$, the expected extinction time will asymptotically approach infinity.

3.6 Discussion

The model studied in this chapter is simpler than the ones studied by Tupper [31] and Wedel [32], but it helps explain the behaviour we see in these models. Changing N in our model doesn't affect whether all categories but one eventually become extinct, but it does affect the time it takes to do so. Our results agree with the extinction result demonstrated in [31] for $N = \infty$. However, our work suggests that the model studied in [32] will eventually show

the same behaviour but on a longer time scale. This longer time scale likely explains why category extinction was not observed in Wedel's simulation [32].

One natural direction we can take in future research is to apply our model to real world data. For example, in Section 3.1, Figure 3.1 shows the evolution of the usage of two spellings of the word cider over 200 years [19]. The archaic spelling "cyder" becomes extinct close to the year 1980. Using the corpus of digitized texts put together in [19], one could determine what values of N and λ best models this type of data.

Chapter 4

Creating a Stable Model Which Implements Segmental Bias

4.1 Introduction

In this chapter we propose a model which segments categories in the phonetic space and we believe is stable. The idea of using segmental bias in exemplar models was introduced in [32]. Segmental bias models how speech is broken up into distinct units which are re-used [23]. The easiest way to demonstrate this behaviour in language is to consider phonemes. For example, the words "bat", "bet", "bit", "bot", and "but", all re-use the phonemes /**b**/ and /**t**/ which are pronounced the same way. The result of this phenomenon is that categories end up segmented and aligned in the phonetic space. Based on the results of [31] and Chapter 4 we have reason to believe the model in [32] has category extinction. Because segmentation is an observable and important part of linguistics, we wish to create a model which implements it and is stable.

In [31], Tupper takes the limit in his exemplar model as the rate of production of exemplars approach infinity, and the initial weight of exemplars approach 0, which we will refer to as the field limit. In this field limit, he has shown his model is stable when misclassified exemplars are discarded from memory. His results suggest that stability depends on whether misclassified sounds are discarded; stable when discarding, not stable when not discarding. The model in [32] is similar to the one in [31] (see Section 3.1 for comparison). Because some misclassified sounds are not discarded in the model described in [32], we believe it is not stable, even in the field limit.

Section 4.2 will describe Wedel's model from [32], which was the basis for the research of Chapter 3. In Section 4.3 we create a new model by adding segmental bias into Tupper's exemplar model in [31]. In Section 4.4, we will discuss how the results of Chapter 3 are related to the models in [32] and [31], and also why we believe categories do not become extinct for the model described in Section 4.3 in the field limit.

4.2 Mathematical Description of Wedel's Exemplar Model

Here we will give the full mathematical description for the model proposed in [32]. In Section 4.3, we will implement segmental bias in a similar manner to the way it is applied here. The purpose of explaining this model is to give a background on segmental bias, and so we can better discuss in Section 4.4 why we do not believe this model is stable in the field limit.

Imagine there are two individuals speaking to one another. Each has their own store of exemplars and can speak one of k categories at a time, while the other categorizes the spoken sound. The individuals take turns speaking.

As done in earlier exemplar models in this thesis, we consider the phonetic space of the system to be a subset of $\mathbb{R}^N$, where N is the number of phonetic variables the speaker/listener is using to classify sounds. Each exemplar has an associated weight. Let w_y be the weight of associated exemplar y. The initial weight of exemplars are set equal to $W_0 = 1$. At each time a sound is spoken, we will decay all exemplar weights by a factor $e^{-\lambda \Delta t}$, where λ is the decay rate, and Δt is the difference in time since the last sound was spoken. The weights of person 2's exemplars are decayed when person 1 is speaking, and vice versa. In [32], $\lambda = .07$, and the value for Δt was not given.

When one person speaks, a category $i \in \{1, \ldots, k\}$, is chosen to be spoken uniformly at random. An exemplar x is chosen from the category to be spoken with probability proportional to its weight. The expressed sound will be a biased version of the chosen exemplar.

We first bias the sound at the word level. The word level bias has the effect of shifting the chosen exemplar towards the center of the spoken category. We create a sound p_w which is the weighted average of exemplars close to it, where the weights are a combination of the exemplar weights and an exponential distance function. This sound p_w is calculated by the expression

$$p_w = \frac{\sum_{y \in C_i} y w_y e^{-r|x-y|}}{\sum_{y \in C_i} w_y e^{-r|x-y|}},$$

where C_i is the set of exemplars within the category i for the speaker, and r is a scaling factor which we set equal to 0.2 as in [32]. The sound p_w will be shifted towards the mean in relation to the chosen exemplar x.

The sound is also biased segmentally, which represents a bias towards keeping the categories aligned. The components of the segmental bias vector are calculated using the following expression

$$p_s^d = \frac{\sum_y y_d w_y e^{-r|x_d - y_d|}}{\sum_y w_y e^{-r|x_d - y_d|}}, \tag{4.1}$$

where $p_s = [p_s^1, p_s^2, \ldots, p_s^N]$, and $x = [x_1, x_2, \ldots, x_N]$ [32].

A new vector p, is the mix of the word and segmental population vectors and is expressed as $p = 0.9p_w + 0.1p_s$, as in [32]. Finally, p is biased towards the center of the phonetic space, and a Gaussian random variable with a standard deviation of 3 is added. The components of the bias towards the center of the phonetic space $b = [b_1, \ldots, b_N]$, are calculated using the expression

$$b_j = -\text{sign}(p_j - m_j)\frac{(p_j - m_j)^2}{G},$$

where m_j is the center of the jth dimension of the dimensional space, $p = [p_1, \ldots, p_N]$, and $G = 5000$ [32]. This will henceforth be referred to as the preferred value bias. This models how speakers prefer to produce exemplars with phonetic values which are not too large or small [31].

The expressed sound is the vector

$$s = p + b + \sigma\eta,$$

where η is a standard Gaussian random variable, and $\sigma = 3$ is its standard deviation.

The listener then attempts to categorize the sound s using their exemplars. The sound is classified using a variation on the Generalized Context Model (henceforth GCM) for categorization [22], which has been modified to account for exemplar weights [32]. The probability of classifying expressed sound s as category i is,

$$\mathbf{P}(s \in \text{categ. } i) = \frac{\sum_{y \in C_i} w_y e^{-r|s-y|}}{\sum_y w_y e^{-r|s-y|}}, \tag{4.2}$$

where C_i represents the set of exemplars in category i for the listener. If category j is the category sound s is classified as, and j is the category with the highest probability given by Equation 4.2 we store s as an exemplar, otherwise it is discarded from memory. This is referred to as anti-ambiguity, where if the sound is not classified as the category with the highest probability of being categorized, it is discarded.

It should also be noted in that in Wedel's simulations, the number of exemplars per category was limited to a maximum of 100. If a classification would give a category its 101st exemplar, the exemplar with the lowest weight is discarded.

4.3 Adding Segmental Bias to Tupper's Model

Tupper discussed an exemplar model in [31], which we will describe in this section. The model was investigated in the limit as the rate of production of exemplars approach infinity and the initial weight of exemplars approach 0, which we refer to as the field limit. In the

field limit, when implementing rejection (a process similar to anti-ambiguity in the section above), the model approaches a stable configuration in which all the categories remain.

The field limit represents taking the limit as the number of exemplars contained in the system approaches infinity. The field model, which is the exemplar model when taking the field limit, is used to infer the behaviour of the exemplar model when there is a large number of exemplars. It also helps with computation time in simulations, which grows with the number of exemplars, making it difficult to determine the long term behaviour if not taking the limit [31]. We believe if one does not take the limit of number of exemplars approaching infinity, that there will eventually be category extinction. We believe the field limit is a better representation of the exemplar model when there is a large number of exemplars.

We want to find a model that is stable in the same way as in [31], and implements the segmental bias used in [32], which can be found in Section 4.2 above. Based on the results found in Chapter 3, along with [31], we believe Wedel's idea of anti-ambiguity is not enough to ensure stability, even in the field limit. The reasons for this will be discussed in Section 4.4. As such, in this section we will alter the exemplar model in [31] to include a segmental bias, and propose that it is stable in the field limit.

As before, the system's phonetic space is a subset of $\mathbb{R}^N$, where N is the number of phonetic variables being used to classify sounds. There is one store of exemplars in this model from which sounds are both spoken and classified. We believe this gives the same results as considering two stores of exemplars as in [32], but this still needs to be researched.

Let k be the number of sound categories. Let w_y be the weight of associated exemplar y. The weights of initial exemplars are set equal to $W_0 = 1$. At each time step, the weights of all exemplars are decayed by a factor $e^{-\lambda \Delta t}$, where λ is the decay rate, and Δt is the time difference since the last sound was spoken.

At each time step we choose a category $i \in \{1, \ldots, k\}$, to be spoken randomly with uniform probability. An exemplar x is chosen from category i to be spoken with probability proportional to its weight.

Let C_i be the set of all exemplars in category i. Let $\bar{y}$ be the weighted mean position of all the exemplars in category i,

$$\bar{y} = \frac{\sum_{y \in C_i} w_y y}{\sum_{y \in C_i} w_y}, \tag{4.3}$$

and y^* be a preferred phonetic value within the space. The exemplar will be biased towards $\bar{y}$ (word bias) and y^* (preferred value bias). Following [32], we let y^* be the center of the phonetic space. Note that the word and preferred value biases are implemented differently than in Wedel's model (see Section 4.2).

So far, we have been describing Tupper's exemplar model from [31]. Now we are going to implement segmental bias into his model. Let the segmental bias be $y_s = [p_s^1, p_s^2, \ldots, p_s^N]$,

where p_s^d is as defined in Equation 4.1. We will bias the system towards y_s representing a segmental bias, helping keep the categories aligned.

The expressed sound s, is

$$s = x + \alpha(\bar{y} - x) + \beta(y^* - x) + \gamma(y_s - x) + \sigma\eta,$$

where η is a standard Gaussian random variable, $\sigma = 3$ is its standard deviation, and $\alpha > 0$, $\beta > 0$, and $\gamma \geq 0$ are the strengths of the bias towards the center of the word (entrenchment as in [24]), the preferred value bias, and segmental bias respectively. If $\gamma = 0$, the model above is the same as the exemplar model described in [31]. All we have done to Tupper's model is add segmental bias.

Similarly to Section 4.2, we use the GCM for categorization [22], which has been modified to account for exemplar weights [32]. Once again, the probability of categorizing sound s in category i is given by Equation 4.2,

$$\mathbf{P}(s \in \text{categ. } i) = \frac{\sum_{y \in C_i} w_y e^{-r|s-y|}}{\sum_y w_y e^{-r|s-y|}},$$

where C_i represents the set of exemplars in category i for the listener. If the expressed sound is misclassified (so it is not classified as the intended category), the expressed exemplar is not stored in memory and is discarded. We refer to this as rejection, which is different than anti-ambiguity in the model described in Section 4.2. We will discuss in Section 4.4 why rejection is better than anti-ambiguity at keeping categories from collapsing.

When running simulations we implement a weight threshold. When the weight of a single exemplar goes below the threshold we discard it from memory. A category goes extinct when there are no exemplars left in that category. This is different than the weight threshold in Chapter 3, Section 3.5, where we were concerned about when the total weight of a category went below the weight threshold.

For all our simulations we let y^* be the origin and the phonetic space be $\mathbb{R}^2$. We let $r = 10$, $\lambda = 1$, $\sigma = 1$, $\nu = 1000$ to stay consistent with the parameters used in [31]. We let $\alpha = 0.9$, and will vary γ and β in our simulations.

Exemplars are introduced to the system with a constant rate ν using a Poisson process. Because of this, the time between spoken sounds is exponentially distributed with mean $1/\nu$. If ν is the rate of production of exemplars, then the flow in of weight is on average νW_0. Because the decay rate is λ, we know the flow out of weight on average is $\lambda W_{tot}(t)$, where $W_{tot}(t)$ is the total weight at time t. At equilibrium, $\nu W_0 = \lambda W$, where W is the equilibrium value of the total weight. In the field limit, we let ν approach infinity. To ensure the equilibrium value W remains fixed in this limit let $W_0 = 1/\nu$. As such, the field model is achieved when you take the limit as ν approaches infinity and as the initial weight of all

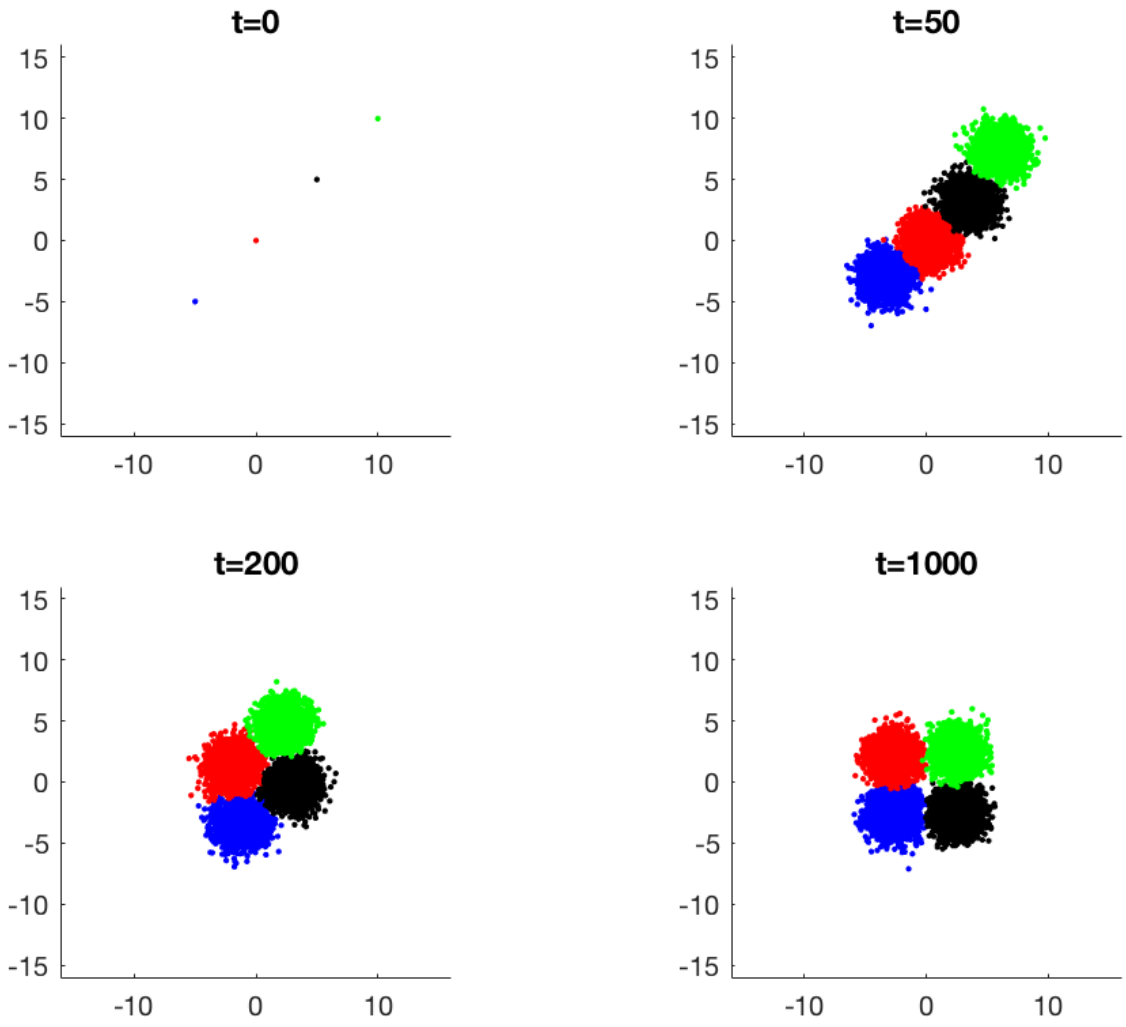

Figure 4.1: Time slices of a single simulation of proposed model with values $k = 4$, $\lambda = 1$, $\nu = 1000$, $r = 10$, $\sigma = 1$, $\alpha = 0.9$, $\beta = 0.01$, $\gamma = 0.9$, and the weight threshold is $10^{-4}W_0$.

exemplars W_0 approaches 0 [31]. For simulations, the system should appear stable for large values of ν.

Figure 4.1 shows time slices of a single simulation of the model when segmental bias is implemented. One can see by time $t = 1000$, that the 4 categories have segmented themselves within the phonetic space. For comparison, Figure 4.2 shows time slices of a single simulation with the same values used in Figure 4.1, but without segmental bias. One can see by the time $t = 1000$ that the categories have not aligned with one another. Comparing Figures 4.1 and 4.2 we can see that the segmental bias when turned on has the effect of segmenting the categories, and when off it does not.

The category clouds need to stay apart for segmental bias to be able to segment the categories. An example of a situation where the clouds do not stay apart, and instead pack

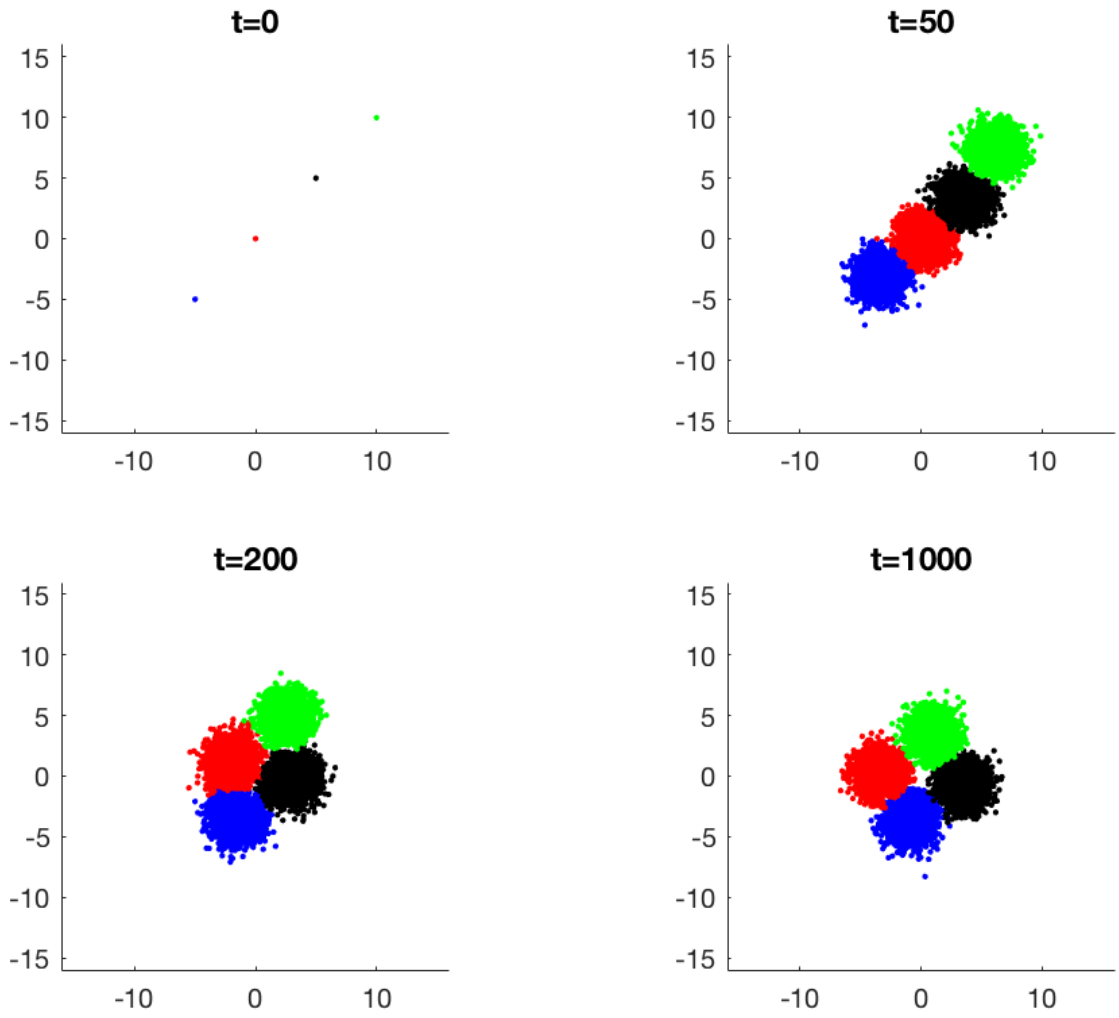

Figure 4.2: Time slices of a single simulation of proposed model with the same parameter values as Figure 4.1, except $\gamma = 0$.

close together is shown in Figure 4.3. Despite that $\gamma = 0.9$ as in Figure 4.1, the categories will never stabilize in an alignment in Figure 4.3, because there is not enough space between the clouds of exemplars to tell the difference between them.

The way to keep the category clouds apart is to make the bias towards the center, β, small. This can be seen in Figure 4.4. The bottom two graphs show the final time slice of the simulations. The top two graphs plot two things. They plot the $e^{-r|x-\bar{y}|}$ for dimension 1 for each category mean, and underneath these plots you can see the exemplars in the same dimension plotted separately. We plot $e^{-r|x-\bar{y}|}$ to give an idea of the size and position of the density function within the segmental bias calculation in Equation 4.1. We have observed that the categories align for the situation on the right, and not for the one on the left. The key to understanding why this happens, is that we sum over all exemplars in the numerator

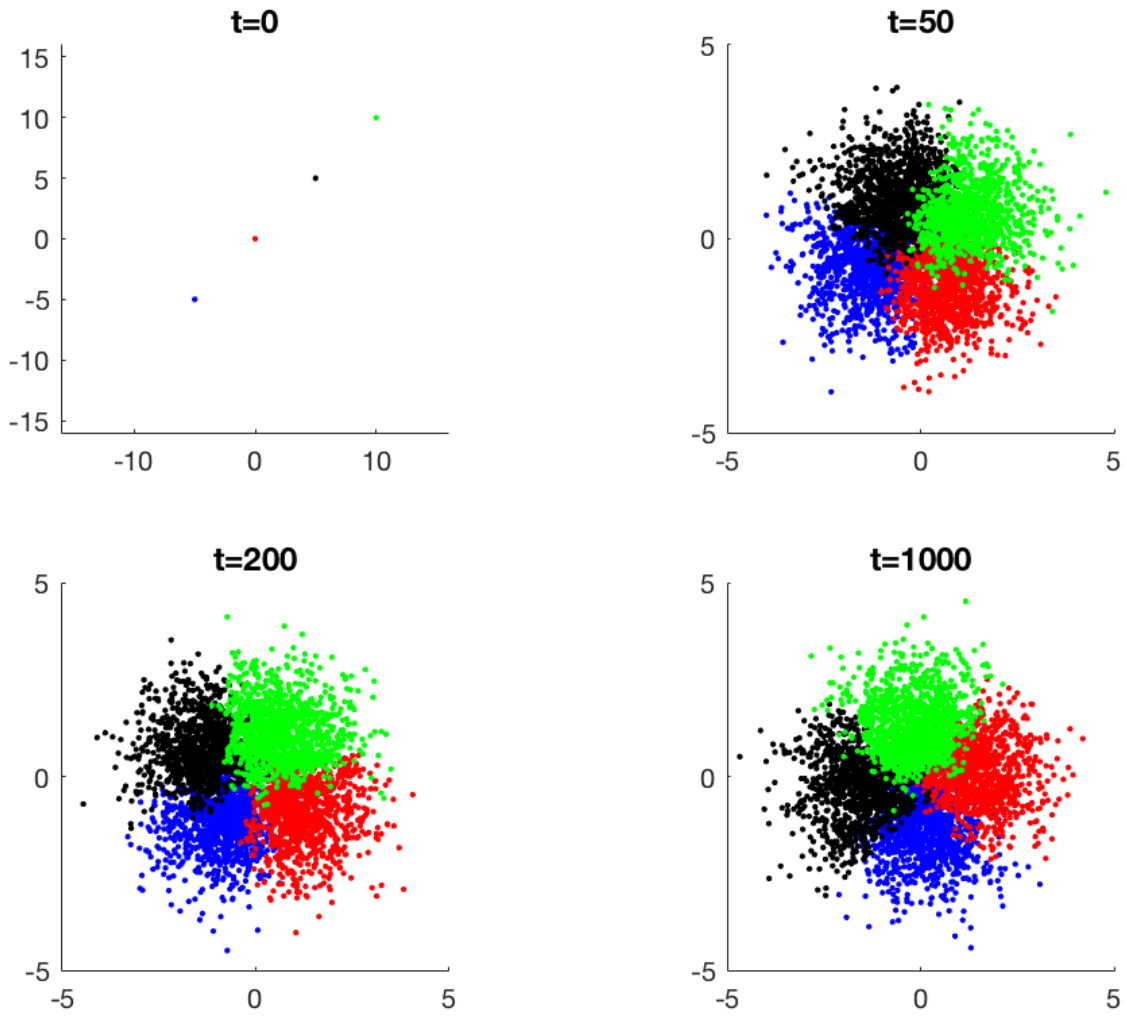

Figure 4.3: Time slices of a single simulation of proposed model with the same parameter values as Figure 4.1, except $\beta = 0.5$.

of Equation 4.1. As such, if the exemplar clouds are too close, then Equation 4.1 cannot tell the difference between category clouds, and cannot segment them. Segmental bias tends to take a long time to segment the categories and requires the category clouds to stay in a separated formation for a long period of time.

4.4 Conclusions

Here we will discuss the results of all the chapters of this thesis and how the results relate to one another. We explain why it is reasonable to believe based on the results of Chapter 3 and [31], that the model in [32] has category extinction. In contrast, we will discuss later in this section why a model like the one studied in Chapter 2 does not have category

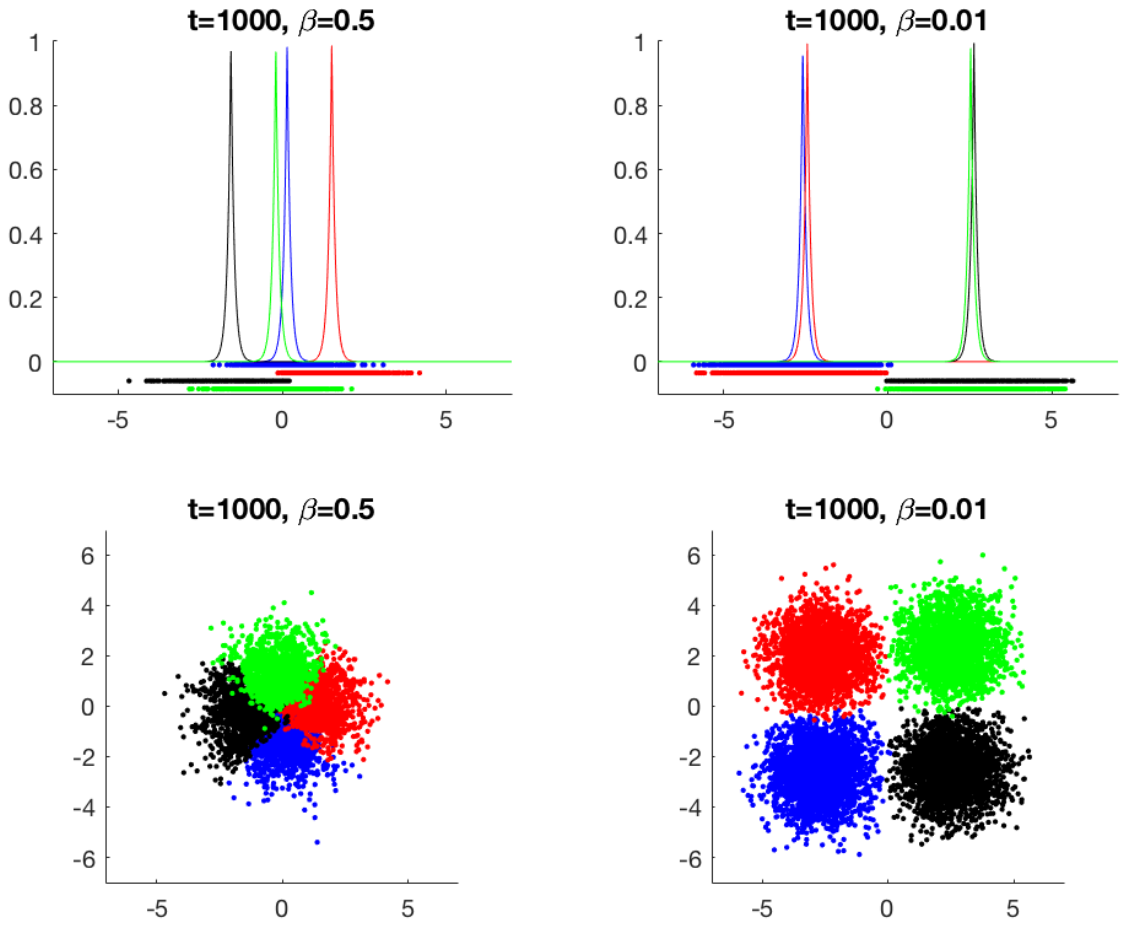

Figure 4.4: Here we plot the distribution of the segmental bias about each category mean, i.e. $e^{-r|x-\bar{y}|}$ in dimension 1, where $\bar{y}$ is given by Equation 4.3, as well as the exemplars. The left two graphs in the figure represent the same values we used in Figure 4.3 ($\beta = 0.5$), while the right two correspond to Figure 4.1 ($\beta = 0.01$).

extinction. In general, we will be able to make the conclusion that category extinction is heavily dependent on the classification method used.

We begin by discussing why the results found in Chapter 3 correspond with the models in Sections 4.2 and 4.3. First we recall sounds are classified in these papers using Equation 4.2 which is:

$$\mathbf{P}(s \in \text{categ. } i) = \frac{\sum_{y \in C_i} w_y e^{-r|s-y|}}{\sum_y w_y e^{-r|s-y|}},$$

where C_i is the set of exemplars in category i (for the listener in [32]). Recall the probability of categorizing a sound as category j in Chapter 3, Section 3.2, is

$$\mathbf{P}(x_n = j|\mathcal{F}_n) = W_j^n / W_{tot}^n,$$

where W_j^n is the total weight of all the exemplars in category j at time n, and W_{tot}^n is the total weight of all exemplars in the system at time n. Note that when $r = 0$, these equations are identical. This similarity, along with reasons yet to be given below, is why we believe the model in [32] has category extinction.

It is important to know how/why collapse is caused in the models described in Sections 4.2 and 4.3. First, remember that extinction occurs when the weights of all the exemplars in a category approach 0. This is caused when no more sounds ever get classified within a category after some point in time. Every time a category is spoken is a chance for an exemplar to be stored within it. The best chance to add an exemplar to a category's store of exemplars is when it is spoken as that category. If it is misclassified, this does two things:

1. It contributes to the weights of that category approaching 0, decreasing the numerator in Equation 4.2.

2. If the sound is stored despite being improperly classified, then the weights of another category increase which increases the denominator in Equation 4.2.

Note, these are the same reasons collapse is caused in the model defined in Chapter 3. This is because it also has the weight of the category in the numerator, and the total weight of all exemplars in the denominator.

Rejection completely gets rid of number 2 (in the list above), and anti-ambiguity gets rid of number 2 partially. There are examples where the intended category spoken will not have the highest probability of being categorized given by Equation 4.2, and thus the category will have no chance of being classified. This is why rejection works better at keeping the system stable than anti-ambiguity, and also contributes to our argument that the model in [32] has category extinction. Note that this agrees with results in [31], which were discussed in Section 4.1. One cannot get rid of number 1 (in the list above) without introducing something else into the model. This contributes to the argument that the model in [32] is not stable.

We believe the model developed in Chapter 4 will be stable in the field limit because adding segmental bias to [31] should not affect whether categories become extinct. The extinction of categories is more related to the classification method used.

Let us go back to the results found in Chapter 2, for which we proved there is never category extinction. This system is stable because the classification method is based primarily on the distance between points in the space, and not on the relative size of a category's weight to the total weight of all exemplars. The only effect that weights have on the categorization of sounds is on the weighted category means, which generate the classification sets via a Voronoi diagram. But the weighted category means only depend on the weights within each category, rather than the weights of all the exemplars. For a category to become extinct in the model described in Chapter 2, the area of classification for a category must approach 0. If the weights of a category get close to 0 in this model, it does not have a

direct affect on that category's chances of categorization. But in the models described in Chapters 3 and 4 it does, because the weights of a category being closer to 0 than to other categories decreases it's chances of getting categorized.

Model	Classification Method	Discard Rule	Category Extinction
Chapter 2	Voronoi Diagrams	None	No
Chapter 3	GCM [22] ($r = 0$)	None	Yes
Wedel [32] (Exemplar Model)	GCM [22]	Anti-Ambiguity	Yes?
Wedel [32] (Field Model)	GCM [22]	Anti-Ambiguity	Yes?
Tupper [31] (Exemplar Model)	GCM [22]	Rejection	Yes?
Tupper [31] (Field Model)	GCM [22]	Rejection	No
Chapter 4 (Field Model)	GCM [22]	Rejection	No?

Table 4.1: A summary of all the exemplar models considered in this thesis. The classification method used, and discard rule for each model is given, and whether we know (or just believe, marked with a question mark) that there is category extinction within the system.

The results of all the models we have discussed within this thesis are summarized in Table 4.1. We can conclude based on our results that the classification method implemented in an exemplar model has a heavy influence on whether the system has category extinction.

Chapter 5

Future Research

Here we will discuss possible future directions for research in studying exemplar models for language change. A significant portion of this will involve using data to see how well our models predict trends seen in language change. We want to know how altering the number of stores of exemplars present in exemplar models changes long term behaviour such as category extinction. There are also ways to expand upon the work done in Chapter 3.

In [9], Jaeger compared evolutionary stable states of simulations to actual languages. We could do a similar analysis with the k-means model studied in Chapter 2, as well as the model proposed in Section 4.3. This would involve directly comparing long term evolutions with vowel arrangements in languages.

We want to use the data found in Google Ngram [19] to create an accurate model for the evolution of different spellings of words in the written lexicon. This can be done using the model developed in Chapter 3. We could determine what values of N (the number of exemplars per category) and λ (decay rate) in the simple weight model best replicates the evolution of alternate spellings of words in the written lexicon.

The differences between exemplar models with a single store of exemplars and 2 stores of exemplars should be determined. The properties of exemplar models are easier to study when there is only one store of exemplars. This would most likely be done via comparisons of simulations.

An exemplar model could be developed/studied which involved two individuals with a different number of categories k. A real life example of this is a conversation between an American English speaker and a British English speaker; American English speakers have 14 or 15 distinct vowels that can contrast monosyllabic words, whereas British English speakers have 20 distinct vowels [12].

The results found in Section 3.5 could be extended by proving that as you decrease N, the extinction time will increase. The proof of this would probably use something similar to the observation made in Section 3.5; classifying a sound as a category increases its probability of classification. A proof for this might involve utilizing stopping times [26,

pg.162]. We could define the stopping time as the time n that $|Z_1^n - 1/2| \geq 1/2 - W_{\text{th}}$, where Z_1^n is as defined in Equation 3.6, and $W_{\text{th}} < 1/2$ is the weight threshold. The main difficulty behind proving this would be that you have to study when the probability of classification goes above or below a specified value.

We want to apply the field limit in the simple weight model considered in Chapter 3. As done in Chapter 3, we would want to show how changing N alters whether categories become extinct in the model. In [31], Tupper already did work on a model where the field limit was taken when $N = \infty$. The proof will likely not be the same if N is finite though. As such, the difficult part of working on this problem will be proving whether categories become extinct when N is finite.

One might be able to determine whether the segmental bias model developed in Chapter 4 is stable in the field limit. The biggest hurdle to overcome here is how segmental bias is implemented within the model, since it is done differently than the other biases used in [31].

Bibliography

[1] Rabi Bhattacharya and Mukul Majumdar. *Random Dynamical Systems: Theory and Applications*. Cambridge University Press, Cambridge, UK, 2007.

[2] Rabi Bhattacharya and Edward C Waymire. *A Basic Course in Probability Theory*. Springer Science & Business Media, New York, NY, USA, 2007.

[3] P. Billingsley. *Probability and Measure*. Wiley series in Probability and Mathematical Statistics. Wiley, San Francisco, CA, USA, 1986.

[4] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, San Francisco, CA, USA, 1995.

[5] Tomas Björk. *Arbitrage Theory in Continuous Time*. Oxford University Press, New York, NY, USA, 2009.

[6] Joan Bybee. Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition*, 24(02):215–221, 2002.

[7] ByoungSeon Choi. *ARMA Model Identification*. Springer Science & Business Media, New York, NY, USA, 2012.

[8] Qiang Du, Vance Faber, and Max Gunzburger. Centroidal voronoi tessellations: Applications and algorithms. *SIAM Review*, 41(4):637–676, December 1999.

[9] Gerhard Jäger. Applications of game theory in linguistics. *Language and Linguistics Compass*, 2(3):406–421, 2008.

[10] Gerhard Jäger, Lars P Metzger, and Frank Riedel. Voronoi languages: Equilibria in cheap-talk games with high-dimensional types and few signals. *Games and Economic Behavior*, 73(2):517–537, 2011.

[11] Keith Johnson. Speech perception without speaker normalization: An exemplar model. *Talker Variability in Speech Processing*, pages 145–165, 1997.

[12] P. Ladefoged and S.F. Disner. *Vowels and Consonants*. Wiley, San Francisco, CA, USA, 2012.

[13] Ilse Lehiste and David Meltzer. Vowel and speaker identification in natural and synthetic speech. *Language and Speech*, 16(4):356–364, 1973.

[14] R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Courier Corporation, Burnaby, BC, Canada, 2005.

[15] J. MacQueen. Some methods for classification and analysis of multivariate observations, 1967.

[16] Henrik Madsen. *Time Series Analysis*. CRC Press, Boca Raton, FL, USA, 2007.

[17] John Maindonald and W. John Braun. *Data Analysis and Graphics Using R*. Cambridge University Press, Cambridge, UK, third edition, 2010. Cambridge Books Online.

[18] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, New York, NY, USA, 2012.

[19] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[20] Robert M. Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39, 1986.

[21] Robert M. Nosofsky. Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(4):700, 1988.

[22] Robert M. Nosofsky. Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1):54, 1988.

[23] Pierre-Yves Oudeyer. The self-organization of speech sounds. *Journal of Theoretical Biology*, 233(3):435–449, 2005.

[24] Janet B. Pierrehumbert. Exemplar dynamics: Word frequency, lenition, and contrast. *Frequency and the Emergence of Linguistic Structure*, 45:137–157, 2001.

[25] Joseph P. Romano and Andrew F. Siegel. *Counterexamples in Probability and Statistics*. CRC Press, Boca Raton, FL, USA, 1986.

[26] Jeffrey S Rosenthal. *A First Look at Rigorous Probability Theory*. second edition.

[27] Walter Rudin. *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc., New York, NY, USA, 1987.

[28] D. Ruppert. *Statistics and Data Analysis for Financial Engineering*. Springer Texts in Statistics. Springer, New York, NY, USA, 2010.

[29] John H. Ryalls and Philip Lieberman. Fundamental frequency and vowel perception. *The Journal of the Acoustical Society of America*, 72(5):1631–1634, 1982.

[30] Mike Steel. Reflections on the extinction–explosion dichotomy. *Theoretical Population Biology*, 101:61–66, 2015.

[31] P. F. Tupper. Exemplar dynamics and sound merger in language. *SIAM Journal on Applied Mathematics*, 75(4):1469–1492, 2015.

[32] Andrew Wedel. Lexical contrast maintenance and the organization of sublexical contrast systems. *Language and Cognition*, 4:319–355, 2012.

[33] Andrew Wedel, Abby Kaplan, and Scott Jackson. High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128(2):179–186, 2013.

[34] Andrew B. Wedel. Exemplar models, evolution and language change. *The Linguistic Review*, pages 247–274, 2006.

[35] Bodo Winter and Andrew Wedel. The co-evolution of speech and the lexicon: The interaction of functional pressures, redundancy, and category variation. *Topics in Cognitive Science*, 8(2):503–513, 2016.

[36] Robert A. Yaffee and Monnie McGee. *Introduction to Time Series Analysis and Forecasting: With Applications of SAS and SPSS*. Academic Press, Inc., Orlando, FL, USA, first edition, 2000.