A STOCHASTIC MODEL OF LANGUAGE CHANGE THROUGH SOCIAL STRUCTURE AND PREDICTION-DRIVEN INSTABILITY

W. GARRETT MITCHENER*

Abstract. Children reliably learn their community's language; consequently human languages are relatively stable on short time scales. However, languages can change dramatically over the course of centuries, and once begun, such changes generally run monotonically to completion. We consider a stochastic model that reproduces this pattern of fluctuations via large deviations. We begin with a Markov chain that represents an age-structured population in which children learn which of two grammars their community prefers, but are aware of age-correlated usage patterns and will use the dispreferred grammar more often if they infer that its use is spreading. The Markov chain is shown to converge in the limit of an infinite population to a stochastic differential equation that generalizes the Wright-Fisher SDE for population genetics. This proof is not routine because the dynamics are only defined in a Cartesian product of simplexes, and it must be verified that trajectories of the SDE cannot escape. Results are proved by changing variables in a way that expands each simplex to an entire plane, yielding reasonable constraints on the dynamics that ensure that a standard but sophisticated theorem for well-posedness of SDEs can be applied. The SDE yields a phase portrait that reveals the mechanism that causes these models to produce sporadic, monotone, population-wide transitions between grammars. A further simplification results in a stochastic functional-delay differential equation that shows how population-level memory effects and the attempt by learners to avoid sounding outdated results in prediction-driven instability.

Key words. language change, prediction-driven instability, population dynamics, stochastic differential equation, noise-activated transitions

AMS subject classifications. 37H10, 60H10, 60J20, 91F20

1. The paradox of language change.

A primary tool in the field of linguistics is the *idealized grammar*, that is, a formalism that distinguishes correctly formed utterances from ill-formed utterances [8, 17]. Historically, much of the research on how children acquire their native language has focused on how they might choose one idealized grammar from many innate possibilities on the basis of example sentences from the surrounding society [1, 57, 60]. From the perspective of idealized grammar, language change is paradoxical: Children acquire their native language accurately and communicate with adults from preceding generations, yet over time, the language can change drastically. Some changes may be attributed to an external event, such as political upheaval, but not every instance of language change seems to have an external cause. Despite their variability, languages do maintain considerable short-term stability, consistently accepting and rejecting large classes of sentences for centuries. The primary challenge addressed by the model discussed in this article is to capture this meta-stability.

Many existing models of language learning in a population focus on characterizing stable properties of languages. For example, naming games and other lexical models focus on the process by which a population forms a permanent consensus on a vocabulary, and how effective that vocabulary is at representing meanings [9, 23, 49, 50, 51, 56, 61]. Related models focus on the structure of lexeme or phoneme inventories once a stable equilibrium is reached [22, 21, 66]. Several algorithms have been proposed as models for the acquisition of idealized grammars [5, 6, 7, 14, 21, 22, 44, 43, 60]. These focus on details of the acquisition process and follow the *probably almost correct* (PAC) learning framework [15], in which the learner's input is a list of grammatically correct utterances called the *primary lin-*

*Department of Mathematics, College of Charleston, SC (MitchenerG@cofc.edu)

guistic data (PLD), and the learner is required to choose a single idealized grammar from a limited set that is somehow maximally consistent with that input. Learners are given no data on social structure, or any negative evidence, that is, information stating that a possible utterance is ungrammatical. The input may be from a single individual [21, 22] or a population, perhaps consisting of adults that collectively use several idealized grammars [43]. Other proposed algorithms address the sensitivity of the PAC framework to noise in the PLD by including means of ignoring rarely occurring constructions [28, 29, 54, 65].

There is room to improve on these models. In contrast to actual human language, these models typically have stable equilibrium states from which the learner or population cannot escape. Furthermore, PAC learning algorithms typically make use of the *subset principle:* Out of all the available idealized grammars, the "correct" choice is the one that generates the smallest set of utterances including the input. The subset principle is frequently included in language learning models because children typically ignore assertions by adults that a particular utterance is ungrammatical. However, there is evidence that the subset principle does not accurately reproduce certain features of child language acquisition, and that children make use of statistical patterns in adult speech to determine that utterances they previously accepted are actually ungrammatical [39, 4].

Many language models for populations are adapted from deterministic, continuous, biological population models and represent language by communication games. These focus on stable behavior in an infinite homogeneous population, although some exhibit ongoing fluctuations [40, 33, 41, 34, 35, 37, 48, 46, 47, 45, 53]. Some are designed to represent a single change [24]. In these models, children learn from an average of speech patterns, and except for [41], these do not model the origins of language changes directly. Instead, an external event must disturb the system and push it from one stable state to another.

As we will see in section 2, a general mean-field model in which children learn from the entire population equally does not lead to spontaneous change, even in the presence of random variation. It appears that spontaneous changes can only arise from random fluctuations in combination with some sort of momentum driven by social structure.

Based on extensive field studies, Labov [26] proposes a model in which phonetic change is driven by females who naturally change their individual speech over time, a force called *incrementation*. A semi-structured approach as in [36] assumes a fully interconnected finite population but agents vary in their influence on learners. These models approximate the time course of a single change, in qualitative agreement with data, but neither addresses the origin of the change.

Some models use network dynamics rather than a mean-field assumption and allow learners to collect input disproportionately from nearby members of the population [12, 59]. These models incorporate observations made by Labov and others that certain individuals tend to lead the population in adopting a new language variant, and the change spreads along the friendship network [25, 26, 27].

In contrast, the model analyzed in this article is built from an alternative perspective in an attempt to resolve the language change paradox. Utterances may be drawn from multiple idealized grammars and classified as more or less archaic or innovative. Such an approach can consider the variation present in natural speech and model it as a *stochastic grammar,* that is, a collection of similar idealized grammars, each of which is used randomly at a particular rate [24, 25, 26, 62]. From this continuous perspective, language change is no longer a paradox, but acquisition requires more than selecting a single idealized grammar as in the PAC framework. Instead, children must learn multiple idealized grammars, plus the usage rates and whatever conditions affect them.

Crucially, instead of limiting learners' input to example sentences, we will assume that children also know something about the ages of speakers and prefer not to sound outdated. They bias their speech against archaic forms by incorporating a prediction step into their acquisition of a stochastic grammar, which introduces incrementation without directly imposing it as in [26]. The age structure and bias against archaic forms introduce momentum into the dynamics, which generates the desired meta-stability. The population tends to hover near a state where one idealized grammar is highly preferred. However, children occasionally detect accidental correlations between age and speech, predict that the population is undergoing a language change, and accelerate the change. This feature will be called *prediction-driven instability.*

The majority of the language modeling literature does not focus on the formal aspects of mathematical models, such as confirming that the dynamics are well-posed or deriving a continuous model as the limit of a discrete model, even though such details are known to be generally important [11]. Numerical simulations of the discrete form of the age-structured stochastic model developed in this article confirm that it has the desired behavior [38] but its continuous form has yet to be placed on a sound theoretical foundation. So, in section 2 we formulate a discrete mean-field model as a Markov chain and discuss its weaknesses. Then in section 3, we extend it to include age-structure, then rigorously consider the limit of an infinitely large population and reformulate the Markov chain as a continuous-time martingale problem.

We rewrite this martingale problem as a system of stochastic differential equations (SDEs), show that it has a unique solution for all initial values, and show that paths of the Markov chain converge weakly to solutions of the SDEs. The proofs make use of theorems in [10] for the existence and uniqueness of solutions to SDEs and convergence of discrete Markov chains to such solutions. However, the SDEs of interest take values in a phase space consisting of Cartesian products of simplexes, and changes-of-variables are required to derive SDEs taking values in a plane as required by the standard theorems. Furthermore, the drift and volatility terms in the resulting SDEs grow too quickly in magnitude at infinity for the most commonly used theorems to be directly applied. Instead, asymptotic estimates must be used to verify that the drift terms push solutions back toward the origin, in which case a more general result presented in [10] guarantees the existence of unique solutions for all time. These results confirm that solutions to the SDEs are at no risk of straying into unrealistic territory where the usage rate of some grammar has escaped from $[0, 1]$. Furthermore, they make minimal assumptions about the vector field and are applicable to other dynamical systems on simplexes.

In the two dimensional case, in which agents use one grammar or the other exclusively, it is possible to see in the phase portrait that proximity of stable equilibria to the boundaries of their basins of attraction is what facilitates spontaneous language change. A final modification to the two-dimensional SDEs allows them to be reformulated as a one-dimensional functional-delay SDE. In this form, it becomes clear that the population switches from one meta-stable state to another when children detect a chance fluctuation in the usage rate of the dominant grammar away from the running average, and amplify it.

2. First stage: An unstructured mean-field model.

Let us suppose initially that individuals have a choice between two similar idealized grammars $\mathcal{G}_1$ and $\mathcal{G}_2$.

Each simulated agent uses $\mathcal{G}_2$ in forming an individual-specific fraction of spoken sentences, and $\mathcal{G}_1$ in forming the rest. Assume that children are always able to acquire both idealized grammars and the only challenge is learning the usage rates. Assume that the population consists of N adult agents, each of which is one of $K+1$ types, numbered 0 to K, where type k means that the individual uses $\mathcal{G}_2$ at a rate k/K and $\mathcal{G}_1$ at a rate $1-k/K$. The state of the chain at time step j is a vector T where $T_n(j)$ is the type of the n-th agent. Define the count vector C where $C_k(j)$ is the number of agents of type k,

$$C_k(j) = \sum_n \mathbf{1}(T_n(j) = k)\,.$$

Dividing the count vector by the population size yields the speech distribution vector $X = C/N$ such that an agent selected at random from the population uniformly at time j is of type k with probability $X_k(j)$.

The mean usage rate of $\mathcal{G}_2$ at step j is therefore

$$M(j) = \sum_{k=0}^{K} \left(\frac{k}{K}\right) X_k(j) \tag{1}$$

Children are assumed to learn the usage rates of the two grammars based only on $M(j)$, the mean usage rate of $\mathcal{G}_2$ in the adult population at time j. Children are assumed to be exposed to enough sample utterances from across the entire population to accurately estimate $M(j)$. The model requires a *mean learning function* $q(m)$ that gives the mean usage rate of children learning from a population with a mean rate m.

The transition process from step j to $j+1$ is as follows. Two additional parameters are required, a birth-and-death rate r_D and a resampling rate r_R. At each time step, each individual agent is examined and one of these three operations is randomly applied to it:

- With probability $p_D = r_D/N$ it dies and is replaced.
- With probability r_R it is resampled.
- With probability $1 - p_D - r_R$ it is unchanged.

Details are given in the following subsections.

2.1. Time, learning, and the birth-death operation. Each time step is interpreted as $1/N$ years. The lifespan of an individual in time steps has a geometric distribution with parameter p_D. The average life span is therefore $1/p_D$ time steps or $1/r_D$ years.

When an agent dies, a replacement agent is created and its type is selected at random based on a discrete distribution vector $Q(M(j))$. That is, $Q_k(m)$ is the probability that a child learning from a population with mean usage rate m is of type k, and therefore uses $\mathcal{G}_2$ at rate k/K. As a specific example, $Q(m)$ could be the mass function for a binomial distribution with parameters $q(m)$ and K,

$$Q_k(m) = \binom{K}{k} q(m)^k (1-q(m))^{K-k}. \tag{2}$$

Since the mean of such a distribution is $q(m)K$, it follows that q and Q satisfy the identity

$$q(m) = \sum_{k=0}^{K} \left(\frac{k}{K}\right) Q_k(m) \tag{3}$$

which confirms that $q(m)$ is indeed the mean usage rate of $\mathcal{G}_2$ by children learning from adults with mean usage rate m.

The mean learning function must be S-shaped to ensure that there are two equilibrium states, representing populations dominated by one grammar or the other. In general, q is assumed to be smooth, strictly increasing, with one inflection point, and

$$\begin{aligned} 0 < q(0) < 1/2 \\ 1/2 < q(1) < 1 \end{aligned} \tag{4}$$

In practice, $q(0)$ will be close to 0 and $q(1)$ will be close to 1. A curved mean learning function means that the more commonly used idealized grammar becomes even more commonly used, until the other grammar all but disappears. This tendency is in agreement with the observation that children regularize language: A growing body of evidence [19] indicates that for the task of learning a language with multiple ways to say something, adults tend to use all the options and match the usage rates in the given data, but children prefer to pick one option and stick with it. Beyond these general properties, this learning model makes no attempt to directly represent the neurological details of language acquisition, although researchers are exploring this area [2, 20, 39, 55, 65].

2.2. Resampling of adults. When an agent is resampled, its new state is copied from another agent picked uniformly at random. The average time an agent spends between resamplings is $1/r_R$ time steps. This feature of the transition process incorporates the fact that as an adult, an individual's language is not entirely fixed [26, 25]. Furthermore, as will be explained in section 4, without this resampling feature, the random fluctuations of this Markov chain diminish to 0 in the limit as $N \to \infty$, which would defeat the purpose of developing a stochastic model. This consideration leads to the peculiar fact that in formulating the Markov chain, p_D must scale as $1/N$ but the probability r_R of an agent being resampled must remain constant. The Wright-Fisher model [10] includes a similar feature: In the discrete formulation, each time step is considered a single generation and each agent is always resampled, akin to setting $r_R = 1$, but when passing to the limit $N \to \infty$, the generation time is taken to scale as $1/N$ without scaling the resampling process.

It is possible that in contrast to standard practice in the population genetics literature, r_D should also scale as $1/N$. That would cause fluctuations in grammar use to shrink as the population size grows, in agreement with anecdotal reports that languages spoken by only a small number of native speakers change rapidly compared to those with larger populations, but in disagreement with other studies [63, 64]. Resolution of this issue is beyond the scope of this article.

2.3. Behavior of the model. This Markov chain is regular. Although it spends most of its time hovering near a state dominated by one idealized grammar or the other, it must eventually exhibit spontaneous language change by switching to the other. However, computer experiments confirm that under this model, a population takes an enormous amount of time to switch dominant grammars. This model is therefore unsuitable for understanding language change on historical time scales. A further undesirable property is that when a population does manage to shift to an intermediate state, it is just as likely to return to the original grammar as to complete the shift to the other grammar. Historical studies [24, 65] show that language changes typically run to completion monotonically and do not reverse themselves partway through (but see [62] for some evidence to the contrary), so again this model is unsatisfactory.

3. Second stage: An age-structured model. One way to remedy the weaknesses of these mean-field models is to introduce social structure into the population. According to sociolinguistics, ongoing language change is reflected in variation, so there is reason to believe children are aware of socially correlated speech variation and use it during acquisition [25].

There are many ways to formulate a socially structured population, and not all formulations apply to all societies. For this article, let us assume that there are two age groups, roughly representing youth and their parents, and that children can detect systematic differences in their speech. We also assume that there are social forces leading children to avoid sounding out-dated.

Let us adapt the Markov chain from section 2 to include age structure. To represent the population at time j, fix the total number of youth and the total number of parents at N, so there are $2N$ agents total. To make the notation systematic, superscript labels Y and A will be used, referring to the youth and adult generations, respectively. Let $T^Y_n(j)$ be the type of the n-th youth and $T^A_n(j)$ be the type of the n-th adult, all between 0 and K. Define $C^Y_k(j)$ to be the number of youth of type k, and define $C^A_k(j)$ to be the number of adults of type k. Let

$$X^Y = \frac{1}{N}C^Y \text{ and } X^A = \frac{1}{N}C^A \tag{5}$$

be the probability distribution vectors of the two generations. Assume that apart from age, children make no distinction among individuals. Thus, they learn essentially from the mean usage rates of the two generations,

$$\begin{aligned} M^Y(j) &= \sum_{k=0}^{K} \left(\frac{k}{K}\right) X^Y_k(j) \\ M^A(j) &= \sum_{k=0}^{K} \left(\frac{k}{K}\right) X^A_k(j) \end{aligned} \tag{6}$$

The modified transition process from time j to $j+1$ is as follows. Each adult is examined:

- With probability $p_D = r_D/N$ it is replaced to simulate death and aging.
- With probability r_R it is resampled from the adult population.
- With probability $1 - p_D - r_R$ it is unchanged.

Each youth is examined:

- With probability $p_D = r_D/N$ it is replaced to simulate birth and aging.
- With probability r_R it is resampled from the youth population.
- With probability $1 - p_D - r_R$ it is unchanged.

Each time step is interpreted as $1/N$ years. The number of time steps spent by an individual in each age group has a geometric distribution with parameter p_D. The average time spent as an adult and as a youth is therefore $1/p_D$ time steps or $1/r_D$ years, so the average life span is now $2/r_D$.

When an agent is resampled, its new state is copied from another agent from the same generation selected uniformly at random. As before, resampling leaves the mean behavior unchanged while introducing volatility.

It is certainly possible to incorporate birth, aging, and death into the model by deleting an adult, directly moving someone from the youth generation to the adult generation, and creating a new youth. However, the calculations are simplified if birth and death are handled separately, resulting in mathematically trivial differences to the Markov chain.

When an adult dies, rather than moving a youth, a replacement is created by sampling from an aging distribution $V\left(X^Y\right)$, that is very close to X^Y but gives at least a minimal probability to every type. This feature allows for innovation in adults, and avoids a technical problem that would cause the model to fall outside the hypotheses of Lemma 4.6. The examples in this article use

$$V_k(X) = X_k(1 - (K+1)\eta) + \eta \tag{7}$$

with $\eta = 1/1000$.

For birth and aging, a randomly selected youth is deleted, and a replacement youth is created based on the discrete probability vector $R(M^Y(j), M^A(j))$. Here, $R(x, y)$ represents the acquisition process, together with prediction: Children hear that the younger generation uses $\mathcal{G}_2$ at a rate x, and the older generation uses a rate y. Based on x and y and any trend those numbers indicate, they predict a rate that their generation should use, and learn based on that predicted target value. Let the predicted mean usage rate be given by a smooth function $r(x, y)$ that is increasing with respect to x, decreasing with respect to y, and satisfies

$$\forall x, y: \quad y < x \implies x < r(x, y)$$

and

$$\forall x, y: \quad y > x \implies x > r(x, y).$$

That is, any trend from the past y compared to the present x should continue to the future $r(x, y)$. Then, our assumptions on learning based on prediction can be incorporated into the mathematics by setting $R(x, y) = Q(r(x, y))$.

For a specific example, let us consider a population of 1000 agents, 500 in each age group, with a birth-death rate of $r_D = 1/20$. Therefore, the mean lifespan of an agent is 40 years. The resampling rate is $r_R = 0.0001$. There are 6 types of agents, representing speech patterns that use $\mathcal{G}_2$ for a fraction $0, 1/5, \ldots, 1$ of spoken sentences.

The learning distribution $Q(m)$ is a binomial distribution with parameters $q(m)$ and 5. The example q in this article is

$$q(m) = \frac{1}{32} + \frac{3600}{751}\left(\frac{33m}{1280} + \frac{161m^2}{320} - \frac{m^3}{3}\right) \tag{8}$$

This polynomial was constructed to be slightly asymmetric and strictly increasing on $[0, 1]$. Its range is $[1/32, 31/32]$, so it satisfies (4) and conditions that will be needed to apply Proposition 4.1.

The example prediction function $r(x, y)$ is based on an exponential sigmoid. Given $s(t) = 1/(1 + \exp(-t))$, define $t_1 = s^{-1}(x)$ and $t_2 = s^{-1}(y)$. Then $h = t_1 - t_2$ is a measure of the trend between the generations. A scale factor α is applied to h, and the scaled trend is added to t_1. After some simplification,

$$r_0(x, y) = s(t_1 + \alpha h) = \frac{1}{1 + \left(\frac{1-x}{x}\right)^{\alpha+1}\left(\frac{y}{1-y}\right)^{\alpha}} \tag{9}$$

For the example calculations in this paper, $\alpha = 3$. Observe that r_0 is a rational function, defined and continuous everywhere in $[0, 1] \times [0, 1]$ except at the corners.

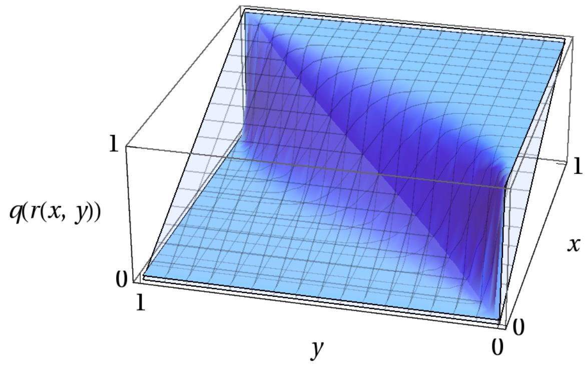

FIG. 1. *The learning-prediction function $q(r(x,y))$ and the plane given by the graph of $(x,y) \mapsto x$.*

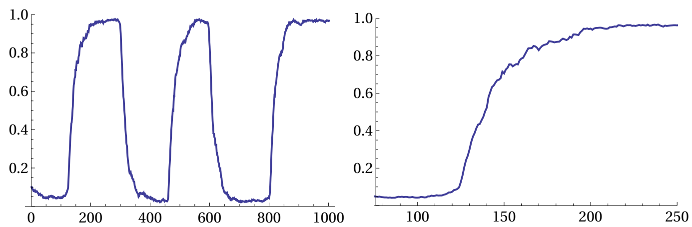

FIG. 2. *Trajectory of the mean usage rate $M^Y(t)$ of $\mathcal{G}_2$ in the young generation from a sample path of the age-structured Markov chain. Left: The path from time 0 to 1000 years, showing several changes between $\mathcal{G}_1$ (low) and $\mathcal{G}_2$ (high). Right: The path from time 75 to 250 years, showing a single grammar change.*

This definition may be smoothly extended to include $r_0(1,0) = 1$ and $r_0(0,1) = 0$, but no extension is possible to $(0,0)$ and $(1,1)$. To remedy this, we will assume that agents have slightly imperfect perception and introduce

$$w(x) = \frac{1}{2} + (1-\delta)\left(x - \frac{1}{2}\right) \tag{10}$$

which maps $[0,1]$ to $[\delta/2, 1-\delta/2]$. Pictures in this article will use $\delta = 1/1000$. Thus a suitable prediction function that is defined and smooth on all of $[0,1] \times [0,1]$ is

$$r(x,y) = r_0\big(w(x), w(y)\big). \tag{11}$$

The combined mean learning-prediction function $q(r(x,y))$ is plotted in Figure 1. An important feature is that since $q(0) > 0$ and $q(1) < 1$, the graph is slightly above the plane given by $(x,y) \mapsto x$ along the edge where $x = 0$, and is slightly below that plane along the edge where $x = 1$. This means that given an initial condition where one of the idealized grammars is not used at all, there is a non-zero probability that it will appear spontaneously.

This model turns out to exhibit the desired properties. The population can spontaneously change from one language to the other and back within a reasonable amount of time, and once initiated the change runs to completion without turning back. See Figure 2 for a graph of the mean usage rate of $\mathcal{G}_2$ among the younger age group as a function of time for a typical run of this Markov chain.

4. Diffusion limit. To better understand why spontaneous change happens in this model, we develop a continuous limit for the Markov chain governing the speech distributions X^Y and X^A of the younger and older generations, respectively, which are points in the open simplex,

$$\mathcal{S}^K = \left\{ (x_0, \ldots, x_K) \,\middle|\, x_k \in (0,1), \sum_{k=0}^{K} x_k = 1 \right\}.$$

In the limit as the population size N increases without bound, the Markov chain $(X^Y(j), X^A(j)) : \mathbf{N} \to \mathcal{S}^K \times \mathcal{S}^K$ ought to converge to the solution $(\xi^Y(t), \xi^A(t)) : [0, \infty) \to \mathcal{S}^K \times \mathcal{S}^K$ of a martingale problem. To formulate it, we must calculate the infinitesimal drift and covariance functions.

4.1. Notation. To reduce notational clutter in this subsection, all time-dependent quantities at time step j will be written without a time index, as in T_n^Y, C_k^Y, and X_k^Y. The learning distribution $Q(r(M^Y, M^A))$ will be written as just Q, and the aging distribution $V(X^Y)$ will be written as just V. Time-dependent quantities at time step $j+1$ will be written with a bar, as in $\bar{T}_n^Y$, $\bar{C}_k^Y$, and $\bar{X}_k^Y$. Expectations and variances with a j subscript are conditioned on the information available at time step j.

4.2. Infinitesimal mean and variance. Conditioning on time step j, $\mathbf{1}(\bar{T}_n^Y = k)$ is a Bernoulli random variable that takes on the value 1 with probability

$$g(n,k) = (1 - p_D - r_R)\, \mathbf{1}(T_n^Y = k) + p_D Q_k + r_R X_k^Y$$

that is, either $T_n^Y = k$ and it didn't change, or it died and was replaced by a child of type k, or it was resampled and became type k. With this observation, the mean and variance of $\bar{C}_k^Y$ conditioned on information known at time step j can be calculated as follows.

$$\mathbf{E}_j\left(\bar{C}_k^Y\right) = \sum_n g(n,k) = (1 - p_D) C_k^Y + p_D N Q_k \tag{12}$$

$$\begin{aligned} \mathrm{Var}_j\left(\bar{C}_k^Y\right) &= \sum_n g(n,k) - g(n,k)^2 \\ &= (1 - p_D - r_R)\, C_k^Y + p_D N Q_k + r_R N X_k^Y \\ &\quad - (1 - p_D - r_R)^2\, C_k^Y \\ &\quad - 2\,(1 - p_D - r_R)\, C_k^Y \left(p_D Q_k + r_R X_k^Y\right) \\ &\quad - N\left(p_D Q_k + r_R X_k^Y\right)^2 \end{aligned} \tag{13}$$

If $m \neq n$ then $\bar{T}_m^Y$ and $\bar{T}_n^Y$ are conditionally independent given the information available at time $j+1$. If $h \neq k$ then $\mathbf{1}(\bar{T}_n^Y = k)\,\mathbf{1}(\bar{T}_n^Y = h) = 0$. Therefore,

$$\begin{aligned} \operatorname{Cov} j\left(\bar{C}_k^Y, \bar{C}_h^Y\right) &= \sum_n \operatorname{Cov} j\left(\mathbf{1}(\bar{T}_n^Y = k), \mathbf{1}(\bar{T}_n^Y = h)\right) \\ &= -\sum_n g(n,k) g(n,h) \\ &= -\Bigg((1 - p_D - r_R)\, C_k^Y \left(p_D Q_h + r_R X_h^Y\right) \\ &\quad + (1 - p_D - r_R)\, C_h^Y \left(p_D Q_k + r_R X_k^Y\right) \\ &\quad + N\left(p_D Q_k + r_R X_k^Y\right)\left(p_D Q_h + r_R X_h^Y\right)\Bigg) \end{aligned} \tag{14}$$

It follows that

$$\mathbf{E}_j\left(\frac{\bar{X}_k^Y - X_k^Y}{1/N}\right) = r_D(Q_k - X_k^Y), \tag{15}$$

which gives the infinitesimal drift component for a martingale problem. We also need an estimate of the covariance matrix for X^Y:

$$\operatorname{Var}_j\left(\bar{X}_k^Y\right) = \frac{1}{N}(2r_R - r_R^2)\left(X_k^Y - \left(X_k^Y\right)^2\right) + O\left(\frac{1}{N^2}\right) \tag{16}$$

$$\operatorname{Cov}_j\left(\bar{X}_k^Y, \bar{X}_h^Y\right) = -\frac{1}{N}(2r_R - r_R^2) X_k^Y X_h^Y + O\left(\frac{1}{N^2}\right) \tag{17}$$

Similar drift and covariance formulas can be derived for X^A,

$$\mathbf{E}_j\left(\frac{\bar{X}_k^A - X_k^A}{1/N}\right) = r_D(V_k - X_k^A) \tag{18}$$

$$\operatorname{Var}_j\left(\bar{X}_k^A\right) = \frac{1}{N}\left(2r_R - r_R^2\right)\left(X_k^A - \left(X_k^A\right)^2\right) + O\left(\frac{1}{N^2}\right) \tag{19}$$

$$\operatorname{Cov}_j\left(\bar{X}_k^A, \bar{X}_h^A\right) = -\frac{1}{N}(2r_R - r_R^2) X_k^A X_h^A + O\left(\frac{1}{N^2}\right) \tag{20}$$

As a further simplification, we can rescale time by a factor of r_D. This finally yields the infinitesimal drift function

$$b : \mathcal{S}^K \times \mathcal{S}^K \to \mathbf{R}^{2K}$$

$$b\begin{pmatrix}\xi^Y \\ \xi^A\end{pmatrix} = \begin{pmatrix} b^Y \\ b^A \end{pmatrix} = \begin{pmatrix} Q - \xi^Y \\ V - \xi^A \end{pmatrix} \tag{21}$$

and the infinitesimal covariance function $\varepsilon^2 A$

$$A : \mathcal{S}^K \times \mathcal{S}^K \to \mathbf{M}\left(\mathbf{R}, 2K \times 2K\right)$$

(22)

$$A\begin{pmatrix}\xi^Y\\ \xi^A\end{pmatrix} = \begin{pmatrix}A^Y & 0\\ 0 & A^A\end{pmatrix} = \begin{pmatrix} \xi_1^Y-\left(\xi_1^Y\right)^2 & -\xi_1^Y\xi_2^Y & \cdots & & & \\ -\xi_1^Y\xi_2^Y & \ddots & & & 0 & \\ \vdots & & & & & \\ & & & \xi_1^A-\left(\xi_1^A\right)^2 & -\xi_1^A\xi_2^A & \cdots \\ & 0 & & -\xi_1^A\xi_2^A & \ddots & \\ & & & \vdots & & \end{pmatrix}$$

and

$$\varepsilon = \sqrt{\frac{2r_R - r_R^2}{r_D}} = \sqrt{\frac{1-(1-r_R)^2}{r_D}} \tag{23}$$

It can be verified by direct calculation that A is positive definite. The dimensions given here use the convention that ξ_0^Y and ξ_0^A are omitted from the dynamics. They will not be considered independent variables because of the population size constraints

$$\begin{aligned} \xi_0^Y &= 1-\left(\xi_1^Y+\cdots+\xi_K^Y\right)\\ \xi_0^A &= 1-\left(\xi_1^A+\cdots+\xi_K^A\right) \end{aligned} \tag{24}$$

The drift function can be augmented by defining

$$b_0^Y = -\sum_{j=1}^{K} b_j^Y \text{ and } b_0^A = -\sum_{j=1}^{K} b_j^A$$

so that deterministic dynamics under the vector field on $\mathbf{R}^{K+1}\times\mathbf{R}^{K+1}$ defined by the augmented b preserve (24).

If the resampling feature is removed by setting $r_R = 0$, then $\varepsilon = 0$ and the dynamics become deterministic. The resampling feature can also be removed from just the older generation by zeroing out A^A, or from just the younger generation by zeroing out A^Y.

4.3. Convergence to system of SDEs. The discrete time Markov chain defined in section 3 converges to a system of stochastic differential equations (SDEs) in the limit as the population size $N\to\infty$ and the physical time of a transition step goes to 0. The time associated with step j of the Markov chain is $t = j/N$, so to properly express the convergence of the Markov chain to a process in continuous time and space, we need the auxiliary processes $\hat{X}^Y$ and $\hat{X}^A$ that map continuous time to discrete steps,

$$\begin{aligned} \hat{X}^Y(t) &= X^Y(\lfloor Nt\rfloor)\\ \hat{X}^A(t) &= X^A(\lfloor Nt\rfloor) \end{aligned} \tag{25}$$

The limiting initial value problem for $\left(\xi^Y, \xi^A\right) \in \mathcal{S}^K \times \mathcal{S}^K$ is built from the infinitesimal vector field (21) and covariance matrix (22):

$$
\begin{aligned}
\mathrm{d}\xi_k^Y(t) &= b^Y(\xi^Y, \xi^A)\,\mathrm{d}t + \varepsilon\sigma^Y(t)\,\mathrm{d}B^Y(t)\\
\xi_0^Y &= 1 - \sum_{k=1}^{K} \xi_k^Y\\
\mathrm{d}\xi_k^A(t) &= b^A(\xi^Y, \xi^A)\,\mathrm{d}t + \varepsilon\sigma^A(t)\,\mathrm{d}B^Y(t)\\
\xi_0^A &= 1 - \sum_{k=1}^{K} \xi_k^A\\
\xi^Y(0) &= \xi^Y_{\text{init}}\\
\xi^A(0) &= \xi^A_{\text{init}}
\end{aligned}
\tag{26}
$$

Here B^Y and B^A are independent K-dimensional Brownian motions, and σ^Y and σ^A are the unique positive-definite, symmetric square-roots of A^Y and A^A. There is no general closed form for σ^Y and σ^A, but the theory turns out to only require A^Y and A^A.

PROPOSITION 4.1. *Suppose* $\left(\hat{X}^Y(0), \hat{X}^A(0)\right)$ *converges to* $\left(\xi^Y_{\text{init}}, \xi^A_{\text{init}}\right)$ *as* $N \to \infty$*. Suppose* (b^Y, b^A) *satisfies the hypotheses of Proposition 4.8. The then for each* $\varepsilon_0 > \varepsilon > 0$*, the process* $\left(\hat{X}^Y(t), \hat{X}^A(t)\right)$ *converges weakly as* $N \to \infty$ *to the solution to (26).*

Proof. We apply theorem 7.1 from Chapter 8 of [10] as follows. The calculations (15), (16), and (17) in section 4 verify that the step-to-step drift, variances, and covariances of the Markov chain converge to the corresponding functions in the SDE (26) as the time step size $1/N$ goes to zero. The remaining condition to check is Durrett's hypothesis (A), which is that the martingale problem associated to the SDE is well posed. The SDE has pathwise-unique strong solutions, as we will prove in Proposition 4.8. That implies uniqueness in distribution [10, §5.4 theorem 4.1] which implies that the martingale problem is well posed [10, §5.4 theorem 4.5] which implies the desired convergence. □

The commonly referenced theorem for existence and uniqueness of solutions to initial value problems for SDEs (see [52, theorem 5.2.1], for example) is not sufficient for (26). It applies to dynamics on Euclidean space, but the dynamics of interest here are restricted to $\mathcal{S}^K \times \mathcal{S}^K$. We can change variables to expand the simplices to whole spaces, but then the global Lipschitz property and global growth constraints required by that theorem are not met. We must therefore apply more general theorems from [10] instead.

4.4. Change of variables. First, we deal with phase space, as (26) only makes sense for $(\xi^Y, \xi^A) \in \mathcal{S}^K \times \mathcal{S}^K$. We change variables so as to push the boundary of the phase space off to infinity. Since the formulas are exactly parallel for each generation, the generation label superscripts will be omitted where possible. To further conserve space, let $\gamma = 1/(K+1)$. Each vector $\xi \in \mathcal{S}^K$ is mapped to a vector λ,

$$
\lambda_k = \tilde{\xi}\,(\xi_k - \gamma) \tag{27}
$$

where

$$\tilde{\xi} = \left(\prod_{k=0}^{K} \xi_k \right)^{-\gamma} \tag{28}$$

The interior of the simplex expands to the entire plane

$$\left\{ \lambda \in \mathbf{R}^{K+1} \;\middle|\; \sum_{k=0}^{K} \lambda_k = 0 \right\}.$$

Let us also define

$$\xi_{\min} = \min_k \xi_k \quad \xi_{\max} = \max_k \xi_k \quad \lambda_{\min} = \min_k \lambda_k \quad \lambda_{\max} = \max_k \lambda_k \tag{29}$$

Note that the extrema for ξ_k and λ_k occur at the same value of the index k. Since $\sum_{k=0}^{K} \xi_k = 1$, it follows immediately that

$$\xi_{\min} \le \gamma \le \xi_{\max} \quad \lambda_{\min} \le 0 \le \lambda_{\max} \tag{30}$$

Furthermore, since $\lambda_{\min} = \tilde{\xi}(\xi_{\min} - \gamma)$,

$$\tilde{\xi} = \frac{-\lambda_{\min}}{\gamma - \xi_{\min}} > -\lambda_{\min}(K+1) \tag{31}$$

LEMMA 4.2. *The change of variables is smooth and smoothly invertible provided none of the ξ_k's are zero, although the inverse does not have a closed form.*

Proof. To prove the existence of the inverse, note that if there is a solution for the ξ_k's in terms of λ_k's, it must hold that

$$\tilde{\xi}^{-(K+1)} = \prod_{k=0}^{K} \xi_k = \prod_{k=0}^{K} \left(\tilde{\xi}^{-1} \lambda_k + \gamma \right) = \tilde{\xi}^{-(K+1)} \prod_{k=0}^{K} \left(\lambda_k + \gamma \tilde{\xi} \right)$$

Thus $f(\tilde{\xi}) = 1$ where f is the polynomial

$$f(x) = \prod_{k=0}^{K} (\lambda_k + \gamma x) \tag{32}$$

Assuming that the λ_k's are known, note that $f\left(-\lambda_{\min}(K+1)\right) = 0$, and for $x > -\lambda_{\min}(K+1)$, $f(x)$ is product of strictly positive terms, all of which are strictly increasing in x, and it is unbounded as $x \to \infty$. There is therefore a unique solution to $f(x) = 1$ with $x > -\lambda_{\min}(K+1)$. Let $\tilde{\xi}$ be this solution, and recover $\xi_k = \tilde{\xi}^{-1}\lambda_k + \gamma$. This change of variables is smooth and locally Lipschitz, but not globally Lipschitz because each partial derivative (40) is unbounded as $\xi_j \to 0$. ∎

Several additional inequalities relating ξ and λ will be required. First, to avoid confusion about whether the 0th element of a vector is included in a dot product or magnitude, let us define

$$\|v\|^2 = \sum_{k=1}^{K} v_k^2 \qquad \|v\|_0^2 = \sum_{k=0}^{K} v_k^2 = \|v\|^2 + (v_0)^2 \tag{33}$$

$$u \cdot v = \sum_{k=1}^{K} u_k v_k \qquad u \odot v = \sum_{k=0}^{K} u_k v_k \tag{34}$$

For a general vector $v = (v_0, \ldots, v_K)^\mathsf{T}$, with extreme elements $v_{\min}$ and $v_{\max}$, it is elementary to verify that

$$\|v\|_0^2 - v_{\max}^2 \le \|v\|^2 \le \|v\|^2 + v_{\min}^2 \le \|v\|_0^2 \le \|v\|^2 + v_{\max}^2 \le 2\|v\|^2 \tag{35}$$

LEMMA 4.3.

$$\tilde{\xi} \le \frac{1 - \lambda_{\min}}{\gamma} \le \frac{1 + \|\lambda\|_0}{\gamma} \le \frac{1 + \sqrt{2}\,\|\lambda\|}{\gamma} \tag{36}$$

Proof. From the definition of $\tilde{\xi}$, it is clear that

$$1 < \xi_{\max}^{-1} \le \tilde{\xi} \le \xi_{\min}^{-1} \tag{37}$$

Building from (37),

$$\tilde{\xi} \le \xi_{\min}^{-1} = \left(\gamma + \frac{\lambda_{\min}}{\tilde{\xi}}\right)^{-1}$$

It follows that

$$\tilde{\xi}\gamma + \lambda_{\min} = \tilde{\xi}\left(\gamma + \frac{\lambda_{\min}}{\tilde{\xi}}\right) \le 1$$

which, in conjunction with (35), yields the bounds (37). □

LEMMA 4.4. *There is a constant $\rho > 0$ such that for all ξ*

$$(K+1)\tilde{\xi} \le \sum_{k=0}^{K} \xi_k^{-1} \le \rho\tilde{\xi}^{K+1} \le \rho\left(\frac{1 + \sqrt{2}\,\|\lambda\|}{\gamma}\right)^{K+1} \tag{38}$$

Proof. The lower bound on $\sum \xi_k^{-1}$ comes from the standard harmonic-geometric mean inequality. For the upper bound, note that

$$f(\xi) = \left(\sum_{k=0}^{K} \frac{1}{\xi_k}\right)\left(\prod_{k=0}^{K} \xi_k\right)$$

is a polynomial, so it has an absolute maximum ρ on the closure of $\mathcal{S}^K$. □

It is important to note that the power $K+1$ of $\|\lambda\|$ in the upper bound (38) is the best possible. Consider the case of $\xi_0 = \delta$, $\xi_k = (1-\delta)/K$ for $k > 0$ and small $\delta > 0$. Then $\sum \xi_k^{-1} \approx \delta^{-1} + K$, $\tilde{\xi} \approx K^{K\gamma}\delta^{-\gamma}$, and $\lambda_k \approx K^{K\gamma}(\xi_k - \gamma)\delta^{-\gamma}$. In this case, $\sum \xi_k^{-1}$ is on the order of $\|\lambda\|^{1/\gamma}$. This power is why so much care must be taken to establish the well-posedness of (42).

4.5. Itô's formula. The following partial derivative formulas are needed in the application of Itô's formula, and are written here assuming $i \ge 1$, $j \ge 1$, $k \ge 1$. Recall that ξ_0 is not considered a separate independent variables because of (24).

$$\partial_{\xi_j}\tilde{\xi} = \gamma\tilde{\xi}\left(\xi_0^{-1} - \xi_j^{-1}\right) \tag{39}$$

$$\begin{aligned}\partial_{\xi_j}\lambda_k &= \gamma\tilde{\xi}\left(\xi_0^{-1} - \xi_j^{-1}\right)(\xi_k - \gamma) + \mathbf{1}(j=k)\,\tilde{\xi} \\ &= \gamma\lambda_k\left(\xi_0^{-1} - \xi_j^{-1}\right) + \mathbf{1}(j=k)\,\tilde{\xi}\end{aligned} \tag{40}$$

$$\begin{aligned}\partial_{\xi_i\xi_j}\lambda_k = {}& \gamma^2\tilde{\xi}\left(\xi_0^{-1} - \xi_i^{-1}\right)\left(\xi_0^{-1} - \xi_j^{-1}\right)(\xi_k - \gamma) + \gamma\tilde{\xi}\xi_0^{-2}(\xi_k - \gamma) \\ &+ \mathbf{1}(i=j)\left(\gamma\tilde{\xi}\xi_j^{-2}(\xi_k - \gamma)\right) \\ &+ \mathbf{1}(i=k)\left(\gamma\tilde{\xi}\left(\xi_0^{-1} - \xi_j^{-1}\right)\right) \\ &+ \mathbf{1}(j=k)\left(\gamma\tilde{\xi}\left(\xi_0^{-1} - \xi_i^{-1}\right)\right)\end{aligned} \tag{41}$$

Applying Itô's formula to change variables to λ yields, for $k \geq 1$,

$$\mathrm{d}\lambda_k = \left(D_\xi\lambda_k \cdot b + \frac{\varepsilon^2}{2}\operatorname{tr}\left(\sigma^\mathsf{T}\left(D_\xi^2\lambda_k\right)\sigma\right)\right)\mathrm{d}t + \left(D_\xi\lambda_k\right)^\mathsf{T}\sigma\,\mathrm{d}B \tag{42}$$

where D_ξ is the gradient with respect to ξ and D_ξ^2 is the Hessian matrix with respect to ξ. No particular form of b is assumed.

Since σ^Y is symmetric and the trace has the general property that $\operatorname{tr}(PQR) = \operatorname{tr}(QRP)$, the trace term may be evaluated as follows despite the fact that no explicit form is possible for σ:

$$\operatorname{tr}\left(\sigma^\mathsf{T}\left(D_\xi^2\lambda_k\right)\sigma\right) = \operatorname{tr}\left(\left(D_\xi^2\lambda_k\right)\sigma\,\sigma^\mathsf{T}\right) = \operatorname{tr}\left(\left(D_\xi^2\lambda_k\right)A\right)$$

After a laborious simplification,

$$\operatorname{tr}\left(\sigma^\mathsf{T}\left(D_\xi^2\lambda_k\right)\sigma\right) = \gamma(\gamma+1)\lambda_k\sum_{j=0}^{K}\xi_j^{-1} \tag{43}$$

4.6. Well-posedness of the SDEs. The drift and volatility terms of (42) are continuously differentiable, so they automatically satisfy a local Lipschitz inequality, as required by the general theorem concerning the existence and uniqueness of solutions in [10, §5.3].

The theorem also requires a growth constraint formulated as follows. Let us adapt the usual big-O notation, using

$$f\left(\lambda^Y, \lambda^A\right) = g\left(\lambda^Y, \lambda^A\right) + \mathcal{O}^2$$

to mean that there exists a constant $H > 0$ such that for all λ^Y and λ^Y,

$$f\left(\lambda^Y, \lambda^A\right) - g\left(\lambda^Y, \lambda^A\right) < H\left(1 + \left\|\lambda^Y\right\|^2 + \left\|\lambda^A\right\|^2\right)$$

The growth constraint required in [10, §5.3] is $\beta^Y + \beta^A = \mathcal{O}^2$ where

$$
\begin{aligned}
\beta^Y &= \sum_{k=1}^{K} \lambda_k^Y \left(D_{\xi^Y} \lambda^Y \cdot b^Y + \frac{\varepsilon^2}{2} \operatorname{tr}\left(\left(\sigma^Y\right)^{\mathsf{T}} \left(D^2_{\xi^Y} \lambda_k^Y \right) \sigma^Y \right) \right) \\
&\qquad + \varepsilon^2 \operatorname{tr}\left(\left(\sigma^Y\right)^{\mathsf{T}} \left(D_{\xi^Y} \lambda^Y \right)^{\mathsf{T}} \left(D_{\xi^Y} \lambda^Y \right) \sigma^Y \right) \\
\beta^A &= \sum_{k=1}^{K} \lambda_k^A \left(D_{\xi^A} \lambda^A \cdot b^A + \frac{\varepsilon^2}{2} \operatorname{tr}\left(\left(\sigma^A\right)^{\mathsf{T}} \left(D^2_{\xi^A} \lambda_k^A \right) \sigma^A \right) \right) \\
&\qquad + \varepsilon^2 \operatorname{tr}\left(\left(\sigma^A\right)^{\mathsf{T}} \left(D_{\xi^A} \lambda^A \right)^{\mathsf{T}} \left(D_{\xi^A} \lambda^A \right) \sigma^A \right)
\end{aligned}
\tag{44}
$$

$D_{\xi^Y} \lambda^Y$ and $D_{\xi^A} \lambda^A$ are Jacobian matrices, and $D^2_{\xi^Y} \lambda^Y_k$ and $D^2_{\xi^A} \lambda^A_k$ are Hessian matrices. The difficulty here is that $\beta^Y + \beta^A$ turns out to contain terms of degree greater than 2, so we must confirm that these are negative for large λ. The following estimates are derived omitting the generation label where possible, as parallel logic applies to β^Y and β^A.

Incorporating (43), the generic β term is

$$
\beta = \sum_{k=1}^{K} \lambda_k \left(D_\xi \lambda \cdot b + \frac{\varepsilon^2}{2} \gamma(\gamma+1) \lambda_k \sum_{j=0}^{K} \xi_j^{-1} \right) + \varepsilon^2 \operatorname{tr}\left(\sigma \left(D_\xi \lambda \right) \left(D_\xi \lambda \right)^{\mathsf{T}} \sigma \right)
\tag{45}
$$

The remaining trace term can be evaluated by cyclically reordering the matrices

$$
\operatorname{tr}\left(\sigma^{\mathsf{T}} \left(D_\xi \lambda \right)^{\mathsf{T}} \left(D_\xi \lambda \right) \sigma \right) = \operatorname{tr}\left(\left(D_\xi \lambda \right)^{\mathsf{T}} \left(D_\xi \lambda \right) \sigma \sigma^{\mathsf{T}} \right) = \operatorname{tr}\left(\left(D_\xi \lambda \right)^{\mathsf{T}} \left(D_\xi \lambda \right) A \right)
$$

After a massive amount of simplification,

$$
\begin{aligned}
\beta &= -\gamma \left\| \lambda \right\|^2 \sum_{j=0}^{K} b_j \xi_j^{-1} + \varepsilon^2 \left(\frac{\gamma(\gamma+1)}{2} \lambda_0 + \gamma^2 \left\| \lambda \right\|^2 \right) \sum_{j=0}^{K} \xi_j^{-1} + \tilde{\xi} \left(\lambda \cdot b + 2\varepsilon^2 \lambda \cdot \xi \right) \\
&\quad + \varepsilon^2 \left(- \left\| \lambda \right\|^2 + \tilde{\xi}^2 \left(1 - \xi_0 - \left\| \xi \right\|^2 \right) + 2\gamma \tilde{\xi} \lambda_0 \right)
\end{aligned}
\tag{46}
$$

The largest magnitude terms are those that include ξ_j^{-1}, and those must be handled carefully. The others are $\mathcal{O}^2$ in light of inequalities proved in subsection 4.4, and the assumption that the b_j's are bounded.

$$
\beta = -\gamma \left\| \lambda \right\|^2 \sum_{k=0}^{K} \frac{b_k}{\xi_k} + \varepsilon^2 \gamma \left\| \lambda \right\|^2 \left(\frac{\gamma+1}{2 \left\| \lambda \right\|^2} + \gamma \right) \sum_{k=0}^{K} \frac{1}{\xi_k} + \mathcal{O}^2
\tag{47}
$$

To express the constraints on b^Y and b^A that are necessary to guarantee that the remaining large magnitude terms in β^Y and β^A are negative overall, the following definitions are required. Given $\mu > 0$, define the *μ-border* of $\mathcal{S}^K$ to be

$$
\mathcal{S}^K_\mu = \left\{ x \in \mathcal{S}^K \;\middle|\; \exists k : \; x_k < \mu \right\}
\tag{48}
$$

The *border class* of $x \in \mathcal{S}^K_\mu$ is $\operatorname{BC}(x; \mu) = \sum_k \mathbf{1}(x_k < \mu)$. The parameter μ will be omitted when it is clear from context.

LEMMA 4.5. *If $\xi \in \mathcal{S}_\mu^K$ then*

$$\tilde{\xi} \geq \mu^{-\gamma\,\mathrm{BC}(\xi)} \geq \mu^{-\gamma} \tag{49}$$

Proof. Let $c = \mathrm{BC}(\xi)$. Then there are c indices k for which $1/\mu < \xi_k^{-1}$ and $K-c$ indices for which $1 < \xi_k^{-1}$. Taking the γ power of the product yields (49). ∎

LEMMA 4.6. *Suppose $\left(b^Y, b^A\right)$ is bounded. Suppose there exist numbers $G > 0$, F, and $\gamma > \mu > 0$ such that*

$$\begin{aligned} \textit{if } \xi^Y \in \mathcal{S}_\mu^K \textit{ then } \sum_{k=0}^K \frac{b_k^Y}{\xi_k^Y} &\geq G \sum_{k=0}^K \frac{1}{\xi_k^Y} + F \\ \textit{and if } \xi^A \in \mathcal{S}_\mu^K \textit{ then } \sum_{k=0}^K \frac{b_k^A}{\xi_k^A} &\geq G \sum_{k=0}^K \frac{1}{\xi_k^A} + F \end{aligned} \tag{50}$$

Then there exists an $\varepsilon_0 > 0$ such that for each $\varepsilon_0 > \varepsilon > 0$, $\beta^Y + \beta^A = \mathcal{O}^2$.

Proof. If $\xi \in \mathcal{S}^K \setminus \mathcal{S}_\mu^K$, then λ is bounded and each ξ_k satisfies $1/\xi_k < 1/\mu$. Since b is assumed to be bounded, it is straightforward to confirm that $\beta = \mathcal{O}^2$ in this case.

Suppose $\xi \in \mathcal{S}_\mu^K$. Then from from (49), $\tilde{\xi} \geq \mu^{-\gamma}$. Consequently, (36) implies $\|\lambda\| \geq (\gamma\mu^{-\gamma} - 1)/\sqrt{2}$. Using the lower bound G to replace the b_k terms and pushing degree 2 terms into $\mathcal{O}^2$,

$$\beta = \gamma \|\lambda\|^2 \left(\sum_{k=0}^K \frac{1}{\xi_k}\right) \left[-G + \varepsilon^2 \left(\frac{\gamma+1}{2\|\lambda\|} + \gamma\right)\right] + \mathcal{O}^2$$

If ε is small enough,

$$\varepsilon \leq \sqrt{\frac{G}{\frac{\gamma+1}{\sqrt{2}(\gamma\mu^{-\gamma}-1)} + \gamma}} = \varepsilon_0$$

then the factor in square brackets is negative and $\beta = \mathcal{O}^2$.

Since the above arguments apply to both β^Y and β^A, the sum satisfies $\beta^Y + \beta^A = \mathcal{O}^2$. ∎

LEMMA 4.7. *If the vector field has the form*

$$\begin{aligned} b^Y\left(\xi^Y, \xi^A\right) &= U^Y\left(\xi^Y, \xi^A\right) - \xi^Y \\ b^A\left(\xi^Y, \xi^A\right) &= U^A\left(\xi^Y, \xi^A\right) - \xi^A \end{aligned}$$

where U^Y and U^A are probability vectors with uniform positive lower bounds

$$\begin{aligned} \forall \xi^Y, \xi^A: \quad U^Y\left(\xi^Y, \xi^A\right) &\geq U_{\min}^Y > 0 \\ \forall \xi^Y, \xi^A: \quad U^A\left(\xi^Y, \xi^A\right) &\geq U_{\min}^A > 0 \end{aligned}$$

then it satisfies (50).

Proof. For either generation, ∎

$$\sum_{k=0}^K \frac{U_k - \xi_k}{\xi_k} \geq U_{\min} \sum_{k=0}^K \frac{1}{\xi_k} - (K+1)$$

The example vector field (21) satisfies (50). Since the example Q from (2) is the probability vector for a binomial distribution, its least element is either Q_0 or Q_K. Therefore, each element of Q satisfies

$$Q_k \geq Q_{\min} = \min \left\{ q(0)^K, (1-q(0))^K, q(1)^K, (1-q(1))^K \right\}.$$

Each element of the distribution vector V as in (7) satisfies $V_k \geq \eta$. If we try to set $V\left(\xi^Y, \xi^A\right) = \xi^Y$, then there is no way to choose G, hence the need for $\eta > 0$.

PROPOSITION 4.8. *If b satisfies the hypotheses of Lemma 4.6, then for each $\varepsilon_0 > \varepsilon > 0$, the SDEs (42) and (26) have pathwise-unique strong solutions for all positive time starting from each suitable initial value.*

Proof. The theorem from [10, §5.3] in conjunction with Lemma 4.6 confirms the result for (λ^Y, λ^A), and the change of variables from subsection 4.4 maps those solutions to solutions of (26). ∎

4.7. Generalizations. The results in this section generalize to many other situations, since many of the proofs make no assumptions on the specific form of b, although they were developed to apply to (21). For example, if there are more than two grammars of interest, the indices 0 through K can be remapped to any mixtures of grammars and the learning function Q can be adjusted accordingly, resulting in a discrete time model that converges to continuous time process with the same form as (26). There's also no need to restrict Q to be the mass function for any particular distribution.

In formulating (26), it was assumed that both generations were subdivided into the same types, that is, everyone of type k uses $\mathcal{G}_1$ with probability k/K. The results in this section do not depend on requiring all sub-populations to have states with same interpretation, or even to lie in simplexes of the same dimension.

These results also generalize immediately to a population divided into any number of sub-populations, such as multiple age groups, geographic regions, or social classes. The key theorem 7.1 from Chapter 8 of [10] requires that the time step size be $\frac{1}{N}$. It would continue to apply if the sub-populations were of different sizes but all were proportional to N.

5. Dynamics in a 2-dimensional case. We will continue by restricting our attention to the case of $K = 1$. That is, simulated individuals use $\mathcal{G}_2$ exclusively or not at all, and in discrete time, X_0^Y is the fraction of the young generation that never uses $\mathcal{G}_2$ and X_1^Y is the fraction that always uses $\mathcal{G}_2$. Since $X_0^Y + X_1^Y = 1$, it is only necessary to deal with X_1^Y. Likewise we may focus on ξ_1^Y, ξ_1^A, and $Q_1 = q(r(X,Y))$.

The covariance function (22) reduces to a 2-by-2 diagonal matrix so it has a very simple square-root:

$$\sigma \begin{pmatrix} \xi^Y \\ \xi^A \end{pmatrix} = \begin{pmatrix} \sqrt{\xi_1^Y - \left(\xi_1^Y\right)^2} & 0 \\ 0 & \sqrt{\xi_1^A - \left(\xi_1^A\right)^2} \end{pmatrix} \tag{51}$$

As $N \to \infty$, the discrete time process converges weakly to the solution (ξ^Y, ξ^A) : $[0, \infty) \to (0,1) \times (0,1)$ of

$$\begin{aligned} \mathrm{d}\xi_1^Y &= \left(q\left(r\left(\xi_1^Y, \xi_1^A\right)\right) - \xi_1^Y\right) \mathrm{d}t + \varepsilon \sqrt{\xi_1^Y \left(1-\xi_1^Y\right)} \mathrm{d}B^Y \\ \mathrm{d}\xi_1^A &= \left(\xi_1^Y (1-\eta/2) + \eta - \xi_1^A\right) \mathrm{d}t + \varepsilon \sqrt{\xi_1^A (1-\xi_1^A)} \mathrm{d}B^A \\ \xi^Y(0) &= \xi_{\text{init}}^Y \text{ and } \xi_1^A(0) = \xi_{\text{init}}^A \end{aligned} \tag{52}$$

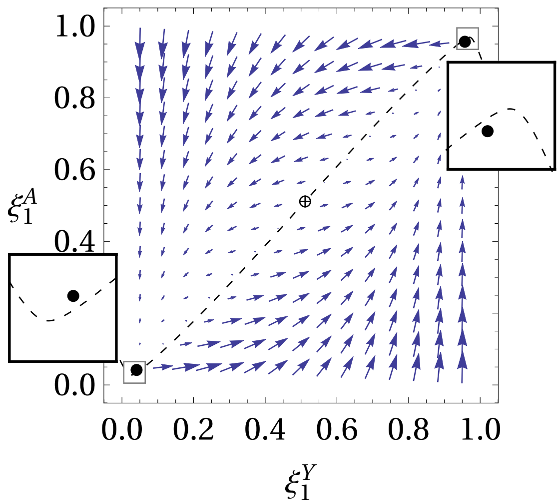

FIG. 3. *Phase portrait for (52) with $\varepsilon = 0$. The crossed dot $\oplus$ is a saddle point, the two dots $\bullet$ are sinks, and the dashed curve is the separatrix between their basins of attraction. The arrows indicate the direction of the vector field. Inset boxes show magnified pictures of the areas around the sinks.*

where B^Y and B^A are independent one-dimensional Brownian motions.

5.1. Comparison to the deterministic limit. In the deterministic limit $\varepsilon = 0$ and returning to the specific learning process described in section 2 and section 3, the dynamical system (52) has two stable equilibria representing populations where both generations are dominated by one grammar or the other. The separatrix forming the boundary between the two basins of attraction passes very close to the stable equilibria. See Figure 3. Under the stochastic dynamics, the population will hover near an equilibrium until random fluctuations cause it to stray across the separatrix, where it will be blown toward the other. It will continue to oscillate irregularly between the two equilibria for all time. These separatrix-crossing events generate spontaneous monotonic language changes separated by reasonably long intervals of temporary stability.

5.2. Memory kernel form. Another way to understand this form of instability is to express ξ_1^A as an average of ξ_1^Y over its past, with an exponential kernel giving greater weight to the recent past. This is accomplished by making two simplifications. First, the resampling step from the Markov chain will be applied only to the younger generation, which removes the random term from $\mathrm{d}\xi_1^A$ in (52) but not from $\mathrm{d}\xi_1^Y$. Second, η in the aging distribution V will be set to 0. This yields a linear ordinary differential equation for ξ_1^A with ξ_1^Y acting as an inhomogeneity

$$\frac{\mathrm{d}\xi_1^A}{\mathrm{d}t} = \xi_1^Y - \xi_1^A, \text{ with solution } \xi_1^A(t) = \mathrm{e}^{-t}\xi_{\text{init}}^A + \int_0^t \mathrm{e}^{-(t-s)}\xi_1^Y(s)\mathrm{d}s.$$

With this simplification, the dynamics for ξ_1^Y take the form of a stochastic functional-delay differential equation

$$\mathrm{d}\xi_1^Y(t) = \left(q\left(r\left(\xi_1^Y(t), K_t\xi_1^Y\right)\right) - \xi_1^Y(t)\right)\mathrm{d}t + \varepsilon\sqrt{\xi_1^Y(t)(1-\xi_1^Y(t))}\mathrm{d}B \quad (53)$$

where the delay appears through convolution with a memory kernel

$$K_t f = \mathrm{e}^{-t}\xi^A_{\mathrm{init}} + \int_0^t \mathrm{e}^{-(t-s)} f(s)\mathrm{d}s.$$

The age structure serves to give the population a memory, so that the speech pattern ξ^Y_1 of the young generation changes depending on how the current young generation deviates from its recent past average. Chance deviations of sufficient size are amplified when children detect them and predict that the trend will continue, yielding prediction-driven instability.

6. Discussion.

6.1. Comparison to other models. The discrete and continuous models as described in sections 2 and 3 are based on the Wright-Fisher model of population genetics as described in [10], which is formulated as a Markov chain and its limit as a stochastic differential equation for an infinite population. The original Wright-Fisher model takes values on an interval, which makes the theoretical analysis much simpler than for (26). A similar derivation to that of section 4 resulting in a Fokker-Planck equation is given in [3], without the theoretical treatment given here. The model in [58] derives a similar model, grounding the learning process in Bayesian inference. Neither these nor the Wright-Fisher model incorporate age structure or forces such as learning and prediction that are not present in biological birth-death processes.

A related dynamical system is the FitzHugh-Nagumo model for a spiking neuron [30, 42], which is a general family of two-variable dynamical systems. Its structure is similar to Figure 3 except that it has only the lower left stable equilibrium, which represents a resting neuron. A disturbance causes the neuron's state to stray away from that rest state and go on a long excursion known as an action potential or spike.

The language change model examined here differs from the stochastic FitzHugh-Nagumo model in several ways. It is derived as a continuous limit of a Markov chain rather than from adding noise to an existing dynamical system. It has two stable equilibria rather than one as long as ε is sufficiently small (although it is conceivable that some linguistic phenomenon might exhibit the single stable equilibrium). It is naturally confined to $\mathcal{S}^K \times \mathcal{S}^K$, where FitzHugh-Nagumo models occupy an entire plane. The random term added to a FitzHugh-Nagumo model is normally Brownian motion multiplied by a small constant. The change of variables $\theta = \arcsin(2\xi - 1)$, $\phi = \arcsin(2\zeta - 1)$ transforms the low dimensional case (52) to that form but the system remains confined to a square, and the change of variables to (42) on the whole plane has a non-constant coefficient on the Brownian motion. Thus, the theory of FitzHugh-Nagumo models must be adapted before it can be applied to this language model.

Population-level memory has been used to model other social trends that exhibit momentum. For example, the authors of [16] develop a model in which parents use a discrete-time memory kernel analogous to (53) to compute running averages of the popularity of given names, and use this information when naming babies. The case of language change is different because children seem to be capable of contributing without decades of accumulated experience. They must get historical information from some other source, and age-correlated differences in speech is a reasonable hypothesis.

7. Conclusion. The main goal of this article is to begin with a discrete time, finite population model that can represent spontaneous language change in a population between meta-stable states, each dominated by one idealized grammar, and connect it via solid theory to a continuous time, infinite population model. Language is represented as a mixture of the idealized grammars to reflect the variability of speech seen in manuscripts and social data. A Markov chain that includes age structure has all the desired properties for the first model. The population can switch spontaneously from one language to the other and the transition is monotonic. Intuitively, the mechanism of these spontaneous changes is that every so often, children pick up on an accidental correlation between age and speech, creating the beginning of a trend. The prediction step in the learning process amplifies the trend, and moves the population away from equilibrium, which suggests the term *prediction-driven instability* for this effect.

Fundamental results were proved. Specifically, in the limit as the number of agents goes to infinity, sample paths of the Markov chain converge weakly to solutions to a system of well-posed SDEs, which have the form of drift terms plus a small stochastic perturbation. The derivation of the correct SDEs and the proof that the convergence happens as intended require a change of variables specifically tailored to the geometry of the simplex, together with theoretical tools more sophisticated than those typically needed for population dynamics models. The proof that the system of SDEs is well-posed relies only on general properties of the drift vector field and the specific form of the infinitesimal covariance matrix.

Looking at a low dimensional case, in the limit of zero noise, the prediction-driven instability comes from the proximity of stable sinks to the separatrix of their basins of attraction. The instability comes from the general geometry of the phase space as in Figure 3. Alternatively, the prediction process may be understood as comparing the current state of the population to an average emphasizing its recent past, and chance deviations trigger the instability. Concrete formulas were given for q, r, and Q, but the interesting behavior is not limited to these examples.

Future studies of this model could include adapting and applying techniques for studying noise-activated transitions among meta-stable states, including exit time problems [13, 31, 32]. For example, it is possible to numerically estimate the time between transitions using a partial differential equation or a variational technique. The change of variables and associated theory may be of use to other dynamical systems whose phase space is a simplex, such as replicator dynamics [18].

REFERENCES

[1] D. Adger, *Core Syntax: A Minimalist Approach*, Oxford University Press, Oxford, May 2003.

[2] A. Alishahi and S. Stevenson, *A Computational Model of Early Argument Structure Acquisition*, Cognitive Science, 32 (2008), pp. 789–834.

[3] G. J. Baxter, R. A. Blythe, W. Croft, and A. J. McKane, *Utterance selection model of language change*, Physical Review E, 73 (2006), p. 046118.

[4] M. Becker, *There Began to Be a Learnability Puzzle*, Linguistic Inquiry, 37 (2006), pp. 441–456.

[5] R. C. Berwick and P. Niyogi, *Learning from Triggers*, Linguistic Inquiry, 27 (1996), pp. 605–622.

[6] T. Briscoe, *Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device*, Language, 76 (2000), pp. 245–296.

[7] T. Briscoe, *Grammatical acquisition and linguistic selection*, in Linguistic Evolution through Language Acquisition: Formal and Computational Models, T. Briscoe, ed., Cambridge University Press, Cambridge, UK, 2002, pp. 255–300.

[8] N. Chomsky, *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA, 1965.

[9] F. Cucker, S. Smale, and D.-X. Zhou, *Modeling Language Evolution*, Foundations of Computational Mathematics, 4 (2004), pp. 315–343.

[10] R. Durrett, *Stochastic Calculus: A Practical Introduction*, Probability and stochastics series, CRC Press, Boca Raton, 1996.

[11] R. Durrett and S. Levin, *The Importance of Being Discrete (and Spatial)*, Theoretical Population Biology, 46 (1994), pp. 363–394.
[12] Z. Fagyal, S. Swarup, A. M. Escobar, L. Gasser, and K. Lakkaraju, *Centers and peripheries: Network roles in language change*, Lingua, 120 (2010), pp. 2061–2079.
[13] M. I. Freĭdlin and A. D. Wentzell, *Random perturbations of dynamical systems*, Springer, 1998.
[14] E. Gibson and K. Wexler, *Triggers*, Linguistic Inquiry, 25 (1994), pp. 407–454.
[15] E. M. Gold, *Language Identification in the Limit*, Information and Control, 10 (1967), pp. 447–474.
[16] T. M. Gureckis and R. L. Goldstone, *How You Named Your Child: Understanding the Relationship Between Individual Decision Making and Collective Outcomes*, Topics in Cognitive Science, 1 (2009), pp. 651–674.
[17] M. D. Hauser, N. Chomsky, and W. T. Fitch, *The Faculty of language: What is it, who has it, and how did it evolve?*, Science, 298 (2002), pp. 1569–1579.
[18] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, 1998.
[19] C. L. H. Kam and E. L. Newport, *Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change.*, Language Learning & Development, 1 (2005), pp. 151–195.
[20] C. Kemp, A. Perfors, and J. B. Tenenbaum, *Learning overhypotheses with hierarchical Bayesian models*, Developmental Science, 10 (2007), pp. 307–321.
[21] S. Kirby, *Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity*, IEEE Transactions on Evolutionary Computation, 5 (2001), pp. 102–110.
[22] S. Kirby and J. R. Hurford, *The Emergence of Structure: An overview of the iterated learning model*, in Simulating the Evolution of Language, 2002, pp. 121–148.
[23] N. L. Komarova, P. Niyogi, and M. A. Nowak, *The Evolutionary dynamics of grammar acquisition*, Journal of Theoretical Biology, 209 (2001), pp. 43–59.
[24] A. Kroch, *Reflexes of Grammar in Patterns of Language Change*, Language Variation and Change, 1 (1989), pp. 199–244.
[25] W. Labov, *Principles of Linguistic Change: Internal Factors*, vol. 1, Blackwell, Malden, MA, 1994.
[26] W. Labov, *Principles of Linguistic Change: Social Factors*, vol. 2, Blackwell, Malden, MA, 2001.
[27] W. Labov, *Transmission and Diffusion*, Language, 83 (2007), pp. 344–387.
[28] D. Lightfoot, *How to Set Parameters: Arguments from Language Change*, MIT Press, Cambridge, MA, 1991.
[29] D. Lightfoot, *The Development of Language: Acquisition, Changes and Evolution*, Blackwell, Malden, MA, 1999.
[30] B. Lindner and L. Schimansky-Geier, *Analytical approach to the stochastic FitzHugh-Nagumo system and coherence resonance*, Physical Review E, 60 (1999), p. 7270.
[31] R. S. Maier and D. L. Stein, *The Escape Problem for Irreversible Systems*, chao-dyn/9303017, (1993).
[32] R. S. Maier and D. L. Stein, *A Scaling Theory of Bifurcations in the Symmetric Weak-Noise Escape Problem*, cond-mat/9506097, (1995).
[33] W. G. Mitchener, *Bifurcation analysis of the fully symmetric language dynamical equation*, Journal of Mathematical Biology, 46 (2003), pp. 265–285.
[34] W. G. Mitchener, *A Mathematical model of the loss of verb-second in Middle English*, in Medieval English and its Heritage, no. 16 in Studies in English Medieval Language and Literature, Peter Lang Pub Inc, 2006, pp. 189–202.
[35] W. G. Mitchener, *Game dynamics with learning and evolution of universal grammar*, Bulletin of Mathematical Biology, 69 (2007), pp. 1093–1118.
[36] W. G. Mitchener, *Inferring leadership structure from data on a syntax change in English*, in Scientific applications of language methods, no. 2 in Frontiers in Mathematical Linguistics and Language Theory, Imperial College Press, London, 2010, pp. 633–662.
[37] W. G. Mitchener, *Mean-field and measure-valued differential equation models for language variation and change in a spatially distributed population*, SIAM Journal on Mathematical Analysis, 42 (2010), pp. 1899–1933.
[38] W. G. Mitchener, *A Mathematical model of prediction-driven instability: How social structure can drive language change*, Journal of Logic, Language and Information, 20 (2011), pp. 385–396.
[39] W. G. Mitchener and M. Becker, *Computational models of learning the raising-control*

distinction, Research on Language and Computation, 8 (2011), pp. 169–207.

[40] W. G. Mitchener and M. A. Nowak, *Competitive exclusion and coexistence of universal grammars*, Bulletin of Mathematical Biology, 65 (2003), pp. 67–93.

[41] W. G. Mitchener and M. A. Nowak, *Chaos and language*, Proceedings of the Royal Society of London B: Biological Sciences, 271 (2004), pp. 701–704.

[42] J. D. Murray, *Mathematical Biology*, vol. 1, Springer-Verlag, New York, 2002.

[43] P. Niyogi, *The Computational Nature of Language Learning and Evolution*, MIT Press, Cambridge, MA, 2006.

[44] P. Niyogi and R. C. Berwick, *A Language learning model for finite parameter spaces*, Cognition, 61 (1996), pp. 161–193.

[45] M. A. Nowak, *Evolutionary Dynamics: Exploring the equations of life*, Harvard University Press, Cambridge, MA, 2006.

[46] M. A. Nowak and N. L. Komarova, *Towards an evolutionary theory of language*, Trends in Cognitive Sciences, 5 (2001), pp. 288–295.

[47] M. A. Nowak, N. L. Komarova, and P. Niyogi, *Evolution of universal grammar*, Science, 291 (2001), pp. 114–118.

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

[49] M. A. Nowak, D. C. Krakauer, and A. Dress, *An error limit for the evolution of language*, Proceedings of the Royal Society of London. Series B: Biological Sciences, 266 (1999), pp. 2131–2136.

[50] M. A. Nowak, J. Plotkin, and D. C. Krakauer, *The Evolutionary Language Game*, Journal of Theoretical Biology, 200 (1999), pp. 147–162.

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

[52] B. Øksendal, *Stochastic Differential Equations*, Universitext, Springer Berlin Heidelberg, Berlin, Heidelberg, 1985.

[53] K. M. Page and M. A. Nowak, *Unifying Evolutionary Dynamics*, Journal of Theoretical Biology, 219 (2002), pp. 93–98.

[54] L. Pearl and A. Weinberg, *Input Filtering in Syntactic Acquisition: Answers From Language Change Modeling*, Language Learning and Development, 3 (2007), pp. 43–72.

[55] A. Perfors, J. B. Tennenbaum, and E. Wonnacott, *Variability, Negative Evidence, and the Acquisition of Verb Argument Constructions*, Journal of Child Language, 37 (2010), pp. 607–642.

[56] J. B. Plotkin and M. A. Nowak, *Language evolution and information theory*, Journal of Theoretical Biology, 205 (2000), pp. 147–159.

[57] A. Radford, *Minimalist Syntax: Exploring the Structure of English*, Cambridge University Press, Cambridge, UK, June 2004.

[58] F. Reali and T. L. Griffiths, *Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift*, Proceedings of the Royal Society of London B: Biological Sciences, 277 (2010), pp. 429–436.

[59] S. Swarup and L. Gasser, *Unifying evolutionary and network dynamics*, Physical Review E, 75 (2007), p. 066114.

[60] B. Tesar and P. Smolensky, *Learnability in Optimality Theory*, MIT Press, Cambridge, MA, May 2000.

[61] P. E. Trapa and M. A. Nowak, *Nash Equilibria for an evolutionary language game*, Journal of Mathematical Biology, 41 (2000), pp. 172–1888.

[62] A. Warner, *Why DO Dove: Evidence for Register Variation in Early Modern English Negatives*, Language Variation and Change, 17 (2005), pp. 257–280.

[63] S. Wichmann and E. W. Holman, *Population Size and Rates of Language Change*, Human Biology, 81 (2009), pp. 259–274.

[64] S. Wichmann, D. Stauffer, C. Schulze, and E. W. Holman, *Do language change rates depend on population size?*, Advances in Complex Systems, 11 (2008), pp. 357–369.

[65] C. D. Yang, *Knowledge and Learning in Natural Language*, Oxford University Press, Oxford, 2002.

[66] W. Zuidema and B. de Boer, *The evolution of combinatorial phonology*, Journal of Phonetics, 37 (2009), pp. 125–144.