Stochastic Stability in Language Evolution

Michael J. Fox and Jeff S. Shamma

***Abstract*— We study a simple game-theoretical model of language evolution in finite populations. This model is of particular interest due to a surprising recent result for the infinite population case: under replicator dynamics, the population game converges to socially inefficient outcomes from a set of initial conditions with non-zero Lesbegue measure. If finite population models do not exhibit this feature then support is lent to the idea that small population sizes are a key ingredient in the emergence of linguistic coherence. We analyze a generalization of replicator dynamics to finite populations that leads to the emergence of linguistic coherence in an absolute sense: After a long enough period of time, linguistic coherence is observed with arbitrarily high probability as a mutation rate parameter is taken to zero. The perturbations are modeled as state-dependent "point mutations". Formally, the stochastically stable action profiles maximize the sum of the individual utilities. Our proofs use the resistance tree method.**

I. Introduction

A. Background

It is difficult to discount the import of language in the success of our species. Human language allows us to spread information at speeds that vastly outstrip the pace of biological evolution. Thus language can be seen as the technology that enables evolutionary change on cultural timescales. Nevertheless, how language first emerged remains somewhat of a mystery. Compounding the issue is a scarcity of physical evidence of the earliest speakers [1]. Two novel approaches to the study of language evolution have emerged in recent decades: genomics [2] and mathematical modeling (for a review, see for instance [3]). We concentrate on the latter. Mathematical modeling of language evolution is especially useful for checking the internal consistency of proposed theories. Alternatively, this endeavor is capable of providing insights into language learning in artificially intelligent systems [4], [5].

A popular approach to explaining language origins is the suggestion that the first languages were simple, possibly gestural [1] linkings from object to symbol. These proto-languages are the predecessors of modern compositional languages. The fundamental problem with the emergence of useful proto-languages is that of cooperation [6]. It is advantageous for many members of a population to associate symbols with objects consistently, but how does such a convention emerge? Invoking a particular symbol to refer to an object is only useful after a significant portion of the population has already adopted such a mapping. Game theory has proven to be a useful framework for studying these simple proto-language models [7], [8], [9], [10], [11], [12], [3].

We mention in passing that the problem we study can also be interpreted as a model of economic signaling [7], [8], [12], [13], although we do not explore this possibility here.

B. The Language Game

We consider a simple language game, first proposed in a substantially similar form in [13], and reformulated more recently in [9]. Each player's strategy (or *language*) is a pair of matrices $(P, Q) \in \mathbb{B}^{m\times n} \times \mathbb{B}^{n\times m} \equiv \mathcal{L}_{m\times n}$, where $\mathbb{B}^{m\times n}$ is the set of binary (having elements from $\{0, 1\}$), row-stochastic $m \times n$ matrices and $\mathcal{L}_{m\times n}$ is the set of *languages*. There are $n^m m^n$ languages in $\mathcal{L}_{m\times n}$. We refer to the two matrices as the *speaker* and *hearer* matrices, respectively. The speaker matrix maps objects to symbols, and the hearer matrix maps symbols to objects. Every player has the same set of languages available to them. The utility of player i with language (P_i, Q_i) is

$$u_i((P_i, Q_i), (\bar{P}, \bar{Q})) \equiv \frac{1}{2}\mathbf{Tr}(P_i\bar{Q}) + \frac{1}{2}\mathbf{Tr}(\bar{P}Q_i)$$

where $(\bar{P}, \bar{Q})$ are the average of the speaker and hearer matrices, respectively, over the entire population. We depart from the more conventional notation of utilities depending on the joint strategies to emphasize that individuals interact with the entire population and do so anonymously. Note that u_i does not depend on i other than through (P_i, Q_i). For this reason, we drop the subscript below. The two terms on the right hand side correspond to speaking and hearing, respectively. We can rewrite one of these terms as

$$\mathbf{Tr}(PQ) = \sum_{k=1}^{n}\sum_{j=1}^{m} p_{kj}q_{jk}$$

where p_{kj} is the kj^{th} element of P and similarly for q. We interpret this as follows: The inner summation is for a fixed object k. Only a single p_{kj} equals one due to the row-stochasticity. This is the symbol j that the speaker matrix P associates with object k. If the hearer Q associates symbol j with object k (i.e. $q_{jk} = 1$) then there is a contribution of one to the utility for object k. The total utility is computed by summing over the objects and weighting contributions from speaking and hearing equally. We include (P_i, Q_i) in $(\bar{P}, \bar{Q})$ in order to streamline the notation, but all of our results can easily be extended to the case where there are no self-interactions.

Research supported in part by AFOSR project #FA9550–08–1–0375 and the DARPA Physical Intelligence program (contract #HR0011–10–1–0009). M.J. Fox is supported by the Department of Defense through the National Defense Science & Engineering Graduate Fellowship Program.

M.J. Fox and J.S. Shamma are with the School of Electrical and Computer Engineering, College of Engineering, Georgia Institute of Technology
{mfox, shamma}@gatech.edu

This model can be augmented to accommodate differing weights for different symbols and events [8] although we do not consider this here. Characterization of various static equilibria for this model are carried out in [10], and corresponding dynamic models are considered in [9]. A discussion of robustness with respect to the specified learning dynamics can be found in [7]. We have up until now left the computation of $(\bar{P}, \bar{Q})$ from the joint strategy intentionally vague so that the same model can be used in both the infinite and finite population settings. We first describe the infinite population case.

1) Infinite Populations: The standard technique for modeling infinite populations is to consider a continuous mass of players [14]. There are n^2m^2 languages in $\mathcal{L}_{m\times n}$ so we define the population state space as $X = \mathbb{S}^{n^2m^2}$ where $\mathbb{S}^r$ is the r-dimensional simplex. We confer any ordering on $\mathcal{L}_{m\times n}$ so that each element x_i of a state $x \in X$ gives the fraction of the population that speaks a particular language. It follows that the subscripts (P_i, Q_i) refer to the i'th language in $\mathcal{L}_{m\times n}$ (not the i'th player) in this setting and similarly for the utilities u_i. We can then compute

$$(\bar{P}, \bar{Q}) = (\sum_{i=1}^{m^2n^2} x_i P_i, \sum_{i=1}^{m^2n^2} x_i Q_i),$$

the average language in the population at large. The standard evolutionary dynamic for studying games of this type is the replicator dynamic

$$\begin{aligned} \dot{x_i} &= x_i[u_i((P_i, Q_i), (\bar{P}, \bar{Q})) - \bar{u}((\bar{P}, \bar{Q}))] \\ &= x_i[\frac{1}{2}\mathbf{Tr}(P_i\bar{Q}) + \frac{1}{2}\mathbf{Tr}(\bar{P}Q_i) - \mathbf{Tr}(\bar{P}\bar{Q})] \end{aligned}$$

for $i = 1, ..., n^2m^2$. The term $\bar{u}((\bar{P}, \bar{Q})) = \mathbf{Tr}(\bar{P}\bar{Q})$ is the payout to the average of the population when it plays against itself. We will use this quantity as our measure of social welfare. The replicator dynamic is imitative: an unused strategies is never subsequently taken up. It follows that each vertex of the simplex is a rest point of the dynamic. What is surprising about the behavior of this system is that there are many neutrally stable strategies (sometimes referred to as weak evolutionarily stable strategies) where social welfare is not maximized that the system will converge to from a set of initial conditions with non-zero Lebesgue measure [12]. This is troubling for proponents of the simple proto-languages explanation of language origins. The retort is that small populations, where mutations can impact the population state, were integral to the formation of the first proto-languages.

2) Finite Populations: In the finite case, we consider N players and a population state space $X = \mathcal{L}_{m\times n}^N$. For the population state $x \in X$ we let $x_i = (P_i, Q_i)$ refer to the language of player i. We can compute

$$(\bar{P}, \bar{Q}) = (\frac{1}{N}\sum_{i=1}^{N} P_i, \frac{1}{N}\sum_{i=1}^{N} Q_i).$$

We reiterate that in this setting the subscript in x_i refers to the player while in the infinite population setting it refers to the language.

One issue with analyzing the language game in finite populations is that there are many different ways to generalize replicator dynamics and evolutionarily stable strategies (the associated static equilibrium concept) to finite populations (see for instance [15]). One particular approach [6] is to consider the limit of weak selection where the contribution of utility to an otherwise uniform reproductive fitness is taken to zero. For some analytical results associated with this solution concept, see for instance [16]. This is the approach taken in [11]. In that model, one player is selected at random proportional to its fitness and then a second randomly chosen player adopts the first player's language. It is shown that, in the limit of weak selection, population states that maximize social welfare are the only states for which no mutant strategy has a fixation probability higher than $1/N$. This analysis is used to argue that evolution directs the system towards linguistic coherence. However, it is clear that this particular model as specified will not, in general, converge to a socially efficient state with high probability. Such would require analyzing a system that exhibits strong selection—this is the idea that is pursued in this paper.

Specifically, in Section III we propose a model of reproduction in populations in which a randomly selected individual adopts the language of one of the players that has the current highest utility. That is, unless a mutation occurs with probability ϵ in which case a random language from a set of "nearby" languages is adopted. We analyze this model in the small mutation rate limit. The resulting prediction of linguistic coherence is in the form of stochastic stability, a concept introduced to study the evolution of social conventions, but not previously suggested in relation to the language game. We review stochastic stability in Section II. This paper makes three novel contributions: we analyze a stochastic, finite population model of the language game exactly for the case of strong selection, we draw a connection between the study of the evolution of social conventions and language evolution, and we suggest that non-equilibrium models like our own are adequate to explain the observed drift in languages over time.

In the next section, we briefly review the concept of stochastic stability.

II. Stochastic Stability

This introduction to the notion of stochastic stability will draw heavily from the presentation of Young [18] in the context of social conventions. We will develop these concepts here with an eye for brevity. We will consider a Markov process P^0 on a finite state space Z. We will restrict our interest to perturbations to this process of a specific form, defined below.

*Definition 2.1: Let P^ϵ be a Markov process on Z for each $\epsilon \in (0, \bar{\epsilon}]$. The process P^ϵ is a **regular perturbed Markov process** if P^ϵ is irreducible and aperiodic for every $\epsilon \in (0, \bar{\epsilon}]$ and for each $z, z' \in Z$ we have*

$$\lim_{\epsilon\to 0} P^\epsilon_{zz'} = P^0_{zz'},$$

and if $P^{\epsilon}_{zz'} > 0$ for some $\epsilon > 0$, then

$$0 < \lim_{\epsilon \to 0} P^{\epsilon}_{zz'} / \epsilon^{r(z,z')} < \infty$$

for some $r(z, z') \geq 0$.

The value $r(z, z') \in \mathbb{R}$ is called the *resistance* of the transition $z \to z'$. Clearly, $r(z, z')$ must be uniquely defined in order to satisfy the condition. Also, $P^0_{zz'} > 0$ if and only if $r(z, z') = 0$. That is, transitions that occur with non-zero probability under P^0 have zero resistance. Transitions that never occur can be considered as having infinite resistance so that $r(z, z')$ is always defined.

For each ϵ, there is a unique stationary distribution, μ^{ϵ}, associated with P^{ϵ} (by its irreducibility and aperiodicity). We can now formally define stochastic stability.

*Definition 2.2: A state z is **stochastically stable** (Young, 1993) if*

$$\lim_{\epsilon \to 0} \mu^{\epsilon}(z) > 0.$$

It has been shown elsewhere that the above limit exists for every z so that every regular perturbed Markov process has at least one stochastically stable state. These states are the ones that the system spends most time in over the long run when ϵ is small. It should be noted that the stochastically stable states correspond to the perturbed process P^{ϵ}. That is, which states survive in the presence of the perturbations will depend on how the perturbations are introduced. It is possible to arrive at different stochastically stable states for the same process P^0 by applying the perturbations differently. Also, the stochastically stable states correspond to the limiting case of ϵ approaching zero, and are not always particularly likely to be observed when ϵ is not small. Next we will describe how to compute the stochastically stable states.

A. Resistance Trees

A recurrent class of a Markov process is a set of states such that from any state in the set one can reach any other state in the set in finite time with positive probability, and no state outside the set is accessible from any state inside it. Let P^0 have K recurrent classes $E_1, E_2, ..., E_K$. We will define for every distinct pair of recurrent classes E_i and E_j, $i \neq j$, a sequence of states $\zeta = (z_1, z_2, ..., z_q), z_1 \in E_i, z_q \in E_j$ called an ij*-path*. The resistance of the path is the sum of resistances in the sequence, $r(\zeta) = r(z_1, z_2) + r(z_2, z_3) + ... + r(z_{q-1}, z_q)$. We further denote $r_{ij} = \min r(\zeta)$ as the ij-path with least resistance. r_{ij} is always positive because there cannot be a zero resistance path between two distinct recurrent classes.

Now, for each recurrent class E_j, construct a tree rooted at a vertex j corresponding to E_j. That is, a set of $K - 1$ directed edges such that each $E_i, i \neq j$ is represented by a vertex i and there is a unique directed path from any vertex different from j to j. The resistance of such a tree is the sum of the resistances r_{ij} on the $K - 1$ edges. The stochastic potential γ_j of the recurrent class E_j is the minimum resistance among all such trees rooted at j. We expect the recurrent classes of minimum stochastic potential to be the most likely when ϵ is small. This result has been formalized [19] as follows:

Theorem 2.1: Let P^{ϵ} be a regular perturbed Markov process, and let μ^{ϵ} be the unique stationary distribution of P^{ϵ} for each $\epsilon > 0$. Then $\lim_{\epsilon \to 0} \mu^{\epsilon} = \mu^0$ exists, and μ^0 is a stationary distribution of P^0. The stochastically stable states are precisely those states that are contained in the recurrent class(es) of P^0 having minimum stochastic potential.

Next we derive a bound on the stochastic potential of a state based on the construction of greedy, or myopic, forests. The bound will be tight for the models we analyze below.

B. Myopic Forests

In this section we introduce a lower bound on the stochastic potential of a recurrent class based on myopic forests. In the case that a myopic forest can be constructed that is itself a resistance tree, the bound is tight and the potential is the minimum over all resistance trees for that recurrent class. A tree has minimum resistance when the sum of all the resistances is minimum. A myopic forest minimizes the resistance of each outgoing edge individually without any connectedness constraint.

Lemma 2.1: Let P^{ϵ} be a regular perturbed Markov process with $E_1, E_2, ..., E_K$ the recurrent classes of P^0, then for any recurrent class j we have

$$\gamma_j \geq \sum_{i \neq j} \min_{k \neq i} (r_{ik}),$$

and the relationship is satisfied with equality whenever there exists a myopic forest $(\{1, ..., K\}, \{(ik) : k \in \textbf{\textit{argmin}}_{k \neq i} r_{ik}\})$ that is a tree rooted at j.

Proof: Assume w.l.o.g. that $j = 1$, then γ_1 is given by the following optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i \neq 1} r_{ik_i} \\ \text{subject to} & (\{1, ..., K\}, \{ik_i : i \in \{2, ..., K\}\}) \\ & \quad \text{is a tree rooted at } 1. \end{array}$$

We arrive at our inequality by dropping the constraint and interchanging the order of summation and minimization. It is obvious that when the condition for equality is added as a constraint then the two optimizations are identical. ■

Next we present our dynamic model.

III. The Dynamic Model

The N players play the language game with the following model of reproduction. At each time t select an agent i at random according to some distribution $F(x)$ satisfying $\mathbf{Pr}[F(x) = i] > 0 \; \forall i \in \{1, ..., N\}, x \in X$. Let

$$x_i[t+1] = \begin{cases} x_{\hat{k}}, & w.p. \quad 1 - \epsilon \\ \textbf{rand}(B_2^H(P_i[t], Q_i[t])), & w.p. \quad \epsilon, \end{cases}$$

where $\hat{k} = \textbf{argmax}_k u_k((P_k[t], Q_k[t]), (\bar{P}[t], \bar{Q}[t]))$, and $\textbf{rand}(B_2^H(P_i[t], Q_i[t]))$ refers to the language given by sampling from the set of accessible mutant languages $B_2^H(P_i[t], Q_i[t]) \subset \mathcal{L}_{m \times n}$ uniformly. We use $B_d^H(l)$ to refer to the element-wise ball of Hamming distance d centered at l. Furthermore let

$$x_j[t+1] = x_j(t) \quad \forall \quad j \neq i.$$

In words, we select a random agent and assign him the language of an individual with a utility that is currently highest, or with small probability we randomly reassign a single row of either $P_i[t]$ or $Q_i[t]$. This dynamic model gives a perturbed Markov processes $\mathcal{P}_{m\times n,N}$ for particular values of m, n, N. We call a state *homogeneous* if for some $l \in \mathcal{L}_{m\times n}$ we have $x = (l, l, ..., l)$. Clearly the absorbing states of the unperturbed process are precisely the homogeneous states. We next compute the stochastically stable states of this process, first for the case where $m = n$ and then for $m > n$ ($n > m$ is then implied by symmetry). Recall that, in the long run, the process spends an arbitrarily large proportion of its time in the stochastically stable states as ϵ goes to zero. We remark that in [20] we analyzed the same system except with the ϵ-probability mutation events causing the selected player i to adopt a new language at random, uniformly from all of $\mathcal{L}_{m\times n}$. The state-dependent "point mutations" we consider here are somewhat more realistic.

A. The $m = n$ Case

We will want to make use of the bound we introduced in Lemma 2.1. In order to do so we must compute some minimum resistances from homogeneous states. First consider homogeneous states that maximize linguistic coherence. These are the states that satisfy $\mathbf{Tr}(\bar{P}\bar{Q}) = n$. This condition implies that $\mathbf{Tr}(P_iQ_i) = n\ \forall i \in \{1, ..., N\}$. We call languages satisfying this condition *aligned*. We call the homogeneous states corresponding to aligned languages *optimal*. The next lemma characterizes the minimum resistance from optimal states.

Lemma 3.1: When $m = n$ the minimum resistance from an optimal state $x = (P, Q)^N$ to any other homogeneous state is N and this is achieved by any homogeneous state in a language differing from x in a single row of one of either P or Q.

Proof: Suppose $\tilde{x}$ is arrived at by reassigning $\tilde{N} < N$ rows among language matrices in x. Let $(\bar{P}, \bar{Q})$ be the average language in $\tilde{x}$. Since each mutation event changes a single row, this implies a resistance of $\tilde{N}$. We will show that

$$u((P, Q), (\bar{P}, \bar{Q})) > u((P', Q'), (\bar{P}, \bar{Q})),$$

for any language (P', Q') in $\tilde{x}$ other than (P, Q). This will imply that, without further mutations, the state will revert back to x. Thus the minimum resistance is at least N. We have

$$u((P, Q), (\bar{P}, \bar{Q})) = m - \frac{\tilde{N}}{2N},$$

because each mutated row gives a loss of $\frac{1}{2N}$. First consider aligned (P', Q'). This implies that (P', Q') differs from (P, Q) in α rows, where $4 \leq \alpha \leq \tilde{N}$ is even. We then have $u((P', Q'), (\bar{P}, \bar{Q}))$

$$\begin{aligned}
&= u((P', Q'), (P, Q)) + u((P', Q'), (\bar{P}, \bar{Q})) - u((P', Q'), (P, Q)) \\
&= m - \frac{\alpha}{2} + u((P', Q'), (\bar{P}, \bar{Q})) - u((P', Q'), (P, Q)) \\
&= m - \frac{\alpha}{2} + u((\bar{P}, \bar{Q}), (P', Q')) - u((P, Q), (P', Q')) \\
&\leq m - \frac{\alpha}{2} + \frac{\tilde{N}}{2N},
\end{aligned}$$

where the inequality follows from noting that communication efficiency with (P', Q') is maximized by having all $\tilde{N}$ mutations applied in a manner consistent with (P', Q'). Combining the expressions for the two languages we get $u((P, Q), (\bar{P}, \bar{Q})) - u((P', Q'), (\bar{P}, \bar{Q})$

$$\geq m - \frac{\tilde{N}}{2N} - m + \frac{\alpha}{2} - \frac{\tilde{N}}{2N} = \frac{\alpha}{2} - \frac{\tilde{N}}{N} > 0.$$

Next consider (P', Q') not aligned, this implies

$$u((P', Q'), (\bar{P}, \bar{Q})) \leq m - \frac{1}{2}$$

so that $u((P, Q), (\bar{P}, \bar{Q})) - u((P', Q'), (\bar{P}, \bar{Q})$

$$\geq m - \frac{\tilde{N}}{2N} - m + \frac{1}{2} = \frac{N - \tilde{N}}{2N}.$$

The minimum resistance is no less than N, but N consecutive, identical mutations to N different users gives a new homogeneous state. So the minimum resistance is N and the new language differs from (P, Q) in a single row. ■

The minimum resistance from a sub-optimal state is one. This is because at least one row of either the speaker or hearer matrix does not contribute to the trace (by sub-optimality). Mutating this row has no utility consequences, so the mutant can fixate in the population without further resistance. We will show that the myopic forest is a tree, i.e. Lemma 2.1 is satisfied with equality. We claim that from any state we can reach any optimal state via a sequence of homogeneous states with each edge having minimal resistance, given the source. That is to say, the sequence will include optimal and sub-optimal states with all edges emanating from the former having resistance N and all edges emanating from the latter having resistance one. The claim follows immediately by induction on the element-wise Hamming distance from the target optimal state once we establish the following Lemma:

Lemma 3.2: Given any homogeneous state $x = (P, Q)^N$ and any target optimal state $\hat{x} = (\hat{P}, \hat{Q})^N$ there exists a homogeneous state $\tilde{x} = (\tilde{P}, \tilde{Q})^N$ that is among the homogeneous states that can be reached from x with minimum resistance and satisfies

$$d_H((\tilde{P}, \tilde{Q}), (\hat{P}, \hat{Q})) < d_H((P, Q), (\hat{P}, \hat{Q})),$$

where d_H is the element-wise Hamming distance.

Proof: If x is optimal then all states in $B_2^H((P, Q))$ are reached with resistance equal to N, which is minimum. We can simply choose any row from either P or Q that does not match the corresponding row in $\hat{P}$ or $\hat{Q}$ and correct it. The resulting homogeneous state satisfies the Lemma.

Now suppose x is sub-optimal. This implies that either P or Q has row(s) not contributing to the trace. If we mutate one of these rows then the mutant will have the same fitness as players with the language (P, Q). It follows that the mutant can fixate without resistance. If one of these rows does not match $(\hat{P}, \hat{Q})$ then we are done, so assume that all rows that do not contribute to trace already match their corresponding row in $(\hat{P}, \hat{Q})$. In this case, there is a zero column in either P or Q. To see this, assume the contrary and let I be the indices of the rows in P that do not contribute to the trace. It follows that I are also indices of columns in Q not contributing to the trace. By assumption Q has no zero columns, so each column in Q with an index in I contains a single one. Let $J(I)$ be the indices of the rows in Q such that for each $j \in J(I)$ there exists $i \in I$ satisfying $Q_{ji} = 1$. This implies

$$Q_{ji} = \hat{Q}_{ji}, \quad \forall j \in J(I), i \in \{1, ..., m\},$$

by assumption. Now since Q has a single one in each column the above is equivalent to

$$Q_{ji} = \hat{Q}_{ji}, \quad \forall i \in I, j \in \{1, ..., m\},$$

i.e. the columns indexed by I in Q match $\hat{Q}$. The rows indexed by I in P also match $\hat{P}$. Recall that $(\hat{P}, \hat{Q})$ is aligned so the rows indexed by I in P actually do contribute to the trace, contradicting our assumptions. Thus, (P, Q) has a zero column.

We now show that the presence of a zero column in either P or Q guarantees the existence of a language $(\tilde{P}, \tilde{Q})$ satisfying the lemma. If any rows not contributing to the trace do not match $(\hat{P}, \hat{Q})$ we are done, so assume all such rows match. Assume without loss of generality that Q has a zero column, i.e. there exists i_1, i_2, and j such that $Q_{i_1 j} = Q_{i_2 j}$. Further assume without loss of generality that row i_1 contributes to trace and row i_2 does not. This implies that row i_2 matches the corresponding row in $\hat{Q}$, so $\hat{P}_{ji_2} = 1$ since all rows in $\hat{Q}$ contribute to trace. Since $P_{ji_1} = 1$ we can mutate this row so that $P_{ji_2} = 1$. This mutant has the same fitness as (P, Q) in a population of players using (P, Q), and is thus a suitable $(\tilde{P}, \tilde{Q})$. ∎

We can now give the main result for the $m = n$ case.

Theorem 3.1: For any $N \geq 3$ and any $m \geq 2$ the stochastically stable states of the process $\mathcal{P}_{m \times m, N}$ are the optimal states.

Proof: First, consider trees rooted at optimal states. Lemma 3.2 implies we can construct a myopic forest per Lemma 2.1 and that this forest is a resistance tree. The bound is therefore satisfied with equality and the stochastic potential of the optimal states is $N(B-1) + \bar{B}$, where B and $\bar{B}$ are the number of optimal and sub-optimal homogeneous states, respectively, with languages in $\mathcal{L}$. Next consider trees rooted at sub-optimal homogeneous states. Applying Lemma 2.1 we find that the stochastic potential is at least $NB + \bar{B} - 1$, which concludes the proof. ∎

Next, we develop a similar result for the case of $m > n$.

B. The $m > n$ Case

Careful inspection of the arguments in Lemma 3.2 reveal that, for (P, Q) sub-optimal, we nowhere assumed $m = n$. The Lemma therefore carries over. However, the minimum resistance targets from optimal states are different in this case.

Lemma 3.3: The minimum resistance between an optimal state and any other state is one, and this is achieved by an optimal language.

Proof: Since $m > n$, Q must have a zero column. The corresponding row in P cannot contribute to fitness so any P' that is the same as P except for this row achieves the same utility. Alternatively, since $m > n$, P must have a column with two or more ones. The corresponding row in Q can have a one in at least two different positions that will maximize utility. ∎

We have shown that every homogeneous (absorbing) optimal state can transition to *some* other homogeneous optimal states with resistance one. Can we reach any optimal state via a sequence of transitions through optimal states, each having resistance one? We answer this question in the affirmative by presenting a constructive algorithm, `Path`.

Example: `Path`

$$(P, Q) = \left(\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right),$$

$$(P', Q') = \left(\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \right),$$

$$\Rightarrow \texttt{Path}((P', Q'), (P, Q)) = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \right.$$

$$\left. \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \right),$$

where we have purposefully omitted the initial and final languages as well as each unchanged matrix in the sequence. The `Path` algorithm always alternates between modifying the speaker and hearer matrix in each step. §

The `Path` algorithm is a recursion that terminates when the last language in the path Φ is the same as the language in the target argument (lines 2-8). If the speaker matrices match, but the hearer matrices do not then we can always reach the target language in one more step (line 6). The idea is to drive down the element-wise hamming distance between the current speaker matrix P and the target P'. Since P can only be part of an aligned language when it has no zero columns, we attempt to identify ones in P that do not match P' (line 9) and are also in columns that sum to more than one. Otherwise changing that ones position in its row would

Algorithm 1 `Path`$((P', Q'), \Phi)$

1: $(P, Q) \leftarrow \Phi_{|\Phi|}$
2: **if** $P = P'$ **then**
3: **if** $Q = Q'$ **then**
4: **return** Φ
5: **else**
6: **return** $(\Phi_0, .., \Phi_{|\Phi|}, (P', Q'))$
7: **end if**
8: **end if**
9: $\mathcal{K} \leftarrow \{(i, j) : P_{ij} > P'_{ij}\}$
10: $\hat{\mathcal{K}} \leftarrow \{(i, j) \in \mathcal{K} : \sum_k P_{kj} > 1\}$
11: **if** $|\hat{\mathcal{K}}| > 0$ **then**
12: $(i^*, j^*) \leftarrow$ **argmin**$_{(i,j)\in\hat{\mathcal{K}}} \sum_k Q_{ki}$
13: **if** $\sum_k Q_{ki^*} = 0$ **then**
14: $\hat{P}_{ij} \leftarrow \begin{cases} P_{ij}, & i \neq i^* \\ P'_{ij}, & i = i^* \end{cases}$
15: **return** `Path`$((P', Q'), (\Phi_0, \Phi_1, ..., \Phi_{|\Phi|}, (\hat{P}, Q)))$
16: **else**
17: **let** $\hat{i} \neq i^*$ satisfy $P_{\hat{i}j^*} = 1$
18: $\hat{Q}_{ij} \leftarrow \begin{cases} Q_{ij}, & i \neq j^* \\ 0, & i = j^*, j \neq \hat{i} \\ 1, & i = j^*, j = \hat{i} \end{cases}$
19: **return** `Path`$((P', Q'), (\Phi_0, \Phi_1, ..., \Phi_{|\Phi|}, (P, \hat{Q})))$
20: **end if**
21: **else**
22: **let** j^* satisfy $\sum_k Q'_{kj^*} = 0$
23: **let** $(\hat{i}, \hat{j}) \in \mathcal{K}$
24: $\hat{P}_{ij} \leftarrow \begin{cases} P_{ij}, & i \neq j^* \\ 0, & i = j^*, j \neq \hat{j} \\ 1, & i = j^*, j = \hat{j} \end{cases}$
25: $\Phi \leftarrow (\Phi_0, \Phi_1, ..., \Phi_{|\Phi|}, (\hat{P}, Q))$
26: **let** $\tilde{i} \neq \hat{i}$ satisfy $\hat{P}_{\tilde{i}\hat{j}} = 1$
27: $\hat{Q}_{ij} \leftarrow \begin{cases} Q_{ij}, & i \neq \hat{j} \\ 0, & i = \hat{j}, j \neq \tilde{i} \\ 1, & i = \hat{j}, j = \tilde{i} \end{cases}$
28: **return** `Path`$((P', Q'), (\Phi_0, \Phi_1, ..., \Phi_{|\Phi|}, (P, \hat{Q})))$
29: **end if**

create a zero column in the new P. If we can find such a one in P whose corresponding column in Q is zero (lines 12-13) then we can move the one so that its row matches P' (lines 14-15). If all such ones correspond to non-zero columns in Q then we instead modify Q (lines 16-19). Whenever a column of P sums to more than one, the corresponding row in Q can change while still maintaining alignment with P. We do this in a manner so as to create a zero column in Q so that in the next recursive call line 13 evaluates true. If we did not find any suitable ones in P (i.e. line 11 evaluates false) then we must actually generate a new P that is further from P' (lines 21-25). We do this in order to add a one to a column that currently contains a mismatched one so that we can later move the mismatched one while maintaining the alignment. We then adjust the corresponding row in Q (lines 26-28) so that we have a zero column in Q corresponding to the mismatched one in P that is in a column that now sums to more than one. Despite the fact that lines 21-28 imply two steps in Φ that do not move P closer to P', the overall sequence does reach P' in finitely many steps.

The behavior of `Path` is described by the following lemma:

Lemma 3.4: The `Path` *algorithm takes two aligned languages, one initial and one final, and returns a sequence of aligned languages linking the associated initial and final optimal states via transitions through the associated optimal states each having resistance one .*

We refer the reader to [20] for a proof. We can now prove our main result for the $m > n$ case.

Theorem 3.2: For any $N \geq 3$ and any $m > n \geq 2$ the stochastically stable states of the process $\mathcal{P}_{m\times n,N}$ are the optimal states.

Proof: For optimal states we can construct a resistance tree where every edge has resistance one. The edges emanating from sub-optimal states can all reach some optimal state via transitions of resistance one. From each optimal state with language (P, Q) we find the path to the candidate optimal state with language (P', Q') via `Path`$((P', Q'), (P, Q))$, eliminating redundancies as needed. This gives only edges of resistance one. What remains is to show that for sub-optimal states, at least one edge of each resistance tree has resistance greater than 1. There must be at least one edge that goes from an optimal state to a sub-optimal state. Suppose that in x there are $N-1$ agents that speak an aligned language (P, Q) and a single agent speaks misaligned language (P', Q'). The resistance between the homogeneous states in these languages is greater than one if $u((P, Q), (\bar{P}, \bar{Q})) > u((P', Q'), (\bar{P}, \bar{Q}))$. We compute $u_(P, Q), (\bar{P}, \bar{Q})) - u_(P', Q'), (\bar{P}, \bar{Q}))$

$$
\begin{aligned}
&= \frac{1}{2}\mathbf{Tr}((P - P')\frac{1}{N}((N-1)Q + Q')) \\
&+ \frac{1}{2}\mathbf{Tr}(\frac{1}{N}((N-1)P + P')(Q - Q')) \\
&= \frac{N-1}{2N}\mathbf{Tr}((P - P')Q + P(Q - Q')) \\
&+ \frac{1}{2N}\mathbf{Tr}(PQ' + P'Q - 2P'Q') \\
&= \frac{N-2}{2N}\mathbf{Tr}((P - P')Q + P(Q - Q')) \\
&+ \frac{1}{N}\mathbf{Tr}(PQ - P'Q')) > 0
\end{aligned}
$$

■

IV. DISCUSSION

We analyzed this process separately for the cases where the number of objects and symbols agree and disagree. In the more natural setting where the number of objects and symbols disagree we showed that we could transit between any two optimal states through a sequence of optimal states requiring only one mutation per transition. This (along with

the non-equilibrium nature of the process) concords with the observed phenomenon of drift in languages. That is, languages seem to change over time (see for instance, [21]) in a manner that is neutral with respect to the expressiveness of the language. The presence of synonyms and homonyms, exploited in our `Path` algorithm, seems a reasonable mechanism for this action.

We note that our results are not especially sensitive to our choice of dynamics. In [20] we show that a number of variations on the model are equivalent with respect to the characterization of stochastically stable states. In particular, our model assumed selection is very strong and only the most fit player reproduces its language. This assumption can be relaxed somewhat without consequence.

A. Future Directions

A key feature of the model that can be generalized is the form of the utility. We computed the utility in a manner reflecting a global interaction. It is possible to instead compute each agents utility based on their ability to communicate with some subset of the total population. This subset could come from either some fixed, exogenous graph or some endogenous considerations. An interesting question that emerges when considering these circumstances is the problem of linguistic diversity. What conditions are needed for heterogeneous states to persist in the population with non-vanishing frequency? Can we quantify network effects on social welfare? These are among the interesting questions that can be studied by considering generalizations to the utility functions of this game that move away from the "everyone talks to everyone" paradigm. These extensions are being pursued by the authors.

REFERENCES

[1] M. Balter, "Animal Communication Helps Reveal Roots of Language," *Science*, vol. 328, no. 5981, pp. 969–971, 2010.

[2] L. Cavalli-Sforza, "Genes, peoples and languages," *Proc. Natl Acad. Sci. USA*, vol. 94, pp. 7719–7724, 1997.

[3] M. Nowak, N. Komarova, and P. Niyogi, "Computational and evolutionary aspects of language," *Nature*, vol. 417, pp. 611–17, 2002.

[4] S. G. Ficici and J. B. Pollack, "Coevolving communicative behavior in a linear pursuer-evader game," in *Proceedings of the Fifth International Conference of the Society for Adaptive Behavior*, K. Pfeifer, Blumberg, Ed. Cambridge: MIT Press, 1998.

[5] C. H. Yong and R. Miikkulainen, "Coevolution of role-based cooperation in multi-agent systems," Department of Computer Sciences, The University of Texas at Austin, Tech. Rep. AI07-338, 2007.

[6] M. Nowak, A. Sasaki, C. Taylor, and D. Fudenberg, "Emergence of cooperation and evolutionary stability in finite populations," *Letters to Nature*, pp. 646–650, April 2004.

[7] S. Huttegger, "Robustness in signaling games," *Philosophy of Science*, vol. 74, pp. 839–847, 2007.

[8] G. Jager, "Evolutionary stability conditions for signaling games with costly signals," *Journal of Theoretical Biology*, vol. 253, no. 1, pp. 131 – 141, 2008.

[9] M. Nowak, J. Plotkin, and D. Krakauer, "The evolutionary language game," *Journal of Theoretical Biology*, vol. 200, no. 2, pp. 147 – 162, 1999.

[10] P. Trapa and M. Nowak, "Nash equilibria for an evolutionary language game," *Journal of Mathematical Biology*, vol. 41, pp. 172–188, 2000, 10.1007/s002850070004.

[11] C. Pawlowitsch, "Finite populations choose an optimal language," *Journal of Theoretical Biology*, vol. 249, no. 3, pp. 606 – 616, 2007.

[12] ——, "Why evolution does not always lead to an optimal signaling system," *Games and Economic Behavior*, vol. 63, no. 1, pp. 203 – 226, 2008.

[13] D. Lewis, *Convention: A Philosophical Study*. Harvard Univ. Press, Cambridge, MA., 1969.

[14] W. H. Sandholm, *Population Games and Evolutionary Dynamics*. MIT Press, 2010.

[15] J. Maynard Smith, "Can a mixed strategy be stable in a finite population?" *J. Theor. Biol.*, vol. 130, pp. 247–251, 1988.

[16] T. Antal, A. Traulsen, H. Ohtsuki, C. E. Tarnita, and M. A. Nowak, "Mutation-selection equilibrium in games with multiple strategies," *Journal of Theoretical Biology*, vol. 258, no. 4, pp. 614 – 622, 2009.

[17] J. Bergin and B. Lipman, "Evolution with state-dependent mutations," *Econometrica*, vol. 64, no. 4, pp. 943–56, July 1996.

[18] H. Young, *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton University Press, 1998.

[19] ——, "The evolution of conventions," *Econometrica*, vol. 61, no. 1, pp. 57–84, January 1993.

[20] M. Fox and J. Shamma, "Language evolution in finite populations," in *Submitted for conference publication*, 2011.

[21] O. Jespersen, *Progress in Language with Special Reference to English*. London: Swan Sonnenschein & Co., 1894.