Modelling Language Evolution [*†]

Felipe Cucker
Department of Mathematics
City University of Hong Kong
83 Tat Chee Avenue, Kowloon
HONG KONG
e-mail: macucker@math.cityu.edu.hk

Steve Smale
Toyota Technological Institute at Chicago
1427 East 60th Street, Chicago, IL 60637
U.S.A.
e-mail: smale@tti-c.org

Ding-Xuan Zhou
Department of Mathematics
City University of Hong Kong
83 Tat Chee Avenue, Kowloon
HONG KONG
e-mail: mazhou@math.cityu.edu.hk

June 2, 2003

1 Introduction

A purpose of this paper is to understand the evolution of the languages used by the agents of a society. We focus on language features in which convexity plays a central role.

In our model a language is a function from a set X of objects or meanings to a set Y of signals belonging to a prespecified class $\mathcal{F}$ in which a distance d is defined. In a linguistic society $\mathcal{P}$ of k agents, a state consists of the set of k languages used by the individual agents. Such a state evolves with time as agents are exposed to meaning-signal pairs produced by other agents in the society and modify their own

[*]This work has been substantially funded by a grant from the Research Grants Council of the Hong Kong SAR (project number CityU 1002/99P). Also, the second named author expresses his appreciation to City University of Hong Kong for its support.

[†]A preliminary version of this paper appears in [Minett and Wang 2003].

languages appropriately. We may say that such an agent learns from the present state of the society and that these learning processes form a learning dynamic. This dynamic depends on the object-signal pairs to which individual agents are exposed. In our model we will eventually assume that these pairs are randomly drawn from $X \times Y$ according to a probability measure in this product space reflecting both the frequency with which different meanings occur in the linguistic setting and the current state as well as noise.

A key role in this dynamic is played by the strength with which each agent affects other agent's language evolution through linguistic encounters. This set of influences is modelled by a $k \times k$ matrix Γ, the *communication matrix* of the society, whose entry γ_{ij}, a non-negative real number, measures the influence of agent j in the development of the language of agent i. Thus, convergence to a common language is related to an irreducibility property of Γ, which we call *weak irreducibility*, and the speed of this convergence to a number associated to Γ. Weak irreducibility ensures the existence of sufficiently many linguistic connections (i.e., non-zero γ_{ij}'s) thus ruling out the possibility of partition the society into two disjoint subgroups which are isolated from one another.

Let $\Delta_{\mathcal{F}} \subset \mathcal{F}^k$ be the diagonal of $\mathcal{F}^k$, i.e., $\Delta_{\mathcal{F}} = \{(f, f, \ldots, f) \mid f \in \mathcal{F}\}$. This is the set of states in which agents of the society share a common language. Also, let $N(\Delta_{\mathcal{F}}, \tau) = \{f \in \mathcal{F}^k \mid d(f, \Delta_{\mathcal{F}}) \leq \tau\}$ be the τ-neighborhood of $\Delta_{\mathcal{F}}$. Our main result can be stated as follows (for a detailed and quantitative statement see Corollary 2 of Theorem 1 below).

Main Result. *Let $\mathcal{P}$ be a society whose communication matrix is weakly irreducible and $\tau > 0$. Assume that at each iteration the agents of $\mathcal{P}$ are exposed to m meaning-signal pairs for a sufficiently large m. Then the learning dynamic converges in a finite number t of steps, with high probability to a state in $N(\Delta_{\mathcal{F}}, \tau)$.*

The Main Result above has obvious interpretations as "evolution to a common language." However, we are using the words "language", "linguistic", etc, in a broad sense and we warn the reader not to read too much into applications to human language. The abstraction is on a level that could express even certain aspects of the internet, biological networks, or networks of economic agents.

We do consider a linguistic model of the emergence of a primitive language using a convex space of utterances, all vowel sounds. Some detail can be given to an interpretation of the Main Result to a theory of convergence of the agents' languages. Here the models and computer simulations of a number of researchers have been an important inspiration to us.

In recent years the subject "Evolution of Language" has seen much activity. We make no attempt to summarize this except to highlight the following five references [De Boer 2001; Ke, Minett, Au, and Wang 2002; Kirby and Hurford 2001; Niyogi 2003a; Nowak, Komarova, and Niyogi 2001]

Conversations with Partha Niyogi have been very useful and the references he has given us have been important for us. Also appreciation to Bill Wang is acknowledged.

2 Language-like functions

In studying the way languages develop to become the shared communication system they are today, a simplified model starts with the representation of a language as a continuous function from a set X of objects (or "meanings") into a set Y of signals (or words).

The choice of the spaces X and Y and the class of functions will depend on the particular language evolution we are attempting to model.

Definition 1 A *linguistic setting* $\mathcal{L}$ is a triple $((X, \rho_X), Y, \mathcal{F})$ where

(1) X is a closed and bounded domain in $\mathbb{R}^n$ and ρ_X is a Borel probability measure on X. The pair (X, ρ_X) is called *space of objects* (or *meanings*).

(2) $Y \subset \mathbb{E}^l$ for some $l \geq 1$ is the *space of signals*, $\mathbb{E}^l$ is Euclidean space of dimension l.

(3) $\mathcal{F}$ is a set of continuous functions from X to Y.

Continuous functions $f : X \to Y$ are called *language-like functions*. We will often refer to a language-like function simply as a *language*.

The measure ρ_X will be interpreted as the relative frequency with which different objects occur in the context at hand.

Example 1 Consider the set of greys, i.e. different intensities of grey varying between white and black. We can model this set of meanings by taking $X = [0, 1]$. Here 0 corresponds to absolute white and 1 to absolute black. In normal speech, we associate absolute black with the word `black` and absolute white with the word `white`. Also, most of the grey tones are associated to the word `grey`. But some dark tones of grey are sometimes described as `grey` and sometimes as `black`, even by the same speaker. The frequency of, say, the former decreases with the darkness of the tone. A similar phenomenon happens with the light grey tones.

To model such a 3-word language we may take Y to be the interval $[0, 2]$ and associate "pure words" `white`, `grey`, and `black` to the points 0, 1 and 2, respectively. A language in this case is a non-decreasing continuous function $f : [0, 1] \to [0, 2]$ which maps $[0, u_1]$ to 0, $[u_2, u_3]$ to 1, and $[u_4, 1]$ to 2, for some $0 < u_1 < u_2 < u_3 < u_4 < 1$. Meanings in the intervals $[u_1, u_2]$ and $[u_3, u_4]$ are mapped to points in the intervals $[0, 1]$ and $[1, 2]$, respectively. For $x \in [u_1, u_2]$, one may interpret the value $f(x)$ as the proportion of times the language uses `white` for x (the value $1 - f(x)$ being the proportion of times it uses `grey`). A similar construction works for $x \in [u_3, u_4]$. In this interpretation one has the phenomenon in which different words may be used for a given meaning.

One can consider a version of the above dealing with colors. Let now X be an equilateral triangle whose vertices are associated with the primary colors blue, red,

and yellow. Any point in this triangle is a convex combination of these three colors and, as such, a definite color. The space Y is also a triangle. The words `blue`, `red`, and `yellow` are associated with the vertices of this triangle and other words are associated to other points in the triangle (e.g. `green` is associated to the midpoint of the side with vertices `blue` and `yellow`, `brown` is in the center of the triangle, etc.).

A language $f : X \to Y$, following the lines of the white-grey-black language above, induces a division of the spaces X and Y into regions such that some regions of X are mapped to a word (a single point in Y) and some others to a convex subset of Y corresponding to the phenomenon described above.

Additionally one can combine the grey and the color linguistic settings in 3 dimensional space to obtain such variations as shades of pink.

Example 2 An idealization (e.g. of William Wang *et al.* [Ke, Minett, Au, and Wang 2002]) in the study of language emergence assumes a situation in which members of a finite society associate utterances to objects. The number of both the objects and the utterances is finite, say r and l respectively. To model this situation one may take $X = \{x_1, \ldots, x_r\}$ and $Y \subset \mathbb{R}^l$ to be the set of l coordinate vectors in $\mathbb{R}^l$. Here l is a number of utterances and the ith utterance is $e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in Y$, the 1 in the ith place. Note that in this case the space Y is finite and therefore non-convex. A primitive language (simple vocabulary) is any $f : X \to Y$, which is of course continuous.

Example 3 We now modify Example 2 to incorporate a convex space of signals Y, to be interpreted as a space of vowel sounds following [De Boer 2001; Fant 1970; Stevens 1998].

The sound spectrum of a vowel sound is represented by a function from frequencies in Hertz (i.e., vibrations per second) to amplitudes. The frequencies corresponding to the peaks (i.e., local maxima of this real function of a real variable) in this graph, are called *resonance frequencies* or *formant frequencies* in acoustic phonetics.

Thus, to a vowel sound is associated an increasing sequence of real numbers, the formant frequencies $F_1, F_2, \ldots$. Following the above references, we limit ourselves to the first four, i.e., $0 < F_1 < F_2 < F_3 < F_4$. These four numbers provide a good idealization of a vowel sound. In other words, we take

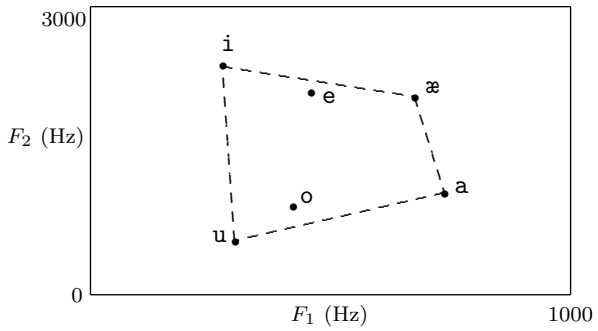
$$Y \subseteq \{(F_1, F_2, F_3, F_4) \in \mathbb{R}^4 \mid 0 < F_1 < F_2 < F_3 < F_4\}$$

a bounded convex closed set, defined by appropriate parameters as in the exposition in [Fant 1970; Stevens 1998]. The space Y is defined by putting bounds on the coordinates coming from physical limits of vocal chords and similar constraints. Thus, we may suppose that every vowel sound is represented by a point in Y and, conversely, each point in Y is *realized* by a vowel sound (i.e., there exist a vowel sound having these formant frequencies).

Of course this model for vowel sounds makes approximations and idealizations; yet four formant frequencies, even two or three, yield a good model. The following table (whose data is taken from [Stevens 1998, Chapter 6]) shows the first three formant frequencies, in Hertz, for the six basic vowels of American English produced by an average male speaker.

Vowel	i	e	æ	a	o	u
F_1	270	460	660	730	450	300
F_2	2290	1890	1720	1090	1050	870
F_3	3010	2670	2410	2440	2610	2240

In the literature, one sees often diagrams with just the first two formant frequencies F_1, F_2. The following is an example (corresponding to the data in the table above).



Now we model a situation where speakers with a primitive vocabulary associate to each object x_i, $i = 1, \ldots, r$, a sound represented by a point in Y characterized by the four real numbers F_1, F_2, F_3, F_4. The space of languages, all functions $f : X = \{x_1, \ldots, x_r\} \to Y$ has some justification from the above acoustic theory of speech, where now Y is convex, and the situation is ripe for the development of the Main Result.

Remark 1 Examples 1, 2, and 3 exhibit a difference between their spaces of signals. In Example 2 the space Y is non-convex while in Examples 1 and 3 it is convex. The space Y in Example 1 is convex but not all points are realizable as words. On the contrary, just a finite number of points are realized as words and the remaining ones are interpreted as probabilities. In Example 3, each point in Y is realizable as a "word" (i.e., a vowel sound). The convexity of the space of signals will be a main hypothesis in our development.

Consider two languages $f, g : X \to Y$. The *distance* between them is given by

$$d(f,g) = \left(\int_X \|f(x) - g(x)\|_Y^2 d\rho_X \right)^{1/2}$$

where $\|y\|_Y$ denotes the norm of y induced by the inner product in $\mathbb{E}^l$. Thus, $\|f(x) - g(x)\|_Y$ is the distance in Y between the signals at $x \in X$ for f and g, and $d(f, g)$ is the average (with respect to ρ_X, which weights objects in terms of occurrences in the environment) of these distances. The distance $d(f, g)$ could be interpreted in terms of the communication ability between two agents using f and g respectively. The smaller the distance, the larger the communication ability. So for an agent using language f, communication with an agent using language g is maximal at $g = f$ or $d(f, g) = 0$. But communication depends also on the "richness" of languages f and g.

Language-like functions belong to the space $\mathscr{L}^2_\rho(X) = \mathscr{L}^2_\rho(X; \mathbb{E}^l)$ of functions $f : X \to \mathbb{E}^l$ whose norm squared is integrable with respect to the measure $\rho = \rho_X$. This Hilbert space is a convenient conceptual framework. For instance, the distance defined above is just the distance in $\mathscr{L}^2_\rho(X)$.

Let $\mathscr{C}(X; Y)$ be the set of continuous functions from X to Y. This is a Banach space with the norm

$$\|f\|_\infty = \sup_{x \in X} \|f(x)\|_Y.$$

Recall that $\mathcal{F}$ is a subset of $\mathscr{C}(X; Y)$. Note that when Y is convex so is $\mathscr{C}(X; Y)$. In this case, we will require $\mathcal{F}$ to be convex as well.

We close this section by remarking that, in a given linguistic setting, information is transmitted in the form of object-signal pairs $(x, y) \in X \times Y$. Such pairs (or data) form the basis for learning a language.

3 Societies

Definition 2 By a *society* (or a *linguistic population*) $\mathcal{P}$ we understand a triple $(\{1, \ldots, k\}, \mathcal{L}, \Gamma)$ where $\mathcal{L}$ is a linguistic setting and Γ is a $k \times k$ matrix with non-negative entries called the *communication matrix* of the society. The k elements in $\{1, \ldots, k\}$ are the *agents* of the society.

The communication matrix helps to model how languages adjust by agents learning from one another. To do so one proceeds by taking into account how the different agents affect each other's language evolution. For each pair $(i, j) \in \{1, \ldots, k\}^2$, the (i, j) entry $\gamma_{ij} \geq 0$ of Γ measures the number of *linguistic encounters* between agents i and j. By "encounters" we mean non-symmetric, effective, encounters, so that γ_{ij} measures the influence of agent j in the development of the language of agent i.

The special case of the diagonal elements γ_{ii} may be interpreted as an *inertia* which could be expected to be small in the case of linguistic immaturity and large for an agent with full language development.

Note that the set of numbers $\{\gamma_{ij} \mid i, j \leq k\}$ defines a weighted oriented graph whose nodes are the agents $\{1, \ldots, k\}$ and having γ_{ij} as weight for the edge from j to i. Edges with weight 0 may be omitted. Thus, an edge can be seen as a data

transmission channel from j to i weighted by the influence that the transmitter has over i.

Example 4 Consider a society consisting of a mother M and a baby B. This is an instance of the problem of language acquisition. The assumption that the mother's language is not affected by the baby's is described by the equality $\gamma_{MB} = 0$ which yields a matrix Γ with the form

$$\begin{array}{c} \\ M \\ B \end{array} \begin{array}{c} \begin{array}{cc} M & B \end{array} \\ \left[\begin{array}{cc} 1 & 0 \\ 1-\theta & \theta \end{array} \right] \end{array}$$

where $\theta > 0$ is small.

Example 5 Consider now a society consisting of two groups of agents, say the inhabitants of two islands, having no direct or indirect contact with each other. If $I = \{1, \ldots, n_i\}$ and $J = \{n_i + 1, \ldots, n_i + n_j\}$ denote these two groups the matrix of linguistic encounters has the form

$$\left[\begin{array}{cc} \Gamma_I & 0 \\ 0 & \Gamma_J \end{array} \right]$$

where Γ_I and Γ_J are square matrices of dimension n_i and n_j respectively. Note that in this situation, one can not expect that all the agents will eventually use the same language. The matrix Γ is said to be "reducible" [Seneta 1973].

Remark 2 Throughout this paper we assume that $\sum_{j=1}^{k} \gamma_{ij} > 0$ for $i = 1, \ldots, k$. This excludes the possibility of a complete immature agent (one with inertia zero) which is not influenced by the rest of the agents.

A *state* of a linguistic society is a k-tuple $(f_1, \ldots, f_k)$, where $f_i : X \to Y$ is the language of the ith agent. The set $\mathcal{F}^k$ of all k-tuples formed by languages from $\mathcal{F}$ will be called *state space*. Recall that $\Delta_{\mathcal{F}} = \{(f, \ldots, f) \in \mathcal{F}^k\}$ is the set of the states in which all agents share a common language precisely.

4 A learning dynamic

Consider a society $\mathcal{P} = (\{1, \ldots, k\}, \mathcal{L}, \Gamma)$. During a given period beginning at time t with a state $f^{(t)} = (f_1^{(t)}, \ldots, f_k^{(t)})$, agents of this society communicate with each other and modify their language so that at the end of the period the state of the society has changed from $f^{(t)}$ to $f^{(t+1)}$. To do so, agent i modifies his language by a *learning algorithm*

$$\mathcal{S}_i^{(t)} \mapsto f_i^{(t+1)}$$

which computes language $f_i^{(t+1)}$ from a collection of data (or *sample*)

$$\mathcal{S}_i^{(t)} = \{(x_1^{(it)}, y_1^{(it)}), \ldots, (x_m^{(it)}, y_m^{(it)})\}$$

consisting of m object-signal pairs transmitted to i by the different agents in the society. This yields a map

$$\begin{array}{ccc} \mathcal{F}^k & \to & \mathcal{F}^k \\ f^{(t)} & \mapsto & f^{(t+1)} \end{array} \tag{A}$$

whose iteration defines a *learning dynamic*.

To complete the construction of the learning dynamic, one needs to specify a learning algorithm and the way the data $\mathcal{S}_i^{(t)}$ is sampled by agent i for $i = 1, \ldots, k$. One does not know much about the actual human mechanisms (i.e., algorithms) of learning. But one can exhibit a simple algorithm with which machines can be efficiently trained and which will be used in our mathematical development.

In the sequel we will assume

$$\text{The space } Y \text{ is convex and the set } \mathcal{F} \text{ is convex and compact} \tag{B}$$

together with a learning algorithm computing

$$f_i^{(t+1)} = \arg\min_{f \in \mathcal{F}} \sum_{(x,y) \in \mathcal{S}_i^{(t)}} (f(x) - y)^2. \tag{C}$$

The existence of $f_i^{(t+1)}$ is ensured by the compactness of $\mathcal{F}$. We remark, though, that $f_i^{(t+1)}$ may not be unique. This ambiguity is suppressed in the following and does not affect the results.

We will assume that the data $\mathcal{S}_i^{(t)}$ is sampled from a probability measure $\rho^{(it)}$ on $Z = X \times \mathbb{R}^l$ supported on $X \times Y$.

Pairs in $\mathcal{S}_i^{(t)}$ are transmitted to agent i from agents j such that $\gamma_{ij} \neq 0$. These pairs are produced by the transmitters; a pair (x, y) is produced by j by randomly selecting an object x from the space of objects X with the probability measure ρ_X and then associating to it $y = f_j^{(t)}(x)$. We will assume that $\rho^{(it)}$ satisfies the conditions (C), (D), and (E) below.

$$\text{The marginal measure of } \rho^{(it)} \text{ on } X \text{ is } \rho_X. \tag{D}$$

This is a natural condition which follows, for instance, from the assumption that all transmitters randomly select points from X according to ρ_X.

Define $F_i^{(t)} : X \to Y$ by

$$F_i^{(t)} = \sum_{j=1}^{k} \lambda_{ij} f_j^{(t)} \qquad \text{where} \qquad \lambda_{ij} = \frac{\gamma_{ij}}{\sum_{j=1}^{k} \gamma_{ij}}.$$

Now recall, the *regression function* of $\rho^{(it)}$ is the function defined by

$$x \mapsto \int_{\mathbb{R}^l} y \, d\rho^{(it)}(y|x)$$

where $\rho^{(it)}(y|x)$ is the conditional (with respect to x) probability measure induced by $\rho^{(it)}$ on Y. The second assumption on $\rho^{(it)}$ is the following.

$$\text{The regression function of } \rho^{(it)} \text{ equals } F_i^{(t)}. \tag{E}$$

This is again a natural condition which follows from the assumption that the transmitter is randomly selected and that the probability of the event "a pair is transmitted by agent j" is λ_{ij}. Indeed, fix $x \in X$ and take a random pair from $X \times Y$ with first coordinate x. Since the probability that this pair was transmitted by agent j is λ_{ij}, the expected value of the second coordinate is

$$\sum_{j=1}^{k} \lambda_{ij} f_j^{(t)}(x).$$

The third condition on $\rho^{(it)}$ is the following.

$$\text{There exists } M \in \mathbb{R} \text{ such that } \max_{\substack{(x,y)\in X\times Y \\ f\in\mathcal{F}}} \|f(x) - y\|_Y \leq M. \tag{F}$$

It is sufficient to assume that $\rho^{(it)}$ has compact support (uniform in t) since $\sup_{\substack{f\in\mathcal{F} \\ x\in X}} \|f(x)\|_Y < \infty$ by the compactness of X and $\mathcal{F}$.

Remark 3 **(i)** The way $F_i^{(t)}$ is defined makes use of both scalar multiplication and addition of languages. These are only formal constructs; they do not have any linguistic interpretation. The linear combination defining $F_i^{(t)}$, however, has a further structure: $F_i^{(t)}$ is a *convex combination* of the languages $f_1^{(t)}, \ldots, f_k^{(t)}$. Therefore, if $\mathcal{F}$ is convex, $F_i^{(t)} \in \mathcal{F}$ for $i = 1, \ldots, k$. This is not true for arbitrary linear combinations of elements in $\mathcal{F}$. These sums are only guaranteed to be in $\mathscr{L}^2_\rho(X)$ showing the contrast between the linguistic nature of $\mathcal{F}$ in our model and the purely formal one of $\mathscr{L}^2_\rho(X)$.

(ii) While conditions (D) and (E) follow from the assumptions that all transmitters randomly select points from x according to ρ_X and that the probability that a pair is transmitted by agent j is λ_{ij}, respectively, one may model with (D–E) more general situations than those satisfying these assumptions. In particular, one may model situations in which noise occurs. This noise may be introduced by the transmitter, the receiver, or present on the communication channel. An

example would be a situation in which the receiver gets pairs $(x, f_j^{(t)}(x) + \varepsilon_j)$ from j with ε_j a random variable centered at 0, instead of noise-free pairs $(x, f_j^{(t)}(x))$. In this case, condition (i) follows as above and condition (ii) follows from the fact that

$$\int_Y y\, d\rho^{(it)}(y|x) = \sum_{j=1}^{k} \lambda_{ij} \int_Y (f_j(x)+\varepsilon_j(y)) = \sum_{j=1}^{k} \lambda_{ij} f_j^{(t)}(x)+\int_Y \varepsilon_j(y) = F_i^{(t)}(x)$$

since the ε_j are centered at 0. But we note that other, more general forms of noise, can lead to condition (E) as well.

(ii) Note that the learning dynamic in (A) is not a simple discrete dynamical system since the map it iterates depends on the random samples $\{(x_1^{(it)}, y_1^{(it)}), \ldots, (x_m^{(it)}, y_m^{(it)})\}$, which vary with t. One may say that it is a stochastic dynamic.

(iii) We remark here that the matrix Γ remains unchanged during the dynamic.

Our definition of language-like function is sufficiently abstract that one may model situations which are not necessarily in the realm of human languages. Our last example, inspired by economic equilibrium theory, exhibits one such case.

Example 6 Consider ℓ commodities $1, \ldots, \ell$ and let $X = [0, K]^\ell$ (here $K > 0$ is a large enough constant) be the space of commodity bundles. We assume all commodities are divisible so that a point $x \in X$ represents a bundle of quantities x_r, $0 \leq x_r \leq K$, of commodity r for each $r = 1, \ldots, \ell$. A *price system* is a map $p : X \to \mathbb{R}_+$ where $p(x)$ is the price of the bundle of goods $(x_1, \ldots, x_\ell)$. The simplest example of price system is given by associating price $p[r] \geq 0$ to r for $r = 1, \ldots, \ell$, interpreted as price $p[r]$ for one unit of r, and $p(x) = \sum_{r=1}^{\ell} x_r p[r]$.

A state $(p_1, \ldots, p_k) \in \mathcal{F}^k$ will have this interpretation. Agent i has the belief or perception that the economy is operating under price system p_i. At a time t, an agent j with his belief in prices p_j will make offers of the type 'buy' or 'sell' a bundle of goods x at price $p_j(x)$. These offers form the sample transmitted to agent i, regulated by λ_{ij}. After receiving this sample, agent i will form a new belief in a price system. The set of these new beliefs for $i = 1, \ldots, k$ will give the state at time $t + 1$ of the learning dynamic.

The Main Result in this paper has an interpretation which gives conditions for the sequence of states to converge to a common belief in a price system. However, there are great limitations to this picture because it takes into account neither the resources nor the preferences of the agents which could play a role in their offers to buy and sell. One might see [Smale 1981] for some background —economic equilibrium theory.

5 Statement of the Main Result

For $f_i \in \mathcal{F}$, $i = 1, \ldots, k$, write

$$F_i = \sum_{j=1}^{k} \lambda_{ij} f_j.$$

Denoting by Λ the matrix

$$\Lambda := \left(\lambda_{ij}\right)_{i,j=1}^{k}$$

and writing $F = (F_1, \ldots, F_k)$ and $f = (f_1, \ldots, f_k)$ we have that $F = \Lambda f$. We will call Λ the *normalized communication matrix* of the society. It is an example of stochastic matrix.

A $k \times k$ matrix $\Lambda = \left(\lambda_{ij}\right)_{i,j=1}^{k}$ is said to be a *stochastic matrix* (or a *Markov matrix*) if $\lambda_{ij} \geq 0$ for all i, j and

$$\sum_{j=1}^{k} \lambda_{ij} = 1, \qquad \forall i = 1, \ldots, k.$$

A main result in the theory of stochastic matrices (cf. [Seneta 1973]) is the following.

Proposition 1 (Perron-Frobenius) *A stochastic matrix Λ has the eigenvalue 1 with the eigenvector $(1, \ldots, 1)$. All its other eigenvalues are not more than 1 in modulus.* □

A stochastic matrix is said to be *weakly irreducible* if 1 is a simple eigenvalue and all its other eigenvalues are less than 1 in modulus. A non-negative square matrix is *weakly irreducible* when its normalization (so that the sum of the elements in each row is 1) is weakly irreducible. For instance, the matrix Λ resulting from normalizing the matrix Γ in Example 4 is weakly irreducible (but not irreducible in the sense of [Seneta 1973]) while that in Example 5 is not weakly irreducible.

If X is a metric space and $\varepsilon > 0$, the *covering number* $\mathcal{N}(X, \varepsilon)$ is defined as the smallest $\ell \in \mathbb{N}$ such that there exist ℓ disks of radius ε covering X.

Also, note that $\mathscr{L}^2_\rho(X)$ induces a metric in $\mathcal{F}^k$ by taking the k-fold product of the distances introduced in Section 2

$$d(f, g) = \|f - g\|_{(\mathscr{L}^2_\rho(X))^k} = \left(\sum_{r=1}^{k} \|f_r - g_r\|^2_{\mathscr{L}^2_\rho(X)} \right)^{1/2}.$$

Using this metric the distance from a state f to the diagonal $\Delta_{\mathcal{F}}$ is defined

$$d(f, \Delta_{\mathcal{F}}) = \inf_{g \in \Delta_{\mathcal{F}}} d(f, g).$$

Theorem 1 *Let $\mathcal{P} = (\{1, \ldots, k\}, \mathcal{L}, \Gamma)$ be a society with a weakly irreducible communication matrix. Let $f^{(t)}$ denote the states obtained by the learning dynamic given by (A), under assumption (B), learning algorithm (C), with initial state $f^{(0)}$ and such that each agent receives, at each iteration, m object-signal pairs sampled from a distribution satisfying conditions (D–F) above. There exist constants $\alpha_* < 1$ and $\mathbf{C} > 0$ (depending on Λ) such that for each $0 < \delta < 1$, and $t \in \mathbb{N}$, if $m \geq m_t$ then*

$$d(f^{(t)}, \Delta_{\mathcal{F}}) \leq \mathbf{C}\alpha_*^t d(f^{(0)}, \Delta_{\mathcal{F}})$$

with confidence at least $1 - \delta$.

One may take

$$m_t = \frac{288kM^2}{(1-\alpha_*)^2\alpha_*^{2t}d(f^{(0)}, \Delta_{\mathcal{F}})^2} \ln\left(tk\mathcal{N}\left(\mathcal{F}, \frac{(1-\alpha_*)^2\alpha_*^{2t}d(f^{(0)}, \Delta_{\mathcal{F}})^2}{24kM}\right) + \ln\left(\frac{1}{\delta}\right)\right).$$

Remark 4 **(i)** A common feature in the main classes $\mathcal{F}$ considered in learning theory (cf. §7.1 below) is the inequality, for some constants $C_{\mathcal{F}}, a > 0$ and all $\varepsilon > 0$,

$$\ln \mathcal{N}(\mathcal{F}, \varepsilon) \leq C_{\mathcal{F}}\left(\frac{1}{\varepsilon}\right)^a. \tag{1}$$

In this case the expression for m_t takes the form

$$m_t = \frac{288kM^2}{(1-\alpha_*)^2\alpha_*^{2t}d(f^{(0)}, \Delta_{\mathcal{F}})^2} \left(\ln tk + C_{\mathcal{F}}\left(\frac{24kM}{(1-\alpha_*)^2\alpha_*^{2t}d(f^{(0)}, \Delta_{\mathcal{F}})^2}\right)^a + \ln\left(\frac{1}{\delta}\right)\right).$$

The constant $C_{\mathcal{F}}$ is independent of ε.

In the rest of the paper we assume that (1) holds.

(ii) Note that, as $t \to \infty$, $m_t \to \infty$. Also, one may replace the condition on m by solving for δ and obtain that

$$\delta \leq tk\, e^{-\frac{m(1-\alpha_*)^2\alpha_*^{2t}d(f^{(0)}, \Delta_{\mathcal{F}})^2}{288kM^2} + C_{\mathcal{F}}\left(\frac{24kM}{(1-\alpha_*)^2\alpha_*^{2t}d(f^{(0)}, \Delta_{\mathcal{F}})^2}\right)^a}.$$

In this form one sees that to have $\delta < 1$ one needs m sufficiently large.

(iii) From the compactness of X and $\mathcal{F}$ it follows that $\operatorname{diam}(\mathcal{F}) = \sup_{f,g\in\mathcal{F}} d(f, g)$ is finite. Without loss of generality we may assume that $\operatorname{diam}(\mathcal{F}) \leq M$. Therefore, $\operatorname{diam}(\mathcal{F}^k) \leq \sqrt{k}M$ and $d(f, \Delta_{\mathcal{F}}) \leq \frac{\sqrt{k}M}{2}$. In particular, we may (crudely) replace $d(f^{(0)}, \Delta_{\mathcal{F}})$ by $\frac{\sqrt{k}M}{2}$ in the expressions for m_t and d.

(iv) Note that constants in Theorem 1 depend on Λ as specified and will be exhibited in the proof.

By appropriately choosing δ as a function of t we can show that, when $t \to \infty$, $f^{(t)}$ tends to $\Delta_{\mathcal{F}}$ almost surely. Let $f^{(t)}_{[m]}$ denote the state after t steps when the number of examples at step t in the dynamic is m.

In the following corollaries, of course, the setting and conditions are as in Theorem 1 (with the addition of Remark 4(i)).

Corollary 1 *Let*

$$m(t) = \frac{288kM^2}{(1-\alpha_*)^2\alpha_*^{2t}d(f^{(0)},\Delta_{\mathcal{F}})^2}\left(\ln t^2k + C_{\mathcal{F}}\left(\frac{24kM}{(1-\alpha_*)^2\alpha_*^{2t}d(f^{(0)},\Delta_{\mathcal{F}})^2}\right)^a\right).$$

Then

$$\sup_{\varepsilon>0}\lim_{t\to\infty}\text{Prob}\left\{d\left(f^{(t)}_{[m(t)]},\Delta_{\mathcal{F}}\right)\le\varepsilon\right\}=1.$$

PROOF. Consider $\varepsilon > 0$. For all t sufficiently large, $\mathbf{C}\alpha_*^t d(f^{(0)},\Delta_{\mathcal{F}}) < \varepsilon$. Therefore, taking $\delta = \frac{1}{t}$, for all such t,

$$\text{Prob}\left\{d\left(f^{(t)}_{[m(t)]},\Delta_{\mathcal{F}}\right)\le\varepsilon\right\}\ge 1-\frac{1}{t}.$$

□

Corollary 2 *Let $0 < \delta < 1$ and $\tau > 0$. If $m \ge m_T$ and T is the first integer greater than or equal to*

$$\frac{\ln(\mathbf{C}d(f^{(0)},\Delta_{\mathcal{F}})) - \ln\tau}{|\ln\alpha_*|}$$

then $f^{(T)} \in N(\Delta_{\mathcal{F}},\tau)$ with confidence $1-\delta$. One may take

$$m_T = \frac{288kM^2\mathbf{C}^2}{(1-\alpha_*)^2\tau^2}\left(\ln Tk + C_{\mathcal{F}}\left(\frac{24kM\mathbf{C}^2}{(1-\alpha_*)^2\tau^2}\right)^a + \ln\left(\frac{1}{\delta}\right)\right).$$

The condition for m can be replaced by one on δ to obtain

$$\delta \le Tk\, e^{-\frac{m(1-\alpha_*)^2\tau^2}{288kM^2\mathbf{C}^2}+C_{\mathcal{F}}\left(\frac{24kM\mathbf{C}^2}{(1-\alpha_*)^2\tau^2}\right)^a}.$$

□

Remark 5 (i) The use of the word "language" in the expressions "the English language" and "John's language" is not the same. Saussure [1983] used the words *langue* and *parole* to distinguish between them. In English, one may

use the term *language-system* to denote the former.[1] We have already used the term "language" to denote the latter and made its definition a key piece in our model. We now give a definition for the Saussurean 'langue.'

We say that a society $\mathcal{P} = (\{1, \ldots, k\}, \mathcal{L}, \Gamma)$ *shares a language-system* when $d((f_1, \ldots, f_k), \Delta_{\mathcal{F}}) \leq \tau$. Here τ is a constant, which may depend on the modelled situation. Agents of a linguistic society sharing a language-system are said to speak an *idiolect* of it.[2]

One might interpret Corollary 2 as giving the time for the emergence of a language-system as quantified by τ in terms especially of the sample size m. We emphasize again that, to have substance, in Corollary 2, m needs to be large enough to ensure $\delta < 1$.

(ii) Note that in Corollary 2 we have supposed that $m_t \to \infty$ as $t \to \infty$. Suppose on the contrary that $m \leq \mathbf{M}$, with $\mathbf{M}$ independent of t (as might follow from the reasonable assumption that in one unit of time the sample size can not become arbitrarily large). In this case, as t increases, the confidence $1-\delta$ decreases and reaches negative values as can be seen from Remark 4(ii). As a consequence the dynamic as t tends to ∞ has the property that any state will eventually move outside a τ-neighborhood of $\Delta_{\mathcal{F}}$. This is a mathematical statement which has an interpretation: a language-system eventually disintegrates. Thus, with the bounded m hypothesis, only a finite t will maintain a shared language-system. That t may be read off from Corollary 2.

One may modify the linguistic model to incorporate the ages of the agents into the states, birth and death. In this setting, with the introduction of immortal agents (e.g. Shakespeare), one may circumvent this property.

6 Proof of Theorem 1

The proof of Theorem 1 relies on tools from two different subjects: stochastic matrices and learning theory.

[1] What Saussure called a 'langue' is any particular language that is the common possession of all the members of a given language community (i.e. of all those who are acknowledged to speak the same language). [...] We will introduce the term language-system in place of it. [...] A *language-system* is a social phenomenon, or institution, which of itself is purely abstract, in that it has no physical existence, but which is actualized on particular occasions in the *language-behaviour* of individual members of the language community. [Lyons 1981, page 10].

[2] "In the last resort, we should have to admit that everyone has his own individual dialect: that he has his own *idiolect*, as linguists put it. Every idiolect will differ from every other, certainly in vocabulary and pronunciation and perhaps also, to a smaller degree, in grammar." [Lyons 1981, pages 26–27].

6.1 Convergence of iterated stochastic matrices

Though the dynamic induced by a weakly irreducible Λ need not be contracting to the diagonal under the Euclidean norm, it is under a modified norm.

Assume Λ is a weakly irreducible stochastic matrix. Let $\mathbf{e} = (1, \ldots, 1)$. Then, $\Lambda\mathbf{e} = \mathbf{e}$ i.e., $\mathbf{e}$ is an eigenvector of Λ with eigenvalue 1. The eigenspace of 1 is thus the diagonal Δ_k in $\mathbb{R}^k$. Let $W \subseteq \mathbb{R}^k$ be the eigenspace corresponding to the remaining eigenvalues of Λ. Also, let V' be an eigenvector of Λ^{T} associated with the eigenvalue 1. Finally, for a point $v \in \mathbb{R}^k$, let $\mathrm{Diag}(v) = (v, \ldots, v) \in \Delta_k$.

Lemma 1 *Let Λ be a stochastic weakly irreducible $k \times k$ matrix. Then*

(i) *The space $\mathbb{R}^k$ decomposes into the direct sum $\mathbb{R}^k = \Delta_k \oplus W$.*

(ii) **(a)** $W = \{w \in \mathbb{R}^k \mid (V')^{\mathrm{T}} w = 0\}$.

(b) $\sum_{i=1}^k V_i' \neq 0$.

(c) *Let $V = \frac{V'}{\sum V'i}$. Then for every $v \in \mathbb{R}^k$, $v = \mathrm{Diag}(V^{\mathrm{T}}v) + (v - \mathrm{Diag}(V^{\mathrm{T}}v))$ with $(v - \mathrm{Diag}(V^{\mathrm{T}}v)) \in W$.*

(iii) *There is a norm $\| \cdot \|_\Lambda$ on $\mathbb{R}^k$ and a number $0 < \alpha_* < 1$ such that, for all $v \in \Delta_k$ and $w \in W$,*

$$d_\Lambda(w, \Delta_k) = \|\Lambda w\|_\Lambda \leq \alpha_* \|w\|_\Lambda,$$

and

$$d_\Lambda(v + w, \Delta_k) = d_\Lambda(v, \Delta_k).$$

Proof. Part (i) is well-known. To prove part (ii) we note that the proof of [Smale 1963, Theorem 3.1], for instance, yields a constant $\alpha_* < 1$ and an inner product $\langle\ ,\ \rangle_\Lambda$ in W satisfying $\|\Lambda w\|_\Lambda \leq \alpha_* \|w\|_\Lambda$ for all $w \in W$. For $w \in W$,

$$(V')^{\mathrm{T}} \Lambda w = (\Lambda^{\mathrm{T}} V')^{\mathrm{T}} w = (V')^{\mathrm{T}} w.$$

Since $V' \neq 0$ and Λ is contractive on W this implies that $(V')^{\mathrm{T}} w = 0$. This proves that $W \subseteq \{w \in \mathbb{R}^k \mid (V')^{\mathrm{T}} w = 0\}$. But since both spaces have dimension $k - 1$ they must coincide. This proves (ii.a).

Since $\mathbf{e} \notin W$ we have that $(V')^{\mathrm{T}} \mathbf{e} \neq 0$. This proves (ii.b). Part (ii.c) is now immediate.

For part (iii) take the extension to $\mathbb{R}^k$ of the inner product in part (ii.a) which makes $\mathbf{e}$ orthogonal to W and satisfies $\|\mathbf{e}\|_\Lambda = 1$. Let $\| \ \|_\Lambda$ be the norm induced by this inner product. We noted in the proof of (ii.a) that $\| \ \|_\Lambda$ is contractive on W thus the first statement in (iii). The second statement follows from the orthogonality of $\mathbf{e}$ and W. □

Remark 6 Let $\alpha_1 = 1, \alpha_2, \ldots, \alpha_k$ be the eigenvalues of Λ with algebraic multiplicity. It follows from the mentioned proof of [Smale 1963, Theorem 3.1] that any $\alpha > \max_{i=2,\ldots,k} |\alpha_i|$ can be taken as α_* and that, if all the eigenvalues are non-degenerate, then one may take $\alpha_* = \max_{i=2,\ldots,k} |\alpha_i|$.

We remarked in Section 5 that $F = \Lambda f$. This defines an action of Λ on $\mathcal{F}^k$ which should not be confused with the action of Λ on $\mathbb{R}^k$ that was the object of Lemma 1. Yet, we would like to extend this lemma to the action of Λ on $\mathcal{F}^k$. To do so, we first consider the extension of this action to $(\mathbb{R}^l)^k$. This extension is given by defining, for $\vec{v} = (\vec{v}_1, \ldots, \vec{v}_k) \in (\mathbb{R}^l)^k$ and $i = 1, \ldots, k$,

$$(\Lambda \vec{v})_i = \sum_{j=1}^{k} \lambda_{ij} \vec{v}_j.$$

Now extend the definitions of Δ_k and W by letting

$$\Delta_{lk} = \left\{ (\vec{v}, \ldots, \vec{v}) \in (\mathbb{R}^l)^k \mid \vec{v} \in \mathbb{R}^l \right\}$$

and

$$W_{lk} = \left\{ \vec{w} \in (\mathbb{R}^l)^k \mid \sum_{i=1}^{k} V_i \vec{w}_i = 0 \right\}$$

and it is easy to check that the following extension of Lemma 1 holds. In what follows, if L is a vector space over $\mathbb{R}$ and $v \in L^k$ we denote by $V^{\mathrm{T}} v$ the element of L given by $\sum_{i=1}^{k} V_i v_i$. Note that, since $\Lambda^{\mathrm{T}} V = V$,

$$V^{\mathrm{T}}(\Lambda v) = V^{\mathrm{T}} v \qquad \text{for all } v \in L^k.$$

Lemma 2 *Let Λ be a stochastic weakly irreducible $k \times k$ matrix. Then*

(i) *The space $(\mathbb{R}^l)^k$ decomposes in the direct sum $\mathbb{R}^k = \Delta_{kl} \oplus W_{kl}$.*

(ii) *The space W_{kl} is invariant under the action of Λ. For every $\vec{v} \in (\mathbb{R}^l)^k$, $\vec{v} = \mathrm{Diag}(V^{\mathrm{T}} \vec{v}) + (\vec{v} - \mathrm{Diag}(V^{\mathrm{T}} \vec{v}))$ with $(\vec{v} - \mathrm{Diag}(V^{\mathrm{T}} \vec{v})) \in W_{kl}$.*

(iii) *There is a norm $\| \cdot \|_\Lambda$ on $(\mathbb{R}^l)^k$ and a number $0 < \alpha_* < 1$ such that, for all $\vec{v} \in \Delta_{kl}$ and $\vec{w} \in W_{kl}$,*

$$d_\Lambda(\vec{w}, \Delta_{kl}) = \|\Lambda \vec{w}\|_\Lambda \leq \alpha_* \|\vec{w}\|_\Lambda,$$

and

$$d_\Lambda(\vec{v} + \vec{w}, \Delta_{kl}) = d_\Lambda(\vec{v}, \Delta_{kl}).$$

PROOF. Parts (i) and (ii) are proved as in Lemma 1. Part (iii) follows by letting, for $\vec{w} = (\vec{w}_1, \ldots, \vec{w}_k) \in (\mathbb{R}^l)^k$,

$$\|\vec{w}\|_\Lambda^2 = \sum_{i=1}^{l} \|(\vec{w}_1)_i, \ldots, (\vec{w}_k)_i\|_\Lambda^2$$

where $(\vec{w}_j)_i$ denotes the ith component of $\vec{w}_j \in \mathbb{R}^l$. □

To further extend the above to $\mathcal{F}$ we consider the linear closure $\overline{\mathcal{F}}$ of $\mathcal{F}$ and the extended action of Λ on $\overline{\mathcal{F}}^k$. Also, we let

$$\Delta_{\overline{\mathcal{F}}} = \left\{(f, \ldots, f) \in (\overline{\mathcal{F}})^k \mid f \in \overline{\mathcal{F}}\right\}$$

and

$$W_{\overline{\mathcal{F}}} = \left\{ g \in \overline{\mathcal{F}}^k \mid \sum_{i=1}^{k} V_i g_i = 0 \right\}$$

and, Lemma 2 now extends as follows.

Lemma 3 *Let Λ be a stochastic weakly irreducible $k \times k$ matrix. Then*

(i) *The space $\overline{\mathcal{F}}^k$ decomposes in the direct sum $\overline{\mathcal{F}}^k = \Delta_{\overline{\mathcal{F}}} \oplus W_{\overline{\mathcal{F}}}$.*

(ii) *The space $W_{\overline{\mathcal{F}}}$ is invariant under the action of Λ. For every $f \in \overline{\mathcal{F}}^k$, $f = \mathrm{Diag}(V^{\mathrm{T}} f) + (f - \mathrm{Diag}(V^{\mathrm{T}} f))$ with $(f - \mathrm{Diag}(V^{\mathrm{T}} f)) \in W_{\overline{\mathcal{F}}}$.*

(iii) *There is a norm $\|\cdot\|_\Lambda$ on $\overline{\mathcal{F}}^k$ and a number $0 < \alpha_* < 1$ such that, for all $f \in \Delta_{\overline{\mathcal{F}}}$ and $g \in W_{\overline{\mathcal{F}}}$,*

$$d_\Lambda(g, \Delta_{\overline{\mathcal{F}}}) = \|\Lambda g\|_\Lambda \leq \alpha_* \|g\|_\Lambda,$$

and

$$d_\Lambda(f + g, \Delta_{\overline{\mathcal{F}}}) = d_\Lambda(g, \Delta_{\overline{\mathcal{F}}}).$$

PROOF. Again, parts (i) and (ii) are proved as in Lemma 1. Part (iii) follows by letting, for $f = (f_1, \ldots, f_k) \in \overline{\mathcal{F}}^k$,

$$\|f\|_\Lambda^2 = \int_X \|(f_1(x), \ldots, f_k(x))\|_\Lambda^2 \, d\rho_X$$

where the norm in the right-hand side is that of Lemma 2. □

Consider the dynamic obtained by iterating the map $\mathcal{T} : \mathcal{F}^k \to \mathcal{F}^k$ defined by $\mathcal{T}(f) = \Lambda f$, $f = (f_1, \ldots, f_k) \in \mathcal{F}^k$. This may be seen as an "ideal" version of the learning dynamic. Our next result asserts this ideal dynamic is contractive with respect to the diagonal $\Delta_{\mathcal{F}}$ for the distance d_Λ.

In what follows we denote $f_\Delta = \mathrm{Diag}(V^{\mathrm{T}} f)$.

Theorem 2 *Let $f \in \overline{\mathcal{F}}^k$, $f = f_\Delta + f_W \in \Delta_{\overline{\mathcal{F}}} \oplus W_{\overline{\mathcal{F}}}$. Then,*

$$d_\Lambda(\mathcal{T}f, f_\Delta) \leq \alpha_* d_\Lambda(f, f_\Delta).$$

PROOF.

$$d_\Lambda(\mathcal{T}f, f_\Delta) = d_\Lambda(f_\Delta + \Lambda f_W, f_\Delta) = \|\Lambda f_W\|_\Lambda \leq \alpha_* \|f_W\|_\Lambda = \alpha_* d_\Lambda(f, f_\Delta).$$

□

Corollary 3 *For all $f \in \mathcal{F}^k$,*

$$d_\Lambda(\mathcal{T}f, \Delta_\mathcal{F}) \leq \alpha_* d_\Lambda(f, \Delta_\mathcal{F})$$

and $\lim_{t\to\infty} \mathcal{T}^t f = f_\Delta = \text{Diag}\left(\sum_{i=1}^k V_i f_i\right)$. □

6.2 Learning theory

The sampling and subsequent learning algorithm described in Section 4 are an instance of the general situation in learning theory. We next briefly describe this general situation (following [Cucker and Smale 2002]).

Let X and Y be as in Section 2 and $Z = X \times \mathbb{R}^l$. Consider a probability measure ρ in Z. For a function $f : X \to \mathbb{R}^l$ we define its *error* (with respect to ρ) by

$$\mathcal{E}(f) := \int_Z \|f(x) - y\|_Y^2 \, d\rho. \tag{2}$$

It is known (see e.g. Proposition 1, Chapter I of [Cucker and Smale 2002]) that, for any function f,

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) + \int_X \|f(x) - f_\rho(x)\|_Y^2 \, d\rho_X. \tag{3}$$

The first term in the right-hand side, in the sequel denoted by σ_ρ^2, is non-negative and independent of f. Therefore, the regression function f_ρ minimizes the error $\mathcal{E}$.

Recall that $\mathscr{C}(X;Y)$ is the Banach space of all continuous functions from X to Y with the norm $\|f\| = \sup_{x\in X}\{\|f(x)\|_Y\}$ and $\mathcal{F} \subset \mathscr{C}(X;Y)$ is a convex and compact subset (which in learning theory is called *hypothesis space*). We may search, among the functions $f \in \mathcal{F}$ the function $f_\mathcal{F}$, called the *target function*, that minimizes the error $\mathcal{E}$,

$$f_\mathcal{F} := \arg\min_{f\in\mathcal{F}} \mathcal{E}(f).$$

A key remark concerning the measure ρ is that this measure is not assumed to be known. Therefore, the minimization problem defining $f_\mathcal{F}$ is not fully explicit and $f_\mathcal{F}$ is not computable. What can be done instead, is to consider a *sample*

$\mathbf{z} = ((x_1, y_y), \ldots, (x_m, y_m))$ of m examples independently drawn from Z according to ρ and to minimize the *empirical error*

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{r=1}^{m} \|f(x_r) - y_r\|_Y^2.$$

The function $f_{\mathbf{z}}$ that minimizes the empirical error in $\mathcal{F}$ is called the *empirical target function*

$$f_{\mathbf{z}} := \arg \min_{f \in \mathcal{F}} \mathcal{E}_{\mathbf{z}}(f). \tag{4}$$

The empirical target function is computable. It approximates the target function well when the number m of samples is large enough (see e.g. [Cucker and Smale 2002; Haussler 1992; Niyogi 1998; Vapnik 1998]). For our purposes we next state Theorem C* of [Cucker and Smale 2002].

Proposition 2 ([Cucker and Smale 2002]) *Let $\mathcal{F}$ be a compact and convex subset of $\mathscr{C}(X;Y)$ and ρ a probability measure on $Z = X \times \mathbb{R}^l$ with support on $X \times Y$. Assume that for all $f \in \mathcal{F}$, $\|f(x) - y\|_Y \leq M$ almost everywhere. Then, for all $\eta > 0$,*

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m} \left\{ \int_X \|f_{\mathbf{z}}(x) - f_{\mathcal{F}}(x)\|_Y^2 \, d\rho_X \leq \eta \right\} \geq 1 - \mathcal{N}\left(\mathcal{F}, \frac{\eta}{24M}\right) e^{-\frac{m\eta}{288M^2}}.$$

□

The individual steps in the learning dynamic described in Section 4 can be seen within the general framework of learning theory. At time t, agent i is exposed to a sample $\mathbf{z} = ((x_1, y_y), \ldots, (x_m, y_m))$ drawn from Z according to $\rho^{(it)}$. Assumption (E) in Section 4 asserts that the regression function $f_{\rho^{(it)}}$ is precisely $F_i^{(t)}$. We note, however, that $f_i^{(t)} \in \mathcal{F}$, for $i = 1, \ldots, k$ and that $\mathcal{F}$ is convex. Thus, $F_i^{(t)} \in \mathcal{F}$ and therefore $f_{\mathcal{F}} = f_{\rho^{(it)}}$. In other words, regression and target functions coincide.

6.3 Proof of Theorem 1

Recall that $F^{(t)} = \mathcal{T} f^{(t)}$. By the triangle inequality,

$$d_\Lambda(f^{(t)}, \Delta_{\mathcal{F}}) \leq d_\Lambda(f^{(t)}, F^{(t)}) + d_\Lambda(F^{(t)}, \Delta_{\mathcal{F}}).$$

Corollary 3 yields the contractivity of the second term, i.e.,

$$d_\Lambda(F^{(t)}, \Delta_{\mathcal{F}}) = d_\Lambda(\mathcal{T} f^{(t)}, \Delta_{\mathcal{F}}) \leq \alpha_* d_\Lambda(f^{(t-1)}, \Delta_{\mathcal{F}}).$$

So we only need to estimate the first term.

We now use Proposition 2 to deduce that, for each $\eta > 0$ and each $i = 1, \ldots, k$,

$$\operatorname*{Prob}_{\mathbf{z}_i^{(t)} \in Z^m} \left\{ \int_X \left\| f_i^{(t)}(x) - F_i^{(t)}(x) \right\|_Y^2 \, d\rho_X \leq \eta \right\} \geq 1 - \mathcal{N}\left(\mathcal{F}, \frac{\eta}{24M}\right) e^{-\frac{m\eta}{288M^2}}.$$

Then with confidence at least $1 - k\mathcal{N}\left(\mathcal{F}, \frac{\eta}{24M}\right) e^{-\frac{m\eta}{288M^2}}$, there holds

$$\left\| f^{(t)} - F^{(t)} \right\|_{(\mathscr{L}^2_\rho(X))^k}^2 \leq k\eta.$$

We need to compare the metrics d_Λ and d. Since the norm $\| \ \|_\Lambda$ on $\mathbb{R}^k$ is equivalent to the Euclidean norm $\| \ \|$, there exist two positive constants C_Λ and C'_Λ such that

$$C'_\Lambda \|v\| \leq \|v\|_\Lambda \leq C_\Lambda \|v\|, \qquad \forall v \in \mathbb{R}^k.$$

It follows that for any $f, g \in \mathcal{F}^k$,

$$C'_\Lambda d(f,g) \leq d_\Lambda(f,g) \leq C_\Lambda d(f,g). \tag{5}$$

This, together with the estimate for $\|f^{(t)} - F^{(t)}\|_{(\mathscr{L}^2_\rho(X))^k}$, implies

$$d_\Lambda(f^{(t)}, F^{(t)}) \leq C_\Lambda \|f^{(t)} - F^{(t)}\|_{(\mathscr{L}^2_\rho(X))^k} \leq C_\Lambda \sqrt{k\eta}.$$

Therefore, with confidence at least $1 - k\mathcal{N}\left(\mathcal{F}, \frac{\eta}{24M}\right) e^{-\frac{m\eta}{288M^2}}$, we have

$$d_\Lambda(f^{(t)}, \Delta_\mathcal{F}) \leq C_\Lambda \sqrt{k\eta} + \alpha_* d_\Lambda(f^{(t-1)}, \Delta_\mathcal{F}). \tag{6}$$

Combining the estimates for $t, t-1, \ldots, 1$, we deduce that, with confidence at least $1 - tk\mathcal{N}\left(\mathcal{F}, \frac{\eta}{24M}\right) e^{-\frac{m\eta}{288M^2}}$,

$$\begin{aligned} d_\Lambda(f^{(t)}, \Delta_\mathcal{F}) &\leq C_\Lambda \sqrt{k\eta} \left(1 + \alpha_* + \ldots + \alpha_*^{t-1}\right) + \alpha_*^t d_\Lambda(f^{(0)}, \Delta_\mathcal{F}) \\ &\leq \frac{C_\Lambda}{1 - \alpha_*} \sqrt{k\eta} + \alpha_*^t d_\Lambda(f^{(0)}, \Delta_\mathcal{F}). \end{aligned}$$

Thus, for

$$m \geq \frac{288M^2}{\eta} \left(\ln \frac{tk\mathcal{N}\left(\mathcal{F}, \frac{\eta}{24M}\right)}{\delta} \right).$$

we have, with confidence at least $1 - \delta$,

$$d(f^{(t)}, \Delta_\mathcal{F}) \leq \frac{C_\Lambda}{C'_\Lambda} \left\{ \frac{\sqrt{k}}{1 - \alpha_*} \sqrt{\eta} + \alpha_*^t d(f^{(0)}, \Delta_\mathcal{F}) \right\}.$$

Taking $\eta = \frac{\alpha_*^{2t} d(f^{(0)}, \Delta_\mathcal{F})^2 (1 - \alpha_*)^2}{k}$ and $\mathbf{C} = 2\frac{C_\Lambda}{C'_\Lambda}$ finishes the proof of Theorem 1.

□

A small variation in the proof of Theorem 1 allows one to prove that the learning dynamic contracts to the diagonal for the distance d_Λ.

Proposition 3 *In the hypothesis of Theorem 1 there exist constants $\alpha_* < \alpha < 1$ and $\mathbf{C}' > 0$ such that for each $0 < \delta < 1$, and $t \in \mathbb{N}$, if $m \geq m_t$ then*

$$d_\Lambda(f^{(t)}, \Delta_\mathcal{F}) \leq \alpha^t d_\Lambda(f^{(0)}, \Delta_\mathcal{F})$$

with confidence at least $1 - \delta$.

One may take

$$m_t = \frac{12M^2k}{\mathbf{C}'\alpha^{2t}d_\Lambda(f^{(0)}, \Delta_\mathcal{F})^2} \ln \frac{tk\mathcal{N}\left(\mathcal{F}, \frac{\mathbf{C}'\alpha^{2t}d_\Lambda(f^{(0)}, \Delta_\mathcal{F})^2}{Mk}\right)}{\delta}.$$

PROOF. Let α be such that $\alpha_* < \alpha < 1$. Choose

$$\eta = \frac{(\alpha^t - \alpha_*^t)^2 d_\Lambda(f^{(0)}, \Delta_\mathcal{F})^2(1 - \alpha_*)}{C_\Lambda^2 k}.$$

Then, by (6),

$$d_\Lambda(f^{(t)}, \Delta_\mathcal{F}) \leq (\alpha^t - \alpha_*^t) d_\Lambda(f^{(0)}, \Delta_\mathcal{F}) + \alpha_*^t d_\Lambda(f^{(0)}, \Delta_\mathcal{F}) = \alpha^t d_\Lambda(f^{(0)}, \Delta_\mathcal{F})$$

and the map $f^{(t)} \mapsto f^{(t+1)}$ is a strict contraction with the d_Λ metric.

Now note that

$$\begin{aligned}
\alpha^t - \alpha_*^t &= \alpha^t \left(\left(1 - \frac{\alpha_*}{\alpha}\right)^t \right) \\
&= \alpha^t \left(1 - \frac{\alpha_*}{\alpha}\right) \left(1 + \frac{\alpha_*}{\alpha} + \left(\frac{\alpha_*}{\alpha}\right)^2 + \ldots + \left(\frac{\alpha_*}{\alpha}\right)^{t-1}\right) \\
&\geq \alpha^t \left(1 - \frac{\alpha_*}{\alpha}\right)
\end{aligned}$$

and therefore

$$\eta \geq \frac{\alpha^{2t} d_\Lambda(f^{(0)}, \Delta_\mathcal{F})^2 \mathbf{C}'}{k}$$

with $\mathbf{C}' = \frac{\left(1 - \frac{\alpha_*}{\alpha}\right)^2 (1 - \alpha_*)}{C_\Lambda^2}$. Therefore, for a given $0 < \delta < 1$, the contraction above holds with confidence $1 - \delta$ as long as

$$\begin{aligned}
m &\geq \frac{288M^2}{\eta} \ln \frac{tk\mathcal{N}\left(\mathcal{F}, \frac{\eta}{24M}\right)}{\delta} \\
&\geq \frac{288M^2k}{\mathbf{C}'\alpha^{2t}d_\Lambda(f^{(0)}, \Delta_\mathcal{F})^2} \ln \frac{tk\mathcal{N}\left(\mathcal{F}, \frac{\mathbf{C}'\alpha^{2t}d_\Lambda(f^{(0)}, \Delta_\mathcal{F})^2}{24Mk}\right)}{\delta}.
\end{aligned}$$

Replacing $\mathbf{C}'$ by $\frac{\mathbf{C}'}{24}$ finishes the proof. □

Remark 7 **(i)** Note that the contraction factor α can be taken arbitrarily close to α_* but as $\alpha \to \alpha_*$ one has $\mathbf{C}' \to 0$.

(ii) One may strengthen Theorem 1. Given $f = f^{(0)} \in \mathcal{F}^k$ and $0 < \delta < 1$ there exists t such that, for appropriate m_t, $d(f^{(t)}, f_\Delta) \leq \tau$. Yet, one can not say that for any $m(t)$ such that $m(t) \to \infty$ when $t \to \infty$, one has $f^{(t)} \to f_\Delta$. To see this, follow the proof of Theorem 1.

(iii) The dynamic defined by $\mathcal{T} : \mathcal{F}^k \to \mathcal{F}^k$ of Theorem 2 corresponds to the limit as $m \to \infty$ for each time period, as can be seen from the proof above.

7 Additional considerations

7.1 On the class $\mathcal{F}$

The space $\mathcal{F}$, present at the origin of a learning dynamic, puts a boundary on the set of learnable languages. In this sense, it plays a role akin to the universal grammar of Chomsky. In our linguistic model, it is a associated to a given society and thus, it is not a parameter we choose.

In contrast, in learning theory, the hypothesis space may be chosen and its choice is an important issue. A general setting to do so is given by a compact embedding of a Hilbert space $\mathbb{H}$,

$$i : \mathbb{H} \hookrightarrow \mathscr{C}(X; Y).$$

For any $R > 0$, the closure $\overline{i(B(\mathbb{H}, R))}$ of the image of the ball $B(\mathbb{H}, R)$ of radius R in $\mathbb{H}$ is a compact, convex subset of $\mathscr{C}(X; Y)$. Choices of $\mathbb{H}$ are spaces of polynomials of bounded degree, say d, Sobolev spaces H^s with $s > n/2$, and reproducing kernel Hilbert spaces arising from a $\mathscr{C}^\infty$ Mercer kernel (for details of these spaces see [Cucker and Smale 2002]). In the last two cases we have the following bound for the logarithm of covering numbers

$$\ln \mathcal{N}(\mathcal{F}, \varepsilon) \leq \left(\frac{RC_s}{\varepsilon}\right)^{n/s} + 1, \qquad \text{and} \qquad \ln \mathcal{N}(\mathcal{F}, \varepsilon) \leq \left(\frac{RC_h}{\varepsilon}\right)^{\frac{2n}{h}}.$$

where h is any number such that $h > n$, and C_s and C_h are constants independent of ε and R. Note, both bounds can be written in the form $\ln \mathcal{N}(\mathcal{F}, \varepsilon) \leq C_{\mathcal{F}} \left(\frac{1}{\varepsilon}\right)^a$ for some $C_{\mathcal{F}} > 0$ and $a > 0$ independent of ε. In the first case, we have

$$\ln \mathcal{N}(\mathcal{F}, \varepsilon) \leq N \ln\left(\frac{4R}{\varepsilon}\right),$$

where N is the dimension of the space of polynomials, i.e., $\binom{n+d}{n}$.

7.2 On the limiting common language

Recall, $V \in \mathbb{R}^k$ is the eigenvector of Λ^{T} associated with the eigenvalue 1 and such that the sum of its components is 1. We proved in Corollary 3 that the limiting language of the dynamic given by the iteration of $\mathcal{T} : \mathcal{F}^k \to \mathcal{F}^k$ with initial state f is given by $\mathrm{Diag}(V^{\mathrm{T}} f)$. Our next result shows a simple closed form for V.

Proposition 4 *If Λ is weakly irreducible then*

$$V = \left(\frac{\widetilde{\Lambda}_{i1}}{\sum_{j=1}^{k} \widetilde{\Lambda}_{j1}} \right)_{i=1}^{k}$$

where $\left(\widetilde{\Lambda}_{i1}\right)_{i=1}^{k}$ are the determinants of the cofactor matrices of the first column of $\Lambda - \mathrm{Id}$. The statement holds true for any column other than the first.

Proof. Let $\Lambda - \mathrm{Id} = \left(\widetilde{\lambda}_{ij}\right)_{i,j=1}^{k}$. We claim that

$$\left[\widetilde{\Lambda}_{11}, \widetilde{\Lambda}_{21}, \ldots, \widetilde{\Lambda}_{k1}\right] (\Lambda - \mathrm{Id}) = [0, \ldots, 0].$$

Indeed,

$$\left[\widetilde{\Lambda}_{11}, \widetilde{\Lambda}_{21}, \ldots, \widetilde{\Lambda}_{k1}\right] \begin{bmatrix} \widetilde{\lambda}_{1j} \\ \widetilde{\lambda}_{2j} \\ \vdots \\ \widetilde{\lambda}_{kj} \end{bmatrix} = \sum_{i=1}^{k} \lambda_{ij} \widetilde{\Lambda}_{i1} = \begin{cases} 0 & \text{if } j \neq 1 \\ \det(\Lambda - \mathrm{Id}) & \text{if } j = 1. \end{cases}$$

But $\det(\Lambda - \mathrm{Id}) = 0$ since 1 is an eigenvalue of Λ. So the claim holds. Since Λ is weakly irreducible, $\mathrm{rank}\,(\Lambda - \mathrm{Id}) = k - 1$. Then, $\widetilde{\Lambda}_{i1} \neq 0$ for some i. Thus, $\left(\widetilde{\Lambda}_{i1}\right)_{i=1}^{k}$ is an eigenvector of Λ^{T} with eigenvalue 1 and therefore

$$V = \left(\frac{\widetilde{\Lambda}_{i1}}{\sum_{j=1}^{k} \widetilde{\Lambda}_{j1}} \right)_{i=1}^{k}.$$

□

8 Some examples revisited

8.1 Two agent societies

In a two agent society the normalized communication matrix has the form

$$\Lambda = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix} \tag{7}$$

for some $0 \leq a, b \leq 1$. By Proposition 4 we have

$$V = \left(\frac{b}{a+b}, \frac{a}{a+b} \right)$$

from which it follows, by Lemma 1, that $W = \{w \in \mathbb{R}^2 \mid bw_1 + aw_2 = 0\}$. An obvious element in this set is $w_0 = (a, -b)$ and from the equation $\Lambda w_0 = \alpha w_0$ it follows that the eigenvalue for w_0 is $1-a-b$. By Remark 6 we can take $\alpha_* = 1-a-b$. Normalizing the two eigenvectors of Λ we obtain a basis

$$\mathcal{B}_\Lambda = \left\{ \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right), \left(\frac{a}{\sqrt{a^2+b^2}}, \frac{-b}{\sqrt{a^2+b^2}} \right) \right\}$$

of $\mathbb{R}^2$ which is orthonormal for $\langle\ ,\ \rangle_\Lambda$.

Let $\mathcal{S}^1$ be the unit circle (with respect to $\|\ \|_\Lambda$) in $\mathbb{R}^2$ and let (x, y) denote the coordinates (in the basis $\mathcal{B}_\Lambda$) of a point in $\mathcal{S}^1$. Then

$$\begin{aligned} C_\Lambda^2 &= \max_{(x,y)\in\mathcal{S}^1} \left(\frac{x}{\sqrt{2}} + \frac{ya}{\sqrt{a^2+b^2}} \right)^2 + \left(\frac{x}{\sqrt{2}} - \frac{yb}{\sqrt{a^2+b^2}} \right)^2 \\ &= \max_{(x,y)\in\mathcal{S}^1} x^2 + y^2 + 2xy \frac{a-b}{\sqrt{(a^2+b^2)2}} \\ &= \max_{(x,y)\in\mathcal{S}^1} 1 + 2xy \frac{a-b}{\sqrt{2(a^2+b^2)}} \end{aligned}$$

and $(C'_\Lambda)^2$ is obtained by minimizing the same expression. The extrema of this expression are given by the solutions of the equation

$$\mu(x, y) = (y, x)$$

where $\mu \in \mathbb{R}$ is a Lagrange multiplier. It follows that $\mu = \pm 1$ and $y = \pm x$ and therefore, that

$$C_\Lambda^2 = 1 + \frac{|a-b|}{\sqrt{2(a^2+b^2)}}$$

and

$$(C'_\Lambda)^2 = 1 + \frac{-|a-b|}{\sqrt{2(a^2+b^2)}}.$$

We conclude that

$$\mathbf{C}^2 = 4 \frac{\sqrt{2(a^2+b^2)} + |a-b|}{\sqrt{2(a^2+b^2)} - |a-b|}.$$

The following result thus follows applying Corollary 2 and Remark 4(iii).

Proposition 5 *Let $\mathcal{P}$ be a 2-agent society with normalized communication matrix $\Lambda = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$ and let $\tau > 0$. Write*

$$\mathbf{C}^2 = 4\frac{\sqrt{2(a^2+b^2)} + |a-b|}{\sqrt{2(a^2+b^2)} - |a-b|}$$

and let T be the smallest integer greater than or equal to $\frac{\ln(\mathbf{C}\frac{\sqrt{2}M}{2}) - \ln\tau}{|\ln\alpha_*|}$. Then, for all $f^{(0)} \in \mathcal{F}^2$, $f^{(T)} \in N(\Delta_{\mathcal{F}}, \tau)$ with confidence at least

$$1 - 2T\, e^{-\frac{m(1-\alpha_*)^2\tau^2}{576M^2\mathbf{C}^2} + C_{\mathcal{F}}\left(\frac{48M\mathbf{C}^2}{(1-\alpha_*)^2\tau^2}\right)^a}$$

where $\alpha_* = 1 - a - b$. □

Remark 8 The limiting language for the "ideal" dynamic given by $f \mapsto \mathcal{T}f$ with initial state (f_1, f_2) is $\left(\frac{bf_1+af_2}{a+b}, \frac{bf_1+af_2}{a+b}\right)$.

8.2 A simple case of language acquisition

We close this section revisiting Example 4. The linguistic society is composed only of a mother and a baby and its normalized communication matrix Λ is given by

$$\begin{pmatrix} 1 & 0 \\ 1-\theta & \theta \end{pmatrix}$$

where $\theta > 0$ is small. Our previous results readily apply to this case since this matrix is the one in (7) with $a = 0$ and $b = 1 - \theta$. In this case, $\alpha_* = \theta$, and

$$\mathbf{C} = 2(\sqrt{2}+1) \approx 4.82842.$$

Proposition 5 then yields the following result.

Proposition 6 *The pair Mother-Baby in Example 4 reaches a language-system (i.e. a state $f \in N(\Delta_{\mathcal{F}}, \tau)$) in at most $T(\theta)$ iterations with probability at least $1-\delta$ where*

$$T(\theta) = \frac{\ln\left(\sqrt{2}(1+\sqrt{2})M\right) - \ln\tau}{|\ln\theta|} \approx \frac{1.28 + \ln M - \ln\tau}{|\ln\theta|}$$

and

$$\delta \leq 2T(\theta)\, e^{-\frac{m(1-\theta)^2\tau^2}{13430M^2} + C_{\mathcal{F}}\left(\frac{1120M}{(1-\theta)^2\tau^2}\right)^a}.$$

□

Remark 9 **(i)** The limiting language for the "ideal" dynamic given by $f \mapsto \mathcal{T}f$ with initial state (f_M, f_B) is, perhaps not surprisingly, (f_M, f_M).

(ii) Note that the numerator in the bound $\frac{2+\ln M - \ln \tau}{|\ln \theta|}$ is common to all Mother-Baby pairs. The speed with which the baby learns the mother's language (an accomplishment which we recognize when $d((f_B^{(t)}, f_M^{(t)}), \Delta_{\mathcal{F}}) \leq \tau$) depends on the denominator $|\ln \theta|$ as well as on conditions on $M, C_{\mathcal{F}}, m$ needed to ensure $\delta < 1$. Variations in θ (interpreted as differences in the innate ability of the baby, frequency of linguistic encounters with the mother, etc.) are an important factor in our model to account for the variations on children's learning speed.

8.3 A final remark on convexity

Consider the finite Y model of [Ke, Minett, Au, and Wang 2002] we described in Example 2. The set Y can be a set of "words" or "sentences". Theorem 1 (and its corollaries) does not directly apply because Y is not convex. One may consider the convex closure of Y in $\mathbb{R}^l$ consisting of the points $\sum w_i e_i$ (where $Y = \{e_1, \ldots, e_l\}$ and $\sum w_i = 1$) and interpret w_i as the probability of using e_i. We have seen this interpretation imbedded quite reasonably in Example 1. But we now remark that languages in this example satisfy some conditions (e.g. mapping the white color to the word `white`, the black color to the word `black`, and being non-decreasing) which make the interpretation above reasonable. Thus, while this interpretation may be appropriate in Example 1, it may give rise as well to languages without communication power; languages in which any utterance in $\{e_1, \ldots, e_l\}$ can be used to describe any object in X.

There is another way, however, to deal with w as probabilities. This is akin to Example 6. During a "learning phase" possibly consisting of several steps of the learning dynamic, agents modify their languages to obtain tentative languages. This learning phase ends when the tentative languages of the society converged to a language system, at which time, for each object x_s the signal e_i with the highest probability w_i is adopted for this object. We will not pursue this model here. We just note that what we have described is not too different from developments in [Niyogi 2003b; Yang 2003].

References

Cucker, F. and S. Smale (2002). On the mathematical foundations of learning. *Bulletin Amer. Math. Soc. 39*, 1–49.

Cucker, F. and R. Wong (2000). *The Collected Papers of Stephen Smale*. World Scientific.

De Boer, B. (2001). *The Origins of Vowel Systems*. Oxford University Press.

Fant, G. (1970). *Acoustic Theory of Speech Production*. Mouton.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation 100*, 78–150.

Ke, J., J. Minett, C.-P. Au, and W. S.-Y. Wang (2002). Self-organization and natural selection in the emergence of vocabulary. To appear in *Complexity*.

Kirby, S. and J. R. Hurford (2001). The emergence of linguistic structure: an overview of the iterated learning model. In D. Parisi and A. Cangelosi (Eds.), *Computational Approaches to the Evolution of Language and Communication*. Springer-Verlag.

Lyons, J. (1981). *Language and Linguistics: An Introduction*. Cambridge University Press.

Minett, J. and W. S.-Y. Wang (2003). *Language Acquisition, Change and Emergence*. To appear.

Niyogi, P. (1998). *The Informational Complexity of Learning*. Kluwer Academic Publishers.

Niyogi, P. (2003a). The computational nature of language learning and evolution. `http://www.cs.uchicago.edu/niyogi`.

Niyogi, P. (2003b). Phase transitions in language evolution. Preprint.

Nowak, M., N. Komarova, and P. Niyogi (2001). Evolution of universal grammar. *Science 404*, 495–498.

Saussure, F. (1983). *Course in General Linguistics*. Duckworth. Translated and annotated by Roy Harris.

Seneta, E. (1973). *Non-Negative Matrices*. John Wiley & Sons.

Smale, S. (1963). Stable manifolds for differential equations and diffeomorphisms. *Annali delle Scuola Normale Sup. di Pisa XVII*, 97–116. Reprinted in [Cucker and Wong 2000, Volume 2].

Smale, S. (1981). Global analysis and economics. In K. Arrow and M. Intrilligator (Eds.), *Handbook of Mathematical Economics*, pp. 331–370. Elsevier Science. Reprinted in [Cucker and Wong 2000, Volume 1].

Stevens, K. (1998). *Acoustic Phonetics*. The MIT Press.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley & Sons.

Yang, C. (2003). *Knowledge and Learning in Natural Language*. Oxford University Press.

Appendix: Fitness Maximization

In this section we present a way to characterize F_i. We assume $\mathcal{F} \subseteq \mathscr{C}(X;Y)$ is convex.

Define the *linguistic fitness* of language f for agent i at the state $(f_1, \ldots, f_k)$ by

$$\Phi_i(f) = -\int_X \left(\sum_{j=1}^{k} \gamma_{ij} \|f(x) - f_j(x)\|_Y^2 \right) d\rho_X(x). \tag{8}$$

Fitness (abuse of language) can be thought to measure the ability of agent i to communicate with members of the society he encounters when he uses the language

f (note that the ith term in the sum above reflects an inertia acting on agent i). This motivates the problem: at a given state, find the language $f \in \mathcal{F}$ that maximizes the linguistic fitness, i.e. compute

$$F_i^* = \arg\max_{f \in \mathcal{F}} \Phi_i(f), \qquad i = 1, \ldots, k. \tag{9}$$

Proposition 7 *For $i = 1, \ldots, k$ the function*

$$F_i(x) = \sum_{j=1}^{k} \lambda_{ij} f_j(x)$$

is a solution of (9). Any other solution of (9) is $\mathscr{L}^2_\rho(X)$-equivalent to F_i.

PROOF. The problem can be solved for each fixed $x \in X$ by minimizing the quantity

$$\sum_{j=1}^{k} \gamma_{ij} \|f(x) - f_j(x)\|_Y^2$$

over the vectors $f(x)$ in $Y = \mathbb{R}^l$. In fact, if we write the vector $f(x)$ as $y = (y_1, \ldots, y_l) \in \mathbb{R}^l$, then the above quantity is the function $\varphi(y) = \sum_{j=1}^{k} \gamma_{ij} \sum_{s=1}^{l} (y_s - (f_j(x))_s)^2$. The only stationary point $y^* = (y_1^*, \ldots, y_l^*)$ of this function satisfies, for $s = 1, \ldots, l$,

$$\frac{\partial \varphi}{\partial y_s}(y^*) = 2 \sum_{j=1}^{k} \gamma_{ij} (y_s^* - (f_j(x))_s) = 2 \left(\sum_{j=1}^{k} \gamma_{ij} \right) y_s^* - 2 \sum_{j=1}^{k} \gamma_{ij} (f_j(x))_s = 0.$$

That is, $(\sum_{j=1}^{k} \gamma_{ij}) y^* = \sum_{j=1}^{k} \gamma_{ij} f_j(x)$. Thus, for each fixed $x \in X$, the function

$$F_i(x) = \frac{\sum_{j=1}^{k} \gamma_{ij} f_j(x)}{\sum_{j=1}^{k} \gamma_{ij}}$$

satisfies that, for all $f \in \mathcal{F}$ and all $x \in X$,

$$-\sum_{j=1}^{k} \gamma_{ij} \|F_i(x) - f_j(x)\|_Y^2 \geq -\sum_{j=1}^{k} \gamma_{ij} \|f(x) - f_j(x)\|_Y^2$$

and the equality holds only when $f(x) = F_i(x)$. Therefore, for all $f \in \mathcal{F}$,

$$\Phi_i(F_i) \geq \Phi_i(f)$$

and the equality holds only when $f = F_i$ almost everywhere on (X, ρ_X). □

Remark 10 We say that the measure ρ_X is *non-degenerate* when, for all open subset $U \subset X$, $\rho_X(U) > 0$. Non-degeneracy is a mild assumption; if ρ_X is degenerate one can replace (X, ρ_X) by $(\overline{X}, \rho_{\overline{X}})$ such that $\overline{X} \subset X$, $\rho_{\overline{X}}$ is non-degenerate and $\rho_{\overline{X}}(\overline{X}) = \rho_X(\overline{X}) = 1$.

We note now that if ρ_X is non-degenerate then, in Proposition 7, F_i is unique since it is continuous.