Part IV

The Origin of Language

Chapter 11

The Origin of Communicative Systems: Communicative Efficiency

In this part of the book, we turn our attention to the question of the *origin* of language, i.e., the genesis of human language from prelinguistic versions of it. The issue is notoriously difficult to investigate as empirical facts that pertain to this matter are particularly hard to come by. At the same time, the genesis of the language faculty certainly counts as one of the major evolutionary transitions – at par with other significant evolutionary events such as the origin of protocells, eukaryotes, sexual reproduction, and the like (see Maynard Smith and Szathmary, 1995). So in some sense, this might seem like the holy grail of linguistic inquiry — how and why did this peculiar recursive communication system of human language arise in biological populations? This particular question has attracted significant attention in recent times (see, for example, Proceedings of the International Conferences on Language Evolution, held every two years since 1996).

Whatever story one chooses to tell about how language arose, one will need to invoke learning theoretic considerations. Thus learning at the individual level and evolution at the population level will continue to be important themes in our narrative. However, we will now need to formally develop certain other notions that have played a minimal role in the models that have been discussed so far.

First, there is the notion of *communicative efficiency*. This has been a recurring theme and a constant subtext in linguistic studies from a func-

tionalist perspective that takes the central defining feature of language to be its communicative function (see Halliday, 1994 and Newmeyer, 1998, for perspectives from different points of view; also Kirby, 1999, for a discussion in the context of language evolution). Human language allows us a rich flexibility to describe and communicate a variety of thoughts and events to each other. While one may seriously debate the degree to which the structure of language reflects its communicative function, there is no denying the fact that among other things, language is also used for communication. There is also no denying the fact that human language allows a richer medium of communication than other animal languages. It is certainly possible that communicative efficiency was important in the evolution of competing linguistic groups where the different groups had different communicative ability.

To fully develop this argument, we need a notion of *differential fitness* and *natural selection*. In much of this book, we have been considering modern[1] human languages where all languages are more or less equally efficient. However, one might imagine populations of organisms that are basically similar except with respect to their communicative language. If communicative efficiency provides biological fitness to individuals in terms of increased ability to reproduce or survive, then would populations converge to coherent linguistic states?

As usual, we will be interested in the stable modes of linguistic populations under such assumptions. In addition to stability, we will examine the closely related notion of *coherence*. A coherent population is one where most members of the population have a shared common language, i.e., the population is linguistically homogeneous. Often, one observes coherent populations in settings where there is no obvious centralizing agency that enforces such coherence. We investigate whether coherence can emerge through individual interactions — an idea that is explored in different contexts in the literature on self organization and emergence in complex systems.

These themes arise in the study of populations of interacting linguistic agents in animal, human, and artificial populations. For example, the notion of emergence is invoked in artificial intelligence where rather than preprogramming structure (design by centralized planning) one asks whether

[1] We mean modern in evolutionary time scales, i.e., the evolution of human language after the origin of the current human species from some pre-human ancestor. It is our assumption that in this modern period, there has been no genetic evolution of the language faculty, but rather cultural evolution of particular natural languages within the scope ordained by the space of possible human languages (universal grammar). This evolutionary process falls within the purview of historical linguistics.

intelligent behavior can evolve or emerge by interaction between simple decentralized agents (e.g. the work of Luc Steels and colleagues in the context of designing languages for robots). Similarly, studies of signalling or animal communication also need to deal with such issues.

This part of the book studies the interplay between communicative efficiency, learning, fitness, and coherence in some abstraction. In the current chapter, we develop the notion of communicative efficiency and fitness. We characterize language as a probabilistic association between form and meaning. We then provide a natural definition of communicative efficiency between two linguistic agents possessing different languages. We consider the implications of this definition for learning and evolution. We also perform an empirical study on large linguistic corpora to find that the structure of the lexicon of several modern languages do not reflect optimality in terms of communicative efficiency. In the light of this, one should take seriously the possibility that communicative efficiency as developed in these chapters plays less of a role in human language and may be a more useful concept in studies of animal or artificial communication systems.

In the next two chapters, we consider a population of linguistic agents evolving with (Chapter 12) and without (Chapter 13) fitness. This fitness is derived from communicative efficiency. In both cases we find that coherent states emerge if the learning fidelity is high, i.e., if the learner is able to acquire the target language with high confidence. In particular, we see that there is a bifurcation point in the dynamics when there is a transition from incoherent to coherent states. We also find that if linguistic agents learn only (primarily) from a single individual in the population (typically parent, e.g., certain animal species that learn in the nesting phase) then fitness is necessary for coherence to emerge. On the other hand, if linguistic agents learn from the population at large, then coherence might emerge even in the absence of fitness.

11.1 Communicative Efficiency of Language

Consider two linguistic agents in a shared world. The agents desire to communicate different messages (meanings) to each other. Such a situation arises in a number of different contexts in natural and artificial communication systems and it is important in such cases to be able to quantify the rate of success in information transfer, in other words, the *mutual intelli-*

gibility or *communicative efficiency*[2] of the agents. Each agent possesses a communicative device or a language that allows it to relate code (signal) and message, form and meaning, syntax and semantics, depending upon the context in which the communication arises. If they share the same language and this language is expressive enough and unambiguous, then mutual intelligibility will be very high. If on the other hand, they do not share the same language, or the languages are inexpressive or ambiguous, the mutual intelligibility will be much lower.

In this chapter we present an analysis of this situation. We view languages as probabilistic associations between form and meaning and develop a natural measure of intelligibility, $F(L_1, L_2)$, between two languages, L_1 and L_2. This is a generalization of a similar function introduced in the context of language evolution by Hurford (1989). We ask the following question: if there is a biological/cultural/technological advantage for an agent to increase its intelligibility with the rest of the population, what are the ways to do this?

The task of increasing intelligibility reduces ultimately to three related sub-problems:

1. Given a language L, what language L' maximizes the mutual intelligibility $F(L, L')$ for two way communication about the shared world?

2. What are some acquisition mechanisms/learning algorithms that can serve the task of improving intelligibility?

3. What are the consequences of individual language acquisition behavior on the population dynamics and the communicative efficiency of an interacting population of linguistic agents?

In the next few chapters, we develop a mathematical framework to address these questions analytically. In this chapter, we address questions (1) and (2) above. We find, surprisingly, that the optimal language L' for maximizing communicative efficiency with a user of L need not be the same as L. In general, if L has ambiguities, then L' will end up being different from L. We will also see that in general, there may be no language L' that achieves the maximal communicative efficiency with L but there always exist languages that come arbitrarily close to this maximum. Our main result

[2] We use the terms mutual intelligibility and communicative efficiency interchangeably in this book.

is an algorithm for constructing such arbitrarily good approximations. Using this result we demonstrate an algorithm that can learn such languages from linguistic exchanges with L. Bounds on the sample complexity of such learning algorithms are then provided. Finally, we conduct empirical studies on linguistic corpora of several languages. These studies suggest that the lexicons of these languages are not structured for optimal communicative efficiency in the sense described here. The implications of this are discussed. In the next chapter, we examine the evolutionary dynamics of populations (question 3 above) within this framework.

11.1.1 Communicability in animal, human and machine communication

The simplest situation where communicability is readily defined corresponds to the case where the "language" may be viewed as an association matrix, A. Such a matrix simply links referents to signals. If there are M referents and N signals, then A is an $N \times M$ matrix. The entries, a_{ij}, define the relative strength of the association between signal i and meaning j. The matrix A thus characterizes the behavior of the linguistic agent in (i) *production mode* where it may produce any of the signals corresponding to a particular meaning in proportion to the strength of the association, and in (ii) *comprehension mode* where it may interpret a particular signal as any of the meanings in proportion to the association strengths.

The specific settings in which such a scheme is a useful description include animal communication, human languages and artificial languages. For instance, it often makes sense to talk about a *lexical matrix* as a formal description of human mental vocabularies. It is introduced to describe the arbitrary relations between discrete words and discrete concepts of human languages (Hurford, 1989; Miller, 1996 ; Regier et al, 2001 ; Regier, 2003; Tenenbaum and Xu, 2000; Komarova and Nowak, 2001). Each column of the lexical matrix corresponds to a particular word meaning (or concept), each row corresponds to a particular word form (or word image). In the Saussurean terminology of arbitrary sign, the lexical matrix provides the link between signifie and signifiant (Saussure, 1983).

An equivalent of a lexical matrix is also at the basis of any animal communication system, where it defines the relation between animal signals and their specific meanings (Hauser, 1997; Smith, 1977; Macedonia and Evans, 1993; Cheney and Seyfarth, 1990). A classic example of this is alarm calls in primates. There are a finite number of referents that are coded using

acoustic signals and decoded appropriately by recipients.

Infinite association matrices can be used as a description of human languages. Human grammars mediate a complex mapping between form and meaning. Here, the space of possible signals is the set of all strings (sentences) over a finite syntactic alphabet and the set of possible meanings is the set of all strings over some semantic alphabet. Most crucially, the sets of possible sentences and meanings are infinite. This accounts for the infinite expressibility of human grammars.

In artificial intelligence, the problem arises in many different settings. A number of studies have emerged where linguistic agents interact with each other in simulated worlds and one studies whether coherent or coordinated communication ultimately emerges (see, for example, Steels, 1996; Steels and Vogt, 1997; Steels and Kaplan, 1998; Oliphant, 1999; Briscoe, 2000; Kirby, 1999). Much of this kind of research employs the simulation methodology of Artificial Life. In this chapter, we create a mathematical framework for these kinds of problems and derive a number of analytic results. Ultimately, we study language coordination (coherence) in a community of linguistic agents and this has natural synergies with research on multi-agent systems (see, for example, Boutilier et al, 1997, for an overview perspective on multi-agent systems.)

In the design of natural language understanding systems, the goal is to develop a computer system that is able to communicate with a human. The statistical approach to this problem assumes an underlying probabilistic model for the human source. This probabilistic model is then recovered or learned from data either by randomly drawn samples as in the case of corpus linguistics or statistical language modeling (see Charniak, 1993; Manning and Schutze, 1999 for overviews of this point of view) or via some interactive exchanges and semantic reinforcement (Levinson, 1991; Kearns et al, 2001). The primary implication of the results here is that optimal communication with a language user might require one to learn a language that is *different* from the target source.

11.2 Communicability for linguistic systems

We will now develop a formal characterization of communicative efficiency between the users of two linguistic systems.

11.2.1 Basic notions

We regard a linguistic system to be an association between form and meaning. Let $\mathcal{S}$ be the set of all possible linguistic forms (sentences or signals) and $\mathcal{M}$ be the set of all possible semantic objects (meanings or referents). Note that depending on the context, the elements of $\mathcal{S}$ can be words, codes, expressions, forms, signals or sentences. The elements of $\mathcal{M}$ can be meanings, messages, events or referents. We will use the general term *signals* for elements of $\mathcal{S}$ and *meanings* for elements of $\mathcal{M}$.

The sets $\mathcal{S}$ and $\mathcal{M}$ need not be finite, but it is essential that they are enumerable. For natural languages, because of its discrete underlying structure, the sets $\mathcal{S}$ and $\mathcal{M}$ may be viewed as countable. In the lexical setting, $\mathcal{S}$ is the set of all words, therefore is naturally countable while the countability of $\mathcal{M}$ (the meanings) is motivated by category formation in semantic space. In the case of human grammars, we may let $\mathcal{S} = \Sigma_1^*$ be the set of all possible strings over a syntactic alphabet (Σ_1) and $\mathcal{M} = \Sigma_2^*$ be the set of all possible strings over a semantic alphabet (Σ_2). Note that in this case $\mathcal{S}$ and $\mathcal{M}$ are infinite.

We define a communication system, or a language, to be a probability measure μ over $\mathcal{S} \times \mathcal{M}$. Note that in the case of finite languages (human or artificial lexicons and animal communication systems), μ is related to the association matrix, A, by means of a simple rescaling.

Let us enumerate all possible signals, i.e. the elements of set $\mathcal{S}$, as $s_1, s_2, s_3, \ldots$ and all possible meanings (elements of $\mathcal{M}$) as $m_1, m_2, m_3, \ldots$. The coding and decoding schemes of the agent are contained in the measure μ in the following manner. Each user of μ is characterized by an *encoding matrix* P and a *decoding matrix* Q where

$$P_{ij} \equiv \mu(s_i|m_j) = \begin{cases} \mu(s_i, m_j)/\sum_p \mu(s_p, m_j), & \text{if } \sum_p \mu(s_p, m_j) > 0, \\ 0, & \text{otherwise,} \end{cases} \quad (11.1)$$

$$Q_{ji} \equiv \mu(m_i|s_j) = \begin{cases} \mu(s_j, m_i)/\sum_p \mu(s_j, m_p), & \text{if } \sum_p \mu(s_j, m_p) > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (11.2)$$

Both P and Q matrices are easily interpreted. P_{ij} is simply the probability of producing the signal s_i given that one wishes to convey the meaning m_j. Similarly, Q_{ij} is the probability of interpreting the expression s_i to mean m_j by the same user.

Matrices analogous to P and Q were introduced in Hurford (1989) , however, they were not explicitly related through a common measure, μ. An

effective connection between P and Q has been employed for a particular learning mechanism, called the *Saussurean* (Hurford, 1989; Oliphant, 1999).

Remarks:

1. The user of a language is characterized in *production mode* by the matrix P and in *comprehension mode* by the matrix Q. This captures the fact that given a particular meaning, there might be many different ways to express it. Correspondingly given a particular signal, there may be no unique interpretation. Thus ambiguities in sentence interpretation or polysemy in lexical semantics are incorporated.

2. We have required that P and Q arise from a common measure μ. This ensures that the user of a language is self consistent in production and comprehension. If P and Q were completely decoupled, one could potentially have pathological situations like the following. A language user could use the word forms $/cat/$ and $/dog/$ to refer to the concepts "cat" and "dog" respectively in production mode. Yet in comprehension mode, the same language user might interpret $/cat/$ as "dog" and $/dog/$ as "cat" respectively. Such contradictory behavior in production and comprehension is eliminated by requiring P and Q to be related to a common association matrix and more generally a probability measure.

3. While a measure μ uniquely defines the corresponding P and Q matrices, the converse is not generally true. Given the P and Q matrices it might be possible to find more than one μ which would have the correct encoding and decoding matrices. An example with 2×2 matrices is $P = Q = I$ and

$$\mu_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}, \qquad \mu_2 = \begin{pmatrix} 1/3 & 0 \\ 0 & 2/3 \end{pmatrix}.$$

   Clearly, both μ_1 and μ_2 lead to the same P and Q. In order to avoid such ambiguities we consider equivalence classes of measures. We will say that two measures μ_1 and μ_2 are equivalent to each other ($\mu_1 \equiv \mu_2$) if and only if the corresponding P and Q matrices are equal, i.e. $P^{(1)} = P^{(2)}$ and $Q^{(1)} = Q^{(2)}$.

4. For a probability measure μ let us introduce

$$\mathcal{S}_\mu = \{s \in \mathcal{S} |\ \exists m \in \mathcal{M} \text{ such that } \mu(s, m) > 0\}.$$

This defines the set of signals that will be used in production or comprehension by a linguistic agent. In the sense of formal language theory, this is the set of well formed syntactic expressions. In a human language setting, $\mathcal{S}_\mu$ corresponds to the traditional notion of language as a set of grammatical expressions generated by some underlying grammar. Our definition of a language as a measure μ contains this traditional notion of language as the *support* of the marginal probability of μ. Similarly, one may define

$$\mathcal{M}_\mu = \{m \in \mathcal{M} | \ \exists s \in \mathcal{S} \text{ such that } \mu(s,m) > 0\}.$$

This defines the set of all meanings that are expressible by the linguistic agent. If $\mathcal{M}_\mu = \mathcal{M}$ then all meanings can be expressed. If $\mathcal{M}_\mu$ is a proper subset of $\mathcal{M}$, this means that some meanings are left unexpressed.

5. The probability measure μ, the sets $\mathcal{S}_\mu$ and $\mathcal{M}_\mu$, and the matrices P and Q in humans and animals presumably arise out of highly structured systems in the brain. In fact, it is clear that in human languages, these objects may not vary arbitrarily. A significant activity in generative linguistics attempts to characterize the nature of this structure and the variation that exists among natural languages of the world. Correspondingly, field work on animal communication systems (see, for example, the large literature on bird songs) characterizes the structure and variation in systems within and across species.

11.2.2 Probability of events and a communicability function

The communicating agents are immersed in a world and the need to communicate messages arises as corresponding events occur in this shared world. Thus one may define a measure σ on the set of possible meanings $\mathcal{M}$ according to which the agents need to communicate these meanings to each other. Given two communication systems, i.e., languages μ_1 and μ_2, the probability that an event occurs whose meaning is successfully communicated from μ_1 to μ_2 is given by

$$\mathbb{P}[1 \to 2] = \sum_i \sigma(m_i) \sum_j \mu_1(s_j|m_i)\mu_2(m_i|s_j).$$

Similarly, one may compute the probability with which an event is successfully communicated from μ_2 to μ_1 as

$$\mathbb{P}[2 \to 1] = \sum_i \sigma(m_i) \sum_j \mu_2(s_j|m_i)\mu_1(m_i|s_j).$$

We may then define the effective *communicability function* of μ_1 and μ_2 as

$$F(\mu_1, \mu_2) = \frac{1}{2}(\mathbb{P}[1 \to 2] + \mathbb{P}[2 \to 1]).$$

This is the *mutual intelligibility* or *communicative efficiency* between a user of μ_1 and a user of μ_2. In matrix notation, this may be written as

$$F(\mu_1, \mu_2) = 1/2\,[tr\,(P^{(1)}\Lambda(Q^{(2)})^T) + tr(P^{(2)}\Lambda(Q^{(1)})^T)], \tag{11.3}$$

where Λ is a diagonal matrix such that $\Lambda_{ii} = \sigma(m_i)$, $tr(A)$ denotes the trace of the matrix A, and $P^{(i)}, Q^{(i)}$ refer to the coding and decoding matrices associated with measure μ_i. Note that $tr(P^{(1)}\Lambda(Q^{(2)})^T)$ is simply the probability that an event occurs and is successfully communicated from user of μ_1 to user of μ_2.

Remarks:

1. The function $F(\mu_1, \mu_2)$ is the average probability with which μ_1 and μ_2 understand each other in two way communication mode. The function $F(\mu_1, \mu_2)$ is symmetrical with respect to its arguments. If $|\mathcal{S}| = |\mathcal{M}|$ and μ_1 is a probability measure with support only on the diagonal elements of $\mathcal{S} \times \mathcal{M}$, then $P = Q = I$ (where I is the identity matrix). The communicative efficiency $F(\mu_1, \mu_1) = 1$.

2. $F(\mu_1, \mu_1)$ is the communicability of two identical linguistic agents. We have

   $$0 < F(\mu_1, \mu_1) \leq 1$$

   For two different agents μ_1 and μ_2 we also have

   $$0 \leq F(\mu_1, \mu_2) = F(\mu_2, \mu_1) \leq 1$$

3. Note that the marginal $\mu(m)$ is not equal to $\sigma(m)$. In other words, the language of an agent is simply given by μ and the conditional

probabilities associated with it. The probability with which agents communicate different meanings is determined not by the language but by the external world in which the agents are grounded. Therefore, two agents might have high communicative efficiency in some world and low communicative efficiency in another one.

4. A function similar to our communicability function was introduced by Hurford (1989). However, all meanings were treated to have equal probabilities (a uniform measure σ), and thus the function was not suitable for infinite matrices.

11.3 Reaching the highest communicability

Let us assume that one of the languages is given and call this language μ_0. According to definition (11.3), for any language μ, we have (where $\sigma_i = \sigma(m_i)$)

$$F(\mu_0, \mu) = \frac{1}{2} \sum_{i,j} \sigma_j [\mu_0(s_i|m_j)\mu(m_j|s_i) + \mu(s_i|m_j)\mu_0(m_j|s_i)]. \tag{11.4}$$

Let us define the *best response* as a language μ_*, such that

$$F(\mu_0, \mu_*) = \sup_{\mu} F(\mu_0, \mu). \tag{11.5}$$

In what follows we will present an algorithm for constructing a best response or a language which in some sense approaches the best response. In particular, we show that the best response need not exist. However, an arbitrarily good response can be constructed. We show how to construct a family of languages (μ_ϵ where $\epsilon > 0$) such that $F(\mu_0, \mu_\epsilon)$ can be made arbitrarily close to $\sup_\mu F(\mu_0, \mu)$ — the maximum possible mutual intelligibility between a user of μ_0 and a user of any allowable language.

11.3.1 A special case of finite languages

In order to keep the argument as transparent as possible, we will first make three simplifying assumptions. We will discuss relaxations later.

i) The languages are finite, and the matrices have the size $N \times M$,

ii) The distribution σ is uniform, i.e. $\sigma_i = 1/M \ \forall i$,

iii) The measure μ_0 satisfies the **property of unique maxima**, i.e. for each i, there exist a unique $p_0(i)$ and a unique $r_0(i)$ such that

$$\mu_0(s_i|m_{p_0(i)}) = \max_p \mu_0(s_i|m_p), \quad \mu_0(m_i|s_{r_0(i)}) = \max_r \mu_0(m_i|s_r). \tag{11.6}$$

The last condition states that there exists strictly one element of each column of $\mu_0(s|m)$ (row of $\mu_0(m|s)$) such that it is the biggest element in the column (row).

Let us maximize each of the two terms in the right hand side of expression (11.4) separately. First, we find a matrix Q^* such that

$$\sum_{i,j} \mu_0(s_i|m_j) Q^*_{ij} = \max_Q \sum_{i,j} \mu_0(s_i|m_j) Q_{ij}, \tag{11.7}$$

where we maximize over all matrices Q whose elements are non-negative and sum up to one within each row. This results in the following definition of Q^*:

$$Q^*_{ij} = \begin{cases} 1, & \mu_0(s_i|m_j) = \max_p \mu_0(s_i|m_p), \\ 0, & \text{otherwise.} \end{cases} \tag{11.8}$$

In other words, in order to construct the *best decoder*, Q^*, we need to find the largest elements in each of the rows of $\mu_0(s|m)$ and put "ones" at the correspondent slots of Q^*. The rest of the entries of the matrix Q^* are zero. This is a well defined operation because of the property of unique maxima. Similarly, we can find the matrix P^* such that

$$\sum_{i,j} P^*_{ij} \mu_0(m_j|s_i) = \max_P \sum_{i,j} P_{ij} \mu_0(m_j|s_i),$$

where we maximize over all matrices P whose elements are non-negative and sum up to one within each column. The *best encoder*, P^*, is given by

$$P^*_{ij} = \begin{cases} 1, & \mu_0(m_j|s_i) = \max_p \mu_0(m_j|s_p), \\ 0, & \text{otherwise,} \end{cases} \tag{11.9}$$

i.e. we maximize each column of the matrix $\mu_0(m|s)$. Now, we have the best encoder and the best decoder for the language μ_0. Finding the matrices P^* and Q^* completes the task of the obverter of Oliphant (1997). However, in

our setting, the two matrices cannot be independent, but they need to be related by a common measure. If a measure μ_* existed such that

$$\mu_*(s|m) = P^*, \quad \mu_*(m|s) = Q^*,$$

then it would satisfy Eq. 11.5, thus defining the best response. It turns out that in general, μ_* does not exist. However, there always exists a measure which approximates the performance of P^* and Q^* arbitrarily well. It is convenient to use the following short hand notation:

$$P^0_{ij} = \mu_0(s_i|m_j), \quad Q^0_{ij} = \mu_0(m_j|s_i).$$

We are ready to formulate the following

Theorem 24 *For any finite language μ_0 satisfying the property of unique maxima, and a uniform probability distribution σ, we have*

$$\sup_\mu F(\mu_0, \mu) = 1/(2M)\,\mathrm{tr}(P^0(Q^*)^T + P^*(Q^0)^T).$$

In order to prove this statement, we need to show that

(a) for all μ, $F(\mu_0, \mu) \leq 1/(2M)\mathrm{tr}(P^0(Q^*)^T + P^*(Q^0)^T)$,

(b) there exists a family of languages, μ_ϵ, such that

$$\lim_{\epsilon\to 0} |\sup_\mu F(\mu_0, \mu) - F(\mu_0, \mu_\epsilon)| = 0.$$

The proof of (a) immediately follows from the definitions of the best decoder and the best encoder. The rest of this subsection is devoted to developing an algorithmic proof of (b). Given the matrices Q^* and P^*, we will build a family of measures, μ_ϵ, such that

$$\lim_{\epsilon\to 0} \mu_\epsilon(s|m) = P^*, \quad \lim_{\epsilon\to 0} \mu_\epsilon(m|s) = Q^*. \tag{11.10}$$

This is not a trivial task, which is demonstrated by the following example. Suppose that the P^* and Q^* matrices are given by

$$P^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad Q^* = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

It is clear that we cannot find a measure μ_ϵ which would satisfy conditions (11.10) for this pair (P^*, Q^*). Fortunately, it turns out that situations like this never arise. In order to prove this we will need to consider some auxiliary matrices.

The auxiliary matrix and the absence of loops

Let us define an auxiliary matrix X in the following way:

$$X_{ij} = \begin{cases} 1, & P^*_{ij} + Q^*_{ij} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

This means that the matrix X contains nonzero entries at the slots where either of the matrices, P^* or Q^*, contains a non-zero entry. Now let us draw lines connecting all the "ones" of the X matrix that belong to the same row, and all the "ones" of the X matrix that belong to the same column. We will obtain some (disjoint) graphs. Let us refer to the "ones" of the X matrix as vertices.

Lemma 4 *Suppose that a finite measure μ_0 has the property of unique maxima. Graphs constructed as described above do not contain any closed loops.*

Proof: Let us assume that there exists a closed loop. It looks like a polygon with right angles. Let us consider its "turning points", i.e. those points that simultaneously belong to a horizontal and a vertical line. Suppose there are $2K$ such vertices (this can only be an even number). We will refer to these vertices as x_{α_i,β_j}, where the pair of integers, (α_i, β_j), gives the coordinates of the vertex. Clearly, $1 \leq i, j \leq K$.

Without loss of generality, let x_{α_1,β_1} be connected with x_{α_1,β_2} with a horizontal line. Then x_{α_1,β_2} is connected with x_{α_2,β_2} with a vertical line, $\ldots$, x_{α_K,β_1} is connected with x_{α_1,β_1} with a vertical line, thus closing the loop (see Fig. 11.1, where we used $K = 3$). It is possible to show that exactly a half of the vertices corresponds to "ones" of the P^* matrix, and the rest - to "ones" of the Q^* matrix. If a vertex correspond to a "one" of the Q^* matrix then the corresponding slot of the P^* matrix is zero, and vice versa. This is a direct consequence of the property of unique maxima.

Let us now suppose that $Q^*_{\alpha_1,\beta_1} = 1$, $P^*_{\alpha_1,\beta_1} = 0$ (the alternative is that $P^*_{\alpha_1,\beta_1} = 1$, $Q^*_{\alpha_1,\beta_1} = 0$, in which case the proof remains very similar). This means that $Q^*_{\alpha_1,\beta_2} = 0$, because by construction (see formula (11.8)), there can be only one nonzero element in the same row of the Q^* matrix. Then the element $P^*_{\alpha_1,\beta_2} = 1$, because the corresponding vertex is present in the X matrix. This leads to $P^*_{\alpha_2,\beta_2} = 0$ (we can only have one positive element in each column of the P^* matrix, Eq. 11.9). This argument can be continued around the loop. The Q^* elements along the loop are alternating between 0 and 1, and so are the elements of the P^* matrix, see Fig. 11.1.

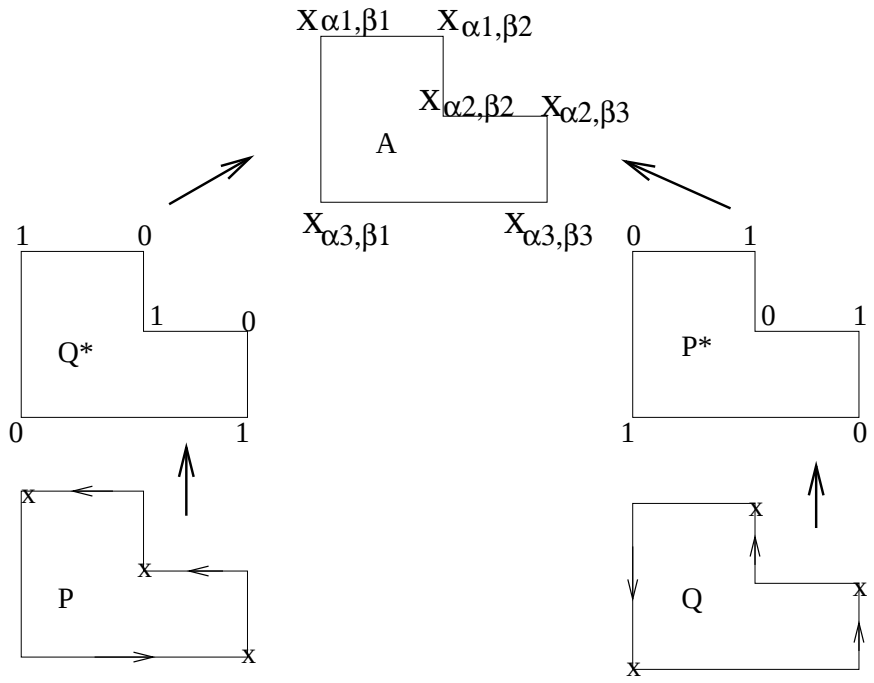

Figure 11.1: No loops in graphs.

We can conclude that $P^0_{\alpha_1,\beta_1} > P^0_{\alpha_1,\beta_2}$, because by construction, positive elements in the Q^* matrix correspond to the largest elements in the corresponding rows of the P^0 matrix. Similarly, we obtain $2K$ inequalities:

$$P^0_{\alpha_i,\beta_i} > P^0_{\alpha_i,\beta_{i+1}}, \tag{11.11}$$

$$Q^0_{\alpha_i,\beta_{i+1}} > Q^0_{\alpha_i,\beta_{i+1}}, \quad 1 \le i \le K \tag{11.12}$$

(here we set $\alpha_{K+1} \equiv \alpha_1$ and $\beta_{K+1} \equiv \beta_1$). In Fig. 11.1, the maximum elements of the rows of P^0 and the columns of Q^0 are marked by crosses. The arrows indicate the direction towards the larger elements.

We will now show that system (11.11-11.12) is incompatible. In order to do this, we write $\mu_0(s_i, m_j) = Q^0_{ij} M_i$, where M_i is the sum of the elements of the ith row of the matrix μ_0: $M_i \equiv \sum_k \mu_0(s_i, m_k)$. Then we can rewrite P^0_{ij} in terms of Q^0 and M:

$$P^0_{ij} = \frac{\mu_0(s_i, m_j)}{\sum_k \mu_0(s_k, m_j)} = \frac{Q^0_{ij} M_i}{\sum_k Q^0_{kj} M_k}.$$

System (11.11-11.12) can be presented as a closed chain of inequalities for Q^0:

$$Q^0_{\alpha_1,\beta_1} > Q^0_{\alpha_1,\beta_2} \frac{\sum_k Q^0_{k\beta_1} M_k}{\sum_k Q^0_{k\beta_2} M_k}, \quad Q^0_{\alpha_1,\beta_2} > Q^0_{\alpha_2,\beta_2}, \tag{11.13}$$

$$Q^0_{\alpha_2,\beta_2} > Q^0_{\alpha_2,\beta_3} \frac{\sum_k Q^0_{k\beta_2} M_k}{\sum_k Q^0_{k\beta_3} M_k}, \quad Q^0_{\alpha_2,\beta_3} > Q^0_{\alpha_3,\beta_3},$$

$$\dots$$

$$Q^0_{\alpha_i,\beta_i} > Q^0_{\alpha_i,\beta_{i+1}} \frac{\sum_k Q^0_{k\beta_i} M_k}{\sum_k Q^0_{k\beta_{i+1}} M_k}, \quad Q^0_{\alpha_i,\beta_{i+1}} > Q^0_{\alpha_{i+1},\beta_{i+1}},$$

$$\dots$$

$$Q^0_{\alpha_K,\beta_K} > Q^0_{\alpha_K,\beta_1} \frac{\sum_k Q^0_{k\beta_K} M_k}{\sum_k Q^0_{k\beta_1} M_k}, \quad Q^0_{\alpha_K\beta_1} > Q^0_{\alpha_1,\beta_1}. \tag{11.14}$$

From the first two inequalities we know that $Q^0_{\alpha_1,\beta_1} > Q^0_{\alpha_2,\beta_2} \frac{\sum_k Q^0_{k\beta_1} M_k}{\sum_k Q^0_{k\beta_2} M_k}$, then using the next pair we similarly derive that $Q^0_{\alpha_1,\beta_1} > Q^0_{\alpha_3,\beta_3} \frac{\sum_k Q^0_{k\beta_1} M_k}{\sum_k Q^0_{k\beta_3} M_k}$. Continuing along the chain, at the Kth step we have $Q^0_{\alpha_1,\beta_1} > Q^0_{\alpha_K,\beta_K} \frac{\sum_k Q^0_{k\beta_1} M_k}{\sum_k Q^0_{k\beta_K} M_k}$. Using the last two inequalities, we finally obtain: $Q^0_{\alpha_1,\beta_1} > Q^0_{\alpha_1,\beta_1}$. This contradiction proves that there can be no closed loops in the matrix X. ■

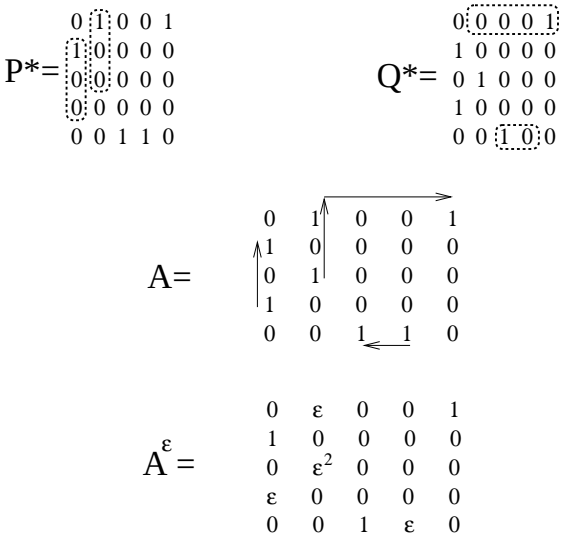

Figure 11.2: Construction of A^ϵ for Example (11.3.1). We first form P^0 and Q^0 matrices by normalizing columns and rows of μ_0 respectively; this step is not shown here. Then we can construct the best encoder, P^*, by identifying the maximal elements in the columns of Q^0, and the best decoder, Q^*, by identifying the maximal elements in the rows of P^0, see the top of the figure. Next, we combine the positive elements (or vertices) of P^* and Q^* to create the auxiliary matrix X. The vertices of X that belong to the same column (row) are connected. In order to define the direction of the arrows, we have to refer to the matrices P^* and Q^*. If two vertices are connected by a vertical line, we find the corresponding elements of the P^* matrix (they are encircled); the direction of the arrow is always towards the "one" of the P^* matrix. Similarly, if two vertices are connected by a horizontal line, we find the corresponding elements of the Q^* matrix (encircled) and direct the arrow towards the "one" of the Q^* matrix. Finally, we build the A^ϵ matrix by replacing the "ones" of the X matrix by powers of ϵ. The powers of ϵ must be arranged in such a way that in each of the connected graphs, the arrows point from a smaller entry to a larger entry. Note that P^* and Q^* are not related by a common measure, i.e., no measure exists such that P^* and Q^* are its conditional probabilities in the sense described in this chapter.

Constructing the matrix μ_ϵ

Now we can systematically build the matrix μ_ϵ. From Lemma 4 it follows that if we connect all the vertices of the matrix X by horizontal and vertical lines, the resulting (disjoint) graphs will contain no closed loops. Some of the graphs may only consist of one vertex.

For each of these graphs we will perform the following procedure. Take a pair of vertices. If they are connected by a horizontal (vertical) line, refer to the corresponding entries of the Q^* matrix (P^* matrix). One of them will be one and the other - zero. Draw an arrow on the graph from the element corresponding to zero to the element corresponding to one. Repeat this for all pairs of vertices. Next, starting from some vertex, replace the corresponding element in the X matrix by ϵ, and then, following the arrows, keep replacing the elements of X by entries of the form ϵ^k, where the integer k increases or decreases from one vertex to the next depending on the direction of the arrow (we can always do this because by Lemma 4, there are no closed loops in the graphs of matrix X). We will call the resulting matrix A^ϵ. The measure μ_ϵ is obtained by re-normalizing the elements of the matrix A^ϵ:

$$\mu_\epsilon(s_i, m_j) = A^\epsilon_{ij} / \sum_{k,l} A^\epsilon_{kl}. \tag{11.15}$$

Remark: In the algorithm above we used powers of the small parameter ϵ, ϵ^k, to assign to vertices of the matrix X. More generally, one can use any functions of ϵ, $f_k(\epsilon)$, such that $\lim_{\epsilon\to 0} f_k(\epsilon)/f_{k+1}(\epsilon) = 0$. Thus, the family μ_ϵ found above is just one of many such families.

Proof of Theorem 24

We are now ready to complete the proof of Theorem 24, part (b).

Proof: Let us show that Eq. 11.10 holds. In order to find entries of $\mu_\epsilon(s|m)$, we need to re-normalize each column of the matrix μ_ϵ so that its elements sum up to one. Obviously, each column will contain at most one segment of one of the graphs. By construction, the biggest element of this segment of the graph corresponds to the positive element of Q^*. In the limit $\epsilon \to 0$, the other elements will be vanishingly small in comparison with the biggest one, and the resulting column of the $\mu_\epsilon(s|m)$ matrix will be identical to the corresponding column of the P^* matrix. The same argument holds for rows of the $\mu_\epsilon(m|s)$ matrix which in the limit become the rows of the Q^* matrix. Thus we conclude that the algorithm of Section 11.3.1 leads to constructing a family of measures μ_ϵ which satisfy the requirements of Theorem 24. ■

Example Consider the following 5×5 matrix:

$$\mu_0 = \frac{1}{1245} \begin{pmatrix} 1 & 64 & 2 & 23 & 90 \\ 92 & 8 & 42 & 81 & 42 \\ 53 & 77 & 60 & 2 & 50 \\ 88 & 15 & 68 & 73 & 59 \\ 39 & 48 & 66 & 65 & 37 \end{pmatrix}. \tag{11.16}$$

For this language, $\sup_\mu F(\mu_0, \mu) = 394/6225$. In Fig. 11.2 we show the calculated P^* and Q^* matrices, and then construct the X and the A^ϵ matrices. The family μ_ϵ is given by

$$\mu_\epsilon = \frac{1}{3(1+\epsilon)+\epsilon^2} \begin{pmatrix} 0 & \epsilon & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & \epsilon^2 & 0 & 0 & 0 \\ \epsilon & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \epsilon & 0 \end{pmatrix}.$$

Remark:
As $\epsilon \to 0$, $F(\mu_0, \mu_\epsilon) \to \sup_\mu F(\mu_0, \mu)$. If we let $\mu_* = \mu_\epsilon|_{\epsilon=0}$, i.e., μ_ϵ evaluated at 0, we note that $\mu_* \neq \mu_0$ in general. Further, $F(\mu_0, \mu_*) < \sup_\mu F(\mu_0, \mu)$. Thus, we have that $\lim_{\epsilon\to 0} \mu_\epsilon = \mu_*$ yet $F(\mu_0, \mu_*) < \lim_{\epsilon\to 0} F(\mu_0, \mu_\epsilon) = \sup_\mu F(\mu_0, \mu)$. This is a consequence of a discontinuity in the definition of the communicability function, $F(L_1, L_2)$. In particular, the conditional probabilities entering definition (11.4) are discontinuous when all the elements of a column or a row of μ are zero, see Equations (11.1)-(11.2). Thus the value of $F(\mu_0, \mu_\epsilon)$ may have a jump at $\epsilon = 0$.

11.3.2 Generalizations

In our discussion in this chapter, we worked with three restrictions on the measures μ. These were (i) the property of unique local maxima in the rows and columns of $\mu(s|m)$ and $\mu(m|s)$ respectively (ii) uniform distribution of events in the world that need to be communicated (iii) finite cardinality of $\mathcal{S}$ and $\mathcal{M}$. Each of these can be dropped. A proper discussion will lead us to technical details that are beyond the scope of this book and so we omit them here. We say just a few words about each of these issues and the interested reader may find a complete treatment in Komarova and Niyogi (2004).

If there are multiple maxima, it turns out that loops may exist and lemma 4 needs to be modified. Fortunately, this can be done and the notion

of a neutral vertex of a graph needs to be introduced. If events do not occur with uniform probability, one simply redefines P^* and Q^* accordingly. For example, Q^* is redefined as

$$Q^*_{ij} = \begin{cases} 1, & \mu_0(s_i|m_j)\sigma_j = \max_p \mu_0(s_i|m_p)\sigma_p, \\ 0, & \text{otherwise,} \end{cases}$$

and P^* is similarly redefined. Everything else works out as before. Finally, to go from finite matrices to infinite matrices, one proceeds by noting that a measure μ on a countably infinite space can be approximated arbitrarily closely by a measure with finite support. This reduces the infinite case to the finite case and the results discussed here apply immediately.

11.4 Implications for learning

From the preceding discussion it is now clear that in order to maximize mutual intelligibility with a language user (characterized by the measure μ), it may be necessary to use a different measure, μ_*, where $\mu_* \neq \mu$. This fact has implications both for learning and evolution of populations of linguistic agents.

Let us first consider the problem of an agent trying to learn a language in order to communicate with some other agent whose language is characterized by the measure μ. Recall that μ_* (the best response) itself may not exist, however, an arbitrarily close approximation μ_ϵ (for any ϵ) does exist. Therefore the best the learner can do is estimate μ_ϵ. What degree of accuracy, ϵ, is useful or necessary will depend upon the particular application in mind. Since the measure μ is unknown to the learner at the outset, there are two natural learning scenarios depending upon how much information is available to the learner on each interaction.

1. *Full Information:* This corresponds to the situation where the learner is able to sample μ directly to get (sentence, meaning) pairs. Thus, when the teacher speaks, both sentence and meaning are directly accessible. The strategy of the learner is to estimate μ as well as it can, derive from it the P^* and Q^* matrices, and ultimately estimate μ_ϵ using the procedure described in the previous sections.

2. *Partial Information:* In most natural settings, however, the meaning may not be directly accessible. In other words, the learner only

hears the sentence while the intended meaning is latent. What the learner reasonably may have access to is whether its interpretation of the sentence was successful or not. On the basis of this information, the learner must somehow derive the optimal communication strategy. We refer to this as learning with partial information.

Thus we see that (1) full information and (2) partial information suggest two different frameworks for learning; in either case, the learner has to estimate P and Q matrices of the teacher.

11.4.1 *Estimating P*

An important task for the learner is to estimate Q^* which is derivable from the P matrix of the teacher. Recall that

$$Q^*_{ij} = \begin{cases} 1, & \mu(s_i|m_j) = \max_p \sigma_p \mu(s_i|m_p), \\ 0, & \text{otherwise.} \end{cases}$$

Learning with full Information

The learner, in this case, has access to (s, m) (sentence,meaning) pairs every time the teacher produces a sentence. We can define the event

$$A_{ij} = \text{Teacher produces } s_i \text{ to communicate } m_j.$$

The probability of event A_{ij} is simply $\sigma_j \mu(s_i|m_j)$. Therefore, if the teacher produces n (sentence, meaning) pairs at random in i.i.d. manner, then the ratio

$$\hat{a}_{ij}(n) = \frac{k_{ij}}{n}$$

is an empirical estimate of the probability of the event A_{ij}. By the law of large numbers, as $n \to \infty$ we have

$$\hat{a}_{ij}(n) \to \sigma_j \mu(s_i|m_j)$$

with probability 1. For the case under consideration, we can even bound the rate at which this convergence occurs. For example, applying Hoeffding's inequality, we have

$$\mathbb{P}[|\hat{a}_{ij}(n) - \sigma_j \mu(s_i|m_j)| > \epsilon] \leq 2e^{-\frac{\epsilon^2 n}{2}}.$$

This convergence is guaranteed for fixed (i, j). In general, the learner must estimate a collection of events. The total number of events are given by the total number of possible (sentence,meaning) pairs. As before, let us assume that there are N possible sentences and M possible meanings. Therefore, there are NM different events whose probabilities need to be estimated. The collection of events A_{ij}, $i = 1, \ldots, N$; $j = 1, \ldots, M$, are disjoint. For a finite collection of such events, we will derive a uniform law of large numbers.

Let event E_{ij} be

$$E_{ij} = |\hat{a}_{ij}(n) - \sigma_j \mu(s_i|m_j)| > \epsilon.$$

Then, by the union bound, we obtain

$$\mathbb{P}[\cup_{i,j} E_{ij}] \leq \sum_{i,j} \mathbb{P}(E_{ij}) \leq NM 2e^{-\frac{\epsilon^2 n}{2}}.$$

Therefore, we have

$$\mathbb{P}[\overline{\cup_{i,j} E_{ij}}] = \mathbb{P}[\forall i, j | \hat{a}_{ij}(n) - \sigma_j \mu(s_i|m_j)| \leq \epsilon] > 1 - NM 2e^{-\frac{\epsilon^2 n}{2}}.$$

Thus, with high probability (depending upon the number of examples, n) all empirical estimates $\hat{a}_{ij}(n)$ are close to $\sigma_j \mu(s_i|m_j)$ respectively. Estimating the $\sigma_j \mu(s_i|m_j)$'s is the step to estimating the Q^* matrix that is required for the optimal communication system.

Learning with partial information

Now consider the setup in (2) where the learner has no access to the meaning directly but has to guess a meaning and is told after the event whether the guess was correct or incorrect. Thus the learner has access to asymmetric information: if the guess was correct, the learner knows the true intended meaning; if the guess was incorrect, the learner merely knows what the meaning was not. As it turns out, this does not dramatically change the state of affairs. To see this, let the learner guess a meaning uniformly at random. Thus with probability $\frac{1}{M}$ the learner chooses a meaning m_j. Each time the teacher produces a sentence, the intended meaning may be successfully communicated or not. Define the event

A_{ij} = Teacher produces s_i; Learner guesses m_j; Communication is successful.

The probability of event A_{ij} is simply $\frac{1}{M}\sigma_j \mu(s_i|m_j)$. The event A_{ij} is observable since the learner knows (i) what sentence has been uttered by the

teacher, (ii) what meaning it (the learner) assigned to the sentence, and (iii) whether communication was successful. Therefore after n sentences have been produced by the teacher, the learner can count k_{ij} – the number of times event A_{ij} has occurred, and can make an empirical estimate of the probability of A_{ij} as

$$\hat{a}_{ij}(n) = \frac{k_{ij}}{n}.$$

By the same argument as before, $\hat{a}_{ij}(n)$ converges in probability to $\frac{1}{M}\sigma_j\mu(s_i|m_j)$ and the rates are provided by the Hoeffding bounds. Since M is fixed in advance and known, this allows the learner to guess $\sigma_j\mu(s_i|m_j)$ for each i, j arbitrarily well. Let us be a little more precise about the rates of convergence. The learner's estimate of $\sigma_j\mu(s_i|m_j)$ is really $M\hat{a}_{ij}$ where $\hat{a}_{ij}$ is defined above. Therefore we have that

$$\mathbb{P}[|M\hat{a}_{ij} - \sigma_j\mu(s_i|m_j)| > \epsilon] = \mathbb{P}[|\hat{a}_{ij} - \frac{1}{M}\sigma_j\mu(s_i|m_j)| > \frac{\epsilon}{M}] \leq 2e^{\frac{-\epsilon^2 n}{2M^2}}.$$

Thus the confidence in the ϵ-good estimate of $\sigma_j\mu(s_i|m_j)$ is poorer than before. By the same argument as in case (2), we have a uniform bound as follows:

$$\mathbb{P}[\forall i, j | M\hat{a}_{ij} - \sigma_j\mu(s_i|m_j)| \leq \epsilon] > 1 - NM2e^{\frac{-\epsilon^2 n}{2M^2}}. \tag{11.17}$$

11.4.2 Estimating Q

Let us now consider the task of estimating P^* which is derivable from the Q matrix of the teacher. The same arguments of the previous section apply. Recall that

$$P^*_{ij} = \begin{cases} 1, & \mu(m_j|s_i) = \max_p \mu(m_j|s_p), \\ 0, & \text{otherwise.} \end{cases}$$

Learning with full information. Here the learner has direct access to the meaning assigned by the teacher to each sentence. Therefore, the learner need only pick a sentence uniformly at random (with probability $\frac{1}{N}$) and produce it for the teacher to hear. Let us define the event

$$A_{ij} = \text{Learner produces } s_i; \text{ Teacher interprets as } m_j.$$

The event A_{ij} is observable on each trial. The probability with which it occurs is given by $\frac{1}{N}\mu(m_j|s_i)$. After n trials (where the learner speaks in

this manner), the learner simply counts the number k_{ij} of times event A_{ij} occurs and its estimate of $\frac{1}{N}\mu(m_j|s_i)$ is $\frac{k_{ij}}{n}$. Therefore, we have

$$\mathbb{P}[|\hat{a}_{ij} - \frac{1}{N}\mu(m_j|s_i)| > \epsilon] \leq 2e^{\frac{-\epsilon^2 n}{2N^2}}.$$

Using the same arguments as before, we have

$$\mathbb{P}[\forall i,j|MN\hat{a}_{ij} - \mu(m_j|s_i)| \leq \epsilon] > 1 - NM2e^{\frac{-\epsilon^2 n}{2N^2}}.$$

Learning with partial information. The learner simply picks a (sentence, meaning) pair uniformly at random (with probability $\frac{1}{NM}$). Define the event

$$A_{ij} = \text{Learner produces } (s_i, m_j);\ \text{Communication is successful.}$$

The event A_{ij} is observable by the learner on each trial. The probability of event A_{ij} is $\frac{1}{NM}\mu(m_j|s_i)$. After n trials (where the learner speaks), the learner counts the number k_{ij} of times event A_{ij} occurs. Therefore, we again have

$$\mathbb{P}[|\hat{a}_{ij} - \frac{1}{NM}\mu(m_j|s_i)| > \epsilon] \leq 2e^{\frac{-\epsilon^2 n}{2M^2N^2}}.$$

Using the same arguments as before, we have

$$\mathbb{P}[\forall i,j|MN\hat{a}_{ij} - \mu(m_j|s_i)| \leq \epsilon] > 1 - NM2e^{\frac{-\epsilon^2 n}{2M^2N^2}}. \tag{11.18}$$

11.4.3 Sample Complexity Bounds

Now we can put the pieces together to determine the number of learning events that need to occur so that with high probability, the learner will be able to develop a language with ϵ-good communicability. Let the teacher's measure be μ. We will assume that μ is such that the P and Q matrices have unique row-wise and column-wise maxima respectively. First let us introduce the *margin* by which the maximum value clears all other values in the row and column respectively. This *margin* will play an important role in determining the number of learning events.

Definition 19 *For each i, let $j_i^* = \arg\max_j \sigma_j\mu(s_i|m_j)$ and for each j, let $i_j^* = \arg\max_i \mu(m_j|s_i)$. Then, we define the margin γ to be the largest real number such that*

$$\sigma_{j_i^*}\mu(s_i|m_{j_i^*}) \geq \sigma_j\mu(s_i|m_j) + \gamma \quad \forall j \neq j_i^*$$

and

$$\mu(m_j|s_{i_j^*}) \geq \mu(m_j|s_i) + \gamma \quad \forall i \neq i_j^*$$

Learning with partial information. We have described how to estimate the Q^* and P^* matrices; the following theorem provides a bound on the number of examples needed to ensure correct estimates:

Theorem 25 *If the total number, n, of interactions between teacher and learner (with partial information) is greater than $\frac{64M^2N^2}{\gamma^2}\log(\frac{4MN}{\delta})$, then with high probability $> 1-\delta$, the learner can construct a measure that will give arbitrarily good communicability with the teacher.*

Proof: Let there be $n/2$ interactions where the teacher speaks and the learner listens and $n/2$ interactions of the other form. The learner constructs estimates of $\sigma_j\mu(s_i|m_j)$ and $\mu(m_j|s_i)$ in the manner described previously. Let the estimates be denoted by $\hat{p}_{ij}$ and $\hat{q}_{ij}$ respectively. By setting $\epsilon = \gamma/4$ in Equations 11.17 and 11.18, we obtain:

$$\mathbb{P}[\forall i,j|\hat{p}_{ij} - \sigma_j\mu(s_i|m_j)| \leq \gamma/4] > 1 - 2NMe^{\frac{-\gamma^2 n}{64M^2}}$$

and

$$\mathbb{P}[\forall i,j|\hat{q}_{ij} - \mu(m_j|s_i)| \leq \gamma/4] > 1 - 2NMe^{\frac{-\gamma^2 n}{64M^2N^2}}.$$

Using the fact that $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$, we can see that with probability greater than $1 - 2NM\left(e^{\frac{-\gamma^2 n}{64M^2N^2}} + e^{\frac{-\gamma^2 n}{64M^2}}\right)$, the estimates $\hat{p}_{ij}$ and $\hat{q}_{ij}$ are both within $\gamma/4$ of the true values. The learner chooses Q^* and P^* using the estimated matrices. Let us first consider the case of Q^*. For each i the learner desires to obtain j_i^* given by

$$j_i^* = \arg\max_j \sigma_j\mu(s_i|m_j).$$

The learner chooses

$$\hat{j}_i = \arg\max_j \hat{p}_{ij},$$

and we claim that $\hat{j}_i = j_i^*$. In order to prove this, assume that this is not the case. Then we get immediately:

$$\sigma_{j_i^*}\mu(s_i|m_{j_i^*}) \geq \sigma_{\hat{j}_i}\mu(s_i|m_{\hat{j}_i}) + \gamma.$$

However, we have the following chain of inequalities:

$$\sigma_{\hat{j}_i}\mu(s_i|m_{\hat{j}_i}) \geq \hat{p}_{i\hat{j}_i} - \gamma/4 \geq \hat{p}_{ij_i^*} - \gamma/4 \geq \sigma_{j_i^*}\mu(s_i|m_{ij_i^*}) - \gamma/2,$$

which leads to a contradiction. This argument holds for every i, therefore, since $\hat{j}_i = j_i^*$ for each i, the Q^* matrix is identified exactly. Similarly, one can show that the P^* matrix is also identified exactly.

The only thing that remains is to ensure that n is large enough so that this occurs with high probability. We have

$$2NM\left(e^{\frac{-\gamma^2 n}{64M^2N^2}} + e^{\frac{-\gamma^2 n}{64M^2}}\right) \leq 4NMe^{\frac{-\gamma^2 n}{64M^2N^2}} \leq \delta.$$

This is satisfied for $n > \frac{64M^2N^2}{\gamma^2}\log(\frac{4MN}{\delta})$. Thus, with probability greater than $1-\delta$, both P^* and Q^* are identified exactly. Now the procedure of approximating the measure may be applied. ■

Remarks:

1. The number of examples is seen to be a function of M, N and γ. The margin γ that depends upon the teacher's language, μ, determines, in some sense, how easy it is to estimate Q^* and P^* matrices for the learner. It therefore characterizes the learning difficulty of μ in this setting.

2. Infinite matrices are not learnable. In fact, infinite dimensional spaces are known to be unlearnable (see Vapnik, 1998) and therefore further constraints will be required on the space of possible measures to which the teacher's language belongs. Such constraints are presumably embodied in the universal grammar (Chomsky, 1986) constraints that learning agents bring to the task.

3. The constants in the bound on sample complexity may be tightened, although the order is essentially correct. For example, we have let the interactions be symmetric, i.e., the numbers of sentences the learner produces and receives are the same. It is easy to check that a more favorable bound is obtained when the learner speaks N^2 times as often as it listens. In this case, it is enough to have $\frac{32M^2(N^2+1)}{\gamma^2}\log(\frac{4MN}{\delta})$ interactions in all.

Learning with full information. For completeness, let us state the number of interactions needed to learn in setting (1). This is given by the following

Theorem 26 *If the total number, n, of interactions between teacher and learner (with full information) is greater than $\frac{64N^2}{\gamma^2}\log(\frac{4MN}{\delta})$, then with high probability $> 1-\delta$, the learner can construct a measure that will give arbitrarily good communicability with the teacher.*

The proof is very similar to that of the previous theorem and we omit it for this reason. It is noteworthy that learning with full information requires M^2 less interactions to learn. This is not surprising since the meanings are accessible, and the larger is the number, M, of different concepts, the greater is the difference between learning with full and partial information.

11.5 Communicative Efficiency and Linguistic Structure

There have often been discussions about the role of communicative efficiency in the structure and possibly the evolution of language. Functionalist perspectives on linguistics hold that the primary function of language is communication or transfer of information from speaker to hearer. Furthermore, it is argued that such communicative pressures shape the structure of language so that over time, the structure of language adapts to facilitate the transfer of information. While there is undoubtedly some merit to this point of view, one must examine the empirical evidence carefully to tease apart the domains in which communicative efficiency may play an explanatory role and those in which it does not.

In this section, we provide an empirical study of the structure of lexical items that suggests that a tight coupling between communicative efficiency and lexical structure may not always be present. More large scale studies of this sort are necessary to fully flesh out this issue in other domains.

11.5.1 Phonemic Contrasts and Lexical Structure

Consider the finite number of words $w_1, \ldots, w_n$ in English. Each word w_i may be viewed as a string of phonemes[3]. Now consider *spoken* communication between a speaker and hearer of English. Assume the speaker produces

[3]More generally, modern linguistic theory views each phoneme to be a complex of *distinctive features*. The segmental phonology of *Sound Pattern of English* (Chomsky and Halle, 1968) decomposes each word into a sequence of feature bundles and phonological operations are defined on feature spaces rather than phoneme spaces. More recently, autosegmental phonology (Goldsmith, 1976 and later work) considers these features to be arranged in an overlapping pattern along multiple tiers rather than along a linear time

a single word. If all the phonemes in the sequence are heard correctly by the hearer, then the word has been successfully transmitted from speaker to hearer. If the speech perception apparatus of the speaker is error-free, then all words (barring homophones) can be distinguished from each other and communication would be perfect. Communicative efficiency would be high.

Now imagine that the hearer can tell all phonemes apart except /*p*/ and /*b*/, i.e., the hearer cannot distinguish between voiced and unvoiced labial stop consonants. As a result, the hearer would not be able to tell apart words that differ by exactly this phonemic contrast — in this case, the voicing feature for labial stop consonants. For example, the hearer would not be able to tell apart the words "pat" (/*pat*/) and "bat" (/*bat*/) or "pit" (/*pit*/) and "bit" (/*bit*/) and so on. Information would no longer be perfectly transmitted from speaker to hearer. By considering all the words in the lexicon that rely on this distinction, one can try to quantify how much information is lost on the whole. Let us do this now.

A natural measure of uncertainty and information content is provided by entropy. If $p_1, \ldots, p_n$ are the probabilities with which the n words are used on average, then the entropy of the entire lexicon is given by

$$H(W) = \sum_i p_i \log(\frac{1}{p_i})$$

To appreciate this measure, note that if all words were equally likely, then

$$H(W) = \sum_{i=1}^{n} \frac{1}{n} \log(n) = \log(n)$$

Thus, the information content of the lexicon grows with the size of the lexicon. The entropic measure accounts for the fact that all words may not be equally likely. If a particular set of three words is used almost all the time and the rest of the words are hardly ever used then it is informally clear that the effective size of this lexicon is closer to 3 that it is to n. The entropic formulation $H(W)$ captures this distinction naturally. $H(W)$ may be regarded as an average measure of the information transmitted from speaker to hearer by transmitting words of the lexicon.

If the phonemic distinction between /*b*/ and /*p*/ collapses, then the lexicon W may be partitioned into *cohorts* (homophone classes after the phonetic collapse) of words that are indistinguishable within each cohort.

sequence. The results presented here may be re-expressed in terms of features rather than phonemes in both segmental and autosegmental frameworks.

The hearer can only distinguish cohorts from each other and the size of the lexicon as perceived by the hearer reduces to the number of cohorts. Let $c_1(\{/p/, /b/\}), \ldots, c_k(\{/p/, /b/\})$ be the k cohorts obtained by losing the distinction between $/b/$ and $/p/$. Let us denote this reduced lexicon as $W(\{/p/, /b/\}) = \{c_1(\{/p/, /b/\}), \ldots, c_k(\{/p/, /b/\})\}$.

The probability with which the hearer will encounter a word that belongs to $c_1(\{/p/, /b/\})$ is given by

$$P_1 = \sum_{w_i \in c_1(\{/p/, /b/\})} p_i$$

Similarly, the probability for each of the other cohorts is calculated as P_i for the ith cohort. The information content of this reduced lexicon is given by

$$H(W(\{/p/, /b/\})) = \sum_i P_i \log(\frac{1}{P_i})$$

The normalised loss of information is given by

$$FL(W(\{/b/, /p/\}) = \frac{H(W) - H(W(\{/p/, /b/\})}{H(W)} \tag{11.19}$$

This is a quantitative measure of the *functional load* carried by the ability to distinguish between $/p/$ and $/b/$. It is easy to check that $0 \leq FL \leq 1$. Thus FL is the fraction of information lost (at the lexical level) by losing the ability to distinguish between $/p/$ and $/b/$. It is the lexical work this phonetic contrast does for the language in question.

11.5.2 Functional Load and Communicative Efficiency

The lexical entropy $H(W)$ is an information theoretic measure of how much information is transmitted on average from speaker to hearer via the lexicon. It is therefore correlated with the communicative efficiency of the lexicon of the language. The functional load $FL(W(\{/p/, /b/\}))$ is a relative measure of how much information is lost as a result of not being able to make the distinction between $/p/$ and $/b/$.

It is clear that for every pair of phonemes $/\pi_i/$ and $/\pi_j/$ one can define the functional load of that phonetic contrast. The functional load is therefore correlated with the loss of communicative efficiency because of being unable to make the phonetic contrast.

In previous sections, we have developed the notion of communicative efficiency in terms of the probability with which the speaker and hearer are able to understand each other on average. One may also formulate functional load in these terms. For example, if all words can be distinguished from each other, then the probability of successful transmission of lexical items is equal to 1. On the other hand, if /b/ and /p/ cannot be distinguished, then what is the new probability of successful transmission? This will depend upon the precise nature of the listener's guessing strategy.

When a speaker produces a particular word, the listener will only be able to perceptually identify the cohort to which the word belongs. Thereafter, the listener will have to use some other strategy to guess a particular word in the cohort. Let us assume a randomized strategy where for each word w_i in cohort c_j, the learner guesses that word with a probability proportional to $f(p_i)$ where p_i is the probability with which the word is used in general. Different choices of f may be made. For example, one could have (i) $f(p_i) = p_i$, (ii) $f(p_i) = \log(p_i)$, (iii) $f(p_i) = 1$(else 0) $\Leftrightarrow p_i > p_k \forall k$ such that $w_k, w_i \in c_j$.

Then the probability of transmission is given by

$$t = \sum_{i=1}^{k} \sum_{w_j \in c_i} p_j g(p_j).$$

Here $g(p_i) = \frac{f(p_i)}{\sum_{\{k|w_i, w_k \in c_j\}} f(p_k)}$. This may be computed for every contrast and provides a measure of communicative efficiency of the contrast. The probability of error is given by $e = 1 - t$ and is a measure of the loss of communicative efficiency as a result of being unable to make the contrast. Unfortunately, the exact numerical value of this quantity depends upon the choice of the functional form f and for this reason, we use Eq. 11.19 in the rest of this discussion.

It is worth noting, however, that all these characterizations of communicative efficiency are correlated with each other. To see the simplest example of this, let us again assume that all n words in the lexicon are equally likely. Then the communicative efficiency may be calculated as

$$t = \sum_{i=1}^{k} \sum_{w_j \in c_i} \frac{1}{n} g(p_j = \frac{1}{n}) = \sum_{i=1}^{k} \frac{1}{n} \sum_{w_j \in c_i} g(p_j = \frac{1}{n}) = \frac{k}{n}$$

Therefore the loss in communicative efficiency is

$$e = 1 - t = 1 - \frac{k}{n}$$

Thus if there are n cohorts, then each cohort consists of a unique word and there is no loss in communicative efficiency. If there is only one cohort, then the only communicative efficiency that remains is through random guessing. For large lexical sizes, the loss is almost 1.

Now consider the information-theoretic measure of functional load. For cohort c_i, one may calculate $P_i = \frac{|c_i|}{n}$ where $|c_i|$ is the size of the ith cohort, i.e., the number of words in that cohort. Then it is easily seen that $H(W) = \log(n)$, $H(W(\{/p/, /b/\})) = \sum_{i=1}^{k} \frac{|c_i|}{n} \log(\frac{n}{|c_i|})$, and

$$FL = \frac{1}{n \log(n)} \sum_{i=1}^{k} |c_i| \log(|c_i|)$$

It is possible to provide a range in which FL must lie. By Jensen's inequality, we have that

$$\sum_{i=1}^{k} \frac{|c_i|}{n} \log(\frac{n}{|c_i|}) \leq \log(k)$$

from which we get $FL \geq 1 - \frac{\log(k)}{\log(n)}$. On the other hand, because there are n words distributed among k cohorts, by a version of the pigeonhole principle, we have

$$|c_i| \leq n - k + 1 \ \forall i$$

Applying this, we have $FL \leq \frac{\log(n-k+1)}{\log(n)}$. Thus

$$1 - \frac{\log(k)}{\log(n)} \leq FL \leq \frac{\log(n-k+1)}{\log(n)}$$

Thus we see that both FL and loss of communicative efficiency decrease as a function of the number of cohorts k.

11.5.3 Perceptual Confusibility and Functional Load

Speakers of a language have some choice regarding the lexical items they wish to use. In general, for optimal communication, it is advantageous to have as few homophones as possible and be able to distinguish as many words as possible.

Therefore, if a particular phonetic contrast is difficult to make perceptually, then an optimally structured lexicon should not rely heavily on making this distinction. For example, if it is very hard to distinguish /m/ (labial

nasal consonant) from /n/ (alveolar nasal consonant), it would seem suboptimal and inefficient to have a whole lot of words in the language that differ from each other by exactly this single distinction. Confusion would be rampant.

In other words, if a phonetic contrast is difficult to make, it should have low functional load. On the other hand, if a contrast is easy to make, it should be utilized in developing lexical distinctions. Thus if communicative efficiency played a role in the evolution of linguistic structure, one should observe a correlation between the perceptual difficulty of making a phonetic contrast and the functional load of that contrast.

Data was collected from several languages (Dutch, English, and Chinese) to examine this question. The perceptual confusibility between phonemes is a psychoacoustic property that depends upon the similarity between the acoustic realizations of those phonemes and the ability of the brain and its perceptual mechanisms to discriminate between them based on their acoustic differences. Psychophysical data on phoneme confusion matrices may be obtained for several languages from the long tradition of research in experimental psycholinguistics. Lexical data where lexical items are annotated with citation form and colloquial pronunciation patterns, frequency of usage, semantic and syntactic information was obtained from the more recent tradition of research in corpus based linguistics.

Surprisingly, we find in all of these languages that there is no significant correlation between functional load and confusibility. The lexicon is therefore not optimally adjusted to our perceptual apparatus. The design of the lexicon compromises on its communicative efficiency.

DUTCH, ENGLISH, CHINESE figures.

These results are robust. The immediate objection that may be raised is that other contextual cues help in identifying the word uniquely — cues that a purely perception based account does not take into account appropriately. To counter this, one may consider subcategories of words based on other attributes and repeat the above experiment on these subcategories. Thus, one may examine words by syntactic category, semantic class (gleaned from the WordNet Project; Fellbaum, 1998), stress patterns and syllable structure, and this lack of correlation remains. Similarly, one may factor in the additional information provided by visual cues in addition to the acoustic ones in perceptual consfusibility, and the lack of correlation remains. Rigorous empirical tests on these issues may be found in Surendran (2003).

What do we conclude from this? It points to the fact that in this particular context, the structure of the lexicon does not display any sign of

having been optimized to suit the perceptual limitations of humans. There are several possible interpretations of this empirical result.

First, it points to the possibility that over historical time scales, communicative efficiency might play little role in the structure of natural languages as they are today. In studies of language evolution and historical linguistics, one is tempted to imagine a scenario where a proto-language originates and then evolves over historical time scales to adapt itself to the communicative needs of humans. If one follows this logic, one would expect that the structure of the phonetic inventory and the structure of the lexicon would co-evolve over historical time scales to yield better adapted lexicons than the ones that are currently attested in modern languages.

Second, it may well be that factors other than the one considered here may need to be incorporated in a significant way into a proper formulation of functional load or communicative efficiency. For example, an important factor that we have not discussed is ease of production. While a lot of mathematical work exists on speech production in general, there are no quantitative characterizations of the ease of production for which empirical data is available. One way around this might be to consider suitable proxies such as developmental data about when different phonemes arise in child speech. We leave this as an open question.

Third, it may be that internal optimization of linguistic interfaces rather than external optimal optimization of communicative efficiency is the key fact driving change and evolution. For example, a modular view of linguistics supported in the generative tradition (see Chomsky, 1995) suggests that the interfaces between the various modules of a linguistic system (phonological, syntactic, semantic) and the conceptual-intensional system of thought need to be well matched for smooth working of the system as a whole. It is these interfaces that may be optimized with respect to each other. Thus, a functionalist perspective on this might suggest that the primary function of language is to aid thought rather than communication. Seeking evidence for optimality in communication may prove to be futile.

These caveats underscore the importance of constantly relating models and theories to empirical facts and phenomena. With that sobering thought, let us continue.

Chapter 12

The Origin of Communicative Systems: Linguistic Coherence and Communicative Fitness

In this chapter, we consider a population of interacting linguistic agents where communicative efficiency provides Darwinian fitness that translates into reproductive success. The basic framework differs from the models in part III in three important respects. First, we assume here that the rate at which an agent produces offspring depends upon its communicative efficiency with the rest of the population at large. This is a particular interpretation of how the forces of natural selection might operate in a communicative setting. In keeping with the fundamental framework of this entire book, we assume that the precise language of the parent is *not* genetically transmitted to the child. Rather, the child will have to learn this language on the basis of linguistic examples. However, unlike most of the models of part III, the parent is assumed to be the dominant source of primary linguistic data for the child — in other words, the child learns mostly from the parents. Finally, while all the models of part III resulted in iterated maps that assumed a generational structure, here we will work with differential equations that provide one way to effectively deal with the overlapping generations issue.

The central theme of this chapter is the interplay between individual learning and population dynamics. We will outline the conditions for the emergence of *coherent* linguistic states where most of the population con-

verges to a single shared language over time. These conditions are seen to be related to the learning *fidelity* of individual children. In particular, if the learning fidelity is high, coherence is achieved. If it is low, no such coherence emerges. More interestingly, the population dynamics undergoes a bifurcation from incoherent to coherent regimes as the learning fidelity is changed. Learning fidelity is the probability with which the child successfully learns the language of its parents. Since this probability will depend upon the complexity of the space of possible grammars, we see that the emergence of coherence is related now to this complexity. In this sense, learnability, evolvability, and grammatical complexity are linked.

12.1 General model

12.1.1 The class of languages

Following the previous chapter, we take a language to be a probability measure μ on a countable set. Let us assume that there are n possible languages[1] given by $\mu_1, \ldots, \mu_n$. We characterize the mutual intelligibility between these languages by an $n \times n$ matrix A. The (i, j) entry of A (denoted by A_{ij}) is the probability that a speaker using language μ_i is understood by a hearer using μ_j. Using the notion of communicative efficiency developed previously, a natural choice for A_{ij} is

$$A_{ij} = \sum_{m \in \mathcal{M}} \sigma(m) \sum_{s \in \mathcal{S}} \mu_i(s|m)\mu(m|s)$$

where $\mathcal{S}$ is the space of possible expressions or signals and $\mathcal{M}$ is the space of possible meanings. σ is a measure on $\mathcal{M}$ that denotes the prior probability with which the need arises to communicate different meanings.

The rest of the chapter does not depend in any crucial sense on how A_{ij} is defined as long as such a matrix of n^2 numbers exists. In general, $0 \leq A_{ij} \leq 1$. However, we will conduct a detailed analysis only for the

[1] In general, one might consider the space of languages to be a continuous space of possible measures. Conventionally, however, language is taken to be categorical, discrete, combinatorial, and therefore countable. In order to engage this traditional conception of language, we have chosen countable spaces of languages. The countably infinite case presents considerable mathematical difficulty and we focus in this chapter only on the case of n languages in competition with each other. The finite n case is particularly suited to analyses in the principles and parameters tradition. However, it is worth noting that the analysis is conducted here for arbitrary n and therefore the large n limit may be viewed by some as a better approximation of reality.

symmetric case where $A_{ii} = 1$ and $A_{ij} = a$ if $i \neq j$. In other words, each language has perfect mutual intelligibility with itself. For any pair of different languages, the mutual intelligibility is a and does not depend upon the pair involved. Thus the languages are all equally fit and there are no clusters within the space of languages. The more complex case where languages have different fitnesses or group into more or less mutually intelligible clusters can also be treated but will not be dealt with in any detail here.

12.1.2 Fitness, reproduction, and learning

Let us consider a population of constant size. Each person uses only one language. The fraction of people who speak the language μ_j is denoted by x_j. Thus the linguistic state of the population is given by n numbers $x_1, \ldots, x_n$ such that $x_i \geq 0$ and $\sum_{j=1}^{n} x_j = 1$.

Fitness: The overall fitness of an individual with language μ_i is its average communicative efficiency with the rest of the population as a whole. Recall that the mutual intelligibility between μ_i and μ_j can be given by the following symmetric formula

$$F(\mu_i, \mu_j) = \frac{1}{2}(A_{ij} + A_{ji})$$

The average communicative efficiency of a speaker of μ_j is then defined to be

$$f_i = f_0 + \sum_{j=1}^{n} F(\mu_i, \mu_j) x_j. \tag{12.1}$$

Here, f_0 is the background fitness which does not depend on the person's language. The language dependent fitness is related not just to the person's own language but also to the proportion of people speaking various languages in the population. Thus a speaker of μ_1 would have a fitness of $f_0 + 1$ if everyone speaks μ_1 in the population but a different (lower) fitness if everyone else speaks μ_2. Ultimately, the evolutionary dynamics will depend upon the F matrix and the tools of the previous chapter are used to provide a plausible measure in terms of the communicative efficiency matrix A.

Differential Reproduction: Following the central tenet of natural selection, we assume that individuals reproduce in proportion to their fitness[2].

[2]It is has been wisely noted (see, e.g. Ariew and Lewontin, 2004) that the notion of fitness has received many disparate and mutually inconsistent treatments in the evolution-

One may argue that successful communication about potentially life threatening events aids the survival of organisms in an uncertain world. Therefore those who have higher communicative efficiency have greater chances of survival and correspondingly greater ability to reproduce.

Learning: Children learn from their parents. For simplicity we assume that each child has only one parent, i.e. each child learns from one teacher. We allow for mistakes during language acquisition. These mistakes may be due to the finite size of the primary linguistic data or the input of non-parents or both. At any rate, it is possible learn from a person with language μ_i and end up speaking language μ_j. The probability of such a transition is denoted by Q_{ij}. The matrix Q depends on the matrix A because the latter defines how close different grammars are to each other (and therefore, how easy it is to confuse them with each other in the learning process). The dependence of Q on A can be modeled if we make assumptions about the precise nature of learning. Much of part II of this book discusses the theory of learning and the tools developed therein may be used to evaluate the dependence of Q on A.

12.1.3 Population dynamics

How will the population evolve? Assume a generational structure and consider two successive generations. Let the state of the population in generation t be given by $x_1(t), \ldots, x_n(t)$. Since reproduction is proportional to fitness, the proportion of the next generation that are offspring of μ_j users is given by $\frac{f_j x_j(t)}{\sum_{i=1}^n f_i x_i(t)}$. Each of these children attempts to learn μ_j. The proportion of μ_j users in the next generation is easily seen to be

$$x_j(t+1) = \frac{\sum_{i=1}^n x_i(t) f_i Q_{ij}}{\sum_{k=1}^n f_k x_k(t)}$$

In order to avoid the assumption of generational structure, we will discretize time finely and consider a natural system of ordinary differential equations that characterize the dynamics. A first candidate for this is

$$\dot{x_j} = \sum_i f_i x_i Q_{ij}, \quad 1 \leq j \leq n,$$

---

ary literature. Therefore it becomes unclear what one means by fitness in evolutionary contexts in general. We will not enter into such philosophical tangles at this point but proceed with our particular interpretation of it for the time being.

This captures the fact that the rate of growth of x_j ought to be proportional to the probability with which children become μ_j users. Unfortunately, the above equation can give rise to unbounded growth for some of the x_j's. The initial state of the population is a point on the n-simplex. The dynamics needs to be defined so as to constrain the evolutionary trajectory to lie on the n-simplex from any such initial condition. In order to do this, we need additional constraints:

1. Total population size is constant, i.e., $\sum_{j=1}^{n} x_j(t) = 1$. This is realized by ensuring that the net differential is 0, i.e, $\sum_{j=1}^{n} \dot{x_j} = 0$.

2. Population sizes are positive, i.e., $\forall j, t\ x_j(t) \geq 0$. In order to achieve this we need that if $x_j = 0$ for some j, then $\dot{x_j} \geq 0$.

Incorporating these constraints, the dynamics of a population $(x_1, \ldots, x_n)$ may be suitably captured by the following general system of ordinary differential equations:

$$\dot{x}_j = \sum_i f_i x_i Q_{ij} - \phi x_j, \quad 1 \leq j \leq n, \tag{12.2}$$

where $\phi = \sum_{m=1}^{n} f_m x_m$. Note that ϕ may be interpreted as the average fitness of the population and its language-dependent part is the *grammatical coherence*. Thus it defines the overall probability that a sentence uttered by a randomly chosen agent is understood by another randomly chosen agent. Eq. 13.1 is similar to a quasi-species equation (Eigen & Schuster, 1979), but has frequency dependent fitness values (Nowak, 2000). We analyze this in some detail for symmetrically distributed language spaces.

12.2 Dynamics of a fully symmetric system

In order to investigate system (13.1), we need to specify the matrices A and Q. Let us consider the simplest case where $A_{ij} = a$, a constant, for all $i \neq j$, and $A_{ii} = 1$. We will refer to such a matrix as a *fully symmetric A* matrix. It corresponds to the situation where all languages have the same communicability with each other. The fitness in this case is simply

$$f_i = (1 - a)x_i + a + f_0. \tag{12.3}$$

Next, we introduce the notion of a *learning fidelity*, q, which is the probability of learning the teacher's (parent's) language, i.e., the probability to learn

language μ_i given that the teacher speaks μ_i. We will assume all languages are equally easy (or hard) to learn, i.e., q does not depend upon i. In keeping with our assumption of an equidistant configuration of languages, we further assume that they are equally confusible on average. Thus, if a mistake is made in learning the parent's language, then it is equally likely that the person will speak μ_j, $j \neq i$, for any j. The probability to be taught μ_i and learn μ_j is therefore $u = (1-q)/(n-1)$, for each $j \neq i$. The quantity u is called the *error rate* of language learning. Therefore the Q matrix is defined by

$$Q_{ii} = q, \quad Q_{ij} = u = (1-q)/(n-1), \;\; i \neq j. \tag{12.4}$$

The learning accuracy satisfies $1/n \leq q \leq 1$. Perfect learning implies $q = 1$, i.e., no mistakes are made. $q = 1/n$ is equivalent to random guessing on the part of the learner. With these assumptions, system (13.1) becomes

$$\dot{x}_j = (1-a)\Big[-x_j^3 + x_j^2 q + \sum_{i \neq j} x_i^2 \left(\frac{1-q}{n-1} - x_j \right) \Big] - \frac{(a+f_0)(1-q)(nx_j - 1)}{n-1}, \tag{12.5}$$

for all $1 \leq j \leq n$.

12.2.1 Fixed points

To begin, we will look for fixed points of system (12.5). These correspond to solutions of $\dot{x_j} = 0$. The dynamics is given by a set of cubic equations. We will exploit symmetry in order to understand the structure of the solutions. In general, one may consider m-language solutions[3] where m languages are used with different frequencies (say $X_1, X_2, \ldots, X_m$) and the rest are used with the same frequency (given by $\frac{1}{(n-m)} \sum_{i=1}^{m} X_i$ each). The most important class of solutions that we will examine in some detail in the rest of this chapter are the one-language solutions.

Factoring the Cubic: Let us set $x_l = X$, $x_m = (1-X)/(n-1)$, $m \neq l$. This corresponds to the case when all languages except one are used with the

[3]It turns out that all m-language solutions have the form that $X_1 = X_2 = \ldots = X_m$. To see this, note if $\bar{x} = (x_1, \ldots, x_n)$ is a fixed point, then each coordinate is a root of the polynomial $(1-a)\left(-x^3 + x^2 q + (\alpha - x)\left(\frac{1-q}{n-1} - x\right)\right) - \frac{(a+f_0)(1-q)(nx-1)}{n-1}$ where $\alpha = \sum_j x_j^2$. This polynomial has three roots given by $\frac{1}{n}, r, s$. Since the constraint $\sum_j x_j = 1$ must be satisfied by the solution, we see that there are two possible solutions (i) all x_j's are equal to $\frac{1}{n}$ (ii) m of the x_j's are equal to r and the rest are equal to s. See Mitchener (2003) for this observation and more developments following from it.

same frequency. Without loss of generality, we can take $l = 1$. From system (12.5) with a zero left hand side, we obtain n equations for the unknown X. They are compatible, because the equations for x_2, x_3, ..., x_n are identical, and their sum is just the equation for x_1 (due to the conservation of the number of people). In other words, each of the equations from the second to the last one is nothing but the first equation divided by $n-1$. Therefore, we only need to solve the first equation,

$$X^3 - X^2 q + \frac{(1-X)^2}{n-1}\left(X - \frac{1-q}{n-1}\right) + \frac{(1-q)(a+f_0)(nX-1)}{(1-a)(n-1)} = 0. \tag{12.6}$$

This is a cubic equation in X and can be factored as

$$(nX-1)(AX^2 + BX + C) = 0$$

where

$$A = \frac{1}{n-1}; B = \frac{2q-1-qn}{(n-1)^2}; C = \frac{(1-q)(a+f_0)}{(1-a)(n-1)} + \frac{1-q}{(n-1)^2}$$

The Solutions: The cubic equation (Eq. 12.6) has three solutions. The factorization provided above makes it clear what these are. One solution is

$$X_0 = 1/n \tag{12.7}$$

and corresponds to the uniform distribution (i.e. all grammars occur in the population equally often). The other two solutions correspond to the two roots of the quadratic factor and are given by $X_\pm = \frac{-B \pm \sqrt{B^2-4AC}}{2A}$. Putting in the values of A, B, C from above we get

$$X_\pm = \frac{-(1-a)(1+(n-2)q) \mp \sqrt{D}}{2(a-1)(n-1)}, \tag{12.8}$$

where

$$D = 4[-1-a(n-2)-f_0(n-1)](1-q)(n-1)(1-a) + (1-a)^2[1+(n-2)q]^2. \tag{12.9}$$

These two solutions describe a less symmetrical situation, when one grammar is the most (least) preferred one and is used with frequency $X_\pm$, and the rest

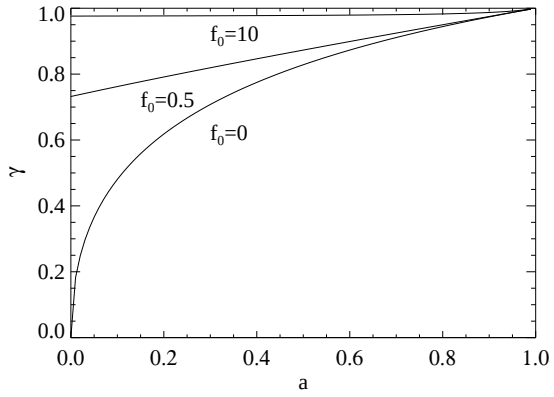

Figure 12.1: The threshold value, γ, of learning accuracy, in the limit of large values of n. For $q > q_1 \approx \gamma$, asymmetric solutions become possible. The coefficient γ is plotted as a function of a for different values of the background fitness, f_0.

of the grammars are used equally often. Therefore there are $2n+1$ solutions in all: (i) uniform solution (ii) n solutions of the form $X_i = X_+$ for some i (ii) n solutions of the form $X_i = X_-$ for some i.

Existence of One-Language Solution: The one-language solutions correspond to the solutions given by Eq. 12.8. Real valued solutions exist only if $D \geq 0$. This, in turn, is equivalent to the existence condition $q \geq q_1$, where

$$q_1 = \frac{4 + 2W(n-1)^{3/2} - 2f_0(n-1)^2 - 3n - a(2n^2 - 7n + 6)}{(1-a)(n-2)^2}, \tag{12.10}$$

and $W = \sqrt{(1+f_0)[1 + a(n-2) + f_0(n-1)]}$. Thus we see that q_1 is a *coherence threshold* above which one-language solutions may exist and below which only the uniform solution is possible. The uniform solution always exists.

Properties of Coherence Threshold: It is worthwhile to reflect on the nature of the coherence threshold q_1 and its dependence on factors such as a (the confusibility between different languages), n (the size of the space of possible languages), and f_0 (the background fitness).

$n = 2$: In the special case of $n = 2$, q_1 is given by $q_1 = (3+a+4f_0)/(4(1+f_0))$. Thus q_1 increases linearly with a and has the value 1 when $a = 1$.

Large n: For $n \gg 1/(a + f_0)$, we have:

$$q_1 = \gamma + O\left(\frac{1}{n}\right), \tag{12.11}$$

where

$$\gamma = \frac{2}{1-a}\left(\sqrt{(a+f_0)(1+f_0)} - (a+f_0)\right). \tag{12.12}$$

We observe (see Fig. 12.1) that γ is a monotonically increasing function of a and it is equal to 1 when $a = 1$. Note that $1-\gamma = \frac{(1-a)(a+f_0)}{(a+f_0+\sqrt{(a+f_0)(1+f_0)})^2}$. If a is close to 1, so that $a = 1-\epsilon$ and $\epsilon \to 0$, we have $\gamma = 1-\epsilon/(4(f_0+1))+O(\epsilon^2)$. The coefficient γ also grows with f_0 reaching 1 as $f_0 \to \infty$. More precisely, we have $\gamma = 1 - (1-a)/(4f_0) + O(1/f_0^2)$.

$a = f_0 = 0$: In the special case of $a = f_0 = 0$, the existence condition looks like

$$q_1 = \frac{4 + 2(n-1)^{\frac{3}{2}} - 3n}{(n-2)^2}. \tag{12.13}$$

For $n \gg 1$ we obtain $q_1 = 2/\sqrt{n} + O(1/n)$, i.e. the asymptotic behavior is quite different.

Summary Remarks: For small values of $q(< q_1)$, only the uniform solution exists. At $q = q_1$, a bifurcation occurs. Solution (12.8) emerges and is shown in Fig. 12.2. For all values of a and f_0, at $q = 1$ we have $X_+ = 1$ and $X_- = 0$. At the point where the solution first appears ($q = q_1$), the value is approximately $X_\pm \approx \frac{q_1}{2} \approx \frac{\gamma}{2}$ for large n. For the setting $f_0 = 0$, we have $X_\pm \approx \sqrt{a}/(1+\sqrt{a})$.

We note that because of the choice of the A and Q matrices, system (12.5) is highly symmetrical and its solutions are degenerate. Namely, by relabeling variables, we can pick any of the n grammars to be the "chosen" one, and then we will have n equivalent solutions of the form

$$x_l = X, \quad x_j = \frac{1-X}{n-1}, \quad \text{where } X = X_0, \ X_+ \text{ or } X_-, \quad \forall j \neq l, \tag{12.14}$$

for any l such that $1 \leq l \leq n$. Perturbations of the A or Q matrix will in general lift the degeneracy, which may result in the following changes: (i)

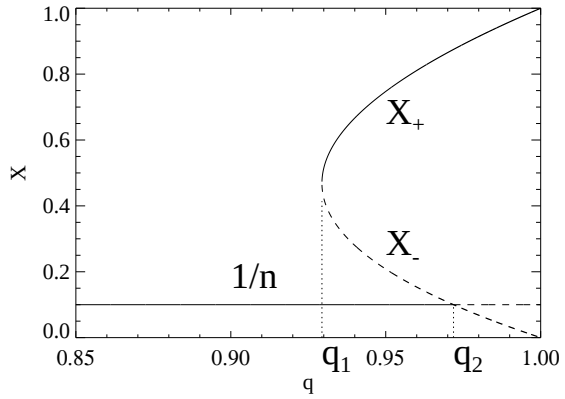

Figure 12.2: The solutions $X = X_0$, X_+ and X_-. Here, $a = 0.5$, $f_0 = 1$ and $n = 10$. Stable solutions are represented by solid lines, and unstable ones by dashed lines (see Section 12.2).

in general, all values of x_j, $j \neq l$, will be different from each other, and (ii) solutions of the form (12.14) will have different shapes for different values of l (in other words, X_0, X_+ and X_- will depend on l).

In the next section we will see that solution (12.14) with $x_l = X_-$ is always unstable and the one with $x_l = X_0$ (the uniform solution) loses stability as q grows further. Only solutions with $x_l = X_+$ remain stable for high values of learning accuracy. When the A matrix is not fully symmetric, the X_+-type solutions have a more complicated form, but one important feature will persist. It turns out that these solutions can be characterized by one language whose share grows as q approaches unity, whereas the frequency of other languages decreases. μ_l will be called the *preferred*, or "chosen", language and the languages μ_j with $j \neq l$ will be called *secondary languages*.

Finally, it is important to re-emphasize that the fixed points found in this section are not the only possible fixed points of system (13.1). It turns out, however, that m-language solutions correspond to saddlenodes and unstable along certain directions. For large values of q, the one language solutions with $X_i = X_+$ are the only attractors of this system. It also turns out that there are no periodic orbits (oscillations) in the symmetric system. In much of this chapter we only concentrate on the three fixed points found above. Later we will briefly examine the case for asymmetric systems where the

dynamics could potentially get much more complicated.

12.2.2 Stability of the fixed points

Given a system of n ordinary differential equations of the form

$$\dot{x_j} = F_j(x_1, x_2, \ldots, x_n);\ j = 1, \ldots, n$$

one can check for the stability of a fixed point $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)^T$ by locally perturbing around $\mathbf{x}^*$ to yield $x_j = x_j^* + \tilde{y}_j$. Taking exponential perturbations of the form $\tilde{y}_j = e^{\Gamma t} y_j$ we get

$$\dot{x_j} = \Gamma y_j e^{\Gamma t} \approx \sum_{i=1}^{n} (\frac{\partial F_j}{\partial x_i}|_{\mathbf{x}^*}) e^{\Gamma t} y_i$$

yielding the following eigenvalue problem

$$\Gamma y_j = \sum_{i=1}^{n} (\frac{\partial F_j}{\partial x_i}|_{\mathbf{x}^*}) y_i; j = 1, \ldots, n$$

Thus the $n \times n$ matrix J where $J_{ij} = \frac{\partial F_j}{\partial x_i}|_{\mathbf{x}^*}$ is the Jacobian whose eigenvalues determine stability. In our case, since we are interested in solutions on the $n-1$ dimensional simplex, we have the additional constraint that $\sum_j y_j = 0$. If there exists a $\Gamma > 0$, the system is unstable. If all eigenvalues are strictly less than 0, the system is stable.

Let us check the stability of solution (12.14); we will take $l = 1$. Following the well developed techniques of a linear stability analysis outlined earlier, we perturb the solution by taking $x_1 = X + \tilde{y}_1$, $x_j = \frac{1-X}{n-1} + \tilde{y}_j$, $j > 1$ (here X can be X_0, X_+ or X_-). We substitute this into system (12.5) and linearize with respect to $\tilde{y}_j$.

Next, we introduce exponential behavior of the perturbations, i.e. $(\tilde{y}_1, \ldots, \tilde{y}_n)^T = e^{\Gamma t}(y_1, \ldots, y_n)^T$. The eigenvalue problem corresponds to $\Gamma \mathbf{y} = J\mathbf{y}$ where $\mathbf{y} = (y_1, \ldots, y_n)^T$. Since the Jacobian J is computed at $\mathbf{x}^* = (X, \frac{(1-X)}{n-1}, \frac{(1-X)}{n-1}, \ldots, \frac{(1-X)}{n-1})^T)$, we see that the elements of J consist of exactly 5 distinct values at each time. As a result, we obtain a system of linear equations for $y_1, \ldots, y_n$ with the following form

$$Ay_1 + B \sum_{m>1} y_m = 0, \qquad Cy_j + D \sum_{\substack{m>1 \\ m \neq j}} y_m + Ey_1 = 0, \quad 2 \leq j \leq n,$$

where A, B, C, D and E are constants given by

$$\begin{array}{ll} A = \frac{\partial F_1}{\partial x_1} - \Gamma & = (1-a)\left(-3X^2 + 2Xq - (n-1)(\frac{1-X}{n-1})^2\right) - \frac{n(a+f_0)(1-q)}{n-1} - \Gamma \\ B = \frac{\partial F_1}{\partial x_j} & = 2(1-a)(\frac{1-X}{n-1})(\frac{1-q}{n-1} - X) \\ C = \frac{\partial F_j}{\partial x_j} - \Gamma & = (1-a)\left(2q(\frac{1-X}{n-1}) - (n+1)(\frac{1-X}{n-1})^2 - X^2\right) - \frac{n(a+f_0)(1-q)}{n-1} - \Gamma \\ D = \frac{\partial F_j}{\partial x_k} & = 2(1-a)(\frac{1-X}{n-1})(\frac{X-q}{n-1}) \\ E = \frac{\partial F_j}{\partial x_1} & = 2(1-a)X(\frac{X-q}{n-1}) \end{array}$$

respectively.

Because of the conservation of the number of people, we have $\sum_{j=1}^{n} y_j = 0$. We therefore need to solve for the above linear system under this constraint. Replacing y_1 by $-\sum_{m=2}^{n} y_m$, we get:

$$(A-B)\sum_{m=2}^{n} y_m = 0, \tag{12.15}$$

$$(C-D)y_j + (D-E)\sum_{m=2}^{n} y_m = 0, \;\; 2 \le j \le n. \tag{12.16}$$

Here, the first equation is the sum of the other $(n-1)$ equations (by construction of equation (13.1)) and is therefore satisfied as long as the other $(n-1)$ equations are satisfied. To ensure the existence of nontrivial solutions of linear system (12.16), we require that the determinant of the corresponding $(n-1)\times(n-1)$ matrix is zero. The matrix $[M_{ij}]$ has the form $M_{ii} = C - D$, $M_{ij} = D - E$ for $i \neq j$, and its determinant is given by

$$(C - 2D + E)^{n-2}(C - D + (n-2)(D-E)). \tag{12.17}$$

Determinant (12.17) is zero if $C = 2D - E$ (the corresponding Γ is denoted as Γ_1) or if $C - D + (n-2)(D-E) = 0$ (the corresponding Γ is denoted as Γ_2). Note that in the special case of $n = 2$ we only have the latter condition. By examining the sign of $\Gamma_{1,2}$, we can study the stability of solutions X_0, X_+ and X_-. If at least one of the growth rates is positive, the corresponding solution is unstable.

The uniform solution.

For $X = X_0 = 1/n$, we have

$$\Gamma_1 = \Gamma_2 = \frac{1}{n(n-1)}\Big[(n(2q-1)-1)(1-a) - n^2(1-q)(f_0+a))\Big]. \tag{12.18}$$

This gives a threshold condition for learning accuracy. Namely, for $q > q_2$, $\Gamma_{1,2}$ become positive and the uniform solution loses stability. The value q_2 is given by

$$q_2 = \frac{n^2(f_0 + a) + (n+1)(1-a)}{n[n(f_0 + a) + 2(1-a)]}. \tag{12.19}$$

The value q_2 corresponds to the point where $X_- = X_0$. Thus, the uniform solution loses stability at the point q where it meets solution X_-. For large n ($n \gg 1/(a + f_0)$), we have

$$q_2 = 1 - \frac{1}{n}\left(\frac{1-a}{a+f_0}\right) + O\left(\frac{1}{n^2}\right). \tag{12.20}$$

Note that in the case $a = f_0 = 0$, we have $q_2 = 1/2 + 1/(2n)$.

The asymmetric solutions.

First, we examine the case $n > 2$. The growth rate for the two asymmetric solutions is presented in Fig. 12.3. It turns out that for the solution X_+, both Γ_1 and Γ_2 are non-positive for all $q \geq q_1$ (the solid lines in Fig. 12.3). This means that the asymmetric solution X_+ is stable everywhere in the domain of its existence. Thus, for higher values of learning accuracy, the system prefers a state when one of the grammars is used very often, whereas the rest of them have an equal (and small) share.

For X_-, the situation is different. In the domain $q_1 \leq q \leq 1$, one of the growth rates is positive whereas the other is negative (at the point $q = q_2$ they are both zero, the dotted lines in Fig. 12.3). This means that the solution X_- is unstable (it is neutrally stable for $q = q_2$). It is instructive to compare the eigenvectors corresponding to the eigenvalues Γ_1 and Γ_2. The former one has $y_1 = 0$, and the latter one has $y_1 \neq 0$. For $q > q_2$, $\Gamma_1 > 0$, which means that the solution X_- loses stability in such a way that x_1 stays the same, but the rest of the grammars fail to keep a uniform distribution.

The complete proof of these stability results is tedious and not reported here. However, to convince the reader of their veracity and to provide some insight into why they hold, let us consider a *large n* analysis for the eigenvalues.

Stability of X_+ for large n:

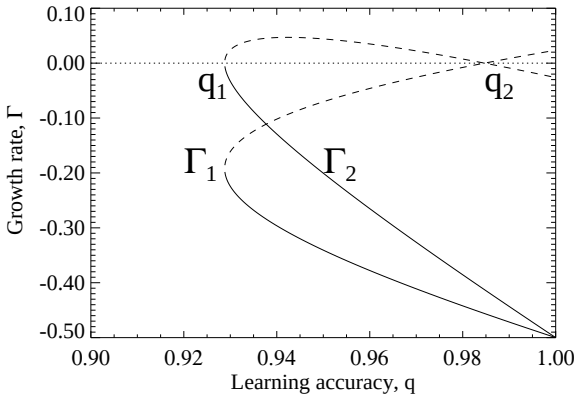

Figure 12.3: The growth rates (eigenvalues) for the one-grammar solutions X_+ (solid lines) and X_- (dashed lines), as functions of q. Here, $a = 0.5$, $f_0 = 1$ and $n = 20$.

Note that

$$X_+ > \frac{1 + (n-2)q}{2(n-1)} \approx \frac{q}{2}$$

where $q \geq q_1 \approx \gamma + O(\frac{1}{n})$.
Now

$$\Gamma_1 = C' - 2D + E$$

where $C' = C + \Gamma = \frac{\partial F_j}{\partial x_j}$. For all n, the following inequalities hold:

$$C' \leq -X^2(1-a) - (a+f_0)(1-q) + 2q(1-a)\frac{1-X}{n-1}$$

$$|D| = |2(\frac{1-X}{n-1})(\frac{X-q}{n-1})(1-a)| \leq \frac{2(1-a)}{(n-1)^2}$$

$$|E| = |2X(\frac{X-q}{n-1})(1-a)| \leq \frac{2(1-a)}{n-1}$$

Therefore, we have

$$\Gamma_1 \leq C' + |2D| + |E| \leq -X^2(1-a) - (a+f_0)(1-q) + O(\frac{1}{n})$$

For sufficiently large n, Γ_1 is therefore seen to be negative.

Consider Γ_2. This is given by the expression

$$\Gamma_2 = C' - D + (n-2)(D-E) \approx C' - nE + nD$$

Since $D/E = \frac{1}{n-1}(\frac{1-X}{X}) = O(\frac{1}{n})$, we have that

$$\Gamma_2 \approx C' - nE$$

For large n (ignoring $O(\frac{1}{n})$ terms), we have

$$\Gamma_2 \approx -(1-a)X^2 - (a+f_0)(1-q) + 2X(q-X)(1-a)$$

But

$$\begin{aligned} &-(1-a)X^2 - (a+f_0)(1-q) + 2X(q-X)(1-a) \\ &\leq -(1-a)X^2 - (a+f_0)(1-q) + 2(\frac{q}{2})(\frac{q}{2})(1-a) \\ &\leq -(1-a)X^2 - (a+f_0)(1-q) + \frac{q^2}{2}(1-a) \end{aligned}$$

Now putting in $X = X_+$, we see

$$X = \frac{1+(n-2)q}{2(n-1)} + \frac{\sqrt{D}}{2(1-a)(n-1)} \approx \frac{q}{2} + \alpha$$

Therefore, we have

$$\Gamma_2 \leq -(1-a)(\frac{q}{2}+\alpha)^2 - (a+f_0)(1-q) + \frac{q^2}{2}(1-a)$$

$$= (1-a)(\frac{q^2}{4}) + q(a+f_0) - (a+f_0) - (1-a)(q\alpha + \alpha^2)$$

Now let us find an approximate expression for α. We see that for large n, we have

$$\alpha \approx \sqrt{\frac{q^2}{4} - \frac{(a+f_0)(1-q)}{1-a}} \tag{12.21}$$

We therefore see that

$$(1-a)(\frac{q^2}{4}) + q(a+f_0) - (a+f_0) - (1-a)\alpha^2 \approx 0$$

and putting this into the approximate upperbound for Γ_2, we have

$$\Gamma_2 \leq -(1-a)q\alpha \leq 0$$

Stability of X_- for large n:

A similar analysis may be conducted for X_-. Here we need to consider two different regions ($q_1 < q < q_2$ and $q_2 < q < 1$ following fig. 12.3). Let us begin by taking q to be slightly larger than q_1 so that

$$\frac{1}{\sqrt{n}} \leq X_- \leq \frac{q_1}{2}$$

In this regime, we show that $\Gamma_2 > 0$. Recall that

$$\Gamma_2 = C' - D + (n-2)(D-E)$$

Note that $\frac{D}{E} = \frac{1-X}{X(n-1)} \leq \frac{(1-q_1)\sqrt{n}}{n-1}$ and for large n this is approximately 0. Therefore, we may take

$$\Gamma_2 \approx C' - nE$$

as before. Following the approximations made earlier, we see that this reduces to

$$\Gamma_2 \approx -(1-a)X^2 - (a+f_0)(1-q) + 2X(q-X)(1-a) + \phi$$

where $\phi = (1-a)\left((\frac{1-X}{n-1})(2q-(1-X))\right)$. A detailed analysis of ϕ suggests that it is either negligible compared to the other terms or is positive (or both). So we ignore ϕ in what follows. We also have $X_- \approx \frac{q}{2} - \alpha$. Putting this into the above equation we have

$$\Gamma_2 \approx -(1-a)(\frac{q}{2}-\alpha)^2 - (a+f_0)(1-q) + 2(\frac{q}{2}-\alpha)(\frac{q}{2}+\alpha)(1-a)$$

Simplifying and cancelling terms, we have

$$\Gamma_2 = (1-a)\frac{q^2}{4} - (a+f_0)(1-q) - (1-a)\alpha^2 + (1-a)q\alpha - 2\alpha^2(1-a)$$

But, from eq. 12.21, this reduces to

$$\Gamma_2 \approx (1-a)\left(q\alpha - 2\alpha^2\right) = \alpha(1-a)(q-2\alpha)$$

Again, from eq. 12.21, we see that

$$0 \leq \alpha \leq \frac{q}{2}$$

Therefore, we have

$$\Gamma_2 \approx \alpha(1-a)(q-2\alpha) \geq 0$$

Thus, Γ_2 is in general positive for large n. By the same argument as before, Γ_1 is negative for large n in this regime.

Now consider the regime where $q > q_2$. Note that for very large n, the value of the threshold $q_2 \approx 1$. It turns out that for finite n, the value of X_- at $q = q_2$ is exactly equal to $\frac{1}{n}$ and correspondingly for all $q \geq q_2$ we have $X_- \leq \frac{1}{n}$. It is easy to check that in this region $|D| \geq |E|$. Therefore, we have

$$\Gamma_1 = C' - 2D + E \geq C' - D \geq C'$$

For convenience, put in $q = 1$. For this value, $X_- = 0$ and we see that

$$\Gamma_1 \geq C' = \phi = \frac{1-a}{n-1} \geq 0$$

Since Γ_1 is a continuous function of q, we see that it must be positive in the neighborhood of $q = 1$. This neighborhood is arbitrarily small for large n and so we conclude that $\Gamma_1 \geq 0$ in this regime.

This concludes our analysis. For completeness we consider the value $n = 2$. In this special case, q_2 coincides with q_1. Therefore, for $q < q_1 = q_2$, the uniform solution $x_{1,2} = 1/2$ is stable, and for higher values of learning accuracy, it loses stability. We have a pitchfork bifurcation with two equivalent stable solutions, $(x_1, x_2) = (X_+, X_-)$ and $(x_1, x_2) = (X_-, X_+)$.

12.2.3 The bifurcation scenario

To sum up the bifurcation picture (Fig. 12.2), we note that for $0 \leq q < q_1$ the only stable solution is the uniform solution $1/n$, then between q_1 and q_2 both the uniform solution and solutions (12.14) with X_+ (the one-grammar solutions) are stable, and finally, for $q > q_2$ the uniform solution loses its stability and the one-grammar solutions remain stable.

At the point $q = q_1$, where the non-uniform solutions first appear, the corresponding average fitness (assuming that n is large) is

$$\phi_{asym} = \frac{\left(\sqrt{(a+f_0)(1+f_0)} - f_0 - a\right)^2}{1-a} + a + f_0, \tag{12.22}$$

whereas the average fitness of the uniform solution (for large n) is

$$\phi_{unif} = a + f_0. \tag{12.23}$$

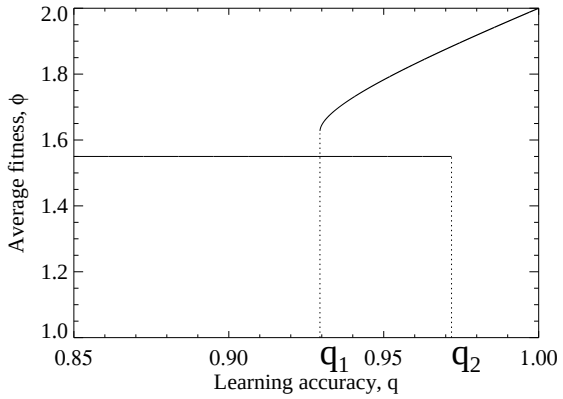

Figure 12.4: Total fitness of the stable solutions of a system with a fully symmetric A matrix, as a function of learning accuracy, q. Parameters of the system are as in Fig. 12.2.

One can see that as the system goes to a one-grammar solution, the average fitness (and the grammatical coherence) experience a jump, $\Delta f = (1 - a)(\gamma/2)^2 + O(1/n)$, see Fig. 12.4. Note that if $a = 1 - \epsilon$, then $\Delta f \sim \epsilon/4$. As q increases to 1, the total fitness of the one-grammar solution monotonically increases to $1 + f_0$, whereas the fitness of the uniform solution stays constant.

It is convenient to present the stability diagram in terms of the error rate, u (see Fig. 12.5). Clearly, as n grows, it becomes harder and harder to maintain one grammar. Also, one can see that there is always a bistability region where the uniform solution and X_+ coexist. Indeed, for the existence of a one-grammar solution we need

$$u \leq u_1 = c_1/n, \quad c_1 \equiv 1 - \gamma. \tag{12.24}$$

For the uniform solution to lose stability we need

$$u \leq u_2 = c_2/n^2, \quad c_2 \equiv (1 - a)/(a + f_0). \tag{12.25}$$

The above inequalities are derived in the case of large n and $a + f_0 > 0$.

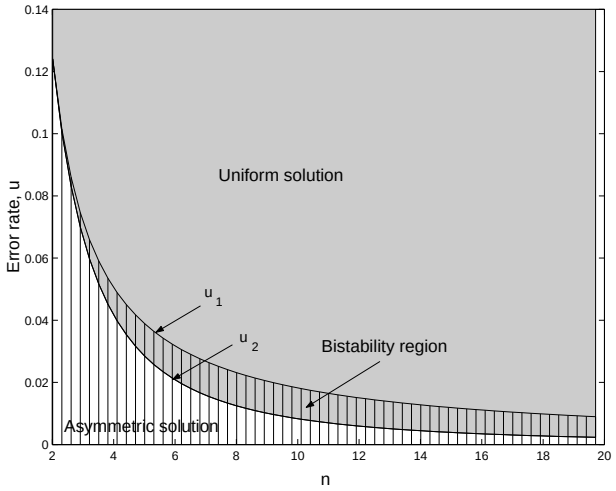

Figure 12.5: The stability diagram in terms of the error rates, $u_{1,2}$. Here, $a = 0.5$, $f_0 = 1$.

12.3 Fidelity of Learning Algorithms

In preceding sections we have examined in some detail how the population dynamics depends upon the learning fidelity q of the individual learner. Bifurcations in the population dynamics are seen to occur at $q = q_1$ and $q = q_2$. In the model, individual learners attempt to learn the parental language based on example sentences they receive. Consequently, the *number* of such linguistic examples is naturally related to the learning fidelity. In general, the greater the number of examples on which learners base their estimates, the higher the probability of learning the parental language correctly, and correspondingly the higher the q value at which the population dynamics operates.

An analysis of the learning algorithm will yield the precise relationship between the number of examples (b) and the learning fidelity (q). To illustrate this point, let us consider two learning algorithms that have been discussed in previous chapters.

12.3.1 Memoryless Learning

In Chapters 3 and 4, we introduced the class of memoryless learners that modify their grammatical hypotheses in an online adaptive fashion. There are many variants but to fix ideas consider the following algorithm:

1. *Initial Hypothesis:* The learner chooses a language uniformly at random from among the n different languages $\mu_1, \ldots, \mu_n$.

2. *Updating Hypothesis:* Let the learner's hypothesis after n sentences be $\mu_{j(n)}$. When the $(n+1)$th sentence s_{n+1} is heard, the learner updates its hypothesis in the following manner:

   (a) **if** s_n is understood, $j(n+1) = j(n)$

   (b) **else** $j(n+1)$ is chosen uniformly at random to be one of the $n-1$ languages not equal to $\mu_{j(n)}$.

Following the analysis of previous chapters, the behavior of such a learner may be characterized by a Markov chain. Such a chain has n states — one for each language. The probability distribution over the states (at each point in time) describes the probability with which the learner will hypothesize each of the n different languages at that time.

The initial probability distribution of the learner is uniform: $\mathbf{p}^{(0)} = (1/n, \ldots, 1/n)^T$, i.e. each of the languages has the same chance to be picked at the initial moment. The discrete time evolution of the vector $\mathbf{p}^{(t)}$ is characterized by a Markov chain with a transition matrix, $T(k)$, which depends on the teacher's language, μ_k. This matrix is defined by

$$T(k)_{ij} = \begin{cases} \frac{(1-A_{ki})}{(n-1)} & \text{if } i \neq j \\ A_{ki} & \text{if } i = j \end{cases}$$

Recall that A_{ki} is the probability with which a learner using language μ_i will understand a sentence from the target language (in this case μ_k). This is therefore the probability with which the learner will retain the hypothesis μ_i for the next time step. With probability $1 - A_{ki}$, the learner will switch its hypothesis to one of the other $n-1$ languages.

After b samplings, the k-th row of matrix Q is given by $(\mathbf{p}^{(b)})^T = (\mathbf{p}^{(0)})^T [T(k)]^b$ obtained with the transition matrix $T(k)$. Therefore we can specify the (i,j) element of Q as

$$Q_{ij} = [(\mathbf{p}^{(0)})^T T(i)^b]_j. \tag{12.26}$$

This expression captures the relation between matrices A and Q. For instance, if we assume that the off-diagonal entries of the A matrix are constant and equal to each other (the fully symmetric case), then, according to Eq. 12.26, the off-diagonal entries of the Q matrix are also equal to each other, and Eq. 12.4 holds. Expression (12.26) can be used to evaluate the learning accuracy and the error rate in terms of a. It is easy to check that

$$q = 1 - \left(1 - \frac{1-a}{n-1}\right)^b \frac{n-1}{n}, \tag{12.27}$$

$$u = \frac{1}{n}\left(1 - \frac{1-a}{n-1}\right)^b. \tag{12.28}$$

Note that $\lim_{b\to\infty} q = 1$ for any fixed n and $0 \leq a < 1$. This simply means that learning fidelity may be arbitrarily close to 1 depending upon how many (b) examples are made available to the learner.

When $a = 1$, all languages are mutually comprehensible. For this case, $q = \frac{1}{n}$ which is the lowest possible value of learning accuracy.

We can use results of the previous section to find conditions for b, the number of sampling sentences per individual, which would allow the population to maintain a particular grammar. We will assume that the number n is large and use inequalities (12.24) and (12.25). In order for solution X_+ to exist, we see that

$$b \geq b_1 = \frac{n}{1-a} \log \frac{1}{c_1}. \tag{12.29}$$

The uniform solution loses stability if

$$b \geq b_2 = \frac{n}{1-a} \log \frac{n}{c_2}. \tag{12.30}$$

The constants c_1 and c_2 are defined in formulas (12.24) and (12.25).

12.3.2 Batch Learning

At another end of the spectrum of learning algorithms lie the batch learning algorithms. Such algorithms form their final decision by globally optimizing over the entire collection of linguistic examples received over the learning period. Consider the following simple instantiation of such an algorithm. Denote the example set by $S = \{s_1, s_2, \ldots, s_b\}$. For each sentence s and language μ let us define the comprehensibility function $c(s, \mu)$ to be a $0-1$

valued function that takes the value 1 if s is comprehensible[4] to a speaker of μ.

1. For each of the candidate languages $\mu_1, \ldots, \mu_n$ determine its total comprehensibility score by

$$C_j(b) = \frac{1}{b} \sum_{i=1}^{b} c(s_i, \mu_j)$$

   Note that $C_j(b)$ is a random variable if the examples are drawn in some random fashion.

2. Determine the set of empirically optimal languages to be

$$\mathcal{U} = \{\mu_j \mid C_j(b) = \max_i C_i(b)\}$$

3. If $|\mathcal{U}| = 1$, then choose the unique optimal language as the guess. If there are multiple optimal languages, i.e., $|\mathcal{U}| > 1$, then choose any of the n languages uniformly at random.

It is worthwhile to make a remark about step (3) of the above algorithm. If there are multiple optimal languages ($|\mathcal{U}| > 1$) it might seem natural to choose one of the elements of $\mathcal{U}$ at random rather than choosing one of the n languages at random as is done by the above algorithm. The reason for considering the above strategy is merely to simplify analysis and to ensure that $Q_{ij} = Q_{ik}$ for all distinct i, j, k. It is easy to check that for symmetric A matrices this property will be satisfied by the above algorithm so that the analysis of the dynamics will go through as before. Furthermore, it is easy to check that the stated algorithm is unbiased in the sense that the learner will converge to the right language as $b \to \infty$.

For a symmetric A matrix, let us now compute bounds on b for this batch learner. We assume $A_{ij} = a$ if $i \neq j$ and $A_{ii} = 1$ $\forall i$. Let the parental language be μ_k. Consider b examples drawn in i.i.d. fashion and presented to the learner. To begin, note that each of the $C_j(b)$'s is an empirical average of

[4]There are many different notions that may be invoked to properly define comprehensibility. One may consider *parsability* of s according to the grammatical rules underlying μ. This is equivalent to determining whether $\sum_m \mu(s, m) > 0$. Alternatively, one might take into account the intended meaning behind s and let $c(s, \mu) = 1$ if a speaker of μ is able to correctly infer the true meaning. In doing so one might follow the treatment of the previous chapter.

b i.i.d random variables (the $c(s_i, \mu_j)$'s) with mean $A_k j$. Therefore, it follows from the simple law of large numbers that for each j the following is true:

$$\lim_{b \to \infty} C_j(b) \to a(\text{ with probability } 1); j \neq k$$

and

$$C_k(b) = 1 \ \forall b$$

Therefore the set $\mathcal{U}$ will always contain μ_k as a member. Let us first compute the probability that it will be the lone member. For this to be the case, it must be that for every other language μ_j, there is at least one sentence $s \in S$ that is incomprehensible to it, i.e., $c(s, \mu_j) = 0$.

Accordingly let us introduce the event E_j which is the event that at least one sentences in S is incomprehensible according to μ_j. The probability that $\mathcal{U}$ has a unique member is therefore given by

$$\mathbb{P}[\cap_{i \neq k} E_i]$$

But

$$\mathbb{P}[\cap_{i \neq k} E_i] = 1 - \mathbb{P}[\cup_{i \neq k} \bar{E}_i] \geq 1 - \sum_{i \neq k} \mathbb{P}[\bar{E}_i]$$

Now $\bar{E}_i$ is simply the probability that every sentence in S is comprehensible according to μ_i. This is simply

$$\mathbb{P}[\bar{E}_i] = a^b$$

Therefore, we have

$$\mathbb{P}(\cap_{i \neq k} E_i) \geq 1 - (n-1)a^b$$

Now let us consider Q_{kk}. This is given by

$$Q_{kk} = \mathbb{P}(\cap_{i \neq k} E_i) + \frac{1}{n}(1 - \mathbb{P}(\cap_{i \neq k} E_i)) = \frac{1}{n} + \frac{n-1}{n}\mathbb{P}(\cap_{i \neq k} E_i) = q$$

In order for q to be greater than q_1, it is *sufficient* for

$$\frac{1}{n} + \frac{n-1}{n}\left(1 - (n-1)a^b\right) > q_1$$

Simplifying, we get

$$b > \frac{\log(\frac{(n-1)^2}{n(1-q_1)})}{\log(\frac{1}{a})} = \Omega(\log(n))$$

It is worthwhile to note that although we have assumed $A_{ki} = a$ for all distinct k, j, the above analysis would work for any arbitrary choice of A matrix as long as it was diagonal dominant, i.e., as long as $A_{kj} > A_{ki}$ for all $i \neq k$. The constants would change and the dependence would be on $\max_{i \neq k} |A_{ii} - A_{ik}|$ rather than a. The dependence on n would remain unchanged.

We thus see that the number of sample sentences needed for a community of batch learners to develop a coherent language grows as $\log n$, whereas memoryless learners need $b \propto n$ sentences (formula (12.29)). This is a consequence of the fact that batch learners have perfect memory, whereas memoryless learners only remember one sentence at a time.

12.4 Asymmetric A matrices

In this chapter we have provided in some detail the analysis of a population of language users where the languages $\mu_1, \ldots, \mu_n$ are in a symmetric configuration with respect to each other. In particular, we have assumed that $A_{ii} = 1$ for all i and $A_{ij} = a$ for all distinct i, j. A natural question that now arises is what happens when the matrix A is not symmetric. It is worth noting that A affects the evolutionary dynamics in two different ways. First, the F matrix is defined via A. By construction, even if the A matrix is not symmetric, the corresponding F matrix will be symmetric. Second, A influences the values of the Q matrix depending upon the learning algorithm used. In the simulations that follow, we have used the memoryless learner in all cases.

Let us now consider some simple examples where some or all symmetries of the A matrix are broken. In order to investigate this we also need to specify the Q matrix which as we saw previously depends both upon the A matrix as well as the learning algorithm. To fix ideas, we will assume in what follows that the learner utilizes the online memoryless algorithm described in the previous section.

The first example assumes that all languages but one are in some sense equivalent, which means that certain symmetries remain in the system, even though the corresponding A matrix is no longer fully symmetric (Section 12.4.1). The second example considers a random configuration of languages leading to a very general system where no symmetries remain (Section 12.4.2).

An analytic consideration of these examples is beyond the scope of the

current chapter. To provide the reader with some insight, however, we present some numerical simulations that reveal the essential character of the results. As before, the population evolution undergoes bifurcations as b changes. For small b, only the uniform solution exists. For large enough b new "one-grammar" solutions (corresponding to X_+) emerge. The bifurcation diagrams are different now and it is seen that some of the languages are suppressed while others are enhanced.

12.4.1 Breaking the symmetry of the A matrix

The A matrices that we have considered so far possessed such symmetries that all one-language solutions (for each language μ_i) were identical. This is not the case in general. All non-symmetrical perturbations of a fully symmetric A matrix lead to the effect of suppressing some languages and enhancing others.

Let us consider the slightest perturbation of the A matrix by replacing one element $a_{ij} = a$ with $a_{ij} = a + \xi$. A simulation was conducted with the following assumptions:

1. Each language μ_i defines a probability distribution over Σ^*. The support of this distribution may be characterized by an underlying grammar G_i. Thus $L_{G_i} \subset \Sigma^*$ is the support of μ_i and a speaker of μ_i produces a sentence $s \in L_{G_i}$ with probability $\mu_i(s)$. With this assumption we have that $a_{ij} = \sum_{s \in L_{G_i} \cap L_{G_j}} \mu_i(s)$.

2. The learning algorithm follows the memoryless procedure described in the previous section.

We observe the following picture. For low values of b, an interior solution (approximately but not exactly uniform) is the only stable one. As b increases, bifurcations occur. The branch of the stable asymmetric solution X_+ corresponding to the grammar G_i will split off from the other one-grammar solutions, whereas solutions with grammars G_l, $l \neq i$, $l \neq j$, will stay together. In other words, the one-grammar solution with G_j as the preferred grammar will deviate ever so slightly from the rest of the grammars. It turns out that if $\xi > 0$, the grammar G_i will be suppressed (and the grammar G_j will be very slightly advantaged), and if $\xi < 0$, the grammar G_i will be enhanced (and G_j will be slightly suppressed).

This means that for $\xi < 0$, the solution with grammar G_i will come into existence earlier (for smaller values of q and b) and will have a larger total fitness (see Fig. 12.6, where $i = 1$, $j = 2$, $a = 0.5$ and $\xi = -0.4$).

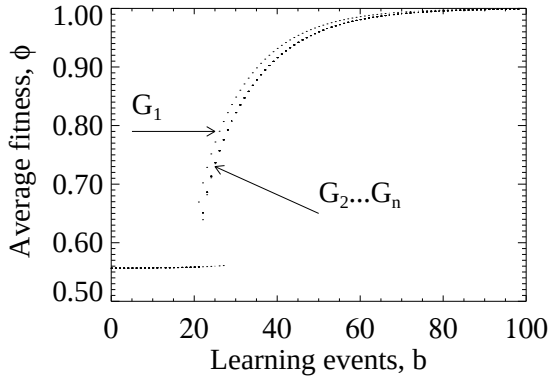

Figure 12.6: The growth rates for the asymmetric matrix A with all off-diagonal $a_{ij} = 0.5$ except $a_{12} = 0.1$. The solution with the G_1 as the preferred grammar is advantageous in comparison with the rest of the one-grammar solutions, it has a higher coherence and comes into existence for smaller values of b. The grammar G_2 is slightly suppressed.

This makes sense because negative (positive) values of ξ mean that the grammar i has a smaller (larger) intersection with the rest of the grammars. When this grammar becomes preferred, it stands out more (less) than other grammars would in its place, i.e. it corresponds to higher (lower) values of X_+ and has a correspondingly larger (smaller) total fitness.

Fig. 12.6 shows a picture of the bifurcation diagram where grammar G_i is the first to emerge and the other grammars emerge later and simultaneously.

12.4.2 Random off-diagonal elements

Let us now consider an example of a non-symmetrical system where the A matrix is composed of random elements. In particular, we take $a_{ii} = 1$ but the off-diagonal elements of the A matrix are random numbers uniformly distributed between zero and one. As a result, no symmetries are left in the system.

Again, bifurcations are seen to occur. For small values of b, an interior solution is the only stable one. All languages are represented in the population. If the number of learning events, b, is high, there are still n stable one-grammar solutions. In Fig. 12.7 one can see 17 of 20 possible stable

solutions.

On the basis of computer simulations, it is possible to make several observations about the dynamics arising from such general matrices.

1. Unlike the symmetric case, one-grammar solutions with different dominant grammars correspond to different values of ϕ, the grammatical coherence of the population. Thus each of the solutions is represented by a separate line. This is consistent with the progression from the symmetric case where all solutions lie on the same line through the partially symmetric case of the previous section where one of the solutions lies on one line and the rest of the solutions lie on another.

2. The number of stable one-grammar solutions grows with b. Some of the grammars become advantaged and have a lower threshold of existence. Some are suppressed until much higher values of b. Such behavior was already present in the previous example where one of the grammars emerged earlier than the rest. The value of b at which the first bifurcation takes place can be roughly predicted by using formula (12.29) with $a = 1/2$, i.e. the average value of the elements a_{ij}. In general, (based on numerical simulations) the first bifurcation point b at which a coherent grammar emerges seems to be roughly estimated with formula (12.29) where we use the average value $a = \langle a \rangle$. Furthermore, the range of the interval over which various grammars emerge (as a result of bifurcations) increases with the range of the distribution of the a_{ij} values.

3. Another interesting feature that can be clearly observed in Fig. 12.7 is that the lowest fitness solution (which corresponds to the uniform solution of the fully symmetric case) flows smoothly into one of the one-grammar solution (the "second best" one for this particular realization of A). This effect can be predicted from standard bifurcation theory. Namely, general perturbations of a pitchfork-like bifurcation will lead to smoothing out sharp edges and avoiding cross-sections, and might also cause the disappearance of "knees" (bistability regions) like those seen in Fig. 12.2.

We conclude that systems with random A matrices behave in a predictable way, and many of the elements of the dynamics can be understood from the analysis of symmetrical systems and their perturbations. However, an extended study of a system with randomly chosen a_{ij} is still needed to

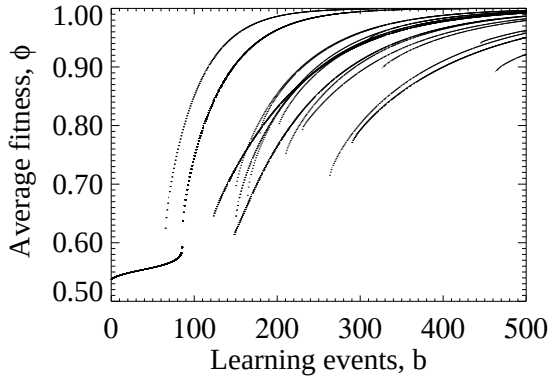

Figure 12.7: The fitness of the system with the A matrix consisting of uniformly distributed random numbers $0 \leq a_{ij} \leq 1$. The number of grammars is $n = 20$, and $f_0 = 0$.

describe the full bifurcation picture. Such a description is beyond the scope of the current chapter. For more analysis, see Komarova and Rivin (2003).

12.4.3 Final Remarks

The general system of equations (Eq. 13.1) may give rise to much more complicated behavior if symmetries in the F and Q matrix are independently controlled (rather than through a common A matrix as in the preceding analysis). For example, choose $n = 3$. Let F be a symmetric matrix as before. For a Q matrix with the following structure

$$Q = \begin{pmatrix} p & q & r \\ r & p & q \\ q & r & p \end{pmatrix}$$

one observes stable oscillations (limit cycles) and spiral attractors depending on the precise values of p, q, r. The reason for this is intuitively clear from the cyclic structure of Q where the errors are such as to induce a rotation. By considering larger systems and having different cycles interact with each other, one may have a path to chaos through a period doubling mechanism. These discussions are beyond the scope of the current chapter but some analysis pertaining to these questions may be found in Mitchener (2003).

12.5 Conclusions

In this chapter, we have studied in some detail the evolution of a population of language users while making two major assumptions about them

1. *Natural Selection:* We assume that communicative fitness confers reproductive success so that the rate at which indivduals produce offspring is related to their communicative efficiency with the rest of the population at large.

2. *Local Learning:* We assume that language is not transmitted genetically from parent to child. Rather, languages have to be learned by children. Crucially, however, we assume that children learn from their parents.

Under these assumptions, the primary question we investigate is when *coherence*, i.e., a shared language, would emerge in the population. As a result of the analysis of this chapter, we now see that grammatical coherence is possible only if the learning accuracy of children is sufficiently high.

Bifurcations are seen to occur in the population dynamics as the learning fidelity changes. For low learning accuracy, the only stable mode is the uniform mode where all languages are used in the population with roughly equal frequency. As the learning accuracy increases, new *one language* modes arise where most of the population speaks one language and a small fraction speaks a smattering of the rest. The system undergoes spontaneous symmetry breaking. There are n stable equilibria. Which one is attained depends upon the initial conditions.

While the general relationship between the fidelity thresholds (q_1 and q_2), the number of learning events b, and the size of the space of languages n is intuitive enough, the details depend upon the precise learning procedure that children use. As an illustrative example we have considered here two learning algorithms that make very different demands on the cognitive ability of individuals — a batch learner and a memoryless learner. A memoryless learner requires $O(n)$ examples for language emergence while the batch learner requires $O(\log(n))$ examples for the same.

The equations of language dynamics studied here are similar to the *quasi-species* equations of evolutionary biology (Eigen & Schuster, 1979) but with frequency dependent fitness terms. In genetic evolution, genes are transmitted from parents to children with possible mutations. The mutation rate

determines the fidelity of such a transmission. In contrast, in linguistic populations, languages are transmitted from parents to children via *learning*. The learning accuracy determines the fidelity of such a transmission.

It is also worthwhile to reflect on the potential role of natural selection in linguistic evolution. Natural selection may be interpreted in two different ways. In its most literal interpretation, it seems unlikely that it has played a role in historical times. It is certainly possible, however, that on evolutionary time scales the ability to communicate effectively affects the chances of survival and therefore reproduction. One can, however, provide alternative interpretations of natural selection in terms of social influence and imitatative behavior on the part of mature adults and child learners respectively. Individuals with higher communicative fitness may have a higher influence on learning children (alternatively they may be imitated more often). As a result the effective number of child learners attempting to learn a particular mature adult's language may depend upon the communicative efficiency of that adult. This interpretation is also consistent with the same set of equations analyzed in this chapter.

Several variations may be considered. We provided a brief discussion of asymmetric configurations of languages. Bifurcations are still seen to occur. Each language now has a different threshold before it can emerge but the essential spirit of the results remains the same as in the symmetric case. The effect of finite populations, spatial distributions of the language speakers, bilingual learning and so on can be studied as in previous chapters. We do not pursue such variations here.

One might choose to tell an evolutionary story based on the results of this chapter. For language to emerge in a population, the learning fidelity must be high. It is possible that over evolutionary time scales, the learning ability of individual children crossed the coherence threshold. This may be due to

1. the separate evolution of learning algorithms
2. the increased maturation time of learners allowing them more examples (b) with which to learn and therefore greater fidelity
3. a decrease in the size (n) of the search space of possible languages.

Each of these is consistent with the evolution of language. In general, though, this raises an important and puzzling question. Why is there linguistic variation at all? It is clear that there is an evolutionary pressure for n to be small.

It is not clear what counteracting pressures there are for n to be large. If such pressures were found, then the tension between the two would support intermediate sizes of universal grammar and would therefore be consistent with the state of affairs in naturally occurring linguistic systems. We leave such questions for future research.

Chapter 13

The Origin of Communicative Systems: Linguistic Coherence and Social Learning

In this chapter, we continue our investigation of the conditions under which linguistic *coherence* might emerge in a population of linguistic agents. We are primarily interested in understanding how a community might arrive at a shared language through the actions of interacting individual learners in the absence of a centralizing agent that enforces linguistic homogeneity in some sense.

From the analysis of the previous chapter, we already appreciate the strong yet subtle relationship between the learning fidelity of individual learners and the emergence of shared languages in the population. We examined a setting with the following assumptions: (i) learning agents (children) learn from one teacher (parent) (ii) communicative efficiency (fitness) translates into reproductive success. Under those conditions, we developed a model where we were able to derive a coherence threshold (corresponding to a bifurcation point) that depended upon the learning fidelity of the individual learner. Below this coherence threshold, the only stable mode was the uniform mode when all linguistic types were equally represented in the population. One may liken this condition to the *Tower of Babel* where many different languages are present and communicative efficiency is low. Above this coherence threshold, *one language* modes that correspond to population

states with shared languages become possible. The communicative efficiency is therefore high.

In this chapter, we consider alternative models in which coherence might emerge. The most significant dimension in which these models differ from each other is in the analysis of individual child learners. Under our model assumptions, we shall see that if children learn only from their parents, then fitness is necessary to ensure coherence. Without fitness, coherence will never emerge, no matter how good the learning ability of individual children. In contrast, if children learn from a wider pool of people than just their parents, coherence might emerge *without* natural selection (fitness). Thus *social learning* may result in shared languages.

Under symmetry conditions that parallel those of the previous chapter, we develop a model for the evolution of n linguistic types in the population. Each linguistic type is equally easy to learn. The learner is exposed, however, to data from a mixture of these types and ends up learning a language that is most consistent with its data set. We analyze the model and characterize equilibria and stability. As before, we find that bifurcations exist. If the number of examples provided to learners is small so that learning fidelity is low, the only equilibrium is the uniform mode where all linguistic types are equally present in the population. As the number of examples increases, a bifurcation point is reached after which the uniform solution becomes unstable and stable one grammar solutions now arise.

To set the stage for the arguments that follow, let us briefly recapitulate our findings for the setting when children learn only from their parents.

13.1 Learning Only From Parents

If there are n linguistic types given by the measures μ_1 through μ_n, then following the arguments of the previous chapter, the linguistic evolution of the population is characterized by the following equation.

$$\dot{x_j} = \sum_i f_i x_i Q_{ij} - \phi x_j, \quad 1 \leq j \leq n, \tag{13.1}$$

Here x_j is the proportion of individuals in the population that speak the language corresponding to μ_j. An individual born to a speaker of μ_i learns a language based on data provided by its parent. The fidelity of the language learning map is provided by the matrix Q where Q_{ij} denotes the probability with which an offspring of a speaker of μ_i ends up learning μ_j. The fitness

f_i of a speaker of μ_i is given by the expression $f_i = \sum_{j=1}^{n} x_j F(\mu_i, \mu_j)$ where $F(\mu_i, \mu_j)$ is the mutual intelligibility between a speaker of μ_i and a speaker of μ_j.

Under the symmetric assumptions of the previous chapter, we have

1. (i) $Q_{ii} = q$ $(ii) Q_{ij} = \frac{1-q}{n-1}; i \neq j$
2. (i) $F(\mu_i, \mu_i) = 1$ $(ii) F(\mu_i, \mu_j) = a; i \neq j$
3. $f_i = \sum_j x_j F(\mu_i, \mu_j) = (1-a)x_i + a + f_0$

When evolution is governed by fitness as in Eq. 13.1, we see there is a coherence threshold (q_1) for learning fidelity (given by q). When $q < q_1$, the only stable mode of the population is the uniform mode in which all linguistic types are equally represented. When $q > q_1$, new *one-grammar* solutions emerge. These correspond to population states in which a majority speak a shared language.

Let us now consider the same evolutionary dynamics but without having a fitness that depends on communicative efficiency F. In other words, we assume

$$\text{for all } i, \quad f_i = f_0$$

Therefore, we have $\phi = f_0 \sum_i x_i = f_0$. For this case, the dynamical equations reduce to

$$\dot{x_j} = \sum_i f_0 x_i Q_{ij} - f_0 x_j, \quad 1 \leq j \leq n, \tag{13.2}$$

Note that this is a set of linear differential equations. Equilibria are computed by setting $\dot{x_j} = 0$. This leads to linear equations for which the only solution is given by the uniform solution $x_j = \frac{1}{n}$ for all j. Thus from all initial conditions, populations move to an equilibrium state in which all languages are equally represented in the population. No bifurcations occur and no shared languages ever emerge.

The above discussion clarifies how fitness based on communicative efficiency is an important ingredient in the emergence of communal or shared languages. In many cases of animal communication, for example, in some species of song birds, the infant of the species learns in the *nesting* phase where it is exposed to primarily one teacher. In such cases, it seems likely that one will need to invoke arguments from natural selection and fitness to provide an explanation for the emergence of shared communication systems.

On the other hand, we shall soon see that if learning is based on input provided by the population at large, i.e., social learning, then shared systems might emerge without natural selection.

13.2 Social Learning: Learning From Everybody

Now consider the case when the individual learner learns on the basis of the linguistic input derived from the entire adult population. This setting was considered in some detail in part III of this book. We now re-examine those models from the point of view of coherence.

13.2.1 The Symmetric Assumption

The analysis of language evolution in the previous chapter was conducted for the symmetric case where $F(L_i, L_i) = 1$ and $F(L_i, L_j) = a$ when $i \neq j$. For ease of analysis we will make an analogous symmetric assumption now.

Let there be n languages $L_1, L_2, \ldots, L_n \subset \Sigma^*$. For each language L_i, we assume there are a set of expressions that are perceptually salient for the purpose of learning. Depending upon one's theoretical persuasion, there are many possible candidates for such a set. Examples of such expressions may be "triggers" of various sorts as discussed in trigger based accounts of language acquisition (e.g. Fodor, 1998; Gibson and Wexler, 1994), "cues" as in Lightfoot (1998), or in general, any expression that provides a linguistic indicator to the child regarding the identity of the grammar that generated the expression. Following the learning-theoretic discussion in Chapter 2, we may even let the set of such expressions to be $L_i \setminus (\cup_{j \neq i} L_j)$. We denote the set of such perceptually salient expressions to be $C_i \subseteq L_i$ such that $C_i \cap C_j = \phi$ for all $i \neq j$.

Speakers of L_i produce sentences according to a probability distribution μ_i. In particular, let

$$\mu_i(C_i) = a_i$$

Our general model of learning will be as follows. The learning child scans its input for cues. If the cues for L_i occur often enough, the child will acquire the language L_i. A number of learning algorithms may be designed around this general principle by taking different computational and cognitive requirements into account. If $a_i > 0$ for every i, i.e., every language has a non-empty (non-measure zero) cue set, then these learning algorithms will be able to identify every language in the family $\{L_1, \ldots, L_n\}$.

In the analysis that follows, we will assume $a_i = a$ for every i. Thus the languages have cue sets of equal measure and are therefore equally easy to learn by unbiased algorithms. This amounts to a symmetric assumption about the ease of learning languages.

13.2.2 Coherence for $n = 2$

We begin by considering the case in which there are two possible languages — L_1 and L_2. Each has a cue set — C_1 and C_2 respectively. Speakers of L_1 produce sentences with a probability distribution μ_1 such that $\mu_1(C_1) = a$. Speakers of L_2 produce sentences according to μ_2 such that $\mu_2(C_2) = a$.

At any point in time, there may be a mixture of L_1 and L_2 speakers. Let the proportion of L_1 speakers (at time t) be $x_1(t)$ and the proportion of L_2 speakers be $x_2(t)$. Children are exposed to both kinds of speakers and therefore potentially hear both kinds of cues. Depending upon the ratio of L_1 and L_2 speakers in their linguistic environment, they are more likely to hear one or the other type of cue.

Let us consider the evolution of this population under the following learning algorithm.

Cue-frequency based batch learner

The learning algorithm receives k examples in all. The algorithm counts

1. k_1: the number of examples that are cues for L_1, i.e., examples that belong to C_1

2. k_2: the number of examples that are cues for L_2, i.e., examples that belong to C_2

3. k_3: the number of examples that are not a cue for either language

Clearly, $k = k_1 + k_2 + k_3$. The learning procedure is empirically driven and simple. If (i) $k_1 > k_2$, the learner chooses L_1 (ii) if $k_2 > k_1$, the learner chooses L_2 (iii) if $k_1 = k_2$, the learner chooses any one of the two languages (with probability $\frac{1}{2}$ each).

If this is the learning algorithm the typical child uses, one may compute the probability with which such a child acquires L_1. Assuming that sentences are produced in i.i.d. fashion, the probability that $k_1 > k_2$ is given by

$$f_1(a, x_1(t), x_2(t), k) = \sum_{\{(k_1,k_2,k_3)\in I_1\}} \binom{k}{k_1 k_2 k_3} (ax_1(t))^{k_1} (ax_2(t))^{k_2} (1-a)^{k_3}$$

where

$$I_1 = \{(k_1, k_2, k_3) | k_1 > k_2;\ k_1 + k_2 + k_3 = k\}$$

Similarly, the probability that $k_2 > k_1$ is given by

$$f_2(a, x_1(t), x_2(t), k) = \sum_{\{(k_1,k_2,k_3)\in I_2\}} \binom{k}{k_1 k_2 k_3} (ax_1(t))^{k_1} (ax_2(t))^{k_2} (1-a)^{k_3}$$

where

$$I_2 = \{(k_1, k_2, k_3) | k_2 > k_1;\ k_1 + k_2 + k_3 = k\}$$

Note that by symmetry, we have

$$f_2(a, x_1(t), x_2(t), k) = f_1(a, x_2(t), x_1(t), k)$$

The probability that a typical child acquires L_1 after k sentences is now given by

$$f_1 + \frac{1}{2}(1 - f_2 - f_1)$$

Therefore, the population dynamics is given by

$$x_1(t+1) = \frac{1}{2}(1 + f_1(a, x_1(t), x_2(t), k) - f_2(a, x_1(t), x_2(t), k)) \tag{13.3}$$

Using the fact that $x_1(t) + x_2(t) = 1$ for all t, we can eliminate $x_2(t)$ from the above equation to obtain a one dimensional map $g : [0,1] \longrightarrow [0,1]$ such that $x_1(t+1) = g(x_1(t))$. For example, g may be expressed in terms of f_1 as

$$g(x) = \frac{1}{2} + \frac{1}{2}\left(f_1(a, x, 1-x, k) - f_1(a, 1-x, x, k)\right) \tag{13.4}$$

Evolutionary Dynamics

Eq. 13.4 determines the evolution of L_1 types in the population. The following observations may now be made:

1. The dynamics depends upon the number of example sentences k that individual children hear before maturation. In particular, g is a kth order polynomial map.

2. A fixed point is provided by $x_1 = x_2 = \frac{1}{2}$. This corresponds to the uniform solution where both languages are spoken in equal proportion.

3. For small values of k, this is the only fixed point. It is stable.

4. As k increases, new coherent states emerge where one of the languages is the dominant language spoken by a majority of the agents. These correspond to stable *one language* modes.

5. As k increases, the uniform state becomes unstable.

6. The critical values of k at which the bifurcations occur depend upon the value of a. In general these critical values become larger as a becomes smaller. Since k takes on integer values, it is more natural to hold k fixed and study the bifurcations as a changes continuously from 0 to 1. When a is close to 0, only the uniform mode is stable. As a increases, the bifurcations occur, the uniform mode becomes unstable and new stable one language equilibria arise.

7. The values of a, k may be related to learning fidelity in a natural way. When a, k have small values, the learner is given too little information on the basis of which language is learned. As a result, learning fidelity is low, the system is noisy and shared languages do not emerge. When a, k have large values, the learner is given a lot of information on the basis of which language is learned, learning fidelity is high, the system is less noisy, and shared languages emerge.

To see (2), simply put in $x_1 = x_2 = \frac{1}{2}$ and notice that $f_1 = f_2$ for this situation.

To see (3), let us evaluate the map g for $k = 2$ and $k = 3$ respectively. This will also provide some insight into the relationship between a and k for bifurcations to occur. Consider $k = 2$. It is easy to check that

$$f_1(a, x_1, x_2, 2) = (ax_1)^2 + 2(ax_1)(1 - a)$$

and

$$f_2(a, x_1, x_2, 2) = (ax_2)^2 + 2(ax_2)(1-a)$$

Putting this into Eq. 13.3, we have

$$g(x_1) = \frac{1}{2} + \frac{2x_1 - 1}{2}(2a - a^2)$$

Clearly $x_1 = \frac{1}{2}$ is the only solution.
Now consider $k = 3$. For this case,

$$f_1(a, x_1, x_2, 3) = (ax_1)^3 + 3(ax_1)^2(ax_2) + 3(ax_1)^2(1-a) + 3(ax_1)(1-a)^2$$

A similar expression holds for f_2. Substituting into Eq. 13.3 we have

$$g(x_1) = \frac{1}{2} + \frac{x_1 - x_2}{2}\left(a^3(x_1^2 + x_1x_2 + x_2^2) + 3a^2x_1x_2 + 3a^2(1-a) + 3a(1-a)^2\right)$$

where $x_1 + x_2 = 1$. This further simplifies to

$$g(x_1) = \frac{1}{2} + \frac{2x_1 - 1}{2}\left(1 - (1-a)^3 + 2a^3x_1(1-x_1)\right)$$

Clearly, g is a polynomial of degree 3 and solving for its fixed points yields three solutions:

$$x = \frac{1}{2}; x = \frac{2a^3 \stackrel{+}{-} \sqrt{4a^6 - 8a^3(1-a)^3}}{4a^3}$$

The second pair of solutions exist only when

$$4a^6 - 8a^3(1-a)^3 \geq 0$$

or

$$(\frac{a}{1-a})^3 \geq 2$$

Thus we see that for $k = 2$, the uniform solution is the *only* solution. For $k = 3$, additional solutions exist only if a is large enough. If a is too small, then k will need to be much higher than 3 for bifurcations to occur. This already reflects the behavior suggested by (4),(5),and (6) in the above list.

We will now provide arguments to gain some insight into the validity of (4),(5), and (6). Let us begin by concentrating on the stability of the fixed point at $x = \frac{1}{2}$. From Eq. 13.4, we have that

$$g'(x) = \frac{1}{2}[f_1'(a, x, 1-x, k) - f_1'(a, 1-x, x, k)]$$

Now

$$f_1'(a,x,1-x,k) = \sum_{k_1>k_2} \binom{k}{k_1k_2k_3} a^{k-k_3}(1-a)^{k_3}\left(k_1x^{k_1-1}(1-x)^{k_2} - k_2x^{k_1}(1-x)^{k_2-1}\right)$$

Putting in $x = \frac{1}{2}$, we see

$$f_1'(a,\frac{1}{2},\frac{1}{2},k) = \sum_{k_1>k_2} \binom{k}{k_1k_2k_3} a^{k-k_3}(1-a)^{k_3}(k_1-k_2)(\frac{1}{2})^{k_1+k_2-1}$$

A similar calculation reveals that

$$f_2'(a,\frac{1}{2},\frac{1}{2},k) = \sum_{k_1>k_2} \binom{k}{k_1k_2k_3} a^{k-k_3}(1-a)^{k_3}(k_2-k_1)(\frac{1}{2})^{k_1+k_2-1}$$

We thus have

$$g'(x=\frac{1}{2},a,k)) = a\sum_{k_1>k_2} \binom{k}{k_1k_2k_3} (1-a)^{k_3}(k_1-k_2)(\frac{a}{2})^{k-k_3-1} \tag{13.5}$$

We immediately see from this expression that $g'(x=\frac{1}{2},a=0,k)) = 0$ for all values of $k \geq 1$. By continuity and differentiability of g', we see that for each k, there exists a sufficiently small a_k such that $|g'(x=\frac{1}{2},a,k))| < 1$ for all $a < a_k$. Now let us turn our attention to $g'(x=\frac{1}{2},a,k))$ for the case $a = 1$. The following proposition is true.

Proposition 5 *For $a = 1$, the uniform solution becomes unstable for large k. In particular, we have*

$$\lim_{k\to\infty} g'(x=\frac{1}{2},a=1,k) = \infty$$

Proof: From Eq. 13.5, we have

$$g'(x=\frac{1}{2},a=1,k)) = \sum_{k_1>k_2;k_3=0} \binom{k}{k_1k_2k_3}(k_1-k_2)(\frac{1}{2})^{k-k_3-1} = 2\sum_{l>\frac{k}{2}} \binom{k}{l}(l-(k-l))(\frac{1}{2})^k.$$

It is sufficient to show that $\sum_{l>\frac{k}{2}} \binom{k}{l}(2l-k)(\frac{1}{2})^k$ grows to ∞ as a function of k. Notice that

$$\sum_{l>\frac{k}{2}} \binom{k}{l}(2l-k)(\frac{1}{2})^k \geq \sum_{l>\frac{k}{2}+k^{\frac{1}{6}}} \binom{k}{l}(2l-k)(\frac{1}{2})^k \geq \sum_{l>\frac{k}{2}+k^{\frac{1}{6}}} \binom{k}{l}k^{\frac{1}{6}}(\frac{1}{2})^k$$

The quantity $\sum_{l>\frac{k}{2}+k^{\frac{1}{6}}} \binom{k}{l}(\frac{1}{2})^k$ is the probability with which an unbiased coin would turn up heads at least $\frac{k}{2}+k^{\frac{1}{6}}$ times in k independent tosses. Denote this probability by P_k. Then by an application of the Central Limit Theorem, we know that

$$\lim_{k\to\infty} P_k = \lim_{k\to\infty} \phi\left(\frac{-k^{\frac{1}{6}}}{\sqrt{\frac{k}{4}}}\right) = \frac{1}{2}$$

where ϕ is the cumulative distribution of the univariate Normal. Therefore, we have

$$\lim_{k\to\infty} g'(x=\frac{1}{2}, a=1, k)) \geq \lim_{k\to\infty} k^{\frac{1}{6}} P_k = \infty$$

■

From this proposition, we see that there exists a K such that for all $k > K$, $g'(x=\frac{1}{2}, a=1, k) > 1$. At the same time, we know that $g'(x=\frac{1}{2}, a=0, k) = 0$. Therefore, by continuity, there exists for each $k > K$ a critical point a_k such that $g'(x=\frac{1}{2}, a=a_k, k) = 1$. This corresponds to the bifurcation point at which the uniform solution becomes unstable as a changes continuously.

Once the uniform solution becomes unstable, new stable solutions must arise. These solutions correspond to situations where one language is spoken by a majority of the population. Thus shared languages emerge. To see this, fix a $k > K$ and choose $a > a_k$ so that the uniform solution is unstable, i.e, $g'(x=\frac{1}{2}, a, k) > 1$. Now notice that $g(x=1, a, k) < 1$ and $g(x=0, a, k) > 0$ while $g(x=\frac{1}{2}, a, k) = \frac{1}{2}$. Since $g'(x=\frac{1}{2}, a, k) > 1$, there exist two points (x_- and x_+) in the neighborhood of $\frac{1}{2}$ such that $g(x_+)-x_+ > 0$ and $g(x_-)-x_- < 0$. Consider the function $h(x) = g(x, a, k) - x$ and note that (i) $h(0) > 0$ and $h(x_-) < 0$ and (ii) $h(1) < 0$ and $h(x_+) > 0$. By continuity of h, we see that there are stable equilibrium points $x_{*,-} \in (0, \frac{1}{2})$ and $x_{*,+} \in (\frac{1}{2}, 1)$. These correspond to situations where a majority speak L_2 and L_1 respectively.

We do not have an analytic form for the relationship between a_k and k. However, from numerical simulations, it seems to be the case that a_k decreases as k increases. Therefore, we see that if $a = a_*$, the one language solutions arise only when k is large enough, i.e., for all k such that $a_k < a_*$. As an example, we show in Fig. 13.1 the bifurcation diagrams for $k = 14$ and $k = 7$ respectively. These correspond to classic pitchfork bifurcations where two stable points arise simultaneously as the uniform fixed point becomes unstable.

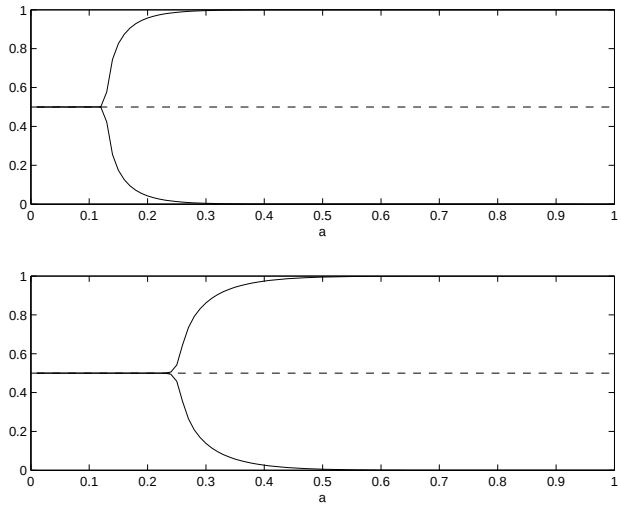

Figure 13.1: Bifurcation diagrams for $k = 14$ (top) and $k = 7$ (bottom) as the cue-frequency a varies.

13.3 Coherence for general n

We now consider the more general case in which there are n linguistic types corresponding to n languages $L_1, L_2, \ldots, L_n$. For each language L_i, there is an associated cue sets $C_i \subset L_i$. Elements of the cue set may be interpreted as expressions that provide a clue to the learner regarding the identity of the grammatical system underlying L_i. Speakers of L_i produce sentences with probability distribution μ_i such that $\mu_i(C_i) = a$.

As before, child learners hear a total of k sentences from the mixture of languages in their linguistic environment. Consider the following learning algorithm they may use to learn a language.

13.3.1 Cue-frequency based batch learner

Children scan the input for cues to each of the languages and choose the language for which maximally many cues occur in their linguistic experience. If there are multiple languages with the same number of cues, then a language is chosen at random.

1. *Count Cues*: Let k_i $(i = 1, \ldots, n)$ be the number of sentences in the data set that belong to C_i. Let k_{n+1} be the number of non-cues that occur in the child's linguistic experience. Clearly, $\sum_{i=1}^{n+1} k_i = 1$.

2. *Find Maximal Languages*: Determine the languages whose cues occur most often. Let $I = \{i | k_i = \max_{1 \leq j \leq n} k_j\}$. Thus I consists of the indices of such maximal languages.

3. *Choose Mature Language*: If there is a single language whose cues occur most often, i.e., $|I| = 1$, then this language is chosen as the mature language. Otherwise, a language is chosen at random. There are two variations:

   (a) Simple minded: choose any one of the n languages with probability $\frac{1}{n}$.

   (b) Careful: Let $L = \{L_i | i \in I\}$. Choose one of the languages in L with probability $\frac{1}{|I|}$.

13.3.2 Evolutionary Dynamics of Batch Learner

One may now try to characterize the population dynamics under this learning algorithm. Let the state of the population at time t be given by $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_n(t))'$ where $x_i(t) \geq 0$ and $\sum_{i=1}^{n} x_i(t) = 1$. The probability distribution with which sentences are presented to the typical child is now given by

$$\mu = \sum_{i=1}^{n} x_i(t)\mu_i$$

With probability $ax_1(t)$, the child will receive a cue for L_1, and in general, with probability $ax_i(t)$, it will receive a cue for L_i. Finally, with probability $1 - a$, a non-cue will be heard. Assuming sentences are presented in i.i.d. fashion according to μ, the probability of hearing k_1 cues for L_1, k_2 cues for L_2 and so on is given by

$$\binom{k}{k_1 k_2 \ldots k_n k_{n+1}} p_1^{k_1} p_2^{k_2} \ldots p_{n+1}^{k_{n+1}}$$

where $p_i = ax_i(t)$ for $i \in \{1, \ldots, n\}$ and $p_{n+1} = 1 - a$. Therefore the probability that k_1 is strictly greater than $k_2, \ldots, k_n$ is given by

$$F_1^k(\mathbf{x}(t)) = \sum_{\mathbf{k} \in I_1} \binom{k}{\mathbf{k}} p_1^{k_1} p_2^{k_2} \ldots p_{n+1}^{k_{n+1}}$$

where we use $\mathbf{k}$ to denote the $n+1$-tuple given by $(k_1, \ldots, k_{n+1})$ and $\binom{k}{\mathbf{k}}$ denotes the multinomial quantity $\binom{k}{k_1 k_2 \ldots k_{n+1}}$. The sum is taken over the set I_1 which consists of all ordered partitions of k into $n+1$ non-negative integers $(k_1, \ldots, k_{n+1})$ such that k_1 is strictly greater than $k_2, \ldots, k_n$. In other words,

$$I_1 = \{\mathbf{k} = (k_1, k_2, \ldots, k_n, k_{n+1}) | k_1 > k_2, \ldots, k_n; \sum_{i=1}^{n+1} k_i = 1\}$$

In a similar way, the probability that k_2 is strictly greater than $k_1, k_3, \ldots, k_n$ can be calculated. Let this be $F_2^k(\mathbf{x}(t))$. We can thus define $F_1^k, \ldots, F_n^k$. Note that for any i, j, we have

$$F_i^k(\ldots, x_i, \ldots, x_j, \ldots) = F_j^k(\ldots, x_j, \ldots, x_i, \ldots)$$

Under the simple minded version of the batch learning algorithm described above, we can compute the probability that the learner will choose L_1 after k examples have been heard. This is given by

$$F_1^k(\mathbf{x}(t)) + (1 - \sum_{i=1}^{n} F_i^k(\mathbf{x}(t)))\frac{1}{n}$$

Thus the population dynamics is given by a map $f^k : \Delta^{n-1} \to \Delta^{n-1}$ where Δ^{n-1} is the $(n-1)$-dimensional simplex in $\mathbb{R}^n$ given by

$$\Delta^{n-1} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \sum_i x_i = 1, (\forall i) x_i \geq 0\}$$

The map $f^k = (f_1^k, f_2^k, \ldots, f_n^k)$ has n components where the jth component is given by

$$f_j^k(\mathbf{x}(t)) = x_j(t+1) = F_j^k(\mathbf{x}(t)) + (1 - \sum_{i=1}^{n} F_i^k(\mathbf{x}(t)))\frac{1}{n} \tag{13.6}$$

One may now investigate the evolutionary dynamics associated with the map f^k with a particular view to the emergence of linguistic coherence. The following observations may now be made.

1. Since f^k is a continuous map from Δ^{n-1} to itself, in general it may have a continuum of fixed points. It is possible to show, however, that due to the special structure of f^k, there can be at most a *finite* number of fixed points (equilibria). In particular, for any $a \in [0, 1]$ and $k > 0$, there are no more than $k(2^n)$ equilibria. Furthermore, all fixed points like on critical lines in the simplex that we describe at length below.

2. For small values of k, the only fixed point of the map is given by the uniform solution $\mathbf{x} = (\frac{1}{n}, \ldots, \frac{1}{n})$. This is stable.

3. For any fixed k that is sufficiently large, the number of fixed points varies with a. Thus as a changes from 0 to 1, there is a bifurcation in the dynamics. For small values of a, there is only one fixed point, the uniform point where all languages are equally represented in the population. As a increases and crosses a critical point, other fixed points arise.

4. For large values of a, only the one language solutions (where one of the languages dominates the population) are stable. The rest (including the uniform solution) are unstable.

5. The same behavior may also be observed by holding a fixed and changing k. For small values of k the uniform point is the only fixed point and the other fixed points arise as k increases beyond a critical value. The values of a, k may be related to learning fidelity in the natural way.

13.4 Proofs of Evolutionary Dynamics Results

In this section, we provide formal proofs and informal arguments supporting the analysis of the evolutionary dynamics outlined earlier.

13.4.1 Preliminaries

Ultimately, we wish to analyze the dynamics associated to the map f^k. It is useful, however, to introduce some intermediate maps along the way.

Definition 20 *Let $\ell \geq 1$. The function $h^\ell : \Delta^{n-1} \to \Delta^{n-1}$ is defined by $h_i^\ell(\mathbf{x})$ = the conditional probability that, given exactly ℓ cues for the languages $L_1, \ldots, L_n$, and the existence of a unique language with the most cues, that L_i is that language.*

Observation: Let $p^\ell(\mathbf{x})$ denote the probability that, given language distribution $\mathbf{x}$, and l cues, there is a unique most-cued language in the learning experience. For $1 \leq i \leq n$, let S_i be the set of ordered partitions of ℓ

into n non-negative integers $(k_1, \ldots, k_n)$ such that k_1 is strictly greater than $k_2, \ldots, k_n$. h^ℓ is given by the formula

$$h_i^\ell(\mathbf{x}) = \frac{1}{p^\ell(\mathbf{x})} \sum_{\mathbf{k} \in S_i} \binom{\ell}{\mathbf{k}} \prod_{j=1}^{n} x_j^{k_j}.$$

Note also that p^ℓ is given by the formula

$$\begin{aligned} p^\ell(\mathbf{x}) &= \sum_{i=1}^{n} \sum_{\mathbf{k} \in S_i} \binom{\ell}{\mathbf{k}} \prod_{j=1}^{n} x_j^{k_j} \\ &= \sum_{\mathbf{k} \in S_1} \binom{\ell}{\mathbf{k}} \sum_{i=1}^{n} \prod_{j=1}^{n} x_j^{k_{\sigma_i(j)}} \\ &= \sum_{\mathbf{k} \in S_1} \binom{\ell}{\mathbf{k}} \left(\prod_{j=1}^{n} x_j^{k_j} \right) \sum_{i=1}^{n} \left(\frac{x_i}{x_1} \right)^{k_1 - k_i} \end{aligned}$$

where $\sigma_i(j) = 1$ if $j = i$, i, if $j = 1$, and j otherwise.
Observation: For $\ell = 1$, h^ℓ is the identity function. For $\ell \geq 2$, h^ℓ has exactly $2^n - 1$ fixed points, namely those inputs $\mathbf{x} = (x_1, \ldots, x_n)$ for which all nonzero coordinates x_i are equal.

Definition 21 *Let $\ell \geq 0$. The function $g^\ell : \Delta^{n-1} \to \Delta^{n-1}$ is defined by $g_i^\ell(\mathbf{x})$ = the conditional probability that, given exactly ℓ cues for the languages $L_1, \ldots, L_n$, that L_i is the selected language. Recall that, if there is no unique most-cued language among $L_1, \ldots, L_n$, then the chosen language is selected uniformly at random from among all n (not just the most-cued).*

Observation: g^ℓ is given by the formula

$$g^\ell = p^\ell h^\ell + (1 - p^\ell)\mathbf{u},$$

where $\mathbf{u}$ is the constant function $(1/n, \ldots, 1/n)$. In particular, g^ℓ is a convex combination of h^ℓ and $\mathbf{u}$.

Definition 22 *Let $k > 0$, $a \in [0, 1]$. The function $f^k : \Delta^{n-1} \to \Delta^{n-1}$ is defined by $f_i^k(\mathbf{x})$ = the probability that L_i is the selected language, given k sentences for the languages $L_1, \ldots, L_n$, each drawn with probability x_i, and having probability a of being a cue. Recall that, if there is no unique most-cued language among $L_1, \ldots, L_n$, then the chosen language is selected uniformly at random from among all n (not just the most-cued).*

Observation: f^k is given by the formula

$$f^k = \sum_{\ell=0}^{k} \binom{k}{\ell} a^\ell (1-a)^{k-\ell} g^\ell.$$

This shows that f^k is a convex combination of $g^0, \ldots, g^k$, which implies that f^k is a convex combination of $h^1, \ldots, h^k$ and $\mathbf{u}$. When $a = 1$, $f^k = g^k$.

13.4.2 Equilibria

In this section, we show that f^k may have at most a finite number of fixed points (Corollary 6). Furthermore, these fixed points have a special structure. If $\mathbf{x} = (x_1, \ldots, x_n)$ is a fixed point of f^k, then there can be at most two distinct values for the x_i's (Corollary 5). We begin with a key lemma.

Lemma 5 *Let $\ell > 1$, and suppose $\mathbf{x} = (x_1, \ldots, x_n)$, where $x_1 > x_2 > x_3$. Let M be the 3×3 matrix*

$$M := \begin{pmatrix} h_1^\ell(\mathbf{x}) & x_1 & 1 \\ h_2^\ell(\mathbf{x}) & x_2 & 1 \\ h_3^\ell(\mathbf{x}) & x_3 & 1 \end{pmatrix}.$$

Then $\det(M) > 0$.

Proof: We will show how to rewrite the left column of M as a positive linear combination of columns of the form

$$\begin{pmatrix} x_1^A \left(x_2^B - x_3^B\right) / (x_2 - x_3) \\ x_2^A \left(x_3^B - x_1^B\right) / (x_3 - x_1) \\ x_3^A \left(x_1^B - x_2^B\right) / (x_1 - x_2) \end{pmatrix},$$

where $A \geq B$ are positive integers. Once this is established, the result follows by multilinearity of the determinant, together with the observation that

$$\begin{vmatrix} x_1^A \left(x_2^B - x_3^B\right) / (x_2 - x_3) & x_1 & 1 \\ x_2^A \left(x_3^B - x_1^B\right) / (x_3 - x_1) & x_2 & 1 \\ x_3^A \left(x_1^B - x_2^B\right) / (x_1 - x_2) & x_3 & 1 \end{vmatrix} = \begin{vmatrix} 1 & x_3^B & x_3^A \\ 1 & x_2^B & x_2^A \\ 1 & x_1^B & x_1^A \end{vmatrix} \geq 0.$$

Equality holds iff $A = B$, since when $A > B$, the second matrix is a submatrix of a doubly increasing Vandermonde matrix ($x_3 < x_2 < x_1$ and $0 < B < A$).

Now, by definition of h^ℓ, we know

$$p^\ell h_1^\ell(\mathbf{x}) = \sum_{\mathbf{k}\in S_1} \binom{\ell}{\mathbf{k}} \prod_{j=1}^{n} x_j^{k_j}$$

Fix values for $k_1, k_4, k_5, \ldots, k_n$. Observe that this fixes the sum $S = k_2 + k_3$, and allows (k_2, k_3) to take on exactly the values $(S-M, M), (S-M+1, M-1), \ldots, (M, S-M)$, where $M = \min\{k_1 - 1, S\}$ is the maximum allowable value for k_2 or k_3. Also note that the coefficient $\binom{\ell}{\mathbf{k}}$ is a decreasing function of $|k_2 - k_3|$. This allows us to rewrite our sum as

$$p^\ell h_1^\ell(\mathbf{x}) = \sum_{k_1,k_4,\ldots,k_n} \prod_{j\neq 2,3} x_j^{k_j} \sum_{J=0}^{M-\lceil S/2\rceil} \alpha(\mathbf{k}, J) \sum_{K=S-M+J}^{M-J} x_2^K x_3^{S-K}, \tag{13.7}$$

where the coefficients $\alpha(\mathbf{k}, J)$ are defined by

$$\alpha(\mathbf{k}, J) := \begin{cases} \binom{\ell}{k_1, S-M, M, k_4, \ldots, k_n} & \text{if } J = 0 \\ \binom{\ell}{k_1, S-M+J, M-J, k_4, \ldots, k_n} - \binom{\ell}{k_1, S-M+J-1, M-J+1, k_4, \ldots, k_n} & \text{otherwise} \end{cases}.$$

For the range of J in the sums above, $\alpha(\mathbf{k}, J)$ is always positive. The innermost sum in (13.7) evaluates to

$$(x_2 x_3)^{S-M+J}(x_2^B - x_3^B)/(x_2 - x_3),$$

where $B = 2M - S - 2J + 1$. This lets us rewrite (13.7) as

$$p^\ell h_1^\ell(\mathbf{x}) = \sum_{k_1,k_4,\ldots,k_n} \sum_{J=0}^{M-\lceil S/2\rceil} \beta(\mathbf{k}, J) x_1^A (x_2^B - x_3^B)/(x_2 - x_3),$$

where $A = k_1 - S + M - J$, and $\beta(\mathbf{k}, J)$ is defined by

$$\beta(\mathbf{k}, J) = \alpha(\mathbf{k}, J)(x_1 x_2 x_3)^{S-M+J} \prod_{j=4}^{n} x_j^{k_j}$$

Noting that $h_2(\mathbf{x}) = h_1(x_2, x_3, x_1, x_4, \ldots, x_n)$ and $h_3(\mathbf{x}) = h_1(x_3, x_1, x_2, x_4, \ldots, x_n)$, and that $\beta(\mathbf{k}, J)$ is symmetric in x_1, x_2, x_3, we can write the left column of the original matrix as

$$\begin{pmatrix} h_1^\ell(\mathbf{x}) \\ h_2^\ell(\mathbf{x}) \\ h_3^\ell(\mathbf{x}) \end{pmatrix} = \sum_{k_1,k_4,\ldots,k_n} \sum_{J=0}^{M-\lceil S/2\rceil} \beta(\mathbf{k}, J) \begin{pmatrix} x_1^A (x_2^B - x_3^B)/(x_2 - x_3) \\ x_2^A (x_3^B - x_1^B)/(x_3 - x_1) \\ x_3^A (x_1^B - x_2^B)/(x_1 - x_2) \end{pmatrix}.$$

This establishes that the original determinant is non-negative. It remains to be checked that, when $\ell > 1$, at least one term satisfying $A > B$ has a nonzero coefficient. The term corresponding to the partition $\mathbf{k} = (\ell, 0, \ldots, 0)$ has $A = \ell, B = 1$ and coefficient $\beta(\mathbf{k}, 0) = 1$, which completes the proof. ■

Corollary 4 *The statement of Lemma 5 remains true if g^ℓ or f^k is substituted for h^ℓ, as long as $\ell, k > 1$ and $a > 0$. For $\ell, k \in \{0, 1\}$, the determinant is zero.*

Proof: As has been observed already, g^ℓ and f^k are convex combinations of $h^0, \ldots, h^k$, where h^0 is taken by convention to equal $\mathbf{u}$. By multilinearity of the determinant, the result follows, with strict inequality from the fact that h^ℓ occurs with nonzero coefficient in g^ℓ, and h^k occurs with nonzero coefficient in f^k when $a > 0$. ■

Corollary 5 *If $\mathbf{x}$ is a fixed point of one of the functions h^ℓ, g^ℓ, f^k, then either $\ell, k \leq 1, a = 0$, or there at most 2 distinct values among $x_1, \ldots, x_n$.*

Proof: If there are at least 3 distinct values among $x_1, \ldots, x_n$, then by relabeling, we may assume $x_1 > x_2 > x_3$. But since $\mathbf{x}$ is a fixed point, the matrix

$$\begin{pmatrix} f_1(\mathbf{x}) & x_1 & 1 \\ f_2(\mathbf{x}) & x_2 & 1 \\ f_3(\mathbf{x}) & x_3 & 1 \end{pmatrix}.$$

has its first two columns identical, and hence has determinant zero, contradicting Corollary 4. ■

Corollary 6 *Fix n, k, a. Then, unless $k = 1$ and $a = 1$, f^k has at most $1 + (k-1)\left(2^{n-1} - 1\right)$ fixed points.*

Proof: When $k \leq 1$ or $a = 0$, the result is obvious. So assume $k > 1$ and $a > 0$. By Corollary 5, all the fixed points lie on the $2^{n-1} - 1$ stable lines corresponding to unordered partitions of $[n]$ into two nonempty subsets. Within each of these lines, the condition of being a fixed point can be expressed by a degree k non-constant polynomial of one variable having a root. Since this can have at most k roots, and $\mathbf{u}$ is a common fixed point on all the stable lines, there can be at most $(k-1)(2^{n-1}-1)$ additional fixed points. ■

13.4.3 Stability

From the previous section, we see that there are only a finite number of fixed points of the dynamical system given by f^k. Our next concern is to understand the stability of these fixed points. Consider a fixed point $\mathbf{x}$ satisfying $\mathbf{x} = f^k(\mathbf{x})$. By Corollary 5, we see that the coordinates ($\mathbf{x}_i$'s) of $\mathbf{x}$ may have at most two distinct values. We may thus classify fixed points into one of three types.

1. **Type A**: $\mathbf{x}$ corresponds to a state where there is a unique dominant language. In other words, $\mathbf{x}$ is such that

$$\mathbf{x}_{\pi(1)} > \mathbf{x}_{\pi(2)} = \mathbf{x}_{\pi(3)} = \ldots = \mathbf{x}_{\pi(n)}$$

   for some permutation π of the set $\{1, 2, \ldots, n\}$ that orders the coordinates in terms of the magnitude of $\mathbf{x}_i$.

2. **Type B**: $\mathbf{x}$ corresponds to a state where there are multiple dominant languages. In other words, for some permutation π and $l > 1$,

$$\mathbf{x}_{\pi(1)} = \ldots = \mathbf{x}_{\pi(l)} > \mathbf{x}_{\pi(l+1)} = \ldots = \mathbf{x}_{\pi(n)}$$

3. **Type C**: $\mathbf{x}$ is the uniform solution given by

$$\mathbf{x} = (\frac{1}{n}, \ldots, \frac{1}{n})'$$

We can now state the following theorem:

Theorem 27 *Fixed points of Type B are always unstable.*

Proof: Consider a fixed point of Type B given by $\mathbf{x}_1 = \mathbf{x}_2 \ldots \mathbf{x}_l > \mathbf{x}_{l+1} = \mathbf{x}_{l+2} \ldots \mathbf{x}_n$. Suppose this is stable. This means there exists a $\delta > 0$ such that for all $\mathbf{y} \in \Delta^{(n-1)} \cap B_\delta(\mathbf{x})$, the sequence given by $\mathbf{y}(t+1) = f(\mathbf{y}(t))$ from the initial condition $\mathbf{y}(0) = \mathbf{y}$ converges to $\mathbf{x}$.

We will now demonstrate the existence of a particular $\mathbf{y} \in \Delta^{(n-1)} \cap B_\delta(\mathbf{x})$ for which the above convergence is not true. Pick

$$\mathbf{y} = \mathbf{x} + (\epsilon, -\frac{\epsilon}{n-1}, \ldots, -\frac{\epsilon}{n-1})'$$

By construction, $\mathbf{y} \in \Delta^{(n-1)}$ and for $\epsilon < \sqrt{\frac{n-1}{n}}\delta$, we see that $\mathbf{y} \in B_\delta(\mathbf{x})$. Note also that

$$\mathbf{y}_1 > \mathbf{y}_2 > \mathbf{y}_n$$

Now by Lemma 6, we have that

$$\lim_{t\to\infty}(\mathbf{y}_2(t) - \mathbf{y}_n(t)) = 0$$

However, since $\mathbf{x}_2 - \mathbf{x}_n > 0$, we see that the sequence $\mathbf{y}(t)$ does not converge to $\mathbf{x}$. ■

Lemma 6 *Let $\mathbf{y} \in \Delta^{n-1}$ be a point where the coordinates are ordered so that (i) $\mathbf{y}_i \geq \mathbf{y}_j$ for all $i > j$ and (ii) $\mathbf{y}_1 > \mathbf{y}_2 > \mathbf{y}_n$. Then the iterative dynamics of f defined by $\mathbf{y}(t+1) = f(\mathbf{y}(t))$ from the initial point $\mathbf{y}$ is such that*

$$\lim_{t\to\infty}(\mathbf{y}_2(t) - \mathbf{y}_n(t)) = 0$$

Proof: Consider the sequence $\{\mathbf{y}, \mathbf{y}(1), \mathbf{y}(2), \dots\}$ where $\mathbf{y}(t) = f(\mathbf{y}(t-1))$. We will study this sequence in the three-dimensional space obtained by applying the coordinate projection $P_3 : \mathbb{R}^n \to \mathbb{R}^3$ given by $\mathbf{x} = P_3\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_n)'$. Thus, we obtain the sequence of 3-tuples $\{\mathbf{x}, \mathbf{x}_1, \dots\}$ where $\mathbf{x}_i = P_3\mathbf{y}(i)$.

In this three-dimensional space, with origin defined by $(0,0,0)'$, let $\mathbf{u} = (1,1,1)'$, $e_1 = (1,0,0)'$ and $e_3 = (0,0,1)'$ respectively. Following usual conventions, we can identify vectors with points in the space and we will do so in our proof. Viewing this space in the direction of $\mathbf{u}$ (so that the point $\mathbf{u}$ is directly above the origin), we get the picture of Fig. 13.2.

Following the picture, consider an arbitrary $\mathbf{x}_t$ and $\mathbf{x}_{t+1} = P_3[f(\mathbf{y}_t)]$. The following observations may be made.

(i) For every t, $\mathbf{x}_t$ and e_3 are on opposite sides of the plane defined by $\mathbf{u}$ and e_1. To see this, simply notice that $|e_1\ \mathbf{x}_t\ \mathbf{u}| \geq 0$ while $|e_1\ e_3\ \mathbf{u}| < 0$ keeping in mind that the equation of the plane is given by $|e_1\ x\ \mathbf{u}| = 0$.

(ii) For every t, e_1 and $\mathbf{x}_t$ are on the same side of the plane defined by $\mathbf{u}$ and e_3. To see this, notice that $|e_3\ e_1\ \mathbf{u}| > 0$ and $|e_3\ \mathbf{x}_t\ \mathbf{u}| \geq 0$.

(iii) For every t, $\mathbf{x}_t$ and e_1 are on opposite sides of the plane defined by $\mathbf{u}$ and $\mathbf{x}_{t+1}$. To see this, notice that $|e_1\ \mathbf{x}_{t+1}\ \mathbf{u}| \geq 0$ while $|\mathbf{x}_t\ \mathbf{x}_{t+1}\ \mathbf{u}| \leq 0$.

This justifies the picture of fig. 13.2 where we view all points from the direction $\mathbf{u}$. Following this picture, we can define for any vector $\mathbf{z} \in \mathbb{R}^3$, the angle that the vector makes with the plane defined by $\mathbf{u}$ and e_1. Let $\theta(\mathbf{z})$ be this angle. Formally, we see that

$$\theta(\mathbf{z}) = \frac{(P_{\mathbf{u}}\mathbf{z}) \cdot (P_{\mathbf{u}}e_1)}{\|(P_{\mathbf{u}}\mathbf{z})\|\|P_{\mathbf{u}}e_1\|}$$

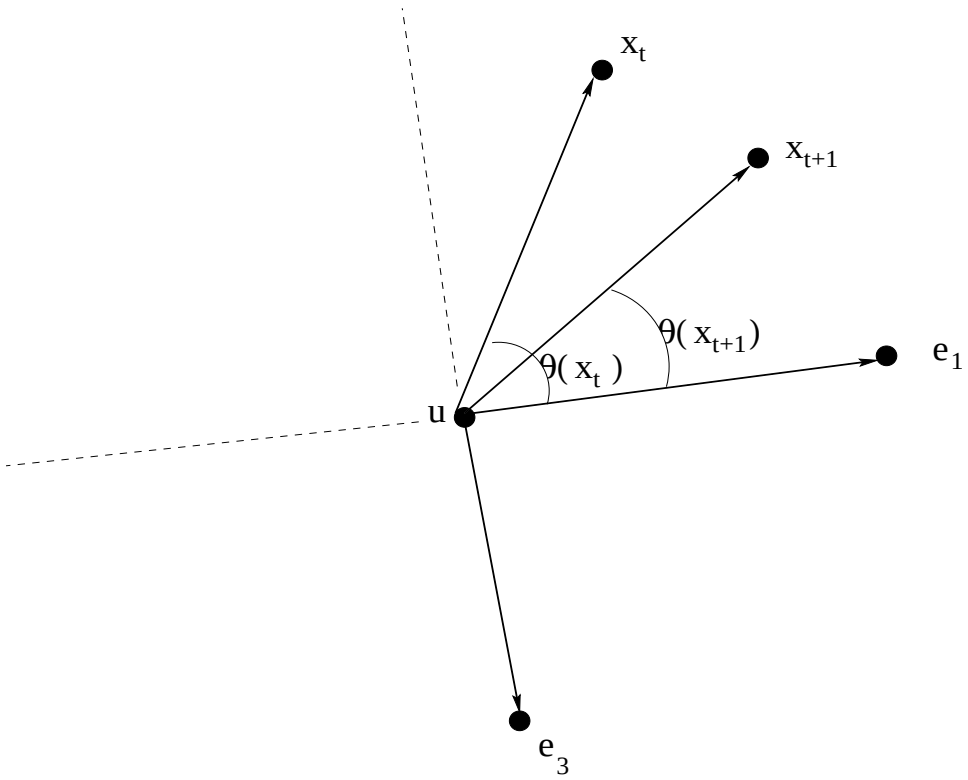

Figure 13.2: A view of the points $\mathbf{x}(t), \mathbf{x}(t+1), \mathbf{u}, e_1, e_3$ in the direction of $\mathbf{u}$. Thus the origin is directly below $\mathbf{u}$ and what is observable in the picture is the projection of each vector onto the plane perpendicular to $\mathbf{u}$.

where $P_{\mathbf{u}}$ denotes the projection onto the two dimensional plane perpendicular to $\mathbf{u}$.

Now consider the set

$$A = \{(z_1, z_2, z_3)' = P_3\mathbf{z} \,|z_1 \geq z_2 \geq z_3; \mathbf{z} \in \Delta^{(n-1)}; 0 \leq \theta((z_1, z_2, z_3)') \leq \theta(-e_3)\}$$

On this set A one can also define the function $\Delta\theta$ to be the change in θ after one step of the dynamics defined by f. In other words, for $P_3\mathbf{z} \in A$

$$\Delta\theta(P_3\mathbf{z}) = \theta(P_3\mathbf{z}) - \theta(P_3 f(\mathbf{z}))$$

Clearly the sequence $\mathbf{x}, \mathbf{x}_1, \ldots$, lies in A. Observations (ii) and (iii) taken together show that $\theta(\mathbf{x}_t)$ is a decreasing function of t. Therefore, we have

$$\forall t \; \theta(\mathbf{x}_t) \geq \theta(\mathbf{x}_{t+1}) \geq 0$$

Therefore $\lim_{t\to\infty} \theta(\mathbf{x}_t)$ must exist. Let this limit be θ_*. Since A is a compact set and $\mathbf{x}_t$ is a sequence in A, we know that it must have a convergent subsequence. Let $\mathbf{x}_{t_n}$ be the convergent subsequence and $\mathbf{x}_* \in A$ be its limit point. Since $\theta(\mathbf{v})$ is a continuous function on A, we have that

$$\theta(\mathbf{x}_*) = \theta_*$$

Since $\Delta\theta(\mathbf{x}_t) = \theta(\mathbf{x}_t) - \theta(\mathbf{x}_{t+1})$, we have

$$\Delta\theta(\mathbf{x}_*) = 0$$

By the latter condition, and Corollary 5, we have that there can be at most two distinct values in $\mathbf{x}_*$. There are therefore three cases:

(a) $\mathbf{x}_*(1) = \mathbf{x}_*(2) = \mathbf{x}_*(3)$

(b) $\mathbf{x}_*(1) > \mathbf{x}_*(2) = \mathbf{x}_*(3)$

(c) $\mathbf{x}_*(1) = \mathbf{x}_*(2) > \mathbf{x}_*(3)$

By observation (i) case (c) cannot be true. Since only case (a) or case (b) is true, therefore, we see that $\mathbf{x}_*$ lies on the plane defined by $\mathbf{u}$ and e_1. Consequently, $\theta_* = \theta(\mathbf{x}_*) = 0$. This, in turn, means that $\lim_{t\to\infty} \mathbf{x}_t(2) - \mathbf{x}_t(3) = 0$ ■

13.4.4 Bifurcations

From the previous sections, we see that there could be at most a finite number of equilibria (fixed points). These have been classified into three types (A,B,C) and the instability of Type B equilibria has already been established. The map $f^k : \Delta^{(n-1)} \to \Delta^{(n-1)}$ is parameterized by a, k and both the number of equilibria and their stability depend upon the values of a and k. We now discuss the bifurcations in the dynamics as a and k change.

Small a, k regime

We first note that for small values of a, k, there is only one fixed point corresponding to the uniform solution $\mathbf{x} = (1/n, \ldots, 1/n)^T$. To see this, consider $a = 0$. Then, $F_i^k = 0$ for all i. Thus, from Eq. 13.6, we have

$$\text{for all } i, \qquad f_i^k = \frac{1}{n}$$

and we see that the population moves to the uniform equilibrium in one time step for all values of k.

It seems that for small values of k, the uniform solution is the only fixed point for all $a \in [0, 1]$. An analytic proof of this with a quantitative characterization of what small k means in this context is missing at the moment. However, let us briefly consider the special cases of $k = 1, 2, 3$ respectively.

For $k = 1$, we see that $F_i^1 = ax_i$ and therefore the dynamical map is given by

$$f_i^1 = ax_i + (1 - a)\frac{1}{n}$$

Solving for $f_i^1(\mathbf{x}) = x_i$, we see that the uniform solution is the only fixed point.

For $k = 2$, we have $F_i^2 = 2ax_i(1 - a) + (ax_i)^2$ and the dynamical map is now given by

$$f_i^2 = 2ax_1(1 - a) + a^2x_i^2 + \left(1 - 2a(1 - a) - a^2\sum_{j=1}^{n} x_j^2\right)\frac{1}{n}$$

The fixed points are given by $f_i^2(\mathbf{x}) = x_i$ and noting that $\left(1 - 2a(1 - a) - a^2\sum_{j=1}^{n} x_j^2\right)\frac{1}{n}$ does not depend upon i, we have that

$$x_i - a^2x_i^2 - 2ax_i(1 - a) = x_j - a^2x_j^2 - 2ax_j(1 - a)$$

for all i, j. From the above equation, we have that either (i) $x_i = x_j$ for all i, j or (ii)$x_i + x_j = \frac{1-2a(1-a)}{a^2}$ for all i, j. It is easy to check that (ii) is not possible so that the uniform solution ($x_i = x_j$ for all i, j) is the only possible solution.

For $k = 3$, we have $F_i^3 = (ax_i)^3 + 3ax_i(1 - a)^2 + 3(ax_i)^2(1 - ax_i)$. Noting that $f_j^3 = F_j^3 + (1 - \sum_{i=1}^{n} F_i^3)\frac{1}{n}$, we have that any fixed point ($f_j^3(\mathbf{x}) = x_j$) must satisfy

$$x_j - F_j^3(\mathbf{x}) = x_i - F_i^3(\mathbf{x})$$

for all i, j. Substituting the expression for F_i^3 in the above equation, we have that for all i, j, the following must be true:

$$x_i - x_j = -2a^3(x_i^3 - x_j^3) + 3a(1-a)^2(x_i - x_j) + 3a^2(x_i^2 - x_j^2)$$

From this we get either (i) $x_i = x_j$ for all i, j or (ii) $3a^2(x_i + x_j) - 2a^3(x_i^2 + x_j^2 + x_i x_j)$ is independent of i, j. Considering case (ii) further, we see that this reduces to either (a) $x_i = x_j$ for all i, j or (b) $3a^2 - 2a^3(x_j + x_k + x_i) = 0$. Clearly $x_j + x_k + x_i \leq 1$ while $\frac{3}{2a} > 1$ so that (b) is impossible. Thus the uniform solution ($x_i = x_j$ for all i, j) remains the only fixed point.

Stability of the Uniform Solution

Let us now consider the uniform solution (Type C) and conduct an analysis of its stability. It is easy to check that the uniform solution $\mathbf{x} = (\frac{1}{n}, \ldots, \frac{1}{n})$ is always a fixed point of the iterated map given by Eq. 13.6 for all values of a, k. This is seen by noting that (by symmetry) $F_i^k(\mathbf{x}) = F_j^k(\mathbf{x})$ for all i, j and then substituting in Eq. 13.6.

To analyze stability for any fixed point $\mathbf{x}_*$ of the map $\mathbf{x}(t+1) = f^k(\mathbf{x})$, we need to check if the map is contracting in all directions along the simplex. Any perturbation of $\mathbf{x}_*$ along the simplex may be given by $\mathbf{x} = \mathbf{y} + \mathbf{x}_*$ where $\mathbf{y} \in R^n$ and $\sum_{i=1}^n y_i = 0$.

Following standard linear stability analysis, we note that $f^k(\mathbf{x}) - f^k(\mathbf{x}_*) \approx \mathbf{J}\mathbf{y}$ where $\mathbf{J}$ is the Jacobian (of f^k) evaluated at $\mathbf{x}_*$. Therefore the stability of $\mathbf{x}_*$ is determined by the value of S where

$$S = \max_{\mathbf{y}^T\mathbf{1}=0} \frac{||\mathbf{J}\mathbf{y}||}{||\mathbf{y}||}$$

Since $f^k : \Delta^{n-1} \to \Delta^{n-1}$ depends upon both a and k (in addition to n), we see that the value of S depends upon n, a, and k. Therefore, we will denote this dependence explicitly by writing $S(n, a, k)$. To evaluate $S(n, a, k)$ for the uniform solution, let us consider the structure of the Jacobian $\mathbf{J}$. Noting that $f_i^k = F_i^k + \frac{1}{n}(1 - \sum_{j=1}^n F_j^k)$, we see that

$$J_{ij} = \frac{\partial f_i^k}{\partial x_j} = \frac{n-1}{n}\frac{\partial F_i^k}{x_j} - \frac{1}{n}\sum_{l \neq i}\frac{\partial F_l^k}{x_j}$$

One may check that by symmetry, we have $\frac{\partial F_i^k}{x_j} = \frac{\partial F_m^k}{x_l}$ for all distinct i, j, m, l when evaluated at the uniform point $\mathbf{x}_* = (1/n, \ldots, 1/n)^T$. Further, for all

i, j, we have $\frac{\partial F_j^k}{x_j} = \frac{\partial F_i^k}{x_i}$. Letting (for all distinct i, j)

$$\frac{\partial F_i^k}{x_i} = A; \frac{\partial F_i^k}{x_j} = B$$

we see that

$$\mathbf{J}_{ii} = \frac{n-1}{n}(A-B); \mathbf{J}_{ij} = \frac{1}{n}(B-A)$$

Because of the special symmetric structure of $\mathbf{J}$, it is easy to check that its eigenvalues and eigenvectors are as follows: the smallest eigenvalue is $\lambda = 0$ and the corresponding eigenvector is $\mathbf{1}$ (the all one's vector). The next eigenvalue is $A - B$ (multiplicity $n-1$) and the eigenspace spanned by the eigenvectors is the $n-1$ dimensional subspace orthogonal to $\mathbf{1}$. Therefore, by a familiar Raleigh-Ritz argument, we see that

$$S(n, a, k) = A - B = \frac{\partial F_i^k}{x_i} - \frac{\partial F_i^k}{x_j}$$

Now note that

$$\frac{\partial F_1^k}{\partial x_1}|_{\mathbf{x}_*} = a \sum_{\mathbf{k} \in I} k_1 \binom{k}{\mathbf{k}} (1-a)^{k_{n+1}} (\frac{a}{n})^{k-k_{n+1}-1} \tag{13.8}$$

and

$$\frac{\partial F_1^k}{\partial x_2}|_{\mathbf{x}_*} = a \sum_{\mathbf{k} \in I} k_2 \binom{k}{\mathbf{k}} (1-a)^{k_{n+1}} (\frac{a}{n})^{k-k_{n+1}-1} \tag{13.9}$$

where $\mathbf{k} = (k_1, k_2, \ldots, k_{n+1})$ refers to an ordered partition of k into $n+1$ positive integers as usual. The set I is given by the following:

$$I = \{(k_1, k_2, \ldots, k_n, k_{n+1}) | k_1 > k_2, \ldots, k_n; \sum_{i=1}^{n+1} k_i = 1\}$$

Therefore, we have

$$S(n, a, k) = a \sum_{\mathbf{k} \in I} \binom{k}{\mathbf{k}} (1-a)^{k_{n+1}} (\frac{a}{n})^{k-k_{n+1}-1} \tag{13.10}$$

The quantity $S(n, a, k)$ determines the stability of the uniform solution of the map f^k for the the parameter value a. We see that $S(n, a = 0, k) = 0$

for all n, k suggesting that the uniform solution is stable at $a = 0$. Now let us consider the value of $S(n, a, k)$ at $a = 1$. This is given by

$$S(n, a = 1, k) = \sum_{(k_1,\ldots,k_n,k_{n+1}=0)\in I} \binom{k}{k_1 k_2 \ldots k_n 0} (\frac{1}{n})^{k-1}$$

We now prove the following theorem:

Theorem 28 *The uniform solution ($a = 1$) becomes unstable for large k. In particular,*

$$\lim_{k\to\infty} S(n, a = 1, k) = \infty$$

Proof: We prove by induction. The base case corresponds to $S(n = 2, a = 1, k)$ and we have already proved that $\lim_{k\to\infty} S(n = 2, a = 1, k) = \infty$. (Proposition 5.)

The induction step is as follows. Suppose $\lim_{k\to\infty} S(n, a = 1, k) = \infty$. We see the following identity.

$$S(n+1, a = 1, k) = \sum_{k_1>k_2,\ldots,k_{n+1}} (k_1 - k_2) \binom{k}{k_1 \ldots k_{n+1}} (\frac{1}{n+1})^k$$

$$= \sum_{k_{n+1}=0}^{k/(n+1)} \sum_{k_1>k_2,\ldots,k_n} (k_1 - k_2) \binom{k}{k_1 \ldots k_{n+1}} (\frac{1}{n+1})^k$$

$$\geq \sum_{k_{n+1}=0}^{k/(n+1)} \sum_{k_1>k_2,\ldots,k_n} (k_1 - k_2) \binom{k}{k_{n+1}} \binom{k - k_{n+1}}{k_1 \ldots k_n} (\frac{1}{n+1})^k$$

$$= \sum_{k_{n+1}=0}^{k/(n+1)} \binom{k}{k_{n+1}} (\frac{1}{n+1})^{k_{n+1}} (\frac{n}{n+1})^{k-k_{n+1}} \sum_{k_1>k_2,\ldots,k_n} (k_1 - k_2) \binom{k - k_{n+1}}{k_1 \ldots k_n} (\frac{1}{n})^{k-k_{n+1}}$$

$$= \sum_{k_{n+1}=0}^{k/(n+1)} \binom{k}{k_{n+1}} (\frac{1}{n+1})^{k_{n+1}} (\frac{n}{n+1})^{k-k_{n+1}} S(n, a = 1, k - k_{n+1})$$

Since (i) $\lim_{l\to\infty} S(n, a = 1, l) = \infty$, and (ii) $k - k_{n+1} \geq \frac{n}{n+1}k$, it follows that $\lim_{k\to\infty} S(n, a = 1, k - k_{n+1}) = \infty$ for each k_{n+1} in the the above expression. Further, since by the Central Limit Theorem, we know that

$\lim_{k\to\infty}\sum_{k_{n+1}=0}^{k/(n+1)}\binom{k}{k_{n+1}}(\frac{1}{n+1})^{k_{n+1}}(\frac{n}{n+1})^{k-k_{n+1}} = \phi(\frac{1}{2}) = \frac{1}{2}$ (where ϕ is the cumulative distribution of the unit normal), we see that

$$\lim_{k\to\infty}\sum_{k_{n+1}=0}^{k/(n+1)}\binom{k}{k_{n+1}}(\frac{1}{n+1})^{k_{n+1}}(\frac{n}{n+1})^{k-k_{n+1}}S(n,a=1,k-k_{n+1})=\infty$$

The theorem is proved. ■

The Emergence of Coherence

We can now piece together a picture of the bifurcations by which shared languages emerge as a result of the dynamics of language evolution in our setting. From the previous section, we see that $\lim_{k\to\infty} S(n,a=1,k)=\infty$. We now make the following observations. Consider any sufficiently large k such that $S(n,a=1,k)>1$. For such a k, we have a bifurcation as a changes continuously from $a=0$ to $a=1$. Specifically,

1. For $a=0$, the only fixed point of the map f^k is the uniform solution and this is stable. Thus, from all initial conditions, the population moves to a situation where all possible languages are equally represented in the population.

2. Since $S(n,a=0,k)=0$ while $S(n,a=1,k)>1$, by continuity, we see that there exists a bifurcation point a_* where $S(n,a_*,k)=1$ and when $a>a_*$, the uniform solution becomes unstable.

3. The bifurcation point $a=a_*$ when the uniform solution becomes unstable corresponds also to the point where new one grammar solutions emerge. To see this fact, consider a critical line segment on the simplex $\Delta^{(n-1)}$ that joins the uniform point $\mathbf{u}=(\frac{1}{n},\ldots,\frac{1}{n})^T\in\Delta^{(n-1)}$ to one of the corners $\mathbf{e}_1=(1,0,\ldots,0)^T\in\Delta^{(n-1)}$. This line may be parameterized by the real number $\alpha\in[0,1]$ and any point $\mathbf{x}\in\Delta^{(n-1)}$ on this line may be represented as $\mathbf{x}=\alpha\mathbf{e}_1+(1-\alpha)\mathbf{u}$. Now consider the map f^k restricted to this critical line. This restriction defines a map from $g:[0,1]\to[0,1]$ as follows. For any $\mathbf{x}=\alpha\mathbf{e}_1+(1-\alpha)\mathbf{u}$, consider $\mathbf{y}=f^k(\mathbf{x})$. It is easy to check that $\mathbf{y}$ lies on the critical line segment so that it can be represented as $\mathbf{y}=\alpha_y\mathbf{u}+(1-\alpha_y)\mathbf{e}_1$. Now we set $g(\alpha)=\alpha_y$. It is easy to check that $g(0)=0$, $g(1)<1$ and $g'(0)>1$. Therefore there exists a point $\alpha_*\in[0,1]$ such that $g(\alpha_*)=\alpha_*$. This corresponds to a point $x_*=\alpha_*\mathbf{u}+(1-\alpha_*)\mathbf{e}_1$ that is a fixed point

of f^k. Further, since $\alpha_* > 0$, it is easy to check that $\mathbf{x}_*$ is such that $\mathbf{x}_*(1) > \frac{1}{n}$. In other words, there is a dominant language in the community. It is also easy to check by standard arguments that this fixed point is stable.

4. It is worth noting that we see the emergence of coherence for large a, k *without* any notion of differential fitness and natural selection. What brings about coherence instead is the fact that all language learners are immersed in the same population and learn from the same source distribution which is a mixture of the languages of the previous generation. When k is large enough, the population eventually settles on a common language. The transition from the *Tower of Babel* mode with a uniform solution to a *one grammar* mode with a shared (majority) language occurs via a bifurcation in the dynamics as we see.

13.5 Coherence for a Memoryless Learner

So far in this chapter, we have spent some time analyzing the evolutionary dynamics of a *batch* learner. We were able to show that coherence would emerge in a population of such learners via a bifurcation. The other important class of learning algorithms we have considered in this book are the *memoryless* learning algorithms. Interestingly, for a prototypical memoryless learning algorithm, we see that coherence *never* emerges.

Formally, the memoryless algorithm is as follows:

1. **Initialize** Choose a language uniformly at random.

2. **Iterate** At time step n, receive a new example sentence. If this sentence is a cue for L_i, then change hypothesis language to L_i. Otherwise, retain current hypothesis language.

It is easy to check that this algorithm satisfies the learnability requirement. By our familiar Markov analysis, the behavior of the learner can be understood as a Markov chain with n states and a transition matrix T whose (i, j) term is given by

$$T_{ij} = \begin{cases} ax_j & \text{if } j \neq i \\ (1-a) + ax_i & \text{otherwise} \end{cases}$$

We see that $T = (1-a)I + a\mathbf{1}\mathbf{x}^T$ where I is the $n \times n$ identity matrix and $\mathbf{x} \in \Delta^{n-1}$ is the state of the population in generation t. The probability

with which learners will settle on the different languages after k examples are drawn is given by our familiar analysis resulting in the following dynamics

$$(\mathbf{x}(t+1))^T = (\frac{1}{n}, \ldots, \frac{1}{n})(T^k)$$

Because of the structure of T, we see that

$$T^k = \left((1-a)I + a\mathbf{1}\mathbf{x}^T\right)^k = (1-a)^k I + (1-(1-a)^k)\mathbf{1}\mathbf{x}^T$$

Using Eq. 13.11, we see that

$$\mathbf{x}(t+1) = (1-a)^k \mathbf{u} + \left(1-(1-a)^k\right)\mathbf{x}(t) \tag{13.11}$$

where $\mathbf{u} = (\frac{1}{n}, \ldots, \frac{1}{n})^T$ is the uniform point. It is straightforward to check that this simple linear dynamics results in the population moving to the uniform solution $\mathbf{u}$ from all initial conditions for all values of $a \in [0, 1)$. For $a = 1$, the dynamics is given by the identity map so that no change is possible. The initial distribution of languages in the population is preserved for all time.

13.6 Learning in Connected Societies: Analogies to Statistical Physics

Our model in this chapter with its bifurcations from incoherent to coherent states is reminiscent of models of phase transitions in Statistical Physics. We have touched upon this analogy in the past and we shall now try to explore this a little more concretely here. In the setting for language evolution considered in this chapter, there are n possible linguistic types. We analyzed the behavior of a population of linguistic agents where each agent belongs to one of these types. Learning by individual agents defined a natural dynamical evolution of the population and we characterized the conditions under which coherent (one-language) modes emerged.

In Statistical Physics, one considers a physical system made up of an ensemble of interacting components (particles). The degree of interaction between these particles is typically governed by temperature in a manner such that the interactions decrease as the temperature increases. One then analyzes some macroscopic property of the system as a function of temperature. In many such systems, one finds a phase transition between two

different regimes of behavior separated by a critical temperature T_c. Examples include the transitions between different states of matter (solid, liquid, gas) or the transitions associated with the loss of permanent magnetization at high temperatures.

Our goals in exploring this analogy are two-fold. First, by developing these connections to Statistical Physics, we hope that intuitions and insights developed in each discipline may be transferred usefully across disciplinary lines. Second, we note that much of our book has focused on the case where agents learn from everybody, i.e., the population is perfectly mixed. This allows for the possibility of long-range interactions between agents and a central focus of this book is on the bifurcations that arise in this setting. We have also considered the case where agents learn from only one person, i.e., the population is largely disconnected and we have noted the lack of bifurcations in this situation. An important question for us is to understand what happens when there are intermediate (local) correlations between agents. One way to formulate these local interactions is by developing a graph structure where vertices are identified with linguistic agents and the edges denote lines of communication along which examples are provided to learning agents situated at these vertices. Once this formulation is adopted, synergies with Statistical Physics become more apparent and consequently worth exploring.

In the next section, we formulate language evolution on a locally connected graph. Following that, we discuss the classical Ising model of spins associated with a graph. Finally, we consider analogies, implications, and directions of future research in Sec. 13.6.3. Our development focuses on the special case in which $n = 2$.

13.6.1 Language Evolution in Locally Connected Societies

Imagine a spatial connectivity pattern in terms of a graph that has the structure of a square lattice ($\sqrt{N} \times \sqrt{N}$ lattice with N vertices in all) as shown in Fig. 13.3. Each site (vertex) is associated with a random variable $X_{i,j}(t) \in \{0, 1\}$. The value of $X_{i,j}(t)$ may be identified with the language of the linguistic agent occupying the location (i, j) at time t in the graph (lattice). Now $X_{i,j}(t+1)$ is the language at that same location (i, j) at the next time step. In our setting, $X_{i,j}(t+1)$ is a random variable such that

$$\mathbb{P}[X_{i,j}(t+1) = 1] = g(a, \mu_{i,j}, k)$$

where

$$\mu_{i,j} = \frac{1}{4}(X_{i+1,j}(t) + X_{i,j+1}(t) + X_{i-1,j}(t) + X_{i,j-1}(t))$$

This corresponds to the following protocol: the language at each location (i, j) is updated by a learning procedure where the learner obtains example sentences at random from the neighboring agents and uses a cue based learning algorithm to determine its language. The function g is obtained by the usual analysis of the previous sections. As before, a is the probability with which speakers of each language produce cues and k is the number of examples each learning agent hears during the learning period. An example of the function g is given in Eq. 13.4.

One may now study the evolution of the following object

$$\alpha_N(t) = \frac{1}{N} \sum_{i,j} X_{i,j}(t)$$

The quantity $\alpha_N(t)$ corresponds to the average number of L_1 speakers at time t.

Remark. If the connectivity pattern of the graph is such that every location is connected to every other location, then all the $X_{i,j}(t)$'s would be identically distributed. Because of the local connectivity described here, the $X_{i,j}(t)$'s are no longer identically distributed. For complete graphs, we believe that $\alpha(t) = \lim_{N\to\infty} \alpha_N(t)$ is well defined and the following is true

$$\alpha(t+1) = \lim_{N\to\infty} \alpha_N(t) = g(a, \alpha(t), k)$$

Thus the bifurcation described in earlier sections of this chapter is recovered as the limiting behavior of large, complete graphs where the graph size goes to infinity. We conjecture that even for locally connected graphs of the sort described in this section, one would observe bifurcations as the graph size was appropriately increased to infinity. The reason for our conjecture rests on analogies to similar phenomena in statistical physics as we describe shortly.

13.6.2 Magnetic Systems: The Ising Model

Now consider the well studied two dimensional Ising model of statistical physics. We discuss this model in the context of phase transitions in magnetic systems. In this setting, each site on the square lattice of Fig. 13.3 is associated with a particle having a *spin* that takes one of two values: *up* or *down*. A state of the magnetic system is denoted by a configuration of spins $\bar{s} \in \{-1, +1\}^N$ where $\bar{s}_{i,j}$ denotes the spin at site (i, j). Given a magnetic

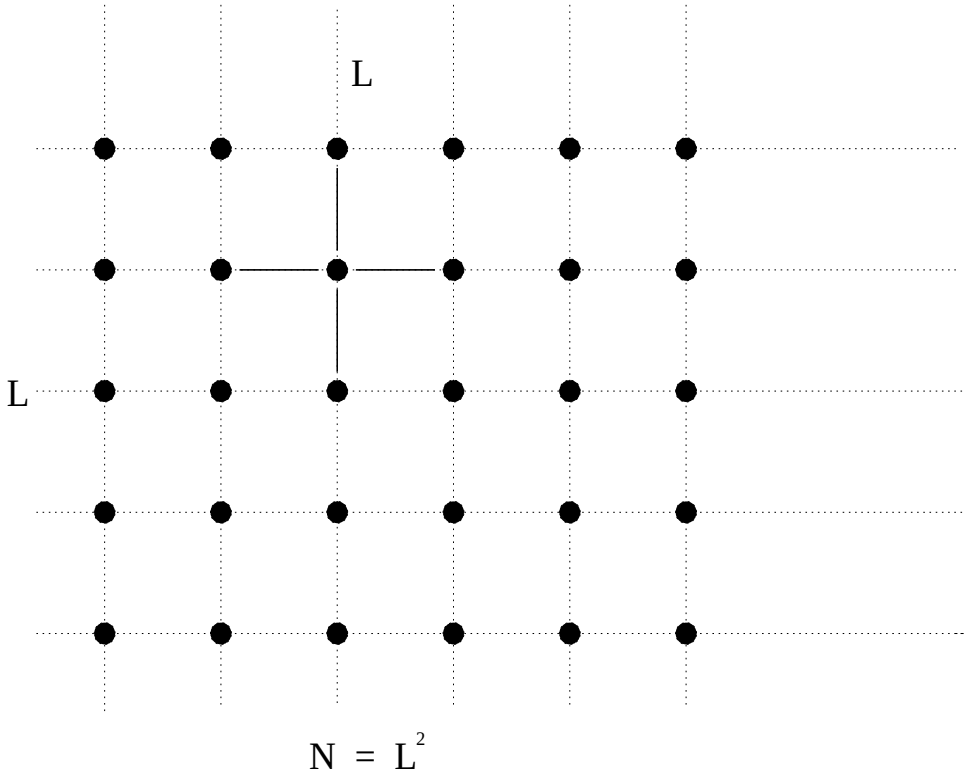

Figure 13.3: A two dimensional lattice with N vertices in all. Each vertex is identified with a linguistic agent with a particular language in the context of language evolution and with a particle with a spin in the context of statistical physics. Each vertex is connected to its immediate neighbors as shown.

field B and a correlation strength J, one may write down the Hamiltonian (energy) of a configuration of spins at each site as

$$H(\bar{s}) = -J \sum_{(i,j)\sim(k,l)} \bar{s}_{i,j}\bar{s}_{k,l} - B\sum_{i,j} \bar{s}_{i,j}$$

In the term $\sum_{(i,j)\sim(k,l)} \bar{s}_{i,j}\bar{s}_{k,l}$, the sum is taken over all neighboring sites and measures the degree of alignment of neighboring spins. The term $B\sum_{i,j} \bar{s}_{i,j}$ measures the alignment of the spins with the direction of the magnetic field. Thus each spin configuration $\bar{s} \in \{-1,+1\}^N$ has an energy value associated with it. The Hamiltonian, in turn, defines an equilibrium probability distribution on the set of all possible spin configurations as

$$\mathbb{P}[\bar{s}] = \frac{e^{-\beta H(\bar{s})}}{Z_N} \tag{13.12}$$

where $Z_N = \sum_{\bar{s}} e^{-\beta H(\bar{s})}$ is the normalizing factor referred to as the partition function. The quantity β is equal to $\frac{1}{kT}$ where T is the temperature and k is Boltzmann's constant. This characterizes the probability with which the particle system will take on different spin configurations at thermal equilibrium.

A parameter of interest is the average spin (magnetization) in the absence of an external magnetic field. It is easy to check that the average magnetization is given by

$$M_N(T) = \frac{1}{N}\sum_{\bar{s}}\sum_{i,j} \bar{s}_{ij}\mathbb{P}[\bar{s}] = \frac{kT}{N}\left(\frac{\partial F}{\partial B}\right)_{B=0}$$

where $F = \ln(Z_N)$ is the Free energy. An important limiting condition is the so called *thermodynamic limit* where one lets $N \to \infty$, i.e., the lattice (graph) size grows to infinity, and in this limiting case, one may define $m(T) = \lim_{N\to\infty} M_N(T)$. Thus $m(T)$ may be interpreted as the average magnetization in a system with an infinite number of particles.

An analysis of this two dimensional Ising model reveals the following phase transition phenomenon. There exists a *critical temperature* T_C such that for all $T > T_C$, the equilibrium state of the material is a state of demagnetization, i.e., $m(T) = 0$, or there are as many *up* spins as *down* spins on average. For $T < T_C$, it turns out that $m(T)$ is non-zero implying that spontaneous magnetization occurs and the material may therefore be in a state of permanent magnetization at that temperature.

Remark 1: A number of technical issues relate to how the thermodynamic limit $N \to \infty$ is taken and the existence of well defined probability measures on the configuration space in this limit. These equilibrium measures correspond to *Gibbs measures* and the phase transitions correspond to whether there is a single Gibbs measure or multiple Gibbs measures in the thermodynamic limit. To provide some intuition, it is worth noting that because of the particular exponential form of the probability distribution in Eq. 13.12, we see that at high temperatures (large T), the value of β is close to 0 and the distribution is relatively flat. Thus all configurations are more or less equally likely, the effect of interactions is less, individual particle spins behave almost at random, and the average magnetization is close (equal) to zero. At $T = 0$, when $\beta = \infty$, the system will settle into a minimum energy configuration where all spins will align. A phase transition may be expected between these two regimes but a rigorous proof is beyond the scope of the current book.

Remark 2: What has been discussed above is a static theory of magnetization that describes the probability distribution of spin configurations at thermal equilibrium at different temperatures. A number of dynamic processes on associated graphs (lattices) have been proposed that correspond to various Markov Chain Monte Carlo methods for sampling from this equilibrium distribution. For example, single spin flip Glauber dynamics is a procedure in which a site is picked at random and using a local rule, the spin at that site remains the same or flips with a probability that is based on the configuration of its neighbors and the strength of interaction. In ferromagnetic materials, spins tend to align with their neighbors.

Remark 3: The formulation and initial development of the theory of Ising systems may be traced back to the first half of the twentieth century as attempts were made to construct theoretical models for the existence of phase transitions in magnetic systems. More recent developments over the last thirty years have elaborated many aspects of such systems. Researchers have studied the behavior of the order parameter (average magnetization) near the critical point leading to insights about power law characteristics and universality. For example, using techniques from the theory of renormalization groups, many now believe that the phenomena of phase transitions and behavior near the critical point depend on some global properties (e.g. the dimension of the lattice — one may consider higher dimensional lattices in general rather than the two dimensional case we described here) rather than the details of the particular interaction. A large number of local interaction behaviors may thus give rise to the same global patterns. For a treatment

of this subject, see Kadanoff (2000).

13.6.3 Analogies and Implications

We can now draw analogies between our model of language evolution on a lattice and the two-dimensional Ising model of statistical physics. Linguistic agents are like particles. Spins are like languages. Each linguistic agent takes on one of two languages while each particle has one of two spins. The linguistic composition of the population (what percentage speaks L_1) is like the average magnetization of the spin system. In the language evolution case, the quantity a behaves like temperature. When $a = 0$, each agent behaves independently of the others and the only equilibrium to be expected is the uniform mode in which both languages are equally represented in the population. This is like the behavior of the magnetic system at very high temperatures when spins are largely uncorrelated and the average magnetization is zero. We conjecture therefore that one ought to observe a phase transition as a increases from 0 to 1 for the specific language evolution model with local interactions described in Sec. 13.6.1. We leave this for future work.

A number of implications of the analogy are worth noting. First, we observe that the models in the first part of this chapter are formulated in the more general case with n linguistic types (languages). Our description in this section, however, has been in terms of the special case when $n = 2$. The corresponding generalization of Ising models with n-valued spins is known as the Potts model and similar issues may be studied in that context.

Second, we believe that our analysis in this chapter (working in the continuum limit of infinitely many linguistic agents) corresponds closely to the situation in which every agent is connected to every other, i.e., the graph of interactions is not the locally connected lattice of Fig. 13.3 but rather is a complete graph where each vertex is connected to all other vertices. The solution of Ising systems with a complete graph correspond rather directly to *mean field theories* where one assumes that the effective field on every particle is the same. Mean field approximations are widely used in statistical physics and have been found to be useful. One may regard many of the models of this book as mean field approximations to more realistic models of language evolution. It is worth noting, however, that our approach in this book has not followed the formalisms or style of statistical mechanical systems. Rather, by taking the continuum limit, we arrive immediately at non-linear dynamical systems and the bifurcation theory of iterated maps plays the central role in our development.

Third, following Remark 3 of the previous section, we believe that the qualitative insights of our "mean field" analysis will translate into other models of language evolution with local (rather than global) interactions as well. In fact, one of the central insights of this book is that bifurcations may arise in language evolution as a result of the interactions of individual language learners. Following Remark 3 (previous section), we believe that this insight will be robust to details of the particular interactions.

Finally, it is worth noting that although there are many analogies to spin systems, there are also some differences. In particular, unlike the case in statistical physics, for models of language evolution, there is no *static* theory. There is no quantity analogous to the Hamiltonian that may describe an equilibrium distribution from first principles. Language evolution is fundamentally a *dynamic* theory where the dynamics is derived from the behavior of individual learning. Second, the details of the model are sufficiently different that a rigorous proof of the behavior of language evolution systems in the thermodynamic limit is not obvious at the present moment.

We thus see that our approach to language evolution may be characterized in terms of constructs that are similar to statistical physics systems. A general setting for language evolution consists of the following ingredients:

1. A graph $G = (V, E)$ where V is the vertex set and E is the edge set represents the linguistic connectivity pattern of the society. Each vertex $v \in V$ may be identified as the *site* of a linguistic agent. The connectivity pattern may reflect social, geographical, economic, or other factors.

2. Let $\mathcal{L}$ be a collection of possible *languages*. These may be identified with sets of expressions, grammars, alternative pronunciations of words, etc. We have seen examples of these over the course of this book. In our current chapter, $\mathcal{L} = \{L_1, \ldots, L_n\}$ where each $L_i \subset \Sigma^*$.

3. A linguistic configuration of the society at time t is given by $x_v(t) \in \mathcal{L}$ for each $v \in V$.

4. A learning algorithm $\mathcal{A}$ describes how the linguistic configuration evolves as each agent is exposed to examples from its neighbors and updates its language. Thus, this defines $x_v(t+1)$ for each $v \in V$.

Depending upon precise choices for each of the four above, one may obtain many versions of the language evolution problem. Some of these

problems may be amenable to study by the techniques developed in probability theory (e.g. percolation processes) and statistical physics. Others may require new techniques. It is our hope that knowledge developed about phase transitions and critical phenomena may be usefully applied in this linguistic setting. The peculiarities of the language evolution problem may also bring fresh perspectives and new problems for theorists interested in complex systems of simple interacting agents. For an approach to language acquisition and evolution based on ideas from statistical physics, see the work of Antonio Galves and colleagues (Cassandro et al, 1999).

13.7 Conclusions

In this chapter, we re-examined the conditions under which linguistic coherence might emerge in a population of linguistic agents. In particular, we considered a setting where individual children learn from the entire adult population at large rather than a single teacher (parent) alone.

We concentrated on the case in which there were n possible languages (linguistic types) that were symmetrically arranged in that all languages were equally easy to learn. Languages were learned on the basis of cues provided by speakers to child learners. The learning algorithm was a batch learner that counted cues and hypothesized a language that was consistent with most cues. Under these assumptions we derived the evolutionary dynamics of the population. This was seen to be a map $f^k : \Delta^{n-1} \to \Delta^{n-1}$ and we showed that this could have only a finite number of fixed points (equilibria). More interestingly, there were two regimes for this map. In the small a, k regime, only the uniform fixed point was stable and populations converged to this from all initial conditions. Thus no shared language emerged. As a, k changed values, a bifurcation led to the loss of stability of the uniform solution and new stable one language modes arose. Shared languages emerged.

Thus, yet again, we see the role of bifurcations in the dynamics of language evolution. The values of a and k may be related to the learning fidelity of the learner. Thus small a, k corresponds to a regime in which there is too little information on the basis of which learners make their linguistic decisions. The system is noisy and shared languages do not emerge. Large a, k corresponds to a regime in which a lot of information is provided to individual learners, learning fidelity is high, the system is less noisy, and shared languages emerge. The small a, k regime may be compared to high thermal

noise in statistical physics systems or high mutation rates in biological evolutionary systems. Correspondingly, the large a, k regime may be compared to low thermal noise in statistical physics or low mutation rates in biological evolution and the bifurcations in language evolution may be qualitatively compared to those that are seen in physics and biology respectively.

It is also worthwhile for us to reflect on the difference between learning from one teacher and learning from the population. Our analysis suggests that when learning from one teacher, differential fitness and natural selection play an important role for the emergence of coherent states corresponding to shared languages in the population. In our particular models, we find that without natural selection, coherence will never emerge, no matter how good the learning ability of individual learners. In contrast, in social learning, as in the model of this chapter, coherence emerges without any recourse to natural selection. This insight must be kept in mind when constructing explanatory paradigms for the evolution of communication systems in the animal kingdom. Thus, we argue that in those evolutionary scenarios (such as in some species of song-birds) when learning is in the nesting phase and primarily from one teacher, natural selection must play an important role in the evolution of shared systems. On the other hand, in social settings such as human linguistic communication, one need not invoke notions of natural selection for language evolution.

While shared communication systems were seen to emerge for the batch learner in a social learning situation, a similar situation was not true for a memoryless learner. In fact, the population dynamics resulting from the individual behavior of the memoryless learner converged to a uniform solution for all values of k. Thus no matter how good the learning ability of the learner, shared languages do not emerge. This suggests that the emergence of shared communication systems is not implied simply by high learning ability. Rather, the details of the learning algorithm and the influence of natural selection interact in a subtle way to bring this about.