Models of Language Evolution: Part III

Matilde Marcolli

MAT1509HS: Mathematical and Computational Linguistics

University of Toronto, Winter 2019, T 4-6 and W 4, BA6180

Main Reference

- Partha Niyogi, *The computational nature of language learning and evolution*, MIT Press, 2006.

Language Evolution and Fitness

- Language evolution modeled on ideas from biological evolution
  - a reproductive process: learning algorithm (individuals of new generation produce language using linguistic input data from previous generation)
  - transmission errors: multilingual models and population dynamics
  - fitness test?
- in biological evolution reproductive fitness drives evolution by natural selection
- Problem: is there a good fitness function in linguistics?
- develop a notion of communicative efficiency

Origins of Language

- how did language arise, from our pre-human ancestors?
- like everything else in the biological world: by Darwinian evolution by natural selection
- evolutionary advantage in terms of reproductive fitness
- communicative efficiency provides biological fitness
- related idea: *coherence* (homogeneous linguistic population) is an *emergent phenomenon* resulting from the behavior of interacting individual linguistic agents
- passage to a coherent state resulting from a bifurcation in the dynamics

Mutual Intelligibility

- $\mathcal{L}_1$ and $\mathcal{L}_2$ two languages: want to define a communicative fitness or mutual intelligibility function $F(\mathcal{L}_1, \mathcal{L}_2)$
- how to improve (or maximize) $F(\mathcal{L}_1, \mathcal{L}_2)$ depends on *ambiguities* in the languages and on learning algorithms
- want an algorithm that identifies $\arg\max_{\mathcal{L}'} F(\mathcal{L}, \mathcal{L}')$ or approximates it arbitrarily well
- if $\mathcal{L}$ has ambiguities $\arg\max_{\mathcal{L}'} F(\mathcal{L}, \mathcal{L}')$ need not be $\mathcal{L}' = \mathcal{L}$

Language as Probabilistic Association between Form and Meaning

- language as an association matrix linking referents to signals
- M referents and N signals: $A = (A_{ij})$ an $M \times N$-matrix, values of entries A_{ij} = strength of the association
- this matrix drives both *production* and *comprehension* (assigning signals to meanings and meanings to signals)
- also used as framework to model "communication in the animal and the machine"
- set of all possible signals is the set of all strings $\mathfrak{A}^\star$; set of all possible meanings is set of strings over a *semantic alphabet*: infinite matrices

- $\mathcal{S}$ = set of signals; $\mathcal{M}$ = set of meanings (finite or countable)
- assume $\mathcal{S} = \mathfrak{A}_1^\star$ and $\mathcal{M} = \mathfrak{A}_2^\star$ (linguistic and semantic alphabets)
- Communication System: probability measure μ on $\mathcal{S} \times \mathcal{M}$
- generalizes the (normalized) "association matrix" to infinite case
- Encoding Matrix (production): $s_i \in \mathcal{S}$ and $m_j \in \mathcal{M}$

$$P_{ij} = \mu(s_i \,|\, m_j) = \frac{\mu(s_i, m_j)}{\sum_k \mu(s_k, m_j)}$$

set equal zero if denominator sum is zero

- Decoding Matrix (comprehension):

$$Q_{ij} = \mu(m_i \,|\, s_j) = \frac{\mu(s_j, m_i)}{\sum_k \mu(s_j, m_k)}$$

and zero if denominator sum is zero

• in other models P and Q not required to come from the same measure μ, but this is a way to ensure *consistency* between active and passive usage of a language

• Note that μ determines P and Q, but these matrices don't determine μ: can consider equivalence classes of measures μ that determine same P, Q

• Useful signals: (actually used in production or comprehension)

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

• this should correspond in formal language theory to the set of well formed (grammatical) sentenced

• Expressible meanings: (can be expressed within the language)

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

Communication

- two systems μ_1 and μ_2
- probability of a meaning being successfully communicated from μ_1 to μ_2 or from μ_2 to μ_1 (with σ distribution on $\mathcal{M}$)

$$\mathbb{P}(1 \to 2) = \sum_i \sigma(m_i) \sum_j \mu_1(s_j \mid m_i)\ \mu_2(m_i \mid s_j)$$

$$\mathbb{P}(2 \to 1) = \sum_i \sigma(m_i) \sum_j \mu_2(s_j \mid m_i)\ \mu_1(m_i \mid s_j)$$

- Communicability $F(\mu_1, \mu_2) = \frac{1}{2}(\mathbb{P}(1 \to 2) + \mathbb{P}(2 \to 1))$

$$F(\mu_1, \mu_2) = \frac{1}{2}(\mathrm{Tr}(P^{(1)} \Lambda\ {}^{\tau}Q^{(2)}) + \mathrm{Tr}(P^{(2)} \Lambda\ {}^{\tau}Q^{(1)}))$$

$\Lambda =$ diagonal matrix entries $\Lambda_{ii} = \sigma(m_i)$

• $F(\mu_1, \mu_2)$ probability of understanding each other in two way communication: $0 \leq F(\mu_1, \mu_2) = F(\mu_2, \mu_1) \leq 1$

• $F(\mu, \mu)$ communicability between linguistic agents with same language: $0 < F(\mu, \mu) \leq 1$

• Note: if had $\mathcal{S} = \mathcal{M}$ and μ supported on diagonal, P and Q would be identity and $F(\mu, \mu) = 1$

• Role of distribution σ on $\mathcal{M}$: not marginal of μ, but determined by "external world", which meanings are more likely to be communicated in a given context, F may be larger or smaller depending on this external context (two linguistic agents may communicate better or worse in different contexts)

Best Response

- suppose one language given $\mu = \mu_0$
- Want to maximize communicability

$$F(\mu_0, \mu_\star) = \sup_{\mu} F(\mu_0, \mu)$$

- Algorithm that approaches the best response $\mu_\star$
- construct a family of languages μ_ϵ such that $F(\mu_0, \mu_\epsilon)$ gets arbitrarily close to $\sup_\mu F(\mu_0, \mu)$ when $\epsilon \to 0$

Finite Languages: simplified model

Assume:

1. Languages are finite with μ an $M \times N$ matrix
2. The distribution σ is uniform $1/M$
3. The measure μ_0 has *unique maximum property*: for all $s \in \mathcal{S}$ there is a unique $m = m(s) \in \mathcal{M}$ and for $m \in \mathcal{M}$ a unique $s = s(m) \in \mathcal{S}$ with

$$\mu_0(s \,|\, m(s)) = \max_{m \in \mathcal{M}} \mu_0(s \,|\, m), \qquad \mu_0(s(m) \,|\, m) = \max_{s \in \mathcal{S}} \mu_0(s \,|\, m)$$

Best Decoder

- Find a matrix $Q_\star$ with

$$\sum_{ij} \mu_0(s_i \mid m_j)\ Q_{\star,ij} = \max_Q \sum_{ij} \mu_0(s_i \mid m_j)\ Q_{ij}$$

maximize over non-negative row-stochastic matrices

- this is given by

$$Q_{\star,ij} = \begin{cases} 1 & \mu_0(s_i \mid m_j) = \max_k \mu_0(s_i \mid m_k) \\ 0 & \text{otherwise} \end{cases}$$

Best Encoder

- Find a matrix $P_\star$ with

$$\sum_{ij} P_{\star,ij}\ \mu_0(m_j \,|\, s_i) = \max_P \sum_{ij} P_{ij}\ \mu_0(m_j \,|\, s_i)$$

maximize over non-negative column-stochastic matrices

- this is given by

$$P_{\star,ij} = \begin{cases} 1 & \mu_0(m_j \,|\, s_i) = \max_k \mu_0(m_j \,|\, s_k) \\ 0 & \text{otherwise} \end{cases}$$

Constrain relating them needs to be satisfied: $\exists \mu_\star$

$$\mu_\star(s_i \,|\, m_j) = P_{\star,ij}, \qquad \mu_\star(m_j \,|\, s_i) = Q_{\star,ij}$$

Problem: this does not always work

Approximations

- define $P^0_{ij} = \mu_0(s_i \mid m_j)$ and $Q^0_{ij} = \mu_0(m_j \mid s_i)$
- Result: for μ_0 finite with unique max and σ uniform

$$\sup_{\mu} F(\mu_0, \mu) = \frac{1}{2M} \mathrm{Tr}(P^0 \, {}^{\tau} Q_\star + P_\star \, {}^{\tau} Q^0)$$

- this follows from two properties:
  - $F(\mu_0, \mu) \leq \frac{1}{2M} \mathrm{Tr}(P^0 \, {}^{\tau} Q_\star + P_\star \, {}^{\tau} Q^0)$ for all μ
  - $\forall \epsilon \, \exists \mu_\epsilon$ with

$$\lim_{\epsilon \to 0} \left(\frac{1}{2M} \mathrm{Tr}(P^0 \, {}^{\tau} Q_\star + P_\star \, {}^{\tau} Q^0) - F(\mu_0, \mu_\epsilon) \right) = 0$$

- first property true by definition of best decoder and best encoder

Construction of the measures μ_ϵ with

$$\lim_{\epsilon \to 0} \mu_\epsilon(s_i \mid m_j) = P_{\star,ij} \qquad \lim_{\epsilon \to 0} \mu_\epsilon(m_j \mid s_i) = Q_{\star,ij}$$

- Auxiliary matrix X:

$$X_{ij} = \begin{cases} 1 & P_{\star,ij} + Q_{\star,ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

- form a Graph: G_X
  - vertices = entries of matrix X that are $= 1$
  - edges = lines connecting 1 entries on the same row and on the same column

- Fact: if measure μ_0 has *unique maximum property* then the graph does not have loops (tree or multiconnected forest)
- then construction of μ_ϵ:
  - for each component of G_X: take each pair of vertices
  - if connected by horizontal (vertical) line: look at corresponding entries of $Q_\star$ (or $P_\star$): one of them is one the other is zero
  - orient the edge from the vertex with entry 0 to the one with entry 1
  - start from one of the vertices: replace corresponding entry of X with ϵ
  - follow oriented path replace successive elements of X with ϵ^k (increasing k along *reverse* orientation of edges: unambiguous because no loops): matrix A^ϵ
  - measure μ_ϵ:

$$\mu_\epsilon(s_i, m_j) = \frac{A^\epsilon_{ij}}{\sum_{k,\ell} A^\epsilon_{k\ell}}$$

Limiting behavior of μ_ϵ

- normalize each column of A^ϵ so that sum adds up to one to get $\mu_\epsilon(s_i \mid m_j)$
- want to show these $\mu_\epsilon(s_i \mid m_j)$ converge to $P_{\star,ij}$
- each column of A^ϵ contains contains *at most one* edge of a connected component of G_X because $P_\star$ (resp. $Q_\star$) has at most one 1 entry per column (resp. row) so X has at most two
- for $\epsilon \to 0$ only lowest power of ϵ dominant, others to zero faster
- dominant term is column entry where $P_\star$ is 1: in the limit it gives the column of $P_\star$
- argument for $\mu_\epsilon(m_j \mid s_i) = Q_{\star,ij}$ is similar using rows

More general cases

• dropping all assumptions of unique maximum for μ_0, uniform distribution σ, and finite N and M:

- N. Komarova, P. Niyogi, *Optimizing the mutual intelligibility of mutual agents in a shared world*, Artificial Intelligence Journal, 154 (2004) 1–42.

Learning

• in this model: trying to communicate with an agent whose language is μ: best response strategy is trying to approximate $\mu_\star$ constructing some μ_ϵ (while $\mu_\star$ itself need not exist)

• but measure μ is unknown to learner: two possible scenarios

- full information: can sample μ directly for (meaning,sentence) pairs; then strategy is sample μ as accurately as possible, construct $P_\star$ and $Q_\star$ and from those μ_ϵ
- partial information: meaning is not directly accessible, only sentences are, and a feedback response on whether interpretation of sentence by learner is correct

Learning with full information

- Event E_{ij}: sentence s_i is produced to communicate meaning m_j
- probability of event E_{ij} is $\sigma(m_j)\,\mu(s_i \mid m_j)$
- for large n events drawn uniformly randomly frequencies k_{ij}/n approximate probability
- can estimate the $\sigma(m_j)\,\mu(s_i \mid m_j)$ using sampling frequencies k_{ij}/n
- use estimated $\sigma(m_j)\,\mu(s_i \mid m_j)$ to compute $P_\star$, $Q_\star$, μ_ϵ

Learning with Partial Information

• learner guesses meaning without direct access to it: if correct guess know meaning, if not only have negative information, asymmetric

• suppose guess meaning uniformly randomly among $M = \#\mathcal{M}$ possible meanings: guess m_j with probability $1/M$

• Event E_{ij}: sentence s_i is produced, meaning m_j is guessed, successfully

• probability of event E_{ij} is $\frac{1}{M}\sigma(m_j)\mu(s_i \mid m_j)$

• also can be empirically estimated from frequencies of correct guessing k_{ij}/n

• so apparently different setup leads to very similar procedure anyway

Communicative Efficiency: a phonetics example

- suppose English words are transmitted from a speaker to a receiver
- each word is a list of phonemes: if every phoneme is received correctly communicative efficiency would be 1
- some phonemes are notoriously difficult to distinguish in transmissions: *p* and *b* for example
- this causes ambiguities in words such as *bit* versus *pit* or *pat* versus *bat*
- since this can cause different meaning associations to sentences there is a loss of communicative efficiency
- subdivide the lexicon into cohorts: equivalence classes of words that become indistinguishable if certain phonemes are no longer distinguished

- p_i = probabilities (frequencies) of words in the original lexicon $\mathcal{W}$
- information content of the lexicon measured by Shannon entropy

$$S(\mathcal{W}) = -\sum_{w_i \in \mathcal{W}} p_i \log(p_i)$$

- after passing to cohorts $\mathcal{W}_{/\sim}$ probabilities $P_k = \sum_{w_i \in C_k} p_i$
- information content of set of cohorts

$$S(\mathcal{W}_{/\sim}) = -\sum_{C_k \in \mathcal{W}_{/\sim}} P_k \log(P_k)$$

- normalized Information Loss

$$\mathcal{IL}(\mathcal{W}, \sim) = \frac{S(\mathcal{W}) - S(\mathcal{W}_{/\sim})}{S(\mathcal{W})}$$

measures the "functional load" in communication carried by the ability to distinguished those phonemes

Communicative Fitness

- $\mathcal{H}$ set of n possible languages
- identify languages with measures μ_k on $\mathcal{M} \times \mathcal{S}$ (meanings and sentences/signals)
- mutual intelligibility matrix

$$\mathbb{A}_{ij} = \sum_{m \in \mathcal{M}} \sigma(m) \sum_{s \in \mathcal{S}} \mu_i(s \,|\, m) \mu_j(m \,|\, s)$$

probability that a speaker of language μ_i is understood by a receiver who speaks language μ_j

- simplified model: assume $A_{ii} = 1$ (each language has perfect intelligibility with itself) and $A_{ij} = a$ for some $0 \leq a \leq 1$ for all pairs $i \neq j$, same for all pairs

- also assume population of constant size with every person speaking only one language

- linguistic distribution of the population: $\alpha_k \geq 0$, with $\sum_{k=1}^{n} \alpha_k = 1$

- individual communicative fitness of a speaker of language μ_k: average communicative efficiency with the rest of the population

- mutual intelligibility of μ_i and μ_j

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

• so average communicative efficiency of a speaker of μ_i

$$f_i = f_0 + \sum_{j=1}^{n} F(\mu_i, \mu_j)\alpha_j$$

f_0 = background, independent of language (but dependent of how much specific environment facilitates communication)

• if everybody spoke the same language μ (assuming $\mathbb{A}_{ii} = 1$) would have $f = f_0 + 1$; if other languages are present, lower value of fitness f

• following basic rule of evolution by natural selection: assume individuals reproduce in proportion to their fitness

• assuming successful communication is an evolutionary advantage in the Darwinian sense

• in this model also make the assumption that children learn language from their parents and not from the entire community

- further simplify the model by assuming each learner has only one teacher (literally "mother tongue")
- also allow for mistakes during language acquisition
- probability of a transition from language μ_i to language μ_j is Q_{ij} (depends on $\mathbb{A}$: on how close the different languages are)
- Population Dynamics

$$\alpha_{t+1,j} = \frac{\sum_{i=1}^{n} \alpha_{t,i} f_i Q_{ij}}{\sum_{k=1}^{n} \alpha_{t,k} f_k}$$

reproduction proportional to fitness: percentage of new generation produced by speakers of language μ_i in previous generation is $f_i \alpha_{t,i}$ (normalized by $\sum_k f_k \alpha_{t,k}$)

ODE: turn difference equation into ordinary differential equation

- normalization condition $\sum_k \alpha_k = 1$ gives $\sum_k \dot{\alpha}_{t,k} = 0$
- positivity $\alpha_k \geq 0$ becomes condition $\dot{\alpha}_{t,k}|_{\alpha_{t,k}=0} \geq 0$
- continuous time differential equation

$$\dot{\alpha}_{t,k} = \sum_{i=1}^{n} f_i \, \alpha_{t,i} \, Q_{ij} - \phi \, \alpha_{t,k}$$

with $\phi(t) = \sum_k f_k \, \alpha_{t,k}$ *average fitness of the population*

- for case with $A_{ii} = 1$ and $A_{ij} = a$ for $i \neq j$ fitness

$$f_i = (1 - a)\alpha_i + a + f_0$$

- *learning fidelity*: probability $\frac{1}{n} \leq q \leq 1$ of learning same language as primary teacher

$$Q_{ii} = q \quad \text{and} \quad Q_{ij} = \frac{(1-q)}{n-1} \text{ when } i \neq j$$

perfect learning $q = 1$; random guessing $q = 1/n$

- then differential equation

$$\dot{\alpha}_{t,k} = (1 - a)\left(-\alpha_{t,k}^3 + \alpha_{t,k}^2 q + \sum_{j \neq k} \alpha_{t,j}^2 \left(\frac{1-q}{n-1} - \alpha_{t,k}\right)\right)$$

$$- \frac{(a + f_0)(1 - q)(n\alpha_{t,k} - 1)}{n - 1}$$

Equilibrium Solutions (critical points $\dot{\alpha}_{t,k} = 0$)

- each x_j root of polynomial (with $\gamma = \sum_j x_j^2$)

$$P_{a,q,n}(x) = (1-a)\left(-x^3 + x^2 q + (\gamma - x)\left(\frac{1-q}{n-1} - x\right)\right)$$

$$- \frac{(a+f_0)(1-q)(nx-1)}{n-1}$$

- if $x_\ell = X$ and all other $x_k = \frac{1-X}{n-1}$ (so $\sum_j x_j = 1$) then equation becomes

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

- Cubic polynomial: three solutions given by $\frac{1}{n}$, and $r_\pm$

$$r_\pm = \frac{-(1-a)(1+(n-2)q) \mp \sqrt{D}}{2(a-1)(n-1)}$$

$$D = 4(-1-a(n-2)-f_0(n-1))(1-q)(n-1)(1-a)+(1-a)^2(1+(n-2)q)^2$$

- So in total $2n+1$ solutions:
  1. uniform solution $x_k = 1/n$ for all k
  2. one $x_\ell = r_+$ and all other $x_k = (1-r_+)/(n-1)$ (n possibilities for x_ℓ)
  3. one $x_\ell = r'_-$ and all other $x_k = (1-r_-)/(n-1)$ (n possibilities for x_ℓ)

the last two cases lead to one preferred language (and all the others with same distribution)

Population Dynamics Model

• following previous model: languages $\mu_1, \ldots, \mu_n$ and linguistic evolution in population modelled by ODE

$$\dot{x}_j = \sum_i x_i \, f_i \, Q_{ij} - \phi \, x_j$$

$x_j = \alpha_j$ proportion of individuals speaking language μ_j

• matrix Q measure fidelity of language map (how much deviation from teacher to learner)

• $f_i =$ fitness

$$f_i = \sum_j x_j \, F(\mu_i, \mu_j)$$

Assumptions

- assuming as before that
  - $Q_{ii} = q$ and $Q_{ij} = \frac{1-q}{n-1}$ for $i \neq j$
  - $F(\mu_i, \mu_i) = 1$ and $F(\mu_i, \mu_j) = a$ for all $i \neq j$
  - $f_i = (1-a)x_i + a + f_0$

Threshold behavior depending on parameter q

- for q small only stable critical point is uniform distribution: all $x_j = 1/n$
- *bifurcation* at some $q = q_1$: two new critical points $r_\pm$

- one-grammar solutions emerge where the majority of population speaks one of the languages

Without fitness

- Note: same equation with $f_i = f_0$ (without fitness function)
- would have $\phi = f_0 \sum_j x_j = f_0$
- equation would be

$$\dot{x}_j = f_0 \sum_i x_i Q_{ij} - f_0 x_j$$

becomes a linear system of ODE

- only equilibrium solution at $x_j = 1/n$, uniform distribution
- no bifurcation and no emergent behavior creating language coherence: those are effects of the presence of the fitness function

Social Learning

• this model was based on assumption that learner takes input only from one teacher (with the possibility of errors in reproduction encoded in Q_{ij})

• consider again other scenario where learner's input is coming from the entire population

• given n languages $\mathcal{L}_1, \ldots, \mathcal{L}_n$ assume a set of expressions is especially useful for language acquisition (triggers, cues, ...)

• this gives subsets $C_i \subseteq \mathcal{L}_i$; assume $C_i \cap C_j = \emptyset$ (these are unambiguous cues)

• speakers of $\mathcal{L}_i$ produce sentences randomly with distribution $\mathbb{P}_i$ and likelihood of producing a cue is

$$a_i = \mathbb{P}_i(C_i)$$

• simplifying assumption: all $a_i = a$ same

Case of two languages

- proportions $\alpha, 1-\alpha$ of speakers: function of time $x_1(t) = \alpha(t)$, $x_2(t) = 1 - \alpha(t)$
- cue-frequency based batch learner: $m = k_1 + k_2 + k_3$
  - k_1 sentences in input that are in C_1
  - k_2 in C_2
  - k_3 are not cues
- probability of $k_1 > k_2$

$$f_{1,a,m}(x_1, x_2) = \sum \binom{m}{k_1 k_2 k_3} (a x_1(t))^{k_1} (a x_2(t))^{k_2} (1-a)^{k_3}$$

sum over (k_1, k_2, k_3) with $m = k_1 + k_2 + k_3$ and $k_1 > k_2$

- probability $f_{2,a,m}$ of $k_1 < k_2$, same with sum over (k_1, k_2, k_3) with $m = k_1 + k_2 + k_3$ and $k_1 > k_2$

- symmetric assumption $a_i = a$ gives $f_{2,a,m}(x_1, x_2) = f_{1,a,m}(x_2, x_1)$
- probability after m inputs of learner acquiring $\mathcal{L}_1$

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

(if no cues received at all: $1/2$ chance of one language or other)

- population dynamics equation

$$x_1(t+1) = \frac{1}{2}(1 + f_{1,a,m}(x_1(t), x_2(t)) - f_{2,a,m}(x_1(t), x_2(t))$$

- a fixed point at $x_1 = x_2 = 1/2$: uniform distribution of population among the two languages

- if number of inputs m small: only fixed point (stable)
- for larger m other fixed points appear (one language becomes dominant)
- for larger m uniform solution $x_1 = x_2 = 1/2$ becomes unstable
- the value of m where bifurcation occurs is a function of parameter a
- can also keep m fixed and vary a:
  - a close to zero: only $x_1 = x_2 = 1/2$ (stable fixed point)
  - bifurcation when a grows: new stable fixed points and $x_1 = x_2 = 1/2$ becomes unstable
  - bifurcation occurs at a value of a dependent on m

Stability of $x_1 = x_2 = 1/2$: more details

- derivative at the fixed point

$$f'_{1,a,m}(1/2,1/2) = \sum_{k_1>k_2} \binom{m}{k_1 k_2 k_3} a^{m-k_3}(1-a)^{k_3}(k_1-k_2)(\frac{1}{2})^{k_1+k_2-1}$$

similar for $f'_{2,a,m}$

- $f'_{1,a,m}(1/2,1/2)|_{a=0} = 0$ so by continuity for small a have

$$|f'_{1,a,m}(1/2,1/2)| < 1$$

stability while in this range

- also see that when $a = 1$, for sufficiently large m have $f'_{1,a,m}(1/2,1/2)|_{a=1} > 1$ so in between will cross value 1: where bifurcation occurs

- emergence of linguistic coherence in the population

Case of n languages

- learner is exposed to a mixture of languages form the environment
- learner scans incoming data for cues and chooses the language from which largest number of cues is received
- if multiple languages with same number of cues: pick one among them randomly
- same simplifying assumption as before $\mathbb{P}_i(C_i) = a$ same for all languages

Algorithm

1. Count cues
   - k_i = number of cues in C_i out of m inputs
   - k_{n+1} = number of non-cues (in any of the languages)
   - $m = k_1 + \cdots + k_n + k_{n+1}$
2. Find maximal languages: languages $\mathcal{L}_i$ with $k_i = \max_j k_j$: $\mathcal{I}$ = set of indices of $\mathcal{L}_i$ maximal
3. Choose language: if $|\mathcal{I}| = 1$ choose that language; if $|\mathcal{I}| > 1$ choose one language randomly in the set $\mathcal{I}$ with probability $1/|\mathcal{I}|$
4. of naive version: just choose a language randomly among all n with probability $1/n$

Population Dynamics in this model

- $\mathbb{P} = \sum_i x_i(t)\mathbb{P}_i$ probability with which input is generated
- $p_i = p_i(t) = a x_i(t)$ probability of receiving a cue from language $\mathcal{L}_i$; $p_{n+1} = 1 - a$
- probability of receiving (strictly) more cues from language $\mathcal{L}_1$ than from any other

$$F_{1,m,a}(x_1, \ldots, x_n) = \sum \binom{m}{k_1 \cdots k_{n+1}} p_1^{k_1} \cdots p_n^{k_n} p_{n+1}^{k_{n+1}}$$

sum over all $(k_1, \ldots, k_{n+1})$ with $m = k_1 + \cdots + k_{n+1}$ and $k_1 > k_j$ for all $j \neq 1$

- similar for other languages with symmetry

$$F_{i,m,a}(\cdots, x_i, \cdots, x_j \cdots) = F_{j,m,a}(\cdots, x_j, \cdots, x_i \cdots)$$

• in this model, probability that learner will choose $\mathcal{L}_i$ after m input data

$$f_{i,m,a}(x_1,\ldots,x_n) = F_{i,m,a}(x_1,\ldots,x_n) + (1 - \sum_{j=1}^{n} F_{j,m,a}(x_1,\ldots,x_n))\frac{1}{n}$$

(with naive version of choice in the cue-less case)

• Recursion relation for population distribution in next generation

$$x_i(t+1) = f_{i,m,a}(x_1(t),\ldots,x_n(t))$$

Fixed Points

- $f = (f_{i,m,a})_{i=1}^{n}$ continuous map $f : \Delta_{n-1} \to \Delta_{n-1}$
- Results
  1. f has finite number of fixed points: at most $m2^n$
  2. for small m only fixed point is $(\frac{1}{n}, \ldots, \frac{1}{n})$, stable
  3. for fixed (sufficiently large) m number of fixed points varies with a: small a only one fixed point (uniform distribution); as a increases bifurcation: other fixed points arise
  4. large values of $a \sim 1$: uniform distribution no longer stable, only the fixed points with one dominant language are

Language Learning and Statistical Physics

• these bifurcations and emergence of linguistic coherence reminiscent of behavior of Ising model and spin glass systems in Statistical Physics

• an ensemble of interacting components

• degree of interaction governed by a thermodynamic parameter $\beta \sim 1/T$ inverse temperature

• these systems often exhibit *phase transitions* between different regimes, at some critical temperature $T = T_c$ (different states of matter, loss of magnetization, etc.)

Language Evolution in Locally Connected Societies

- two possible languages: $\{\mathcal{L}_0, \mathcal{L}_1\} = \{0, 1\}$
- Graph G representing linguistic agents and their interaction
  - each vertex $v \in V(G)$ has an associated random variable $X_v(t)$
  - $X_v(t) \in \{0, 1\}$: language of agent occupying position v
  - $X_v(t+1)$ language occupying same position at next step (generation)
  - $\mathbb{P}(X_v(t+1) = 1) = g_{a,m}(\mu_v(t))$

$$\mu_v = \frac{1}{\mathrm{val}(v)} \left(\sum_{e \in E(G): \partial(e) = \{v, v'\}} X_{v'}(t) \right)$$

- nearest neighbor interaction considered only

- as before assuming $a = \mathbb{P}_i(C_i)$ same for both languages
- a possible choice for the function $g_{a,m} : [0,1] \to [0,1]$:

$$g(x) = \frac{1}{2} + \frac{1}{2}(f_{1,a,m}(x, 1-x) - f_{1,a,m}(1-x, x))$$

with $f_{1,a,m}$ as before counting probability of set of cues $k_1 > k_2$

$$f_{1,a,m}(x, 1-x) = \sum \binom{m}{k_1 k_2 k_3} (ax)^{k_1} (a(1-x))^{k_2} (1-a)^{k_3}$$

sum over (k_1, k_2, k_3) with $m = k_1 + k_2 + k_3$ and $k_1 > k_2$

- study evolution of

$$\alpha_G(t) = \frac{1}{\#V(G)} \sum_{v \in V(G)} X_v(t)$$

average number of $\mathcal{L}_1$-speakers at time/generation t

- for a complete graph have all language users connected to all others: recover model in which learning from whole community

- can consider asymptotic behaviors when size of graph becomes large $\#V(G) = N \to \infty$

- can simplify the geometry making special assumptions on the graph: e.g. a square lattice

The Ising Model of spin systems on a graph G

- configurations of spins $s : V(G) \to \{\pm 1\}$
- magnetic field B and correlation strength J: Hamiltonian

$$H(s) = -J \sum_{e \in E(G): \partial(e) = \{v, v'\}} s_v s_{v'} - B \sum_{v \in V(G)} s_v$$

- first term measures degree of alignment of nearby spins
- second term measures alignment of spins with direction of magnetic field
- see previous discussion of Spin Glass Models of Language Evolution though syntactic parameters

Questions: enrich the spin glass model with models of language acquisition/evolution as discussed here; enrich dynamics of language evolution used in phylogenetic (Markov models on trees) with some more refined information about language acquisition and language change