LI 511: Computational Models of Sound Change

James Kirby and Morgan Sonderegger

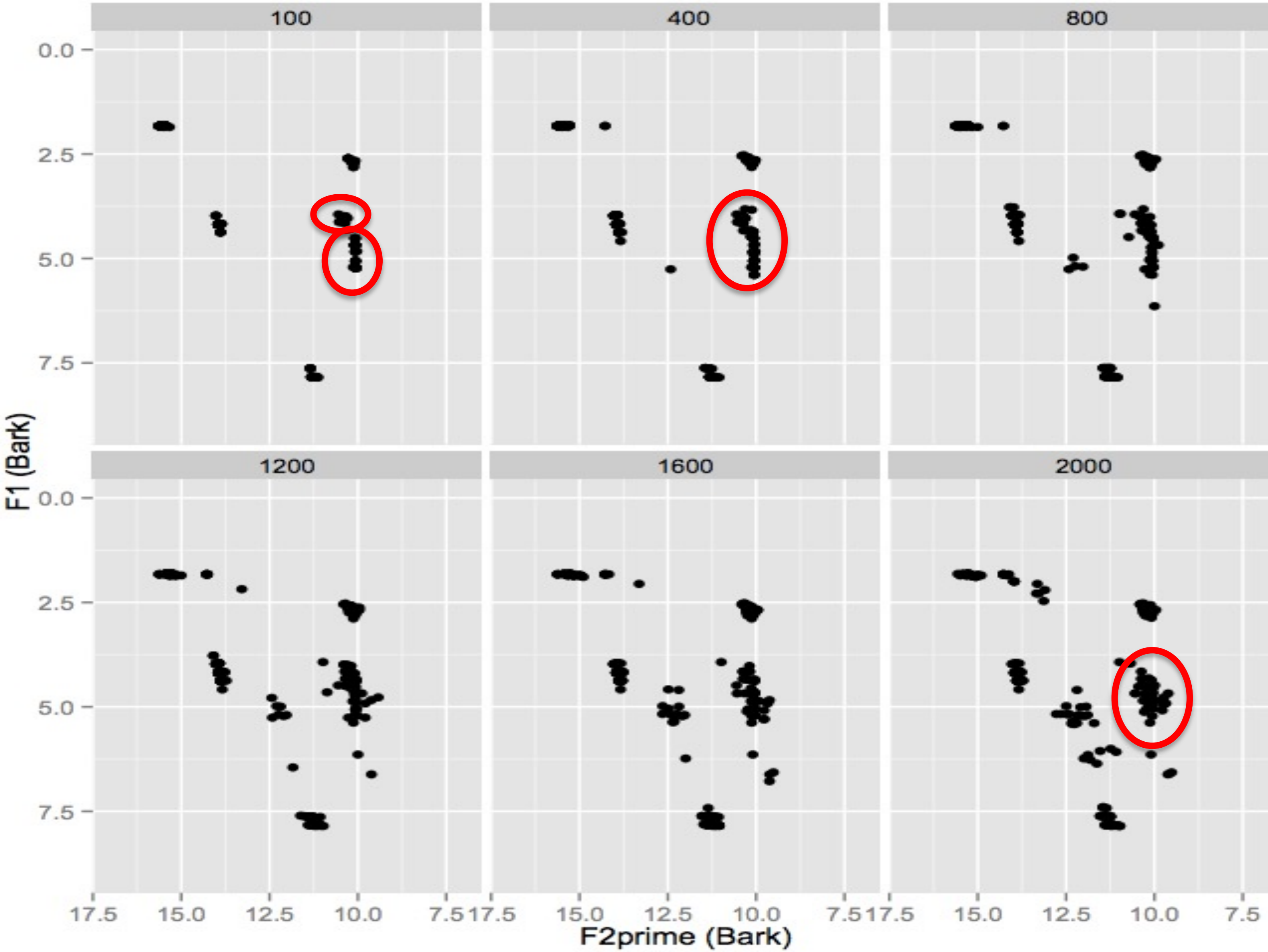
9 July 2013

Administrative

- Office Hours: Wed, 5:30-6:30 (Espresso Royale)
- Short email (250 words max) on topic + plan: Friday
- Phonetics-Phonology Social Hour: Tonight, 5-6:45 PM, Dominick's

Dialect contact 1 (Christina & Rebecca)

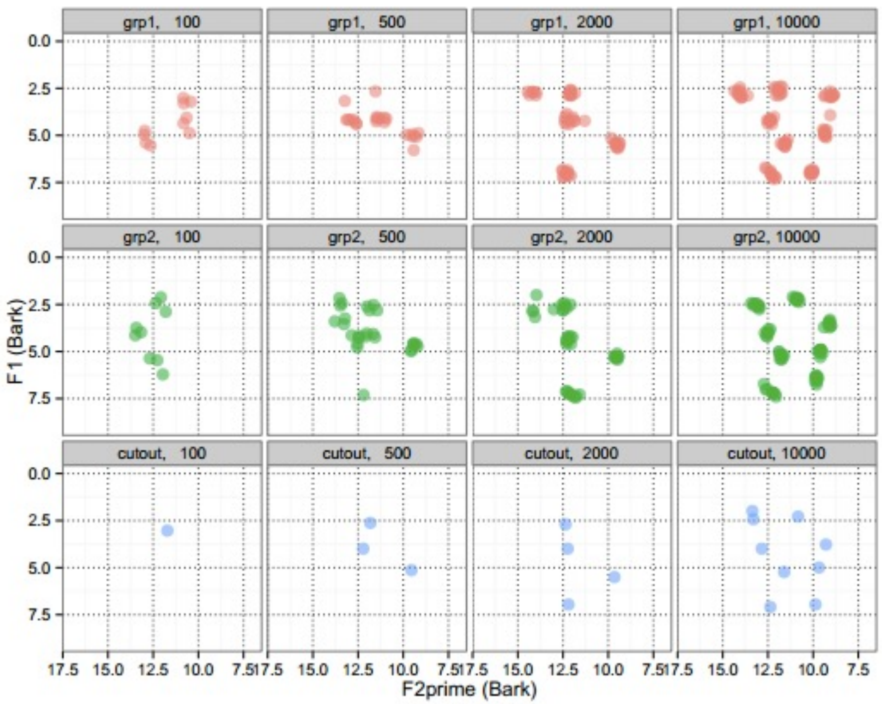
- Two groups with 5-vowel systems with 1 vowel different put into contact at t=0
- Herzog's Principle (mergers > splits)

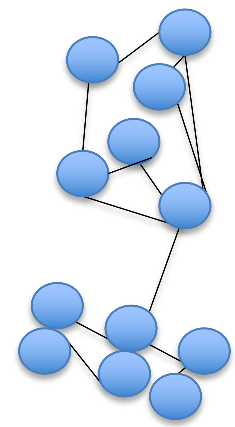

100
400
800
1200
1600
2000
0.0
2.5
5.0
7.5
F1 (Bark)
17.5
15.0
12.5
10.0
7.5
F2prime (Bark)

Dialect contact II (Jon F. & Meghan)

1. Two groups, connected by one ‘cutout’ agent
2. Two groups suddenly put into full contact

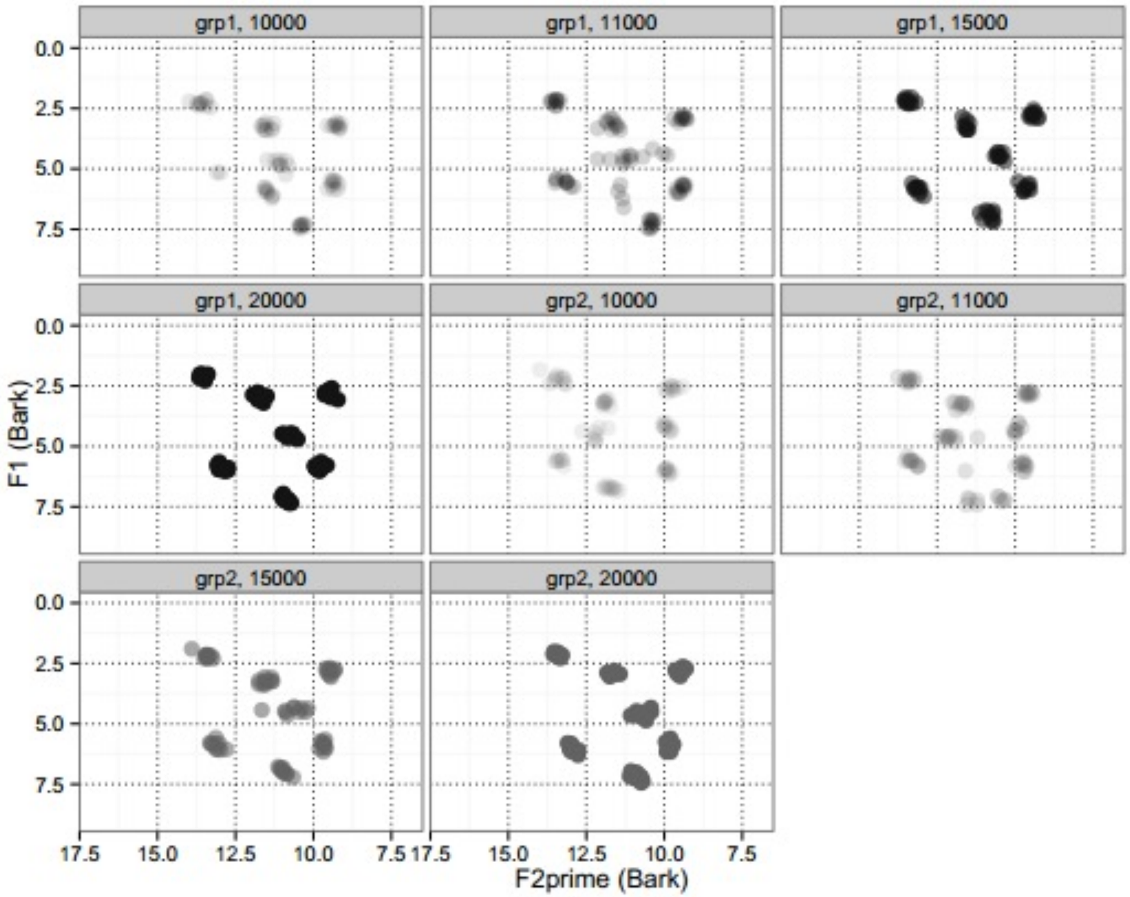
1





Q: Would convergence still happen with large enough group size?

2.



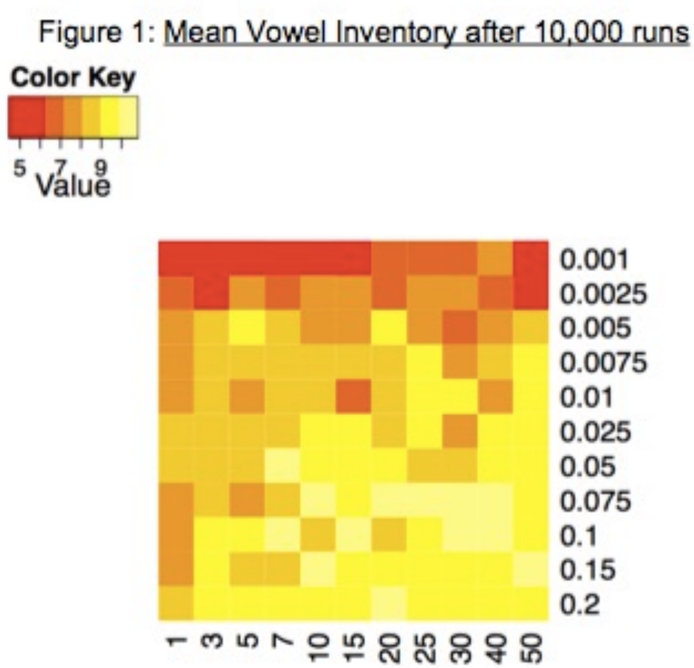
grp1, 10000
grp1, 11000
grp1, 15000
grp1, 20000
grp2, 10000
grp2, 11000
grp2, 15000
grp2, 20000
0.0
2.5
5.0
7.5
F1 (Bark)
17.5
15.0
12.5
10.0
7.5
F2prime (Bark)

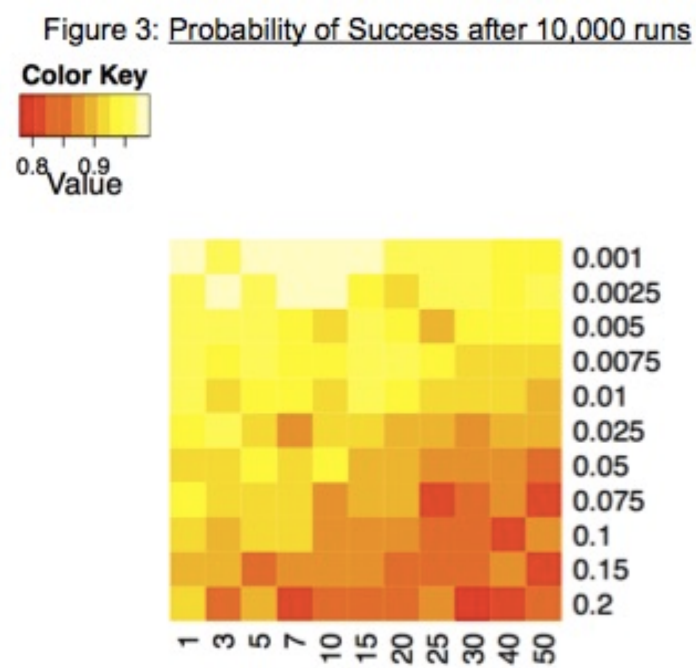
Varying $n_{minUses}$ and $p_{addition}$ (Jon H. and Russell)

- Increased $n_{minUses}$ or $p_{addition}$:
  - Higher energy, vowel inventory size, P(success)

Varying $n_{minUses}$ and $p_{addition}$ (Jon H. and Russell)

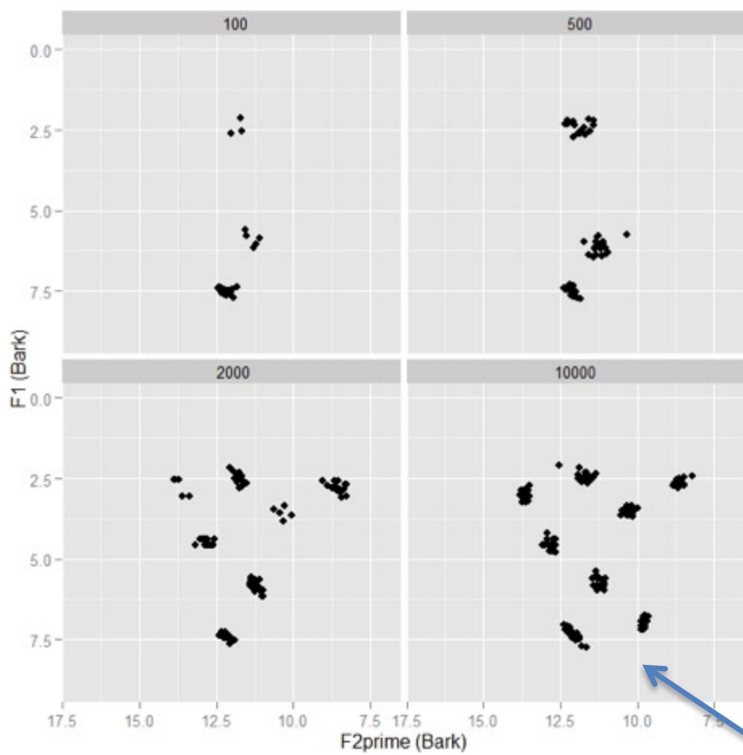
- Increased $n_{minUses}$ or $p_{addition}$:
  - Higher energy, vowel inventory size, P(success)

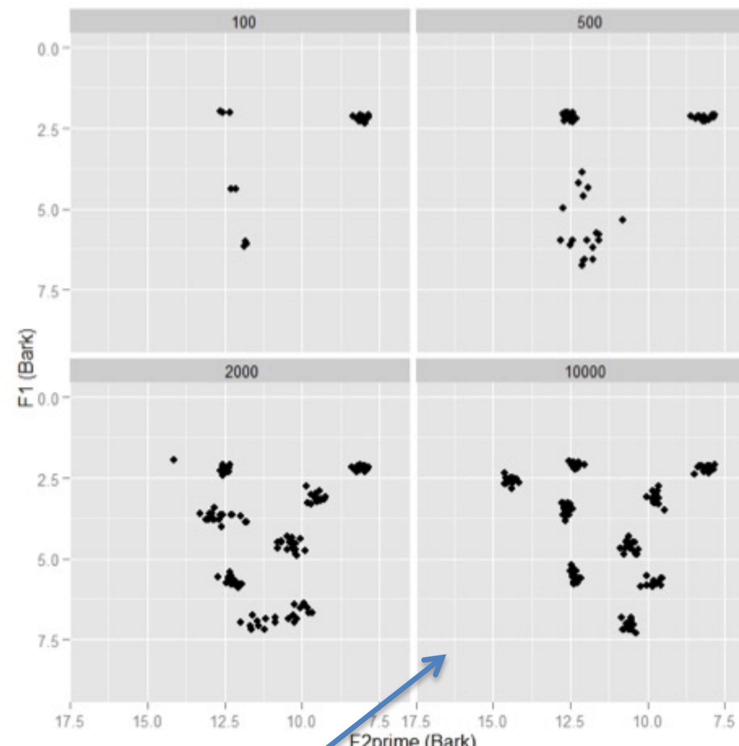




Varying the initial vowel inventory (Jiang, Kouros, Mingxing)

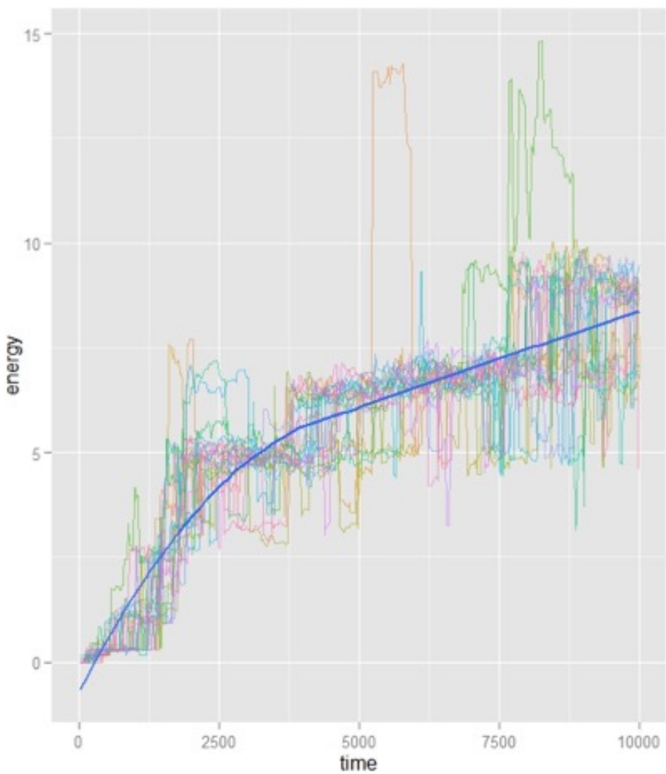
- Agents start with just /a/ or /u/:

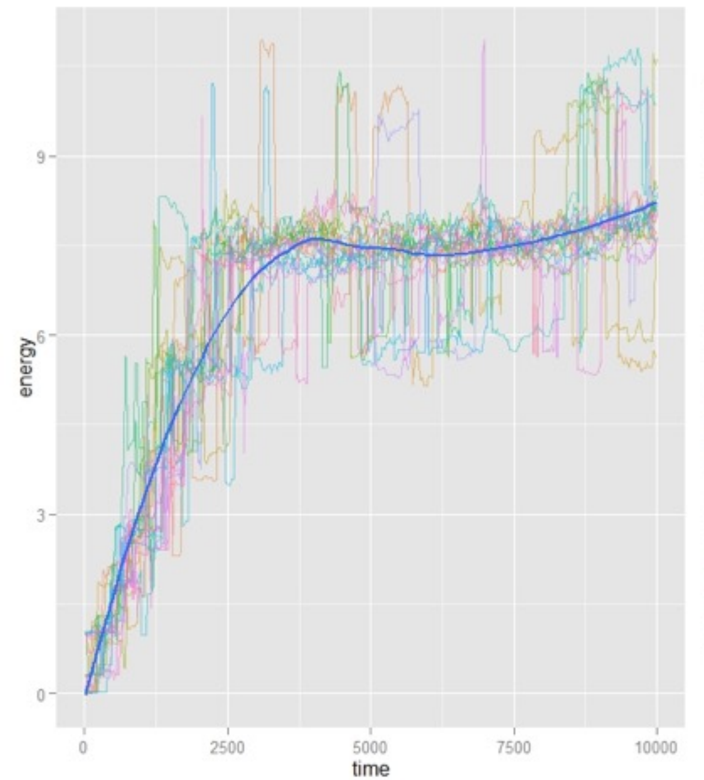




n=9 vs n=8, would need to check with more runs

• Start with n=1 vs n=5: faster convergence





Varying acoustic and articulatory merge thresholds (Jevon & Sarah)

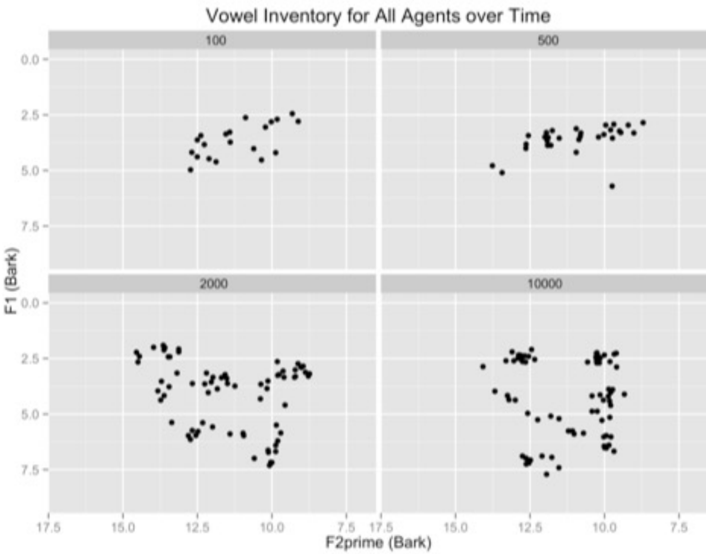
- Start with high articulatory threshold
  → one-vowel system
- "Individual runs of the same threshold yield different results…"
- Higher artic or acoust threshold ⇒ larger, more diffuse categories
  - Larger of the two thresholds dominates

Mortal agents with evolving ε, varying n_{agents} (Anthony, Emily, Reza)

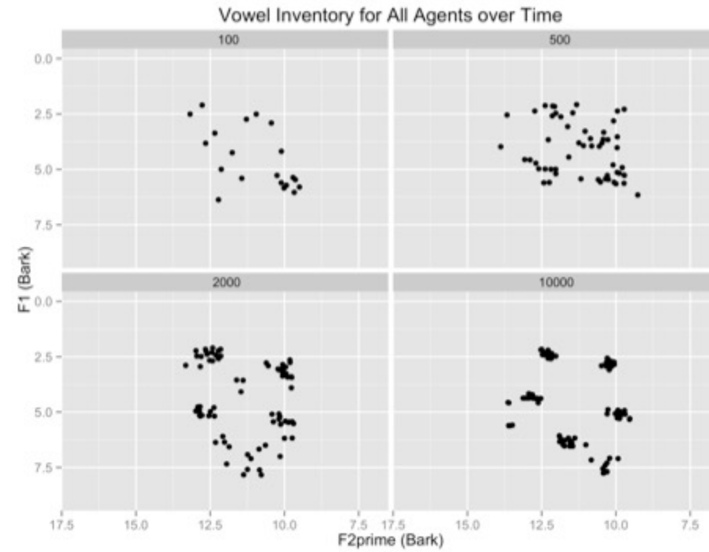
- Agent dies and is replaced with p_{bd} (each round)
- ε↓ t=0-300, then ↑300-800

- Higher p_{bd}
  - Fewer, larger V categories
  - Still get stable system! (c.f. de Boer book, Sec. 5.4)

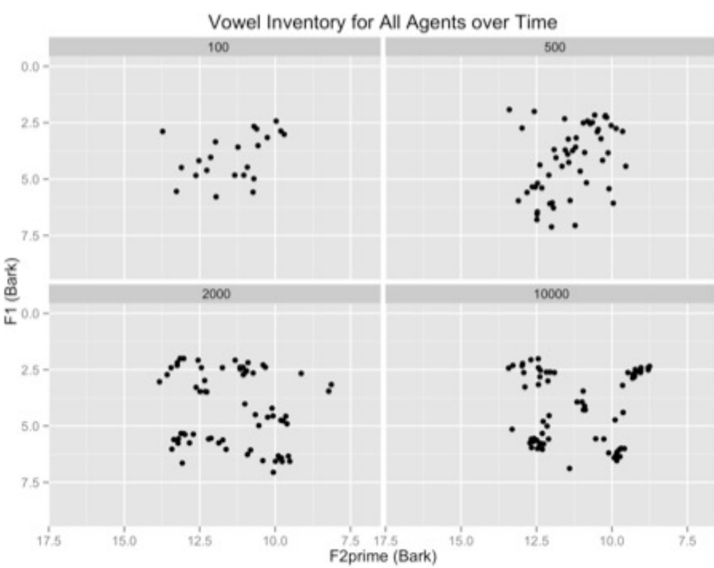
p_{bd}=0.1, n_{agents} = 20



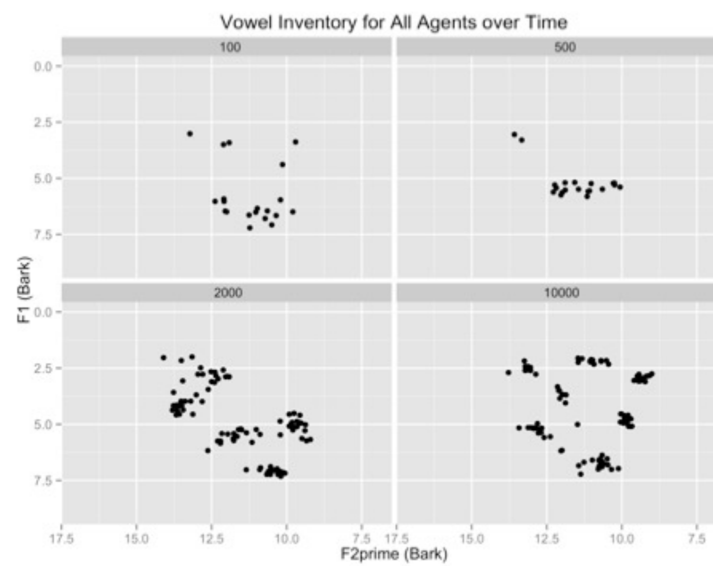
p_{bd}=0.005, n_{agents} = 20



p_{bd}=0.1, n_{agents} = 40



p_{bd}=0.005, n_{agents} = 40



Explorations, paper: Summary

- $p_{addition}$, $n_{minUses}$, acoust/artic thresholds, birth/death rate, ε (more in book)
  - ↑n, category variance; ↓energy

- n_{agents} :↑prototype stability

- Starting inventory : No qualitative diff

- Systems likely to converge with any contact

Explorations, paper: Summary

- $p_{addition}$, $n_{minUses}$, acoust/artic thresholds, birth/death rate,ε (more in book)
  - ↑n, category variance; ↓energy
- n_{agents} :↑prototype stability
- Starting inventory : No qualitative diff
- Systems likely to converge with any contact
  - Desirable? Any ideas to alleviate?

- In general: What's here, missing vs. empirical facts?
  - From sense of model so far, what directions to be explored to fill gaps?

- In general: What's here, missing vs. empirical facts?
  - From sense of model so far, what directions to be explored to fill gaps?

- Methodology for doing a 'real' extension
  - Many simulations per parameter setting
  - Sweep parameter subspace
  - Consider alternative explanations for observations

(keep in mind for projects..)

de Boer (1999) and extensions: What have we learned?

- Existence proofs?
- Explicitness / implementation?
- Counterintuitive results?
- Qualitative predictions?
- Baseline?

Analytic vs. simulation approaches

- Recap: Pros and cons
- This week: More analytic models

Interlude: Some math

- Recurrent tools:
  - Conditional probability
  - "Bayesian inference"

- Mathematical objects used (here) to analyze systems changing over time
  1. Stochastic evolution: Markov chains
  2. Deterministic evolution: Dynamical systems

Conditional probability

- Refresher?

Bayesian inference

- Hypotheses: $h \in H$
- Data: $x \in D$

- Prior: How likely are different *h*, a priori?
  - Equally likely: "Flat prior"

- Posterior: How likely are different *h*, after seeing the data?

Bayesian inference: Example

- Guess vowel category given F_1, F_2
- Guess word in "I ate some fis*S*", where *S* = ambiguous sibilant
  - – Prior: likelihood of word in this context
  - – Data: Signal actually heard

Models: Overview

- Dynamical systems, iterated learning models

- Main concern: Relationship of population-level diachronic dynamics to assumptions about
  1. Knowledge state of individuals
  2. Learning algorithm
  3. Assumptions about communication
  4. Network structure

- Usually: Examine <u>long-term</u> behavior <u>analytically</u>

Today: Discrete variables, knowledge state

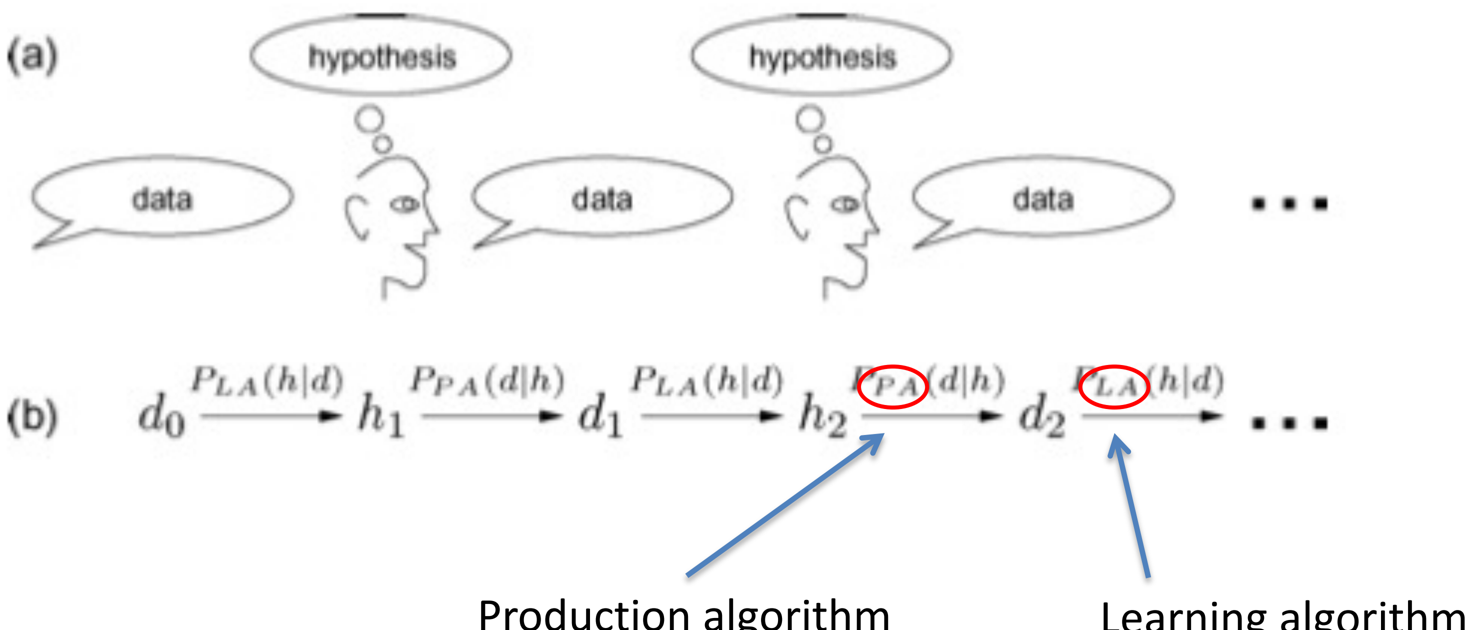
- Network structure
  - Generation size = 1, ∞
  - Number of parents
- Knowledge state: Categorical (0/1)
- Learning algorithm
  - Various

(Continuous variables, knowledge state: Thursday)

Iterated learning: Griffiths & Kalish (2007)

- IL* scenario: chain of learners/teachers
  - H_i : Hypothesis
  - d_i : Data

For agent *i*



Interlude: Markov chains

- Markov chain: Sequence of random variables $v_0, v_1, \ldots$ s.t.

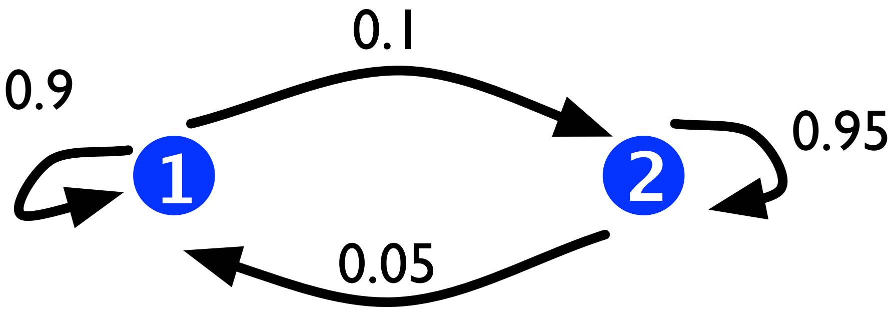
$$P(v_n | v_0, v_1, \ldots, v_{n-1}) = P(v_n | v_{n-1})$$

- States: (Finite) set V
- Transition probabilities:

$$q_{ij} = P(v_n = i | v_{n-1} = j), \quad i, j \in V$$

- Transition matrix $Q = \{q_{ij}\}$

Example



$$V = 1, 2$$

$$Q = \begin{pmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{pmatrix}$$

Note: Each row of *Q* sums to 1

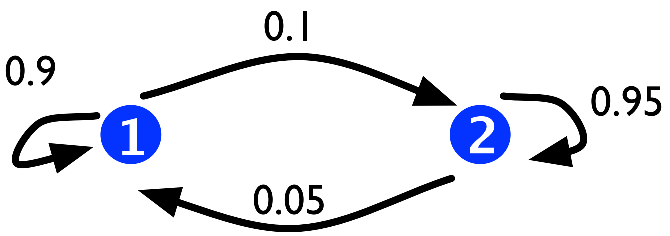
- Stationary distribution: vector p s.t.

$$\vec{p} = Q\vec{p}$$

- Theorem: Under some conditions on Q, there is a unique stationary distribution.

- As $t \longrightarrow \infty$, always end up in the stationary distribution.

Example



$$Q = \begin{pmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{pmatrix}$$

Stationary distribution: $\vec{p} = \begin{pmatrix} 0.333 \\ 0.666 \end{pmatrix}$

- Griffiths & Kalish:
  - Hypothesis h_i (and data d_i) is a Markov Chain
  - Goal: Determine stationary state*

* exists if some conditions apply

GK Model 1

- Two languages, 0/1 output (“Gavagai!”)
- Data: $d_i \in X$ (one data point)
- Hypotheses: 1, 2 (L_1, L_2)
- Grammars: G_1, G_2

- Production algorithm:

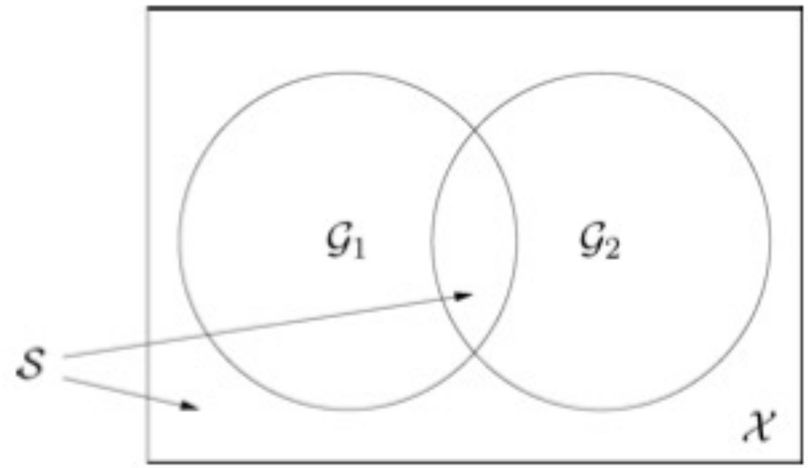
$$P_{PA}(d = (x, y)|h = i) = \begin{cases} P(x)(1-\epsilon) & \text{if } y = I(x \in \mathcal{G}_i) \\ P(x)\epsilon & \text{otherwise} \end{cases}$$

- Noise: ε

- Prior: $P(h_1) = \alpha, \quad P(h_2) = 1 - \alpha$

- Ambiguous objects: *S*

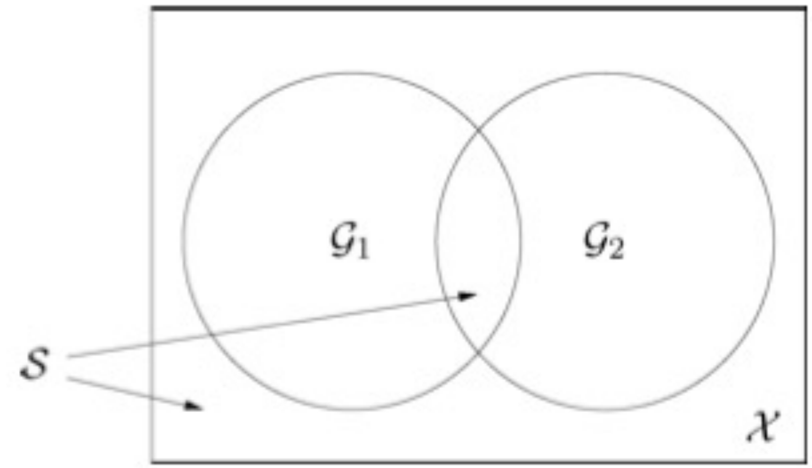
$$s = P(x \in \mathcal{S})$$



- Prior: $P(h_1) = \alpha, \quad P(h_2) = 1 - \alpha$

- Ambiguous objects: *S*

$$s = P(x \in \mathcal{S})$$



- Learner *i* gets data, ⇒ posterior

- How do they proceed?
  - Model 1.1 : Sampling from the posterior
  - Model 1.2 : Maximum a posteriori estimation

Sampling from the posterior

- Learner *i* :
  - Picks h_i with $P(h_i|d_{i-1})$
  - Generates d_i from h_i

- We can calculate $P(h_t = 1|h_{t-1} = 2)$, etc.
  - These are the q_{ij} ⇒ transition matrix Q

Sampling from the posterior

- Learner i :
  - Picks h_i with $P(h_i|d_{i-1})$
  - Generates d_i from h_i
- We can calculate $P(h_t = 1|h_{t-1} = 2)$, etc.
  - These are the q_{ij} ⇒ transition matrix Q
- Stationary state: $(\alpha, 1-\alpha)$
  - Convergence to the prior
  - No dependence on ε, s !

MAP estimation

- Learner *i* :
  - Picks h_i which <u>maximizes</u> $P(h_i|d_{i-1})$
  - Generates d_i from h_i

- Can again work out *Q*
  - Stationary distribution now depends on all system parameters

$P(h_1)$ in stationary distribution

Properties of the Markov chain on hypotheses for iterated learning with MAP estimation

Condition	q_{12}	q_{21}	θ_1
$\epsilon < 1 - \alpha$	$s + (1 - s)\epsilon$	$(1 - s)\epsilon$	$\frac{s+(1-s)\epsilon}{s+2(1-s)\epsilon}$
$\epsilon = 1 - \alpha$	$s + (1 - s)(1 + \epsilon)/2$	$(1 - s)\epsilon/2$	$\frac{s+(1-s)(1+\epsilon)/2}{s+(1-s)(1+2\epsilon)/2}$
$\epsilon > 1 - \alpha$	1	0	1

- $P(h_1)$ always > 0.5 : L_1 favored as $t \longrightarrow \infty$
- By how much depends on $s, \ \alpha, \ \epsilon$
  - Note: <u>No</u> dependence on α !

GK Model 1: Discussion

- Preference encoded in prior maintained…
- … by how much depends heavily on
  - Learning algorithm
  - Assumptions about communication
- What relevance for sound change?
  - (what kind of situation would this model?)

Iterated learning

- This type of single-chain model (usually)

- Evolang, cultural evolution literatures
  - Simulation, analytical
    (Griffiths & Kirby, 2007; Kirby, 2000 et seq; Kirby et al., 2007; Smith et al., 2003…)
  - <u>Experimental</u>
    (Kalish et al, 2007; Kirby et al, 2008; Sanborn & Griffiths, 2008…)

Iterated learning

- This type of single-chain model (usually)

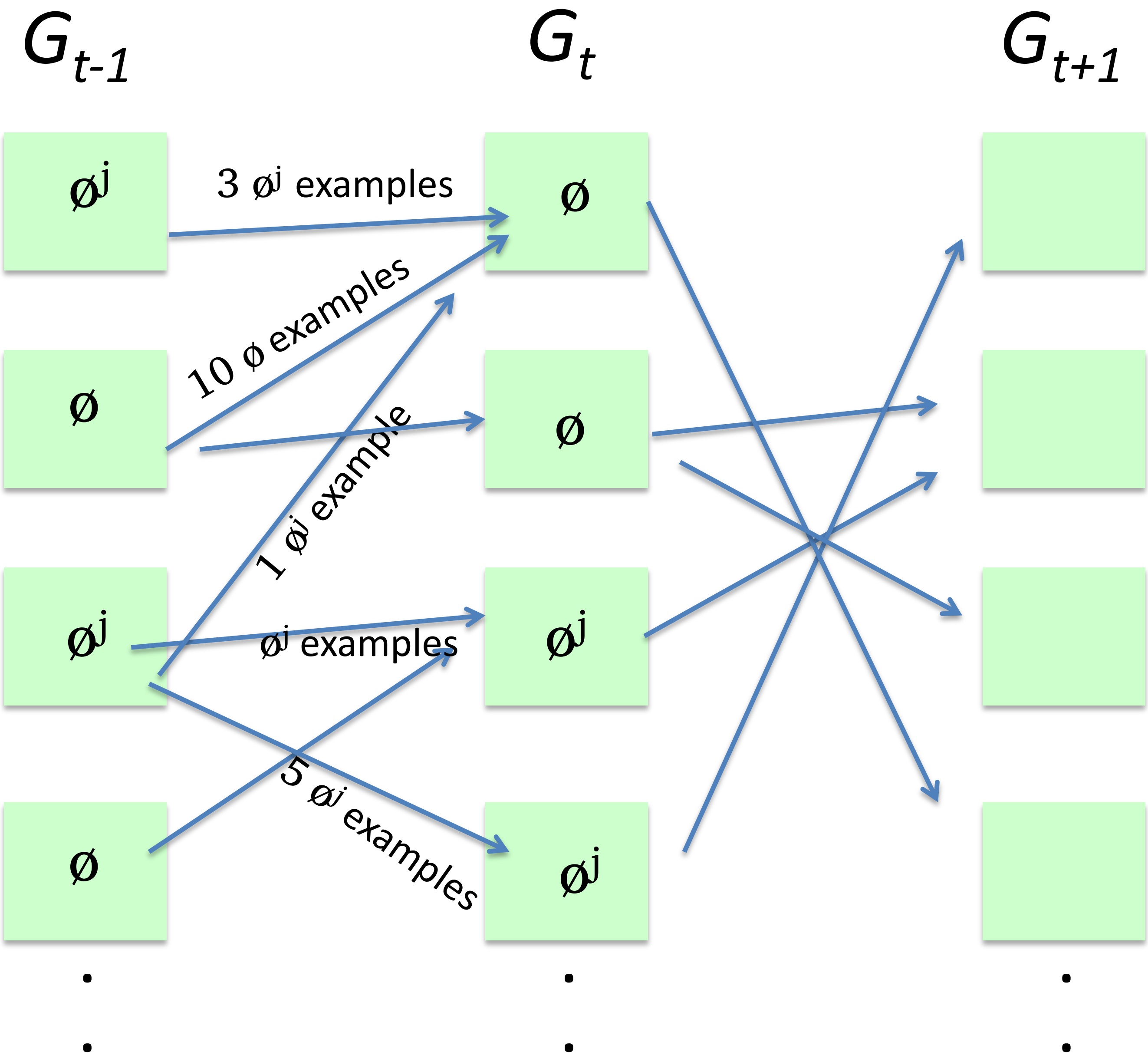
- Evolang, cultural evolution literatures
  - Simulation, analytical
    (Griffiths & Kirby, 2007; Kirby, 2000 et seq; Kirby et al., 2007; Smith et al., 2003…)
  - <u>Experimental</u>
    (Kalish et al, 2007; Kirby et al, 2008; Sanborn & Griffiths, 2008…)

Dynamical systems: Niyogi (2006)

- DS models: ∞ **learners per** generation
  (Komarova et al., 2002; Mitchener, 2003; Niyogi & Berwick, 1995 *et seq*; Niyogi, 2006; Yang, 2003…)

- Learners in gen. *i* (G_t) learn from data drawn from teachers in G_{t-1}

- Must specify:
  – Learning algorithm
  – Network structure
  – Assumptions about communication



Learning algorithm: Map from examples to øj or ø

Interlude: Dynamical systems

- System state at time *t*: α_t
  - Typically continuous, e.g. $\alpha_t \in [0, 1]$

- Rule for going from *t* to *t+1*: evolution equation

$$\alpha_{t+1} = f(\alpha_t)$$

- Fixed point :

$$\alpha_* \text{ s.t. } f(\alpha_*) = \alpha_*$$

- Fixed points can be
  - Stable: system returns to α_* when perturbed from it
  - Unstable: system doesn't `` ``

- Stable if $|f'(x^*)| < 1$, unstable otherwise

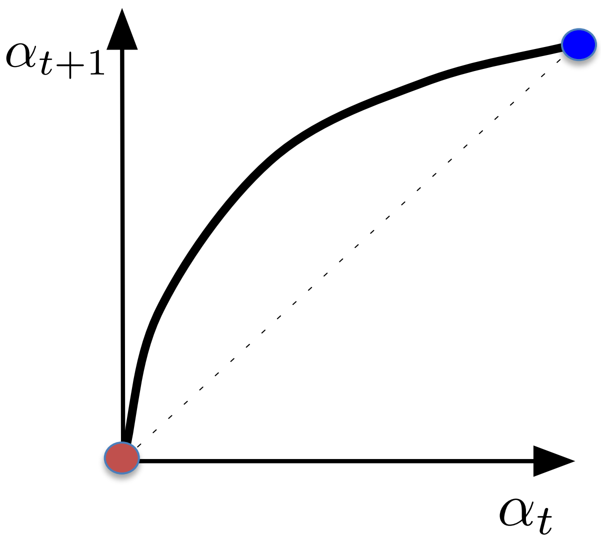
  - Slope of evolution equation

- Bifurcation: Change in the number or stability of FPs change as system parameters are varied

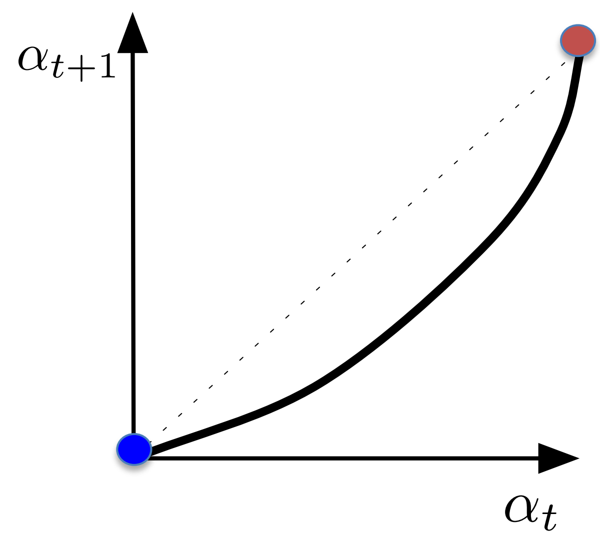
- Goals of a DS analysis:
  1. For a given evolution equation, find FPs & stabilities
  2. Determine bifurcations as parameters are varied

Example

$$\alpha_{t+1} = a\alpha_t^2 + (1-a)\alpha_t$$



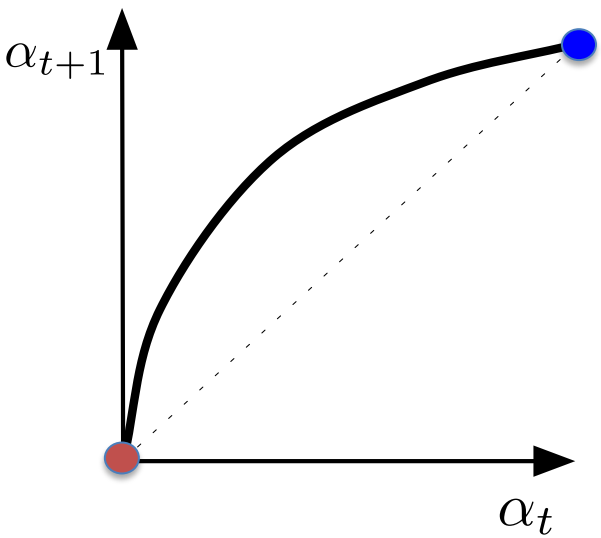
a < 0
unstable FP at 0
stable FP at 1

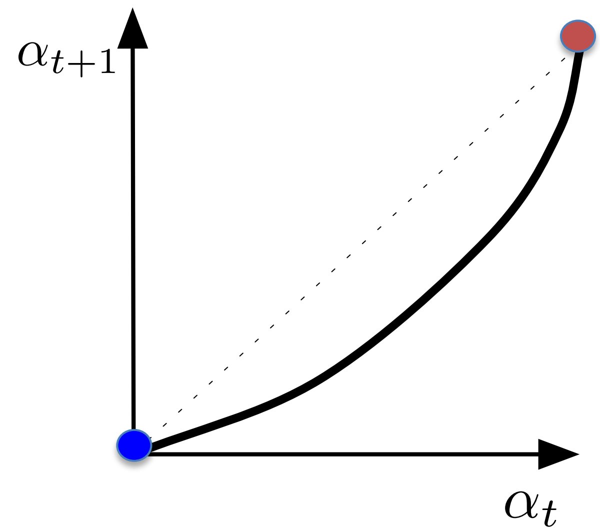


a > 0
stable FP at 0
unstable FP at 1

Example

$$\alpha_{t+1} = a\alpha_t^2 + (1-a)\alpha_t$$





bifurcation at a=0

$a < 0$
unstable FP at 0
stable FP at 1

$a > 0$
stable FP at 0
unstable FP at 1

- Because each generation is infinite, evolution of system state (α_t) is deterministic $\Rightarrow$ a dynamical system *f*

- Goals:
  - Determine FPs of *f* , bifurcation structure as parameters changed

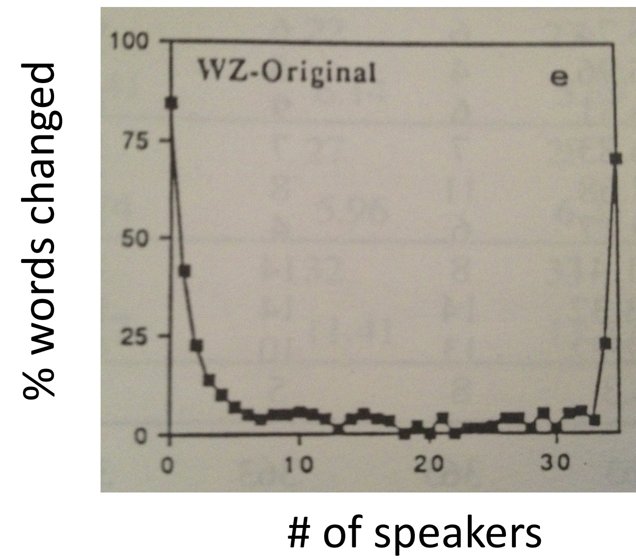
- Because each generation is infinite, evolution of system state (α_t) is deterministic $\Rightarrow$ a dynamical system *f*

- Goals:
  - Determine FPs of *f*, bifurcation structure as parameters changed

- Variation/change interpretation?
  - What differs from IL?

Shen (1997): Merger in Wenzhounese

- Unconditioned merger: /$ø^j$/ → [ø]
- Data: 0/1 (unchanged/changed)
  - 363 informants (!), ages 17-75
  - 35 *$ø^j$ words
- Phonetic motivation
- Results
  - Apparent-time change: S-shaped curves
  - Some lexical diffusion (Wang, 1969)
  - But, words tend to change together



Niyogi (2006) models

- Learning pronunciation of <u>one</u> $*ø^j$ word

- Dimensions:
  - \# of parents: 2, or ∞
  - Error in production/perception: None vs. some
  - Result of learning: categorical (0/1) or prob. (*p*)

- Correspond to different assumptions about…

Niyogi (2006) models

- Learning pronunciation of <u>one</u> $*ø^j$ word

- Dimensions:
  - – # of parents: 2, or ∞
  - – Error in production/perception: None vs. some
  - – Result of learning: categorical (0/1) or prob. (p)

today

- Correspond to different assumptions about…

Model 1

- Network: Child draws *N* words from <u>all</u> teachers in G_{t-1}

- Knowledge state: 0/1

- Algorithm: Choose [$ø^j$] if heard more than K times

- System state: α_t
  - Percentage of G_t with [$ø^j$],

- For 1 child in G_{t+1}

$$P(\text{hear ø}^{\text{j}}\ k \text{ times}) = \binom{N}{k} (\alpha_t)^k (1-\alpha_t)^{N-k}$$

$$P(\text{choose ø}^{\text{j}}) = \sum_{k=N/2}^{N} \binom{N}{k} (\alpha_t)^k (1-\alpha_t)^{N-k}$$

- For 1 child in G_{t+1}

$$P(\text{hear ø}^{\text{j}}\ k \text{ times}) = \binom{N}{k} (\alpha_t)^k (1 - \alpha_t)^{N-k}$$

$$P(\text{choose ø}\underline{^{\text{j}}}) = \sum_{k=N/2}^{N} \binom{N}{k} (\alpha_t)^k (1 - \alpha_t)^{N-k}$$

- Evolution equation:

$$\alpha_{t+1} = \sum_{k=N/2}^{N} \binom{N}{k} (\alpha_t)^k (1 - \alpha_t)^{N-k}$$

- Fixed points:
  - Stable: 0 and 1
  - Unstable: One between 0 and 1
- No bifurcations
  - No dependence on threshold K
- Interpretation
  - 100% or 0% [ø^{j}] are stable
  - [ø^{j}] $\sim$ [ø] variation possible, but unstable

Model 2: Categorical, 2 parents

- Same as model 1, except child draws examples from 2 teachers ("parents")
- Draw equally from each parent

Model 2: Categorical, 2 parents

- Same as model 1, except child draws examples from 2 teachers ("parents")
- Draw equally from each parent

- Fixed points:
  - Depends on constant *c* which increases with *K*
  - $c < 0.5$: 0% [ø^j] stable, 100% [ø^j] unstable
  - $c > 0.5$: 0% [ø^j] unstable, 100% [ø^j] stable
  - Bifurcation at $c = 0.5$, when $K=N/2$

Model 3: Model 1 + noise

- Model 1, plus <u>asymmetric mistransmission</u>
  - Every token of [$ø^j$] misheard with prob. ε
  - Every token of [ø] heard correctly

Model 3: Model 1 + noise

- Model 1, plus <u>asymmetric mistransmission</u>
  - Every token of [øʲ] misheard with prob. ε
  - Every token of [ø] heard correctly

- Critical value *C*, such that:
  - ε< C: Two stable FPs: 0% [øʲ] and k% [øʲ] (for some k ≤ 100)
  - ε> C: Only 0% [øʲ] FP is stable

- Bifurcation at ε= C
  - Critical value depends on *N*

- Interpretation
- Relationship to actuation

N 2006 models: Discussion

- Network structure
  - Big effect (Model 2 vs. Model 1)
- Assumptions about communication
  - Asymmetric noise
- Interpretation in terms of variation and change?

IL and DS models: Discussion

- Broad results
  - Can say exactly what long-term dynamics are from synchronic assumptions (c.f. simulation)
  - Some pieces matter a lot, others don't

- Dynamics: Linear (IL), nonlinear (DS)
  - Nonlinear only: Multiple stable states, bifurcation
  - In general, make different predictions (Dediu, 2009; Niyogi & Berwick, 2009; Smith, 2009)
  - DS more general, realistic (?)

- But:
  - IL feasibility in lab is important
  - ∞ **generation size clearly also** wrong.
  - Maybe IL OK as an approximation, in some sense?

- Need empirical evidence (historical):
  - Sudden change once a parameter (e.g. frequency) passes a threshold?
  - Multiple stable states, for same parameter values?