A model of

Language variation & change:

An age-structured population with prediction-driven instability

by Garrett Mitchener, *MitchenerG@cofc.edu*
College of Charleston

For *Applications of Analysis to Mathematical Biology*
Duke University, Durham, NC
May 21–23, 2007

Abstract. Children learn language accurately and instinctively, yet languages still change. Once a change begins, it tends to run to completion. I will develop a mathematical model of a population that can switch between two languages in several stages. A one-dimensional ODE serves as a base: Its fixed points represent populations dominated by one language or the other. The addition of stochastic noise allows for the possibility of spontaneous change from one to the other, but the time scale is too long. A two-dimensional model is required: The addition of social structure in the form of age groups leads to the desired behavior. *Supported by NSF grant 0734783.*

Background: Language Change

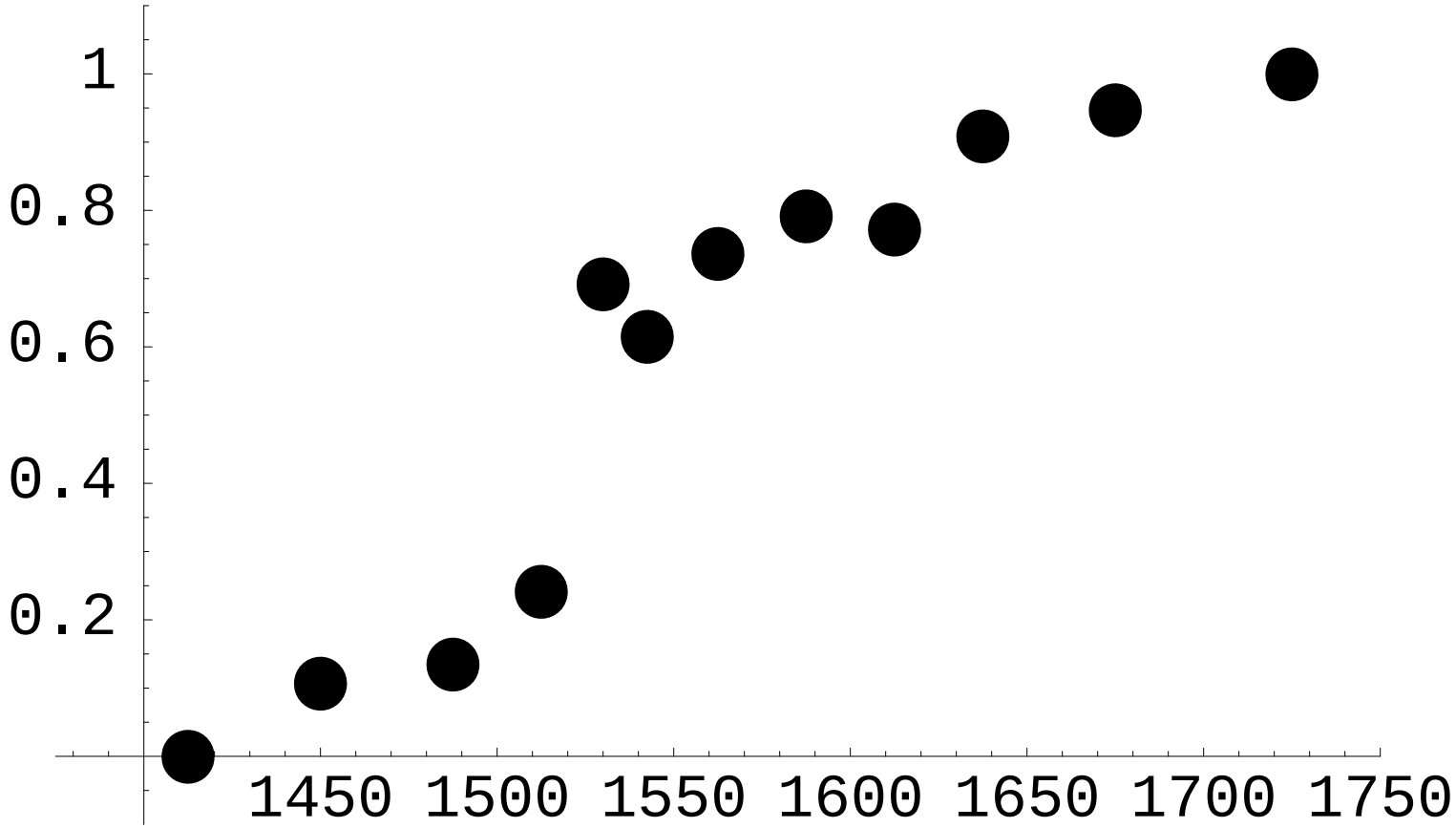
Verb raising
(Old & Middle English)

- I know the muffin man.
- Know you the muffin man?
- Know you not the muffin man?
- I know not the muffin man.

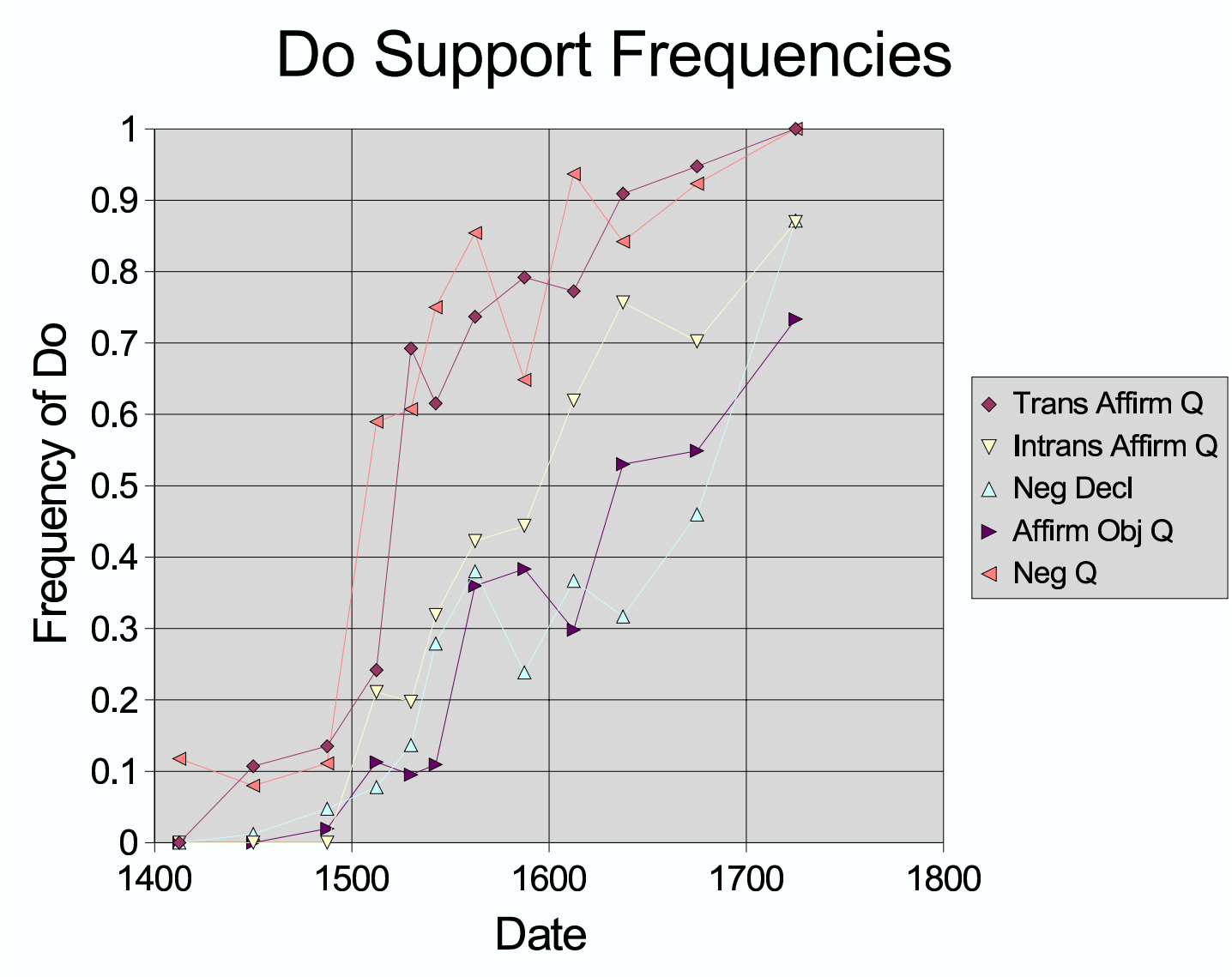
***Do*-support**
(Late Middle & Modern English)

- I know the muffin man.
- Do you know the muffin man?
- Do you not know the muffin man?
- I don't know the muffin man.

The syntax of English verbs has changed over the centuries. According to manuscript data, the transition from a verb raising grammar to a *do*-support grammar was smooth and sigmoid-shaped. Individual manuscripts use a mixture of both constructions, so during the transition, an individual's spoken grammar was a mixture of both idealized grammars. Observe that the change is monotonic: it begins and runs to completion without turning back, a feature typical of language changes [Yang, 2002].

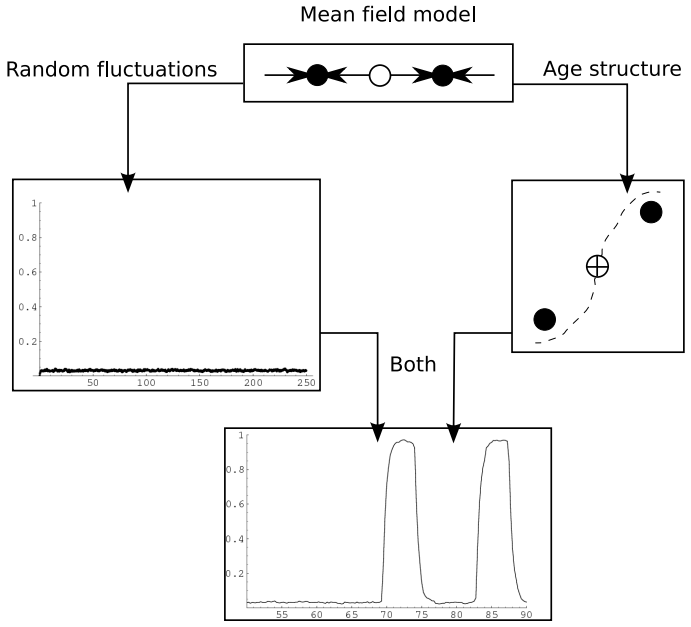


Change over time of the fraction of transitive affirmative questions using *do*-support. (As in "Who know you?" vs. "Who do you know?") Data from manuscripts collected by Ellegård [1953], analyzed by Kroch [1989].



More data from Ellegård [1953], Kroch [1989].

Road map



We will develop a model of spontaneous language change with the properties observed in the history of English:

- Simplification: Two idealized grammars
- Individuals speak using mixtures of the two idealized grammars
- Two nearly steady states representing dominance by each of the two grammars
- Spontaneous transitions between them in both directions

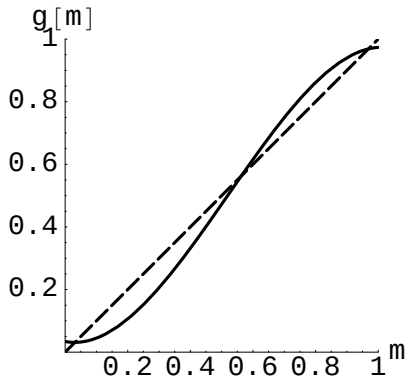
We will begin with a simple dynamical system on an interval with two stable fixed points separated by an unstable fixed point. From there we will add random fluctuations and an added dimension through social structure. Neither of these additions alone suffices. However, a combination of the two gives a model with qualitatively correct behavior.

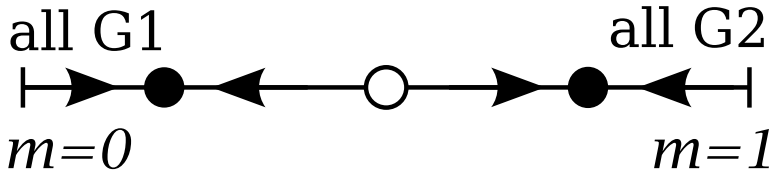
The Mean Field Foundation

We assume that there are two idealized grammars G_1 and G_2. The population is large and well mixed. Children learn from everyone—their speech is a function of the mean usage rate of G_2. Therefore, the population may be represented by its mean usage rate

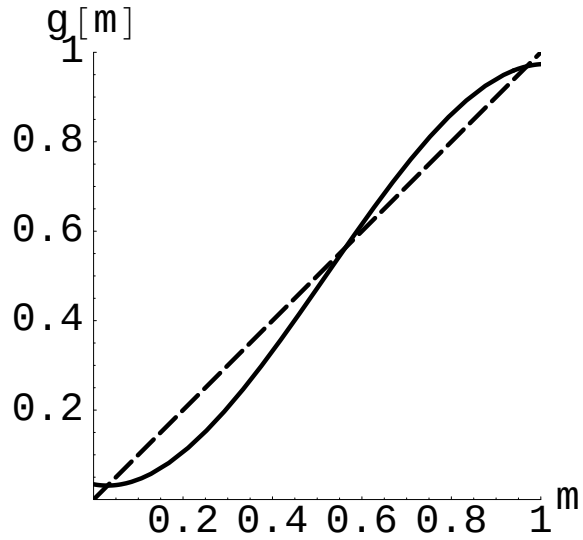
$$m' = \beta \, (\underbrace{g(m)}_{\text{birth \& learning}} - \underbrace{m}_{\text{death}}) \tag{1}$$

- $m(t)$ = mean usage rate of G_2 at time t, $0 \leq m \leq 1$
- β = birth & death rate
- $g(m)$ = learning function = mean usage rate of children after acquiring language from a population with usage rate m





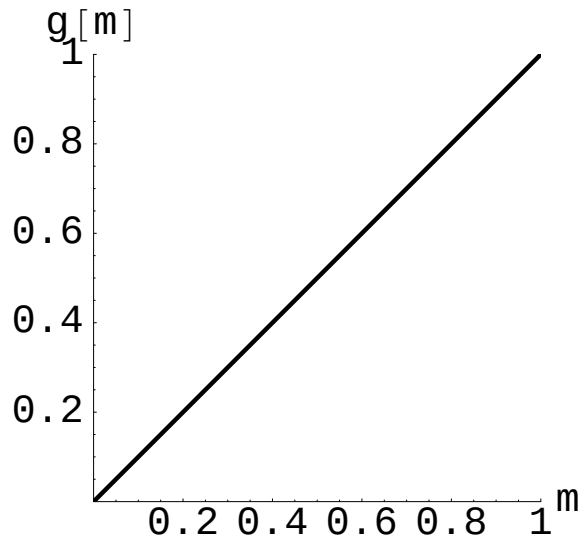
In most situations, languages prefer to use one grammatical alternative most of the time. Therefore, we want m to have stable fixed points near 0 and 1, representing populations that prefer G_1 or G_2, respectively. That constrains g to have the following form:



Intuitively: If G_1 dominates, then children should prefer G_1. (Similarly for G_2.) Children will also amplify that preference and use the dominant grammar even more than data suggests. These assumptions give the shape of $g(m)$ near $m = 0$ and $m = 1$, and the remainder of the sigmoid comes from joining these bits smoothly.

Perfect learning

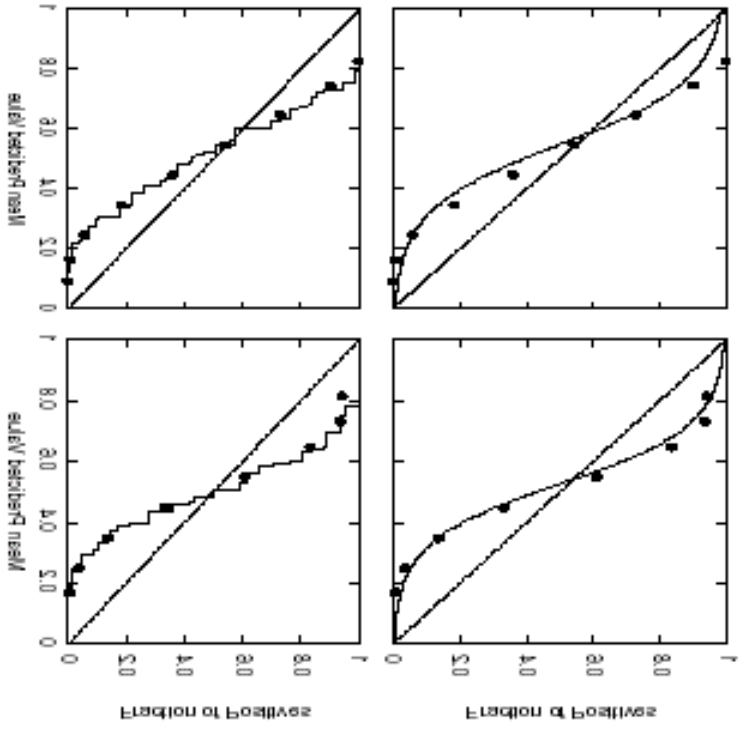
If learning were perfect, then $g(m)$ would look like this:



But then $m' = 0$ and all states are fixed points. This would be appropriate for modeling cases where any mixture of the two grammars is maintained indefinitely, but can be shifted by an external influence.

Statistical Learning Theory

Abstractly, the learning problem is to estimate $\mathbb{P}(\text{use } G_2)$ from data consisting of sample sentences that can be parsed only by G_1 or only by G_2, or both. Oddly, many probability estimating algorithms from statistical learning theory (SLT) have exactly the wrong shape.



Calibration plots for some popular SLT algorithms applied to problems where they must estimate the a probability between 0 and 1 from data. These pictures are from Caruana and Niculescu-Mizil [2005]. They have been rotated to match my axis convention, where the horizontal axis is the input probability used to generate the data (analogous to m), and the vertical axis is the probability estimated by the algorithm (analogous to $g(m)$).

Intuitively: SLT algorithms for probability estimation are typically based on binary classifiers, in which points are labeled $+1$ or -1 depending on which side of a decision boundary the fall on. In trying to predict a point's label, classifiers are most uncertain near the decision boundary, leading to a flat spot in the middle of their calibration curves and a reverse-sigmoid shape.

Probability estimators such as these can be post-processed with additional training data to eliminate their inherent reverse-sigmoid bias.

The lesson here is that human language acquisition is doing something quite different from these SLT algorithms.

A Stochastic Model

From the deterministic dynamical system (1) we deduced the rough shape of the learning algorithm, but such an ODE cannot represent spontaneous change from speech dominated by G_1 to speech dominated by G_2 and back. So, we add random fluctuations. We begin with a Markov chain and go to the limit of an infinite population, similar to the Wright-Fisher model of population genetics [Durrett, 1996, Ethier and Kurtz, 1986].

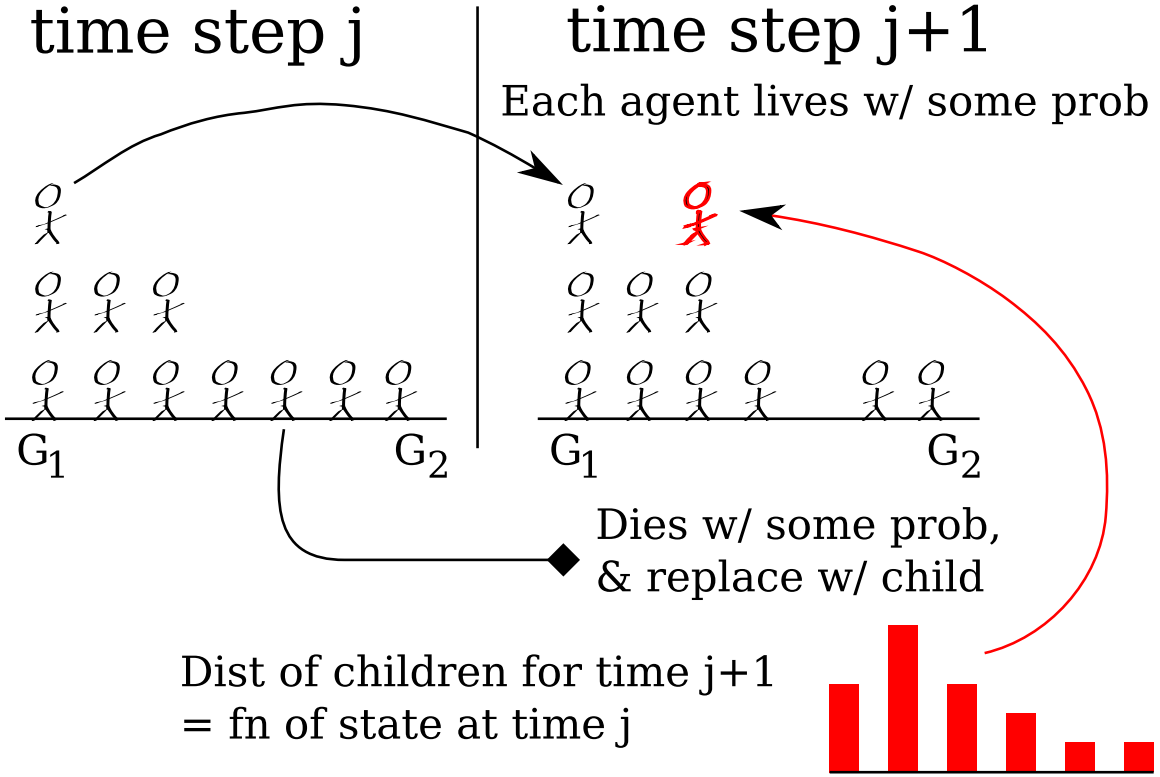
- We discretize speech by assuming $K+1$ types of agents, labeled $0, \ldots, K$

$$\mathbb{P}\left(\text{sentence from agent of type } k \text{ uses } G_2\right) = \frac{k}{K}$$

- The population state is a vector Z where

$$Z_k(j) = \#\text{ speakers of type } k \text{ at time step } j.$$

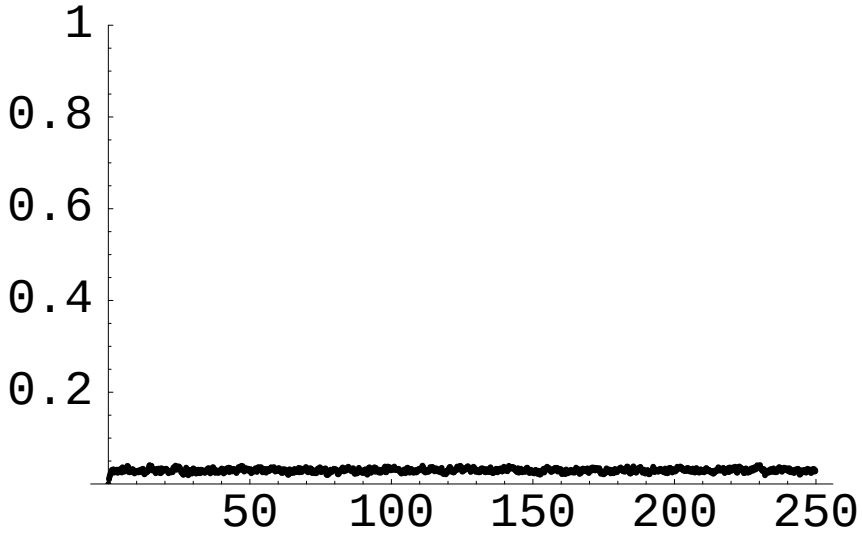
- The total population $N = \sum_k Z_k(j)$ is fixed.
- Dividing the state vector Z by N gives the speech distribution vector Y where $Y(j) = Z(j)/N$.
- Each discrete time step represents an elapsed time of $h = 1/N$.



The transition from one time step to the next is as follows.

- For each agent
  - Keep the agent with probability $1 - \frac{\beta}{N} = 1 - \beta\, h$
  - *Death & birth.* Replace the agent with probability $\frac{\beta}{N} = \beta\, h$ using the distribution from the learning function Q:

$$\begin{aligned} Q_k(j) &= \mathbb{P}\,(\text{child learns to use } G_2 \text{ at rate } k/K) \\ &= \text{sigmoid function of mean of } Y(j) \end{aligned}$$



Mean usage rate of G_2 as a function of time for a trajectory of the Markov chain.

Trajectories for this Markov chain hover around one of two equilibrium states: one with mean near 0 and one with mean near 1. These are analogous to the stable fixed points from the ODE. Transitions from one to the other are extremely rare. A population in an intermediate state is just as likely to go to one as to the other, so once begun, a change is fairly likely to reverse itself instead of running to completion.

Why is the change so rare under this model?

We may approximate the Markov chain by a stochastic differential equation (SDE) by letting the population size N go to infinity. In the simplest case $K = 1$, we find

$$dX_t = \beta\,(Q(X_t) - X_t)\,dt + \sqrt{X_t(1 - X_t)}\,dB_t \tag{2}$$

where X_t = fraction of the population that uses G_2 exclusively. We may then solve a differential equation for the stationary distribution of X.

How to derive the SDE

By letting the population size N go to infinity, we derive a system of Itô stochastic differential equations for the distribution vector $X = \lim Y$:

$$Y(j) = \frac{Z(j)}{N} \text{ and let } N \to \infty, \text{ timestep } h = \frac{1}{N} \to 0$$

Infinitesimal drift:

$$\mathbb{E}\left(\frac{Y(j+1) - Y(j)}{h}\middle| Y(j)\right) = \beta\left(Q(j) - Y(j)\right)$$

Infinitesimal variance of type k:

$$\text{Var}\left(\frac{Y_k(j+1)}{\sqrt{h}}\middle| Y(j)\right) = P - P^2 \to Y_k(j) - Y_k(j)^2$$

where

$$P = \frac{\beta}{N} Q_k(j) + \left(1 - \frac{\beta}{N}\right) Y_k(j)$$

Infinitesimal covariances ≈ 0:

$$\operatorname{Cov}\left(Y_k(j+1), Y_r(j+1) \,|Y(j)\,\right) = O\left(h^2\right)$$

The resulting system of SDEs is

$$\begin{aligned} dX_k(t) &= \beta\left(Q_k(X(t)) - X_k(t)\right) dt + \sqrt{X_k(t)(1 - X_k(t))}\, dB_k(t) \\ &\qquad \text{where } k = 1, \ldots, K; \\ X_0(t) &= 1 - X_1(t) - \cdots - X_K(t) \end{aligned} \tag{3}$$

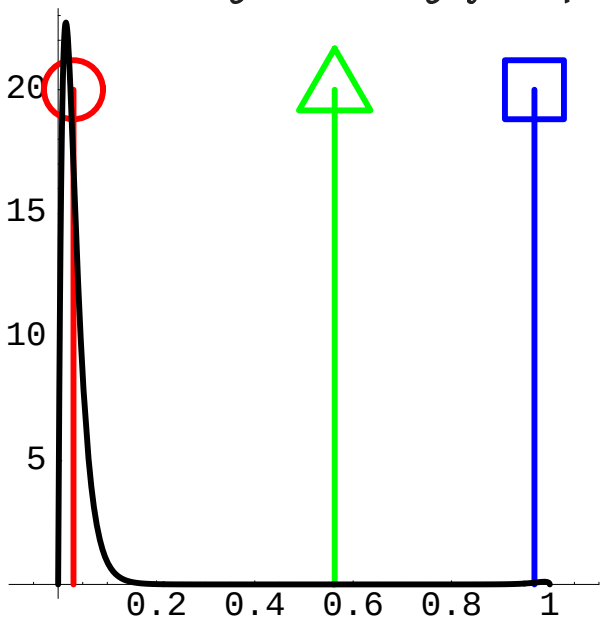
We focus on a simple case, $K = 1$:

- $X_0(t) =$ usage rate of $G_1 = 1 - X_1(t)$
- $X_1(t) =$ usage rate of G_2
- To simplify notation, let $X_t = X_1(t)$, $X_t \in [0, 1]$

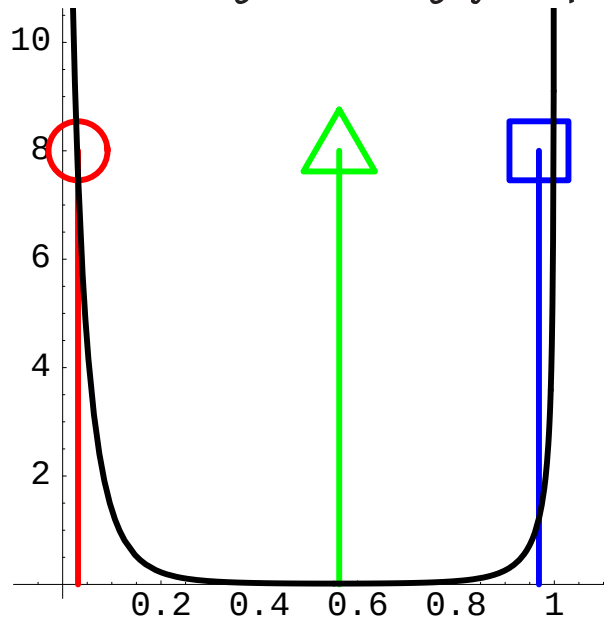
$$dX_t = \underbrace{\beta\,(Q(X_t) - X_t)}_{b(X_t)}\,dt + \underbrace{\sqrt{X_t(1 - X_t)}}_{\sigma(X_t)}\,dB_t \tag{4}$$

- $a(x) = \sigma(x)^2 = x(1 - x) =$ infinitesimal variance
- $Q(x) =$ sigmoid learning function

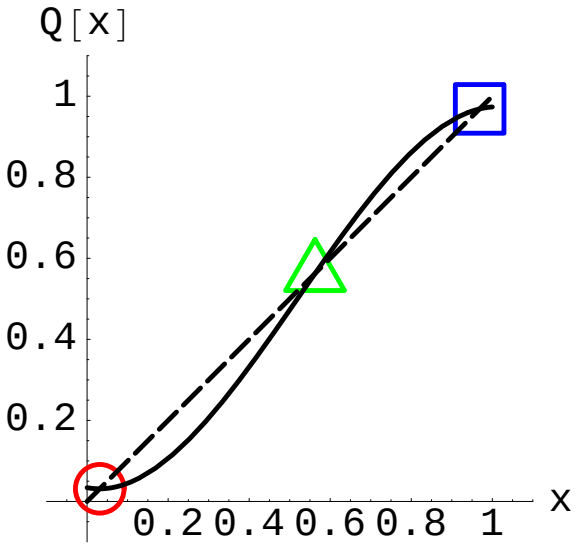
Stationary density for $\beta = 30$



Stationary density for $\beta = 10$



Learning function



The drift term from the SDE (4) can be considered alone as an ODE of the form in (1); it has two stable fixed points (marked ◯ and □) separated by an unstable fixed point (marked △). As shown, the stationary density of X_t is concentrated around the two stable fixed points, particularly for large β. In between ◯ and □, the density is exponentially small, so it is extremely rare for X_t to be found an intermediate state.

Solving for the stationary density

- $p(t,x,y)$ = transition density

$$\mathbb{P}\left(X_t \in A \,|X_0 = x\right) = \int_A p(t,x,y)\,dy$$

- The transition density obeys the forward Kolmogorov PDE (also known as the Fokker-Planck PDE):

$$\frac{\partial}{\partial t}p(t,x,y) = \frac{\partial}{\partial y}\left(\underbrace{\overbrace{\frac{1}{2}\frac{\partial}{\partial y}\big(a(y)p(t,x,y)\big)}^{\text{diffusion}} - \overbrace{b(y)p(t,x,y)}^{\text{transport}}}_{S=\text{probability current}}\right) \tag{5}$$

To find the equilibrium density $p(y)$, we seek a solution that does not depend on the initial state x or the time t. Those assumptions reduce the PDE to an ODE:

$$0 = S'(y) = \frac{1}{2}\big(a(y)p(y)\big)'' - \big(b(y)p(y)\big)' \tag{6}$$

The solution for $p(y)$ is straightforward because (6) is the derivative of a first-order linear ODE. One constant of integration is fixed at zero because of no-flux boundary conditions. The constraint that p must have total mass 1 determines the other constant of integration C_2.

$$p(y) = C_2 \frac{1}{\mu(y)} \text{ where } \mu = \exp\left(\int_{y_0}^{y} \left(\frac{\frac{1}{2}a'(s) - b(s)}{\frac{1}{2}a(s)}\right) ds\right) \tag{7}$$

In addition, the solution has Frobenius series about 0 and 1.

$$p(y) = y^{\alpha_0} \times \text{ analytic in } y$$

where

$$\alpha_0 = -1 + 2b(0) = -1 + 2\beta Q(0) > -1$$

$$p(y) = (1-y)^{\alpha_1} \times \text{ analytic in } 1-y$$

where

$$\alpha_1 = -1 - 2b(1) = -1 - 2\beta(Q(1) - 1) > -1$$

Boundary behavior and Feller's test

$$dX_t = \beta\,(Q(X_t) - X_t)\,dt + \sqrt{X_t(1-X_t)}\,dB_t$$

Feller theory [Durrett, 1996] is about determining whether a stochastic process on an interval can hit a boundary point, and whether it can escape from a boundary point. The tests are based on whether particular integrals converge. For this SDE, the results of the test are

- ❄ Large $\beta \implies \alpha_0 \geq 0, \alpha_1 \geq 0$:
  Entrance boundaries = can get out & can't get in
- ❄ Small $\beta \implies -1 < \alpha_0 < 0, -1 < \alpha_1 < 0$:
  Regular boundaries = can get out & can get in

The results of Feller's test agree with whether the stationary density has fractional order poles or zeros at the endpoints: The process can always escape from the boundary points, which means neither grammar ever goes permanently extinct. The process can hit a boundary point exactly when the stationary density has a pole there.

Integrals for Feller's test

- Natural scale ϕ

$$s(x) = \phi'(x) = \exp \int_{y_0}^{y} -\frac{b(z)}{\frac{1}{2}a(z)} dz = \frac{\mu(x)}{a(x)}$$

- Speed measure m

$$m(x) = \frac{1}{\phi'(x)a(x)} = \frac{1}{\mu(x)}$$

- Near 0:

$$I(y) = \int_0^y \int_0^z s(w)\, dw\, m(z)\, dz = \text{finite if can get in}$$

$$J(y) = \int_0^y \int_0^z m(w)\, dw\, s(z)\, dz = \text{finite if can get out}$$

Example: compute J for boundary point 0

- $\int_0^z m(w)dw = w^{\alpha_0+1} \times$ analytic
- $\int_0^z m(w)dw\ s(z) =$ analytic
- $J = \int_0^y \int_0^z m(w)dw\ s(z)dz =$ finite
- Can always escape from the boundary at 0

- $a(w) = w - w^2$
- $\mu(w) = w^{-\alpha_0} \times$ analytic
- $s(w) = w^{-\alpha_0-1} \times$ analytic
- $m(w) = w^{\alpha_0} \times$ analytic

Add a dimension: Age structure

The SDE (3) adds random fluctuations to the ODE (1) but they are insufficient to drive the population across the bottleneck between steady states. For the population to shift from G_1 to G_2 without turning back, the dynamics need some sort of momentum or trendiness, which requires a second dimension.

According to sociolinguistics [Aitchinson, 1987, Labov, 1994] ongoing language change is reflected in social variation (economic class, native vs. outsider, men vs. women, conversational vs. formal). Speakers alter their speech consciously and subconsciously to fit in.

ODE with age structure

We add age structure by splitting the population into a young group (parents) and an old group (grandparents). In this revised model, children can hear differences in the two age groups' speech patterns, and predict where the population is going. They base their speech on the prediction, the intuition being that they don't want to sound outdated.

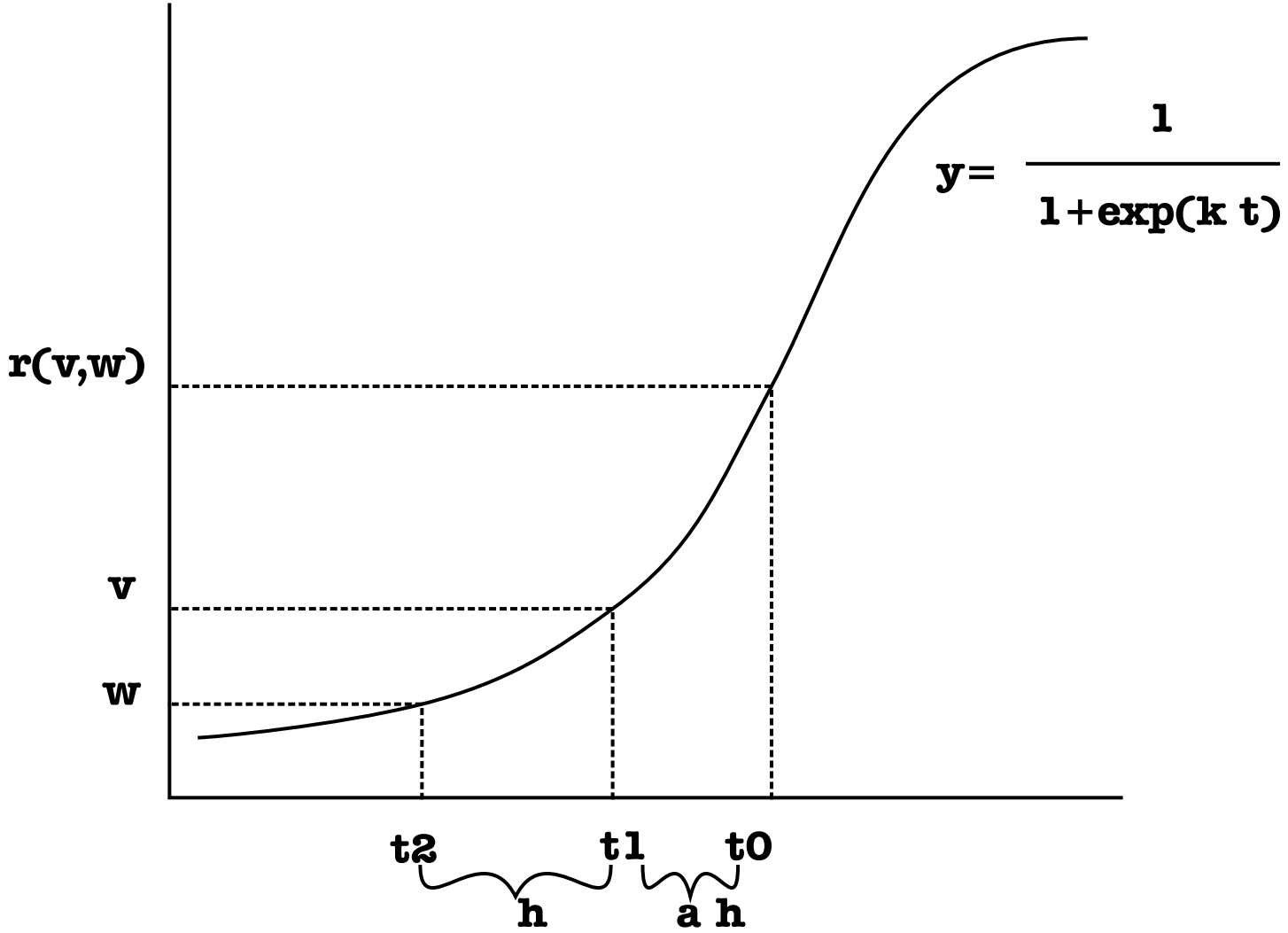
- Still two idealized grammars G_1 and G_2
- $v =$ mean usage rate of G_2 in the young group, $0 \leq v \leq 1$
- $w =$ mean usage rate of G_2 in the old group, $0 \leq v \leq 1$
- Death removes agents from the older group, and age shifts agents from the young group to the old group

$$w' = \beta\,(v - w)$$

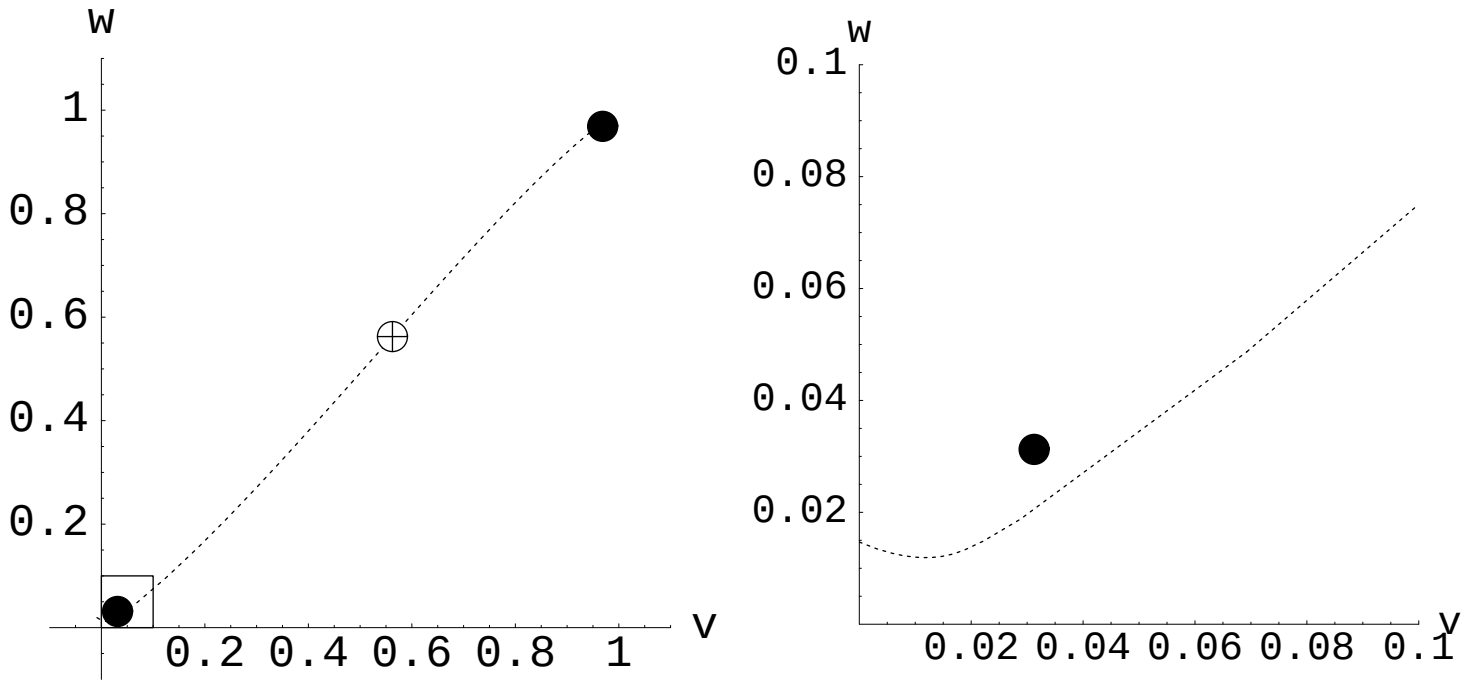
- We define $r(v, w)$ to be a function that predicts the usage rate for the next generation given the usage rates of the young and old generations. The birth and learning term is modified from (1) to include r:

$$v' = \beta\left(g\big(r(v, w)\big) - v\right)$$

Definition of the prediction function $r(v, w)$



Given parameters k and a, the prediction function takes v and w, finds their pre-images t_1 and t_2 under a sigmoid function, then advances from t_1 by a fraction of the difference $t_1 - t_2$.



Left: Phase portrait in v and w. *Right:* Phase portrait, enlargement of the lower left corner, marked by a box in the left figure.

Even with age structure, the dynamics are simple. There are two stable fixed points (marked by dots) that represent populations dominated by G_1 (lower left) and G_2 (upper right). In the middle is a saddle point (marked $\oplus$) whose stable manifold (dotted line) forms the separatrix between the two basins of attraction. There is an important difference with respect to the one-dimensional model: The stable fixed points are *very* close to the separatrix. A disturbance can knock a population at equilibrium across the separatrix and eventually send it to the other fixed point.

Combined model: Stochastic with age structure

We now combine age structure with random fluctuations. The phase portrait from the age structure ODE suggests that this will give the desired behavior.

- We discretize speech by assuming $K+1$ types of agents, labeled $0, \ldots, K$

$$\mathbb{P}\,(\text{sentence from agent of type } k \text{ uses } G_2) = \frac{k}{K}$$

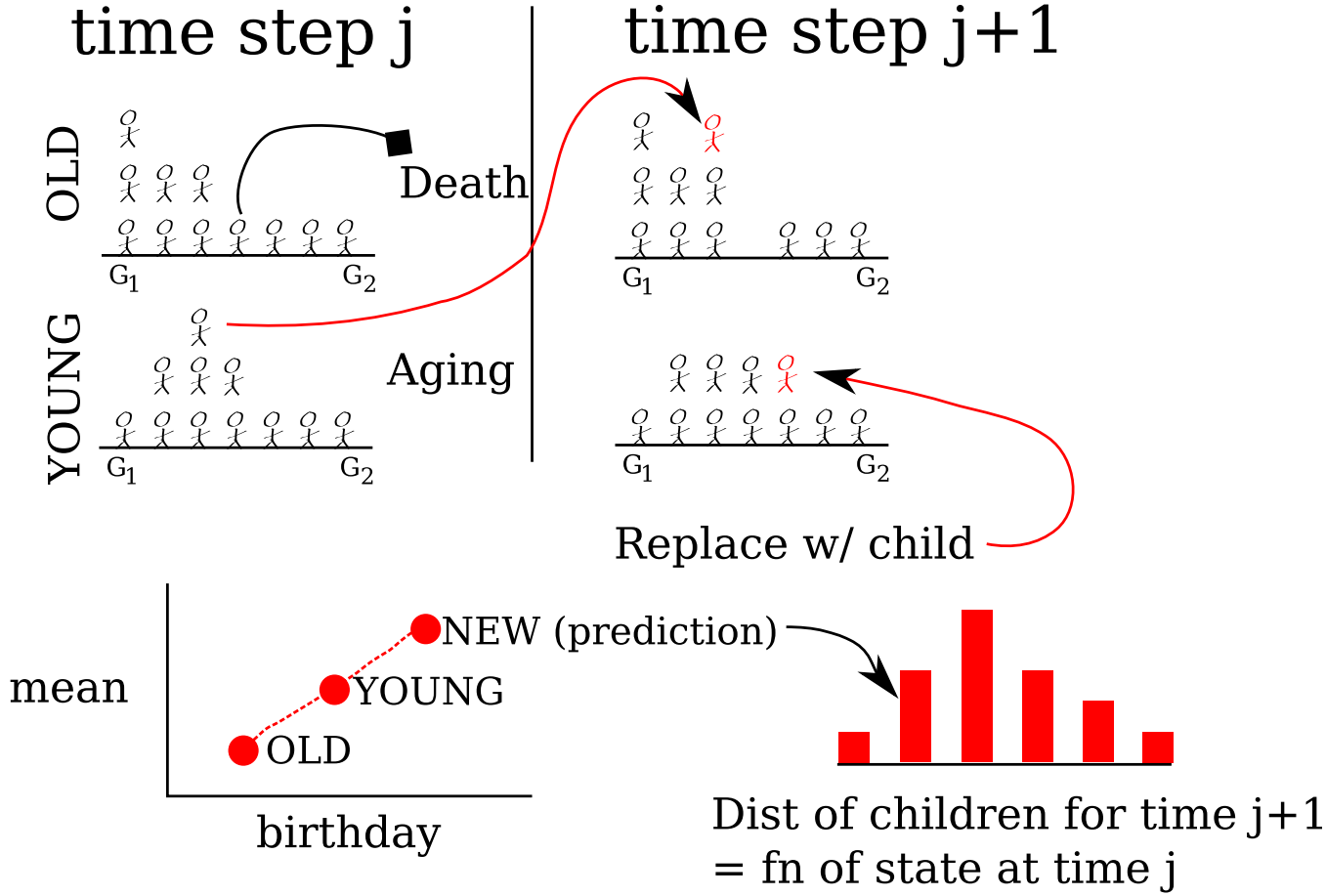
- The population state is a pair of vectors V and W where

$$V_k(j) = \#\text{ young speakers of type } k \text{ at time step } j$$
$$W_k(j) = \#\text{ old speakers of type } k \text{ at time step } j$$

- The total population N is fixed.

- Each discrete time step represents an elapsed time of $h = 1/N$.

Let's assume that there are social reasons to avoid sounding out-dated. If children detect a trend (ex: younger people make more use of G_2 than older people) they can predict whether one grammar will become more popular. They accelerate the trend by choosing to speak more like the younger generation.



The transition from one time step to the next is as follows.

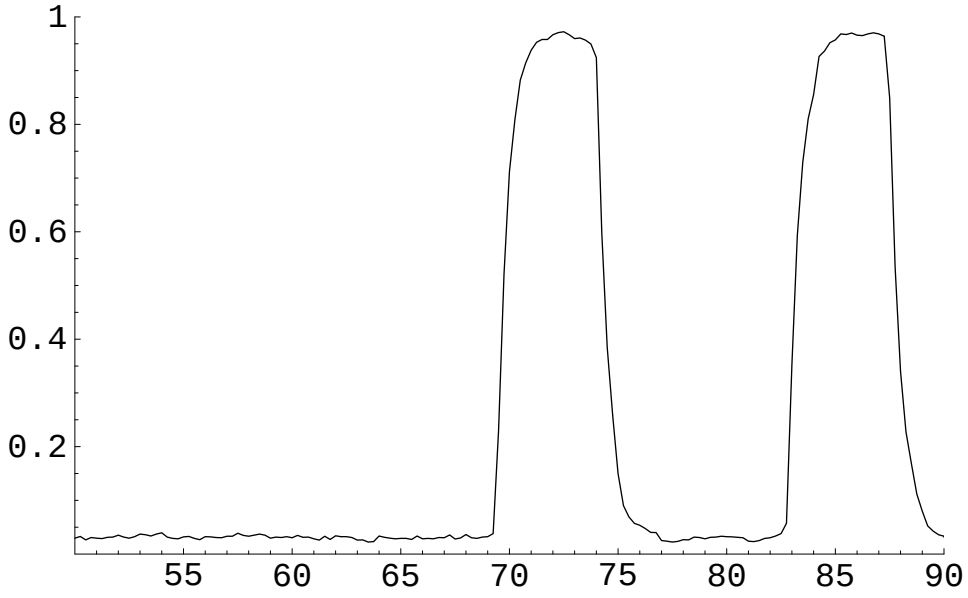
- For each old agent
  - Keep the agent with probability $1 - \frac{\beta}{N} = 1 - \beta\, h$
  - *Death & aging.* Replace the agent with probability $\frac{\beta}{N} = \beta\, h$ using the young group distribution given by V/N.

- For each young agent
  - Copy the agent with probability $1 - \frac{\beta}{N} = 1 - \beta\, h$
  - *Birth & aging.* Replace the agent with probability $\frac{\beta}{N} = \beta\, h$ using the distribution from the learning function Q:

$$\begin{aligned} Q_k(j) &= \mathbb{P}\,(\text{child learns to use } G_2 \text{ at rate } k/K) \\ &= \text{sigmoid function of } r(v, w) \end{aligned}$$

where $v = \text{mean}\, V(j)$ and $w = \text{mean}\, W(j)$

Sample trajectory



Sample trajectory of this Markov chain, mean of young plotted as a function of time. Notice how the population changes spontaneously from G_1 (low) to G_2 (high), and it can changes back. Once a change begins, it runs to completion. Changes also happen in a reasonable amount of time.

Discussion & conclusion

We began with the simplest possible mean-field ODE model of a population that can be dominated by either of two languages. Based on its properties, the learning function must be sigmoid.

We reformulated the model as a Markov chain and SDE. However, the random fluctuations of this model were too weak to drive the population from one language to another: such changes, though possible, are extremely rare, and they are generally not monotonic.

We reformulated the model as a mean-field ODE model with age structure, so that children would have enough information to identify trends in the population's language and base their speech on predictions.

The fourth formulation, a Markov chain with age structure, has the desired properties. Children can detect accidental trends in speech and predict how such trends will continue. Their choice of speech may amplify the trend, leading to language change. Thus in the case of an age-structured population with random fluctuations, language may change due to *prediction driven instability* from an otherwise stable state.

References

J. Aitchinson. *Words in the Mind: An Introduction to the Mental Lexicon*. Basil Blackwell, Oxford, 1987.

Rich Caruana and Alex Niculescu-Mizil. Predicting good probabilities with supervised learning. In *Proceedings of the American Meteorology Conference*, San Diego, 2005. URL `http://ams.confex.com/ams/pdfpapers/88928.pdf`.

Richard Durrett. *Stochastic Calculus: A Practical Introduction*. CRC Press, New York, 1996.

Alvar Ellegård. *The Auxiliary do: The Establishment and Regulation of Its Use in English*. Gothenburg Studies in English. Almqvist and Wiksell, 1953.

Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, New York, 1986.

Anthony Kroch. Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1:199–244, 1989.

William Labov. *Principles of Linguistic Change: Internal Factors*. Blackwell, Cambridge, MA, 1994.

Charles D. Yang. *Knowledge and Learning in Natural Language*. Oxford University Press, Oxford, 2002.