Constraint breeding during on-line incremental learning*

Elliott Moreton
University of North Carolina, Chapel Hill
moreton@unc.edu

Abstract

An evolutionary algorithm for simultaneously inducing and weighting phonological constraints (Winnow-MaxEnt-Subtree Breeder) is described, analyzed, and illustrated. Implementing weights as sub-population sizes, reproduction with selection executes a new variant of Winnow (Littlestone, 1988), which is shown to converge. A flexible constraint schema, based on the same prosodic and autosegmental trees used in representations, is described, together with algorithms for mutation and recombination (mating). The algorithm is applied to explaining abrupt learning curves, and predicts an empirical connection between abruptness and language-particularity.

1 Introduction

This paper aims to unite, within the framework of Harmonic Grammar (Legendre et al., 1990), two facts about phonological learning. One is that not all constraints can be innate; some are parochial and must be induced from language data (e.g., Prince and Smolensky 1993, 101), raising the question of how new constraints are induced. The other is that phonological learning can be abrupt, in the sense that a flat learning curve can later accelerate, both in nature (Smith, 1973; Macken and Barton, 1978; Vihman and Velleman, 1989; Barlow and Dinnsen, 1998; Levelt and van Oostendorp, 2007; Gerlach, 2010; Becker and Tessier, 2011; Guy, 2014) and in the lab (Moreton and Pertsova, 2016), raising the question of what is going on during the period of apparent stagnation .

The proposed answer to both is that constraint induction and constraint reweighting happen simultaneously via a single mechanism. Constraint weights are represented as sub-population sizes, i.e., the number of copies of a given constraint ("micro-constraints" in a "macro-constraint"). Error-driven reweighting of the macro-constraints happens when micro-constraints reproduce with fitness dependent on their contribution to error reduction. This process is shown to implement a modestly new incremental HG learning algorithm, Winnow-MaxEnt (§2), which is then analyzed (§§3, 4, 5). A flexible prosodic and Feature-Geometric constraint schema, the Subtree Schema(§6), is used to add mutation and recombination (§7), so that fitter constraint variants can evolve and rapidly supersede their predecessors, leading to abrupt changes in performance (§8). The paper ends with discussion (§9).

2 Weights as population sizes

The first step towards constraint breeding is to replace constraint weights with population sizes so that differential reproductive success implements up- or down-weighting. Without changing the har-

*The author is indebted to Brian Hsu, Katya Pertsova, and Jen Smith, Chris Wiesen of the Odum Institute at UNC-CH, and three anonymous SCiL reviewers for comments on previous drafts. The paper benefited from audience comments on portions of Sections 6 and 7 at the Workshop on Computational Modelling of Sound Pattern Acquisition (University of Alberta, February 14, 2010), at an MIT departmental colloquium (April 9, 2010), and at the Workshop on Grammar Induction (Cornell University, May 14, 2010). The research was supported in part by NSF BCS 1651105, "Inside phonological learning", to E. Moreton and K. Pertsova.

mony of any candidate, we can replace any constraint of weight w with k "micro-constraints", i.e., clones of that constraint, each with weight w/k. Across an entire grammar, we can fix a parameter ζ to be the quantum of harmony, and replace every ("macro-") constraint of weight w with a population of w/ζ microconstraints of fixed weight ζ.

Changes in weight of a macro-constraint result from changes in population size of a micro-constraint. When an error occurs, each micro-constraint produces an offspring with probability $(1+\epsilon)^d$, where d is the difference between the winner's and loser's score on that constraint and ϵ is the learning-rate parameter. If $(1+\epsilon)^d > k$ for some integer $k \geq 1$, the constraint produces k offspring with certainty, and another with probability $(1+\epsilon)^d - k$. The new generation replaces the current generation.

The weight update at the macro level is therefore not described by a Perceptron-like algorithm, in which weights change by a fixed absolute increment, implementing gradient ascent on log-likelihood (Rosenblatt, 1958; Sutton and Barto, 1981; Jäger, 2007; Boersma and Pater, 2016), but rather by one in which the increment is proportional to the weight, i.e., by a variant of Winnow-2 (Littlestone, 1988). This algorithm is analyzed in Sections 3–5 below.

The weights grow exponentially in the number of mistakes (Propositions 1 and 2, below), so a population explosion may threaten to overwhelm the learner's limited computational substrate. That problem can be addressed using weight decay: On each update (or on each trial), the learner can delete each micro-constraint with a fixed probability, causing the macro-constraint weights to decay by an amount proportional to their magnitude and slowing population growth. Alternatively, the decay rate can be adjusted dynamically, randomly deleting or duplicating micro-constraints to maintain a fixed total micro-constraint population size.

3 The Winnow-MaxEnt algorithm

Winnow-MaxEnt, the learning algorithm induced by the constraint-breeding algorithm, is similar to Winnow-2 (Littlestone, 1988), but with the following differences. Winnow-MaxEnt (1) models k-alternative forced choices rather than yes-no classification, (2) responds probabilistically rather than deterministically, and (3) supports non-binary constraint scores. The original Winnow-2 was first suggested as a possible HG learner by Magri (2013). This section describes the two-alternative Winnow-MaxEnt. Generalization to $k > 2$ and to negative constraints is discussed in Section 5 below.

Each of n constraints (macro-constraints) gives a non-negative score to any candidate. The algorithm sees only the score vectors, and so is equally applicable to pure phonotactic learning (where each candidate is a surface form and there are no faithfulness constraints) and to alternation learning (where each candidate is an input-output pair and there are faithfulness constraints). We write x_i for the score given by C_i to Candidate x. Each constraint C_i has weight $w_i > 0$, so that the state of the learner is described by the weight vector $\mathbf{w} = (w_1, \ldots, w_n)$.

Given the experimenter's intended winner x^+ and intended loser x^-, the learner chooses x^+ with a probability that depends on the harmonies of the candidates. The version of the model discussed here uses the Luce choice rule applied to the exponentiated harmonies:

$$\Pr(x^+ \mid x^+, x^-) = \frac{\exp(\sum_{i=1}^n x_i^+ w_i)}{\exp(\sum_{i=1}^n x_i^+ w_i) + \exp(\sum_{i=1}^n x_i^- w_i)} \quad (1)$$

This rule (Luce, 1959, 23) is an independently-justified model of human choice behavior across a wide range of domains (Bradley and Terry, 1952; Luce, 1977; Strauss, 1992; Macmillan and Creelman, 2004). Its use with exponentiated harmonies yields a conditional Maximum Entropy (MaxEnt) model (Goldwater and Johnson, 2003; Jäger, 2007; Hayes and Wilson, 2008). The losers ("negative evidence") may be presented to the learner explicitly in the form of a 2AFC experimental trial, or implicitly in the form of an internally-generated candidate set.

If x^+ is chosen, nothing changes. If x^- is chosen, the weights of winner-preferring constraints grow, and those of loser-preferring constraints shrink, according to the update rule

$$w_i' = w_i \alpha^{d_i} \quad (2)$$

where $\alpha = 1+\epsilon$ for some fixed learning-rate parameter $\epsilon > 0$, and $d_i = x_i^+ - x_i^-$.

4 Convergence of Winnow-MaxEnt

Winnow-MaxEnt is different enough from Winnow-2 that convergence cannot be assumed on the basis of Littlestone (1988)'s proof for Winnow-2, though ideas from that proof are useful here. In fact, since the probability of an error cannot be zero, Winnow-MaxEnt does not converge at all, in the sense of ceasing to make mistakes. However, we will see that the error rate can be made arbitrarily small.

4.1 Consequences of the update rule

The Propositions presented in this subsection are derived from the update rule (Equation 2) and do not depend on the response rule, the candidate-set size, or the sign of the marks awarded.

We proceed as usual (Novikoff, 1963) by first assuming that the target concept is representable in the learner, and then finding lower and upper bounds on a function of the weights in terms of the number of mistakes. Let D be the (multi-)set of candidate pairs used in the experiment. Each pair consists of an intended winner x^+ and an intended loser x^-. The same pair may occur multiple times, and not all possible pairs need occur. A candidate that is the positive member of one pair may be the negative member of another. Suppose that there exist nonnegative weights $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and a $\delta_{\boldsymbol{\mu}}$ such that for every candidate pair $(x^+, x^-) \in D$,

$$\sum_{i=1}^{n} \mu_i (x_i^+ - x_i^-) > \delta_{\boldsymbol{\mu}} > 0 \tag{3}$$

Proposition 1 (analogous to Littlestone (1988)'s Lemma 9)**.** *Let $W = \sum_{i=1}^{n} w_i$, and let a target concept satisfying Inequality 3 be given. Let $A_{\boldsymbol{\mu}} = \delta_{\boldsymbol{\mu}} / \sum_{i=1}^{n} \mu_i$. Then after the t-th update,*

$$\log W(t) > \min_i (\log w_i(0)) + t \cdot A_{\boldsymbol{\mu}} \log(1+\epsilon) \tag{4}$$

Proof. Taking the logarithm of Equation 2 yields $\log w_i' = \log w_i + (x_i^+ - x_i^-) \log \alpha$. Hence

$$\sum_{i=1}^{n} \mu_i \log w_i' = \sum_{i=1}^{n} \mu_i \log w_i + (\log \alpha) \sum_{i=1}^{n} \mu_i (x_i^+ - x_i^-) \tag{5}$$

Substituting from Equation 3 we have

$$\sum_{i=1}^{n} \mu_i \log w_i' > \sum_{i=1}^{n} \mu_i \log w_i + (\log \alpha) \cdot \delta_{\boldsymbol{\mu}} \tag{6}$$

and so after t updates,

$$\sum_{i=1}^{n} \mu_i \log w_i(t) > \sum_{i=1}^{n} \mu_i \log w_i(0) + (\log \alpha) \cdot \delta_{\boldsymbol{\mu}} \cdot t \tag{7}$$

Since all the μ_i's are nonnegative, the sum on the left doesn't get smaller if we replace all the weights with the largest weight, and the sum on the right doesn't get larger if we replace all the weights with the smallest weight. Let $i^* = \arg\max_i w_i(t)$ and $\hat{i} = \arg\min_i w_i(0)$; then

$$\log w_{i^*}(t) \sum_{k=1}^{n} \mu_k > \log w_{\hat{i}}(0) \sum_{k=1}^{n} \mu_k + (\log \alpha) \cdot \delta_{\boldsymbol{\mu}} \cdot t \tag{8}$$

Since the μ_i's are all nonnegative, we can divide through by their sum to conclude that

$$\log w_{i^*}(t) > \log w_{\hat{i}}(0) + t \cdot A_{\boldsymbol{\mu}} \cdot \log \alpha \tag{9}$$

Since $\log W(t) = \log \sum_{i=0}^{n} w_i(t) > \log w_{i^*}(t)$, the claim is proven. □

Proposition 2. *Let a target concept satisfying Inequality 3 be given. Let $\Sigma^+ = \sum_{i=1}^{n} x_i^+ w_i$ and $\Sigma^- = \sum_{i=1}^{n} x_i^- w_i$ for a given winner-loser pair (x^+, x^-). Then for any $\epsilon \leq 1/(d_{max} - 1)$,*

$$\log W(t) \leq \log W(0) + \epsilon \cdot \sum_{\tau=0}^{t-1} \frac{\Sigma^+(\tau) - \Sigma^-(\tau)}{W(\tau)} + t \cdot \frac{d_{max}^2 \epsilon^2}{1 - (d_{max} - 1)\epsilon} \tag{10}$$

Proof. From the update rule in Equation 2 plus the binomial theorem,

$$\begin{aligned} W' &= W + \sum_{i=1}^{n} (\alpha^{d_i} - 1) w_i \\ &= W + \sum_{i=1}^{n} (d_i \epsilon + O(\epsilon^2)) w_i \\ &= W + \epsilon(\Sigma^+ - \Sigma^-) + O(\epsilon^2) W \end{aligned} \tag{11}$$

To bound the $O(\epsilon^2)$ term explicitly, we rewrite Equation 11 as

$$W' = W + \sum_{i|d_i>0} (\alpha^{d_i} - 1) w_i + \sum_{i|d_i<0} (\alpha^{d_i} - 1) w_i \tag{12}$$

Since $\epsilon(\Sigma^+ - \Sigma^-) = \sum_{i=1}^{n} \epsilon d_i w_i$, we can rewrite that again as

$$W' = W + \epsilon(\Sigma^+ - \Sigma^-) + \underbrace{\sum_{i|d_i>0} (\alpha^{d_i} - 1 - \epsilon d_i) w_i}_{Y} + \underbrace{\sum_{i|d_i<0} (\alpha^{d_i} - 1 - \epsilon d_i) w_i}_{Z} \tag{13}$$

By Theorem 2 of Mitrinović (1970, p. 34),

$$(1+x)^n - 1 \leq \frac{nx}{1-(n-1)x} \tag{14}$$

for $n > 1$ and $-1 \leq x \leq 1/(n-1)$. This clearly also holds for $n = 0$ and $n = 1$ as long as $x \geq 0$. Hence for $d_i \geq 0$ and $\epsilon \leq 1/(d_{\max} - 1)$,

$$\alpha^{d_i} - 1 \leq \frac{d_i \epsilon}{1-(d_i-1)\epsilon} \tag{15}$$

with strict equality if $d_i = 0$ or $d_i = 1$. Therefore,

$$\begin{aligned} Y &\leq \sum_{i|d_i>0} d_i \epsilon \left(\frac{1}{1-(d_i-1)\epsilon} - 1 \right) w_i \\ &\leq \sum_{i|d_i>0} d_i \epsilon \left(\frac{(d_i-1)\epsilon}{1-(d_i-1)\epsilon} \right) w_i \end{aligned} \tag{16}$$

Likewise,

$$Z = \sum_{i|d_i<0} (\alpha^{-|d_i|} - 1 + \epsilon |d_i|) w_i \tag{17}$$

If $n, x \geq 0$, then by the binomial theorem, $(1+x)^n \geq 1 + nx$, so

$$(1+x)^{-n} - 1 = \frac{1}{(1+x)^n} - 1 \leq \frac{1}{1+nx} - 1 = \frac{-nx}{1+nx} \tag{18}$$

Hence,

$$Z \leq \sum_{i|d_i<0} \left(\frac{-|d_i|\epsilon}{1+|d_i|\epsilon} + |d_i|\epsilon \right) w_i \leq \sum_{i|d_i<0} \frac{(|d_i|\epsilon)^2}{1+|d_i|\epsilon} w_i \tag{19}$$

The sum $Y + Z$ is therefore bounded by

$$Y + Z \leq \sum_{i=0}^{n} \max \left\{ \begin{array}{c} \frac{|d_i|(|d_i|-1)\epsilon^2}{1-(|d_i|-1)\epsilon} \\ \frac{|d_i|^2\epsilon^2}{1+|d_i|\epsilon} \end{array} \right\} w_i \tag{20}$$

Combining the larger numerator and smaller denominator to get a fraction that is larger than either, we have

$$\begin{aligned} Y + Z &\leq \sum_{i=1}^{n} \frac{d_i^2 \epsilon^2}{1-(|d_i|-1)\epsilon} w_i \\ &\leq \frac{d_{\max}^2 \epsilon^2}{1-(d_{\max}-1)\epsilon} W \end{aligned} \tag{21}$$

Combining Inequalities 13 and 21 yields

$$W' \leq W \left(1 + \epsilon \frac{\Sigma^+ - \Sigma^-}{W} + \frac{d_{\max}^2 \epsilon^2}{1-(d_{\max}-1)\epsilon} \right) \tag{22}$$

Since $\log(1+x) \leq x$, we have

$$\log W' \leq \log W + \epsilon \frac{\Sigma^+ - \Sigma^-}{W} + \frac{d_{\max}^2 \epsilon^2}{1-(d_{\max}-1)\epsilon} \tag{23}$$

from which the proposition follows by summation from $\tau = 0$ to $\tau = t-1$. □

Proposition 3. *Let a target concept satisfying Inequality 3 be given, and let A_{max} be the least upper bound on $A_{\boldsymbol{\mu}}$ over all $\boldsymbol{\mu}$. Let $V = \log W(0) - \min_i(\log w_i(0))$, and let $a = (\Sigma^+ - \Sigma^-)/W$ for a given winner-loser pair (x^+, x^-). Then for any $\theta > 0$, there exist ϵ_θ and t_θ such that when Winnow-MaxEnt is run with $\epsilon = \epsilon_\theta$,*

$$\frac{1}{t} \cdot \sum_{\tau=0}^{t-1} a(\tau) \geq A_{max} - \theta \tag{24}$$

for all $t \geq t_\theta$.

Proof. Propositions 1 and 2 together imply that for all $t \geq 0$ and $\epsilon \leq 1/(d_{\max} - 1)$,

$$\epsilon \sum_{\tau=0}^{t-1} a(\tau) \geq -V + tA \log(1+\epsilon) - \frac{t d_{\max}^2 \epsilon^2}{1-(d_{\max}-1)\epsilon} \tag{25}$$

Since $\log 1 + x \geq x - x^2/2$,

$$\frac{1}{t} \cdot \sum_{\tau=0}^{t-1} a(\tau) \geq A - \frac{1}{2} A\epsilon - \frac{d_{\max}^2 \epsilon}{1-(d_{\max}-1)\epsilon} - \frac{V}{\epsilon t} \tag{26}$$

Sufficiently small ϵ and large t make the right-hand side as close to A as desired. □

To bound t_θ, we note that the remainder in Inequality 26 is bounded above by

$$\begin{aligned} f(\epsilon, t) &= \frac{1}{2} A\epsilon + \frac{d_{\max}^2 \epsilon}{1-(d_{\max}-1)\epsilon} + \frac{V}{\epsilon t} \\ &< \frac{1}{2} A\epsilon + \frac{d_{\max}^2 \epsilon}{1-d_{\max}\epsilon} + \frac{V}{\epsilon t} \end{aligned} \tag{27}$$

so that $f(\epsilon, t) < g(d_{\max}\epsilon, 1/t)$,

$$g(x, y) = Fx + d_{\max} \frac{x}{1-x} + \frac{Gy}{x} \tag{28}$$

where $F = A/2d_{\max}$ and $G = V d_{\max}$. Any pair (ϵ, t) that satisfies $g(d_{\max}\epsilon, 1/t) = \theta$ also satisfies

$f(\epsilon, t) < \theta$. Setting $g(x, y) = \theta$ and solving for y yields

$$y = h(x) = \frac{1}{G}\left(\theta x - Fx^2 - d_{\max}\frac{x^2}{1-x}\right) \quad (29)$$

for $x \in (0, 1)$. We want to choose $x(= d_{\max}\epsilon)$ so as to maximize $y(= 1/t)$. The function $h(x)$ is hard to maximize analytically, so instead we maximize a more tractable minorant $i(x)$ to bound the maximum of $h(x)$ below. Using the fact that $1/(1-x) \leq 1 + 2x$, $x \in [0, 1/2]$, we have

$$i(x) = \frac{1}{G}\left(\theta x - Fx^2 - d_{\max}(x^2 + 2x^3)\right) \quad (30)$$

Then $h(x) \geq i(x)$ for all $x \in (0, 1/2]$. Differentiation shows that $i(x)$ attains a maximum at

$$x_\theta = \frac{\sqrt{(d_{\max} + F)^2 + 6d_{\max}\theta} - (d_{\max} + F)}{6d_{\max}} \quad (31)$$

Whatever the global maximum of $h(x)$ might be, it is at least as big as $i(x_\theta)$, i.e. $\max_{x\in(0,\infty)} h(x) \geq \max_{x\in(0,1/2]} h(x) \geq h(x_\theta) \geq i(x_\theta)$, which is

$$i(x_\theta) = \frac{\left(T + U^2\right)\left(\sqrt{T + U^2} - U\right) - \frac{1}{2}TU}{54V} \quad (32)$$

where $T = 6\theta/d_{\max}$ and $U = 1 + A/2d_{\max}^2$. Therefore, there exist an $\epsilon_\theta = x_\theta/d_{\max}$ and a $t_\theta = 1/i(x_\theta)$ such that $f(\epsilon_\theta, t_\theta) \leq g(x_\theta, i(x_\theta)) = \theta$. Thus, $t_\theta = O(\theta^{-3/2})$. The smallest possible V occurs when all the initial weights are equal, in which case $V = \log n$ and $t_\theta = O(\log n)$; i.e., the time bound is not very sensitive to the number of constraints.

4.2 2AFC performance

This subsection addresses the question of how the bound on the relative harmony gap (Proposition 3) translates into a bound on 2AFC error probability. From Equation 1, the log-odds of choosing the correct candidate is

$$\text{log odds}(x^+ \mid x^+, x^-, \mathbf{w}) = \Sigma^+ - \Sigma^- = aW \quad (33)$$

where a is defined as in Proposition 3. The cumulative average log-odds of a correct response across all trials where an error actually occurred is therefore

$$L(t) = \frac{1}{t}\sum_{\tau=0}^{t-1} a(\tau)W(\tau) \quad (34)$$

where τ indexes errors as in Propositions 2 and 3. Let θ be given and let ϵ_θ and t_θ be as in Proposition 3, and $t \geq t_\theta$. From Proposition 1, for any $\tau \geq 0$,

$$\begin{aligned} W(\tau) &\geq \exp(\min_i(\log w_i(0)) + A_{\max}(\log \alpha_\theta)\tau) \\ &\geq \min_i(w_i(0)) \exp(A_{\max} \log(\alpha_\theta)\tau) \\ &\geq \min_i(w_i(0)) \exp(A_{\max}\epsilon_\theta\tau) \end{aligned} \quad (35)$$

since $1 + x \geq \log x$. This lower bound on $W(\tau)$ is a strictly increasing function of τ. The same is not necessarily true of $a(\tau)$, but we can see that for a fixed value of $A(t) = (1/t)\sum_{\tau=0}^{t-1} a(\tau)$, the lower bound on $L(t)$ is minimized when $a(0), a(1), \ldots, a(\tau^*)$ are as big as possible — i.e., equal to $d_{\max}$ — and the rest of the $a(\tau)$ are zero. Thus $\tau^* = t \cdot A(t)/d_{\max}$. To skirt complications when τ^* is not an integer, we switch to a continuous approximation, using the fact that $\sum_{k=0}^{n} e^k \geq \int_0^n e^x \mathrm{d}x$:

$$\begin{aligned} L(t) &\geq \frac{1}{t}\int_0^{\tau^*} d_{\max} W(\tau)\mathrm{d}\tau \\ &\geq \frac{d_{\max}}{t}\int_0^{\tau^*} \mu_0 \exp(A_{\max}\epsilon_\theta\tau)\mathrm{d}\tau \\ &\geq \frac{\mu_0 d_{\max}}{A_{\max}\epsilon_\theta}\frac{1}{t}\left(\exp\left(\frac{A_{\max}\epsilon_\theta}{d_{\max}} tA(t)\right) - 1\right) \end{aligned} \quad (36)$$

where $\mu_0 = \min_i(w_i(0))$. From Proposition 3, we know that $A(t) \geq A_{\max} - \theta$, so

$$L(t) \geq \frac{\mu_0 d_{\max}}{A_{\max}\epsilon_\theta}\frac{1}{t}\left(\exp\left(\frac{A_{\max}\epsilon_\theta(A_{\max} - \theta)}{d_{\max}} t\right) - 1\right) \quad (37)$$

Thus, the cumulative mean log-odds of a correct response on error trials (i.e., the log-odds of a correct response immediately before each error was committed) is bounded below by a function that is only slightly less than exponential in the number of mistakes. (This of course causes the mistakes themselves to become less and less frequent, so the log-odds grows slower in terms of the number of trials.)

This is a worst-case bound that does not depend on how the training sequence is constructed. Since error trials oversample error-prone 2AFC pairs, the cumulative mean log-odds of a correct response on all trials is expected to be greater than that on the error trials.

4.3 Simulation results

The analysis was checked by a simulation whose parameters were chosen to roughly approximate a typical Harmonic Grammar phonological analysis. For each replication of the simulation, n was sampled uniformly from $\{2, \ldots, 20\}$, and $d_{\max}$ from $\{1, 2, 3, 4\}$. Numbers m and r were uniformly sampled from 4 to 64 and from 8 to 256, respectively. A weight vector $\boldsymbol{\mu}$ of length n was made by uniformly sampling each entry from the interval $(0, 1)$. The cells of an $m \times n$ tableau (candidates $\times$ constraints) were filled by uniformly sampling each from $\{-d_{\max}, \ldots, 0\}$, and one was randomly (uniformly) chosen to be the most-harmonic positive stimulus, so long as it was not the most- or least-harmonic of all. The other candidates' harmonies thus determined their positive/negative status. The least upper bound $A_{\max}$ was approximated by maximizing $A_{\boldsymbol{\mu}}$ over all $\boldsymbol{\mu}$ consistent with the concept using the quasi-Newton method of Byrd et al. (1995) as implemented in the `optim` function of Version 3.2.2 of the `stats` package in R (R Core Team, 2015). A number r of winner-loser pairs was made by randomly sampling (uniformly, with replacement) from the positive and negative candidates, provided that some intended winners had less-than-perfect scores. A θ was sampled uniformly from $[1/32, 1/8]$, ϵ_θ was chosen as in Equation 31, and t_θ was chosen as in Equation 32. Initial weights were all set to 1. Winnow-MaxEnt was trained until the learner's cumulative average relative harmony gap (the left-hand side of Inequality 26) reached or exceeded $A_{\max} - \theta/2$, or until $5t_\theta$ errors had occurred. (If that criterion was already met before any training, the simulation was discarded and replaced.)

In 10,000 replications, the bound of Inequality 26 always underestimated the cumulative average harmony gap at every time point (error) in every replication by a margin of at least 0.0823 (median, 0.9732). The bound of Inequality 32 always overestimated the actual number of errors required to reach $A_{\max} - \theta$ by a factor of at least 3.93, and usually by very much more (the median was a factor of 52.60). The bound in Inequality 37 always underestimated the actual log-odds at $t = t_\theta$ by at least a margin of 1.084 (median, 280.3). Average-case performance in actual applications may therefore be much better.

5 Beyond 2AFC with positive constraints

Because the Luce choice rule describes how to choose one item out of a set of alternatives on the basis of nonnegative harmony values, k-AFC for $k > 2$ requires no amendment for positive constraints. For negative constraints, we let the alternatives be, not individual candidates $x_i \in X$, competing to be the winning individual on the basis of their harmonies $h_{\mathbf{w}}(x_i)$, but rather sets of $k-1$ candidates $X_i = X - \{x_i\}$, competing to be the losing set on the basis of their harmonies $H_{\mathbf{w}}(X_j) = \sum_{x \in X_j} h_{\mathbf{w}}(x) = (\sum_{x \in X} h_{\mathbf{w}}(x)) - h_{\mathbf{w}}(x_j)$. Then

$$\begin{aligned}\Pr(x_i \mid X, \mathbf{w}) &= \frac{\exp\left(\sum_{x \in X} h_{\mathbf{w}}(x)\right) / \exp(h_{\mathbf{w}}(x_j))}{\sum_{j=1}^{k} \exp\left(\sum_{x \in X} h_{\mathbf{w}}(x)\right) / \exp(h_{\mathbf{w}}(x_j))} \\ &= \frac{\exp(-h_{\mathbf{w}}(x_i))}{\sum_{j=1}^{k} \exp(-h_{\mathbf{w}}(x_j))}\end{aligned} \tag{38}$$

In other words, negative (penalizing) constraints can be implemented by simply inverting the sign of the marks awarded.

6 Constraints as representation subtrees

The next step is a constraint schema that enables breeding and mutation. We can define markedness constraints as subtrees of autosegmental representations, such that every representation is itself a constraint (Burzio, 1999). The representational system in the implemented model is a hierarchical prosodic and featural tree structure simplified from Gussenhoven and Jacobs (2005, Ch. 5) by omitting feet and moras. Figure 6 shows an example.

A markedness constraint is a representation, rooted at a PrWd, which awards a mark to a candidate for each time it matches part of that candidate. Examples of familiar markedness constraints expressed in this schema are shown in Figure 2. The symbols L and R mark left and right constituent boundaries.

The Subtree Schema (Moreton, 2010b,a,c) differs from previous explicitly described constraint schemas used in implemented inductive learning models (Hayes and Wilson, 2008; Adriaans and Kager, 2010; Pizzo, 2013; Rasin and Katzir, 2016) in that it imposes no extra limits on constraint structure beyond those inherited from representational structure; it integrates autosegmental tier structure

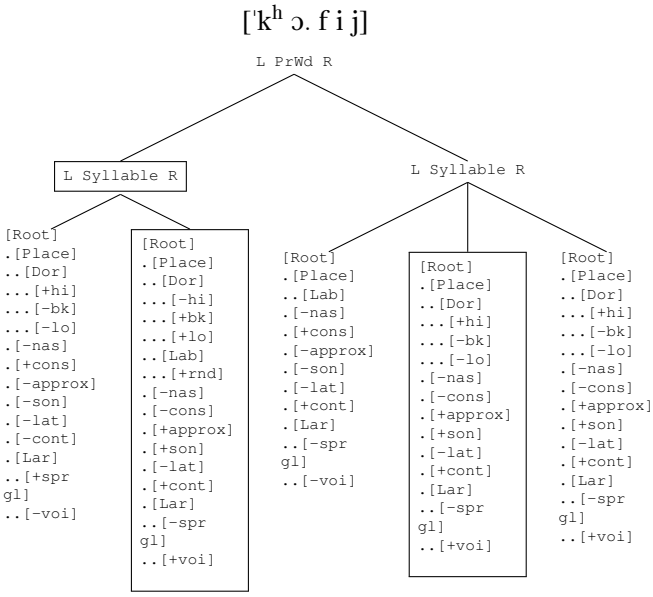

Figure 1: Example: *coffee*. Heads are enclosed in a box.

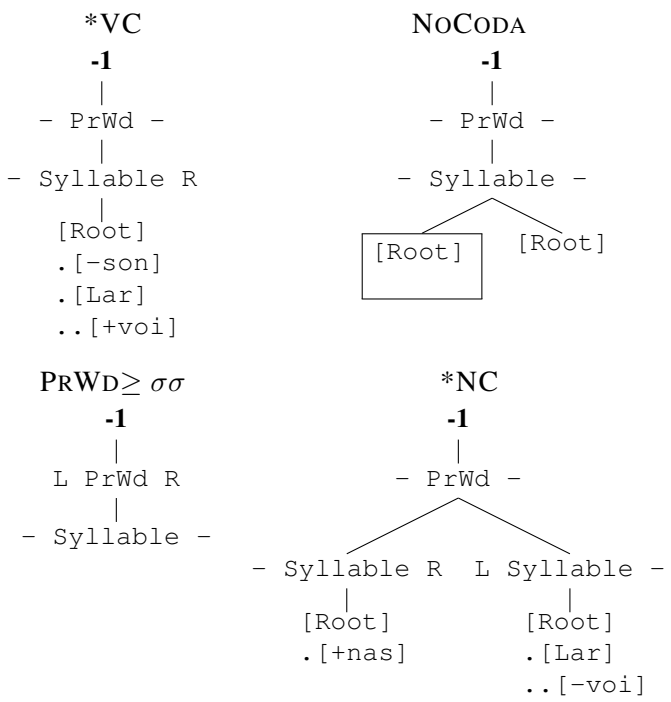

Figure 2: Some familiar markedness constraints in the Subtree Schema: *VC, final-obstruent devoicing (Ito and Mester, 2003); NoCoda (Prince and Smolensky, 1993); a constraint enforcing a disyllabic minimal prosodic word; *NC (Pater, 2004).

with prosodic constituent structure, thus supporting positional constraints; it accommodates non-adjacent dependencies; it allows variables (not discussed here) for (e.g.) Agree, OCP, and reduplication; and it provides continuity between constraints and representations (Golston, 1996; Burzio, 1999). The tree structure also lends itself to a recursive breeding algorithm, as described in the next section.

7 Mutation, recombination, and selection

We now let the micro-constraints reproduce with variation, so that the Winnow-MaxEnt rule acts as a selective force in an evolutionary algorithm. Evolutionary algorithms have long been applied to problems closely related to the ones addressed here, including evolving receptive fields for inputs to the single-layer perceptron (Nakano et al., 1995) and evolving tree structures (Cramer, 1985; Koza, 1989). Replication of mental representations with variation, recombination, and selection is a leading theory of human creativity in other domains (Simonton, 1999, 2004; Dietrich and Haider, 2015).

Each constraint which is chosen to breed is randomly paired with another chosen breeder of equal or greater fitness (reproductive probability). The two constraints are mated recursively. The offspring of two prosodic-category nodes randomly copies node-level properties (left and right anchors) from the parents. Immediate dependents of each parent node are randomly paired, preserving left-to-right order, and leaving some dependents unpaired if one parent node has more than the other. Each pair of dependents then breeds to make one node in the offspring. An unpaired dependent is either inherited intact or deleted, with probability 1/2. The offspring of paired compatible unary-feature nodes (e.g., [+Cor] bred with [+Cor]) is computed analogously: Subfeature nodes common to both parents are paired and bred recursively; unpaired nodes are either copied intact or deleted, with probability 1/2. The offspring of paired compatible binary-feature nodes (e.g., [+voice] bred with [−voice]) is identical to each of the parents with probability 1/2. An example is shown in Figure 3.

The offspring then undergoes undirected mutation. Mutation is recursive (when a node is exposed to the hazard, so are its dependents). Mutations include gaining, losing, or duplicating a dependent node; designating, undesignating, or redesignating a constituent as a prosodic head; setting or unsetting the left and right prosodic anchors; and invert-

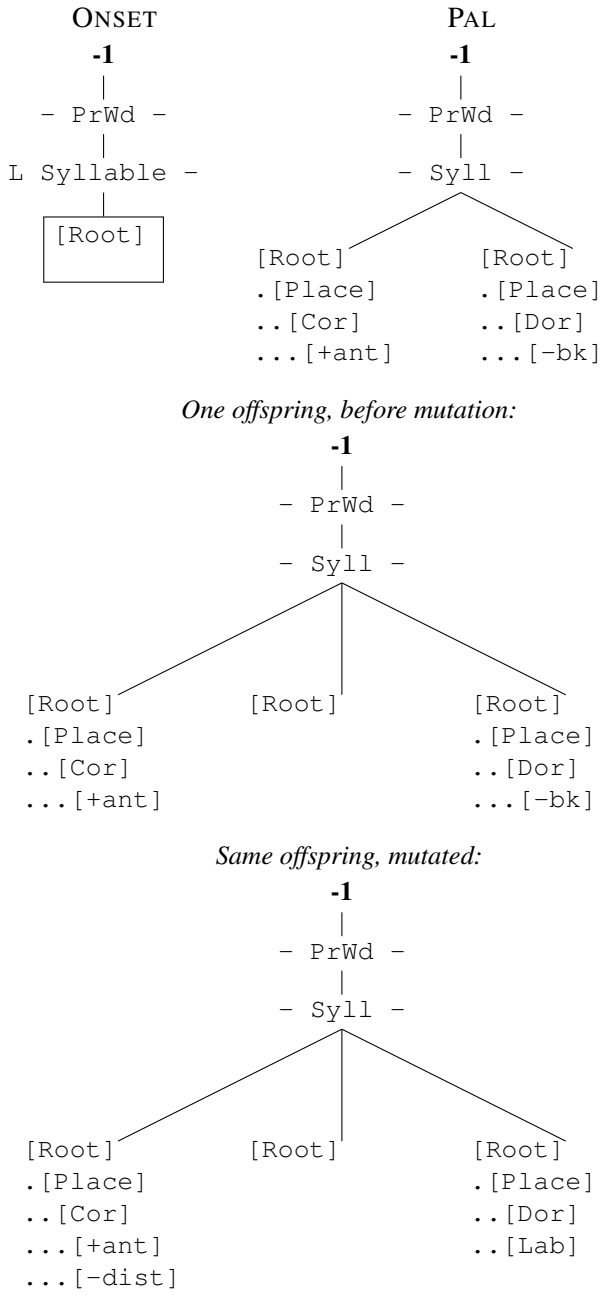

Figure 3: Breeding and mutation, illustrated with parents ONSET (à la Smith 2003, 2012) and PAL (McCarthy, 1999).

ing the coefficient on a binary feature. The probability of each is controlled by a separate parameter. A macro-constraint thus becomes an equivalence class of formally diverse micro-constraints which assign the same marks to all of the candidates.

8 Constraint breeding in practice

The combination of Winnow-MaxEnt with the Subtree Schema is illustrated using an artificial phonotactic pattern. Candidates were the initial syllables from the stimuli of Saffran and Thiessen (2003, 494). The positive stimuli (winners) had the form $\{p,t,k\}V\{b,d,g\}$ and the negative stimuli (losers) $\{b,d,g\}V\{p,t,k\}$. The population size was fixed at $n = 1000$ (micro-)constraints, which initially were identical clones that gave -1 mark to every PrWd. On each trial, a random candidate pair was presented for 2AFC judgement. When a mistake happened, the quantity $r_i = \alpha^{x_i^+ - x_i^-}$ was calculated for each constraint C_i. If $r_i \geq 1$, the constraint made one offspring with certainty, then another with probability $1 - r_i$. If $r_i < 1$, the constraint made one offspring with probability r_i. If the offspring violated a "hard" restriction on representations (e.g., a ban on [+high +low]), or scored all candidates alike, breeding was retried up to 100 times before giving up and accepting the undesirable offspring. The new generation then completely replaced the old.

For one set of 50 simulations, the harmony quantum ζ was set to 0.01, the learning rate ϵ to 0.25, the mutation probability to 0.25, and the probabilities of each individual mutation type to 0.1. In 17 of them, performance on the 1000th trial was above 0.90 correct. Two examples are shown in Figure 4 to illustrate the variety of simulation behavior.

In the top panel, performance is initially at chance on all pairs. As the constraint population diversifies, so do the error probabilities of the individual pairs, but average performance stays at chance. That changes after two near-simultaneous innovations, the fell-swoop constraint $C_9 = *[-\text{voice}]]_\sigma$ (Trial 378) and a parochial version $C_{11} = *V : [-\text{voice}]]_\sigma$ (Trial 384) that applies only when the vowel is long (tense). Both macro-constraints prosper, but greater generality of C_9 gives it a reproductive advantage (it breeds whenever C_{11} does, but not vice versa). By Trial 999, C_9 is represented 591 times in the population (equivalent to a weight of $591 \cdot \zeta = 5.91$). The slight bifurcation at the end, visible as a thickening of the gray line, is due to 48 instances of C_{11} that cause slightly better accuracy for long vowels.

In the bottom panel, the fell-swoop constraint $*[-\text{voice}]]_\sigma$ does not arise until Trial 732, by which time two parochial constraints, one for long- and one for short-vowelled syllables, have already established themselves and slowed the learning rate. The

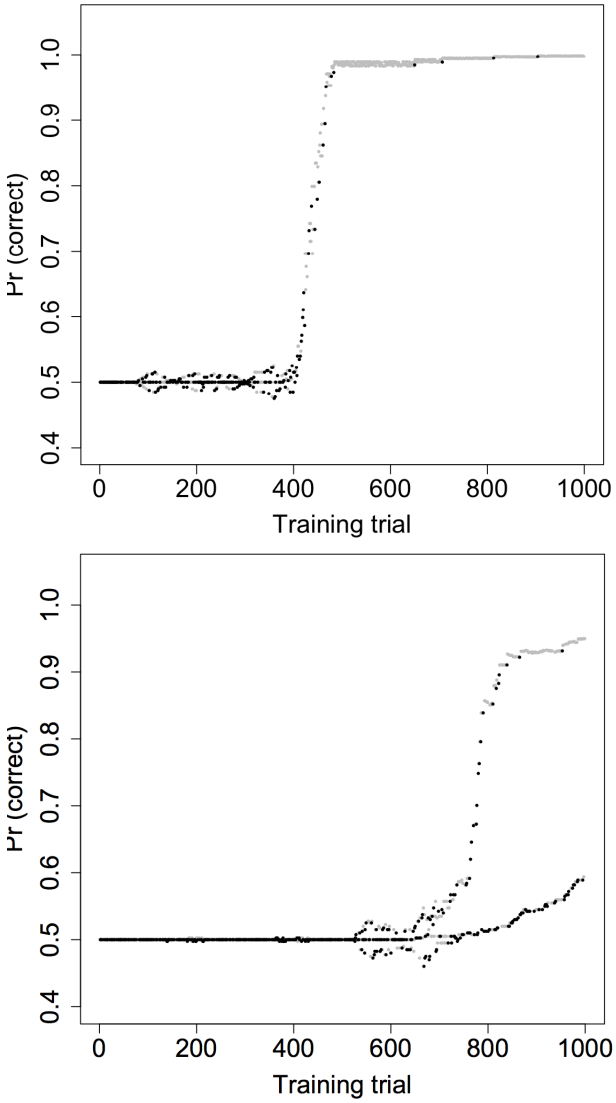

Figure 4: Probability of correct response to the specific winner-loser pair presented on each trial, for one run of the simulation. Black points show errors.

two have prospered unequally ($284\times$ vs. $28\times$ on Trial 999), so that the learning curves for the two syllable types diverge. The fell-swoop constraint is still feeble ($7\times$) because the learner nearly stopped making errors on long-vowel syllables before it was discovered, so its weight now grows in tandem with that of the short-vowel constraint. When the simulation ends the learner is near-perfect when the vowel is long, but only a bit above chance when it is short. The early discovery of a solution for a subset of the data has inhibited a more general solution.

9 Discussion

Sigmoidal abruptness in an observed learning curve has often been taken as distinguishing "rule-based" learning by serial hypothesis testing from "cue-based" associative learning by gradual weight changes (Ashby et al., 1998; Love, 2002; Maddox and Ashby, 2004; Smith et al., 2012; Kurtz et al., 2013). The theory is that while the curve is flat, the learner is serially testing and discarding incorrect rule hypotheses, and the jump occurs when the correct rule is found. The Winnow-MaxEnt-Subtree Breeder model shows that the same observation is consistent with an incremental constraint-based learner. While the curve is flat, this learner is exploring the space of possible constraints, and the jump occurs when a useful mutant arises and prospers.

This behavior leads to a hypothesis. Becker and Tessier (2011) have proposed a correlation between abruptness and innateness (or at least pre-existingness), viz., that a U-shaped kink in an L1 learning curve means that the learner has innovated a constraint and added it at the top of the hierarchy, causing a transient drop in adult-like performance. Analogously, the behavior of Winnow-MaxEnt-Subtree Breeder implies that patterns which depend only on preexisting constraints (supplied by Universal Grammar, or transferred from a previously-acquired language) should be acquired less abruptly than patterns which depend on constraints that are specific to the particular natural or artificial language.

Another consequence is an emergent bias in favor of more-general constraints. In Winnow-MaxEnt-Subtree Breeder, general constraints automatically outcompete parochial ones because they are more fit (see discussion of Figure 4, above; Pater and Moreton 2012; Moreton et al. 2017, §4.1). In that respect, the model is akin to the Minimum Description Length learner of Rasin and Katzir (2016), which adds, removes, and changes constraints without trying to anticipate their effects, rather than learners in which a drive towards generalization is hard-wired into the constraint-generation component (Hayes and Wilson 2008, §4.2.2, Adriaans and Kager 2010, 317f.).

References

Adriaans, F. and R. Kager (2010). Adding generalization to statistical learning: the induction of phonotactics from continuous speech. *Journal of Memory and Language 62*(3), 311–331.

Ashby, F. G., L. A. Alfonso-Reese, A. U. Turken, and E. M. Waldron (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review 105*(3), 442–481.

Barlow, J. A. and D. A. Dinnsen (1998). Asymmetrical cluster development in a disordered system. *Language Acquisition 7*(1), 1–49.

Becker, M. and A. Tessier (2011). Trajectories of faithfulness in child-specific phonology. *Phonology 28*, 163–196.

Boersma, P. and J. Pater (2016). Convergence properties of a gradual learning algorithm for Harmonic Grammar. In J. J. McCarthy and J. Pater (Eds.), *Harmonic Grammar and Harmonic Serialism*, pp. 389–434. Sheffield, England: Equinox.

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39*(3/4), 324–345.

Burzio, L. (1999). Surface-to-surface morphology: when your representations turn into constraints. MS, Department of Cognitive Science, Johns Hopkins University. ROA-341.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing 16*, 1190–1208.

Cramer, N. L. (1985). A representation for the adaptive generation of simple sequential programs. In J. Grefenstette (Ed.), *Proceedings of the First International Conference on Genetic Algorithms*, pp. 183–187.

Dietrich, A. and H. Haider (2015). Human creativity, evolutionary algorithms, and predictive representations: the mechanics of thought trials. *Psychonomic Bulletin and Review 22*, 897–915.

Gerlach, S. R. (2010). *The acquisition of consonant feature sequences: harmony, metathesis, and deletion patterns in phonological development*. Ph. D. thesis, University of Minnesota.

Goldwater, S. J. and M. Johnson (2003). Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Erkisson, and O. Dahl (Eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pp. 111–120.

Golston, C. (1996). Direct Optimality Theory: Representation as pure markedness. *Language 72*(4), 713–748.

Gussenhoven, C. and H. Jacobs (2005). *Understanding phonology* (2nd ed.). Understanding Language Series. London: Hodder Arnold.

Guy, G. R. (2014). Linking usage and grammar: generative phonology, exemplar theory, and variable rules. *Lingua 142*, 57–65.

Hayes, B. and C. Wilson (2008). A Maximum Entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry 39*(3), 379–440.

Ito, J. and R. A. Mester (2003). On the sources of opacity in OT: coda processes in German. In C. Féry and R. van de Vijver (Eds.), *The syllable in Optimality Theory*, pp. 271–303. Cambridge, England: Cambridge University Press.

Jäger, G. (2007). Maximum Entropy models and Stochastic Optimality Theory. In J. Grimshaw, J. Maling, C. Manning, J. Simpson, and A. Zaenen (Eds.), *Architectures, rules, and preferences: a festschrift for Joan Bresnan*, pp. 467–479. Stanford, California: CSLI Publications.

Koza, J. R. (1989). Hierarchical genetic algorithms operating on populations of computer programs. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, Volume 1, San Mateo, California, pp. 768–774. Morgan Kaufmann.

Kurtz, K. J., K. R. Levering, R. D. Stanton, J. Romero, and S. N. Morris (2013). Human learning of elemental category structures: revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition 39*(2), 552–572.

Legendre, G., Y. Miyata, and P. Smolensky (1990). Can connectionism contribute to syntax? Harmonic Grammar, with an application. In M. Ziolkowski, M. Noske, and K. Deaton (Eds.), *Proceedings of the 26th Regional Meeting of the Chicago Linguistic Society*, Chicago, pp. 237–252. Chicago Linguistic Society.

Levelt, C. and M. van Oostendorp (2007). Feature

co-occurrence constraints in L1 acquisition. *Linguistics in the Netherlands 24*(1), 162–172.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning 2*, 285–318.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review 9*(4), 829–835.

Luce, R. D. (1977). The Choice Axiom after twenty years. *Journal of Mathematical Psychology 15*, 215–233.

Luce, R. D. (2005 [1959]). *Individual choice behavior: a theoretical analysis*. New York: Dover.

Macken, M. A. and D. Barton (1978, March). The acquisition of the voicing contrast in English: a study of voice-onset time in word-initial stop consonants. Report from the Stanford Child Phonology Project.

Macmillan, N. A. and C. D. Creelman (2004). *Detection Theory: A User's Guide*. Cambridge, England: Lawrence Erlbaum.

Maddox, W. T. and F. G. Ashby (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes 66*, 309–332.

Magri, G. (2013). HG has no computational advantages over OT: toward a new toolkit for computational OT. *Linguistic Inquiry 44*(4), 569–609.

McCarthy, J. J. (1999). Introductory OT on CD-ROM. Graduate Linguistic Students' Association, University of Massachusetts, Amherst.

Mitrinović, D. S. (1970). *Analytic inequalities*. New York: Springer-Verlag.

Moreton, E. (2010a, April). Connecting paradigmatic and syntagmatic simplicity bias in phonotactic learning. Department colloquium, Department of Linguistics, MIT.

Moreton, E. (2010b, February). Constraint induction and simplicity bias. Talk given at the Workshop on Computational Modelling of Sound Pattern Acquisition, University of Alberta.

Moreton, E. (2010c, May). Constraint induction and simplicity bias in phonotactic learning. Handout from a talk at the Workshop on Grammar Induction, Cornell University.

Moreton, E. (2018). Conditions on abruptness in a gradient-ascent Maximum Entropy learner. In G. Jarosz and J. Pater (Eds.), *Proceedings of the Society for Computation in Linguistics*, Volume 1, pp. Article 13.

Moreton, E., J. Pater, and K. Pertsova (2017). Phonological concept learning. *Cognitive Science 41*(1), 4–69.

Moreton, E. and K. Pertsova (2016). Implicit and explicit processes in phonotactic learning. In TBA (Ed.), *Proceedings of the 40th Boston University Conference on Language Development*, Somerville, Mass., pp. TBA. Cascadilla.

Nakano, K., H. Hiraki, and S. Ikeda (1995). A learning machine that evolves. In *Proceedings of ICEC–95*, pp. 808–813.

Novikoff, A. B. (1963). On convergence proofs for perceptrons. Technical report, Stanford Research Institute.

Pater, J. (2004). Austronesian nasal substitution and other $*NC$ effects. In J. J. McCarthy (Ed.), *Optimality Theory in phonology: a reader*, Chapter 14, pp. 271–289. Malden, Mass.: Blackwell.

Pater, J. and E. Moreton (2012). Structurally biased phonology: complexity in learning and typology. *Journal of the English and Foreign Languages University, Hyderabad 3*(2), 1–44.

Pizzo, P. (2013, January 19). Learning phonological alternations with online constraint induction. Slides from a presentation at the 10th Old World Conference on Phonology (OCP 10).

Prince, A. and P. Smolensky (1993). *Optimality Theory: constraint interaction in generative grammar*. Department of Linguistics, Rutgers University.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasin, E. and R. Katzir (2016). On evaluation metrics in optimality theory. *Linguistic Inquiry 47*(2), 235–282.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review 65*(6), 386–408.

Saffran, J. R. and E. D. Thiessen (2003). Pattern induction by infant language learners. *Developmental Psychology 39*(3), 484–494.

Simonton, D. K. (1999). Creativity as blind variation and selective retention: is the creative process Darwinian? *Psychological Inquiry 10*(4), 309–

328.

Simonton, D. K. (2004). *Creativity in science: chance, logic, genius, and Zeitgeist*. Cambridge University Press.

Smith, J. D., M. E. Berg, R. G. Cook, M. S. Murphy, M. J. Crossley, J. Boomer, B. Spiering, M. J. Beran, B. A. Church, F. G. Ashby, and R. C. Grace (2012). Implicit and explicit categorization: a tale of four species. *Neuroscience and Biobehavioral Reviews 36*(10), 2355–2369.

Smith, J. L. (2003). Onset sonority constraints and subsyllabic structure. MS, Department of Linguistics, University of North Carolina, Chapel Hill. ROA-602.

Smith, J. L. (2012). The formal definition of the ONSET constraint and implications for Korean syllable structure. In T. Borowsky, S. Kawahara, T. Shinya, and M. Sugahara (Eds.), *Prosody matters: essays in honor of Elisabeth Selkirk*, pp. 73–108. Equinox.

Smith, N. V. (1973). *The acquisition of phonology: a case study*. Cambridge, England: Cambridge University Press.

Strauss, D. (1992). The many faces of logistic regression. *American Statistician 46*(4), 321–327.

Sutton, R. S. and A. G. Barto (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review 88*(2), 135–170.

Vihman, M. M. and S. Velleman (1989). Phonological reorganization: a case study. *Language and Speech 32*, 149–170.