Proceedings of the Society for Computation in Linguistics

Volume 1 Article 13

2018

Conditions on abruptness in a gradient-ascent Maximum Entropy learner

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

Conditions on abruptness in a gradient-ascent Maximum Entropy learner*

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

Abstract

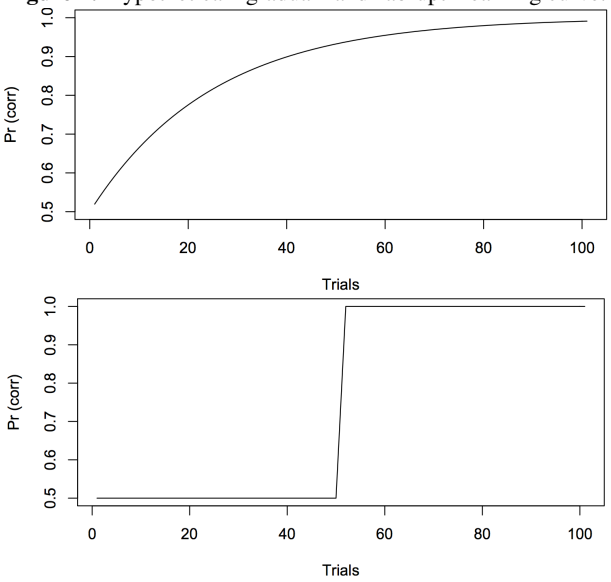
When does a gradual learning *rule* translate into gradual learning *performance*? This paper studies a gradient-ascent Maximum Entropy phonotactic learner, as applied to two-alternative forced-choice performance expressed as log-odds. The main result is that slow initial performance cannot accelerate later if the initial weights are near zero, but can if they are not. Stated another way, abruptness in this learner is an effect of transfer, either from Universal Grammar in the form of an initial weighting, or from previous learning in the form of an acquired weighting.

1 Introduction

An important class of constraint-based phonological learning models responds to training by making small changes in the weight or rank of constraints (reviewed in Jarosz 2016). The gradualness of the learning rule seems to suggest that performance ought to change gradually as well, resembling the first rather than the second panel in Figure 1. In work on non-linguistic pattern learning, abrupt improvement has been cited as diagnostic of an explicit, "rule-based" learning algorithm which serially tests hypotheses, as opposed to a "cue-based" one which slowly learns association weights (Ashby et al., 1998; Love, 2002; Maddox and Ashby, 2004; Smith et al., 2012; Kurtz et al., 2013). Abruptness has been found to correlate with other indicia of explicitness by humans learning artificial phonology (Moreton and Pertsova, 2016).

In fact, performance can change abruptly in gradual learners (Elman et al. 1996, Ch. 3–4; GLA examples in Boersma 1998, Figure 14.25; Boersma and Levelt 2000; Jesney 2016). When does a gradual learning rule entail gradual learning performance? Could the model spend many trials invisibly inching its way around to some point in weight space from which it can suddenly accelerate? Conversely, if we observe abrupt improvement in human learners, does that disconfirm the model?

Figure 1: Hypothetical "gradual" and "abrupt" learning curve.



This paper addresses the question in a particularly basic case, that of a Maximum Entropy phonotactic learner with a fixed constraint set that uses

*The author is indebted to Jen Smith, Joe Pater, Katya Pertsova, and Chris Wiesen for comments and suggestions. Any errors are of the author's own making. The research was supported in part by NSF BCS 1651105, "Inside phonological learning", to E. Moreton and K. Pertsova.

Proceedings of the Society for Computation in Linguistics (SCiL) 2018, pages 113-124.
Salt Lake City, Utah, January 4-7, 2018

gradient ascent on log-likelihood, no prior, and no restrictions on weights, and that makes two-alternative forced-choice (2AFC) decisions using the Luce choice rule. Gradient ascent Max-Ent is of interest not only in its own right, but because of its close relation to the Gradual Learning Algorithms for Stochastic Optimality Theory, Harmonic Grammar and Noisy Harmonic Grammar, and models of non-linguistic learning such as the Perceptron (Boersma and Hayes, 2001; Fischer, 2005; Jäger, 2007; Johnson, 2007; Pater, 2008; Pater and Moreton, 2012; Boersma and Pater, 2016; Moreton et al., 2017).

The results can be summarized as follows: Regardless of what the constraints actually are, if the initial weights are exactly zero then — provided that the training and test distributions are chosen in a particular way — 2AFC performance improves fastest at the outset of learning, making abrupt learning impossible. Even if, instead, the initial weights are only *near* zero, the 2AFC learning curve tracks that of a learner whose initial weights are *exactly* zero, in that the two learners' trajectories in weight space steadily converge, and the closer they are in weight space, the more similar their 2AFC performance is. An example is given to show that large *non*-zero initial weights can, but need not, lead to abrupt 2AFC performance.

2 Learner and experimental scenario

The universe of candidates is a finite set $X = \{x_1, \ldots, x_n\}$, known to the experimenter. The model uses an unobservable set of constraints $c_1, \ldots, c_m$ and an unobservable weight vector $\mathbf{w} = (w_1, \ldots, w_m)$ to assign unobservable probabilities $\mathbf{p} = (p_1, \ldots, p_n)$ to the candidate. This is done as follows (Goldwater and Johnson, 2003; Jäger, 2007; Hayes and Wilson, 2008):

The *harmony* of a candidate x_j is defined as the sum of its score vector, weighted by the current weights:

$$h_{\mathbf{w}}(x_j) = \sum_{i=1}^{m} w_i c_i(x_j) \tag{1}$$

The model's estimate of the probability p_j of candidate x_j is the exponential of its harmony, divided by the summed exponentials of the harmonies of all representations:

$$Z_{\mathbf{w}} = \sum_{j=1}^{n} \exp h_{\mathbf{w}}(x_j) \tag{2}$$

$$\Pr(X = x_j \mid \mathbf{w}) = \frac{\exp h_{\mathbf{w}}(x_j)}{Z_{\mathbf{w}}} \tag{3}$$

The experimenter can at any time give the model a two-alternative forced-choice *test*, in which two candidates x_i and x_j are presented to the model, which chooses x_i with probability

$$\Pr(x_i|(x_i, x_j)) = \frac{p_i}{p_i + p_j} \tag{4}$$

This is the Luce choice rule (Luce, 1959, 23). The test is assumed not to change the state of the model. At the beginning of the experiment, the experimenter chooses two probability distributions $\mathbf{r}^+$ and $\mathbf{r}^-$. On each test trial, one candidate is sampled from X with probabilities given by $\mathbf{r}^+$, and the other is sampled from X with probabilities given by $\mathbf{r}^-$.

The experimenter can also *train* the model by giving it a candidate x_i as an example of a legal word. Instead of training on individual candidates (stochastic gradient ascent), we instead run the learner in batch mode (gradient ascent); i.e., instead of a candidate on each trial, the learner receives a distribution $\mathbf{p}^+$, where p_i^+ corresponds to the probability of presenting x_i on a stochastic gradient ascent training trial.

The model updates its weights according to the following rule:

$$\Delta w_i = \theta \cdot (E_{\mathbf{p}^+}[c_i] - E_{\mathbf{w}}[c_i]) \tag{5}$$

This the Maximum Entropy gradient-ascent update rule, as described by Jäger (2007). Its contribution to the update is independent of $\mathbf{p}^-$, the probabilities of the negative training candidates; i.e., the learner does "unsupervised" learning.

Below a continuous approximation to this discrete update rule is used, substituting $\mathrm{d}w_i/\mathrm{d}t$ for Δw_i. The learning rate parameter η is omitted by setting it to 1; i.e., the training-time unit is defined to be how long it takes a constraint to change its weight by one weight unit when $E_{emp}[c_i] - E_{\mathbf{w}}[c_i] = 1$. The step size in weight space is thus fixed, rather than decreasing on a preset schedule (Boersma and Hayes,

2001) or adaptively (Boyd and Vandenberghe, 1999, Section 5.2.1).

In this paper, "abrupt" is used to mean that performance improves slowly at the outset of the experiment, then accelerates later (e.g., a sigmoid). Performance is expressed here as log-odds rather than proportion correct because (A) log-odds is more transparently related both to the learning model (Jäger, 2007) and to the statistical models fit to experimental results (Jaeger, 2008), and (B) proportion correct acts as a squashing function, reducing the visible influence of changing large weights and thus exaggerating the effect whose existence we are arguing for on other grounds.

3 Improvement in log-likelihood decelerates monotonically

We begin by establishing a result that is almost what we want:

Proposition 1. *Let $L(t) = \sum_{j=1}^{n} p_j^+ \log p_j(t)$ denote the model's expectation of the log-likelihood of the empirical distribution at time t (Berger et al., 1996). Then $L(t)$ is always increasing but never accelerating; i.e., for any $t \geq 0$, $dL/dt \geq 0$ and $d^2L/dt^2 \leq 0$.*

Proof. We convert the learner to its Replicator form (Moreton et al., 2017):

$$\frac{\mathrm{d}}{\mathrm{d}t} \log p_i = (C^T C \mathbf{e})_j - \mathbf{p}^T C^T C \mathbf{e} \tag{6}$$

where C is the matrix[1] whose (i, j)-th entry is $c_i(x_j)$, and $\mathbf{e} = \mathbf{p}^+ - \mathbf{p}$. Differentiating the definition of $L(t)$ then yields

$$\begin{aligned} \frac{\mathrm{d}L}{\mathrm{d}t} =& \sum_{j=1}^{n} p_j^+ \frac{\mathrm{d}}{\mathrm{d}t} \log p_j \\ =& \sum_{j=1}^{n} p_j^+ ((C^T C \mathbf{e})_j - \mathbf{p}^T C^T C \mathbf{e})) \\ =& (\mathbf{p}^+ - \mathbf{p})^T C^T C \mathbf{e} \\ =& \mathbf{e}^T C^T C \mathbf{e} \\ =& \|C\mathbf{e}\|^2 \end{aligned} \tag{7}$$

[1]Note difference from familiar tableaus: *Rows* of C correspond to constraints, and *columns* to candidates.

(In this paper, $\|\cdot\|$ is the usual Euclidean norm.) Since $C^T C$ is positive semidefinite, $dL/dt \geq 0$. That confirms what we already know, since the learner does gradient ascent on L. The second derivative is

$$\begin{aligned} \frac{\mathrm{d}^2 L}{\mathrm{d}t^2} =& \frac{\mathrm{d}}{\mathrm{d}t} \mathbf{e}^T C^T C \mathbf{e} \\ =& 2\mathbf{e} C^T C \left(\frac{\mathrm{d}}{\mathrm{d}t} \mathbf{e} \right) \\ =& -2 \left(\sum_j p_j (C^T C \mathbf{e})_j^2 - \sum_j (p_j (C^T C \mathbf{e})_j)^2 \right) \\ =& -2 \sum_j p_j (1 - p_j) (C^T C \mathbf{e})_j^2 \end{aligned} \tag{8}$$

Since $0 < p_j < 1$, the sum is positive unless $\mathbf{e} = \mathbf{0}$. Hence $\mathrm{d}^2 L/\mathrm{d}t^2 \leq 0$ — the log-likelihood is always increasing, but always more and more slowly, until it stops. □

Regardless of the constraint set, initial state, target pattern, or model parameters, learning, measured as log-likelihood, only ever slows down. Abrupt, sigmoidal, or U-shaped $L(t)$ curves are not possible. However, what the experiments measure is not log-likelihood, which depends only on the probability assigned by the model to the winners, but rather 2AFC performance, which depends in part on how it distributes probability among the losing candidates. The next section addresses that complication.

4 When initial weights are all zero, initial improvement bounds later improvement

In typical "artificial-language" experiments, the training and testing stimuli are, or approximate, random samples from the same distribution, and are presented to participants with equal frequency. We consider here a slightly more general possibility:

Proposition 2. *Suppose the experimenter chooses $\mathbf{p}^+$, $\mathbf{r}^+$, and $\mathbf{r}^-$ such that at time $t = 0$, we have*

$$\mathbf{p}^+ - \mathbf{p}(0) = \alpha(\mathbf{r}^+ - \mathbf{r}^-) \tag{9}$$

for some $\alpha > 0$. Let $\lambda_{+,-}$ be the log-odds of a correct 2AFC response. Then at any time $t \geq 0$,

$$\frac{d}{dt}E_{\mathbf{w}}[\lambda_{+,-}]\Big|_t \leq \frac{d}{dt}E_{\mathbf{w}}[\lambda_{+,-}]\Big|_0 \tag{10}$$

Proof. For a given weight vector $\mathbf{w}$, the expected harmony of a positive test candidate is

$$\begin{aligned} E_{\mathbf{w},\mathbf{r}^+}[h_{\mathbf{w}}(x^+)] &= \sum_j r_j^+ h_{\mathbf{w}}(x_j) \\ &= \sum_j r_j^+ \left(\sum_i w_i c_i(x_j) \right) \\ &= \sum_i w_i \left(\sum_j r_j^+ c_i(x_j) \right) \\ &= \sum_i w_i E_{\mathbf{r}^+}[c_i] \end{aligned} \tag{11}$$

i.e, the expected harmony of a positive test candidate is the weighted-by-the-weights sum of the average score on each constraint among positive test candidates. In terms of C, the matrix whose (i,j)-th entry is $c_i(x_j)$, we can write this as

$$\begin{aligned} E_{\mathbf{w},\mathbf{r}^+}[h_{\mathbf{w}}(x^+)] &= \sum_i w_i (C\mathbf{r}^+)_i \\ &= \mathbf{w}^T C\mathbf{r}^+ \end{aligned} \tag{12}$$

The same holds, *mutatis mutandis*, for negative test candidates. For any test pair (x_i, x_j), the log-odds $\lambda_{i,j}$ of choosing x_i is, by Equation 3, just the difference in harmony scores given the current weighting:

$$\lambda_{i,j} = h_{\mathbf{w}}(x_i) - h_{\mathbf{w}}(x_j) \tag{13}$$

That gives us the following expression for the expected value of $\lambda_{+,-}$, the log-odds in favor of a correct test response:

$$\begin{aligned} E_{\mathbf{w}}[\lambda_{+,-}] &= E_{\mathbf{w}}[h_{\mathbf{w}}(x^+) - h_{\mathbf{w}}(x^-)] \\ &= \sum_i w_i (E_{\mathbf{r}^+}[c_i] - E_{\mathbf{r}^-}[c_i]) \\ &= \mathbf{w}^T C(\mathbf{r}^+ - \mathbf{r}^-) \end{aligned} \tag{14}$$

We differentiate that to get

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}E_{\mathbf{w}}[\lambda_{+,-}] &= \sum_i \left(\frac{\mathrm{d}}{\mathrm{d}t} w_i \right) (C(\mathbf{r}^+ - \mathbf{r}^-))_i \\ &= \left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{w} \right)^T C(\mathbf{r}^+ - \mathbf{r}^-) \end{aligned} \tag{15}$$

By the update rule in Equation 5, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} w_i = (E_{\mathbf{p}^+}[c_i] - E_{\mathbf{w}}[c_i]) \tag{16}$$

If $\mathbf{q}$ is any probability distribution over the candidates $X = \{x_1, \ldots, x_n\}$, then $C\mathbf{q}$ is a vector with m elements in which the ith element is $E_{\mathbf{q}}[c_i]$, the expected score on Constraint c_i when a candidate is sampled from X under the distribution $\mathbf{q}$. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{w} = C(\mathbf{p}^+ - \mathbf{p}) \tag{17}$$

Substituting back into Equation 15 then yields

$$\frac{\mathrm{d}}{\mathrm{d}t}E_{\mathbf{w}}[\lambda_{+,-}] = (C(\mathbf{p}^+ - \mathbf{p}))^T C(\mathbf{r}^+ - \mathbf{r}^-) \tag{18}$$

Setting $\mathbf{e} = \mathbf{p}^+ - \mathbf{p}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}E_{\mathbf{w}}[\lambda_{+,-}] = (C\mathbf{e})^T C(\mathbf{r}^+ - \mathbf{r}^-) \tag{19}$$

Applying the Cauchy-Schwarz inequality to Equation 19 yields:

$$\left| \frac{\mathrm{d}}{\mathrm{d}t}E_{\mathbf{w}}[\lambda_{+,-}] \right| \leq \|C\mathbf{e}\| \cdot \|C(\mathbf{r}^+ - \mathbf{r}^-)\| \tag{20}$$

with strict equality if and only if $C(\mathbf{r}^+ - \mathbf{r}^-)$ is a scalar multiple of $C\mathbf{e}$. Because the experimenter chose $\mathbf{p}^+, \mathbf{r}^+$, and $\mathbf{r}^-$ to satisfy the hypothesis in Equation 9, $C(\mathbf{r}^+ - \mathbf{r}^-)$ is a scalar multiple of $C\mathbf{e}(0)$, and so strict equality holds at $t = 0$:

$$\left| \frac{\mathrm{d}}{\mathrm{d}t}E_{\mathbf{w}}[\lambda_{+,-}]|_{t=0} \right| = \|C\mathbf{e}(0)\| \cdot \|C(\mathbf{r}^+ - \mathbf{r}^-)\| \tag{21}$$

From Proposition 1, we know that $\|C\mathbf{e}\|$ decreases monotonically as t increases. Since $\|C(\mathbf{r}^+ - \mathbf{r}^-)\|$ is constant, the product is never bigger than at $t = 0$:

$$\begin{aligned} \left| \frac{\mathrm{d}}{\mathrm{d}t}E_{\mathbf{w}}[\lambda_{+,-}] \right| &\leq \|C\mathbf{e}(0)\| \cdot \|C(\mathbf{r}^+ - \mathbf{r}^-)\| \\ &\leq \left| \frac{\mathrm{d}}{\mathrm{d}t}E_{\mathbf{w}}[\lambda_{+,-}]|_{t=0} \right| \end{aligned} \tag{22}$$

□

In other words, if the learner starts with $\mathbf{w} = \mathbf{0}$ at $t = 0$, then 2AFC performance can't later on improve (or deteriorate) any faster than it did at $t = 0$. In particular, the abrupt learning curve of Figure 1 is impossible for such a learner.[2]

This result was checked numerically by simulation. Each replication of the simulation was done as follows: m and n were sampled uniformly from $\{4, \ldots, 30\}$. A probability s was sampled uniformly from the interval $(0, 1)$, and an $m \times n$ constraint-by-candidate matrix C was generated by randomly setting each entry to 1 with probability s, else to 0. A random concept was generated by uniformly sampling an integer k from $\{1, \ldots, m\}$, and decreeing Candidates $\{x_1, \ldots, x_k\}$ to be positive. The training and test distributions were set thus: $\mathbf{p}^+ = \mathbf{r}^+ = (1/k, \ldots, 1/k, 0, \ldots, 0)^T$; $\mathbf{r}^- = (0, \ldots, 0, 1/(n-k), \ldots, 1/(n-k))^T$. The learning rate η was set to 1/100, $\mathbf{w}(0)$ was set to $\mathbf{0}$, and the learner was run for 300 update cycles. The change in the model's log-likelihood on each cycle was measured, and the index of the largest increase was recorded. Ten thousand such replications were run. The largest increase always occurred on the first cycle, as predicted.

5 Adjacent learning trajectories converge in weight space

In this section, we show that learning erases small perturbations in the state of the model: Two learners that now have slightly different weights will in the future draw closer and closer together. This is not altogether unexpected; after all, the learners are climbing the same convex hill, and eventually they will both arrive at the summit. But what if a small initial difference somehow leads to paths that diverge before converging, or causes one to lag further and further behind the other for a while? Fortunately, this is not the case.

Proposition 3. *Consider two otherwise identical learning simulations such that at a given time t_1, one is in state $\mathbf{w}(t_1)$ and the other is in a nearby state $\mathbf{w}'(t_1)$. For any $t_2 > t_1$, we have $\|\mathbf{w}(t_2) - \mathbf{w}'(t_2)\| \leq \|\mathbf{w}(t_1) - \mathbf{w}'(t_1)\|$.*

[2]The rate of improvement is bounded above by a monotonically decreasing quantity (Equation 20), but that does not guarantee that the rate itself is monotonically decreasing. Hence, U-shaped curves are not excluded by this result. None were found in the simulations described on this page.

Proof. The rate of change in the squared distance between the two learners in weight space at any time t is

$$\begin{aligned} D = \frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{w} - \mathbf{w}'\|^2 &= 2(\mathbf{w} - \mathbf{w}')^T \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{w} - \mathbf{w}') \\ &= 2(\mathbf{w} - \mathbf{w}')^T (C\mathbf{e} - C\mathbf{e}') \\ &= -2(\mathbf{w} - \mathbf{w}')^T C(\mathbf{p} - \mathbf{p}') \end{aligned} \tag{23}$$

Since the harmonies of the candidates are the weighted sums of their constraint scores,

$$(\mathbf{w} - \mathbf{w}')^T C = (\mathbf{h} - \mathbf{h}')^T \tag{24}$$

It will be convenient to set $\boldsymbol{\gamma} = \mathbf{h}' - \mathbf{h}$ and write

$$D = 2\boldsymbol{\gamma}^T(\mathbf{p} - \mathbf{p}') \tag{25}$$

We will show that D attains a local maximum at $\boldsymbol{\gamma} = \mathbf{0}$, using the usual second-derivative test.

When $\mathbf{w} = \mathbf{w}'$, $\boldsymbol{\gamma} = \mathbf{0}$, and so of course $D = 0$. We now find the first and second partial derivatives of D with respect to the elements of $\boldsymbol{\gamma}$, evaluated at $\boldsymbol{\gamma} = \mathbf{0}$. From Equation 3, we have

$$p'_i = \frac{e^{h_i + \gamma_i}}{\sum_k e^{h_k + \gamma_k}} \tag{26}$$

The effect on $\mathbf{p}'$ of small changes in $\boldsymbol{\gamma}$ is given by the derivatives

$$\frac{\partial}{\partial \gamma_i} p'_i = p'_i(1 - p'_i) \qquad \text{and} \qquad \frac{\partial}{\partial \gamma_i} p'_{j \neq i} = -p'_i p'_j \tag{27}$$

Hence the first-order partials of D are

$$\begin{aligned} \frac{\partial}{\partial \gamma_i} D =& 2 \sum_k \frac{\partial}{\partial \gamma_i} \left(\gamma_k (p_k - p'_k) \right) \\ =& 2 \left(p_i - p'_i - \gamma_i p'_i + p'_i \sum_k \gamma_k p'_k \right) \end{aligned} \tag{28}$$

These are all zero at $\boldsymbol{\gamma} = \mathbf{0}$, since then $p_i = p'_i$. The second-order partials at $\boldsymbol{\gamma} = \mathbf{0}$ turn out (after considerable algebra, omitted here) to be

$$\frac{\partial^2}{\partial \gamma_i^2} D = -4 p_i (1 - p_i) \tag{29}$$

and

$$\frac{\partial^2}{\partial\gamma_i\partial\gamma_j}D = 4p_ip_j \tag{30}$$

The Hessian matrix H of D at $\boldsymbol{\gamma} = \mathbf{0}$ is thus[3]

$$H = -4(\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T) \tag{31}$$

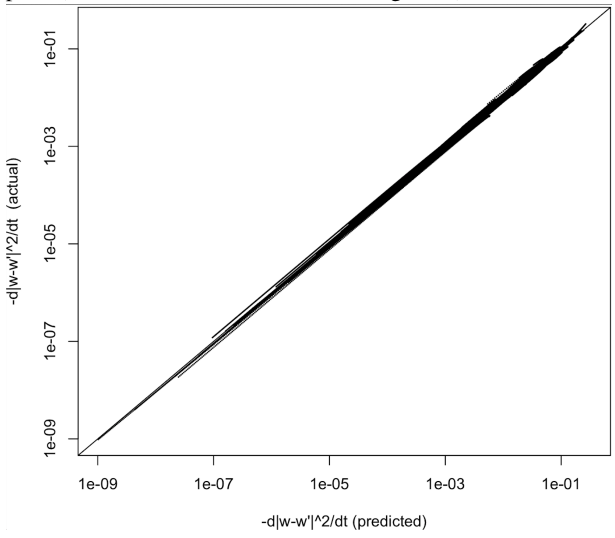
The row sums of H are all zero, the diagonal entries are all negative, and the off-diagonal entries are positive, so by the Gershgorin circle theorem (Horn and Johnson, 1985, 344–345), none of the eigenvalues of H are positive. We now show that exactly one of the eigenvalues is zero: Suppose $H\mathbf{x} = \mathbf{0}$. Then $\text{diag}(\mathbf{p})\mathbf{x} = \mathbf{p}(\mathbf{p}^T\mathbf{x})$, i.e., a scalar multiple of $\mathbf{p}$. Consequently, for every i, it is true that $p_ix_i = p_i\mathbf{p}^T\mathbf{x}$. We can cancel the p_is to get $x_i = \mathbf{p}^T\mathbf{x}$, so $\mathbf{x} = \mathbf{1}(\mathbf{p}^T\mathbf{x})$. Hence, $\mathbf{1}$ is the *only* eigenvector whose eigenvalue is zero. The other eigenvalues are all negative.

Thus the value that D attains at $\boldsymbol{\gamma} = \mathbf{0}$ is a local maximum in every direction except along the line where $\boldsymbol{\gamma}$ is a scalar multiple of $\mathbf{1}$. We now show that along this line, D is constantly zero: Let $\boldsymbol{\gamma} = t\mathbf{1}$. Then, from the original definition of D in Equation 25, $D = 2t\mathbf{1}^T(\mathbf{p} - \mathbf{p}') = 2t(\mathbf{1}^T\mathbf{p} - \mathbf{1}^T\mathbf{p}') = 0$. All derivatives of D along that line are therefore constantly zero.[4]

If $\mathbf{w}'$ differs by a small amount from $\mathbf{w}$, then $\boldsymbol{\gamma}$ differs by a small amount from $\mathbf{0}$. The component of the difference along the line $t\mathbf{1}$ has no effect on D; all other components make D negative. Since D is the rate of change in the squared distance between $\mathbf{w}(t)$ and $\mathbf{w}'(t)$, that distance must be stable or decreasing over time. Thus for any $t_2 > t_1$, we have $\|\mathbf{w}(t_2) - \mathbf{w}'(t_2)\| \leq \|\mathbf{w}(t_1) - \mathbf{w}'(t_1)\|$, as claimed. □

To check this result, the simulations from Section 4 were repeated, except this time the initial weights $\mathbf{w}(0)$ were not zero, but were sampled from a normal distribution (mean 0, s.d. 1). Then, for each of those 10,000 simulations, a perturbed mate was made by adding normally distributed noise (mean zero, s.d. 0.1) to $\mathbf{w}(0)$ to get $\mathbf{w}'(0)$. The second-order Taylor approximation to D around $\boldsymbol{\gamma} = \mathbf{0}$ is $D \approx (1/2)\boldsymbol{\gamma}^T H\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = C^T(\mathbf{w}' - \mathbf{w})$. This was used to predict D on each of the 300 update steps for each pair. Results, shown in Figure 2, verify the accuracy of the approximation (and hence corroborate the analysis), and show that D was in every case negative, i.e., that each learner consistently converged with its mate over time.

Figure 2: Actual vs. predicted rate of convergence between pairs in weight space. Values have been multiplied by -1 so that logarithmic axes can be used. Each streak is one of $N = 1000$ pairs (9,000 more omitted to reduce image size).



6 Similar initial weights imply similar 2AFC learning curves

A learner that starts with weights *near* $\mathbf{0}$ follows a trajectory that is close to, and convergent with, that of the learner that started *at* $\mathbf{0}$ (Proposition 3). 2AFC

[3]The Hessian H is -4 times the variance-covariance matrix of the multinomial distribution parametrized by $\mathbf{p}$ (Agresti 1990, 423; Chris Wiesen, p.c., 2017), i.e., the Max Ent distribution parametrized by $\mathbf{w}$. Hence D is approximately a scalar multiple of the variance in the difference between the two learners in harmony, sampled under the distribution of the original learner.

[4]Moving $\boldsymbol{\gamma}$ along the line $\mathbf{1}t$ has the effect of adding the same fixed amount t to the harmony of every candidate. Doing that does not change the model's candidate-probability estimates (e^t cancels in the numerator and denominator of Equation 3), and so it also does not change the model's expectations of the constraint scores, and so it also does not change the update to the weights. Two learners whose weights differ by a multiple of $\mathbf{1}$ are indistiguishable by any experiment. That is a special case of a general consequence of Max-Ent/Replicator equivalence, which is that if two learners assign the same probabilities to all candidates at some time t, they will continue to do so at all later times if exposed to identical training data.

performance in a learner that starts at $\mathbf{0}$ never improves faster than it did at $t = 0$ (Proposition 2). Just how closely is the 2AFC learning curve of the near-$\mathbf{0}$ learner tethered to the bounded learning curve of the at-$\mathbf{0}$ learner?

Proposition 4. *For any* $\mathbf{w}$ *and* $\mathbf{w}'$*, the difference* $\Delta\lambda(\mathbf{w}, \mathbf{w}') = E_{\mathbf{w}'}[\lambda_{+,-}] - E_{\mathbf{w}}[\lambda_{+,-}]$ *is bounded by* $|\Delta\lambda| \leq \|\mathbf{w}' - \mathbf{w}\|\sqrt{m}c_{range}$*, where* c_{range} *is the largest absolute difference between any two entries in* C*.*

Proof. From Equation 14, the difference between the two learners in the expected log-odds of a correct test response is given by $\Delta\lambda(\mathbf{w}, \mathbf{w}') = (\mathbf{w}' - \mathbf{w})^T C(\mathbf{r}^+ - \mathbf{r}^-)$. By Cauchy-Schwarz, this is no greater than $\|\mathbf{w}' - \mathbf{w}\|\|C(\mathbf{r}^+ - \mathbf{r}^-)\|$. Each entry in $C(\mathbf{r}^+ - \mathbf{r}^-)$ is the difference between the average scores of the positive versus negative test stimuli on one of the constraints, which is at most c_{range}. Thus $\|C(\mathbf{r}^+ - \mathbf{r}^-)\| \leq \sqrt{m}c_{\text{range}}$. □

Since $\|\mathbf{w}'(t) - \mathbf{w}(t)\| \leq \|\mathbf{w}'(0) - \mathbf{w}(0)\|$ by Proposition 3, the 2AFC learning curve of a learner that started at $\mathbf{w}'(0) \neq \mathbf{0}$ cannot stray further than $\|\mathbf{w}'(0)\|\sqrt{m}c_{\text{range}}$ from one that started at $\mathbf{0}$.

In actual practice, the experimenter will often divide the candidates into positive and negative test sets in a controlled way, so that the two sets receive, on average, the same score from all but some small number $m^\star$ of the m constraints, which we can suppose are Constraints c_1 through $c_{m^\star}$. Since the ith entry of $C(\mathbf{r}^+ - \mathbf{r}^-)$ is the average difference between the positive and the negative test sets in their score on the ith constraint, all but the first $m^\star$ of the entries will be as close to zero as the experimenter is able to arrange. In that case, we can truncate $\mathbf{w}$, $\mathbf{w}'$, and C to their first $m^\star$ entries or rows, tightening the bound to $|\Delta\lambda| \leq \|\mathbf{w}'^\star - \mathbf{w}^\star\|\sqrt{m^\star}c^\star_{\text{range}}$.

Proposition 5. *Let* $M = \max_i \|C_{i,\cdot}\|$ *be the norm of the row of* C *with the largest norm, and let* $R = \max_i |C_{i,\cdot}\mathbf{1}|$ *be the largest absolute row sum in* C*. Then the difference between the initial rate of 2AFC improvement of a learner that starts at* $\mathbf{w} = \mathbf{0}$ *and one that starts at* $\mathbf{w}'$ *near* $\mathbf{0}$ *is bounded by*

$$|d\Delta\lambda/dt|_{\mathbf{w}=\mathbf{0}} \leq \|\mathbf{w}'\| \left(\frac{M^2}{n} + \frac{R^2}{n^2}\right) m\sqrt{m^\star}c_{range} \tag{32}$$

Proof. Use Equation 19, approximating $\mathbf{p}' - \mathbf{p} \approx JC^T(\mathbf{w}' - \mathbf{w})$, where $J_{i,j} = \partial p_i/\partial\gamma_j$ is the Jacobian of $\mathbf{p}'$ as a function of $\boldsymbol{\gamma}$. From Equation 27 it follows that $J = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T$ (i.e., $J = -H/4$; see Eqn. 31). Then

$$\begin{aligned} |\mathrm{d}\Delta\lambda/\mathrm{d}t| &\approx \left|(\mathbf{w}' - \mathbf{w})^T CJC^T C(\mathbf{r}^+ - \mathbf{r}^-)\right| \\ &\leq \|(\mathbf{w}' - \mathbf{w})^T\|\|CJC^T\|\|C(\mathbf{r}^+ - \mathbf{r}^-)\| \end{aligned} \tag{33}$$

where $\|\cdot\|$, for matrices, is the operator norm, the maximum factor by which the matrix can stretch a vector (Strang, 1980, 284). Since CJC^T is symmetric, its operator norm is simply its largest eigenvalue, which is no larger than ma, where $a = \max_{i,j} |(CJC^T)_{i,j}|$ (Zhan, 2006, Corollary 2). What is a?

For $\mathbf{w} = \mathbf{0}$, we have $J = \frac{1}{n}I - \frac{1}{n^2}\mathbf{1}\mathbf{1}^T$, so $CJC^T = \frac{1}{n}CC^T - \frac{1}{n^2}C\mathbf{1}\mathbf{1}^T C^T$. Then

$$\begin{aligned} \left|\frac{1}{n}(CC^T)_{i,j}\right| &= \frac{1}{n}|C_{i,\cdot} \cdot C_{j,\cdot}| \\ &\leq \frac{1}{n}\max_i |C_{i,\cdot} \cdot C_{i,\cdot}| \\ &\leq \frac{1}{n}M^2 \end{aligned} \tag{34}$$

Likewise,

$$\begin{aligned} \left|\frac{1}{n^2}C\mathbf{1}\mathbf{1}^T C^T\right| &\leq \frac{1}{n^2}\max_{i,j} |(C\mathbf{1})_i \cdot (C\mathbf{1})_j| \\ &\leq \frac{1}{n^2}\max i(C\mathbf{1})_i^2 \\ &\leq \frac{1}{n^2}R^2 \end{aligned} \tag{35}$$

Hence

$$a = \max_{i,j} |(CJC^T)_{i,j}| \leq \frac{1}{n}M^2 + \frac{1}{n^2}R^2 \tag{36}$$

and so, since $\|CJC^T\| \leq ma$,

$$\|CJC^T\| \leq m\left(\frac{1}{n}M^2 + \frac{1}{n^2}R^2\right) \tag{37}$$

Substituting $\|C(\mathbf{r}^+ - \mathbf{r}^-)\| \leq \sqrt{m^\star}c_{\text{range}}$ from the discussion of Proposition 4 completes the proof. □

A row of C corresponds to a constraint, each entry being the score that that constraint gives to one candidate. Since all of those scores are less than $c_{\max}$ (the largest absolute value of any element in C), we have $M^2 \leq nc_{\max}^2$ and $R^2 \leq (nc_{\max})^2$. Hence

$$|\mathrm{d}\Delta\lambda/\mathrm{d}t|_{\mathbf{w}=\mathbf{0}} \leq \|\mathbf{w}'\| m\sqrt{m^\star} c_{\max}^2 c_{\text{range}} \tag{38}$$

This is a worst-case estimate, based on very weak hypotheses about C and on the blunt instrument of the vector and matrix norms, which ignore exploitable structure. Stronger hypotheses permit improvment. For example, suppose C is binary, and let $d_i = \frac{1}{n}C\mathbf{1}$ be the proportion of 1's in Row i. Each entry $(CC^T)_{i,j}$ is the number of 1's that appear in the same column in Rows i and j, and hence is at most the smaller of the two row sums, so $\frac{1}{n}(CC^T)_{i,j} \leq \min\{d_i, d_j\}$. Each entry $(C\mathbf{1}\mathbf{1}^T C^T)_{i,j}$ is the product of the row sums of Rows i and j, so $\frac{1}{n^2}(C\mathbf{1}\mathbf{1}^T C^T)_{i,j} = d_i d_j$. Consequently,

$$\begin{aligned} a = \max_{i,j} |(CJC^T)_{i,j}| &\leq \min\{d_i, d_j\} - d_i d_j \\ &\leq \min\{d_i(1-d_j), d_j(1-d_i)\} \\ &\leq 1/4 \end{aligned} \tag{39}$$

To justify this last step, suppose without loss of generality that $d_i(1-d_j) \leq d_j(1-d_i)$. Then $a^2 = (d_i(1-d_j))^2 \leq d_i(1-d_j)d_j(1-d_i) = d_i(1-d_i)d_j(1-d_j) \leq (1/4)(1/4)$, so unsquaring on both sides yields $a \leq 1/4$. It follows that

$$|\mathrm{d}\Delta\lambda/\mathrm{d}t|_{\mathbf{w}=\mathbf{0}} \leq \|\mathbf{w}'\| \frac{1}{4} m\sqrt{m^\star} \tag{40}$$

Suppose further that the entries of C are modelled as i.i.d. Bernoulli trials with $\Pr(C_{i,j} = 1) = s$. Then $M^2 = \max_i \sum_{j=1}^n C_{i,j}^2 = \sum_{j=1}^n |C_{i,j}| = R$. If n is large, the row sums are approximately samples from a normal distribution with mean ns and standard deviation $\sqrt{ns(1-s)}$. The expected value of the maximum of a sample of size m from the standard normal distribution $N(0,1)$ is approximately $\sqrt{2\log m}$ (Cramér, 1946, 374). Hence $E[M^2] = E[R] = (ns + \sqrt{ns(1-s)2\log m})$. As $n \to \infty$, $E[M^2]/n \to s$, while $E[R^2]/n^2 = (E[R]/n)^2 = (E[M^2]/n)^2 \to s^2$. Thus $\|(CJC^T)\| \to (s + s^2)m = s(1+s)m$, and $c_{\max} = c_{\text{range}} = 1$, so from Equation 40, we have

$$E\left[|\mathrm{d}\Delta\lambda/\mathrm{d}t|_{\mathbf{w}=\mathbf{0}}\right] \leq \|\mathbf{w}'\| m\sqrt{m^\star} \min\left\{\frac{1}{4}, s(1+s)\right\} \tag{41}$$

Equation 41 was checked against 10,000 simulations, generated as described in Section 4. The yoked pairs consisted of one learner that started at $\mathbf{w}(0) = \mathbf{0}$, and one that started at $\mathbf{w}'(0)$ with entries sampled from a normal distribution with mean 0 and standard deviation 0.1. In calculating the bound, $m^\star$ was set equal to m. The actual value was always less than the bound, with the minimum difference being 0.01405. The bound was usually a substantial overestimate, the median difference being 5.372 and the maximum 30.51. In the subset where the near-zero learner's initial performance was near chance ($|\lambda'(0)| \leq 1/10$) and initial improvement was near zero ($|\mathrm{d}\lambda'/\mathrm{d}t| \leq 1/10$), a total of 1075 cases, the bound proved much tighter, overestimating by a median of 0.591 and a maximum of 0.998. These cases tended to have either small m or extreme s.

7 Putting the bounds together

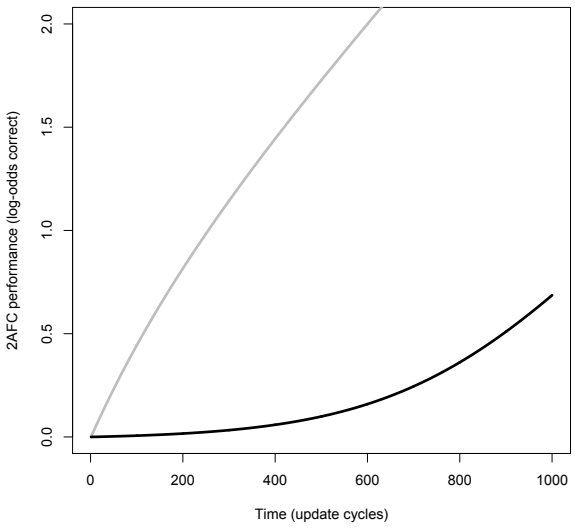
One way that these bounds might be applied in practice is as follows. Suppose we hypothesize that the learner's initial weights $\mathbf{w}'(0)$ are such that $\|\mathbf{w}'(0)\| \leq w$, and we experimentally measure initial performance to be $\lambda'(0) = 0$ and the initial improvment rate to be $d\lambda'/dt(0) = 0$. Proposition 5 then gives us a bound — call it b — on the initial improvement rate for an otherwise identical learner with $\mathbf{w}(0) = \mathbf{0}$. By Proposition 2, the slope of the hypothetical $\mathbf{0}$-learner's 2AFC curve $\lambda(t)$ never exceeds b. That curve would have started at $\lambda(0) = 0$, since $\mathbf{w}(0) = \mathbf{0}$ makes all candidates equally harmonic. Hence the hypothetical $\mathbf{0}$-learner's 2AFC curve is bounded by $\lambda(t) \leq bt$. By Proposition 3, the observed and hypothetical learner converge in weight space, which by Proposition 4 means that $\lambda'(t) \leq bt + w\sqrt{m^\star} c_{\text{range}}$. Conversely, if $\lambda'(t)$ ever exceeds this value, we know that $\mathbf{w}'(0)$ must have been more than w, contrary to hypothesis.[5]

[5] We also have to assume that the experiment has sufficient time resolution that learning cannot begin with an undetectably

8 When initial weights are far from zero, 2AFC performance can accelerate

If the initial weights are far from $\mathbf{0}$, then even a simple constraint set can yield abrupt learning. For $n = 4$, let $C = I_4$, the identity matrix of order 4 (i.e., 4 candidates, 4 constraints, each constraint gives a 1 to just one candidate). If we set $\mathbf{w}(0) = (x, -x, 0, 0)^T$, $\mathbf{p}^+ = \mathbf{r}^+ = (1/2, 1/2, 0, 0)^T$, and $\mathbf{r}^- = (0, 0, 1/2, 1/2)^T$, the 2AFC curve starts out flat at 0 and stays that way (longer the bigger x is), then starts climbing rapidly as shown in the black curve on Figure 3.

Figure 3: 2AFC learning curves for $n = 4$, $C = I_4$, and $\mathbf{w}_0 = (6, -6, 0, 0)^T$ (black curve) or $(6, -6, 6, -6)^T$ (gray curve), with $\mathbf{p}^+ = \mathbf{r}^+ = (1/2, 1/2, 0, 0)^T$, and $\mathbf{r}^- = (0, 0, 1/2, 1/2)^T$. Other parameters: $\eta = 1/100$.



The idea behind this construction is that initially, every negative candidate is much less probable than half of the positive candidates, and much more probable than the other half, so that the outcome of a 2AFC trial is 50% likely to be correct (0 logits). The learner then spends ages laboriously hauling up the low-frequency half of the positive candidates, and letting down the negative candidates, until the low-frequency positive candidates finally start winning a noticeable number of 2AFC competitions. The construction can be carried out for any C that makes it possible to sandwich the initial probabilities of the positive (negative) candidate between those of the negative (positive) ones by artful choice of $\mathbf{w}(0)$.

brief but huge improvment rate that would satisfy the hypothesis of Proposition 2 and thus allow later unexpected sudden improvement.

9 Discussion

Since abrupt learning has been observed in human phonological acquisition in nature Smith (1973); Macken and Barton (1978); Vihman and Velleman (1989); Barlow and Dinnsen (1998); Levelt and van Oostendorp (2007); Gerlach (2010); Guy (2014) and in the lab Moreton and Pertsova (2016), the question of when a gradual learning *rule* translates into gradual learning *performance* is pertinent. For the learner and experimental paradigm studied here, transfer from UG or from previous learning is a necessary condition for abruptness. This result spawns many further questions, among them:

▷ The non-abrupt gray curve in Figure 3 shows that not just any set of large non-zero initial weights, paired with just any training and test distribution, leads to abrupt learning in the model. Which ones do? What are the most general sufficient conditions? Phonological theory offers many proposals about the initial state of L1 or L2 learning (e.g., Demuth (1995); Gnanadesikan (1995); Smolensky (1996); Pater (1997); Broselow et al. (1998); Boersma and Levelt (2000); Curtin and Zuraw (2002); Hayes (2004); Wilson (2006); Hayes et al. (2009); Jesney and Tessier (2011); White (2014)); what predictions follow for abruptness?

▷ In human learners, is abrupt learning associated with transfer of constraint weights from UG, L1, or previous training in the lab? What is going on during apparent initial stagnation? Does it actually consist of steady *un*learning of a pre-existing grammar?

▷ Do the present results extend to other learners that are algorithmically related to this one? That is a sizable class, including not only elaborations of Max Ent gradient ascent, but also the Gradual Learning Algorithms for Stochastic OT and Harmonic Grammar (recent reviews: Boersma and Pater (2016); Pater (2016); Jarosz (2016)). Abrupt learning has been seen in some of them (see Introduction above). Differences in the conditions under which they admit abrupt learning may provide a hitherto unused way to them empirically.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley Interscience.

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.

Berger, A. L., S. A. Della Pietra, and V. J. Della Pietra (1996). A maximum entropy approach to natural language processing. *Computational Linguistics 22*(1), 39–71.

Boersma, P. (1998). *Functional Phonology: formalizing the interactions between articulatory and perceptual drives*. Ph. D. thesis, University of Amsterdam.

Boersma, P. and B. Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry 32*, 45–86.

Boersma, P. and C. Levelt (2000). Gradual constraint-ranking learning algorithm predicts acquisition order. In *Proceedings of Child Language Research Forum 30*, Stanford, California, pp. 229–237.

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.

Boyd, S. and L. Vandenberghe (1999). *Convex optimization*. Cambridge University Press.

Broselow, E., S.-I. Chen, and C. Wang (1998). The emergence of the unmarked in second language phonology. *Studies in Second Language Acquisition 20*, 261–280.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, New Jersey: Princeton University Press.

Curtin, S. and K. R. Zuraw (2002). Explaining constraint demotion in a developing system. In B. Skerabela, S. Fish, and A. H.-J. Do (Eds.), *Papers from the 26th Boston University Conference on Language Development (BUCLD 26)*, Somerville, pp. 118–129. Cascadilla Press.

Demuth, K. (1995). Markedness and the development of prosodic structure. In J. Beckman (Ed.), *Proceedings of the 25th Meeting of the North-East Linguistics Society*, Amherst, Mass., pp. 13–26. Graduate Linguistics Students Association.

Elman, J. L., E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett (1996). *Rethinking innateness*. Cambridge, Massachusetts: MIT Press.

Fischer, M. (2005). A Robbins-Monro type learning algorithm for an entropy maximizing version of Stochastic Optimality Theory. Master's thesis, Humboldt-Universität, Berlin.

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

Gnanadesikan, A. (1995, October). Markedness and faithfulness constraints in child phonology. Manuscript # 67, Rutgers Optimality Archive (`roa.rutgers.edu`).

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.

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

Hayes, B. (2004). Phonological acquisition in Optimality Theory: the early stages. In R. Kager, J. Pater, and W. Zonneveld (Eds.), *Constraints in phonological acquisition*, Chapter 5, pp. 158–203. Cambridge, England: Cambridge University Press.

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

Hayes, B., K. Zuraw, P. Siptár, and Z. Londe (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language 85*(4), 822–863.

Horn, R. A. and C. R. Johnson (1985). *Matrix analysis*. Cambridge, England: Cambridge University Press.

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language 59*, 434–446.

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.

Jarosz, G. (2016). Learning with violable constraints. To appear in: Jeff Lidz, William Snyder, and Joe Pater (eds.), *The Oxford handbook of developmental linguistics*. Oxford, England: Oxford University Press.

Jesney, K. (2016). On the relationship between learning sequence and rate of acquisition. In G. lafur Hansson, A. Farris-Trimble, K. McMullin, and D. Pulleyblank (Eds.), *Proceedings of the Annual Meeting on Phonology 2015*, Volume 3. Linguistic Society of America.

Jesney, K. and A.-M. Tessier (2011). Biases in Harmonic Grammar: the road to restrictive learning. *Natural Language and Linguistic Theory 29*(1), 251–290.

Johnson, M. (2007, November). A gentle introduction to Maximum Entropy models and their friends. Slides from a talk, accessed at `web.science.mq.edu.au/~mjohnson/papers/CompPhon07-slides.pdf` on 2013 August 6.

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.

Levelt, C. and M. van Oostendorp (2007). Feature co-occurrence constraints in L1 acquisition. *Linguistics in the Netherlands 24*(1), 162–172.

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

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.

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

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.

Pater, J. (1997). Minimal violation in phonological development. *Language Acquisition 6*(3), 201–253.

Pater, J. (2008). Gradual learning and convergence. *Linguistic Inquiry 39*(2), 334–345.

Pater, J. (2016). Universal Grammar with weighted constraints. To appear in: John McCarthy and Joe Pater (eds.), Harmonic Grammar and Harmonic Serialism.

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.

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

Smolensky, P. (1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry 27*, 720–731.

Strang, G. (1980). *Linear algebra and its applications*. Orlando, Florida: Academic Press.

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

White, J. (2014). Evidence for a learning bias against saltatory phonological alternations. *Cognition 130*, 96–115.

Wilson, C. (2006). Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science 30*(5), 945–982.

Zhan, X. (2006). Extremal eigenvalues of real symmetric matrices with entries in an interval. *SIAM Journal of Matrix Analysis and Applications 27*(3), 851–860.