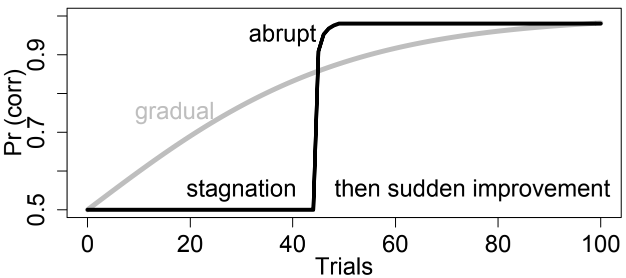
Conditions on abruptness in a gradient-ascent Max Ent learner

SCiL, 2017.01.04 *Elliott Moreton, University of North Carolina, Chapel Hill* Salt Lake City

Q: When does an incremental learning algorithm yield abrupt learning performance?



A: A gradient-ascent Max Ent learner needs **nonzero initial weights** for abrupt improvement in this two-alternative forced-choice experiment — i.e., **abruptness is a transfer effect**.

1. The learning scenario

A. Grammatical model: Basic Max Ent [6]:

Constraints	$\{c_1, \ldots, c_m\}$
Weights	$(w_1, \ldots, w_m) = \mathbf{w}$
Candidates	$\{x_1, \ldots, x_n\}$
Harmonies	$h_{\mathbf{w}}(x_j) = \sum_{i=1}^m w_i c_i(x_j)$
Probabilities	$\Pr(x_j \mid \mathbf{w}) = p_j = \dfrac{\exp(h_{\mathbf{w}}(x_j))}{\sum_{j=1}^n \exp(h_{\mathbf{w}}(x_j))}$

B. Training: Positive (= legal) stimuli from empirical distribution $\mathbf{p}^+$. Gradual update using the Delta Rule:

$$\Delta w_i = \eta(E_{\mathbf{p}^+}[c_i] - E_{\mathbf{w}}[c_i]) \qquad (1)$$

i.e., batch-mode gradient ascent on log-likelihood [6].

C. Testing: Two-alternative forced-choice (2AFC) test using the Luce choice rule [9]:

$$\Pr(x_i^+ | (x_i^+, x_j^-)) = \frac{p_i}{p_i + p_j} \qquad (2)$$

with the positive and negative test items (candidates) sampled from complementary flat distributions $\mathbf{r}^+$ and $\mathbf{r}^-$:

$$\begin{aligned} \mathbf{r}^+ &= (\tfrac{1}{k}, \ \ldots, \ \tfrac{1}{k}, \ 0, \ \ldots, \ 0\ )^T = \mathbf{p}^+ \\ \mathbf{r}^- &= (0, \ \ldots, \ 0, \ \tfrac{1}{n-k}, \ \ldots, \ \tfrac{1}{n-k})^T \end{aligned} \qquad (3)$$

2. Log-likelihood improves non-abruptly...

Let C be a matrix such that $C_{i,j} = c_i(x_j)$, the score that Constraint i gives Simulus j.

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

Furthermore, $dL/dt = \|C(\mathbf{p}^+ - \mathbf{p})\|^2$.

These claims follow straightforwardly from the Replicator representation of the Max Ent learner [11].

3. ...but log-likelihood isn't 2AFC performance

▷ Log-likelihood depends only on the probabilities assigned by the model to the positive stimuli.
▷ 2AFC performance depends as well on the probability mass in the negative stimuli.

4. Main result: If initial weights are 0, abruptness is impossible

Proposition 2. *Suppose that at time $t = 0$, $\mathbf{p}^+ - \mathbf{p}(0) = \alpha(\mathbf{r}^+ - \mathbf{r}^-)$ for some $\alpha > 0$. Let λ be the log-odds of a correct 2AFC response. Then at $t \geq 0$,*

$$\left| \frac{d}{dt} E_{\mathbf{w}}[\lambda] \Big|_t \right| \leq \left| \frac{d}{dt} E_{\mathbf{w}}[\lambda] \Big|_0 \right| \qquad (4)$$

Proof. (Sketch): We show that $\frac{d}{dt}E_{\mathbf{w}}[\lambda] = (C(\mathbf{p}^+ - \mathbf{p}))^T C(\mathbf{r}^+ - \mathbf{r}^-)$. Then by Cauchy-Schwarz,

$$\left| \frac{d}{dt} E_{\mathbf{w}}[\lambda] \right| \leq \underbrace{\|C(\mathbf{p}^+ - \mathbf{p})\|}_{\substack{\text{nonincreasing} \\ \text{by Prop. 1}}} \cdot \underbrace{\|C(\mathbf{r}^+ - \mathbf{r}^-)\|}_{\text{constant}} \qquad (5)$$

with strict equality if and only if $C(\mathbf{r}^+ - \mathbf{r}^-)$ is a scalar multiple of $C(\mathbf{p}^+ - \mathbf{p})$.

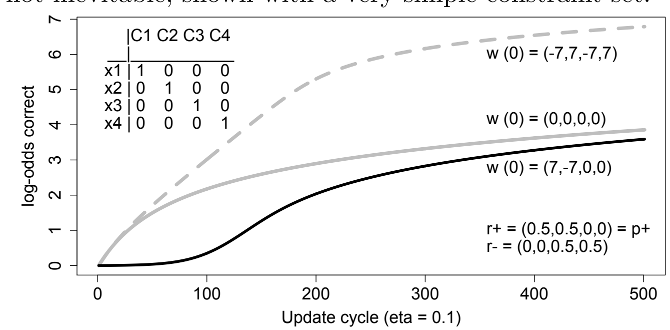
□

A. Application: $\mathbf{w}(0) = \mathbf{0} \Rightarrow \mathbf{p}(0) = (1/n, \ldots, 1/n)^T$. Then by Eqn. 3, $C(\mathbf{r}^+ - \mathbf{r}^-) = C(\mathbf{p}^+ - \mathbf{p}) \cdot (n-k)/n$, so by Prop. 2, **2AFC performance improves fastest at $t = 0$.**

B. Generalization: In weight space, learners that start *near* and *at* $\mathbf{0}$ converge monotonically. We derive bounds on 2AFC difference as a function of initial weight-space distance.

5. Abruptness can happen with non-zero initial weights

For nonzero initial weights, abrupt learning is possible but not inevitable, shown with a very simple constraint set:



6. Abruptness = transfer

Abrupt learning happens in acquisition of natural [14, 10, 15, 1, 8, 4, 5] and artificial [12] languages, but when and why? In this case, transfer from UG or previous learning (or noise) is a necessary condition for abruptness.
▷ Is abruptness associated with transfer in humans too? Is apparent initial stagnation really *un*learning of previous grammar?
▷ Not just any non-zero initial weights plus any training and test distribution, leads the model to abrupt learning. Which ones do?
▷ What conditions abruptness in algorithmically related learners [3, 13, 7]?

Thanks to Jen Smith, Joe Pater, Katya Pertsova, and UNC-CH's P-Side caucus. Supported in part by NSF BCS 1651105, "Inside phonological learning", to E. Moreton and K. Pertsova. **QR code for paper:**

The accompanying paper can be found at http://www.unc.edu/∼moreton/Papers/Moreton2018SCiL.pdf. Address for correspondence: moreton@unc.edu.

References

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

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

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

[4] Gerlach, S. R. *The acquisition of consonant feature sequences: harmony, metathesis, and deletion patterns in phonological development.* PhD thesis, University of Minnesota, 2010.

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

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

[7] Jarosz, G. 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, 2016.

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

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

[10] Macken, M. A., and Barton, D. 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, March 1978.

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

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

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

[14] Smith, N. *The acquisition of phonology: a case study.* Cambridge University Press, Cambridge, England, 1973.

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