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 \(w_1, \ldots, w_m = w\)
Candidates \(x_1, \ldots, x_n\)
Harmonies \(h_w(x_j) = \sum_{i=1}^n w_i c_i(x_j)\)
Probabilities \(Pr(x_j) = \frac{\exp(h_w(x_j))}{\sum_{j=1}^n \exp(h_w(x_j))}\)

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

\[ \Delta w_i = \eta(E_{p^+}[c_i] - E_{w[c_i]}), \]

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

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

\[ Pr(x_j^+|(x_l^+, x_j)) = \frac{p_j}{p_l + p_j} \]

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

\[ r^+ = (\frac{1}{2}, \ldots, \frac{1}{2}, 0, \ldots, 0)^T = p^+ \]
\[ r^- = (0, \ldots, 0, \frac{1}{n-1}, \ldots, \frac{1}{n-1})^T \]

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 Stimulus \( j \).

**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, \( dL/dt = ||C(p^+ - p)||^2 \).

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

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

\( \triangleright \) Log-likelihood depends only on the probabilities assigned by the model to the positive stimuli.

\( \triangleright \) 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 \( t = 0 \), \( p^+ - p(0) = \alpha(r^+ - r^-) \) for some \( \alpha > 0 \). Let \( \lambda \) be the log-odds of a correct 2AFC response. Then at \( t \geq 0 \),

\[ \left| \frac{d}{dt} E_w[\lambda] \right|_{|t} \leq \left| \frac{d}{dt} E_w[\lambda] \right|_{|0} \]

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

\[ \left| \frac{d}{dt} E_w[\lambda] \right|_{|t} \leq \|C(p^+ - p)\| \cdot \|C(r^+ - r^-)\| \]

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

□

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.

\( \triangleright \) Is abruptness associated with transfer in humans too? Is apparent initial stagnation really unlearning of previous grammar?

\( \triangleright \) Not just any non-zero initial weights plus any training and test distribution, leads the model to abrupt learning. Which ones do?

\( \triangleright \) What conditions abruptness in algorithmically related learners [3, 13, 7]?

**A. Application:** \( w(0) = 0 \Rightarrow p(0) = (1/n, \ldots, 1/n)^T \).

Then by Eqn. 3, \( C(r^+ - r^-) = C(p^+ - p) \cdot (n-k)/n \), so by Prop. 2, 2AFC performance improves fastest at \( t = 0 \).

**B. Generalization:** In weight space, learners that start near and at 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:

\[ \begin{align*}
C1 & : x_1 \leq 0, x_2 \leq 0, x_3 \leq 0, x_4 \leq 0, \\
C2 & : x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0, \\
C3 & : x_1 \leq x_2, x_3 \leq x_4, \\
C4 & : x_1 \geq x_2, x_3 \geq x_4.
\end{align*} \]

\[ \frac{d}{dt} E_w[\lambda] \leq \left| \frac{d}{dt} E_w[\lambda] \right|_{|0} \]

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


