MaxEnt Learners are Biased Against Giving Probability to Harmonically Bounded Candidates

Charlie O'Hara
cohara@umich.edu

1. Overview

MaxEnt Harmonic Grammar (MaxEnt) assigns some probability to harmonically bounded candidates (Jesney 2007)

- A candidate in a tableau is **harmonically bounded** if there exists another candidate that performs at least as well on every constraint (and better on at least one)
- OT, HG, Noisy HG can not select harmonically bounded candidates.

/CV/	Onset w=1	Max w=1	Harm	Prob
[CV]			0	.88
[V]	-1	-1	-2	.12

The typological predictions of a grammar are not only influenced by what is *generable,* but what is *learnable*.

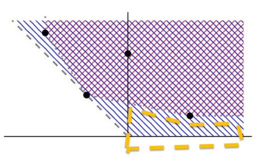
CLAIM: A learning bias exists in the truncated Perceptron algorithm leading to harmonically bounded candidates receiving little probability.

- Learners converge toward a different grammar than their target pattern
- MaxEnt's typological picture is similar to Noisy HG

2. Harmonically Bounded Candidates in MaxEnt

MaxEnt generates a different set of possible patterns of variation than Noisy HG. (Jesney 2007, Anttila & Magri 2018, Magri 2018, Anttila et al 2019)

- Assigning any probability to harmonically bounded candidates is a major difference
- Magri & colleagues focus on *t-orders:* *(x→y) → (x'→y')*



p(x→y) ≤ p(x'→y') for all weights

In Noisy HG, the t-order holds if the difference vectors between x', y' and all candidates z' fall in the blue region of the figure above (generated by the difference vectors of x, y, and candidates for x.)
In MaxEnt, the t-order holds if:

- /x/ and /x'/ have the same number of output candidates
- /x'/-[y']-[z'] all fall in the red region.

Any candidate whose difference vector falls in the top right region must be harmonically bounded by /x'/-[y']

- Bounded candidates receive more probability when the bounding constraints are weighted near 0.

If MaxEnt is biased against giving probability to harmonically bounded candidates, the impact of these candidates shrinks.

Collectively Harmonically Bounded candidates act differently, happy to talk more

3. Truncated Perceptron

The Truncated Perceptron algorithm is a stochastic gradient ascent algorithm that is restricted to non-negative constraint weights. (Magri 2015, Pater 2008, Jäger 2003)

Update Rule

$$\Delta w_C = \mu(C(x, y_T) - C(x, y_L))$$

- Sample an input at random
- Sample outputs based on the probabilities in both the teacher and learner's MaxEnt grammar
  - Update weights:
  - Increase weights of teacher favoring constraints
  - Decrease weights of learner favoring constraints
  - Reset all negative weights to zero.

4. Generational Stability

Generational Simulations (Staubs 2014, Hughto 2019, O'Hara 2021)

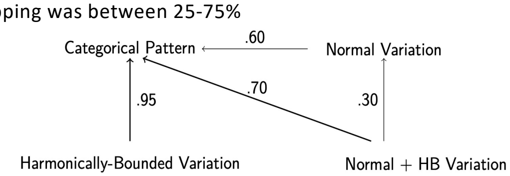
- Agent-based simulation
- 15 generations exposed to 5000 forms from previous gen
- All learners initialized with Markedness at 50, Faithfulness at 1.

(Staubs 2014, Hughto 2019, O'Hara 2021)

Three Patterns of Onset/Coda Typology using Max, Dep, NoCoda and Onset.

- Normal Variation (/V/→[CV] 50%, [V] 50%)
- Harmonically Bounded Variation (/CV/→[CV] 50%, [V] 50%)
- Normal + HB Variation (50% onset and coda deletion)

Classified final generation learners as variable if probability for a mapping was between 25-75%

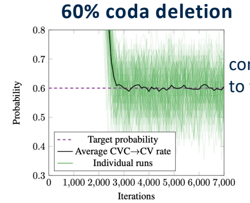


3. Truncated Perceptron

The average learner converges to the target pattern in normal variation, but converges to the wrong pattern in harmonically bounded variation

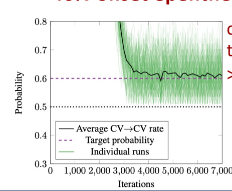
- Simulated 100 learners trained with both:

60% coda deletion



converges to target

40% onset epenthesis



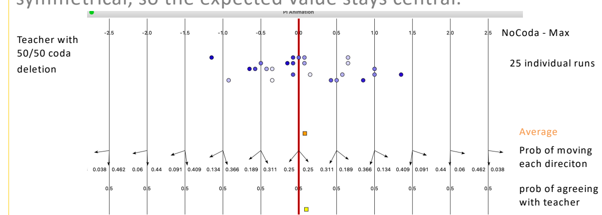
converges to prob > target

6. Discussion

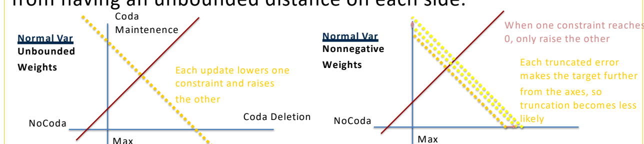
Why does MaxEnt converge incorrectly to HB-variation?
Perceptron *weakly converges* for MaxEnt (Fischer 2006).

- Each individual learner continues to oscillate around variable targets.
- The expected value converges as the number of iterations increases

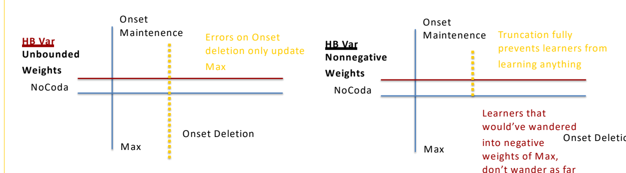
In the simplest case: 50%-50% variation, uniform error size, no bounds on constraint weights, two candidates, oscillation is symmetrical, so the expected value stays central.



The restriction to non-negative constraint weights prevents learners from having an unbounded distance on each side.



Truncation has a different effect on normal vs. HB variation.
In normal variation as iterations increase, the probability of truncated updates approaches zero, so convergence resembles a system without truncation.



Truncation has a more significant effect on HB variation.

- Learners that were "supposed" to go below the axis cannot, and have higher weights of Max, failing to counteract the learners with high weights of Max.

If Harmonically Bounded Candidates are available to an online MaxEnt learner--- even if they are observed in data--- learners will be biased against near zero constraint weights.
Learning bias against harmonically bounded candidates