MaxEnt Learners are Biased Against Giving Probability to Harmonically Bounded Candidates

Charlie O'Hara
University of Michigan, charleso@usc.edu

Follow this and additional works at: https://scholarworks.umass.edu/scil

Part of the Computational Linguistics Commons, Phonetics and Phonology Commons, and the Typological Linguistics and Linguistic Diversity Commons

Recommended Citation
DOI: https://doi.org/10.7275/sc8a-rf84
Available at: https://scholarworks.umass.edu/scil/vol5/iss1/24

This Extended Abstract is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Proceedings of the Society for Computation in Linguistics by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
MaxEnt Learners are Biased Against Giving Probability to Harmonically Bounded Candidates

Charlie O’Hara
Department of Linguistics
University of Michigan
cohara@umich.edu

1 Overview

One of the major differences between MaxEnt Harmonic Grammar (Goldwater and Johnson, 2003) and Noisy Harmonic Grammar (Boersma and Pater, 2016) is that in MaxEnt harmonically bounded candidates are able to get some probability, whereas in most other constraint-based grammars they can never be output (Jesney, 2007). The probability given to harmonically bounded candidates is taken from other candidates, in some cases allowing MaxEnt to model grammars that subvert some of the universal implications that are true in Noisy HG and categorical forms of HG (Anttila and Magri, 2018). Magri (2018) argues that the types of implicational universals that remain valid in MaxEnt are phonologically implausible, suggesting that MaxEnt overgenerates Noisy HG in a problematic way.

However, a variety of recent work has shown that some of the possible grammars in a constraint based grammar may be unlikely to be observed because they are difficult to learn (Staubs, 2014; Stanton, 2016; Pater and Moreton, 2012; Hughto, 2019; O’Hara, 2021). Here, I show that grammars that give too much weight to harmonically bounded candidates, and violate the implicational universals that hold in Noisy HG are significantly harder to learn than those grammars that are also possible in Noisy HG. With learnability applied, I claim that the typological predictions of MaxEnt and Noisy HG are in fact much more similar than they would seem based on the grammars alone. This paper focuses on the classically harmonically bounded candidates, because collectively bounded candidates reflect a different type of constraint weighting, and are more often observed typologically (see local optionality Riggle and Wilson (2005); Hayes (2017)).

2 The Problem

Anttila and Magri (2018) show that MaxEnt overpredicts Noisy HG. Specifically, given a specific set of constraints, there are probabilistic universals in Noisy HG that are not maintained in MaxEnt; in other words for all Noisy HG grammars the probability of one mapping \( /x/ \rightarrow [y] \) is always less than or equal to the probability of some other mapping \( /\hat{x}/ \rightarrow [\hat{y}] \), but in MaxEnt the former mapping can be more probable. They characterize the difference between MaxEnt and Noisy HG geometrically, showing that the probabilistic universals generated by Noisy HG are a superset of those generated by MaxEnt for any particular set of tableaux.

Figure 1 shows an example of this difference in a system with two constraints. Each node represents a difference vector between the antecedent mapping \( /x/ \rightarrow [y] \) and one of its competitors \( /\hat{x}/ \rightarrow [\hat{z}] \), calculated by subtracting the violations of \( /\hat{x}/ \rightarrow [\hat{z}] \) from \( /x/ \rightarrow [y] \) (assuming violations are counted negatively). Anttila and Magri (2018) show that in order for some consequent mapping \( /\hat{x}/ \rightarrow [\hat{y}] \) to never receive a lower probability than the antecedent \( /x/ \rightarrow [y] \) mapping under all weightings of constraints, all difference vectors between \( /\hat{x}/ \rightarrow [\hat{y}] \) and its competitors must have fall in the region greater than the convex hull generated by the antecedent difference vectors in MaxEnt (corresponding...
Table 1: Universal-Subverting Pattern in MaxEnt

<table>
<thead>
<tr>
<th>/CV/</th>
<th>ONSET</th>
<th>DEP</th>
<th>HARM</th>
<th>PROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weights</td>
<td>w = 2</td>
<td>w = 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. CV</td>
<td>-1</td>
<td>0</td>
<td>-5</td>
<td>0.05</td>
</tr>
<tr>
<td>b. V</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>0.95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>/V/</th>
<th>ONSET</th>
<th>DEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weights</td>
<td>w = 2</td>
<td>w = 5</td>
</tr>
<tr>
<td>a. CV</td>
<td>-1</td>
<td>-5</td>
</tr>
<tr>
<td>b. V</td>
<td>-1</td>
<td>-2</td>
</tr>
</tbody>
</table>

Candidate violates a superset of the violations of some other candidate. If the candidate is bounded by the target mapping, the difference vector between the target vector and the candidate will be non-negative for all constraints, placing it in the first quadrant (top right) of the graph. If a candidate is harmonically bounded by some other candidate, it will be at least as large (component by component) than the candidate that harmonically bounds it. The second case is less problematic in MaxEnt because if /x/-[y]-[z] harmonically bounds /x/-[y]-[z], and the difference vector for [z] falls outside of the cross hatched region for some antecedent vector, so must the difference vector for [z]. On the other hand, as seen in the example above, when the target mapping harmonically bounds a candidate, that candidate can fall in the first quadrant, but below the convex hull generated by the set of antecedent difference vectors. We can see that a large portion of the difference vectors that behave differently in MaxEnt and Noisy HG are of this subtype—they fall in the region in the first quadrant under the convex hull. Harmonically bounded candidates only receive probability under certain restricted weighting conditions—as the weight of the harmonically bounded constraints increases, the probability assigned to candidates bounded by those constraints becomes vanishingly small. If not all weighting conditions are equally easy to learn, is it possible that it is particularly hard to learn constraint weightings that would assign a significant probability to harmonically bounded candidates?}

---

1As long as the number of competitors for the antecedent and consequent are the same.

2So that this system can be represented two-dimensionally, here I am excluding MAX, as well any constraints or candidates with codas. These will be introduced later in the paper for the simulations. This situation is the same as if MAX was weighted zero, and NOCODA was weighted very high.
Table 2: Patterns under consideration

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>/CV/</td>
<td>1 0 0 0</td>
</tr>
<tr>
<td>/V/</td>
<td>0 1 0 0</td>
</tr>
<tr>
<td>/CVC/</td>
<td>0 0 1 0</td>
</tr>
<tr>
<td>/VC/</td>
<td>0 0 0 1</td>
</tr>
</tbody>
</table>

(2a), one variable grammar that is consistent with the implicational universals (2b), and one variable grammar that subverts the implicational universals (2c). Notably, only the last pattern gives any probability to harmonically bounded candidates.

4 Simulation Results

The simulations show that the categorical patterns are learned most consistently, followed by the universal-respecting variation patterns. The universal-subverting patterns available only in MaxEnt are learned consistently worse than the other types of patterns on multiple metrics. First, the universal subverting patterns require much more data to be learned accurately, as shown by the number of iterations it took to learn the pattern on average in the first generation (Table 3). Further we can look the end result of the 20 runs performed for each simulation to see how stably the pattern is learned across generations, which allows us to see how likely a pattern is to change, and how likely a pattern is to be innovated. Figure 3 presents the results after fifteen generations, classified according to what the initial target pattern was, and what the pattern the final generation learned would be classified as. It can be seen that the categorical pattern is learned fully stably under these parameters; whereas the universal respecting variation changes in 12 of the 20 runs, often reducing the variability of the pattern. Finally, the universal subverting patterns are learned very unstably, changing into a type of pattern that can be modeled in Noisy Harmonic Grammar in all 20 runs.

Table 3: # of iterations needed to learn each pattern.

<table>
<thead>
<tr>
<th>Grammar Type</th>
<th>Iterations Needed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td>2000</td>
</tr>
<tr>
<td>Respecting</td>
<td>2200</td>
</tr>
<tr>
<td>Subverting</td>
<td>5000</td>
</tr>
</tbody>
</table>
5 Discussion

The universal-subverting patterns are harder to learn because it is necessary for the weights of some constraints to approach zero, rather than simply becoming lower or higher than some conflicting constraint. In this case, the only evidence that would force a constraint close to zero is from observing harmonically bounded candidates in the target grammar. The difficulty learners have learning typology subverting patterns is due to the convergence properties of online MaxEnt learner that restrict constraint weights to non-negative numbers. While this learning algorithm is weakly convergent (Fischer, 2005), I show that the expected weighting of a learner upon convergence differs from the target weighting substantially more when that target weighting has constraint weights close to zero—a necessary property of typology subverting variation patterns, but not typology respecting variation.

When learning a variable pattern, individual learners do not ever stop updating, because even if the learner and teacher have the same grammar, errors still occur. Each individual learner ends up oscillating around the target pattern. When this variation is symmetrical, the average across many learners converges to the target pattern. However, when the target pattern requires a constraint being weighted particularly close to zero, learners oscillate asymmetrically—some learners

This learning bias is of a stronger sort than many considered in the learning literature, rather than simply requiring more time to converge, learners trained on typology subverting patterns converge on a grammar different from the target grammar. To demonstrate a basic example of how the learning algorithm converges more accurately to normal variation than harmonically bounded variation, I ran 100 learners on two variable patterns. In the normal variation pattern, onset consonants neither epenthesized or deleted (100% faithful) and coda consonants deleted 60% of the time. In the harmonically bounded variation pattern, coda consonants deleted categorically, but onsets deleted 40% of the time. Each simulation ran according to the parameters of the above simulations. Figures 4 and 5 show the results of these simulations. The dark black line represents the average probability of the target variable mapping across all 100 learners, whereas the lighter green lines represent each individual run. The dashed gray line shows the target probability of the mapping. In normal variation (Figure 5), the learners oscillate symmetrically around the target pattern, with the average staying very close to the target probability. In harmonically bounded variation (Figure 6), the average remains notably above the target probability. Harmonically bounded variation acts differently because learners assigning less probability to the target mapping end up “bouncing” off of a wall, because the harmonically bounded CV→V mapping can never receive more than 50% because constraint weights must remain nonnegative.

If phonological learners are biased against assigning probability to harmonically bounded candidates even when weightings exist in MaxEnt that assign probability to them, a major source of typological difference between MaxEnt and Noisy HG appears to be less significant. Future work will investigate the other geometrical regions of difference between MaxEnt and Noisy HG, and explore whether they also require very low constraint weights that are difficult to learn.
Acknowledgments

Thanks to Karen Jesney, Caitlin Smith, Rachel Walker and audiences at USC’s PhonLunch, AMP 2019, and University of Michigan’s Phondi meetings for comments on this project.

References


