MaxEnt Learners are Biased Against Giving Probability to Harmonically Bounded Candidates

Charlie O'Hara
Department of Linguistics
Univerisity 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 over-predicts 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/→[y]) is always less than or equal to the probability of some other mapping (/$\hat{x}$/→[$\hat{y}$]), 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: In order for a particular mapping /x/→[y] to be always assigned a lower or equal probability than the mapping /$\hat{x}$/→[$\hat{y}$]: in Noisy HG all difference vectors between /$\hat{x}$/→[$\hat{y}$] and its competitors must fall in the dashed region, whereas in MaxEnt, they must fall in the crosshatched region. The dots represent the difference vectors of /x/→[y] compared to its competitors. Adapted from Anttila and Magri (2018).

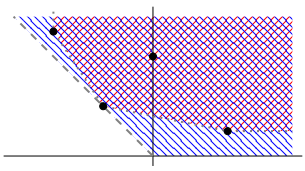


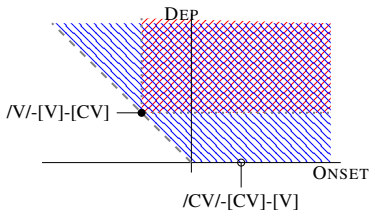
Figure 1 shows an example of this difference in a system with two constraints. Each node represents a difference vector between the antecedent mapping /x/→[y] and one of its competitors /x/-[z], calculated by subtracting the violations of /x/→[z] from /x/→[y] (assuming violations are counted negatively).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

Proceedings of the Society for Computation in Linguistics (SCiL) 2022, pages 229-234.
Held on-line February 7-9, 2022

Table 1: Universal-Subverting Pattern in MaxEnt

/CV/	ONSET	DEP		
Weights	$w = 2$	$w = 5$	HARM	PROB
a. CV			0	0.88
b. V	-1		-2	0.12
/V/	ONSET	DEP		
Weights	$w = 2$	$w = 5$	HARM	PROB
a. CV		-1	-5	0.05
b. V	-1		-2	0.95

Figure 2: Geometric representation of the onset typology with DEP and ONSET.



ing to the crosshatched region).[1] In Noisy HG, the consequent's difference vectors can fall anywhere in the region greater than the convex *cone* generated by the antecedent difference vectors (also including the dashed regions). Here, I will argue that many of these cases are caused by the fact that MaxEnt assigns probability to harmonically bounded candidates but Noisy HG does not.

A simple concrete example emerges in syllable structure—using the constraints and candidates in Table 1, it is quite obvious in noisy HG that onsetful syllables map faithfully (/CV/-[CV]) at least as often as onsetless syllables do (/V/-[V]), since /CV/-[CV] harmonically bounds its competitor. However, in MaxEnt it is possible for the onsetless faithful mapping to receive more probability than the onsetful mapping, see Table 1.[2] This difference between MaxEnt and Noisy HG is directly caused by the harmonically bounded candidate /CV/-[V] being able to take probability from the /CV/-[CV] mapping only in MaxEnt. This type of *classically harmonically bounded* candidate can only receive any probability when the bounding constraints (here MAX and ONSET) are sufficiently low-weighted. This difference is geometrically represented in Figure 2. The filled dot represents the difference vector between /V/→[V] and /V/→[CV], whereas the unfilled dot represents the difference vector between /CV/→[CV] and /CV/→[V]. Crucially, the unfilled dot falls only in the dashed region, but not the crosshatched region.

Harmonically bounded candidates show particular geometric properties. A harmonically bounded 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]-[$\hat{z}$], and the difference vector for [$\hat{z}$] 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.[3] 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?

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

[2] So 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.

[3] There are two other regions that differentiated MaxEnt and Noisy HG—in this two-dimensional representation, the triangle generated by the origin, the y-axis and the left edge of the convex cone, and the triangle generated by the leftmost difference vector, the left edge of the cone, and the left edge of the region larger than the convex hull. I save characterization of these regions for future work.

2a. Categorical Pattern				
	Output			
Input	[CV]	[V]	[CVC]	[VC]
/CV/	1	0	0	0
/V/	0	1	0	0
/CVC/	0	0	1	0
/VC/	0	0	0	1
2b. Universal Respecting Pattern				
	Output			
Input	[CV]	[V]	[CVC]	[VC]
/CV/	1	0	0	0
/V/	.5	.5	0	0
/CVC/	.5	0	.5	0
/VC/	.25	.25	.25	.25
2c. Universal Subverting Pattern				
	Output			
Input	[CV]	[V]	[CVC]	[VC]
/CV/	.5	.5	0	0
/V/	0	1	0	0
/CVC/	.25	.25	.25	.25
/VC/	0	.5	0	.5

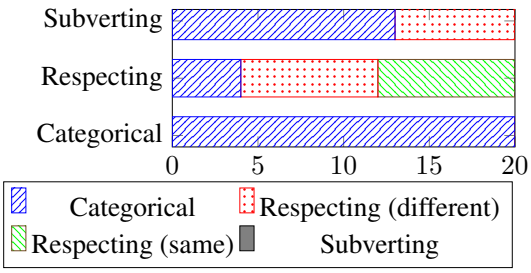
Table 2: Patterns under consideration

3 Learnability

To evaluate the learnability of different classes of grammars, I make us of agent-based generational learning simulations (Kirby and Huford, 2002; Kirby, 2017). These simulations make use of a series of learning agents using the Perceptron learning algorithm (Rosenblatt, 1958; Jäger, 2003; Boersma and Pater, 2016); each initialized following conventional assumptions in the phonological learning literature (i.e. markedness constraints weighted high faithfulness low (Gnanadesikan, 2004; Tesar and Smolensky, 2000; Jesney and Tessier, 2011)). Learners are exposed to a limited number of input-output mappings randomly chosen from their target grammar (each underlying syllable type is sampled equally frequently, surface forms sampled according to the target grammar). After the learner is exposed to the number of forms (here 7000 forms per generation), the learner *matures* and whatever grammar it learned is used as the target grammar for the next learner. Each run of the simulation consists of 15 generations, with the first generation exposed to whatever grammar is being tested.

Three types of patterns were tested: one fully categorical pattern available in MaxEnt and Noisy HG (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.

Figure 3: Resulting patterns after 15 generations.



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.

Grammar Type	Iterations Needed
Categorical	2000
Respecting	2200
Subverting	5000

Figure 4: 100 learners trained on normal variation with 60% coda deletion.

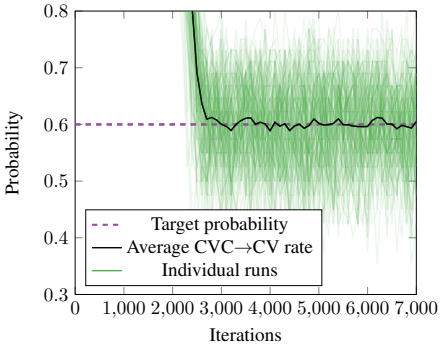
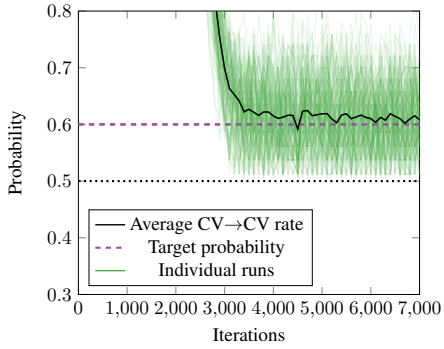


Figure 5: 100 learners trained on harmonically bounded variation with 40% onset deletion



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 cannot oscillate symmetrically around the target pattern—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

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Arto Anttila and Giorgio Magri. 2018. Does maxent overgenerate? implicational universals in maximum entropy grammar. In *Proceedings of the 2017 Annual Meeting on Phonology*. Linguistics Society of America.

Paul Boersma and Joe Pater. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In John J. McCarthy and Joe Pater, editors, *Harmonic Grammar and Harmonic Serialism*. Equinox.

Benjamin Börschinger and Mark Johnson. 2011. A particle filter algorithm for Bayesian wordsegmentation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 10–18, Canberra, Australia.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Markus Fischer. 2005. A Robbins-Monro type learning algorithm for an entropy maximizing version of stochastic optimality theory. Master's thesis, Humboldt University, Berlin.

Amalia Gnanadesikan. 2004. Markedness and faithfulness in child phonology [ROA-67]. In René Kager, Joe Pater, and Wim Zonneveld, editors, *Fixing Priorities: Constraints in Phonological Acquisition*, pages 73–108. Cambridge University Press, Cambridge.

Sharon Goldwater and Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory*. Stockholm University.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 1, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Bruce Hayes. 2017. Varieties of noisy harmonic grammar. In *Proceedings of the 2016 Annual Meeting on Phonology*, Washington, DC. Linguistics Society of America.

Coral Hughto. 2019. *Emergent Typological Effects of Agent-based learning models in Maximum Entropy Grammar*. Ph.D. thesis, University of Massachusetts Amherst.

Gerhard Jäger. 2003. Learning constraint subhierarchies: the bidirectional gradual learning algorithm. In Henk Zeevat and Reinhard Blutner, editors, *Optimality Theory and pragmatics*, pages 251–287. Palgrave Macmillan, Basingstoke.

Karen Jesney. 2007. The locus of variation in weighted constraint grammars. Handout for poster presented at the Workshop on Variation, Gradience and Frequency in Phonology, Stanford University.

Karen Jesney and Anne-Michelle Tessier. 2011. Biases in harmonic grammar: The road to restrictive learning. *Natural Language & Linguistic Theory*, 29.

Simon Kirby. 2017. Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin and Review*, 24:118–137.

Simon Kirby and James Huford. 2002. The emergence of linguistic structure: An overview of the iterated learning model. In A Cangelosi and D. Parisi, editors, *Simulating the Evolution of Language*, chapter 6, pages 121–148. Springer Verlag, London.

Giorgio Magri. 2018. Implicational universals in stochastic constraint-based phonology implicational universals in stochastic constraint-based phonology implicational universals in stochastic constraint-based phonology. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Charlie O'Hara. 2021. *Soft Biases in Phonology: Learnability meets grammar*. Ph.D. thesis, University of Southern California.

Joe Pater and Elliott Moreton. 2012. Structurally biased phonology: complexity in language learning and typology. *The EFL Journal*, 3(2):1–44.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Jason Riggle and Colin Wilson. 2005. Local optionality. In *Proceeding of the North Eastern Linguistics Society*, volume 35.

F Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408.

Juliet Stanton. 2016. Learnability shapes typology: the case of the midpoint pathology. *Language*, 92(4):753–791.

Robert Staubs. 2014. *Computational modeling of learning biases in stress typology*. Ph.D. thesis, University of Massachusetts Amherst, Amherst.

Bruce Tesar and Paul Smolensky. 2000. *Learnability in Optimality Theory*. MIT Press.