Learning Prevents MaxEnt from Giving Probability to Harmonically Bounded Candidates

Overview. One of the major differences between MaxEnt Harmonic Grammar (Goldwater & Johnson, 2003) and Noisy Harmonic Grammar (Boersma & 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 & 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 & Moreton, 2012; Hughto, 2018; O'Hara, 2017). 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.

The Problem. Anttila & Magri (2018) show that MaxEnt overpredicts Noisy HG. Specifically, given a specific set of constraints, there are 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 greater or equal to the probability of some other mapping (/x'/-[y']), but in MaxEnt the former mapping can be less probable. A simple concrete example emerges in syllable structure—with a classic set of syllable structure constraints, it is found in noisy HG that onsetful syllables map faithfully (/CV/-[CV]) at least as often as onsetless syllables do (/V/-[V]). However, in MaxEnt it is possible for the onsetless faithful mapping to receive more probability than the onsetful mapping, see the tableaux (1).

(1) *Universal-Subverting Pattern in MaxEnt*

/CV/	NoCoda	Onset	Max	Dep		
Weights	$w = 5$	$w = 2$	$w = .01$	$w = 5$	Harm	Prob
a. CV					0	0.88
b. V		-1	-1		-2.01	0.12
c. CVC	-1			-1	-10	∼0
d. VC	-1	-1	-1	-1	-12.01	∼0

/V/	NoCoda	Onset	Max	Dep		
Weights	$w = 5$	$w = 2$	$w = .01$	$w = 5$	Harm	Prob
a. CV				-1	-5	0.05
b. V		-1			-2	0.95
c. CVC	-1			-2	-10.02	∼0
d. VC	-1	-1		-1	-7.02	0.01

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 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 & Wilson (2005); Hayes (2017).

Learnability. To evaluate the learnability of different classes of grammars, I make us of agent-based generational learning simulations (Kirby & Huford, 2002; Kirby, 2017). These simulations make use of a series of learning agents using the Perceptron learning algorithm (Rosenblatt, 1958; Jäger, 2003; Boersma & Pater, 2016); each initialized following conventional assumptions in the phonological learning literature (i.e. markedness constraints weighted high faithfulness low (Gnanadesikan, 2004; Tesar & Smolensky, 2000; Jesney & 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.

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

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 1). 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. Table 2 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. (The fractions within the diagonal cells shows both how many runs maintained the same grammar in the numerator, and the number that maintained the same type in the denominator.) 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 1: Number of iterations needed to learn each pattern.

Grammar Type	Iterations Needed
Categorical	2000
Respecting	2200
Subverting	5000

Table 2: Resulting patterns after 15 generations.

	Pattern after 15 gens		
Initial Pattern	Categorical	Respecting	Subverting
Categorical	$\frac{20}{20}$	0	0
Respecting	4	$\frac{8}{16}$	0
Subverting	13	7	$\frac{0}{0}$

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. This means that innovating a grammar that gives probability to harmonically bounded candidates is unlikely, but once harmonically bounded candidates start becoming less popular across generations, they are lost rather quickly. Thus, as long as markedness constraints start substantially far from zero, grammars with classically harmonically bounded candidates will very rarely receive substantial probability in the languages of the world. Rather than harmonically bounded candidates being a weakness of MaxEnt, giving them very small amounts of probability may be a strength: harmonically bounded candidates emerge in speech errors, and can perform better in gradient well-formedness tasks than some nonbounded candidates, (Hayes, 2017).