Stochastic Harmonic Grammars as Random Utility Models

Edward Flemming (flemming@mit.edu)
Massachusetts Institute of Technology

Introduction

- Stochastic grammars assign probabilities to outputs, making it possible to analyze variation and gradient acceptability in phonology.
- 'Maximum Entropy' Grammar is a form of stochastic grammar that is widely used in phonology (Goldwater & Johnson 2003, Hayes & Wilson 2008).
  - It builds on Harmonic Grammar (Legendre et al 2006) rather than classical Optimality Theory
- But MaxEnt grammar is not the only proposal for 'stochasticizing' Harmonic Grammar – an alternative is Noisy Harmonic Grammar (Boersma & Pater 2016)
- Identify a uniform framework for comparing and analyzing Stochastic Harmonic Grammars: Random Utility Models.
  - Use it to draw out similarities and differences between MaxEnt and NHG.

Stochastic Harmonic Grammars

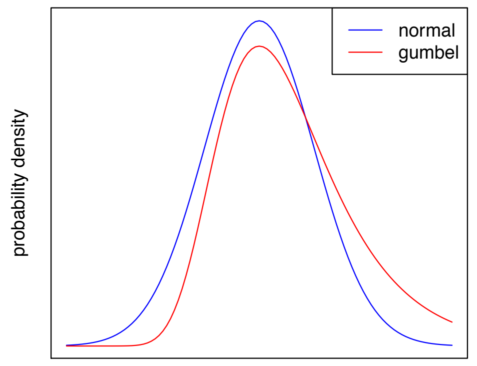
weights:	15 $+n_1$	8 $+n_2$	8 $+n_3$		NHG		MaxEnt	
/input/	C_1	C_2	C_3	h_i	ε_i	P_i	ε_i	P_i
a	-1			-15	n_1	0.6	ε_1	0.58
b		-2		-16	$2n_2$	0.26	ε_2	0.21
c		-1	-1	-16	n_2+n_3	0.14	ε_3	0.21

- Noisy Harmonic Grammar (NHG): Random noise, n_k, is added to constraint weights at each evaluation (Boersma & Pater 2016).
  - n_k are independent and normally distributed, with mean 0 and variance $\sigma^2 = 1$.
- MaxEnt: Probability of a candidate depends on its harmony, h_i: $P_i = e^{h_i} / \sum_j e^{h_j}$
  - where j ranges over the set of candidates (e.g. Hayes & Wilson 2008).
  - e.g. $P_a = \frac{e^{-15}}{e^{-15} + e^{-16} + e^{-16}} = 0.58$

NHG and MaxEnt as Random Utility Models

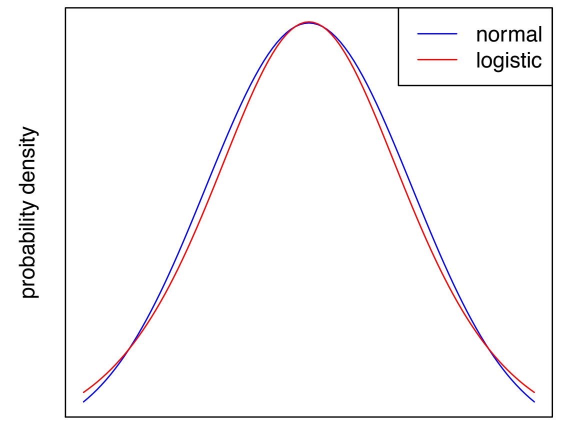
- Although NHG and MaxEnt are superficially very different they can both be formulated as NHGs where the harmony of candidate i is $h_i + \varepsilon_i$, where ε_i is a random variable ('noise').
  - HG is made stochastic by adding noise to harmony
  - Referred to in economics as a Random Utility Model (e.g. Train 2009).

- In NHG noise is added to the constraint weights, but the resulting harmony expression can be separated into $h_i + \varepsilon_i$
  - ε_i is the sum of noise components associated with the constraint violations of candidate i.
- It has been proven that the MaxEnt (multinomial logit) model follows from a RUM where the noise components, ε_i, are:
  - independent
  - all drawn from the same Gumbel (a.k.a Extreme Value Type I) distribution (e.g. Train 2009:75f.)



Shape of the noise distribution

- Noise terms (ε_i) follow a Gumbel distribution in MaxEnt and a normal distribution in NHG (normal + normal ⇒ normal)
- This is not an important difference because independent Gumbel ε_i's are essentially a tractable approximation to independent normal ε_i's.
- The probability of a candidate having the highest harmony depends on the difference in harmony between it and competing candidates.
  - ➢ The distribution of a difference in harmony depends on the distribution of differences between noise terms ε_i-ε_j
  - ➢ $(h_i+\varepsilon_i)-(h_j+\varepsilon_j) = (h_i-h_j)+(\varepsilon_i-\varepsilon_j)$
- Distributions of differences between random variables:
  - ➢ Gumbel - Gumbel ⇒ logistic
  - ➢ Normal - normal ⇒ normal
- The logistic distribution is very similar to the normal distribution



- So MaxEnt is difficult to distinguish from a variant of NHG in which ε_i are independent and normal (cf. Hayes 2017, Train 2009:35).
  - ➢ **The simplest form of NHG**
- However the Gumbel formulation (MaxEnt) has the practical advantage of a simple closed-form solution for candidate probabilities.

Independence of ε_i's and harmonic bounding

- Important differences:
  - In MaxEnt, ε_i's are independent and all drawn from the same distribution.
  - In NHG, ε_i's are not independent – candidates that violate constraint k share a noise component n_k – and are drawn from distributions with different variances.
    - ➢ Variance of ε_i is $\sigma^2 \sum_{k=1}^{N} c_{ik}^2$
- If noise terms ε_i are independent, as in MaxEnt, then all candidates receive non-zero probability, including harmonically bounded candidates (cf. Hayes 2017).
  - a harmonically bounded candidate cannot win under any fixed weighting of the constraints.
- In NHG all violations of a given constraint are perturbed by the same noise so shared violations cancel out precisely,
  - ➢ So a harmonically bounded candidate always has lower harmony than the candidate that bounds it (as long as noise is not permitted to make constraint weights negative (Jesney 2007))

weights:	10 $+n_1$	1 $+n_2$		NHG		MaxEnt
/input/	C_1	C_2	h_i	ε_i	P_i	P_i
a	-1		-10	n_1	1	0.73
b	-1	-1	-10-1	n_1+n_2	0	0.27

- Should harmonically bounded candidates be assigned $P = 0$?
  - Assigning probability to bounded candidates is central to the MaxEnt analysis of local optionality (Hayes 2017), and to 'markedness only' analyses of gradient phonotactics (Hayes & Wilson 2008).
  - The NHG mechanism for assigning zero probability to harmonically bounded candidates has additional effects – 'partial harmonic bounding'.

MaxEnt is simple

- In MaxEnt, probability is directly related to harmony: P_i is proportional to e^{h_i}
- The relative probabilities of two candidates are independent of the rest of the candidate set because ε_i's are independent

$$\frac{P_i}{P_j} = \frac{e^{h_i}/\sum_k e^{h_k}}{e^{h_j}/\sum_k e^{h_k}} = \frac{e^{h_i}}{e^{h_j}} = e^{h_i - h_j}$$

- ➢ In choice models this property is referred to as 'Independence from Irrelevant Alternatives' (Train 2009:45ff.)

NHG is complicated

- In NHG, ε_i's are not independent so the relationship between harmony and probability is complex, and the relative probability of pairs of candidates can depend on other candidates.
- The same difference in harmony translates into different relative probabilities, depending on how many constraint violations are shared – 'partial harmonic bounding'

weight:	$30+n_1$	$28+n_2$	$5+n_3$	$5+n_4$		NHG	
	C1	C2	C3	C4	h	ε_i	P
cand1		-1	-1		-33	n_2+n_3	0.92
cand2	-1		-1		-35	n_1+n_3	0.08

$\varepsilon_1 - \varepsilon_2 = n_2 - n_1$
var = $2\sigma^2 = 2$

weight:	$30+n_1$	$28+n_2$	$5+n_3$	$5+n_4$		NHG	
	C1	C2	C3	C4	h	ε_i	P
cand1		-1	-1		-33	n_2+n_3	0.84
cand2	-1			-1	-35	n_1+n_4	0.16

$\varepsilon_1 - \varepsilon_2 = n_2+n_3-n_1-n_4$
var = $4\sigma^2 = 4$

- Candidates with the same harmony can have different probabilities:

	weights:	$30+n_1$	$30+n_2$	$5+n_3$	$5+n_4$			NHG
	/spa/	C//V	DEPV	CONTIG	*CC	h	ε	P
a.	spa	-1			-1	-35	n_1+n_4	0.375
b.	əspa		-1		-1	-35	n_2+n_4	0.25
c.	səpa		-1	-1		-35	n_2+n_3	0.375
d.		-1		-1		-35	n_1+n_3	

  - ➢ (b) has lower probability because it shares violations with (a) and (c).
  - ➢ n_1 only affects the probability of (a), while n_3 affects the probabilities of (b) and (c) equally.
- The relative probabilities of a pair of candidates depends on all other candidates that violate the same constraints (more probable candidates have stronger effects)
  - If only (b) and (c) are considered, they are assigned equal probabilities.
  - If a candidate with violation profile (d) were possible, then all four candidates would be equally probable.

Conclusions

- In spite of superficial differences, MaxEnt is actually a variety of NHG.
  - Essentially the simplest form of NHG
- In the absence of empirical evidence in favor of NHG, MaxEnt is to be preferred for its simplicity and tractability.

Boersma, P. & Pater, J. (2016) Convergence properties of a gradual learning algorithm for Harmonic Grammar. McCarthy & Pater, eds., *Harmonic Grammar and Harmonic Serialism*, Equinox.
Goldwater, S. & Johnson, M. (2003) Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*.
Hayes, B. (2017). Varieties of Noisy Harmonic Grammar. Proceedings of AMP 2016.
Hayes, B. & Wilson, C. (2008) A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.
Jesney, K. (2007) The locus of variation in weighted constraint grammars. Poster presented at the Workshop on Variation, Gradience and Frequency in Phonology, Stanford University.
Train, K. (2009) *Discrete Choice Methods with Simulation, 2nd Edition*. CUP.