Bruce Hayes
October 21, 2016

Department of Linguistics
UCLA

Varieties of Noisy Harmonic Grammar[1]

Handout for talk at the Annual Meeting in Phonology (AMP)
University of Southern California, Los Angeles

BACKGROUND

1. Stochastic constraint-based grammar frameworks in modern linguistics

- These generally originate with Optimality Theory (Prince and Smolensky 1993) and represent an effort to **stochasticize** it:
    - Generate not a single winner but rather a **probability distribution** over GEN.
- The earliest frameworks of this sort were Partial Ordering OT (Anttila 1997) and Stochastic OT (Boersma 1997), but now there are a fair number of other approaches (below).
- The pursuit of such frameworks is motivated by several considerations:

2. Gradience in language is pervasive (Bod et al. 2003)

- Free variation outputs
- Gradient well-formedness intuitions
- Quantitative phonological patterns in the lexicon — often captured quantitatively by language learners and experimentally detectible ("Law of Frequency Matching"; Hayes/Zuraw/Siptár/Londe 2009, with lit. review)

3. Aside: is belief in non-gradience the consequence of traditional methodologies?

- Often, fieldworker and theorist never meet.
- The latter takes the former's best-estimate conjecture as the analytical target.
- Different results are obtained if the theorist is able to work with corpora (text and dictionary) or do experiments.

4. Stochastic grammar: analytic research is paired with learnability research

- I.e. efforts at computational modeling of human language learning.

[1] This talk develops material from a course at UCLA I co-taught with Kie Zuraw, to whom many thanks for help on this project. Thanks also to the members of the UCLA Phonology Seminar for listening to a rehearsal version of this talk and suggesting improvements.

- Typically each stochastic framework has coupled with it a computational procedure for learning its grammars.
- Stochastic frameworks are well adapted for this: isolated exceptions and speech errors don't faze them.

SOME CONSTRAINT-BASED STOCHASTIC GRAMMAR FRAMEWORKS

5. I will focus on a subset of them here

- I focus on the frameworks that performed well in the testing carried out by Zuraw and Hayes.
    - Kie Zuraw and Bruce Hayes (in press, *Lg.*), "Intersecting constraint families: an argument for Harmonic Grammar"
- Ability to predict the result when a single candidate competition responds to two independent constraint families strongly distinguishes the theories.
- The winners are all stochastic forms of …

6. Harmonic Grammar

- Legendre et al. (1990), Legendre et al. (2006), Potts et al. (2010), Pater (2016)
- Uses the same **GEN-cum-EVAL** architecture as Optimality Theory (Prince and Smolensky 1993).
- Constraints are not ranked but have numerical **weights**.
- For each candidate's row in the tableau: multiply all violation counts by corresponding constraint weights and add up the total across constraints = **Harmony**, a kind of penalty.[2]
- Winning candidate is the least penalized one.

/Input/	CONSTRAINT1	CONSTRAINT 2
weights:	2	1
Candidate 1		**
Candidate 2	*	*

/Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
weights:	2	1	
☞ Candidate 1		** × 1 = 2	0 + 2 = **2**
Candidate 2	* × 2 = 2	* × 2 = 2	2 + 2 = **4**

7. Stochasticized versions of Harmonic grammar

- Two main approaches have been taken in "stochasticizing" this framework: NHG, and maxent.

[2] Different scholars deploy minus signs differently, so for some Harmony is a negative quantity. The nomenclature/choice of signs followed here is taken from Wilson (2006).

8. Noisy Harmonic Grammar

- Ref.: Boersma and Pater (2008/2016)
- NHG **stochasticizes** Harmonic Grammar:
  - at each "evaluation time," constraint weights are nudged upward or downward by a random amount drawn from a Gaussian distribution (bell curve) — the **noise**.[3]
  - other than this preliminary step, selection of winner works the same as in non-stochastic Harmonic Grammar

9. Classical Noisy Harmonic Grammar as a procedure, shown as a tableau

- We add in the noise factors at the very start — to the constraint weights.
- Calculating, we get noisy harmonies.

/Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
weight:	$2 + N_1$	$1 + N_2$	
☞ Candidate 1		$2 \times (1 + N_2)$	$2 \times (1 + N_2)$
Candidate 2	$1 \times (2 + N_1)$	$1 \times (1 + N_2)$	$3 + N_1 + N_2$

- By iterating the noise-addition process, output probabilities of Candidate 1 and Candidate 2 can be estimated.

10. Varieties of Noisy Harmonic Grammar

- what I just described will be "Classical NHG"
- I will explore other variants, created by altering the procedure slightly:
  - Cell-specific noise
  - Addition of noise post violation-multiplication.[4]

11. NHG with cell-specific noise (Goldrick and Daland 2009)

- We put in a **fresh noise value for every cell** (cells with no violations can safely be skipped).

/Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
weight:	2	1	
☞ Candidate 1		$2 \times (1 + N_2)$	$2 \times (1 + N_2)$
Candidate 2	$1 \times (2 + N_1)$	$1 \times (1 + N_3)$	$4 + N_1 + N_3$

- The difference is one of *scope*: at what level of calculation is the noise added in?
- We can think of (11) as displaying **cell-granularity** and (9) **constraint-granularity**.

[3] This distribution is given an arbitrary standard deviation, which is the same for all constraints.
[4] There are others; see Boersma and Pater (2016) for the option of letting weights go negative.

12. A completely different (non-NHG) approach to stochasticizing Harmonic Grammar: *maxent grammars*

- Sources: Goldwater and Johnson (2003), Wilson (2006), Jaeger (2007), Hayes and Wilson (2008)
- Again, carries forward the GEN-cum-EVAL architecture of OT.
- Again, carries forward Harmony, as calculated above, as the key basis of the theory.
- Intellectual ancestry: 19th century physics; 1980's cognitive science (Smolensky 1986)
- A simple procedure converts Harmony to output probabilities:

eHarmony:[5]	$e^{-\text{Harmony}(x)}$	(e to the minus Harmony of candidate x)
Z:	$\Sigma_j\, eHarmony(j)$	(sum eHarmony over all candidates)
probability:	$\frac{\text{eHarmony}(x)}{Z}$	(candidate x's share of Z)

- Intuitively:
  - Larger weights will have a greater role in lowering probability of violators.
  - Multiple violations will have more effect in lowering probability of violating candidate.

13. Where we are going

- We now have three stochastic theories in hand (2 NHGs and 1 maxent); more to come.
- Goal: locate cases in which these variant theories **behave differently**, in the hopes of connecting with empirical data that would bear on these differences.

HARMONIC BOUNDING

14. A very simple case of harmonic bounding

/Input/	CONSTRAINT1	CONSTRAINT 2
☞ Candidate 1		*
Candidate 2	*	*

- Candidate 2 has a proper superset of the violations of Candidate 1; i.e. Candidate 1 harmonically bounds Candidate 2.

15. Harmonic bounding in Classical OT

- Taught to beginners everywhere: harmonically bounded candidates never win under any ranking.

[5] This clear and useful term from Colin Wilson is, alas, also a joke (eHarmony is a dating site on the web).

16. Harmonic bounding in Classical Noisy Harmonic Grammar

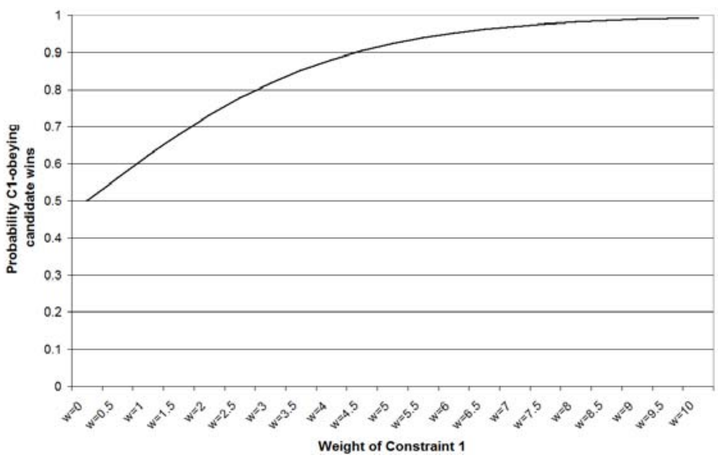
- Likewise, harmonically bounded candidates never win (frequency derived = zero).
- Why? In, e.g., (14) Candidate 2 must have a higher Harmony penalty than Candidate1 *on every evaluation time*, and *under any weighting*, since the only cases where violations don't match are to the disadvantage of Candidate 2.
  - Exception: if weight of Constraint 1 is zero you get a 50/50 tie.

17. Maxent

- Harmonically bounded candidates *can* win in maxent (Jesney 2007)
- Maxent imposes a **stochastic version of Harmonic Bounding**:
  - "A harmonically bounded candidate can never receive a higher probability than the candidate that bounds it."

18. Relating constraint weights to probability of the harmonic-bounder candidate in maxent

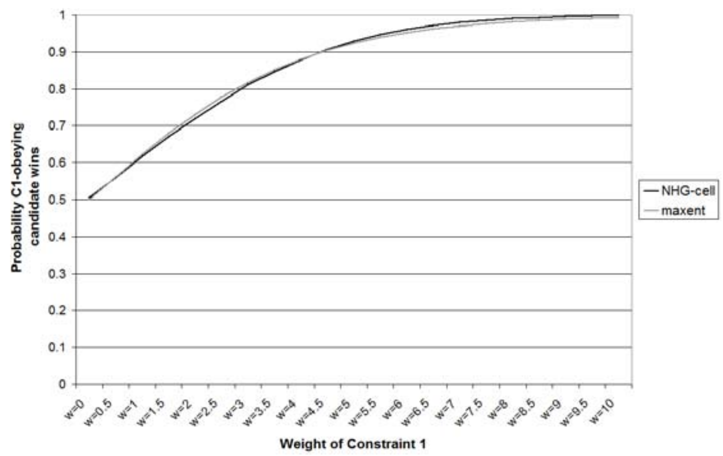
- Easy to work out the math on a spreadsheet.
- For (14), weight of Constraint 2 does not matter at all — it cancels out.
- If the weight of Constraint1 is zero, then candidates are 50/50.
- As the weight of Constraint1 approaches infinity, probability of Candidate1 **asymptotes at 1**.



19. Cell-granularity NHG

- Like maxent: harmonically bounded candidates are not shut out completely (Goldrick and Daland 2009)
- Because noise is computed separately for each candidate, in (14) Candidate 2 can sometimes win. These conditions must be met:
  - By chance, Candidate 1 is penalized on CONSTRAINT 2 more than Candidate 2 is, for this particular evaluation time.
  - And, the difference must exceed the weight of (randomly perturbed) CONSTRAINT 1.

- Working out the math, this too yields a rising curve asymptoting at 1.
- Maxent curve is given in gray for comparison. They are *almost* identical, when scaled appropriately.



- Again, it is not possible for the more-penalized candidate to get a higher probability than its harmonic-bounder.
- So Maxent and Cell-granularity NHG stand out as frameworks that implement the probabilistic, rather than absolute interpretation of harmonic bounding.

A SPECIAL CASE OF HARMONIC BOUNDING: LOCAL OPTIONALITY

20. Premise

- The same phonological process, guided by the same constraints, is applicable in more than one place in the input.

21. Literature

- Vaux (2003, 2008), Riggle and Wilson (2005), Kaplan (2011), Kimper (2011)
- There is a small pile of useful empirical cases to work with: French schwa deletion, Makonde vowel harmony, Bengali intonation phrasing, Pima reduplication ...
- AFAIK none of these has enough quantitative data to test the differences I will demonstrate.

22. I'll use a schematic example: intervocalic voicing of /p/ in long strings

/apapapapa/	IDENT(voice)	*VpV
[apapapapa]		****
[apapapa**b**a] et al.	*	***
[apapa**b**a**b**a] et al.	**	**
[apa**b**a**b**a**b**a] et al.	***	*
[a**b**a**b**a**b**a**b**a]	****	

- We see **double pyramids** of asterisks, since reducing Markedness violates Faithfulness.
- "Double pyramids" are characteristic, since multiple application will generally involve tradeoff of Markedness and Faithfulness violations.

23. Behavior of classical NHG

- Derives *only the extreme candidates*.
- Why? Harmonic bounding, as previous scholars have noted.

24. Multiple-violation patterns: what of the theories with probabilistic harmonic bounding?

- Maxent and Cell-granularity NHG *can* assign probability to the medial candidates.
- But let's take this further: what sort of *probability distributions* to the theories characteristically assign?

25. An abstract example to work with

- We scale up (22) to *six* /p/'s, with underlying /apapapapapapa/.
- Assume 50% probability of intervocalic voicing in all positions, with independent outcome for each locus.[6]

[6] I am aware that a real-life counterpart of this language, Warao (Osborn 1966), is claimed to have obligatory across-the-board-or-none application. On the slimness of the Warao data see Riggle and Wilson (2005).

- Here is a tableau for simulation/study, with the 64 equiprobable candidates.

Input	*Candidate*	*Freq.*	*VpV	IDENT(voice)
/apapapapapapa/	apapapapapapa	1	6	
	apapapapapaba	1	5	1
	apapapapabapa	1	5	1
	apapapabapapa	1	5	1
	apapabapapapa	1	5	1
	apabapapapapa	1	5	1
	abapapapapapa	1	5	1
	apapapapababa	1	4	2
	apapapabapaba	1	4	2
	(48 not listed)	…	…	…
	abababababapaba	1	1	5
	ababababababapa	1	1	5
	abababababababa	1	0	6

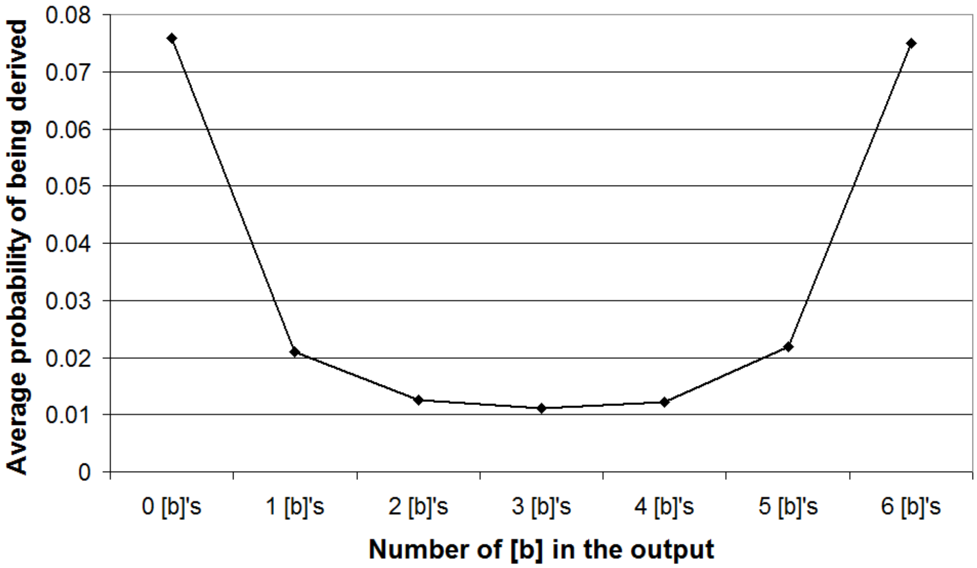
26. Maxent

- Under suitable training conditions,[7] we get the most minimal analysis conceivable.
  - Both constraints are weighted **zero**.
  - Equiprobability is predicted — a perfect fit.
- Indeed, this is why the theory is called maximum entropy!
  - = Minimal commitments in the absence of any sort of pattern in the training data.

27. Cell-granular NHG

- I have implemented Cell-granular NHG on a private copy of my "OTSoft" software,[8] with the Boersma-Pater learning algorithm.
- Curiously, probability is allocated to all candidates, but with *priority to the extremes*, and lowest probabilities to the midpoints.
  - The candidates are sorted here by number of /p/ → [b] changes; i.e. in order of descending Faithfulness.

[7] Specifically, the general penalty for constraint weights (Gaussian prior) often employed for maxent.
[8] www.linguistics.ucla.edu/people/hayes/otsoft/



28. Naming this pattern

- **Upsilonism**: the property of learning a U-shaped probability distribution when trained on flat data.
  - ➢ This is a property of cell-granular NHG (and no other theory I know of)

29. An intuitive explanation of upsilonism in cell-granular NHG

- Think of just the candidate [apapapapapapa].
- Suppose that the roll of the noise-dice is lucky for this candidate: the constraint that penalizes it, *VpV, happens to get a low weight.
- [apapapapapa] benefits *sixfold* from the lowness of this weight.
- [abababababa] likewise benefits sixfold when IDENT(voice) gets a low weight for this cell.
- "Medial" candidates benefit less (fivefold, fourfold, etc.) from their lucky moments — so they have trouble standing out; the more so the closer to the center they lie.

30. Flipping the terms of the deal: maxent deals with U-shaped training data

- Experiment: I trained a maxent grammar
  - ➢ same constraints, candidates, violations
  - ➢ frequencies were the very frequencies output by the upsilonic cell-granular NHG
- Result: flatline! — equal frequencies assigned across the board
- We might call this deviation-from-training-data **rectilinearism**.

31. A pattern that might someday lead to empirical testability

- Multiple-locus phonology
- Opposed constraints of Markedness and Faithfulness

- Training data is *overall* symmetrical between faithful and altered outputs.
- Maxent predicts rectilinearism; cell-granular NHG predicts upsilonism.

STILL MORE VARIETIES OF NHG:
POST-MULTIPLICATIVE VS. PRE-MULTIPLICATIVE NOISE ADDITION

32. The system of noise-addition for Classical NHG (repeated)

/Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
weight:	$3 + N_1$	$1 + N_2$	
☞ Candidate 1		$2 \times (1 + N_2)$	$2 + (2 \times N_2)$
Candidate 2	$1 \times (3 + N_1)$	$1 \times (1 + N_2)$	$4 + N_1 + N_2$

- Notice that N_2 got doubled in the final Harmony computation for Candidate1.

33. Changing scope again: noise addition follows violation-multiplication

/Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
basic constraint weight:	2	1	
select and remember a noise value:	N_1	N_2	
☞ Candidate 1		$\mathbf{(2 \times 1) + N_2}$	$2 + N_2$
Candidate 2	$(\mathit{1} \times 2) + N_1$	$(\mathit{1} \times 1) + N_2$	$4 + N_1 + N_2$

- We pick a noise value for each constraint.
- But add it in only after the violations have been multiplied by weights.
- The harmonies emerge as different: **unmultiplied** weights
  - ➢ Difference here: Harmony(Candidate1) is $2 + N_2$ instead of $2 + 2N_2$

34. Terminology

- Classical NHG has **premultiplicative noise**.
- (33) is **postmultiplicative noise**.

35. A scenario in which a difference can be observed

- Assume binary competitions in which two viable candidates, 1 and 2, compete for each input.
- Assume one constraint that always incurs one violation and penalizes Candidate 1.
- Assume an opposed **scalar** constraint, with a range of integer values, penalizing Candidate 2 depending on "how many" of something it has.

36. McPherson and Hayes's (2016) example: Tommo So vowel harmony

- The single-violation constraint is IDENT(feature) (Faithfulness)

- The scalar constraint is AGREE(feature) — for seven degrees of "morphological closeness"
  - Formalizing the "levels" of classical Lexical Phonology (Kiparsky 1982)
  - Inner levels give more violations of AGREE.

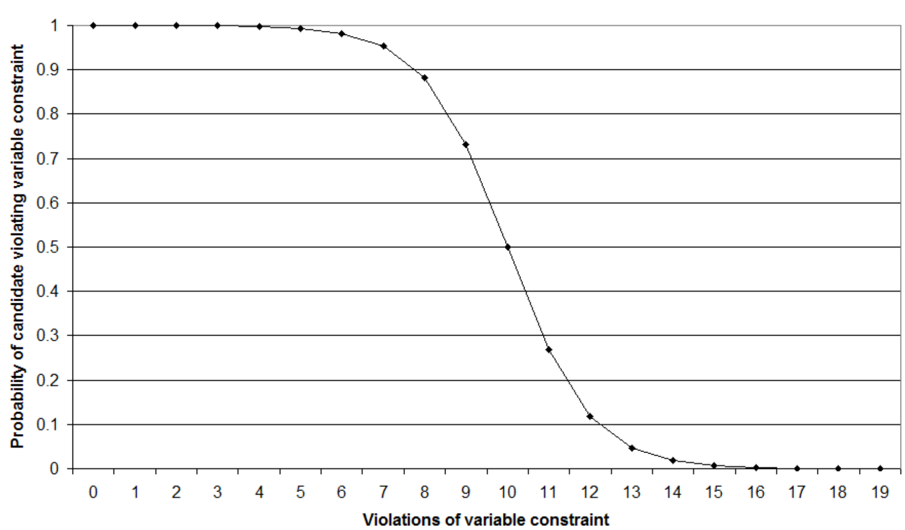
37. Toy example for the comparisons to follow

- We'll do a highly schematic tableau; real-life cases tend to be cut off owing to lack of the full range of input.

		VARIABLE CONSTRAINT	CONSTANT CONSTRAINT
Input 1	Cand1		1
	Cand2	0	
Input 2	Cand1		1
	Cand2	1	
Input 3	Cand1		1
	Cand2	2	
Input 4	Cand1		1
	Cand2	3	
…	…	…	…
Input 20	Cand1		1
	Cand2	19	

38. The essential plot — maxent version

- How does the probability of Candidate 1 go down as we increase the number of violations of VARIABLE CONSTRAINT?
- The following chart is for $\text{weight}_{variable} = 1$, $\text{weight}_{constant} = 10$

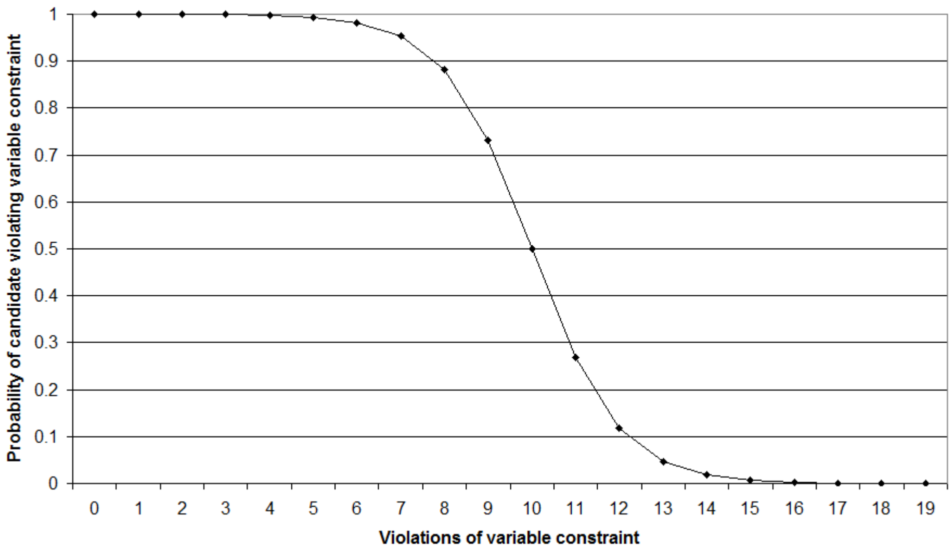


- This is a **sigmoid** curve.
- Mathematically it is a version of the **logistic function**.

- It is **symmetrical** about the point of 50% probability.
- It asymptotes at **one** and **zero**.
- It fits the Tommo So data pretty well; see McPherson/Hayes.

39. Same example, NHG with post-multiplicative noise

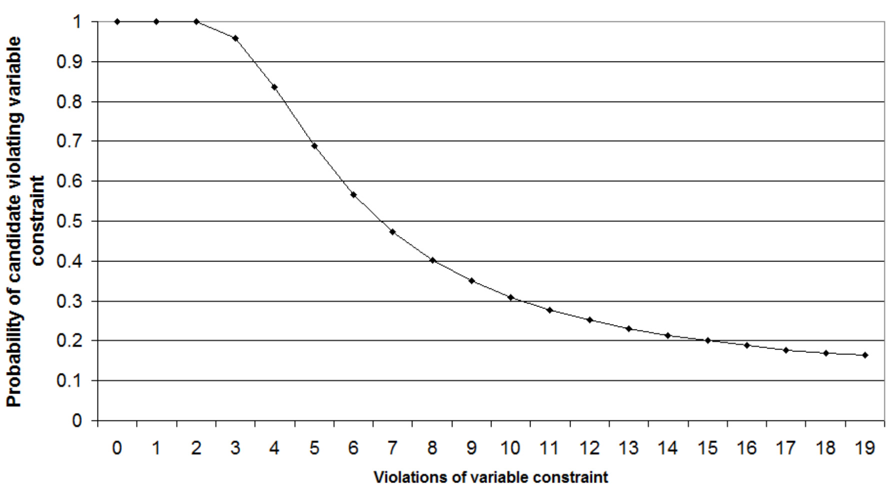
- The following chart is for $\text{weight}_{\text{variable}} = 2$, $\text{weight}_{\text{constant}} = 20$ (twice the maxent weights).



- This, too, is a sigmoid curve.
- Mathematically it is a version of the **cumulative normal distribution** — amazingly similar to the logistic function but of completely different mathematical origin.
- It too is symmetrical about the 50% point and asymptotes at one and zero.
- It also fits the Tommo So data pretty well.

40. Classical NHG (pre-multiplicative noise)

- Weights again 2 for constant constraint and 20 for variable constraint.



- This one is different:
  - Not symmetrical.
  - Asymptotes at one on the left, but **not at zero** on the right.
  - The asymptote at the right turns out to be about 0.07.

41. Why these differences? I: maxent

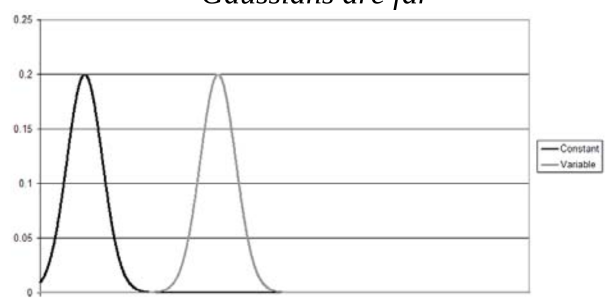
- Maxent: symmetricalness about 50% point and zero/one asymptotes are direct consequencies of the math.
  - Supplemental Materials for McPherson and Hayes (2016) gives all the details.[9]

42. Why these differences? II: post-multiplicative NHG

- Imagine two Gaussians;
  - the one for the variable constraint "slides rightward" as the number of violations of the variable constraint goes up.
  - They keep the same standard deviation, because noise is not multiplied.
  - Probability of a less-likely outcome ("reversed" harmony) will asymptote to zero, as the two Gaussians separate:

Gaussians are close

Gaussians are far



43. Why these differences? III: Classical (pre-multiplicative) NHG

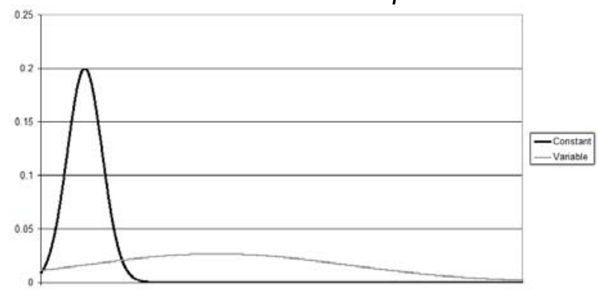
- Again, sliding Gaussians for probability distributions of weights.
- But the rightward-sliding Gaussian also *expands its standard deviation* as it slides.
- So some "left tail" that has some reasonable chance of overlapping with the constant-constraint Gaussian remains in place, so probability eventually settles on an above-zero minimum.[10]

---

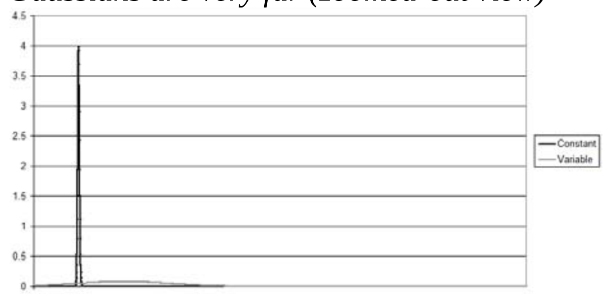
[9] www.linguistics.ucla.edu/people/hayes/papers/SupplementalFilesMcPhersonAndHayes2016.zip
[10] For the math involved here consult http://mathworld.wolfram.com/NormalDifferenceDistribution.html.

Gaussians are close

Gaussians are far



Gaussians are very far (zoomed-out view)



REVISITING UPSILONISM/RECTILINEARISM

44. Our tweaks are freely combinable

- It's perfectly feasible to set the noise afresh in every tableau cell, *and* add it in after weights are multiplied by violations, thus:

/Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
weight:	3	1	
☞ Candidate 1		$(2 \times 1) + N_2$	$2 + N_2$
Candidate 2	$(1 \times 3) + N_1$	$(1 \times 1) + N_3$	$4 + N_1 + N_3$

- This creates yet another variety of Noisy Harmonic Grammar — post-multiplicative-noise, cell-granularity

45. The /apapapapapa/ language in post-multiplicative-noise, cell-granularity NHG

- Result:
  - Moving to post-multiplicative noise *cancels* upsilonism: output distribution is **flat**, just like in maxent.
  - This makes sense: the intuitive explanation for upsilonism in (29) depended on the multiplication of the noise.
- Rectilinearism:
  - Like maxent, post-multiplicative-noise, cell-granularity NHG has a rectilinearity bias: it outputs a flat distribution when trained on an upsilonic one.

46. Accidental twins?

- I have yet to find any qualitative differences between maxent and post-multiplicative noise, cell-granularity NHG; though their math is entirely different.[11]
- As noted, there *are* differences but they rest on the subtle difference between logistic vs. normal-distribution sigmoids — far beyond the resolution of any sort of empirical work being done today.

47. Summary: classifying the differences

Framework	*Effects of harmonic bounding*	*Upsilonism/ rectilinearism/ extremes only*	*Type of sigmoids derived*
Classical NHG	absolute	extremes only	asymmetrical with non-zero asymptote
NHG with cell-specific noise	probabilistic	upsilonism	asymmetrical with non-zero asymptote
NHG with post-multiplicative noise addition	absolute	extremes only	symmetrical with 1/0 asymptotes
NHG with cell-specific noise and post-multiplicative weight addition	probabilistic	rectilinearism	symmetrical with 1/0 asymptotes
Maxent	probabilistic	rectilinearism	symmetrical with 1/0 asymptotes

THE EMPIRICAL SIDE

48. What I have something to say about ..

- I have no data to bear on upsilonism/rectilinearism.
- I will address only the issues of
  - statistical harmonic bounding
  - sigmoid shape

49. Inventory theory

- Linguistics is filled with cases where derivations are of dubious value.
- I.e. we simply want to indicate the legal members of an inventory (possible words, possible sentences, possible metrical lines of verse)
  - Cf. phonology, where the task of unifying the members of a paradigm give us reason to derive surface forms from underlying forms.

[11] I just had time to check Jesney's (2007) case (CCVC can lose its coda or simplify its onset independently). Again they behave alike.

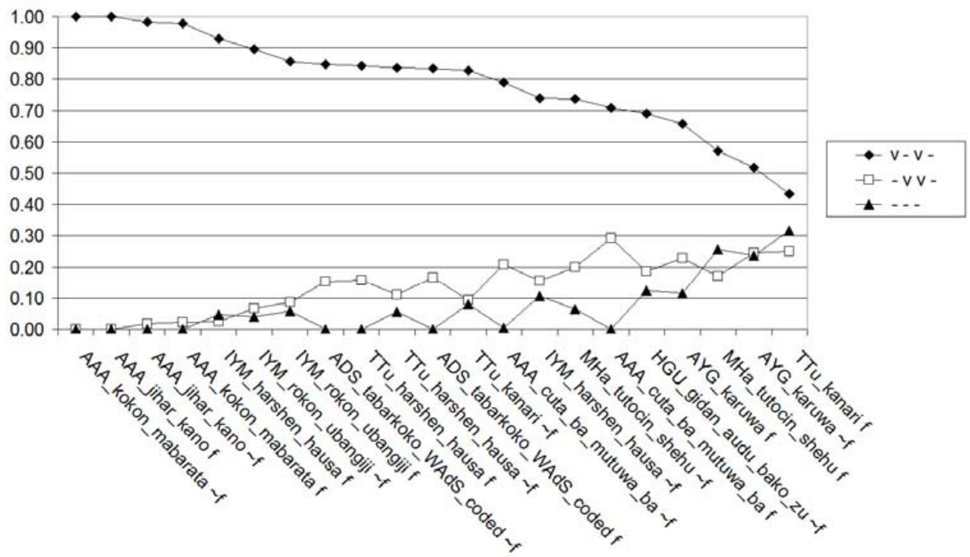
- The Rich Base concept does permit “pseudo-derivations” to underlie inventories, with a literature addressing this (Keer and Baković 1997, Hayes 2004, Prince and Tesar 2014, etc.)
- An alternative is simply to develop a Faithfulness-free stochastic grammar that assigns a probability to every member of GEN.
- Such a grammar would assign, I suspect, a lower probability to [pra] than [pa];
    - [pra] is harmonically bounded, and indeed should by any criterion receive a lower probability score than [pa].
    - Pseudo-derivational theories face the awkward fact that [pra] is more marked than [pa] *but not repaired* (other than in rare speech errors) to [pa].
- References on “evaluate-all-of-GEN” grammars: Hayes and Wilson (2008) on phonotactics; Hayes and Moore-Cantwell (2011) on metrics of Hopkins; Hayes, Wilson and Shisko (2012) on metrics of Shakespeare/Milton
- All such work must assign some probability to harmonically bounded candidates, and thus works only with some of the frameworks above.

50. An apparent case of statistical harmonic bounding

- Hayes and Schuh (in progress) evaluate three of the principal “metra” of the Hausa *rajaz* quantitative meter.

	STRONG POSITIONS MUST INITIATE HEAVY SYLLABLE	HEAVY SYLLABLE MUST INITIATE STRONG POSITION
x x x x x x x x ˘ — ˘ —		
x x x x x x x x — ˘ ˘ —	*	*
x x x x x x x x — — —	*	**

- A near-perfect pattern: no matter what the actual frequencies, the relative frequencies respect harmonic bounding relations for various poets/poems/stanza positions



- Why? because ◡ – ◡ – bounds – ◡ ◡ – and – ◡ ◡ – bounds – – –; and in the frameworks adopted here, subsets of violation implies *invariant differences of probability*; rather than outright absence of the candidate with superset violations.

51. The Goldrick/Daland theory of speech errors

- If Goldrick/Daland are right that a "misweighted" grammar occasionally outputs harmonically-bounded candidates.

52. Multiple locus cases

- Already discussed.
- To be sure, there are other remedies available. (Riggle-Wilson; Kaplan; Kimper)
- I think there is something appealing about getting multiple-locus cases — including, ideally, their fine-grained statistical detail — from the fundamental architecture of the theory (how probability distributions are computed)

CHARACTERISTIC SHAPE OF SIGMOIDS IN GRAMMAR

53. McPherson and Hayes's (2016) result

- We tried to fit the Tommo So data using classical NHG (asymmetrical sigmoids) and got a somewhat inferior fit to the data.
- Reason: our empirical sigmoids were quite symmetrical.

54. The world in general

- Zuraw and Hayes (in press) looked at a fair number of empirical sigmoids in phonology

- We repeatedly found symmetrical sigmoid curves asymptoting — if the empirically observed range permitted it — at zero and one.
- See Zuraw and Hayes (in press) for a brief survey of such sigmoids elsewhere in language.
  - speech perception
  - syntactic change (work of Kroch and colleagues)[12]
  - possibly even cognition in general

CONCLUSION

55. The research agenda

- Earlier research on stochastic constraint theories emphasized:
  - simple feasibility (ability to fit particular data sets; e.g. Anttila 1997; Boersma and Hayes 2001)
  - effectiveness of the associated learnability theory (e.g. Pater 2008)

- But it's also possible to treat such theories as **abstract characterizations of possible grammars**, which permit/favor only particular data patterns.
  - Zuraw and Hayes (in press) — only some theories capture the parallel sigmoid curves that characterize "intersecting constraint families"
  - Here: empirical picture less clear, but the theories are quite different in their predictions re. harmonic bounding, upsilonism/rectilinearism, sigmoid shape

- Empirical work can now keep its eyes open for data bearing on relatively abstract principles of grammatical organization.

[12] Indeed, I suspect that serious study of synchronic syntactic patterns from a stochastic constraint-based viewpoint would also reveal abundant sigmoidal phenomena.

References

Anttila, Arto. 1997. Deriving variation from grammar. In Frans Hinskens, Roeland van Hout & W. Leo Wetzels (eds.), *Variation, change, and phonological theory*, 35–68. Amsterdam: John Benjamins.

Bod, Rens, Jennifer Hay & Stephanie Jannedy (eds.). 2003. *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45–86.

Boersma, Paul and Joe Pater. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In John J. McCarthy and Joe Pater, eds., *Harmonic Grammar and Harmonic Serialism*. Sheffield: Equinox, pp. 389-434. [N.B.: publication date misleading; in samizdat since 2008]

Della Pietra, Stephen, Vincent J. Della Pietra & John D Lafferty. 1997. Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 19. 380–393.

Goldrick, Matt and Robert Daland (2009) Linking speech errors and phonological grammars: Insights from Harmonic Grammar networks. *Phonology* 26: 147-185.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, ed. Jennifer Spenader, Anders Eriksson, and Osten Dahl, 111–120.

Hayes, Bruce and Claire Moore-Cantwell (2011). Gerard Manley Hopkins's sprung rhythm: corpus study and stochastic grammar. *Phonology* 28:235–282

Hayes, Bruce and Russell Schuh (in progress) Metrical structure and sung rhythm of the Hausa rajaz. Ms., Department of Linguistics, UCLA.

Hayes, Bruce and Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39: 379-440.

Hayes, Bruce, Colin Wilson and Anne Shisko (2012) Maxent grammars for the metrics of Shakespeare and Milton, *Language* 88:691-731.

Jäger, Gerhard. 2007. Maximum entropy models and stochastic Optimality Theory. Architectures, rules, and preferences. Variations on themes by Joan W. Bresnan, ed. by Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Chris Manning, 467-479. Stanford: CSLI Publications.

Jesney, Karen. 2007. The locus of variation in weighted constraint grammars. *Workshop on Variation, Gradience and Frequency in Phonology*. Stanford University.

Kaplan, Aaron F. 2011. Variation through markedness suppression. *Phonology* 28(03). 331–370.

Keer, Edward and Eric Baković. 1997. Have FAITH in syntax. In Emily Curtis, James Lyle, and Gabriel Webster, eds., *Proceedings of the Sixteenth West Coast Conference on Formal Linguistics*, pp. 255-269. Stanford, CA: Center for the Study of Language and Information.

Kimper, Wendell A. 2011. Locality and globality in phonological variation. *Natural Language & Linguistic Theory* 29:423–465.

Kiparsky, Paul (1982). Lexical phonology and morphology. In I. S. Yang (ed.), *Linguistics in the Morning Calm.* Seoul: Hanshin. 3-91.

Kroch, Anthony (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change* **1**. 199–244.

Legendre, Geraldine, Yoshiro Miyata & Paul Smolensky. 1990. Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness: An Application. Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, 884–891. Mahwah, NJ: Lawrence Erlbaum Associates.

Legendre, Géraldine, Antonella Sorace & Paul Smolensky. 2006. The Optimality Theory–Harmonic Grammar Connection. In Paul Smolensky & Géraldine Legendre (eds.), *The Harmonic Mind*, 339–402. Cambridge, MA: MIT Press.

Liphola, Marcelino M. 2001. Aspects of phonology and morphology of Shimakonde. Ohio State University PhD dissertation.

McPherson, Laura and Bruce Hayes (2016) Relating application frequency to morphological structure: the case of Tommo So vowel harmony. *Phonology* 33:125-167.

Osborn, Henry A. 1966. Warao I: Phonology and Morphophonemics. *International Journal of American Linguistics* 32(2). 108–123.

Pater, Joe (2008). Gradual learning and convergence. *Linguistic Inquiry* **39.2**: 334-345.

Pater, Joe (2016). Universal grammar with weighted constraints. In John J. McCarthy and Joe Pater, eds., *Harmonic Grammar and Harmonic Serialism*. Sheffield: Equinox, pp. 1-46.

Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt & Michael Becker. 2010. Harmonic Grammar with linear programming: from linear systems to linguistic typology. *Phonology* 27: 77–117.

Prince, Alan, and Bruce Tesar. 2004. Learning phonotactic distributions. In *Fixing priorities: Constraints in phonological acquisition*, ed. by René Kager, Joe Pater, and Wim Zonneveld, 245–291. Cambridge: Cambridge University Press.

Riggle, Jason & Colin Wilson. 2005. Local optionality. In Leah Bateman & Cherlon Ussery (eds.), *NELS 35*.

Smolensky, Paul. 1986. Information processing in dynamical systems: foundations of Harmony Theory. In *Parallel distributed processing: explorations* in *the microstructure of cognition*, ed. David E. Rumelhart, James L. McClelland, and the PDP Research Group, volume 1, 194–281. Cambridge, Mass.: MIT Press/Bradford Books.

Vaux, Bert (2003) Why the phonological component must be serial and rule-based. Paper presented at the Linguistic Society of America.

Vaux, Bert (2008) Why the phonological component must be serial and rule-based. In Bert Vaux and Andrew Nevins, eds., *Rules, constraints, and phonological phenomena*. Oxford: Oxford University Press.

Wilson, Colin. 2006. Learning phonology with substantive bias: an experimental and computational investigation of velar palatalization. *Cognitive Science* 30:945–982

Zuraw, Kie, and Bruce Hayes (in press) "Intersecting constraint families: an argument for Harmonic Grammar." To appear in *Language*.