A survey of Wug-shaped curves: phonology, phonetics, syntax, semantics, language change

Bruce Hayes
UCLA

1. Notes about this talk

- I’m giving a précis of a paper that exists in full first draft (see my website).
  ➢ Intended for Annual Review of Linguistics
- To save time I will stint on detail, rigor, citations — please see the paper for a less sloppy version.
- The purpose of this talk from my point of view is to get your feedback, so please give some! :=)

2. Plan

- Present a “quantitative signature” — an arithmetic pattern — that evidently arises in data throughout linguistics.
- The pattern will be called the “wug-shaped curve”
- It is a natural consequence of the MaxEnt framework (Goldwater and Johnson 2003, Zuraw and Hayes 2017).

3. Background: what we need probabilistic frameworks for

- Alternative surface forms in free variation, with patterns and preferences
- Frequency-matching of the lexicon (since Zuraw 2000, much other work now; also syntactic matching of selection properties; Jurafsky 2003, Linzen et al. 2016)
- Generate gradient native speaker judgments (e.g. Hayes 2000)

- All three will have examples in this talk.

4. Theories

- As a start, assume Optimality Theory (OT, Prince and Smolensky 1993), with the architecture of GEN, candidates, constraints, winners.
- Here we explore evolved, probabilistic versions of OT, meant to cover (3) above.
  ➢ given an input and its set of output candidates (from GEN)
    - look at the sets of constraint violations for the candidates
    - from it calculate and assign a probability to every candidate
• all but a few candidates get essentially zero probability
  • There are many ways to do this.

5. Candidate theories
  • MaxEnt (Goldwater and Johnson 2003), Noisy Harmonic Grammar (Boersma and Pater 2010), Stochastic OT (Boersma 1998) — and some other less obvious candidates, to be mentioned below
  • We will use MaxEnt as our hallmark theory, then compare the others with it.
  • We need a somewhat more thorough review of MaxEnt than usual, since the details — the intuitive details — will matter a lot.

HOW MAXENT WORKS, WITH CONNECTIONS TO INTUITION

6. Theme
  • Think of selecting the output candidate (or candidates) as a rational decision.
  • Think of the constraint violations as evidence we might you to make this decision.
  • From this perspective: every part of the MaxEnt math can be regarded as matching our intuitions about rational, optimized decision-making.

7. The MaxEnt formula deriving probability of candidate \( x \) from its tableau

\[
Pr(x) = \frac{\exp(-\sum_i w_i f_i(x))}{Z}, \text{ where } Z = \sum_j \exp(-\sum_i w_i f_i(x_j))
\]

• “The probability of candidate \( x \) is derived from the tableau information as …”
• We will cover the formula one step at a time.

8. Weights
  • Every constraint has a nonnegative number, its weight, which tells you how strong it is.
    ➢ More specifically, how much it lowers the probability of candidates that violate it.
    ➢ In (7), this is \( w_i \) for each constraint \( i \).
    ➢ Weights are intuitive — we know that reasons differ in cogency.

9. MaxEnt, Step 1
  • For each tableau cell, multiply the number of violations by the weight of the constraint.
    ➢ In (7), this is \( w_i f_i(x) \) (\( x \) is candidate, \( f \) is number of violations)
    ➢ This is intuitive, in the sense that two violations are is plausibly “twice the evidence” of one violation.

10. MaxEnt, Step 2
  • For each candidate:
➢ For each constraint:
   ➢ Add up the result of Step 1, across constraints, to get a single value for this candidate.

• This is an aggregate penalty score for this candidate, called the **Harmony**.
• In formula (7), Harmony is represented by $\sum w_i f_i(x)$
• Harmony is *intuitive*:
   ➢ When we make rational decisions, we appropriately weigh all the evidence.
   ➢ Classical OT is bravely counterintuitive: the decision between two rival candidates is made *solely* by the highest ranked constraint that distinguishes them.
   ➢ The view taken here is: yes, brave, but empirically wrong

11. **MaxEnt, Step 3**

• Take every Harmony value and compute from it the corresponding $e$Harmony.$^1$
• In formula (7), eHarmony is: $\exp(-\sum w_i f_i(x))$.
• That is, you negate the Harmony, then take $e$ (about 2.7) to the result (graphed here).

   ➢ Graphing eHarmony against Harmony:

   ![Graph](image)

   • eHarmony performs a sort of “squishing”: If Harmony gets very big, eHarmony is already close to zero and gets only slightly smaller.
   • I claim that eHarmony is *intuitive*:
     ➢ if we are at probability .5 for choosing a candidate, we welcome evidence to help decide and are seriously influenced by it (steep region of curve)
     ➢ But for a candidate already heavily penalized (e.g. .001), even a great deal of evidence may only move us to .0005.
     ➢ Same for candidates close to one: their rivals are already penalized by a lot of Harmony and increase will only move the top candidate e.g. .999 $\rightarrow$ .9995
     ➢ The principle: *certainty is evidentially expensive*.
     ➢ This will matter below.

12. **MaxEnt, Last step**

• Sum the eHarmony for every candidate for this input, call the result Z.

---

$^1$ Term comes from Wilson (2014), who was joking (eHarmony is a dating website), but I like the mnemonic.
In (7), this is: \( \sum_j \exp(-\sum_i w_i f_i(x_j)) \).

- The probability of a candidate is its eHarmony divided by \( Z \); i.e. its share in \( Z \)
- This is also intuitive: a candidate is less likely if it has strong rivals.
- Formula (7) is now explicated in full.

13. **Summing up: the intuitivity of MaxEnt**

a. Constraints differ in their evidential force.
b. Multiple violations of the same constraint are predicted to make a candidate less probable.
c. All evidence from the constraints is duly considered in proportion to their weights; and no evidence is thrown out.
d. Evidence is scaled to make it have less effect as we approach certainty.
e. Candidates become less probable when they compete with powerful rivals.

**MAXENT AND THE SIGMOID CURVE**

14. **Next step**

- Let’s start to go (a bit) empirical, looking at sigmoids as ways of modeling real data.

15. **The analytic scenario of interest here**

- Imagine a setup with:
  - One single constraint, called ONOFF, conflicting with
  - A constraint, or set of constraints, defining a scale.
- Some scales:
  - A family of assimilation triggers of varying strength — e.g. vowels, triggering vowel harmony
  - A set of phonology-triggering affixes that vary in their propensity-to-trigger
- Imagine a theory that takes these ingredients and computes a probability for all possible outcomes along the scale.
- Simplest case first: the scale is defined by one single constraint, VARIABLE, with multiple violation levels.

16. **Concretizing a bit**

- Let VARIABLE have seven values, 1-7.
- It is opposed by ONOFF.
- As throughout this talk, each input has only two viable candidates:
  - One obeys VARIABLE, violates ONOFF
  - One obeys ONOFF, violates VARIABLE to some degree, depending
  - All other candidates violate powerful constraints that in effect rule them out.
- We plot a probability function:
  - Horizontal axis: value for VARIABLE
  - Vertical axis: probability that the candidate that obeys ONOFF wins.
• We will plot for all values, not just the integers 1-7, since the curve emerges more clearly that way.

17. Do this in MaxEnt — you will get a sigmoid

![Sigmoid Curve](image)

- The sigmoid asymptotes at its extremes to 1 and 0 — assuming that empirical cases exist covering enough of the horizontal axis.
- It is **symmetrical** about the 50% probability mark.
- The slope transitions gracefully from near-level, to steep, to near-level.

18. Relating the properties of the sigmoid to the MaxEnt constraint weights

- High weight for VARIABLE = steep maximum slope
- Weight of ONOFF: moves the curve from side to side

A SIGMOID CURVE IN PHONETICS

19. A MaxEnt grammar for perception, after Boersma (1998 et seq.)

- Inputs: physical values along one single parameter — like stop closure duration.
- Output: the probability of a particular phonemic percept, like [p] vs. [b].

20. A classical speech-perception experiment from Kluender et al. (1988)

- Vary closure duration in a synthetic intervocalic bilabial stop.
- Ask participants if they hear a [p] or a [b]
• Result is a widely-found, quasi-sigmoid pattern:

![Graph showing probability of percepts versus closure duration in milliseconds.]

21. Spelling out the MaxEnt perception grammar

• Constraints
  ➢ VARIABLE: penalize the percept of [b] to the extent that closure duration deviates from the extreme value of 20 ms.
  ➢ ONOFF: penalizes the [p] candidate

• Weights
  ➢ As throughout this talk, I computed these by fitting them to the available data.
  ➢ I used a standard method (maximum likelihood estimation)
  ➢ I used the “Solver” utility that comes with Excel.

  VARIABLE: 0.088 (units of Harmony per millisecond)
  ONOFF: 4.34

• Inputs
  ➢ Every closure duration value in the relevant range, by intervals of 10 msec.

• Candidates for each input
  ➢ The percept [b]
  ➢ The percept [p]
22. The Kluender et al. (1988) data replotted, superposed with model predictions

![Graph showing a sigmoid curve with data points]

23. A plotting gimmick I will be employing throughout

- Take any tableaux cell with $n$ violations for the [p] candidate.
- Remove them, and instead install $-n$ violations for the [b] candidate.
- As can be determined from formula (7), you get the same answer.
- Convenient, because now the [p] candidate has zero harmony, and all relevant information lies in the harmony of the [b] candidate — plotted on the x-axis above.

24. A little bit on sigmoids in phonetics

- This use of MaxEnt, under its more proper label of logistic regression, is a standard procedure in continuum experiments for speech perception (Morrison 2007).
- These experiments commonly yield sigmoid curves.

A MORE COMPLEX QUANTITATIVE SIGNATURE: THE WUG-SHAPED CURVE

25. Scenario

- Let us augment the primal case of (17); ONOFF vs. VARIABLE (where Variable is either a gradiently violable constraint, or a constraint family).
- Now, double the input set:
  - Add a new batch of inputs, identical to the first, except that one of their two candidates violates the PERTURBER — a constraint defined on an independent dimension.
26. Effect of perturbers in Maxent

- They create a **second sigmoid**, shifted over relative to the original sigmoid by a particular amount, namely the weight of PERTURBER.
- Here is the double sigmoid resulting from the presence of a PERTURBER constraint:

![Diagram of double sigmoid](image)

27. I will refer to this configuration as a **wug-shaped curve**

![Diagram of wug-shaped curve](image)

- Thanks to Dustin Bowers for thinking of this name.

28. The wug-shaped curve and constraint weights

- A highly weighted VARIABLE or VARIABLE family will make the wug stand up straighter (relative to an observable variable like closure duration).
- A highly weighted PERTURBER makes the wug **fatter**.

29. Can there be multiple perturbers?

- Indeed, and you get multiple sigmoids, spaced as expected.
• Visually, perhaps a flock of wugs marching in parallel?²
• I will simply use the term “wug-shaped curve” for these cases as well.

THE SEARCH FOR WUG-SHAPED CURVES

30. Premises

• Suppose that the right framework for linguistics in general is something like OT, with GEN, EVAL, constraints …
• Suppose that the right way to engage OT with probability theory is MaxEnt.
• Suppose that constraints frequently come in distinct families with distinct teleologies (hence we can choose VARIABLES and PERTURBERS, compute, plot …)
• Then we should find wug-shaped curves in all fields of linguistics; it should be the default pattern observed VARIABLE-PERTURBER constraint interaction.

31. What got me into searching for wug-shaped curves

• This started when I was second author of Zuraw and Hayes (2017), “Intersecting constraint families: an argument for Harmonic Grammar.”
• Our “intersecting constraint families” were VARIABLE and PERTURBER families from the phonologies of Tagalog, French, and Hungarian.
• In all three cases, we found wug-shaped curves.
• We showed that such curves are the natural consequence of MaxEnt.

32. One example from Zuraw and Hayes (2017): Hungarian Vowel Harmony

• Most suffixes have two variants, one with a phonetically back vowel and the other with a front vowel (e.g. [-nɔk]/[-nɛk] for the Dative).
• The most important factor in determining the choice of suffix variant is the character of the last vowel, or last few vowels, of the stem.

² Thanks to an anonymous talk audience member for this appealing image.
33. The **VARIABLE** constraint family of Hungarian

- Standard taxonomy of Hungarian vowels is Back (B), Front Rounded (F), and Neutral (N).
- Different vowels, at different distances, help determine the choice of back or front suffix.
- On the chart below you will see formulae denoting various relevant arrangements of stem vowels.
- It’s complicated, and no real need to go into the details (Hayes and Londe 2006; Hayes, Zuraw, Siptár, and Londe 2009; Zuraw and Hayes 2017)

34. The **PERTURBER** constraint family of Hungarian

- As discovered by Hayes, Zuraw, Siptár, and Londe (2009), Hungarian vowel harmony is probabilistically influenced by consonants.
- If a stem ends in
  - two consonants
  - a bilabial stop
  - a sibilant, or
  - a coronal sonorant
  it is more likely to take front suffixes than otherwise.
- This is backed by both lexicon-searching and by our online wug-test (demonstrating speakers internalize the pattern).

35. Plotting this all as a wug-shaped curve

- On a spreadsheet I fitted a MaxEnt grammar to the Hungarian data, using the Zuraw/Hayes constraints.
- On my spreadsheet I *segregated* the doses of Harmony coming from **VARIABLE** (vowel-based constraints) and **PERTURBER** (consonant-based constraints)
- Thus, I could plot separate sigmoids for each level of the **PERTURBER** family, as in (26) above.
- Following Zuraw and Hayes (2017), I aggregated data, to get enough points to trust:
  - forms with *no* consonant environment present
  - forms with *1* consonant environment present
  - forms with *2* consonant environments present
36. Some key points on this and other graphs

- Sigmoid curves are roughly parallel — e.g., same maximum slope.
- Sigmoid curves asymptote to zero on one side, one on the other (assuming enough data are present; not always so).
- The vertical “columns” of points are spread in the middle, bunched (or even superposed) as the periphery.
- There are no instances of convexity (up-then-down) or concavity (down-then-up).
- All of these are consequences of the MaxEnt formula (7).

37. My graphs are slightly different from the Zuraw/Hayes (2017) graphs

- I have horizontally spaced the data points according to the Variable harmony assigned to them in the analysis.
- This permits us to visually check for match to the MaxEnt sigmoid.

MY SUMMER TOUR OF PROBABILISTIC LINGUISTICS

38. Why I was led to follow up Zuraw and Hayes (2017) more broadly

- I volunteered my February 2020 tutorial presentation to the Berkeley Linguistic Society as a review article for the Annual Review of Linguistics.
- Editor Mark Liberman noted that the readership of the journal is broad; might I consider moving beyond phonology?

39. What I did in looking further

- Browse through classic works of probabilistic linguistics.
40. If you want to see it all

- I don’t want to be seen to be cherry-picking nice cases, so I put all of my plots on a website: the Gallery of Wug-Shaped Curves, https://linguistics.ucla.edu/people/hayes/GalleryOfWugShapedCurves/

41. Criteria for choosing cases/picking what is VARIABLE, PERTURBER

- Probability of candidates had to approach one at one end, or zero at the other, or ideally both.
- Examples abundant enough so that each data point would represent multiple observations
- The PERTURBER set (and where possible, the VARIABLE set as well) should have a unified, intuitive rationale.

42. Other cases in phonology

- I’ll skip here, referring you to the Gallery, which contains wug-shaped curves of various degree of convincingness for:
  - French Liaison (Zuraw and Hayes 2017)
  - Ernestus and Baayen (2003) on undoing Dutch Final Devoicing
  - Anttila (1997) on Finnish genitive plural allomorphs
  - Schwa-zero alternations in French (Smith and Pater 2020)
  - Stress in Hupa (Ryan 2019)

SOCIOLINGUISTICS

43. Sociolinguists as pioneers of the ideas discussed here

- Theoretical sociolinguists, working in the 1970’s:
  - First noticed sigmoids, perturbers, and wug-shaped curves
  - First employed (a form of) MaxEnt for their analyses (work of David Sankoff and colleagues)

44. Labov (1969) on copula-contraction in Black English

- This paper discovered perturbers and their tendency to have across-the-board effects as MaxEnt predicts.
- Field data was from Black English, as spoken in Harlem.
• Binary choice is whether copula is realized as a syllable (e.g. Boot is) or as just a consonant (Boot’s)
• VARIABLE constraints:
  ➢ Is copula preceded by a consonant, or by a vowel, or is it part of a lexically-listed portmanteau like it’s?
• PERTURBERS:
  ➢ Four distinct syntactic environments coming after the copula.

![Graph showing probability of uncontracted variant vs base harmony]

• Observe:
  ➢ Vertical compression of the spacing of the data columns as we move to right periphery.
  ➢ Language does not provide data points on left; hence a wug-tail, found elsewhere.

45. Bailey (1973) discovers the wug-shaped curve

• Not with this notation, nor with any math to derive it
  ➢ The math came a few years later from David Sankoff.
• Data were borrowed from a ms.\(^3\) by Gillian Sankoff: deletion of [l] in function words in Montréal French.
• VARIABLE: each function word has its own lexical propensity (Zuraw/Hayes 2017, Zymet 2018) to lose its [l] in running speech.
• PERTURBER: speaking style; where sex and class are a proxy for style.
  ➢ Careful speech, I suggest, has a higher weight for MAX(l).
  ➢ People speak more carefully to an interviewer if they are female or professional class, as opposed to male or working class.

\(^3\) Possibly published, but I’ll need to use the library to find out…
46. A remark on how MaxEnt is used in classical sociolinguistics

- Their key theoretical device is called the **variable rule**.
- It works like this:
  - Set up a rule in the style of *SPE* Chomsky and Halle 1968)
  - Set up a MaxEnt grammar with two output candidates *Apply Rule* and *Don’t Apply Rule*.
  - Its constraints are the Perturbers.
  - The MaxEnt grammar informs the rule of its *probability of application*.
- Nowadays we are more inclined to use OT in some form. So the MaxEnt grammar can do the whole job, without any rules.
- The older theory could be validated if we could show that the little MaxEnt grammars attached to each rule each need to have different constraint weights …

47. More sociolinguistics in the online Gallery

- Deletion of the copula in Black English (Labov 1969, Cedergren and Sankoff 1974)
- Omission of *que* in Québec French (Cedergren and Sankoff 1974)
- R-Spirantization in Panamanian Spanish (Cedergren and Sankoff 1974)
- R-Dropping in New York City English (Labov, via Cedergren and Sankoff 1974)
- Cluster Simplification in Detroit Black English (Wolfram 1969)

**SYNTAX**

48. A method for using MaxEnt analysis in syntax

- Find a *choice*, like English datives
  - NP NP (*Mary gave John a book*)
NP PP (Mary gave a book to John)
- Same communicative intent, two syntactic realizations.
- Form a constraint-based grammar intended predict how speakers make that choice.
- More broadly: OT syntax lets there be an underlying speaking intent; GEN and EVAL chose a realization for that intent.

49. Empirical results in probabilistic syntax

- This has been a major research activity of Joan Bresnan and her colleagues over the last 15 years. Two main results:
  - choices such as NP NP/ NP PP respond systematically to many factors
  - Thus, these choices are *semipredictable*, with surprising accuracy.
  - These choices are made on a *language-specific basis* (e.g. U.S. vs. Australian English), and thus evidence that children learn probabilistic grammars and match the frequencies of what they are hearing, just as similar research has shown in phonology.

50. Szmrecsanyi et al. (2017) on English datives

- Constraints collectively embodying VARIABLE here:
  - those which depend on verb semantics, distinguishing “transfer,” “communication,” and “abstract”
  - those which depend on properties of the *recipient* NP, such as animacy, definiteness, and pronounhood
  - a constraint based on relative *length* (in words), which prefers placing longer phrases second
- **PERTURBER** constraints:
  - These single out the three common categories of the *theme* NP (that which is given): indefinite full NP, definite full NP, and pronoun.
51. **More syntax in the online Gallery**

- Szmrecsanyi et al.’s Genitive cases (NP’s N, N of NP), plotted with two choices of PERTURBER.

**WUG SHAPED CURVES IN PHONETICS**

52. **Perturbers in classical speech perception experiments**

- Researcher seldom if ever seek a single sigmoid curve, as in (22) above.
- Rather, they seek to demonstrate that some particular PERTURBER matters, and to assess the size of its effect.
- Let us return to the Kluender et al. (1988) example from earlier, doing it in full.
- Their PERTURBER is based on length of the vowel preceding the [b]/[p].
  - *VOICELESS PERCEPT AFTER LONGER VOWEL*
- Now we can look at all of the data from the original experiment:

![Graph showing wug-shaped curves in phonetics](image)

53. **This setup in the context of speech perception in general**

- This is quite standard, and I suspect the greatest number of wug-shaped curves in linguistics come from speech perception.
- You can do a simple calculation from the constraint weights to calculate the size of the effect of the PERTURBER, which here is about 9 msec.

**HISTORICAL LINGUISTICS**

54. **Main source of this research tradition**

- Kroch (1989)
- Kroch inspects old texts across time in order to track the relative frequencies of syntactic variants, as a language gradually changes.
55. One key finding

- Syntactic changes characteristically follow a sigmoid path (start slow, speed up, end slow).
- Below, with Portuguese possessed noun phrases: we see e.g. *seus livros* ‘his books’ gradually replaced by *os seus livros* ‘(the) his books’

![Graph showing sigmoid curve](image)

- See Blythe and Croft (2012:279-280) for a whole bunch of sigmoid curves in language change.

56. Why should be get sigmoids here? I. Mathematically

- If this is to be attributed to constraint-based linguistics, we must be seeing the weight of some constraint increasing at a constant rate.

57. Why should be get sigmoids here? II. In the speech community

- There is some nice modeling work on the market: Blythe and Croft (2012) and Stadler et al. (2016).
- A plausible conjecture attributes the mechanism to key speakers: younger adolescents.
- They imitate, *and exaggerate* (in terms of Harmony? I think so) the way in which the older adolescents differ from adults, and this keeps on going for centuries.

58. Adding in Perturbers

- These express syntactic preferences that are stable over time (and hence, are exempt from the mechanism just mentioned).
- Since we have a steadily rising *VARIABLE* and constant *PERTURBERS*, you get wug-shaped curves.
- Kroch, subtly, calls this the *constant rate hypothesis* — constant across contexts, *but only* if you measure with Harmony, not raw probability.
59. A recent example from Zimmermann (2017)

- Variable: English *have* is treated ever more often as main verb, rather than an Aux.
- Four contexts where this distinction matters:

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>As Aux</th>
<th>As Main Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>I haven’t any</td>
<td>I don’t have any.</td>
</tr>
<tr>
<td>Inversion</td>
<td>Have you a penny?</td>
<td>Do you have a penny?</td>
</tr>
<tr>
<td>Ellipsis</td>
<td>You have a flair; you really have.</td>
<td>...do.</td>
</tr>
<tr>
<td>Adverbs</td>
<td>He has already the approval of the nation</td>
<td>... already has</td>
</tr>
</tbody>
</table>

- Zimmermann treats this as a change in the weight of a VARIABLE constraint, with four invariant PERTURBERS, getting a wug-shaped curve:

- In right panel: plot the four changes in terms of Harmony values: you get near-perfect parallel lines, as Kroch (1989) predicted.

SEMANTICS/PRAGMATICS

60. Quantifier scope from probabilistic constraint interaction: AnderBois, Brasoveanu, and Henderson (2012)

- Quantifier scope is the ambiguity in sentences like:
  
  *A student saw every professor*

- Quantifier scope judgments reflect a whole set of conflicting constraints.
  - Lexically-specific constraints for each quantifier
  - Also, constraints for linear order, grammatical relations.
➢ I find this approach satisfying since quantifier scope judgments when I make them feel extremely gradient to me.

61. A curve for a tiny subset of AnderBois et al.’s data

- Alas I couldn’t borrow the whole dataset since it employs the Law School Aptitude Test which is copyrighted.
- VARIABLE: linear order (leftward favors broad scope)
- PERTURBER: grammatical relations (subject position favors broad scope)

![Diagram of Wug-shaped curve]

- This is just four data points, yet even these show the narrowing of “vertical” range as we approach the periphery.

WHAT FORMAL MODELS OTHER THAN MAXENT CAN GENERATE WUG-SHAPED CURVES?

62. Preliminary note: models often get lucky

- In a particular case, we can sometimes tweak the parameters to get a good fit — sometimes even better than MaxEnt.
- We need to inspect the models across a broad range of cases to evaluate them rigorously — a good model will work for all cases.

63. Long ago, scholars pondered some very simple models to go from weights to probabilities

- See Cedergren and Sankoff (1974), Sankoff and Labov (1979) for examples and discussion.
- These authors quickly adopted MaxEnt instead once they had learned about it.
64. Multiplication-cum-Cutoff model

- Every constraint violation has the effect of multiplying candidate probability by a particular value, namely the constraint’s weight.
- We will allow constraints to have values greater than one, so that they can increase as well as reduce probability.
- To avoid impossible probabilities, we will impose a ceiling at one by fiat.
- Here is the quantitative signature of the Multiplication-cum-Cutoff model: the truncated fan:

![Graph showing the truncated fan](image)

65. Another version of the quantitative signature: multiple violations of a single VARIABLE constraint yield an asymmetrical sigmoid

- Fitting the Multiplication-cum-Cutoff model to the data of Kluender et al. (post-long vowel series)

![Graph showing the quantitative signature](image)

66. Addition-cum-Cutoff model

- Weights are basically themselves probabilities, and they add up.
- But when you add them, do not go below a floor of 0 or above an ceiling of 1.
- Quantitative signature is the Z-shaped curve:
67. **Returning to the earlier claim of intuitive sensibleness for MaxEnt**

- Neither Multiplication-cum-Cutoff nor Addition-cum-Cutoff makes certainty evidentially expensive.
- On top of this, Multiplication-cum-Cutoff treats evidence differently at different locations on the scale.

68. **Stochastic Optimality Theory (Boersma 1998, Boersma and Hayes 2001)**

- The first effort to make OT a probabilistic theory.
- Trashed in detail, on the basis of its poor performance on wug-shaped curve cases, by Zuraw and Hayes (2017).
- Stochastic OT *throws away* evidence, because it uses the classical OT decision procedure (highest ranking constraint that cares decides; others ignored).
- This is fatal when a PERTURBER has effects all along the Harmony scale — which is exactly what happens in many of the curves you have seen.
- Just to rub it in, here is the bad prediction Stochastic OT makes for Quebec French [l] deletion ((45) above):
The ill-fitting horizontal lines fall where a necessary PERTURBER is too “far away” on the ranking scale to make any difference.

69. There are statistical models that are more sophisticated than MaxEnt

- MaxEnt, viewed from the viewpoint of statistics, is just simple logistic regression — cutting edge statistics in about 1979!\(^4\)
- Here are more recent developments, advocated by experts as more accurate and reliable:
  - **Mixed-effects logistic regression** (now well-entrenched for experimentalists; see Baayen 2008 and Johnson 2011 for textbook coverage)
  - **Random forest models** (Tagliamonte and Baayen 2012)
  - **Neural network models** (Goldberg 2017)
- Any of these models could be used as the decision procedure (EVAL) for probabilistic OT.
- Would any or all of them do better in this role?
  - By this I mean: mimic human behavior/intuitions better
- I have neglected them here because:
  - I’m busy …
  - I know for sure that MaxEnt generates wug-shaped curves (i.e. the math itself makes this plain).
  - I need to study/learn more to know the whether these newer models generate wug-shaped curves as part of their inherent behavior.

---

\(^4\) The MaxEnt sigmoid was discovered in 1845 by Pierre-François Verhulst. But logistic regression (fitting sigmoids to data) had to wait for computers, since you need one to set the weights.
CONCLUSIONS

70. Is the widespread occurrence of wug-shaped curves meaningful, and if so, how?

- Let’s go out on a limb:
  - I raise the possibility that there exist general quantitative principles, along the lines of MaxEnt, that establish the normal patterning of variation in human languages.
  - This is what leads to the repeated appearance of wug-shaped curves when we plot data from the various fields of linguistics.

- I steel myself, though, for gloomy possibilities:
  - Maybe I’ve left out all the nonconfirming cases? (But look at the Gallery: my worst cases are not actually terrible.)
  - The result can be shown to be obvious or trivial (but then why do all those bad models do so badly?)

71. Are the MaxEnt principles specific to language?

- I doubt it.
- See e.g. Smolensky (1986), an early presentation of MaxEnt meant for use in cognitive science in general.
- A decision procedure that effectively weighs and integrates evidence is a fine thing, no matter when in the mind/brain it gets used.
- Indeed, nothing should stop us from seeing wug-shaped curves that combine linguistic with pragmatic/real world principles — which is what AnderBois and Brasoveanu (email correspondence) think is going on with quantifier scope.

References


Jurafsky, Dan and James H. Martin (2019) *Speech and Language Processing* (3rd ed. draft), web.stanford.edu/~jurafsky/slp3/


Sankoff, Gillian S. (ms.) A quantitative paradigm for studying communicative competence. Paper given at the Conference on the Ethnography of Speaking, Austin, Texas.


Smolensky, Paul (1995). On the internal structure of the constraint component Con of UG. Talk given at UCLA.


