Appendix D: Deriving the sigmoid curve from first principles

Appendix to "Deriving the Wug-shaped curve:
A criterion for assessing formal theories of linguistic variation"

Bruce Hayes
UCLA

March 2021

The main body of the paper attempted to demonstrate that the "wug-shaped curve," a combination of two or more sigmoid probability functions, is a pervasive pattern seen in all areas of probabilistic linguistics, and moreover is a natural consequence of adopting MaxEnt Harmonic Grammar as the basis for probabilistic linguistics. However, we are entitled to ask, "if the MaxEnt math is so pervasively correct, *why* should this be so?" The most spectacular explanation would be to say that the MaxEnt formula is innate in humans, part of UG; and try to spot it somehow on our chromosomes. I find this implausible, and the purpose of this Appendix is to lay out a more modest alternative. This would be to say that at least in the computations of language humans respect sensible principles of inductive learning, and that MaxEnt happens to be a effective and explicit mathematical formalization of these underlying principles.

We can break the problem down into two parts. The first concerns how to express bodies of evidence numerically. The second how to interpret the resulting numbers are probability. Harmonic Grammar gives a clear answer to the first question, and in xxxx of the main article I argued that it satisfies four of the five criteria for sensible inductive reasoning given above in (**Error! Reference source not found.**):

(1) *Reviewing the four principles of sensible inductive reasoning*

a. Constraints differ in their evidential force (xxx §**Error! Reference source not found.**).
b. Multiple violations of the same constraint are predicted to make a candidate less probable (xxx §**Error! Reference source not found.**).
c. All evidence from the constraints is duly considered in proportion to their weights; and no evidence is thrown out (xxx §**Error! Reference source not found.**).
e. Candidates become less probable when they compete with powerful rivals (xxx §**Error! Reference source not found.**).

Consider next the question of how to map from Harmony differences to probability; call this function *P*(H). Once again, we can appeal to intuition in various ways that will constrain which of the vast array of functions we might pick to carry out this mapping. To begin:

(2) More evidence is always more persuasive than less evidence.

In our miniature world in evidence is expressed as the Harmony difference between two candidates, it follows that *P*(H) must be *monotonic*, rising steadily upward toward one (as evidence for one candidate accumulates) and falling steadily downward toward zero (as evidence for the opposite candidate accumulates).

We can also appeal to the fourth of the five principles discussed in (§**Error! Reference source not found.**):

(3) Evidence is scaled to make it have less effect as we approach certainty.

This tells about the *derivative* of *P*(H): slope is greatest medially, and declines steadily toward the peripheries. Since probability is constrained to be between zero and one, the slope must indeed; asymptote to zero; else a monotonic function would break out of the 0-1 range.

I further suggest the following as sensible:

(4) Evidence in large quantities is ultimately decisive.

In every case, there exists a quantity of evidence (perhaps unrealistically vast, but even so) that is persuasive. Hence, the values at which the curve asymptotes are indeed zero and one, as in MaxEnt, and not some intermediate value.

Another principle that seems intuitive to me is in (5):

(5) Infinitesimal differences in evidence may only yield infinitesimal difference in probability.

In other words, the curve must be *continuous*, in the mathematical sense, having no sudden leaps.

Taking this a step further, we might suppose that *sensitivity to small differences in harmony*, mentioned in fn. **Error! Bookmark not defined.** of the main text xxx in connection to speech perception, is likewise continuous. This means that infinitesimal differences in harmony cannot create supra-infinitesimal differences in sensitivity. In mathematical terms, the derivative of our curve must be continuous.

At this point, we are in a position to rule out some quasi-serious alternatives: any model that implements the probability minima and maxima at 0 and 1 with a cutoff (e.g. Addition-cum-Cutoff, Multiplication-cum-Cutoff, as in xxx of the main text) will have a discontinuous derivative at the point when the function gets truncated, and thus will fail this criterion.

Lastly, we go out further on a limb, groping for a reason to make our curve *symmetrical.* I suggest:

(6) Nothing other than proximity to certainty may influence the probability function.

It follows that the function must be symmetrical, and thus that we would have to reject the Multiplicative Model given above in (**Error! Reference source not found.**) xxx of the main text (even in a version without cutoff), as well as all but the candidate-noise version of Noisy Harmonic Grammar. I note that here we are probably on the weakest ground empirically; it is not easy to prove that the right sigmoid is always symmetrical.

Putting all of these principles together, we end up with a fairly narrow criterion for the function that maps Harmony onto probability: it must be monotonic, asymptote at zero and one, be continuous with a continuous derivative that has its maximum in the middle, and be symmetrical. This is a fairly stringent description, which is satisfied by two functions discussed here: MaxEnt's logistic function and the cumulative distribution function of the normal distribution used in Noisy Harmonic Grammar; these were shown plotted together in (**Error! Reference source not found.**)xxx of the main text. I am curious to know if there are other candidate functions that satisfy these criteria and are empirically implausible.

Summing up, it seems fruitful and sensible to me not to attribute the MaxEnt sigmoid directly to innate principles. Rather, when we articulate sensible principles of probability-based induction, we are led toward functions that look like the MaxEnt sigmoid. Not all of these principles are equally plausible (e.g. the principle of symmetry), and to the extent that they are not, we expect to find additional probability functions in the real world.