Stochastic constraint ranking

24.964—Fall 2004
Modeling phonological learning

Class 9 (4 Nov, 2004)

BUCLD this weekend

http://www.bu.edu/linguistics/APPLIED/BUCLD/

- Many interesting talks; some even relevant to this course

Review of last time

- The superset problem, as seen in the *azba* language
  - Differences between RCD, BCD, and LFCD
  - LFCD is the only one that words straightforwardly in this case (why?)

- Bayes' Theorem
  - $P(B|A) = \frac{P(B) \times P(A|B)}{P(A)}$

- A goal: try to use Bayes' Theorem to guide probabilistic constraint ranking

Today's agenda

Current approaches to stochastic constraint ranking

- Gradual Learning Algorithm (Boersma 1997, Boersma & Hayes 2001)
- Its use in a larger Bayesian model by Zuraw (2000)

The Gradual Learning Algorithm

Some problems with the RCD, EDCD, etc.

- They don't yield suitably restrictive grammars (*azba* problem)
- They aren't robust to noise (errors)
- They can't handle free variation

The Gradual Learning Algorithm

Making ranking algorithms more robust

- Albro (2000): rankings are shaped by probability of the datum by Bayesian reasons
  - Very rare data (including errors) have little effect on the overall ranking
- Boersma (1997): rankings are shaped by probability of the datum by conservatively adjusting them only tiny amounts in response to each individual token

The Gradual Learning Algorithm

Boersma (1997) *How we learn variation, optionality, and probability*

- Rather than demoting constraints categorically and irrevocably below others, try just nudging them a little each time

/sa/	*si	*s	*ʃ	$\mathcal{F}$(ant)
sa ∼ *ʃa		L	W	W

The Gradual Learning Algorithm

Boersma (1997) *How we learn variation, optionality, and probability*

- Rather than demoting constraints categorically and irrevocably below others, try just nudging them a little each time

/sa/	*si	*s	*ʃ	$\mathcal{F}$(ant)
sa ∼ *ʃa		L→	W	W

The Gradual Learning Algorithm

Implications

- Assumes that constraints are ranking along a continuous scale, rather than in discrete strata
- Suggests that there's a time when the constraints must meet (to cross each other)

The Gradual Learning Algorithm

What do you do when constraints have the same ranking?

The Gradual Learning Algorithm

What do you do when constraints have the same ranking?

- Tesar and Smolensky: they cancel each other out
- Anttila, and others: you get optionality/free variation

Since allowing OT to capture free variation is one of the goals here, it makes sense to try the second possibility

The Gradual Learning Algorithm

But continuous ranking scale is not enough by itself

- If constraints are just points on the scale, then a continuous or discrete scale doesn't even matter from the point of view of using the grammar
- The only way for two constraints to tie (producing optionality) is to be ranked at exactly the same point
- Can only model 50%/50% free variation; but this is certainly not the only pattern we observe in language!

Conclusion: constraints are not just points, but rather *probability distributions*

The Gradual Learning Algorithm

Boersma (1997), Hayes & MacEachern (1998) “Quatrain form in English folk verse”, Hayes (2000) “Gradient Well-formedness in OT”

- Constraints occupy ranges of possible

Image removed due to copyright considerations.
Please see:
Boersma, Paul, and Bruce Hayes. "Empirical tests of the Gradual Learning Algorithm." *Linguistic Inquiry* 32 (2001): 45-86. The MIT Press, Cambridge, MA.

The Gradual Learning Algorithm

Boersma (1997), Hayes & MacEachern (1998) "Quatrain form in English folk verse", Hayes (2000) "Gradient Well-formedness in OT"

When ranges overlap, free variation ensues

Image removed due to copyright considerations.
Please see:

Boersma, Paul, and Bruce Hayes. "Empirical tests of the Gradual Learning Algorithm." *Linguistic Inquiry* 32 (2001): 45-86. The MIT Press, Cambridge, MA.

The Gradual Learning Algorithm

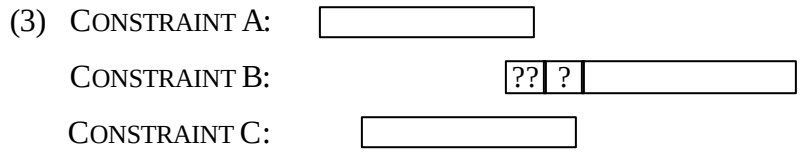
Boersma (1997), Hayes & MacEachern (1998) "Quatrain form in English folk verse", Hayes (2000) "Gradient Well-formedness in OT"

On individual production occasions, actual ranking values are selected from within the ranges

The Gradual Learning Algorithm

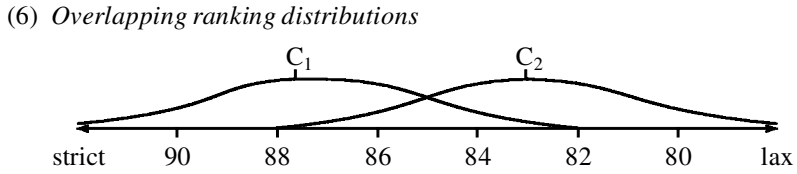
Another crucial insight: not all sections of the range are equally likely

- Hayes (2000): "fringes" of marginal acceptability



Courtesy of Dr. Bruce P. Hayes. Used with permission.

- Boersma (1997): ranges are actually normally distributed probability curves



Courtesy of Dr. Paul Boersma. Used with permission.

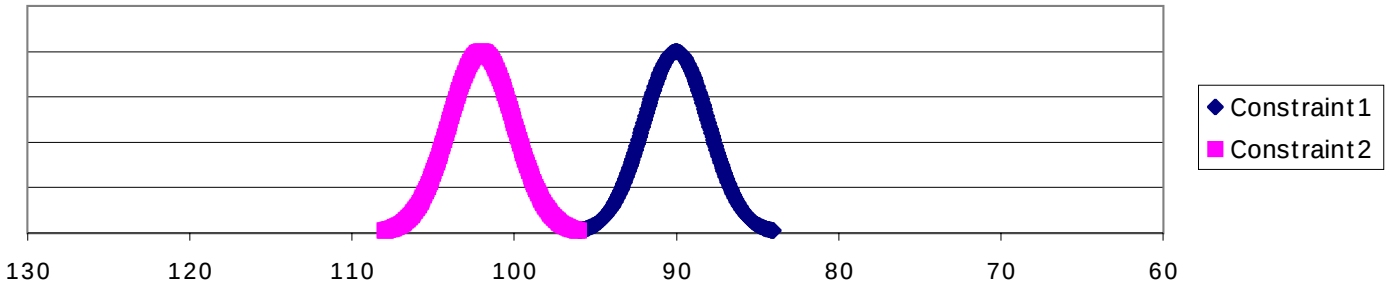
(Assumption: curve shape is the same for all constraints; S.D. = 2)

The Gradual Learning Algorithm

The Gradual Learning Algorithm

Relation between overlap and production probability

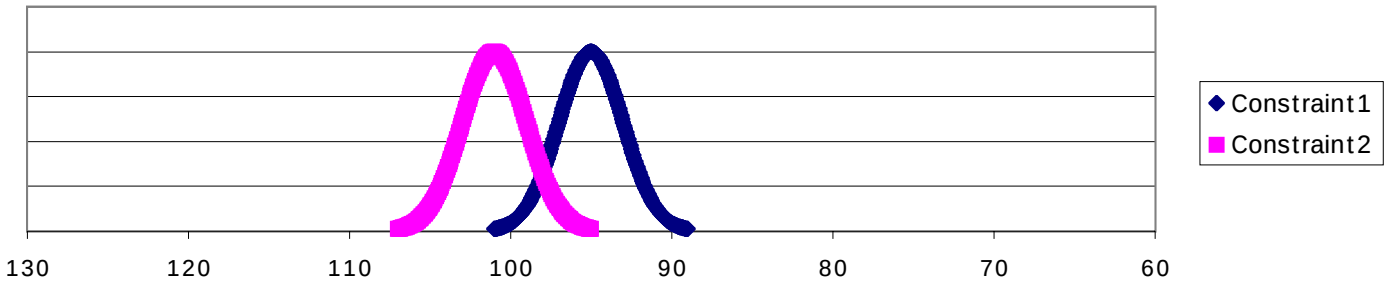
- No overlap: (essentially) no variation



The Gradual Learning Algorithm

Relation between overlap and production probability

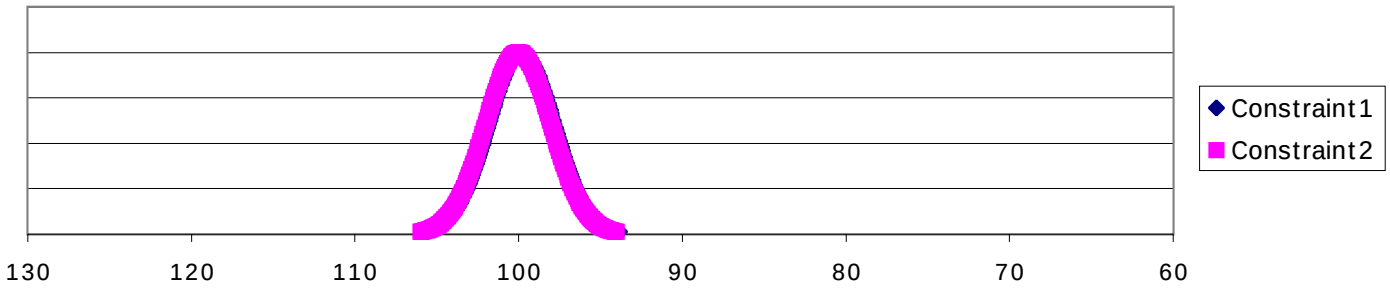
- Partial overlap: occasional variation



The Gradual Learning Algorithm

Relation between overlap and production probability

- Total overlap: free variation



The Gradual Learning Algorithm

Learning in the GLA

The Gradual Learning Algorithm

Learning in the GLA

- All constraints start out equal (or random)
- Datum is heard
- Grammar tries to derive an output for it
- Compares generated output with actual output (= input)

The Gradual Learning Algorithm

If actual output is different from predicted output:

- Construct mark-data pair
- Perform mark cancellation (or comparative tableau)
- Find loser-preferrers and winner-preferrers

(Error-driven learning; exactly the same so far as EDCD)

The Gradual Learning Algorithm

Ranking adjustment:

- Loser-preferrers are nudged downwards slightly
- Winner-preferrers are nudged upwards slightly

The Gradual Learning Algorithm

How much do you nudge the relevant constraints?

- *Plasticity*: the amount a constraint can change by per trial
- A sensible (but perhaps not crucial) idea:
  - Plasticity starts out low, when the learner knows nothing
  - Gradually decreases with age/wisdom
  - In practice, sims tend to start out w/2, end at .002

The Gradual Learning Algorithm

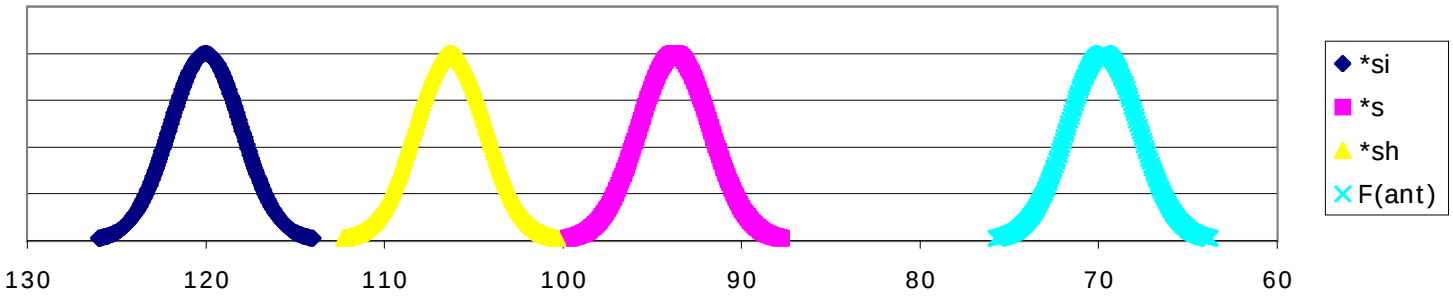
Training cycles:

- Since each datum only has a tiny effect, you need to repeat this many times!
- Number of training cycles depends on number of different input forms, and relative frequency
- OTSoft default is really big (hundreds of thousands of times through data)

The Gradual Learning Algorithm

Results for SaShiAllophonic language

- Recall that SaShiAllophonic is a language with [sa] and [ʃi], but no *[ʃa] or *[si]
- Training file SaShiAllophonic.txt has /sa/, /ʃa/ → [sa], and /si/, /ʃi/ → [ʃi]
- Results:



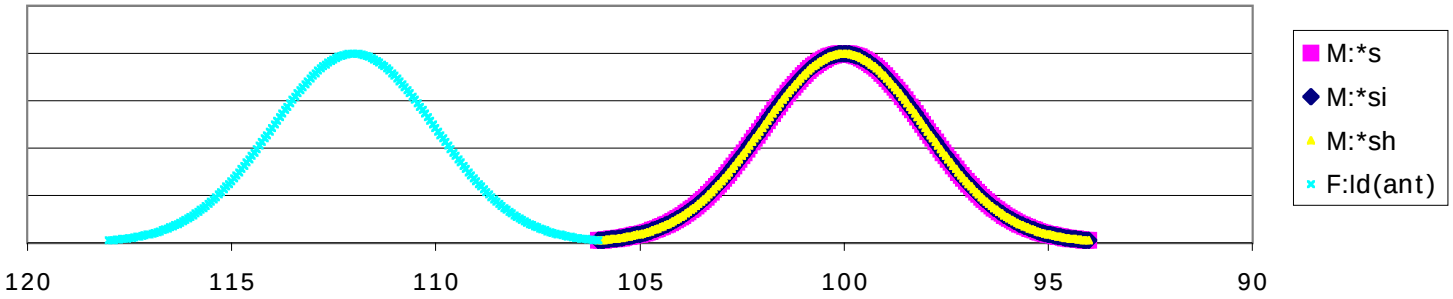
- Why is faithfulness so much lower? (i.e., why not immediately adjacent to *s) Will it continue to fall forever?

The Gradual Learning Algorithm

What about a more realistic training set?

- SaShiAllophonic3: same surface pattern ([sa] and [ʃi]), but this time only UR's /sa/ and /ʃi/ are considered

- Results: same as RCD (why?)



The Gradual Learning Algorithm

Fixing the GLA, a la BC/LFCD

- Hayes (1999) mentions that this is a problem, but that it is unsolved
- OTSoft does give an option for initial ranking states (but Prince & Tesar argue this is not sufficient)
- As far as I know, the more general problem is still unsolved

(How did it do on *azba* for you?)

The Gradual Learning Algorithm

Modeling variation

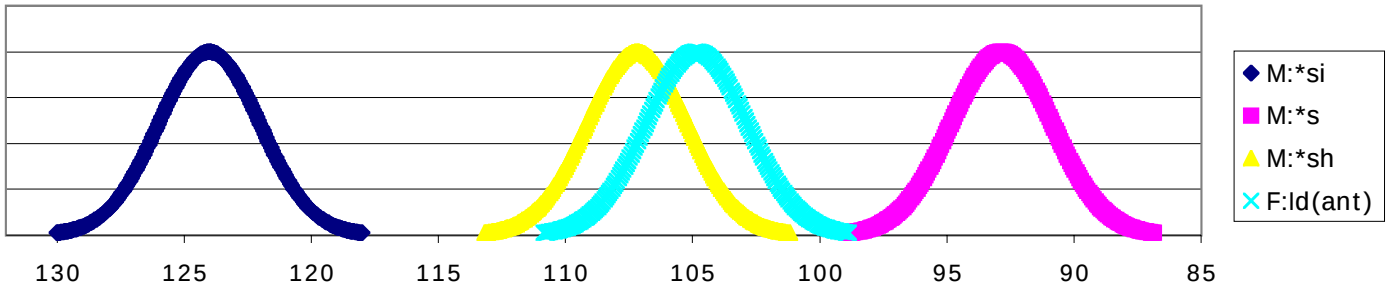
- Suppose /ʃa/ is pronounced as [sa] 80% of the time, but emerges faithfully as [ʃa] 20% of the time

The Gradual Learning Algorithm

Modeling variation

- Suppose /ʃa/ is pronounced as [sa] 80% of the time, but emerges faithfully as [ʃa] 20% of the time

- Ranking needed:



The Gradual Learning Algorithm

Why is nudging symmetrical?

- In cases of variation, learning never really ceases (grammar often makes wrong prediction for a particular instance)
- So, competing constraints are always being nudged around slightly
  - Amount does get smaller and smaller, according to plasticity schedule
- If all you do is demote the offenders, they will continue to push each other downwards forever (pushing other, lower constraints ahead of them)

The Gradual Learning Algorithm

Critique: Keller & Asudeh (2002) "Probabilistic Learning Algorithms and Optimality Theory" *Linguistic Inquiry* 33, 225-244.

- Complain that the GLA is not tested appropriately
- It is not able to learn all attested data patterns
- It doesn't come with proofs
- It has a fundamental misconception about the relationship between grammaticality and corpus frequency

Keller & Asudeh (2002)

Criticism 1: The model hasn't been tested correctly

> "However, no tests on unseen data are reported for the GLA by Boersma and Hayes (2001). The absence of such tests leaves open the possibility that the algorithm overfits the data (i.e., that it achieves a good fit on the training set, but is unable to generalize to unseen data). This problem of overfitting is potentially quite serious. In Boersma and Hayes's (2001) model of light versus dark /l/, six free parameters (viz., the strictness values of the six constraints in the model) are used to fit seven data points (viz., the seven mean acceptability ratings that are being modeled). Overfitting seems very likely in this situation."

(Suggest using held-out (unseen) test data, or k-fold cross-validation (ideally, leave-one-out)

Keller & Asudeh (2002)

But wait a minute...

- For one thing "six free parameters" doesn't seem like an accurate characterization
  - Constraints are not independent entities; they act in grammars
  - Factorial typology of these 6 constraints is unlikely 6! possible grammars
  - In fact, grammar depends on just a few factors (pre-vocalic, pre-tonic, in OO relation)
  - Constraints are not parameters. The rankings are the parameters.

Keller & Asudeh (2002)

But wait a minute...

- Furthermore, simulations are being done on idealized data
- "Overfitting" means that an unseen form is not handled correctly; but here, an unseen form means an *unseen patter*
- So the only way to test on an unseen form is to see if the grammar behaves right on a *type* of word that it's never seen before
- Not clear if we should consider it a failing if it does not...
  - Not even clear what an unseen pattern would be here, actually, since all possibilities according to relevant constraints have already been considered

Keller & Asudeh (2002)

Criticism 3 (going out of order): no proofs

- Boersma has made some sketches, but there is no proof provided that the GLA is guaranteed to converge on a certain class of data patterns.

- Relatedly, there is no way to tell that learning is failing
  - This means that if there is no possible ranking, the GLA will continue to try forever anyway.

Keller & Asudeh (2002)

This is a genuine criticism (and the termination problem is especially worrisome)—but is it fatal to the approach?

- Non-termination, in particular, may not be fatal
- A practical problem, but does it make it implausible as a model of humans?
- What do human children do when faced with an unlearnable language?
  - Note that this doesn't mean a pattern which can't be captured without resorting to all-out faithfulness (a superset language)
  - This means a language that can't be captured under ANY ranking, using $\mathcal{M}$ or $\mathcal{F}$

Keller & Asudeh (2002)

Criticism 2: Counterexamples (things the GLA can't learn)

Table 3
Data set that the GLA cannot learn (log-transformed mean acceptability scores for word order in German; Keller 2000b, experiment 10)

Image removed due to copyright considerations.
Please see:
Keller and Dr. Ash Asudeh. "Probabilistic Learning Algorithms and OT."
Linguistic Inquiry 33, no. 2 (2002): 225-244. The MIT Press, Cambridge, MA.

Keller & Asudeh (2002)

Example:

- S = *der Polizeibeamte* ‘the policeman’
- O = *der Dieb* ‘the thief’
- V = *erwischt* ‘captures’
- pron = *er, ihn* ‘he, him’

. . . daß ihn der Polizeibeamte erwischt	+0.2412
. . . daß den Dieb er erwischt	−0.0887
. . . daß erwischt er den Dieb	−0.1861

Keller & Asudeh (2002)

Boersma (2004) "A stochastic OT account of paralinguistic tasks such as grammaticality and prototypicality judgments"

- Points out that K & A are comparing candidates which are arguably not even of the same underlying form
  - Pronominalize the subject vs. the object
  - At the very least, there must be additional constraints involved in choosing which to pronominalize, which would remove the harmonic bounding problem

Keller & Asudeh (2002)

Boersma (2004) points out there's still an interesting puzzle:

Candidate b. is judged marginally plausible, even though it has the same corpus frequency as totally impossible things (0)

Boersma (2004)

Boersma (2004): suggests that these judgments are done by pairwise comparison

(11) *Candidate* b *has zero frequency but is not the least grammatical*

|S = 'the policeman$_i$', O = 'the thief', V = 'capture', topic = 'the policeman$_i$'|	VERB 105.0	NOM 98.0	PRO 98.0	accept-ability	corpus freq.	pairwise freq.
☞ a. *dass er den Dieb erwischt*				√	100%	100%
b. *dass den Dieb er erwischt*		*	*	??	0%	66%
c. *dass erwischt er den Dieb*	*			*	0%	34%
d. *dass erwischt den Dieb er*	*	*	*	*	0%	0%

Candidate a. is better than everyone, but candidate b. is at least better than 2/3 of the others

- Similar to current work by Coetzee (UMass diss in progress)

Boersma (2004)

Note that this alone can't be totally right

- Imagine a language with epenthesis in some contexts
- For an input that demands epenthesis, the winner is the epenthesizing candidate
- But there are infinitely many losers that epenthesize more
- All of these will be better than any candidate that violates a higher ranked constraint

Keller & Asudeh (2002)

Although K&A's example isn't that great, there are cases which the GLA doesn't handle all that well

- (If you are interested in seeing some, and perhaps trying to understand why not, I can provide some from my own experience!)

Keller & Asudeh (2002)

Criticism 4: Wrong to assume monotonic relation between well-formedness and frequency

- Things other than grammaticality/WF play a role in determining corpus frequency

Keller & Asudeh (2002)

Boersma (2004): of course they do

- Subcategorization example isn't even relevant, because once again, it is comparing the relative frequencies of things that correspond to different UR's

- The purpose of the GLA is to determine relative frequency of competing options for the same word, not to explain why some words are more common than others

Keller & Asudeh (2002)

This does, however, raise a larger question

- What other factors may influence the choice of a variant?
- Should they be modeled in the phonology?
- Once they are included, is there actually such a thing as free variation?

Is there free variation at all?

My conjecture: a tentative yes

- In many cases, variation can be explained by reference to sociolinguistic variables
  - A lot of careful work in recent decades has gone into showing how that this is true, and how it can be modeled (e.g., the VARBRUL framework)

- However, we still need to constrain the possible variants
  - Connection to sociolinguistic variables can tell us things about why one is chosen over the other, but not why those were the choices to start with
  - Something probabilistic or non-deterministic in phonology may be needed to create those forms in the first place, which then get latched on to as social markers

The issue of errors

The problem of distinguishing errors from licit (but rare) patterns

- Parallel to problem of distinguish impossible patterns from marginally possible but rare (or even non-occurring) patterns

Looking ahead

- Discussion of Zuraw
- You should be thinking about projects (and talking with me, if you haven't already)
- No class next week; this should give some time to be pondering and starting to do something concrete
- New readings: Boersma (2004) "A stochastic OT account…", and Everett and Berent (1997) "Comparative optimality…"