Investigating the consequences of iterated learning in phonological typology

Coral Hughto

University of Massachusetts Amherst

Society for Computation in Linguistics (SCiL)
6 January 2018

Introduction

- Traditional goal of typology: predict divide between attested and unattested patterns
  - Grammar should be able to represent all and only attested patterns
- Some recent work combines a theory of grammar with a theory of learning to generate probabilistic typological predictions
  - Pater 2012, Staubs 2014, Stanton 2016, O'Hara 2018, among others
- This approach draws on differences in learnability to explain differences in frequency of attestation

In This Talk

- I examine the predictions of combining Maximum Entropy (MaxEnt; Goldwater & Johnson 2003) grammar with one of two agent-based learning models
  - Reviewing previous work with **Interactive** learning model
  - Introducing follow-up work with **Iterated** learning model
- Emergent learning biases from both learning models:
  - Bias away from constraint cumulativity (gang effects)
  - Bias away from variability (such that agents accumulate probability on one output per input)
    - See Zuraw (2016) on Polarized Variation
- With Iterated learning model, bias away from variability only occurs for longer learning times

MaxEnt

- My (and much other) work assumes a weighted-constraint grammatical theory as its base (but see Stanton 2016)

/*In*$_1$/	3 X	2 Y	*H*	*p*
→A		-1	-2	0.73
B	-1		-3	0.27

/*In*$_2$/	3 X	2 Y	*H*	*p*
→C	-1		-3	0.73
D		-2	-4	0.27

- Harmony score (H) = weighted sum of constraint violations
  - $H(x) = \sum_{i=1}^{n} W(C_i) * C_i(x)$
- Probability (p) = proportion of exponentiated Harmony out of sum over competing candidate set
  - $p(x) = \frac{e^{H(x)}}{e^{H(x)}+e^{H(y)}+e^{H(z)}\ldots}$

Gang Effects

/In_1/	3 X	2 Y	H	p
→A		-1	-2	0.73
B	-1		-3	0.27

/In_2/	3 X	2 Y	H	p
→C	-1		-3	0.73
D		-2	-4	0.27

- Weighted constraint grammars allow for cumulative constraint interaction (a.k.a. gang effects)
- Multiple violations of (a) lower-weighted constraint(s) can cumulatively outweigh one violation of a higher-weighted constraint

Gang Effects

/In_1/	3 X	2 Y	H	p
→A		-1	-2	0.73
B	-1		-3	0.27

/In_2/	3 X	2 Y	H	p
→C	-1		-3	0.73
D		-2	-4	0.27

- This property of weighted constraint grammars has been criticized for overpredicting the space of typological possibilities (e.g. Legendre et al. 2006, but see Pater 2009)
- Despite overprediction, the extra representational power may be desirable, e.g.:
  - stress windows (Staubs 2014)
  - "general-case" neutralization (Hughto and Pater 2017)

Previous work: Hughto and Pater 2017

- How to limit overprediction of gang effects with weighted constraints?
- Perhaps considerations of *learnability*
  - Gang effect patterns require a particular balance between the constraint weights
- Paired MaxEnt with an agent-based, interactive learning model to generate gradient typological predictions
- Interactive learning model: simulated learning agents play a kind of imitation game

Previous work: Hughto and Pater 2017

- In the interactive learning model, two agents take turns in the roles of teacher and learner
  - Agents know: constraints, initial weights, inputs and corresponding output candidates
  - There is no target grammar
- In each run of the simulation, the agents exchange data for some number of learning steps
- $A_1 \leftrightarrow A_2$
- Agents' final grammars are categorized as belonging to a pattern in the typology
- The distribution of languages learned across multiple runs is taken as the predicted typology

Palatalization Typology

- Palatalization typology: possible contrast patterns between /s/ and /ʃ/ (before [i] vs other vowels [a]) (Carroll 2012)
- Constraints: No[ʃ], No[si], Ident
- With these constraints, 5 possible patterns:
  - (44%) Total Neutralization
    [si], [sa]
  - (37%) Full Contrast
    [si], [ʃi], [sa], [ʃa]
  - (10.3%) Complementary Distribution
    [ʃi], [sa]
  - (8.2%) Special-Case Neutralization
    [ʃi], [sa], [ʃa]
  - (0.5%) General-Case Neutralization (gang effect)
    [si], [ʃi], [sa]

General-Case Neutralization (GCN; gang effect)

weights	3	2	2	
/sa/	NO[ʃ]	NO[si]	IDENT	*H*
☞sa				0
ʃa	-1		-1	-5
/ʃa/	NO[ʃ]	NO[si]	IDENT	
☞sa			-1	-2
ʃa	-1			-3
/si/	NO[ʃ]	NO[si]	IDENT	
☞si		-1		-2
ʃi	-1		-1	-5
/ʃi/	NO[ʃ]	NO[si]	IDENT	
si		-1	-1	-4
☞ʃi	-1			-3

Results: Avoids gang effect

- Zero: Agents initialized with constraint weights at zero
- Random: Agents initialized with sampled weights, 0-10
- Sampling: Just sampling constraint weights, no interaction

Type	Observed	Zero	Random	Sampling
Total Neut.	44%	46.6%	25.7%	16.8%
Full Contrast	37%	48%	47.5%	41.3%
Comp. Dist.	10.3%	2.6%	7.7%	8.3%
Contextual Neut.	8.2%	2.7%	8%	8.4%
General-case Neut.	0.5%	0.1%	11.1%	25%
r^2		0.96	0.63	0.17

Discussion

- Combining MaxEnt + learning model:
  - Keeps the representational power of weighted constraints
  - Restricts typological overprediction by assigning low probability to typologically rare or unobserved patterns, including gang effects
- The Interactive learning model additionally tends towards accumulating probability on one output candidate over its competitors
- Effects are robust across different parameter settings tested
- Potential issue: in the interactive learning model, agents are not working towards a target grammar
- Do these biases still emerge in a model where agents are tasked with learning a target grammar?

Iterated Learning Model

- Staubs 2014: Iterated learning reduced the predicted probability of gang effects in stress window systems
- The Iterated learning model approximates the transmission of a language across generations
- One agent serves as the "teacher" (the target grammar) for a "learner agent"
- After a period of learning, the learner becomes the teacher for a new learner, and the process repeats for some number of generations
- $A_1 \rightarrow A_2$, then $A_2 \rightarrow A_3$, then $A_3 \rightarrow A_4$ ...

How it works

- $A_1 \rightarrow A_2$, then $A_2 \rightarrow A_3$, then $A_3 \rightarrow A_4$ ...
- Each agent begins with a set of initial constraint weights (e.g. zero, or randomly sampled)
- In each learning step:
  - An input is randomly selected, and each agent samples an output according to its current grammar
  - If the outputs are different, the learner updates its constraint weights using the Perceptron update rule (see also Stochastic Gradient Descent, HG-GLA)
    - New Weights = Old Weights + (Teacher's Violations - Learner's Violations) * Learning Rate
- From initial teacher to final learner = 1 run of the simulation
- The distribution of languages learned across multiple runs of the simulation is taken as the predicted typology

Minimal Working Example

/In_1/	X	Y
A		-1
B	-1	
/In_2/		
D		-2
C	-1	

- Three possible patterns:
  - BC : w(Y) > w(X)
  - AD : w(X) > 2w(Y)
  - AC : 2w(Y) > w(X) > w(Y) (Gang effect)

Minimal Working Example

$/In_1/$	1 X	3 Y	H
A		-1	-3
→B	-1		-1
$/In_2$			
→C	-1		-1
D		-2	-6

$/In_1/$	3 X	1 Y	H
→A		-1	-1
B	-1		-3
$/In_2$			
C	-1		-3
→D		-2	-2

$/In_1/$	3 X	2 Y	H
→A		-1	-2
B	-1		-3
$/In_2$			
→C	-1		-3
D		-2	-4

- Three possible patterns:
  - BC : w(Y) > w(X)
  - AD : w(X) > 2w(Y)
  - AC : 2w(Y) > w(X) > w(Y) (Gang effect)

Iterated Learning Simulations

- The iterated learning model was run 1,000 times
- Two initial constraint weight conditions were tested:
  - (Zero-Init) Agents' initial constraint weights were zero
  - (Rand-Init) Agents' initial constraint weights were randomly sampled from a uniform distribution between 0-10
- Each learner agent learned from its teacher for 1,000 learning steps, and there were 50 generations in each run
- Baseline prediction is the proportion of possible weights that generate each pattern type
  - BC: 0.5, AD: 0.25, AC: 0.25

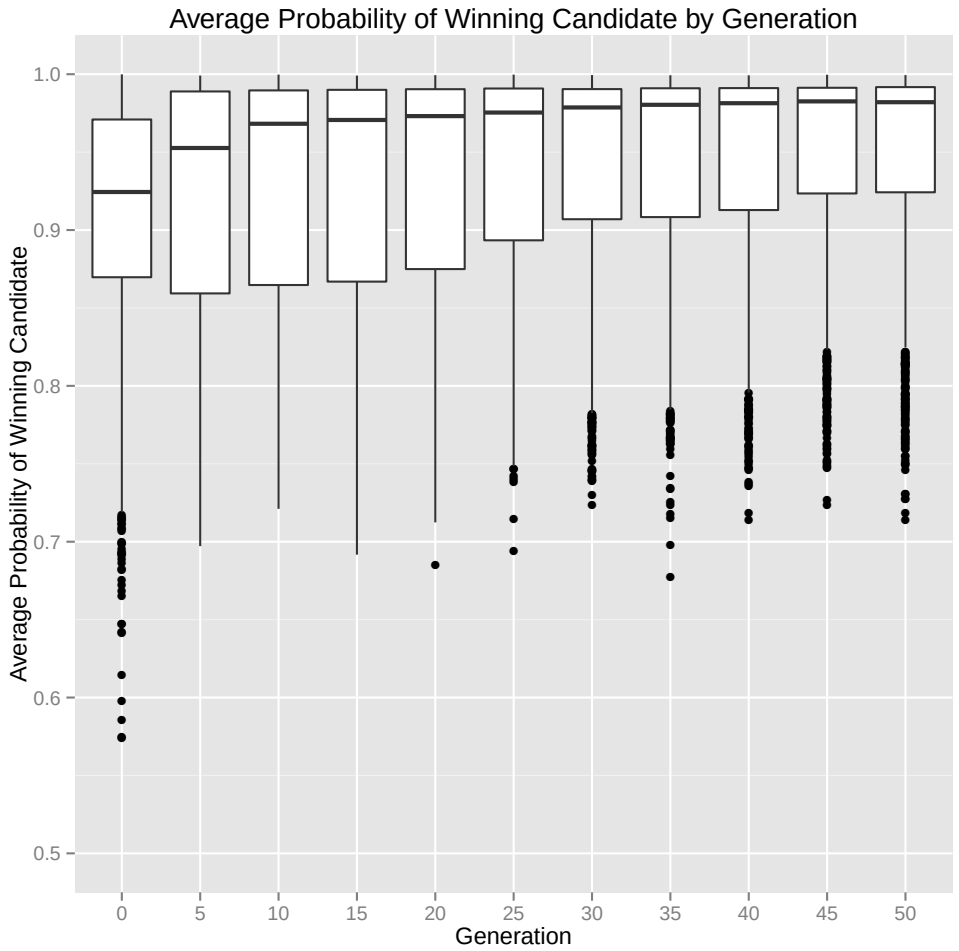
Simulation Results (1,000 learning steps)

- Like the interactive learning model, the iterated learning model results show a bias away from the gang effect pattern
- In both the Zero-Init and Rand-Init initial weighting conditions, the model results reduced the predicted probability of the gang effect AC pattern, relative to the sampled baseline estimation

Pattern	Sampling	Zero-Init	Rand-Init
BC	0.50	0.55	0.55
AD	0.25	0.43	0.30
AC	0.25	0.03	0.15

Simulation Results (1,000 learning steps)

- Results show a bias away from variation (graph shows Rand-Init)



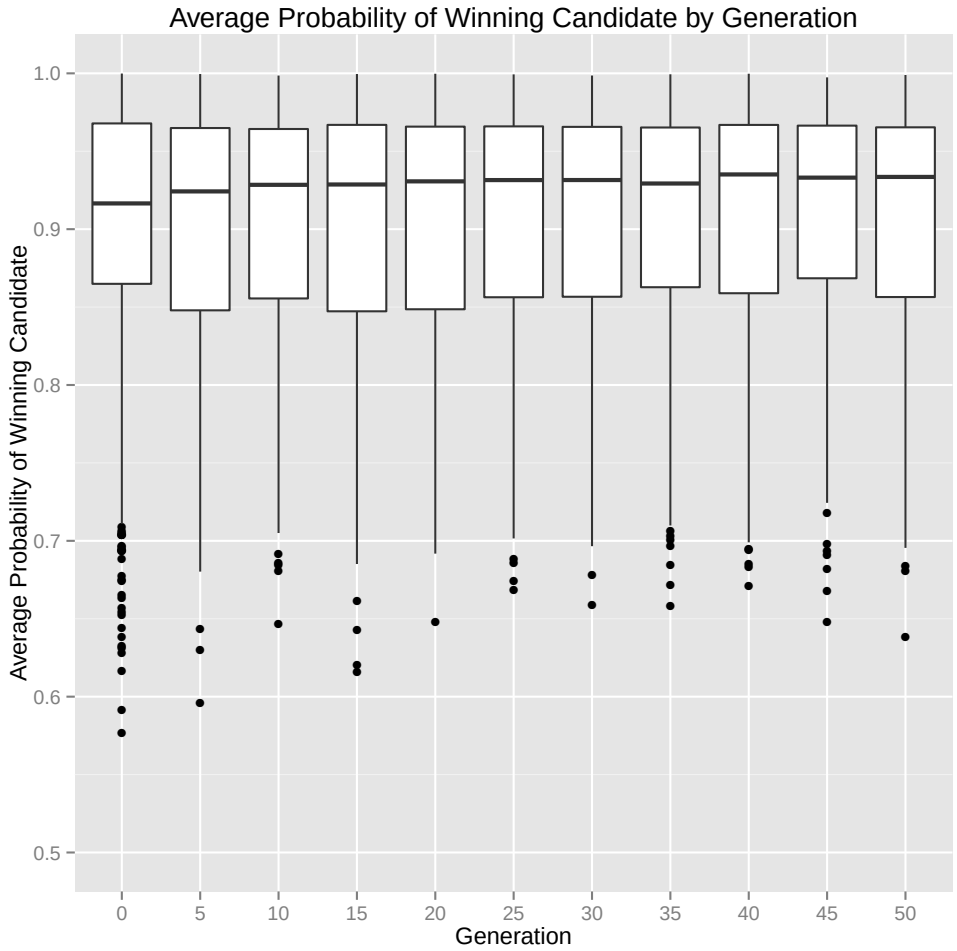
Simulation Results (200 learning steps)

- In the iterated learning model, the bias away from variation is sensitive to the learning step parameter
- With shorter learning time, 200 learning steps per generation, the bias away from the gang effect AC pattern still emerges:

Language	Sampling	Rand-Init
BC	0.50	0.60
AD	0.25	0.24
AC	0.25	0.16

Simulation Results (200 learning steps)

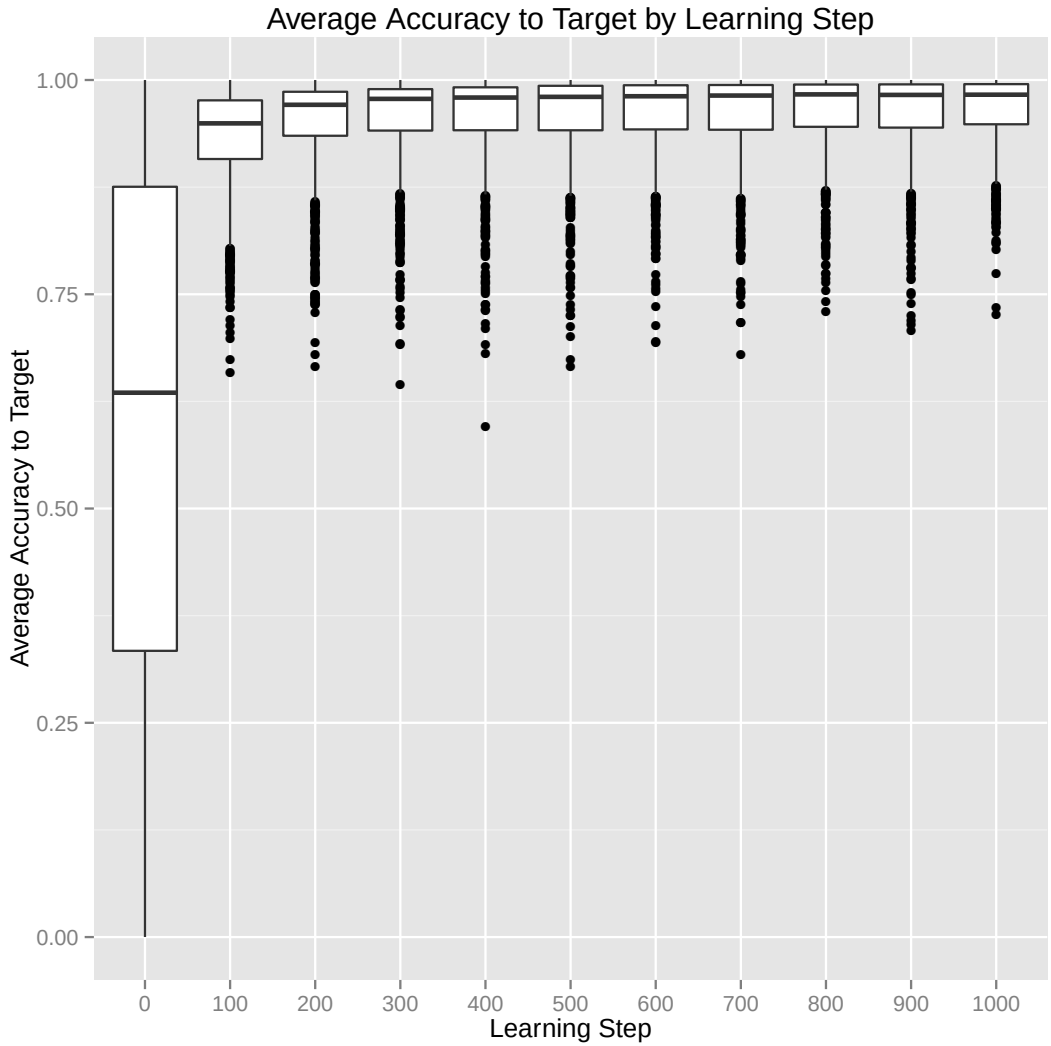
- However, the bias away from variation does not visibly emerge:



Variation depends on number of learning steps?

- Not sure why more learning steps correlates with decreasing variability across generations
- More learning steps expected to correlate with higher accuracy to the target distribution
  - So, more learning steps should mean less deviation from the initial distribution, not more
- But, the difference in accuracy acheived between 200 and 1,000 learning steps in this system isn't that significant anyway

Target Accuracy



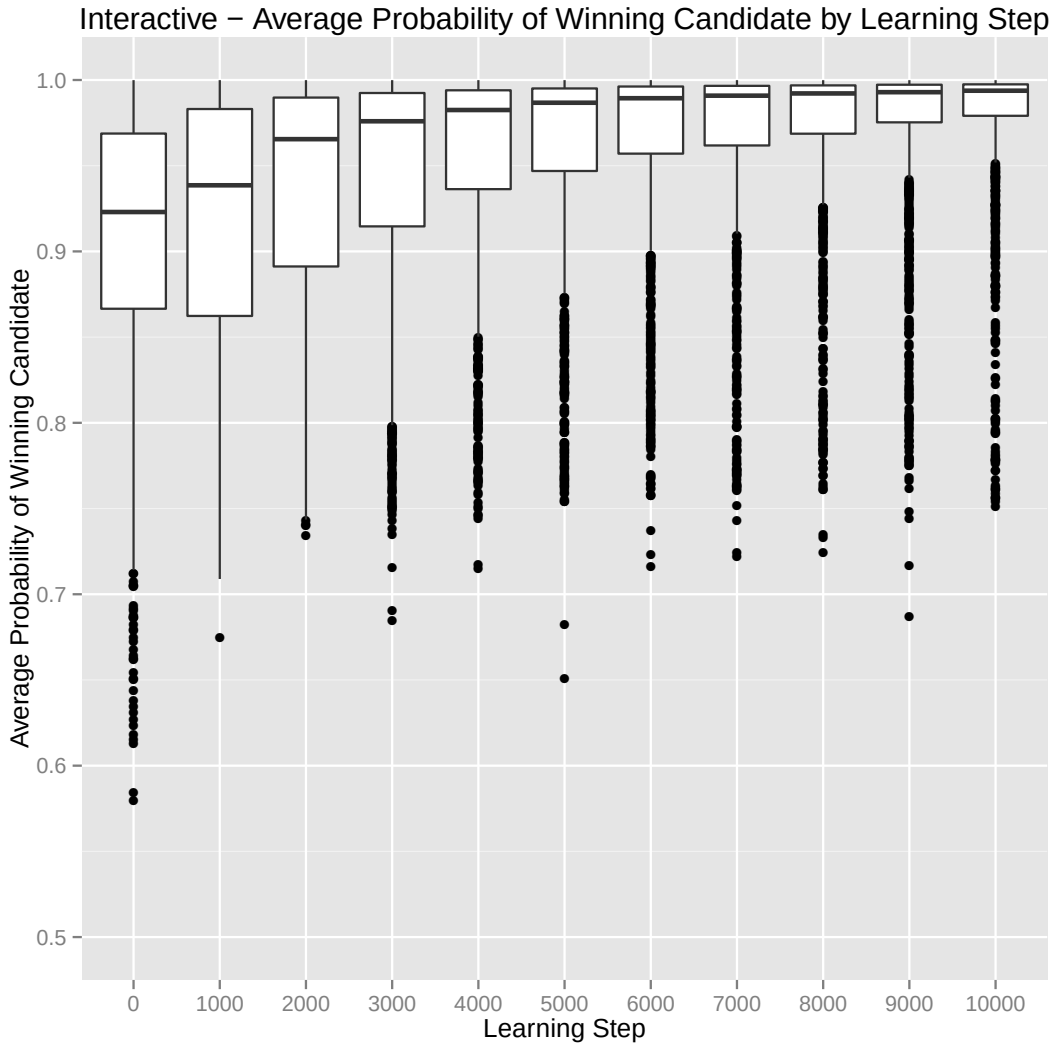
Summary

- Traditional goal of typology: predict divide between attested and unattested patterns
- A growing line of research additionally investigates the role of learning biases in shaping typology
- Can generate probabilistic typological predictions by combining a grammar theory with a learning theory
  - e.g. MaxEnt and an agent-based learning model
- Both Interactive and Iterated learning models demonstrate:
  - Bias away from gang effects (cumulative constraint interaction)
  - Bias away from variation
- The iterated learning model only produces a bias away from variation at higher learning step values

Thanks!

Thanks to Joe Pater, Gaja Jarosz, audiences at UMass, PhoNE, NECPhon, mfm, CLS, and everyone here.

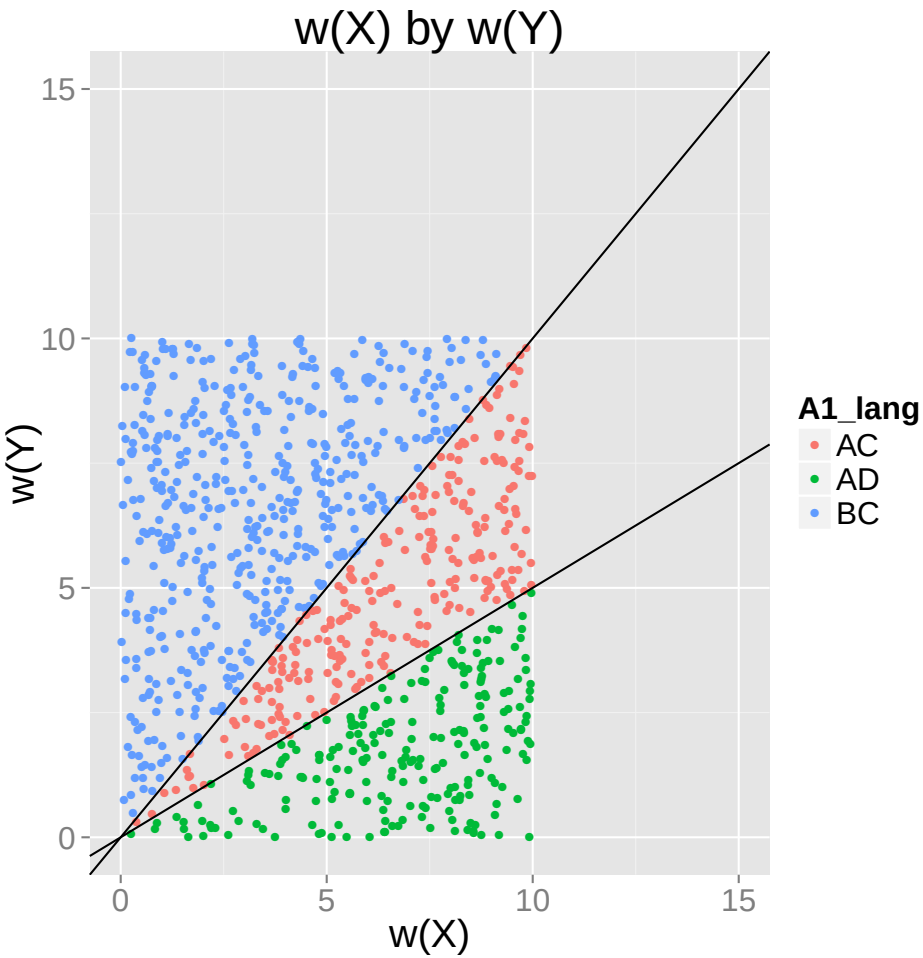
Avoid variability (Interactive)



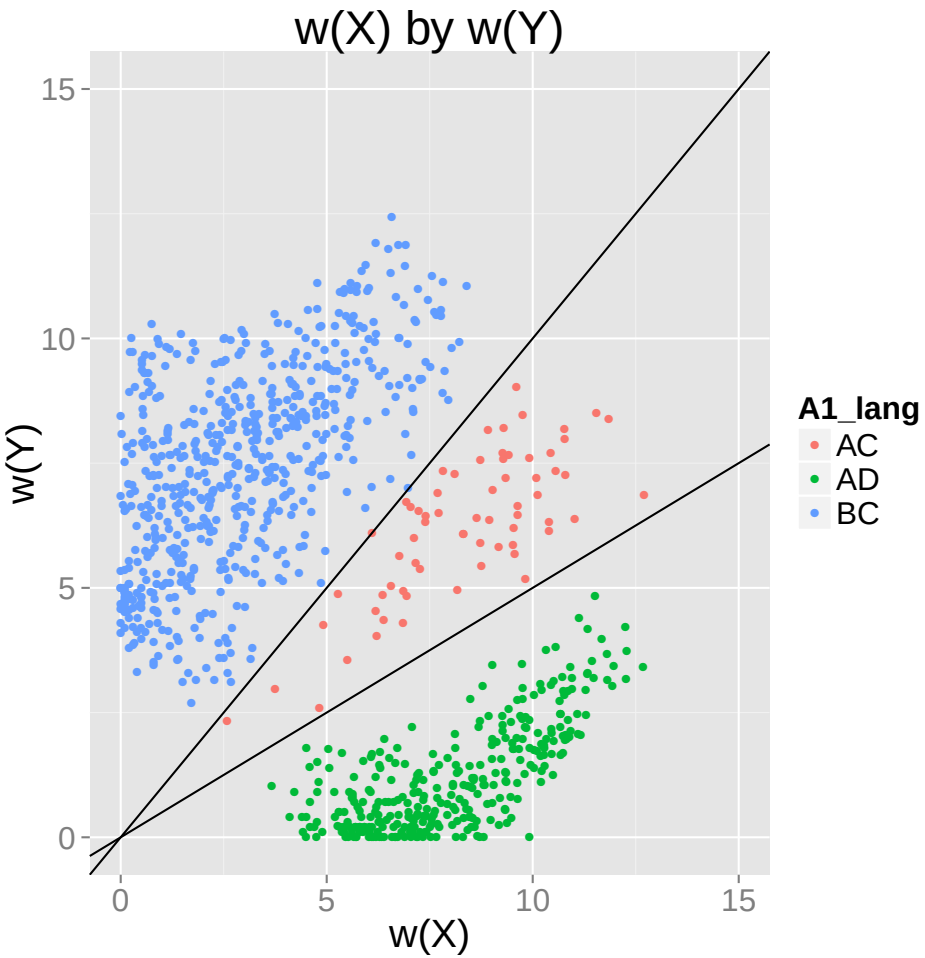
Learning Simulations

- The Interactive learning model was run 1,000 times
- Agents were initialized with random constraint weights sampled from a uniform distribution ranging 0-10
- Agents interacted for 10,000 learning steps
- Baseline prediction is the proportion of possible weights that generate each language type
  - BC: 0.5, AD: 0.25, AC: 0.25

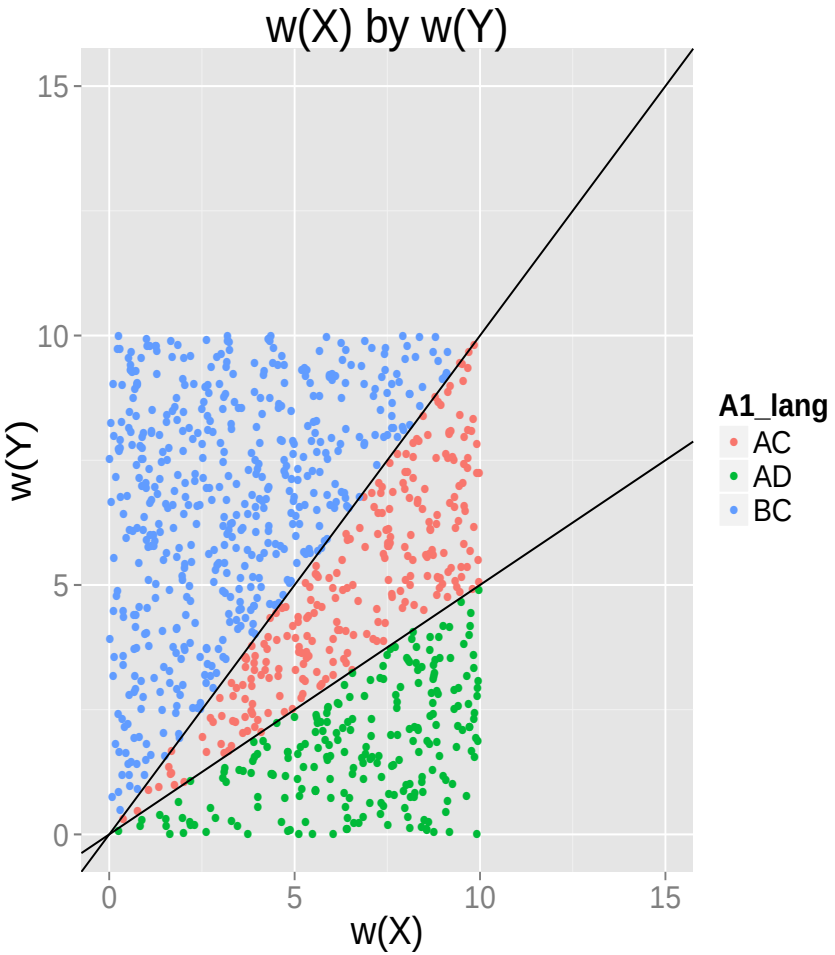
Simulation Start

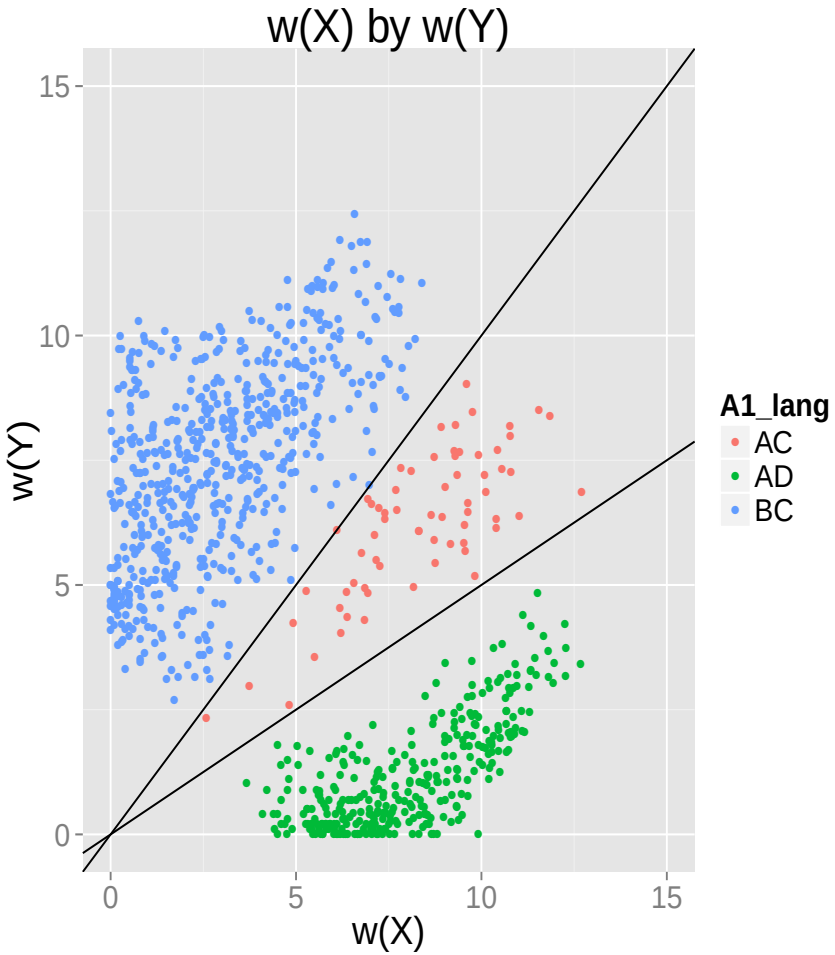


Simulation End

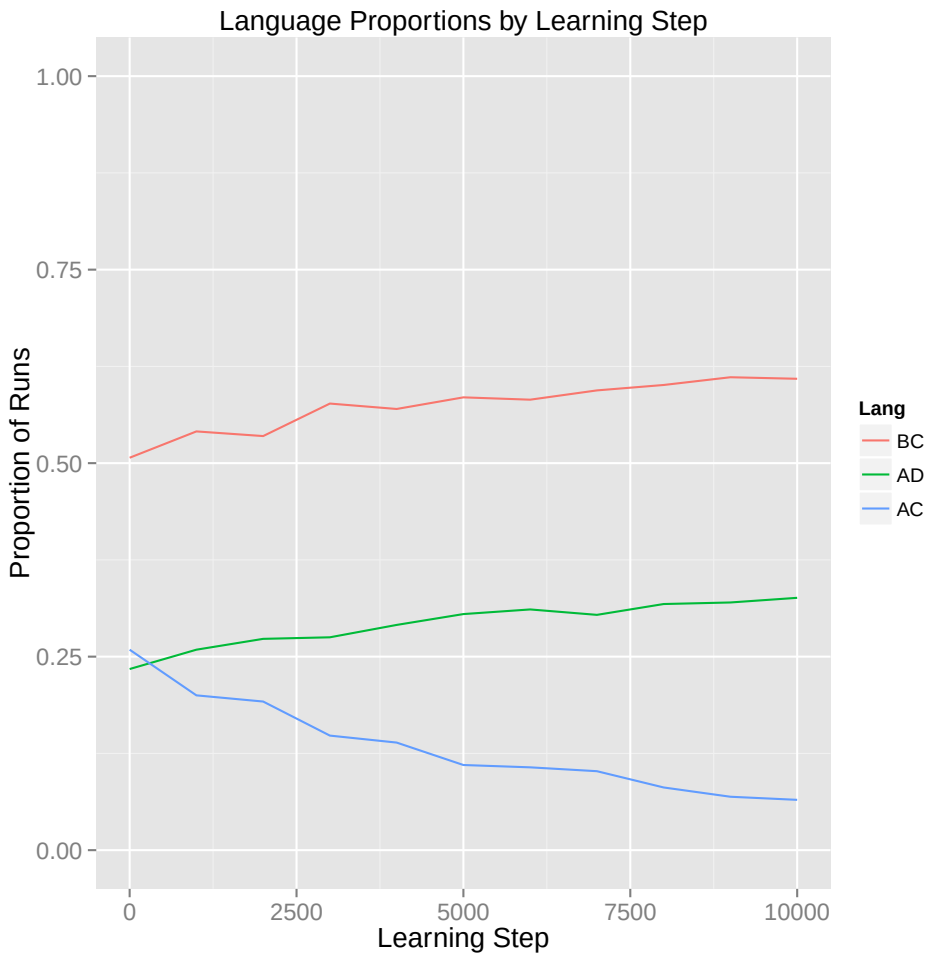


Simulation Results





Simulation Results (Interactive)



Simulation Results (Interactive)