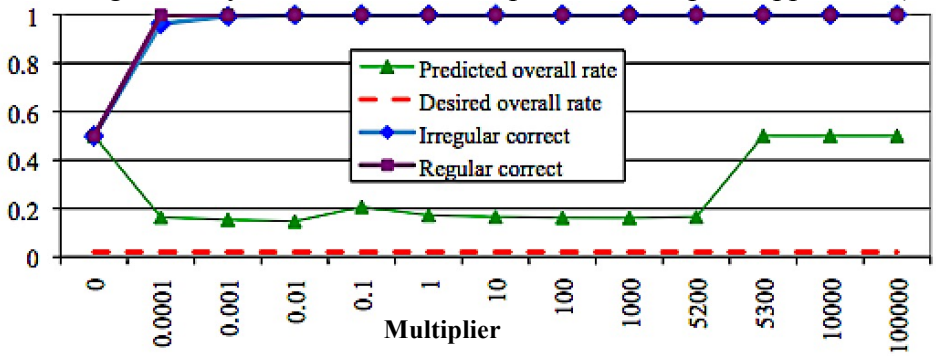
Learning lexical trends together with idiosyncrasy: MaxEnt versus the mixed logit

Jesse Zymet, UC Berkeley

In experiments testing for knowledge of phonological variation, speakers have been found to frequency match lexical trends *as well as* display knowledge of item-specific idiosyncrasies. How can the two kinds of knowledge be modeled together? Zuraw (2000) used OT with variable rankings and the Gradual Learning Algorithm to learn fixed pronunciations together with trends across the lexicon, but recent findings suggest that OT is inadequate for capturing certain cases of variation relative to Harmonic Grammar (Zuraw & Hayes 2017). A recent MaxEnt-based approach recruits general constraints for lexical trends and lexical(ly indexed) constraints for lexical idiosyncrasies (Moore-Cantwell & Pater 2016, Zuraw & Hayes 2017, Tanaka 2017). Using learning simulations, I argue that this approach overfits lexical constraints to the data. Suppose we have a set of 50 regular forms, and 1 irregular form. We can vary frequencies in the dataset by some multiplier *m*, keeping the **2%** irregularity rate constant (e.g., 5000 regulars, 100 irregulars)—such paradigms are attested, with irregularity rates found to be matched by speakers in nonce probe studies (*cf.* Hayes & Londe 2006 on *hi:d*-stems in Hungarian vowel harmony). We have two constraints: BEREGULAR, governing regularity rate across the data, and BELEXICAL, determining which words are regular versus irregular. The simulations were run in Excel's Solver: in early learning trials we input small *m*, while in later trials we input large *m*; in each trial, weights were initialized at 0, then optimized to match frequencies. The graph below reveals the problem: before BEREGULAR reaches a weight that frequency-matches irregularity rate, the learner acquires lexical idiosyncrasies with such accuracy using BELEXICAL that BEREGULAR's weight vacillates and plummets to zero, rendered ineffective (see Albright & Hayes 2006 on a similar problem with prior approaches).

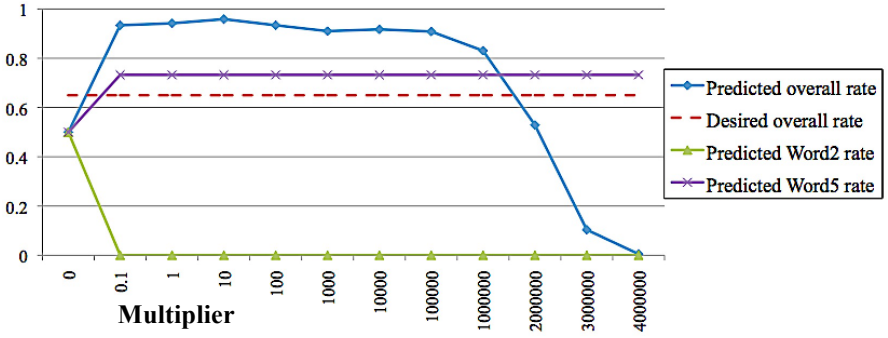
UR	SR	Freq.	BEREGULAR	BELEXICAL
/Regular/	[Regular]	50*m*		
	[Irregular]	0	1	1
/Irregular/	[Regular]	0		1
	[Irregular]	1*m*	1	



We try with another dataset: twelve words undergoing a variable process at different rates (*cf.* Zuraw 2016, Zuraw & Hayes 2017, Tanaka 2017, Zymet 2018), with the overall regularity rate being **65%**. Word3 has 1*m* tokens that undergo out of 100*m* tokens; Word4 has 12*m* tokens that undergo out of 100*m*; etc. We have a general constraint APPLY, whose weight should predict the overall regularity rate, and APPLY1, APPLY2, ..., APPLY12 (with weights allowed to go negative), whose weights should predict lexical rates. Using the same methods as before, we find that lexical constraints fit to lexical rates perfectly, while the

Word	Rate	Word	Rate	Word	Rate
1	0.00	5	0.73	9	0.99
2	0.00	6	0.88	10	0.99
3	0.01	7	0.98	11	0.99
4	0.12	8	0.99	12	0.99

grammar behaves erratically without experiencing sustained periods of frequency matching to the overall rate.



It will be shown that varying settings for the MaxEnt penalty term—e.g., setting strong penalty for nonzero lexical weights and weak penalty for nonzero grammatical weight—delays the overfitting problem, but fails to circumvent it. To avoid overfitting, it is proposed that the model be endowed with a **generality bias**, beyond MaxEnt's penalty term: it must privilege general constraints over lexical constraints to avoid overfitting of the latter (*cf.* Boersma 1998, Albright & Hayes 2006). MaxEnt, essentially a canonical logit, is replaced with the mixed-effects logit, i.e. **Mixed-Effects MaxEnt**. General, grammatical constraints, insensitive to what the sample dataset is, are encoded as fixed effects, whereas lexical constraints are encoded as levels of a random intercept. Generality bias is rooted in the fixed-random distinction: coefficients of the levels of the random intercept are determined by a weighted average the word-specific rate *and* the overall rate in the dataset. Using R, the mixed logit is shown to succeed in modeling the propensity dataset: it finds a grammatical weight that closely predicts overall rate (actual rate = 65%; predicted rate = 69%), and lexical weights that predict each word rate down to the two decimal points.

The mixed logit is used to model lexical variation in Slovenian velar palatalization and French liaison. Below we illustrate the mixed logit analysis for velar palatalization (/obla**k**-itsa/, [obla**ʧ**-itsa], cloud-DIM). Jurgec (2016)'s data—a corpus of ~5.7 million word tokens extracted from the *Gigafida* (Logar-Berginc et al. 2012)—reveal that it is triggered only by a certain set of suffixes, with the suffixes triggering at different rates; moreover, the variation appears to be conditioned by phonological factors (e.g., target velar identity, whether palatalizing would avoid a geminate, whether suffix begins with a front vocoid). I use data collection methods similar to Jurgec's, obtaining a corpus of ~3 million tokens extracted from the same database. I ran a mixed logit on the data, encoding all of Jurgec's phonological factors and log word frequency as fixed effects, and suffix identity (in the table below), stem identity, and whole word as random intercepts. Including stem and suffix identities as random intercepts resulted in a superior model according to AIC and BIC, relative to a baseline model with only word as a random intercept, or with only word and stem or word and suffix as intercepts. Two main effects were significant predictors: target identity and geminate avoidance. In my data, *k* undergoes palatalization more regularly than *g*; and palatalization applies nearly obligatorily to avoid /{k, g}+k/ sequences. The model predicts the two phonological trends closely: for example, while *k* palatalizes in my data nearly categorically in frequent forms, the model predicts that it should palatalize at roughly a 94% rate; moreover, *g* palatalizes 52% of the time in frequent forms, while the model predicts a 51% rate. The model predicts distinctions in morphemic gradience as well, as the table below reveals—observed suffix rates below are averages over the rates across stem types. In all, the mixed logit is promising for modeling lexical trends together with lexical idiosyncrasy.

Suffix:	-ovje	-ʦ	-nat	-itsa	-ina	-itʃ	-je	-n	-k
Obs. rate:	18%	41%	42%	70%	71%	78%	88%	94%	96%
Pred. rate:	1%	7%	30%	63%	72%	86%	97%	97%	80%