FREQUENCY-DEPENDENT REGULARIZATION IN ITERATED LEARNING

ANONYMOUS AUTHOR 1
University Department, University Name
City, Country
email@university

Binomial expressions are more *regularized*—their ordering preferences (e.g. "bread and butter" vs. "butter and bread") are more extreme—the higher their frequency. Although standard iterated-learning models of language evolution can encode overall regularization biases, the stationary distributions in these standard models do not exhibit a relationship between expression frequency and regularization. Here we show that introducing a frequency-*independent* regularization bias into the data-generation stage of a 2-Alternative Iterated Learning Model yields frequency-*dependent* regularization in the stationary distribution. We also show that this model accounts for the distribution of binomial ordering preferences seen in corpus data.

1. Introduction

Languages are shaped both by the cognitive architectures of individual speakers and by the process of cultural transmission that acts across generations. In this paper we ask how these two factors jointly contribute to a key dichotomy in language structure: the trade-off between broadly-applicable compositional knowledge and knowledge of item-specific idiosyncrasies. Specifically, we take up the case of frequency dependence in *regularization*—the extremity of a preference for a given form among multiple alternatives. Although regularization is a well-attested phenomenon in statistical learning, *frequency-dependent* regularization is not. Here we demonstrate that frequency dependence of regularization can arise as an emergent property of a frequency-*independent* regularization bias in language production, combined with the bottleneck effect of cultural transmission.

Item-specific idiosyncrasies (i.e exceptions to the rules) are well known to be frequency-dependent. For example, more frequent verbs are more likely to have irregular conjugations (?). More recently, ? (?) have demonstrated a different type of frequency-dependent idiosyncrasy at the level of multi-word phrases, specifically *binomial expressions* of the form "X and Y" (?, ?). Word order preferences for these expressions are gradient; for example, "radio and television" is preferred to "television and radio" in a 63 to 37 ratio, while "bread and butter" is preferred to "butter and bread" 99 to 1 (?). These ordering preferences are partially determined

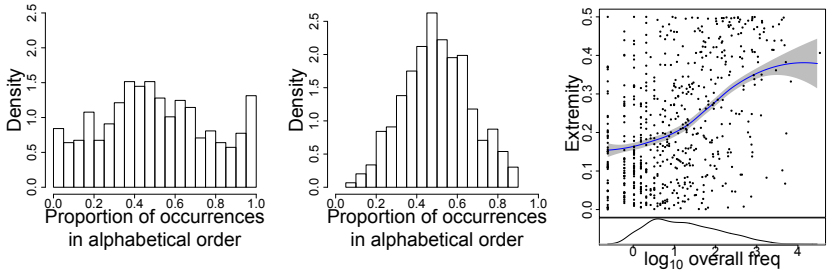
by productive, violable constraints, e.g. a constraint to put shorter words before longer words. But these expressions are also subject to learned item-specific idiosyncrasies, e.g. despite a generally strong constraint to put men before women, "ladies and gentlemen" is preferred over "gentlemen and ladies". In addition to the possibility of the complete reversal of compositional preferences, item-specific idiosyncrasies can also be gradient, e.g. a binomial whose compositional preference predicts a 60/40 distribution might instead be used in a 90/10 ratio. ? (?) showed that, as is the case with irregular verbs, the distribution of idiosyncrasies in binomial ordering preference is frequency-dependent: more frequent binomial expressions deviate more from compositional preferences. In particular, more frequent binomials are more strongly regularized.

Regularization is a well-established phenomenon in statistical learning. In a variety of tasks, both linguistic and non-linguistic, in which participants learn and reproduce probability distributions over alternates, both children and adults tend to regularize their productions (?, ?, ?). For example, ? (?) found that when exposed to two labels for a novel object, subjects reproduced the more frequent label *even more frequently* than that label was seen in training. Although this tendency was weak, they demonstrated that even such a small bias towards regularization can have significant long-term impacts, as the bias acts across successive generations to shape language over time. ? (?), ? (?), and others have argued that children's tendency to regularize is an important mechanism of language change, e.g. for forming more consistent languages out of pidgins.

However, standard iterated-learning theories of language evolution do not, in general, lead to frequency-dependent regularization. Thus Morgan and Levy's finding is unexpected, and poses a challenge to models of language evolution. In this paper, we review the key data (Section **??**) and show that standard iterated-learning models fail to account for frequency-dependent regularization (Section **??**). We then show that frequency-dependent regularization emerges when the data-generation stage of a standard iterated learning model is augmented with a frequency-independent regularization bias, and that this augmented model accounts for the empirical distribution of binomial ordering preferences (Section **??**). Section **??** concludes.

2. Dataset

We take advantage of a uniquely appropriate real-world data set: ? (?)'s corpus of 594 binomial expression types hand-annotated for a range of semantic, phonological, and lexical constraints known to affect binomial ordering preferences, and with frequencies of each ordering extracted from the Google Books corpus (?). Morgan and Levy also reported a model estimating the quantitative compositional ordering preference for each binomial expression, as expected on the basis of the above constraints (independent of actual occurrence frequencies). The dataset and model thus give us three key measures for these expressions:



(a) (b) (c)

Figure 1. Results from Morgan and Levy (2015). (a) Histogram of binomial types' observed preferences. (b) Histogram of binomial types' compositional preferences. (c) We define an expression's *extremity* as the absolute difference between its observed preference and 0.5. More frequent expressions have more extreme/regularized preferences; see Morgan & Levy (2015) for alternative ways to quantify extremity that yield similar conclusions. Lower panel shows density of overall frequency counts (scaled as described in Section **??**). The distribution is non-Zipfian because the corpus is restricted to binomial types with at least 1000 occurrences in the Google Books corpus to ensure accurate observed preference estimates.

- The *overall (unordered) frequency* of an expression: freq("X and Y")+freq("Y and X")
- The *observed preference* for occurrence in a given order, expressed as a number between 0 and 1: freq("X and Y")/(freq("X and Y")+freq("Y and X"))
- The *compositional preference* for occurrence in a given order, expressed as a number between 0 and 1, given by Morgan and Levy's model.

Observed preferences are multimodally distributed, with modes at the extremes as well as around 0.5 (Fig. **??**). Crucially, this pattern is not predicted by compositional preferences, which predict only a single mode (Fig. **??**). This pattern reflects the key generalization to be accounted for in the present paper: that expressions with higher overall frequency diverge most from compositional preferences, and are more regularized (Fig. **??**).

3. Regularization is Frequency-Independent in Standard Iterated Learning

We use 2-alternative iterated learning (?, ?) to simulate the evolution of binomial expressions over generations of speakers. A learner hears N tokens of a binomial expression, with x_1 of them in a given order—we use alphabetical order as a neutral reference order—and then infers a hypothesis $\theta_1 \in [0, 1]$ which is the proportion of time a binomial should be produced in alphabetical order. The learner then generates new data using θ_1.

The prior probability $P(\theta_1)$ of a binomial being preferred in a given order can

be expressed using the beta distribution. We can treat the compositional preference as a form of prior knowledge of ordering preferences for a binomial. To incorporate this prior knowledge, we use a parameterization of the beta distribution with a parameter μ that determines the mean of draws and a concentration parameter ν that determines how tightly clustered around the mean those draws are. (ν can also be thought of as reflecting how confident in the prior we are, e.g. $\nu = 10$ would indicate confidence equivalent to having seen ten instances of a given binomial expression type before.) Under this parameterization,

$$P(\theta_1) = \frac{\theta_1^{\mu\nu-1}(1-\theta_1)^{(1-\mu)\nu-1}}{\mathrm{B}(\mu\nu,(1-\mu)\nu)} \tag{1}$$

where B is the beta function. Because μ represents compositional ordering preferences, it varies for each binomial, and is set according to Morgan and Levy's model. All learners are assumed to have the same μ value for a given binomial. ν is constant for all binomial expressions for all learners, and is a free parameter. Given θ_1, data is generated binomially:

$$P(x_1|\theta_1) = \binom{N}{x_1}\theta_1^{x_1}(1-\theta_1)^{N-x_1} \tag{2}$$

We define a chain of learners under this model by initializing a single learner with some hypothesis. This first generation produces N utterances according to the distribution defined in Eq. **??**. The learner in the next generation applies Bayes rule and chooses a hypothesis from the resulting posterior distribution over hypotheses. This process continues iteratively.

? (?) have demonstrated that regularization occurs in iterated learning models with sparse priors (i.e. those that favor hypothesis close to 0 and 1); given our parameterization of the beta distribution, these are hypothesis with $\nu < 2$. However, this regularization is not dependent on the expression's overall frequency: the stationary distribution is . We demonstrate this by modeling chains of learners with different values of N. We model a single binomial expression with prior probability $\mu = 0.6$. We explore different values of ν, specifically $\nu = 1$ (a sparse prior) and $\nu = 10$ (a dense prior), and values of $N = 10, 100, 200, 500$. For each combination of ν and N, we approximate the distribution over expression preferences by running 500 chains of learners for 500 generations each and taking the hypothesis of the final generation in each chain. (For all chains in all simulations in this paper, we initialize $\theta_1 = 0.5$ and use MAP estimation to choose θ_1 in each new generation. Results are qualitatively similar under posterior sampling.) Regularization in the resulting distributions does not depend on N (Fig. **??**, dashed lines). The apparent sensitivity to N for a given value of ν, most clearly in the $N = 500, \nu = 1$ case, is due to the finite number of chains used in the simulations, and because convergence to the stationary distribution is slower for higher

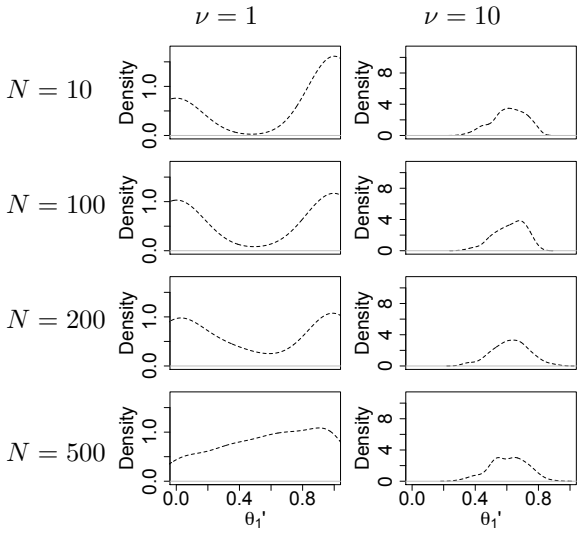


Figure 2. Simulated distribution of binomial ordering preferences for a single expression type with $\mu = 0.6$ in a standard 2-Alternative Iterated Learning Model (dotted lines) and one with an explicit regularization bias in data production of $\alpha = 1.1$ (solid lines). Note that $\theta_1' = \theta_1$ in the standard model. Regularization depends upon N only in the model with an explicit regularization bias.

values of N. The number of times an expression is seen in each generation does not affect its ultimate degree of regularization.

4. Emergence of Frequency-Dependent Regularization in Iterated Learning

The standard 2-Alternative Iterated Learning Model does not predict frequency-dependent regularization. We now demonstrate that we can predict frequency-dependent regularization by introducing a frequency-*independent* regularization bias into our model. Under this model, frequency-dependent regularization is an emergent property of the interaction of the frequency-independent regularization bias with the bottleneck effect of cultural transmission.

We augment the learning and transmission process as follows. After hearing data, the learner chooses a hypothesis θ_1 as before, then applies a regularization function to produce a new hypothesis θ_1', then generates data from θ_1'.

The regularization function is the regularized incomplete beta function (equivalently, the cumulative distribution function of the beta distribution), restricted to be symmetric such that it has a single free parameter α:

$$f(x;\alpha) = \frac{\int_0^x t^{\alpha-1}(1-t)^{\alpha-1}\,\mathrm{d}t}{\mathrm{B}(\alpha,\alpha)} \tag{3}$$

As shown in Fig. **??**, the bias parameter α controls strength of regularization. When $\alpha = 1$, this is the identity function, i.e. no explicit regularization is added. As α increases, the regularization bias grows stronger.

4.1. *Results: Frequency-dependent regularization*

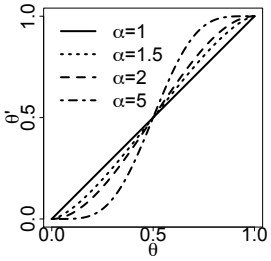


Figure 3. Regularization function with different values of α

When we repeat the simulations from above using a non-trivial regularization bias $\alpha = 1.1$, we see frequency-dependent regularization in the case with a dense prior (Fig. **??**). Although the regularization bias itself is frequency-independent, frequency-dependence emerges from the interaction of the regularization bias with the process of cultural transmission: At lower frequencies, there is not sufficient data for the regularization bias to overcome the prior. At higher frequencies, the regularization bias becomes increasingly dominant as there is increasingly enough data for the effects of this bias to be carried across generations. Even a relatively weak bias ($\alpha = 1.1$) can produce noticeable regularization when compounded across generations. However, the prior always continues to exert some influence; thus, even the highest frequency expressions do not become completely regularized.

Another linguistically accurate property of this model is that for sufficiently high values of N, the distribution over hypotheses includes a mode on the opposite side of 0.5 from the prior. Thus the model correctly predicts that at high enough frequencies, an expression can become idiosyncratically preferred in the opposite of its compositionally predicted direction (as in "ladies and gentlemen").

4.2. *Results: Simulating corpus data*

Having demonstrated that our augmented model produces frequency-dependent regularization, we now show that it additionally predicts the true language-wide distribution of binomial preference strengths seen in corpus data. The target distribution to be accounted for is shown in Fig. **??**.

We take the true properties of each binomial expression in the corpus: its compositional preference determines μ and its overall frequency determines N. We scale overall frequency counts based on estimated lifetime exposure to 300 million total words (?, footnote 10). The resulting distribution of values N is shown in Fig. **??**. For each binomial in the corpus, we approximate the stationary distribution by modeling 10 chains of learners for 200 generations each and take the hypothesis θ'_1 of the final generation of each chain.

Our model has two free parameters, ν and α. We model the corpus data as described above for a range of values of both of these parameters. As shown in Fig. **??**, our model displays a trade-off between the prior and the regularization

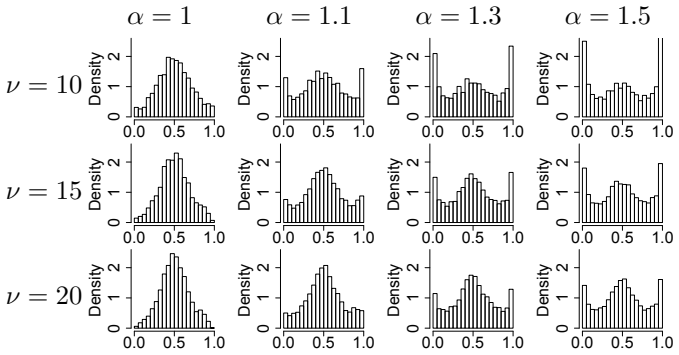


Figure 4. Predicted distribution of θ_1'. We see a trade-off between effects of the prior and the regularization bias. When the prior is stronger (high ν, low α), we see a unimodal distribution of preferences, similar to Fig. **??**. When the regularization bias is stronger (low ν, high α), we see too much regularization. At appropriate values of α and ν, we see the correct multimodal distribution of preferences as seen in corpus data (Fig. **??**).

bias as a function of these parameters. At appropriate values, our model correctly predicts the multimodal distribution of corpus data as seen in Fig. **??**.

5. Conclusion

We have demonstrated that a frequency-independent regularization bias in data generation, combined with cultural transmission, can produce the pattern of frequency-dependent regularization of binomial ordering preferences seen in corpus data. Cultural transmission creates frequency-dependence by introducing a bottleneck effect that favors prior knowledge at lower frequencies while allowing the regularization bias to be increasingly well transmitted at higher frequencies. This finding sheds light on the origins of linguistic structure in two important ways: one, it confirms earlier demonstrations of a bias to regularize when learning stochastic linguistic items. Second, it shows that this bias can apply equally across all levels of frequency, but that the distribution of idiosyncrasy seen in the language emerges from the interaction of individuals' cognitive biases with the bottleneck effect of cultural transmission. Additionally, we have expanded the empirical coverage of iterated learning models, showing that they can account not only for qualitative generalizations in natural language and data from laboratory experiments, but also detailed patterns of naturalistic corpus data. As we hope to have shown, binomial ordering preferences are a particularly suitable test case for iterated learning models, at once theoretically interesting, data-rich, and computationally tractable.

Acknowledgements

[Acknowledgments withheld in submitted version to maintain author anonymity during review]