CHAPTER 2

Constraint interaction in maximum entropy grammars

2.1 Introduction

Speakers of any language have systematic and impressively detailed intuitions of the sounds and sound patterns that are probable and improbable in their language. Such intuitions have substantially contributed to developments in phonological theory, offering new insights into the analytic, substantive, and learning biases that shape the distribution of attested linguistic structures across the world's languages (Steriade, 2001; Wilson, 2006; Moreton, 2008; Pater and Moreton, 2012; White, 2014, among others). But, only relatively recently have we started to understand the effects that the multiple presence of phonotactically illegal structures have on the distribution of words in a language and speaker's intuitions about these. Recent studies find that forms with two or more marked sound patterns are less frequent, repaired more often, and receive lower ratings (and are therefore more marked), than forms with one marked structure, suggesting that constraint violations accumulate and interact in certain ways (Albright, 2008, 2012; Yang et al., 2018; Shih, 2016, 2017). And, other studies find that speakers generalize cumulative effects from distributional phonotactic patterns in their language (Pizzo, 2015; Smith and Pater, 2020; Breiss, 2020; Kawahara, 2020, 2021), and that these cumulative effects across different sound patterns are, at least in part, learnable from artificial languages in the lab (Breiss and Albright, 2022).

However, current phonological theories within Optimality Theory (Prince and Smolensky,

1993) make diverging predictions about how constraint violations interact. The central example of this is in the comparison of constraint-based theories of phonology with a strict ranking among constraints against those with weighted constraints (Classic OT vs. Standard Harmonic Grammar; Legendre et al. (1990)). Namely, Classic OT predicts that forms with one violation of a constraint are no different than forms with two or more constraint violations as long as there is a higher ranked constraint that distriguishes among two candidates. However, In Standard HG, multiple constraint violations may have additive effects on the grammar's predictions since these grammars consider all constraint violations, no matter how low-weighted, in the evaluation of the outcome.

This chapter evaluates the nature of constraint interaction in Maximum Entropy (MaxEnt) models of phonological grammar (Goldwater and Johnson, 2003; Hayes and Wilson, 2008). MaxEnt grammars, being probabilistic extensions of Standard HG, already predict that multiple violations have cumulative effects: these grammars predict that a form with more violations is somehow more marked (rather, has a lower probability) than a similar form with less violations. However, this chapter finds that different constraint violations interact in different ways across different kinds of MaxEnt model structures, with this being a result of the different ways in which these grammars organize inputs and candidates and therefore assign probability distributions. This chapter focuses on the comparison of two MaxEnt grammar structures widely used in modeling phonotactic and alternation patterns with and without cumulative constraint interactions. The first grammar structure assigns a single probability distribution over all surface forms evaluated in the grammar. This is grammar structure, which I call *single competition* models in this work, is typically assumed in modeling phonotactics and phonotactic learning given their success at capturing gradient and probabilistic phonotactic patterns from the surface distributions of forms in the data (Hayes and Wilson, 2008). The second grammar structure assigns multiple probability distributions among two candidates, such that candidates that are otherwise in the same probability distribution in single competition grammars are now in separate probability distributions.

This grammar structure, which I call *multiple competitions* models in this work, is often used in modeling patterns as a binary decision, such as the rate at which certain processes apply or don't apply (Shih, 2016, 2017; Zuraw and Hayes, 2017; Kawahara, 2020; Smith and Pater, 2020; Kim, 2022), or rates at which a given input is accepted or rejected (Breiss and Albright, 2022; Hayes, 2022). Crucially, this chapter finds that constraint violations in single competition models only have an *increasingly weaker* effect on a form's markedness as violations accumulate, while in multiple competitions models constraint violations may have a *stronger or weaker* effect on a form's markedness along with other constraint violations even if the independent effects of these violations are rather weak.

The findings have direct consequences for our theories of phonotactics and phonotactic learning. First, these grammar structures predict vastly different typologies of constraint interaction. Previous literature finds stronger effects of constraint violations in the environment of other violations versus when they are violated independently in various patterns ("superlinear" or "superadditive" effects; Green and Davis, 2014; Shih, 2016, 2017; Smith and Pater, 2020; Kim, 2022; Milenković, to appear), and a few of these cases include purely phonotactic patterns (Albright, 2008, 2012; Yang et al., 2018; Kawahara, 2020; Breiss and Albright, 2022). Therefore, single competition probabilistic models fail to predict these stronger cumulative effects without some additional mechanisms that directly penalize forms with two or more violations. And, importantly, these models structures also make predictions about the mechanisms by which speakers learn and generalize cumulative phonotactic effects. Single competition models, those akin to the Hayes and Wilson (2008) learner, predict that cumulative effects are not learnable from data – the structure of the grammar already predicts the nature of constraint interaction regardless of how singly- and multiply-violating forms distribute relative to each other in the language. Therefore, their intuitions of simplex phonotactic patterns are derived from their distribution in the data, but their intuitions of interacting constraint violations are derived a priori from the universal combinatorial properties that underlie the phonological grammar. On the other hand, multiple competitions model structures predict that speakers fully build intuitions about cumulative and non-cumulative effects from the statistical properties in their data – they are able to learn and generalize patterns in which constraint violations have a stronger or weaker effect on a form's markedness compared to its independent effect, with this stronger or weaker effect being a piece of phonological information the speakers must learn from their language.

I begin this chapter by defining the two MaxEnt model structures I evaluate in the rest of the dissertation. Section 2.3 surveys how previous literature evaluates constraint interaction in empirical patterns and in the predictions of phonological grammars. I argue in this section that the standard evaluation metric used in prevous literature to assess constraint interaction, namely the *linearity* of cumulativity, faces a few critical problems given its reliance on the notion of an "expected" penalty or probability. As a result, I define the evaluation metric I will be using in this chapter and in the result of the dissertation to evaluate constraint interaction, which I call the *concavity* of cumulativity. Sections 2.4 and 2.5 lay out the predictions of single and multiple competitions MaxEnt models, respectively, using this proposed metric of constraint interaction to show that they make different predictions regarding the effects accumulating violations have on their predicted probability distribution(s). I show their predicted cumulative effects for both interacting violations of the same constraint (called "counting" cumulativity; (Jager and Rosenbach, 2006)) and interacting violations of different constraints (called "ganging" cumulativity). This chapter concludes by briefly discussing the implications of these findings for phonotactics and phonotactic learning, ultimately motivating the behavioral experiments and computational modeling of subsequent chapters.

2.2 MaxEnt grammars: single vs. multiple competitions

Maximum Entropy (MaxEnt) grammars are stochastic log-linear models based on Standard Harmonic Grammar (Legendre et al., 1990) used in modeling and learning phonological grammars. Starting with Goldwater and Johnson (2003) with alternation patterns, they have been extended to the modeling and learning of surface phonotactic patterns (Hayes and Wilson, 2008). MaxEnt grammars are particularly successful at modeling gradient and probabilistic phonological patterns because they generate probability distributions over all output candidates of an input. This is in opposition to generating categorical outcomes, as is the case with Classic OT (Prince and Smolensky, 1993) and Standard HG. MaxEnt models additionally allow for mathematically explicit comparisons of phonological theories and they are closely related to general statistical models of data analysis (Jurafsky and Martin, 2019).

MaxEnt grammars assign numerical weights to phonological constraints, and the probability of each candidate output is a function of these constraint weights. For any given input x, MaxEnt assigns a probability to each potential output candidate y of input x according to Equations 2.1 and 2.2 below.

$$P(x \mid y) = \frac{exp(-\sum_{i=1}^{m} w_i C_i(y,x))}{Z} \tag{2.1}$$

$$where\ Z = \sum_{y \in Y(x)} exp(-\sum_{i=1}^{m} w_i C_i(y,x)) \tag{2.2}$$

As with Standard HG, each candidate's *harmony value* in MaxEnt is calculated by multiplying the weight w_i of each constraint C_i by the number of times a given input-output pair violates the constraint, then summing over all m constraints in the grammar. Then, each candidate's *MaxEnt value*, the numerator in Equation 2.1 above, is the base of the natural logarithm e raised to the negative harmony value. Finally, the probability of each candidate is calculated by dividing the MaxEnt value by the sum of the MaxEnt values of all output candidates for a given input y. So, the difference between MaxEnt and Standard HG is in the nature of EVAL, namely the component of optimality-theoretic grammars that evaluates a grammar's outcomes given all candidates' violation profiles. While Standard HG grammars select a single winning candidate with the highest harmony, MaxEnt grammars generate a probability distribution over all possible output candidates for each input, allowing for probability mass to be unevenly distributed across the different candidates as a function of their constraint violations.

The weights w_i for all constraints in the grammar are induced via a learning algorithm that provably converges on the "best" grammar, which is the grammar that maximizes the (log) likelihood of the observed data and in turn minimizes the probability of the unobserved data. Importantly, this calculation is sensitive to the observed frequency of each candidate, which can be either type or token frequency, and to the number of candidates outputs for each input (see Hayes and Wilson (2008) for more details). So, the goal of the learner is to induce constraint weights for all constraints in the grammar such that the probability of the data is maximized.

$$P(D) = \prod_{i=1}^{n} Pr(y_i \mid x_i) \tag{2.3}$$

This objective function also often includes a prior term that is subtracted from the (log) probability of the data. This prior is a Gaussian distribution with free parameters for mean μ and standard deviation σ, and each constraint weight w_i is associated with its own prior term.

$$\sum_{i=1}^{m} \frac{(w_i - \mu_i)^2}{2\sigma_i^2} \tag{2.4}$$

These priors bias the model's learning of constraint weights in particular ways. The μ_i value for each constraint is its preferred weight a priori. Since μ_i is subtracted from the learned constraint weight w_i and this difference is then squared, the penalty imposed by the prior increases as the constraints weights deviate from their a priori μ value. The σ_i^2 term determines the degree to which deviations from a constraint's μ value is penalized. Being in the denominator, a high value of σ_i^2 decreases the value of the prior, hence allowing for more freedom to deviate from μ_i. Finally, since the sum of the constraint priors is subtracted from the log probability of the data, the penalty imposed on the model increases the more the learned constraint weights diverge from their a priori μ values.

Although the constraint-based analysis of alternation patterns involves the use of inputs and constraints that regulate input-output mappings (i.e. OT-style faithfulness constraints), no role is played by inputs or faithfulness constraints in the modeling phonotactic patterns in MaxEnt (Hayes and Wilson, 2008). Instead, the typical model structure for phonotactic modeling in MaxEnt is one in which all observed forms are in a single (and possibly giant) competition, meaning that all candidates' exponentiated harmonies are normalized by the same constant Z. Recall that this normalizing constant is the sum of all exponentiated harmonies of all candidate forms in the grammar, therefore all predicted probabilities sum to one across all output forms. For clarity, I'll call these kinds of grammar structures **single competition** MaxEnt models since constraint weights are learned from a single collection of outputs that form parts of the same probability distribution or competition. The single competition MaxEnt tableau in Figure 2.1 below shows how we might model the gradient pattern of nasal-place assimilation in English. In English and other languages, nasal consonants must agree in place of articulation with the following obstruent, but it might be favorable to model such pattern with a probabilistic grammar such as MaxEnt since the English lexicon does have some exceptions to the pattern (such as "input" and "Camden").

Figure 2.1: Example single competitions grammar in MaxEnt modeling nasal place assimilation in English.

	AGREE[PLACE] w_m	H	e^{-H}	predicted prob.
hæ**mp**ɹ̩	0	0	1	1 / **Z**
ɪ**ŋg**lɪʃ	0	0	1	1 / **Z**
...	...	...	...	...
ɪ**mp**ʌt	1	$-w_m$	e^{-w_m}	e^{-w_m} / **Z**
kæ**md**n̩	1	$-w_m$	e^{-w_m}	e^{-w_m} / **Z**

However, since input-output mappings and OT-style faithfulness constraints do play a role in alternation patterns, MaxEnt grammars modeling alternation patterns have a different structure: a particular input lexical item has multiple candidates with different violation profiles for markedness and faithfulness constraints. A MaxEnt grammar modeling alternation patterns assigns separate probability distributions over all the output candidates for a given input, therefore multiple distributions are assigned in the grammar and candidates in different competitions may be normalized by different Z constants. Going forward, I'll call these kinds of grammar structures **multiple competitions** models.

The tableau in Figure 2.2 below shows an example grammar modeling the English plural suffix alternation pattern. In English, the plural suffix assimilates in voicing to its preceding non-sibilant obstruent, therefore alternating from underlying /-z/ to surface [-s] when attached to roots ending in a voiceless non-sibilant obstruent such as “cat”, and remaining as [-z] when attached to roots ending in a voiced non-sibilant obstruent such as “dog”. This simplified grammar only assumes two candidates for each input, but in principle we would compare more candidates in each competition and more accurately predict their distribution with additional markedness and faithfulness constraints.

Figure 2.2: Example alternation grammar in multiple competitions MaxEnt modeling the English plural suffix alternation.

inputs	cands.	Ident[voi] w_{id}	*[αvoi][$-\alpha$voi] w_{voi}	e^{-H}	predicted probability
/dɔg + z/	dɔgz			1	$1\ /\ (1 + e^{-w_{id}})$
	dɔgs	1		$e^{-w_{id}}$	$e^{-w_{id}}\ /\ (1 + e^{-w_{id}})$
/kæt + z/	kætz		1	$e^{-w_{voi}}$	$e^{-w_{voi}}\ /\ (e^{-w_{voi}} + e^{-w_{id}})$
	kæts	1		$e^{-w_{id}}$	$e^{-w_{id}}\ /\ (e^{-w_{voi}} + e^{-w_{id}})$

This dissertation investigates a more specific multiple competitions model structure that is often used in modeling patterns with cumulative constraint interactions (Shih, 2016, 2017; Zuraw and Hayes, 2017; Smith and Pater, 2020; Kim, 2022; Kawahara, 2020, 2021; Breiss and Albright, 2022; Hayes, 2022; Milenković, to appear). The model structure is one wherein each competition has only two candidates and the candidates are in an *asymmetric trade-off* (Pater, 2009). An asymmetric trade-off occurs when a candidate with zero or more violations of a given constraint (usually the input candidate) competes against a candidate with a single violation of an opposing constraint, as shown in the schematic tableau in Figure 2.3 below. For each input c_n in this grammar, n violations of Variable "trade off" against a single violation of OnOff.

Figure 2.3: Schematic tableau with an asymmetric trade-off between Variable and OnOff. Constraint names adapted from Hayes (2022).

		OnOff w_{oo}	Variable w_v	H	predicted probability
c_0	a. c_0		0	0	$1 \ / \ e^{-w_{oo}}$
	b. c_{oo}	1		$-w_{oo}$	
c_1	a. c_1		1	$-w_v$	$e^{-w_v} \ / \ (e^{-w_v} + e^{-w_{oo}})$
	b. c_{oo}	1		$-w_{oo}$	
c_2	a. c_2		2	$-2w_v$	$e^{-2w_v} \ / \ (e^{-2w_v} + e^{-w_{oo}})$
	b. c_{oo}	1		$-w_{oo}$	
...	...	...	...	...	...

Asymmetric trade-offs crucially predict categorical *gang effects* in Standard HG (Jager and Rosenbach, 2006; Pater, 2009; Kaplan, 2018). These grammars predict categorical gang effects when a constraint is satisfied at the cost of n violations of some lower weighted contraint(s) but not at the cost of $n+1$ violations. In other words, a categorical shift in winner in an asymmetric trade-off occurs between a candidate with n violations of $\mathbb{C}2$ and a different candidate with $n+1$ violations of $\mathbb{C}2$ when $w_{\mathbb{C}1} > nw_{\mathbb{C}2}$ but $w_{\mathbb{C}1} < (n+1)w_{\mathbb{C}2}$. The Standard HG tableaus in Figure 2.4 below show a categorical gang effect occurring between the second and third violations of Variable: at the given constraint weights, a candidate with two violations of Variable wins against the opposing candidate, but an additional violation of Variable makes such candidate the loser against the opposing candidate. Note that these grammar structures still involve an asymmetric trade-off: n violations of Variable trade off against a single violation of OnOff in both tableaus.[1] This equivalently holds for multiple

[1] Asymmetric trade-offs are in opposition to a *symmetric trade-offs*, which occur when variable violations of one constraint are avoided by violating the opposing constraint by the same amount. Symmetric trade-offs fail to predict categorical gang effects in Standard HG since every violation of a lower-weighted constraint will

single violations of different Variable constraints as long as the same weighing conditions hold across n and $n+1$ violations of different constraints.

Figure 2.4: Schematic Standard HG tableaus showing a categorical gang effect. OnOff is satisfied at the cost of 3 violations of Variable: candidate c_n wins at 2 but not 3 violations of Variable.

	OnOff $w_{oo} = 5$	Variable $w_v = 2$	H
☞ a. c_2		2	-4
b. c_{oo}	1		-5

	OnOff $w_{oo} = 5$	Variable $w_v = 2$	H
a. c_3		3	-6
☞ b. c_{oo}	1		-5

Note that this condition of an asymmetric trade-off is not required to predict categorical gang effects in Standard HG. For example, these may still be predicted with a conjoined constraint that is violated when its conjunct constraints are all violated by the same candidate (Smolensky, 2003, 2006; Itô and Mester, 1998, 2003; Łubowicz, 2005; Crowhurst, 2011; Shih, 2016, 2017). Figure 2.5 below shows the effect of constriant conjunction at predicting categorical gang effects even when the sum of the weights of the lower-ranked constraints $\mathbb{C}2$ and $\mathbb{C}3$ do not surpass the weight of OnOff. Conjoined constraints are also used in predicting categorical gang effects in strict-ranking Classic OT: in these grammars, multiple violations of lower-ranked constraints have no effect on the outcome of the grammar if there is some higher-ranked constraint violated by an opposing candidate, but they have an effect with a violation of a conjoined constraint that is ranked higher than the opposing constraint.

incur the same number of violations of the higher weighted constraint (Pater, 2009; Milenković, to appear), unless some additional mechanisms such as constraint conjunction directly increase the penalty associated with one of the variably-violating forms but not the other. Therefore, symmetric trade-offs in Standard HG are equivalent to strict domination in Classic OT grammars.

Figure 2.5: Conjoined constraints predict categorical gang effects in Standard HG as long as $w_{\mathbb{C}1} < w_{oo}$, $w_{\mathbb{C}2} < w_{oo}$, and $w_{[\mathbb{C}1\&\mathbb{C}2]} + w_{\mathbb{C}1} + w_{\mathbb{C}2} > w_{oo}$, assuming single violations for all constraints in the grammar.

	$[\mathbb{C}1$ & $\mathbb{C}2]$ $w = 1.5$	OnOff $w = 5$	$\mathbb{C}1$ $w = 2$	$\mathbb{C}2$ $w = 2$	H
☞ a. c_1			1		-2
b. c_{ooo}		1			-5
a. c_2	1		1	1	-5.5
☞ b. c_{oo}		1			-5

Still, this multiple competitions Maxent grammar structure with the asymmetric trade-off is crucial because it has been used to model cumulative *phonotactic* effects (Kawahara, 2020, 2021; Breiss and Albright, 2022), in addition to its use in modeling cumulative and non-cumulative alternation patterns.[2] Instead of all structural candidates belonging to a single probability distribution as in single competition models, in multiple competitions phonotactic models probability distributions are assigned only between a structural candidate and its opposing candidate, represented as $\odot$, such that structural candidates are now in separate competitions and probability distributions. Figure 2.6 below shows how the phonotactic pattern of English nasal place assimilation, previously modeled under a single competition grammar structure in Figure 2.1, is modeled under a multiple competitions model. Note that this grammar maintains an asymmetric trade-off: candidates with variable violations of Agree[pl] (Variable) compete against the opposing candidate $\odot$ in each competition, and any $\odot$ candidate violates MParse (OnOff) only once.

[2] By "cumulative alternation patterns", I refer to patterns wherein markedness constraints accumulate against a faithfulness constraint, as well as patterns where violations of faithfulness constraints may also accumulate (Farris-Trimble, 2008, 2010). And, by "cumulative phonotactic effects" I refer to patterns where accumulating violations of markedness affects a form's frequency and acceptability.

Figure 2.6: English nasal place assimilation modeled under multiple competitions MaxEnt with an asymmetric trade-off.

inputs	cands.	MParse w_{mp}	Agree[pl] w_a	e^{-H}	predicted probability
hæ**mp**ɹ̩	hæ**mp**ɹ̩		0	1	$1 \;/\; 1 + e^{-w_{mp}}$
	⊙	1		$e^{-w_{mp}}$	
ɪ**np**ʌt	ɪ**np**ʌt		1	e^{-w_a}	$e^{-w_a} \;/\; (e^{-w_a} + e^{-w_{mp}})$
	⊙	1		$e^{-w_{mp}}$	
kæ**md**n̩	kæ**md**n̩		1	e^{-w_a}	$e^{-w_a} \;/\; (e^{-w_a} + e^{-w_{mp}})$
	⊙	1		$e^{-w_{mp}}$	

The ⊙ candidate can be thought of as the null candidate in the grammar, namely, a candidate that has no phonetic realization. Following Prince and Smolensky (1993), this candidate, being phonologically null, satisfies all markedness and faithfulness constraints and only violates MParse, which no other candidate violates. So, we can think of multiple competitions models as modeling phonotactics as a binary choice between producing and not producing a word at certain rates, or as the probabilistic choice between accepting or rejecting an input surface form, therefore parallel to modeling alternation patterns as the rate at which a certain process applies or fails to apply.[3]

In addition to their use in analyzing cumulative phonotactic effects and alternation patterns in general, multiple competitions models with asymmetric trade-offs are the critical grammar structures that derive MaxEnt's "wug-shaped" curves, argued to be the "quantitative signature" of this grammar formalism (Zuraw and Hayes, 2017; Kawahara, 2020; Hayes, 2022). Wug-shaped curves are two or more parallel sigmoidal (logistic) curves that

[3] The null candidate ⊙ and MParse are also used in analyzing paradigm gaps, which have been modeled as cases wherein the phonological grammar has no structural output for a particular morphological or lexical item (McCarthy and Wolf, 2005).

map probability against some measure of markedness (such as baseline harmony, number of violations of VARIABLE, among others) and are argued to be present in a variety of empirical domains ranging from categorical speech perception to probabilistic alternation patterns.[4]

In summary, this chapter compares the predictions of single competition and multiple competitions Maximum Entropy models of phonology and phonological learning. Single competitions grammars are akin to the Hayes and Wilson (2008) phonotactic learner, in which a phonotactic grammar is learned from a single probability distribution over all observable surface forms in the data. On the other hand, multiple competitions MaxEnt grammar model phonological patterns as the binary decision between two candidates, therefore the model assigns multiple probability distributions, one for each input. Multiple competitions models, although more straightforwardly applied to alternation patterns, have also been applied to phonotactics, where the surface input form is evaluated against a null candidate, therefore modeling phonotactics as the binary choice between producing or not producing, or accepting or rejecting, each input form. Multiple competitions grammars are especially well-suited for modeling various phonotactic and alternation patterns with and without cumulative constraint interactions, and they even reveal properties of MaxEnt grammars that seem fundamental to capturing a variety of empirical patterns in different domains.

2.3 Evaluating constraint interaction

Before turning to the different predictions of single competition and multiple competitions MaxEnt grammars regarding constraint cumulativity, it's important to define how constraint interaction is and will be evaluated in this dissertation both across model predictions and in empirical data. Previous literature on cumulativity uses to notion of **linearity** or **additivity** in assessing constraint interaction, which generally involves a comparison of observed and

[4] See Bruce Hayes's Gallery of Wug-Shaped Curves: https://brucehayes.org/GalleryOfWugShapedCurves/index.htm.

"expected" penalties associated with different degrees of violations. Namely, *sublinear* (or *subadditive*) patterns of cumulativity are those wherein the observed penalty is higher than expected, and *superlinear* (or *superadditive*) patterns are those wherein the observed penalty is lower than expected. Previous works use the linearity of cumulativity to characterize constraint interaction as evidenced in a variety of empirical patterns – such as the distribution of words in the lexicon of a language, in the rates of repair of marked structures in a language's lexicon or in experimental tasks, in acceptability judgments, etc. – and in the predictions of grammar models.

For example, Albright (2008) finds superlinear cumulative interactions in the Lakhota lexicon among fricatives, ejectives, aspirated consonants, and consonant clusters, which are all semi-marked structures in the language. He finds that the observed frequency of words in the Lakhota lexicon with two of these marked structures is lower than the frequency that is "expected" given the independent combination of the frequencies of words with only one of these structures. For example, 32% of bisyllabic CVCV words in Lakhota have a word-initial fricative and 18% of them have a word-medial fricative, but only 1% of words have fricatives in both positions. This pattern is superlinear because the expected proportion of words with two fricatives is approximately 6% given the joint proportion of words with fricatives in word-initial and word-medial positions (0.32 x 0.18 = 0.06). Smith and Pater (2020) take a similar approach to assessing the linearity of cumulativity among the marked structures that lead to increasesd schwa realization in French. In a production study, they find that French speakers produce an underlying schwa in 91% of forms where the production of the schwa breaks up a three-consonant cluster (as in mã**ʒ l**[œ] **ɡ**ato 'eat the cake'), and they produce a schwa in 65% of forms when the realization of schwa breaks up a stress clash (as in ˈlo s[œ] ˈvã 'the water was sold'). But, the rate of schwa realization is signficantly higher, namely 94%, when the realization of schwa simultaneously breaks up a three-consonant cluster and a stress clash (as in yn ˈvɛ**st**[œ] ˈ**ʁuʒ** 'a red jacket'). They characterize this pattern as superlinear since the observed rate of schwa realization in doubly-marked forms, 94%, is higher than what is expected by the independent combination of the rates of schwa realization in singly-marked forms (0.91 x 0.65 = 0.59).[5]

However, this work takes a different approach in evaluating cumulative constraint interaction in empirical data and grammar models. This is because the linearity of cumulativity as defined in previous literature faces a few critical problems. First, what is taken to be the "expected" penalty of multiply-violating forms, against which the observed penalty is compared, varies substantially across the literature on cumulativity. Therefore, any conclusion about the manner and degree to which constraints interact in some representative data sample of a language, in acceptability judgments, or in a grammar model's predictions may vary depending on the particular definition assumed. This in turn does not allow for controlled comparisons of how constraint violations accumulate crosslinguistically and in behavioral tasks, or for controlled theory and model comparison. Second, the notion of the "expected penalty" faces logical inconsistencies when extending it to evaluating the predicted linearity of probabilistic grammar models. This boils down to the fact that constraint-based models define the outcome (or, a candidate's probability) over the constraints candidates violates as well as the constraints they do not violate. Lastly, this "expected" penalty is not a prediction of any optimality-theoretic model of phonological grammar. So, there seems to be no principled or theoretically motivated reason why we should assume that the independent combination of penalties, or any other "theory-independent" baseline of comparison similar to this, should be the expected distribution.

Regarding the first problem identified here, as previously discussed, the more standard definition of linearity assumed in previous literature follows that of Albright (2008, 2012) and Smith and Pater (2020). Here, constraint interaction is determined via a comparison of the "expected" penalty or probability distribution against the observed, where the "expected" penalty is the joint combination of the penalties of singly-marked structures. But, other

[5] A higher rate of schwa realization indicates increased penalty for such form, which means that the penalty associated with doubly-marked forms is higher than expected and hence superlinear.

works define the "expected" penalty in different ways, hence introducing inconsistencies in the definition of linearity even within the same work.

Shih (2016, 2017) and Kim (2022) take the "expected" probabilities to be the predicted probabilities of a MaxEnt model which they then compare against the observed distribution to determine the nature of constraint interaction (linearity) in their data. Under this definition of linearity, superlinear patterns are those wherein the observed frequencies of multiply-violating forms in their data are lower than the probabilities a MaxEnt model predicts for those forms. Note that this is different from concluding that the data patterns are superlinear via the comparison against the independent combination of penalties in their data.[6] The MaxEnt models assumed in these works are the same multiple competitions models with asymmetric trade-offs introduced in the previous section: they model their data as the binary competition between a fully faithful candidate with variable violations of markedness constraints against an opposing candidate violating faithfulness only once. Shih (2016, 2017) and Kim (2022) find that their proposed multiple competitions MaxEnt models fail to predict a low enough probability for multiply-violating candidates, therefore concluding that their data shows superlinearity. To this end, they propose additional mechanisms and model parameters to allow MaxEnt to predict the "superlinear" patterns in their data. Shih (2016, 2017) propose conjoined constraints for MaxEnt grammars to directly lower the harmony and ultimately the probability of multiply-violating candidates. And, to predict "superlinear" cumulativity for multiple violations of the same constraint, Kim (2022) introduces a parameter for each constraint in the grammar that exponentates a candidate's number of violations of each constraint, hence also directly lowering the harmony and probability of multiply-violating candidates.[7] Figures 2.7 and 2.8 below shows how constraint conjunction

[6] I refer the reader to these works for an in-depth description and characterization of their data.

[7] Kim (2022) proposes the exponentiation of the number of violations as a way of avoiding the self-conjunction of constraints. Self-conjunction is arguably problematic for constraint-based phonological theories. For example, a stringency hierarchy needs to be assumed among self-conjoined constraints (de Lacy, 2002), where it must be universally stipulated that a self-conjoined constraint that is violated with $n+1$ violations of a simplex constraint has higher rank or weight than a self-conjoined constraint that is violated

and exponentiation directly lower the predicted probability of multiply-violating candidates in multiple competitions MaxEnt models.

Figure 2.7: Example multiple competitions MaxEnt grammars with and without constraint conjunction. The conjoined constraint [ℂ2 & ℂ3] directly lowers the harmony and probability of the doubly-violating candidate since it now incurs a violation of an additional constraint. Adapted from Shih (2017, p. 248-249).

						without CC		**with CC**	
		[ℂ2 & ℂ3]	ℂ1	ℂ2	ℂ3	H	p	H	p
		2.16	0.84	0.44	0.44				
i_1	a. non-violating					0	70%	0	70%
	b. loser		1			-0.84	30%	-0.84	30%
i_2	a. ℂ2-violating			1		-0.44	60%	-0.44	60%
	b. loser		1			-0.84	40%	-0.84	40%
i_3	a. ℂ3-violating				1	-0.44	60%	-0.44	60%
	b. loser		1			-0.84	40%	-0.84	40%
i_4	a. doubly-violating	1		1	1	**-0.88**	**49%**	**-3.04**	**10%**
	b. winner		1			**-0.84**	**51%**	**-0.84**	**90%**

with n or less violations of the simplex constraint. Self-conjunction has otherwise been argued to be useful in analyzing, for example, OCP-like patterns that discourage the multiple presence of a semi-marked structure in some domain. But, in this literature, self-conjunction is proposed only for two violations of the same constraint and not more. See Crowhurst (2011) for a general review on constraint conjunction and more details on self-conjunction.

Figure 2.8: Example multiple competitions MaxEnt grammars with and without the exponentiation of number of violations. Such exponentiation directly lowers the harmony and probability of multiply-violating candidates since their number of violations are now scaled up. Adapted from Kim (2022, p. 11).

				without *b*		**with *b***	
		$\mathbb{C}1$ $w = 0.5$ $b = 1$	$\mathbb{C}2$ $w = 0.2$ $b = 4.5$	H	p	H	p
i_1	a. non-violating			0	62%	0	63%
	b. loser	1		-0.5	38%	-0.5	37%
i_2	a. singly-violating		$1^{4.5}$	-0.4	52%	-0.2	59%
	b. loser	1		-0.5	48%	-0.5	41%
i_3	a. doubly-violating		$2^{4.5}$	**-0.8**	**41%**	**-3.7**	**4%**
	b. winner	1		**-0.5**	**59%**	**-0.5**	**96%**
i_4	a. triply-violating		$3^{4.5}$	**-1.2**	**31%**	**-23**	**0%**
	b. winner	1		**-0.5**	**69%**	**-0.5**	**100%**

This approach to assessing the linearity of cumulative constraint interactions is particularly problematic because previous literature, under different definitions of linearity, find that multiple competitions MaxEnt models already predict superlinear cumulativity without additional mechanisms or model parameters that introduce such superlinearity at harmony. For example, Smith and Pater (2020) report superlinear predictions for Noisy Harmonic Grammar and multiple competitions MaxEnt and they claim that these models provide a better fit to their data given their more flexible hypothesis space. Breiss and Albright (2022) also find superlinear predictions for the same MaxEnt grammar structures assumed in Shih (2016, 2017), Smith and Pater (2020), and Kim (2022) at particular weighing conditions. Specifically, they show that, as shown in Figure 2.9 below, the "expected" probability based on the independent combination of the predicted probabilities for singly-violating candidates is 77.4% (from 0.88 x 0.88), but the MaxEnt grammar predicts for the doubly-violating candidate a probability that is significantly lower. From these findings we can conclude that the baseline MaxEnt models in Shih (2016, 2017) and Kim (2022) fail to provide a good fit to their data given that multiple competitions MaxEnt does not predict enough "superlinearity" without additional mechanisms. Crucially, their results do not suggest that baseline multiple competitions MaxEnt fails to predict superlinearity altogether.

Figure 2.9: Example superlinear prediction in multiple competitions MaxEnt. At the given constraint weights, the predicted probability of the doubly-violating candidate is lower than expected from the probabilities of the singly-violating candidates. Adapted from Breiss and Albright (2022; p. 7).

		$\mathbb{C}1$	$\mathbb{C}2$	$\mathbb{C}3$	H	p
		$w = 5$	$w = 3$	$w = 3$		
i_1	a. $\mathbb{C}$2-violating		1		-3	88%
	b. loser	1			-5	12%
i_2	a. $\mathbb{C}$3-violating			1	-3	88%
	b. loser	1			-5	12%
i_3	a. doubly-violating		1	1	-6	**27%**; expected: **77.4%**
	b. winner	1			-5	73%

Alternatively, Yang et al. (2018) return to the notion of expected penalty as the independent combination of the constraint penalties in their investigation of the English and Mandarin lexicons of monosyllables. They create baseline lexicons assuming all constraints are independent and compare the observed distributions of English and Mandarin monosyllables to these "expected" baseline lexicons. They find that the English and Mandarin lexicons are more well-formed than expected given their superlinearity - the log-probabilities of English and Mandarin monosyllables are lower in the real lexicons than in the sampled lexicons that assume constraint independence.

Other works use linear mixed effects models as the baseline of comparison for observed penalties to characterize the linearity of their data (Pizzo, 2015; Durvasula and Liter, 2020; Breiss, 2020; Breiss and Albright, 2022). “Linear” cumulativity in this case would be when the fixed effects (the independent constraints) are statistically significiant but there is no significant interaction among the effects or constraints. Therefore, signficant interaction terms suggest that the combined effect of the constraints is stronger (superlinear) or weaker (sublinear) than their independent effects, and the polarity of its coefficient determines if the combined effect is stronger or weaker. For example, Pizzo (2015) claims the acceptability scores collected for English-like noncewords with various combinations of complex onsets and codas are *sublinear* because there is a significant and positive interaction term among an onset violation and a coda violation in the same form, indicating that the penalty for doubly-violating forms is less than can be accounted for based on the combination of the independent penalties introduced by each marked structure. Similarly, Breiss and Albright (2022) claim the acceptability scores collected from artificial grammar learning experiments are *superlinear* because the significant interaction term among the two phonotactic patterns participants learned in the study is negative, indicating that the penalty for doubly-violating forms is greater than can be accounted for based on the penalties independently introduced by each violation. This evaluation metric for linearity is similar to the independence-based definition of linearity used by Albright (2008) and others, since a significant interaction term of either polarity in these linear mixed effects models indicates that the relationship among the fixed effects is not strictly additive.

Lastly, Smith and Pater (2020) actually use a different definition of linearity in evaluating constraint interaction in various probabilistic phonological models than the one used to evaluate the rate of schwa realization in their data. They evaluate constraint interaction for these models by identifying each constraint’s *contribution* to the penalty or probability of candidates with varying degrees of violations. They define constraint contribution as the difference between the probability of a candidate that violates a given constraint and a minimally different candidate in which the given constraint violation does not occur. So, the linearity of cumulativity is determined by comparing the contribution to penalty of a constraint in a non-cumulative context compared to its contribution in a cumulative context, as opposed to comparing an observed probability distribution to an "expected" distribution of some kind. For example, the non-cumulative contribution of a constraint such as *CCC is the probability of producing a schwa in a non-violating context (for example, in a form such as [yn bɔt(œ) ʃinˈwaz] 'a Chinese boot') minus the probability of producing a schwa in a form where only *CCC is violated (as in [ˈmã**ʒ** **l**(œ) **g**aˈto] 'eat the cake'). And, its cumulative contribution is the difference between the probability of a form that violates *CCC in the environment of another constraint (in this case, *Clash as [yn ˈvɛ**st**(œ) ˈ**ʁ**uʒ] 'a red jacket') and the probability in non-cumulative contexts (in [ˈmã**ʒ** **l**(œ) **g**aˈto] 'eat the cake'). So, following Smith and Pater (2020), two constraint are in a *sublinear* interaction if a constraint's contribution in a cumulative context (in the environment of another constraint violation) is less than its independent contribution, and two constraints have a *superlinear* interaction if a constraint's cumulative contribution is greater than its independent contribution. Using this definition of linearity, Smith and Pater (2020) conclude that Stochastic OT (Boersma, 1997; Boersma and Hayes, 2001) can only predict sublinear cumulativity while Noisy Harmonic Grammar (Boersma and Pater, 2016) and multiple competitions MaxEnt predict both sublinear and superlinear cumulativity.[8] Note that this definition of linearity is quite different from how constraint interaction was evaluated in their data – they evaluate the linearity of cumulativity in the rate of schwa production using the more traditional definition used

[8] Stochastic OT is a probabilistic variant of Classic OT where each constraint's ranking value, a real number that is eventually converted to ordinal strict ranking, is perturbed by some normally distributed noise. The noise introduced to the ranking values of constraints may generate flexible rankings, hence affecting the frequency at which a given candidate is predicted as the winner. Noisy Harmonic Grammar is a probabilistic variant of Standard HG where weights, rather than ranking values and eventually the strict ranking, is perturbed by noise in the same manner as in Stochastic OT. Like Stochastic OT, the noise introduced to constraint weights affects the harmony of each outcome and potentially the frequency at which the candidate is predicted as the winner.

by Albright (2008) and others, but they use the notion of constraint contribution when evaluating linearity in the predictions of probabilistic grammar models.[9]

So, in summary, the first problem with the linearity of cumulativity, which is the mainstream way of evaluating constraint interaction both in empirical data and in the predictions of gramamr models, is that the literature varies considerably in what is considered to be the "expected" penalty against which the observed penalty is compared. And, there is also variation in the definition of linearity used to evaluate the same phenomena. Therefore, any conclusion about linearity depends on the particular definition assumed. This doesn't allow for principled and controlled comparison of the data against model predictions, or across various empirical patterns and comparison of theories and models of grammar.

The second problem with the notion of "expected" penalty is that it faces some logical inconsistencies. This becomes especially apparent when evaluating constraint interaction in the predictions of probabilistic phonological models such as MaxEnt. As an example, assume the minimal single competition MaxEnt grammar in Figure 2.10 below. The grammar has a zero-violating candidate $c_{(0,0)}$, two singly-violating candidates $c_{(0,1)}$ and $c_{(1,0)}$, one for each separate violation of the two constraints, and a doubly-violating candidate $c_{(1,1)}$ with single violations of each of the two constraints, therefore an example of ganging cumulativity. Recall that, in single competitions MaxEnt grammars, all candidates are in the same probability distribution and so they are all are normalized by the same constant Z.

[9] Pizzo (2015) also loosely uses this notion of constraint contribution when evaluating constraint interaction in their data compared to the predictions of (single competitions) MaxEnt. They note that MaxEnt models, compared to Standard HG's harmony, predicts sublinear cumulativity as each violation introduces a penalty that is lower than the penalty introduced by the preceding violation. However, the linearity of cumulativity in their data is determined with linear mixed effects models as discussed above, and not via the notion of a constraint's contribution to penalty.

Figure 2.10: Schematic single competitions MaxEnt grammar for ganging cumulativity among two constraints.

candidates	$\mathbb{C}1$ $w_{\mathbb{C}1}$	$\mathbb{C}2$ $w_{\mathbb{C}2}$	H	predicted probability
$c_{(0,0)}$	0	0	0	$1\ /\ Z$
$c_{(1,0)}$	1	0	$-w_{\mathbb{C}1}$	$e^{-w_{\mathbb{C}1}}\ /\ Z$
$c_{(0,1)}$	0	1	$-w_{\mathbb{C}2}$	$e^{-w_{\mathbb{C}2}}\ /\ Z$
$c_{(1,1)}$	1	1	$-w_{\mathbb{C}1} - w_{\mathbb{C}2}$	$e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}}\ /\ Z$

Following the more widely adopted definition of linearity, the "expected" probability of the doubly-violating candidate $c_{(1,1)}$ is the multiplication of the predicted probabilities of $c_{(0,1)}$ and $c_{(1,0)}$, the singly-violating candidates. However, the probability of each candidate in MaxEnt is defined by the constraints that the candidate violates and the constraints that the candidate *does not* violate. For example, the probability of the singly-violating candidate $c_{(1,0)}$ is defined over a violation of $\mathbb{C}1$ *and no violations* of $\mathbb{C}2$, and not just over a violation of $\mathbb{C}1$. The same goes for the other singly-violating candidate: the probability of $c_{(0,1)}$ is defined over a violation of $\mathbb{C}2$ *and no violations* of $\mathbb{C}1$. Therefore, the "expected" probability of $c_{(1,1)}$ is logically inconsistent: its expected probability is the probability of violating $\mathbb{C}1$ and not $\mathbb{C}2$ times the probability of violating $\mathbb{C}2$ and not $\mathbb{C}1$ – an apparent contradiction. This logical inconsistency is true of both single competition and multiple competitions MaxEnt grammars, and also for multiple violations of the same constraint (counting cumulativity).

The last problem with the notion of linearity as defined in previous works is that what is taken to be the "expected" penalty is often not a prediction of any theoretically motivated model of phonological grammar. In other words, no mainstream theory of phonological grammar explictly predicts strictly linear constraint interactions, therefore there is no theoretically principled reason for why the independent combinations of penalties is the default assumption we should have about constraint interaction. This is true of the general notion of the expected penalty of multiply-violating forms as being the independent combination of their simplex penalties, which applies to the use of linear mixed effects models as well.

The harmony values in Standard HG and its variants are indeed linear by definition (a candidate's harmony is the negative sumproduct of the number of violations of each constraint and the constraint weights, therefore harmony is strictly additive), but Standard HG's output is the most harmonic candidate as opposed to these linear raw harmony scores, so its prediction is not strictly linear by definition. And, as previously discussed, the output of MaxEnt grammars are probabilities derived via a non-linear (exponential) transformation of negative harmony scores. So, even if Standard HG and its successors do have a component in the grammar that defines a strictly linear interaction among constraint violations, the output of the grammar is not such linear component.

Two possible exceptions to this generalization are Linear Optimality Theory (Keller, 2000, 2006) and Linear Harmonic Grammar (Pizzo, 2015). Linear OT assumes that the grammaticality of a given structure is proportional to its harmony, and Linear HG assumes that the probability of a given structure is proportional to its harmony. Specifically, both models assume the difference in harmony among two candidates of the same input is proportional to their acceptability or probability difference. In Linear OT, these acceptability differences among candidates serve directly as the ranking arguments to learning constraint weights, therefore affecting the predicted harmonies scores of all candidates.[10] Linear HG predicts harmony values as opposed to predicting a single winning candidate, then evaluates differences in harmony against differences in probability to set the constraint weights of the grammar.

However, these two grammar models might not be true exceptions to this generalization

[10] In Linear OT, constraint weights are learned via Least Squared Estimation, namely by minimizing the sum of the squared differences between the acceptability differences and harmony differences among all possible candidate pairs in the grammar.

because Linear OT and Linear HG do not directly predict acceptability scores or probability distributions from harmony. These models instead predict harmony values and simply assume a *linear linking hypothesis* between harmony and acceptability and harmony and probability, namely that acceptability and harmony differences are proportional to harmony differences among candidates. As it stands, there is little to no empirical evidence directly in support of the hypothesis that acceptability and probability are linearly proportional to harmony. In fact, the very existence of non-linear patterns of cumulativity in a variety of languages and in behavioral tasks could be argued as evidence against such hypothesis, as is argued explicitly by Pizzo (2015) and Yang et al. (2018).

Given these existing problems with the notion of linearity, I argue for a different evaluation metric of constraint interaction. I propose to evaluate constraint interaction in a similar manner to Smith and Pater (2020)'s measure of *constraint contribution*, where the nature of interaction among constraints is determined by comparing a constraint's effect across cumulative and non-cumulative contexts. However, I will characterize the different types of constraint interaction here using the notion of **concavity** to avoid conflating this metric with the more standard measure of linearity that compared observed and "expected" penalties. Specifically, I define *concave-up* patterns are those wherein a constraint violation introduces a weaker penalty in the environment of another constraint violation compared to its effect on its own ("sublinear" in Smith and Pater (2020)), and *concave-down* patterns are those wherein a constraint violation introduces a stronger penalty in the environment of another constraint violation compared to its effect on its own ("superlinear" in Smith and Pater (2020)). This allows us assess both the manner and degree to which a constraint's contribution to penalty potentially varies given its interaction with other constraints violations.

The definition of concavity differs only slightly across accumulating violations of the same constraint vs. single violations of multiple constraints (namely, across *counting* vs. *ganging* cumulativity, respectively). In counting cumulativity, *concave-up* patterns are those wherein an n^{th} violation introduces a stronger penalty than the $n+1^{\text{st}}$ violation, and *concave-down* patterns are those wherein the $n+1^{\text{st}}$ violation introduces a stronger penalty than the n^{th} violation. In ganging cumulativity, *concave-up* patterns are those wherein the penalty introduced by the violation of a given constraint is *smaller* in the environment of a violation of a different constraint compared to the penalty it introduces when violated independently, and *concave-down* patterns are those wherein the penalty introduced by a violation of the given constraint is *larger* in the environment of a violation of a different constraint compared to its effect when violated independently. We can also equivalently evaluate the concavity of cumulativity via the difference in penalty across subsequent violations and across non-cumulative and cumulative contexts: if their difference in penalty is positive then the pattern is concave-up and if they're negative then the pattern is concave-down, and the patterns are more strongly concave-up or concave-down as these differences increase. The table in Figure 2.11 below summarizes the definition of concavity across counting and ganging cumulativity.

Figure 2.11: The definition of concavity across counting and ganging cumulativity. c indicates a candidate with no constraint violations, c_n a candidate with n violations of the same constraint, and c_X a candidate with one violation of Constraint X.

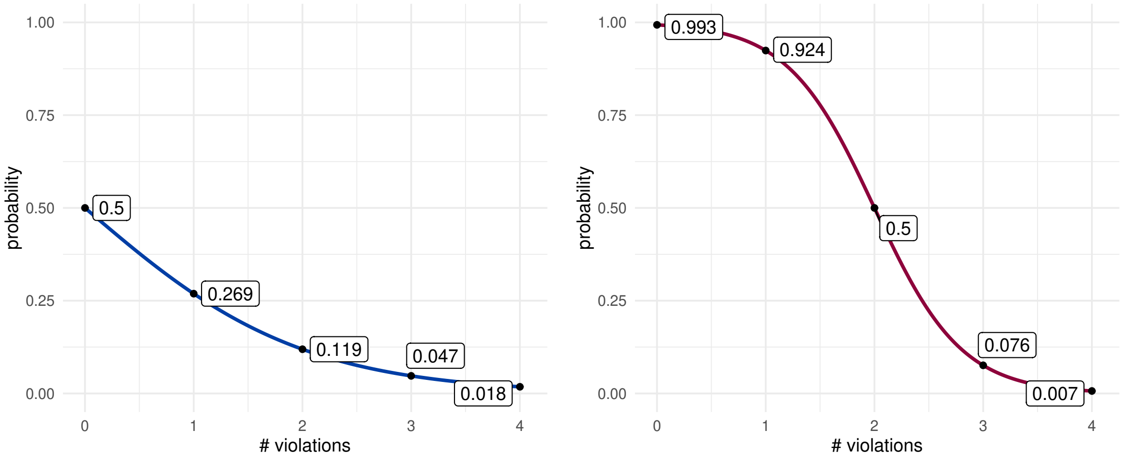
Concavity	**Counting cumulativity**	**Ganging cumulativity**
Concave-up	$P(c_{n-1}) - P(c_n) > P(c_n) - P(c_{n+1})$	$P(c) - P(c_A) > P(c_A) - P(c_{A,B})$
	or	or
	$[P(c_{n-1}) - P(c_n)] - [P(c_n) - P(c_{n+1})] > 0$	$[P(c) - P(c_A)] - [P(c_A) - P(c_{A,B})] > 0$
		(along Constraint A)
Concave-down	$P(c_{n-1}) - P(c_n) < P(c_n) - P(c_{n+1})$	$P(c) - P(c_A) < P(c_A) - P(c_{A,B})$
	or	or
	$[P(c_{n-1}) - P(c_n)] - [P(c_n) - P(c_{n+1})] < 0$	$[P(c) - P(c_A)] - [P(c_A) - P(c_{A,B})] < 0$
		(along Constraint A)

Note from the table above that concavity in ganging cumulativity is evaluated along a specific constraint (Constraint A) even when two constraints are involved in the interaction (Constraints A and B). Crucially, in ganging cumulativity, *all* constraints in the cumulative interaction must be evaluated for their contribution to penalty to gain a full understanding of constraint interaction in the given data. It is logically possible to observe patterns of *mixed concavity*, which are patterns of ganging cumulativity where some constraint's effect is concave-up given a violation of the other constraint, but the other constraint's effect is concave-down given a violation of the original constraint.[11]

The notion of concavity becomes visually identifiable when plotting probabilities on a graph. The pattern of counting cumulativity in Figure 2.12a is concave-up since each decrease in probability at each subsequent violation is smaller than the decrease in probability at the immediately preceding violation: the constraint's effect on probability is increasingly weaker as its violations accumulate. For example, the probability decrease introduced by the first violation, 0.231 (0.5 - 0.269), is greater than the penalty introduced by the second violation, 0.15 (0.269 - 0.119), and this pattern continues on throughout subsequent violations. On the other hand, the distribution in Figure 2.12b is concave-down since the probability decrease at the second violation is larger than the decrease in probability at the first violation. The penalty introduced by the third violation is roughly the same as the penalty introduced by the second (0.05 - 0.076 = 0.424), and at later violations the pattern becomes concave-up. Going forward, I'll assume that concave-down patterns are those wherein some portion of the distribution, especially at the first few violations, shows concave-down cumulativity, while concave-up patterns are those wherein the penalties across all accumulating violations are concave-up.

[11] More concretely, this would be a case where $P(c) - P(c_A) > P(c_A) - P(c_{A,B})$, but $P(c) - P(c_B) < P(c_B) - P(c_{A,B})$, or vice versa.

Figure 2.12: Example concave-up and concave-down cumulativity for accumulating violations of the same constraint (counting cumulativity).



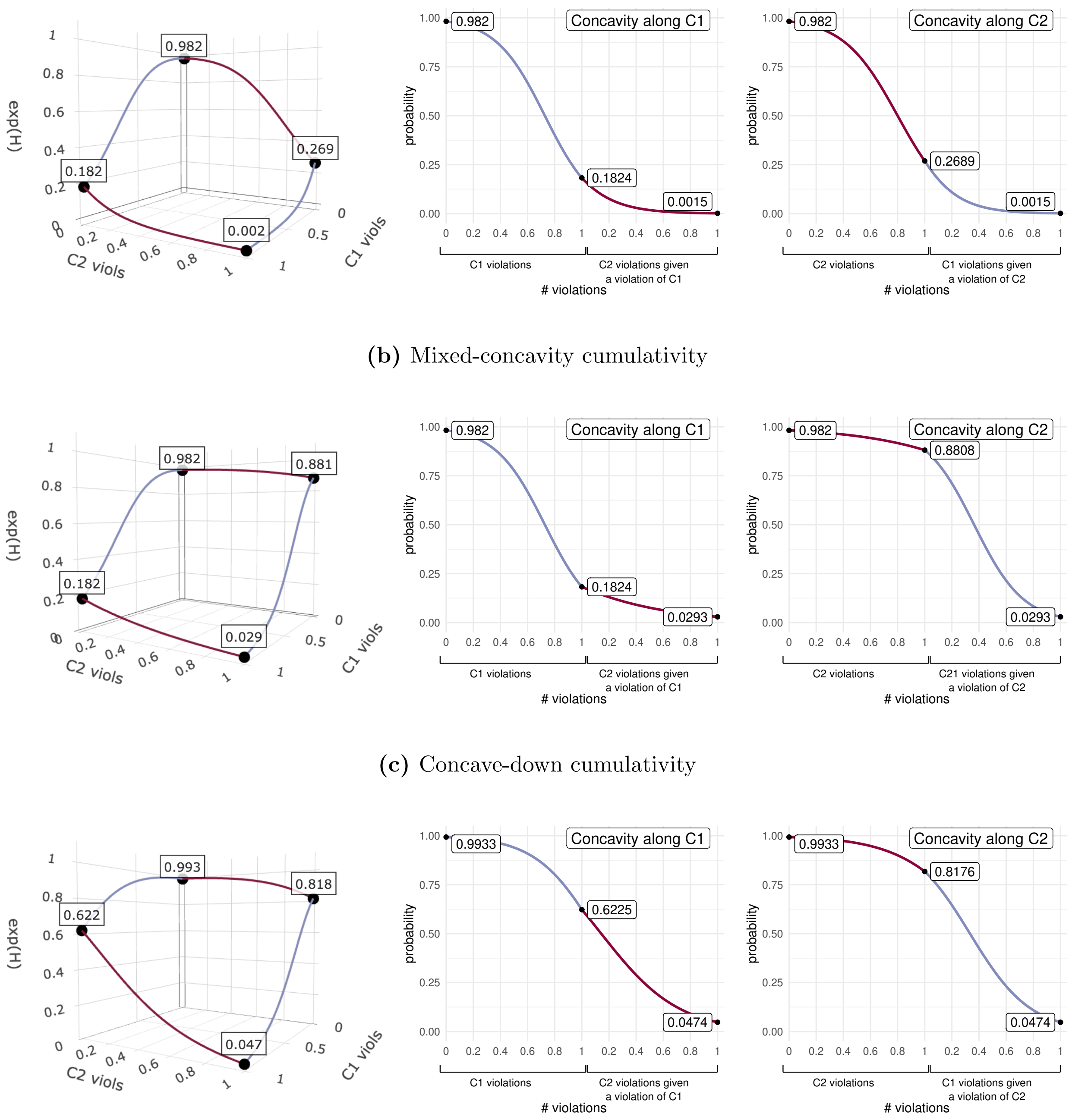
(a) Concave-up cumulativity

(b) Concave-down cumulativity

Figure 2.13 below illustrates example concave-up, concave-down, and mixed-concavity patterns of ganging cumulativity. The probability distribution in Figure 2.13a is concave-up since both constraints in the interaction introduce a smaller decrease in probability relative to the decrease in probability they introduce when violated separately.[12] This is also easily observable in the middle and rightmost graphs that plot each constraint's effect on probability independently and along with a violation of the other constraint. Next, Figure 2.13b shows an example pattern of mixed concavity: a violation of $\mathbb{C}1$ introduces a smaller penalty with a violation of $\mathbb{C}1$ compared to its effect when violated independently (concave-up), but $\mathbb{C}2$ introduces a stronger penalty when violated with $\mathbb{C}2$ compared to when it is violated independently (concave-down). Finally, the distribution in Figure 2.13c is concave-down since, for both $\mathbb{C}1$ and $\mathbb{C}2$, the probability decreases both constraints introduce when violated together is larger than the probability decrease they introduce when violated separately.

[12] The distribution along $\mathbb{C}1$ is concave-up since $0.982 - 0.182 > 0.182 - 0.002$, and the distribution along $\mathbb{C}2$ is also concave-up since $0.982 - 0.269 > 0.269 - 0.002$.

Figure 2.13: Examples of concave-up and concave-down cumulativity for accumulating violations of two different constraints (ganging cumulativity).



(a) Concave-up cumulativity

(b) Mixed-concavity cumulativity

(c) Concave-down cumulativity

In summary, I propose an alternative measure of constraint interaction similar to that used by Smith and Pater (2020) in evaluating the effects of accumulating constraint violations. This is because the more standard measure of constraint interaction faces some critical problems given the notion of the "expected" penalty: what is often taken as the "expected" penalty varies significantly within and across previous works, the definition is logically inconsistent since constraint-based grammars define outcomes and probability distributions according to the constraints that candidates do and don't violate, and this expected penalty is not a prediction of any theory of phonological grammar (with some possible exceptions discussed). Like the more standard measure of linearity, the concavity of cumulativity still assesses the greater or lesser effect of constraint violations in the environment of other violations without relying on the notion of an "expected" penalty. An additional benefit to this evaluation metric is that it can be used in evaluating constraint interaction in various kinds of empirical patterns and model predictions without needing to adjust its definition. Finally, as will become apparent in the next few sections, the concavity of cumulativity clearly demarcates the boundary between the predictions of single competition and multiple competitions MaxEnt models.

The subsequent sections of this chapter show that single competition MaxEnt models only predict *concave-up* cumulativity for both counting and ganging cumulativity and for ganging cumulativity of *mixed concavity*, while multiple competitions MaxEnt models predict concave-up, mixed concavity, and *concave-down* cumulativity for both counting and ganging cumulativity.

2.4 Single competition MaxEnt grammars

2.4.1 Counting cumulativity

Recall that single competition MaxEnt models assign a single probability distribution over all candidate forms in the grammar, therefore all candidates' exponentiated harmonies are normalized by the same constant Z. This is the typical model structure assumed in phonotactic modeling and learning, wherein a form's harmony and probability is strictly defined by its surface properties and not in relation to a phonological input (Hayes and Wilson, 2008). In analyzing the predictions of single competition MaxEnt models for counting cumulativity, multiple violations of the same constraint, we can assume the schematic grammar in Figure 2.14 below.

Figure 2.14: Schematic single competition MaxEnt grammar for counting cumulativity – accumulating violations of the same constraint.

inputs	$\mathbb{C}$	H	predicted prob.
	w		
c_0	0	0	$1\ /\ \mathbf{Z}$
c_1	1	$-w_m$	$e^{-w}\ /\ \mathbf{Z}$
c_2	2	$-2w_m$	$e^{-2w}\ /\ \mathbf{Z}$
c_3	3	$-3w_m$	$e^{-3w}\ /\ \mathbf{Z}$
...	...	...	...

In general, we can think of single competition MaxEnt models as defining a geometric sequence that sums to one, as in Equation 2.5 below, where each candidate in the tableau above represents a term in a geometric series. In counting cumulativity under this model, the predicted probability of each subsequent candidate is the preceding candidate's predicted probability multiplied by the common ratio e^{-w}, and the initial term of the series is the predicted probability of the non-violating candidate, which is $\frac{1}{Z}$. The geometric series sums to one since all candidate forms are in the same probability distribution.

$$\frac{1}{Z} + \left(\frac{1}{Z}\right) e^{-w} + \left(\frac{1}{Z}\right) \left(e^{-w}\right)^2 + \left(\frac{1}{Z}\right) \left(e^{-w}\right)^3 + ... = 1 \tag{2.5}$$

Since the common ratio among the terms in the series is e^{-w}, then the predicted probabilities fall by a factor of e^{-w} at each additional violation. In other words, the predicted probability of a candidate with $n+1$ violations is $\frac{1}{e^w}$ of the probability of the candidate with n violations. For example, if we assume $w = \ln 3$, which is around 1.0986, then probability falls by a factor of 3 at each additional violation, which means that the probability at a subsequent violation is $\frac{1}{3}$ of the probability at the immediately preceding violation.[13] This can also be deduced from the proportional probability of a candidate with $n+1$ violations and a candidate with n violations, as in Equation 2.6 below.

$$\frac{P(c_{n+1})}{P(c_n)} = \frac{(\frac{1}{Z})(e^{-w(n+1)})}{(\frac{1}{Z})(e^{-wn})} = \frac{e^{-w(n+1)}}{e^{-w}} = \frac{e^{-wn\ +\ (-w)}}{e^{-wn}} = \frac{e^{-wn} \times e^{-w}}{e^{-wn}} = e^{-w} \tag{2.6}$$

Therefore, single competition MaxEnt's predictions for counting cumulativity are all *concave-up*: a later violation always causes a smaller decrease in probability relative to the decrease in probability associated with an earlier violation. Naturally, the extent to which probability decreases is a function of the constraint weight w, such that, at larger values of w, probability decreases by a larger multiplicative factor.

This is also deducible from the definition of concavity outlined in the previous section in Figure 2.11. To more concretely show that all counting cumulativity predicted by single competitions MaxEnt is concave-up, we want to show that the difference between the probability of a candidate with $n-1$ violations and a candidate with n violations is larger than the difference in probability between a candidate with n violations and a candidate with $n+1$ violations. Therefore, we must find that the inequality below is true for any constraint

[13] Assuming $w = \ln 3$, probability falls by a factor of $e^{\ln 3}$. Since $e^{-\ln 3} = \frac{1}{e^{\ln 3}} = \frac{1}{3}$, then probability falls by a factor of $\frac{1}{\frac{1}{3}} = 3$.

weight $w > 0$ and any violation level $n \geq 1$.[14,15]

$$P(c_{n-1}) - P(c_n) > P(c_n) - P(c_{n+1}) \tag{2.7}$$

$$e^{-w(n-1)} - e^{-wn} > e^{-wn} - e^{-w(n+1)} \tag{2.8}$$

$$e^{-w(n-1)}(1 - e^{-w}) > e^{-wn}(1 - e^{-w}) \tag{2.9}$$

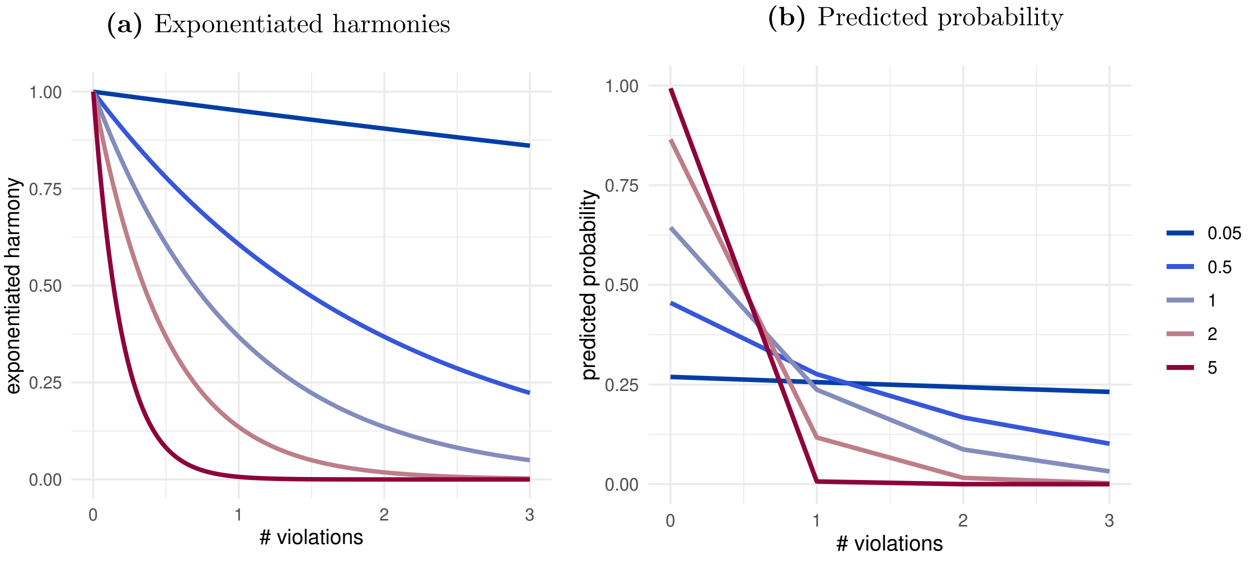
$$e^{-w(n-1)} > e^{-wn} \tag{2.10}$$

The inequality $e^{-w(n-1)} > e^{-wn}$ in Equation 2.10 is clearly true for all $w > 0$ and for all $n \geq 1$: the exponent of the left-hand term, $-w(n-1)$, is always smaller than the exponent of the right-hand term, $-wn$, at these values for n and w. So, these inequalities also show that single competition MaxEnt's predictions are all *concave-up* since the difference between two subsequent earlier violations is always larger than the difference between two subsequent later violations at any non-negative constraint weight and violation level. Figure 2.15 below visually shows this with example predictions at different constraint weights, both before and after normalizing by $\frac{1}{Z}$.

[14] The inequalities in Equations 2.7 to 2.10 ignore the normalizing constant $\frac{1}{Z}$ since it is a common factor among all terms in the inequality. Therefore, whether a pattern is concave-up or concave-down does not vary across exponentiated harmonies and probability. See Figure 2.15 below.

[15] As is the standard in Standard HG and MaxEnt, I assume in this work that constraints and violation levels can only be assigned a non-negative real number – constraints only assign penalties, and not rewards, to candidates that violate them. Negative weights are argued to have the undesirable property of reversing harmonic bounding. See Pater (2009); Boersma and Pater (2016); Hayes and Kaplan (2025) for more details. For visualization purposes I also assume continuous violation levels, though violation levels are whole numbers and not decimals in Standard HG and its variants.

Figure 2.15: Predicted exponentiated harmonies and probabilities of single competition MaxEnt models at various constraint weights, across zero to three violations of the same constraint.


(a) Exponentiated harmonies

(b) Predicted probability

Naturally, a constraint's contribution to penalty at any level of violation is a function of the weight of the variably-violated constraint. As the constraint weight increases, predicted probability falls by a larger factor.[16] Figure 2.16 below shows how the strength of the predicted concave-up patterns changes as a function of the constraint weight by plotting the difference $(e^{-w(n-1)} - e^{-wn}) - (e^{-wn} - e^{-w(n+1)})$, which compares the penalty introduced by the n^{th} violations against that introduced by the $n+1^{\text{st}}$, at $n = 1$ and $n = 2$ violation levels.

[16] Recall that in single competition MaxEnt models probability falls by a factor of e^{-w}. So, for example, at $w = \ln 3$, which is approximately 1.0986, predicted probability in single competition MaxEnt decreases by a factor of 3 at each additional violation. But when $w = \ln 5$, which is approximately 1.60944, probability falls by a factor of 5 at each additional violation.

Figure 2.16: Concavity as a function of the constraint weight w. The $n = 1$ curve plots the difference in exponentiated harmony between candidates with zero vs. one and one vs. two violations, and the $n = 2$ curve between one vs. two and two vs. three violations of the same constraint.

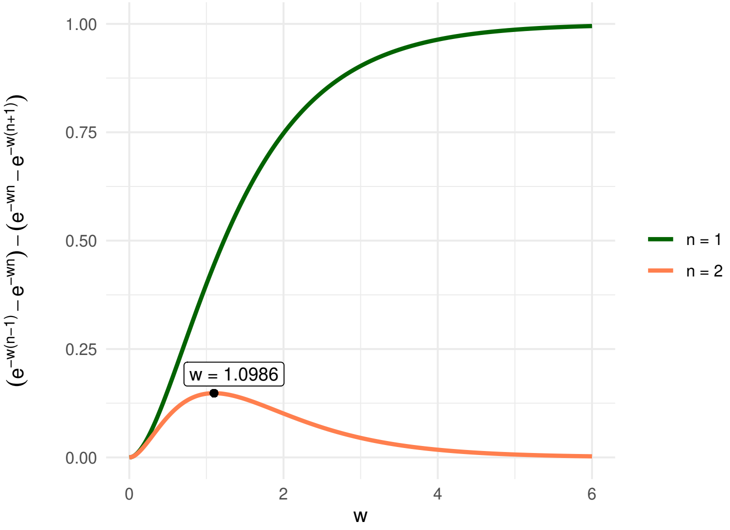


Figure 2.16 above shows that, as the constraint weight increases, the difference in exponentiated harmony between candidates with 0 and 1 violation vs. candidates with 1 vs. 2 violations ($n = 1$) increases. This means that predicted probabilities are *increasingly concave-up* as the constraint weight increases: the first violation's penalty becomes stronger and the second violation's penalty becomes weaker, hence increasing their difference. This increase in probability difference asymptotes on 1 at larger values of w since the probability of candidates with even one violation of the constraint approximate zero (and therefore the probability of c_0, the candidate with zero violations of the constraint, approximates 1). However, the strength of concavity at later violations (the $n = 2$ curve) show a different pattern. The difference between candidates with 1 vs. 2 violations and 2 vs. 3 violations increases at low constraint weights, therefore also predicting an increasingly concave-up pattern. But at $w = 1.0986$, the maximum of the $n = 2$ curve, the difference starts to decrease as the

probabilities of candidates with two or more violations approximate zero.[17] Note that all penalty differences in this plot are positive since all probability distributions predicted by single competition MaxEnt models are concave-up: constraint violations have an increasingly weaker effect on probability as violations accumulate.

2.4.2 Ganging cumulativity

I now turn to describing the range of predicted constraint interactions of single competition MaxEnt grammars for ganging cumulativity – accumulating violations of different constraints. As a starting point, we can assume the minimal grammar in Figure 2.17 below. The grammar has a zero-violating candidate $c_{(0,0)}$, two singly-violating candidates $c_{(0,1)}$ and $c_{(1,0)}$, one for each separate violation of two constraints, and a doubly-violating candidate $c_{(1,1)}$ with single violations of two constraints. As with counting cumulativity, the predicted probability of a given candidate is its exponentiated harmony normalized by the sum of the exponentatiated harmonies of all candidates in the grammar.

Figure 2.17: Schematic single competition MaxEnt grammar with ganging cumulativity among two constraints.

candidates	$\mathbb{C}1$ $w_{\mathbb{C}1}$	$\mathbb{C}2$ $w_{\mathbb{C}2}$	H	predicted probability
$c_{(0,0)}$	0	0	0	$1 \ / \ Z$
$c_{(1,0)}$	1	0	$-w_{\mathbb{C}1}$	$e^{-w_{\mathbb{C}1}} \ / \ Z$
$c_{(0,1)}$	0	1	$-w_{\mathbb{C}2}$	$e^{-w_{\mathbb{C}2}} \ / \ Z$
$c_{(1,1)}$	1	1	$-w_{\mathbb{C}1} - w_{\mathbb{C}2}$	$e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}} \ / \ Z$

[17] Although not shown in Figure 2.16, violations levels at $n \geq 3$ show the same pattern as at $n = 2$ but with overall lower probabilities. The maximum weight at which the probability difference starts to fall is also lower for greater violation levels.

Since ganging cumulativity now involves the interaction of two (or more) constraints, we must separately assess the effect on probability of each constraint in cumulative and non-cumulative contexts. So, assuming the interaction of only two constraints as in the tableau in Figure 2.17 above, we must investigate the differences in predicted probability along $\mathbb{C}1$, namely $P(c_{(0,0)}) - P(c_{(1,0)})$ and $P(c_{(1,0)}) - P(c_{(1,1)})$, and also along $\mathbb{C}2$, namely $P(c_{(0,0)}) - P(c_{(0,1)})$ and $P(c_{(0,1)}) - P(c_{(1,1)})$. If both differences among the penalties of singly-violating candidates and the doubly-violating candidate are positive then the predicted patterns are concave-up, if only one of the differences is positive and the other is negative then the predicted patterns are of mixed concavity, and if both differences are negative then the model's predictions are concave-down.[18]

Concavity along $\mathbb{C}1$:

$$[P(c_{(0,0)}) - P(c_{(1,0)})] - [P(c_{(1,0)}) - P(c_{(1,1)})] \tag{2.11}$$

$$(1 - e^{-w_{\mathbb{C}1}}) - (e^{-w_{\mathbb{C}1}} - e^{-w_{\mathbb{C}1} - w_{\mathbb{C}2}}) \tag{2.12}$$

Concavity along $\mathbb{C}2$:

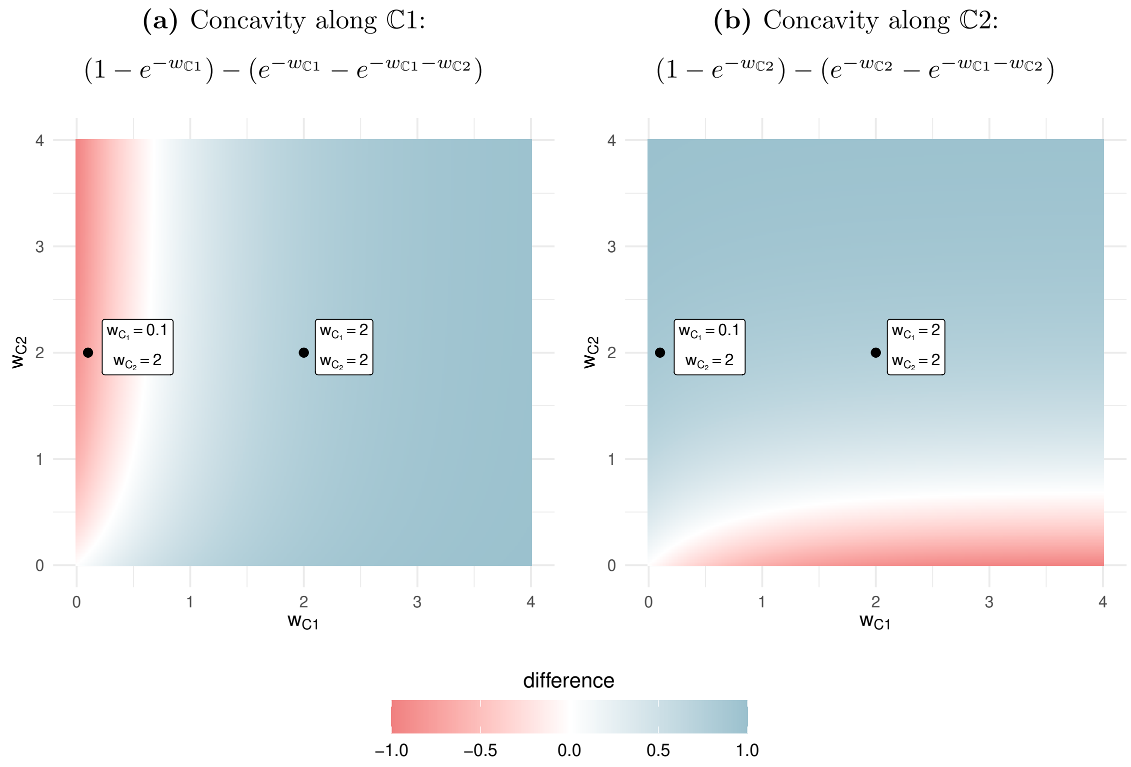
$$[P(c_{(0,0)}) - P(c_{(0,1)})] - [P(c_{(0,1)}) - P(c_{(1,1)})] \tag{2.13}$$

$$(1 - e^{-w_{\mathbb{C}2}}) - (e^{-w_{\mathbb{C}2}} - e^{-w_{\mathbb{C}1} - w_{\mathbb{C}2}}) \tag{2.14}$$

Figure 2.18 below shows the predicted strength and concavity of cumulativity at a continuous range of constraint weights for the two interacting constraints.

[18] Again, the normalizing constant Z in the differences above are ignored since it is a common factor among all terms, and normalizing by Z does not change the predicted concavities among any of the constraint violations.

Figure 2.18: Predicted concavities in exponentiated harmony space for ganging cumulativity among two constraints in single competition MaxEnt.

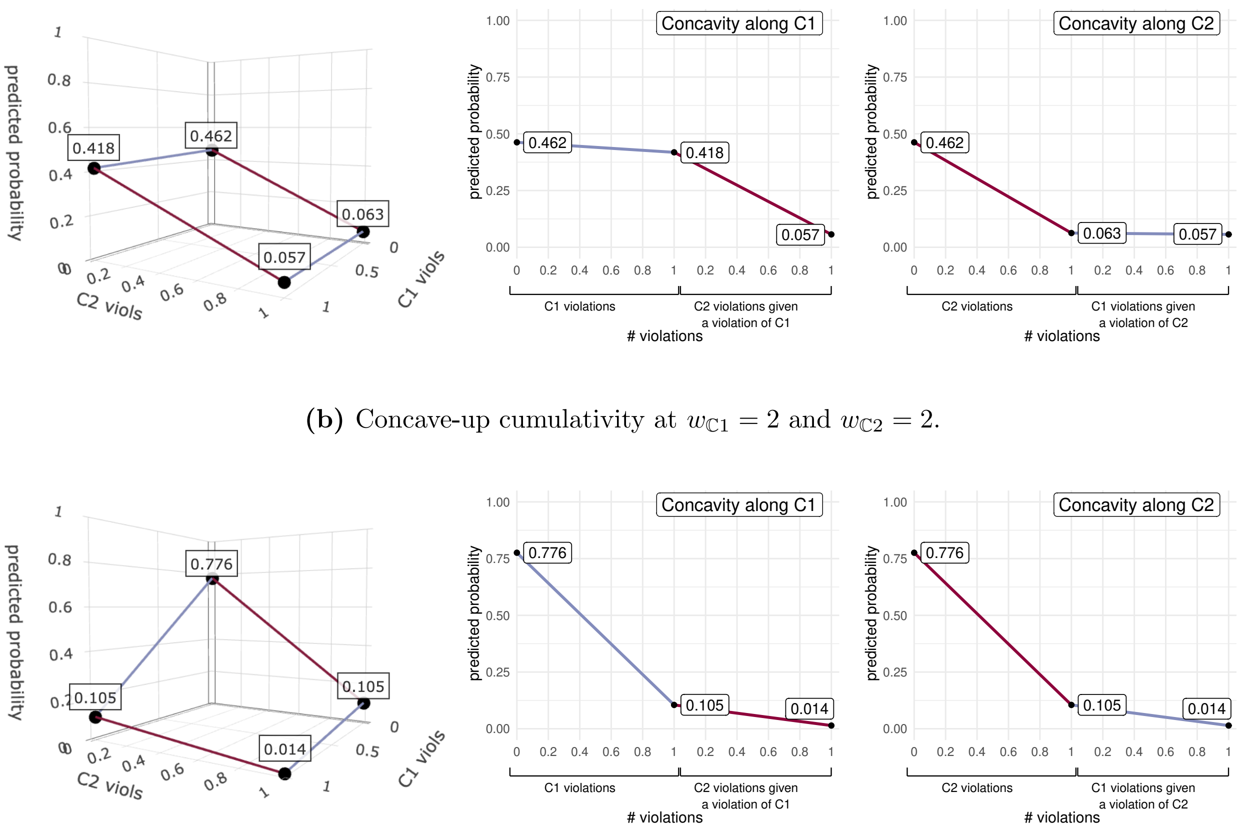


(a) Concavity along $\mathbb{C}1$: $(1 - e^{-w_{\mathbb{C}1}}) - (e^{-w_{\mathbb{C}1}} - e^{-w_{\mathbb{C}1} - w_{\mathbb{C}2}})$

(b) Concavity along $\mathbb{C}2$: $(1 - e^{-w_{\mathbb{C}2}}) - (e^{-w_{\mathbb{C}2}} - e^{-w_{\mathbb{C}1} - w_{\mathbb{C}2}})$

This Figure shows that single competition MaxEnt predicts either concave-up ganging cumulativity, or ganging cumulativity of mixed concavity at certain weighing conditions. For example, at $w_{\mathbb{C}1} = 2$ and $w_{\mathbb{C}2} = 2$, single competition MaxEnt predicts concave-up cumulativity along both constraints since both differences are positive, meaning that the cumulative contribution of both constraints is smaller than their independent contributions. However, at $w_{\mathbb{C}1} = 0.1$ and $w_{\mathbb{C}2} = 2$, the difference in probability along $\mathbb{C}1$ is negative, so the predicted pattern is concave-down along $\mathbb{C}1$, but the predicted probabilities are concave-up along $\mathbb{C}2$ since the difference along $\mathbb{C}2$ is positive at these constraint weights. This means that, at $w_{\mathbb{C}1} = 0.1$ and $w_{\mathbb{C}2} = 2$, the predicted probabilities are of *mixed concavity*: a violation of $\mathbb{C}1$ introduces a greater penalty in the environment of $\mathbb{C}2$ violation compared to the penalty it introduces without a violation of $\mathbb{C}2$, but a violation of $\mathbb{C}2$ introduces a smaller penalty when violated with $\mathbb{C}1$ compared to the penalty it introduces without a violation of $\mathbb{C}1$. Figure 2.19 below shows the predicted concavities at these constraint weights in probability space (namely, after normalizing by Z).

Figure 2.19: Predicted probability distributions for ganging cumulativity in single competition MaxEnt at select weights of $\mathbb{C}1$ and $\mathbb{C}2$.



(a) Mixed concavity at $w_{\mathbb{C}1} = 0.1$ and $w_{\mathbb{C}2} = 2$.

(b) Concave-up cumulativity at $w_{\mathbb{C}1} = 2$ and $w_{\mathbb{C}2} = 2$.

It's important to note that, in patterns of mixed concavity, the "concave-down" pattern along one of the constraints occurs because the other constraint violated in the cumulative context has a relatively high weight. Therefore, the larger penalty introduced by such constraint violation is due to the behavior of the other constraint in the cumulative context, and not because the two violations interact in such a way that increases their penalty when violated together and not independently. Therefore, single competition MaxEnt models fail to predict true concave-down patterns of ganging cumulativity. Notice in Figure 2.18 above that the red-colored areas are disjoint across the grids: it is always the case that, if one of the constraints has a concave-down contribution to penalty, the other will show a concave-up contribution. Therefore, the patterns that are predicted by single competition MaxEnt for ganging cumulativity are either concave-up along both constraints or are patterns of mixed concavity, and the grammar model fails to predict true concave-down distributions along both constraints in the cumulative interaction. See Appendix A of this dissertation for a small proof showing that the differences in Equations 2.12 and 2.14 above can never both be negative (or, concave-down) for the same $w_{\mathbb{C}1}$ and $w_{\mathbb{C}2}$.

2.5 Multiple competitions MaxEnt grammars

2.5.1 Counting cumulativity

Recall that in multiple competitions models each input is assigned its own probabilility distribution over its candidates, therefore the model assigns multiple probability distributions and candidates in different competitions are normalized by different constants. And, the constraints in the grammar are in an *asymmetric trade-off*: zero or more violations of a given constraint trade off against a single violation of another constraint. This asymmetric trade-off is the crucial structure that predicts categorical gang effects in Standard HG and its variants (Pater, 2009), and it is often implemented in modeling both phonotactic and alternation patterns that show and don't show evidence of cumulativity (Kaplan, 2018; Shih, 2016, 2017; Zuraw and Hayes, 2017; Kawahara, 2021; Breiss and Albright, 2022; Hayes, 2022; Kim, 2022; Magri, 2025).

I now evaluate the range of predicted kinds of constraint interactions in multiple competitions MaxEnt models. Assume the minimal grammar structure in Figure 2.20 below. As with the single competitions model for counting cumulativity, the input candidates show zero or more violations of the constraint (Variable), but now each candidate competes against an opposing candidate that violates OnOff only once at every competition in the grammar. Since the candidates with variable violations are now in separate competitions, they form part of different probability distributions and are normalized by different constants.

Figure 2.20: Basic structure of the multiple competitions model in counting cumulativity.

		OnOff	Variable	H	predicted probability
		w_{oo}	w_v		
i_0	a. c_0		0	0	$1 \ / \ e^{-w_{oo}}$
	b. c_{oo}	1		$-w_{oo}$	
i_1	a. c_1		1	$-w_v$	$e^{-w_v} \ / \ e^{-w_v} + e^{-w_{oo}}$
	b. c_{oo}	1		$-w_{oo}$	
i_2	a. c_2		2	$-2w_v$	$e^{-2w_v} \ / \ e^{-2w_v} + e^{-w_{oo}}$
	b. c_{oo}	1		$-w_{oo}$	
i_3	a. c_3		3	$-3w_v$	$e^{-3w_v} \ / \ e^{-3w_v} + e^{-w_{oo}}$
	b. c_{oo}	1		$-w_{oo}$	

Recall that the concavity of constraint interaction for patterns of counting cumulativity is assessed by measuring the penalty introduced by n violations of the same constraint compared to the penalty introduced by $n+1$ violations. Specifically, a probability distribution is concave-up if the penalty introduced by the n^{th} violation is larger than the penalty introduced by the $n+1^{\text{st}}$, and concave-down if the penalty introduced by the n^{th} violation is smaller than the penalty introduced by the $n+1^{\text{st}}$. Equations 2.15 and 2.16 below show the definition of concavity for multiple competitions MaxEnt models with counting cumulativity. At any $w_v > 0$, and $w_{oo} > 0$, the grammar's predicted probability distributions are concave-up if the differences below are positive across all $n \geq 1$, and concave-down if the differences below are negative for at least one $n \geq 1$.

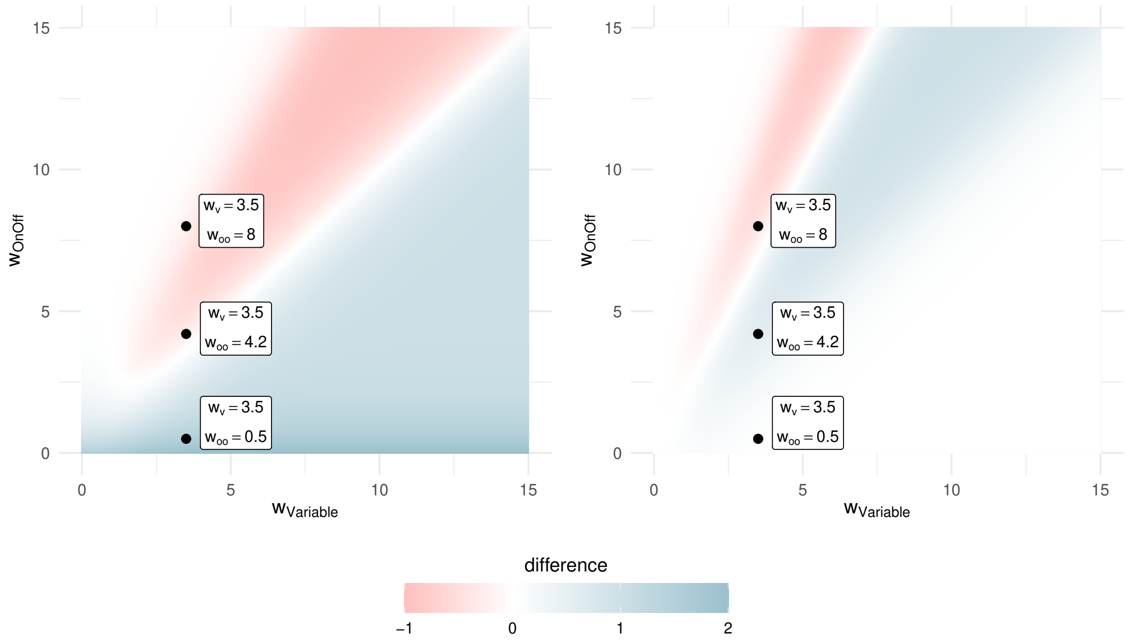
$$[P(c_{n-1}) - P(c_n)] - [P(c_n) - P(c_{n+1})] \tag{2.15}$$

$$\left(\frac{e^{-w_v(n-1)}}{e^{-w_v(n-1)} + e^{-w_{oo}}} - \frac{e^{-w_v n}}{e^{-w_v n} + e^{-w_{oo}}}\right) - \left(\frac{e^{-w_v n}}{e^{-w_v n} + e^{-w_{oo}}} - \frac{e^{-w_v(n+1)}}{e^{-w_v(n+1)} + e^{-w_{oo}}}\right) \tag{2.16}$$

Figure 2.21 below plots the differences in Equation 2.16 above at violation levels $n = 1$ and $n = 2$ and at varying weights weights of OnOff and Variable.[19]

[19] Recall that $n = 1$ compares the probabilities of candidates with zero, one, and two violations of Variable, and $n = 2$ compares the probabilities of candidates with one, two, and three violations of Variable.

Figure 2.21: Predicted concave-up and concave-down counting cumulativity of multiple competitions MaxEnt at $n = 1$ and $n = 2$ violation levels. Differences are plotted in probability space, following Equation 2.16.



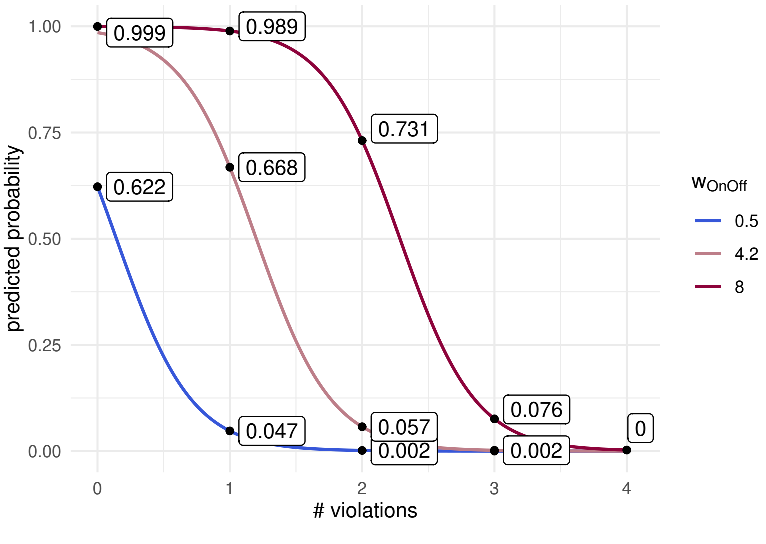
(a) Difference in Equation (2.16) at $n = 1$.

(b) Difference in Equation (2.16) at $n = 2$.

The Figure above shows that multiple competitions models predict both concave-up and concave-down cumulativity: the difference in penalty incurred by the n^{th} violation and the $n + 1^{\text{st}}$ violation can be positive (concave-up) or negative (concave-down) across both $n = 1$ and $n = 2$ violation levels. For example, at $w_v = 3.5$ and $w_{oo} = 0.5$, the penalty incurred by the first violation is larger than the penalty incurred by the second (Figure 2.21a), therefore the predicted distribution is concave-up. The predicted distribution at these constraint weights is also weakly concave-up in the penalties introduced by the second and third violations of Variable (Figure 2.21b). And, notice that various concave-down patterns are predicted at higher weights of OnOff. For example, at $w_v = 3.5$ and $w_{oo} = 4.2$, the predicted distribution is concave-down in the penalty introduced by the first vs. second violation (Figure 2.21a), but concave-up in the penalty introduced by the second vs. third violation (Figure 2.21b). Finally, at $w_v = 3.5$ and $w_{oo} = 8$, the penalties among the first, second, and third violations are all concave-down, as seen across both subfigures.

The model's predicted range of concave-up and concave-down distributions is more visually identifiable when plotting its predicted probabilities for the candidates with varying violations of Variable. Figure 2.22 below plots the predicted distributions at our three example weights of Variable and OnOff from the Figure above. Notice that these curves show the same patterns of concavity as previously described: with the weight of Variable held constaint at 3.5, at $w_{oo} = 0.5$ the distribution is concave-up at both $n = 1$ and $n = 2$ violations, at $w_{oo} = 4.2$ the distribution is concave-down at $n = 1$ but concave-up at $n = 2$, and at $w_{oo} = 8$ the distributions are concave-down at both $n = 1$ and $n = 2$.

Figure 2.22: Predicted probabilities at different weights of OnOff with weight of Variable held constant at 3.5.



We can think of the effect of manipulating the weight of OnOff as shifting MaxEnt's sigmoid curve along the x-axis, creating identical curves shifted apart along the x-axis. This is exactly the manipulation that creates MaxEnt's "wug-shaped curves" in previous literature

(Zuraw and Hayes, 2017; Kawahara, 2020; Hayes, 2022; Magri, 2025).

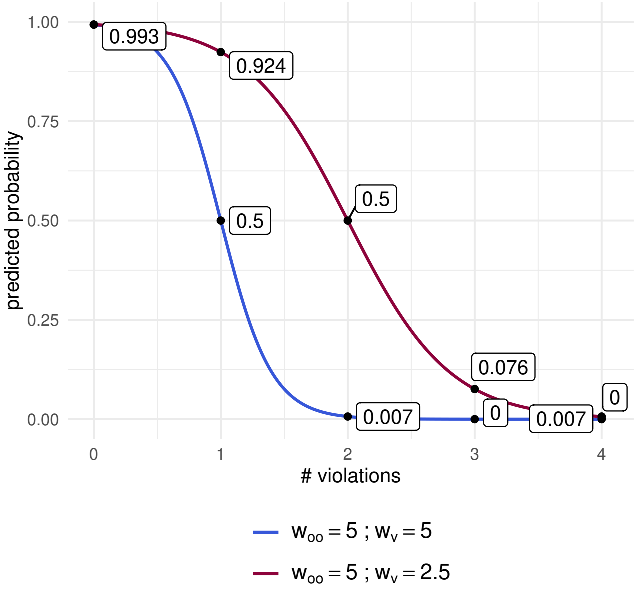
Additionally, the grids in Figure 2.21 above show that the type of concavity of the grammar's predicted probabilities is a function of the ratio of the constraint weights. Specifically, the predicted distributions switch concavity when the weights of OnOff and Variable are at an $n : 1$ ratio. At $n = 1$ (Figure 2.21a above), the predicted distribution changes concavity at a $1 : 1$ ratio of the constraint weights: at $w_{oo} = w_v$, the first and second violations of Variable introduce the same probability decrease, hence their difference is zero. Additionally, at $n = 2$ (Figure 2.21b), the switch in concavity occurs at a $2 : 1$ ratio between the weights of OnOff and Variable. So, at $2w_{oo} = w_{var}$, the penalty introduced by the second violation of Variable is the same as the penalty introduced by its third violation, hence the difference in probability decrease introduced by these violations is zero.[20] The ratio of the constraint weights also determines the violation level at which the two candidates in the competition have equal probability. At $w_{oo} = w_v$, the candidate with one violation of Variable has the same probability as its opposing candidate (0.5), and at $2w_{oo} = w_v$, the candidate with 2 violations of Variable and its opposing candidate have the same probability. Therefore, when c_n candidates have a probability of 0.5, the distribution is neither concave-up or concave down in comparing the effect of the n^{th} and $n + 1^{\text{st}}$ violations.

Figure 2.23 below more clearly shows the relationship between the ratio of the constraint weights and the concavity of the model's predicted probability distributions. The inflection point of the predicted probabilities at $w_{oo} = 5$ and $w_v = 5$ occurs at the first violation since the constraint weights are at a $1 : 1$ ratio, and the inflection point of the predicted probabilities at $w_{oo} = 5$ and $w_v = 2.5$ occurs at the second violation since the constraint weights are at a $2 : 1$ ratio. The inflection point of the curve defines the point at which the predicted distributions switch from concave-down to concave-up as violations of Variable accumulate. Also notice that the inflection point occurs at probability 0.5, which is when

[20] Although not shown in Figure 2.21, this $n : 1$ ratio between the weights of OnOff and Variable at determining the type of concavity generalizes to larger values of n.

c_n candidates with accumulating violations of Variable have the same probability as their opposing candidate.[21]

Figure 2.23: Example constraint weights at a $1:1$ and $2:1$ ratio between the weight of OnOff and the weight of Variable.

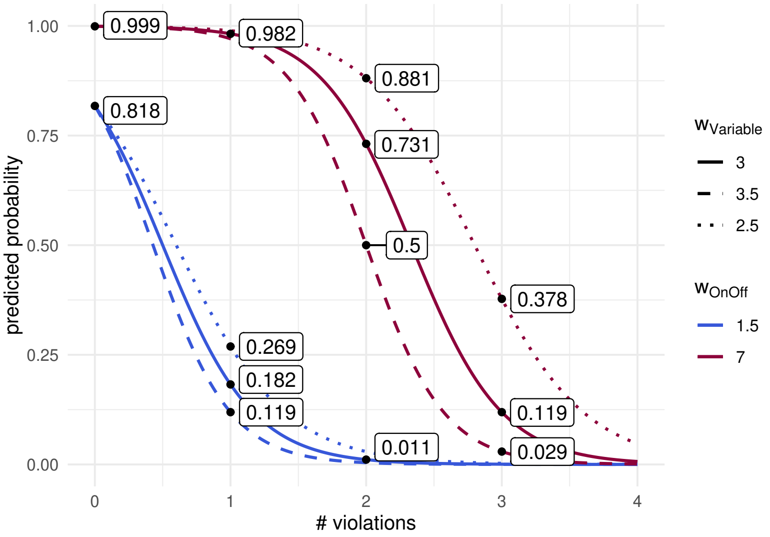


Also notice from Figure 2.23 above that the steepness of the curve is a function of the weight of the variably violated constraint. In this Figure, the curve with the higher weight of Variable (blue) falls faster than the curve with the lower weight of Variable (red). In general, multiple competitions MaxEnt predicts stronger concave-up or concave-down distributions at higher weights of Variable and weaker concavities at lower weights of Variable. This is more clearly shown in Figure 2.24 below. For example, the first violation of Variable incurs the largest penalty at $w_v = 3.5$ and it incurs the smallest penalty at $w_v = 2.5$ for the concave-up distribution predicted at $w_{oo} = 1.5$. This also holds for the

[21] Although in Figure 2.23 below Variable has the same contribution to probability decrease at the second and third violation of the constraint, this distribution is still concave-down: the second violation of Variable introduces a probability decrease that is larger than the probability decrease at the first violation of Variable.

second and third violations of the concave-down distributions predicted at $w_{oo} = 7$, where the both second and third violations very clearly incur the largest probability decrease at $w_v = 3.5$. Again, notice that the predicted distribution at $w_v = 7, w_{oo} = 3.5$ is 0.5 at the second violation since the constraint weights in this predicted distribution are at an $n : 1$ ratio.

Figure 2.24: Predicted probabilities at various weights of OnOff and Variable. The steepness of the curve is a function of the weight of Variable.



In summary, multiple competitions MaxEnt models with an asymmetric trade-off predict both concave-up and concave-down constraint interactions for accumulating violations of the same constraint. At specific weighing conditions, multiple competitions models predict that a violation could have a stronger (concave-down) or weaker (concave-up) effect in the environment of other violations of the same constraint. This stronger or weaker effect (or, the type of concavity) is a function of the ratio of the constraint weights: in general, an $n : 1$ ratio between the weights of OnOff and Variable the predicted distributions shift from concave-up to concave-down at the nth violation of Variable. Furthermore, the strength of the concavity of the predicted distribution is a function of the weight of Variable, such that, at higher weights of Variable, the probability differences across candidates with varying violations of Variable are bigger and therefore more strongly concave-up or concave-down.

2.5.2 Ganging cumulativity

Lastly, I now evaluate the predictions of multiple competitions MaxEnt models for accumulating violations of different constraints. We assume the simple grammar in Figure 2.25 below. As with our assessment of the predicted concavities of single competition MaxEnt models with ganging cumulativity, we compare candidates with different combinations of violations of ℂ1 and ℂ2, but now each input candidate competes against an opposing candidate with one violation of OnOff. And, ℂ1 and ℂ2 together trade-off asymmetrically against OnOff, therefore candidates with violations of these constraints may have a higher or lower probability than the opposing candidate at certain weighing conditions, parallel to the effect of asymmetric trade-offs in Standard HG (Pater, 2009).

Figure 2.25: Schematic multiple competitions MaxEnt grammar for ganging cumulativity – accumulating violations of different constraints.

		OnOff w_{oo}	C1 $w_{\mathbb{C}1}$	C2 $w_{\mathbb{C}2}$	Harmony	predicted probability
$i_{(0,0)}$	a. $c_{(0,0)}$		0	0	0	$1 \;/\; (1+e^{-w_{oo}})$
	b. c_{oo}	1			$-w_{oo}$	
$i_{(1,0)}$	a. $c_{(1,0)}$		1	0	$-w_{\mathbb{C}1}$	$e^{-w_{\mathbb{C}1}} \;/\; (e^{-w_{\mathbb{C}1}}+e^{-w_{oo}})$
	b. c_{oo}	1			$-w_{oo}$	
$i_{(0,1)}$	a. $c_{(0,1)}$		0	1	$-w_{\mathbb{C}2}$	$e^{-w_{\mathbb{C}2}} \;/\; (e^{-w_{\mathbb{C}2}}+e^{-w_{oo}})$
	b. c_{oo}	1			$-w_{oo}$	
$i_{(1,1)}$	a. $c_{(1,1)}$		1	1	$-w_{\mathbb{C}1}-w_{\mathbb{C}2}$	$e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}} \;/\; (e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}}+e^{-w_{oo}})$
	b. c_{oo}	1			$-w_{oo}$	

We assess the concavity of this model's predicted probability distributions by evaluating each constraint's effect on probability across cumulative and non-cumulative contexts. Following our definition of concavity, this model's predicted distributions are concave-up if the differences in Equations 2.17 to 2.20 are positive for both constraints, of mixed concavity of one of the differences is positive and the other is negative, and concave-down if both differences along both constraints are negative.

Concavity along $\mathbb{C}1$:

$$(P(c_{(0,0)}) - P(c_{(1,0)})) - (P(c_{(1,0)}) - P(c_{(1,1)})) \tag{2.17}$$

$$\left(\frac{1}{1+e^{-w_{oo}}} - \frac{e^{-w_{\mathbb{C}1}}}{e^{-w_{\mathbb{C}1}} + e^{-w_{oo}}}\right) - \left(\frac{e^{-w_{\mathbb{C}1}}}{e^{-w_{\mathbb{C}1}} + e^{-w_{oo}}} - \frac{e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}}}{e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}} + e^{-w_{oo}}}\right) \tag{2.18}$$

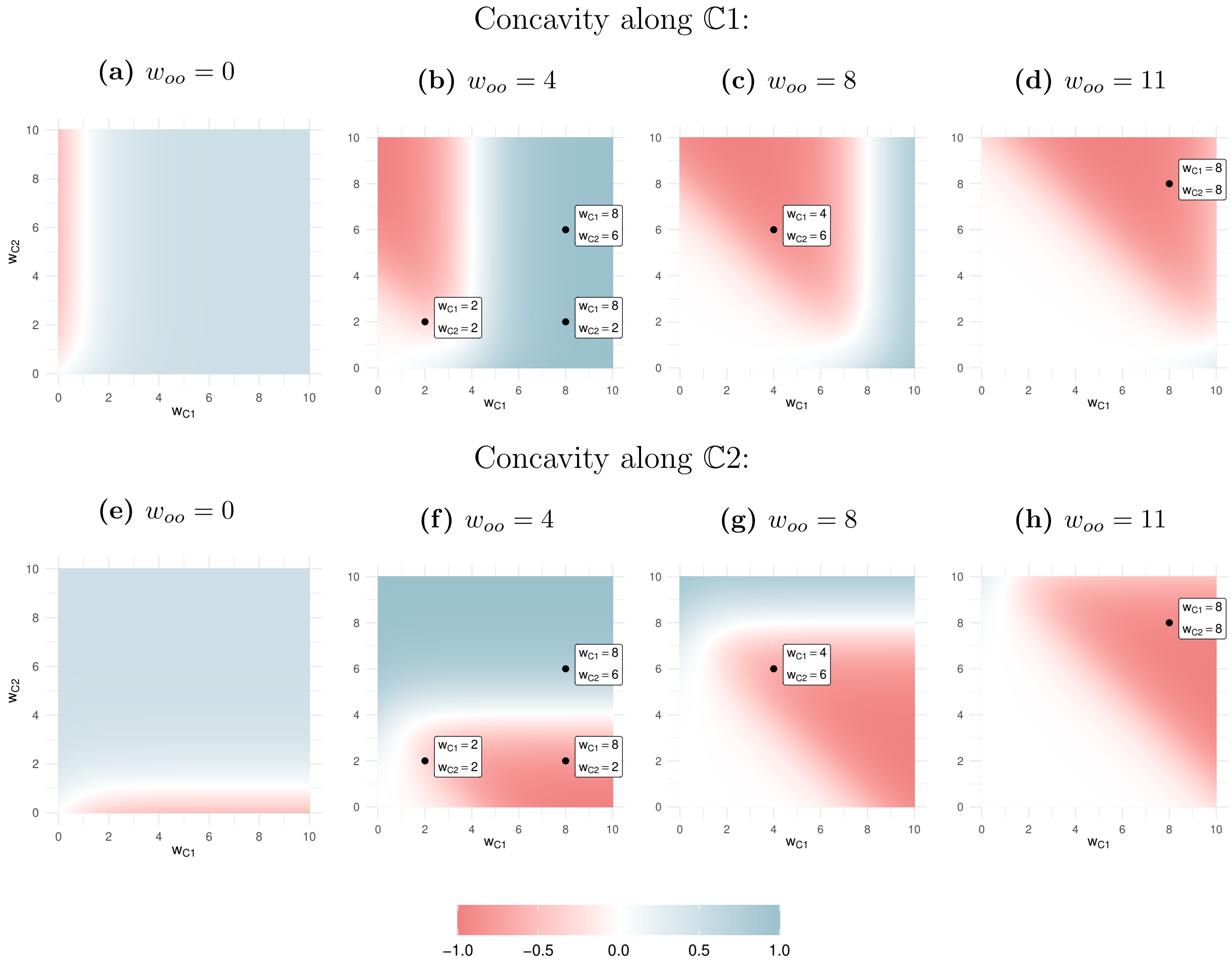
Concavity along $\mathbb{C}2$:

$$(P(c_{(0,0)}) - P(c_{(0,1)})) - (P(c_{(0,1)}) - P(c_{(1,1)})) \tag{2.19}$$

$$\left(\frac{1}{1+e^{-w_{oo}}} - \frac{e^{-w_{\mathbb{C}2}}}{e^{-w_{\mathbb{C}2}} + e^{-w_{oo}}}\right) - \left(\frac{e^{-w_{\mathbb{C}2}}}{e^{-w_{\mathbb{C}2}} + e^{-w_{oo}}} - \frac{e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}}}{e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}} + e^{-w_{oo}}}\right) \tag{2.20}$$

Figure 2.26 below plots these differences at varying constraint weights for $\mathbb{C}1$ and C2, and at select weights for OnOff.

Figure 2.26: Concavity at different constraint weights. This is probability space.



Concavity along $\mathbb{C}1$:

(a) $w_{oo} = 0$

(b) $w_{oo} = 4$

(c) $w_{oo} = 8$

(d) $w_{oo} = 11$

Concavity along $\mathbb{C}2$:

(e) $w_{oo} = 0$

(f) $w_{oo} = 4$

(g) $w_{oo} = 8$

(h) $w_{oo} = 11$

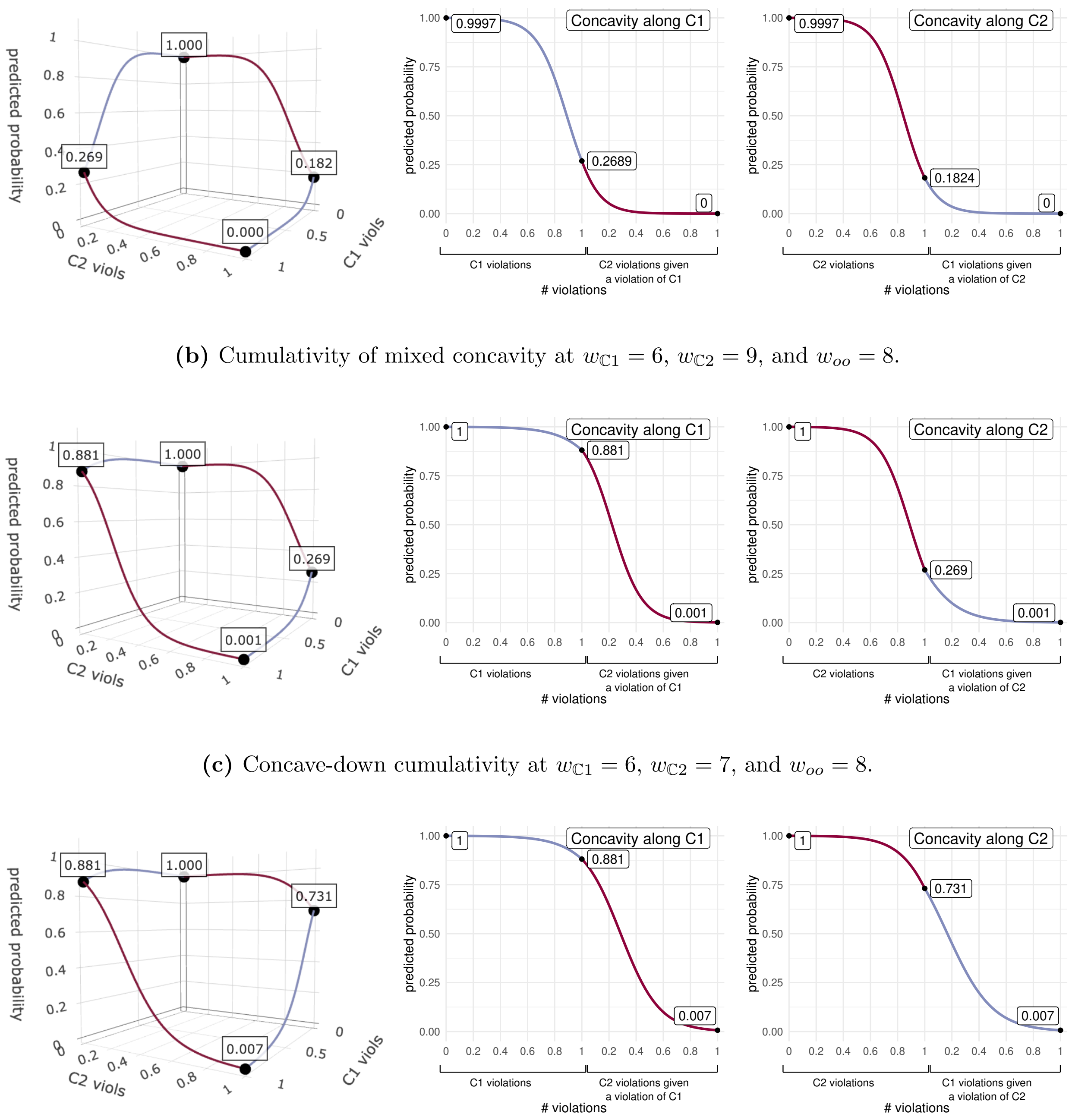
First, at $w_{oo} = 0$, the predicted probabilities of multiple competitions MaxEnt exactly match those predicted by single competition MaxEnt models: both model structures predict either concave-up cumulativity or cumulativity of mixed concavity, but they fail to predict concave-down patterns along both constraints. In other words, subplots 2.26a and 2.26e are the same as Figure 2.18 of the previous section. However, as the weight of OnOff increases, the model predicts more overlapping concave-down patterns along both constraints. At intermediate weights of OnOff, multiple competitions MaxEnt predicts all kinds of constraint interactions. For example, at $w_{oo} = 4$ (Figures 2.26b and 2.26f), the model predicts concave-up distributions at $w_{\mathbb{C}1} = 8$ and $w_{\mathbb{C}2} = 6$, distributions of mixed concavity at $w_{\mathbb{C}1} = 4$ and $w_{\mathbb{C}2} = 6$, and concave-down distributions at $w_{\mathbb{C}1} = 8$ and $w_{\mathbb{C}2} = 8$. At $w_{oo} = 8$ the model predicts exclusively concave-down distributions across the weights of $\mathbb{C}1$ and $\mathbb{C}2$ plotted.

Additionally, as with multiple competitions MaxEnt grammars for counting cumulativity, the concavity of the predicted distributions is a function of the relative weights of $\mathbb{C}1$, $\mathbb{C}2$, and OnOff. Specifically, the pattern changes from concave-down to concave-up along a particular constraint when the weight of such constraint surpasses the weight of OnOff. For example, Figures 2.26b to 2.26d show that, once the weight of $\mathbb{C}1$ surpasses the weight of OnOff, the predicted patterns switch from concave-down along $\mathbb{C}1$ to concave-up along $\mathbb{C}1$. The same holds in Figures 2.26f to 2.26h for $\mathbb{C}2$: once the weight of $\mathbb{C}2$ surpasses the weight of OnOff, the predicted distributions switch from concave-down to concave-up along $\mathbb{C}2$. Therefore, concave-down probability distributions are predicted when the weights of $\mathbb{C}1$ and $\mathbb{C}2$ are both less than the weight of OnOff, and concave-up distributions when the weights of $\mathbb{C}1$ and $\mathbb{C}2$ are both greater than the weight of OnOff.

Lastly, notice that at $w_{oo} = w_{\mathbb{C}1} + w_{\mathbb{C}2}$ the predicted distributions are concave-down along both constraints in the interaction when $w_{oo} > 0$. This reflects the non-linear transformation from harmony to probability in MaxEnt: at these constraint weights the doubly-violating candidate and its opposing candidate have equal harmony, but the model's predicted probability distributions at these constraints weights are concave-down, therefore non-linear.

Figure 2.27 below shows example concave-up, concave-down, and cumulativity of mixed concavity at selects weights of $\mathbb{C}1$, $\mathbb{C}2$, and OnOff. Notice that these distributions are consistent with the observations discussed above: concave-up distributions are predicted when the weights of $\mathbb{C}1$ and $\mathbb{C}2$ are both greater than the weight of OnOff (Figure 2.27a), distributions of mixed concavity are predicted when only one of the weights is greater than the weight of OnOff (Figure 2.27b), and concave-down distributions are predicted when both weights are lower than the weight of OnOff (Figure 2.27c).

Figure 2.27: Concave-up, mixed concavity, and concave-down cumulativity at various weights of $\mathbb{C}1$ and $\mathbb{C}2$, with the weight of OnOff held constant at 8.



(a) Concave-up cumulativity at $w_{\mathbb{C}1} = 9$, $w_{\mathbb{C}2} = 9.5$, and $w_{oo} = 8$.

(b) Cumulativity of mixed concavity at $w_{\mathbb{C}1} = 6$, $w_{\mathbb{C}2} = 9$, and $w_{oo} = 8$.

(c) Concave-down cumulativity at $w_{\mathbb{C}1} = 6$, $w_{\mathbb{C}2} = 7$, and $w_{oo} = 8$.

2.6 Discussion and conclusion

This chapter described the diverging modes of constraint interaction in two types of Maximum Entropy (MaxEnt) model structures used in the modeling and learning of both phonotactic and alternation patterns. I showed that single competition MaxEnt grammars, which are models that assign a single probability distribution across all candidate forms in the grammar (Hayes and Wilson, 2008), can only predict that constraint violations have an increasingly weaker effect on the probability of a multiply-violating candidate, which I call *concave-up* and *mixed concavity* distributions in this work. On the other hand, I showed that multiple competitions MaxEnt grammars predict a wider range of constraint interactions, including patterns of cumulativity where two constraint violations can have a very strong effect on the probability of the multiply-violating forms even when these constraint violations introduce weak penalties independently. These findings are true for both accumulating violations of the same constraint (counting cumulativity) and accumulating violations of different contraints (ganging cumulativity). This chapter additionally argued that constraint interaction should be evaluated by assessing the penalty constraint violations introduce across cumulative vs. non-cumulative contexts, as opposed to evaluating constraint interaction by comparing observed distributions against distributions that are "expected" under some definition. The definition of linearity has varied substantially in previous literature and faces logical problems, therefore not allowing for principled and controlled comparisons across empirical data and different theories of phonological grammar.

The primary contribution of this chapter is that it gives us concrete, quantitative, and testable predictions about what the appropriate theory of phonotactics and phonotactic learning should be. Both single and multiple competitions MaxEnt models are particularly successful at capturing a wide range of phonological patterns with and without cumulative constraint interactions, but it remains unclear which model structure best captures the range of constraint interactions and speakers' intuitions of such patterns. The following chapters of this dissertation address these questions empirically via corpus studies and behavioral experiments with infants and adults, as well as via further computational modeling of these two MaxEnt grammar grammar structures.

The next chapter of this dissertation reports the findings of a series of artificial grammar learning studies in which adult participants learn toy languages with concave-up and concave-down patterns of ganging cumulativity, following the experiment designs in Breiss (2020) and Breiss and Albright (2022). In the artificial grammars with concave-down cumulativity, the frequency of multiply-violating forms is very low compared to the frequency of singly-violating forms, therefore we test if participants are able to detect that a constraint's contribution to penalty is strong only when it occurs along other marked sound patterns and not when it is violated independently. This is with the overall goal of understanding if participant intuitions support a model of phonotactic learning wherein distributional patterns are represented and learned more globally across the language (single competitions models), or a model in which distributional phonotactic knowledge is represented on a binary accept vs. reject basis (multiple competitions models). Subsequent chapters of this dissertation investigate if phonotactic learning is possible under multiple competitions MaxEnt model structures. The opposing candidate, which is considered the null candidate when using multiple competitions MaxEnt grammars in modeling phonotactics, is unobservable from the surface distribution of sound patterns in the data, but by assuming that any input candidate is possible (namely, an unrestricted Gen), the null candidate is assigned a frequency and therefore phonotactic patterns under this model are learnable.

References

Albright, Adam. 2008. Cumulative violations and complexity thresholds. Ms., MIT.

Albright, Adam. 2012. Additive markedness interactions in phonology. Colloquium talk at UCLA.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam.*

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32:45–86.

Boersma, Paul, and Joe Pater. 2016. Convergence properties of a gradual learning algorithm for harmonic grammar. In *Harmonic Grammar and Harmonic Serialism*, ed. John McCarthy and Joe Pater. Equinox Press.

Breiss, Canaan. 2020. Constraint cumulativity in phonotactics: evidence from artificial grammar learning. *Phonology* 37:551–576.

Breiss, Canaan, and Adam Albright. 2022. Cumulative markedness effects and (non-)linearity in phonotactics. *Glossa: a journal of general linguistics* 7:1–32.

Crowhurst, Megan J. 2011. Constraint conjunction. In *The Blackwell Companion to Phonology*, ed. Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume, and Keren Rice. John Wiley & Sons.

Durvasula, Kathik, and Adam Liter. 2020. There is a simplicity bias when generalizing fro ambiguous data. *Phonology* 37:177–213.

Farris-Trimble, Ashley W. 2008. Cumulative faithfulness effects in phonology. Doctoral Dissertation, Indiana University.

Farris-Trimble, Ashley W. 2010. Nothing is better than being unfaithful in multiple ways. In *Proceedings from the Annual Meeting of the Chicago Linguistics Society*, volume 44, 79–93. Chicago Linguistics Society.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory*, ed. Jennifer Spenader, Anders Eriksson, and Osten Dahl, 111–120. Stockholm University.

Green, Christopher R., and Stuart Davis. 2014. Superadditivity and limitations on syllable complexity in Bambara words. In *Perspectives on Phonological Theory and Acquisition: papers in honor of Daniel A. Dinnsen*, ed. Ashley W. Farris-Trimble and Jessica A. Barlow, 223–247. Benjamins.

Hayes, Bruce. 2022. Deriving the wug-shaped curve: A criterion for assessing formal theories of linguistic variation. *Annual Review of Linguistics* 8:473–494.

Hayes, Bruce, and Aaron Kaplan. 2025. Zero-weighted constraints in noisy harmonic grammar. *Linguistic Inquiry* 56.

Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistics Inquiry* 39:379–440.

Itô, Junko, and Armin Mester. 1998. Markedness and word structure: OCP effects in Japanese. Ms., University of California, Santa Cruz.

Itô, Junko, and Armin Mester. 2003. *Japanese morphophonemics: markedness and word structure*. MIT Press.

Jager, Gerhard, and Anette Rosenbach. 2006. The winner takes it all - almost: Cumulativity in grammatical variation. *Linguistics* 44:937–971.

Jurafsky, Daniel, and James H. Martin. 2019. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Available online at: https://web.stanford.edu/~jurafsky/slp3/.

Kaplan, Aaron. 2018. Positional licensing, asymmetric trade-offs and gradient constraints in harmonic grammar. *Phonology* 35:247–286.

Kawahara, Shigeto. 2020. A wug-shaped curve in sound symbolisms: the case of Japanese Pokémon names. *Phonology* 37(3):383–418.

Kawahara, Shigeto. 2021. Testing MaxEnt with sound symbolisms: a stripy wug-shaped curve in Japanese Pokémon names. *Language* 97:e341–e359.

Keller, Frank. 2000. Gradience in grammar: experimental and computational aspects of degrees of grammaticality. Doctoral Dissertation, University of Edinburgh.

Keller, Frank. 2006. Linear ot as a model of gradience in grammar. In *Gradience in Grammar: Generative Perspectives*, ed. Gisbert Fanselow, Caroline Féry, Ralph Vogel, and Matthias Schlesewsky, 270—-288.

Kim, Seoyoung. 2022. A MaxEnt learner for super-additive counting cumulativity. *Glossa: a journal of general linguistics* 7:1–34.

de Lacy, Paul. 2002. The formal expression of markedness. Doctoral Dissertation, University of Massachusetts Amherst.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1990. Harmonic Grammar: a formal multi-level connectionist theory of linguistic well-formedness: an application. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, 884–891. Erlbaum.

Łubowicz, Anna. 2005. Locality łf conjunction. In *Proceedings of the 24th West Coast Conference on Formal Linguistics*, 254–262. Cascadilla Press.

Magri, Giorgio. 2025. Constraint interaction in probabilistic phonology: deducing MaxEnt from Hayes and Zuraw's shifted sigmoids generalization. *Linguistic Inquiry* URL https://doi.org/10.1162/LING.a.66.

McCarthy, John J., and Matthew Wolf. 2005. Less than zero: correspondence and the null output. *Modeling ungrammaticality in Optimality Theory* 22.

Milenković, Aljoša. to appear. Superadditive cumulativity in categorical prosodic patterns: prosodic minimality in Bosnian/Croatian/Montenegrin/Serbian. *Phonology* .

Moreton, Elliot. 2008. Analytic bias and phonological typology. *Phonology* 25(1):83–127.

Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33:999–1035.

Pater, Joe, and Elliot Moreton. 2012. Structurally biased phonology: complexity in learning and typology. *Journal of English and Foreign Languages University, Hyderabad* 3(2):1–44.

Pizzo, Presley. 2015. Investigating properties of phonotactic knowledge through web-based experimentation. Doctoral Dissertation, University of Massachusetts Amherst.

Prince, Alan, and Paul Smolensky. 1993. Optimality Theory: constraint interaction in generative grammar. ROA.

Shih, Stephanie. 2017. Constraint conjunction in weighted probabilistic grammar. *Phonology* 34:243–268.

Shih, Stephanie S. 2016. Super additive similarity in Dioula tone harmony. In *Proceedings of the 33rd West Coast Conference on Formal Linguistics*, 361–370.

Smith, Brian, and Joe Pater. 2020. French schwa and gradient cumulativity. *Glossa: a journal of general linguistics* 5:1–33.

Smolensky, Paul. 2003. Harmony, markedness, and phonological activity. Paper presented at Rutgers Optimality Workshop. Available as ROA-87 from the Rutgers Optimality Archive.

Smolensky, Paul. 2006. Optimality in phonology ii: harmonic completeness, local constraint conjunction, and feature domain markedness. In *The harmonic mind: from neural computation to optimality theoretic grammar*, ed. Paul Smolensky and Géraldine Legendre, 27–160. MIT Press.

Steriade, Donca. 2001. The phonology of perceptibility effects: the P-Map and its consequences for constraint organization. Ms., UCLA/MIT.

White, Jamie. 2014. Evidence for a learning bias against saltatory phonological alternations. *Cognition* 130(1):95–115.

Wilson, Colin. 2006. Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science* 30:945–982.

Yang, Shiying, Chelsea Sanker, and Uriel Cohen Priva. 2018. The organization of lexicons: a cross-linguistic analysis of monosyllabic words. In *Proceedings of the Society for Computation in Linguistics*, 164–173.

Zuraw, Kie, and Bruce Hayes. 2017. Intersecting constraint families: an argument for Harminic Grammar. *Language* 93:497–548.

Appendix A

To more concretely show that single competitions models do not predict concave-down ganging cumulativity, we want to show that, for any non-negative weights of constraints, the probability differences along the two constraints can never be negative simultaneously. Below is the definition of concavity for single competition ganging cumulativity along the two constraints.

Concavity along $\mathbb{C}1$:

$$[P(c_{(0,0)}) - P(c_{(1,0)})] - [P(c_{(1,0)}) - P(c_{(1,1)})] \tag{2.21}$$

$$(1 - e^{-w_{\mathbb{C}1}}) - (e^{-w_{\mathbb{C}1}} - e^{-w_{\mathbb{C}1} - w_{\mathbb{C}2}}) \tag{2.22}$$

Concavity along $\mathbb{C}2$:

$$[P(c_{(0,0)}) - P(c_{(0,1)})] - [P(c_{(0,1)}) - P(c_{(1,1)})] \tag{2.23}$$

$$(1 - e^{-w_{\mathbb{C}2}}) - (e^{-w_{\mathbb{C}2}} - e^{-w_{\mathbb{C}1} - w_{\mathbb{C}2}}) \tag{2.24}$$

We can use a simple algebraic fact to show that neither of these differences can both be negative for the same $w_{\mathbb{C}1}$ and $w_{\mathbb{C}2}$: if $x > 0$ and $y > 0$, then $x + y > 0$. The contrapositive of this claim is also true: if $x + y > 0$, then it's impossible that both x and y are negative. So, if we show that the sum of these two differences is positive, then it must be the case that the differences are not concave-down for both constraints.

We add these differences and simplify the expression as follows:

$$[(1 - e^{-w_{\mathbb{C}1}}) - (e^{-w_{\mathbb{C}1}} - e^{-w_{\mathbb{C}1} - w_{\mathbb{C}2}})] + [(1 - e^{-w_{\mathbb{C}2}}) - (e^{-w_{\mathbb{C}2}} - e^{-w_{\mathbb{C}1} - w_{\mathbb{C}2}})] \tag{2.25}$$

$$(1 - e^{-w_{\mathbb{C}1}} - e^{-w_{\mathbb{C}1}} + e^{-w_{\mathbb{C}1} - w_{\mathbb{C}2}}) + (1 - e^{-w_{\mathbb{C}2}} - e^{-w_{\mathbb{C}2}} + e^{-w_{\mathbb{C}1} - w_{\mathbb{C}2}}) \tag{2.26}$$

$$(1 - 2e^{-w_{\mathbb{C}1}} + e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}}) + (1 - 2e^{-w_{\mathbb{C}2}} + e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}}) \tag{2.27}$$

$$2 - 2e^{-w_{\mathbb{C}1}} - 2e^{-w_{\mathbb{C}2}} + 2e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}} \tag{2.28}$$

$$2(1 - e^{-w_{\mathbb{C}1}} - e^{-w_{\mathbb{C}2}} + e^{-(w_{\mathbb{C}1}+w_{\mathbb{C}2})}) \tag{2.29}$$

$$2(1 - e^{-w_{\mathbb{C}1}})(1 - e^{-w_{\mathbb{C}2}}) \tag{2.30}$$

So, if we show that $2(1 - e^{-w_{\mathbb{C}1}})(1 - e^{-w_{\mathbb{C}2}})$ is always positive for any value of $w_{\mathbb{C}1}$ and $w_{\mathbb{C}2}$ greater than zero, then it cannot be the case that the differences are both negative (or, concave-down) for the same values of $w_{\mathbb{C}1}$ and $w_{\mathbb{C}2}$.

In order for this expression to be positive, $e^{-w_{\mathbb{C}1}}$ and $e^{-w_{\mathbb{C}2}}$ must be greater than zero and less than 1. First, e^x is always positive for any value of x (either positive or negative), so it must be the case that $e^{-w_{\mathbb{C}1}}$ and $e^{-w_{\mathbb{C}2}}$ are greater than zero. And, we know that $e^0 = 1$, and since the exponential function e^x is always increasing, then it must be the case that, for any $w_{\mathbb{C}1}$ and $w_{\mathbb{C}2}$, $e^{-w_{\mathbb{C}1}} < 1$ and $e^{-w_{\mathbb{C}2}} < 1$.

Therefore, the differences $(1 - e^{-w_{\mathbb{C}1}}) - (e^{-w_{\mathbb{C}1}} - e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}})$ and $(1 - e^{-w_{\mathbb{C}2}}) - (e^{-w_{\mathbb{C}2}} - e^{-w_{\mathbb{C}1}-w_{\mathbb{C}2}})$ cannot both be negative simultaneously, and we proved this by showing that their sum is always positive. So, single competitions MaxEnt cannot predict concave-down patterns for ganging cumulativity: it can be the case that one of the differences along one of the constraints is negative, but both differences cannot be negative simultaneously.