How many maximum entropy grammars are predicted by a constraint set when we ignore small differences among grammars?

Giorgio Magri
CNRS, SFL, University of Paris 8 / 59 rue Pouchet, 75005 Paris, France
magrigrg@gmail.con

Abstract

All constraint-based probabilistic phonological typologies considered in the recent literature consist of uncountably many different grammars. Yet, what if two grammars that differ only slightly are coarsely counted as only one grammar when assessing the finiteness of a probabilistic typology? This paper formalizes various notions of coarse identity between probabilistic grammars and corresponding notions of coarse finiteness. It then shows that typologies of maximum entropy grammars are stubbornly infinite even when their grammars are counted coarsely (and even when the constraint set is simple, in the sense that the corresponding categorical harmonic grammar typology is finite). A companion paper shows that typologies of noisy or stochastic harmonic grammars are instead always coarsely finite (as long as the constraint set is simple). Coarse finiteness thus provides further evidence that maximum entropy is a richer, less restrictive framework.

1 Introduction

Probabilistic phonological grammars assign probabilities to phonological mappings. Probabilities take continuous values between zero and one. Hence, a probabilistic typology can contain uncountably many probabilistic grammars when we count two grammars as different in the standard sense, namely as soon as they assign different probabilities to some mapping, no matter how small the difference between those probabilities.

What if we instead tolerate some differences among probabilities as negligible? What if we count grammars **coarsely** because we count two different grammars that have only negligible differences as only one grammar? Do uncountably infinite probabilistic typologies turn finite when we count their grammars coarsely? Section 2 formalizes coarse identity between probabilistic grammars in a couple of different ways.

According to one formalization, two **ϵ-identical** probabilistic grammars can assign different probabilities to the same mapping, as long as the difference is negligible because smaller than some threshold ϵ (in absolute value). Equivalently, the ℓ_∞ distance between the two grammars is smaller than ϵ. This definition can be generalized by replacing the ℓ_∞ distance with other measures of the difference between probabilistic grammars, such as the ℓ_1 distance and the KL and χ^2 divergences.

According to another formalization of coarse identity, two **order-identical** probabilistic grammars can assign different probabilities to the same mapping as long as the difference is negligible because it does not affect the predicted probability inequalities. In other words, a mapping has a larger probability than another mapping according to one of the two grammars if and only if the same inequality holds according to the other grammar.

These notions of coarse identity yield corresponding notions of coarse finiteness. A probabilistic typology is called **ϵ-finite** or **order-finite** when it contains only finitely many grammars when we count two ϵ-identical grammars or two order-identical grammars as only one grammar. Section 3 investigates the coarse finiteness of typologies of maximum entropy (ME; Hayes and Wilson 2008) grammars.

Obviously, ME typologies always contain uncountably many grammars when we count two grammars as different in the standard sense, namely as soon as they assign different probabilities to some mapping. That is the case even when we consider only a handful of phonological mappings, no matter the constraints employed. Indeed, ME typologies are parametrized by uncountably many weight vectors and any two different weight vectors yield two ME grammars that differ because they assign different probabilities. Let us now turn from standard to coarse infinity.

To start, we consider the case of finitely many

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 190-204.
Irvine, California, June 27-29, 2024

phonological mappings. In this case, it is straightforward to verify that ME typologies are ϵ-finite and order-finite, no matter the choice of the constraints. Thus, we focus on the case of infinitely many phonological mappings, say the mappings corresponding to all underlying strings of finite but arbitrary length that can be constructed out of a finite alphabet of segments. Do ME typologies remain coarsely finite also in this case, no matter the choice of the constraints? Or can we construct counterexample constraints whose corresponding ME typologies contain infinitely many grammars even when we count grammar coarsely?

This paper shows that, for every threshold $\epsilon < 1$, it is possible to construct counterexample constraints such that the corresponding ME typology is ϵ-**in**finite. To illustrate, even if we choose $\epsilon = 0.999$ and are therefore willing to ignore pretty much all differences among probabilistic grammars, it is possible to construct a counterexample ME typology that is so so rich to qualify as infinite even at this level of coarseness.

Crucially, this richness is intrinsic to the ME mode of constraint interaction and does not require particularly complex constraint violation profiles. Indeed, the counterexample constraints can be chosen so simple that the corresponding categorical HG typology consists of a single grammar.

Furthermore, this result is robust: it does not depend on the specific way we measure differences among probabilities to adjudicate whether they are smaller than ϵ. Indeed, this result holds no matter whether ϵ-identity between probabilistic grammars is defined in terms of the ℓ_∞ distance or other measures of the difference between grammars, such as the ℓ_1 distances and the KL and χ^2 divergences.

Finally, this result extends from ϵ-identity to order-identity. Indeed, it is possible to construct counterexample constraints that are so simple that the corresponding categorical HG typology consists of a single grammar and yet the ME typology is order-**in**finite: its grammars order the infinitely many mappings made available by the phonological domain in infinitely many different ways.

The proofs of these results on ME coarse infiniteness consist of straightforward linear algebra manipulations detailed in the final appendix. The counterexample constraints constructed in these proofs are abstract and do not admit any readily available phonological interpretation. Although abstract, these counterexamples have substantial implications for the comparison between ME versus noisy or stochastic HG (SHG; Boersma and Pater 2016; Hayes 2017; Magri and Anttila in preparation), along the following lines.

ME and SHG look *prima facie* as very similar probabilistic extensions of categorical HG. They share the formalism of weighted constraints and have been shown to make very similar empirical predictions on a variety of test cases (Hayes 2017, Flemming 2021, and Breiss and Albright 2022, among others). Alderete and Finley (2023) indeed submit that ME and SHG "make use of relatively similar mathematical foundations, and often have very similar predictions. [...]. [They] produce very similar results, raising questions about what can be learned from different versions of Harmonic Grammar when the results are relatively similar. [...] It can be a challenge to compare differences between versions of Harmonic Grammar because they are so similar."

Yet, when we look beyond empirical predictions on a simple test cases and dig deeper into the underlying mathematics, we see that SHG and ME have very different formal properties. Coarse finiteness is indeed one of the mathematical properties on which ME and SHG come apart. In fact, Magri and Anttila (in preparation) show that SHG typologies are always ϵ-finite and always order-finite, no matter the number of mappings considered, as long as the constraints are simple, in the sense that thecorresponding categorical HG typology consists of only finitely many categorical grammars, which is usually the case (Pater 2009, 2016).

In other words, in the case of SHG, it is impossible to construct some counterexample constraints like those constructed here for ME, that yield an unrestrictive probabilistic typology (coarsely infinite) but the most restrictive categorical HG typology (a singleton). As summarized in the concluding section 4, the results on ME coarse infinity obtained in this paper show that ME is a richer, less restrictive probabilistic extension of categorical HG than SHG is.

2 Coarse finiteness

This section develops coarse notions of finiteness for probabilistic typologies that ignore "small" differences among probabilistic grammars.

2.1 Underlying and surface forms

A **phonological mapping** is a pair (x, y) consisting of an underlying form x and a surface realization y.

The description of the phonological system of interest starts by listing into a **phonological domain** $\mathfrak{D}$ all the relevant phonological mappings. $B_{\mathfrak{D}}$ denotes the **base set** of underlying forms listed by the phonological domain $\mathfrak{D}$. And $\mathfrak{D}(\mathsf{x})$ denotes the set of **candidate** surface realizations listed by $\mathfrak{D}$ for that underlying forms x.[1]

To circumvent the problem of defining probabilities on infinite sets, a candidate set $\mathfrak{D}(\mathsf{x})$ is usually assumed to be finite (but see Daland 2015). The base set $B_{\mathfrak{D}}$ is instead allowed to be countably infinite, say because it lists all the strings of finite but arbitrary length that can be constructed out of a finite alphabet of segments.

To illustrate, the phonological domain $\mathfrak{D}$ in figure 1 consists of the sixteen phonological mappings constructed out of the four strings CV, CVC, V, and VC, that differ for whether the onset or the coda are filled or empty. The base set $B_{\mathfrak{D}}$ consists of the underlying forms /CV/, /CVC/, /V/, /VC/. All candidate sets list the surface forms [CV], [CVC], [V], [VC].

2.2 Grammars and typologies

A **probabilistic (phonological) grammar** G assigns to each mapping (x, y) listed by the phonological domain $\mathfrak{D}$ a non-negative number $G(\mathsf{y} \,|\, \mathsf{x}) \geq 0$. We interpret this number as the probability of realizing the underlying form x as the surface candidate y. In order for this interpretation to make sense, these numbers $G(\mathsf{y} \,|\, \mathsf{x})$ must be **normalized** across candidate sets, as stated in (1).

$$\sum_{\mathsf{y} \in \mathfrak{D}(\mathsf{x})} G(\mathsf{y} \,|\, \mathsf{x}) = 1 \tag{1}$$

Equivalently, a probabilistic grammar G assigns to each underlying form x in the base set $B_{\mathfrak{D}}$ a **probability histogram** $G(\mathsf{x})$ on the corresponding candidate set $\mathfrak{D}(\mathsf{x})$. This reformulation highlights the fact that a probabilistic grammar G only models the probability of a surface realization y of a given underlying form x, as made explicit by the notation $G(\mathsf{y} \,|\, \mathsf{x})$ for **conditional probability**. A probabilistic grammar G does not model the probability of the underlying form x itself.

To illustrate, figure 2 provides two probabilistic grammars G_1 and G_2 for the phonological domain $\mathfrak{D}$ in figure 1. Grammar G_1 takes, say, the underlying form /CV/ and returns the leftmost probability histogram G_1(/CV/) over the candidate set $\mathfrak{D}$(/CV/). This probability histogram assigns to the surface candidate [CV] the probability 0.6.

Finally, a **probabilistic (phonological) typology** $\mathfrak{T}$ is a collection of probabilistic phonological grammars for the same phonological domain $\mathfrak{D}$. Throughout this section, we ignore how exactly typologies and grammars are defined (as ME grammars, as SHG grammars, and so on). The crucial point is that, no matter the choice of the framework, a probabilistic typology $\mathfrak{T}$ usually contains uncountably many different probabilistic grammars when two probabilistic grammars are counted as different in the standard sense, namely as soon as they assign slightly different probabilities to some mapping. The rest of this section thus develops coarser notions of identity between probabilistic grammars and spells out the corresponding coarser notions of finiteness for probabilistic typologies.

2.3 ϵ-finiteness

Given a threshold $\epsilon \geq 0$, two probabilistic grammars G_1 and G_2 are called **ϵ-identical** provided they assign to every mapping (x, y) in the phonological domain $\mathfrak{D}$ two probabilities $G_1(\mathsf{y} \,|\, \mathsf{x})$ and $G_2(\mathsf{y} \,|\, \mathsf{x})$ that differ by at most ϵ (in absolute value). To illustrate, the grammars G_1 and G_2 in figure 2 are not identical in the standard sense because, say, they assign different probabilities 0.6 and 0.55 to the mapping (/CV/, [CV]). Yet, these probabilities 0.6 and 0.55 differ by only $\epsilon = 0.05$. Analogous considerations hold for all mappings in the phonological domain $\mathfrak{D}$. These grammars G_1 and G_2 are therefore ϵ-identical with $\epsilon = 0.05$. If we ignore differences between probabilities up to $\epsilon = 0.05$, we can count these two probabilistic grammars as the "same" grammar.

A probabilistic typology $\mathfrak{T}$ is called **ϵ-finite** provided it contains a finite subset $T \subseteq \mathfrak{T}$ such that any grammar in the typology $\mathfrak{T}$ is ϵ-identical to some grammar in T. This condition is schematized in figure 3, where the red dots represent the grammars in the finite subset T, the blue dots represent all other grammars of the typology $\mathfrak{T}$, the lines represent ϵ-identity. In conclusion, if we ignore differences between probabilities up to ϵ, the finite subset T provides as much phonological information as the original (possibly infinite) typology $\mathfrak{T}$.

2.4 How to choose the threshold ϵ

When $\epsilon = 0$, two grammars are ϵ-identical only if they are identical in the standard sense, namely they assign exactly the same probability to every

[1] In the realm of OT, $\mathfrak{D}$ is notated *Gen*. I have changed notation to underscore the generality of the discussion.

$$\mathfrak{D} = \left\{ \begin{array}{cccc} (/\text{CV}/, [\text{CV}]) & (/\text{CVC}/, [\text{CV}]) & (/\text{V}/, [\text{CV}]) & (/\text{VC}/, [\text{CV}]) \\ (/\text{CV}/, [\text{CVC}]) & (/\text{CVC}/, [\text{CVC}]) & (/\text{V}/, [\text{CVC}]) & (/\text{VC}/, [\text{CVC}]) \\ (/\text{CV}/, [\text{V}]) & (/\text{CVC}/, [\text{V}]) & (/\text{V}/, [\text{V}]) & (/\text{VC}/, [\text{V}]) \\ (/\text{CV}/, [\text{VC}]) & (/\text{CVC}/, [\text{VC}]) & (/\text{V}/, [\text{VC}]) & (/\text{VC}/, [\text{VC}]) \end{array} \right\}$$

Figure 1: A phonological domain for basic syllable phonology

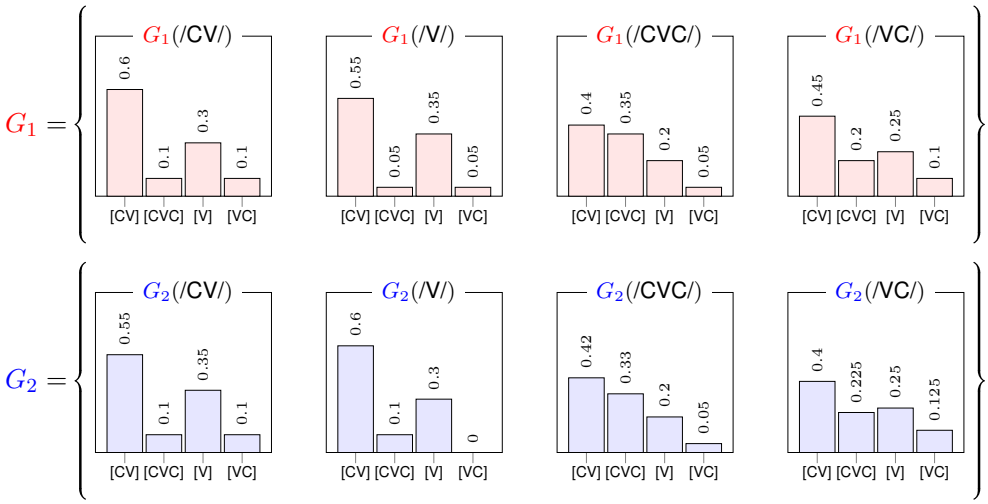

Figure 2: Two different grammars G_1 and G_2 that are nonetheless ϵ-identical with $\epsilon = 0.05$.

mapping. Hence, when $\epsilon = 0$, a probabilistic typology is ϵ-finite only if it is finite in the standard sense. In other words, ϵ-finiteness generalizes the standard notion of finiteness. As the threshold ϵ increases, we obtain coarser notions of finiteness.

When $\epsilon > 0$, the probability interval between 0 and 1 can be partitioned into finitely many disjoint intervals $I_1, I_2, \ldots, I_N$ of length at most ϵ. Suppose that the phonological domain $\mathfrak{D}$ lists only finitely many mappings (say, because the base set $B_{\mathfrak{D}}$ lists only finitely many underlying forms and all candidate sets are finite). In this case, any probabilistic typology $\mathfrak{T}$ is ϵ-finite because there are only finitely many ways of assigning one of the finitely many mappings from $\mathfrak{D}$ to one of the finitely many intervals $I_1, I_2, \ldots, I_N$. In other words, we can make infinitely many probability distinctions only when we distinguish among infinitely many mappings (namely, $\mathfrak{D}$ is infinite) or allow arbitrarily fine grained distinctions (namely, $\epsilon = 0$).

Finally, when $\epsilon \geq 1$, any two probabilistic grammars are ϵ-identical and any probabilistic typology is therefore ϵ-finite (just choose as the subset T a singleton consisting of a unique grammar from $\mathfrak{T}$).

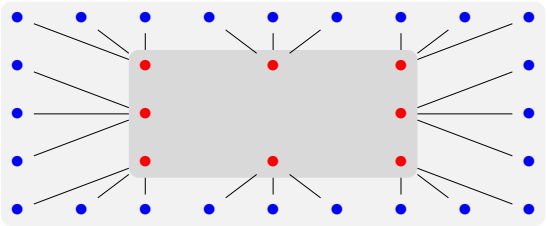

Figure 3: Schematic representation of ϵ-finiteness

In conclusion, it makes sense to investigate whether an infinite probabilistic typology $\mathfrak{T}$ is nonetheless ϵ-finite only for ϵ between zero and one (both excluded) and when the phonological domain $\mathfrak{D}$ lists infinitely many mappings.

2.5 Generalizing ϵ-finiteness

The notion of ϵ-finiteness introduced in subsection 2.3 can be generalized as follows. We denote by D any function that takes two probabilistic grammars G_1 and G_2 for the same phonological domain $\mathfrak{D}$ and returns a non-negative number $D(G_1, G_2) \geq 0$ subject to the only condition that $D(G_1, G_2) = 0$ if and only if the grammars G_1 and G_2 are identical

in the standard sense, namely they assign the same probability to any mapping in the phonological domain $\mathfrak{D}$. We will interpret the quantity $D(G_1, G_2)$ as a measure of the difference between G_1 and G_2 and thus refer to D as a **distance** between probabilistic grammars (this is a slight abuse as D need not even be symmetric: $D(G_1, G_2)$ and $D(G_2, G_1)$ can be different quantities).

Two probabilistic grammars G_1 and G_2 are then called **ϵ-identical according to** D provided their distance measured by D is at most ϵ, namely $D(G_1, G_2) \leq \epsilon$. Furthermore, a probabilistic typology $\mathfrak{T}$ is called **ϵ-finite according to** D provided it contains some finite subset $T \subseteq \mathfrak{T}$ such that any grammar in the typology $\mathfrak{T}$ is ϵ-identical according to D to some grammar in T. Since the distance $D(G_1, G_2)$ is equal to zero if and only if the two grammars G_1 and G_2 are identical in the standard sense, the notion of ϵ-finiteness according to D with $\epsilon = 0$ coincides with the standard notion of finiteness. In conclusion, ϵ-finiteness generalizes the standard notion of finiteness, no matter the distance D used to compare grammars.

Here is a simple strategy to define a distance between two probabilistic grammars G_1 and G_2. First, we define a distance $D(G_1(\mathsf{x}), G_2(\mathsf{x}))$ between the probability histograms $G_1(\mathsf{x})$ and $G_2(\mathsf{x})$ assigned by the two grammars G_1 and G_2 to an arbitrary underlying form x in the base set $B_{\mathfrak{D}}$ of the phonological domain. Then, we define the distance $D(G_1, G_2)$ between the two grammars G_1, G_2 as the largest distance between their probability histograms, as stated in (2).

$$D(G_1, G_2) = \sup_{\mathsf{x} \in B_{\mathfrak{D}}} D\big(G_1(\mathsf{x}), G_2(\mathsf{x})\big) \tag{2}$$

The initial notion of ϵ-finiteness from subsection 2.3 fits into this scheme when the distance D is the ℓ_∞ (or supremum) distance D_∞ recalled in (3). It measures the distance between two probability histograms in terms of the largest difference between two bars for the same candidate.

$$\begin{aligned} &D_\infty\big(G_1(\mathsf{x}), G_2(\mathsf{x})\big) \\ &= \sup_{\mathsf{y} \in \mathfrak{D}(\mathsf{x})} \big| G_1(\mathsf{y} \,|\, \mathsf{x}) - G_2(\mathsf{y} \,|\, \mathsf{x}) \big| \end{aligned} \tag{3}$$

Another natural distance that can be used to define ϵ-finiteness is the ℓ_1 distance D_1 recalled in (4). It measures the distance between two probability histograms in terms of the sum of the differences between two bars for the same candidates (by Scheffé's theorem, it is equal to twice the total variation distance; see Tsybakov 2009, lemma 21, page 84).

$$\begin{aligned} &D_1\big(G_1(\mathsf{x}), G_2(\mathsf{x})\big) \\ &= \sum_{\mathsf{y} \in \mathfrak{D}(\mathsf{x})} \big| G_1(\mathsf{y} \,|\, \mathsf{x}) - G_2(\mathsf{y} \,|\, \mathsf{x}) \big| \end{aligned} \tag{4}$$

Other natural choices for the distance D are so called f-divergences (Tsybakov 2009, section 2.4) such as the Kullback-Leibler (KL) and the χ^2 divergences. When no mapping in the phonological domain has zero probability (as is the case for ME), these two divergences are defined as in (5) and (6).

$$\begin{aligned} &D_{\mathrm{KL}}\big(G_1(\mathsf{x}), G_2(\mathsf{x})\big) \\ &= \sum_{\mathsf{y} \in \mathfrak{D}(\mathsf{x})} G_1(\mathsf{y} \,|\, \mathsf{x}) \log \frac{G_1(\mathsf{y} \,|\, \mathsf{x})}{G_2(\mathsf{y} \,|\, \mathsf{x})} \end{aligned} \tag{5}$$

$$\begin{aligned} &D_{\chi^2}\big(G_1(\mathsf{x}), G_2(\mathsf{x})\big) \\ &= \sum_{\mathsf{y} \in \mathfrak{D}(\mathsf{x})} \frac{\big(G_1(\mathsf{y} \,|\, \mathsf{x}) - G_2(\mathsf{y} \,|\, \mathsf{x})\big)^2}{G_2(\mathsf{y} \,|\, \mathsf{x})} \end{aligned} \tag{6}$$

2.6 From sheer sizes to inequalities

The notion of ϵ-identity looks at the sheer size of the probabilities and it is coarse because it ignores small differences in size. Various authors have suggested that we should focus not on the sheer size of the probabilities but on the inequalities they satisfy. For instance, Coetzee (2004, 2006) argues that probabilistic phonology should only model relative empirical frequencies, not absolute frequencies. In other words, a probabilistic grammar should be evaluated by comparing the inequalities among the probabilities it predicts with the inequalities among the empirical frequencies, not by fitting the predicted probabilities to the empirical frequencies.

Furthermore, the generalizations uncovered in probabilistic phonology usually consist of probability inequalities. A representative example is the famous generalization that word final t-deletion (the deletion of a stop at the end of a word preceded by another consonant) is more frequent when the following word starts with a consonant than when it starts with a vowel (see Guy 1980 and Coetzee and Kawahara 2013 for overviews). This generalization indeed consists of an inequality between the frequencies of deletion for *cost#me* versus *cost#us*. The generalization says nothing about the absolute frequencies of deletion. Indeed, Anttila and Magri (2018) and Magri and Anttila (in preparation) capture such generalizations by extending the

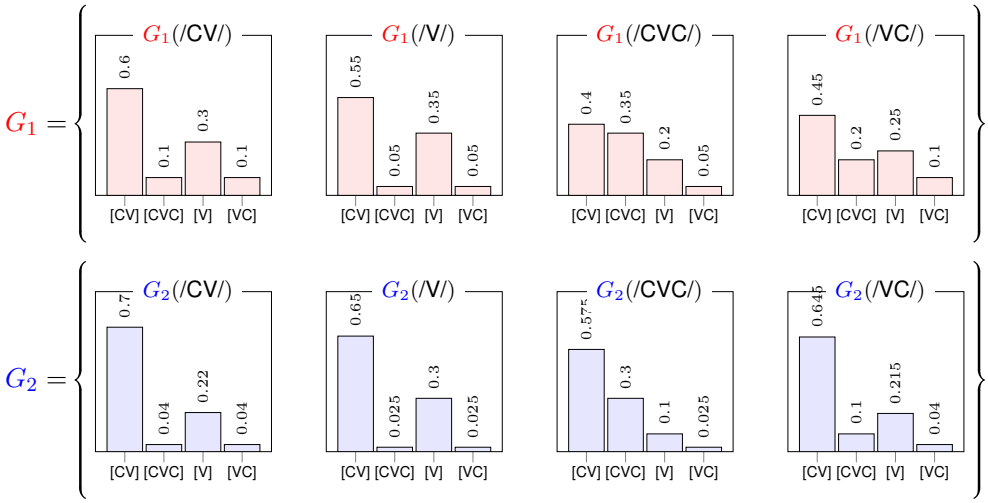

Figure 4: Two different grammars G_1 and G_2 that are nonetheless order-identical

Greenbergian implicational universals from the categorical to the probabilistic setting in terms of probability inequalities that hold uniformly across all the grammars in a probabilistic typology.

2.7 Order-finiteness

Based on these considerations, we say that two probabilistic grammars G_1 and G_2 for some phonological domain $\mathfrak{D}$ are **order-identical** provided they agree on how they order the mappings in $\mathfrak{D}$ in terms of the size of their probabilities: for any two mappings (x, y) and $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ from $\mathfrak{D}$, grammar G_1 satisfies the inequality $G_1(\mathsf{y} \,|\, \mathsf{x}) > G_1(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{x}})$ if and only if the other grammar G_2 satisfies the same inequality $G_2(\mathsf{y} \,|\, \mathsf{x}) > G_2(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{x}})$.

To illustrate, the two grammars G_1 and G_2 in figure 2 are ϵ-identical because they assign probabilities that differ by at most $\epsilon = 0.05$. Yet, they are not order-identical because these small differences in probabilities impact the inequalities. For instance, G_1 assigns more probability to (/CV/, [CV]) than to (/V/, [CV]) while G_2 does the reverse.

The situation is different for the two grammars G_1 and G_2 in figure 4. They are not ϵ-identical with $\epsilon = 0.05$ (for instance because the probabilities they assign to (/CV/, [CV]) differ by 0.1). Yet, both G_1 and G_2 assign more probability to (/CV/, [CV]) than to (/V/, [CV]). Analogous considerations hold for any pair of mappings in the phonological domain $\mathfrak{D}$: G_1 and G_2 induce the same order of the sixteen mappings according to the size of their probabilities (with ties broken in some arbitrary but fixed way), as made explicit in figure 5. We conclude that these grammars G_1 and G_2 are order-identical. If we ignore sheer differences between probabilities and only care about the inequalities they satisfy, as argued in subsection 2.6, we can count these two probabilistic grammars G_1 and G_2 as the "same" grammar.

A probabilistic typology $\mathfrak{T}$ is called **order-finite** provided it contains some finite set $T \subseteq \mathfrak{T}$ such that any grammar in the typology $\mathfrak{T}$ is order-identical to some grammar in T. In other words, this finite subset T provides as much phonological information as the original (possibly infinite) typology $\mathfrak{T}$ when we ignore sheer probabilities and only care about the inequalities they satisfy.

When the phonological domain $\mathfrak{D}$ lists only finitely many mappings, any probabilistic typology $\mathfrak{T}$ is order-finite, because there are only finitely many ways of ordering finitely many mappings. Thus, it makes sense to investigate whether an infinite probabilistic typology $\mathfrak{T}$ is nonetheless order-finite only when the phonological domain $\mathfrak{D}$ lists infinitely many mappings.

2.8 Summary

An infinite probabilistic typology is called **coarsely finite** if it is ϵ-finite relative to some distance D for some threshold ϵ between zero and one as in subsection 2.5 or order-finite as in subsection 2.7. In other words, the typology contains only finitely many grammars when we count grammars coarsely by ignoring differences between probabilities that are negligible because smaller than ϵ or because too small to affect the inequalities among probabilities.

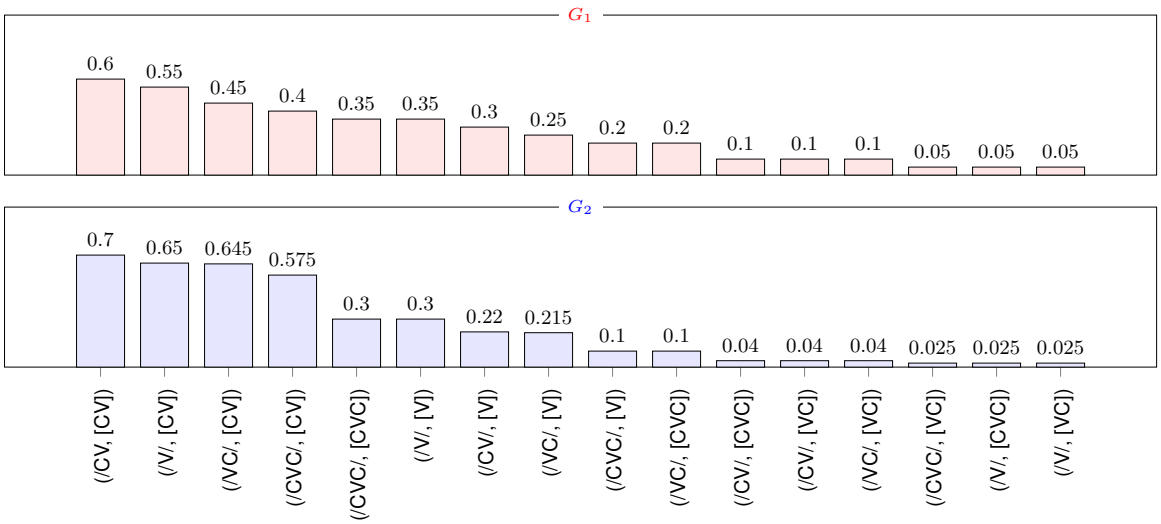

Figure 5: Grammars G_1 and G_2 in figure 4 order the mappings based on their probabilities in the same way

3 Coarse finiteness of ME typologies

This section applies these notions of coarse finiteness to the analysis of ME typologies.

3.1 Probabilistic ME typologies

So far, we have worked with arbitrary probabilistic grammars, extensionally defined as in subsection 2.2 as collections of probability histograms. Now, we focus on ME grammars, briefly recalled here. We start with a set $\boldsymbol{C}$ consisting of a finite number n of phonological **constraints** $C_1, \ldots, C_n$ for the phonological domain $\mathfrak{D}$.[2] A constraint C_k assigns to each phonological mapping (x, y) a number $C_k(\mathsf{x}, \mathsf{y})$. This number is integral and non-negative because it is the result of counting the number of occurrences of some specific marked structure in the surface form y or the number of occurrences of some specific discrepancy between the underlying and surface forms x and y. Each constraint C_k is assigned a non-negative **weight** $w_k \geq 0$ that quantifies its importance. These weights are collected into a vector $\boldsymbol{w} = (w_1, \ldots, w_n)$.

The probabilistic **ME grammar** $G^{\mathrm{ME}}_{\boldsymbol{w}}$ corresponding to this weight vector $\boldsymbol{w}$ assigns to each mapping (x, y) a probability proportional to the exponential of the opposite of the weighted sum of constraint violations, as stated in (7). The proportionality constant is univocally determined by the normalization condition (1).

$$G^{\mathrm{ME}}_{\boldsymbol{w}}(\mathsf{y} \,|\, \mathsf{x}) \propto \exp\left\{ -\sum_{k=1}^{n} w_k C_k(\mathsf{x}, \mathsf{y}) \right\} \qquad (7)$$

The probabilistic **ME typology** defined by a phonological domain $\mathfrak{D}$ and a constraint set $\boldsymbol{C}$ is the family $\mathfrak{T}^{\mathrm{ME}}(\mathfrak{D}, \boldsymbol{C})$ consisting of the probabilistic ME grammars (7) corresponding to all vectors $\boldsymbol{w}$ of non-negative constraint weights.

3.2 Categorical HG typologies

Probabilistic ME grammars are closely related to categorical HG grammars recalled here as well. A weight vector $\boldsymbol{w}$ is called **proper** provided it satisfies the following condition: for every underlying form x in the base set $B_{\mathfrak{D}}$, there exists a unique surface form y (called the **winner**) in the candidate set $\mathfrak{D}(\mathsf{x})$ that is assigned by the corresponding ME grammar $G^{\mathrm{ME}}_{\boldsymbol{w}}$ a probability strictly larger than the probability assigned to each other surface form z (dismissed as a **loser**) in the candidate set $\mathfrak{D}(\mathsf{x})$, namely $G^{\mathrm{ME}}_{\boldsymbol{w}}(\mathsf{y} \,|\, \mathsf{x}) > G^{\mathrm{ME}}_{\boldsymbol{w}}(\mathsf{z} \,|\, \mathsf{x})$. The categorical **HG grammar** corresponding to a proper weight vector $\boldsymbol{w}$ realizes each underlying form in the base set $B_{\mathfrak{D}}$ as the corresponding winner candidate with largest ME probability. The categorical **HG typology** defined by a phonological domain $\mathfrak{D}$ and a constraint set $\boldsymbol{C}$ is the family $\mathfrak{T}^{\mathrm{HG}}(\mathfrak{D}, \boldsymbol{C})$ consisting of the categorical HG grammars corresponding to all proper vectors $\boldsymbol{w}$ of non-negative weights.

[2] In the realm of OT, $\boldsymbol{C}$ is notated Con. I have changed notation to underscore the generality of the discussion.

3.3 ME typologies are not coarsely finite

Both categorical HG typologies and probabilistic ME typologies are parametrized by uncountably many weight vectors. Since categorical grammars make only binary choices, many different weight vectors yield the same categorical HG grammar. Since probabilities can instead take any continuous value between zero and one, any two different weight vectors yield two probabilistic ME grammars that are different in the standard sense, namely assign different probabilities to the same phonological mapping. As a result probabilistic ME typologies are always uncountably infinite, even when the phonological domain $\mathfrak{D}$ lists only one underlying form with only two candidate surface realizations.

Yet, in subsection 2.3 we have said that two probabilistic grammars are ϵ-identical or order-identical when the differences between the probabilities they assign are negligible because they are smaller than some threshold ϵ or they do not affect the inequalities among probabilities. We have then observed that an infinite probabilistic typology can nonetheless qualify as ϵ-finite or order-finite when indeed we count multiple ϵ-identical or multiple order-identical grammars as one single grammar.

Are ME typologies always ϵ-finite or order-finite, no matter the choice of the constraints? The following two main results provide a negative answer to this question. The proofs of these two facts consist of straightforward linear algebra manipulations detailed in the final appendix.

Result 1 *For every positive threshold $0 < \epsilon < 1$ strictly smaller than one, it is possible to construct an infinite phonological domain $\mathfrak{D}$ and a constraint set $\boldsymbol{C}$ such that the corresponding ME typology $\mathfrak{T}^{ME}(\mathfrak{D}, \boldsymbol{C})$ is ϵ-**in**finite while the corresponding HG typology $\mathfrak{T}^{HG}(\mathfrak{D}, \boldsymbol{C})$ is a singleton.* □

Result 2 *It is possible to construct an infinite phonological domain $\mathfrak{D}$ and a constraint set $\boldsymbol{C}$ such that the corresponding ME typology $\mathfrak{T}^{ME}(\mathfrak{D}, \boldsymbol{C})$ is order-**in**finite while the corresponding HG typology $\mathfrak{T}^{HG}(\mathfrak{D}, \boldsymbol{C})$ is a singleton.* □

A few remarks are in order. **(A)** Let us consider a threshold ϵ very close to one, say $\epsilon = 0.999$. This means that we are willing to ignore as negligible pretty much all disagreements among probabilities. In other words, we are willing to count as one single grammar even multiple grammars that are very different in the standard sense. And yet, even at this highest degree of coarseness, result 1 says that we can construct ME typologies that are ϵ-infinite. **(B)** This typological richness is a direct consequence of the ME mode of constraint interaction and does not require a particularly complex pattern of constraint violations. Indeed, both results guarantee that the constraints used in the ME counterexamples are very simple, in the sense that the corresponding categorical HG typology is simplest, namely consists of a single grammar. **(C)** Finally, result 1 is robust: appendix 5.3 shows that it straightforwardly extends from the original basic notion of ϵ-finiteness from subsection 2.3 to its generalization in subsection 2.5 in terms of other measures of the difference between probabilistic grammars such as the ℓ_1 distance and the KL and χ^2 divergences.

3.4 Comparison with SHG

To appreciate the significance of these results for phonological theory, we briefly turn to SHG phonology, recalled here. The probabilistic **SHG grammar** corresponding to a non-negative weight vector $\boldsymbol{w}$ assigns to each mapping (x, y) a probability equal to the probability of sampling according to the normal distribution with mean $\boldsymbol{w}$ some non-negative proper weight vector such that the corresponding categorical HG grammar indeed realizes the underlying form x as the surface candidate y.[3] The probabilistic **SHG typology** is the family of the probabilistic SHG grammars $G^{\text{SHG}}_{\boldsymbol{w}}$ corresponding to all vectors $\boldsymbol{w}$ of non-negative constraint weights

ME and SHG look *prima facie* as similar probabilistic extensions of categorical HG. Indeed, both ME and SHG are defined in terms of weighted sums of constraint violations. Furthermore, ME and SHG have been shown to fit equally well various patterns of empirical frequencies (Hayes 2017, Flemming 2021, and Breiss and Albright 2022, among others). Yet, SHG behaves very differently from ME in terms of coarse finiteness, as follows.

Typologies of categorical HG grammars can be infinite (Legendre et al. 2006), contrary to typologies of categorical OT grammars, that are instead always finite. Yet, OT and HG make such diver-

[3] Thus defined, SHG grammars can unfortunately flout the normalization condition (1): the normal distribution with mean $\boldsymbol{w}$ can assign some probability to vectors that are negative or non-proper and therefore correspond to no categorical HG grammar. Hayes and Kaplan (2023) and Magri and Anttila (in preparation) discuss various modifications of the basic definition of SHG to deal with this problem. These modifications have no implications for the coarse finiteness of SHG typologies.

gent typological predictions only for very special (and possibly unwarranted) constraint configurations (Pater 2009, 2016). In general, categorical HG typologies are finite, just as OT typologies.

Magri and Anttila (in preparation) then show that, whenever the categorical HG typology corresponding to some constraint set is finite, the probabilistic SHG typology corresponding to that constraint set, although uncountably infinite, is nonetheless ϵ-finite and order-finite. It is therefore impossible to construct for SHG some counterexample constraints like those constructed here for ME, that yield a very complex probabilistic typology (coarsely infinite) but a very simple categorical HG typology (a singleton).

To illustrate, let us consider a threshold ϵ very close to zero, say $\epsilon = 0.0001$. This means that we are willing to ignore as negligible only the smallest differences among probabilities. In other words, we are willing to count as one single grammar only two grammars that are indeed very close to being identical in the standard sense. And yet, even at this lowest degree of coarseness, we cannot construct SHG typologies that are ϵ-infinite, unless we resort to special (and possibly unwarranted) constraint sets that yield infinite categorical HG typologies. We conclude that the results on ME coarse infinity obtained in this paper show that ME is a richer, less restrictive probabilistic extension of categorical HG than SHG is.

4 Conclusions

This paper has developed techniques to discretize an uncountably infinite probabilistic typology down to a finite core by ignoring small differences among probabilities. The notion of ϵ-finiteness arises when we ignore differences smaller than ϵ between the probabilities assigned by two grammars. The notion of order-finiteness arises when we ignore differences that do not compromise the inequalities among the probabilities assigned by two grammars. Magri and Anttila (in preparation) show that SHG typologies are always ϵ-finite and order-finite, as long as the constraints are simple, in the sense that the corresponding categorical HG typology is finite. This paper has shown that ME typologies can instead be ϵ-**in**finite and order-**in**finite, even when the constraints are so simple that the corresponding categorical HG typology is a singleton. We conclude that ME is a richer, less restrictive probabilistic extension of categorical HG.

References

John Alderete and Sara Finley. 2023. Probabilistic phonology: a review of theoretical perspectives, applications, and problems. *Language and Linguistics*, 24:565–610.

Arto Anttila and Giorgio Magri. 2018. Does MaxEnt overgenerate? Implicational universals in Maximum Entropy grammar. In *AMP 2017: Proceedings of the 2017 Annual Meeting on Phonology*, Washington, DC. Linguistic Society of America.

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

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

Andries W. Coetzee. 2004. *What it Means to be a Loser: Non-Optimal Candidates in Optimality Theory*. Ph.D. thesis, University of Massachusetts, Amherst.

Andries W. Coetzee. 2006. Variation as assessing 'non-optimal' candidates. *Phonology*, 23:337–385.

Andries W. Coetzee and Shigeto Kawahara. 2013. Frequency biases in phonological variation. *Natural Language and Linguistic Theory*, 31(1):47–89.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing.

Robert Daland. 2015. Long words in maximum entropy phonotactic grammars. *Phonology*, 32.3:353–383.

Edward Flemming. 2021. Comparing maxent and noisy harmonic grammar. *Glossa: a journal of general linguistics*, 6:1–42.

Gregory R. Guy. 1980. Variation in the group and the individual: The case of final stop deletion. In William Labov, editor, *Locating language in time and space*, pages 1–36. Academic Press, New York.

Bruce Hayes. 2017. Varieties of Noisy Harmonic Grammar. In *Proceedings of the 2016 Annual Meeting in Phonology*, Washington, DC. Linguistic Society of America.

Bruce Hayes and Aaron Kaplan. 2023. Zero-weighted constraints in Noisy Harmonic Grammar. *Linguistic Inquiry*, pages 1–14.

Bruce Hayes and Colin Wilson. 2008. A Maximum Entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39:379–440.

Géraldine Legendre, Antonella Sorace, and Paul Smolensky. 2006. The optimality theory/harmonic grammar connection. In Paul Smolensky and Gèraldine Legendre, editors, *The Harmonic Mind*, pages 903–966. MIT Press, Cambridge, MA.

Giorgio Magri and Arto Anttila. in preparation. Principles of probabilistic phonology.

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

Joe Pater. 2016. Universal grammar with weighted constraints. In Joe Pater and John J. McCarthy, editors, *Harmonic Grammar and Harmonic Serialism*, pages 1–46. Equinox, London.

Alexandre B. Tsybakov. 2009. *Introduction to Nonparametric Estimation*. Springer Verlag, New York.

5 Appendices

Throughout this appendix, $\boldsymbol{a} \cdot \boldsymbol{b}$ denotes the scalar product $\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{k=1}^{n} a_k b_k$ between two vectors $\boldsymbol{a} = (a_1, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, \ldots, b_n)$; furthermore, $\|\boldsymbol{a}\|$ denotes the 2-norm $\|\boldsymbol{a}\| = \sqrt{\sum_{k=1}^{n} a_k^2}$. The proofs in this appendix consist of straightforward linear algebra manipulations.

5.1 A lemma for the proof of result 1

Lemma 1 *Consider $k-1$ vectors $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{k-1}$ with positive integral components and a vector $\boldsymbol{w}_k$ with positive rational components such that $\boldsymbol{w}_k \cdot \boldsymbol{c}_1 \leq 1, \ldots, \boldsymbol{w}_k \cdot \boldsymbol{c}_{k-1} \leq 1$. For any $\Delta > 0$, there exist a vector $\boldsymbol{c}_k$ with positive integral components and a vector $\boldsymbol{w}_{k+1}$ with positive rational components such that $\boldsymbol{w}_{k+1} \cdot \boldsymbol{c}_1 \leq 1, \ldots, \boldsymbol{w}_{k+1} \cdot \boldsymbol{c}_{k-1} \leq 1$ and furthermore $\boldsymbol{w}_{k+1} \cdot \boldsymbol{c}_k \leq 1$ while $\boldsymbol{w}_k \cdot \boldsymbol{c}_k \geq \Delta$.*

Indeed, since the vector $\boldsymbol{w}_k$ has positive rational components, it has the shape $\boldsymbol{w} = (\frac{a_1}{b_1}, \frac{a_2}{b_2}, \ldots, \frac{a_n}{b_n})$, where $a_1, \ldots, a_n, b_1, \ldots, b_n$ are positive integers. Let M be the smallest common multiple of the denominators $b_1, \ldots, b_n$. Hence, $M\boldsymbol{w}$ is a vector with positive integral components. We choose a positive integer $\ell > 0$ and a positive rational number $\xi > 0$ as in (8).

$$\ell \geq \frac{\Delta}{M\|\boldsymbol{w}_k\|^2}, \qquad \xi \leq \min\left\{1, \frac{1}{M\ell\|\boldsymbol{w}_k\|^2}\right\} \tag{8}$$

We define the vector $\boldsymbol{c}_k$ with positive integral components and the vector $\boldsymbol{w}_{k+1}$ with positive rational components as in (9).

$$\boldsymbol{c}_k = \ell M \boldsymbol{w}_k \qquad\qquad \boldsymbol{w}_{k+1} = \xi \boldsymbol{w}_k \tag{9}$$

These positions satisfy the inequalities (10) and (11) as well as the inequality (12) for every $h = 1, \ldots, k-1$, completing the proof of the lemma.

$$\begin{aligned} \boldsymbol{w}_k \cdot \boldsymbol{c}_k &= \ell M \|\boldsymbol{w}_k\|^2 \geq \Delta && (10)\\ \boldsymbol{w}_{k+1} \cdot \boldsymbol{c}_k &= \ell \xi M \|\boldsymbol{w}_k\|^2 \leq 1 && (11)\\ \boldsymbol{w}_{k+1} \cdot \boldsymbol{c}_h &= \xi \boldsymbol{w}_k \cdot \boldsymbol{c}_h \leq \xi 1 \leq 1 && (12) \end{aligned}$$

5.2 Proof of result 1

Given a threshold $0 < \epsilon < 1$, we choose two constants $0 < \epsilon_1 < \epsilon_2 < 1$ more than ϵ apart, namely $\epsilon_1 + \epsilon < \epsilon_2$. Furthermore, we choose a positive integer $m > 0$ and a positive constant $\Delta > 0$ that satisfy the inequalities in (13).

$$m \geq \frac{1-\epsilon_1}{\epsilon_1} e, \qquad \Delta \geq \log\left(m \frac{\epsilon_2}{1-\epsilon_2}\right) \tag{13}$$

We start with an arbitrary vector $\boldsymbol{w}_1$ with positive rational components. By applying lemma 1 with $k = 1$ to this vector $\boldsymbol{w}_1$, we conclude that there exist a vector $\boldsymbol{c}_1$ with positive integral components and a vector $\boldsymbol{w}_2$ with positive rational components that validate the red inequalities in the first step of the reasoning in figure 6. By applying again lemma 1 with $k = 2$ to the vectors $\boldsymbol{c}_1$ and $\boldsymbol{w}_2$ in the bottom line of this first step, we conclude that there exist a vector $\boldsymbol{c}_2$ with positive integral components and a vector $\boldsymbol{w}_3$ with positive rational components that validate the red inequalities in the second step of the reasoning in figure 6. By applying once again lemma 1 with $k = 3$ to the vectors $\boldsymbol{c}_1, \boldsymbol{c}_2$ and $\boldsymbol{w}_3$ in the bottom line of this second step, we conclude that there exist a vector $\boldsymbol{c}_3$ with positive integral components and a vector $\boldsymbol{w}_4$ with positive rational components that validate the red inequalities in the third step of the reasoning in figure 6. And so on and so forth.

In conclusion, we have established the existence of a sequence of vectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k, \ldots$ with positive rational components and a sequence of vectors $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_k, \ldots$ with positive integral components that satisfy the k inequalities in (14) for every $k = 1, 2, \ldots$.

$$\begin{aligned} \boldsymbol{w}_k \cdot \boldsymbol{c}_1 &\leq 1 \qquad\qquad \boldsymbol{w}_k \cdot \boldsymbol{c}_k \geq \Delta \\ &\vdots \\ \boldsymbol{w}_k \cdot \boldsymbol{c}_{k-1} &\leq 1 \end{aligned} \tag{14}$$

To construct the desired counterexample, we consider the infinite phonological domain $\mathfrak{D}$ described in (15). For every index $k = 1, 2, \ldots$, the base set $B_{\mathfrak{D}}$ of the phonological domain contains the underlying form x_k. Its candidate set $\mathfrak{D}(\mathsf{x}_k)$ consists of $m+1$ candidates $\mathsf{y}, \mathsf{z}_1, \ldots, \mathsf{z}_m$. For concreteness, we refer to y as the **winner** candidate and to $\mathsf{z}_1, \ldots, \mathsf{z}_m$ as the **loser** candidates.

$$\begin{aligned} B_{\mathfrak{D}} &= \{\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_k, \ldots\} \\ \mathfrak{D}(\mathsf{x}_k) &= \{\mathsf{y},\ \mathsf{z}_1, \ldots, \mathsf{z}_m\} \end{aligned} \tag{15}$$

Furthermore, we define the constraint set $\boldsymbol{C}$ in such a way that, for every underlying form x_k and for each loser candidate z_i with $i = 1, \ldots, m$, the difference between the constraint violation vector $\boldsymbol{C}(\mathsf{x}_k, \mathsf{z}_i)$ of this loser candidate minus the constraint violation vector $\boldsymbol{C}(\mathsf{x}_k, \mathsf{y})$ of the winner candidate y is equal to the vector $\boldsymbol{c}_k$ constructed in (14), as stated in (16).

$$\boldsymbol{C}(\mathsf{x}_k, \mathsf{z}_i) - \boldsymbol{C}(\mathsf{x}_k, \mathsf{y}) = \boldsymbol{c}_k \tag{16}$$

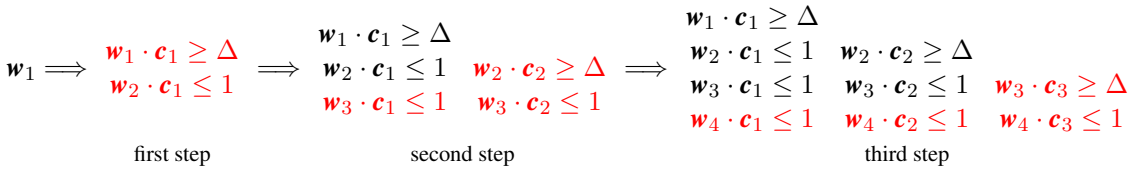

Figure 6

This position (16) makes sense because the vector $\boldsymbol{c}_k$ has integral components that can can therefore be interpreted as differences between numbers of constraint violations under the usual assumption that constraints assign integral numbers of violations. Furthermore, the integral components of the vector $\boldsymbol{c}_k$ are all positive. The position (16) thus says that every constraint in the constraint set $\boldsymbol{C}$ assigns less violations to the winner mapping $(\mathsf{x}_k, \mathsf{y})$ than to each of the loser mappings $(\mathsf{x}_k, \mathsf{z}_i)$. Equivalently, the winner mapping $(\mathsf{x}_k, \mathsf{y})$ always beats each loser mapping $(\mathsf{x}_k, \mathsf{z}_i)$ in HG, no matter the choice of the non-negative constraint weights. We conclude that the HG typology $\mathfrak{T}^{\text{HG}}(\mathfrak{D}, \boldsymbol{C})$ corresponding to the phonological domain $\mathfrak{D}$ in (15) and the constraint set $\boldsymbol{C}$ in (16) consists of a unique HG grammar, namely the grammar that realizes each underlying form x_k as its winner candidate y.

We now switch from categorical HG to probabilistic ME. We focus on the ME grammar $G^{\text{ME}}_{\boldsymbol{w}_k}$ corresponding to the weight vector $\boldsymbol{w}_k$ in (14). We want to bound the probability it assigns to the mappings $(\mathsf{x}_h, \mathsf{y})$ with $h = 1, \ldots, k-1$ and to the mapping $(\mathsf{x}_k, \mathsf{y})$. As explained below, the inequalities $\boldsymbol{w}_k \cdot \boldsymbol{c}_h \leq 1$ with $h = 1, \ldots, k-1$ on the lefthand side of (14) ensure that the ME grammar $G^{\text{ME}}_{\boldsymbol{w}_k}$ assigns to the mapping $(\mathsf{x}_h, \mathsf{y})$ with $h = 1, \ldots, k-1$ a probability that is small, namely at most ϵ_1, as stated in (17). Analogously, the inequality $\boldsymbol{w}_k \cdot \boldsymbol{c}_k \geq \Delta$ on the righthand side of (14) ensures that the ME grammar $G^{\text{ME}}_{\boldsymbol{w}_k}$ assigns to the mapping $(\mathsf{x}_k, \mathsf{y})$ a probability that is instead large, namely at least ϵ_2, as stated in (18).

$$G^{\text{ME}}_{\boldsymbol{w}_k}(\mathsf{y} \,|\, \mathsf{x}_h) \;\leq\; \epsilon_1 \tag{17}$$

$$G^{\text{ME}}_{\boldsymbol{w}_k}(\mathsf{y} \,|\, \mathsf{x}_k) \;\geq\; \epsilon_2 \tag{18}$$

Indeed, the computation in (19) establishes the inequality (17) and an analogous computation establishes the inequality (18). Step (19a) holds because of the definition (7) of the ME probability $G^{\text{ME}}_{\boldsymbol{w}_k}(\mathsf{y} \,|\, \mathsf{x}_h)$ as proportional to the exponential of the opposite of the weighted sum of the constraint violations of the winner candidate y. The proportionality constant is univocally determined by the normalization condition (1) and has been made explicit in the denominator. Step (19b) holds by dividing both the numerator and the denominator by $\exp\{-\boldsymbol{w}_k \cdot \boldsymbol{C}(\mathsf{x}_h, \mathsf{y})\}$. Step (19c) holds because of the assumption (16) that the difference $\boldsymbol{C}(\mathsf{x}_h, \mathsf{z}_i) - \boldsymbol{C}(\mathsf{x}_h, \mathsf{y})$ is equal to the vector $\boldsymbol{c}_h$ for every $i = 1, \ldots, m$. Step (19d) holds because of the assumption (14) that the scalar product $\boldsymbol{w}_k \cdot \boldsymbol{c}_h$ between the two vectors $\boldsymbol{w}_k$ and $\boldsymbol{c}_h$ is at most one for every $h = 1, \ldots, k-1$. Finally, step (19e) holds because of the assumption that the integer m is large enough, as in (13).

$$\begin{aligned}
G^{\text{ME}}_{\boldsymbol{w}_k}(\mathsf{y} \,|\, \mathsf{x}_h) = & \qquad (19)\\
&\overset{(a)}{=} \frac{e^{-\boldsymbol{w}_k \cdot \boldsymbol{C}(\mathsf{x}_h, \mathsf{y})}}{e^{-\boldsymbol{w}_k \cdot \boldsymbol{C}(\mathsf{x}_h, \mathsf{y})} + \sum_{i=1}^{m} e^{-\boldsymbol{w}_k \cdot \boldsymbol{C}(\mathsf{x}_h, \mathsf{z}_i)}}\\
&\overset{(b)}{=} \frac{1}{1 + \sum_{i=1}^{m} e^{-\boldsymbol{w}_k \cdot \left(\boldsymbol{C}(\mathsf{x}_h, \mathsf{z}_i) - \boldsymbol{C}(\mathsf{x}_h, \mathsf{y})\right)}}\\
&\overset{(c)}{=} \frac{1}{1 + m e^{-\boldsymbol{w}_k \cdot \boldsymbol{c}_h}}\\
&\overset{(d)}{\leq} \frac{1}{1 + m e^{-1}} \overset{(e)}{\leq} \epsilon_1
\end{aligned}$$

To complete the proof of result 1, we now consider the ME grammars $G^{\text{ME}}_{\boldsymbol{w}_{k_1}}$ and $G^{\text{ME}}_{\boldsymbol{w}_{k_2}}$ corresponding to two weight vectors $\boldsymbol{w}_{k_1}$ and $\boldsymbol{w}_{k_2}$ with $k_1 > k_2$. By (17), the ME grammar $G^{\text{ME}}_{\boldsymbol{w}_{k_1}}$ assigns a probability smaller than ϵ_1 to the mapping $(\mathsf{x}_{k_2}, \mathsf{y})$, namely $G^{\text{ME}}_{\boldsymbol{w}_{k_1}}(\mathsf{y} \,|\, \mathsf{x}_{k_2}) \leq \epsilon_1$. Furthermore, by (18), the ME grammar $G^{\text{ME}}_{\boldsymbol{w}_{k_2}}$ instead assigns a probability larger than ϵ_2 to this mapping $(\mathsf{x}_{k_2}, \mathsf{y})$, namely $G^{\text{ME}}_{\boldsymbol{w}_{k_2}}(\mathsf{y} \,|\, \mathsf{x}_{k_2}) \geq \epsilon_2$. Since ϵ_1 and ϵ_2 are more than ϵ apart, we conclude that these two ME grammars $G^{\text{ME}}_{\boldsymbol{w}_{k_1}}$ and $G^{\text{ME}}_{\boldsymbol{w}_{k_2}}$ are not ϵ-identical because they assign to the mapping $(\mathsf{x}_{k_2}, \mathsf{y})$ two probabilities that differ by more than ϵ. In conclusion, the ME typology $\mathfrak{T}^{\text{ME}}(\mathfrak{D}, \boldsymbol{C})$ corresponding to the phonological domain $\mathfrak{D}$ in (15) and the constraint set $\boldsymbol{C}$ in (16) is ϵ-infinite because it contains an infinite sequence

of ME grammars $G^{\text{ME}}_{\boldsymbol{w}_1}, G^{\text{ME}}_{\boldsymbol{w}_k}, \ldots, G^{\text{ME}}_{\boldsymbol{w}_k}, \ldots$ which are pair-wise ϵ-different.

5.3 Generalization of result 1 to other distances

The distance D_∞ in (3) is obviously never larger than the distance D_1 in (4), as stated in (20a). Furthermore, we recall (see for instance Cover and Thomas 1991, page 300 and Tsybakov 2009, lemma 2.5, page 88) that the distance D_1 is never larger than twice the square root of the KL divergence D_{KL} in (5), as stated by **Pinsker's inequality** (20b). Finally, we recall (see for instance Tsybakov 2009, lemma 2.7, page 90) that the KL divergence D_{KL} is never larger than the χ^2 divergence D_{χ^2} in (6), yielding the inequality (20c).

$$\begin{aligned} D_\infty(G_1, G_2) &\overset{(a)}{\leq} D_1(G_1, G_2) \\ &\overset{(b)}{\leq} 2\sqrt{D_{\text{KL}}(G_1, G_2)} \\ &\overset{(c)}{\leq} 2\sqrt{D_{\chi^2}(G_1, G_2)} \end{aligned} \tag{20}$$

It follows from these inequalities (20) that, if a probabilistic typology is ϵ-**in**finite relative to the distance D_∞, then it is also ϵ-**in**finite relative to the distance D_1 as well as relative to the divergences D_{KL} and D_{χ^2}. Since the ME typology $\mathfrak{T}^{\text{ME}}(\mathfrak{D}, \boldsymbol{C})$ constructed in appendix 5.2 is ϵ-infinite relative to D_∞, it is also ϵ-infinite relative to D_1, D_{KL}, and D_{χ^2}. In other words, the result proved in appendix 5.2 is robust because it does not depend on how we measure the difference between probabilistic grammars.

5.4 A lemma for the proof of result 2

Lemma 2 *Consider $k-1$ vectors $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{k-1}$ with integral components (without restrictions on their sign) and a vector $\boldsymbol{w}_k$ with positive rational components such that $\boldsymbol{w}_k \cdot \boldsymbol{d}_1 > 0, \ldots, \boldsymbol{w}_k \cdot \boldsymbol{d}_{k-1} > 0$. There exist a vector $\boldsymbol{d}_k$ with integral components (without restrictions on their sign) and a vector $\boldsymbol{w}_{k+1}$ with positive rational components such that $\boldsymbol{w}_{k+1} \cdot \boldsymbol{d}_1 > 0, \ldots, \boldsymbol{w}_{k+1} \cdot \boldsymbol{d}_{k-1} > 0$ and furthermore $\boldsymbol{w}_{k+1} \cdot \boldsymbol{d}_k > 0$ while $\boldsymbol{w}_k \cdot \boldsymbol{d}_k < 0$.* □

This lemma admits the following geometric interpretation. We start from some vectors $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{k-1}$, represented as blue dots in figure 7. They all sit in the interior of some half-space, represented as the blue region in figure 7a. We can always slightly tilt the surface that defines this half-space in such a way that the new half-space, represented as the

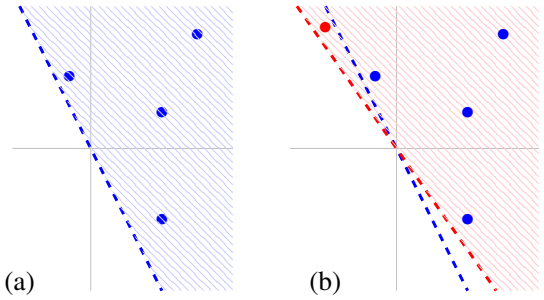

Figure 7

red region in figure 7b, satisfies the following two conditions. First, the original vectors $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_k$ sit in the interior of the tilted half-space as well. Second, we have made room for some new vector $\boldsymbol{d}_{k+1}$, represented by the red dot in figure 7b, that sits in the interior of the tilted red half-space but not of the original blue half-space.

To establish the lemma, we observe that, since the strict inequality $\boldsymbol{w}_k \cdot \boldsymbol{d}_h > 0$ holds for every $h = 1, \ldots, k-1$, there exists a positive rational constant $\epsilon > 0$ such that $\boldsymbol{w}_k \cdot \boldsymbol{d}_h \geq \epsilon$ for every $h = 1, \ldots, k-1$. Since the vector $\boldsymbol{w}_k$ has rational components, there exists a vector $\boldsymbol{v}$ with rational components orthogonal to $\boldsymbol{w}_k$, namely such that $\boldsymbol{v} \cdot \boldsymbol{w}_k = 0$. Let $M_1 > 0$ be the smallest common multiple of the denominators of the components of $\boldsymbol{v}$, whereby $M_1\boldsymbol{v}$ has integral components. Let $M_2 > 0$ be the smallest common multiple of the denominators of the components of $\boldsymbol{w}_k$, whereby $M_2\boldsymbol{w}_k$ has positive integral components. We choose a positive rational constant $\alpha > 0$ and a positive integer ℓ as in (21).

$$\begin{aligned} \alpha &= \begin{cases} 1 & \text{if } \beta \geq 0 \\ -\dfrac{\epsilon}{2\beta} & \text{if } \beta < 0 \end{cases} \quad \text{with } \beta = \min_{h=1}^{k-1} \boldsymbol{v} \cdot \boldsymbol{d}_h \\ \ell &\geq \frac{M_2 \|\boldsymbol{w}_k\|^2}{\alpha M_1 \|\boldsymbol{v}\|^2} \end{aligned} \tag{21}$$

We define the vector $\boldsymbol{w}_{k+1}$ with positive rational components and the vector $\boldsymbol{d}_k$ with integral components as in (22).

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \alpha\boldsymbol{v} \qquad \boldsymbol{d}_k = \ell M_1 \boldsymbol{v} - M_2 \boldsymbol{w}_k \tag{22}$$

These positions satisfy the inequalities (23) and (24) as well as the inequality (25) for $h =$

$1, \ldots, k-1$, completing the proof of the lemma.

$$\begin{aligned} \boldsymbol{w}_k \cdot \boldsymbol{d}_k &= \boldsymbol{w}_k \cdot \big(\ell M_1 \boldsymbol{v} - M_2 \boldsymbol{w}_k\big) \qquad (23) \\ &= -M_2 \|\boldsymbol{w}_k\|^2 < 0 \end{aligned}$$

$$\begin{aligned} \boldsymbol{w}_{k+1} \cdot \boldsymbol{d}_k &= \big(\boldsymbol{w}_k + \alpha \boldsymbol{v}\big) \cdot \big(\ell M_1 \boldsymbol{v} - M_2 \boldsymbol{w}_k\big) \qquad (24) \\ &= \alpha \ell M_1 \|\boldsymbol{v}\|^2 - M_2 \|\boldsymbol{w}_k\|^2 > 0 \end{aligned}$$

$$\begin{aligned} \boldsymbol{w}_{k+1} \cdot \boldsymbol{d}_h &= \big(\boldsymbol{w}_k + \alpha \boldsymbol{v}\big) \cdot \boldsymbol{d}_h \qquad (25) \\ &= \boldsymbol{w}_k \cdot \boldsymbol{d}_h + \alpha \boldsymbol{v} \cdot \boldsymbol{d}_h \geq \epsilon + \alpha\beta > 0 \end{aligned}$$

5.5 Proof of result 2

We start with an arbitrary vector $\boldsymbol{w}_1$ with positive rational components. By applying lemma 2 with $k = 1$ to this vector $\boldsymbol{w}_1$, we conclude that there exist a vector $\boldsymbol{d}_1$ with integral components and a vector $\boldsymbol{w}_2$ with positive rational components that validate the red inequalities in the first step of the reasoning in figure 8. By applying again lemma 2 with $k = 2$ to the vectors $\boldsymbol{d}_1$ and $\boldsymbol{w}_2$ in the bottom line of this first step, we conclude that there exist a vector $\boldsymbol{d}_2$ with integral components and a vector $\boldsymbol{w}_3$ with positive rational components that validate the red inequalities in the second step of the reasoning in figure 8. By applying once again lemma 2 with $k = 3$ to the vectors $\boldsymbol{d}_1, \boldsymbol{d}_2$ and $\boldsymbol{w}_3$ in the bottom line of this second step, we conclude that there exist a vector $\boldsymbol{d}_3$ with integral components and a vector $\boldsymbol{w}_4$ with positive rational components that validate the red inequalities in the third step of the reasoning in figure 8. And so on and so forth.

In conclusion, we have established the existence of a sequence of vectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k, \ldots$ with positive rational components and a sequence of vectors $\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_k, \ldots$ with integral components that satisfy the k inequalities in (26) for every index $k = 1, 2, \ldots$.

$$\begin{aligned} \boldsymbol{w}_k \cdot \boldsymbol{d}_1 &> 0 \qquad\qquad \boldsymbol{w}_k \cdot \boldsymbol{d}_k < 0 \qquad (26) \\ &\vdots \\ \boldsymbol{w}_k \cdot \boldsymbol{d}_{k-1} &> 0 \end{aligned}$$

To construct the desired counterexample, we consider the infinite phonological domain $\mathfrak{D}$ described in (27). For every index $k = 1, 2, \ldots$, the base set $B_{\mathfrak{D}}$ of this phonological domain contains the two underlying forms x_k and $\widehat{\mathsf{x}}_k$. Their candidate sets consist of only two surface forms, namely y, z and $\widehat{\mathsf{y}}, \widehat{\mathsf{z}}$, respectively. For concreteness, we refer to y and $\widehat{\mathsf{y}}$ as the **winner** candidate and to z and $\widehat{\mathsf{z}}$ as the **loser** candidates.

$$B_{\mathfrak{D}} = \left\{ \begin{matrix} \mathsf{x}_1\ \mathsf{x}_2\ \ldots\ \mathsf{x}_k\ \ldots \\ \widehat{\mathsf{x}}_1\ \widehat{\mathsf{x}}_2\ \ldots\ \widehat{\mathsf{x}}_k\ \ldots \end{matrix} \right\} \quad \begin{matrix} \mathfrak{D}(\mathsf{x}_k) = \{\mathsf{y}, \mathsf{z}\} \\ \mathfrak{D}(\widehat{\mathsf{x}}_k) = \{\widehat{\mathsf{y}}, \widehat{\mathsf{z}}\} \end{matrix} \qquad (27)$$

Furthermore, we define the constraint set $\boldsymbol{C}$ in such a way that the identity (28) holds for every index $k = 1, 2, \ldots$. The first difference $\boldsymbol{C}(\widehat{\mathsf{x}}_k, \widehat{\mathsf{z}}) - \boldsymbol{C}(\widehat{\mathsf{x}}_k, \widehat{\mathsf{y}})$ on the righthand side compares the constraint violations of the loser and winner candidates $\widehat{\mathsf{z}}$ and $\widehat{\mathsf{y}}$ of the underlying form $\widehat{\mathsf{x}}_k$. The second difference $\boldsymbol{C}(\mathsf{x}_k, \mathsf{z}) - \boldsymbol{C}(\mathsf{x}_k, \mathsf{y})$ compares the constraint violations of the loser and winner candidates z and y of the underlying form x_k that bears the same index k. The identity (28) says that the difference between these two differences must be equal to the vector $\boldsymbol{d}_k$ in (26).

$$\begin{aligned} \boldsymbol{d}_k &= \big(\boldsymbol{C}(\widehat{\mathsf{x}}_k, \widehat{\mathsf{z}}) - \boldsymbol{C}(\widehat{\mathsf{x}}_k, \widehat{\mathsf{y}})\big) \\ &\quad - \big(\boldsymbol{C}(\mathsf{x}_k, \mathsf{z}) - \boldsymbol{C}(\mathsf{x}_k, \mathsf{y})\big) \end{aligned} \qquad (28)$$

This position (28) makes sense because the vector $\boldsymbol{d}_k$ has integral components that can therefore be inetrpreted as differences between integral numbers of constraint violations. Furthermore, despite the fact that the components of this vector $\boldsymbol{d}_k$ can be positive or negative, the identity (28) can always be satisfied by choosing constraint violation vectors such that $\boldsymbol{C}(\widehat{\mathsf{x}}_k, \widehat{\mathsf{z}}) \geq \boldsymbol{C}(\widehat{\mathsf{x}}_k, \widehat{\mathsf{y}})$ and $\boldsymbol{C}(\mathsf{x}_k, \mathsf{z}) \geq \boldsymbol{C}(\mathsf{x}_k, \mathsf{y})$. This means that every constraint in the constraint set $\boldsymbol{C}$ assigns less violations to the winner mapping $(\mathsf{x}_k, \mathsf{y})$ than to the loser mapping $(\mathsf{x}_k, \mathsf{z})$; analogously, it assigns less violations to the winner mapping $(\widehat{\mathsf{x}}_k, \widehat{\mathsf{y}})$ than to the loser mapping $(\widehat{\mathsf{x}}_k, \widehat{\mathsf{z}})$. Equivalently, the winner mappings $(\mathsf{x}_k, \mathsf{y})$ and $(\widehat{\mathsf{x}}_k, \widehat{\mathsf{y}})$ always beat in HG the loser mappings $(\mathsf{x}_k, \mathsf{z})$ and $(\widehat{\mathsf{x}}_k, \widehat{\mathsf{z}})$ respectively, no matter the choice of the non-negative constraint weights. The HG typology $\mathfrak{T}^{\mathrm{HG}}(\mathfrak{D}, \boldsymbol{C})$ corresponding to the phonological domain $\mathfrak{D}$ in (27) and the constraint set $\boldsymbol{C}$ in (28) therefore consists of a single HG grammar, namely the grammar that maps all the underlying forms x_k and $\widehat{\mathsf{x}}_k$ to the candidates y and $\widehat{\mathsf{y}}$, respectively.

We now switch from categorical HG to probabilistic ME. We focus on the ME grammar $G^{\mathrm{ME}}_{\boldsymbol{w}_k}$ corresponding to the weight vector $\boldsymbol{w}_k$ in (26). We want to compare the probabilities it assigns to the two mappings $(\mathsf{x}_h, \mathsf{y})$ versus $(\widehat{\mathsf{x}}_h, \widehat{\mathsf{y}})$ with $h = 1, \ldots, k-1$ as well as to the two mappings $(\mathsf{x}_k, \mathsf{y})$ versus $(\widehat{\mathsf{x}}_k, \widehat{\mathsf{y}})$. As explained below, the inequalities $\boldsymbol{w}_k \cdot \boldsymbol{d}_h > 0$ with $h = 1, \ldots, k-1$ on the lefthand side of (26) ensure that the ME grammar $G^{\mathrm{ME}}_{\boldsymbol{w}_k}$ assigns less probability to the mapping $(\mathsf{x}_h, \mathsf{y})$ than to the mapping $(\widehat{\mathsf{x}}_h, \widehat{\mathsf{y}})$ for every $h = 1, \ldots, k-1$, as stated in (29). Analogously, the $\boldsymbol{w}_k \cdot \boldsymbol{d}_k < 0$ on the righthand side of (26) ensures

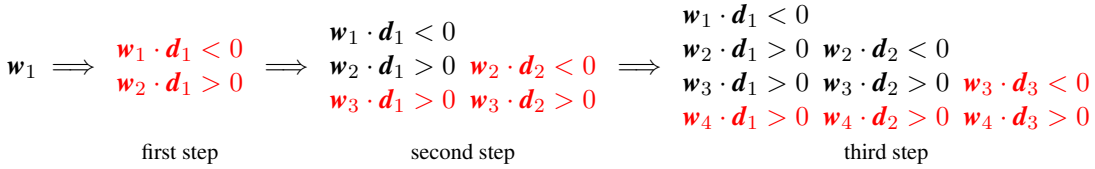

Figure 8

that the ME grammar $G^{\text{ME}}_{\boldsymbol{w}_k}$ assigns more probability to the mapping $(\mathsf{x}_k, \mathsf{y})$ than to the mapping $(\widehat{\mathsf{x}}_k, \widehat{\mathsf{y}})$, as stated in (30).

$$G^{\text{ME}}_{\boldsymbol{w}_k}(\mathsf{y} \,|\, \mathsf{x}_h) \;<\; G^{\text{ME}}_{\boldsymbol{w}_k}(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{x}}_h) \tag{29}$$

$$G^{\text{ME}}_{\boldsymbol{w}_k}(\mathsf{y} \,|\, \mathsf{x}_k) \;>\; G^{\text{ME}}_{\boldsymbol{w}_k}(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{x}}_k) \tag{30}$$

Indeed, the reasoning in (31) establishes the inequality (29) and an analogous reasoning establishes the inequality (30). Step (31a) holds by unpacking the ME probability as in steps (19a)-(19b) above. And tep (31b) holds because of the definition (28) of the vector $\boldsymbol{d}_h$. The condition $\boldsymbol{w}_k \cdot \boldsymbol{d}_h > 0$ arrived at is ensures by the choice of the vectors $\boldsymbol{w}_k$ and $\boldsymbol{d}_h$ in (26).

$$\begin{aligned}
&G^{\text{ME}}_{\boldsymbol{w}_k}(\mathsf{y} \,|\, \mathsf{x}_h) < G^{\text{ME}}_{\boldsymbol{w}_k}(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{x}}_h) \qquad (31)\\
&\overset{(a)}{\iff} \quad \frac{1}{1 + e^{-\boldsymbol{w}_k \cdot (\boldsymbol{C}(\mathsf{x}_k, \mathsf{z}) - \boldsymbol{C}(\mathsf{x}_k, \mathsf{y}))}} < \\
&\qquad\qquad < \frac{1}{1 + e^{-\boldsymbol{w}_k \cdot \left(\boldsymbol{C}(\widehat{\mathsf{x}}_k, \widehat{\mathsf{z}}) - \boldsymbol{C}(\widehat{\mathsf{x}}_k, \widehat{\mathsf{y}})\right)}}\\
&\iff \quad e^{-\boldsymbol{w}_k \cdot \left(\boldsymbol{C}(\mathsf{x}_k, \mathsf{z}) - \boldsymbol{C}(\mathsf{x}_k, \mathsf{y})\right)}\\
&\qquad\qquad > e^{-\boldsymbol{w}_k \cdot \left(\boldsymbol{C}(\widehat{\mathsf{x}}_k, \widehat{\mathsf{z}}) - \boldsymbol{C}(\widehat{\mathsf{x}}_k, \widehat{\mathsf{y}})\right)}\\
&\iff \quad \boldsymbol{w}_k \cdot \left(\boldsymbol{C}(\mathsf{x}_k, \mathsf{z}) - \boldsymbol{C}(\mathsf{x}_k, \mathsf{y})\right)\\
&\qquad\qquad < \boldsymbol{w}_k \cdot \left(\boldsymbol{C}(\widehat{\mathsf{x}}_k, \widehat{\mathsf{z}}) - \boldsymbol{C}(\widehat{\mathsf{x}}_k, \widehat{\mathsf{y}})\right)\\
&\overset{(b)}{\iff} \quad \boldsymbol{w}_k \cdot \boldsymbol{d}_h > 0
\end{aligned}$$

To complete the proof of result 2, we now consider the two ME grammars $G^{\text{ME}}_{\boldsymbol{w}_{k_1}}$ and $G^{\text{ME}}_{\boldsymbol{w}_{k_2}}$ corresponding to two weight vectors $\boldsymbol{w}_{k_1}$ and $\boldsymbol{w}_{k_2}$ with $k_1 > k_2$. By (29), the ME grammar $G^{\text{ME}}_{\boldsymbol{w}_{k_1}}$ assigns less probability to the mapping $(\mathsf{x}_{k_2}, \mathsf{y})$ than to the mapping $(\widehat{\mathsf{x}}_{k_2}, \widehat{\mathsf{y}})$, namely $G^{\text{ME}}_{\boldsymbol{w}_{k_1}}(\mathsf{y} \,|\, \mathsf{x}_{k_2}) < G^{\text{ME}}_{\boldsymbol{w}_{k_1}}(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{x}}_{k_2})$. By (30), the ME grammar $G^{\text{ME}}_{\boldsymbol{w}_{k_2}}$ instead assigns more probability to the mapping $(\mathsf{x}_{k_2}, \mathsf{y})$ than to the mapping $(\widehat{\mathsf{x}}_{k_2}, \widehat{\mathsf{y}})$, namely $G^{\text{ME}}_{\boldsymbol{w}_{k_2}}(\mathsf{y} \,|\, \mathsf{x}_{k_2}) > G^{\text{ME}}_{\boldsymbol{w}_{k_2}}(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{x}}_{k_2})$. These probability inequalities say that these two ME grammars $G^{\text{ME}}_{\boldsymbol{w}_{k_1}}$ and $G^{\text{ME}}_{\boldsymbol{w}_{k_2}}$ are not order-identical because they order the two mappings $(\mathsf{x}_{k_2}, \mathsf{y})$ and $(\widehat{\mathsf{x}}_{k_2}, \widehat{\mathsf{y}})$ differently. In conclusion, the ME typology $\mathfrak{T}^{\text{ME}}(\mathfrak{D}, \boldsymbol{C})$ corresponding to the phonological domain $\mathfrak{D}$ in (27) and the constraint set $\boldsymbol{C}$ in (28) is order-infinite because it contains an infinite sequence of ME grammars $G^{\text{ME}}_{\boldsymbol{w}_1}, G^{\text{ME}}_{\boldsymbol{w}_k}, \ldots, G^{\text{ME}}_{\boldsymbol{w}_k}, \ldots$ which are pair-wise order-different.