Equiprobable mappings in weighted constraint grammars

Arto Anttila
Stanford University
anttila@stanford.edu

Scott Borgeson
Stanford University
borgeson@stanford.edu

Giorgio Magri
CNRS
magrigrg@gmail.com

Abstract

We show that MaxEnt is so rich that it can distinguish between any two different mappings: there always exists a nonnegative weight vector which assigns them different MaxEnt probabilities. Stochastic HG instead does admit equiprobable mappings and we give a complete formal characterization of them. We compare these different predictions of the two frameworks on a test case of Finnish stress.

1 Introduction

This paper compares two frameworks for probabilistic constraint-based phonology: *Stochastic Harmonic Grammar* (SHG; Boersma and Pater, 2016)[1] and *Maximum Entropy* (ME; Goldwater and Johnson, 2003; Hayes and Wilson, 2008). Recent literature has documented a few realistic quantitative patterns which seem to admit a better fit in ME than in SHG (Smith and Pater, 2017; Zuraw and Hayes, 2017; Hayes, 2017). These findings suggest that ME is a richer probabilistic framework than SHG (relative to the same constraint set). But how much richer? Can these anecdotal observations reported in the literature be systematized into a principled formal comparison between SHG and ME probabilistic typologies? This paper is part of a larger project trying to address this question. In particular, this paper compares ME and SHG from the perspective of their *equiprobable mappings*. That is phonological mappings which are always assigned the same probability and are therefore phonologically equivalent despite being distinguished by the constraint set.

Section 2 motivates this notion of equiprobability within phonological theory. Section 4 then shows that the ME typology is so rich that it admits no equiprobable mappings: for any two mappings distinguished by the constraints, there exists an ME grammar that distinguishes between them, namely assigns them different probabilities. This typological richness is peculiar to ME and does *not* extend to other implementations of probabilistic constraint-based phonology such as SHG. Indeed, Section 5 shows that the equiprobable SHG mappings are exactly those mappings which are indistinguishable by categorical *Harmonic Grammars* (HG; Legendre *et al.*, 1990a,b; Smolensky and Legendre, 2006) and thus provides a complete characterization of SHG equiprobability.

These formal results are presented informally. A detailed proof of the ME result is provided in a final appendix. The proof of the SHG result is analogous and it is omitted for reasons of space (see the longer version of this paper available on the authors' websites). Our discussion rests on some earlier results on uniform SHG and ME probability inequalities from Anttila and Magri (2018), recalled in Section 3.

Is the richness of ME relative to SHG typologies an empirical advantage or a case of unmotivated overgeneration? Section 6 provides some preliminary evidence that the latter might be the case, by looking at the case of Finnish stress. We compute SHG equiprobable mappings using the formal characterization obtained in Section 5. We show that a large corpus of Finnish provides preliminary empirical support for these mappings indeed being equiprobable. Finally, we show that ME breaks up these equiprobabilities in a way that is phonologically counterintuitive.

2 Equiprobability

A typical phonological process applies uniformly to all forms that share some relevant property, but

[1] Boersma and Pater (2016) actually use the term "noisy HG" instead of "stochastic HG". We prefer "stochastic HG" to stress the complete analogy with Boersma's (1997; 1998) earlier framework of stochastic OT. Furthermore, we prefer to use "stochastic" to describe a property of the framework, reserving "noisy" to describe a property of the learning scenario (as opposed to noise-free).

Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 125–134
Florence, Italy. August 2, 2019 ©2019 Association for Computational Linguistics

ignores the irrelevant ways in which they differ. For example, in Latin, stress targets heavy syllables, but ignores vowel quality; in English, aspiration targets voiceless stops, but ignores place of articulation; in Finnish, vowel harmony targets [±back], but ignores the number of syllables. This means that words with the same distribution of heavy and light syllables are stressed alike; voiceless stops are aspirated alike; and words of any length harmonize alike. These phonological *equivalences* are a key property of phonological systems.

Derivational phonology captures these equivalences straightforwardly: phonological rules are allowed to refer to only the shared property that defines a natural class, ignoring everything else. To illustrate, the Finnish vowel harmony rule can be simply written as $\mathrm{V} \rightarrow [\alpha\mathrm{back}]/\mathrm{V}[\alpha\mathrm{back}]\mathrm{C}_0\_$. This rule directly encodes the fact that harmony targets [±back] but ignores any other properties such as, say, the number of syllables. Thus, the monosyllabic /maa/ 'country' and the disyllabic /kaava/ 'formula' trigger back harmony on the suffix /-nä/ 'ESSIVE' in exactly the same way. In other words, they are equivalent for vowel harmony.

The situation is *prima facie* less obvious in constraint-based phonology. A candidate may contain multiple constraint violations, some relevant, some irrelevant, but all simultaneously visible and potentially interacting. Yet, categorical implementations of constraint-based phonology are well known to readily predict these desired phonological equivalences. To illustrate, consider an HG grammar for Finnish vowel harmony based on the constraints in Table 1, from Ringen and Heinämäki (1999). The back harmony mappings /maa-nä/ $\rightarrow$ [maana] and /kaava-nä/ $\rightarrow$ [kaavana] can be shown to be HG equivalent: no matter the weighting, no HG grammar succeeds on one but fails on the other.

*INT[+back]:	No vowel between [+back] and right word edge
*INT[-back]:	No vowel between [-back] and right word edge
IDENT-ROOT:	Be faithful to /a, ä/ in roots
IDENT:	Be faithful to /a, ä/

Table 1: Constraints for Finnish vowel harmony

How should phonological equivalence be extended from the categorical to the probabilistic setting? We submit that equiprobability provides an answer to this question. In fact, let us recall that a *probabilistic phonological grammar* is a function which assigns to each underlying representation (UR) x a probability distribution $\mathbb{P}(\mathsf{y}\,|\,\mathsf{x})$ over the corresponding set of candidate surface representations (SRs) y. We consider two mappings (x, y) and $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ of the two URs $\mathsf{x}, \widehat{\mathsf{x}}$ to the two SRs $\mathsf{y}, \widehat{\mathsf{y}}$. We say that these two mappings are *(uniformly) equiprobable* provided there is no probabilistic grammar in the typology considered which assigns a different probability to those two mappings, namely such that $\mathbb{P}(\mathsf{y}\,|\,\mathsf{x}) \neq \mathbb{P}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})$. To illustrate, the equivalence between the two mappings /maa-nä/ $\rightarrow$ [maana] and /kaava-nä/ $\rightarrow$ [kaavana] is captured in a probabilistic setting through the requirement that their probabilities $\mathbb{P}(\text{[maana]}\,|\,\text{/maa-nä/})$ and $\mathbb{P}(\text{[kaavana]}\,|\,\text{/kaava-nä/})$ always coincide. In other words, the probability of vowel harmony does not depend on the number of syllables.[2]

As we will see in Section 5, two mappings are equivalent according to categorical HG if and only if they are equiprobable in SHG. This result suggests that equiprobability is indeed the right extension of the notion of phonological equivalence from the categorical to the probabilistic setting. Surprisingly, we will see in Section 4 that ME instead allows for no equiprobable mappings and thus fails to capture the notion of phonological equivalence.

3 Formal background

Our characterization of ME and SHG equiprobability in sections 4-5 rests on some results from Anttila and Magri (2018; A&M) recalled here.

HG A *weight vector* $\mathbf{w} = (w_1, \ldots, w_n)$ assigns nonnegative weights $w_1, \ldots, w_n \geq 0$ to n underlying phonological constraints $C_1, \ldots, C_n$. The phonological quality of a phonological mapping (x, y) of a UR x and a candidate SR y is quantified by its *harmony* $H_{\mathbf{w}}(\mathsf{x}, \mathsf{y})$. This quantity is defined as the weighted sum of the constraint vi-

[2] Note that this is quite different from the well-known case of Hungarian vowel harmony where suffixes show different degrees of back-front variation after stems with both back and neutral vowels depending on the number of neutral vowels; see, e.g., Hayes and Londe (2006), Hayes et al. (2009), and Zymet (2015). In our Finnish example, all the stem vowels are unambiguously back, yet our Proposition 1 below says that ME fails to guarantee that the suffix harmony is invariably back.

olations multiplied by -1, namely $H_{\mathbf{w}}(x, y) = -\sum_{k=1}^{n} w_k C_k(x, y)$. Mappings with large harmony have small constraint violations. The HG grammar corresponding to a weight vector $\mathbf{w}$ maps a UR x to the candidate SR y such that the mapping (x, y) has a larger harmony than the mapping (x, z) corresponding to any other candidate z of x. In this case, we say that y is the *winner* while any other candidate z is a *loser*.

HG thus has an intrinsic comparative nature: absolute numbers of violations are irrelevant, what matters is only the comparison between the violations of the loser and those of the winner. To bring out this intuition, we define the *difference vector* $\mathbf{C}(x, y, z)$ for a UR x, an intended winner candidate y, and an intended loser candidate z as in (1). This vector has a component for each constraint C_k defined as the difference between the number $C_k(x, z)$ of violations assigned by C_k to the loser mapping (x, z) minus the number $C_k(x, y)$ of violations assigned to the winner mapping (x, y).

$$\mathbf{C}(x, y, z) = \begin{bmatrix} C_1(x, z) - C_1(x, y) \\ \vdots \\ C_k(x, z) - C_k(x, y) \\ \vdots \\ C_n(x, z) - C_n(x, y) \end{bmatrix} \qquad (1)$$

SHG and ME are two probabilistic extensions of this underlying categorical HG model.

SHG The SHG probability $\mathbb{P}^{\text{SHG}}_{\mathbf{w}}(y \mid x)$ that a UR x is mapped to a SR y according to the weight vector $\mathbf{w}$ is the probability of sampling n numbers $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ independently according to a distribution $\mathcal{D}$ in such a way that the HG grammar corresponding to the weight vector $\mathbf{w} + \boldsymbol{\epsilon} = (w_1 + \epsilon_1, \ldots, w_n + \epsilon_n)$ indeed maps x to y. A&M prove the following Lemma 1 about *uniform* probability inequalities in SHG, namely inequalities which hold for every choice of the weight vector.

Lemma 1 *Consider two mappings (x, y) and $(\widehat{x}, \widehat{y})$. Assume that the UR x comes with only a finite number m of loser candidates $z_1, \ldots, z_m$ (besides the winner candidate y) and that the mapping (x, y) is possible in HG (namely, y beats the losers $z_1, \ldots, z_m$ relative to some nonnegative weight vector). The SHG probability inequality $\mathbb{P}^{SGH}_{\mathbf{w}}(y \mid x) \leq \mathbb{P}^{SGH}_{\mathbf{w}}(\widehat{y} \mid \widehat{x})$ holds uniformly for every choice of the nonnegative weight vector $\mathbf{w}$ if and only if for every loser candidate $\widehat{z}$ of the UR $\widehat{x}$, there exist m nonnegative coefficients $\lambda_1, \ldots, \lambda_m \geq 0$ (one for each loser candidate $z_1, \ldots, z_m$ of the UR x) such that*

$$\boldsymbol{C}(\widehat{x}, \widehat{y}, \widehat{z}) \geq \sum_{i=1}^{m} \lambda_i \boldsymbol{C}(x, y, z_i) \qquad (2)$$

namely the difference vector $\boldsymbol{C}(\widehat{x}, \widehat{y}, \widehat{z})$ is at least as large (constraint by constraint) as the sum of the difference vectors $\boldsymbol{C}(x, y, z_i)$ each rescaled by a corresponding nonnegative coefficient λ_i.[3] □

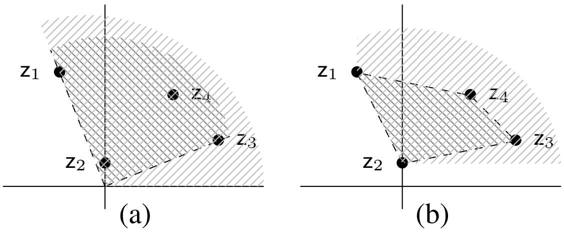


Figure 1: Geometric representation of (a) the SHG Lemma 1 and (b) the ME Lemma 2.

Lemma 1 admits the following geometric interpretation, which will be used below. Suppose there are only $n = 2$ constraints and $m = 4$ losers z_i. The difference vectors $\mathbf{C}(x, y, z_i)$ which appear on the right hand side of (2) can therefore be represented as the four black dots in Fig. 1. The region $\{\sum_{i=1}^{m} \lambda_i \mathbf{C}(x, y, z_i) \mid \lambda_i \geq 0\}$ is the *convex cone* generated by these four difference vectors $\mathbf{C}(x, y, z_i)$, depicted in dark gray in Fig. 1a. The region in light gray singles out the points which are at least as large as some point in this cone. Condition (2) thus says that the difference vector $\mathbf{C}(\widehat{x}, \widehat{y}, \widehat{z})$ belongs to this light gray region.

ME The ME probability $\mathbb{P}^{\text{ME}}_{\mathbf{w}}(y \mid x)$ that a UR x is mapped to a SR y according to a nonnegative weight vector $\mathbf{w}$ is the exponential of the harmony

[3] The two assumptions made by the lemma—-that the UR x comes with only a finite number of losers and that the mapping (x, y) is possible in HG—-are non-restrictive. In fact, if a mapping (x, y) is impossible in HG, then its SHG probability $\mathbb{P}^{\text{SGH}}_{\mathbf{w}}(y \mid x)$ can be shown to be equal to zero for every choice of the nonnegative weight vector $\mathbf{w}$. The probability inequality $\mathbb{P}^{\text{SGH}}_{\mathbf{w}}(y \mid x) \leq \mathbb{P}^{\text{SGH}}_{\mathbf{w}}(\widehat{y} \mid \widehat{x})$ thus holds uniformly, because its left hand side is always equal to zero. The assumption made by the lemma that the mapping (x, y) is possible in HG is therefore non-restrictive. Furthermore, HG has the property that only a finite number of candidates of any given UR win according to some weights (Magri, 2019). All other candidates are redundant because impossible no matter how the weights are chosen. Since HG impossible mappings have zero SHG probability, the candidate set of any underlying form can always be assumed to be finite without loss of generality in SHG. The assumption made by the lemma that the UR x comes with only a finite number of losers is therefore non-restrictive.

$H_{\mathbf{w}}(\mathsf{x},\mathsf{y})$ of that mapping, normalized through a constant $Z = Z(\mathbf{w},\mathsf{x})$, namely $\mathbb{P}^{\text{ME}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x}) = e^{H_{\mathbf{w}}(\mathsf{x},\mathsf{y})}/Z$. A&M show that also in ME uniform probability inequalities can be characterized in terms of difference vectors, as stated by Lemma 2 below. This ME Lemma is analogous to the SHG Lemma 1 above, but for two differences. The first difference is that condition (2) is only necessary in ME while it is also sufficient in SHG. The second difference is that ME requires the *normalization condition* (3) on the coefficients λ_i.

Lemma 2 *Consider two mappings* (x,y) *and* $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$. *Assume that the UR* x *comes with a finite number* m *of loser candidates* $\mathsf{z}_1,\ldots,\mathsf{z}_m$ *(besides the winner candidate* y*). If the ME probability inequality* $\mathbb{P}^{ME}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x}) \leq \mathbb{P}^{ME}_{\mathbf{w}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})$ *holds uniformly for every choice of the nonnegative weight vector* $\mathbf{w}$, *then for every loser candidate* $\widehat{\mathsf{z}}$ *of the UR* $\widehat{\mathsf{x}}$, *there exist* m *nonnegative coefficients* $\lambda_1,\ldots,\lambda_m \geq 0$ *(one for each loser candidate* $\mathsf{z}_1,\ldots,\mathsf{z}_m$ *of the UR* x*) which add up to 1*

$$\lambda_1 + \ldots + \lambda_m = 1 \tag{3}$$

and furthermore satisfy condition (2). □

The normalization condition (3) admits the following geometric interpretation. As seen above, the region $\{\sum_i \lambda_i \mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}_i)\,|\,\lambda_i \geq 0\}$ is the convex cone generated by the difference vectors $\mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}_i)$, represented by the dark gray region in Fig. 1a. The smaller region $\{\sum_i \lambda_i \mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}_i)\,|\,\lambda_i \geq 0, \boxed{\sum_i \lambda_i = 1}\}$, which differs for the (boxed) normalization condition (3) on the coefficients λ_i, is instead the *convex hull* generated by the difference vectors $\mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}_i)$, represented by the smaller dark gray region in Fig. 1b. The effect of the normalization condition (3) is thus to shrink from the larger convex cone to the smaller convex hull. Finally, the region in light gray in Fig. 1b singles out the points which are at least as large as some point in this convex hull. Lemma 2 thus requires the difference vector $\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}},\widehat{\mathsf{z}})$ to belong to this light gray region.

4 ME has no equiprobable mappings

Lemmas 1 and 2 say that ME differs from SHG because of the normalization condition (3). This apparently small technical difference has substantial phonological implications. Indeed, this Section shows that the normalization condition (3) makes the ME typology so rich that it can distinguish between any two mappings. In other words, equiprobability is impossible in ME. The reasoning is presented here informally, split up into three steps formalized in the final appendix.

Step 1 Let us suppose that the two mappings (x,y) and $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ are equiprobable in ME, namely that the ME probability identity $\mathbb{P}^{\text{ME}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x}) = \mathbb{P}^{\text{ME}}_{\mathbf{w}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})$ holds for every choice of the nonnegative weight vector $\mathbf{w}$. Let $\mathsf{z}_1,\ldots,\mathsf{z}_m$ be the loser candidates of the UR x. They define a light gray region as in Fig. 1b, namely the region of points which are at least as large as the points in the convex hull generated by the difference vectors $\mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}_i)$. Let us denote this light gray region as $LGR^{\text{ME}}(\mathsf{z}_1,\ldots,\mathsf{z}_m)$. Analogously, let $\widehat{\mathsf{z}}_1,\ldots,\widehat{\mathsf{z}}_{\widehat{m}}$ be the loser candidates of the other UR $\widehat{\mathsf{x}}$. They as well define the light gray region of points which are at least as large as the points in the convex hull generated by the difference vectors $\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}},\widehat{\mathsf{z}}_j)$. Let us denote this light gray region as $LGR^{\text{ME}}(\widehat{\mathsf{z}}_1,\ldots,\widehat{\mathsf{z}}_{\widehat{m}})$.

The probability identity $\mathbb{P}^{\text{ME}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x}) = \mathbb{P}^{\text{ME}}_{\mathbf{w}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})$ is equivalent to the two reverse inequalities $\mathbb{P}^{\text{ME}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x}) \leq \mathbb{P}^{\text{ME}}_{\mathbf{w}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})$ and $\mathbb{P}^{\text{ME}}_{\mathbf{w}}(\mathsf{y}\,|\,\mathsf{x}) \geq \mathbb{P}^{\text{ME}}_{\mathbf{w}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})$. By lemma 2 above, the former inequality requires each difference vector $\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}},\widehat{\mathsf{z}}_j)$ to belong to $LGR^{\text{ME}}(\mathsf{z}_1,\ldots,\mathsf{z}_m)$. And the latter inequality requires each difference vector $\mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}_i)$ to belong to $LGR^{\text{ME}}(\widehat{\mathsf{z}}_1,\ldots,\widehat{\mathsf{z}}_{\widehat{m}})$. A simple convexity argument deduces from these two facts the identity $LGR^{\text{ME}}(\mathsf{z}_1,\ldots,\mathsf{z}_m) = LGR^{\text{ME}}(\widehat{\mathsf{z}}_1,\ldots,\widehat{\mathsf{z}}_{\widehat{m}})$ between the two light gray regions.

Step 2 To proceed, let us suppose for concreteness that $m = 4$ and that the light gray region $LGR^{\text{ME}}(\mathsf{z}_1,\mathsf{z}_2,\mathsf{z}_3,\mathsf{z}_4)$ is the one plotted in light gray in Fig. 1b. The difference vectors corresponding to the two losers z_1 and z_2 are *extreme points* (or *vertices*) of this light gray region. In the sense that they crucially contribute to shape it: if these two points were shifted even slightly in any direction, the corresponding light gray region would change. The identity between the two light gray regions established in step 1 thus entails that the two light gray regions share the same set of extreme points. In conclusion, the two difference vectors corresponding to losers z_1 and z_2 which are extreme points of the light gray region in figure Fig. 1b must be shared by the two equiprobable mappings considered. Since these difference vectors are shared by the two equiprobable mappings, they can be "peel off" the two sides of the

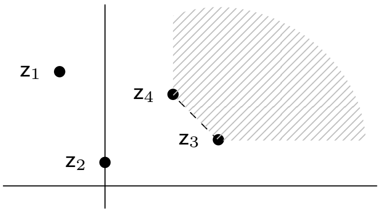


Figure 2: Steps 1-2 for the remaining losers z_3 and z_4.

ME probability identity.

Step 3 We are thus left with the difference vectors corresponding to the other two losers z_3 and z_4 in Fig. 1b. These latter two vectors are not extreme points of the original light gray region but rather sit in the interior of the light gray region. Indeed, they can be shifted around without affecting the shape of the light gray region. Yet, once the two losers z_1 and z_2 have been "peeled off" at step 2, we can repeat the reasoning in steps 1 and 2 ignoring the two losers z_1 and z_2 and instead considering only the other two losers z_3 and z_4.

Thus, we construct the convex hull of the difference vectors corresponding to just these two remaining losers z_3 and z_4. This convex hull is the segment which connects the two corresponding dots. Next, we construct the light gray region of points which are at least as large as some point in that segment, as depicted in Fig. 2. Now the difference vectors corresponding to the two losers z_3 and z_4 are extreme points of the new light gray region. We can therefore repeat the reasoning in steps 1-2 and conclude that these two difference vectors as well must be shared by the two equiprobable mappings considered. And so on.

The reasoning informally sketched above leads to the following Proposition 1, which is the first main result of this paper. It says that two mappings are equiprobable in ME if and only if they share all difference vectors. This entails in particular that the two mappings must have the same number of loser candidates. In other words, the ME typology is so rich that the only case where ME fails to come up with at least one weight vector which assigns different probabilities to the two mappings (x, y) and $(\widehat{x}, \widehat{y})$ is when the two mappings are the same mapping, in the sense that they are indistinguishable by the constraints, as they have the same difference vectors.[4]

Proposition 1 *Two mappings* (x, y) *and* $(\widehat{x}, \widehat{y})$ *are equiprobable in ME if and only if the corresponding sets of difference vectors coincide.* □

5 SHG allows for equiprobable mappings

The preceding Section has shown that ME is so rich that it can distinguish between any two different mappings. Crucially, this typological richness is peculiar to ME, not intrinsic to probabilistic constraint-based phonology. In this section, we illustrate this point with the case of SHG. As in the preceding section, the discussion is kept informal. The formalization rests on the same convex geometric tools used for ME in the final appendix. The details are omitted here for reasons of space (see the longer version of this paper available on the authors' website).

Let us consider two mappings (x, y) and $(\widehat{x}, \widehat{y})$. Again, let $z_1, \dots, z_m$ be the loser candidates of the UR x. They define a light gray region as in Fig. 1a, namely the region of points which are at least as large as the points in the convex cone generated by the difference vectors $\mathbf{C}(x, y, z_i)$. Let us denote this light gray region as $LGR^{\text{SHG}}(z_1, \dots, z_m)$. This region is different from (and larger than) the light gray region $LGR^{\text{ME}}(z_1, \dots, z_m)$ considered above for ME, because the latter ME region is restricted through the normalization condition (3) and therefore defined in terms of convex hulls rather than convex cones. Analogously, let $\widehat{z}_1, \dots, \widehat{z}_{\widehat{m}}$ be the loser candidates of the other UR $\widehat{x}$ and let $LGR^{\text{SHG}}(\widehat{z}_1, \dots, \widehat{z}_{\widehat{m}})$ be the corresponding SHG light gray region.

Again as in the case of ME, Lemma 1 says that the uniform SHG probability identity $\mathbb{P}^{\text{SHG}}_{\mathbf{w}}(y \mid x) = \mathbb{P}^{\text{SHG}}_{\mathbf{w}}(\widehat{y} \mid \widehat{x})$ entails that the two SHG light gray regions coincide, namely that $LGR^{\text{SHG}}(z_1, \dots, z_m) = LGR^{\text{SHG}}(\widehat{z}_1, \dots, \widehat{z}_{\widehat{m}})$. Yet, these SHG light gray regions have different geometric properties than the ME light gray regions. As a result, in the case of SHG the identity between the two light gray regions tells us much less about the difference vectors that generate them than in the case of ME.

To see that concretely, let us consider for instance the SHG light gray region in Fig. 1a. The loser candidates z_2, z_3 and z_4 have difference vectors which sit in the interior of this light gray region. These losers thus contribute nothing to shape

[4] To illustrate, suppose that the constraint set only consists of the two constraints NoVoicedObstruent and Ident(voice). The mappings $(x, y) = (\text{/mab/}, \text{[map]})$ and $(\widehat{x}, \widehat{y}) = (\text{/bam/}, \text{[pam]})$ will always have the same ME probability, because they and their losers have the same constraint violation profiles.

the light gray region: their difference vectors can be shifted around without affecting the shape of the region. Identity of the light gray regions thus tells us nothing about identity of these difference vectors which sit in the interior.

Interestingly, the loser candidates whose difference vectors sit in the interior of the SHG light gray region can be characterized phonologically as those losers which are *HG redundant* given the rest of the losers. In the sense that, for every nonnegative weight vector $\mathbf{w}$, if the HG harmony of the winner y is larger than that of the nonredundant losers, then it is in particular larger than the harmony of the redundant losers. In other words, these redundant losers carry no interesting phonological content as they do not in any way affect the weight vectors consistent with the mapping (x, y).

The case of the loser z_1 in Fig. 1a is instead different. Its difference vector sits on the border of the light gray region and therefore contributes to its shape. Yet, its position is not completely determined by the shape of the region. In fact, the shape of the region is not affected if this difference vector is slid closer to or further away from the origin. Equivalently, the shape of the region is not affected if the difference vector corresponding to the nonredundant loser z_1 is rescaled by a nonnegative constant $\lambda \geq 0$. This means that the identity of the two SHG light gray regions does not entail identity of the difference vectors which generate them, not even for those difference vectors which sit on the boundary of the regions and therefore correspond to nonredundant losers. The identity of the two SHG light gray regions only entails that the difference vectors of the nonredundant losers are one the rescaling of the other. This informal reasoning leads to the following Proposition, which is our second main result.

Proposition 2 *Two mappings* (x, y) *and* $(\widehat{x}, \widehat{y})$ *are equiprobable in SHG if and only if each nonredundant difference vector* $\boldsymbol{C}(\mathsf{x}, \mathsf{y}, \mathsf{z}_i)$ *is a rescaling of some nonredundant difference vector* $\boldsymbol{C}(\widehat{x}, \widehat{y}, \widehat{z}_j)$*, namely* $\boldsymbol{C}(\mathsf{x}, \mathsf{y}, \mathsf{z}_i) = \lambda \boldsymbol{C}(\widehat{x}, \widehat{y}, \widehat{z}_j)$ *for some* $\lambda \geq 0$*; analogously, each nonredundant difference vector* $\boldsymbol{C}(\widehat{x}, \widehat{y}, \widehat{z}_j)$ *is a rescaling of some nonredundant difference vector* $\boldsymbol{C}(\mathsf{x}, \mathsf{y}, \mathsf{z}_i)$. □

Interestingly, this characterization of SHG equiprobability coincides with the characterization of equivalence in categorical HG obtained by A&M. We conclude that two mappings are equiprobable in SHG (namely are always assigned the same probability) if and only if they are equivalent in categorical HG (namely no HG grammar succeeds on one but fails on the other).

FTBIN	Feet are disyllabic.
PKPROM	No unstressed light syllables.
ALIGN-L	All feet left.
*REV	No trochees with sonority reversal.
*FLAT	No trochees with a flat sonority profile.
*H.X	No stress next to a heavy syllable.
WSP	No unstressed heavy syllables.
WSP/VV	No unstressed heavies with long vowel.

Table 2: Constraints for foot structure in Finnish nouns

6 Equiprobability in Finnish stress

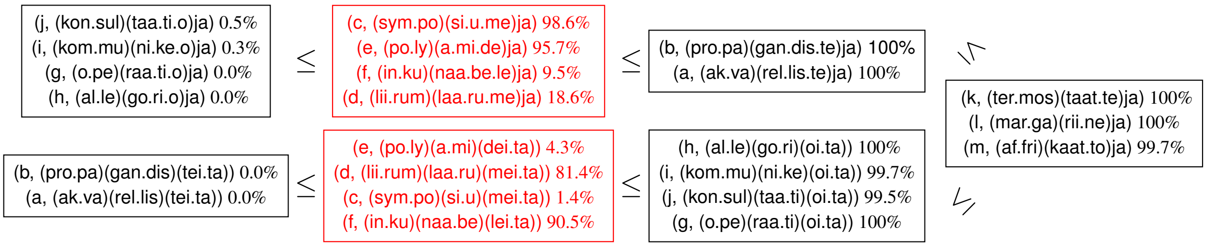
This section brings the preceding formal results to bear on Finnish word stress.

The phonological system The basic generalizations about Finnish word stress can be stated as follows (Carlson, 1978; Hanson and Kiparsky, 1996; Elenbaas, 1999; Elenbaas and Kager, 1999; Karvonen, 2005): (a) primary stress falls on the initial syllable; (b) secondary stress falls on every other syllable after that, (c) except that a light syllable is skipped if the syllable after that is heavy, unless the heavy syllable is final. Examples are íl.moit.tàu.tu.mì.nen 'registering' and íl.moit.tàu.tu.mi.sès.ta 'from registering'.

However, the skipping clause turns out to be a coarse approximation of the actual facts. Skipping is sometimes optional and we find variable stress in cases like pró.fes.so.rìl.la∼pró.fes.sò.ril.la 'professor-ADE', where the basic rule fails at the second variant. This optional pattern turns out to depend on two additional conditions that affect the outcome in a gradient manner (Anttila, 2012): (a) low vowels (/a, ä, o, ö/) attract stress and high vowels (/e, i, u, y/) repel stress; (b) stress is avoided next to a heavy syllable.[5]

In addition to native speaker intuitions about syllable prominence, empirical support for these soft conditions can be obtained from the optional rule of *Stop Deletion* (Keyser and Kiparsky, 1984) which deletes singleton stops in extrametrical syllables (Anttila, 2012). In particular, the /t/ in the partitive suffix /-tA/ is deleted vs. retained

[5] The categories "low" and "high" are morphophonemic, not phonetic. In Finnish, low vowels alternate morphophonologically with rounded mid vowels (a ∼ o, ä ∼ ö) and the unrounded high vowel alternates with the unrounded mid vowel (i ∼ e). For this reason we consider o, ö low and e high.



(j, (kon.sul)(taa.ti.o)ja) 0.5%
(i, (kom.mu)(ni.ke.o)ja) 0.3%
(g, (o.pe)(raa.ti.o)ja) 0.0%
(h, (al.le)(go.ri.o)ja) 0.0%
≤
(c, (sym.po)(si.u.me)ja) 98.6%
(e, (po.ly)(a.mi.de)ja) 95.7%
(f, (in.ku)(naa.be.le)ja) 9.5%
(d, (lii.rum)(laa.ru.me)ja) 18.6%
≤
(b, (pro.pa)(gan.dis.te)ja) 100%
(a, (ak.va)(rel.lis.te)ja) 100%
≤
(k, (ter.mos)(taat.te)ja) 100%
(l, (mar.ga)(rii.ne)ja) 100%
(m, (af.fri)(kaat.to)ja) 99.7%

(b, (pro.pa)(gan.dis)(tei.ta)) 0.0%
(a, (ak.va)(rel.lis)(tei.ta)) 0.0%
≤
(e, (po.ly)(a.mi)(dei.ta)) 4.3%
(d, (lii.rum)(laa.ru)(mei.ta)) 81.4%
(c, (sym.po)(si.u)(mei.ta)) 1.4%
(f, (in.ku)(naa.be)(lei.ta)) 90.5%
≤
(h, (al.le)(go.ri)(oi.ta)) 100%
(i, (kom.mu)(ni.ke)(oi.ta)) 99.7%
(j, (kon.sul)(taa.ti)(oi.ta)) 99.5%
(g, (o.pe)(raa.ti)(oi.ta)) 100%
≤
(k, (ter.mos)(taat.te)ja) 100%
(l, (mar.ga)(rii.ne)ja) 100%
(m, (af.fri)(kaat.to)ja) 99.7%

Table 3: Seven blocks of equiprobable mappings predicted by SHG

(c, (sym.po)(si.u)(mei.ta)) 1.4%	≤	(e, (po.ly)(a.mi)(dei.ta)) 4.3%	≤	(d, (lii.rum)(laa.ru)(mei.ta)) 81.4%	≤	(f, (in.ku)(naa.be)(lei.ta)) 90.5%
(c, (sym.po)(si.u.me)ja) 98.6%	≤	(e, (po.ly)(a.mi.de)ja) 95.7%	≤	(d, (lii.rum)(laa.ru.me)ja) 18.6%	≤	(f, (in.ku)(naa.be.le)ja) 9.5%

Table 4: SHG's two red blocks are split into two chains of uniform inequalities in ME

depending on the location of secondary stress feet. Given the input /professori-i-tA/ 'professor-PL-PAR' we have two possible foot structures: (pró.fes.so)(rèi.ta) where /t/ falls inside a foot and is retained vs. (pró.fes)(sò.re)ja where /t/ falls outside a foot and is deleted. The metrical free variation is thus reflected in segmental free variation. This provides a valuable diagnostic for foot structure, especially because the segmental variation is present even in the written standard language readily available in large quantities.

The constraints necessary for deriving the foot structure in Finnish nouns are shown in Table 2. These constraints were applied to 48 types of partitive plural nouns, systematically varying stem length, syllable weight, and vowel sonority. All in all, the test set contains 4 types of three-syllable stems, 12 types of 4-syllable stems, and 32 types 5-syllable stems (stem types are briefly denoted as "(a), (b), . . . " in what follows).

SHG We computed the uniform probability inequalities predicted by SHG for this Finnish stress test case using ℂo𝔾e𝕋o (Magri and Anttila, 2019), a suite of tools for studying constraint-based typologies of categorical and probabilistic phonological grammars based on their underlying rich convex geometry. The key observation is that SHG predicts seven blocks of equiprobable mappings, shown in Table 3. These blocks are furthermore organized into two chains of uniform probability inequalities. The predicted probabilities increase from left to right. The symbol "≤" between two boxes means that the candidates in the box on the left are predicted to have a probability at most as large as the candidates in the box on the right.

To evaluate the empirical accuracy of the equiprobabilities predicted by SHG, we examined Finnish /t/-deletion in a corpus of approximately 9 million nouns (tokens) harvested from Finnish internet sites on April 12, 2005. The percentages reported in Table 3 represent the token frequency of /t/-retention vs. /t/-deletion variants for each phonologically distinct stem type. The corpus data are consistent with the equiprobability prediction in five out of the seven blocks, namely those in black. These blocks turn out to be empirically nearly categorical, with almost all stems undergoing either /t/-deletion or /t/-retention, consistently with the equiprobability prediction.

However, the two red blocks in Table 3 bundle together the stem types (c)-(f) despite them showing rather different empirical frequencies, providing *prima facie* evidence against SHG's equiprobability prediction. The stem types are illustrated by /symposiumi/ 'symposium', /polyamidi/ 'polyamide', /liirumlaarumi/ 'nonsense', and /inkunaabeli/ 'incunable'. The stems differ in the weight and quality of the preantepenultimate and antepenultimate syllables (heavy vs. light, [+low] vs. [−low]), which results in constraint violation differences, yet HG predicts that all four should undergo /t/-deletion/retention at identical rates. In order to reconcile SHG's equiprobability predicitions with corpus frequencies, we make the following observations. First, the difference between types (d) /liirumlaarumi/ and (f) /inkunaabeli/ is not statistically significant (χ^2 = 2.9849, *df* = 1, p = 0.08404). Second, type (c) contains only two stems: /symposiumi/ 'symposium' and /imperiumi/ 'empire', both potentially syllabifiable as four-syllable stems, e.g., im.pe.ri.u.mi ∼ im.pe.riu.mi (Anttila and Shapiro, 2017), which is consistent with their unexpectedly high /t/-deletion rate. This leaves us with type (e) /polyamidi/ 'polyamide' (N = 69), again with an unexpectedly high deletion rate for which we have no plausible explanation. We conclude that by and large our Finnish corpus data support SHG's equiprobability predictions.

ME One might wonder whether ME with its ability to make fine-grained distinctions might actually offer a more principled solution to the difficulties just discussed. This turns out *not* to be the case. On the retention side, ME predicts the uniform probability inequalities in the top row of Table 4. For example, the retention probability of /polyamidi/ is predicted to be at most as high as that of /liirumlaarumi/, no matter the choice of the weight vector. That seems initially promising: these inequalities are in fact exactly what we observe in the data. Puzzlingly, on the deletion side, ME reverses the probabilities, yielding the uniform probability inequalities in the bottom row of Table 4. For example, the deletion probability of /polyamidi/ is predicted to be at most as high as that of /liirumlaarumi/. This is exactly the opposite of what we observe in the data. We submit there is simply no way to reconcile ME's predictions with the corpus data. Such counterintuitive probability reversals appear in other blocks as well.

7 Summary and conclusions

We have shown that ME predicts typologies so rich that ME grammars can distinguish between any two different mappings and therefore admit no equiprobable mappings (Proposition 1). This richness does not extend to other implementations of probabilistic constraint-based phonology, such as SHG (Proposition 2), revealing a fundamental difference between the two frameworks.

We have then applied these results to the test case of Finnish word stress. Our corpus data provide preliminary evidence in favor of SHG's equiprobability predictions. In the two blocks where SHG appeared to run into problems, ME did not help refine the analysis empirically, but instead split the SHG equiprobable stem types apart in a counterintuitive fashion. Our study thus provides some preliminary empirical support in favor of SHG, which permits equiprobable mappings, against ME, which does not.

Acknowledgements

For very useful discussion, we would like to thank the participants of the workshop *Analyzing typological structure: from categorical to probabilistic phonology* (Stanford, September 2018; website). The research reported in this paper has been supported by a Collaborative Projects Grant from the France-Stanford Center for Interdisciplinary Studies (project title: *The Mathematics of Language Universals*) as well as by a JCJC grant from the Agence Nationale de la Recherche (project title: *The mathematics of segmental phonotactics*).

A Proof of Proposition 1

We write $\mathbf{c}_i$ and $\widehat{\mathbf{c}}_j$ as shorthands for the difference vectors $\mathbf{C}(\mathsf{x}, \mathsf{y}, \mathsf{z}_i)$ and $\mathbf{C}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}, \widehat{\mathsf{z}}_j)$ corresponding to the losers z_i and $\widehat{\mathsf{z}}_j$. The ME probability inequality $\mathbb{P}^{\text{ME}}_{\mathbf{w}}(\mathsf{x}, \mathsf{y}) = \mathbb{P}^{\text{ME}}_{\mathbf{w}}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ can be made explicit as in (4) through some elementary manipulations. As usual, $\mathbf{a}^{\mathsf{T}}\mathbf{b}$ denotes the scalar product of $\mathbf{a}$ and $\mathbf{b}$.

$$\sum_{i=1}^{m} e^{\mathbf{w}^{\mathsf{T}}\mathbf{c}_i} = \sum_{j=1}^{\widehat{m}} e^{\mathbf{w}^{\mathsf{T}}\widehat{\mathbf{c}}_j} \qquad (4)$$

Once the ME probability identity $\mathbb{P}^{\text{ME}}_{\mathbf{w}}(\mathsf{x}, \mathsf{y}) = \mathbb{P}^{\text{ME}}_{\mathbf{w}}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is made explicit as in (4), it is obvious that it holds uniformly for every weight vector $\mathbf{w}$ when the two sets of difference vectors coincide, namely $\{\mathbf{c}_1, \ldots, \mathbf{c}_m\} = \{\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_{\widehat{m}}\}$. To complete the proof of Proposition 1, we thus only have to prove the reverse. We split the proof into three steps, corresponding to those in Section 4.

Step 1. We start from the assumption that the ME probability identity $\mathbb{P}^{\text{ME}}_{\mathbf{w}}(\mathsf{x}, \mathsf{y}) = \mathbb{P}^{\text{ME}}_{\mathbf{w}}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ holds uniformly. This means in particular that the probability inequality $\mathbb{P}^{\text{ME}}_{\mathbf{w}}(\mathsf{x}, \mathsf{y}) \leq \mathbb{P}^{\text{ME}}_{\mathbf{w}}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ holds uniformly. The necessary condition for this uniform ME inequality provided by Proposition 2 can be rewritten as the inclusion (1). As usual, $\mathbb{R}_+$ is the set of nonnegative real numbers and $A + B = \{\mathbf{a} + \mathbf{b} \,|\, \mathbf{a} \in A, \mathbf{b} \in B\}$ is the vector sum of two sets A and B of $\mathbb{R}^n$. The region on the right hand side of (1) is the light gray region in Fig. 3.b.

$$\text{(1)} \quad \{\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_{\widehat{m}}\} \subseteq conv(\mathbf{c}_1, \ldots, \mathbf{c}_m) + \mathbb{R}^n_+$$

The set $conv(\mathbf{c}_1, \ldots, \mathbf{c}_m) + \mathbb{R}^n_+$ on the right hand side of (1) is convex because the two sets $conv(\mathbf{c}_1, \ldots, \mathbf{c}_m)$ and $\mathbb{R}^n_+$ are both convex and the sum of two convex sets is convex (Boyd and Vandenberghe, 2004, Section 2.3.2). The inclusion (1) thus extends from the points $\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_{\widehat{m}}$ to their convex hull $conv(\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_{\widehat{m}})$, yielding the inclusion $conv(\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_{\widehat{m}}) \subseteq conv(\mathbf{c}_1, \ldots, \mathbf{c}_m) + \mathbb{R}^n_+$. Finally, by adding $\mathbb{R}^n_+$ at both sides, the latter inclusion entails $conv(\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_{\widehat{m}}) + \mathbb{R}^n_+ \subseteq conv(\mathbf{c}_1, \ldots, \mathbf{c}_m) + \mathbb{R}^n_+$. Analogously, the reverse probability inequality $\mathbb{P}^{\text{ME}}_{\mathbf{w}}(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}) \leq \mathbb{P}^{\text{ME}}_{\mathbf{w}}(\mathsf{x}, \mathsf{y})$ requires the reverse inclusion $conv(\mathbf{c}_1, \ldots, \mathbf{c}_m) + \mathbb{R}^n_+ \subseteq conv(\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_{\widehat{m}}) + \mathbb{R}^n_+$, yielding (5).

$$\underbrace{conv(\mathbf{c}_1 \ldots \mathbf{c}_m) + \mathbb{R}^n_+}_{P} = \underbrace{conv(\widehat{\mathbf{c}}_1 \ldots \widehat{\mathbf{c}}_{\widehat{m}}) + \mathbb{R}^n_+}_{\widehat{P}} \quad (5)$$

Step 2. This identity (5) says in particular that the two sets P and $\widehat{P}$ on its left and right hand side have the same set of extreme points, namely $ext(P) = ext(\widehat{P})$. The set $ext(P)$ of extreme points of the set P is nonempty. In fact, a set which is closed, convex, nonempty, and does not contain a line admits at least an extreme point (Bertsekas, 2009, Proposition 2.1.2). Indeed, P is closed, because $conv(\mathbf{c}_1, ..., \mathbf{c}_m)$ is compact, $\mathbb{R}^n_+$ is closed, and the sum of a compact set with a closed set is closed (Bertsekas, 2009, Section 1.3). Furthermore, P is convex, because $conv(\mathbf{c}_1, ..., \mathbf{c}_m)$ and $\mathbb{R}^n_+$ are both convex and the sum of two convex sets is convex. Finally, P is obviously nonempty and it does not contain a line.

The set $ext(P)$ of extreme points of the set P is a subset of the set of difference vectors $\{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$. In fact, the set of extreme points of the finitely generated polyhedron $conv(\mathbf{c}_1, \ldots, \mathbf{c}_m)$ is a subset of $\{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$ (by the Krein-Milman theorem). The set of extreme points of the pointed cone $\mathbb{R}^n_+$ only consists of the zero vector $\mathbf{0}$. And the set $ext(A + B)$ of extreme points of the vector sum $A + B$ of any two polyhedra A and B is a subset of the vector sum $ext(A) + ext(B)$ of the two sets $ext(A)$ and $ext(B)$ of extreme points of A and B, namely $ext(A + B) \subseteq ext(A) + ext(B)$ (Bertsimas and Tsitsiklis, 1997, exercise 2.22). Analogously, the set $ext(\widehat{P})$ of extreme points of the set $\widehat{P}$ is a nonempty subset of the set $\{\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_m\}$.

In conclusion, the two sets of difference vectors $\{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$ and $\{\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_m\}$ share the vectors in the nonempty set $\Omega = ext(P) = ext(\widehat{P})$. Without loss of generality, we assume that these shared vectors are those corresponding to the first $h \geq 1$ losers, so that $\{\mathbf{c}_1, \ldots, \mathbf{c}_m\} = \Omega \cup \{\mathbf{c}_{h+1}, \ldots, \mathbf{c}_m\}$ and $\{\widehat{\mathbf{c}}_1, \ldots, \widehat{\mathbf{c}}_m\} = \Omega \cup \{\widehat{\mathbf{c}}_{h+1}, \ldots, \widehat{\mathbf{c}}_{\widehat{m}}\}$.

Step 3. The terms on the left and the right hand side of the ME probability identity (4) which correspond to the shared difference vectors in Ω cancel out. The ME probability identity thus reduces to $\sum_{i=h+1}^{m} e^{\mathbf{w}^\intercal \mathbf{c}_i} = \sum_{j=h+1}^{\widehat{m}} e^{\mathbf{w}^\intercal \widehat{\mathbf{c}}_j}$, where the sums start at $h + 1$ rather than at 1. The claim follows by iterating the reasoning above, starting from the latter simplified ME probability identity.

References

Arto Anttila. 2012. Modeling phonological variation. In Abigail C. Cohn, Cécile Fougeron, and Marie Huffman, editors, *The Oxford Handbook of Laboratory Phonology*, pages 76–91. Oxford University Press, Oxford.

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.

Arto Anttila and Naomi Tachikawa Shapiro. 2017. The interaction of stress and syllabification: Serial or parallel? In *Proceedings of the 34th West Coast Conference on Formal Linguistics*, pages 52–61, Somerville, MA, USA. Cascadilla Proceedings Project.

Dimitri P. Bertsekas. 2009. *Convex Optimization Theory*. Athena Scientific, Belmont, MA, USA.

Dimitris Bertsimas and John N. Tsitsiklis. 1997. *Linear Optimization*. Athena Scientific.

Paul Boersma. 1997. How we learn variation, optionality and probability. In *Proceedings of the Institute of Phonetic Sciences (IFA) 21*, pages 43–58, University of Amsterdam. Institute of Phonetic Sciences.

Paul Boersma. 1998. *Functional Phonology*. Ph.D. thesis, University of Amsterdam, The Netherlands. The Hague: Holland Academic Graphics.

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.

Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Lauri Carlson. 1978. Word stress in Finnish. Massachusetts Institute of Technology, Cambridge, Massachusetts.

Nine Elenbaas. 1999. *A unified account of binary and ternary stress. Considerations from Sentani and Finnish*. Ph.D. thesis, LOT: Netherlands Graduate School of Linguistics.

Nine Elenbaas and René Kager. 1999. Ternary rhythm and the lapse constraint. *Phonology*, 16:273–329.

Sharon Goldwater and Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the Stockholm Workshop on Variation Within Optimality Theory*, pages 111–120, Stockholm University.

Kristin Hanson and Paul Kiparsky. 1996. A parametric theory of poetic meter. *Language*, 72:287–335.

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 Zsuzsa Cziráky Londe. 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology*, 23.1:59–104.

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

Bruce Hayes, Kie Zuraw, Péter Siptár, and Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language*, 85:822–863.

Dan Karvonen. 2005. *Word Prosody in Finnish*. Ph.D. thesis, University of California, Santa Cruz.

Samuel Jay Keyser and Paul Kiparsky. 1984. Syllable structure in Finnish phonology. In Mark Aronoff and Richard Oehrle, editors, *Language Sound Structure*, pages 7–31. MIT Press, Cambridge, Massachusetts.

Géraldine Legendre, Yoshiro Miyata, and Paul Smolensky. 1990a. 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*, pages 884–891, Hillsdale, NJ. Lawrence Erlbaum Associates.

Géraldine Legendre, Yoshiro Miyata, and Paul Smolensky. 1990b. Harmonic Grammar: A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the 12th annual conference of the Cognitive Science Society*, pages 388–395, Hillsdale, NJ. Lawrence Erlbaum.

Giorgio Magri. 2019. Finiteness of optima in constraint-based phonology. Manuscript, CNRS.

Giorgio Magri and Arto Anttila. 2019. ℂ𝕠𝔾𝕖𝕋𝕠: Convex geometry tools for typological analysis in categorical and probabilistic constraint-based phonology (version 1.0). Available at https://cogeto.stanford.edu.

Catherine Ringen and Orvokki Heinämäki. 1999. Variation in Finnish vowel harmony. *Natural Language and Linguistc Theory*, 17:303–337.

Brian W. Smith and Joe Pater. 2017. French schwa and gradient cumulativity. Manuscript. University of California, Berkeley and University of Massachusetts, Amherst.

Paul Smolensky and Géraldine Legendre. 2006. *The Harmonic Mind*. MIT Press, Cambridge, MA.

Kie Zuraw and Bruce Hayes. 2017. Intersecting constraint families: an argument for Harmonic Grammar. *Language*, 93.3:497–546.

Jesse Zymet. 2015. Distance-based decay in long-distance phonological processes. In *Proceedings of the 32nd West Coast Conference on Formal Linguistics*, pages 72–81, Somerville, MA. Cascadilla Proceedings Project.