Implicational Universals in Stochastic Constraint-Based Phonology

Giorgio Magri
CNRS / University of Paris 8
magrigrg@gmail.com

Abstract

This paper focuses on the most basic implicational universals in phonological theory, called *T-orders* after Anttila and Andrus (2006). It shows that the T-orders predicted by stochastic and categorical Optimality Theory coincide. Analogously, the T-orders predicted by stochastic and categorical Harmonic Grammar coincide. In other words, these stochastic constraint-based frameworks do not tamper with the typological structure induced by the corresponding categorical frameworks.

1 Introduction

Phonology has traditionally focused on *alternations* revealed by paradigms such as the German final devoicing examples [baːt]/[bɛːdə] ('bath-SG/PL') and [tsuːk]/[tsyːgə] ('train-SG/PL'). These alternations are usually modeled through *phonological grammars* which map from *underlying representations* (URs) to *surface representations* (SRs) (Chomsky and Halle, 1968). Constraint-based implementations of this combinatorial phonological theory include *Optimality Theory* (OT; Prince and Smolensky, 1997, 2004) and *Harmonic Grammar* (HG; Legendre et al., 1990; Smolensky and Legendre, 2006), reviewed below in sections 4 and 5.

More recently, phonology has extended its empirical coverage from categorical alternations to patterns of phonologically conditioned variation and gradient phonological (or phonotactic) judgements (see for instance Anttila, 2012 and Coetzee and Pater, 2011). This extension of the empirical coverage has required a corresponding extension of the theoretical framework. A phonological grammar cannot be construed anymore as a categorical function from URs to SRs. Instead, it must be construed as a function from URs to probability distributions over the entire set of SRs. Constraint-based implementations of this stochastic theory include *partial order OT* (Anttila, 1997b), *stochastic OT* (SOT; Boersma, 1997, 1998), and *stochastic HG* (SHG; Boersma and Pater, 2016)[1], recalled below in sections 4 and 5. Another framework explored in the recent literature on probabilistic constraint-based phonology is *MaxEnt* (ME; Goldwater and Johnson, 2003; Hayes and Wilson, 2008). Its T-orders are discussed in a companion paper (Anttila and Magri, 2018).

How can we investigate and understand the typological structure encoded by a probabilistic phonological framework? In the case of a categorical framework such as OT or HG, the predicted typological structure can be investigated directly by exhaustively listing all the grammars predicted for certain constraint and candidate sets. That is possible because the predicted typology of grammars is usually *finite*. The situation is rather different for probabilistic frameworks: the predicted typology always consists of an *infinite* number of probability distributions which therefore cannot be exhaustively listed and directly inspected. A more indirect strategy is needed to chart the predicted typological structure.

A natural indirect strategy that gets around the problem raised by an infinite typology is to enumerate, not the individual languages in the typology, but the set of implicational universals predicted by the typology. An *implicational universal* is an implication $P \xrightarrow{\mathfrak{T}} \widehat{P}$ which holds of a given typology $\mathfrak{T}$ whenever *every* language in the typology that satisfies the antecedent property P also satisfies the consequent property $\widehat{P}$ (Greenberg, 1963). Since implicational universals take

[1] Boersma and Pater (2016) actually use the term "noisy HG" instead of "stochastic HG". We prefer "stochastic HG" to stress the analogy with Boersma's 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). Hayes (2017) discusses further stochastic variants of categorical HG.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3265–3274
Brussels, Belgium, October 31 - November 4, 2018.

into account every language in the typology, they chart the boundaries and measure the richness of the typological structure predicted by $\mathfrak{T}$.

Which antecedent and consequent properties P and $\widehat{P}$ should we focus on? To start from the simplest case, let us consider a typology $\mathfrak{T}$ of *categorical* phonological grammars, construed traditionally as mappings from URs to SRs. Within this categorical framework, the simplest, most basic, most atomic antecedent property P is the property of mapping a certain specific UR x to a certain specific SR y. Analogously, the simplest consequent property $\widehat{P}$ is the property of mapping a certain specific UR $\widehat{\mathsf{x}}$ to a certain specific SR $\widehat{\mathsf{y}}$. We thus focus on the following class of implications:

Definition 1 *The implicational universal* $(\mathsf{x},\mathsf{y}) \xrightarrow{\mathfrak{T}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ *holds relative to a categorical typology* $\mathfrak{T}$ *provided each grammar in* $\mathfrak{T}$ *which succeeds at the antecedent mapping (i.e., it maps the UR* x *to the SR* y*), also succeeds at the consequent mapping (i.e., it maps the UR* $\widehat{\mathsf{x}}$ *to the SR* $\widehat{\mathsf{y}}$*).* □

The relation $\xrightarrow{\mathfrak{T}}$ thus defined over mappings is a partial order (under mild additional assumptions). It is called the *T-order* induced by the typology $\mathfrak{T}$ (Anttila and Andrus, 2006). For example, any dialect of English that deletes t/d at the end of a coda cluster before a vowel also deletes it before a consonant (Guy, 1991; Kiparsky, 1993; Coetzee, 2004). The implication $(/\mathsf{cost.us}/, [\mathsf{cos.us}]) \rightarrow (/\mathsf{cost.me}/, [\mathsf{cos.me}])$ thus holds relative to the typology $\mathfrak{T}$ of English dialects.

Implicational universals can also be statistical. For instance, in dialects of English where t/d deletion applies variably, deletion has been found to be more frequent before consonants than before vowels. To model these frequency effects, we need to consider a typology $\mathfrak{T}$ of *probabilistic* phonological grammars, construed as functions from URs to probability distributions over SRs. We propose to extend the notion of T-orders from the categorical to the probabilistic setting as follows:

Definition 2 *The implicational universal* $(\mathsf{x},\mathsf{y}) \xrightarrow{\mathfrak{T}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ *holds relative to a probabilistic typology* $\mathfrak{T}$ *provided each grammar in* $\mathfrak{T}$ *assigns a probability to the consequent mapping* $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ *which is at least as large as the probability it assigns to the antecedent mapping* (x,y)*.* □

To illustrate, the implication $(/\mathsf{cost.us}/, [\mathsf{cos.us}]) \rightarrow (/\mathsf{cost.me}/, [\mathsf{cos.me}])$ also holds relative to the typology $\mathfrak{T}$ of English dialects with variable deletion because the probability of the consequent $(/\mathsf{cost.me}/, [\mathsf{cos.me}])$ (i.e., the frequency of deletion before a consonant) in any dialect is at least as large as the probability of the antecedent $(/\mathsf{cost.us}/, [\mathsf{cos.us}])$ (i.e., the frequency of deletion before a vowel).

The original categorical definition 1 of T-orders is a special case of the probabilistic definition 2. In fact, suppose that a categorical grammar succeeds on the antecedent mapping (x,y). That grammar construed probabilistically thus assigns probability 1 to the antecedent mapping. Definition 2 then requires that grammar to also assign probability 1 to the consequent mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$. In other words, the grammar construed categorically succeeds on the consequent mapping, as required by the original definition 1 of categorical T-orders.

T-orders are defined at the level of mappings from URs to SRs. They thus allow for cross-framework comparisons, even bridging across categorical and probabilistic frameworks. This paper (together with the companion Anttila and Magri 2018) thus uses T-orders to compare the probabilistic implementations of constraint-based phonology with the original categorical implementations.

The main result reported in this paper is that the T-orders predicted by stochastic OT (and by partial order OT) coincide with those predicted by categorical OT, no matter what the candidate and constraint sets look like, as shown in section 4. Analogously, the T-orders predicted by stochastic HG coincide with those predicted by categorical HG, as shown in section 5. In other words, these stochastic frameworks do not tamper with the typological structure induced by the original categorical frameworks, at least when that structure is measured in terms of T-orders. These specific results about OT and HG are derived as a special case of a more general result on stochastic typologies, developed in sections 2 and 3.

As discussed in a companion paper (Anttila and Magri, 2018), the situation is very different for ME. Both ME and stochastic HG can be construed as probabilistic variants of categorical HG. Stochastic and categorical HG share the same T-orders. The ME T-orders instead obey a rather different underlying convex geometry and turn out to be much sparser. In other words, ME yields a much richer probabilistic extension of categorical HG than stochastic HG does. Section 6 concludes the paper by discussing these results in the context

of the recent literature on probabilistic constraint-based phonology.

2 Categorical and stochastic phonology

We assume a relation Gen which pairs each UR x with a set $Gen(\mathsf{x})$ of *candidate* SRs. As recalled above, a *categorical* phonological grammar G takes a UR x and selects a corresponding SR $\mathsf{y} = G(\mathsf{x})$ from the candidate set $Gen(\mathsf{x})$. A stochastic phonological grammar G instead takes a UR x and returns a probability distribution $G(\cdot\,|\,\mathsf{x})$ over $Gen(\mathsf{x})$ which assigns a probability $G(\mathsf{y}|\,\mathsf{x})$ to each candidate SR y in $Gen(\mathsf{x})$. This section illustrates a general method to leverage a given typology $\mathfrak{T}$ of categorical grammars into a typology of stochastic grammars. Sections 4 and 5 will then show that various stochastic frameworks in the recent constraint-based literature (such as partial order OT, stochastic OT, and stochastic HG) all fit within this general scheme.

Following common practice in constraint-based phonology, we assume that the categorical typology $\mathfrak{T}$ only contains a finite number of grammars.[2] We consider a probability mass function p over $\mathfrak{T}$. Thus, p assigns to each categorical grammar G in $\mathfrak{T}$ a nonnegative probability mass $p(G) \geq 0$ and these masses sum up to 1, namely $\sum_{G\in\mathfrak{T}} p(G) = 1$. We can then define the stochastic grammar G_p corresponding to the probability mass function p as the function which takes a UR x and returns the probability distribution $G_p(\cdot\,|\,\mathsf{x})$ over the candidate set $Gen(\mathsf{x})$ defined as in (1). It says that the probability $G_p(\mathsf{y}\,|\,\mathsf{x})$ that the UR x is mapped to the SR y is he probability mass allocated by p to the region $\{G \in \mathfrak{T}\,|\,G(\mathsf{x}) = \mathsf{y}\}$ of the typology $\mathfrak{T}$ consisting of those categorical grammars which succeed on the mapping (x, y).

$$G_p(\mathsf{y}\,|\,\mathsf{x}) = \sum_{\{G\in\mathfrak{T}\,|\,G(\mathsf{x})=\mathsf{y}\}} p(G) \qquad (1)$$

We assume next that each categorical grammar in the typology $\mathfrak{T}$ returns a *unique* SR y for each UR x.[3] This assumption suffices to ensure that G_p is indeed a probability distribution, namely that the sum of the probabilities $G_p(\mathsf{y}\,|\,\mathsf{x})$ over the candidates y in $Gen(\mathsf{x})$ is equal to 1, as shown in (2).

$$\sum_{\mathsf{y}\in Gen(\mathsf{x})} G_p(\mathsf{y}\,|\,\mathsf{x}) \overset{(a)}{=} \sum_{\mathsf{y}\in Gen(\mathsf{x})} \sum_{\{G\in\mathfrak{T}\,|\,G(\mathsf{x})=\mathsf{y}\}} p(G) \overset{(b)}{=} \sum_{G\in\mathfrak{T}} p(G) \overset{(c)}{=} 1 \qquad (2)$$

In step (2a), we have used the definition (1) of $G_p(\mathsf{y}\,|\,\mathsf{x})$. In step (2b), we have used the fact that every grammar in $\mathfrak{T}$ maps x to a unique SR y, so that the sets $\{G \in \mathfrak{T}\,|\,G(\mathsf{x}) = \mathsf{y}\}$ partition the typology $\mathfrak{T}$ into disjoint sets as y spans the candidate set $Gen(\mathsf{x})$. In step (2c), we have used the fact that p is a probability mass function over $\mathfrak{T}$ and thus adds up to 1.

A family $\mathcal{P}$ of probability mass functions $p_1, p_2, \ldots$ over the finite categorical typology $\mathfrak{T}$ thus induces a typology $\{G_{p_1}, G_{p_2}, \ldots\}$ of stochastic grammars. It is called the *stochastic* typology corresponding to the categorical typology $\mathfrak{T}$ and the probability family $\mathcal{P}$, and it is denoted by $\mathfrak{T}_\mathcal{P}$. We denote by $\xrightarrow{\mathfrak{T}}$ the T-order relative to the categorical typology $\mathfrak{T}$ in the sense of definition 1 and by $\xrightarrow{\mathfrak{T}_\mathcal{P}}$ the T-order relative to the stochastic typology $\mathfrak{T}_\mathcal{P}$ in the sense of definition 2. We want to investigate the relationship between these categorical and stochastic T-orders.

3 Relationship between categorical and stochastic T-orders

Let us suppose that the implication $(\mathsf{x}, \mathsf{y}) \xrightarrow{\mathfrak{T}} (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ holds between an antecedent mapping (x, y) and a consequent mapping $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ relative to a categorical typology $\mathfrak{T}$. By definition 1, this means that every categorical grammar G in the typology $\mathfrak{T}$ that maps the antecedent UR x to the antecedent SR y (namely, $G(\mathsf{x}) = \mathsf{y}$) also maps the consequent UR $\widehat{\mathsf{x}}$ to the consequent SR $\widehat{\mathsf{y}}$ (namely, $G(\widehat{\mathsf{x}}) = \widehat{\mathsf{y}}$), yielding the inclusion (3).

$$\underbrace{\{G \in \mathfrak{T}\,|\,G(\mathsf{x})=\mathsf{y}\}}_{\substack{\text{grammars consistent}\\ \text{with the antecedent}\\ \text{mapping}}} \subseteq \underbrace{\{G \in \mathfrak{T}\,|\,G(\widehat{\mathsf{x}})=\widehat{y}\}}_{\substack{\text{grammars consistent}\\ \text{with the consequent}\\ \text{mapping}}} \qquad (3)$$

By (1), this inclusion (3) entails that the probability assigned by G_p to the consequent mapping $(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ is at least as large as the probability assigned to the antecedent mapping (x, y), as stated in (4). This entailment follows from the sheer fact that

[2] This assumption always holds in OT, as the number of constraint rankings is finite. It might fail in HG, but only in rather pathological situations which do not seem germane to natural language phonology.

[3] Suppose instead that a categorical grammar were to return two different SRs y_1 and y_2 for some UR x. How should we interpret such a scenario? Plausibly, we should interpret the two SRs y_1 and y_2 as free variants with equal probability of 0.5 (while all other candidates have probability 0). But this means that our grammar is stochastic, not categorical.

probabilities are monotonic relative to set inclusion. The entailment from the inclusion (3) to the inequality (4) thus holds under no assumptions whatsoever on the probability mass function p used to define the stochastic grammar G_p.

$$\underset{\substack{\text{probability of the}\\ \text{antecedent mapping}}}{G_p(\mathsf{y}\,|\,\mathsf{x})} \;\leq\; \underset{\substack{\text{probability of the}\\ \text{consequent mapping}}}{G_p(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})} \tag{4}$$

The latter inequality (4) finally says that the implication $(\mathsf{x},\mathsf{y}) \xrightarrow{\mathfrak{T}_\mathcal{P}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ holds also relative to the stochastic typology $\mathfrak{T}_\mathcal{P}$ in the sense of definition 2. In conclusion, a categorical T-order always entails the corresponding stochastic T-order, no matter the shape of the family $\mathcal{P}$ of probability mass functions used to derive the stochastic typology $\mathfrak{T}_\mathcal{P}$ from the categorical typology $\mathfrak{T}$.

We now turn to the reverse entailment. Suppose that an implication $(\mathsf{x},\mathsf{y}) \xrightarrow{\mathfrak{T}_\mathcal{P}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ holds between an antecedent mapping (x,y) and a consequent mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ relative to the stochastic typology $\mathfrak{T}_\mathcal{P}$. By definition 2, this means in turn that the inequality (4) holds between the probabilities $G_p(\mathsf{y}\,|\,\mathsf{x})$ and $G_p(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})$ of the antecedent and the consequent mappings relative to any probability mass function p in the family $\mathcal{P}$. Suppose by contradiction that the corresponding implication $(\mathsf{x},\mathsf{y}) \xrightarrow{\mathfrak{T}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ relative to the original categorical typology $\mathfrak{T}$ instead fails. By definition 1, this means that the set inclusion (3) fails because there exists some grammar G_0 with the properties in (5): G_0 succeeds on the antecedent mapping, namely it maps x to y; but G_0 fails on the consequent mapping, namely it maps $\widehat{\mathsf{x}}$ to some loser candidate $\widehat{\mathsf{z}}$ different from the intended winner candidate $\widehat{\mathsf{y}}$.

$$G_0(\mathsf{x}) = \mathsf{y}, \qquad G_0(\widehat{\mathsf{x}}) = \widehat{\mathsf{z}} \neq \widehat{\mathsf{y}} \tag{5}$$

We would like to derive a contradiction from the assumption (4) that the stochastic implication $(\mathsf{x},\mathsf{y}) \xrightarrow{\mathfrak{T}_\mathcal{P}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ holds and the assumption (5) that the categorical implication $(\mathsf{x},\mathsf{y}) \xrightarrow{\mathfrak{T}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ fails.

Yet, no contradiction arises in the general case. Indeed, suppose that the probability mass functions in the family $\mathcal{P}$ all happen to assign zero (or tiny) probability mass to this grammar G_0 which flouts the categorical implication because of (5). This problematic grammar G_0 thus bears no (or only a tiny) effect on the total probabilities $G_p(\mathsf{y}\,|\,\mathsf{x})$ and $G_p(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}})$ of the two mappings (x,y) and $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$. The probability inequality (4) is therefore not necessarily compromised by the offensive behavior (5) of G_0, as long as the other grammars in the typology comply.

In order to derive a contradiction from these two conditions (4) and (5), we need to make some assumptions on the family $\mathcal{P}$ of probability mass functions. Indeed, the problem just discussed arises when *every* probability mass function p in $\mathcal{P}$ assigns zero (or tiny) probability to the problematic grammar G_0. We need to rule out this scenario. We propose to achieve that through the assumption that the family $\mathcal{P}$ satisfies the following

Definition 3 *The family $\mathcal{P}$ of probability mass functions over the finite categorical typology $\mathfrak{T}$ is* sufficiently rich *in the sense that for every categorical grammar G in $\mathfrak{T}$ and for any two URs x and $\widehat{\mathsf{x}}$, the following inequalities*

$$G_p(G(\mathsf{x})|\,\mathsf{x}) > 1/2, \quad G_p(G(\widehat{\mathsf{x}})|\,\widehat{\mathsf{x}}) > 1/2 \tag{6}$$

hold for some probability mass function p in $\mathcal{P}$. □

Here is the intuition behind this definition. Suppose that for every categorical grammar G, the family $\mathcal{P}$ contains a probability mass function p which assigns all the probability mass to that grammar G. By (1), the corresponding stochastic grammar G_p assigns probability 1 to the mappings enforced by G, as stated in (7).

$$G_p(G(\mathsf{x})|\,\mathsf{x}) = 1 \text{ for every UR } \mathsf{x} \tag{7}$$

In other words, the stochastic grammar G_p "coincides" with the categorical grammar G and the stochastic typology $\mathfrak{T}_\mathcal{P}$ thus "contains" or "extends" the original categorical typology $\mathfrak{T}$. In this special case, we obviously expect the stochastic implication $(\mathsf{x},\mathsf{y}) \xrightarrow{\mathfrak{T}_\mathcal{P}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ to entail the categorical implication $(\mathsf{x},\mathsf{y}) \xrightarrow{\mathfrak{T}} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$, as desired.

Condition (6) required by definition 3 is a weaker version of the latter condition (7). First, it is weaker because the requirement $G_p(G(\mathsf{x})|\,\mathsf{x}) = 1$ is replaced with the weaker requirement $G_p(G(\mathsf{x})|\,\mathsf{x}) > 1/2$: the probability assigned to the mappings enforced by G needs not be 1, as long as it is large enough, namely larger than 1/2. Second, this requirement $G_p(G(\mathsf{x})|\mathsf{x}) > 1/2$ needs not be satisfied by a unique mass p for all URs: it suffices to look at just two URs at the time.

If the family $\mathcal{P}$ is sufficiently rich in the sense of definition 3, the two conditions (4) and (5) are indeed contradictory. In fact, condition (5) now ensures that $\mathcal{P}$ contains a probability mass function

p_0 such that the corresponding stochastic grammar G_{p_0} maps x to y with probability larger than 1/2 and it maps $\widehat{\mathsf{x}}$ to $\widehat{\mathsf{z}}$ with probability larger than 1/2. The latter fact means in turn that G_{p_0} maps $\widehat{\mathsf{x}}$ to $\widehat{\mathsf{y}}$ with probability smaller than 1/2, because the probabilities of the various candidates $\widehat{\mathsf{y}}, \widehat{\mathsf{z}}, \ldots$ in $Gen(\widehat{\mathsf{x}})$ must add up to 1. In conclusion, we have obtained $G_{p_0}(\mathsf{y} \,|\, \mathsf{x}) > 1/2$ and $G_{p_0}(\widehat{\mathsf{y}} \,|\, \widehat{\mathsf{x}}) < 1/2$, in blatant contradiction of (4).

The preceding reasoning is summarized in the following proposition 1, which says that the T-order relative to a categorical typology $\mathfrak{T}$ and the T-order relative to the corresponding stochastic typology $\mathfrak{T}_{\mathcal{P}}$ coincide, no matter what the family $\mathcal{P}$ of probability mass functions looks like, as long as it is sufficiently rich, in the sense of definition 3. Identity of T-orders holds even when the family $\mathcal{P}$ is infinite, so that the stochastic typology $\mathfrak{T}_{\mathcal{P}}$ contains an infinite number of stochastic grammars, while the categorical typology $\mathfrak{T}$ contains only a finite number of grammars.

Proposition 1 *Consider a finite typology $\mathfrak{T}$ of categorical grammars and a family $\mathcal{P}$ of probability mass functions on $\mathfrak{T}$. Let $\mathfrak{T}_{\mathcal{P}}$ be the typology of the corresponding stochastic grammars, defined through (1). If $\mathcal{P}$ is sufficiently rich in the sense of definition 3, the T-order $\xrightarrow{\mathfrak{T}}$ relative to the categorical typology $\mathfrak{T}$ and the T-order $\xrightarrow{\mathfrak{T}_{\mathcal{P}}}$ relative to the stochastic typology $\mathfrak{T}_{\mathcal{P}}$ coincide.* □

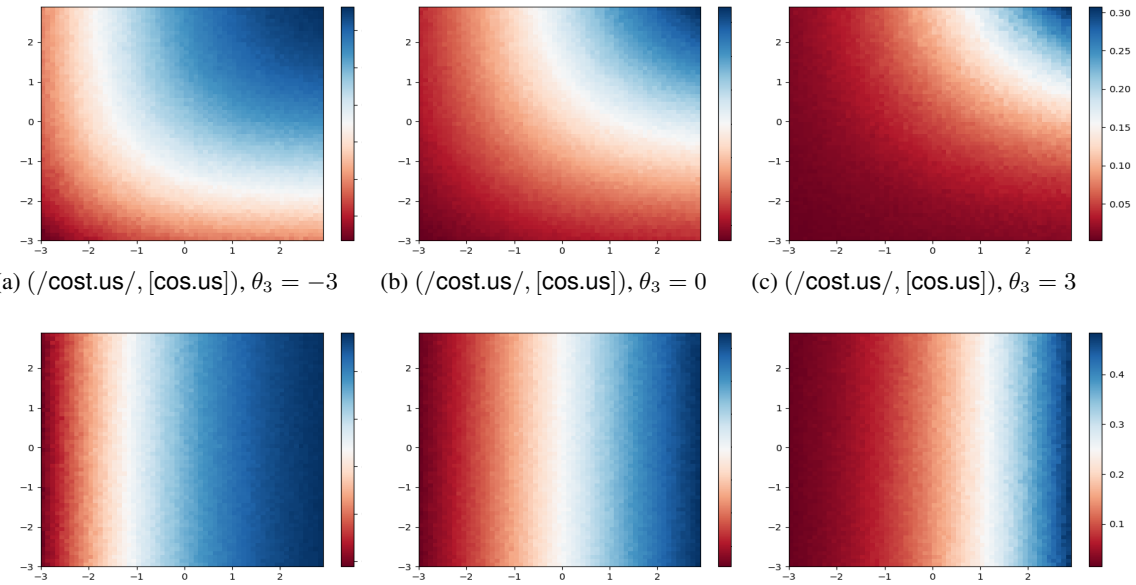
In the rest of the paper, we apply this result to various categorical and stochastic frameworks for constraint-based phonology.

4 Categorial OT, partial order OT, and stochastic OT induce the same T-orders

In this section, we focus on categorical and stochastic OT. We assume a set of n constraints $C_1, \ldots, C_k, \ldots, C_n$ and some candidacy relation Gen. We recall that a constraint C_k *prefers* a mapping (x, y) to another mapping (x, z) provided C_k assigns less violations to the former than to the latter, namely $C_k(\mathsf{x}, \mathsf{y}) < C_k(\mathsf{x}, \mathsf{z})$. A *constraint ranking* is an arbitrary linear order $\gg$ over the constraint set. A constraint ranking $\gg$ *prefers* a mapping (x, y) to another mapping (x, z) provided the highest $\gg$-ranked constraint which distinguishes between the two mappings (x, y) and (x, z) prefers (x, y). The *categorical OT grammar* corresponding to a ranking $\gg$ maps a UR x to that SR y such that $\gg$ prefers the mapping (x, y) to the mapping (x, z) corresponding to any other candidate z in $Gen(\mathsf{x})$ (Prince and Smolensky, 2004). We denote by $\xrightarrow{\text{OT}}$ the T-order corresponding to the typology $\mathfrak{T}$ of the categorical OT grammars corresponding to all constraint rankings, in the sense of definition 2.

To illustrate, consider the following three constraints (from Kiparsky, 1993) for the process of t/d deletion mentioned in section 1: C_1 = SYLLABLEWELLFORMEDNESS (SWF) penalizes codas and tautosyllabic consonant clusters; C_2 = ALIGN penalizes resyllabification across word boundaries; and C_3 = MAX penalizes segment deletion. Suppose that the UR /cost us/ comes with the three candidate SRs [cost.us] (faithful), [cos.us] (with deletion), and [cos.tus] (with resyllabification). Analogously, suppose that the UR /cost me/ comes with the three candidate SRs [cost.me], [cos.me], and [cos.tme]. It is easy to verify that the implication $(/\text{cost.us}/, [\text{cos.us}]) \xrightarrow{\text{OT}} (/\text{cost.me}/, [\text{cos.me}])$ holds relative to the OT typology generated by constraints C_1, C_2, C_3 in the sense of definition 1: every ranking of the three constraints which succeeds on the antecedent mapping $(/\text{cost.us}/, [\text{cos.us}])$ also succeeds on the consequent mapping $(/\text{cost.me}/, [\text{cos.me}])$. In other words, t/d deletion before a vowel entails deletion before a consonant.

We now turn to the stochastic counterpart of this categorical framework. A *ranking vector* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k, \ldots, \theta_n) \in \mathbb{R}^n$ assigns a numerical *ranking value* θ_k to each constraint C_k. The *stochastic ranking vector* $\boldsymbol{\theta} + \boldsymbol{\epsilon} = (\theta_1 + \epsilon_1, \ldots, \theta_n + \epsilon_n)$ is obtained by adding to the ranking values $\theta_1, \ldots, \theta_n$ some numbers $\epsilon_1, \ldots, \epsilon_n$ sampled independently from each other according to some distribution $\mathcal{D}$ on $\mathbb{R}$. If the distribution $\mathcal{D}$ is continuous, the probability that two stochastic ranking values $\theta_h + \epsilon_h$ and $\theta_k + \epsilon_k$ coincide is equal to zero. The stochastic ranking vector $\boldsymbol{\theta} + \boldsymbol{\epsilon}$ thus describes the unique ranking $\gg_{\boldsymbol{\theta}+\boldsymbol{\epsilon}}$ which respects the relative size of the stochastic ranking values: a constraint C_h is ranked above a constraint C_k according to $\gg_{\boldsymbol{\theta}+\boldsymbol{\epsilon}}$ (namely, $C_h \gg_{\boldsymbol{\theta}+\boldsymbol{\epsilon}} C_k$) if and only if the stochastic ranking value of the former is larger than that of the latter (namely, $\theta_h + \epsilon_h > \theta_k + \epsilon_k$). A ranking vector $\boldsymbol{\theta}$ thus induces the probability mass function $p^{\mathcal{D}}_{\boldsymbol{\theta}}$ defined in (8) over the categorical OT typology $\mathfrak{T}$. Obviously, this definition yields a probability mass, namely the sum of the masses $p^{\mathcal{D}}_{\boldsymbol{\theta}}(G)$ over all the



(a) (/cost.us/, [cos.us]), $\theta_3 = -3$ (b) (/cost.us/, [cos.us]), $\theta_3 = 0$ (c) (/cost.us/, [cos.us]), $\theta_3 = 3$

(d) (/cost.me/, [cos.me]), $\theta_3 = -3$ (e) (/cost.me/, [cos.me]), $\theta_3 = 0$ (f) (/cost.me/, [cos.me]), $\theta_3 = 3$

Figure 1: SOT probabilities of the antecedent mapping (/cost.us/, [cos.us]) and the consequent mapping (/cost.me/, [cos.me]) as a function of θ_1 (horizontal axis) and θ_2 (vertical axis) for three choices of θ_3

categorical OT grammars G in $\mathfrak{T}$ is indeed 1.

$$p^{\mathcal{D}}_{\boldsymbol{\theta}}(G) = \text{the probability of sampling } \epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} \mathcal{D} \text{ such that the OT grammar corresponding to the ranking } \gg_{\boldsymbol{\theta}+\boldsymbol{\epsilon}} \text{ is indeed } G \tag{8}$$

The typology of stochastic grammars $\mathfrak{T}_{\mathcal{P}}$ obtained as in section 2 from the categorical OT typology $\mathfrak{T}$ and the family $\mathcal{P} = \{p^{\mathcal{D}}_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathbb{R}^n\}$ of probability mass functions $p^{\mathcal{D}}_{\boldsymbol{\theta}}$ corresponding to all ranking vectors $\boldsymbol{\theta}$ is called *stochastic OT* (SOT; Boersma, 1997, 1998). We denote by $\overset{\text{SOT}}{\longrightarrow}$ the T-orders corresponding to SOT in the sense of definition 2.

What is the typological structure encoded by SOT's T-orders? Given that the original OT typology is finite (because there are only a finite number of constraint rankings) while the SOT typology is infinite (it contains an infinite number of grammars which assign different probabilities), how much of OT's typological structure is preserved in SOT? These questions are crucial for phonological theory but technically nontrivial. To illustrate, figure 1 plots the SOT probability of the mappings (/cost.us/, [cos.us]) and (/cost.me/, [cos.me]) relative to the three constraints C_1, C_2, C_3 listed above as a function of the ranking value θ_1 of constraint C_1 (horizontal axis) and the ranking value θ_2 of constraint C_2 (vertical axis) for three choices of the ranking value θ_3 of constraint C_3.[4] These plots suggest that the implication (/cost.us/, [cos.us]) $\overset{\text{SOT}}{\rightarrow}$ (/cost.me/, [cos.me]) holds in SOT: the probability of the consequent (/cost.me/, [cos.me]) (plotted in the bottom row) seems to be always larger than the probability of the antecedent (/cost.us/, [cos.us]) (plotted in the top row). But how can this conjecture be checked, given that SOT probabilities seem not to admit a closed-form expression?

The result obtained in section 3 provides a straightforward solution to this problem. Suppose that there exists a positive constant Δ large enough that the distribution $\mathcal{D}$ concentrates most of the probability mass on the interval $[-\Delta, +\Delta]$, as stated in (9). This assumption holds in particular when $\mathcal{D}$ has a bounded support or it is defined through a density (such as a gaussian, as assumed in Boersma, 1997, 1998).

$$(\mathcal{D}([-\Delta, +\Delta]))^n > 1/2 \tag{9}$$

For any constraint ranking $\gg$, consider a ranking vector $\boldsymbol{\theta}$ such that the top $\gg$-ranked constraint has the largest ranking value; the second top $\gg$-ranked constraint has the second largest ranking value; and so on. Furthermore, assume that these ranking values are spaced apart by more than 2Δ. Since the numbers $\epsilon_1, \ldots, \epsilon_n$ are all bounded between $-\Delta$ and $+\Delta$ with probability at least 1/2

[4] These plots (as well as those in figure 2) are obtained by sampling for 10,000 times from each stochastic grammar. The distribution $\mathcal{D}$ is a gaussian with mean 0 and variance 2.

and since the ranking values are spaced apart by more than 2Δ, the constraint ranking $\gg_{\boldsymbol{\theta}+\boldsymbol{\epsilon}}$ corresponding to the stochastic ranking vector $\boldsymbol{\theta}+\boldsymbol{\epsilon}$ coincides with the original ranking $\gg$ with probability at least 1/2. In other words, the probability mass function $p^{\mathcal{D}}_{\boldsymbol{\theta}}$ corresponding to this ranking vector $\boldsymbol{\theta}$ according to (8) assigns more than half of the probability mass to the OT grammar corresponding to the ranking $\gg$. The family $\mathcal{P} = \{p^{\mathcal{D}}_{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \in \mathbb{R}^n\}$ is therefore sufficiently rich in the sense of definition 3. Proposition 1 thus yields the following

Corollary 1 *Under the mild assumption (9) on the distribution $\mathcal{D}$, the T-order $\xrightarrow{SOT}$ relative to SOT is identical to the T-order $\xrightarrow{OT}$ relative to categorical OT for any constraint and candidate set.* □

In conclusion, despite the SOT typology being infinite, SOT induces the same typological structure as categorical OT, at least when typological structure is measured in terms of T-orders. Furthermore, the technical problem of computing T-orders relative to SOT is reduced to the much easier problem of computing T-orders relative to categorical OT, which indeed admits an efficient solution (Magri, 2018a). This result extends to *partial order OT* (Anttila, 1997a), as the latter is a special case of SOT.

5 Categorial HG and stochastic HG induce the same T-orders

This section shows that completely analogous considerations hold for HG. A weight vector $\mathbf{w} = (w_1, \dots, w_k, \dots, w_n) \in \mathbb{R}^n_+$ assigns a nonnegative weight $w_k \geq 0$ to each constraint C_k. The $\mathbf{w}$-*harmony* of a mapping (x, y) is the weighted sum of the constraint violations multiplied by -1, namely $-\sum_{k=1}^n w_k C_k(\mathsf{x}, \mathsf{y})$. Because of the minus sign, mappings with a large harmony have few constraint violations. The categorical HG grammar corresponding to a weight vector $\mathbf{w}$ maps a UR x to the surface form y such that the mapping (x, y) has a larger $\mathbf{w}$-harmony than the mapping (x, z) corresponding to any other candidate z in $Gen(\mathsf{x})$ (Legendre et al., 1990; Smolensky and Legendre, 2006). We denote by $\xrightarrow{HG}$ the T-order corresponding to the typology $\mathfrak{T}$ of the categorical HG grammars corresponding to all nonnegative weight vectors, in the sense of definition 2.

To illustrate, it is easy to verify that the implication $(/\mathsf{cost.us}/, [\mathsf{cos.us}]) \xrightarrow{HG} (/\mathsf{cost.me}/, [\mathsf{cos.me}])$ considered above holds also relative to the HG typology in the sense of definition 1: every weighting of the three constraints which succeeds on the antecedent mapping $(/\mathsf{cost.us}/, [\mathsf{cos.us}])$ also succeeds on the consequent mapping $(/\mathsf{cost.me}/, [\mathsf{cos.me}])$. In general, the HG typology is a proper superset of the OT typology (when the set of URs is finite). The HG T-order is therefore a subset of the corresponding OT T-order.

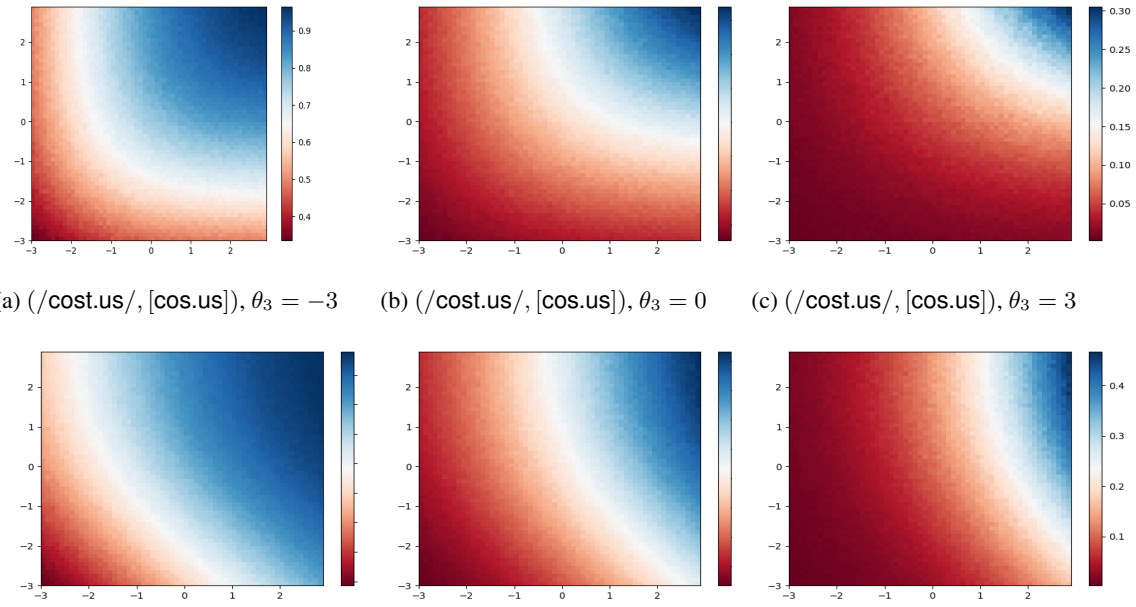
We now turn to the stochastic counterpart of this categorical framework. The *stochastic weight vector* $\mathbf{w}+\boldsymbol{\epsilon} = (w_1+\epsilon_1, \dots, w_n+\epsilon_n)$ is obtained by adding to the weights $w_1, \dots, w_n$ some numbers $\epsilon_1, \dots, \epsilon_n$ sampled independently from each other according to some distribution $\mathcal{D}$ on $\mathbb{R}$.[5] A weight vector $\mathbf{w}$ induces the corresponding probability mass function $p^{\mathcal{D}}_{\mathbf{w}}$ on the categorical HG typology $\mathfrak{T}$ defined in (10). Obviously, this definition yields a probability mass, namely the sum of the masses $p^{\mathcal{D}}_{\mathbf{w}}(G)$ over all the categorical grammars G in the HG typology $\mathfrak{T}$ is equal to 1.

$$p^{\mathcal{D}}_{\mathbf{w}}(G) = \text{the probability of sampling } \epsilon_1, \dots, \epsilon_n \overset{i.i.d.}{\sim} \mathcal{D} \text{ such that the HG grammar corresponding to the weight vector } \mathbf{w}+\boldsymbol{\epsilon} \text{ is } G \tag{10}$$

The typology of stochastic grammars $\mathfrak{T}_{\mathcal{P}}$ obtained as in section 2 from the categorical HG typology $\mathfrak{T}$ and the family $\mathcal{P} = \{p^{\mathcal{D}}_{\mathbf{w}} \,|\, \mathbf{w} \in \mathbb{R}^n_+\}$ of probability mass functions $p^{\mathcal{D}}_{\mathbf{w}}$ corresponding to all nonnegative weight vectors $\mathbf{w}$ is called *stochastic HG* (SHG; Boersma and Pater, 2016). We denote by $\xrightarrow{SHG}$ the T-orders corresponding to SHG in the sense of definition 2.

To illustrate, figure 2 plots the SHG probability of the mappings $(/\mathsf{cost.us}/, [\mathsf{cos.us}])$ and $(/\mathsf{cost.me}/, [\mathsf{cos.me}])$ as a function of the ranking values of the three constraints C_1, C_2, C_3 listed above. These plots suggest that the implication $(/\mathsf{cost.us}/, [\mathsf{cos.us}]) \xrightarrow{SHG} (/\mathsf{cost.me}/, [\mathsf{cos.me}])$ holds in SHG as well: the probability of the consequent $(/\mathsf{cost.me}/, [\mathsf{cos.me}])$ (plotted in the bottom row) seems to be always larger than the probability of the antecedent $(/\mathsf{cost.us}/, [\mathsf{cos.us}])$ (plotted in the top row). The result obtained in section 3 makes sense of this observation, as follows.

[5] Some component $w_k+\epsilon_k$ of the corrupted weight vector $\mathbf{w}+\boldsymbol{\epsilon}$ could be negative. In this case, $\mathbf{w}+\boldsymbol{\epsilon}$ could correspond to no HG grammar in $\mathfrak{T}$ and the probability mass defined in (10) could therefore add up to less than 1. This problem can be avoided simply by truncating the corrupted weights at zero, namely by replacing $w_k+\epsilon_k$ with $\max\{w_k+\epsilon_k, 0\}$ in the definition of the corrupted weight vector $\mathbf{w}+\boldsymbol{\epsilon}$ (Magri, 2015).



(a) $(/\mathsf{cost.us}/, [\mathsf{cos.us}])$, $\theta_3 = -3$ (b) $(/\mathsf{cost.us}/, [\mathsf{cos.us}])$, $\theta_3 = 0$ (c) $(/\mathsf{cost.us}/, [\mathsf{cos.us}])$, $\theta_3 = 3$

(d) $(/\mathsf{cost.me}/, [\mathsf{cos.me}])$, $\theta_3 = -3$ (e) $(/\mathsf{cost.me}/, [\mathsf{cos.me}])$, $\theta_3 = 0$ (f) $(/\mathsf{cost.me}/, [\mathsf{cos.me}])$, $\theta_3 = 3$

Figure 2: SHG probabilities of the antecedent mapping $(/\mathsf{cost.us}/, [\mathsf{cos.us}])$ and the consequent mapping $(/\mathsf{cost.me}/, [\mathsf{cos.me}])$ as a function of θ_1 (horizontal axis) and θ_2 (vertical axis) for three choices of θ_3

We consider two URs x and $\widehat{\mathsf{x}}$. We assume that x comes with a finite number $m+1$ of candidates $\mathsf{y}, \mathsf{z}_1, \ldots, \mathsf{z}_m$ and that $\widehat{\mathsf{x}}$ comes with a finite number $\widehat{m}+1$ of candidates $\widehat{\mathsf{y}}, \widehat{\mathsf{z}}_1, \ldots, \widehat{\mathsf{z}}_{\widehat{m}}$. This assumption is nonrestrictive. In fact, each UR admits only a finite number of optima in HG (Magri, 2018b). Candidate sets can thus be assumed to be finite without loss of generality. We consider a categorical HG grammar in the typology $\mathfrak{T}$ and assume that it maps x and $\widehat{\mathsf{x}}$ to y and $\widehat{\mathsf{y}}$, respectively. This means that any weight vector $\mathbf{w} = (w_1, \ldots, w_n)$ corresponding to this HG grammar assigns a larger harmony to the winner mappings (x, y) and $\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ than to any of the loser mappings $(\mathsf{x}, \mathsf{z}_i)$ and $(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j)$ respectively, as stated in (11).

$$\begin{aligned} &\underbrace{\min_{i=1,\ldots,m} \sum_k w_k (C_k(\mathsf{x}, \mathsf{z}_i) - C_k(\mathsf{x}, \mathsf{y})) > 0}_{\xi} \\ &\underbrace{\min_{j=1,\ldots,\widehat{m}} \sum_k w_k (C_k(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j) - C_k(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})) > 0}_{\widehat{\xi}} \end{aligned} \tag{11}$$

Let B be an upper bound on the constraint violation differences, so that $|C(\mathsf{x}, \mathsf{z}_i) - C(\mathsf{x}, \mathsf{y})| \leq B$ and $|C(\widehat{\mathsf{x}}, \widehat{\mathsf{z}}_j) - C(\widehat{\mathsf{x}}, \widehat{\mathsf{y}})| \leq B$ for every $i = 1, \ldots, m$ and $j = 1, \ldots, \widehat{m}$. Suppose again that there exists a positive constant Δ large enough that the distribution $\mathcal{D}$ concentrates most of the probability mass on $[-\Delta, +\Delta]$, in the sense that it satisfies the inequality (9). We consider the weight vector $\lambda\mathbf{w} = (\lambda w_1, \ldots, \lambda w_n)$ obtained by rescaling the weight vector $\mathbf{w}$ by a positive scalar $\lambda > 0$ sufficiently large, in the sense of (12).

$$\lambda > \max\left\{\frac{n\Delta B}{\xi}, \frac{n\Delta B}{\widehat{\xi}}\right\} \tag{12}$$

Whenever $\boldsymbol{\epsilon} \in [-\Delta, +\Delta]^n$, the HG grammar corresponding to the stochastic rescaled weight vector $\lambda\mathbf{w} + \boldsymbol{\epsilon}$ maps the UR x to the SR y, as shown in (13). An analogous reasoning shows that it also maps $\widehat{\mathsf{x}}$ to $\widehat{\mathsf{y}}$. In step (13a), we have used the definition (11) of ξ. In step (13b), we have lower bounded $C(\mathsf{x}, \mathsf{z}_i) - C(\mathsf{x}, \mathsf{y})$ with $-B$. In step (13c), we have used the definition (12) of λ.

$$\begin{aligned} &\sum_k (\lambda w_k + \epsilon_k)(C_k(\mathsf{x}, \mathsf{z}_i) - C_k(\mathsf{x}, \mathsf{y})) = \\ &= \lambda \sum_k w_k (C_k(\mathsf{x}, \mathsf{z}_i) - C_k(\mathsf{x}, \mathsf{y})) + \\ &\qquad + \sum_k \epsilon_k (C_k(\mathsf{x}, \mathsf{z}_i) - C_k(\mathsf{x}, \mathsf{y})) \\ &\overset{(a)}{\geq} \lambda\xi + \sum_k \epsilon_k (C_k(\mathsf{x}, \mathsf{z}_i) - C_k(\mathsf{x}, \mathsf{y})) \\ &\overset{(b)}{\geq} \lambda\xi - n\Delta B \overset{(c)}{>} 0 \end{aligned} \tag{13}$$

The intuition behind this reasoning (13) is as follows. The rescaled weight vector $\lambda\mathbf{w}$ generates

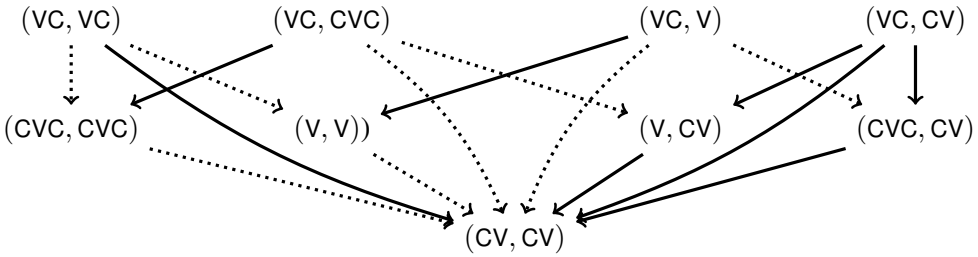


Figure 3: Solid arrows are entailments that hold in OT, HG, SOT, SHG, and ME; dotted arrows are entailments that fail in ME.

the same HG grammar as the weight vector $\mathbf{w}$. If λ is large, the nonzero weights of the rescaled vector $\lambda\mathbf{w}$ are very large (in absolute value). On the other hand, the stochastic values $\epsilon_1, \ldots, \epsilon_n$ are instead small (because bounded between $-\Delta$ and $+\Delta$) and therefore negligible relative to the rescaled weights. The original weight vector $\mathbf{w}$ and the stochastic rescaled vector $\boldsymbol{\epsilon} + \lambda\mathbf{w}$ thus generate the same HG grammar.

In conclusion, the SOT grammar $G_{p^{\mathcal{D}}_{\lambda\mathbf{w}}}$ corresponding to the probability mass function $p^{\mathcal{D}}_{\lambda\mathbf{w}}$ (corresponding to the rescaled weight vector $\lambda\mathbf{w}$) satisfies the identities $G_{p^{\mathcal{D}}_{\lambda\mathbf{w}}}(\mathsf{y}\,|\,\mathsf{x}) > 1/2$ and $G_{p^{\mathcal{D}}_{\lambda\mathbf{w}}}(\widehat{\mathsf{y}}\,|\,\widehat{\mathsf{x}}) > 1/2$. This shows that the family $\mathcal{P} = \{p^{\mathcal{D}}_{\mathbf{w}} \,|\, \mathbf{w} \in \mathbb{R}^n_+\}$ of probability masses is sufficiently rich in the sense of definition 3. Proposition 1 thus yields the following:

Corollary 2 *Under the mild assumption (9) on the distribution $\mathcal{D}$, the T-order $\overset{SHG}{\longrightarrow}$ relative to SHG is identical to the T-order $\overset{HG}{\longrightarrow}$ relative to categorical HG for any constraint set and any candidate set which assigns a finite number of candidates to each UR (while the number of URs can be infinite).* □

6 Conclusions

Phonology has traditionally focused on patterns of categorical alternations modeled within categorical frameworks such as OT and HG. More recently, phonology has extended its empirical coverage to quantitative data such as gradient judgments and patterns of variation. This move has required a parallel extension from categorical to stochastic frameworks, such as partial order OT, stochastic OT, and stochastic HG. These stochastic frameworks are "extensions" of the original categorical frameworks in the sense discussed in section 3. One might thus expect the stochastic frameworks to be typologically less restrictive than the original categorical frameworks. This paper has shown that is not the case, at least when typological restrictiveness is measured in terms of the most basic implicational universals, namely T-orders. Indeed, the T-orders induced by partial order and stochastic OT coincide with those induced by categorical OT. Analogously, the T-orders induced by stochastic HG coincide with those induced by categorical HG.

As discussed in a companion paper (Anttila and Magri, 2018), the situation is very different in ME. To illustrate, consider the *basic syllable system* of Prince and Smolensky (2004). The set of forms consists of the four syllable types CV, CVC, V, and VC. Each of them is a candidate of each other. The constraint set consists of the four constraints ONSET, NOCODA, MAX, and DEP. The HG and OT T-orders coincide and consist of 16 entailments with a feasible antecedent, plotted in figure 3. These entailments extend to SOT and SHG, by virtue of the corollaries 1 and 2 obtained above.

ME instead misses the eight dotted entailments. Of the eight entailments which do survive in ME, seven are such that the antecedent and the consequent surface form coincide, plus the entailment $(\mathsf{VC}, \mathsf{VC}) \rightarrow (\mathsf{CV}, \mathsf{CV})$, which is a quirk due to the fact that VC is the most marked syllable type. This restriction to entailments whose antecedent and consequent surface forms coincide is not phonologically plausible. Anttila and Magri (2018) conclude that the ME formalism imposes typological restrictions at odds with phonological intuition.

Acknowledgments

The research reported in this paper has been funded by the *Agence National de la Recherche* (project title: 'The mathematics of segmental phonotactics'). This paper is part of a larger project on T-orders, developed in collaboration with Arto Anttila. His comments on this paper are gratefully acknowledged.

References

Arto Anttila. 1997a. Deriving variation from grammar: A study of Finnish genitives. In Frans Hinskens, Roeland van Hout, and Leo Wetzels, editors, *Variation, change and phonological theory*, pages 35–68. John Benjamins, Amsterdam. Rutgers Optimality Archive ROA-63, http://ruccs.rutgers.edu/roa.html.

Arto Anttila. 1997b. *Variation in Finnish phonology and morphology*. Ph.D. thesis, Stanford University.

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 Curtis Andrus. 2006. T-orders. Manuscript and software (Stanford).

Arto Anttila and Giorgio Magri. 2018. T-orders across categorical and probabilistic constraint-based phonology. Manuscript (Stanford, CNRS).

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.

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper and Row, New York.

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 and Joe Pater. 2011. The place of variation in phonological theory. In John Goldsmith, Jason Riggle, and Alan Yu, editors, *Handbook of phonological theory*, pages 401–434. Blackwell, Cambridge.

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.

Joseph H. Greenberg. 1963. *Universals of Language*. MIT Press, Cambridge, MA.

G. Guy. 1991. Explanation in variable phonology. *Language Variation and Change*, 3:1–22.

Bruce Hayes. 2017. Varieties of Noisy Harmonic Grammar. In *Proceedings of the 2016 Annual Meeting in Phonology*, pages –.

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

Paul Kiparsky. 1993. An OT perspective on phonological variation. Handout (Stanford).

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

Giorgio Magri. 2015. How to keep the HG weights non-negative: the truncated Perceptron reweighing rule. *Journal of Language Modeling*, 3.2:345–375.

Giorgio Magri. 2018a. Efficient computation of implicational universals in constraint-based phonology through the Hyperplane Separation Theorem. Manuscript (CNRS).

Giorgio Magri. 2018b. Finiteness of optima in constraint-based phonology. Manuscript (Stanford, CNRS).

Alan Prince and Paul Smolensky. 1997. Optimality: From neural networks to universal grammar. *Science*, 275:1604–1610.

Alan Prince and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in generative grammar*. Blackwell, Oxford. Original version, Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, April 1993. Available from the Rutgers Optimality Archive as ROA 537.

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