Efficient Computation of Implicational Universals in Constraint-Based Phonology Through the Hyperplane Separation Theorem

Giorgio Magri
CNRS, SFL, UPL
`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 develops necessary and sufficient constraint characterizations of T-orders within *Harmonic Grammar* and *Optimality Theory*. These conditions rest on the rich convex geometry underlying these frameworks. They are phonologically intuitive and have significant algorithmic implications.

1 Introduction

A *typology* $\mathfrak{T}$ is a collection of grammars $G_1, G_2, \ldots$ For instance, $\mathfrak{T}$ could be the set of syntactic grammars corresponding to all possible combinations of values of a set of parameters (Chomsky, 1981). Or the set of phonological grammars corresponding to all possible orderings of an underlying set of phonological rules (Chomsky and Halle 1968). Or the set of grammars corresponding to all rankings of an underlying constraint set (Prince and Smolensky, 2004).

The structure induced by a typology $\mathfrak{T}$ can be investigated though its *implicational universals* of the form (1). This implication holds provided every grammar in the typology $\mathfrak{T}$ that satisfies the *antecedent* property P also satisfies the *consequent* property $\widehat{P}$ (Greenberg 1963).

$$P \xrightarrow{\mathfrak{T}} \widehat{P} \qquad (1)$$

To illustrate, suppose that $\mathfrak{T}$ is the typology of syntactic grammars. Consider the antecedent property P of having VSO as the basic word order. And the consequent property $\widehat{P}$ of having prepositions (as opposed to postpositions). In this case, (1) is Greenberg's implicational universal #3.

In this paper, we are interested in typologies of *phonological grammars*. We assume a representational framework which distinguishes between two representational levels: *underlying representations* (URs), denoted as $x, \widehat{x}, \ldots$; and *surface representations* (SRs), denoted as $y, \widehat{y}, \ldots$ or $z, \widehat{z}, \ldots$. A *phonological grammar* G is a function which takes a UR x and returns a SR y. For instance, the phonology of German maps the UR $x = $ /bɛːd/ to the SR $y = $ [bɛːt] ('bath'). A *phonological typology* $\mathfrak{T}$ is a collection of phonological grammars $G_1, G_2, \ldots$ that we assume are all defined over the same set of URs (*Richness of the Base* assumption; Prince and Smolensky 2004).

Since phonological grammars are functions from URs to SRs, the most basic or atomic antecedent property P of an implicational universal (1) is the property of mapping a certain UR x to a certain SR y. Analogously, the most basic consequent property $\widehat{P}$ is the property of mapping a certain UR $\widehat{x}$ to a certain SR $\widehat{y}$. We thus focus on implicational universals of the form (2). This implication holds provided every grammar in the typology $\mathfrak{T}$ that succeeds on the antecedent mapping (i.e., it maps the antecedent UR x to the antecedent SR y), also succeeds on the consequent mapping (i.e., it also maps the consequent UR $\widehat{x}$ to the consequent SR $\widehat{y}$). This definition makes sense because every grammar in the typology $\mathfrak{T}$ is defined on every UR, so that every grammar can be applied to the two URs x and $\widehat{x}$.

$$(x, y) \xrightarrow{\mathfrak{T}} (\widehat{x}, \widehat{y}) \qquad (2)$$

The relation $\xrightarrow{\mathfrak{T}}$ thus defined over mappings turns out to be a partial order (under mild additional assumptions). It is called the *T-order* induced by the typology $\mathfrak{T}$ (Anttila and Andrus, 2006).

A familiar example concerns coda cluster simplification in English. Suppose that a coda t/d deletes before vowels in a certain dialect, so that the UR /cost us/ is realized as the SR [cos' us].

Then the coda also deletes before consonants in that same dialect, so that the UR /cost me/ is realized as the SR [cos' me] (Guy, 1991; Kiparsky, 1993; Coetzee, 2004). In other words, the implication $(/tV/, [V]) \xrightarrow{\mathfrak{T}} (/tC/, [C])$ holds relative to the typology $\mathfrak{T}$ of English dialects.

Two important phonological frameworks explored in the literature are *Harmonic Grammar* (HG; Legendre et al., 1990; Smolensky and Legendre, 2006; Potts et al., 2010) and *Optimality Theory* (OT; Prince and Smolensky, 2004). The crucial idea shared by HG and OT is that the relevant properties of phonological mappings are extracted by a set of n phonological *constraints* that effectively represent discrete phonological mappings as points of $\mathbb{R}^n$. The goal of this paper is to express an implication $(x, y) \rightarrow (\widehat{x}, \widehat{y})$ in HG and OT in terms of the constraint violations of the two mappings (x, y) and $(\widehat{x}, \widehat{y})$ and their competitors.

Section 2 presents the constraint condition for HG T-orders. It rests on the rich geometry underlying HG, as it follows from a classical result of convex geometry (the *Hyperplane Separation Theorem*), as detailed in section 3. Section 4 presents the constraint condition for OT T-orders. It rests on an equivalence between OT and HG T-orders established in section 5.

These constraint conditions admit a straightforward interpretation and thus help us better understand the phonological import of T-orders. Furthermore, they allow us to compute T-orders efficiently, circumventing the laborious computation of the entire HG or OT typology (as it is currently done in the literature; see for instance the OT *T-order Generator* by Anttila and Andrus, 2006).

2 Constraint Conditions for HG T-orders

HG assumes a relation Gen which pairs each UR x with a set $Gen(x)$ of *candidate* SRs. It also assumes a set of n phonological constraints $C_1, \ldots, C_n$. Each *constraint* C_k takes a phonological mapping (x, y) of a UR x and a candidate SR y in $Gen(x)$ and returns the corresponding *number of violations* $C_k(x, y) \in \mathbb{N}$, a nonnegative integer which quantifies the "badness" of that mapping (x, y) from the phonological perspective encoded by that constraint C_k. A *weight vector* $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n_+$ 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 multi-plied by -1, namely $-\sum_{k=1}^n w_k C_k(x, y)$. Because of the minus sign, mappings with a large harmony have few constraint violations. The *HG grammar* corresponding to a weight vector $\mathbf{w}$ maps a UR x to the candidate SR y in $Gen(x)$ such that the mapping (x, y) has a larger $\mathbf{w}$-harmony than the mapping (x, z) corresponding to any other candidate z in $Gen(x)$ (Legendre et al., 1990; Smolensky and Legendre, 2006; Potts et al., 2010). The *HG typology* (relative to a candidate relation and a constraint set) consists of the HG grammars corresponding to all weight vectors.

We denote by $(x, y) \xrightarrow{\text{HG}} (\widehat{x}, \widehat{y})$ the implication between an antecedent mapping (x, y) and a consequent mapping $(\widehat{x}, \widehat{y})$ relative to the HG typology. We assume that the antecedent UR x comes with only a finite number m of *antecedent loser* candidates $z_1, \ldots, z_m$ besides the *antecedent winner* candidate y. Analogously, we assume that the consequent UR $\widehat{x}$ comes with only a finite number $\widehat{m}$ of *consequent loser* candidates $\widehat{z}_1, \ldots, \widehat{z}_{\widehat{m}}$ besides the *consequent winner* candidate $\widehat{y}$. This assumption is nonrestrictive. In fact, a UR admits only a finite number of HG optimal candidates (Magri, 2018). Candidate sets can thus be assumed to be finite without loss of generality.

For each antecedent loser z_i, we define the *antecedent difference vector* $\mathbf{C}(x, y, z_i)$ as in (3). It has a component for each constraint C_k defined as the *violation difference* $C_k(x, y, z_i)$ between the number $C_k(x, z_i)$ of violations assigned by C_k to the loser mapping (x, z_i) minus the number $C_k(x, y)$ of violations assigned to the antecedent winner mapping (x, y).

$$\mathbf{C}(x, y, z_i) = \begin{bmatrix} C_1(x, z_i) - C_1(x, y) \\ \vdots \\ C_k(x, z_i) - C_k(x, y) \\ \vdots \\ C_n(x, z_i) - C_n(x, y) \end{bmatrix} \quad (3)$$

The *consequent difference vector* $\mathbf{C}(\widehat{x}, \widehat{y}, \widehat{z}_j)$ is defined analogously, as pitting the consequent winner mapping $(\widehat{x}, \widehat{y})$ against one of its losers $(\widehat{x}, \widehat{z}_j)$.

The definition of the HG implication $(x, y) \xrightarrow{\text{HG}} (\widehat{x}, \widehat{y})$ requires every HG grammar which succeeds on the antecedent mapping to also succeed on the consequent mapping. This condition is trivially satisfied if no HG grammar succeeds on the antecedent mapping, namely the mapping (x, y) is HG unfeasible. Thus, let's suppose that is not the case. The following proposition then provides a
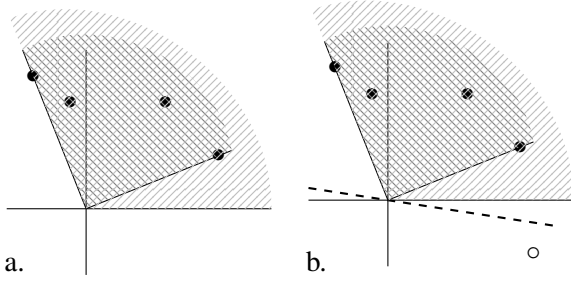
Figure 1: Geometric representation of condition (4).

		ONSET	NOCODA	MAX	DEPV	DEPC
/CC/	[null]	0	0	2	0	0
	[CV.CV]	0	0	0	2	0
	[CVC]	0	1	0	1	0
/CCC/	[null]	0	0	3	0	0
	[CV.CV.CV]	0	0	0	3	0
	[CV.CVC]	0	1	0	2	0
	[CVC]	0	1	1	1	0

Table 1: Violation profiles for the mappings of the URs /CC/ and /CCC/ to their non-harmonically bounded candidates.

complete (both necessary and sufficient) characterization of the HG implication $(x, y) \stackrel{HG}{\to} (\widehat{x}, \widehat{y})$ in terms of condition (4) stated entirely in terms of antecedent and consequent difference vectors.

Proposition 1 *If the antecedent mapping* (x, y) *is HG feasible, the HG implication* $(x, y) \stackrel{HG}{\to} (\widehat{x}, \widehat{y})$ *holds if and only if for every consequent loser candidate* $\widehat{z}_j$ *with* $j = 1, \ldots, \widehat{m}$, *there exist* m *nonnegative coefficients* $\lambda_1, \ldots, \lambda_m \geq 0$ *(one for each antecedent loser candidate* $z_1, \ldots, z_m$*) such that*

$$\mathbf{C}(\widehat{x}, \widehat{y}, \widehat{z}_j) \geq \sum_{i=1}^{m} \lambda_i \, \mathbf{C}(x, y, z_i) \qquad (4)$$

and furthermore at least one of these coefficients $\lambda_1, \ldots, \lambda_m$ *is different from zero.* □

Proposition 1 admits the following phonological interpretation. Condition (4) says that each consequent loser $\widehat{z}_j$ violates the constraints at least as much as (some conic combination of) the antecedent losers $z_1, \ldots, z_m$. In other words, the consequent losers are "worse" than the antecedent losers. The consequent winner $\widehat{y}$ thus has an "easier" time beating its losers than the antecedent winner y, as required by the definition of T-order.

Proposition 1 has important algorithmic implications. In fact, checking the definition of T-order (in general, of any implicational universal) directly is costly, because it requires computing the entire typology, which can be large. But proposition 1 says that, in the case of HG, T-orders can be determined locally, by only looking at the antecedent and consequent mappings together with their losers. Indeed, this proposition effectively reduces the problem of computing HG T-orders to the problem of finding coefficients λ_i which satisfy the inequality (4). The latter is a polyhedral feasibility problem that can be solved efficiently with standard linear programming technology. A Python package to compute HG T-orders using condition (4) will be released shortly.

Proposition 1 admits the following geometric interpretation. Suppose there are only $n = 2$ constraints and $m = 4$ antecedent difference vectors $\mathbf{C}(x, y, z_i)$, represented as the black dots in figure 1a. 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 antecedent difference vectors, depicted in dark gray in figure 1a. The region in light gray singles out the points which are at least as large (component by component) as some point in this cone. Condition (4) thus says that each consequent difference vector $\mathbf{C}(\widehat{x}, \widehat{y}, \widehat{z}_j)$ must belong to this light gray region.

Indeed, suppose that some consequent difference vector does not belong to this light gray region, as represented by the white dot in figure 1b. The dashed line leaves the antecedent difference vectors (black dots) and the consequent difference vector (white dot) on two different sides. This means that the HG grammar corresponding to a nonnegative weight vector orthogonal to this line succeeds on the antecedent mapping (x, y) but it fails on the consequent mapping $(\widehat{x}, \widehat{y})$, defying the implication $(x, y) \stackrel{HG}{\to} (\widehat{x}, \widehat{y})$.

The existence of (a weight vector corresponding to) a dashed line such as the one depicted in figure 1b is geometrically obvious in the case with only $n = 2$ constraints. For an arbitrary number n of constraints, a fundamental result of convex geometry, the *Hyperplane Separation Theorem* (HST; Rockafellar, 1970, §11; Boyd and Vandenberghe, 2004, §2.5), indeed guarantees the existence of a weight vector which separates the cone generated by the antecedent difference vectors from the outlier consequent difference vector. This is the core of the proof of proposition 1 provided in section 3.

Let's finally look at a couple of examples (based on Bane and Riggle 2009). We assume $n = 5$ con-

$$\begin{array}{l}
\text{ONSET} \\ \text{NOCODA} \\ \text{MAX} \\ \text{DEPV} \\ \text{DEPC}
\end{array}
\begin{bmatrix} 0 \\ 0 \\ 3 \\ -3 \\ 0 \end{bmatrix} \geq 1.5
\begin{bmatrix} 0 \\ 0 \\ 2 \\ -2 \\ 0 \end{bmatrix} + 0
\begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \quad
\begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{bmatrix} \geq 0
\begin{bmatrix} 0 \\ 0 \\ 2 \\ -2 \\ 0 \end{bmatrix} + 1
\begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \quad
\begin{bmatrix} 0 \\ 1 \\ 1 \\ -2 \\ 0 \end{bmatrix} \geq 0.5
\begin{bmatrix} 0 \\ 0 \\ 2 \\ -2 \\ 0 \end{bmatrix} + 1
\begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}$$

Table 2: Verifying that condition (4) holds for the HG implication $(\text{CC}, \text{CV.CV}) \rightarrow (\text{CCC}, \text{CV.CV.CV})$.

straints: ONSET, which penalizes surface syllables starting with a vowel (V); NOCODA, which penalizes surface syllables ending with a consonant (C); MAX, which penalizes deletion of underlying segments; and DEPV and DEPC, which penalize epenthetic vowels and consonants, respectively. We focus on the two URs /CC/ and /CCC/. We only consider their non-harmonically bounded candidates, listed in table 1 with their constraint violations (the candidate [CVC.CV] is omitted because indistinguishable by the constraints from [CV.CVC]).

We focus on the implication $(\text{CC}, \text{CV.CV}) \rightarrow (\text{CCC}, \text{CV.CV.CV})$. The antecedent UR $\text{x} = $ /CC/ comes with the winner candidate $\text{y} = $ [CV.CV] and the $m = 2$ loser candidates $\text{z}_1 = $ [null] and $\text{z}_2 = $ [CVC]. There are therefore two antecedent difference vectors $\mathbf{C}(\text{x}, \text{y}, \text{z}_i)$, repeated on the right hand side of each of the three inequalities in table 2. The consequent UR $\widehat{\text{x}} = $ /CCC/ comes with the winner candidate $\widehat{\text{y}} = $ [CV.CV.CV] and the $\widehat{m} = 3$ loser candidates $\widehat{\text{z}}_1 = $ [null], $\widehat{\text{z}}_2 = $ [CV.CVC], and $\widehat{\text{z}}_3 = $ [CVC]. There are therefore three consequent difference vectors $\mathbf{C}(\widehat{\text{x}}, \widehat{\text{y}}, \widehat{\text{z}}_j)$, which appear on the left hand side of the three inequalities in table 2. Condition (4) holds: each consequent difference vector is at least as large as a conic combination of the antecedent difference vectors, as shown in table 2. Proposition 1 thus establishes the HG implication $(\text{CC}, \text{CV.CV}) \overset{\text{HG}}{\rightarrow} (\text{CCC}, \text{CV.CV.CV})$.

Proposition 1 can also be used to show that an implication fails in HG. To illustrate, we focus on the implication $(\text{CC}, \text{CVC}) \rightarrow (\text{CCC}, \text{CV.CVC})$. We consider the consequent difference vector $\mathbf{C}(\text{/CCC/}, [\text{CV.CVC}], [\text{null}])$, which appears on the left hand side of (5). There are two antecedent difference vectors $\mathbf{C}(\text{/CC/}, [\text{CVC}], [\text{null}])$ and $\mathbf{C}(\text{/CC/}, [\text{CVC}], [\text{CV.CV}])$, which appear on the right hand side of (5).

$$\begin{array}{l}
\text{ONSET} \\ \text{NOCODA} \\ \text{MAX} \\ \text{DEPV} \\ \text{DEPC}
\end{array}
\begin{bmatrix} 0 \\ -1 \\ 3 \\ -2 \\ 0 \end{bmatrix} \not\geq \lambda_1
\begin{bmatrix} 0 \\ -1 \\ 2 \\ -1 \\ 0 \end{bmatrix} + \lambda_2
\begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \qquad (5)$$

Condition (4) fails: the consequent difference vector is not larger than any conic combination of the two antecedent difference vectors, no matter the choice of the coefficients $\lambda_1, \lambda_2 \geq 0$. In fact, the inequality (5) for DEPV requires $\lambda_1 \geq 2$, whereby the inequality fails for MAX. Proposition 1 thus establishes that the implication $(\text{CC}, \text{CVC}) \overset{\text{HG}}{\not\rightarrow} (\text{CCC}, \text{CV.CVC})$ fails in HG.

3 Proof of Proposition 1

The HST has a number of algebraic consequences known as *theorems of the alternatives*.[1] One of these theorems is the *Motzkin Transposition Theorem* (MTT; Bertsekas, 2009, proposition 5.6.2), which is particularly suited to our needs. It states that conditions (C1) and (C2) below are mutually exclusive (one and only one of them holds) for any two matrices $\mathbf{A} \in \mathbb{R}^{p \times n}$ and $\mathbf{B} \in \mathbb{R}^{q \times n}$.

(C1) There exists a vector $\mathbf{w} \in \mathbb{R}^n$ such that $\mathbf{A}\mathbf{w} < \mathbf{0}$ and $\mathbf{B}\mathbf{w} \leq \mathbf{0}$.

(C2) There exist two nonnegative vectors $\boldsymbol{\xi} \in \mathbb{R}_+^q$ and $\boldsymbol{\mu} \in \mathbb{R}_+^p$ with $\boldsymbol{\mu} \neq \mathbf{0}$ such that $\mathbf{A}^{\mathsf{T}}\boldsymbol{\mu} + \mathbf{B}^{\mathsf{T}}\boldsymbol{\xi} = \mathbf{0}$.

It is useful to specialize the MTT as follows. Consider some vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m, \mathbf{b} \in \mathbb{R}^n$. Let $\mathbf{A}$ be the matrix whose $p = m$ rows are $-\mathbf{a}_1^{\mathsf{T}}, \ldots, -\mathbf{a}_m^{\mathsf{T}}$. Let $\mathbf{B}$ be the matrix whose $q = n+1$ rows are $-\mathbf{e}_1^{\mathsf{T}}, \ldots, -\mathbf{e}_n^{\mathsf{T}}, \mathbf{b}^{\mathsf{T}}$ (where $\mathbf{e}_i \in \mathbb{R}^n$ has all components equal to 0 but for the ith component which is equal to 1). The two conditions (C1) and (C2) thus become (C1′) and (C2′).

(C1′) There exists a nonnegative vector $\mathbf{w} \in \mathbb{R}_+^n$ such that $\mathbf{a}_1^{\mathsf{T}}\mathbf{w} > 0, \ldots, \mathbf{a}_m^{\mathsf{T}}\mathbf{w} > 0$ but $\mathbf{b}^{\mathsf{T}}\mathbf{w} \leq 0$.

(C2′) There exist some nonnegative coefficients $\mu_1, \ldots, \mu_m, \xi \geq 0$ with at least one of the coefficients $\mu_1, \ldots, \mu_m$ different from 0 such that $\xi \mathbf{b} \geq \sum_{i=1}^m \mu_i \mathbf{a}_i$.

---

[1] Throughout this section, all vectors are column vectors; T stands for matrix transposition. Vector inequalities must hold component-wise.

With these preliminaries in place, we now consider the HG implication $(\mathsf{x},\mathsf{y}) \overset{\text{HG}}{\to} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$. Suppose that the HG grammar corresponding to some non-negative weight vector $\mathbf{w} \in \mathbb{R}^n_+$ succeeds on the antecedent mapping (x,y). This means that the $\mathbf{w}$-harmony of this mapping (x,y) is larger than that of every antecedent loser mapping $(\mathsf{x},\mathsf{z}_i)$. This condition can be stated in terms of the antecedent difference vectors as in (6), taking advantage of the linearity of the HG harmony.

$$\mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}_i)^\mathsf{T}\mathbf{w} > 0, \qquad i = 1,\ldots,m \quad (6)$$

The implication $(\mathsf{x},\mathsf{y}) \overset{\text{HG}}{\to} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ then requires the HG grammar corresponding to that weight vector $\mathbf{w}$ to also succeed on the consequent mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$. This means that the $\mathbf{w}$-harmony of this mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ is larger than that of every consequent loser mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{z}}_j)$. This condition can be stated in terms of the consequent difference vectors as in (7).

$$\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}},\widehat{\mathsf{z}}_j)^\mathsf{T}\mathbf{w} > 0, \qquad j = 1,\ldots,\widehat{m} \quad (7)$$

In other words, the HG implication $(\mathsf{x},\mathsf{y}) \overset{\text{HG}}{\to} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ holds if and only if every nonnegative weight vector $\mathbf{w}$ which satisfies (6) also satisfies (7). Equivalently, the HG T-order holds if and only if for every $j = 1,\ldots,\widehat{m}$, it is false that there exists a nonnegative weight vector $\mathbf{w} \in \mathbb{R}^n_+$ such that $\mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}_i)^\mathsf{T}\mathbf{w} > 0$ for every $i = 1,\ldots,m$ but $\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}},\widehat{\mathsf{z}}_j)^\mathsf{T}\mathbf{w} \leq 0$. In other words, for every $j = 1,\ldots,\widehat{m}$, condition (C1′) is false, with the positions $\mathbf{a}_i = \mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}_i)$ and $\mathbf{b} = \mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}},\widehat{\mathsf{z}}_j)$. By the MTT, condition (C2′) must therefore be true for every $j = 1,\ldots,\widehat{m}$. This means that there exist some non-negative coefficients $\mu_1,\ldots,\mu_m,\xi \geq 0$ such that at least one of the coefficients $\mu_1,\ldots,\mu_m$ is strictly positive and furthermore the inequality (8) holds.

$$\xi\,\mathbf{C}(\widehat{\mathsf{x}},\widehat{\mathsf{y}},\widehat{\mathsf{z}}_j) \geq \sum_{i=1}^{m} \mu_i\,\mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}_i) \quad (8)$$

The coefficient ξ in (8) must be strictly positive. In fact, suppose by contradiction that $\xi = 0$, whereby inequality (8) becomes (9).

$$\mathbf{0} \geq \sum_{i=1}^{m} \mu_i\,\mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}_i) \quad (9)$$

Consider a weight vector $\mathbf{w}$ whose corresponding HG grammar maps the antecedent UR x to the antecedent winner y, which exists by hypothesis. This weight vector $\mathbf{w}$ thus satisfies condition (6). Since $\mathbf{w}$ is non-negative, the scalar product of both sides of (9) with $\mathbf{w}$ preserves the inequality, yielding (10). But the latter inequality requires $\mu_1 = \cdots = \mu_m = 0$, contradicting the assumption that at least one of the nonnegative coefficients $\mu_1,\ldots,\mu_m \geq 0$ is strictly positive.

$$0 \geq \sum_{i=1}^{m} \mu_i\,\underbrace{\mathbf{C}(\mathsf{x},\mathsf{y},\mathsf{z}_i)^\mathsf{T}\mathbf{w}}_{>0} \quad (10)$$

Since the coefficient ξ is strictly positive, both sides of (8) can be divided by ξ, yielding the inequality (4) with the position $\lambda_i = \mu_i/\xi$.

4 Constraint Conditions for OT T-orders

This section extends the convex geometric analysis of T-orders developed in the preceding sections from HG to OT. We start by recalling that in OT a constraint C_k is said to *prefer* 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 fact that the highest $\gg$-ranked relevant constraint defines the preference of the entire ranking, irrespectively of the preferences of lower $\gg$-ranked constraints, is captured by saying that the former constraint *strictly dominates* the latter constraints. The *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). The OT typology (for a given candidate relation and constraint set) consists of the OT grammars corresponding to all rankings.

We denote by $(\mathsf{x},\mathsf{y}) \overset{\text{OT}}{\to} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ the implication between an antecedent mapping (x,y) and a consequent mapping $(\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ relative to the OT typology. By definition, this implication holds provided every constraint ranking that succeeds on the antecedent mapping also succeeds on the consequent mapping. Thus, a natural strategy to check the OT implication $(\mathsf{x},\mathsf{y}) \overset{\text{OT}}{\to} (\widehat{\mathsf{x}},\widehat{\mathsf{y}})$ would be to use Recursive Constraint Demotion (RCD; Tesar and Smolensky, 1998) to check that for every $j =$

$1, \ldots, \widehat{m}$, no ranking is consistent simultaneously with the two mappings (x, y) and $(\widehat{x}, \widehat{z}_j)$. In this section, we develop instead an alternative strategy which uses the HG-to-OT-portability result of Magri (2013) to extend to OT the convex geometric characterization of HG T-orders developed in sections 2-3.

To start, we recall that an OT grammar can be construed as an HG grammar (as long as the constraint violations are bounded, which is the case when the set of URs and the candidate sets are finite). In fact, OT's strict domination can be mimicked through HG weights which decrease exponentially. Indeed, if a weight is much larger than every smaller weight, the preferences of the constraint with the larger weight cannot be overcome by the preferences of the constraints with smaller weights (Prince and Smolensky, 2004; Keller, 2006). Since the OT typology is a subset of the HG typology, whenever an implication $(x, y) \overset{\text{HG}}{\to} (\widehat{x}, \widehat{y})$ holds in HG, the implication $(x, y) \overset{\text{OT}}{\to} (\widehat{x}, \widehat{y})$ holds in OT.

Lemma 1 slightly strengthens this conclusion. In fact, OT only cares about constraints' preferences. Equivalently, about the sign of the violation differences. Thus, the HG implication $(x, y) \overset{\text{HG}}{\to} (\widehat{x}, \widehat{y})$ entails not only the corresponding OT implication $(x, y) \overset{\text{OT}}{\to} (\widehat{x}, \widehat{y})$ but also any other OT implication $(x^*, y^*) \overset{\text{OT}}{\to} (\widehat{x}^*, \widehat{y}^*)$ whose antecedent and consequent mappings (x^*, y^*) and $(\widehat{x}^*, \widehat{y}^*)$ yield violation differences with the same sign as the original antecedent and consequent mappings (x, y) and $(\widehat{x}, \widehat{y})$. The proof of this lemma recalled in the appendix is a straightforward variation on the observation that exponentially decaying HG weights mimic OT strict domination.

Lemma 1 *Given an antecedent mapping (x, y) with its m antecedent loser candidates $z_1, \ldots, z_m$, consider another mapping (x^*, y^*) with the same number m of loser candidates $z_1^*, \ldots, z_m^*$ such that the m corresponding violation differences have the same sign, in the sense that condition (11) holds for $k = 1, \ldots, n$ and $i = 1, \ldots, m$.*

$$C_k(x, y, z_i) \gtreqqless 0 \iff C_k(x^*, y^*, z_i^*) \gtreqqless 0 \quad (11)$$

Analogously, given the consequent mapping $(\widehat{x}, \widehat{y})$ with its $\widehat{m}$ consequent loser candidates $\widehat{z}_1, \ldots, \widehat{z}_{\widehat{m}}$, consider another mapping $(\widehat{x}^*, \widehat{y}^*)$ with the same number $\widehat{m}$ of loser candidates $\widehat{z}_1^*, \ldots, \widehat{z}_{\widehat{m}}^*$ such that the $\widehat{m}$ corresponding violation differences have the same sign, in the sense

that condition (12) holds for $k = 1, \ldots, n$ and $j = 1, \ldots, \widehat{m}$.

$$C_k(\widehat{x}, \widehat{y}, \widehat{z}_j) \gtreqqless 0 \iff C_k(\widehat{x}^*, \widehat{y}^*, \widehat{z}_j^*) \gtreqqless 0 \quad (12)$$

The HG implication $(x, y) \overset{\text{HG}}{\longrightarrow} (\widehat{x}, \widehat{y})$ then entails the OT implication $(x^*, y^*) \overset{\text{OT}}{\longrightarrow} (\widehat{x}^*, \widehat{y}^*)$.

The preceding lemma establishes an entailment from HG to OT implications. We now want to investigate the reverse entailment from OT to HG implications. Thus, we suppose that an implication $(x, y) \overset{\text{OT}}{\to} (\widehat{x}, \widehat{y})$ holds in OT. Of course, that does not entail that the implication $(x, y) \overset{\text{HG}}{\to} (\widehat{x}, \widehat{y})$ between the same two mappings also holds in HG. That is because the HG typology is usually a *proper* superset of the OT typology. And a larger typology yields sparser T-orders. Thus, it makes no sense to try to establish that the OT implication $(x, y) \overset{\text{OT}}{\to} (\widehat{x}, \widehat{y})$ entails the HG implication $(x, y) \overset{\text{HG}}{\to} (\widehat{x}, \widehat{y})$ between the same two mappings.

We will try to establish something weaker instead: the OT implication $(x, y) \overset{\text{OT}}{\to} (\widehat{x}, \widehat{y})$ entails an HG implication $(x^{\text{dif}}, y^{\text{dif}}) \overset{\text{HG}}{\to} (\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$ between an antecedent mapping $(x^{\text{dif}}, y^{\text{dif}})$ different from (x, y) and a consequent mapping $(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$ different from $(\widehat{x}, \widehat{y})$. And we will choose this new antecedent mapping $(x^{\text{dif}}, y^{\text{dif}})$ and this new consequent mapping $(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$ in such a way that the new HG implication $(x^{\text{dif}}, y^{\text{dif}}) \overset{\text{HG}}{\to} (\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$ is "more likely to hold" than the original implication $(x, y) \overset{\text{HG}}{\to} (\widehat{x}, \widehat{y})$ and thus validates the entailment from OT to HG implications.

What does it mean that an implication is "more likely to hold"? Intuitively, an implication from an antecedent to a consequent mapping is "likely to hold" when the antecedent mapping is "difficult" to obtain, namely it is consistent with very few grammars. In the limit, the implication holds trivially when the antecedent mapping is consistent with no grammars at all. Thus, we want to define the new antecedent mapping $(x^{\text{dif}}, y^{\text{dif}})$ in such a way that it is "more difficult" to obtain in HG than the original antecedent mapping (x, y), whereby the superscript "diff". Analogously, an implication from an antecedent to a consequent mapping is intuitively "likely to hold" when the consequent mapping is "easy" to obtain, namely it is consistent with very many grammars. In the limit, the implication holds trivially when the consequent mapping is consistent with every grammar. Thus, we want to define the new consequent

mapping $(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$ in such a way that it is "easier" to obtain in HG than the original consequent mapping $(\widehat{x}, \widehat{y})$, whereby the superscript "easy".

Let us now turn to the details. As discussed above around (6), it suffices to define the difference vectors corresponding to the new difficult mapping antecedent $(x^{\text{dif}}, y^{\text{dif}})$. Given the original antecedent mapping (x, y) with its m loser candidates $z_1, \ldots, z_m$, we assume that the new antecedent mapping $(x^{\text{dif}}, y^{\text{dif}})$ comes with the same number m of loser candidates $z_1^{\text{dif}}, \ldots, z_m^{\text{dif}}$ whose violation differences are defined as in (13). Here, Ω_i is the total number of constraints C_k such that C_k prefers the original antecedent winner mapping (x, y) to the original antecedent loser mapping (x, z_i), in the sense that $C_k(x, y, z_i) > 0$.

$$C_k(x^{\text{dif}}, y^{\text{dif}}, z_i^{\text{dif}}) =$$
$$= \begin{cases} 1 & \text{if } C_k(x, y, z_i) > 0 \\ 0 & \text{if } C_k(x, y, z_i) = 0 \\ -\Omega_i - 1 & \text{if } C_k(x, y, z_i) < 0 \end{cases} \quad (13)$$

The intuition behind this definition (13) is as follows. OT only cares about the sign of the violation differences. Thus, the new violation difference $C_k(x^{\text{dif}}, y^{\text{dif}}, z_i^{\text{dif}})$ is defined in such a way that it has the same sign as the original violation difference $C_k(x, y, z_i)$: one is positive or negative if and only if the other is as well. HG also cares about the size of the violation differences, not only about their sign. In order for the mapping $(x^{\text{dif}}, y^{\text{dif}})$ to be "difficult" in HG, we want its positive violation differences to be as small as possible. For this reason, the positive violation differences in (13) have been set equal to 1, which is the smallest positive integer. Analogously, in order for the mapping $(x^{\text{dif}}, y^{\text{dif}})$ to be "difficult" in HG, we want its negative violation differences to be large (in absolute value) relative to the strength of the positive violation differences they have to "fight off". Since the positive entries are all equal to 1 in (13), the "strength" of the positive entries only depends on their number Ω_i. For this reason, the absolute value of the negative violation differences in (13) has been set equal to $\Omega_i + 1$. In conclusion, this definition (13) ensures that the mapping $(x^{\text{dif}}, y^{\text{dif}})$ is "difficult" in HG, because the positive violation differences are small and the negative ones are large (in absolute value).

We now turn to the consequents. Given the original consequent mapping $(\widehat{x}, \widehat{y})$ with its $\widehat{m}$ loser candidates $\widehat{z}_1, \ldots, \widehat{z}_{\widehat{m}}$, we assume that the new

consequent mapping $(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$ comes with the same number $\widehat{m}$ of loser candidates $\widehat{z}_1^{\text{easy}}, \ldots, \widehat{z}_{\widehat{m}}^{\text{easy}}$ whose violation differences are defined as in (14). Here $\widehat{\Lambda}_j$ is the total number of constraints C_k such that C_k prefers the original consequent loser mapping $(\widehat{x}, \widehat{z}_j)$ to the original consequent winner mapping $(\widehat{x}, \widehat{y})$, in the sense that $C_k(\widehat{x}, \widehat{y}, \widehat{z}_j) < 0$.

$$C_k(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}}, \widehat{z}_j^{\text{easy}}) =$$
$$= \begin{cases} \widehat{\Lambda}_j + 1 & \text{if } C_k(\widehat{x}, \widehat{y}, \widehat{z}_j) > 0 \\ 0 & \text{if } C_k(\widehat{x}, \widehat{y}, \widehat{z}_j) = 0 \\ -1 & \text{if } C_k(\widehat{x}, \widehat{y}, \widehat{z}_j) < 0 \end{cases} \quad (14)$$

The intuition behind this definition (14) is as follows. Whenever the original violation difference $C_k(\widehat{x}, \widehat{y}, \widehat{z}_j)$ is positive or negative, the new violation difference $C_k(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}}, \widehat{z}_j^{\text{easy}})$ is positive or negative as well, so that the original and the new violation differences have the same sign. The size of the new violation differences has been chosen as follows. In order for the mapping $(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$ to be "easy" in HG, we want its negative violation differences to be as small as possible (in absolute value). For this reason, the negative violation differences in (14) have been set equal to -1, which is the negative integer smallest in absolute value. Analogously, in order for the mapping $(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$ to be "easy" in HG, we want its positive violation differences to be large relative to the strength of the negative violation differences they have to "fight off". Since the negative entries are all equal to -1 in (14), the "strength" of the negative entries only depends on their number $\widehat{\Lambda}_j$. For this reason, the positive violation differences in (14) have been set equal to $\widehat{\Lambda}_j + 1$. In conclusion, this definition (14) ensures that the mapping $(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$ is "easy" in HG, because the positive violation differences are large and the negative violation differences are small (in absolute value).

We are now ready to put the pieces together. As anticipated, the OT implication $(x, y) \stackrel{\text{OT}}{\to} (\widehat{x}, \widehat{y})$ might not entail the HG implication $(x, y) \stackrel{\text{HG}}{\to} (\widehat{x}, \widehat{y})$ with the same antecedent and consequent mappings. Nonetheless, the following lemma 2 ensures that the OT implication $(x, y) \stackrel{\text{OT}}{\to} (\widehat{x}, \widehat{y})$ does entail the HG implication $(x^{\text{dif}}, y^{\text{dif}}) \stackrel{\text{HG}}{\to} (\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$. The intuition is that the latter is less demanding than the HG implication $(x, y) \stackrel{\text{HG}}{\to} (\widehat{x}, \widehat{y})$, because its antecedent is "difficult" (namely, consistent with few HG grammars) and its consequent is "easy" (namely, consistent

with many HG grammars). The proof of this lemma is provided in section 5, mimicking a reasoning in Magri (2013).

Lemma 2 *The OT implication* $(x, y) \overset{\text{OT}}{\to} (\widehat{x}, \widehat{y})$ *entails the HG implication* $(x^{dif}, y^{dif}) \overset{\text{HG}}{\to} (\widehat{x}^{easy}, \widehat{y}^{easy})$ *between the antecedent mapping* (x^{dif}, y^{dif}) *and the consequent mapping* $(\widehat{x}^{easy}, \widehat{y}^{easy})$ *whose violation differences are defined in (13) and (14).* ☐

As remarked explicitly above, (13) ensures that the original antecedent violation differences $C_k(x, y, z_i)$ and the new antecedent violation differences $C_k(x^{\text{dif}}, y^{\text{dif}}, z_i^{\text{dif}})$ have the same sign. In other words, condition (11) holds with the positions $x^* = x^{\text{dif}}$, $y^* = y^{\text{dif}}$, and $z_i^* = z_i^{\text{dif}}$. Analogously, (14) ensures that the original consequent violation differences $C_k(\widehat{x}, \widehat{y}, \widehat{z}_j)$ and the new consequent violation differences $C_k(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}}, \widehat{z}_i^{\text{easy}})$ have the same sign. In other words, condition (12) holds with the positions $\widehat{x}^* = \widehat{x}^{\text{easy}}$, $\widehat{y}^* = \widehat{y}^{\text{easy}}$, and $\widehat{z}_i^* = \widehat{z}_i^{\text{easy}}$. The two lemmas 1 and 2 can therefore be combined into the following conclusion: the OT implication $(x, y) \overset{\text{OT}}{\to} (\widehat{x}, \widehat{y})$ holds if and only if the HG implication $(x^{\text{dif}}, y^{\text{dif}}) \overset{\text{HG}}{\to} (\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$ holds. We can thus extend to OT the characterization of HG T-orders provided by the HG proposition 1 above, obtaining the following:

Proposition 2 *If the antecedent mapping* (x, y) *is OT feasible, the OT implication* $(x, y) \overset{\text{OT}}{\to} (\widehat{x}, \widehat{y})$ *holds iff for every* $j = 1, \ldots, \widehat{m}$, *there exist* m *nonnegative coefficients* $\lambda_1, \ldots, \lambda_m \geq 0$ *such that*

$$\mathbf{C}(\widehat{x}^{easy}, \widehat{y}^{easy}, \widehat{z}_j^{easy}) \geq \sum_{i=1}^{m} \lambda_i \, \mathbf{C}(x^{dif}, y^{dif}, z_i^{dif}) \tag{15}$$

and furthermore at least one of these coefficients $\lambda_1, \ldots, \lambda_m$ *is different from zero.* ☐

To illustrate, we have seen at the end of section 2 that the HG implication $(\text{cc}, \text{cvc}) \overset{\text{HG}}{\to} (\text{ccc}, \text{cv.cvc})$ fails in HG because condition (4) fails, as shown in (5). But this entailment $(\text{cc}, \text{cvc}) \overset{\text{OT}}{\to} (\text{ccc}, \text{cv.cvc})$ does hold in OT. In fact, the three "easy" consequent difference vectors $\mathbf{C}(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}}, \widehat{z}_j^{\text{easy}})$ in this case are listed on the left hand side of the three inequalities in table 3. The two "difficult" antecedent difference vectors $\mathbf{C}(x^{\text{dif}}, y^{\text{dif}}, z_i^{\text{dif}})$ are repeated on the right hand side of the three inequalities. The table thus shows that condition (15) holds.

5 Proof of Lemma 2

We assume that the OT implication $(x, y) \overset{\text{OT}}{\to} (\widehat{x}, \widehat{y})$ holds. We consider an arbitrary nonnegative weight vector $\mathbf{w} = (w_1, \ldots, w_n)$ which succeeds on the "difficult" antecedent mapping $(x^{\text{dif}}, y^{\text{dif}})$ and we prove that it is also succeeds on the "easy" consequent mapping $(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$, thus securing the HG implication $(x^{\text{dif}}, y^{\text{dif}}) \overset{\text{HG}}{\longrightarrow} (\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$.

The assumption that the weight vector $\mathbf{w}$ succeeds on the "difficult" antecedent mapping $(x^{\text{dif}}, y^{\text{dif}})$ means that $\sum_{k=1}^{n} w_k C_k(x^{\text{dif}}, y^{\text{dif}}, z_i^{\text{dif}}) > 0$ for every $i = 1, \ldots, m$. The latter inequality can be unpacked as in (16). In step (16a), we have used the definition (13). Here $W(x, y, z_i)$ and $L(x, y, z_i)$ are the sets of winner-preferring and loser-preferring constraints relative to the winner (x, y) and the loser (x, z_i). In step (16b), we have upper bounded the sum $\sum_{h \in W(x, y, z_i)} w_h$ with its largest term $\max_{h \in W(x, y, z_i)} w_h$ times the number Ω_i of its addenda. In step (16c), we have lower bounded the sum $\sum_{k \in L(x, y, z_i)} w_k$ with one of its terms, as the addenda are all non-negative.

$$\sum_{k=1}^{n} w_k C_k(x^{\text{dif}}, y^{\text{dif}}, z_i^{\text{dif}}) > 0$$

$$\overset{(a)}{\Longleftrightarrow} \sum_{h \in W(x, y, z_i)} w_h > (\Omega_i + 1) \sum_{k \in L(x, y, z_i)} w_k$$

$$\overset{(b)}{\Longrightarrow} \Omega_i \max_{h \in W(x, y, z_i)} w_h > (\Omega_i + 1) \sum_{k \in L(x, y, z_i)} w_k$$

$$\overset{(c)}{\Longrightarrow} \Omega_i \max_{h \in W(x, y, z_i)} w_h > (\Omega_i + 1) w_k$$
$$\text{for every } k \in L(x, y, z_i)$$

$$\Longrightarrow \max_{h \in W(x, y, z_i)} w_h > w_k \tag{16}$$

We now show that the conclusion reached in the last line of (16) entails that the strict inequality (17) holds for every $j = 1, \ldots, \widehat{m}$.

$$\max_{h \in W(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_h > \max_{k \in L(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_k \tag{17}$$

In fact, suppose by contradiction that (17) fails for some $j = 1, \ldots, \widehat{m}$. Consider a ranking $\gg$ which respects the relative size of the weights, in the sense that conditions [A] and [B] hold for any two constraints C_s, C_t with weights w_s, w_t.

[A] If $w_s > w_t$, then C_s is $\gg$-ranked above C_k.

[B] If $w_s = w_t$ and $C_s \in L(\widehat{x}, \widehat{y}, \widehat{z}_j)$ and $C_t \in W(\widehat{x}, \widehat{y}, \widehat{z}_j)$, then C_s is $\gg$-ranked above C_k.

$$
\begin{array}{r}
\text{ONSET} \\ \text{NOCODA} \\ \text{MAX} \\ \text{DEPV} \\ \text{DEPC}
\end{array}
\begin{bmatrix} 0 \\ -1 \\ 3 \\ -1 \\ 0 \end{bmatrix} \geq 0.5
\begin{bmatrix} 0 \\ -2 \\ 1 \\ -2 \\ 0 \end{bmatrix} + 0
\begin{bmatrix} 0 \\ -2 \\ 0 \\ 1 \\ 0 \end{bmatrix}
\qquad
\begin{bmatrix} 0 \\ -1 \\ 0 \\ 2 \\ 0 \end{bmatrix} \geq 0
\begin{bmatrix} 0 \\ -2 \\ 1 \\ -2 \\ 0 \end{bmatrix} + 0.5
\begin{bmatrix} 0 \\ -2 \\ 0 \\ 1 \\ 0 \end{bmatrix}
\qquad
\begin{bmatrix} 0 \\ 0 \\ 2 \\ -1 \\ 0 \end{bmatrix} \geq 0.5
\begin{bmatrix} 0 \\ -2 \\ 1 \\ -2 \\ 0 \end{bmatrix} + 0
\begin{bmatrix} 0 \\ -2 \\ 0 \\ 1 \\ 0 \end{bmatrix}
$$

Table 3: Verifying that condition (15) holds for the OT implication $(\text{CC}, \text{CVC}) \to (\text{CCC}, \text{CV.CVC})$.

The ranking $\gg$ succeeds on the antecedent mapping (x, y). In fact, the condition obtained in the last line of (16) says that there exists a constraint which prefers the winner (x, y) to the loser (x, z_i) whose weight is strictly larger than the weight of every constraint which instead prefers the loser (x, z_i) to the winner (x, y). By [A], this means that a constraint which prefers the winner (x, y) is $\gg$-ranked above every constraint that instead prefers the loser (x, z_i). The ranking $\gg$ therefore prefers the winner (x, y) to the loser (x, z_i). Since this conclusion holds for every $i = 1, \ldots, m$, the ranking $\gg$ succeeds on the antecedent mapping (x, y).

On the other hand, the ranking $\gg$ fails on the consequent mapping $(\widehat{x}, \widehat{y})$. In fact, the contradictory assumption that (17) fails means that $\max_{h \in W(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_h \leq \max_{k \in L(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_k$. In other words, there exists a constraint which prefers the loser $(\widehat{x}, \widehat{z}_j)$ to the winner $(\widehat{x}, \widehat{y})$ whose weight is strictly larger than or equal to the weights of the constraints which instead prefer the winner $(\widehat{x}, \widehat{y})$ to the loser $(\widehat{x}, \widehat{z}_j)$. By [A] and [B], the ranking $\gg$ cannot prefer $(\widehat{x}, \widehat{y})$ to $(\widehat{x}, \widehat{z}_j)$.

The conclusion that $\gg$ succeeds on the antecedent (x, y) but fails on the consequent $(\widehat{x}, \widehat{y})$ contradicts the assumption that the implication $(x, y) \overset{\text{OT}}{\to} (\widehat{x}, \widehat{y})$ holds in OT, thus establishing the inequality (17). This inequality can in turn be unpacked as in (18). In step (18a), we have lower bounded $\widehat{\Lambda}_j \max_{k \in L(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_k$ with the sum $\sum_{k \in L(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_k$, because $\widehat{\Lambda}_j$ is the number of addenda in the sum. In step (18b), we have upper bounded the maximum $\max_{h \in W(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_h$ with the sum $\sum_{h \in W(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_h$, because the weights being summed over are all non-negative. In step (18c), we have used the definition (14) of the constraint differences $C_k(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}}, \widehat{z}_j^{\text{easy}})$.

$$
\begin{aligned}
\max_{h \in W(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_h &> \max_{k \in L(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_k \implies \\
&\implies (\widehat{\Lambda}_j + 1) \max_{h \in W(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_h > \widehat{\Lambda}_j \max_{k \in L(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_k \\
&\overset{(a)}{\implies} (\widehat{\Lambda}_j + 1) \max_{h \in W(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_h > \sum_{k \in L(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_k \\
&\overset{(b)}{\implies} (\widehat{\Lambda}_j + 1) \sum_{h \in W(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_h > \sum_{k \in L(\widehat{x}, \widehat{y}, \widehat{z}_j)} w_k \\
&\overset{(c)}{\implies} \sum_{k=1}^{n} w_k C_k(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}}, \widehat{z}_j^{\text{easy}}) > 0
\end{aligned}
\tag{18}
$$

The inequality in the last line of (18) holds for every $j = 1, \ldots, \widehat{m}$, ensuring that the weights $\mathbf{w}$ succeed on the consequent mapping $(\widehat{x}^{\text{easy}}, \widehat{y}^{\text{easy}})$.

6 Conclusions

A central task of linguistic theory is to characterize the typological structure predicted by a grammatical formalism in order to match it to linguistic data. A classical strategy to characterize typological structure is to chart the implicational universals predicted by the formalism. In this paper, we have focused on the two constraint-based phonological formalisms of HG and OT. And we have considered the simplest type of implicational universals, namely *T-orders*. The main result of this paper has been a complete constraint characterization of T-orders in HG and OT. These constraint conditions rely on an elegant underlying convex geometry. These conditions are phonologically intuitive and have important algorithmic implications.

References

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

Maximilian Bane and Jason Riggle. 2009. Evaluating Strict Domination: The typological consequences of weighted constraints. In *Proceedings of the 45th annual meeting of the Chicago Linguistics Society*, pages 13–27.

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

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

Noam Chomsky. 1981. *Lectures on Government and Binding*. Mouton de Gruyter.

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.

Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, Cambridge, MA.

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

Frank Keller. 2006. Linear Optimality Theory as a model of gradience in grammar. In Gisbert Fanselow, Caroline Féry, Ralph Vogel, and Matthias Schlesewsky, editors, *Gradience in Grammar: Generative Perspectives*, pages 270–287. Oxford University Press, Oxford.

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. 2013. HG has no computational advantages over OT: towards a new toolkit for computational OT. *Linguistic Inquiry*, 44.4:569–609.

Giorgio Magri. 2018. Finiteness of optima in constraint-based phonology. Manuscript, CNRS.

Christopher Potts, Joe Pater, Karen Jesney, Rajesh Bhatt, and Michael Becker. 2010. Harmonic Grammar with Linear Programming: From linear systems to linguistic typology. *Phonology*, 27(1):1–41.

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.

R. Tyrrell Rockafellar. 1970. *Convex Analysis*. Princeton Landmarks in Mathematics; Princeton University Press.

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

Bruce Tesar and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry*, 29:229–268.

Appendix: Proof of Lemma 1

We assume that the HG implication $(\mathsf{x}, \mathsf{y}) \overset{\text{HG}}{\longrightarrow} (\widehat{\mathsf{x}}, \widehat{\mathsf{y}})$ holds. In order to establish the OT implication $(\mathsf{x}^*, \mathsf{y}^*) \overset{\text{OT}}{\longrightarrow} (\widehat{\mathsf{x}}^*, \widehat{\mathsf{y}}^*)$, we consider an arbitrary ranking $\gg$ which succeeds on the antecedent mapping $(\mathsf{x}^*, \mathsf{y}^*)$ and we establish that $\gg$ also succeeds on the consequent mapping $(\widehat{\mathsf{x}}^*, \widehat{\mathsf{y}}^*)$. Without loss of generality, we assume that $\gg$ ranks constraint C_n at the top, followed by constraint C_{n-1}, and so on, until the bottom ranked constraint C_1. We consider the corresponding weight vector $\mathbf{w} = (w_1, \ldots, w_n)$ defined in (19). It spaces the weights exponentially apart and assigns the largest weight w_n to the top ranked constraint C_n, the next largest weight w_{n-1} to the second top ranked constraint C_{n-1}, and so on.

$$w_n = \Delta^n, \;\; w_{n-1} = \Delta^{n-1}, \;\ldots\; w_1 = \Delta \quad (19)$$

To space the weights sufficiently apart, we chose the base Δ in (19) large enough, namely larger than the largest (in absolute value) antecedent or consequent violation difference, as in (20).

$$\Delta = 1 + \max \left\{ \begin{matrix} |C_k(\mathsf{x}, \mathsf{y}, \mathsf{z}_i)|, \\ |C_k(\widehat{\mathsf{x}}, \widehat{\mathsf{y}}, \widehat{\mathsf{z}}_j)| \end{matrix} \left| \begin{matrix} k = 1, \ldots, n \\ i = 1, \ldots, m \\ j = 1, \ldots, \widehat{m} \end{matrix} \right. \right\} \quad (20)$$

We now consider an arbitrary loser z_i with $i = 1, \ldots, m$. Since the ranking $\gg$ is by assumption OT consistent with the mapping $(\mathsf{x}^*, \mathsf{y}^*)$, $\gg$ prefers the winner $(\mathsf{x}^*, \mathsf{y}^*)$ to the loser $(\mathsf{x}^*, \mathsf{z}_i^*)$. This means that there exists a constraint C_k such that the constraints $C_n, \ldots, C_{k+1}$ ranked above C_k assign the same number of violations to the two mappings $(\mathsf{x}^*, \mathsf{y}^*)$ and $(\mathsf{x}^*, \mathsf{z}_i^*)$ while the constraint

C_k assigns less violations to the winner (x^*, y^*) than to the loser (x^*, z_i^*). By assumption (11), the same holds relative to the two original mappings (x, y) and (x, z_i): the constraints $C_n, \ldots, C_{k+1}$ ranked above C_k assign the same number of violations to (x, y) and (x, z_i) while the constraint C_k assigns less violations to the winner (x, y) than to the loser (x, z_i).

The chain of inequalities in (21) thus holds. In step (21a), we have dropped the terms corresponding to $h = k + 1, \ldots, n$ because the constraint violation difference $C_h(x, y, z_i)$ is equal to zero for every $h = k + 1, \ldots, n$. In step (21b), we have used the definition (19) of the weights w_h. In step (21c), we have used the fact that $C_k(x, y, z_i)$ is at least $+1$ (because C_k prefers the winner) and that $C_h(x, y, z_i)$ is at least $-(\Delta - 1)$ for every $h = 1, \ldots, k - 1$, because of the definition (20) of the constant Δ.

$$\sum_{h=1}^{n} w_h C_h(x, y, z_i) =$$
$$\overset{(a)}{=} \sum_{h=1}^{k-1} w_h C_h(x, y, z_i) + w_k C_k(x, y, z_i)$$
$$\overset{(b)}{=} \sum_{h=1}^{k-1} \Delta^h C_h(x, y, z_i) + \Delta^k C_k(x, y, z_i)$$
$$\overset{(c)}{\geq} -(\Delta - 1) \sum_{h=1}^{k-1} \Delta^h + \Delta^k$$
$$= \Delta > 1 \qquad (21)$$

Since the chain of inequalities (21) holds for an arbitrary $i = 1, \ldots, m$, we conclude that the HG grammar corresponding to the weight vector $\mathbf{w}$ succeeds on the antecedent mapping (x, y). The assumption $(x, y) \overset{\text{HG}}{\longrightarrow} (\widehat{x}, \widehat{y})$ thus ensures that $\mathbf{w}$ also succeeds on the consequent mapping $(\widehat{x}, \widehat{y})$. We now derive the conclusion that the ranking $\gg$ succeeds on the consequent mapping $(\widehat{x}^*, \widehat{y}^*)$.

In fact, suppose by contradiction that there exists $j = 1, \ldots, \widehat{m}$ such that $\gg$ prefers the loser $(\widehat{x}^*, \widehat{z}_j^*)$ to the winner $(\widehat{x}^*, \widehat{y})$. This means that there exists a constraint C_k such that the constraints $C_n, \ldots, C_{k+1}$ ranked above C_k assign the same number of violations to the two mappings $(\widehat{x}^*, \widehat{y}^*)$ and $(\widehat{x}^*, \widehat{z}_j^*)$ while the constraint C_k assigns less violations to the loser $(\widehat{x}^*, \widehat{z}_j^*)$ than to the winner $(\widehat{x}^*, \widehat{y}^*)$. By assumption (12), the same holds relative to the two original mappings $(\widehat{x}, \widehat{y})$ and $(\widehat{x}, \widehat{z}_j)$: the constraints $C_n, \ldots, C_{k+1}$ ranked above C_k assign the same number of violations to $(\widehat{x}, \widehat{y})$ and $(\widehat{x}, \widehat{z}_j)$ while the constraint C_k assigns less violations to the loser $(\widehat{x}, \widehat{z}_j)$ than to the winner $(\widehat{x}, \widehat{y})$.

The chain of inequalities in (22) thus holds. In steps (22a) and (22b), we have reasoned exactly as in steps (21a) and (21b) above. In step (22c), we have used the fact that $C_k(\widehat{x}, \widehat{y}, \widehat{z}_j)$ is at most -1 (because C_k prefers the loser) while $C_h(\widehat{x}, \widehat{y}, \widehat{z}_i)$ is at most $(\Delta - 1)$ for every $h = 1, \ldots, k - 1$, because of the definition (20) of the constant Δ.

$$\sum_{h=1}^{n} w_h C_h(\widehat{x}, \widehat{y}, \widehat{z}_j) =$$
$$\overset{(a)}{=} \sum_{h=1}^{k-1} w_h C_h(\widehat{x}, \widehat{y}, \widehat{z}_j) + w_k C_k(\widehat{x}, \widehat{y}, \widehat{z}_j)$$
$$\overset{(b)}{=} \sum_{h=1}^{k-1} \Delta^h C_h(\widehat{x}, \widehat{y}, \widehat{z}_j) + \Delta^k C_k(\widehat{x}, \widehat{y}, \widehat{z}_j)$$
$$\overset{(c)}{\leq} (\Delta - 1) \sum_{h=1}^{k-1} \Delta^h - \Delta^k$$
$$= -\Delta < -1 \qquad (22)$$

The chain of inequalities (22) yields the contradictory conclusion that the HG grammar corresponding to the weight vector $\mathbf{w}$ fails on the consequent mapping $(\widehat{x}, \widehat{y})$.