Giorgio Magri (IJN, ENS)
magrigrg@gmail.com

EALING 2010, ENS
Paris, September 9, 2010

The OT online model of the acquisition of phonotactics

Class 3: convergence again, from the perspective of the computational equivalence between OT and HG

Summary — This class introduces the current debate on OT versus Harmonic Grammar (HG) from a computational perspective; shows that HG is not computationally superior to OT, as any algorithms for HG can be readapted to OT; exploits this perspective to obtain a new proof of convergence of the OT online algorithm introduced in class 2; further exploits this perspective to discuss the worst-case number of updates and to obtain further variants of the OT online algorithm that are provably convergent.

1 Introduction

- **Optimality Theory.** OT uses *contraint ranking* and thus enforces *strict domination*. Thus, OT seems to have no close correspondent within core Machine Learning.[1] Hence, computational OT has been developed so far as in (1).

  (1) Computational problems that arise in modeling the acquisition of phonology within OT are tackled by means of *ad hoc* combinatorial algorithms, specifically tailored to OT, developed from scratch with no connections to methods and results in Machine Learning.

  Classes 1 and 2 exemplify the approach (1) to computational OT.

- **Harmonic Grammar.** In order to bridge this gap between Computational Phonology and Machine Learning, various scholars have started to entertain variants of OT that replace constraint ranking with *constraint weighting* and strict domination with *additive interaction*, and thus fall within the general class of *linear models* well studied in Machine Learning. An important such model is Harmonic Grammar (HG); see Legendre et al. (1990b,a). Claim (2) has thus become rather common.

  (2) HG is computationally superior to OT, as it comes with algorithms from Machine Learning (i.e. algorithms for linear classification), contrary to OT.

  See for instance Potts et al. (to appear), Pater (2009), Hayes and Wilson (2008), Coetzee and Pater (2008), Boersma and Pater (2007, 2008), Jesney and Tessier (2007, 2008), among others. Here is a quote from Pater (2009) that exemplifies (2):

  (3) "[I will] illustrate and extend existing arguments for the replacement of OT's ranked constraints with HG's weighted ones: [...] that it can make use of well understood algorithms for the modeling of learning and for other computational implementations. [...] The strengths of HG over OT in this area are of considerable importance".

- **HG has no computational advantages over OT.** In this class, I *prove* that (2) is false. In fact, I show that algorithms for HG can be extended to OT. Thus, HG has no computational advantages over OT. This result opens the way to the new approach to computational OT in (4), radically different from the approach (1) pursued so far.

  (4) Computational problems that arise in Computational OT are tackled by adapting well known algorithms from Machine Learning.

  In this class, I illustrate the benefits of this new approach (4) to computational OT, by showing that it leads to further developments in the theory of the OT online algorithm.

[1] Yet, a framework close to OT was popular in Operation Research in the Seventies; see for instance Fishburn (1974). As a metter of fact, Tesar and Smolensky's (1998) Recursive Constraint Demotion (see class 2) was later re-descovered within the Operation Research literature; see Dombi et al. (2007).

2 OT and the Ranking problem

- **Rankings.** An OT-grammar is parameterized by a *ranking*, which is a linear order $\gg$ over the constraint set $\mathcal{C}$, as illustrated in (5). A constraint C_h is $\gg$*-ranked above* another constraint C_k iff $C_h \gg C_k$.

  (5) top ranked
  $$C_1 \gg C_2 \gg \ldots \gg C_n$$

- **OT-compatibility.** Consider an underlying/winner/loser form data triplet:

  (6) winner
  $$(\; x, \quad \hat{y}, \quad y \;)$$
  loser

  A ranking $\gg$ is *OT-compatible* with $(x,\ \hat{y},\ y)$ iff condition (7a) holds, explained in (7b).

  (7) a. There exists a constraint C_k such that:
  i. $C_k(x, y)$ is strictly larger than $C_k(x, \hat{y})$
  ii. C_k is ranked above every constraint C_h such that $C_h(x, y) \neq C_h(x, \hat{y})$.
  b. The loser y violates the constraints "more severely" than the winner $\hat{y}$, in the sense that, among those constraints that distinguish between them, the top ranked one assigns more violations to the loser than to the winner.

  A ranking is OT-compatible with a set of triplets iff it is OT-compatible with every triplet in the set. A set of triplets is called OT-compatible iff it is compatible with a ranking.

■ **The Ranking problem.** The simplest computational problem in OT is (8). I denote by RP($\mathcal{D}$) the instance of (8) corresponding to a set of triplets $\mathcal{D}$, or the set of its solutions.

(8) *given:* a finite OT-compatible data set $\mathcal{D}$ of underlying/winner/loser form data triplets;
find: a ranking $\gg$ OT-compatible with the set of data triplets $\mathcal{D}$, according to condition (7).

■ **Comparative notation.** Given an underlying/winner/loser form data triplet $(x, \hat{y}, y)$, the notion of OT-compatibility (7) has the following property:

(9) OT-compatibility only depends on the sign of the constraint differences $C_k(x,y) - C(x,\hat{y})$, not on the actual numbers of constraint violations $C_k(x,y)$, $C(x,\hat{y})$.

Thus, the information contained in a triplet that is needed to compare the winner and the loser in OT can be distilled into an n-tuple $\mathbf{a}$ as in (10), called an *OT-comparative row*.

(10) $$\begin{matrix}\text{winner}\\ |\\ (x,\ \hat{y},\ y)\\ |\\ \text{loser}\end{matrix} \quad\Longrightarrow\quad \mathbf{a} = (a_1, \ldots, a_n) \quad a_k \stackrel{\text{def}}{=} \begin{cases} \text{W} & \text{if } C_k(x,y) - C_k(x,\hat{y}) > 0 \\ \text{L} & \text{if } C_k(x,y) - C_k(x,\hat{y}) < 0 \\ \text{E} & \text{if } C_k(x,y) - C_k(x,\hat{y}) = 0 \end{cases}$$

If we have many OT-comparative rows, we can stack them into an *OT-comparative tableau* $\mathbf{A}$. A ranking $\gg$ is *OT-compatible* with $\mathbf{A}$ iff it satisfies (11).

(11) Once the n columns of $\mathbf{A}$ are reordered from left to right in decreasing order according to $\gg$, then the leftmost non-E entry is a W in every row.

The Ranking problem restated in terms of OT-comparative tableaux is (12). I denote by RP($\mathbf{A}$) the instance of (12) corresponding to a tableau $\mathbf{A}$, or the set of its solutions.

(12) *given:* an OT-compatible OT-comparative tableau $\mathbf{A}$;
find: a ranking $\gg$ OT-compatible with the tableau $\mathbf{A}$, according to (11).

A ranking is a solution of the instance of the original Ranking problem (8) for a given set of data triplets $\mathcal{D}$ iff it is a solution of the instance of the problem (12) for the corresponding comparative tableau $\mathbf{A}(\mathcal{D})$, namely:

(13) RP($\mathcal{D}$) = RP($\mathbf{A}(\mathcal{D})$).

3 HG and the Weighting problem

■ **Weight vectors.** An HG grammar is parameterized by a *weight vector*, which is a tuple $\boldsymbol{\theta}$ with n numerical components $\theta_1, \ldots, \theta_n$ (one for every constraint), as in (14).

(14) $$\boldsymbol{\theta} = \begin{matrix} C_1 & \ldots & C_k & \ldots & C_n \end{matrix}\\ \begin{pmatrix} \theta_1, & \ldots & \theta_k, & \ldots & \theta_n \end{pmatrix}$$

The kth component θ_k is called the *weight* of the corresponding constraint C_k.

■ **HG-compatibility.** Consider an underlying/winner/loser form data triplet:

(15) $$\begin{matrix}\text{winner}\\ |\\ (\ x,\ \ \hat{y},\ \ y\ )\\ |\\ \text{loser}\end{matrix}$$

A weight vector $\boldsymbol{\theta}$ is *HG-compatible* with the triplet $(x,\ \hat{y},\ y)$ iff condition (16a) holds, explained in (16b).

(16) a. $$\sum_{k=1}^{n} \theta_k \cdot C_k(x,y) > \sum_{k=1}^{n} \theta_k \cdot C_k(x,\hat{y})$$
b. the loser y violates the constraints "more severely" than the winner $\hat{y}$, in the sense that the sum of the constraint violations of y weighted by $\boldsymbol{\theta}$ is (strictly) larger than the sum of the constraint violations of $\hat{y}$ weighted by $\boldsymbol{\theta}$.

A weight vector $\boldsymbol{\theta}$ is HG-compatible with a set of triplets iff it is HG-compatible with each. A set of data triplets is HG-compatible iff it is compatible with a weight vector.

■ **Non-negativity.** If the weights are allowed to be negative, unwanted typological consequences follows (e.g. unmarked forms can be mapped to marked ones). This problem is avoided if weights are required to be nonnegative, as in (17).

(17) $\theta_1, \ldots, \theta_n \geq 0$

Assumption (17) is not part of the core computational description of the model, and can therefore be relaxed, if we so wish.

■ **The Weighting problem.** The simplest computational problem in HG is (18). I denote by WP($\mathcal{D}$) the instance of (18) corresponding to a data set $\mathcal{D}$, or the set of its solutions.

(18) *given:* a finite HG-compatible data set $\mathcal{D}$ of underlying/winner/loser form data triplets;
find: a nonnegative weight vector $\boldsymbol{\theta}$ HG-compatible with the set of data triplets $\mathcal{D}$, according to condition (16).

I denote by $\text{WP}_{\text{unr}}(\mathcal{D})$ the corresponding problem without the non-negativity restriction.

■ **Comparative notation.** Given an underlying/winner/loser form data triplet $(x,\ \hat{y},\ y)$, the notion of HG-compatibility (16) has the following property:

(19) HG-compatibility only depends on the constraint differences $C_k(x,y) - C(x,\hat{y})$, not on the actual numbers of constraint violations $C_k(x,y)$, $C(x,\hat{y})$.

Thus, the information contained in a triplet that is needed to compare the winner and the loser in HG can be distilled into an n-tuple $\mathbf{a}$ as in (20), called an *HG-comparative row*.

(20) $$\underset{(x,\ \hat{y},\ y)}{\overset{\text{winner}}{\overset{|}{\ }}\overset{}{\overset{|}{\underset{\text{loser}}{\ }}}} \implies \overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n) \qquad \overline{a}_k \stackrel{\text{def}}{=} C_k(x,y) - C_k(x,\hat{y})$$

An example is provided in (21).

(21) $$\underset{(/\text{rad}/,\ [\text{rad}],\ [\text{rat}])}{\overset{\text{winner}}{\overset{|}{\ }}\overset{}{\overset{|}{\underset{\text{loser}}{\ }}}} \implies \begin{bmatrix} C_1 & C_2 & C_3 \\ 0 & 1 & -1 \end{bmatrix}$$

C_1 = IDENT[VOICE]/ONSET
C_2 = IDENT[VOICE]
C_3 = *[VOICE]

A weight vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is *HG-compatible* with a triplet according to (16) iff it satisfies condition (22) w.r.t. the corresponding HG-comparative row $\overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n)$.

(22) $$\sum_{k=1}^{n} \theta_k \overline{a}_k > 0.$$

If we have many HG-comparative rows, then we can stack them one above the other into an *HG-comparative tableau* $\overline{\mathbf{A}}$. The Weighting problem restated in terms of HG-comparative tableaux is (23). I denote by WP($\overline{\mathbf{A}}$) the instance of (23) corresponding to a tableau $\overline{\mathbf{A}}$, or the set of its solutions.

(23) *given:* an HG-compatible HG-comparative tableau $\overline{\mathbf{A}}$;
find: a nonnegative weight vector $\boldsymbol{\theta}$ HG-compatible with the tableau $\overline{\mathbf{A}}$, according to condition (22).

A weight vector is a solution of the instance of the original Weight problem (18) for a given data set $\mathcal{D}$ iff it is a solution of the instance of the problem (23) for the corresponding HG-comparative tableau $\overline{\mathbf{A}}(\mathcal{D})$, namely:

(24) WP($\mathcal{D}$) = WP($\overline{\mathbf{A}}(\mathcal{D})$).

Let WP$_{\text{unr}}$($\overline{\mathbf{A}}$) be the corresponding problem (25), with no non-negativity restriction.

(25) *given:* an HG-compatible HG-comparative tableau $\overline{\mathbf{A}}$;
find: a weight vector $\boldsymbol{\theta}$ (with no restriction on the sign of the weights) HG-compatible with the tableau $\overline{\mathbf{A}}$.

4 What is known about the relationship between HG and OT

■ **Claim 1** Prince and Smolensky (2004) and Keller (2000, 2005) show that:

(26) If a set $\mathcal{D}$ of underlying/winner/loser form triplets is OT-compatible, then it is also HG-compatible.

In fact, let $\gg$ be a ranking OT-compatible with $\mathcal{D}$; wlg, assume that it is (27a); then the weight vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ in (27b) is HG-compatible with $\mathcal{D}$, where δ is the largest constraint difference (ignoring sign) over all constraints and all triplets in $\mathcal{D}$.

(27) a. $C_n \gg C_{n-1} \gg \ldots \gg C_2 \gg C_1$
b. $\theta_n = (\delta+1)^n,\ \theta_{n-1} = (\delta+1)^{n-1}\ \ldots\ \theta_2 = (\delta+1)^2,\ \theta_1 = (\delta+1)$

The idea is that the "highest-takes-all" behavior of OT can be mimicked in HG with exponentially spaced weights.

■ **Comparative restatement.** OT-comparative rows are defined in (10) in terms of the sign of constraint differences. Constraint differences are encoded into the entries of HG-comparative rows. Thus, (10) really says how to construct OT-comparative rows out of HG-comparative rows, as repeated in (28).

(28) $$\overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n) \implies \mathbf{a} = (a_1, \ldots, a_n) \qquad a_k \stackrel{\text{def}}{=} \begin{cases} \text{W} & \text{if } \overline{a}_k > 0 \\ \text{L} & \text{if } \overline{a}_k < 0 \\ \text{E} & \text{if } \overline{a}_k = 0 \end{cases}$$

Consider an HG-comparative tableau $\overline{\mathbf{A}}$ of constraint differences and the corresponding OT-comparative tableau $\mathbf{A}$ defined row by row as in (28). Claim (26) becomes:

(29) If the OT-comparative tableau $\mathbf{A}$ is OT-compatible, then the HG-comparative tableau $\overline{\mathbf{A}}$ we started from is HG-compatible.

In fact, let $\gg$ be a ranking OT-compatible with the OT-comparative tableau $\mathbf{A}$; wlg, assume it is (27a); then the weight vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ in (27b) is HG-compatible with the HG-comparative tableau $\overline{\mathbf{A}}$, where δ is the largest entry in $\overline{\mathbf{A}}$ (ignoring sign).

■ **Example.** Given the HG-comparative tableau $\overline{\mathbf{A}}$ in (30), the corresponding OT-comparative tableau according to (28) is $\mathbf{A}$ in (30).

(30) $$\overline{\mathbf{A}} = \begin{matrix} \end{matrix}\begin{matrix} C_1 & C_2 & C_3 \end{matrix}\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \implies \mathbf{A} = \begin{matrix} C_1 & C_2 & C_3 \end{matrix}\begin{bmatrix} \text{W} & \text{L} & \text{W} \\ \text{E} & \text{W} & \text{L} \end{bmatrix}$$

The OT-comparative tableau $\mathbf{A}$ is OT-compatible with the ranking in (31a). Since in this case $\delta = 1$, the corresponding weight vector according to (27b) is (31b).

(31) a. $C_1 \gg C_2 \gg C_3$
b. $\boldsymbol{\theta} = (8, 4, 2)$

The vector $\boldsymbol{\theta}$ in (31b) is indeed HG-compatible with the HG-comparative tableau $\overline{\mathbf{A}}$ (30).

■ **Reverse.** The reverse of claim 1 is false, i.e. there exist data sets $\mathcal{D}$ that are HG-compatible but not OT-compatible. Here is a counterexample in comparative notation:

(32) $$\overline{\mathbf{A}} = \begin{matrix} \end{matrix}\begin{matrix} C_1 & C_2 & C_3 \end{matrix}\begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \qquad \mathbf{A} = \begin{matrix} C_1 & C_2 & C_3 \end{matrix}\begin{bmatrix} \text{L} & \text{W} & \text{W} \\ \text{W} & \text{L} & \text{E} \\ \text{W} & \text{E} & \text{L} \end{bmatrix}$$

The HG-tableau $\overline{\mathbf{A}}$ is HG-compatible, say with $\boldsymbol{\theta} = (2, 1, 1)$; but the corresponding OT-tableau $\mathbf{A}$ is not OT-compatible (it has a diagonal of L's).

■ *Proof.* Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ be the weight vector in (27b). Consider a row $\overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n)$ of the given HG-comparative tableau. Let $\mathbf{a} = (a_1, \ldots, a_n)$ be the corresponding OT-comparative row by (28). The ranking (27a) thus satisfies the OT-compatibility condition (11) for $\mathbf{a}$. Thus $a_n = a_{n-1} = \ldots = a_{k+1} = \text{E}$ and $a_k = \text{W}$ for some $k \in \{n, n-1, \ldots, 1\}$. To simplify notation, let $B = \delta + 1$. I can then compute as follows:

(33)
$$\sum_{i=1}^{n} \theta_i \overline{a}_i \overset{(a)}{=} \sum_{i=1}^{k-1} \theta_i \overline{a}_i + \theta_k \overline{a}_k + \sum_{i=k+1}^{n} \theta_i \overline{a}_i$$
by splitting up the set $\{1, \ldots, n\}$ that the index i runs over into the three subsets $\{1, \ldots, k-1\}$, $\{k\}$ and $\{k+1, \ldots, n\}$
$$\overset{(b)}{=} \sum_{i=1}^{k-1} \theta_i \overline{a}_i + \theta_k \overline{a}_k$$
because $a_n = a_{n-1} = \ldots = a_{k+1} = \text{E}$, which means in turn that $\overline{a}_n = \overline{a}_{n-1} = \ldots = \overline{a}_{k+1} = 0$
$$\overset{(c)}{\geq} \sum_{i=1}^{k-1} \theta_i \overline{a}_i + \theta_k$$
lower-bounding by replacing $\overline{a}_k$ with the smallest value it can take, which is 1 (in fact, since $a_k = \text{W}$, then $\overline{a}_k > 0$ and thus $\overline{a}_k \geq 1$, since constraint differences are integers)
$$\overset{(d)}{\geq} -\sum_{i=1}^{k-1} \theta_i (B-1) + \theta_k$$
lower-bounding by replacing $\overline{a}_i$ with the smallest value it can take, namely $-\delta = -(B-1)$
$$\overset{(e)}{=} -\sum_{i=1}^{k-1} B^i (B-1) + B^k$$
by definition (27b) of the weight vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, restated in terms of $B = \delta + 1$
$$= -\sum_{i=1}^{k-1} B^{i+1} + \sum_{i=1}^{k-1} B^i + B^k$$
$$= -\sum_{i=2}^{k} B^i + \sum_{i=1}^{k-1} B^i + B^k$$
$$= -B^k + B + B^k > 0$$

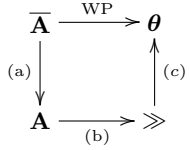
The chain of inequalities in (33) shows that the weight vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ defined in (27) does indeed satisfy the HG-compatibility condition $\sum_{k=1}^{n} \theta_k \overline{a}_k > 0$ in (22). □

5 But we want the opposite algorithmic perspective

■ **The algorithmic perspective of claim 1.** Roughly, claim 1 says that:

(34) An instance of the WP (23) can be paired up with an instance of the RP (12) such that a solution to the former is obtained by solving the latter instead.

Let me say this more precisely. Consider a Weighting problem WP($\overline{\mathbf{A}}$). Suppose the OT-tableau $\mathbf{A}$ corresponding to $\overline{\mathbf{A}}$ by (28) happens to be OT-compatible. Claim 1 says that, instead of solving WP($\overline{\mathbf{A}}$) *directly*, we can solve it *indirectly*, through (35a)-(35c).

(35)


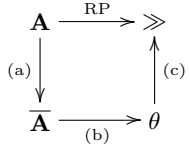
a. construct the OT-comparative tableau $\mathbf{A}$ corresponding to the HG-comparative tableau $\overline{\mathbf{A}}$ by (28);
b. solve the corresponding instance RP($\mathbf{A}$) of the Ranking problem (12);
c. obtain a solution of the given Weighting problem WP($\overline{\mathbf{A}}$) through (27b).

But this algorithmic perspective is *useless*: we already know how to solve the WP, since we can draw on the very large literature on linear models; see e.g. Potts et al. (to appear); what we need is instead good methods to solve the RP.

■ **The algorithmic perspective we want.** Thus, what we want is the reverse of (34):

(36) An instance of the RP (12) can be paired up with an instance of the WP (23) such that a solution to the former is obtained by solving the latter instead.

Let me say this more precisely. Given a Ranking problem RP($\mathbf{A}$), we would like to find one (or, even better, all) of its solutions without solving the problem *directly* but rather *indirectly* through the three steps (37a)-(37c).

(37)


a. pair up the given OT-comparative tableau $\mathbf{A}$ with some properly derived HG-comparative tableau $\overline{\mathbf{A}}$;
b. find a solution $\boldsymbol{\theta}$ of the corresponding Weighting problem WP($\overline{\mathbf{A}}$);
c. pair up that solution $\boldsymbol{\theta}$ with a ranking $\gg$ that refines $\boldsymbol{\theta}$.

Diagram (37) is the reverse of diagram (35) corresponding to claim 1. But the reverse of claim 1 does not hold, as seen above. Thus, how the diagram (35) be inverted into (37)? Let me illustrate the idea with a couple of examples, before I turn to the details.

■ **An example.** Consider the OT-comparative row $\mathbf{a}$ in (38). Suppose that the corresponding HG-comparative row is $\overline{\mathbf{a}}$.

(38)
$$\mathbf{a} = \begin{bmatrix} \overset{C_1}{\text{E}} & \overset{C_2}{\text{W}} & \overset{C_3}{\text{L}} \end{bmatrix} \quad \Longrightarrow \quad \overline{\mathbf{a}} = \begin{bmatrix} \overset{C_1}{0} & \overset{C_2}{1} & \overset{C_3}{-1} \end{bmatrix}$$

I can reason as follows:

(39) a. $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ is HG-compatible with $\overline{\mathbf{a}}$ $\iff$
$\iff$ $\theta_2 - \theta_3$ is strictly positive
$\iff$ the weight θ_2 is strictly larger than the weight θ_3
$\iff$ every ranking that refines $\boldsymbol{\theta}$ ranks C_2 above C_3
$\iff$ every ranking that refines $\boldsymbol{\theta}$ is OT-compatible with $\mathbf{a}$

The OT-comparative row $\mathbf{a}$ in (38) has a unique W, and its L has been mapped to -1 in the corresponding derived HG-comparative row $\overline{\mathbf{a}}$.

■ **Another example.** Consider the OT-comparative row $\mathbf{a}$ in (40). Suppose the corresponding HG-comparative row is $\overline{\mathbf{a}}$. Note that here the L is replaced with -2.

$$(40)\quad \mathbf{a} = \begin{array}{c} \begin{matrix} C_1 & C_2 & C_3 \end{matrix} \\ \begin{bmatrix} \text{W} & \text{W} & \text{L} \end{bmatrix} \end{array} \quad \Longrightarrow \quad \overline{\mathbf{a}} = \begin{array}{c} \begin{matrix} C_1 & C_2 & C_3 \end{matrix} \\ \begin{bmatrix} 1 & 1 & -2 \end{bmatrix} \end{array}$$

I can reason as follows:

(41) $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ is HG-compatible with $\overline{\mathbf{a}} \iff$
- $\iff$ $\theta_1 + \theta_2 - 2\theta_3$ is strictly positive
- $\iff$ $(\theta_1 - \theta_3) + (\theta_2 - \theta_3)$ is strictly positive.
- $\iff$ either $(\theta_1 - \theta_3)$ or $(\theta_2 - \theta_3)$ is strictly positive (or both)
- $\iff$ every ranking that refines $\boldsymbol{\theta}$ ranks C_1 above C_3 or C_2 above C_3 (or both)
- $\iff$ every ranking that refines $\boldsymbol{\theta}$ is OT-compatible with $\mathbf{a}$.

The OT-comparative row $\mathbf{a}$ in (40) has two W's, and its L has been mapped to -2 in the derived HG-comparative row $\overline{\mathbf{a}}$. This is the idea: map the L's to the total number of W's.

---

6 A new observation on the relationship between HG and OT

■ **Derived HG-comparative tableaux.** Given an OT-comparative row $\mathbf{a} = (a_1, \ldots, a_n)$, let me say that an HG-comparative row $\overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n)$ is *derived* from $\mathbf{a}$ iff it satisfies (42): W, E and L's in $\mathbf{a}$ are replaced with positive, null non-positive entries in $\overline{\mathbf{a}}$.

$$(42)\quad \mathbf{a} = (a_1, \ldots, a_n) \longrightarrow \overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n) \text{ such that } \overline{a}_k \begin{cases} > 0 & \text{if } a_k = \text{W} \\ = 0 & \text{if } a_k = \text{E} \\ \leq 0 & \text{if } a_k = \text{L} \end{cases}$$

Given an OT-comparative tableau $\mathbf{A}$, an HG-comparative tableau $\overline{\mathbf{A}}$ with the same number of rows and columns is *derived* from $\mathbf{A}$ iff each row of $\overline{\mathbf{A}}$ is derived from the corresponding row of $\mathbf{A}$ according to (42). In this case, I also write:

$$(43)\quad \overline{\mathbf{A}} = \mathit{derived}(\mathbf{A})$$

■ **A special case of derived tableaux.** Given an OT-comparative row $\mathbf{a}$, consider the derived HG-comparative row $\overline{\mathbf{a}}$ in (44), where $w(\mathbf{a})$ is the number of W's in $\mathbf{a}$.

$$(44)\quad \mathbf{a} = (a_1, \ldots, a_n) \longrightarrow \overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n) \text{ s.t. } \overline{a}_k \doteq \begin{cases} 1 & \text{if } a_k = \text{W} \\ 0 & \text{if } a_k = \text{E} \\ -w(\mathbf{a}) & \text{if } a_k = \text{L} \end{cases}$$

The definition (44) is illustrated in (45).

$$(45)\quad \mathbf{A} = \begin{array}{c} \begin{matrix} C_1 & C_2 & C_3 & C_4 \end{matrix} \\ \begin{bmatrix} \text{W} & \text{E} & \text{W} & \textcircled{\text{L}} \\ \text{E} & \text{W} & \textcircled{\text{L}} & \text{E} \end{bmatrix} \end{array} \longrightarrow \overline{\mathbf{A}} = \begin{array}{c} \begin{matrix} C_1 & C_2 & C_3 & C_4 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 1 & \textcircled{-2} \\ 0 & 1 & \textcircled{-1} & 0 \end{bmatrix} \end{array}$$

■ **Derived rankings.** A ranking $\gg$ is *derived* from a weight vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ iff it satisfies all the rankings implicit in the relative size of the weights, in the sense that condition (46) holds for every pair of constraints C_h and C_k.

$$(46)\quad \theta_h > \theta_k \longrightarrow C_h \gg C_k$$

Here are two examples:

$$(47)\quad \text{a. } \boldsymbol{\theta} = \begin{array}{c} \begin{matrix} C_1 & C_2 & C_3 \end{matrix} \\ \begin{pmatrix} 100 & 10 & 50 \end{pmatrix} \end{array} \longrightarrow C_1 \gg C_3 \gg C_2$$

$$\text{b. } \boldsymbol{\theta} = \begin{array}{c} \begin{matrix} C_1 & C_2 & C_3 \end{matrix} \\ \begin{pmatrix} 100 & 50 & 50 \end{pmatrix} \end{array} \begin{array}{l} \nearrow \; C_1 \gg C_3 \gg C_2 \\ \searrow \; C_1 \gg C_2 \gg C_3 \end{array}$$

This is of course the same notion we already encountered in class 1. Given a set W of weight vectors, I will also write:

$$(48)\quad \mathit{derived}(W) = \text{the set of all rankings derived from weight vectors in } W$$

■ **Claim 2** Given an instance RP($\mathbf{A}$) of the Ranking problem (12) corresponding to an OT-comparative tableau $\mathbf{A}$, consider the instance WP($\overline{\mathbf{A}}$) of the Weighting problem (23) corresponding to the HG-comparative tableau $\overline{\mathbf{A}} = \mathit{derived}(\mathbf{A})$ derived from $\mathbf{A}$ according to (44). Then, the following inclusion holds:

$$(49)\quad \mathit{derived}(\text{WP}(\mathit{derived}(\mathbf{A}))) \subseteq \text{RP}(\mathbf{A})$$

namely, every ranking derived from any solution of WP($\overline{\mathbf{A}}$) is a solution of RP($\mathbf{A}$).

■ *Proof.* Let $\boldsymbol{\theta}$ be a solution of WP($\overline{\mathbf{A}}$). Consider an arbitrary row $\mathbf{a}$ of the OT-comparative tableau $\mathbf{A}$; let $\overline{\mathbf{a}}$ be the corresponding HG-comparative row, as defined in (44). Let:

$$(50)\quad \begin{array}{rcl} W(\mathbf{a}) & = & \text{set of constraints that have a W in the OT-comparative row } \mathbf{a} \\ L(\mathbf{a}) & = & \text{set of constraints that have an L in the OT-comparative row } \mathbf{a} \end{array}$$

The chain of inequalities (51) holds for every loser-preferring constraint $k \in L(\mathbf{a})$:

(51) $0 \overset{(a)}{<} \sum_{h=1}^{n} \theta_h \overline{a}_h$

by the hypothesis that $\boldsymbol{\theta}$ is a solution of WP($\overline{\mathbf{A}}$) and thus satisfies condition (22)

$\overset{(b)}{=} \sum_{h \in W(\mathbf{a})} \theta_h \overline{a}_h + \sum_{h \in L(\mathbf{a})} \theta_h \overline{a}_h + \sum_{h \notin W(\mathbf{a}) \cup L(\mathbf{a})} \theta_h \overline{a}_h$

by splitting the set $\{1, \ldots, n\}$ that h runs over into the sets $W(\mathbf{a})$, $L(\mathbf{a})$ and their complement

$\overset{(c)}{=} \sum_{h \in W(\mathbf{a})} \theta_h - w(\mathbf{a}) \sum_{h \in L(\mathbf{a})} \theta_h$

because $\overline{a}_h = 1$ for every $h \in W(\mathbf{a})$, $\overline{a}_h = -w(\mathbf{a})$ for every $h \in L(\mathbf{a})$ and $\overline{a}_h = 0$ for every $h \notin W(\mathbf{a}) \cup L(\mathbf{a})$, by definition (44)

$\overset{(d)}{\leq} w(\mathbf{a}) \max_{h \in W(\mathbf{a})} \theta_h - w(\mathbf{a}) \sum_{h \in L(\mathbf{a})} \theta_h$

by upper bounding the sum $\sum_{h \in W(\mathbf{a})} \theta_h$ with its biggest term $\max_{h \in W(\mathbf{a})} \theta_h$ multiplied by the total number $w(\mathbf{a})$ of terms

$\overset{(e)}{\leq} w(\mathbf{a}) \max_{h \in W(\mathbf{a})} \theta_h - w(\mathbf{a}) \theta_k$

because all the components of $\boldsymbol{\theta}$ are nonnegative and thus $\sum_{h \in L(\mathbf{a})} \theta_h \geq \theta_k$ provided that $k \in L(\mathbf{a})$

By reordering the inequality obtained in (51), I have:

(52) $\max_{h \in W(\mathbf{a})} \theta_h > \theta_k$

Since (52) holds for every $k \in L(\mathbf{a})$, then in particular I have (53).

(53) $\max_{h \in W(\mathbf{a})} \theta_h > \max_{k \in L(\mathbf{a})} \theta_k$

By (53), the largest weight among winner-preferring constraints is strictly larger than the largest weight over loser-preferring constraints. Thus, any refinement of $\boldsymbol{\theta}$ ranks a winner-preferring constraint above all loser-preferring constraints. Hence, any refinement of $\boldsymbol{\theta}$ is OT-compatible with $\mathbf{a}$. □

7 Corollaries

■ **Claim 3** A ranking solves an instance RP($\mathbf{A}$) of the Ranking problem (12) corresponding to an OT-comparative tableau $\mathbf{A}$, iff it is derived from a weight vector that solves the Weighting problem WP($\overline{\mathbf{A}}$) corresponding to the HG-comparative tableau $\overline{\mathbf{A}} = \mathit{derived}(\mathbf{A})$ derived from $\mathbf{A}$ according to (44). In a fancy notation:

(54) *derived*(WP(*derived*($\mathbf{A}$))) = RP($\mathbf{A}$)

■ *Proof.* Follows straightforwardly from claims 1 and 2:

(55) a. Inclusion "$\subseteq$":
is guaranteed by (49) in the new claim 2.

b. Inclusion "$\supseteq$":
follows from claim 1, as ranking (27a) is derived from weight vector (27b).

□

■ **Remark.** Claim 3 says that OT does not raise any new computational challenges beyond HG. Hence, the popular claim (2) that I quoted at the beginning of this class is wrong.

■ **Claim 4** A comparative tableau $\mathbf{A}$ is OT-compatible iff *every* HG-comparative tableau $\overline{\mathbf{A}}$ derived from $\mathbf{A}$ according to the general scheme (42) is HG-compatible.

■ *Proof.* Follows straightforwardly from claims 1 and 2:

(56) a. OT-compatibility ⇒ HG-compatibility:
follows from claim 1

b. HG-compatibility ⇒ OT-compatibility:

i. assume every HG-comparative tableau $\overline{\mathbf{A}}$ derived from the OT-compatible tableau $\mathbf{A}$ according to (42) is HG-compatible;

ii. then in particular the HG-comparative tableau $\overline{\mathbf{A}}$ derived from $\mathbf{A}$ according to (44) is HG-compatible;

iii. then, any ranking derived from any weight vector HG-compatible with $\overline{\mathbf{A}}$ is OT-compatible with $\mathbf{A}$, by claim 2;

iv. hence, $\mathbf{A}$ is OT-compatible.

□

■ **Remark.** Claim 4 fulfills the promise made in class 1 of contributing to the search for new characterizations of the notion of OT-compatibility.

8 A digression on the non-negativity restriction

■ **The role of the nonnegativity requirement.** Claim 2 fails if we drop the nonnegativity restriction (17), namely if we switch from the Weighting problem (23) to the unrestricted variant (25). Here is a counterexample:

(57) a. OT-comparative row:

$$\mathbf{a} = \begin{array}{c} \begin{matrix} C_1 & C_2 & C_3 \end{matrix} \\ \begin{bmatrix} \text{W} & \text{L} & \text{L} \end{bmatrix} \end{array}$$

b. HG-comparative row derived from $\mathbf{a}$ according to (44):

$$\overline{\mathbf{a}} = \begin{array}{c} \begin{matrix} C_1 & C_2 & C_3 \end{matrix} \\ \begin{bmatrix} 1 & -1 & -1 \end{bmatrix} \end{array}$$

c. weight vector HG-compatible with $\overline{\mathbf{a}}$, that doesn't satisfy the nonnegativity restriction (17):

$$\boldsymbol{\theta} = \begin{array}{c} \begin{matrix} C_1 & C_2 & C_3 \end{matrix} \\ \begin{bmatrix} -4 & -3 & -3 \end{bmatrix} \end{array}$$

d. ranking derived from $\boldsymbol{\theta}$ that is not OT-compatible with $\mathbf{a}$:
$C_2 \gg C_3 \gg C_1$

Yet, claim 2 holds also without the nonnegativity restriction (17), provided the given OT-comparative tableau is slightly pre-processed. Here come the details.

■ **Tableaux with a unique** L **per row.** Two OT-comparative tableaux $\mathbf{A}$ and $\mathbf{A}'$ (with the same number of columns but a possibly different number of rows) are *OT-equivalent* iff are OT-compatible with exactly the same rankings, as in (58).

(58) $\text{RP}(\mathbf{A}) = \text{RP}(\mathbf{A}')$

A row with multiple L's is OT-equivalent to multiple rows with a unique L, as in (81).

(59) $\mathbf{A} = \begin{bmatrix} \text{E} & \text{W} & \text{L} & \text{L} \end{bmatrix} \implies \mathbf{A}' = \begin{bmatrix} \text{E} & \text{W} & \text{L} & \text{E} \\ \text{E} & \text{W} & \text{E} & \text{L} \end{bmatrix}$

Given an OT-comparative tableau $\mathbf{A}$, I can construct an equivalent tableau $\mathbf{A}'$ with a unique L per row. Without loss of generality, I can thus assume that a given OT-comparative tableau has only one L per row.

■ **Claim 5** A ranking solves an instance $\text{RP}(\mathbf{A})$ of the Ranking problem (12) corresponding to an OT-comparative tableau $\mathbf{A}$ *that has a unique* L *per row* iff it is derived from a weight vector that solves the *unrestricted* Weighting problem $\text{WP}_{\text{unr}}(\overline{\mathbf{A}})$ corresponding to the HG-comparative tableau $\overline{\mathbf{A}} = \mathit{derived}(\mathbf{A})$ derived from $\mathbf{A}$ according to (44). In a fancy notation:

(60) $\mathit{derived}(\text{WP}_{\text{unr}}(\mathit{derived}(\mathbf{A}))) = \text{RP}(\mathbf{A})$

■ *Proof.* The nonnegativity restriction (17) was used only once in the proof of claim 2, namely to obtain the following inequality in step (51e):

(61) $\sum_{h \in L(\mathbf{a})} \theta_h \leq \theta_k \qquad \text{for some } k \in L(\mathbf{a})$

If we drop the nonnegativity restriction (17), the inequality (61) does not hold any more in the general case. Yet, it trivially holds if $\mathbf{a}$ has a unique L corresponding to some constraint C_k, as in this case $L(\mathbf{a}) = \{k\}$ and thus $\sum_{h \in L(\mathbf{a})} \theta_h = \theta_k$. □

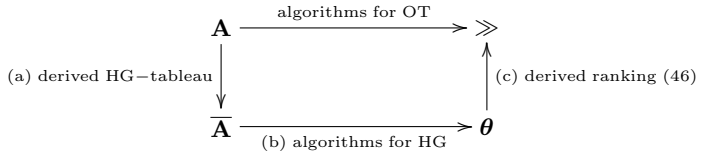
9 A new approach to computational OT

■ **The general idea.** Classes 1 and 2 have illustrated the classical approach (1) to computational OT, repeated in (62).

(62) Computational problems that arise in modeling the acquisition of phonology within OT are tackled by means of *ad hoc* combinatorial algorithms, specifically tailored to OT, developed from scratch with no connections to methods and results in Machine Learning.

This approach to computational OT corresponds to the top horizontal arrow in (63).

(63)



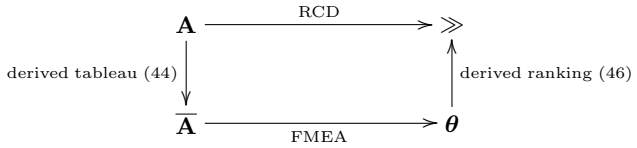
The elementary result presented in sections 6-8 shows that Machine Learning algorithms for HG can be "translated" into OT, as in (63a)-(63c). This is the alternative algorithmic strategy anticipated in (4), repeated in (64).

(64) Computational problems that arise in Computational OT are tackled by adapting well known HG algorithms from Machine Learning.

In the rest of this class, I will illustrate this alternative algorithmic strategy (64).

■ **Application to batch algorithms.** In class 2, we saw T&S's RCD algorithm. It turns out that RCD for the RP "corresponds" to the *Fourier-Motzkin Elimination Algorithm* (FMEA) for the WP, as in (65); see Bertsimas and Tsitsiklis (1997, pp. 70-74).

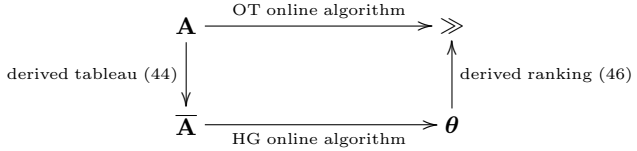
(65)



Scheme (65) is a special case of scheme (63). I will not present this application here.

■ **Applications to online algorithms.** In classes 1 and 2, we developed a theory of the OT online algorithm. It turns out that OT online algorithms "correspond" to HG online algorithms, as in (66); see Cesa-Bianchi and Lugosi (2006).

(66)



This perspective can be exploited to further develop the theory of OT online algorithms along various directions:

(67) a. *Convergence*: a new proof of convergence of the cautious promotion/demotion update rule of class 2 can be obtained, using the convergence theorem of the Perceptron Algorithm.

b. *Worst-case number of updates:* a (unfortunately, rather loose) bound on the worst case number of updates of OT online algorithms can be obtained, using such bounds for HG online algorithms.

c. *New update rules:* convergent OT update rules can be obtained, that update multiplicatively rather than additively, using the theory of Bregman online algorithms.

And perhaps also:

(68) a. *Margin*: develop a notion of online complexity measures of a given comparative tableau, using the recent notion of margin in the theory of classifiers.

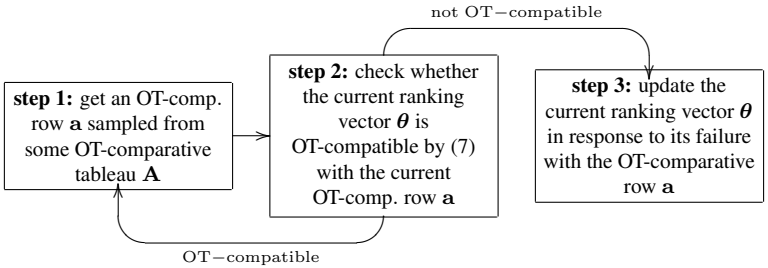
b. *child variation as an algorithmic strategy:* try to connect variation in child phonology with a large body of results that say that randomization helps a big deal: variation in child phonology is not a consequence of a transient imperfection of the developing phonology rather a smart learning strategy.

In this class, I only illustrate (67a).

10 Derived OT update rules

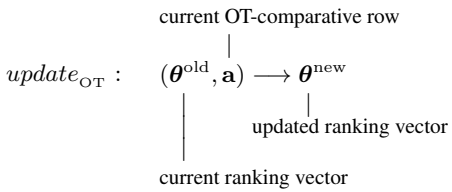
■ **OT online algorithms.** As seen in class 1, OT online algorithms have the shape (69).

(69)



An update rule for an OT online algorithm is a function (70): it takes a current ranking vector $\boldsymbol{\theta}^{\text{old}}$ and an OT-comparative row $\mathbf{a}$ and it returns an updated ranking vector $\boldsymbol{\theta}^{\text{new}}$.

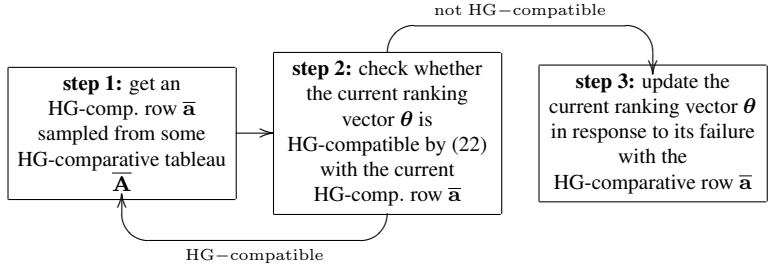
(70)



The OT online algorithm (69) *converges* with a given OT update rule (70) iff it can only make a finite number of updates for any OT-compatible OT-comparative input tableau.

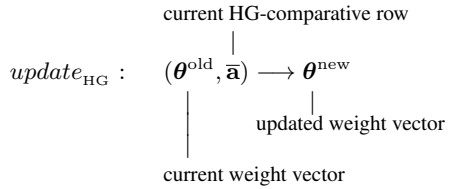
■ **HG online algorithms.** HG online algorithms have the general shape (71).

(71)



An update rule for an HG online algorithm is a function (72): it takes a weight vector $\boldsymbol{\theta}^{\text{old}}$ and an HG-comparative row $\overline{\mathbf{a}}$ and it returns an updated weight vector $\boldsymbol{\theta}^{\text{new}}$.

(72)



The HG online algorithm (71) *converges* with a given HG update rule (72) iff it can only make a finite number of updates for any HG-compatible HG-comparative input tableau.

■ **Derived OT update rules.** Since weight and ranking vectors are the same (apart from the optional non-negativity of weights), OT and HG update rules only differ because:

(73) a. OT update rules take *OT*-comparative rows $\mathbf{a} = (a_1, \ldots, a_n)$;

b. HG update rules take *HG*-comparative rows $\overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n)$.

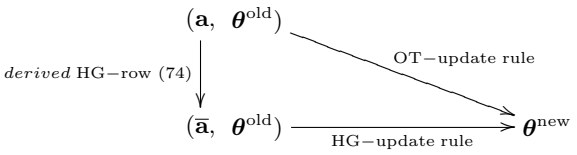
Recall we have introduced in section 6 the notion (42) of a mapping from OT-comparative rows into *derived* HG-comparative rows, repeated in (74).

(74) $\mathbf{a} = (a_1, \ldots, a_n)$

$\overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n)$ such that $\overline{a}_k \begin{cases} > 0 & \text{if } a_k = \text{W} \\ = 0 & \text{if } a_k = \text{E} \\ \leq 0 & \text{if } a_k = \text{L} \end{cases}$

Using a mapping (74) from OT-comparative rows $\mathbf{a}$ into derived HG-comparative rows $\overline{\mathbf{a}}$, we can thus translate HG update rules into OT update rules as in (75).

(75)



Scheme (75) is a special instance of the general algorithmic scheme (63). An OT update rule defined through (75) is called *derived* from the corresponding HG update rule.

■ **Preservation of convergence.** We are interested in instances of the scheme (75) that *preserve convergence*, in the sense that:

(76) If the HG online algorithm converges with a given HG update rule, then the OT online algorithm converges with the OT update rule derived through (75).

Non all schemes (75) preserve convergence. Here is a counterexample. As a mapping (74) from OT-comparative into derived HG-comparative rows consider (77).

(77) $\mathbf{a} = (a_1, \ldots, a_n) \longrightarrow \overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n)$ s.t. $\overline{a}_k = \begin{cases} 1 & \text{if } a_k = \text{W} \\ 0 & \text{if } a_k = \text{E} \\ -1 & \text{if } a_k = \text{L} \end{cases}$

The HG online algorithm (71) converges with the HG update rule (78), by the convergence theorem of the *Perceptron Algorithm*; see Cristianini and Shawe-Taylor (2000).

(78) If $\boldsymbol{\theta}^{\text{old}}$ is not HG-compatible with the HG-comparative row $\overline{\mathbf{a}}$:
$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta}^{\text{old}} + \overline{\mathbf{a}}$

The update rule derived from (78) using the mapping (77) according to the scheme (75) is (79). But this is Boersma's (1997) OT update rule, shown in class 1 not to converge.

(79) If $\boldsymbol{\theta}^{\text{old}}$ is not OT-compatible with the OT-comparative row $\mathbf{a}$:
a. promote each winner-preferrer by 1;
b. demote each loser-preferrer by 1.

Thus, the scheme (75) from HG update rules into derived OT update rules corresponding to the mapping from OT- to HG-comparative rows in (77) does not preserve convergence.

11 Preservation of convergence: case of a unique L per row

■ **A temporary restriction.** Let me assume that:

(80) Every row of the input comparative tableau contains only one entry equal to L.

As seen in section 8, assumption (80) is not too restrictive: a row with multiple L's is equivalent to multiple rows with a unique L, as illustrated in (81).

(81) $\mathbf{A} = \begin{bmatrix} \text{W} & \text{L} & \text{L} \end{bmatrix} \implies \mathbf{A}' = \begin{bmatrix} \text{W} & \text{L} & \\ \text{W} & & \text{L} \end{bmatrix}$

By (80), the cautious promotion/demotion OT update rule studied in class 2 becomes:

(82) If $\boldsymbol{\theta}^{\text{old}}$ is not OT-compatible with the OT-comparative row $\mathbf{a}$:
a. promote each winner-preferrer by 1;
b. demote the unique loser-preferrer by the total number of WPCs.

I now present an alternative proof of convergence of the OT online algorithm with this cautious update rule. Assumption (80) is made temporarily, for the sake of clarity. In the next section, I will drop (80) and turn to the general update rule considered in class 2.

■ **Claim 6** Consider again a mapping (74) from OT- into derived HG-comparative rows:

(83) $\mathbf{a} = (a_1, \ldots, a_n) \longrightarrow \overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n)$ s.t. $\overline{a}_k$ is $\begin{cases} > 0 & \text{if } a_k = \text{W} \\ 0 & \text{if } a_k = \text{E} \\ \leq 0 & \text{if } a_k = \text{L} \end{cases}$

Suppose that this mapping satisfies (84) for every weight vector $\boldsymbol{\theta}$ and every OT-comparative row $\mathbf{a}$.

(84) If the OT-comparative row $\mathbf{a}$ is *not* OT-compatible with a ranking represented by $\boldsymbol{\theta}$, then the derived HG-comparative row $\overline{\mathbf{a}}$ is *not* HG-compatible with $\boldsymbol{\theta}$.

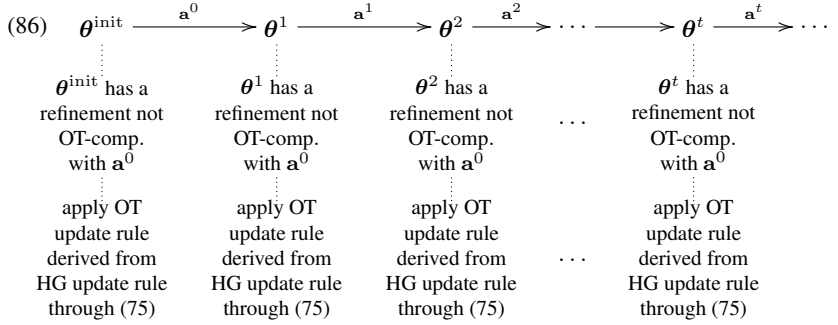
If a mapping (83) from OT- into derived HG-comparative rows satisfies (84), then the corresponding scheme (75) from HG update rules into derived OT update rules preserves convergence.

■ **Remark.** The mapping (77) that corresponds to Boersma's update rule (79) does not satisfy (84), as shown by the counterexample in (85).

(85) a. OT-comparative row:
$\mathbf{a} = \begin{bmatrix} C_1 & C_2 & C_3 \\ \text{W} & \text{W} & \text{L} \end{bmatrix}$
b. HG-comparative row derived from $\mathbf{a}$ according to (77):
$\overline{\mathbf{a}} = \begin{bmatrix} C_1 & C_2 & C_3 \\ 1 & 1 & -1 \end{bmatrix}$
c. weight vector HG-compatible with $\overline{\mathbf{a}}$:
$\boldsymbol{\theta} = \begin{bmatrix} C_1 & C_2 & C_3 \\ 2 & 2 & 3 \end{bmatrix}$
d. refinement of $\boldsymbol{\theta}$ not OT-compatible with $\mathbf{a}$:
$C_3 \gg C_2 \gg C_1$

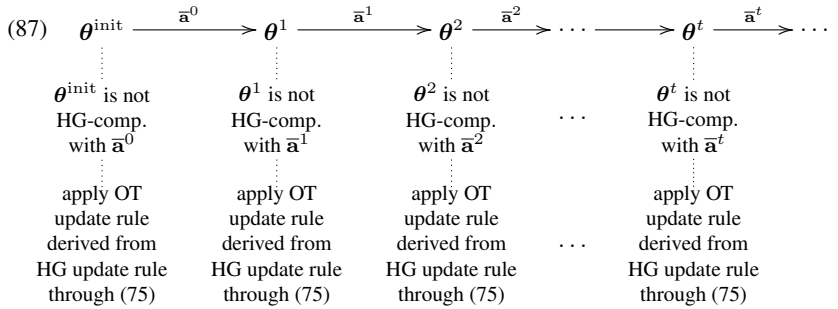
■ *Proof.* By contradiction, suppose the OT online algorithm fails to converge with the derived OT update rule. This means that (86) holds, where the OT-comparative rows $\mathbf{a}^0, \mathbf{a}^1, \mathbf{a}^2, \ldots$ are sampled from some input OT-compatible OT-comparative tableau $\mathbf{A}$.

(86)



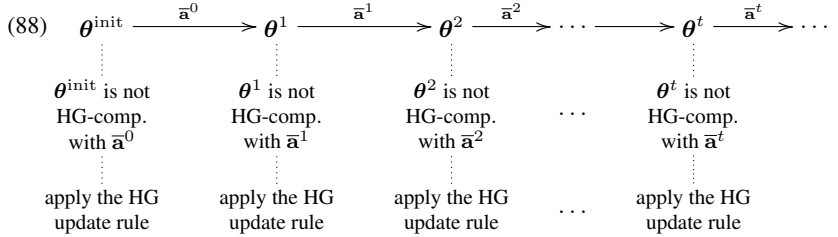
Let $\overline{\mathbf{a}}^t$ be the HG-comparative row derived from the OT-comparative row $\mathbf{a}^t$ with a mapping (74). Since that mapping satisfies condition (84) by hypothesis, then $\overline{\mathbf{a}}^t$ is not HG-compatible with the weight vector $\boldsymbol{\theta}^t$. The situation (86) thus entails (87).

(87)



Update by the OT-comparative row $\mathbf{a}^t$ according to the OT update rule derived through (75) coincides by definition with update by the HG-comparative row $\overline{\mathbf{a}}^t$ according to the original HG update rule. The situation (87) thus entails (88).

(88)



Let $\overline{\mathbf{A}}$ be the HG-comparative tableau derived from the input OT-comparative tableau. Claim 1 from ensures that $\overline{\mathbf{A}}$ is HG-compatible, since it is derived from an OT-compatible tableau. The diagram in (88) thus contradicts the hypothesis that the HG online algorithm converges with the given HG update rule. □

■ **Claim 7** Suppose that the input tableau contains a unique L per row. The OT online algorithm (69) with the cautious promotion/demotion OT update rule (82) converges.

■ *Proof.* The OT update rule (82) is derived from the convergent HG update rule (78) through scheme (75) via the mapping from OT-comparative rows into derived HG-comparative rows defined in (89).

(89)

$$\mathbf{a} = (a_1, \ldots, a_n) \qquad \overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n) \qquad \text{such that } \overline{a}_k = \begin{cases} 1 & \text{if } a_k = \text{W} \\ 0 & \text{if } a_k = \text{E} \\ -w(\mathbf{a}) & \text{if } a_k = \text{L} \end{cases}$$

Claim 5 ensures that the mapping (89) satisfies condition (84) — recall I am assuming that the input row has a unique L, hence the non-negativity of weights does not matter. Convergence thus follows from claim 6 □

12 Preservation of convergence: general case

■ **Idea.** I want to drop assumption (80) that input rows have a unique L, and prove convergence for the general cautious promotion/demotion update rule of class 3, namely:

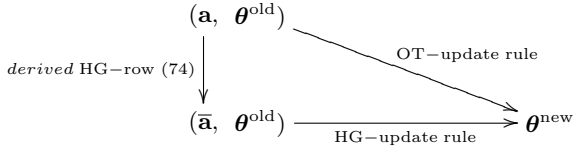
(90)
a. Promote each WPC by the total number of undominated LPCs;
b. demote each LPC by the total number of WPCs.

This is not straightforward, in fact:

(91)
a. I cannot use claim 2, because it requires weights to be nonnegative;
b. I cannot use claim 5, because the input rows don't have a unique L per row.

Yet, there is a very simple way out. Consider again scheme (75) used so far to derive an OT update rule from an HG update rule, repeated below:

(92)



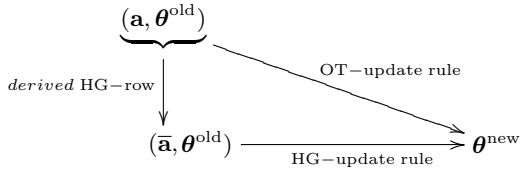
Here, I am assuming a mapping from an OT-comparative row $\mathbf{a}$ into a derived HG-comparative row $\overline{\mathbf{a}}$ of the form (93), independent of the current ranking vector $\boldsymbol{\theta}^{\text{old}}$.

(93)

$$\mathbf{a} = (a_1, \ldots, a_n) \qquad \overline{\mathbf{a}} = (\overline{a}_1, \ldots, \overline{a}_n) \qquad \text{such that } \overline{a}_k \begin{cases} > 0 & \text{if } a_k = \text{W} \\ = 0 & \text{if } a_k = \text{E} \\ \leq 0 & \text{if } a_k = \text{L} \end{cases}$$

Let's switch to a slight variant of (92) as follows:

(94)



where I am using a mapping from an OT-comparative row $\mathbf{a}$ into a derived HG-comparative row $\overline{\mathbf{a}}$ of the form (95), that depends on the current ranking vector $\boldsymbol{\theta}^{\text{old}}$.

(95) $(\mathbf{a}, \boldsymbol{\theta}^{\text{old}})$

$\downarrow$

$\overline{\mathbf{a}}$ such that $\overline{a}_k \begin{cases} > 0 & \text{if } a_k = \text{W} \\ = 0 & \text{if } a_k = \text{E} \\ \leq 0 & \text{if } a_k = \text{L} \end{cases}$

■ **Claim 8** Consider a mapping (95) from OT comparative rows and ranking vectors into derived HG-comparative rows. Suppose that this mapping satisfies (84) for every ranking vector $\boldsymbol{\theta}$ and every OT-comparative row $\mathbf{a}$.

(96) If the OT-comparative row $\mathbf{a}$ is *not* OT-compatible with a ranking represented by $\boldsymbol{\theta}$, then the HG-comparative row $\overline{\mathbf{a}}$ derived through (83) from that OT-comparative row and that ranking vector is *not* HG-compatible with $\boldsymbol{\theta}$.

If a mapping (95) satisfies (96), then the corresponding scheme (94) from HG update rules into derived OT update rules preserves convergence.

Proof. Identical to the proof of claim 6. □

■ **Claim 9** The OT online algorithm (69) with the cautious promotion/demotion OT update rule (90) converges.

Proof. The OT update rule (90) is derived from the convergent HG update rule (78) through scheme (94) via the following mapping of the form (95):

(97) $\overline{a}_k = \begin{cases} \text{\# of undominated LPCs} & \text{if } a_k = \text{W} \\ 0 & \text{if } a_k = \text{E} \\ -\text{\#of WPCs} & \text{if } a_k = \text{L} \end{cases}$

(This mapping depends both on the current comparative row $\mathbf{a}$ and on the current ranking vector $\boldsymbol{\theta}^{\text{old}}$, as it involves the number of currently *undominated* LPCs).

The mapping (97) satisfies condition (84), as shown by the chain of inequalities in (98), that is analogous to (51).

(98)
$$\sum_{h=1}^{n} \theta_h \overline{a}_h \overset{(a)}{=} \sum_{h \in \text{WPC}} \theta_h \overline{a}_h + \sum_{h \in \text{ULPC}} \theta_h \overline{a}_h + \sum_{h \notin \text{WPC} \cup \text{ULPC}} \theta_h \overline{a}_h$$

by splitting the set $\{1, \ldots, n\}$ that h ranges over into the the set WPC of currently winner-preferring constraints, the set ULPC of currently undominated loser-preferring constraints and their complement

$$\overset{(b)}{=} \ell \sum_{h \in \text{WPC}} \theta_h - w \sum_{h \in \text{ULPC}} \theta_h$$

by the definition (97) of the components $\overline{a}_1, \ldots, \overline{a}_n$ of the derived HG-comparative row, where w and ℓ are the numbers of WPCs and of undominated LPCs

$$\overset{(c)}{\leq} \ell w \max_{h \in \text{WPC}} \theta_h - w \sum_{h \in \text{ULPC}} \theta_h$$

by upper bounding the sum $\sum_{h \in \text{WPC}} \theta_h$ with its biggest term $\max_{h \in \text{WPC}} \theta_h$ multiplied by the number w of terms

$$\overset{(d)}{\leq} \ell w \max_{h \in \text{WPC}} \theta_h - w\ell \min_{h \in \text{ULPC}} \theta_h$$

by lower bounding the sum $\sum_{h \in \text{WPC}} \theta_h$ with its smallest term $\min_{h \in \text{ULPC}} \theta_h$ multiplied by the number ℓ of terms

$$= w\ell \underbrace{\left(\max_{h \in \text{WPC}} \theta_h - \min_{k \in \text{ULPC}} \theta_k \right)}_{(*)}$$

$$\overset{(e)}{\leq} 0$$

the hypothesis that the ranking vector $\boldsymbol{\theta}$ admits a refinement which is not OT-compatible with $\mathbf{a}$ entails that (*) is non-positive.

Since the mapping in (97) satisfies condition (84), claim 8 then ensures convergence. □

References

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

Boersma, Paul. 1997. "How We Learn Variation, Optionality and Probability". In *IFA Proceedings 21*, 43–58. University of Amsterdam: Institute for Phonetic Sciences.

Boersma, Paul, and Joe Pater. 2007. "Convergence Properties of a Gradual Learner for Harmonic Grammar". In *Proceedings of NELS 38*, –. .

Boersma, Paul, and Joe Pater. 2008. "Convergence Properties of a Gradual Learning Algorithm for Harmonic Grammar". Ms.

Cesa-Bianchi, Nicolò, and Gábor Lugosi. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

Coetzee, Andries W., and Joe Pater. 2008. Weighted constraints and gradient restrictions on place co-occurrence in muna and arabic. *Natural Language and Linguistic Theory* 26:289–337.

Cristianini, Nello, and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Methods*. Cambridge University Press.

Dombi, József, Csanád Imreh, and Nándor Vincze. 2007. "Learning Lexicographic Orders". *European Journal of Operational Research* 183.2:748–756.

Fishburn, P. C. 1974. "Lexicographic Orders, Utilities and Decision Rules: A Survey". *Management Science* 20:1442–1471.

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

Jesney, Karen, and Anne-Michelle Tessier. 2007. "Re-evaluating learning biases in Harmonic Grammar". In *University of massachusetts occasional papers 36: Papers in theoretical and computational phonology*, ed. Michael Becker.

Jesney, Karen, and Anne-Michelle Tessier. 2008. "Gradual learning and faithfulness: consequences of ranked vs. weighted constraints". In *Proceedings of NELS38*, –.

Keller, Frank. 2000. Gradience in Grammar. Experimental and Computational Aspects of Degrees of Grammaticality. Doctoral Dissertation, University of Edinburgh.

Keller, Frank. 2005. "Linear Optimality Theory as a Model of Gradience in Grammar". In *Gradience in Grammar: Generative Perspectives*, ed. Gisbert Fanselow, Caroline Féry, Ralph Vogel, and Matthias Schlesewsky, –. Oxford: Oxford University Press.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1990a. "Harmonic Grammar: A formal multi-level connectionist theory of linguistic well-formedness: An application". In *Proceedings of the twelfth annual conference of the Cognitive Science Society*, 884–891. Cambridge, MA: Lawrence Erlbaum.

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

Pater, Joe. 2009. "Weighted Constraints in Generative Linguistics". *Cognitive Science* 33:999–1035.

Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt, and Michael Becker. to appear. "Harmonic Grammar with Linear Programming: From linear systems to linguistic typology". Ms, University of Massachusetts, Amherst; to appear in *Phonology*.

Prince, Alan, and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell. As 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, New Brunswick, NJ, April 1993. Rutgers Optimality Archive 537 version, 2002.

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