The error-driven ranking model of the acquisition of phonotactics: how to keep the faithfulness constraints at bay

Giorgio Magri
SFL (CNRS and University of Paris 8)
UiL OTS (Utrecht University)
magrigrg@gmail.com

Abstract

A problem which arises in the theory of the error-driven ranking model of the acquisition of phonotactics is that the faithfulness constraints need to be promoted but should not be promoted too high. This paper motivates this technical problem and shows how to tune the promotion component of the re-ranking rule so as to keep the faithfulness constraints at bay.

Sections 1-2 introduce the algorithmic framework considered in the paper, namely the *error-driven ranking model* of the acquisition of phonotactics. Section 3 motivates a specific problem which arises in the design and analysis of this model, namely the problem of *controlling* the height reached by the *faithfulness* ($\mathcal{F}$) constraints. Sections 4-6 sketch the theory of $\mathcal{F}$-controlling. Magri (2014a) presents the theory in more detail.

1 The acquisition of phonotactics

Generative linguistics assumes that the learner is provided with a typology of grammars G_1, G_2... The language-learning problem thus consists of individuating the target adult grammar G^* within the typology, on the basis of a finite set of data generated by that grammar. Various formulations of this problem differ for the structural assumptions about the underlying typology, for the type of data fed to the learner, and for the criteria of success used to evaluate the grammar $\hat{G}$ chosen by the learner relative to the target grammar G^*.

In this paper, I focus on the following specific formulation of this general language learning problem. The typology consists of the phonological grammars defined in *Optimality Theoretic* (OT) terms through the rankings of a given set of constraints (Prince and Smolensky, 2004). The data fed to the learner consist of surface forms sampled from the *language* L^* generated by the target OT grammar G^*, namely the set of surface forms which are the phonological realizations of some underlying forms according to G^*. The criteria for success is that the OT grammar $\hat{G}$ chosen by the learner generates a language $\hat{L}$ which coincides with the target one: $\hat{L} = L^*$.

This specific formulation is called the *problem of the acquisition of phonotactics*. In fact, phonotactics is the knowledge of the distinction between licit and illicit forms. Assuming that the distinction is categorical (Gorman, 2013), knowledge of phonotactics reduces to knowledge of the set of licit forms (the set of illicit forms is just the complement). And the set of licit forms relative to an OT grammar G is the corresponding language L_G.

2 The EDRA model

In this paper, I focus on a specific algorithmic approach to the problem of the acquisition of phonotactics, based on *error-driven ranking algorithms* (EDRAs). This approach is summarized below and explained in the rest of this section.

Algorithm 1 *The EDRA model*

```
Initialize
  the ranking values of $\mathcal{F}$ constraints to zero
  the ranking values of $\mathcal{M}$ constraints to $\theta^{\text{init}} > 0$
Repeat
  1 get a surface form $[y]$ from the target language
  2 pick a loser form $\cancel{[z]}$
  3 check whether the current ranking vector $\boldsymbol{\theta}$ is
    consistent with the underlying/winner/loser
    form triplet $(/y/, [y], \cancel{[z]})$
  4 if it isn't, update the current ranking vector $\boldsymbol{\theta}$
until no more mistakes are made at step 3
```

The EDRA model maintains a current hypothesis of the target adult grammar, namely a current constraint ranking. This ranking is represented numerically through a *ranking vector* $\boldsymbol{\theta} = (\theta_1, , \theta_n)$

Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, pages 10–18,
Baltimore, Maryland USA, June 27 2014. ©2014 Association for Computational Linguistics

which assigns to each constraint C_k a numerical *ranking value* θ_k. A constraint C_k is ranked above another constraint C_h according to a ranking vector $\boldsymbol{\theta}$ provided the ranking value θ_k of the former is (strictly) larger than the ranking value θ_h of the latter (Boersma, 1998; Boersma, 2009).

The current ranking (vector) is initialized in such a way that the corresponding initial language is as small as possible. OT constraints come in two varieties: *faithfulness* ($\mathcal{F}$) and *markedness* ($\mathcal{M}$) constraints. A smallest language corresponds to a ranking which assigns all $\mathcal{F}$ constraints underneath all $\mathcal{M}$ constraints. Thus, the $\mathcal{F}$ constraints are assigned a small initial ranking value, say zero for concreteness; and the $\mathcal{M}$ constraints start with a large positive initial ranking value $\theta^{\text{init}} > 0$. The algorithm then loops through the three steps **1**-**4**.

At step **1**, the EDRA model receives a piece of data, namely a surface form $[y]$ sampled from the target language L^*. Assuming that the underlying typology satisfies Tesar's (2013) *Surface Orientedness Condition*, this piece of data provides evidence that the target grammar G^* maps this phonological form (construed as the underlying form $/y/$) into itself (construed as the surface form $[y]$) rather than reducing it to some non-faithful candidate ~~$[z]$~~ (as a mnemonic, I strike out a candidate construed as a loser). In other words, the target adult ranking (vector) $\boldsymbol{\theta}^*$ is *consistent* with the underlying/winner/loser form triplet ($/y/, [y],$ ~~$[z]$~~) for any loser ~~$[z]$~~, namely it satisfies condition (1). Here, W (L) is the set of *winner-preferring* (*loser-preferring*) constraints, namely those which assign less (more) violations to the faithful mapping of $/y/$ to $[y]$ than to the neutralization of $/y/$ to ~~$[z]$~~.

(1) $$\max_{C_k \in W} \theta^*_k > \max_{C_h \in L} \theta^*_h$$

This consistency condition (1) says that there is at least a winner-preferring constraint which is ranked above all loser-preferring constraints by the target ranking (vector) $\boldsymbol{\theta}^* = (\theta^*_1, \ldots, \theta^*_n)$.

At steps **2** and **3**, the EDRA model thus picks a specific loser ~~$[z]$~~ and checks whether its current ranking vector $\boldsymbol{\theta}$ satisfies the corresponding consistency condition (1). Failure to satisfy this condition means that the current ranking values of the loser-preferring (winner-preferring) constraints are too large (too small). The algorithm thus promotes the winner-preferring constraints by a small *promotion amount* and demotes the loser-preferring constraints by a small *demotion amount*. What matters is not the actual values of the promotion and demotion amounts, but rather their ratio. Thus, the demotion amount can be set equal to 1 for concreteness, letting instead the promotion amount be equal to an arbitrary nonnegative constant $p \geq 0$, as in (2).

(2) a. Increase the ranking value of each winner-preferring constraint by $p \geq 0$;
b. decrease the ranking value of each *undominated* loser-preferring constraint by 1.

Crucially, not *all* loser-preferring constraints are demoted by (2b), but only those that need to be demoted, namely the *undominated* ones (Tesar and Smolensky, 1998), whose current ranking value is at least as large as the ranking value of all winner-preferring constraints and thus are responsible for flouting the consistency condition (1).

3 The problem of $\mathcal{F}$-control

The crucial implementation parameter of the EDRA model is the promotion amount $p \geq 0$ used in the promotion component (2a) of the re-ranking rule. How should this parameter be tuned so as to optimize the performance of the EDRA model of the acquisition of phonotactics? This section explains how this question leads to the problem of controlling the height of the $\mathcal{F}$ constraints.

3.1 Some initial guarantees

The problem of the acquisition of phonotactics in OT is intractable: no algorithm can solve efficiently an arbitrary instance of the problem corresponding to an arbitrary constraint set (Magri, 2013a). Prompted by this intractability result, Magri (2013b) starts to tackle the problem by looking at a class of "easy" cases.

The intuitive idea is that the relative ranking of the $\mathcal{F}$ constraints might often be irrelevant for phonotactics, namely for drawing the line between licit and illicit forms (although it is of course always crucial for phonology, namely for the specific way in which illicit forms are repaired). This intuition that the relative ranking of the $\mathcal{F}$ constraints is not relevant to describe a certain phonotactic pattern can be formalized as follows. A *partial* constraint ranking is any partial order on the constraint set. A partial ranking *generates* a language provided each one of its total refinements generates that language in the usual OT sense (Yanovich, 2012). A language is called *$\mathcal{F}$-irrelevant* provided it can be generated in this technical sense by a partial ranking which does not rank any two $\mathcal{F}$ constraints relative to each other (see subsection 3.2 for an example).

Suppose that the EDRA model is trained on a target language L^* which is $\mathcal{F}$-irrelevant. The $\mathcal{M}$ constraints start out high, with an initial ranking value θ^{init} usually larger than the number m of markedness constraints. The $\mathcal{F}$ constraints instead start out low, with a null initial ranking value. Throughout learning, the $\mathcal{F}$ constraints will raise, if the algorithm adopts a non-null promotion amount $p > 0$. Theorem 1 provides guarantees that the EDRA model learns the target phonotactics, as long as the $\mathcal{F}$ constraints don't raise too high, namely their ranking values remain smaller by at least m than the initial ranking value θ^{init} of the $\mathcal{M}$ constraints, as stated in (3).

Theorem 1 *Suppose that the underlying OT typology satisfies the following two assumptions. First, if a surface form $[y]$ is a non-faithful candidate of an underlying form $/x/$, then there exists at least one faithfulness constraint which assigns at least one violation to the mapping of $/x/$ into $[y]$ (*$\mathcal{F}$-discernibility assumption*). Second, a form $[y]$ is a candidate of an underlying form $/x/$ if and only if the latter form construed as the surface form $[x]$ is vice versa a candidate of the former form construed as the underlying form $/y/$ (*symmetric candidacy assumption*). Consider a language in this OT typology which is $\mathcal{F}$-irrelevant. Suppose that the EDRA model only makes a finite number of errors and then converges to a final ranking vector which is never updated again. Suppose furthermore that the ranking value θ_F of any $\mathcal{F}$ constraint F at any time in the run satisfies condition (3), where m is the number of $\mathcal{M}$ constraints and $\theta^{init} > m$ their initial ranking value.*

(3) $\boxed{\theta_F \leq \theta^{init} - m}$

Then, the language generated by (an arbitrary refinement of) the final ranking vector learned by the EDRA model coincides with the target language the EDRA model has been trained on. ■

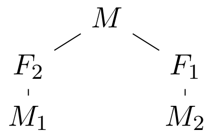
The two assumptions of $\mathcal{F}$-discernibility and symmetric candidacy required by theorem 1 are extremely mild. Magri (2013b; 2014b) conjectures that the relative ranking of the faithfulness constraints turns out to matter for phonotactics only in very special configurations, so that the $\mathcal{F}$-irrelevancy assumption might plausibly hold in the vast majority of cases. Theorem 1 thus provides guarantees that the EDRA model succeeds at the problem of the acquisition of phonotactics in a large class of cases under two crucial assumptions. One assumption is that it can only make a finite number of errors before it *converges* to a final ranking which is consistent with any form and thus never updated. The other assumption is the condition (3) that the height of the $\mathcal{F}$-constraints can be properly controlled.

3.2 Some examples

To illustrate the issues raised by convergence and $\mathcal{F}$-control, consider the following OT typology. The set of forms consists of only four forms $\{apsa, apza, absa, abza\}$. The faithfulness constraints are the two identity constraints for voicing in stops and fricatives (F_1, F_2). The markedness constraints are the two corresponding constraints against stop and fricative voicing (M_1, M_2) plus an additional constraint M which bans sequences of stops and fricatives which agree in voicing, namely it is violated by the two forms *apsa* and *abza*. The candidacy relation is total: the four forms are all candidates of each other.

The OT typology just described contains in particular the language $L =$ {[absa], [apza]}. This language is generated by any ranking which satisfies the ranking conditions (4).

(4)



These ranking conditions (4) say nothing about the relative ranking of the two $\mathcal{F}$ constraints F_1 and F_2. The language L thus qualifies as $\mathcal{F}$-irrelevant.

When trained on this language, the EDRA model will be provided at step 1 with a sequence of the two licit forms [absa] and [apza]. It will then complete them into an underlying/winner/loser form triplet at steps 2 and 3 by assuming a faithful underlying form and a non-faithful loser form. The list of all possible such triplets that the algorithm can consider is provided in (5).

(5)

	F_1	F_2	M_1	M_2	M
(/absa/, [absa], ~~[apsa]~~)	W	*e*	L	*e*	W
(/absa/, [absa], ~~[abza]~~)	*e*	W	*e*	W	W
(/absa/, [absa], ~~[apza]~~)	W	W	L	W	*e*
(/apza/, [apza], ~~[abza]~~)	W	*e*	W	*e*	W
(/apza/, [apza], ~~[apsa]~~)	*e*	W	*e*	L	W
(/apza/, [apza], ~~[absa]~~)	W	W	W	L	*e*

Each triplet is described here in *ERC notation* (Prince, 2002): constraints which are winner- or loser-preferring or even relative to a triplet are marked with a corresponding W or L or e.

The triplets where the constraint M is winner-preferring will trigger virtually no update, since that constraint starts high and is never demoted, and will thus always ensure consistency with those triplets. The learning run is thus driven by the two remaining triplets, boldfaced in (5), which I assume are fed one after the other to the algorithm. Suppose the promotion amount is non-null, say equal to the demotion amount: $p = 1$. The resulting learning run is described in (6).

(6)
$$\begin{array}{c} F_1 \\ F_2 \\ M_1 \\ M_2 \\ M \end{array}\begin{bmatrix}0\\0\\5\\5\\5\end{bmatrix}\to\begin{bmatrix}1\\1\\4\\6\\5\end{bmatrix}\to\begin{bmatrix}2\\2\\5\\5\\5\end{bmatrix}\to\begin{bmatrix}3\\3\\4\\6\\5\end{bmatrix}\to\begin{bmatrix}4\\4\\5\\5\\5\end{bmatrix}\to\begin{bmatrix}5\\5\\4\\6\\5\end{bmatrix}\to\begin{bmatrix}6\\6\\5\\5\\5\end{bmatrix}$$

The two $\mathcal{F}$ constraints end up too high, namely with a final ranking value $\theta_{F_1} = \theta_{F_2} = 6$ which is larger than the initial ranking value $\theta^{\text{init}} = 5$ of the $\mathcal{M}$ constraints. And indeed the EDRA has failed at learning the target phonotactics: since the $\mathcal{F}$ constraints are ranked at the top, the model has incorrectly learned that any form is licit.

A trivial strategy to enforce the $\mathcal{F}$-control condition (3) would be to threshold the promotion component (2a) of the re-ranking rule, as in (2a′).

(2) a′. Increase the ranking value of each winner-preferring constraint by p, *except for an $\mathcal{F}$ constraint which is already close to the forbidden threshold θ^{init}–m.*

Yet, suppose we tried to remedy to the failure in (6) by thresholding the promotions as in (2a′). When an $\mathcal{F}$ constraint reaches the height $\theta^{\text{init}} - m = 5 - 3 = 2$, we stop promoting it, as boldfaced in the learning run (7).

(7)
$$\begin{array}{c} F_1 \\ F_2 \\ M_1 \\ M_2 \\ M \end{array}\begin{bmatrix}0\\0\\5\\5\\5\end{bmatrix}\to\begin{bmatrix}1\\1\\4\\6\\5\end{bmatrix}\to\begin{bmatrix}2\\2\\5\\5\\5\end{bmatrix}\to\begin{bmatrix}\mathbf{2}\\\mathbf{2}\\4\\6\\5\end{bmatrix}\to\begin{bmatrix}\mathbf{2}\\\mathbf{2}\\5\\5\\5\end{bmatrix}\to\begin{bmatrix}\mathbf{2}\\\mathbf{2}\\4\\6\\5\end{bmatrix}\to\begin{bmatrix}\mathbf{2}\\\mathbf{2}\\5\\5\\5\end{bmatrix}\ \ldots$$

In this run, the constraint M stays put at its initial position. The constraints M_1 and M_2 oscillate up and down, because promoted and demoted by the two boldfaced triplets in (5). The constraints F_1 and F_2 raise a bit until they hit the threshold, and then settle. The EDRA model will thus keep making mistakes forever, without ever converging to a ranking vector consistent with the data.

3.3 Against a null promotion amount

These difficulties with convergence and the $\mathcal{F}$-control condition (3) would disappear if the promotion amount p was set equal to zero, so that the EDRA performs no constraint promotion at all. In fact, Tesar and Smolensky (1998) guarantee convergence for the demotion-only case. And the $\mathcal{F}$ constraints could not possibly be promoted too high, as they would not be promoted at all.

Unfortunately, the option of a null promotion amount is not viable, as argued in Magri (2012; 2014b). In fact, recall that the EDRA model at step **3** always considers underlying/winner/loser form triplets $(/y/, [y], \cancel{[z]})$ which have an underlying form $/y/$ faithful to the winner $[y]$. This means that the $\mathcal{F}$ constraints are never loser-preferring and are therefore never demoted. If the promotion amount is set equal to zero, then they will not be promoted either. In the end, the $\mathcal{F}$ constraints will thus never be re-ranked. This hampers the ability of the EDRA model to learn the correct relative ranking of the $\mathcal{F}$ constraints when trained on a $\mathcal{F}$-relevant language, namely when it needs to learn a phonotactic pattern which crucially does require a specific relative ranking of the $\mathcal{F}$ constraints.

3.4 Convergence through calibration

As recalled above, Tesar and Smolensky (1998) show that the EDRA model converges when the promotion amount is null and the algorithm performs only constraint demotion. It could in principle be the case that convergence does not extend to the demotion/promotion case, because any amount of promotion disrupts convergence. But Magri (2012) shows that is not the case: convergence extends to EDRAs which perform constraint promotion as well, as long as the promotion amount is small enough. In particular, consider a promotion amount p which scales as in (8) with the numbers ℓ and w of currently undominated loser-preferring constraints and of winner-preferring constraints.

(8) $$p = \frac{\ell}{w + \sigma}$$

It turns out that the EDRA model converges efficiently if (and only if) the promotion amount is *calibrated*, namely has the shape in (8) corresponding to some strictly positive *calibration constant* $\sigma > 0$. The larger the calibration constant σ, the smaller the promotion amount. The case of a null promotion amount corresponds to the limiting case $\sigma = \infty$.

3.5 $\mathcal{F}$-control through calibration as well

Let's take stock. Theorem 1 provides some initial guarantees of success of the EDRA model of the acquisition of phonotactics. These guarantees hold under two crucial assumptions: convergence and the $\mathcal{F}$-control condition (3). Do these assumptions hold when the promotion amount is non-null? Convergence does hold, if the promotion amount, although not null, is nonetheless small, namely calibrated as in (8). What about the $\mathcal{F}$-control condition (3)? Can we play the same trick of a small promotion amount? Or is it the case that, no matter how small the promotion amount, as soon as it is allowed to be non-null, the $\mathcal{F}$ constraints raise too high through a long sequence of very small promotions? Section 4 shows that the latter scenario can never arise: the $\mathcal{F}$ constraints can never raise too high if the promotion amount is small enough. More precisely, it assumes a calibrated promotion amount as in (8). And it shows that the $\mathcal{F}$-control condition (3) holds when the calibration constant σ is large enough, namely it grows as the number m of $\mathcal{M}$ constraints.

As the calibration constant increases as the number m of markedness constraints, the promotion amount decreases quickly. Is it possible to improve on the analysis of section 4 and guarantee the $\mathcal{F}$-control condition (3) with a calibration constant σ which does not grow with the number m of markedness constraints? Unfortunately, section 5 shows that the calibration constant must increase with m. More precisely, this section considers the very simple case where there is a single $\mathcal{F}$ constraint and where the $\mathcal{M}$ constraints are always loser-preferring (or even) but never winner-preferring. In this case, the $\mathcal{F}$-control condition fails if the calibration constant σ does not grow with m at least as $\frac{m}{\log m}$.

Interestingly, the derivative of the function $\frac{m}{\log m}$ goes to zero as m grows. In other words, although the function increases with m, the rate of increase becomes smaller and smaller, making this function as close as possible to a constant. Is this particularly favorable choice of the calibration constant only possible in the peculiar case considered in section 5? or does this favorable choice of the calibration constant ensure $\mathcal{F}$-calibration also in the general case? Section 6 shows how to relax at least one of the two restrictive assumptions made in section 5, namely the assumption that the $\mathcal{M}$ constraints are never winner-preferring.

4 $\mathcal{F}$-controlling with a non-null promotion amount

The most basic question of the theory of $\mathcal{F}$-control is as follows: is it possible to guarantee the $\mathcal{F}$-control condition (3) despite a non-null promotion amount? This section provides a positive answer to this question. In particular, assume the promotion amount p is calibrated as in (8), through the calibration constant σ. The $\mathcal{F}$-control condition then holds provided the calibration constant σ satisfies the bound (9), where m is the number of $\mathcal{M}$ constraints and θ^{init} is their initial ranking value.

$$(9)\quad \sigma \geq \frac{2m + m\theta^{\text{init}}}{\theta^{\text{init}} - m}$$

To get a sense of the bound (9), assume that the initial ranking value θ^{init} of the $\mathcal{M}$ constraints is some power of the number m of $\mathcal{M}$ constraints: $\theta^{\text{init}} = m^k$ for some $k > 1$. The bound (9) thus becomes $\frac{m^k+2}{m^{k-1}-1}$, which is approximately m.

At each update, each of the ℓ currently undominated loser-preferring constraints is demoted by 1 and each of the w winner-preferring constraints is promoted by p. Because of the specific shape (8) of the promotion amount p, the sum of the current ranking values decreases by $\ell - wp = \ell - \frac{w\ell}{w+\sigma} = \frac{\ell\sigma}{w+\sigma}$. And the latter is at least $\frac{\sigma}{w+\sigma}$, as every update requires at least one undominated loser-preferring constraint, namely $\ell \geq 1$. Let α_i be the number of updates triggered by the ith ERC in the run considered up to the time considered (note that there is only a finite number of ERCs relative to a finite number of constraints). Thus, the sum $\sum_k \theta_k$ of the current ranking values has overall decreased by at least $\sum_i \alpha_i \frac{\sigma}{w_i+\sigma}$ relative to the the sum $\sum_k \theta_k^{\text{init}}$ of the initial ranking values, as stated in (10).

$$(10)\quad \sum_k \theta_k \leq \sum_k \theta_k^{\text{init}} - \sum_i \alpha_i \frac{\sigma}{w_i + \sigma}$$

The sum $\sum_k \theta_k^{\text{init}}$ of the initial ranking values can be computed explicitly as in (11), as the m $\mathcal{M}$ constraints start with the initial ranking value θ^{init} while the $\mathcal{F}$ constraints start with a null initial ranking value.

$$(11)\quad \sum_k \theta_k^{\text{init}} = m\,\theta^{\text{init}}$$

The sum $\sum_k \theta_k$ of the current ranking values can be lower bounded as in (12).

(12) $$\sum_k \theta_k \overset{(a)}{=} \sum_F \theta_F + \sum_M \theta_M \overset{(b)}{\geq} 0 + \sum_M \theta_M \overset{(c)}{>} 0 + m(-2) = -2m$$

In step (11a), I have split the sum over all constraints into the sum over the faithfulness and the markedness constraints. In step (11b), I have noted that the ranking value θ_F of any faithfulness constraint F is always at least as large as 0. In fact, the faithfulness constraints start with a null initial ranking value and are never demoted, because the EDRA model always assumes an underlying form faithful to the winner, so that the faithfulness constraints are never loser-preferring. In step (11c), I have noted that the ranking value θ_M of a markedness constraint M can never get smaller than -2. In fact, suppose by contradiction that M managed to be demoted that low. That would imply that some ERC triggers an update that demotes M despite the fact that its current ranking value is strictly smaller than 0. And that is impossible. In fact, at least one faithfulness constraint F must be winner-preferring relative to that ERC, because of the $\mathcal{F}$-discernibility assumption. Furthermore, that constraint F must already dominate M, because F has a non-negative current ranking value while M has a negative current ranking value.

Using the expressions for the sum of the initial and the current ranking values obtained in (11) and (12) respectively, the original inequality (10) yields the bound in (13).

(13) $$\sum_i \alpha_i \frac{1}{w_i + \sigma} < \frac{2m + m\theta^{\text{init}}}{\sigma}$$

The ranking value θ_F of a generic faithfulness constraint F can now be bound as in (14).

(14) $$\theta_F \overset{(a)}{\leq} \sum_i \alpha_i \frac{1}{w_i + \sigma} \overset{(b)}{<} \frac{2m + m\theta^{\text{init}}}{\sigma} \overset{(c)}{\leq} \theta^{\text{init}} - m$$

In step (14a), I have used the fact that the faithfulness constraint F starts with a null initial ranking value and is promoted by $\frac{1}{w_i+\sigma}$ for each one of the α_i updates triggered by the ith ERC, as long as F is winner-preferring relative to that ERC. In step (14b), I have used the bound computed in (13). And in step (14c), I have used the choice (9) of the calibration constant σ.

The bound obtained in (14) guarantees that the generic faithfulness constraint F never raises above the forbidden threshold $\theta^{\text{init}} - m$, thus complying with the $\mathcal{F}$-control condition (3). In other words, we have obtained the following sufficient solution to the problem of $\mathcal{F}$-controlling.

Theorem 2 *Suppose the underlying typology satisfies the $\mathcal{F}$-discernibility assumption. Consider a run of the EDRA model on an arbitrary language in that typology. Assume that the $\mathcal{F}$ constraints start out with a null initial ranking value while the m $\mathcal{M}$ constraints start out with an initial ranking value $\theta^{\text{init}} > m$. Assume furthermore that the promotion amount is calibrated as in (8) and that the calibration constant σ is large enough to satisfy the bound (9). Then, the ranking values of the $\mathcal{F}$ constraints remain smaller than the forbidden threshold $\theta^{\text{init}} - m$ throughout the entire run.* ■

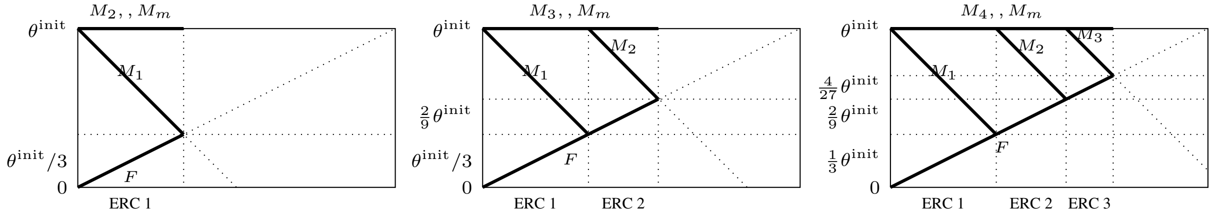
5 $\mathcal{F}$-controlling on the diagonal case

The preceding section has established the $\mathcal{F}$-control condition (3) when the promotion amount is not null, provided it is small enough, namely it corresponds to a calibration constant which grows as the number m of $\mathcal{M}$ constraints. Is it possible to do better? In particular, is it possible to guarantee $\mathcal{F}$-control when the calibration constant does not increase with m? This section sketches a counterexample which provides a negative answer to this question; see Magri (2014a) for details.

At every iteration, the EDRA model receives a winner form sampled from the target language, assumes a corresponding faithful underlying form and picks a corresponding loser candidate. At every iteration, the model thus constructs an underlying/winner/loser form triplet, which can be described in terms of the corresponding ERC, as exemplified in (5) above. Since there are only a finite number of ERCs corresponding to a finite number of constraints, the ERCs considered in a run of the model can be stacked one on top of the other into an *input ERC matrix*.

Without loss of generality, assume that each input ERC has a unique loser-preferring constraint. Next, let me make two crucial assumptions. First, assume that the constraint set contains a single faithfulness constraint F – plus of course a certain number m of markedness constraints $M_1, \ldots, M_m$. Second, assume that $M_1, \ldots, M_m$

Figure 1: First three stages of the learning dynamics where each diagonal ERC is fed persistently in turn



are either loser-preferring or even in the input ERCs, but never winner-preferring. The input ERC matrix thus is (a subset of) the matrix (15).

(15)
$$\begin{array}{r|c|ccc|} & F & M_1 & \dots & M_m \\ \text{ERC } 1 & \text{W} & \text{L} & & e \\ \vdots & | & & \ddots & \\ \text{ERC } m & \text{W} & e & & \text{L} \end{array}$$

The column corresponding to F consists of all W's. The entries corresponding to $M_1, \ldots, M_m$ are all equal to e's but for the diagonal of L's. This ERC matrix is thus called *diagonal*.

What is the maximum height that the constraint F can reach in a run of the EDRA model on the input diagonal ERC matrix (15)? To address this question, consider the following special run. To start, we persistently feed ERC 1 to the algorithm, until the markedness constraint M_1 is demoted underneath the faithfulness constraint F and that ERC cannot trigger any further update. Only at that point, we stop feeding ERC 1 to the algorithm, and persistently feed ERC 2 instead, again until it cannot trigger any further update. Only at that point, we stop feeding ERC 2 and persistently feed ERC 3. And so on.

Assume that the promotion amount has the shape (8) and suppose for concreteness that the calibration constant is $\sigma = 1$, so that the faithfulness constraint is promoted by $1/2$ with each update. The dynamics of the ranking values is depicted in Figure 1 for the first three learning stages. Throughout stage 1, it is ERC 1 that triggers updates, whereby the markedness constraint M_1 is demoted and the faithfulness constraint is promoted by $\frac{1}{3}\theta = \frac{1}{2+\sigma}\theta^{\text{init}}$, until the two constraints meet. Throughout stage 2, it is ERC 2 that triggers updates, whereby the markedness constraint M_2 is demoted and the faithfulness constraint is promoted by another $\frac{2}{9}\theta = \frac{1+\sigma}{(2+\sigma)^2}\theta^{\text{init}}$, until the two constraints meet. Throughout the generic kth stage, it is the kth ERC that triggers updates, whereby the markedness constraint M_k is demoted and the faithfulness constraint promoted by an amount that turns out to be equal to $\frac{(1+\sigma)^{k-1}}{(2+\sigma)^k}\theta^{\text{init}}$. The height θ_F reached by the faithfulness constraint at the end of the special run considered is thus $\sum_{k=1}^{m} \frac{(1+\sigma)^{k-1}}{(2+\sigma)^k}\theta^{\text{init}}$. It turns out that this is indeed the maximum height reacheable by the faithfulness constraint F on *any* run on the diagonal ERC matrix (15).

The $\mathcal{F}$-control condition (3) thus boils down to the inequality $\sum_{k=1}^{m} \frac{(1+\sigma)^{k-1}}{(2+\sigma)^k}\theta^{\text{init}} \leq \theta^{\text{init}} - m$. Assume that the m markedness constraints start out with the initial ranking value $\theta^{\text{init}} = m^k$. This inequality can then be solved analytically yielding $\sigma(m) = (1 - \exp\{(-k \log m)/m\})^{-1}$. By a first order Taylor expansion $exp(x) \sim 1 + x + o(x^2)$ of the exponential function, the latter expression can be approximated as in (16).

(16)
$$\sigma = \sigma(m) \sim \frac{m}{k \log m}$$

The latter bound for the calibration threshold is substantially smaller than the linear bound $\sigma(m) \sim m$ obtained through the elementary analysis of section 6. In particular, although (16) is not bounded as a function of m, its derivative goes to zero as $1/\log m$.

6 $\mathcal{F}$-controlling when the promotion amount decreases *slowly*

The preceding section has made two restrictive assumptions. First, that there is a unique $\mathcal{F}$ constraint. Second, that the $\mathcal{M}$ constraints are never winner-preferring. Under these assumptions, it has shown that the $\mathcal{F}$-control condition (3) holds when the calibration constant grows only very slowly with m, namely as in (16). Does this favorable result also hold when we relax the two restrictive assumptions? This section shows how to relax one of the two assumptions, namely the assumption that the $\mathcal{M}$ constraints cannot be winner-preferring. At this stage, I do know how to relax the other assumption that there is a unique $\mathcal{F}$ constraint. Again, the reasoning here is only sketched; see Magri (2014a) for details.

To illustrate the core idea, suppose that the EDRA model is trained on the input ERC matrix (17a) and walks through the run (18a). Here, I am assuming that the promotion amount p has the shape in (8), with the calibration constant $\sigma = 0$ set equal to zero for concreteness.

(17) a. $$\begin{array}{r} \\ \text{ERC 1} \\ \text{ERC 2} \end{array} \begin{array}{c} \begin{array}{ccc} F & M_1 & M_2 \end{array} \\ \left[\begin{array}{c|cc} \text{W} & \text{L} & e \\ \text{W} & \text{W} & \text{L} \end{array}\right] \end{array}$$

b. $$\begin{array}{r} \\ \text{ERC 1} \\ \text{ERC 2} \end{array} \begin{array}{c} \begin{array}{ccc} F & M_1 & M_2 \end{array} \\ \left[\begin{array}{c|cc} \text{W} & \text{L} & e \\ \text{W} & e & \text{L} \end{array}\right] \end{array}$$

(18) a. $$\begin{array}{r} F \\ M_1 \\ M_2 \end{array}\begin{bmatrix} 0 \\ 10 \\ 10 \end{bmatrix} \xrightarrow{\text{ERC 1}} \begin{bmatrix} 1 \\ 9 \\ 10 \end{bmatrix} \xrightarrow{\text{ERC 1}} \begin{bmatrix} 2 \\ 8 \\ 10 \end{bmatrix} \xrightarrow{\text{ERC 2}} \begin{bmatrix} 2.5 \\ 8.5 \\ 9 \end{bmatrix} \xrightarrow{\text{ERC 2}} \begin{bmatrix} 3 \\ 9 \\ 8 \end{bmatrix}$$

b. $$\begin{array}{r} F \\ M_1 \\ M_2 \end{array}\begin{bmatrix} 0 \\ 10 \\ 10 \end{bmatrix} \xrightarrow{\text{ERC 1}} \begin{bmatrix} 1 \\ 9 \\ 10 \end{bmatrix} \xrightarrow{\qquad\qquad\text{ERC 2}\qquad\qquad} \begin{bmatrix} 2 \\ 9 \\ 9 \end{bmatrix} \xrightarrow{\text{ERC 2}} \begin{bmatrix} 3 \\ 9 \\ 8 \end{bmatrix}$$

Consider the diagonal ERC matrix (17b) corresponding to $m = 2$ markedness constraints. The original run (18a) on the original ERC matrix (17a) can be simulated with the run (18b) on the diagonal ERC matrix (17b) in such a way that all constraints end up at the same high in the two runs.

This reasoning holds in complete generality. Indeed, under the assumption that there is a unique $\mathcal{F}$ constraint but no restrictions on the $\mathcal{M}$ constraints, the input ERC matrix looks like (19).

(19) $$\begin{array}{c} \begin{array}{cccc} F_1 & M_1 & \dots & M_m \end{array} \\ \left[\begin{array}{c|ccc} | & \ddots & & \iddots \\ \text{W} & & \text{L}, e, \text{W} & \\ | & \iddots & & \ddots \end{array}\right] \end{array}$$

Any run of the EDRA model on this input ERC matrix (19) can be mimicked by a corresponding run on the diagonal ERC matrix (15). This reduction to the diagonal case holds provided the promotion amount is calibrated, namely has the shape in (8), no matter the choice of the calibration constant $\sigma \geq 0$. This reduction fails if the promotion amount is not calibrated.

Another crucial condition needed for the reduction to the diagonal case is the following: in the original run, the markedness constraints are allowed to raise only slightly above their initial ranking value θ^{init}. Indeed, if a markedness constraint could raise arbitrarily high above its initial ranking value in the original run, there would be no way to mimic that increasing ranking dynamics with a derived run on the diagonal ERC matrix (15), as the latter only demotes but never promotes the markedness constraints. The fact that the markedness constraints can raise by a small amount does not threaten the reduction to the diagonal case, because the markedness constraints can be assigned a slightly larger initial ranking value in the derived run on the diagonal ERC matrix.

Fortunately, the markedness constraints $M_1, \dots, M_m$ indeed can raise above their initial ranking value θ^{init} only by a small amount, namely never by more than m, as stated in (20).

(20) $$\theta_1, \dots, \theta_m \leq \theta^{\text{init}} + m$$

Obviously, this bound (20) holds at the beginning of the run. It is thus sufficient to prove that this bound is an *invariant* of the algorithm: if it holds of the current ranking values at some time $t-1$, then it also holds at the subsequent time t. The challenge is that a winner-preferring markedness constraint M_1 sitting right at $\theta^{\text{init}} + m$ at time $t-1$ could in principle be promoted above that forbidden threshold, so that the bound (20) would hold at time $t-1$ but fail at time t. Yet, in order for such an update to happen, there has got to exist another constraint M_2 which is loser-preferring and is ranked at time $t-1$ at least as high as the winer-preferring constraint M_1. This means in turn that the sum $\theta_1^{t-1} + \theta_2^{t-1}$ of the two ranking values of M_1 and M_2 at time $t-1$ is at least $(\theta^{\text{init}} + m) + (\theta^{\text{init}} + m)$. This suggests to cope with the difficulty just highlighted by strengthening the invariant. Not only a *single* ranking value cannot get larger than $\theta^{\text{init}} + m$, but also the sum of any two ranking values can never reach $(\theta^{\text{init}} + m) + (\theta^{\text{init}} + m)$. For instance, let's say it can never get larger than $(\theta^{\text{init}} + m) + (\theta^{\text{init}} + m - 1)$. But now again, in order to prove that the latter bound on the sum of *two* ranking values holds at time t, I need an assumption about the sum of *three* ranking values at time $t-1$. And so on. Indeed, the sum $\theta_{i_1} + \ldots + \theta_{i_k}$ of the current ranking values of any number k of different markedness constraints $M_{i_1}, \dots, M_{i_k}$ can be bound as in (21). This bound holds for any promotion amount with the shape (8) corresponding to a calibration constant σ which is not too small, namely $\sigma \geq 1$.

(21) $$\sum_{h=1}^{k} \theta_{i_h} \;\leq\; \sum_{h=1}^{k} (\theta^{\text{init}} + m - h + 1)$$

For $k = 1$, (21) yields the desired bound (20).

Acknowledgments

This research was supported by a Marie Curie Intra European Fellowship within the 7th European

Community Framework Programme.

References

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

Paul Boersma. 2009. Some correct error-driven versions of the constraint demotion algorithm. *Linguistic Inquiry*, 40:667–686.

Kyle Gorman. 2013. *Generative phonotactics*. Ph.D. thesis, University of Pennsylvania.

Giorgio Magri. 2012. Convergence of error-driven ranking algorithms. *Phonology*, 29(2):213–269.

Giorgio Magri. 2013a. The complexity of learning in OT and its implications for the acquisition of phonotactics. *Linguistic Inquiry*, 44.3:433468.

Giorgio Magri. 2013b. An initial result on the restrictiveness of the error-driven ranking model of the early stage of the acquisition of phonotactics. In Hsin-Lun Huang, Ethan Poole, and Amanda Rysling, editors, *Proceedings of NELS 43: the 43rd annual meeting of the North East Linguistic Society*.

Giorgio Magri. 2014a. The error-driven ranking model of the acquisition of phonotactics: how to control the height of the faithfulness constraints. CNRS, UiL-OTS ms.

Giorgio Magri. 2014b. Error-driven versus batch models of the acquisition of phonotactics: David defeats Goliath. In John Kingston, Claire Moore-Cantwell, Joe Pater, and Robert Staubs, editors, *Supplemental Proceedings of Phonology 2013*, Washington DC. Linguistic Society of America.

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

Alan Prince and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in generative grammar*. Blackwell, Oxford. 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. Also available as ROA 537 version.

Alan Prince. 2002. Entailed ranking arguments. Ms., Rutgers University, New Brunswick, NJ. Rutgers Optimality Archive, ROA 500. Available at http://www.roa.rutgers.edu.

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

Bruce Tesar. 2013. *Output-Driven Phonology: Theory and Learning*. Cambridge Studies in Linguistics.

Igor Yanovich. 2012. The logic of OT rankings. MIT manuscript.