An online model of the acquisition of phonotactics within Optimality Theory

Giorgio Magri (magrigrg@gmail.com)
Institut Jean Nicod, École Normale Supérieure, 29 rue d'Ulm
75005 Paris, France

Abstract

Within the mainstream phonological framework of Optimality Theory (OT), grammars are parameterized by how they prioritize or rank a given set of constraints. OT online learning consists of slight re-rankings triggered by exposure to a single piece of data at the time. This paper presents a new online model for the acquisition of phonotactics in OT. Convergence and correctness are analytically investigated and the proposed model is shown to be superior to existing OT online models.

Keywords: Language Acquisition; Phonotactics; Optimality Theory; Online algorithms.

Description of the model

English speakers know that *blik* would be a possible word while *bnik* would not, despite the fact that both are unattested in the English lexicon. The knowledge of this distinction between licit vs. illicit sound combinations is called *phonotactics*. In carefully controlled experimental conditions, one-year olds already react differently to licit vs. illicit sound combinations. They thus display knowledge of phonotactics at an early stage, when other linguistic abilities (most notably *morphology*) are still lagging behind (Hayes, 2004). How is that possible? This paper tackles the problem of the acquisition of phonotactics from a computational perspective.

The acquisition of phonotactics is *gradual*: the target adult grammar is approached through a path of conservative intermediate stages. This gradualness is illustrated in (1) with some spontaneous productions of two children attempting to say *clock(s)*, from McLeod, Doorn, and Reed (2001).

(1)

2:3	**2:5**	**2:6**	**2:8**	**2:8**	**2:10**	**2:11**	**3:1**
tʌk	lʌk	dk	flʌk	kʌk	kəlʌ:k	klʌk	klʌk
tʌk	lʌk	dʌk	θlʌk	kʌk	kʌk	klʌk	klʌk
	flʌkθ	klʌkθ	θlʌk	kʌk	kəlʌk	klʌks	
	klʌkθ				kəlʌk	kʌk	

We see reduction of the target cluster /kl/ with sonority-driven preservation of the obstruent (/kl/ → [k]); we see reduction to the fronted obstruent (/kl/ → [t]); etcetera. We need a computationally sound model of the acquisition of phonotactics that is able to describe the observed gradualness.

Assume that each attempted phonological form (say, the cluster /kl/) comes with a preassigned set of *candidate* productions (say, the cluster [kl] itself, the two singleton consonants [k] and [l]; and variants thereof such as [f] or [t]). The relevant properties of a target phonological structure and a corresponding candidate are extracted by a set of phonological *constraints*, that measure how that pair deviates from the ideal along various dimensions. There are two types of constraints. *Markedness constraints* measure how much a candidate violates wellformedness conditions. For instance, the constraint *DORSAL assigns violations to dorsal consonants (i.e. it is violated by the candidates [kl] and [k]). *Faithfulness constraints* measure how much a candidate differs from the corresponding target. For instance, the constraint MAX assigns a violation for every deleted target segment (i.e. it is violated by the candidate [k] but not by [kl] for the target /kl/). Two or more constraints can conflict. For example, MAX prefers the candidate [kl] over [t] as the production of the target /kl/, while *DORSAL prefers [t] over [kl]. Grammars differ in how they prioritize or *rank* these constraints. And conflicts among constraints are resolved by a grammar in favor of the constraint it top ranks, in the sense that a phonological target is mapped to the (provably unique) *winner* candidate that satisfies condition (2) for any other *loser* candidate.

(2) Among those constraints that assign to the pair of the target and the winner a different number of violations than to the pair of the target and the loser, the top ranked one assigns less violations to the former.

These ideas are formalized in the mainstream phonological framework of *Optimality Theory* (OT) developed by Prince and Smolensky (2004), that I assume in this paper. This framework ties up well with the recent bloom of interest for models based on orders and rankings in Machine Learning (see the 2009 NIPS Workshop on Learning with rankings).

At this early stage of research on the acquisition of phonotactics, it makes sense to keep the learning problem as simple as possible. I thus assume that the set of constraints is universal, shared by both developing children and adults and thus needs not be learned. For instance, the constraint *DORSAL is motivated both by the process of *fronting* in child phonology (/k/ → [t]) as well as by languages that lack velars entirely (e.g. Tahitian). The constraint is available also to English speakers, but low ranked. The typology of adult phonotactics and the typology of child intermediate stages thus coincide, consisting of the collection of all possible rankings.

The gradualness illustrated in (1) suggests a model whereby the learner entertains a current hypothesis of the target phonotactics that gets updated over time based on exposure to phonotactically licit adult forms, describing a path within the space of possible phonotactics. This intuition can be formalized as follows. At every time t, the model maintains a current ranking, that represents its current hypothesis. This current ranking is represented as a numerical vector $\theta^t = (\theta^t_1, \ldots, \theta^t_n)$, where n is the number of constraints and θ^t_k is the *ranking value* of constraint C_k at time t. Constraint C_h is ranked above constraint C_k at time t iff $\theta^t_h > \theta^t_k$. The initial ranking vector top ranks the markedness constraints, and thus corresponds to a smallest language. The current ranking

vector is updated over time as follows. The model receives a piece of data from the target adult phonotactics, say the word *clock*. It thus infers that the target /kl/ should be faithfully mapped to the winner candidate [kl]; see (Hayes, 2004) for discussion. The model then picks (at random or according to some refined procedure) a non faithful candidate, say [t]. If the current ranking prefers the loser unfaithful candidate [t] over the faithful winner candidate [kl] according to (2), then the model updates its current ranking vector according to the general scheme (3). For instance, if the target is /kl/ and the current grammar prefers the unfaithful production [t] over [kl], then the algorithm might demote *DORSAL (that prefers the loser) and promote MAX (that prefers the winner).

(3) a. Decrease the ranking value of *loser-preferring constraints* (LPCs), i.e. constraints that prefer the unfaithful loser candidate over the faithful winner one;
b. increase the ranking value of *winner-preferring constraints* (WPCs), i.e.constraints that prefer the faithful winner candidate over the unfaithful loser one.

What is relevant about loser and winner candidates for a given target is which constraints are LPCs and which WPCs. The relevant information can thus be summarized with a row of W's, L's and E's corresponding to WPCs, LPCs and *even* constraints (that assign the same number of violations to winner and loser), called an *Elementary Ranking Condition* (ERC).

(4) $\begin{pmatrix} \text{target,} \\ \text{winner,} \\ \text{loser} \end{pmatrix} \Rightarrow \begin{array}{c} \begin{array}{ccccccc} \ldots & C_h & \ldots & C_k & \ldots & C_\ell & \ldots \end{array} \\ \left[\begin{array}{ccccccc} \ldots & \text{W} & \ldots & \text{L} & \ldots & \text{E} & \ldots \end{array} \right] \end{array}$

Algorithms that fit into this broad scheme are called *OT online models*. They are very simple and widely assumed, thus deserving a close investigation as the null hypothesis.

The core ingredient in the definition of OT online models are the details of the re-ranking rule (3) used to update the current ranking vector. Two main options have been considered in the literature. One option is (5), due to Tesar and Smolensky (1998). A LPC is *undominated* if it is ranked "too high", namely above all WPCs. This update rule (5) demotes (undominated) LPCs but does not promote WPCs, whose ranking values are not updated. The resulting OT online algorithm is called (gradual) *Constraint Demotion* (CD).

(5) a. Demote each *undominated* LPC by 1;
b. but do nothing to the WPCs.

Boersma (1997) argues (withinin a framework slightly different from the one considered here) that promotion is needed, and thus considers the update rule (6). The resulting OT online model is called *Gradual Learning Algorithm* (GLA).

(6) a. Demote each (undominated?) LPCs by 1;
b. and promote each WPC by 1.

This paper argues that neither re-ranking rule (5) or (6) yields a proper online model of the acquisition of phonotactics, and defends a new update rule. To introduce the idea, let me distinguish two cases, depending on whether the current ERC contains a unique W, as in (7a); or else multiple W's, say two as in (7b). The former case (7a) is simple: we know that the unique WPC must in the end be ranked above the LPCs, irrespectively of the rest of the data. The case (7b) is instead delicate: we don't know which of the two WPCs needs in the end to be ranked above the LPCs, as one of them might have to be ranked low, depending on the rest of the data.

(7) $\begin{array}{cc} & \begin{array}{ccccccc} \ldots & C_h & \ldots & C_k & \ldots & C_\ell & \ldots \end{array} \\ \text{a.} & \left[\begin{array}{ccccccc} \ldots & \ldots & \ldots & \text{W} & \ldots & \text{L} & \ldots \end{array} \right] \\ \text{b.} & \left[\begin{array}{ccccccc} \ldots & \text{W} & \ldots & \text{W} & \ldots & \text{L} & \ldots \end{array} \right] \end{array}$

According to Boersma's re-ranking rule (6), WPCs get promoted by 1, both in the case of a simple ERC (7a) with a unique W and in the case of a challenging ERC (7b) with multiple W's. This does not look like a good idea though, as it does not capture the crucial difference between the two cases. Here is a more principled alternative. In the simple case (7a), the unique WPC can be confidently promoted by the same amount LPCs are demoted, say 1. But in the challenging case (7b), we should be *cautious* and split our confidence between the two WPCs, promoting each one just by $1/2$. As uncertainty scales with the total number w of WPCs, each WPC should be promoted just by $1/w$ in the general case, as in the new *cautious* promotion/demotion re-ranking rule (8).

(8) a. Demote each *undominated* LPC by 1;
b. promote each WPC by $1/w$.

Under the plausible conjecture that actual language learning strategies employed by humans have been selected by evolution because of their computational efficiency and soundness, computational considerations gain currency within cognitive science. From a computational perspective, there are two basic desiderata on sound OT online models of the acquisition of phonotactics. One is *convergence*: the model needs to eventually entertain a hypothesis consistent with the target adult phonotactics, so that only a finite number of updates are performed. A convergent online model is *correct* provided the corresponding final grammar entertained at convergence is not only consistent with the target adult language but also restrictive enough to capture the target phonotactics. This paper argues that the new re-ranking rule (8) is computationally superior to the existing rules (5) and (6). I show that the corresponding OT online model is convergent, contrary to the case of (6). Furthermore, I consider a very simple OT model for *segmental phonotactics*, and I sketch an argument that the OT online model with the new re-ranking rule (8) is always correct, contrary to the case of (5) and (6).

Convergence

Tesar and Smolensky (1998) proved convergence for their *demotion-only* re-ranking rule (5). But convergence for Boersma's *promotion/demotion* re-ranking rule (6) has remained an open issue, until Pater (2008) provided a counterexample. It is thus currently an open problem whether convergent constraint promotion is possible at all. Theorem 1 settles the issue with a positive answer. A sketch of the proof follows.

Theorem 1 *The OT online model with the new promotion/demotion reranking rule (8) converges (provided the data fed to the model are consistent with some OT grammar).* ■

Tesar and Smolensky show that the current ranking values in the case of their demotion-only re-ranking rule (5) are always larger than a constant (that depends on the number n of constraints). In fact, as constraints are only demoted when needed (i.e. when *undominated*), they cannot be demoted too much. A careful look at their proof reveals that lower boundness of the current ranking values extends to any promotion/demotion update rule that only demotes undominated LPCs. In particular, the following fact thus holds true.

Fact 1 *The current ranking values entertained by the OT online model with the new promotion/demotion re-ranking rule (8) cannot get* smaller *than a constant (provided the data fed to the model are consistent with some OT grammar).* ■

Having established that the current ranking values cannot get too *small*, we now ask whether they can get too *large*. This is precisely what happens when Boersma's update rule (6) is run on Pater's counterexample: the ranking values increase indefinitely. It turns out that that cannot happen with the new update rule (8). The reason is as follows. As we never promote more than we demote, the sum of the promotion amounts and the demotion amounts is always negative or null. Thus, the sum of the current ranking values at any time is always equal to or smaller than the sum of the initial ranking values. As the single ranking values cannot become too small (by Fact 1) and as their sum cannot get too large, then the single ranking values cannot become too large either.

Fact 2 *The current ranking values entertained by the OT online model with the new promotion/demotion re-ranking rule (8) cannot get* larger *than a constant (provided the data fed to the model are consistent with some OT grammar).* ■

The sequence of ranking vectors entertained by the OT online model with a demotion-only update rule such as (5) cannot have the shape (9), whereby the same ranking vector θ is entertained twice but with some other ranking vector $\theta' \neq \theta$ entertained in between. In fact, (9) would require some ranking value to first decrease and then increase back to its original value, which is impossible if only demotion is performed.

$$(9) \quad \ldots \longrightarrow \theta \longrightarrow \ldots \longrightarrow \theta' \longrightarrow \ldots \longrightarrow \theta \longrightarrow \ldots$$

Thus, demotion-only OT online models never loop back to a ranking vector that had been previously deemed unsuitable and thus updated. Fact 3 ensures that this property extends to promotion-demotion update rules. The proof (postponed to the Appendix) rests on the following fact: the hypothesis that the data be *OT-consistent* entails that the vectors of promotion and demotion amounts are *conically independent*.

Fact 3 *The OT online model with the new promotion/demotion re-ranking rule (8) cannot loop back to a ranking vector previously dismissed (provided the data fed to the model are consistent with some OT grammar).* ■

The convergence theorem 1 now follows straightforwardly. Fact 1 guarantees that the current ranking values cannot get too small, namely cannot live in the shaded region in Fig.1a. And Fact 2 guarantees that the current ranking values cannot get too large either, namely cannot live in the shaded region in Fig.1b. Taken together, Facts 1 and 2 thus guarantee that the current ranking values must live in a bounded region, namely in the non-shaded region of Fig.1c. Furthermore, the algorithm can only entertain ranking vectors in a lattice, namely the dots in Fig.1d. Thus, the search space of the algorithm is finite, as there is only a finite number of points in a bounded lattice. As the algorithm cannot loop by Fact 3, finiteness of the search space entails convergence, namely ensures that the algorithm can only perform a finite number of updates.

Correctness

A crucial component of the acquisition of the target adult language is the acquisition of its *segmental phonotactics*, i.e. of the inventory of licit segments and of their licit concatenations. Some elementary examples of OT typologies for segmental phonotactics are provided in (10) and (11).

$$(10) \quad \text{a.} \ \{ \ \text{t} \quad \text{d} \quad \text{t}^h \quad \text{d}^h \ \}$$

$$\text{b.} \ \left\{ \begin{array}{cc} F_1 = \textsc{Ident}[\textsc{voice}] & F_2 = \textsc{Ident}[\textsc{asp}] \\ M_1 = *[+\textsc{voice}] & M_2 = *[+\textsc{asp}] \\ M_{1,2} = *[+\textsc{voice}, +\textsc{asp}] & \end{array} \right\}$$

$$(11) \quad \text{a.} \ \{ \ \text{ps} \quad \text{bs} \quad \text{pz} \quad \text{bz} \ \}$$

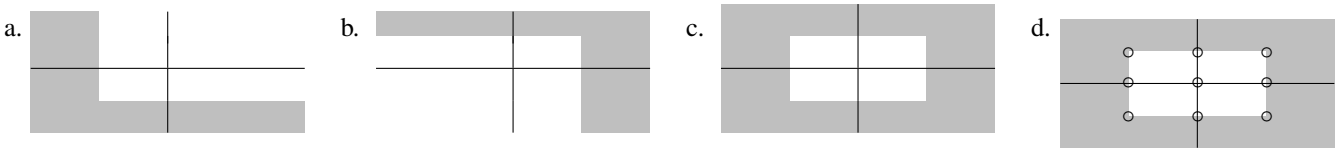
$$\text{b.} \ \left\{ \begin{array}{cc} F_1 = \textsc{Idnt}[\textsc{fric-voi}] & F_2 = \textsc{Idnt}[\textsc{stp-voi}] \\ M_1 = *[+\textsc{fric-voi}] & M_2 = *[+\textsc{stp-voi}] \\ M_{1,2} = \textsc{Agree}[\textsc{stp-voi}, \textsc{fric-voi}] & \end{array} \right\}$$

The set of forms (10a) consists of obstruents described by the features VOICE and ASPIRATION. The set of forms (11a) consists of two adjacent obstruents, described by the features STOP-VOICING and FRICATIVE-VOICING. The constraint sets (10b) and (11b) contain identity faithfulness constraints F_1, F_2 for the two features; markedness constraints M_1, M_2 that punish the marked value of the two features; and a markedness constraint $M_{1,2}$ that punishes certain marked combinations of values of the two features.

I now sketch a formal OT framework for segmental phonotactics that generalizes examples such as (10)-(11). The construction starts with N partial binary phonological features $\varphi_1, \ldots, \varphi_i, \ldots, \varphi_N$, such as VOICE or ASPIRATION in (10). Each feature φ_i takes a phonological form and returns the value 0 or 1, or else # in case it is undefined. Segmental phonology is feature-based: a segment can be identified with the corresponding N-tuple $\langle x_1, \ldots, x_i, \ldots, x_N \rangle$ of feature values $x_i \in \{0, 1, \#\}$. The set of segments is thus defined as some set of such N-tuples, as in (12). As it is usual in phonotactics, I assume no distinction between underlying and surface forms. Nonetheless, I will use the symbol $\mathbf{x}$ (or $\mathbf{y}$) when a form is construed as an underlying (or surface) form.

$$(12) \quad \begin{array}{c}\text{set of under-} \\ \text{lying forms}\end{array} = \begin{array}{c}\text{set of sur-} \\ \text{face forms}\end{array} \subseteq \{0, 1, \#\}^N$$

Figure 1: Sketch of the proof of the convergence theorem 1 in the case with $n = 2$ constraints



The set of candidates corresponding to an underlying form $\mathbf{x}$ is the set of all forms defined for the same features that $\mathbf{x}$ is defined for. Finally, the constraint set can contain three types of constraints, listed in (13). The *faithfulness constraint* F_i corresponding to feature φ_i is violated by an underlying form and a candidate that differ w.r.t. feature φ_i. The *(simple) markedness constraint* M_i corresponding to feature φ_i is violated by a form that has the marked value for feature φ_i. I assume w.l.g. that the marked value is always 1. Finally, the *binary markedness constraint* (BMC) $M^{\mu}_{i,j}$ corresponding to features φ_i and φ_j and a *markedness pattern* μ is violated by a form whose pair of values for features φ_i, φ_j belongs to the designated set $\mu \subseteq \{0,1\} \times \{0,1\}$ of marked combinations of feature values.

(13) $F_i(\mathbf{x},\mathbf{y}) = 1 \iff x_i \neq y_i$
$M_i(\mathbf{y}) = 1 \iff y_i = 1 =$ the marked value
$M^{\mu}_{i,j}(\mathbf{y}) = 1 \iff \langle y_i, y_j \rangle \in \mu =$ set of marked feature value combinations

There are sixteen possible markedness patterns μ and thus as many BMCs. Here are some examples: the markedness pattern (14a) corresponds to the "doubly marked" constraint $M_{1,2}$ in (10b); the markedness pattern (14b) corresponds to the "agreement" constraint $M_{1,2}$ in (11b); the complementary pattern (14c) corresponds to an "OCP" constraint; and so on.

(14) a. $\mu = \{\langle 1,1 \rangle\}$ doubly marked constraint
b. $\mu = \{\langle 0,1 \rangle, \langle 1,0 \rangle\}$ agreement constraint
c. $\mu = \{\langle 0,0 \rangle, \langle 1,1 \rangle\}$ OCP constraint

BMCs are important because they model *feature interaction*. As the learning complexity intuitively depends on feature interaction, we are led to the following question: which restrictive assumptions on feature interaction guarantee correctness of OT online models? I assume that the amount of feature interaction is *limited*. The case with no feature interaction is trivial. The simplest non-trivial case is thus that each feature interacts with at most another feature, so that (15) holds.[1]

(15) The constraint set does not contain any two BMCs that both target the same feature.

I also assume that the mode of feature interaction is *phonologically plausible*, in the sense that the constraint set contains no BMCs with a phonologically implausible markedness pattern, namely a markedness pattern that has cardinality 3, as in (16a);[2] or that only punishes forms unmarked w.r.t. the two features targeted by the BMC, as in (16b).

[1] In fact, if there were two BMCs $M_{i,j}$ and $M_{i,k}$ targeting the same feature φ_i, then φ_i would interact with two features φ_j and φ_k.

[2] The markedness pattern of cardinality 4 yields a trivial BMC.

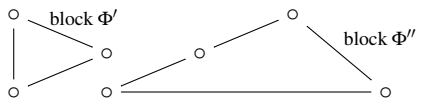
(16) a. ~~$\mu = \{\langle 1,1 \rangle, \langle 0,1 \rangle, \langle 1,0 \rangle\}$~~, ~~$\mu = \{\langle 0,0 \rangle, \langle 0,1 \rangle, \langle 1,0 \rangle\}$~~
~~$\mu = \{\langle 0,0 \rangle, \langle 1,0 \rangle, \langle 1,1 \rangle\}$~~, ~~$\mu = \{\langle 0,0 \rangle, \langle 0,1 \rangle, \langle 1,1 \rangle\}$~~
b. ~~$\mu = \{\langle 0,0 \rangle\}$~~.

The following result starts the investigation of correctness of OT online models of the acquisition of phonotactics. The proof is only sketched here; see (Magri, 2011) for details and for extensions beyond the overly restrictive assumption (15).

Theorem 2 *Consider an OT typology (12)-(13) corresponding to N features. Assume that feature interaction is* limited *according to (15) and* phonologically plausible *according to (16).*[3] *Then the OT online model with the new re-ranking rule (8) is correct on any language and for any sequence of data fed to the algorithm. On the contrary, for both re-ranking rules (5) and (6), there exist languages for which the model is incorrect for any sequence of data fed to the algorithm.* ■

Draw a circle for every feature and an edge between any two features that interact through a BMC. Suppose that the resulting graph looks like (17): the features can be split into two disjoint sets Φ' and Φ'' with no interactions between them.

(17)



We intuitively expect that the "difficult" problem of correctness of the OT online model on the "large" original typology with N features can be *reduced* to the two "simpler" problems of correctness on the two "smaller" typologies corresponding to the two sets of features Φ', Φ''. This intuition can indeed be formalized and shown to hold true.

By assumption (15) that feature interaction is limited, each feature interacts with at most another feature and the feature interaction graph (17) thus consists of connected components of cardinality at most 2. I thus need to prove Theorem 2 only in the case with $N = 2$ features. To this end, let's sort the languages into three types, based on what needs to be ranked above the faithfulness constraints. To start, let languages of *type I* be those languages that do not require any constraint to be ranked above the faithfulness constraints. It is easy to check that any OT online model is correct on such languages.

Next, let languages of *type II* be those languages that require markedness constraints to be ranked above faithfulness constraints, but do not require the two faithfulness constraints to be ranked relative to each other. To illustrate, suppose that the BMC corresponds to the OCP markedness pattern (14c).

[3] A further technical assumption on the set of candidates is needed, omitted here for space; see (Magri, 2011) for details.

The corresponding typology contains language (18a), which is is of type II, as it corresponds to the ranking (18b). The complete set of ERCs is (18c).[4]

(18) a. $L = \left\{ \begin{matrix} \langle 0,1 \rangle \\ \langle 1,0 \rangle \end{matrix} \right\}$ b. $\begin{matrix} & M_{1,2} & \\ & | & \\ F_1 & & F_2 \\ | & & | \\ M_1 & & M_2 \end{matrix}$ c.

F_1	F_2	M_1	M_2	$M_{1,2}$
W		L		W
	W		W	W
W	W	L	W	
W		W		W
	W		L	W
W	W	W	L	

A learning path using Boersma's update rule (6) is (19).[5] The final ranking vector incorrectly ranks F_1 above $M_{1,2}$, so that the model has failed to learn the target ranking (18b). It can be shown that the model fails for any possible learning path.

(19)

F_1	0		1		2		3		4		5	6
F_2	0		1		2		3		4		5	6
M_1	5	$\overset{3}{\rightarrow}$	4	$\overset{6}{\rightarrow}$	5	$\overset{3}{\rightarrow}$	4	$\overset{6}{\rightarrow}$	5	$\overset{3}{\rightarrow}$	4	$\overset{6}{\rightarrow}$ 5
M_2	5		6		5		6		5		6	5
$M_{1,2}$	5		5		5		5		5		5	5

Crucially, it is impossible for F_1 to get incorrectly ranked above $M_{1,2}$ in the case of the new re-ranking rule (8). Suppose by contradiction that it did. Suppose $M_{1,2}$, M_1 and M_2 start out at 5 and F_1 and F_2 at 0, as in (19). As $M_{1,2}$ is never a LPC, its ranking value cannot decrease. In order for F_1 to make it above $M_{1,2}$, its ranking value must thus increase to at least 5. The last update that brings F_1 that high requires one of M_1 or M_2 (call it M_i) to be a LPC and to be ranked even higher. Recall that the sum of the ranking values cannot increase over time. As the sum of the ranking values of $M_{1,2}$, F_1 and M_i is (almost) equal to the sum of the initial ranking values, the sum of the ranking values of the two remaining constraints F_2 and M_j must be (almost) smaller than zero. And this is easily shown to be impossible. A formalization of this heuristic reasoning shows that the OT online model with the new re-ranking rule (8) is always correct on languages of type II.

Finally, let languages of *type III* be the remaining languages, i.e. those that require the faithfulness constraints to be ranked relative to each other. To illustrate, suppose that the BMC corresponds to the Agree markedness pattern (14b). The corresponding typology contains the language (20a). This language is of type III, as it corresponds to ranking (20b). The corresponding set of ERCs is (20c).[6]

(20) a. $L = \left\{ \begin{matrix} \langle 1,1 \rangle \\ \langle 0,0 \rangle \end{matrix} \right\}$ b. $\begin{matrix} & F_2 & & M_{1,2} \\ M_2 & \diagup & \diagup & \\ & M_1 & & \\ & | & & \\ & F_1 & & \end{matrix}$ c.

F_1	F_2	M_1	M_2	$M_{1,2}$
W		L		W
	W		L	
W	W	L	L	

Tesar and Smolensky's demotion-only re-ranking rule (5) is never correct on this language: as the faithfulness constraints F_1 and F_2 are never LPCs, they are never re-ranked by demotion-only; there is thus no way that a demotion-only re-ranking rule can rank one of them on top of the other. Constraint promotion is needed in order to move around F_1 and F_2 too. But will an OT online model that performs promotion too be able to converge to the correct relative ranking of F_2 above F_1? It can be shown that the first ERC in (20c) can trigger at most one update, as it has a W corresponding to the BMC $M_{1,2}$ whose column does not have any L's. Thus, updates are triggered just by the 2nd and 3rd ERCs. As the former only promotes F_2 while the latter promotes both F_1 and F_2, F_2 will raise faster and thus be ranked above F_1 throughout learning, thus allowing a model that performs constraint promotion to converge to the correct final ranking (20b). In the case with $N = 2$ features, it can be shown that any language of type III that requires a faithfulness constraint to be ranked above the other has more ERCs that push the former over the latter, thus allowing promotion/demotion re-ranking rules to converge to the correct ranking for any learning path.

Appendix: proof of Fact 3

For the sake of clarity, assume that the ERCs fed to the model contain a unique LPC and that the initial ranking vector is the null vector; the reasoning below trivially extends to the general case. The contribution of an input ERC $\mathbf{a}$ to the current ranking vector according to the update rule (8) can be summarized by pairing it up with the corresponding *update vector* $\overline{\mathbf{a}}$ in (21): the entry corresponding to the LPC is -1; entries corresponding to WPCs are $1/w$, where w is the total number of WPCs in the ERC $\mathbf{a}$; all other entries are 0. The total number of possible input ERCs is always finite; call it m. Let $\overline{\mathbf{a}}_i$ be the update vector corresponding to the ith ERC.

(21) $\mathbf{a} = [a_1, \ldots, a_n] \rightarrow \overline{\mathbf{a}} = \begin{bmatrix} \overline{a}_1 \\ \vdots \\ \overline{a}_n \end{bmatrix}$ with $\overline{a}_k = \begin{cases} \frac{1}{w} & \text{if } a_k = \text{W} \\ -1 & \text{if } a_k = \text{L} \\ 0 & \text{otherwise} \end{cases}$

As updates consist of adding update vectors, the current ranking vector θ^t entertained by the model at some time t can be described as a combination (22) of the update vectors $\overline{\mathbf{a}}_1, \ldots, \overline{\mathbf{a}}_m$, each multiplied by the number of updates α_i^t triggered by the corresponding ith input ERC up to time t.

(22) $\theta^t = \alpha_1^t \overline{\mathbf{a}}_1 + \ldots + \alpha_i^t \overline{\mathbf{a}}_i + \ldots + \alpha_m^t \overline{\mathbf{a}}_m$

As the coefficients α_i^t are by definition non-negative, (22) says that the current ranking vector is a *conic combination* of the update vectors. We thus need to study the *conic geometry* of the update vectors, namely the properties of their *conic combinations*. Here is an important conic property: the update vectors $\overline{\mathbf{a}}_1, \ldots, \overline{\mathbf{a}}_m$ are called *conically independent* provided that there are no coefficients $\alpha_1, \ldots, \alpha_m$ that satisfy the conditions in (23): it is impossible to synthesize the null vector as a conic combination of the update vectors, unless the non-negative coefficients are all null.

[4] The first three ERCs in (18c) correspond to the underlying form $\mathbf{x} = \langle 0,1 \rangle$, the faithful winner-form $\mathbf{y} = \langle 0,1 \rangle$ and the three unfaithful candidate losers $\langle 1,1 \rangle$, $\langle 0,0 \rangle$ and $\langle 1,0 \rangle$. The bottom three rows are obtained analogously from the underlying/winner form $\langle 1,0 \rangle$.

[5] Numbers above arrows specifie the row triggering the update.

[6] These are the ERCs corresponding to the winner/underlying form $\langle 1,1 \rangle$, constructed as in footnote 4. The ERCs corresponding to $\langle 0,0 \rangle$ are omitted for space, as they do not contain any L.

(23) a. $\alpha_1 \bar{\mathbf{a}}_1 + \ldots \alpha_m \bar{\mathbf{a}}_m = \mathbf{0}$;
b. $\alpha_i \geq 0$ for all $i = 1, \ldots, m$;
c. $\alpha_i \neq 0$ for some $i = 1, \ldots, m$.

Fact 4 guarantees conic independence of the update vectors. And Fact 5 in turn says that conic independence entails that the algorithm cannot loop. Fact 3 follows straightforwardly.

Fact 4 *The update vectors (21) corresponding to ERCs consistent with some OT grammar are conically independent.* ■

Proof. A set of ERCs OT-consistent with the ranking $C_1 \gg \ldots \gg C_n$ can be stacked as in (24): there is a top block of ERCs that start with a W; followed by a second block of ERCs that start with an E followed by a W; etcetera, until a final dth block of ERCs that start with $d-1$ E's followed by a W.

(24)

	C_1	C_2	...	C_{d-1}	C_d	...		C_n
1st block	W							
	W	...	...	...	...	...	...	...
2nd block	E	W						
	E	W	...	...	...	...	...	...
⋮			⋱	...	...	...	...	...
dth block	E	E	—	E	W			
	E	E	—	E	W	...	...	...

As the mapping (21) from ERCs into update vectors replaces W's with $1/w$ and E's with 0, the update vectors corresponding to the stack of ERCs (24) look like (25).

(25) $\underbrace{\begin{bmatrix} \frac{1}{w} \\ \vdots \\ \frac{1}{w} \end{bmatrix}, \ldots, \begin{bmatrix} \frac{1}{w} \\ \vdots \\ \frac{1}{w} \end{bmatrix}}_{\text{1st block}}, \underbrace{\begin{bmatrix} 0 \\ \frac{1}{w} \\ \vdots \\ \frac{1}{w} \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ \frac{1}{w} \\ \vdots \\ \frac{1}{w} \end{bmatrix}}_{\text{2nd block}}, \ldots, \underbrace{\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{1}{w} \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{1}{w} \end{bmatrix}}_{\text{dth block}}$

Suppose that a conic combination of the vectors (25) yields the null vector, namely that conditions (23a) and (23b) hold. The first component of the update vectors corresponding to the 1st block is positive while the first component of the remaining update vectors is null. In order for the first components to sum to zero, the nonnegative coefficients α_i that multiply the vectors corresponding to the 1st block must be all null. As their coefficients are null, the vectors corresponding to the 1st block can be ignored. By reasoning analogously for the second components, I conclude that also the coefficients α_i that multiply the vectors corresponding to the 2nd block are all null. By repeating this reasoning d times, I conclude that all the coefficients α_i in the combination are null, contradicting condition (23c). □

Fact 5 *If the update vectors are conically independent, then the OT online model can never loop back to a current ranking vector that it had previously dismissed.* ■

Proof. Suppose by contradiction that the algorithm loops back at time t' to a ranking vector that it had dismissed at a previous time t, as stated in (26).

(26) a. The ranking vectors at two times t and t' coincide;
b. time t precedes time t';
c. a different ranking vector is entertained at some time in between t and t'.

Assumption (26a) that the ranking vectors θ^t and $\theta^{t'}$ entertained at times t and t' coincide, can be expressed as the identity (27a), using the general characterization (22) of the current ranking vector. Assumption (26b) that time t' follows time t entails that the number of updates $\alpha_i^{t'}$ triggered by the ith ERC up to time t' is larger than or equal to the number of updates α_i^t triggered up to time t, as stated in (27b). Finally, assumption (26c) entails that some update has happened at some time in between t and t', so that at least one of the coefficients has increased from time t to time t', as stated in (27c).

(27) a. $\alpha_1^t \bar{\mathbf{a}}_1 + \ldots + \alpha_m^t \bar{\mathbf{a}}_m = \alpha_1^{t'} \bar{\mathbf{a}}_1 + \ldots + \alpha_m^{t'} \bar{\mathbf{a}}_m$
b. $\alpha_i^{t'} \geq \alpha_i^t$ for all $i = 1, \ldots, m$
c. $\alpha_i^{t'} \neq \alpha_i^t$ for some $i = 1, \ldots, m$

Introducing the coefficients $\alpha_i \stackrel{\text{def.}}{=} \alpha_i^{t'} - \alpha_i^t$, conditions (27) can be rewritten as in (28), which contradict the hypothesis that the update vectors $\bar{\mathbf{a}}_1, \ldots, \bar{\mathbf{a}}_m$ are conically independent.

(28) a. $\alpha_1 \bar{\mathbf{a}}_1 + \ldots + \alpha_m \bar{\mathbf{a}}_m = \mathbf{0}$
b. $\alpha_i \geq 0$ for all $i = 1, \ldots, m$
c. $\alpha_i \neq 0$ for some $i = 1, \ldots, m$

Acknowledgments

I wish to thank A. Albright. This work was supported in part by a 'Euryi' grant from the ESF to P. Schlenker.

References

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

Hayes, B. (2004). "Phonological Acquisition in Optimality Theory: The Early Stages". In R. Kager, J. Pater, & W. Zonneveld (Eds.), *Constraints in Phonological Acquisition* (pp. 158–203). Cambridge University Press.

Magri, G. (2011). *"A computational investigation of OT online models of the early stage of the acquisition of phonotactics. Part 2: correctness"*. (Manuscript, IJN, ENS.)

McLeod, S., Doorn, J. van, & Reed, V. (2001). "Normal acquisition of consonant clusters". *American Journal of Speech-Language Pathology*, *10*, 99–110.

Pater, J. (2008). "Gradual Learning and Convergence". *Linguistic Inquiry*, *39.2*, 334–345.

Prince, A., & Smolensky, P. (2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell.

Tesar, B., & Smolensky, P. (1998). "Learnability in Optimality Theory". *Linguistic Inquiry*, *29*, 229–268.