Statistical learning theory and linguistic typology: a learnability perspective on OT's strict domination

Émile Enguehard
ENS
emile.enguehard@ens.fr

Edward Flemming
MIT
flemming@mit.edu

Giorgio Magri
CNRS
magrigrg@gmail.com

Abstract

This paper develops a learnability argument for strict domination by looking at the generalization error of learners trained on OT and HG target grammars. The argument is based on both a review of error bounds in the recent statistical learning literature and simulation results on realistic phonological test cases.

1 Introduction

According to **Optimality Theory** (OT; Prince and Smolensky 2004), constraint interaction in natural language phonology is severely constrained by the hypothesis of **strict domination**. According to this hypothesis, "the constraints [are] arranged in a hierarchy" and "each constraint is strictly more important than — takes absolute priority over — all the constraints lower-ranked in the hierarchy. [...] Strict domination thus limits drastically the range of possible strength-interactions between constraints to those representable with the algebra of total order" (Prince and Smolensky, 1997). This hypothesis of strict domination has been challenged in the recent phonological literature (Pater, 2009; Potts et al., 2010; Pater, 2016), which has therefore started to explore an implementation of constraint-based phonology which does away with strict domination, known as **Harmonic Grammar** (HG; Legendre *et al.*, 1990a,b; Smolensky and Legendre, 2006). Section 2 re-assesses the OT versus HG debate, concluding that HG over-generates for many natural constraint sets and that natural language phonology thus supports OT's hypothesis of strict domination.

Why should constraint interaction in natural language phonology display strict domination? Legendre *et al.* (2006) conjecture that "demands of learnability [might] provide a pressure for strict domination among constraints" although they admit that "it remains an open problem to formally characterize exactly what is essential about strict domination to guarantee efficient learning." Riggle *et al.* (2009; 2010) take a closer look at this alleged connection between strict domination and learnability. They look at **error bounds** in terms of a classical measure of the learning complexity of a hypothesis class, namely its **Vapnik-Chervonenkis** (VC) dimension (Vapnik and Chervonenkis, 1971). But they find that the VC dimension is the same for OT and HG, despite OT typologies being smaller than HG typologies because of strict domination. They conclude that, "though there may be factors that favor one model [OT or HG] over the other, the complexity of learning [...] is not one of them."

Yet, VC dimension is an old measure of learning complexity (it dates back to the seventies) which is inevitably coarse as it applies to completely arbitrary classifiers. Since Schapire *et al.* (1998), statistical learning theory has instead focused on a special class of classifiers, namely **voting classifiers** which aggregate the "votes" of more basic classifiers scaled through corresponding weights. For this special class of classifiers, better error bounds have been developed, which take into account the **margin** of "confidence" with which a classifier succeeds on the data. More recently, Koltchinskii and Panchenko (Koltchinskii and Panchenko, 2002; Koltchinskii et al., 2003b; Koltchinskii et al., 2003a; Koltchinskii

Proceedings of the Society for Computation in Linguistics (SCiL) 2018, pages 1-11.
Salt Lake City, Utah, January 4-7, 2018

and Panchenko, 2005) have further refined margin theory through error bounds which depend not only on the margin but also on the rate of decay of the weights of the basic classifiers: the bounds get better (that is provide guarantees for a smaller generalization error) as the rate of decay increases.

Crucially, HG and OT grammars can be construed as voting classifiers with the phonological constraints playing the role of the basic classifiers. Section 3 thus brings Koltchinskii and Panchenko's result to bear on the debate between HG and OT, through the well known characterization of OT as a special case of HG with weights decreasing fast, specifically exponentially.

Section 4 complements these theoretical results with simulation-based estimates of the generalization error (codes and data are provided as online supplements). We look at two test cases related to vowel harmony and syllable types. We compute the corresponding typologies of OT grammars and HG-non-OT grammars (namely HG grammars with no OT correspondent). For both types of target grammars, we compute the generalization error of the hypothesis that performs better (that is, has the largest margin) on a training set of cardinality n. We show that on average the generalization error decreases faster as a function of n for the OT targets than for the HG-non-OT ones. Section 5 concludes the paper and discusses various issues to explore in future research.

2 The OT versus HG debate

As reviewed above, HG fundamentally differs from OT because it does away with strict domination and therefore allows for **gang effects** in which multiple violations of lower-weighted constraints outweigh a violation of a higher-weighted constraint (see section 3 for details). Bane and Riggle (2009) show that sets of constraints drawn from the phonological literature yield much richer typologies in HG than in OT as a result of gang effects, and that many of the additional patterns derived under HG are unattested. The same point is made by the investigation of Kaun's (2004) analysis of the typology of rounding harmony discussed in section 4. However these constraint sets were developed in the context of OT, so these results leave open the possibility that a revised HG constraint set could provide a closer match to natural language typology. In this section we see that there is reason to doubt that the problem of typological over-generation faced by HG phonology can be solved in this way. The evidence comes from classes of problematic gang effects that arise from basic and uncontroversial constraints.

For example AGREE(place) penalizes heterorganic clusters, and *g penalizes voiced velar stops. The weighting of these constraints in figure 1a derives a pattern in which only [g] undergoes place assimilation because IDENT(place) outweighs each markedness constraint individually, but heterorganic [g] violates both constraints, which together outweigh IDENT(place). This pattern cannot be derived by any ranking of these constraints in OT: to block general place assimilation, IDENT(place) must outrank AGREE(place), but that ranking prevents assimilation of [g] as well.

Place assimilation targeting only [g] is unattested (velars resist place assimilation more than coronals and labials and voicing does not affect place assimilation (Jun, 2004)), but once HG is adopted, it is hard to avoid predicting the existence of this process because its derivation does not depend on the specific formulations of AGREE(place) and *g. The prediction follows as long as there is some constraint that penalizes heterorganic consonant clusters over homorganic clusters, which is necessary to account for place assimilation, and some constraint that penalizes [g] more than [b, d] and voiceless stops, which is necessary to account for a variety of phenomena, including languages such as Thai that allow voiced stops but not [g] (Ohala, 1983).

Variants of this configuration are easy to generate, e.g. *p (Hayes, 1999) can replace *g to derive place assimilation that only targets [p], or AGREE(place) can be replaced by AGREE(voice) to derive a pattern in which mixed-voicing clusters are tolerated unless they contain [g], in which case devoicing applies. Neither pattern has been reported in spite of thorough investigations of the typologies of place and voicing assimilation. More generally, HG predicts that any markedness constraints that mention the same feature specification in compatible contexts should be able to gang up on faithfulness constraints regulating that feature.

Furthermore, in HG any set of markedness constraints that can penalize a single segment should be

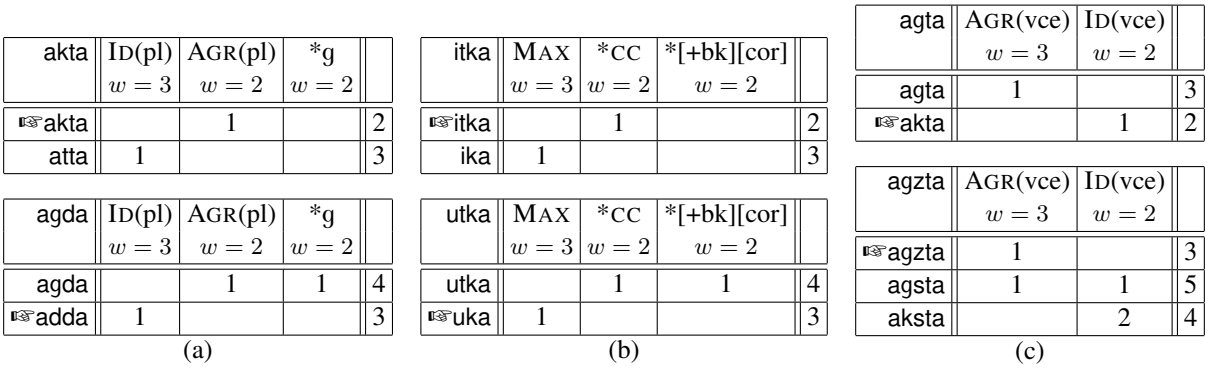

Figure 1: Examples of unattested phonological patterns predicted by HG gang effects

able to gang up on a MAX constraint because deletion of a segment eliminates all of its constraint violations. For example, a constraint against consonant clusters, *CC, and a markedness constraint that penalizes particular VC sequences, e.g. *[+back][cor] (cf. Flemming 2003), can together derive the unattested pattern in figure 1b: pre-consonantal coronals are deleted only if the preceding vowel is back.

Many potential gang effects involving deletion are likely to be ruled out by independent principles. E.g. an alternative repair may be universally preferred due to a fixed ranking among faithfulness constraints (Steriade, 2008). This cannot be the case in the current example because it is a variant of a well-attested process of cluster simplification. On this basis, we can make the generalization that HG predicts the existence of variants of attested deletion processes in which deletion applies only in the presence of additional constraint violations. This set includes many unattested processes.

Another general class of problematic predictions of HG concerns iterative processes in which a markedness constraint can motivate multiple violations of faithfulness. For example, if voicing assimilation is motivated by a constraint like AGREE(voice), then mappings like /agta/→[akta] and /agzta/→[aksta] eliminate just one violation of AGREE(voice) at the cost of $n-1$ violations of IDENT(voice) with a cluster of n obstruents. In HG, the relative weighting of these two constraints establishes a maximum number of consonants that will undergo assimilation (a maximum of 1 in figure 1c) — an unattested phenomenon. In OT, the equivalent ranking derives unbounded assimilation because one violation of AGREE(voice) is worse than any number of violations of IDENT(voice).

Examples of gang effects have been posited by analysts (see Pater 2016 for a review), but alternative OT analyses have been proposed in a number of cases, as in the much discussed case of Japanese loanword devoicing (Pater, 2009; Kawahara, 2006). On balance, the evidence for HG gang effects is weak compared to the evidence that they result in substantial typological over-generation, supporting OT's hypothesis of strict constraint domination.

3 The perspective of statistical learning

We turn now to results from statistical learning theory and bring them to bear on OT's hypothesis of strict domination. The presentation is kept informal with technical details relegated to the final appendix.

3.1 Binary classification

The statistical learning framework of binary classification assumes a **set of instances** $\mathcal{X}$ and a **set of labels** $\mathcal{Y} = \{+1, -1\}$. A classifier can then be construed as a function which assigns a label $y = +1$ or $y = -1$ to an instance x in the set $\mathcal{X}$. We are interested in classifiers with a special shape, as follows.

We start with a collection $\mathcal{H}$ of functions $h : \mathcal{X} \to [-1, +1]$ that take an instance and return a number between -1 and $+1$. Using the functions in $\mathcal{H}$, we construct the collection $\mathcal{F}$ of all weighted sums $f = \sum_{k=1}^{K} w_k h_k$ of an arbitrary finite number K of functions h_k in $\mathcal{H}$ through some corresponding weights w_k. We restrict ourself to weights which are non-negative and sum up to 1 (whereby $\mathcal{F}$ is the convex hull of $\mathcal{H}$). A function $h \in \mathcal{H}$ or a function $f \in \mathcal{F}$ maps instances in $\mathcal{X}$ to numbers between -1 and $+1$. The sign of these numbers can in turn be interpreted as a classification label. Thus, $\mathrm{sign}(h)$ with $h \in \mathcal{H}$ is called a **basic classifier** and $\mathrm{sign}(f)$

with $f = \sum_{k=1}^{K} w_k h_k \in \mathcal{F}$ is called a **voting** (or an **ensemble**) **classifier**, because it aggregates and averages the "votes" of the basic classifiers.

We consider a probability distribution $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$ that generates labels from instances according to the conditional probability $\mathbb{P}(y|x)$. The **generalization error** $Err_{\mathbb{P}}(f)$ of a classifier $f \in \mathcal{F}$ relative to $\mathbb{P}$ is the probability of misclassification of f, namely the probability under $\mathbb{P}$ of a labeled instance (x, y) such that f assigns to the instance x a label $\text{sign}(f(x))$ different from the intended label y:

$$Err_{\mathbb{P}}(f) = \mathbb{P}\big[\text{sign}(f(x)) \neq y\big]$$

As the generalization error measures the probability of misclassification, a classifier with a smaller generalization error is better than a classifier with a larger generalization error. The learner's ideal goal would be to find a classifier $f \in \mathcal{F}$ with the smallest possible generalization error, that is a classifier which maps instances to their most probable label. Unfortunately, the generalization error $Err_{\mathbb{P}}(\cdot)$ cannot be minimized directly, because it is defined in terms of the probability $\mathbb{P}$ which is unknown to the learner. Indeed, the learner only has at its disposal a training set $T = ((x_1, y_1), \ldots, (x_n, y_n))$ consisting of n labeled instances $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ sampled independently according to $\mathbb{P}$.

The goal of statistical learning theory is to provide **error bounds**, that is bounds on the generalization error $Err_{\mathbb{P}}(f)$ of an arbitrary classifier $f \in \mathcal{F}$ based on parameters such as the shape of f or its performance on the training set T. Of course, we want our error bounds to be as low as possible, thus providing guarantees for the smallest possible generalization error. In this section, we focus on a state-of-the-art error bound due to Koltchinskii and Panchenko (2005, theorem 2, page 1464; henceforth KP), recalled in appendix A.1. Sections 3.2 and 3.3 discuss the two crucial properties of KP's bound qualitatively. This will suffice to make a connection with OT's strict domination in section 3.4.

3.2 KP's bound depends on the margin

The condition $\text{sign}(f(x_i)) = y_i$ that a voting classifier $sign(f)$ classifies correctly the data pair (x_i, y_i) is equivalent to the inequality $y_i f(x_i) > 0$. Thus, the size of the real number $y_i f(x_i)$ can be intuitively interpreted as the margin of confidence with which f succeeds at assigning the correct label y_i to the instance x_i: the larger $y_i f(x_i)$ is above zero, the larger the confidence. Given a training set $T = ((x_1, y_1), \ldots, (x_n, y_n))$ that f classifies correctly, we focus on the most dangerous training pair, namely the one that f classifies with the smallest confidence. That smallest margin of confidence is called the **margin** $\delta_T(f)$ of f on the training set T:

$$\delta_T(f) = \min_{i=1,\ldots,n} y_i f(x_i) \tag{1}$$

Since the margin $\delta_T(f)$ represents the worst-case confidence of f on the training set T, it is intuitive that KP's bound (like earlier bounds, since Schapire *et al.* 1998) depends on the margin in such a way that the error bound is large (that is, worse) when the margin $\delta_T(f)$ is small (namely close to 0). See appendix A.2 for details on the dependence of KP's bound on the margin. In conclusion, KP's bound says that, all else being equal, the learner should pick a classifier in $\mathcal{F}$ which correctly classifies the training set T with the largest margin $\delta_T(f)$. We will use this fact extensively in section 4.

3.3 KP's bound depends on the effective dimension

Consider a representation of a voting classifier $f \in \mathcal{F}$ as a sum of basic classifiers in $\mathcal{H}$, namely $f = \sum_{k=1}^{K} w_k h_k$ with non-negative weights w_k which sum up to 1 and are therefore each smaller than 1. We assume without loss of generality that $w_1 \geq w_2 \geq \cdots \geq w_K$. Intuitively, the number K of basic classifiers in the representation of f can be interpreted as the **dimension** of f. Yet, the weights in the tail of the representation of f might be tiny whereby the corresponding basic classifiers contribute only little and should be discounted when determining the dimension of f. KP thus consider the alternative notion (2) of **effective dimension** $d_T(f)$ of the classifier f. Intuitively, we split K as $K = d + (K - d)$ and replace $K - d$ with the sum $\sum_{j=d+1}^{K} w_j$ of the $K - d$ weights in the tail, thus taking into account the smallness of the smallest weights. If the weights decrease fast, the tail weights will be small and the effective dimension $d_T(f)$ will therefore be small.

$$d_T(f) = \min_{0 \leq d \leq K} \left[d + \left(\sum_{j=d+1}^{K} w_j \right)^2 \frac{2 \log n}{\delta_T(f)^2} \right] \tag{2}$$

The novelty of KP's error bound is that it depends not only on the margin $\delta_T(f)$ of the classifier f but also on its effective dimension $d_T(f)$ and thus on the decay of the weights in a representation of f. In the sense that (for a fixed margin) KP's bound is small (that is, better) when the effective dimension is small because of a fast decay of the weights. For instance, the bound is smaller for exponentially decaying weights than for polynomially decaying weights (assuming that the margin is the same in the two cases). See appendix A.3 for details on the dependence of KP's bound on the decay of the weights. In conclusion, KP's error bound says that, all else being equal, the learner should pick a classifier in $\mathcal{F}$ which correctly classifies the training set T and whose weights decay fastest, possibly exponentially. We now make explicit the implications of this conclusion for the OT versus HG debate.

3.4 KP's bound and OT's strict domination

The connection between the classification framework reviewed above and the framework of constraint-based phonology can be drawn as follows. Let the space of instances consist of triplets (u, s, s') where u is an underlying form and s, s' are corresponding candidate surface forms. We interpret s as the intended **winner** and s' as the intended **loser**. The **HG grammar** relative to constraints $C_1, \ldots, C_K$ and weights $w_1, \ldots, w_K \geq 0$ is consistent with the triplet (u, s, s') provided $\sum_{k=1}^{K} w_k h_k(u, s, s') > 0$ where $h_k(u, s, s')$ is the constraint violation difference

$$h_k(u, s, s') = C_k(u, s') - C_k(u, s) \qquad (3)$$

Without loss of generality, we assume the weights w_k sum up to 1. Furthermore, we assume that there are a finite number of underlying forms and a finite number of surface forms (for discussion of this assumption, see Alber *et al.* 2015). Thus, we can assume without loss of generality that

$$-1 \leq h_k(u, s, s') \leq +1 \qquad (4)$$

for every triplet (u, s, s'). In fact, if the inequalities (4) fail for the original constraints, we can divide them by the largest number of constraint violations without affecting the typological predictions. In conclusion, an HG grammar can be construed as a classifier $f \in \mathcal{F} = \mathrm{conv}(\mathcal{H})$ in the convex hull of the collection $\mathcal{H}$ of the constraint violation differences h_k in (3) which take values in $[-1, +1]$ by (4).

The **OT grammar** relative to constraints $C_1, \ldots, C_K$ and a constraint ranking π is consistent with the triplet (u, s, s') provided there exists a constraint C_k such that each of the constraints π-ranked above C_k assigns the same number of violations to the two mappings (u, s) and (u, s') while the constraint C_k assigns less violations to the winner mapping (u, s) than to the loser mapping (u, s'). The following well known result says that the latter condition is equivalent to the HG consistency condition relative to exponentially decaying weights (Prince and Smolensky, 2004; Keller, 2000; Keller, 2005). The constant Z in (5b) is arbitrary and can be used to normalize the weights.

Theorem 1 *Consider an arbitrary ranking π. Without loss of generality, assume that π is (5a), whereby C_1 is ranked at the top, C_2 is ranked below it and so on, until the bottom ranked C_K.*

$$\begin{array}{lll} a.\ C_1 & b.\ w_1 = \frac{1}{Z}\left(\frac{\Delta+\delta}{\delta}\right)^{-1} \\ \downarrow & \\ C_2 & w_2 = \frac{1}{Z}\left(\frac{\Delta+\delta}{\delta}\right)^{-2} \\ \downarrow & \\ \vdots & \vdots \\ \downarrow & \\ C_K & w_K = \frac{1}{Z}\left(\frac{\Delta+\delta}{\delta}\right)^{-K} \end{array} \qquad (5)$$

The HG grammar corresponding to the weights in (5b) for an arbitrary $Z > 0$ and

$$\Delta = \max\left\{|h_k(u, s, s')| \,\middle|\, k = 1 \cdots K\right\}$$
$$\delta = \min\left\{h_k(u, s, s') \,\middle|\, h_k(u, s, s') > 0\right\}$$

is consistent with a triplet (u, s, s') if and only if the OT grammar corresponding to π is.

Theorem 1 says that OT's strict domination corresponds to a restriction to the subset of the HG typology corresponding to weights which decay exponentially, as in (5b). KP's bound provides a learnability rationale for this restriction: fast decaying weights ensure a smaller effective dimension (as long as the margin does not shrink) and thus a smaller (that is, better) error-bound. Thus, a learner of an OT grammar would have a better guarantee of a low generalization error, and we may conjecture that it will actually have a lower generalization error in practice.

4 Empirical simulations

To complement the theoretical perspective of section 3, we now turn to simulations of margin-based learning on two test cases. Our experiments found OT target grammars to be easier, on average, to learn than HG-non-OT ones. Furthermore, we found that this learning procedure yields weights with a lower effective dimension on OT targets than on HG-non-OT ones.

4.1 Test cases

Our first test case is based on the analysis of rounding harmony by Kaun (2004). It models progressive harmony between two vowels of the same backness. As it posits two levels of height and backness, it assumes 8 underlying forms consisting of one of 4 possible triggers (i.e., the four rounded vowels which differ for height and backness) and of one of 2 possible targets (the unrounded vowels of corresponding backness of both possible heights). Each underlying form has 2 candidate surface forms, one with harmony and one without. The constraint set consists of 7 constraints (see the online supplementary materials). The typology (computed with OT-Help2; Staubs *et al.* 2010) consists of 37 OT grammars and 26 HG-non-OT grammars.

Our second test case is based on the analysis of syllable structure by Prince and Smolensky (2004, Part II). This analysis involves 5 constraints in its simpler variant. As in Bane and Riggle (2009), the set of underlying forms consists of all 13 strings of length 1 to 3 of symbols in $\{C, V\}$ (except CV which has only one possible output). Furthermore, we used their procedure to precompute all possibly optimal outputs, yielding a total of 56 surface forms.[1] The typology (computed with OT-Help2) consists of 12 OT and 13 HG-non-OT grammars.

[1] Note that what we call underlying and surface forms do not really correspond to actual forms but to patterns of constraint violations. For instance in our second test case, the underlying forms /tat/ and /bat/ are a single "underlying form" /CV/, and the surface forms (say for /tat/) [ta] and [da] are a single "surface form" [CV]. This means that the admittedly low number of data points we have should not be compared to the number of words human learners are exposed to; our data points exemplify all the possible patterns of small length for each phenomenon.

4.2 Procedure

Algorithm 1 features the pseudo-code for our simulation procedure. For each grammar G in the typology, we build the set of instances $\mathcal{X}_G$ in (6). We consider all triplets (u, s, s') where: u is an underlying form; s is the corresponding winner surface form according to the grammar G; and s' is a loser candidate for u different from s. We represent (u, s, s') as the vector $h(u, s, s')$ whose components are the constraint violation differences $h_k(u, s, s')$ in (3).

$$\mathcal{X}_G = \{x = h(u, s, s') \,|\, G \text{ maps } u \text{ to } s\} \quad (6)$$

We sample a training set T by drawing uniformly with replacement n data points from $\mathcal{X}_G$ (we assume all labels are equal to $y = 1$, because we only generate positive data). Based on the considerations in section 3.2, we compute the weights w^* which maximise the empirical margin on the training set T over all non-negative weight vectors $w \geq 0$. The margin (1) can be made explicit as in (7) in the specific case considered

$$\delta_T(w) = \min\{w^\mathsf{T} x \,|\, x \in T\} \quad (7)$$

We do this for n ranging from 3 to an arbitrary number N. This procedure is repeated 250 times, so we can compute the average generalization error $Err(n, G)$ that a margin-based learner trying to learn G makes after seeing n data points.

Algorithm 1: Learning simulation procedure.

```
for G in the typology do
    for n = 3, ..., N do
        for m = 1, ..., 250 do
            Randomly select T ∈ X_G^n
            w* ← arg max_{w≥0} δ_T(w)
            Err(m) ← P(w*ᵀ x ≤ 0 | x ∈ X_G)
        Err(n, G) ← (1/250) Σ_m Err(m)
```

4.3 Results

Figure 2 plots the error $Err(n, G)$ averaged over target OT-grammars G (solid red lines) and averaged over target HG-non-OT grammars (dashed blue lines). We observe a learnability advantage for OT grammars in practice, as the generalization error of

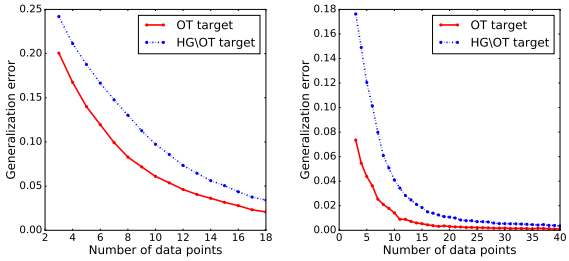

Figure 2: Average of the generalization error $Err(n, G)$ over OT and over HG-non-OT target grammars as a function of n, for rounding harmony (left) and syllable types (right) data.

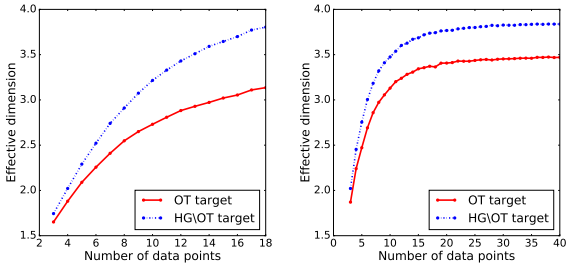

Figure 3: Average effective dimension of the learner's weights over OT and HG-non-OT target grammars as a function of n, for rounding harmony (left) and syllable types (right) data.

a margin-based learner on OT target grammars is lower for any given number n of data points than that of the same learner on HG-non-OT targets.

The error obtained in the simulations cannot be straightforwardly compared to Koltchinskii and Panchenko's error bound (8), as we do not know the value of the constant K which appears in the bound. Yet, figure 3 shows a lower effective dimension — as defined in (2) — of the weights w^* selected by the learner when trained on OT target grammars (red solid line) than on HG-non-OT targets (blue dotted line). Thus, we can speculate that the easier learnability of OT grammars compared to HG-non-OT grammars is related to the lower effective dimension of the HG weights that generate them.

Of course, the advantage of OT that we observe *on average* could be due to just a couple of very "easy" OT grammars that drag the average down. For instance, in the case of harmony, both the grammar with systematic harmony and the one with no harmony at all only depend on only one constraint (respectively ALIGN-L/R([RD]) and DEP(LINK)) and both belong to the OT typology. Figure 4 thus plots the generalization error $Err(n, G)$ for each individual OT (red dashed lines) and each individual HG-non-OT (blue dotted lines) target grammar G. The overall pattern is that most OT grammars are easier to learn than most HG-non-OT grammars. In the case of syllable structure, there are indeed only a few exceptions to this general pattern. The pattern is admittedly somewhat less clear in the case of vowel harmony, as discussed below in section 5.A.

5 Conclusions and open issues

This paper has argued that OT's strict domination seems to be warranted by phonological typology (section 2) and that strict domination might provide a learnability advantage (*pace* Riggle *et al.* 2009; 2010). This learnability argument is twofold: first, a review of recent results in the statistical learning literature (section 3) lets us conclude that learners of OT grammars will infer them with greater chance of success for the same amount of data. Second, simulation results on realistic test cases (section 4) show that OT target grammars are indeed easier to learn under certain assumptions. We conclude with various open issues that we would like to address in future research.

(A) As remarked above, figure 4 shows that several of the "hardest" grammars are part of the OT typology in the harmony case. As a tentative explanation, we note that in this test case, there are few underlying and surface forms, and many constraints, some of which are closely related. For instance, there are three different variants of ALIGN-L/R([RD]) for different features of the trigger vowel. Thus, in most grammars of the HG typology, not all constraints have to be active (in the sense of having non-zero weights). Certain OT grammars are harder than certain HG-non-OT grammars by virtue of requiring more active constraints. Future work will try to get a cleaner picture by comparing only OT and HG-non-OT grammars that require a comparable number of active constraints.

(B) For consistency with the classification framework of section 3.1, the simulations described in section 4 define the error in terms of the number of triplets (u, s, s') where the loser s' incorrectly beats the winner s (see line 6 in algorithm 1). We might instead want to redefine the error in terms of the number of underlying forms u mapped to a winner different from s. For the results from statistical learning theory in section 3 to still be relevant, we would need to extend them from classifiers of the form $f(x) =$

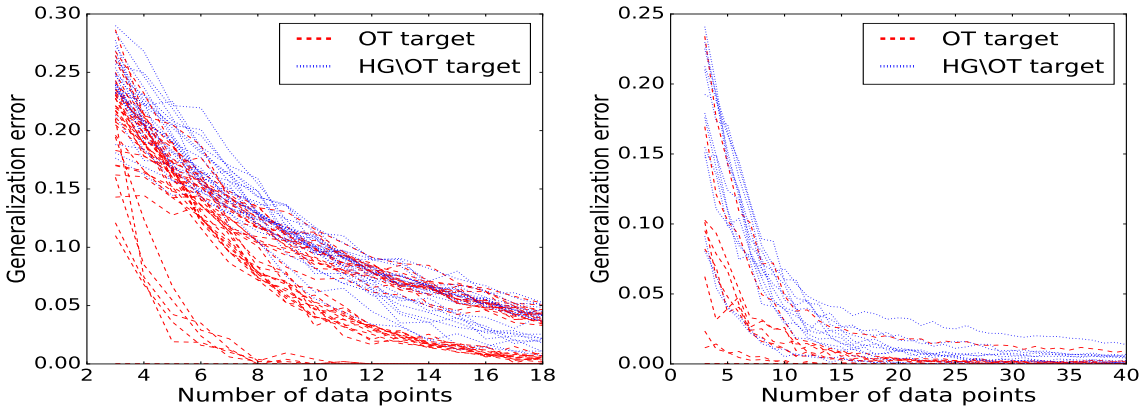

Figure 4: Generalization error $Err(n, G)$ as a function of n, for rounding harmony (left) and syllable types (right) data, for each OT target grammar G (red line) and each HG-non-OT target grammar G (blue line).

$\sum_k w_k h_k(x)$ to $f(x) = \min_{t \in S(x)} \sum_k w_k h_k(x, t)$, where S is a function from x to some finite set.

(C) The simulations reported in section 4 assume a uniform distribution over triplets (u, s, s') all consistent with some target grammar G. Future research will look at different data distributions (e.g., a Zipfian distribution over the underlying forms u) and the addition of some noise in the training data.

(D) The learner tested in section 4 simply looks for weights which maximize the margin but is oblivious to whether the target grammar is an OT or an HG-non-OT grammar. For OT targets, theorem 1 suggests the more specific learning strategy in algorithm 2. We consider each ranking π, construct the corresponding exponentially decaying weights w_π in (5), and determine the ranking $\pi*$ whose weights w_{π^*} maximize the margin. We denote by $Err_{\mathrm{OT}}(n, G)$ the average error of the OT grammar corresponding to π^* on the target grammar G.

$Err_{\mathrm{OT}}(n, G)$ is generally quite high when G is a HG-non-OT grammar. This is not surprising, since we're trying to learn a grammar outside the search space. Yet, figure 5 shows that even when the target grammar G is OT, $Err_{\mathrm{OT}}(n, G)$ (red solid line) is slightly higher than the error $Err(n, G)$ (dashed blue line) obtained with the general learning procedure in algorithm 1. This is puzzling, as one might have expected that the restriction of the search space in algorithm 2 should have led to a lower error. Towards a possible explanation, we observe that the weights w_{π^*} obtained by algorithm 2 result in a very low margin, and thus a very high effective dimension compared to the weights w^* obtained through algorithm 1, as shown in figure 6.

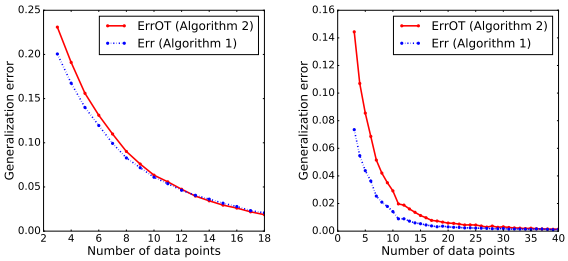

Figure 5: Average over OT target grammars of the generalization errors $Err(n, G)$ in algorithm 1 and $Err_{\mathrm{OT}}(n, G)$ in algorithm 2, for harmony (left) and syllable types (right) data.

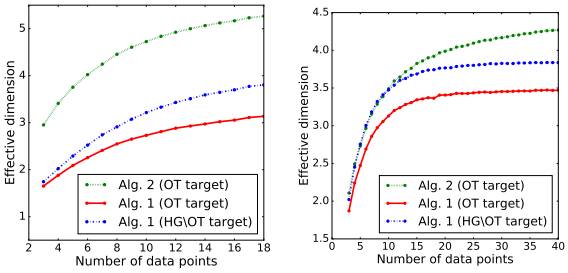

Figure 6: Effective dimension of the weights in algorithm 1 averaged over OT and over HG-non-OT target grammars; effective dimension of the weights in algorithm 2 averaged over OT target grammars.

Evidently, margin-based learning is incompatible with the strategy (5) for computing exponentially-decaying weights corresponding to OT rankings. One possibility for future research is to base weights not on full rankings, but on RCD's (Tesar and Smolensky, 1998) hierarchy $H_1 \gg H_2 \gg \cdots$ (H_1 consists of constraints never loser preferring in T; H_2 consists of constraints which are only loser-preferring on triplets (u, s, s') of T where some constraint in H_1 is winner-preferring; and so on). For instance, one could pick weights w_H so that the constraints in H_1 all have the same weight, the

constraints in H_2 all have the same exponentially smaller weight, and so on. A strategy of this kind might reach a compromise between fast decay and large margin.

Algorithm 2: Learning simulation for OT targets.

for *G in the typology* **do**
  for $n = 3, \ldots, N$ **do**
    for $m = 1, \ldots, 250$ **do**
      Randomly select $T \in \mathcal{X}_G^n$
      $\pi^* \leftarrow \arg\max_\pi \delta_T(w_\pi)$
      $Err_{\text{OT}}(m) \leftarrow \mathbb{P}(w_{\pi^*}^T x \leq 0 | x \in \mathcal{X}_G)$
    $Err_{\text{OT}}(n, G) \leftarrow \frac{1}{250} \sum_m Err_{\text{OT}}(m)$

Acknowledgments

The research reported in this paper was partially supported by the *MIT France Seed Fund* (project title: 'Phonological Typology and Learnability') and by the *Agence National de la Recherche* (project title: 'The mathematics of segmental phonotactics'). We thank an anonymous reviewer for helpful comments.

Appendix: more details on KP's bound

A.1 The exact formulation of Koltchinskii and Panchenko's (2005, theorem 2, p. 1464) error bound discussed in section 3 is as follows:

Theorem 2 *Suppose that $\mathcal{H}$ is a VC-subgraph class with VC-dimension V (see for instance Mohri* et al. *2012). Consider a voting classifier $f = \sum_{k=1}^K w_k h_k \in \mathcal{F} = conv(\mathcal{H})$ which classifies correctly a training set $T = ((x_1, y_1), \ldots, (x_n, y_n))$ sampled i.i.d. according to a distribution $\mathbb{P}$. For every $t > 0$, the generalization error $Err_{\mathbb{P}}(f)$ of f is bound as follows with probability at least $1 - e^{-t}$:*

$$Err_{\mathbb{P}}(f) \leq K \left(\frac{V d_T(f)}{n} \log \frac{n}{\delta_T(f)} + \frac{t}{n} \right) \quad (8)$$

where K is a universal constant, $\delta_T(f)$ is the margin of the classifier f on the training set T defined in (1) and $d_T(f)$ is its effective dimension defined in (2).

A.2 Since $\sum_{k=1}^K w_k \leq 1$, the choice $d = 0$ in the definition (2) of the effective dimension yields $d_T(f) \leq \frac{2}{\delta_T(f)^2} \log n$. KP's bound (8) thus becomes

$$Err_{\mathbb{P}}(f) \leq K \left(\frac{V \log n}{n \delta^2} \log \frac{n}{\delta} + \frac{t}{n} \right) \quad (9)$$

which decreases as $1/n$ when $n \to \infty$ and increases as $1/\delta^2$ when $\delta \to 0$.

A.3 The effective dimension $d_T(f)$ which appears in KP's bound (8) depends on the decay of the weights in a representation of f. The following corollary (see Koltchinskii and Panchenko 2005, example on p. 1465) details the dependence of the bound on the decay. The proof of the corollary is provided in the online supplement, based on class notes by Panchenko (2004, class 21), as it has not appeared in the literature.

Corollary 1 *Consider a classifier $f = \sum_{i=1}^K w_k h_k$ in $\mathcal{F}$ which classifies correctly a training set $T = ((x_1, y_1), \ldots, (x_n, y_n))$ with margin $\delta = \delta_T(f)$, namely $y_1 f(x_1), \ldots, y_n f(x_n) > \delta$.*

- *If the weights w_k decay polynomially, i.e. $w_k \leq k^{-B}$ for some $B > 1$, KP's bound (8) becomes:*

$$Err_{\mathbb{P}}(f) \leq K \left(\frac{C_B}{\delta^{2/(2B-1)}} \frac{V}{n} \log^2 \frac{n}{\delta} + \frac{t}{n} \right) \quad (10)$$

where $C_B \to 1$ as $B \to \infty$.

- *If the weights w_k decay exponentially, namely $w_k \leq e^{-k}$, KP's bound (8) becomes:*

$$Err_{\mathbb{P}}(f) \leq K \left(\frac{V}{n} \log^2 \frac{n}{\delta} + \frac{t}{n} \right) \quad (11)$$

The two bounds (10) and (11) decrease as $1/n$ when $n \to \infty$, just as in the general case (9). The substantial improvement concerns the growth of the bound when $\delta \to 0$. The general bound (9) grows as $1/\delta^2$ when $\delta \to 0$. The bound (10) for the case of polynomial decay instead grows only as $1/\delta^{2/(2B-1)}$, which is slower than $1/\delta^2$ because $2/(2B-1) \leq 2$ as $B > 1$. Furthermore, the bound (11) for the case of exponential decay grows only as $\log 1/\delta$ when $\delta \to 0$, which is substantially slower than $1/\delta^2$. When $B \to \infty$, the bound (10) for the case of polynomial decay becomes the bound (11) for the case of exponential decay.

References

Birgit Alber, Natalie DelBusso, and Alan Prince. 2015. From intensional properties to universal support. Università degli Studi di Verona and Rutgers University.

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

Max Bane, Jason Riggle, and Morgan Sonderegger. 2010. The VC dimension of constraint-based grammars. *Lingua*, 120.5:1194–1208.

Edward Flemming. 2003. The relationship between coronal place and vowel backness. *Phonology*, 20:335–373.

Bruce Hayes. 1999. Phonetically-driven phonology: the role of Optimality Theory and inductive grounding. In Michael Darnell, Edith Moravscik, Michael Noonan, Frederick Newmeyer, and Kathleen Wheatly, editors, *Functionalism and Formalism in Linguistics*, volume 1: General Papers, pages 243–285. John Benjamins, Amsterdam.

Jongho Jun. 2004. Place assimilation. In B. Hayes, R. Kirchner, and D. Steriade, editors, *Phonetically Based Phonology*, pages 58–86. Cambridge University Press.

Abigail Kaun. 2004. The typology of rounding harmony. In Bruce Hayes, Robert Kirchner, and Donca Steriade, editors, *Phonetically based phonology*, pages 87–116. Cambridge University Press.

Shigeto Kawahara. 2006. A faithfulness ranking projected from a perceptibility scale: The case of [+voice] in Japanese. *Language*, 82:536–574.

Frank Keller. 2000. *Gradience in Grammar. Experimental and Computational Aspects of Degrees of Grammaticality*. Ph.D. thesis, University of Edinburgh, England.

Frank Keller. 2005. Linear Optimality Theory as a model of gradience in grammar. In Gisbert Fanselow, Caroline Féry, Ralph Vogel, and Matthias Schlesewsky, editors, *Gradience in Grammar: Generative Perspectives*, pages 270–287. Oxford University Press, Oxford.

Vladimir Koltchinskii and Dmitry Panchenko. 2002. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, 30:1–50.

Vladimir Koltchinskii and Dmitry Panchenko. 2005. Complexities of convex combinations and bounding the generalization error in classification. *Ann. Statist.*, 33.4:1455–1496.

Vladimir Koltchinskii, Dmitry Panchenko, and Savina Andonova. 2003a. Generalization bounds for voting classifiers based on sparsity and clustering. In *Lecture Notes in Artificial Intelligence 2777*, pages 492–505.

Vladimir Koltchinskii, Dmitry Panchenko, and Lozano. 2003b. Bounding the generalization error of convex combinations of classifiers: Balancing the dimensionality and the margins. *Ann. Appl. Probab.*, 13:213–252.

Gèraldine Legendre, Yoshiro Miyata, and Paul Smolensky. 1990a. Harmonic Grammar: A formal multi-level connectionist theory of linguistic well-formedness: An application. In Morton Ann Gernsbacher and Sharon J. Derry, editors, *Annual conference of the Cognitive Science Society 12*, pages 884–891, Mahwah, New Jersey. Lawrence Erlbaum Associates.

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

Gèraldine Legendre, Antonella Sorace, and Paul Smolensky. 2006. The optimality theory/harmonic grammar connection. In Paul Smolensky and Gèraldine Legendre, editors, *The Harmonic Mind*, pages 903–966. MIT Press, Cambridge, MA.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of Machine Learning*. MIT Press, Cambridge, MA.

John J. Ohala. 1983. The origin of sound patterns in vocal tract constraints. In Peter F. MacNeilage, editor, *The production of speech*, pages 189–216. Springer-Verlag, New York.

Dmitry Panchenko. 2004. Statistical learning theory. Lecture notes for the class 18.465 (Topics in Statistics), Department of Mathematics, MIT.

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

Joe Pater. 2016. Universal grammar with weighted constraints. In Joe Pater and John J. McCarthy, editors, *Harmonic Grammar and Harmonic Serialism*, pages 1–46. Equinox, London.

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

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

Alan Prince and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in generative grammar*. Blackwell, Oxford. 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.

Jason Riggle. 2009. The complexity of ranking hypotheses in Optimality Theory. *Computational Linguistics*, 35(1):47–59.

Robert E. Shapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. 1998. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26.5:1651–1686.

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

Robert Staubs, Michael Becker, Christopher Potts, Patrick Pratt, John J. McCarthy, and Joe Pater. 2010. OT-Help 2.0. Software package. Software Package. University of Massachussetts, Amherst.

Donca Steriade. 2008. The phonology of perceptibility effects: the P-map and its consequences for constraint organization. In Kristin Hanson and Sharon Inkelas, editors, *The nature of the word: essays in honor of Paul Kiparsky*, pages 151–179. MIT Press, Cambridge.

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

Vladimir N. Vapnik and Alexey Y. Chervonenkis. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280.