The Complexity of Ranking Hypotheses in Optimality Theory

Jason Riggle*
University of Chicago

Given a constraint set with k *constraints in the framework of Optimality Theory (OT), what is its capacity as a classification scheme for linguistic data? One useful measure of this capacity is the size of the largest data set of which each subset is consistent with a different grammar hypothesis. This measure is known as the Vapnik-Chervonenkis dimension (VCD) and is a standard complexity measure for concept classes in computational learnability theory. In this work, I use the three-valued logic of Elementary Ranking Conditions to show that the VCD of Optimality Theory with* k *constraints is* k−1*. Analysis of OT in terms of the VCD establishes that the complexity of OT is a well-behaved function of* k *and that the 'hardness' of learning in OT is linear in* k *for a variety of frameworks that employ probabilistic definitions of learnability.*

1. Introduction

Given a set CON of k constraints in the framework of Optimality Theory (OT; Prince and Smolensky 1993), what is the capacity of CON as a classification scheme for samples of language data? In OT, constraints are functions that map candidates to natural numbers, where each candidate is a member of the (possibly infinite) set of possible derivations of an input form i supplied by the candidate generating function GEN(i). The number that a constraint $\mathbb{C}_i$ assigns to a candidate indicates how many times that candidate violates $\mathbb{C}_i$. A grammar is a **ranking** of the constraints that imposes a total ordering on CON, $\mathcal{R}_{\text{CON}}$ (or simply $\mathcal{R}$ when CON is clear from the context), and the **language** that is generated by grammar $\mathcal{R}$ is the set of candidates that are optimal according to $\mathcal{R}$ as in Definition 1.

Definition 1

a. Candidate a is more **harmonic** than candidate b according to $\mathcal{R}$, written $a \succ b$, if they share the same input and a is assigned fewer violations by the highest-ranked constraint that assigns different numbers of violations to a and b.

b. Candidate a is **optimal** according to ranking $\mathcal{R}$ iff no other candidate generated by GEN is more harmonic than a.

Because each of the $k!$ rankings of CON is a different grammar that generates a potentially unique language, one natural measure of the classificatory capacity of CON

* Department of Linguistics, University of Chicago, 1010 E. 59th St., Chicago, IL 60637. jriggle@uchicago.edu. Many thanks to Alan Prince, Jeff Heinz, Greg Kobele, Colin Wilson, and two anonymous *Computational Linguistics* reviewers for helpful comments and suggestions. Any errors are, of course, my own.

Submission received: 22 December 2006; revised submission received: 9 October 2007; accepted for publication: 17 December 2007.

is the upper bound of $k!$ languages in what Prince and Smolensky (1993, page 27) dub the **factorial typology** of the constraint set. Another complexity metric that is useful in analyses of learnability (especially for non-finite concept classes) is the cardinality of the largest data set of which each subset corresponds to a different ranking hypothesis. The idea of measuring the complexity of a concept class (in the case at hand, a set of grammars) in this way comes from the work of Vapnik and Chervonenkis (1971) and is known as the Vapnik-Chervonenkis dimension (VCD). In OT, the VCD of a constraint set CON (i.e., the concept class consisting of languages generated by rankings of CON) is the size of the largest sample (set of candidates) that is shatterable as in Definition 2.

Definition 2
A sample S is **shatterable** by a constraint set CON iff, for every partitioning of S into two disjoint sets T and F (including the null/whole partition), there is at least one ranking $\mathcal{R}_{\text{CON}}$ that makes every $s \in T$ optimal but no $s \in F$ optimal.

Vapnik and Chervonenkis's definition of shatterability has interesting implications for samples consisting of OT candidates. For instance, each candidate in a shatterable sample S must be an *input* $\rightarrow$ *output* mapping for a unique input form because two candidates a and b with the same input would either *tie* with identical sets of violations or show harmonic inequality. In the case of a tie, no ranking could realize a partitioning that separates a and b and, in the case of harmonic inequality, no ranking could realize a partitioning in which a and b are simultaneously optimal. More generally, the VCD places an upper bound on the number of distinct grammar hypotheses that can be realized over any sample of linguistic data consisting of OT candidates, and thus provides a ready measure of the complexity of the hypothesis space in Optimality Theory.

The VCD of a concept class is obviously not independent of its size. As Blumer et al. (1989) point out, for any finite concept class C, the VCD is bounded at $\log_2 |C|$ because it takes at least 2^d hypotheses to associate a unique hypothesis with every subset of a sample of size d. Thus, because the number of grammars (hypotheses) over k constraints is finite—one grammar for each of the $k!$ rankings—the VCD of OT is bounded at $\log_2 k!$. Or, put more simply, because $\log_2 x! \leq x \log_2 x$, this establishes $k \log_2 k$ as an upper bound on the VCD of OT. In this article, I will show how the structure of the hypothesis space in Optimality Theory provides a tighter bound on the VCD of OT than the bound established by the finitude of the hypothesis space. I will improve upon the inherent bound of $k \log_2 k$ by showing that the VCD of OT with k constraints is actually bounded at $k - 1$ and thus grows linearly with the size of $|\text{CON}|$.

The complexity measured by the VC dimension has a number of ramifications for learning in Optimality Theory. For instance, the VCD of a concept class places an absolute lower bound on the number of mistakes that any error-driven learning algorithm can be guaranteed of achieving (Littlestone 1988). This fact tells us that it may yet be possible to improve upon the quadratic mistake bound of $(k^2 - k)/2$ for Recursive Constraint Demotion (Tesar and Smolensky 1993, 2000; Tesar 1995, 1997, 1998), the reigning mistake bound for any OT learning algorithm. The VCD of a concept class also provides a very general bound on the number of data samples that are required for learning in probabilistic models of learning that will be discussed in Section 5.

2. Elementary Ranking Conditions

The main result for the VC dimension of OT will be given in Section 4. First, some supporting results will be established showing that there is an upper bound of $k - 1$ on

shatterable sets of statements about constraint rankings that are expressed with Prince's (2002) Elementary Ranking Conditions.

If our sample space X consists of candidates, then any $x \in X$ can be described in terms of the set of constraint rankings under which x is optimal. Prince (2002) provides a scheme for encoding this kind of ranking information called an Elementary Ranking Condition (ERC). In this section, I will review some formal properties of ERCs that are relevant for establishing the VC dimension of OT. Prince demonstrates many formal properties of ERCs beyond those covered here and shows that ERCS are equivalent to the implication-negation fragment of the three-valued relevance logic RM3 (cf. Anderson and Belnap 1975). This section will review properties of ERCs that are most relevant for the results at hand. For formal proofs and a complete exposition of the logic of ERCs, see Prince (2002).

For a constraint set CON containing k constraints, ERCs are k-length vectors that use the symbols L, e, and W to encode logical statements about rankings. Each constraint is assigned an arbitrary numeric index, and in each ERC α, the i^{th} coordinate α_i refers to the constraint with i^{th} index $\mathbb{C}_i$. The meaning of an ERC is that at least one constraint whose corresponding coordinate contains a W outranks *all* of the constraints whose coordinates contain L's. Thus, $\langle$W, e, L, L$\rangle$ means that $\mathbb{C}_1$ outranks both $\mathbb{C}_3$ and $\mathbb{C}_4$, while $\langle$L, L, W, W$\rangle$ means that either $\mathbb{C}_3$ or $\mathbb{C}_4$ outranks both $\mathbb{C}_1$ and $\mathbb{C}_2$. ERCs can be constructed by comparing candidates as in Definition 3. Note that $\mathbb{C}_i(a)$ denotes the number of times candidate a violates the constraint with index i.

Definition 3
Given a constraint set CON with k constraints indexed $\{1 \ldots k\}$ and two candidates that share the same input, the function $erc_{\text{CON}}(a, b)$ returns an ERC $\alpha = \langle \alpha_1, \ldots, \alpha_k \rangle$ that describes the rankings under which $a \succ b$.[1]

$$erc(a,b) = \langle \alpha_1, \ldots, \alpha_k \rangle \text{ where } \begin{cases} \alpha_i = \text{W} \text{ if } \mathbb{C}_i(a) < \mathbb{C}_i(b) \\ \alpha_i = \text{L} \text{ if } \mathbb{C}_i(a) > \mathbb{C}_i(b) \\ \alpha_i = e \text{ if } \mathbb{C}_i(a) = \mathbb{C}_i(b) \end{cases}$$

The symbol W in α_i of $erc(a, b) = \alpha$ is a mnemonic for the fact that $\mathbb{C}_i$ favors a (the *winner*), whereas an L in coordinate i is a mnemonic for the fact that $\mathbb{C}_i$ favors b (the *loser*). An e in α_i indicates that the candidates are *equivalent* according to $\mathbb{C}_i$.

Example 1

input	$\mathbb{C}_1$	$\mathbb{C}_2$	$\mathbb{C}_3$
cand. a	*	**	*
cand. b	**	*	

(*'s indicate number of violations)

$erc(b, a) = \langle \text{L}, \text{W}, \text{W} \rangle = b \succ a$ if $\mathbb{C}_2$ or $\mathbb{C}_3$ outranks $\mathbb{C}_1$
$erc(a, b) = \langle \text{W}, \text{L}, \text{L} \rangle = a \succ b$ if $\mathbb{C}_1$ outranks $\mathbb{C}_2$ and $\mathbb{C}_3$

Note the symmetry between $erc(a, b) = \langle \text{W}, \text{L}, \text{L} \rangle$, which says that candidate a is more harmonic than b under any ranking where $\mathbb{C}_1$ outranks both $\mathbb{C}_2$ and $\mathbb{C}_3$, and $erc(b, a)$,

1 The function $erc_{\text{CON}}(a, b)$, or simply $erc(a, b)$ when CON is clear from context, is undefined for candidate *input* $\rightarrow$ *output* mappings with different inputs because they cannot be meaningfully compared.

which says that b is more harmonic than a under any ranking where either $\mathbb{C}_2$ or $\mathbb{C}_3$ outranks $\mathbb{C}_1$. This symmetry reflects the fact that $erc(a, b)$ and $erc(b, a)$ encode antithetical ranking conditions. The opposition between these ERCs follows straightforwardly from the fact that only one of the two candidates can be optimal under any given ranking.

The illustrative tableaux presented with OT analyses can be turned into sets of ERCs by making pairwise comparisons between the violations for one designated (or observed) winner and the violations for each other candidate.

Example 2

input	$\mathbb{C}_1$	$\mathbb{C}_2$	$\mathbb{C}_3$	
cand. *a*	*	**		*winner*
cand. *b*	**	*		$erc(a,b) = \langle \text{W}, \text{L}, e \rangle = a \succ b$ if $\mathbb{C}_1$ outranks $\mathbb{C}_2$
cand. *c*	*		**	$erc(a,c) = \langle e, \text{L}, \text{W} \rangle = a \succ c$ if $\mathbb{C}_3$ outranks $\mathbb{C}_2$
cand. *d*	*	*	*	$erc(a,d) = \langle e, \text{L}, \text{W} \rangle = a \succ d$ if $\mathbb{C}_3$ outranks $\mathbb{C}_2$
cand. *e*	**	**		$erc(a,e) = \langle \text{W}, e, e \rangle = a \succ e$ under every ranking
cand. *f*		***	*	$erc(a,f) = \langle \text{L}, \text{W}, \text{W} \rangle = a \succ f$ if $\mathbb{C}_2$ or $\mathbb{C}_3$ outranks $\mathbb{C}_1$

The comparison of candidate a with candidate e in Example 2, $erc(a, e) = \langle \text{W}, e, e \rangle$, yields an odd ranking condition that does not actually express a particular ranking (no constraint has an L), but instead indicates that $\mathbb{C}_1$ favors candidate a and no constraint favors candidate e. In this case, candidate e is said to be **harmonically bounded** by candidate a because there can be no ranking under which e is more harmonic than a. Conversely, if candidate e were designated the winner, then $erc(e, a) = \langle \text{L}, e, e \rangle$. This ERC also does not encode a specific ranking, but rather indicates that the mere existence of candidate a as an alternative means that no ranking can make candidate e optimal.

Like most OT tableaux, Example 2 is an illustration of how a handful of candidates fare with respect to one another according to a particular set of constraints. To know which rankings (if any) make candidate a globally optimal, it would be necessary to define the candidate generating function GEN in order to obtain a representation of the entire set of ERCs $\{erc(a, x) \mid x \in \text{GEN}(input)\} = ERCS(a)$. This is not as daunting as it might appear because, even though $|\text{GEN}(input)|$ may be infinite, the fact that the number of k-length ERCs is finite guarantees that each of the candidates in GEN(*input*) will map to one of a finite number of ERC sets. Furthermore, as Riggle (2004) demonstrates, the standard OT assumption of the universality of **faithfulness** constraints that penalize changes to the input guarantees that all but finitely many of the members of GEN(*input*) will be harmonically bounded. Riggle also presents an algorithm for computing this finite set of **contenders** (i.e., candidates that are not harmonically bounded) that can be used in cases where GEN is restricted so that it is a rational function.[2] Regardless of how optimization is computed, what is relevant for the assessment of the VCD of OT is the definition of optimality. Following Definition 1, a ranking $\mathcal{R}_{\text{CON}}$ can be seen as a function from candidates to *True* (if they are optimal) or *False* (if they are

2 GEN is **rational** if it is representable as a finite state transducer. Riggle's (2004) CONTENDERS algorithm is an extension of Ellison's (1994) application of Dijkstra's (1959) "shortest paths" algorithm to optimization in OT that operates over finite-state representations of GEN and EVAL. Ellison showed that if harmony is used as the "distance" to be optimized, then optimal outputs can be efficiently found. The CONTENDERS algorithm follows a similar strategy but, instead of finding the shortest (i.e., most harmonic) path for one ranking, the algorithm finds all non-harmonically-bounded paths and thereby optimizes for all rankings.

not). The entire ERC set for a candidate $ERCS(a)$ describes exactly the rankings under which candidate a is a globally optimal candidate.

The reduction of candidates to ERC sets makes it possible to use the logic of ERCs to reason about candidates. Most of the time, the ERCs of interest are those that contain at least one L and one W—what Prince calls **nontrivial** ERCs. ERCs that contain W's but no L's are generated when a candidate is compared with another candidate that it harmonically bounds, such as $erc(a, e) = \langle \text{W}, e, e \rangle$ in Example 2. This ERC reveals that candidate e cannot be optimal but yields no information about what rankings make candidate a optimal. Similarly, no ranking information can be gleaned from the all-e ERC that results from comparing "tied" candidates that have the same violations. Finally, ERCs like $erc(e, a) = \langle \text{L}, e, e \rangle$, with L's but no W's reveal nothing other than the fact that candidate e cannot be optimal under any ranking.

The most relevant logical relation for ERCs is that of entailment. The entailment relation among nontrivial ERCs is given in Definition 4 (Prince 2002, page 6, Proposition 1.1).

Definition 4
For nontrivial ERCs α and β, $\alpha \rightarrow \beta$ iff each $\alpha_i \in \alpha$ entails $\beta_i \in \beta$ where $\text{L} \rightarrow e \rightarrow \text{W}$.

Because nontrivial ERCs encode disjunctions of conjunctions (i.e., [$\mathbb{C}_1$ **or** ... $\mathbb{C}_n$] outranks [$\mathbb{C}_{1'}$ **and** ... $\mathbb{C}_{n'}$]), entailments of the form $\alpha \rightarrow \beta$ line up with the logical operations of disjunction introduction (whenever β has W where α has an L or an e) and conjunction elimination (whenever β has an e where α has an L).

Example 3

$\langle \text{W}, \text{L}, \text{L}, e \rangle \rightarrow \langle \text{W}, e, \text{L}, e \rangle$	i.e., If $\mathbb{C}_1$ outranks $\mathbb{C}_2$ and $\mathbb{C}_3$ then $\mathbb{C}_1$ outranks $\mathbb{C}_3$.
$\langle \text{W}, e, \text{L}, e \rangle \rightarrow \langle \text{W}, e, \text{L}, \text{W} \rangle$	i.e., If $\mathbb{C}_1$ outranks $\mathbb{C}_3$ then $\mathbb{C}_1$ or $\mathbb{C}_4$ outranks $\mathbb{C}_3$.
$\langle \text{W}, \text{L}, \text{L}, e \rangle \rightarrow \langle \text{W}, e, \text{L}, \text{W} \rangle$	i.e., If $\mathbb{C}_1$ outranks $\mathbb{C}_2$ and $\mathbb{C}_3$ then $\mathbb{C}_1$ or $\mathbb{C}_4$ outranks $\mathbb{C}_3$.

In addition to revealing entailments among individual ranking conditions, the logic of ERCs makes it possible to derive new ranking conditions that are entailed by the combination of other ERCs. Prince (2002, page 8) provides a logical operation called fusion that derives entailments from sets of ERCs.

Definition 5
The **fusion** of ERC set Φ is a single ERC ϕ that is entailed by Φ where:
$\phi_i = \text{L}$ if any ERC in Φ has an L in its i^{th} coordinate,
$\phi_i = e$ if every ERC in Φ has an e in its i^{th} coordinate,
$\phi_i = \text{W}$ otherwise.

Every ERC entailed by Φ is entailed by the fusion of a subset of Φ (Prince 2002, page 14). Thus, the operation of fusion can reveal nonobvious entailments among ERCs. Consider $\Phi = \{\langle \text{W}, \text{W}, e, \text{L} \rangle, \langle \text{L}, \text{W}, \text{W}, e \rangle, \langle \text{W}, e, \text{L}, \text{W} \rangle\}$. The ERCs in Φ denote, respectively, "$\mathbb{C}_1$ or $\mathbb{C}_2$ outranks $\mathbb{C}_4$," "$\mathbb{C}_2$ or $\mathbb{C}_3$ outranks $\mathbb{C}_1$," and "$\mathbb{C}_1$ or $\mathbb{C}_4$ outranks $\mathbb{C}_3$." The fusion of Φ is $\langle \text{L}, \text{W}, \text{L}, \text{L} \rangle$, which encodes the inference from Φ that $\mathbb{C}_2$ outranks $\mathbb{C}_1$, $\mathbb{C}_3$, and $\mathbb{C}_4$.

The operation of fusion can also reveal inconsistencies in ERC sets. Consider the set $\Psi = \{\langle \text{W}, \text{L}, \text{W} \rangle, \langle \text{L}, \text{W}, \text{W} \rangle, \langle \text{W}, \text{W}, \text{L} \rangle\}$. Fusing Ψ yields $\langle \text{L}, \text{L}, \text{L} \rangle$. As with harmonically bounded candidates, this ERC shows that no constraint ranking is consistent with the statements in Ψ (in fact, they are circular). Prince refers to the class of ERCs with

L's but no W's as $\mathcal{L}^+$. He shows that these ERCs arise from fusion if and only if the fused set contains incompatible ranking conditions.

Definition 6
An ERC set is **consistent** iff it has no subset that fuses to an ERC in $\mathcal{L}^+$ (Prince 2002, page 11).

For any consistent ERC set there is a constraint ranking (often several) of which all of its ERCs are true statements (Prince 2002, page 21). The ERCs in an inconsistent set, on the other hand, can never all be true of a single ranking. Inconsistency can arise from a single pair of candidates (e.g., $ERCS(e)$ in Example 2 contains $erc(e,a) = \langle \mathrm{L}, e, e \rangle$). Inconsistency can also arise across multiple candidate comparisons (e.g., $ERCS(d)$ in Example 2 contains $erc(d,a) = \langle e, \mathrm{W}, \mathrm{L} \rangle$ and $erc(d,c) = \langle e, \mathrm{L}, \mathrm{W} \rangle$). This latter type of inconsistency, where several of the ERCs associated with a candidate fuse to $\mathcal{L}^+$, arises from what Samek-Lodovici and Prince (1999) call **collective harmonic bounding**. Finally, it is possible for inconsistencies to arise when ERCs for several candidates with distinct inputs are combined. For example, if $ERCS(x) = \{\langle \mathrm{W}, \mathrm{L}, \mathrm{W} \rangle\}$, $ERCS(y) = \{\langle \mathrm{L}, \mathrm{W}, \mathrm{W} \rangle\}$, and $ERCS(z) = \{\langle \mathrm{W}, \mathrm{W}, \mathrm{L} \rangle\}$ then, even though x, y, and z may be candidates for distinct inputs (i.e., come from different tableaux), the union of their ERCs fuses to $\langle \mathrm{L}, \mathrm{L}, \mathrm{L} \rangle \in \mathcal{L}^+$ and thereby reveals that there is no ranking under which all three candidates are simultaneously optimal.

Inverting the W's and L's of an ERC produces its antithetical counterpart that is true whenever the original ERC is false and vice versa. This opposition can be exploited in describing the range of consistent ERC sets.

Definition 7
The negation of α is $\overline{\alpha}$ where: $\overline{\alpha}_i = \mathrm{W}$ if $\alpha_i = \mathrm{L}$, $\overline{\alpha}_i = \mathrm{L}$ if $\alpha_i = \mathrm{W}$, and $\overline{\alpha}_i = e$ if $\alpha_i = e$.

Provided that α is not all e's, every ranking is described by either α or $\overline{\alpha}$ but not both (Prince 2002, page 42). In this way, ERC negation is just the standard notion of negation in three-valued logics. The opposition between α and $\overline{\alpha}$ makes a binary partition on the space of rankings. This is intuitively obvious for simple statements like $\langle \mathrm{W}, \mathrm{L}, e \rangle$ and $\langle \mathrm{L}, \mathrm{W}, e \rangle$. The opposition is a bit less intuitive for more complex conditions like $\langle \mathrm{W}, \mathrm{L}, \mathrm{L} \rangle$ and $\langle \mathrm{L}, \mathrm{W}, \mathrm{W} \rangle$, but the fact that $erc(a,b)$ is the antithesis of $erc(b,a)$ makes it abundantly clear (i.e., if a and b are not tied, then every ranking must prefer one or the other). The antithetical relationship between an ERC and its negation is reflected in the operation of fusion by the fact that fusing antithetical ERCs will always yield an ERC in $\mathcal{L}^+$.

3. The VCD of Elementary Ranking Conditions

Before turning to the question of the VC dimension of the sample space in OT, it will be helpful to define shatterability purely in terms of ERCs and thereby to establish a bound on the VCD of sets of ERCs. We will say that an ERC α is **true** of a given ranking $\mathcal{R}$ if the condition imposed by α is consistent with the linear ordering of the constraints defined by $\mathcal{R}$.

Definition 8
An ERC set Ω over constraints CON is **shatterable** iff for every subset $\Delta \subseteq \Omega$, there is a ranking $\mathcal{R}_{\mathrm{CON}}$ of which all ERCs in Ω–Δ are true while all the ERCs in Δ are false.

From this definition of shatterability for sets of ERCs, it is immediately clear that only nontrivial ERCs can occur in shatterable sets.

Lemma 1
Every ERC in a shatterable set must contain at least one L and one W.

Proof: The ERCs of $\mathcal{L}^+$ cannot occur in a shatterable set because there is no ranking of which they are true. Conversely, ERCs with no L's cannot occur in shatterable sets because there is no ranking of which they are false. □

With Definition 8 in hand, and having excluded the trivial ERCs from the picture, it will be possible to reduce shatterability for ERC sets to consistency under negation. First, a definition of negation for sets of ERCs.

Definition 9
A *partial negation* of ERC set Ω is obtained by negating every ERC in a subset $\Delta \subseteq \Omega$.

For example: $\Omega = \{\langle \mathrm{W}, \mathrm{L}, \mathrm{L}\rangle, \langle e, \mathrm{W}, \mathrm{L}\rangle\}$ has four partial negations: one per subset.

$$\left\{\begin{array}{l}\alpha = \langle \mathrm{W}, \mathrm{L}, \mathrm{L}\rangle \\ \gamma = \langle e, \mathrm{W}, \mathrm{L}\rangle\end{array}\right\} \quad \left\{\begin{array}{l}\overline{\alpha} = \langle \mathrm{L}, \mathrm{W}, \mathrm{W}\rangle \\ \gamma = \langle e, \mathrm{W}, \mathrm{L}\rangle\end{array}\right\} \quad \left\{\begin{array}{l}\alpha = \langle \mathrm{W}, \mathrm{L}, \mathrm{L}\rangle \\ \overline{\gamma} = \langle e, \mathrm{L}, \mathrm{W}\rangle\end{array}\right\} \quad \left\{\begin{array}{l}\overline{\alpha} = \langle \mathrm{L}, \mathrm{W}, \mathrm{W}\rangle \\ \overline{\gamma} = \langle e, \mathrm{L}, \mathrm{W}\rangle\end{array}\right\}$$

Theorem 1
An ERC set Ω is shatterable iff every partial negation of Ω is consistent.

Proof: Suppose every partial negation of Ω is consistent. Thus, for any partial negation in which Δ is the negated subset of Ω and Γ is the rest of Ω, it is the case that there is a ranking $\mathcal{R}$ of which all the ERCs of $\Delta + \Gamma$ are true. Because a nontrivial ERC and its negation are never both true of the same ranking and trivial ERCs cannot occur in shatterable sets, the ERCs in $\Omega - \Gamma$ are false of ranking $\mathcal{R}$ while the ERCs of Γ are true. Because Δ was arbitrary, it is the case that for every subset of Ω, there is a ranking of which the ERCs in that subset are false while the rest are true, and thus consistency under partial negation is sufficient for shatterability. If, on the other hand, there is a partial negation that is not consistent, then there is a subset of Ω such that if the ERCs in that subset are negated, the resulting $\Delta + \Gamma$ is not consistent. However, because there is no ranking of which the members of an inconsistent ERC set are all true, Ω is not shatterable because there is no ranking of which the ERCs in Γ are true while the ERCs in $\Omega - \Gamma$ are false. Thus, consistency under partial negation is both necessary and sufficient for shatterability. □

Corollary 1
Every subset of a shatterable ERC set is itself shatterable.

Proof: Because each partial negation of a shatterable ERC set must, by definition, be consistent and because every subset of a consistent set must also be consistent, it is the case that every subset of a shatterable set is consistent under every partial negation and is thus shatterable. □

Defining shatterability in terms of partial negation lines up with the commonsense observation that no set containing α and γ where $\alpha \rightarrow \gamma$ is shatterable because there can be no ranking of which the former is true while the latter is false. This is neatly captured by the fact that if $\alpha \rightarrow \gamma$, no superset of $\{\alpha, \gamma\}$ can be shattered because fusing $\{\alpha, \overline{\gamma}\}$ is guaranteed to yield an ERC in $\mathcal{L}^+$. The requirement of consistency under partial negation also shows why relatively weak conditions like $\langle \mathrm{W}, \mathrm{W}, \mathrm{L}\rangle$ and $\langle \mathrm{W}, \mathrm{L}, \mathrm{W}\rangle$ cannot co-occur in shatterable sets even though neither entails the other. In this case, fusing the

negation of both ERCs yields $\langle \text{L}, \text{L}, \text{L} \rangle \in \mathcal{L}^+$. This follows transparently from the fact that either the statement "$\mathbb{C}_1$ or $\mathbb{C}_2$ outranks $\mathbb{C}_3$" or the statement "$\mathbb{C}_1$ or $\mathbb{C}_3$ outranks $\mathbb{C}_2$" is true of any ranking of three constraints.

The definition of shatterability for ERC sets in terms of consistency under partial negation makes it easy to demonstrate that for $|\text{CON}| = k$, there are shatterable ERC sets of size $k - 1$. Diagonal ERC sets provide a particularly simple example of a class of shatterable ERC sets of this size.

Definition 10
ERC set Ω is **diagonal** if its members can be given as a list L^{Ω} in which each n^{th} ERC in the list has a W in its n^{th} coordinate, an L in its $n + 1^{\text{th}}$ coordinate, and *e*'s everywhere else.

$$\text{E.g., } \Omega = \left\{ \begin{array}{l} \langle \text{W}, \text{L}, e, e, e \rangle \\ \langle e, \text{W}, \text{L}, e, e \rangle \\ \langle e, e, \text{W}, \text{L}, e \rangle \\ \langle e, e, e, \text{W}, \text{L} \rangle \end{array} \right\}$$

Lemma 2
Diagonal ERC sets are shatterable.

Proof: Assume that Ω is a diagonal ERC set and Ω' is an arbitrary subset of an arbitrary partial negation of Ω. If n is the number of ERCs in Ω' then, by the definition of diagonal ERC sets, there must be at least $n + 1$ coordinates (columns) in Ω' that are filled with L or W for some ERC in Ω' (i.e., are not all-*e* columns). Because each of the n ERCs has only one L, at most n columns contain L's, thus the fusion of Ω' contains at least one W. Because Ω' was an arbitrary subset, no subset fuses to $\mathcal{L}^+$. Because the partial negation was arbitrary, every partial negation is consistent and thus Ω is shatterable. □

From the shatterability of diagonal ERC sets (with $k - 1$ members if $|\text{CON}| = k$), we obtain a lower bound of $k - 1$ on the VCD of ERC sets. Having established that there are shatterable sets of *k*-length ERCs with $k - 1$ members, what remains to be shown is that no set larger than $k - 1$ is shatterable.

Definition 11
Coordinate $\mathbb{C}_i$ is W-**unique** in ERC set Φ if Φ has a partial negation Φ' such that in the fusion of Φ', $\phi = \langle \phi_1, ..., \phi_k \rangle$, the only coordinate that contains a W is ϕ_i.

Definition 12
The minor $\Omega_{\alpha,j}$ of an ERC set Ω is a new set Ω' in which ERC α has been removed and the j^{th} coordinate has been removed from the remaining ERCs.

$$\text{For example, if } \Omega = \left\{ \begin{array}{l} \alpha : \langle \text{L}, \text{L}, \text{W}, e, \text{W} \rangle \\ \beta : \langle e, e, e, \text{L}, \text{W} \rangle \\ \gamma : \langle e, e, \text{W}, \text{L}, e \rangle \\ \delta : \langle \text{L}, \text{W}, e, e, e \rangle \end{array} \right\} \text{ then } \Omega_{\gamma,3} = \left\{ \begin{array}{l} \alpha : \langle \text{L}, \text{L}, e, \text{W} \rangle \\ \beta : \langle e, e, \text{L}, \text{W} \rangle \\ \delta : \langle \text{L}, \text{W}, e, e \rangle \end{array} \right\}$$

As illustrated in Definition 12, the term "minor" used here is analogous to the standard notion of the minor of a matrix. It is straightforward to show that every shatterable ERC set contains shatterable minors that can be obtained by removing one constraint's coordinate (column) and one ERC (row).

Lemma 3
Reduction Lemma. – If Ω is a shatterable ERC set, then it has a shatterable minor $\Omega_{\alpha,j}$.

Proof: By Corollary 1, for any $\alpha \in \Omega$, $\Omega - \{\alpha\}$ is shatterable. In $\Omega - \{\alpha\}$ there must be at least one coordinate $\mathbb{C}_j$ that is not W-unique. If this were not the case and *every* coordinate in $\Omega - \{\alpha\}$ was W-unique, then one of the L's in α would occlude the only W in a partial negation of Ω, making it inconsistent contra the assumption that Ω is shatterable (α must have at least one L by Lemma 1). Because $\mathbb{C}_j$ is not W-unique, for every partial negation of every subset of $\Omega - \{\alpha\}$, there is a coordinate other than $\mathbb{C}_j$ that fuses to W. This being the case, shatterability is preserved if $\mathbb{C}_j$ is eliminated. Thus, the minor $\Omega_{\alpha,j}$ is shatterable as required. □

Theorem 2
For $k > 1$, the largest shatterable set of k-length ERCs has $k - 1$ members.

Proof: If x is the size of the largest shatterable set of k-length ERCs and y is the size of the largest shatterable set of $(k+1)$-length ERCs, then y is not greater than $x + 1$. This must be so because if $y \geq x + 2$ then x could not be the size of the largest shatterable set of k-length ERCs because a set of $(k+1)$-length ERCs would have a shatterable minor larger than x. Because $\langle \text{W}, \text{L} \rangle$ and $\langle \text{L}, \text{W} \rangle$ are the only nontrivial ERCs for $k = 2$ and because they are antithetical and thus cannot co-occur in a shatterable set, the largest shatterable ERC set at $k = 2$ consists of a single ERC. This base case establishes an upper bound of $k - 1$ on the size of shatterable ERC sets and the diagonal ERC sets provide a lower bound of $k - 1$. Together these bounds place the cardinality of the largest shatterable set at exactly $k - 1$. □

Along with the diagonal ERC sets, there are many shatterable ERC sets with $k - 1$ members, but no shatterable sets with more than $k - 1$ members. What remains now is to connect this result for ERC sets back to the realm of candidates.

4. The VCD of Optimality Theory

The question posed at the outset of this article was: for a constraint set CON with k constraints that map candidates to natural numbers, what is the cardinality of the largest set of candidates S such that, for each subset $T \subseteq S$, there is at least one ranking $\mathcal{R}_{\text{CON}}$ under which every t in T is optimal, but no s in $S - T$ is optimal? Clearly, the answer to this question depends greatly on details of the constraints in CON. However, if we reduce candidates to the ERC sets associated with them, it is possible to place an upper bound on the size of S without knowing anything about CON other than its size k.

Recall that a candidate c is mapped to *True* by ranking $\mathcal{R}_{\text{CON}}$ just in case every ERC in $ERCS(c)$ is consistent with $\mathcal{R}$. Conversely, c is mapped to *False* by $\mathcal{R}$ if any of the ERCs in $ERCS(c)$ is not consistent with $\mathcal{R}$. This notion can be extended to sets of candidates as follows. If S is a set of candidates, then $ERCS(S)$ is the union of $ERCS(s)$ for all $s \in S$. A sample S is **accepted** by ranking $\mathcal{R}$ just in case every α in $ERCS(S)$ is consistent with $\mathcal{R}$. Conversely, S is **rejected** by $\mathcal{R}$ if any α in $ERCS(S)$ is not consistent with $\mathcal{R}$. Furthermore, if $ERCS(S)$ is consistent, then there must be at least one ranking that accepts S. In this case, we will refer to S as a **consistent sample**. The concepts of partial negation and W-uniqueness also have analogs for candidate sets.

Definition 13
For a consistent sample S, a **partial exclusion** is a partial negation of $ERCS(S)$ that rejects some $F \subseteq S$ by rendering $ercs(f)$ inconsistent for each $f \in F$ while preserving the consistency of $ercs(s)$ for every $s \in (S - F)$.

Definition 14
$\mathbb{C}_i$ is **w-unique** in S if there is a partitioning of S into T and F under which $\mathbb{C}_i$ is the only coordinate that fuses to W in $ERCS(T)$ for every partial exclusion that rejects F.

The property of W-uniqueness in samples crucially contrasts with what one might call being semi-unique—the case where for at least one, but not all, of the partial exclusions that reject F, $\mathbb{C}_i$ is the only column that fuses to W in $ERCS(T)$.

Definition 15
Given a constraint set CON and a sample S, the minor $S_{x,j}$ is obtained by removing candidate x from S and removing constraint $\mathbb{C}_j$ from CON.

By extending partial negation, W-uniqueness, and the concept of minors to the realm of samples, it is straightforward to show that shatterable samples have shatterable minors.

Lemma 4
Shatterable samples have shatterable minors.

Proof: Assume that S is a shatterable sample. Because removing candidate x from S has no effect on whether the remainder of S can be shattered, $S - \{x\}$ is also shatterable. Sample $S - \{x\}$ must have at least one coordinate that is not W-unique. If this were not the case, then, because $ERCS(x)$ must contain at least one coordinate with an L (else there would be no way to reject $\{x\}$), the presence of x in S would place an L in a W-unique coordinate in $S - \{x\}$. However, this would make it impossible to associate a ranking with at least one partitioning of S into accepted and rejected subsets contra the assumption that S is shatterable. If $\mathbb{C}_j$ is a coordinate in $S - \{x\}$ that is not W-unique, then, for every partial exclusion that rejects F, under each partitioning of $S - \{x\}$ into T and F, there is at least one other coordinate $\mathbb{C}_i$ that fuses to W. Thus, $\mathbb{C}_j$ could be removed from CON while preserving the shatterability of $S - \{x\}$. Therefore, $S_{x,j}$ is a shatterable minor as required. □

The crucial piece of the proof in Section 3 that the VCD of ERC sets is $k - 1$ was the illustration of a one-to-one relationship between k and the bound on shatterable sets by showing that removing an ERC from a shatterable set makes it possible to remove a coordinate from the remaining ERCs while preserving shatterability. If shatterable ERC sets could be larger than $k - 1$, then it would have been necessary to remove several ERCs before it was possible to safely remove a coordinate from the remaining ERCs. Because ERC sets and candidate samples both have shatterable minors, a similar strategy will show that shatterable samples must also grow at a one-to-one rate with k.

Theorem 3
If $|\text{CON}| = k$, then the size of shatterable sample sets is bounded at $k - 1$.

Proof: If $k = 2$, a sample consisting of a single candidate can be shattered if $ERCS(s)$ is $\{\langle \text{W}, \text{L}\rangle\}$ or $\{\langle \text{L}, \text{W}\rangle\}$, but no larger sample can be shattered. If there were such a sample, it would contain at least two candidates a and b and there would be a ranking under which both candidates were optimal, a ranking under which neither candidate was optimal, a ranking that made a but not b optimal, and another ranking that made b but not a optimal. This state of affairs requires at least four distinct rankings, which is impossible with only two constraints. Thus, it is established that, at $k = 2$, the largest shatterable sample set has at most one candidate.

If S is the largest shatterable sample for k constraints, then, at $k+1$, the size of the largest shatterable sample is $|S|+1$. If this were not the case, there would be a shatterable sample X such that $|X| \geq |S|+2$ for $k+1$ constraints. However, because shatterable samples have shatterable minors (Lemma 4), this would mean that there was a shatterable sample of size $|S|+1$ for k constraints, contrary to the assumption that S was largest. Given the base case that $|S|=1$ when $k=2$, the cardinality of shatterable samples is thus bounded at $k-1$ as required. □

The bound of $k-1$ defines the limiting case that is obtained when there can be candidates in the sample space for any ERC set. In actual practice, the specific details of the constraints in CON and the range of ways that they interact will determine which elements of the powerset of the set of k-length ERCs are associated with candidates in the sample space. This means that the VC dimension of a specific constraint set CON can be much lower than $|\text{CON}|-1$. Nonetheless, the result that the VCD of OT can be at most $|\text{CON}|-1$ is propitious for the learnability of Optimality Theoretic grammars.

5. Conclusions

Bounding the VC dimension of OT according to the number of constraints in CON establishes a general property of the sets of ranking hypotheses that can be associated with sets of candidates. This bound is independent of any assumptions about how the ERC sets for candidates are computed, independent of any assumptions about how optimizations are computed, and independent of any assumptions about the formal properties of constraints other than that they map candidates to $\mathbb{N}$.

The linear growth of the VCD with $|\text{CON}|=k$ provides a very general positive learnability result for OT. Blumer et al. (1989), building on the learning model of Valiant (1984), define a concept class C as **uniformly learnable** if there is a learning algorithm $\mathcal{A}$ such that, for any error threshold ϵ and confidence level δ, if $\mathcal{A}$ is given m training samples randomly drawn according to a probability distribution π over the sample space, then $\mathcal{A}$ has at least probability δ of generating a hypothesis whose likelihood of misclassifying any point in the sample space drawn randomly according to π is less than ϵ. Blumer et al. link the VC dimension to learnability by showing that concept classes are uniformly learnable if and only if they have a finite VCD. Moreover, they show that upper bounds on m can be established for learning that depend only on the VC dimension of the concept class to be learned. The bound on m according to d =VCD from Blumer et al. is given in Equation (1).

$$m \leq \left\lceil \frac{4}{\epsilon}\left(d \ln \frac{12}{\epsilon} + \ln \frac{2}{\delta}\right)\right\rceil \tag{1}$$

This is a worst-case bound that holds for the most adversarial probability distributions over the sample space and the worst consistent learning algorithms (i.e., algorithms that are consistent in that they correctly classify all data in the training set, but worst-case in that they err maximally on all unobserved data). Specific OT learning algorithms that have tighter bounds and non-worst-case probability distributions over samples will certainly present a different picture.

For a concrete example of OT learning, consider a version of Prince and Smolensky's (1993) basic CV syllable theory in which candidates are mappings from {C, V}* to {C, V, .}*, and for each input $i \in$ {C, V}*, the candidate set produced by GEN(i) represents

all ways of modifying i through deletion and insertion of **.**, C, and V.[3] If CON contains (i) a constraint against deletion, (ii) a constraint against V insertion, (iii) a constraint against C insertion, (iv) a constraint against syllables with codas, and (v) a constraint against syllables without onsets, then the range of possible rankings of these five constraints allows for 120 different grammars which in turn define twelve different languages (i.e., twelve subsets of the sample space $X = \{\text{C, V}\}^* \times \{\text{C, V, .}\}^*$).

If learners are trained with positive evidence in the form of optimal *input* → *output* mappings, then the probability distribution over the sample space can be characterized in terms of the probability distribution over the input strings in $\{\text{C, V}\}^*$. Each optimal candidate $a = (i \rightarrow o)$ provides information about the teacher's ranking in the form of $ERCS(a) = \{erc(a, b) | b = (i \rightarrow x) \in \text{GEN}(i)\}$. Riggle (2004) shows that, because the functions in this system are all rational (i.e., finite state representable), the set $ERCS(a)$ can be derived via an algorithm called CONTENDERS. In this system, $ERCS(a)$ can contain from zero to twelve ERCs. The zero-ERC cases arise for input strings that share the same optimal output under all rankings (i.e., /CV/→[.CV.]). The sets top out at twelve because there are never more than twelve contenders (i.e., non-harmonically-bounded candidates) for any given input string. The twelve ERC bound is a consequence of the fact the 120 rankings only realize twelve distinct languages.[4]

As noted in Section 1, candidates with the same input cannot co-occur in shatterable sets. Because of this, the bound on shatterable samples established in Section 4 carries transparently over to the more general case where learners are trained with optimal (i, o) mappings and then tested with novel inputs. Because the set of contenders is determined solely by GEN and CON (which the learner is presumed to have access to) if the learner can compute CONTENDERS(i), then testing on novel inputs reduces to having the learner select one optimal candidate from the set of contenders, which in turn reduces to binary questions of harmonic inequality between pairs a and b in CONTENDERS(i), which in turn reduces to the question of which of $erc(a, b)$ or $erc(b, a)$ is consistent with the ERCs gleaned from previous observations.

This is merely one sketch of how the learning problem in OT can be formulated so that the VC dimension can predict its success. There are undoubtedly other possible formulations. Furthermore, as noted, real-world cases will often contain details that are more relevant than the VC dimension in predicting learnability. For instance, in syllable structure grammar just described, there are inputs for which the CONTENDERS algorithm generates one candidate per language in the factorial typology. In such a case, the ERC set for a single optimal candidate can serve as a "global trigger" that is sufficient to uniquely identify the teacher's language. Further analysis with specific constraints and OT learning algorithms like Recursive Constraint Demotion (Tesar 1995, 1997, 1998; Tesar and Smolensky 1993, 2000), the Gradual Learning Algorithm (Boersma 1997, 1998; Boersma and Hayes 2001), and the ERC-Union learner (Riggle 2004) will surely yield further insights and a less abstract picture of learning in Optimality Theory.

The VCD is an extremely robust metric that characterizes hardness in many learning frameworks (Haussler, Kearns, and Schapire 1992) and is applicable without any assumptions other than that the learner is consistent. Any learner that bases its hypotheses on the union of the ERCs associated with the data on which it is trained is guaranteed to be consistent, and thus an extremely simple ERC-union learner can learn OT grammars

3 C and V represent consonants and vowels respectively and "**.**" represents a syllable boundary marker.

4 Riggle (2004) extends Prince and Smolensky's nine-way factorial typology to twelve with a slightly looser GEN. In this case, because $\lceil \log_2 12 \rceil = 4$, the ERC-based VCD bound is the same as that obtained by the finitude of the typology. Usually, however, we do not have the luxury of knowing the size of C.

from random training texts whose size m is linear in k. This linear bound on the relationship between k and sample complexity is a nice tightening of the $k \log_2 k$ bound that follows from the finitude of $k!$ and contrasts starkly with pessimistic assessments of learnability suggested by the factorial relationship between k and the number of possible grammars.

References

Anderson, Alan R. and Nuel D. Belnap, Jr. 1975. *Entailment - The Logic of Relevance and Necessity*. Princeton University Press.

Blumer, Anselm, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. 1989. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences*, 21:43–58.

Boersma, Paul. 1998. *Functional Phonology: Formalizing the Interactions between Articulatory and Perceptual Drives*. Ph.D. thesis, The Hague.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32:45–86.

Dijkstra, Edsger. W. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

Ellison, T. Mark. 1994. Phonological derivation in optimality theory. In *Proceedings of the Fifteenth Conference on Computational Linguistics*, pages 1007–1013, Kyoto, Japan. doi:dx.doi.org/10.3115/991250.991312.

Haussler, David, Michael Kearns, and Robert Schapire. 1992. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. Technical Report UCSC-CRL-91-44.

Littlestone, Nick. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318.

Prince, Alan. 2002. Entailed ranking arguments. ROA 500. Available at http://roa.rutgers.edu.

Prince, Alan and Paul Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell, Malden, MA.

Riggle, Jason. 2004. *Generation, Recognition, and Learning in Finite State Optimality Theory*. Ph.D. thesis, University of California, Los Angeles.

Samek-Lodovici, Vieri and Alan Prince. 1999. Optima. ROA 785. Available at http://roa.rutgers.edu.

Tesar, Bruce. 1995. *Computational Optimality Theory*. Ph.D. thesis, University of Colorado.

Tesar, Bruce. 1997. Multi-recursive constraint demotion. ROA 197. Available at http://roa.rutgers.edu.

Tesar, Bruce. 1998. Error-driven learning in Optimality Theory via the efficient computation of optimal forms. In *Is the Best Good Enough? Optimality and Competition in Syntax*, ed. Pilar Barbosa, Danny Fox, Paul Hagstran, Martha J. McGinnis, and David Pesetsky. MIT Press, Cambridge, MA.

Tesar, Bruce and Paul Smolensky. 1993. The learnability of optimality theory: An algorithm and some basic complexity results. Unpublished manuscript. Department of Computer Science & Institute of Cognitive Science, University of Colorado at Boulder.

Tesar, Bruce and Paul Smolensky. 2000. *Learnability in Optimality Theory*. MIT Press, Cambridge, MA.

Valiant, Leslie G. 1984. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vapnik, V. N. and A. Chervonenkis. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applicaions*, 16:264–280.