completed 12-20-2006

The complexity of 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 (i.e. set of input-output pairs) of which each subset is consistent with a different grammar hypothesis. This metric 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 OT with *k* constraints is *k*–1. This result establishes that the complexity of Optimality Theory is well behaved as a function of *k* and that the hardness of OT learning is linear in *k* for a variety of frameworks that employ probabilistic definitions of learnability.

1. Introduction

Given a constraint set CON containing *k* constraints in the framework of Optimality Theory (OT; Prince and Smolensky 1993/2004), what is the capacity of CON as a classification scheme for language data consisting of (*input*, *output*) pairs? In Optimality Theory each constraint is a function from (*i*, *o*) pairs to the natural numbers where the number that a constraint $\mathbb{C}$ assigns to a given (*i*, *o*) pair indicates how many times that pair violates $\mathbb{C}$. Given a constraint set CON, a grammar is a ranking $\mathcal{R}$ (a total ordering) of the constraints in CON and the language generated by $\mathcal{R}$ is the set of (*i*, *o*) pairs that are optimal according to $\mathcal{R}$ as defined in (1).

(1) (*i*, *o*) is *optimal* according to $\mathcal{R}$ if and only if no (*i*, *o*') is assigned fewer violations by the highest ranked constraint that assigns different numbers of violations to (*i*, *o*) and (*i*, *o*').

Because each of the *k*! rankings of CON is a different grammar that generates a potentially unique language of (*i*, *o*) pairs, it might be natural consider *k*! to be the classificatory capacity of CON. This is, however, a bit too simple. A metric that more directly reflects the complexity of the set of hypotheses that CON allows for a sample of language data is the size of the largest data set of which each subset corresponds to a different ranking hypothesis. To make this concrete we will call the sets of symbols from which inputs and outputs are built Σ and Θ, and the sets of all possible inputs and all possible outputs Σ* and Θ* respectively. If Σ*×Θ* is the *sample space* then what we are after is the cardinality of the largest sample set that is *shatterable* as defined in (2).

(2) Set *S* is *shatterable* by CON if and only if, for every division of *S* into two sets *In* and *Out*, there's a ranking $\mathcal{R}$ of CON that makes every $(i, o) \in In$ optimal but no $(i, o) \in Out$ optimal.

The idea of measuring the complexity of a concept class (in this case a set of grammars) in terms of the largest shatterable set of points in the sample space comes from the work of Vapnik and Chervonenkis (1971) and is known as the Vapnik-Chervonenkis Dimension (VCD). In the case of OT, the VCD of CON tells us the largest set of (*i*, *o*) pairs of which each subset yields a different ranking hypothesis and thus bounds the size of the set of hypotheses that can correspond to any data set. In section 3 I prove that the VCD of Con with k constraints is k–1. Before presenting this proof, I review in section 2 a formal representation scheme for encoding information about OT constraint rankings that will help in establishing the VCD of OT.

2. Elementary Ranking Conditions

Each (i, o) pair in the sample space can be described in terms of the constraint rankings under which it is optimal. Prince (2002) provides a scheme for encoding this kind of 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. For a complete exposition of the logic of ERCs see Prince (2002).

For Con containing k constraints, ERCs are k-length vectors that use the symbols E, L, and W to encode disjunctions of partial orderings of the constraints. Each constraint is arbitrarily assigned a 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 of the constraints whose corresponding coordinate contains a W outranks all of the constraints whose corresponding coordinates are filled with L's.

(3) E.g. $\langle$L, E, W, E$\rangle$ means that constraint $\mathbb{C}_3$ outranks constraint $\mathbb{C}_1$
$\langle$W, L, W, E$\rangle$ means that either $\mathbb{C}_3$ or $\mathbb{C}_1$ outranks constraint $\mathbb{C}_2$
$\langle$W, W, L, L$\rangle$ means that either $\mathbb{C}_1$ or $\mathbb{C}_2$ outranks both $\mathbb{C}_3$ and $\mathbb{C}_4$

The illustrative tableaux usually presented with OT analyses can be turned into sets of ERCs by making pair-wise comparisons between the violations for the designated (or observed) winner o and the violations for each other candidate o'. Each such comparison yields a single ERC α that describes the rankings under which o is preferred over o' as the output for input i. If o' has more violations than o for the i^{th} constraint then the α_i is W (a mnemonic for the fact that the winner is favored by $\mathbb{C}_i$). If o' has fewer violations than o of the i^{th} constraint then α_i is L (a mnemonic for the fact that that the loser is favored by the $\mathbb{C}_i$). If o and o' have the same number of violations of the i^{th} constraint then α_i is E (a mnemonic for the fact that o and o' are equivalent for $\mathbb{C}_i$).

(4)

input	Constraint 1	Constraint 2	Constraint 3	
☞***output* A**	*	**		– A violates $\mathbb{C}_1$ once and $\mathbb{C}_2$ twice
output B	W **	L *	E	← A beats B if $\mathbb{C}_1$ outranks $\mathbb{C}_2$
output C	E *	L	W **	A beats C and D if $\mathbb{C}_3$ outranks $\mathbb{C}_2$
output D	E *	L *	W *	
output E	W **	E **	E	(A beats E under every ranking)
output F	L	W ***	W *	← A beats F if $\mathbb{C}_3$ or $\mathbb{C}_2$ outranks $\mathbb{C}_1$

The comparison of A with E in (4) produces a 'ranking condition' that doesn't actually express a particular ranking (no constraint has an L) but instead indicates that $\mathbb{C}_1$ favors output A while no constraint favors output E. In this case output E is said to be *harmonically bounded* by A because there is no ranking under which E can be optimal when compared to A. If, on the other hand, E were designated the winner, then the ERC generated by comparing E with A would be ⟨L, E, E ⟩. Like ⟨W, E, E ⟩, this ERC does not encode a specific constraint ranking. Instead, ⟨L, E, E ⟩ indicates that the mere existence of A as an alternative means that no ranking will ever make E optimal.

Because ERCs encode disjunctions of conjunctions (i.e. [$\mathbb{C}_1$ or ... $\mathbb{C}_n$] outranks [$\mathbb{C}_{1'}$ and ... $\mathbb{C}_{n'}$]) it is possible to *weaken* an ERC by adding W's or removing L's. Adding W's to ERCs is the logical operation of disjunction introduction while removing L's is conjunction elimination. Because these operations are logically valid, weaker ERCs are entailed by their stronger counterparts. Put differently, $\alpha \rightarrow \beta$ iff each α_i entails β_i where L→E→W. The prose in (5) makes this obvious.

(5) ⟨W, L, L, E⟩ → ⟨W, E, L, E⟩ i.e. If $\mathbb{C}_1$ outranks $\mathbb{C}_2$ and $\mathbb{C}_3$ then $\mathbb{C}_1$ must outrank $\mathbb{C}_3$.
⟨W, E, L, E⟩ → ⟨W, E, L, W⟩ i.e. If $\mathbb{C}_1$ outranks $\mathbb{C}_3$ then $\mathbb{C}_1$ or $\mathbb{C}_4$ must outrank $\mathbb{C}_3$.
⟨W, L, L, E⟩ → ⟨W, E, L, W⟩ i.e. If $\mathbb{C}_1$ outranks $\mathbb{C}_2$ and $\mathbb{C}_3$ then $\mathbb{C}_1$ or $\mathbb{C}_4$ must outrank $\mathbb{C}_3$.

In addition to revealing entailments from strong to weak ERCs, the logic of ERCs shows how groups of weaker ERCs can combine to entail stronger ones. Prince (2002: 8) provides a logical operation on ERCs called *fusion* to derive entailments from sets of ERCs.

(6) The *fusion* of an ERC set Φ is a single ERC ϕ in which:

ϕ_i = L if any ERC in Φ has an L in its i^{th} coordinate,
ϕ_i = E if every ERC in Φ has an E in its i^{th} coordinate,
ϕ_i = W otherwise.

The operation of fusion can reveal nonobvious entailments among sets of ERCs. Consider the set Φ = {⟨W, W, E, L⟩, ⟨L, W, W, E⟩, ⟨W, E, L, W⟩}. The ERCs in Φ respectively denote that "$\mathbb{C}_1$ 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, ⟨L, W, L, L⟩, is stronger than each individual ERC in Φ and reveals that $\mathbb{C}_3$ outranks $\mathbb{C}_1$, $\mathbb{C}_2$, and $\mathbb{C}_4$.

Fusion can also reveal internal inconsistencies in ERC sets. For example, consider the set Ψ = {⟨W, L, W⟩,⟨L, W, W⟩,⟨W, W, L⟩}. Fusing Ψ produces ⟨L, L, L⟩, a strange 'ranking condition' that has no W. Prince refers to the class of ERCs with an L but no W's as $\mathcal{L}^+$ and shows that these ERCs arise from fusion just in case the ERCs in the fused set encode incompatible statements about the ranking of the constraints. In the case of Ψ, the ERCs are circular, so there can be no ranking of which they are all true. Fusion can thus be used to define the range of ERC sets that are free from internal contradictions as in (7).

(7) ERC set Φ is *consistent* if and only if no subset of Φ fuses to an ERC in $\mathcal{L}^+$.

For any consistent ERC set there is a constraint ranking (often several) of which all of its ERCs are true statements (Prince 2002: 21). The ERCs in an inconsistent set, on the other hand, cannot all be true of any ranking. Inconsistency can arise from a single comparison between candidates (e.g. the ERC set for E in (4) contains $\langle$L, E, E$\rangle$ because output A harmonically bounds E), and inconsistency can also arise from multiple comparisons. For instance, the ERCs that describe the rankings required in (4) for D to beat A and C are $\langle$E, W, L$\rangle$ and $\langle$E, L, W$\rangle$ respectively. But these fuse to $\langle$E, L, L$\rangle$, so no ranking allows D to simultaneously beat A and C – a situation that Samek-Lodovici and Prince (1999) call *complex harmonic bounding*. Finally, inconsistencies can arise across (i, o) pairs. If (i_1, o_1) is optimal under $\{\langle$W, L, W$\rangle\}$, (i_2, o_2) is optimal under $\{\langle$L, W, W$\rangle\}$, and (i_3, o_3) is optimal under $\{\langle$W, W, L$\rangle\}$ then, because these ERCs are inconsistent (they fuse to $\langle$L, L, L$\rangle$), there can be no ranking under which all three (i, o) pairs are simultaneously optimal.

Any ERC α can be negated to produce a statement that is true of the complement of the rankings of which α is true. This property can be exploited in describing the range of consistent ERC sets.

(8) The *negation* of α is $\overline{\alpha}$ where: $\overline{\alpha}_i$=W if α_i=L, $\overline{\alpha}_i$=L if α_i=W, and $\overline{\alpha}_i$=E if α_i=E. If α is not all E's then every ranking is described by either α or $\overline{\alpha}$ but not both (Prince 2002: 42).

The opposition between an ERC and its negation is intuitively obvious for simple statements like α=$\langle$W, L, E$\rangle$ and $\overline{\alpha}$=$\langle$L, W, E$\rangle$. For more complex statements such as γ=$\langle$W, L, L$\rangle$ and $\overline{\gamma}$=$\langle$L, W, W$\rangle$, however, the opposition is a bit less intuitive. (Working out the six rankings makes it clear.) The antithetical relationship between γ and $\overline{\gamma}$ is reflected in the operation of fusion by the fact that fusing an ERC and its negation will always yield in an ERC in $\mathcal{L}^+$.

3. The VC-Dimension of OT

The question that we started with was: for a constraint set CON containing k constraints that map (i, o) pairs drawn from $\Sigma^*\times\Theta^*$ to the natural numbers, what is the largest set $S\subseteq\Sigma^*\times\Theta^*$ such that for each subset $P\subseteq S$ there is at least one ranking $\mathcal{R}$ under which every (i, o) pair in P is optimal but no (i, o) in S–P is optimal? Clearly the answer depends greatly on details of the constraints in CON. However, now that we have associated ERC sets with (i, o) pairs, it is possible to place an upper bound on the size of S without knowing anything about CON other than its size k.

(9) An ERC set Φ expressed over constraints CON is *shatterable* if and only if for every subset $\Psi\subseteq\Phi$ there is a ranking $\mathcal{R}$ of which all ERCs in $\Phi-\Psi$ are true while all ERCs in Ψ are false.

From this definition it immediately follows that every ERC in a shatterable ERC set must have at least one L and one W. The ERCs of $\mathcal{L}^+$ are excluded because there is no ranking of which they are true and conversely ERCs with no L's are excluded because there is no ranking of which they are false. To bound shatterable sample sets it will be sufficient to bound the size of shatterable ERC sets. This can be done by exploiting the opposition between ERCs and their negations.

(10) A *partial negation* of an ERC set Φ is obtained by negating every ERC in a subset of Φ. E.g. If $\Phi=\{\langle \text{W, L, L}\rangle,\langle \text{E, W, L}\rangle\}$ there are four partial negations of Φ – one per subset.

The four partial negations of Φ: $\begin{Bmatrix}\alpha{:}\langle \text{W, L, L}\rangle \\ \gamma{:}\langle \text{E, W, L}\rangle\end{Bmatrix}$ $\begin{Bmatrix}\overline{\alpha}{:}\langle \text{L, W, W}\rangle \\ \gamma{:}\langle \text{E, W, L}\rangle\end{Bmatrix}$ $\begin{Bmatrix}\alpha{:}\langle \text{W, L, L}\rangle \\ \overline{\gamma}{:}\langle \text{E, L, W}\rangle\end{Bmatrix}$ $\begin{Bmatrix}\overline{\alpha}{:}\langle \text{L, W, W}\rangle \\ \overline{\gamma}{:}\langle \text{E, L, W}\rangle\end{Bmatrix}$

(11) **theorem 1**: Φ is shatterable if and only if every partial negation of Φ is consistent.

proof: Suppose that every partial negation of Φ is consistent. By the definition of partial negation, for each $\Psi\subseteq\Phi$ there is a partial negation Φ' in which Ψ is the subset of ERCs that are negated. Because Φ' is consistent, there is a ranking $\mathcal{R}$ of which all the ERCs of Φ' are true. Because an ERC and its negation are never both true of the same ranking (the all-E ERC cannot occur in shatterable sets), the un-negated versions of the ERCs in Ψ are false of ranking $\mathcal{R}$. Because this holds of every $\Psi\subseteq\Phi$ it's the case that for every subset of Φ 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 a sufficient condition for shatterability.

If, on the other hand, there is a partial negation of Φ that is not consistent then there is a subset $\Psi\subseteq\Phi$ such that if the ERCs of Ψ are negated the resulting Φ' is not consistent. But because there is no ranking of which the members of an inconsistent ERC set are all true, Φ is not shatterable because there is no ranking of which the un-negated counterparts of Ψ are false while the rest of the ERCs in Φ are true. Thus consistency under partial negation is a necessary condition for shatterability. □

(12) **corollary 1**: Every subset of a shatterable ERC set is itself shatterable.
Because each partial negation of a shatterable 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 common sense observation that no set containing α and γ where α entails γ can be shattered because there is no ranking of which α is true but γ false. This is captured by the fact that no superset of {α, γ} can be shattered because fusing $\{\alpha, \overline{\gamma}\}$ yields an ERC in $\mathcal{L}^+$ whenever $\alpha\rightarrow\gamma$. Consistency under partial negation also illustrates why relatively weak ERCs like $\varphi=\langle \text{W, W, L}\rangle$ and $\mu=\langle \text{W, L, W}\rangle$ cannot cooccur in a shatterable ERC set even though neither entails the other. In this case the fusion of $\overline{\varphi}=\langle \text{L, L, W}\rangle$ and $\overline{\mu}=\langle \text{L, W, L}\rangle$ is $\langle \text{L, L, L}\rangle$ which follows transparently from the fact that one of the statements "$\mathbb{C}_1$ or $\mathbb{C}_2$ outranks $\mathbb{C}_3$" or "$\mathbb{C}_1$ or $\mathbb{C}_3$ outranks $\mathbb{C}_2$" is true of any ranking of three constraints.

(13) **theorem 2**: Given shatterable Φ and Ψ, a shatterable set Θ can be built as follows: If Φ has m ERCs of length x and Ψ has n ERCs of length y, $p=\max(1,x)$, and $q=\max(1,y)$, then Θ contains $1+m+n$ ERCs where m of Θ's ERCs are Φ padded with p E's at the left side, n of Θ's ERCs are Ψ padded with q E's at the right side, and the final ERC of Θ is α with nonempty sets of W's and L's, one set in the first p and the other in the last q coordinates.

proof: Suppose Θ' is a partial negation of Θ and Ω is a subset of Θ'. For a contradiction, assume that the fusion of Ω is in $\mathcal{L}^+$. I'll call the ERCs that Φ and Ψ contribute Φ' and Ψ' respectively. It cannot be the case that Ω contains no ERCs from Φ' or Ψ' because then Ω would be $\varnothing$, $\{\alpha\}$, or $\{\overline{\alpha}\}$ none of which is in $\mathcal{L}^+$. Because Φ and Ψ are shatterable (i.e. on any partial negation no subset fuses to $\mathcal{L}^+$) if Ω contains ERCs from Φ' or Ψ' then the fusion of these ERCs has all E's in the first p coordinates and at least one W in the last q coordinates, or at least one W in the first p coordinates and E's in the last q coordinates. By the same logic, if Ω contains ERCs from both Φ' and Ψ' then the fusion of these ERCs has at least one W in both the first p and the last q coordinates. None of these alternatives is in $\mathcal{L}^+$ and adding $\{\alpha\}$ or $\{\overline{\alpha}\}$ cannot change this fact because both $\{\alpha\}$ and $\{\overline{\alpha}\}$ have at least one W and no L's in either the first p or last q coordinates. This exhausts the subsets of Θ' and thus contradicts the assumption that Ω fuses to $\mathcal{L}^+$. Because Ω was arbitrary, no subset of Θ' fuses to $\mathcal{L}^+$, so Θ' is consistent. Finally, because Θ' was an arbitrary partial negation of Θ, all partial negations of Θ are consistent and thus Θ is shatterable. □

(14) **corollary 2:** For $k \geq 2$ constraints, there are shatterable ERC sets with k–1 members.

Following the construction scheme in (13), the shatterable set $\{\langle \text{W, L} \rangle\}$ can be recursively extended by padding each ERC in the set with a new E-filed coordinate α_1 and adding a new ERC γ (the same length as the others in the set) in which γ_1=W, γ_2=L, and $\gamma_{n>2}$=E.

For example, if k=4 we can extend $\langle \text{W, L} \rangle$ three times to build Φ=
To illustrate the procedure, the padded E's are written as ***e***.

$$\Phi=\begin{Bmatrix}\langle \text{W, L, E, E, E} \rangle \\ \langle \boldsymbol{e}, \text{W, L, E, E} \rangle \\ \langle \boldsymbol{e}, \boldsymbol{e}, \text{W, L, E} \rangle \\ \langle \boldsymbol{e}, \boldsymbol{e}, \boldsymbol{e}, \text{W, L} \rangle\end{Bmatrix}$$

Having established that there are shatterable sets of k length ERCs with k–1 members, it remains to be shown that no set larger than k–1 is shatterable. Some new terminology will aid in this task.

(15) For the ERCs in a set Φ, the i^{th} coordinate is *W-unique* if and only if there is a partial negation of Φ under which α_i is the sole W in the fusion of Φ.

(16) The *minor* $\Phi_{\alpha j}$ of Φ is $\Phi-\{\alpha\}$ in which the j^{th} coordinate of each ERC has been removed.

It is straightforward to show that every shatterable ERC set contains shatterable minors obtained by removing a coordinate and an ERC. (The term *minor* is used analogously to the minor of a matrix.)

(17) **reduction lemma**: If Ψ is a shatterable ERC set then Ψ has a shatterable minor $\Psi_{\alpha j}$.

proof: For any arbitrary $\alpha \in \Psi$, by corollary 1, Ψ–$\{\alpha\}$ is shatterable. In Ψ–$\{\alpha\}$ there must be at least one coordinate $\mathbb{C}_j$ that is not W-unique because if this were not the case then one of the L's in α (of which there must be one because every ERC in a shatterable set has at least one W and at least one L) would lie in a coordinate that is W-unique in Ψ–$\{\alpha\}$ and as such Ψ would fuse to $\mathcal{L}^+$. Because $\mathbb{C}_j$ is not W-unique, for every partial negation of every subset of Ψ–{α} there is a coordinate other than $\mathbb{C}_j$ that fuses to W. This being the case, $\mathbb{C}_j$ can be removed from Ψ–{α} without affecting the shatterability of Ψ–{α} and thus the *minor* $\Psi_{\alpha j}$ is shatterable as required. □

(18) **theorem 3:** For $k\geq 2$, no shatterable set of k-length ERCs has cardinality greater than k–1.

proof: The largest shatterable set of k length ERCs cannot have more than one more member than the largest shatterable set of k–1 length ERCs because, if this were not the case then for some k the largest shatterable set of k–1 length ERCs has x members while largest shatterable set of k length ERCs has more than x+2 members. But this contradicts the assumption that x is the size of the largest shatterable set of k–1 length ERCs because we can derive a shatterable set of k–1 length ERCs with x+1 members via the reduction lemma from the set of k length ERCs, so the largest shatterable set at length k has at most one more member than the largest shatterable set at length k–1. {⟨W, L⟩} and {⟨L, W⟩} are the largest shatterable sets at k=2 and thus serve as a base case that establishes that the largest shatterable set of k-length ERCs has cardinality no greater than k–1 as required. □

Taking corollary 2 and theorem 3 together establishes that the largest shatterable sets of k-length ERCs have $k-1$ members. Because each (i, o) pair in the sample space maps to at least one ERC this bounds the VC-Dimension of rankings as a classification scheme over $\Sigma^*\times\Theta^*$ at |CON|–1.

4. Conclusions

ERCs are most useful for characterizing (i, o) pairs if they are efficiently computable. But, given that any element of Θ^* could be paired with i, generating the ERC set for any (i, o) is potentially hard. For OT in which all constraints are regular (expressible as finite state machines), Riggle (2004) gives an algorithm that generates sets of *contenders* – the non-harmonically-bounded (i, o) pairings for a given i. Under standard OT assumptions these sets are guaranteed to be finite and from them can be derived finite sets of ERCs that totally describe the rankings that make a given (i, o) optimal. The number of contenders can grow factorially with k in the worst case, so there is no guarantee that they can all be feasibly generated. Modulo the number of contenders, however, Riggle's algorithm efficiently generates ERC sets for the regular fragment of OT.

Bounding the VCD of OT according to the size of |CON| establishes a very general property of the ranking hypothesis space associated with sets of (i, o) pairs. This bound is independent of any assumptions about how ERC sets are computed (or that they are computable), 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 (i, o) pairs to $\mathbb{N}$.

Using only the linear growth of the VCD with k, it is possible to establish a very general positive learnability result for OT. In Valliant's (1984) Probably Approximately Correct (PAC) model of learning, a concept class (in our case rankings of CON) is said to be PAC-learnable just in case there is a learning algorithm $\mathcal{A}$ 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 probability at least δ of generating a hypothesis whose chance of misclassifying any point in the sample space drawn randomly according to π is less than ε. Haussler and Welzl (1987) and Blumer, Ehrenfeucht, Haussler, and Warmuth (1989), link the VC-Dimension to PAC learnability by showing that concept classes are PAC-learnable if and only if they have a finite VCD. Moreover, they show that bounds on *m* can be established for PAC 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* (1989) is given in (19).

$$(19)\qquad m \le \left\lceil \frac{4}{\varepsilon}\left(d \ln(12/\varepsilon) + \ln(2/\delta)\right) \right\rceil$$

This is a worst-case bound on *m* 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). Tighter bounds can be established for specific OT learning algorithms and this is an obvious next step in analyzing OT learning. Nonetheless, the VCD is an extremely robust metric that characterizes hardness in many learning frameworks (c.f. Haussler *et al* 1994) and is applicable with no assumptions other than that the leaner is consistent. Any learner that collects ERC sets for the training data is guaranteed to be consistent and thus a simple ERC-union learner can learn OT grammars from random training texts whose size *m* is linear in *k*. This linear relationship between *k* and sample complexity starkly contrasts with the factorial relationship between *k* and the number of possible grammars and shows that, in OT, the complexity of the set of grammar hypotheses that arise from sets of training data is surprisingly well behaved.

References

Blumer A, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. 1989. "Learnability and the Vapnik-Chervonenkis dimension." *Journal of the ACM*, 36(4):929--865

Haussler, D and E Welzl. *Epsilon-nets and simplex range queries*. Discrete and Computational Geometry, 2:127--151, 1987

Haussler, D, M. Kearns, and R. E. Schapire. 1994. "Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension." Machine Learning, 14(1):83--113.

Pitt, Leonard and Leslie G. Valiant. 1988. *Computational limitations on learning from examples*. Journal of the ACM, 35(4):965--984

Prince, Alan. 2002. Entailed ranking arguments. Ms., Rutgers University. ROA-500.

Prince, Alan, and Paul Smolensky. 1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell: Malden-Oxford-Carlton. Also available as ROA-537.

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

Samek-Lodovici, Vieri and Alan Prince. 1999. Optima. Rutgers Center for Cognitive Science, RuCCS-TR-57.

Valiant, Leslie. G. 1984. "A theory of the learnable." *Communications of the ACM 1984*, pp1134--1142.

Vapnik, Vladimir N. 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