Sampling Rankings

Jason Riggle, University of Chicago

jriggle@uchicago.edu

Draft 5/28/10, comments welcome

Abstract

In this work, I present a recursive algorithm for computing the number of rankings consistent with a set of optimal candidates in the framework of Optimality Theory. The ability to measure this quantity, which I call the r-volume, allows a simple and effective Bayesian strategy in learning: *choose the candidate preferred by a plurality of the rankings consistent with previous observations.* With k constraints, this strategy is guaranteed to make fewer than $k \log_2(k)$ mistaken predictions. This improves the k^2 bound on mistakes for Tesar and Smolensky's Constraint Demotion algorithm, and I show that it is within a logarithmic factor of the best possible mistake bound for learning rankings. Though, the recursive algorithm is vastly better than brute-force enumeration in vastly many cases, the counting problem is inherently hard (#P-complete), so the worst cases will be intractable for large k. This complexity can, however, be avoided if r-volumes are estimated via sampling. In this case—though it is never computed—the r-volume of a candidate is proportional to its likelihood of being selected by a given sampled ranking. In addition to polling rankings to find candidates with maximal r-volume, sampling can be used to make predictions whose probability matches r-volume. This latter mechanism has been independently used to model linguistic variation, so the use of sampling in learning offers a formal connection between tendencies in variation and asymmetries in typological distributions. The ability to compute r-volumes makes it possible to assess this connection and to provide a precise quantitative evaluation of the sampling model of variation. The second half of the paper reviews a range of cases in which r-volume is correlated with frequency of typological attestation and frequency of use in variation.

Sampling Rankings

Jason Riggle, University of Chicago

Draft May 28, 2010

1 Introduction

In this work, I present a strategy for computing the *number* of rankings that generate a set of optimal candidates in the framework of Optimality Theory (OT; Prince and Smolensky 1993/2004), a quantity I call the r-volume. Given a pair of candidates a b, the r-volume of a is simply the ratio $\frac{A}{A+B}$, where A is the number of constraints that *prefer* candidate a (i.e., those for which a has fewer violations) and B is the number that prefer b.

(1)

input	c_1	c_2	c_3	c_4	
candidate a		*	*	*	← the r-volume of a is 1/3 (i.e., 8 rankings)
candidate b	**		*		← the r-volume of b is 2/3 (i.e., 16 rankings)

In (1), a single constraint $\{c_1\}$ prefers candidate a while two constraints $\{c_2, c_4\}$ prefer b, thus the r-volume of a is 1/3 (and conversely the r-volume of b is 2/3).

The r-volume is somewhat similar to Tesar and Prince's (1999) r-measure, but instead of measuring the restrictiveness of a grammar in terms of the markedness violations it permits, r-volume quantifies the restriction imposed on the space of possible grammars (rankings) by sets of optimal candidates. r-volume is probably most familiar from models of variation like Antilla's (1997a) in which a pattern's frequency is proportional to its r-volume.

In this work, I link models of variation like Anttila's to learning models where r-volume influences predictions. I show how learning can connect frequency distributions observed in phonological variation with distributions of patterns observed in linguistic typology. The tools that I provide for measuring r-volume make it possible to evaluate the hypothesis that r-volume influences frequency using constraint sets that are too large to evaluate by enumeration of cases. One notable consequence of considering larger constraint sets is that they allow finer granularity of predictions in models of variation where r-volume $\propto$ frequency.

In §2, I review Prince's (2002) system for representing ranking information from tableaux in terms of Elementary Ranking Conditions (ERCs) and provide a recursive algorithm that operates over ERCs to compute r-volumes. Though r-volumes in pairs of candidates are easily calculated, computing r-volumes for sets of winners among many competitors can be exponentially more complex. Approximation by sampling can overcome this problem.

In §3, I show that access to r-volumes provides a simple and effective Bayesian heuristic for learning in OT: *All else equal, pick candidates that are preferred by a plurality of rankings consistent with previous observations*; this is basically Littlestone's (1988) *halving algorithm.*

If the hypothesis space $\mathcal{H}$ is finite—as it is in OT—the upper bound on mistakes for this strategy is logarithmic in the size of $|\mathcal{H}|$. With k constraints, $|\mathcal{H}| = k!$ can be quite large, however $\log_2(k!)$ is less than $k \log_2(k)$, so the mistake-bound of the halving algorithm is only *quasilinear* in k. This contrasts favorably with Constraint Demotion (Tesar & Smolensky 1993) whose mistake bound of k^2 is quadratic in k. Error bounds are illuminating because, in addition to limiting worst case performance, they can be translated into bounds on the number of random samples that need to be observed to guarantee that the learner arrives at a good hypothesis with high likelihood.

Unfortunately, this strategy is not practical for large constraint sets because the problem of measuring r-volume—like many combinatoric counting problems—is #P complete. To overcome this complexity I propose learning strategies that make predictions by sampling rankings from those consistent with previous observations. Sampling can be used in an *approximate* halving algorithm but it can also be used to make predictions with probabilities that match r-volume (rather choosing candidates that maximize it). The matching approach implements a prediction strategy for learning that shares essential properties with Anttila's model of variation. In the remainder of the paper this connection is explored in depth.

In §4, I discuss a range of cases where r-volume seems to be useful in predicting/modeling the frequency distributions of linguistic patterns. Throughout this section, I evaluate the fit between r-volume and frequency for sets of ERCs rather than stratified hierarchies, which have been the (sometimes tacit) object of analysis in most previous work on this issue. This is relevant because ERCs are more expressive than stratified hierarchies and this expressivity allows them to make a wider range of more finely grained frequency predictions.

2 Counting Rankings

In this section, algorithms are presented for computing r-volumes. In hopes of providing useful analytical tools, appendix A contains examples of several of the algorithms discussed here implemented in the programming language Python (available from `www.python.org`).

2.1 Definitions

In Optimality Theory, a *candidate* is an alignment of an underlying form and a surface form (i.e., a particular pairing between the units that make up the forms). A constraint c is a function from candidates to integers where $c(x)$ denotes the number of times that candidate x violates constraint c. Given two candidates x y, constraint c is said to *prefer* candidate x if $c(x) < c(y)$ and similarly a *ranking* (i.e., total ordering) of a set of constraints is said to *prefer* candidate x if at least one constraint that prefers x outranks all constraints that

prefer y. A *tableau* is a sequence of candidates with the same underlying form. In tableau T, candidate x is *optimal* under ranking $\mathcal{R}$ if no other candidate $y \in T$ is preferred by $\mathcal{R}$.

Given a set of k constraints indexed $[c_1, c_2, ..., c_k]$, the *violation profile* of candidate x is a vector of integers $\langle c_1(x), c_2(x), ..., c_k(x)\rangle \in \mathbb{Z}^k$. If each candidate in tableau T has a unique violation profile then every constraint ranking will select exactly one candidate as optimal (otherwise, T may contain ties). A candidate is a *contender* in tableau T if it is optimal under at least one ranking, and T is *complete* if it contains all and only the contenders for a given input form. Though the range of possible candidates may be infinite, there can be no more than $k!$ contenders with distinct violation profiles in any tableau T, and thus all tableaux consisting of contenders are guaranteed to be finite.[1]

If the candidate generating function *Gen* and all the constraints obey certain complexity restrictions then it is possible to algorithmically generate complete tableaux (Riggle 2004). Moreover, this can be done with algorithms whose complexity is linear in the number of contenders generated (Riggle 2009b). In general, however, it is most often the case that OT analyses use hand-crafted tableaux, so I make no assumption here that tableaux are complete and define r-volume purely in terms a given set of tableaux. As is true in all OT analyses, results obtained with incomplete tableaux must be taken with the caveat that they depend crucially on having considered all the relevant candidates.

Given a tableau, each candidate it contains can be described in terms of the rankings under which it is optimal. Prince (2002) gives a representation for this information called an Elementary Ranking Condition. ERCs are k-tuples of the symbols {L, e, W} in which the i^{th} component represents information about the relative ranking of constraint c_i. The meaning of an ERC is that at least one constraint associated with a W outranks *all* of the constraints associated with L's. Thus $\langle$ W e L L$\rangle$ means that c_1 outranks c_3 and c_4, while $\langle$ W W e L$\rangle$ means that c_1 or c_2 outranks c_4, and $\langle$ W W L L$\rangle$ means that either c_1 or c_2 outranks both c_3 and c_4. For a pair of candidates a b, an ERC describing the rankings under which a is optimal can be obtained as in (2).

$$(2) \quad erc(a,\ b) = \langle \varepsilon_1 \ldots \varepsilon_k \rangle \text{ where } \begin{cases} \varepsilon_i = \text{L} & \text{if } c_i(a) > c_i(b) \\ \varepsilon_i = e & \text{if } c_i(a) = c_i(b) \\ \varepsilon_i = \text{W} & \text{if } c_i(a) < c_i(b) \end{cases}$$

The symbols in $erc(a, b)$ are intended as a mnemonic: a W in the i^{th} component of the ERC indicates that constraint c_i prefers the *winner* (i.e., candidate a) while L indicates that c_i

[1] This follows from the fact that constraints penalizing epenthesis allow only finitely many candidates in T to share a violation profile. Moreover, even if such constraints are omitted and infinite-way ties do occur, the output of optimization will have to contain a finite representation of the range of forms sharing a given violation profile and no more than $k!$ representations of this kind can be contenders in T. Thus, OT presents a *combinatorial* optimization problem over a finite set of feasible solutions that happen to lie among an infinite range of possible solutions (as is often the case in combinatorial optimization).

prefers the *loser* and e indicates that the candidates are *equivalent* according to c_i. For some concrete examples, consider the sequence of tableaux in (3).

(3)

input 1	c_1	c_2	c_3	c_4	
candidate a				*	$\langle$ W e e L $\rangle$: a wins iff c_1 outranks c_4
candidate b	*				$\langle$ L e e W $\rangle$: b wins iff c_4 outranks c_1

input 2	c_1	c_2	c_3	c_4	
candidate c				*	$\langle$ W W e L $\rangle$: c wins iff c_1 or c_2 outranks c_4
candidate d	*	*			$\langle$ L L e W $\rangle$: d wins iff c_4 outranks c_1 and c_2

input 3	c_1	c_2	c_3	c_4	
candidate e		*		*	$\langle$ W L W L $\rangle$: e wins iff c_1 or c_3 outranks c_2 and c_4
candidate f	*		*		$\langle$ L W L W $\rangle$: f wins iff c_2 or c_4 outranks c_1 and c_3

Note how the ERCs for each candidate pair in (3) are mirror images with W swapped for L. This is the standard notion of negation in three-valued logic. If $erc(a, b) = \varepsilon$ then its logical antithesis $erc(b, a)$ is denoted $\bar{\varepsilon}$. If a and b have distinct violation profiles then ε and $\bar{\varepsilon}$ partition the set of rankings into two disjoint subsets, those that agree with the former and those that agree with the latter (i.e., the sum of the r-volumes is one in $\frac{A}{A+B} + \frac{B}{A+B} = 1$).

2.2 Computing r-volumes

In a tableau T with n candidates, a winner $x \in T$ provides up to $n-1$ distinct ERCs, one for each comparison of the winner to an alternative.

(4) $\quad ercs(x \,|\, T) = \{erc(x, y) : x, y \in T\}$

Given a sequence of tableaux $\mathcal{T} = [T_1 \ldots T_n]$, a *language* can be defined as a sequence $\ell = [l_1 \ldots l_n]$ consisting of one optimal candidate from each tableau where l_i is the winner in T_i for $1 \leq i \leq n$. The rankings that generate language ℓ are described by the ERCs in (5).

(5) $\quad ercs\,(\ell \,|\, \mathcal{T}) = \big\{erc(l_i, x) : 1 \leq i \leq n \text{ and } x \in T_i\big\}$

The r-volume of language ℓ is simply the number of rankings that generate ℓ divided by $k!$.

For example, if $\mathcal{T}$ is the sequence of tableaux in (3) and $\ell = [a, c, e]$ consists of the first candidate in each one, then $ercs(\ell \,|\, \mathcal{T}) = \big\{\langle \text{W e e L} \rangle, \langle \text{W W e L} \rangle, \langle \text{W L W L} \rangle\big\}$. This set can be simplified because $\langle$ W e e L $\rangle$ "c_1 outranks c_4" entails $\langle$ W W e L $\rangle$ "c_1 or c_2 outranks c_4". Thus the r-volume for ℓ in $\mathcal{T}$ can be computed from ERC set E in (6).

(6) $\quad E = \left\{ \begin{array}{c|c|c|c} \langle \text{W} & e & e & \text{L} \rangle \\ \hline \langle \text{W} & \text{L} & \text{W} & \text{L} \rangle \end{array} \right\}$

Each of the two ERCs in E is consistent with 12 rankings but, crucially, not the *same* twelve rankings. In fact, $r(E)$ is 9. This can be verified by enumerating and checking each of the 24 rankings, but computing $r(\cdot)$ by exhaustive search is not generally feasible given that 10 constraints allow 3.6 million rankings and 60 constraints allow 1.3×10^{81} rankings (a quantity on par with estimates of the number of atoms in the observable universe).

Stated informally, the r-volume of the ERC-set in (6) can be calculated as follows: First, note that if c_1 is top-ranked both ERCs are satisfied under all 6 rankings of the remaining 3 constraints. Second, note that if c_3 is ranked highest then $\langle \text{W L W L} \rangle$ is satisfied and thus it imposes no further conditions. This leaves $\{c_1, c_2, c_4\}$ which, according to the remaining ERC, can be satisfied by $\langle \text{W } e \text{ L} \rangle = \frac{1}{2} \cdot 3! = 3$ rankings. Finally, note that neither c_2 nor c_4 can be ranked highest and thus a total of $6 + 3 = 9$ rankings satisfy the conditions in E.

The computation of $r(E)$ can be reduced to two rules. The first rule is quite simple: for each constraint c_i where all ERCs have W's (i.e., $\forall \varepsilon \in E$, $\varepsilon_i = \text{W}$) there are $(len(E) - 1)!$ rankings, where $len(E)$ denotes the length of the ERCs. This follows from the fact that if an all-W column is ranked on top then the other columns can be in any order. The second rule is recursive: among the columns not covered by the first rule, for each c_i where no ERC has an L (i.e., $\forall \varepsilon \in E$, $\varepsilon_i \neq \text{L}$) there are $r(E_{-i})$ rankings, where E_{-i} is what remains after removing from E any ERC satisfied by ranking c_i at the top (i.e., $\varepsilon \in E$ for which $\varepsilon_i = \text{W}$), and then removing the column for c_i from the remaining ERCs.

(7) $$E_{-i} = \left\{ \langle \varepsilon_1 \ldots \varepsilon_{i-1}, \varepsilon_{i+1} \ldots \varepsilon_k \rangle : \varepsilon \in E \text{ and } \varepsilon_i \neq \text{W} \right\}$$

Because E_{-i} consists of just those conditions that remain to be satisfied after removing the conditions satisfied by ranking c_i at the top, constraint c_i is no longer relevant and reference to it can be removed from the remaining ERCs. In (8), I given an example that shows the process by which ERC set A yields A_{-3}.

(8) If $A =$

c_1	c_2	c_3	c_4	c_5
⟨ W	e	e	e	L ⟩
⟨ W	W	W	e	L ⟩
⟨ W	L	e	W	L ⟩
⟨ W	L	W	W	e ⟩

then $A_{-3} =$

c_1	c_2	c_4	c_5
⟨ W	e	e	L ⟩
⟨ W	L	W	L ⟩

Using this notation, r-volume is computed via the function in (9), as illustrated in (10).

(9) $$r(E) = \sum_{1 \leq i \leq k} \begin{cases} (len(E) - 1)! & \text{if } \varepsilon_i = \text{W for all } \varepsilon \in E \\ r(E_{-i}) & \text{if } \varepsilon_i \neq \text{L for all } \varepsilon \in E \text{ and } \varepsilon_i = e \text{ for some } \varepsilon \in E \\ 0 & \text{if } \varepsilon_i = \text{L for any } \varepsilon \in E \end{cases}$$

(10) For ERC set $A = \{\langle \text{W}\,e\,e\,e\,\text{L}\rangle, \langle \text{W}\,\text{W}\,\text{W}\,e\,\text{L}\rangle, \langle \text{W}\,\text{L}\,e\,\text{W}\,\text{L}\rangle, \langle \text{W}\,\text{L}\,\text{W}\,\text{W}\,e\rangle\}$, $\boldsymbol{r(A) = 45}$

$$A = \begin{array}{|c|c|c|c|c|}\hline c_1 & c_2 & c_3 & c_4 & c_5 \\ \hline \text{W} & e & e & e & \text{L} \\ \hline \text{W} & \text{W} & \text{W} & e & \text{L} \\ \hline \text{W} & \text{L} & \text{e} & \text{W} & \text{L} \\ \hline \text{W} & \text{L} & \text{W} & \text{W} & e \\ \hline \end{array} \boldsymbol{24+}$$

$$A_{-3} = \begin{array}{|c|c|c|c|}\hline c_1 & c_2 & c_4 & c_5 \\ \hline \text{W} & e & e & \text{L} \\ \hline \text{W} & \text{L} & \text{W} & \text{L} \\ \hline \end{array} \boldsymbol{6+} \qquad A_{-4} = \begin{array}{|c|c|c|c|}\hline c_1 & c_2 & c_3 & c_5 \\ \hline \text{W} & e & e & \text{L} \\ \hline \text{W} & \text{W} & \text{W} & \text{L} \\ \hline \end{array} \boldsymbol{6+}$$

$$A_{-3,4} = \begin{array}{|c|c|c|}\hline c_1 & c_2 & c_5 \\ \hline \text{W} & e & \text{L} \\ \hline \end{array} \boldsymbol{2+} \qquad A_{-4,2} = \begin{array}{|c|c|c|}\hline c_1 & c_3 & c_5 \\ \hline \text{W} & e & \text{L} \\ \hline \end{array} \boldsymbol{2+} \qquad A_{-4,3} = \begin{array}{|c|c|c|}\hline c_1 & c_2 & c_5 \\ \hline \text{W} & e & \text{L} \\ \hline \end{array} \boldsymbol{2+}$$

$$A_{-3,4,2} = \begin{array}{|c|c|}\hline c_1 & c_5 \\ \hline \text{W} & \text{L} \\ \hline \end{array} \boldsymbol{1+} \qquad A_{-4,2,3} = \begin{array}{|c|c|}\hline c_1 & c_5 \\ \hline \text{W} & \text{L} \\ \hline \end{array} \boldsymbol{1+} \qquad A_{-4,3,2} = \begin{array}{|c|c|}\hline c_1 & c_5 \\ \hline \text{W} & \text{L} \\ \hline \end{array} \boldsymbol{1}$$

$$\begin{array}{r} \boldsymbol{24+} \\ \boldsymbol{6+} \\ \boldsymbol{6+} \\ \boldsymbol{2+} \\ \boldsymbol{2+} \\ \boldsymbol{2+} \\ \boldsymbol{1+} \\ \boldsymbol{1+} \\ \boldsymbol{1} \\ \hline \boldsymbol{45} \end{array}$$

Because all references in A to c_1 are W, there are at least $4! = 24$ rankings consistent with A. There are no L's for c_3 or c_4, which adds $r(A_{-3}) + r(A_{-4})$ rankings to $r(A)$. These each add $3! = 6$ rankings and yield $A_{-3,4}$, $A_{-4,2}$, $A_{-4,3}$ (extending the notation in the obvious way), which each add $2! = 2$ rankings and yield another set with 1 more ranking, for a total of 45.

This computation can be made more efficient by using the closed form for single ERCs, removing entailed ERCs, and storing partial results to avoid duplicate computations, which all seem to keep the complexity fairly tame for $k \leq 40$. Nonetheless, in the worst cases the depth of recursion can grow factorially with k. In fact, under the P $\neq$ NP assumption, no algorithm can efficiently compute $r(\cdot)$ in all cases because ERCs can encode partial orders and the problem of counting linear extensions of partial orders has been shown to be #P-complete (see, Brightwell and Winkler 1991).[2] Fortunately, in scenarios where an exact count is not necessary, the number of linear extensions of a partial order *can* be efficiently approximated using the methods of Huber (2006). Approximation will be discussed in §3.4.

The computation in (9) is vastly better than a brute force search in vastly many cases. The fact that exact computation can be infeasible in the worst cases is not that problematic for the use of r-volume as a tool for understanding distributions of phenomena such as those in §4 provided that cases of interest are within the limits of resources available to researchers. On the other hand, if r-volume is to be used by humans (or algorithms) in producing or learning language then the ability to efficiently compute or approximate it is essential.

[2] #P is a complexity class introduced by Valiant (1979) for counting problems that is analogous to the NP class of decision problems (but note that solutions to easy decision problems can be #P-hard to count).

3 Learning

Consider an 'on-line' learning scenario in which a learner tries to predict optimal candidates in a sequence of tableaux $\mathcal{T}$ where, after each prediction, the correct candidate is revealed and the knowledge that the learner uses to make future predictions may be updated. In this framework, the *mistake bound* of a learning strategy is defined as the greatest number of mistakes it can make for any sequence $\mathcal{T}$. For a given class of problems—such as learning languages over tableaux—a strategy is said to have an *optimal mistake bound* if no other strategy can possibly have a better mistake bound (Littlestone 1989). Mistake bounds are illuminating because they reveal general limits on the efficacy of learning strategies, including whatever strategies humans learners actually use (an interesting but independent issue).

3.1 Choosing popular candidates

In predicting optima, the strategy of choosing whichever candidate is preferred by a plurality of the rankings consistent with previous observations yields a nearly optimal mistake bound. If Y is the set of ERCs for previously observed optima, and $X = ercs(x \mid T)$ are the ranking conditions under which candidate x is optimal in tableau T, then this strategy amounts to choosing the candidate with the largest *conditional* r-volume: $r(X|Y) = r(X \cup Y)/r(X)$. The predictions made by this strategy are illustrated for a sequence of tableaux in (11).

(11)

T_1: input-1	c_1	c_2	c_3	c_4
candidate a				*
candidate b	*	*		

← Choose candidate a with r-volume of $\frac{16}{24} = \frac{2}{3}$.
For $E = \varnothing$, a's conditional r-volume is also $\frac{2}{3}$.

Suppose this prediction is wrong and candidate b is optimal. Once this error is revealed, E is updated to $\{\langle \text{L L } e \text{ W}\rangle\}$ and only 8 of the 24 possible rankings remain viable.

T_2: input-2	c_1	c_2	c_3	c_4
candidate c			*	
candidate d	*			

← Choose d, whose conditional r-volume is $\frac{5}{8}$.

Though $r(d \mid T_2) = 1/2$, 5 of the 8 rankings consistent with E prefer d, so it is chosen. Assuming this is wrong, E is set to $\{\langle \text{L L } e \text{ W}\rangle, \langle \text{W } e \text{ L } e\rangle\}$ and 3 viable rankings remain.

T_3: input-3	c_1	c_2	c_3	c_4
candidate e		*		
candidate f	*		*	

← Choose e, whose conditional r-volume is $\frac{2}{3}$.

Two of the three viable rankings prefer candidate e. If predicting e is an error then E is set to $\{\langle \text{L L } e \text{ W}\rangle, \langle \text{W } e \text{ L } e\rangle, \langle \text{L W L } e\rangle\}$ and only one viable ranking remains.

This approach implements a well-known on-line learning strategy called the *halving algorithm* (so named by Littlestone 1988, but also see Angluin 1988, Barzdin & Freivald 1972, and Mitchell 1982 for foundational ideas). Though the strategy in (11) uses ERCs and recursively computes r-volumes, a brute-force approach could make exactly the same predictions by starting with a list of all rankings and deleting those that prefer the loser after each tableau. Regardless of implementation, the strategy of choosing candidates preferred by a plurality of the rankings consistent with previous observations can never make more than $\log_2(k!)$ erroneous predictions. This follows from the fact that, whenever a mistake is made, the candidate that turns out to have been optimal will be consistent with half or fewer of the viable rankings else it would have been chosen, so each correction will reduce the viable rankings by at least half.

This strategy can be thought of as choosing the candidate that is most likely if the prior probability over rankings is uniform. For a given pair of candidates a b, Bayes' rule defines the posterior probability for $X = \{erc(a,b)\}$ given $Y = ercs(\ell_{[1\,\ldots\,n-1]} \,|\, \mathcal{T}_{[1\,\ldots\,n-1]})$ (i.e., the probability that candidate a is optimal given the ERCs for the previous optima). Under the uniform prior, the probability of an ERC set is just its r-volume: $P(E) = r(E)$ and the conditional probability $P(X|Y)$ is the conditional r-volume $r(X|Y)$, i.e., the fraction of the rankings consistent with the latter that are also consistent with the former.

$$(12)\quad \underbrace{P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}}_{\textit{Bayes' Rule}} \;=\; \frac{\frac{P(X\wedge Y)}{P(X)}\,P(X)}{P(Y)} \;=\; \frac{P(X\wedge Y)}{P(Y)} \;=\; \frac{r(X\cup Y)}{r(Y)} = r(X|Y)$$

For example, suppose the ERCs for previously observed optima are $Y = \{\langle \text{W L L } e\rangle\}$ with an r-volume of $1/3$ and the learner sees tableau T in (13) where $X = ercs(a,T) = \{\langle \text{W L } e\, e\rangle\}$ has an r-volume of $1/2$. Among the twelve rankings consistent with X there are eight that are also consistent with Y, so $P(Y|X) = \frac{1/3}{1/2} = 2/3$, and $P(X) = 1/2$, making $P(X|Y) = 1$.

(13)

input	c_1	c_2	c_3	c_4
candidate a		*		
candidate b	*			

$$P(X|Y) = \frac{\frac{r(X\cup Y) = 1/3}{r(X) = 1/2}\cdot r(X) = 1/2}{r(Y) = 1/3} = 1$$

The probability $P(X|Y) = 1$ reflects the fact that $\langle \text{W L L } e\rangle$ entails $\langle \text{W L } e\, e\rangle$. Conversely, $P(\neg X|Y) = 0$ because $\neg X$ and Y are contradictory (i.e., $r(erc(b,a)\wedge Y) = 0$).

Under this approach, the prior expectation of language ℓ over tableaux $\mathcal{T}$ is just $r(\ell \,|\, \mathcal{T})$. Because different languages can have vastly different r-volumes, this constitutes a substantive bias in learning (albeit one highly dependent on the constraints and candidates in $\mathcal{T}$). In §4.2, I discuss methods that can simulate the effects of nonuniform priors.

3.2 The r-volume learner

Another useful property of the halving algorithm is that it can provide an upper bound on the amount of information learners need to store to make good hypotheses. Using ERCs already imposes a significant 'dimensionality reduction' on the problem by reducing all knowledge about optima to finite sets of ranking conditions. Even so, there are 3^k ERCs and, even if all redundancies and entailments are factored out using methods like those of Prince & Brasoveanu (2005), the size of the resulting ERC set can still be exponential in k.[3]

Fortunately, ERCs need not be stored for *all* optima. The $\log_2(k!)$ bound on mistakes can double as a bound on memory by storing ERCs just in cases of error and using only these to make predictions. To formalize this strategy, a measure of *r-inequality* will be useful.

(14) **Prediction by r-inequality**
$a \geq_E b \iff r(\{erc(a,b)\} \cup E) \geq r(\{erc(b,a)\} \cup E))$

If the ERCs in E make one candidate optimal, r-inequality reduces to *harmonic inequality*, but otherwise, relative r-volume will adjudicate when optimality is underdetermined. If predictions are made in each tableau T (recall that T is a list of candidates) by choosing the first candidate x such that $x \geq_E y$ for all $y \in T$, then any erroneous predictions can be corrected one ERC at a time, and consequently the size of E can be bounded at $\log_2(k!)$. Combining predictions based on r-inequality with the rule of updating E only on errors yields an on-line error-driven learning strategy that I will refer to as the r-volume learner.

(15) **The r-volume learner (RVL)**
For each T in $\mathcal{T}$, return the first $x \in T = [x_1 \ldots x_n]$ where $x \geq_E x_i$ for all $1 \leq i \leq n$.
On an error, if $y \neq x$ is optimal, reassess T adding $erc(y,z)$ to E whenever $z \geq_E y$.

In general E will start as an empty list, but initial ranking conditions can also be imposed (in which case the bounds given here hold for additions to E). Adding ERCs that are consistent with half or fewer of the viable rankings one-at-a-time guarantees that the size of E will not exceed $\log_2(k!)$. It would be simpler to add $ercs(x \mid T)$ to E on each mistake, but this could result in the addition of nearly $k!$ conditions to E in the most extreme case.[4]

[3] For instance, if there are 9 constraints then the set E of 8-choose-4 = 70 ERCs that split W and e over 8 positions and share an L in the 9^{th} cannot be reduced (i.e., $\forall \varepsilon \in E$, $\forall F \subseteq E/\{\varepsilon\}$ members of F have a W for some c_i where $\varepsilon_i = e$), and similar sets can be constructed for all of the central binomial coefficients.

[4] RVL can be used as a loss-less compression scheme for $\mathcal{T}$ because the optimal candidates of $\mathcal{T}$ can be recovered from the ERC list produced by RVL. For example, if E_1 is the output of RVL for tableau-sequence $\mathcal{T}$, then the optimal candidates can be reconstructed by running RVL again on $\mathcal{T}$. Call the ERC-list for this 'reconstruction' step E_2 and, as usual, let it start out empty. During reconstruction, each time RVL is about to return a candidate $x \in T_i \in \mathcal{T}$, the ERC for each competitor $erc(x,y)$ for $y \in T_i$ is checked against ε, the first ERC in E_1. There will be a conflict just in case RVL is about to make a bad prediction based on the accumulated conditions in E_2—this is precisely the point where ε was added to E_1—so instead of returning that candidate, ε is moved from E_1 to E_2, and a new prediction is made which is checked the same way.

The mistake bound and memory bound for RVL are both within a logarithmic factor of the absolute lower limits set by the combinatoric complexity of OT as measured by the Vapnik-Chervonenkis Dimension (Vapnik and Chervonenkis 1971). Roughly speaking, the VC dimension measures limits on the independence in data for a set of classifiers (in this case, rankings) in terms of the largest set of data where all classifications are possible. Riggle (2009a) showed that the VC dimension of OT is $k-1$. What this means in concrete terms is that there are sets of $k-1$ tableaux in which every possible pattern of winners is realized by at least one ranking, but there is no set of k or more tableaux that have this property. Suppose that $\mathcal{T}_4$ consists of the three tableaux in (16).

(16)

input 1	c_1	c_2	c_3	c_4
cand. a		*		
cand. b	*			

input 2	c_1	c_2	c_3	c_4
cand. c			*	
cand. d		*		

input 3	c_1	c_2	c_3	c_4
cand. e				*
cand. f			*	

In $\mathcal{T}$, each of the 2^3 choices of winners is supported by a ranking, but there is no set of four binary tableaux in which all of the 2^4 choices are supported. Proving the upper bound is a bit complicated (see Riggle 2009a), but for the lower bound, $\mathcal{T}_4$ demonstrates that the VC dimension of OT is at least $k-1$ at $k=4$ (with an obvious extension to all $k \geq 2$).

Examples like $\mathcal{T}_4$ make it easy to see why the VC dimension sets a lower limit on mistake and memory bounds for learners. For k constraints it is simple to construct a sequence $\mathcal{T}_k$ of $k-1$ tableau following the pattern in (16), and, because each of the 2^{k-1} ways to choose winners is supported by a ranking, the target language could be chosen by flipping a fair coin for each tableau. In this scenario, knowing the optima in the first $n-1$ tableaux provides no information about the winner in the n^{th} tableau and thus no learning strategy can guarantee fewer than $k-1$ mistakes or the need for fewer than $k-1$ bits of information to represent optima in $\mathcal{T}_k$. Compared to $\frac{k^2-k}{2}$, the bound of $k \log k$ looks more like $k-1$ the higher k gets, so the $k \log k$ mistake and memory bounds of RVL are very nearly optimal.[5]

It is important to note that specific cases can have mistake bounds drastically lower than these general worst-case bounds. For instance, preconditions on rankings or the use

[5]Littlestone (1988) proposes another algorithm he calls the Standard Optimal Algorithm (SOA) that can outperform the halving algorithm. The SOA achieves the absolutely minimal mistake bound by enumerating all binary decision trees over future predictions and choosing candidates so as to maximize the size of the remaining decision tree. This is analogous to choosing candidates that maximize the set of viable hypotheses (i.e., those with maximal r-volume). This strategy bounds the mistakes for the SOA to the height of the tallest complete binary decision tree. Ben-David et al. (2009) call this quantity the *Littlestone Dimension* (LD) of a learning problem. For any finite hypothesis space $\mathcal{H}$, $\text{VCD}(\mathcal{H}) \leq \text{LD}(\mathcal{H}) \leq \lceil \log_2(|\mathcal{H}|) \rceil$. If there is no structure in $\mathcal{H}$ these three are equal (for instance, if $\mathcal{H}$ is all subsets of a finite set). The hypothesis space in OT is highly structured and thus the inequalities are strict. For example, when learning rankings of $k=5$ constraints, the VCD is four, $\lceil \log_2(k!) \rceil$ is six, and an exhaustive search over decision trees for 5 constraints reveals that the LD is five. Enumerating decision trees is much more computationally costly than the halving algorithm, which is itself already impractical for large k. Though the SOA can shave a bit off the mistake bound of RVL, the improvement is a logarithmic factor at best.

of constraints that lie in a stringency relationship can reduce the set of distinct rankings at the outset of learning. Furthermore, it is possible for a given 'lexicon' to lack a set of underlying forms that achieves the maximal mistake bound for a given constraint set. In such cases, performance can be much better. Happily, the general worst-case mistake bounds are already rather tame, and there is no way for them to get worse.

3.3 Comparison with Constraint Demotion strategies

One of the best known learning strategies for OT is the family of constraint demotion algorithms developed by Tesar, Smolensky, and a few others (Tesar 1995, 1997, 1998, Tesar & Smolensky 1993, 1996, 2000, Prince & Tesar 2004, Boersma 2009). Constraint demotion is a method whereby ranking information is used to update a *stratified hierarchy* $\mathcal{H} = [H_1 \ldots H_n]$ that consists of $n \leq k$ nonempty partitions of a set of k constraints where the constraints in stratum H_i outrank the constraints in H_{i+1} but are unranked with respect to each other. Technically, $\mathcal{H}$ is a *weak order*, a partial order where incomparability is transitive: if $a \sim b$ (i.e., a and b are in the same stratum) and $b \sim c$ then $a \sim c$. The constraint demotion rule for updating $\mathcal{H}$ given by Tesar (1995:84) can be stated in terms of ERCs as in (17).

(17) **Constraint Demotion (CD)**
Given an ERC ε and a stratified hierarchy $\mathcal{H}$, if H_w is the highest $H \in \mathcal{H}$ containing at least one constraint c_i such that $\varepsilon_i = \text{W}$ then any c_j in a stratum above H_w such that $\varepsilon_j = \text{L}$ is demoted to the stratum below H_w (creating a new stratum if needed).

An appealing property of CD is that at most $k^2 - k$ demotions can be made in response to ERCs consistent with a given ranking (see Tesar & Smolensky 1996:56). To obtain a mistake bound for on-line learning all that is needed is a rule for using the learner's current $\mathcal{H}$ to make predictions.

The only complication is that whenever $\mathcal{H}$ is less than a total order it does not impose a total order on violation profiles, so a strategy for predicting optima is needed in such cases. Tesar (1995:96) suggests that candidates be selected by *pooling* violations within strata.

(18) **Prediction by stratal inequality**
Denoting the sum of $c(a)$ over constraints in H as $H(a)$, for a pair of candidates $a\,b$, $a \geq_{\mathcal{H}} b$ if every H_j where $H_j(a) \geq H_j(b)$ is below a stratum H_i where $H_i(a) \leq H_i(b)$.

Pooling has the advantage that it will deterministically choose a candidate in many cases of uncertainty—though ties are still possible—but it has a few drawbacks discussed §3.4 along with Boersma's (2003, 2009) suggestion of sampling as an alternative to pooling.

Using the pooling strategy, Tesar (1995:95) defines Error-driven Constraint Demotion (EDCD), an error-driven strategy whereby the grammaticality of an input-output pair (x, y)

is evaluated by comparing it to the candidate (x, z) that is optimal under the learner's current $\mathcal{H}$, rather than, say, all contenders for x. In actual practice, however, most analyses of OT leaning have used fixed sets of candidates rather than algorithmically generated optima. Given a fixed set of candidates, EDCD simply chooses the one that is minimal by stratal inequality (see, for instance, Boersma 2009).

In all the variants of CD, if $\mathcal{H}$ is starts out as a single stratum with all the constraints, it is only possible to make $\frac{k^2-k}{2}$ changes on the way to a total ranking and thus EDCD has a quadratic mistake bound. Though it is pleasingly polynomial, this bound is significantly worse than the quasilinear mistake bound of RVL. In terms of memory requirements, RVL has a similar advantage. Though EDCD's stratified hierarchies are quite parsimonious, they are a 'lossy' representation in the sense that accidentally correct predictions can be undone by subsequent observations and thus a stratified hierarchy can make mistakes on the very sequence of examples that created it. This can be remedied by storing each ERC that causes an update to $\mathcal{H}$ (as in Tesar 1997) and using a strategy like the one in *fn.* 4. However, such an approach makes ECDC's memory requirements match its mistake bound. The area where EDCD outshines RVL is computational complexity; the worst-case complexity of the former grows linearly with k while the worst-case complexity of the latter grows factorially.

3.4 Sampling rankings

One way to avoid the worst-case complexity of RVL is to make predictions based on estimates of r-volume rather than exact counts. In fact, all the learner really needs to do is fairly reliably select candidates that are preferred by more of the rankings consistent with the known ERCs. This can be done by *sampling*. The strategy in (19) produces steps for a random walk over the set of rankings consistent with a given ERC set.[6]

(19) $\mathcal{R}$-STEP: Given an ERC set E and a ranking $\mathcal{R} = [c_1 \ldots c_k]$ consistent with E,
generate $\mathcal{R}' = [\ldots c_i, c_j \ldots]$ by taking a random transposition of $\mathcal{R} = [\ldots c_j, c_i \ldots]$.
If $\mathcal{R}'$ is consistent with E, return $\mathcal{R}'$ with probability $\frac{k-1}{k}$, otherwise return $\mathcal{R}$.

This random walk is a particularly simple instance of Markov chain Monte Carlo (MCMC) sampling using the Metropolis-Hastings algorithm. MCMC is remarkable in that it allows one to sample uniformly from a set that has an unknown normalizing constant over the members' probabilities (i.e., to sample uniformly from E without knowing $r(E)$).

The effectiveness of using MCMC sampling to estimate probabilities in Bayesian inference problems depends on the *mixing time* of the Markov chain (e.g., how many $\mathcal{R}$-STEPs it takes

[6] Conceptually, this can be thought of as a random walk that traverses a *permutohedron*: a graph with a vertex for each permutation of a set of elements and an edge between each pair of vertices that differ by a single transposition of a pair of adjacent elements.

for the system to 'forget' the ranking it started with). In some cases, such as the *bounding chains* introduced by Huber (2006) to sample linear extensions of partial orders, it is possible to prove that the chains mix rapidly. This allows Huber's method to efficiently generate random samples drawn perfectly uniformly from the set of linear extensions.

Though some ERC sets represent partial orders, ERCs are actually a generalization of partial orders best characterized by what is known as an *antimatroid* in the combinatorics literature. It may be possible to extend a variant of Huber's bounding chains approach to sample from antimatroids, but as of yet no uniform sampler has been proven to be efficient for this case. If such an extension is possible or if learners are restricted to ERC sets that encode partial orders, an efficient uniform sampling algorithm can be used to implement an efficient *approximate halving algorithm* (Goldman et al. 1989) that is guaranteed to make fewer than $k \lg k + (\lg e) \lg k$ mistakes with hight probability.[7]

In lieu of formal bounds on mixing time, heuristics are often used to guide MCMC (see Resnik and Hardisty 2009 for an accessible introduction to the practical details). In order to provide a general sense of the effectiveness of sampling, the performance of learners who draw samples of various fixed sizes will be considered. Learners that estimate r-volume by sampling will be denoted RVL^n with n indicating the use of k^n samples.

(20) **$\text{RVL}^{\boldsymbol{n}}$**: pick the candidate preferred by more rankings output by k^n calls to $\mathcal{R}$-STEP (after $2 \times k$ steps starting from $\mathcal{R}_S$ built by adding a random member of $\{c_i : \forall \varepsilon \in E, \varepsilon_i \neq \text{L}\}$ to the bottom of $\mathcal{R}_S$ and setting E to E_{-i} until $E = \varnothing$ then adding any c not yet in $\mathcal{R}_S$).

The fine-print in (20) specifies an initial *burn in* period where the rankings from $2 \times k$ steps are discarded in hopes of eliminating bias introduced by the process that generates the initial 'seed' ranking $\mathcal{R}_S$ which is constructed by adding one constraint at a time to the bottom of the hierarchy in a fashion that does not generate rankings uniformly. By choosing a sample size that is a polynomial function of k the efficiency of these learners is guaranteed at the expense of the representativeness of the samples, especially for small n.

[7] *Approximate halving algorithms* make predictions in accord with the preference of at least $0 < \varphi \leq \frac{1}{2}$ of the hypotheses (e.g., rankings) consistent with previous observations (or corrections). Setting φ to less than 1/2 simply lowers the base of the log in the halving algorithm's mistake bound to $(1 - \varphi)^{-1}$. For example, $(1 - 1/2)^{-1}$ is 2, but $(1 - 1/5)^{-1}$ is 1.25, so if a learner selects candidates that are preferred by at least 20% of the viable rankings then each error will reduce the viable rankings by at least 20% allowing no more than $\log_{1.25} k!$ errors to occur before just one ranking remains. Goldman et al. (1989) show that approximate halving algorithms can circumvent #P-hard counting problems by replacing exact counts with approximations provided by a fully polynomial randomized approximation scheme (FPRAS) that returns cardinality estimates that are correct to within a factor of $1 + \epsilon$ with probability at least $1 - \delta$ and does so with computation that is polynomial in $1/\epsilon$ and $\lg(1/\delta)$. Germane to the task at hand, they show that total orders of k elements can be efficiently learned (despite the #P-completeness of the counting problem) using an FPRAS that samples nearly uniformly from partial orders to generate predictions for an approximate halving algorithm. For any confidence level δ their strategy runs in time polynomial in k and $\lg(1/\delta)$ and makes at most $k \lg k + (\lg e) \lg k$ mistakes with probability at least $1/\delta$.

The case of RVL^0 warrants special attention. Many learning strategies make predictions that do not necessarily correspond to *any* of the concepts (e.g., rankings) they are designed to learn. This is true of the halving algorithm—and by extension RVL. Consider (21):

(21)

input 1	c_1	c_2	c_3
cand. a	*		
cand. b		*	*

input 2	c_1	c_2	c_3
cand. c		*	
cand. d	*		*

input 3	c_1	c_2	c_3
cand. e			*
cand. f	*	*	

Candidates a, c, and e each have an r-volume of $2/3$ in their respective tableaux and thus (absent any prior ranking conditions) they will be chosen by RVL. Yet, there is no ranking that selects $\ell = [a, c, e]$ because, taken together, their ranking conditions are circular. Interestingly, EDCD chooses precisely the same set of candidates when minimizing the sum of the pooled violations in the singe stratum in $\mathcal{H} = \big[\{c_1, c_2, c_3\}\big]$.

One way to avoid this kind of collective inconsistency is to make predictions based on a single randomly chosen ranking (as in, e.g., RVL^0). The randomness this strategy introduces has an additional benefit of allowing a shift in focus from worst-case to expected mistake bounds.[8] These two properties—randomness and predictions that represent target concepts—have proven useful in several approaches akin to RVL^0. For example, Haussler et al. (1991) propose a learner that makes predictions using a random hypothesis drawn uniformly from those consistent with all previous observations which they call a *Gibbs* learner. They show that, in a Bayesian setting, the learning curve of the Gibbs learner is nearly optimal and that it is especially robust in the face of incorrect priors. In another framework, Maass (1991) proposes a learner that, on each error, randomly selects a new hypothesis consistent with all previous corrections; this approach is often called *randomized halving*. Maass shows that the *expected* mistake bound of this approach is $k \ln(k)$, which is actually lower than that of the halving algorithm (i.e., the base of the log is e rather than 2). RVL^0 is not quite either of these approaches; its sample is not necessarily uniform and it is error-driven like approximate halving, but chooses a new ranking for each prediction like Gibbs.

A single sampled ranking can also be used to generate candidates for RVL and RVL^n in a fashion similar to that of EDCD, which mitigates the need to generate complete tableaux in cases where candidates are not antecedently given. In fact, Boersma (2003, 2009) suggests that EDCD should make predictions using random linearizations of $\mathcal{H}$ rather than pooling. This *sampling*-EDCD strategy is essentially duplicated by RVL^0 save for the fact that the latter replaces $\mathcal{H}$ with a set of ERCs. Comparing these approaches allows a direct evaluation of the utility of ERCs vs. stratified hierarchies that will be taken up in §3.5.

[8] Randomization offers protection from the worst case (e.g., an adversarial teacher) because there is no training sequence that is guaranteed to produce the maximal number or errors.

3.5 Performance in simulations

Figure 1 gives the number of mistakes with $k = 2 \ldots 12$ constraints averaged over 1,000 trials at each value of k. In each trial, a random target ranking $\mathcal{R}_T$ was obtained by shuffling the constraints and EDCD, RVL^0, and RVL^3 were tested in parallel with their error-counts incremented for each disagreement with $\mathcal{R}_T$ in a sequence of $k^2 \times 100$ random *binary tableau* (i.e., a pair of candidates with violation profiles drawn uniformly from $\{0,1\}^k$).[9] The use of $k^2 \times 100$ tableaux per trial reflects the order of EDCD's mistake bound plus a factor of 100.

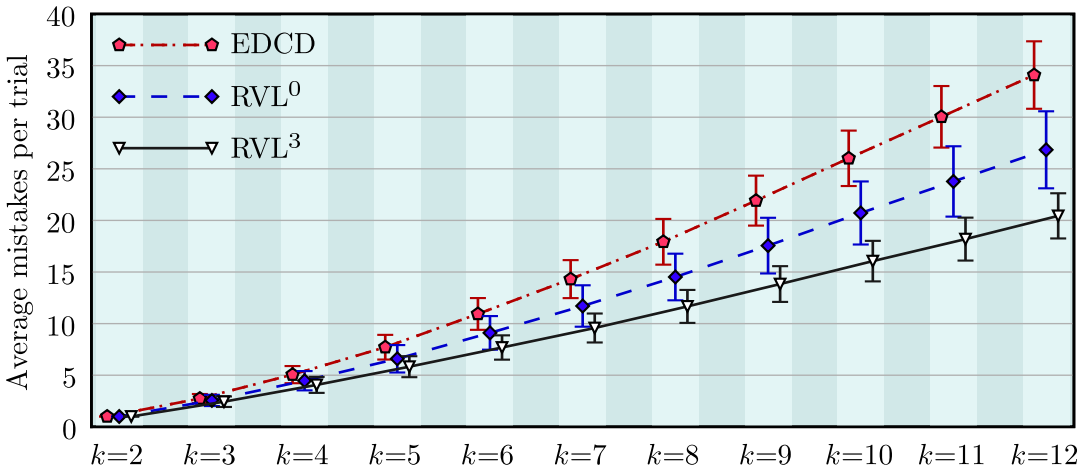

Figure 1: 1,000 trials of $k^2 \times 100$ binary tableaux for three learners

Figure 2 provides a closer view at $k = 12$ with RVL^n taking a wider range of sample sizes.[10]

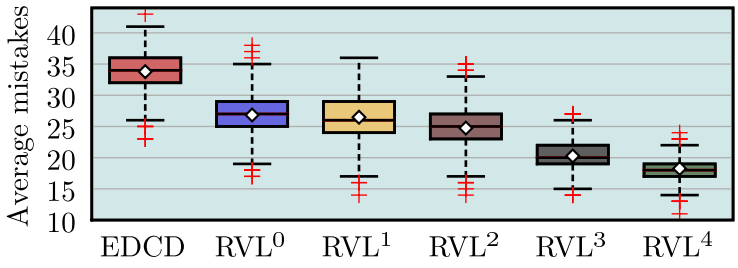

Figure 2: Detail view of six learners at $k = 12$

In these simulations, RVL^n performs significantly better than EDCD. The comparison gets a bit less abstract if the algorithms' predictions are taken seriously as empirical models.

[9] Binary tableaux might seem overly artificial, but any tableau can be decomposed into candidate pairs and using violation profiles draw uniformly from $\{0,1\}^k$ produces competitions that are impressionistically fairly realistic in that, only about half the constraints differentiate an average pair of candidates. If anything, ERCs with more e values might be more realistic but binary tableaux have the advantage of simplicity.

[10] Note that the error-bars in Fig. 1 mark one standard deviation but, as is conventional in box plots, the whiskers in Fig. 2 surround the data within 1.5 times the inner quartile range (diamonds mark the means).

One salient difference is that EDCD can select harmonically bounded candidates (e.g., in a stratum, the profile $\langle 1, 1\rangle$ is chosen over $\langle 3, 0\rangle$ and $\langle 0, 3\rangle$ despite being harmonically bounded by them). Moreover, Boersma (2009) shows that violation pooling breaks the convergence proof for constraint demotion (Tesar & Smolensky 1998:264) because convergence fails when pooling occludes critical rankings among constraints in strata where an optimal candidate happens to have the fewest pooled violations. Conversely, Boersma shows that the proof *does* work (in a probabilistic sense) if EDCD selects candidates using a random ranking sampled from those described by $\mathcal{H}$. This strategy, *sampling*-EDCD, offers an ideal comparison with RVL^0 because they differ only in the representation of ranking information (with the caveat that uniform samples can be drawn from $\mathcal{H}$ much more easily than $\mathcal{E}$).

Figure 3 gives average errors (1,000 trials of $k^2 \times 100$ binary tableaux at each $k = 2 \ldots 12$) for learners that sample rankings from either a weak order $\mathcal{H}$ or an ERC-set $\mathcal{E}$ using either *constant sampling* (i.e., a new ranking for each prediction) or *error-driven sampling* (i.e., a new ranking with each error). Also, a run of pooling-EDCD is included for comparison.

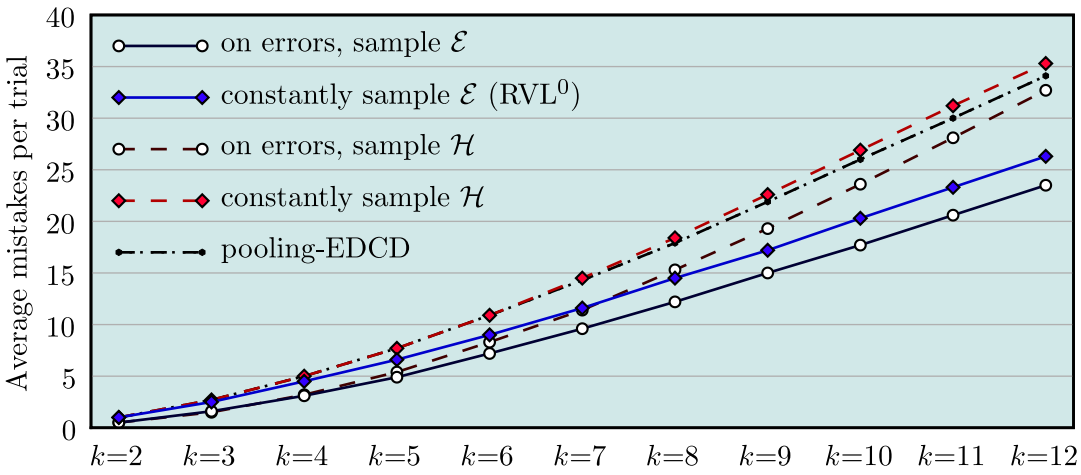

Figure 3: Constant vs. error-driven sampling from $\mathcal{H}$ vs. $\mathcal{E}$

With error-driven sampling, an early lucky guess can curtail future errors. Factorial growth in the set of rankings mitigates this advantage as k grows (as can be seen for $\mathcal{H}$-sampling). $\mathcal{H}$-pooling fares better than constant $\mathcal{H}$-sampling (nonconvergence notwithstanding), but error-driven sampling is better and $\mathcal{E}$-sampling better still. Error-driven $\mathcal{E}$ sampling performs best, with a curve fitting its expected error bound of $k \ln(k)$ (Maass 1991).

This kind of sampling has been explored by several linguists as a way to model variation (e.g., Anttila 1997 uses constant $\mathcal{H}$ sampling). The remainder of the paper will focus on the various sampling strategies and their empirical ramifications for typology and variation.

4 Frequency, Typology, and Variation

It is a laudable and challenging endeavor to build linguistic models that can fit frequencies observed in variation (e.g., Labov 1969, Boersma 1997, Boersma and Hayes 2001, Goldwater and Johnson 2003). If, however, those same frequencies follow as consequences of interactions among independently motivated components of a linguistic model then variation becomes a window on the core grammar rather than a complication that obscures the grammar (e.g., Kiparsky 1993, Anttila 1997a, Coetzee 2002). Moreover, the idea that a single mechanism might account for frequency distributions over variable patterns in individual languages *and* the frequency distributions over static patterns across languages is too elegant not to pursue.

4.1 Typological variation

Bane and Riggle (2008) analyze the link between r-volume and typological frequency in quantity insensitive stress systems using Gordon's (2002) non-foot-based metrical constraints and a database of 306 languages assembled by Heinz (2007).[11] Gordon's constraints generate 152 stress systems of which 26 are attested in the data. The r-volumes of the predicted systems and frequencies of the attested systems both obey somewhat Zipfian distributions. For example, 2/3 of the languages have one of the 3 most common patterns but the 13 least common are attested in 1 language each. Figure 4 gives frequencies of the most commonly attested patterns and r-volumes of those most voluminously predicted. Ten patterns are in the top 20 by both measures; these are marked with the number of languages they represent.

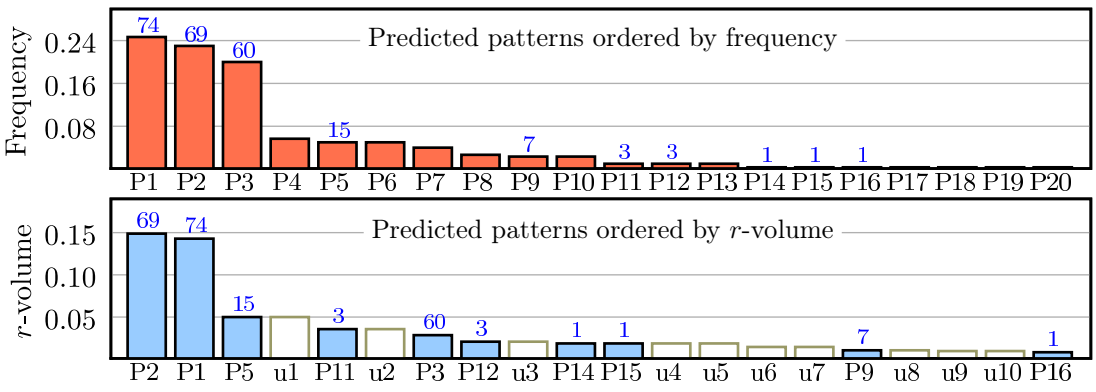

Figure 4: Top 20 patterns by frequency of attestation and by r-volume

Bane and Riggle found that the r-volume of a predicted pattern is significantly correlated with whether it is attested ($p < 0.01$), and moreover that frequency of attestation shows an exponential correlation with the log of the r-volume ($R^2 = 0.704,\ p < 0.001$).

[11] Heinz' database combines material from Bailey (1995) and Gordon (2002) which themselves draw from Hyman (1977), Halle & Vergnaud (1987), Hayes (1980, 1995), and many others. It contains 422 genetically and geographically diverse languages of which 306 have stress systems that are QI and non-variable.

This distribution is what one would expect if human learners were doing something like error-driven $\mathcal{E}$ sampling. In such a scenario, typological frequency will be shaped by two linguistic factors (along with many extralinguistic factors outside the scope of this analysis). First the relative r-volumes of the patterns will bias the choice of the learner's 'final' grammar when the choice is underdetermined by the observed data. Second, and equally important, the set of grammars that the learner chooses from will be a function of the probability of observing forms that distinguish the teacher's language from each of its neighbors (where two languages are *neighbors* if they share the same outputs for some input forms).[12]

4.2 Typological priors

Coetzee (2002) was the first to investigate the possibility that the frequency of a linguistic pattern's attestation might be linked to the frequency with which it is generated among the set of all possible rankings (of the universal constraint set CON).[13] His analysis is based on Anttila's (1997a) model of free variation as the result of sampling from the rankings that 'linearize' the relatively un-ranked constraints in a stratified hierarchy.

In principle, the set of rankings sampled in Anttila's model is defined by a partial order over constraints; I will refer to this as a Partially Ordered Grammar (POG). In practice, however, most ostensibly POG-based analyses use stratified hierarchies (e.g., Anttila (1997a, 2002), Anttila and Cho (1998), Coetzee (2002)), which are less expressive than partial orders. Stratified hierarchies are more generally known as *weak orders* and I will refer to a grammar specified in this way as a Weakly Ordered Grammar (WOG). Conversely, Elementary Ranking Conditions are more expressive than partial orders (which they properly include) because ERCs allow disjunctions (e.g., $\langle \text{W W L} \rangle =$ 'c_1 or c_2 outranks c_3'). I will refer to a grammar specified with ERCs as an ERC-Set Grammar (ESG). These classes form an inclusion hierarchy WOG $\subset$ POG $\subset$ ESG that will be discussed in §4.4.

Unlike Bane and Riggle's assessment of QI stress systems, Coetzee found distributional asymmetries among segment inventories that do not appear to pattern according to r-volume

[12] The relevance of the second factor is nicely illustrated by grammars that include so-called 'economy' constraints. For instance, if a syllable structure grammar utilizes a constraint representing Selkirk's (1981) Syllable Minimization Principle or the constraint *STRUC-σ (Zoll 1998) to derive syncope, then one of the languages in the typology is the Null language. Gouskova (2003) argues that such predictions should exclude economy constraints from CON. In the current context, this argument seems to be bolstered by the fact that the Null language has massive r-volume—when *STRUC-σ is top-ranked nothing else matters—and thus is highly likely to be selected if included among the languages a learner has to choose from. But this is where the second factor comes in, the Null language is nearly impossible to to confuse with its neighbors, so the chances for it to occur among the languages consistent with the learner's training data will be vanishingly small. It is not necessary to rule out the Null language by excluding *STRUC-σ from CON or by appealing to meta-constraints because its apparent typological absence can be derived from a model of learning and the fact that it is wildly different from its neighbors.

[13] Coetzee's constraint set is implicitly CON, but as is the norm in OT analyses, only a few constraints are included in the analysis of each typological pattern he considers. As I will show below, this is significant because the r-volumes for a set of patterns are wholly contingent on the set of constraints one considers.

among the languages in the UCLA Phonological Segment Inventory Database (UPSID; Maddieson 1984, Maddieson and Precoda 2002). One of the main examples Coetzee discusses is the distribution of front round and back unround vowels cross-linguistically. Among the 451 languages in the database, %7 allow both, %3 allow front round, %18 allow back unround, and the remaining %72 allow neither. Both vowels are relatively uncommon across languages but front round are more so (%10 attestation) than back unround (%25 attestation). This is illustrated in Fig. 5.

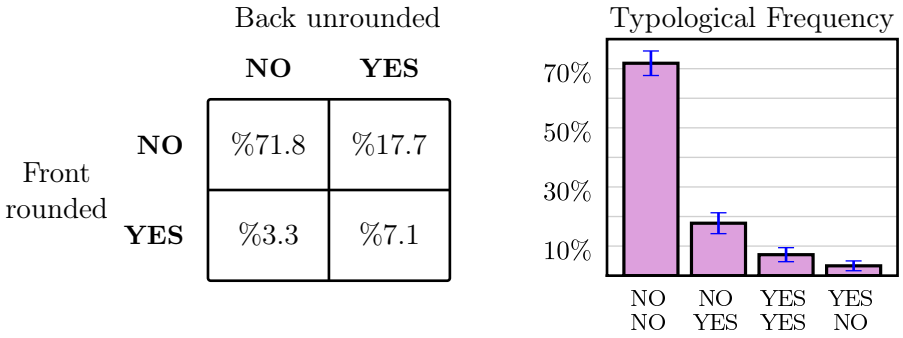

Figure 5: Distribution of front-rounded and back-unrounded vowels

Coetzee suggests that the asymmetry in Fig. 5 arises because both vowels are dispreferred on perceptual grounds due to insufficient distinctiveness (see Liljencrants & Lindblom 1972),[14] but front rounded vowels are also dispreferred on articulatory grounds because they use two active articulators, tongue and lips, while back unrounded vowels use only the tongue.

Coetzee models the typology in Fig. 5 as the interaction between a faithfulness constraint protecting rounding, IDENT(RND), and markedness constraints, *FRRD and *BKUNRD, that respectively penalize front round and back unround vowels. These constraints might be viewed as a grammatical encoding of the observations of Stevens et al. (1986) that feature combinations tend to enhance perceptual contrasts.

Coetzee provides a simple negative result regarding the correlation between frequency and r-volume with the constraints {IDENT(RND), *FRRD, *BKUNRD}. In the 2 rankings with IDENT on bottom neither vowel is allowed, in the 2 rankings with it on top, both are allowed, and in each of the remaining 2 rankings only one vowel is allowed. This distribution is far from that of Fig. 5 and it totally fails to capture the asymmetry between front round and back unround vowels. Coetzee notes that attempting to capture the asymmetry by positing a universally fixed ranking, *FRRD $\gg$ *BKUNRD, actually makes things worse because it allows zero rankings that generate languages that have front round but lack back unround vowels. In Fig. 6 the distribution of r-volumes for each proposal are set against the observed frequency distribution (dashed lines) and δ indicates the Kullback-

[14] I.e., Rounding of the lips lowers the value of F2 and thus the most distinct pair of F2 values is obtained with front unrounded and back rounded vowels.

Leibler divergence between the two. For the fixed ranking, $\delta = \infty$ because zero probability (i.e., 0 r-volume) is assigned to one of the attested values.

Coetzee identifies the crux of the problem as the challenge of modeling nearly universal tendencies without ruling out the rare cases—Goldsmith (1990) refers to this as the problem of *soft universals*. To overcome this problem he posits a new kind of constraint called a *preference constraint* that simultaneously evaluates candidates and rankings. A preference constraint $[c_1 \gg c_2]$ is basically inert when its ranking preference is met but, under any ranking where c_2 dominates c_1 it assigns violations to every candidate that violates c_1. Adding the preference constraint [*FRRD $\gg$ *BKUNRD] to the constraint set brings the r-volumes closer to the observed frequencies. This is illustrated in Fig. 6.

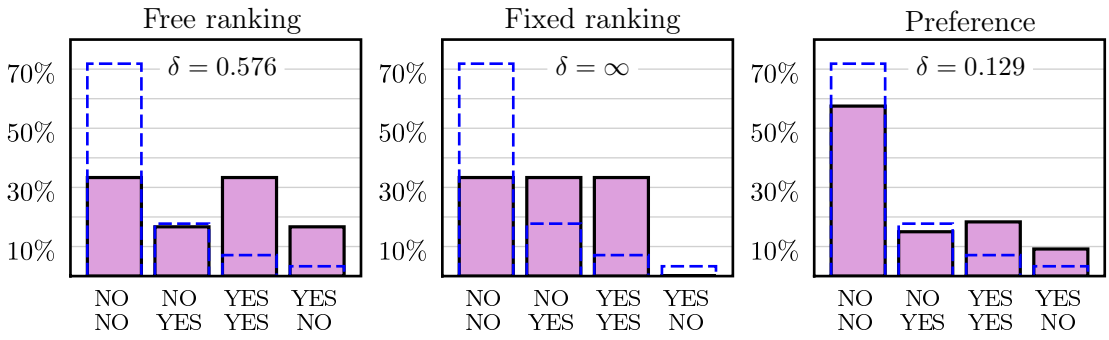

Figure 6: Typological frequency vs. r-volume for three models

Coetzee's innovation does indeed yield a better fit to the data, but it does so at the expense of introducing a new and complex mechanism that I believe is not necessary. Instead, if we take Coetzee's explanation of why front round are worse than back unround vowels literally and suppose that there are simply *more* constraints that disprefer the former, then the very mechanism that generates variation will generate the desired asymmetry.

4.3 Strength in numbers

The basic idea is that redundancy in the constraint set will give more influence to the constraints that (seem to) occur multiple times. This kind of redundancy can arise whenever distinct constraints are rendered circumstantially equivalent through the action of other high-ranked constraints.[15]

If the constraint set were to contain two tokens of *FRRD, then 10/24 of the rankings would allow neither of the marked vowels, 6/24 would allow only back unround, 6/24 would

[15] Alternatively, redundancy could be viewed as a fundamental and pervasive property of grammars under the thesis that *Con* is actually a multiset. This might seem like a significant departure but it would not actually change the factorial typology in any way other than changing the r-volumes of the languages generated. For now I will evaluate the fit of r-volume to frequency as if there were multiple tokens of some constraints, while remaining agnostic about precisely what this means.

allow both, and 2/24 would allow only front round vowels. This is easy enough to calculate with four constraints, but an enumeration of the cases is obviously not generally feasible. This is where the methods of §2.2 are especially useful. Consider, in (22), the ERCs for surface forms derived from a hypothetical underlying form containing a front round and back unround vowel.

(22)

/i.y.u.ɯ/	IDT	*FR	*BUR
a. [i.i.u.u]	**		
b. [i.i.u.ɯ]	*		*
c. [i.y.u.ɯ]		*	*
d. [i.y.u.u]	*	*	

y : NO ɯ : NO	candidate a wins if $E = \left\{ \begin{smallmatrix} \langle \text{L } e \text{ W} \rangle \\ \langle \text{L W } e \rangle \end{smallmatrix} \right\}$
y : NO ɯ : YES	if $E = \left\{ \begin{smallmatrix} \langle \text{L L W} \rangle \\ \langle \text{L W } e \rangle \end{smallmatrix} \right\}$, candidate d wins
y : YES ɯ : YES	candidate c wins if $E = \left\{ \begin{smallmatrix} \langle \text{W } e \text{ L} \rangle \\ \langle \text{W L } e \rangle \end{smallmatrix} \right\}$
y : YES ɯ : NO	if $E = \left\{ \begin{smallmatrix} \langle \text{L W L} \rangle \\ \langle \text{W } e \text{ L} \rangle \end{smallmatrix} \right\}$, candidate b wins

To compute the change that redundant constraints make to the r-volume, it suffices to duplicate one of the columns in E. For instance if there is 1 token of IDENT(RND), 2 tokens of *FRRN, and 1 token of *BKUNRD, which I will denote $(1 \cdot 2 \cdot 1)$, then the ERC set for candidate a is $\left\{ \begin{smallmatrix} \langle \text{L } e \text{ } e \text{ W} \rangle \\ \langle \text{L W W } e \rangle \end{smallmatrix} \right\}$. The r-volume for this pair can be computed as in (23).

$$(23)\quad \text{For } |E| = 2,\ r(E) = \frac{\left|\begin{smallmatrix} \text{W} \\ \text{W} \end{smallmatrix}\right|}{p} + \frac{\left|\begin{smallmatrix} \text{W} \\ e \end{smallmatrix}\right|}{p} \times \frac{\left|\begin{smallmatrix} \star \\ \text{W} \end{smallmatrix}\right|}{\left|\begin{smallmatrix} \star \\ \text{W} \end{smallmatrix}\right| + \left|\begin{smallmatrix} \star \\ \text{L} \end{smallmatrix}\right|} + \frac{\left|\begin{smallmatrix} e \\ \text{W} \end{smallmatrix}\right|}{p} \times \frac{\left|\begin{smallmatrix} \text{W} \\ \star \end{smallmatrix}\right|}{\left|\begin{smallmatrix} \text{W} \\ \star \end{smallmatrix}\right| + \left|\begin{smallmatrix} \text{L} \\ \star \end{smallmatrix}\right|}$$

For a pair of ERCs ε γ, the notation $\left|\begin{smallmatrix} e \\ \text{W} \end{smallmatrix}\right|$ denotes the number of columns where $\varepsilon = e$ and $\gamma = \text{W}$, while $\left|\begin{smallmatrix} \text{W} \\ \star \end{smallmatrix}\right|$ denotes the number of columns where $\varepsilon = \text{W}$ and so on. The variable p is the number of columns with a 'polar' value (non-e) in at least one of the ERCs.[16]

Though the addition of a single redundant constraint begins to capture the asymmetry between the vowels, even more redundancy does an even better job. Adding another token of each markedness constraint better captures the relative weakness of IDENT(RND) and the best fit of all is obtained with 1 token of IDENT(RND), 8 of *FRRD, and 3 of *BKUNRD, which yields r-volumes that are well within the %99 confidence intervals for the frequencies given that the sample contains 451 data points (shown in Fig. 7). This observation is not intended as a claim that there are necessarily eight different constraints that penalize front unrounded vowels—though this does not seem out of the question if the constraint set is fairly complete—but rather to make it clear that a large constraint set with a good deal of redundancy can predict the typological distribution with great precision. In fact, the

[16] The closed form can be solved for a particular r-volume so as to fit the frequency data. Two ERCs can describe variation among $n \leq 6$ forms and four ERCs can describe variation among $n \leq 24$ forms.

$(1 \cdot 8 \cdot 3)$ model seems more to illustrate the potential for overfitting, especially given the number of factors outside the scope of the model that are likely to be relevant.

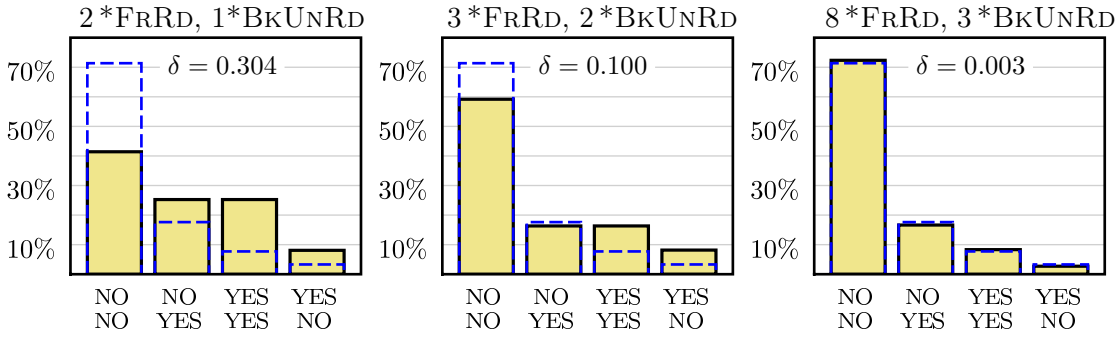

Figure 7: Typological frequency vs. r-volume with redundant constraints

The observed frequencies and r-volumes for all the models discussed are given in Table 1.

	attested	free	fixed	pref.	$(1 \cdot 2 \cdot 1)$	$(1 \cdot 3 \cdot 2)$	$(1 \cdot 8 \cdot 3)$
NO NO	0.718	0.333	0.333	0.575	0.416	0.583	0.722
NO YES	0.177	0.166	0.333	0.150	0.250	0.166	0.166
YES YES	0.071	0.333	0.333	0.183	0.250	0.166	0.083
YES NO	0.033	0.166	0	0.092	0.083	0.083	0.027

Table 1: Attested frequencies vs. predicted frequencies for six models

The fact that the presence of redundant constraints makes it possible to fit distributions so well suggests that, though typological asymmetries might raise questions about asymmetries in the markedness of some structures, the redundancy in the model should be justified by well grounded constraints and not chosen simply to fit the observed distributions.

On the other hand, if the idea that the sampling model should *predict* the distributions was abandoned, then redundancy could make it possible to fit frequencies in much the same fashion as the constraint weights of the MaxEnt model of Goldwater & Johnson (2003) or the noisy-HG model of Boersma & Pater (2007). Such a move would, however, undermine the most appealing aspect of the sampling model, so I will assume that redundancy is a consequence of distinct perceptual, articulatory, paradigmatic, and/or cognitive constraints that happen to share the same set of violating structures in a given context.

With or without redundant constraints, using a learning strategy where grammars are chosen by sampling from those that are consistent with previous observations, creates a

kind of *analytic bias* akin to that discussed by Wilson (2006).[17] The distribution of vowel inventories might also arise from *channel bias* via systematic misperception of small F2 contrasts. Boersma (2008) offers a model of learning with an explicit phonetic layer that could capture this bias, but an explanation of the asymmetry between front round and back unround vowels would likely invoke multiple factors—articulatory and perceptual—that disproportionately degrade the former, much like the analysis given here. Regarding the QI stress systems, however, it becomes clear that channel bias alone cannot provide a complete explanation for typological asymmetries because it cannot explain patterns that arise from critical distinctions being omitted during learning. As opposed to bias that arises from the data being *noisy* (i.e., because misarticulation and/or misperception systematically overrepresents some patterns), bias that surfaces when the data is *sparse* shows how analytical predispositions of learners (plus the contents of CON) skews the typology.

4.4 The sampling model of variation

Kiparsky (1993) first proposed that competition (i.e., sampling) among a range of different constraint rankings could link grammatical structure to frequencies observed in variation.

> [*V*]*ariation comes from competition of grammatical systems (in the individual or in the community), not from a probabilistic component in the rules of the language …*
>
> *General prediction: the frequency of a variant is a function of the number of allowed constraint rankings in which it is the optimal output.* (Kiparsky 1993:2,4)

This approach has been extensively developed by Anttila (e.g., Anttila 1997a, 1997b, 2002, Anttila and Cho 1998) for OT grammars consisting of constraints that are partially, rather than totally, ranked. In principle, Kiparsky's proposal can be construed more generally as deriving variation from arbitrary collections of grammars (including those that cannot be described with partial orders and ostensibly those that do not use constraints at all). Anttila calls this most general construal the Multiple Grammars Theory of Variation (another proposal in this vein is advanced in Kroch 1989).

Though most analyses that adopt this strategy—henceforth the *sampling model*—describe the system in terms of sampling rankings from partial orders over constraints, actual practice has overwhelmingly used stratified hierarchies of constraints, which are a bit less expressive. Elementary Ranking Conditions, on the other hand, are a bit more expressive than partial

[17] Wilson's analytic bias—the generalization of patterns to similar elements—is more theory-neutral than the case discussed here. The common factor is a model of bias in the way that learners *generalize* from data (or the absence thereof) rather than a model of how the communication channel itself introduces bias through systematic distortion of the data used to make generalizations. Moreton (2008) discusses this distinction at length and offers several other cases of analytic bias that have no plausible perceptual/articulatory basis.

orders, but with the exception of Bane (2009), they have not often been used to model variation. Anttila (2007:6, *fn.*4) offers yet another possibility, suggesting that sampling models might use ranking *multisets* (i.e., sets with repetitions). These can fit frequencies arbitrarily well but, concomitantly, make the fewest predictions. The critical question is which representation scheme is expressive enough to fit observed frequencies while still offering principled restrictions on variation.

representation	$k=3$	description
ranking multisets	$\infty/64$	a 'bag' of rankings with repetition allowed
sets of rankings	64	any arbitrary subset of the possible rankings
ERCs (antimatroids)	23	any co-convex subset of the possible rankings
partial orders	19	any convex subset of the possible rankings
weak orders	13	a total ordering of a partitioning of Con
total orders	6	a total ordering of Con (i.e., 'classic' OT)

Table 2: A hierarchy of grammar representations for the sampling model

Table 2 lists representations in order of nested inclusion. The value for $k=3$ is the number of different grammars and the number of the $2^6 = 64$ subsets of the 6 rankings that can be expressed for a set of 3 constraints (in all but the multiset case, these are the same thing).

The factorial typology of any constraint set Con contains a set of *implicational universals* (i.e., $X \underset{Con}{\rightarrow} Y$ if every ranking that generates structure or mapping X also generates Y). Though the representations in Table 2 have different levels of flexibility in picking out subsets of the factorial typology, they all draw from the same typology and thus they all preserve the implicational universals and extend them to the probabilities in sampling (i.e., if $X \underset{Con}{\rightarrow} Y$ then sampling from any subset of the rankings preserves $P(X) \geq P(Y)$, where $P(X)$ is the probability of drawing a ranking that generates X).[18] Where the representations differ is in the range of relative values they allow for $P(X)$ and $P(Y)$. I will refer to universals on the relative magnitude of $P(X)$ and $P(Y)$ as *stochastic implicational universals*.

4.5 Stochastic implicational universals

Anttila (2007) identifies a case where stochastic implicational universals appear to bear on the choice of representations in Table 2. Vowel hiatus gives rise to a range of different variable and non-variable outcomes for the merger of /ea/ to [ee] and /ia/ to [ii] in five Finnish dialects. Across five dialects surveyed by Paunonen (1995) the change /ea/→[ee] is more common, it is optional in two and obligatory in two others while /ia/→[ii] is optional in two dialects. Most interestingly, in the dialect where both changes are optional, Anttila

[18] This holds for stochastic OT as well. In Boersma & Hayes (2001) the scheme for drawing rankings is different but, like all the cases here, the rankings are drawn from the same set.

notes that Paunonen observes that /ea/→[ee] is more frequent than /ia/→[ii] thus paralleling typological frequency.

	Literary		Töölö		Häme		Uusimaa		Colloquial	
merge?	no	yes	no	yes	no	yes	no	yes	no	yes
/ia/→	[ia]	–	[ia]	–	[ia]	–	[ia]	[ii]	[ia]	[ii]
/ea/→	[ea]	–	[ea]	[ee]	–	[ee]	–	[ee]	[ea]	[ee]

Figure 8: Vowel merger in five Finnish dialects

To generate the typology Anttila uses a generic FAITH constraint and a pair of constraints against diphthongs with a universal ranking *EA ≫ *IA. This allows three rankings:

(24) a. FAITH ≫ *EA ≫ *IA : no coalescence
b. *EA ≫ FAITH ≫ *IA : /ea/→[ee]
c. *EA ≫ *IA ≫ FAITH : /ea/→[ee] and /ia/→[ii]

Anttila notes that any partial order of these constraints that allows both (a) and (b) while respecting the universal *EA ≫ *IA must also allow ranking (c) and thus, because sampling from {a,b,c} will produce coalescence of [ea] more often than [ia], the greater frequency of /ea/→[ee] in the Colloquial Helskini dialect follows as a consequence of partial ordering. Crucially, this prediction does not follow if sampling uses arbitrary sets of rankings because sampling from {a,c} would generate both types of coalescence at the same rate.

While it is definitely true that partial orders are more restrictive than ranking-sets in generating sampling distributions, the claim that the asymmetry follows as a necessary consequence of partial ordering must be taken with the (now familiar) caveat that the choice of constraints is integral. If we were to adopt Casali's (1995) analysis of hiatus resolution with the constraints *DIPHTHONG, IDENT(HI), and IDENT(LOW) then the same typology would be generated without any stipulated universal rankings (i.e., /ia/→[ii] implies /ea/→[ee] because the former violates ID(HI) and ID(LOW) while the latter only violates ID(LOW)). The partial ordering, IDENT(HI) ≫ IDENT(LOW), describes the three rankings in (25).

(25) a. IDENT(HI) ≫ IDENT(LOW) ≫ *DIPHTHONG : no coalescence
b. IDENT(HI) ≫ *DIPHTHONG ≫ IDENT(LOW) : no coalescence
c. *DIPHTHONG ≫ IDENT(HI) ≫ IDENT(LOW) : /ea/→[ee] and /ia/→[ii]

Sampling from (25) produces coalescence one third of the time for both vowels and thus fails to predict the observed asymmetry. Though the major appeal of the sampling model is the promise that additional empirical facts from variation can be used in adjudicating among competing proposals, this does not seems like a good example of such a case.[19]

[19]The culprit in this case is the stringency relationship between the faithfulness constraints, but taking this case as evidence that IDENT(HI) and IDENT(LOW) are 'bad' constraints seems rather extreme.

A more illuminating response to this apparent failure is provided by asking what mechanism gives rise to the partial order. If arbitrary partial orders are allowed then, as (25) shows, variation frequency need not resemble typological frequency. But, if one assumes that sampling grammars are based on observed data, the correct predictions emerge. Consider, in (26), the examples Anttila offers of forms that show variable coalescence.

(26)

/ruotsi-a/	ID(HI)	ID(LO)	*DIPH
a. [ruotsi-a]			*
b. [ruotsi-i]	*	*	

/suome-a/	ID(HI)	ID(LO)	*DIPH
c. [suome-a]			*
d. [suome-e]		*	

If one assumes that the ranking arguments (partial orders, ERCs, ... whatever) that come from pairs of variants are ignored or somehow cancel each other out, then upon observing variation between $a \sim b$ and $c \sim d$ the learner would be left with no information at all about the ranking of these three constraints. In this case the asymmetry between mid and high vowels is again predicted (i.e., under the 6 rankings, 3 allow no coalescence, 2 coalesce both vowels, and 1 coalesces only [ea] diphthongs). Thus we see that though there are partial orders that fail to predict the asymmetry, every partial order consistent with the observations does predict the asymmetry. This is true of both ERC sets and partial orders.

4.6 Bane's generalization

Bane (2008) offers an intriguing analysis of how an explicit model of the way that the sampling grammar is derived from observations can predict frequencies. The data Bane considers is drawn from Anttila's (2008) prosodic analysis of the English dative alternation as attested in a corpus of 1,580 prosodically annotated dative constructions drawn in 2004 from the blogs hosted by `www.blogspot.com` (The Blogspot Corpus).[20]

Pattern	Example	frequency
double object construction	give [my sister] [the old book]	69.8%
prepositional construction	give [the old book] [to my sister]	26.9%
heavy NP shift construction	give [to my sister] [the old book]	3.3%

Table 3: Frequency of attestation in The Blogspot Corpus

Anttila (2009, in press) proposes a set of 9 constraints that refer to properties such as footedness, stress, and phrasing (among others). He works out the T-orders—a graph of the implicational universals—and asks how adding partial rankings can restrict the T-order so

[20] Anttila credits Philipp Angermeyer, Rahul Balusu, and Peter Liem with creation of the corpus.

that the rank of the mappings matches the rank of their frequencies. This is like Coetzee's (2004) move to link the rank-order of r-volume to frequencies.

Bane takes Anttila's stratified hierarchy (i.e., weak order) and, using the techniques from §2.2, computes the goodness of fit between the r-volumes it generates and the observed frequency. Bane then proposes a search algorithm that performs a random walk on the set of partial orders that attempts to maximize the fit by hill-climbing. By actually using the expressive power of the partial order he achieves a sampling grammar with an R^2 value of 0.981 and an accuracy of 80.6%.[21]

Presumably, a search of the space of possible ERC-set grammars would produce an even tighter fit. The problem is that, in both cases, the search is computationally expensive and grows rapidly more so for larger constraint sets. The reason that stratified hierarchies are so often used is that we have an efficient way of deriving them from data. Bane then asks whether the same might hold for ERCs. He takes the ERC set that corresponds only to the non-variable data and uses that as his sampling grammar. It turns out that the match between the r-volumes this set imposes on the variants and their frequency is on par with the partial ordering derived at great expense. This leads Bane to the following observation.

(27) **Bane's generalization:** the categorical data can explain much of the variable data.

In other words, non-variable forms predict the overall shape of variation. The solution proposed for the Finnish dialects can be seen as a simple case that fits Bane's generalization.

4.7 Redundancy and goodness of fit with sampling

An obvious challenge for the sampling model is the 'quantal' character of the probability distributions that arise from sampling over what is a fundamentally combinatoric system. The crucial question is whether sampling can produce distributions that are (relatively) close to those observed in empirical data on variation.

Kiparsky (1993) introduced the idea of modeling variation by sampling from constraint rankings using data from several dialects of English for rates of coronal deletion in word final clusters before C-initial words, V-initial words, and phrase finally. This phenomenon is known in the sociolinguistics literature as the variable (TD) and it offers a rich testing ground for any proposed model of variation because it is one of the longest studied (e.g., Labov et al. 1968) and one of the most often studied aspects of variation in English (Guy 1980, Santa Ana 1992, and Patrick 1999 provide a broad perspective on the literature).

[21] Frequency is influenced by a wide range of extra-grammatical factors (e.g., register, speech rate, target audience) as well as a range of grammatical factors (e.g., pronominality, animacy, givenness) that are outside the scope of Anttila's—and by extension Bane's—treatment of the phenomenon. What's interesting is that the prosodic factors (e.g., footedness, stress, phrasing) can also do a good job explaining the variance.

Coronal deletion also provides clear tests for any proposed model's restrictiveness and explanatory adequacy because it exhibits robust stochastic implicational universals (e.g., in all dialects studied to date, deletion is most frequent before C-initial words) but also exhibits significant variation (e.g., whether pause patterns with C or V varies across dialects).

The analysis of variable coronal deletion in OT has been discussed by Kiparsky (1993), Anttila (2002), Coetzee (2004), and Pater & Coetzee (2010) among many others. In Table 4 rates of deletion are reported from sociolinguistic studies on (TD) in five dialects of English.[22]

English Dialect	C_#C	C_#V	C_##	data source
AAVE	76%	29%	73%	Fasold (1972:76)
Tejano	62%	25%	46%	Bayley (1995:310)
Chicano	62%	45%	37%	Santa Ana (1991:76)
British	46%	8%	5%	Tagliamonte & Temple (2005:288)
Jamaican	85%	63%	71%	Patrick (1992:181)

Table 4: Rates of t/d-deletion by environment for five dialects of English

In analyzing this data, I will adopt the constraint set of Coetzee & Pater (2010) {*CT, MAX, MAX-PREVOWEL (MPV), MAX-PREFINAL (MPF)}, but I will add to it two additional markedness constraints in (28). These are inspired by Coetzee's (2004) discussion of coronal deletion in terms of the licensing-by-cue paradigm (see Steriade 1997) and the discussion in Kirchner & Varelas (2002) of the licensing of non-post-vocalic consonants.

(28) a. **POST-C RELEASE** (RELS): postconsonantal consonants should have a good release. This penalizes Ct#C and Ct#V.

b. **POST-C TRANSITION** (TRANS): postconsonantal consonants should have good transitions to the following segment. This penalizes Ct#C and Ct##.

These six constraints assign violations to the candidates in the three cases as shown in (29).

(29)

input	→ *output*	*CT	RELS	TRANS	MAX	MPV	MPF	ERC for deletion
a. Ct#C	→ Ct#C	*	*	*				
b.	→ C#C				*			⟨WWWL *e e*⟩
c. Ct#V	→ Ct#V	*	*					
d.	→ C#V				*	*		⟨WW *e* L L *e*⟩
e. Ct##	→ Ct##	*		*				
f.	→ C##				*		*	⟨W *e* W L *e* L⟩

[22] I omit here data from Kiparsky (1993) which mistakenly gives *weights* of VARBRUL factors as deletion frequencies (e.g., compare the numbers Kiparsky attributes to Santa Ana (1991) with Santa Ana (1996:66) or consider Guy (1980:29), which gives 0.86 as a rate of deletion *modified* by the factors Kiparsky reports).

Using ERCs describing the rankings that select the deletion candidates, in (29), it is possible to construct a sampling grammar G_s defined by a pair of ERCs such that sampling uniformly from the rankings consistent with G_s yields deletion frequencies close to those observed. Table 5 contrasts deletion rates in the five dialects with those predicted by the best fitting G_s consisting of a pair of ERCs.

English Dialect		environment for deletion C_#C	C_#V	C_##	partial grammar G_s
AAVE	total tokens	143	202	37	$\left\{\begin{array}{l}\langle \text{w l ww } e \text{ l} \rangle \\ \langle e \ e \text{ l l w l} \rangle\end{array}\right\}$
	del. observed:	109 (.76)	59 (.29)	27 (.73)	
	del. expected:	109 (.76)	59 (.29)	28 (.76)	$X^2 = 0.07$
Tejano	total tokens	1,738	974	564	$\left\{\begin{array}{l}\langle \text{w l wwww} \rangle \\ \langle e \text{ l } e \text{ ww } e \rangle\end{array}\right\}$
	del. observed:	1,078 (.62)	244 (.25)	259 (.46)	
	del. expected:	1,086 (.63)	244 (.25)	259 (.46)	$X^2 = 0.12$
Chicano	total tokens	3,693	1,574	1,024	$\left\{\begin{array}{l}\langle \text{w } e \ e \text{ w l } e \rangle \\ \langle \text{l l w } e \ e \text{ w} \rangle\end{array}\right\}$
	del. observed:	2,290 (.62)	708 (.45)	379 (.37)	
	del. expected:	2,283 (.62)	715 (.45)	372 (.36)	$X^2 = 0.44$
British	total tokens	523	570	136	$\left\{\begin{array}{l}\langle e \text{ l l } e \ e \text{ w} \rangle \\ \langle \text{l l wwww } e \rangle\end{array}\right\}$
	del. observed:	241 (.46)	46 (.08)	7 (.05)	
	del. expected:	234 (.45)	45 (.08)	7 (.05)	$X^2 = 0.46$
Jamaican	total tokens	1,252	793	252	$\left\{\begin{array}{l}\langle \text{w } e \text{ l l w l} \rangle \\ \langle e \text{ w l w } e \text{ l} \rangle\end{array}\right\}$
	del. observed:	1,064 (.85)	500 (.63)	179 (.71)	
	del. expected:	1,073 (.86)	510 (.64)	180 (.71)	$X^2 = 0.55$

Table 5: Using ERCs to fit [t/d]-deletion rates in five dialects of English

There is no literal redundancy among the constraints used here but there are multiple ways to generate each pattern. In this case, representing the sampling grammar with a pair of ERCs allows the sampling model to fit the observed frequencies extremely well. Even in the case with the poorest fit, a difference as large as the one between the predicted and observed distributions would be expected to arise by chance 91% of the time.[23]

In Fig. 9, I plot the rates of deletion before vowel-initial and consonant-initial words in the five dialects in table 5 along with the rates reported for eleven more dialects in a range of studies of (TD) that are collected in Patrick (1999:162). The pre-pause environment is omitted in this case because many of the early studies did not explicitly track this condition. The ellipses around the data points represent the 95% confidence region based on the sample

[23]The X^2 statistic is the sum of $2\frac{(O-E)^2}{E}$ for each observed deletion rate O and expected rate E, there are three degrees of freedom because there are three conditions in which the rate of retention is 1 minus that of deletion, and the p-value for a chi-square statistic of 0.55 with three degrees of freedom is 0.91.

sizes in each study (but note that the dashed regions around the dialects in Wolfram (1969) are based on Patrick's (1999) estimate of Wolfram's sample sizes).

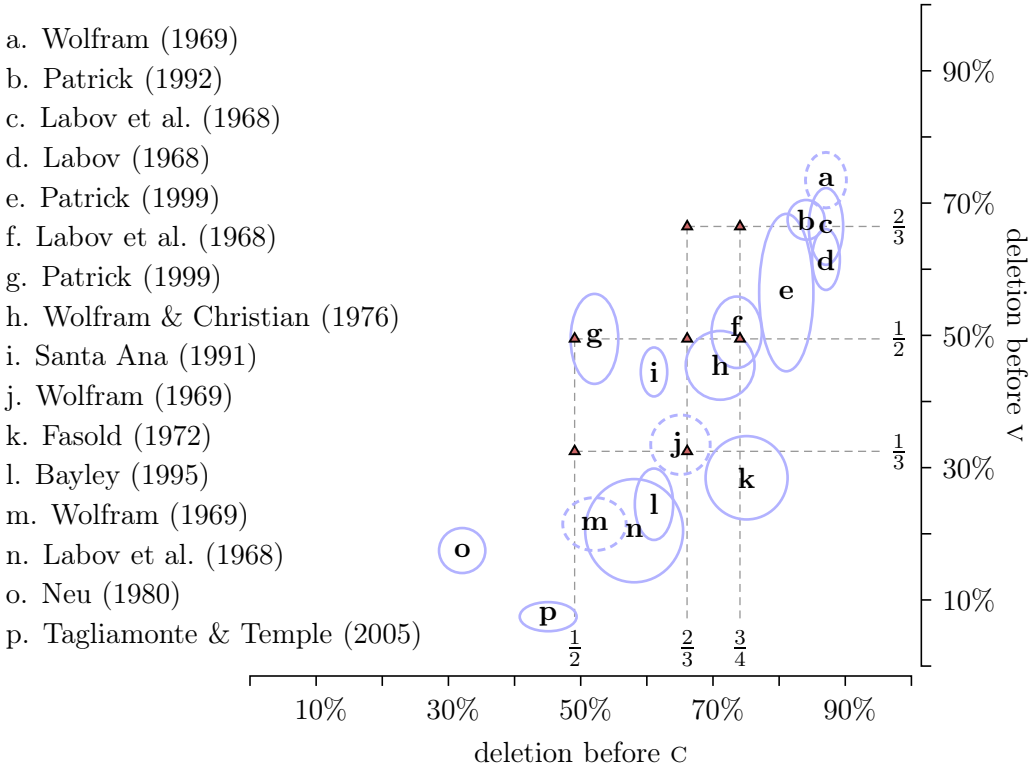

Figure 9: Ratios of deletion before V-initial and C-initial words

The main arguments that have been made against the sampling model question its adequacy in terms of the frequency ratios it can generate and the granularity of its predictions. These arguments assume that sampling draws from weak orders, which is understandable given that stratified hierarchies have been used almost exclusively sampling-based analyses to date. For instance, Boersma & Hayes (2001:72) point out that a 99-to-1 frequency ratio would suggest that 99 constraints must favor one of the variants and Bender (2000:226) discusses the fact that a small set of constraints can produce only a handful of frequency ratios and thus could not (in general) describe a pair of dialects whose only difference was a small across-the-board increment in the likelihood of a set of variants. Though both of these critiques are couched in terms of hypothetical patterns of variation, Fig. 9 shows that (for the six constraints considered) the sampling model generates just seven points in the variation space (marked ▲) and these offer a very poor fit for the range of observed patterns.

Similar critiques are offered by Guy (1997:345), Biro (2006:22), Clark (2008:35), and many others, including Coetzee & Pater (2010) in their discussion of (TD), who write "[t]here

are therefore only three probabilities of deletion in this context that the POC [Partially Ordered Constraints] theory can derive: 0, .50, and 1. One could always increase the size of the constraint set to yield other probability distributions ... but this strategy becomes implausible very quickly." Despite the 'partially ordered' moniker, Coetzee & Pater are clearly referring to models that use weak orders. Figure 10, plots the 16 empirical cases over the range of frequency ratios (rounded to two decimal places) that can be described using a sampling grammar comprising a pair of ERCs. The darker points correspond to ERC-pairs that describe partial orders (i.e. pairs in which each ERC has exactly one W).

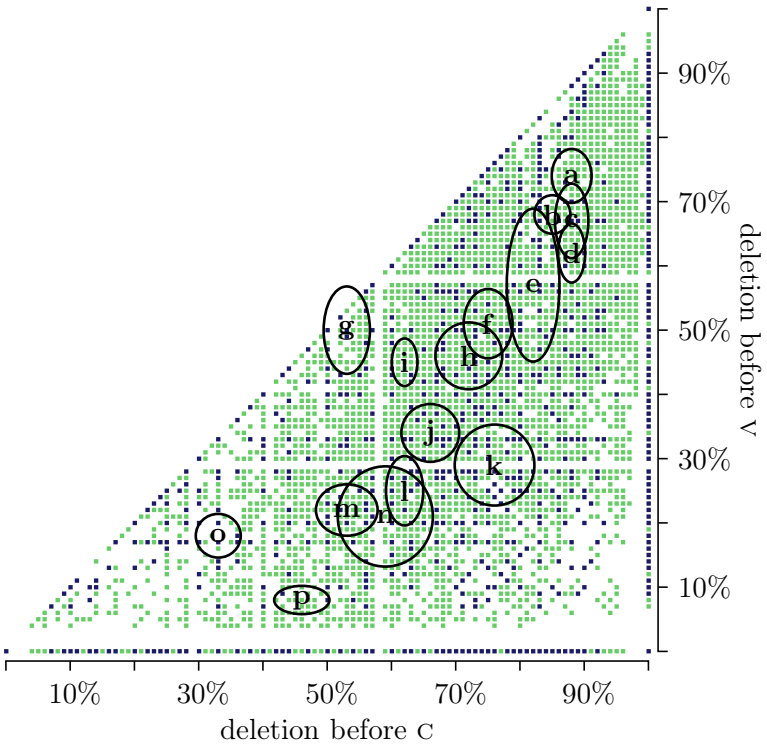

Figure 10: Ratios of deletion before V-initial and C-initial words

Figure 10 shows that moving beyond stratified hierarchies allows the sampling model to generate distributions that fit the empirical data quite well. In general, ERCs allow a much wider range of more finely grained predictions than partial orders but, even if sampling grammars are restricted to the coarser predictions of partial orders, the distance between the predicted and observed frequencies are not statistically significant in any of the 16 cases.

Despite initial appearances, the sampling model's ability to fit frequency distributions does not seem appreciably less powerful than that of models which encode probabilities using real-valued parameters such as those proposed by Boersma & Hayes (2001), Goldwater & Johnson (2003), Boersma & Pater (2007). These models can fit the infinite range of

frequency distributions generated their respective grammar formalisms arbitrarily well, so there is a sense in which they are, by definition, more powerful than models that distinguish only finitely many frequency ratios. Nonetheless, given the unavoidability of sampling error, the fact that numerous non-grammatical and grammatical factors influence all instances of variation studied to date, and the fact that incorporating constraints for these additional factors will give the sampling model even wider and more fine-grained predictions, it seems highly unlikely that empirical distributions will arise for which the coarseness of the sampling model's frequency predictions is a statistically significant liability.

Somewhat surprisingly, these models might be distinguishable along the one dimension where sampling has greater descriptive power. In (30), three constraints allow each of six hypothetical candidates to win under one of the six possible rankings. Consider the sets of candidates that can be assigned non-zero probability in three classes of models of variation.

(30)

/input/	C_1	C_2	C_3
candidate *a*		*	**
candidate *b*		**	*
candidate *c*	*		**
candidate *d*	**		*
candidate *e*	*	**	
candidate *f*	**	*	

WOG ⊆ POG ⊆ ESG

⟨W L L⟩ ⟨W *e* L⟩ ⟨W W L⟩

also: Nagy & Reynolds (1997)

also: Boersma & Hayes (2001), Goldwater & Johnson (2003)

Sampling from weakly ordered grammars (stratified hierarchies) can pick out six pairs of candidates (e.g., *a* & *b* alternate under the ranking $C_1 \gg \{C_2, C_3\}$ and *a* & *c* alternate under the ranking $\{C_1, C_2\} \gg C_3$). StOT and MaxEnt grammars restrict variation to precisely the same candidate pairs—though they can assign them infinitely more probability ratios.

In StOT, this restriction follows from the axiomatic assumption that variation among rankings is the same for all constraints. Boersma & Hayes (2001:fn3) expressly prohibit unequal variance, asserting that the gaussian noise is a global property of the model and not a parameter associated with each constraint. If the variance was a parameter of each constraint, then StOT would be able generate patterns like those of Nagy & Reynolds (1997) where one constraint can 'float' relative to two others with a fixed ranking (as in, e.g., the partial order $C_1 > C_3$ that picks out candidates $\{a, b, c\}$ in (30)). This kind of pattern is strictly impossible in MaxEnt models because there is no constraint weighting that will give a bimodal distribution to $\{a, b, c\}$ vs. $\{d, e, f\}$.

If sampling grammars are expressed with ERCs that are not restricted to partial orders, a single ERC ⟨W W L⟩, can pick out candidates $\{a, b, c, d\}$ to the exclusion of $\{e, f\}$. In this case, the ranking for *b* is the *opposite* of that for *d* (cf., $C_1 \gg C_3 \gg C_2$ vs. $C_2 \gg C_3 \gg C_1$)

and thus neither StOT, MaxEnt, nor floating constraints can select a set of candidates that includes b and d without allowing all six rankings and thus selecting all candidates.

This scenario is not some outlandish hypothetical but rather something that comes up constantly due to the inherent uncertainty in tableaux. For instance, in (31) I repeat the example of coalescence in Finnish. If the non-coallescing candidate is optimal we know only that *Diph is not ranked highest among these three constraints (i.e., ⟨W W L⟩).

(31)

/ruotsi-a/	Id(hi)	Id(lo)	*Diph
a. [ruotsi-a]			*
b. [ruotsi-i]	*	*	

The crucial question is then whether there are cases of variation involving sets of candidates best described with POGs or ESGs rather than WOGs. As usual, everything hinges on the choice of constraints. Just as Boersma & Hayes recast the constraint set that Nagy & Reynolds (1997) use to argue for (what is in essence) a POG in order to argue that the same phenomena can be captured by a WOG (with stochastic ranking-noise), it seems likely that any phenomenon that suggests analysis with an ESG will be able to be recast in terms of weakly ordered constraints.

Ultimately, the choice of models will depend on whether well-motivated constraints can be used to characterize the range of attested variation in terms of the kinds of weak orders that emerge in StOT and MaxEnt and on whether it is deemed more elegant (in terms of parsimony or transparency) to admit real-valued parameters or to describe sampling grammars in terms of ERCs. Regarding the latter, ERCs seem to have the advantage in that they are independently necessary descriptions of the rankings that logically follow from the optimality of a given candidate.

5 Conclusions

In this paper I have presented methods for computing r-volumes and methods for sampling from sets of rankings specified in terms of ERCs. The latter can be used to approximate r-volumes and can also be used directly in models of learning or variation in which the candidates are chosen with probability proportional to their conditional r-volume. In the case of learning, conditional r-volume is determined relative to the set of rankings consistent with previous observations (or corrections), and in the case of variation conditional r-volume is determined relative to the set of rankings described by a sampling grammar.

Exact computation of r-volume makes it possible to formulate learning algorithms for OT based on Littlestone's (1988) halving algorithm. The mistake-driven version of this approach

to learning rankings, which I call the r-volume learner, has a mistake bound and memory bound of $k \log_2 k$. This mistake bound is better than the quadratic bound of EDCD (Tesar & Smolensky 1996) and is within a logarithmic factor of the absolute bound for learning rankings, which is set at $k - 1$ by the VC-dimension of OT (see Riggle 2009a).

Exact computation is, however, not feasible in the general case because the problem of counting rankings—like many combinatoric counting problems—is #P-complete. One way to sidestep this complexity is to approximate r-volumes via sampling. For ERC sets that correspond to partial orders (i.e., those in which each ERC has exactly one W), the r-volume can be approximated using the polynomial-complexity uniform sampling strategy provided by Huber (2006). Under the partial-order restriction, efficient uniform sampling can serve as the foundation for an FPRAS (fully-polynomial randomized approximation scheme) in an approximate-halving algorithm, which Goldman et al. (1989) have shown to have an expected mistake bound of $k \lg k + o(k \lg k)$.

For the general case, where ERC sets are not restricted to those that correspond to partial orders, it is not yet known whether methods like Huber's can sample uniformly in polynomial time. If so, the approximate halving algorithm can be extended to this case, but if not, polynomial-complexity samples that are not perfectly uniform can still be used as a heuristic in learning algorithms. Heuristic versions of the r-volume learner, RVL^n, that make predictions according to k^n samples generated by random walk on the set of rankings consistent with a given ERC set, seem to compare quite favorably to EDCD. The results of the simulations in §3.5 suggest that RVL^n is superior to EDCD, even in the case where just $k^0 = 1$ sample is drawn. This is of particular interest because the single-sample strategy of RVL^0 corresponds precisely to the kind of sampling that has been independently argued to provide a model of variation that can use properties of the constraint set to makes good predictions about frequencies (see, e.g., Anttila 1997a).

Incorporating the same kind of sampling in the learning process provides a mechanism that links the frequency with which linguistic structures are attested typologically and the frequency distributions over forms that vary in individual languages. In evaluating this connection the algorithm for exactly computing r-volumes is quite useful. This is a good example of the fact that, though exact computation might not be a practical strategy for learners in all cases, it can still be quite useful in the analysis of linguistic models, which can be done 'off-line' with fewer constraints on computational resources.

One of the most intriguing things that emerges from algorithmic exploration of the link between typological frequency and r-volume is the observation that (circumstantial) redundancy can increase the apparent strength of a constraint. Thus the observation that not all marked structures are equally bad cross-linguistically can be seen to follow straightforwardly from the presence of asymmetries in the number of constraints that disprefer a given

structure. Under the sampling model, these effects are predicted to manifest in the patterns of variation within individual languages as well.

A wide range of models have been proposed for variation (see Coetzee & Pater 2010 for a recent overview). Computation of r-volumes makes it possible to give a precise quantitive evaluation of the sampling model alongside models in which probabilities follow more directly from real-valued parameters. Furthermore, the analysis given here of the kinds or probability distributions generated by sampling from weak orders, partial orders, and ERC-sets reveals some fundamental differences among the constraint-based models of variation. Overall, the sampling model seems to deliver on its promise of providing a model of variation in which frequencies are partially predictable from properties of the universal constraint set. However, it is also clear that evaluation of the model crucially depends on assumptions about the contents of the constraint set. In this sense, however, constraint-based models of variation are just like constraint based models of every other linguistic phenomenon.

Appendix A Python Code

`tab2ERCs`$(\cdot)$ takes a list of (candidateName, winnerBoolean, violationProfile) triples and returns ERCs for the winner (i.e., the first candidate where winnerBoolean equals `True`).

```
def tab2ERCs(T):
    for (candName, winnerBool, vPrf) in T:
        if winnerBool == True:
            return [tuple([cmp(W,L) for (W,L) in zip(vPrf,V)]) for (N,B,V) in T]
```

`rVol`$(\cdot)$ takes a list of ERCs and returns the r-volume.

```
def rVol(ercs):
    vol, C = 0.0, [c for c in zip(*ercs) if 1 in c or -1 in c]
    if len(ercs) == 0 or len(C) == 0: return 1.0
    for col in C:
        if -1 in col: continue
        vol += 1.0/len(C) * rVol([ercs[i] for (i,v) in enumerate(col) if v==0])
    return vol
```

`condVol`$(\cdot,\cdot)$ returns the conditional r-volume of the first ERC list given the second.

```
def condVol(E1,E2): return rVol(E1+E2)/rVol(E1)
```

An example of usage ('dogs' are the winner in the tableau `exTabx`):

```
In [1]: exTabx = [('cats',0,(0,1,2,2)),('dogs',1,(1,1,2,1)),('bats',0,(1,0,2,2))]

In [2]: tab2ERCs(exampleTabx)
Out[2]: [(1, 0, 0, -1), (0, 0, 0, 0), (0, 1, 0, -1)]

In [3]: rVol(tab2ERCs(exampleTabx))
Out[3]: 0.33333333333333331
```

References

Angluin, Dana (1988) Queries and concept learning. *Machine Learning* **2**: 319–342.

Anttila, Arto (1997a) *Deriving variation from grammar : A study of Finnish genetives*. John Benjamins, 35–68.

Anttila, Arto (1997b) *Variation in Finnish Phonology and Morphology*. Ph.D. thesis, Stanford.

Anttila, Arto (2002) Variation and phonological theory. In *The Handbook of Language Variation and Change*, Peter Trudgill J.K. Chambers & Natalie Schilling-Estes, eds., OxfordL Blackwell, 206–243.

Anttila, Arto (2007) *Variation and optionality*, chap. The Cambridge Handbook of Phonology. Cambridge University Press, 519–536.

Anttila, Arto & Young mee Yu Cho (1998) Variation and change in Optimality Theory. *Lingua* **104**: 31–56.

Bailey, Tod (1995) *Nonmetrical constraints on stress*. Ph.D. thesis, University of Minnesota.

Bane, Max & Jason Riggle (2008) Three correlates of the typological frequency of quantity-insensitive stress systems. In *Proceedings of the Tenth Workshop of the Association for Computational Linguistics' Special Interest Group in Morphology and Phonology*, rOA-966.

Barzdin, J. & R. Freivald (1972) On the prediction of general recursive functions. *Soviet Mathematics Doklady* **13**: 1224 – 1228.

Bayley, Robert (1995) Consonant cluster reduction in Tejano English. *Language Variation and Change* **6**: 303–326.

Ben-David, Shai, Dávid Pál, & Shai Shalev-Shwartz (2009) Agnostic Online Learning. In *Proceedings of COLT 2009*.

Bender, Emily M. (2000) *Syntactic Variation and Linguistic Competence: The Case of AAVE Copula Absence*. Ph.D. thesis, Stanford University.

Biro, Tamas (2006) *Finding the Right Words*. Ph.D. thesis, RIJKSUNIVERSITEIT GRONINGEN.

Boersma, Paul (1997) How We Learn Variation, Optionality, and Probability. *Proceedings of the Institute of Phonetic Sciences, Amsterdam* **21**: 43–58.

Boersma, Paul (2003) Bruce Tesar and Paul Smolensky (2000). Learnability in Optimality Theory. Cambridge, Mass.: MIT Press. Pp. vii+140. *Phonology* **20**(03): 436–446.

Boersma, Paul (2008) Emergent ranking of faithfulness explains markedness and licensing by cue. Rutgers Optimality Archive 954, available at www.fon.hum.uva.nl/paul/papers/EmergeFaith.pdf.

Boersma, Paul (2009) Some Correct Error-Driven Versions of the Constraint Demotion Algorithm. *Linguistic Inquiry* **40**(4): 667–686.

Boersma, Paul & B. Hayes (2001) Empirical Tests of the Gradual Learning Algorithm. *Linguistic Inquiry* **32**(1): 45–86.

Boersma, Paul & Joe Pater (2007) Testing gradual learning agorithms. Ms, University of Amsterdam and UMass Amherst.

Brightwell, Graham & Peter Winkler (1991) Counting linear extensions is #P-complete. In *STOC '91: Proceedings of the twenty-third annual ACM symposium on Theory of computing*, New York, NY, USA: ACM, 175–181.

Casali, Roderic F. (1995) *Resolving Hiatus*. Ph.D. thesis, UCLA.

Clark, Lynn (2008) *Variation, Change and the Usage-based Approach*. Ph.D. thesis, University of Edinburgh.

Coetzee, Andries (2004) *What it means to be a loser: Non-optimal candidates in Optimality Theory*. Ph.D. thesis, UMass Amherst.

Coetzee, Andries W (2002) Between-Language Frequency Effects in Phonological Theory. Rutgers Optimality Archive, available at: roa.rutgers.edu/view.php3?id=1399.

Fasold, Ralph W. (1972) *Tense marking in Black English*. Washington, D.C.: Center for Applied Linguistics.

Goldman, S. A., R. L. Rivest, & R. E. Schapire (1989) Learning binary relations and total orders. In *SFCS '89: Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, Washington, DC, USA: IEEE Computer Society, 46–51.

Goldsmith, John (1990) *Autosegmental and Metrical Phonology*. Oxford: Blackwell.

Goldwater, Sharon & Mark Johnson (2003) Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory*, Stockholm.

Gordon, Matthew (2002) A Factorial Typology of Quantity-Insensitive Stress. *Natural Language and Linguistic Theory* **20**(3): 491–552.

Gouskova, Maria (2003) *Deriving Economy: Syncope in Optimality Theory*. Ph.D. thesis, University of Massachusetts Amherst.

Guy, G. (1980) Variation in the Group and the Individual. In *Locating Language in Time and Space*, W. Labov, ed., New York: Academic Press, 1–36.

Halle, Morris & Jean-Roger Vergnaud (1987) *An Essay on Stress*. Cambridge Mass.: MIT Press.

Haussler, David, Michael Kearns, & Robert Schapire (1991) Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension. In *Proceedings of the fourth annual workshop on Computational learning theory*, UCSC-CRL-91-44, 61–74.

Hayes, Bruce (1980) *A metrical theory of stress rules*. New York: Garland, 1985, distributed by the Indiana University Linguistics Club.

Hayes, Bruce (1995) *Metrical Stress Theory : Principles and case studies*. Chicago: The University of Chicago Press.

Heinz, Jeffrey (2007) *Inductive Learning of Phonotactic Patterns*. Ph.D. thesis, UCLA.

Huber, Mark (2006) Fast perfect sampling from linear extensions. *Discrete Mathematics* **306**(4): 420–428.

Hyman, Larry (1977) *On the nature of linguistic stress*. SCOPIL 4, Los Angeles: Southern California University Press, 37–82.

Kiparsky, Paul (1993) An OT perspective on phonological variation. Handout, Rutgers Optimality Workshop 1993, available at http://www.stanford.edu/ kiparsky/Papers/nwave94.pdf.

Kirchner, Robert & Eleni Varelas (2002) A Cue-Based Approach to the Phonotactics of Upper Necaxa Totonac. paper presented at the 7th annual meeting of the Workshop on Structure and Constituency in the Languages of the Americas, Edmonton, Alberta., available:.

Kroch, A. (1989) Reflexes of grammar in patterns of language change. *Language Variation and Change* **1**: 199–244.

Labov, W (1968) *The Reflection of Social Processes in Linguistic Structures*, chap. Readings in the Sociology of Language,. Mouton Publishers, 240–251.

Labov, William (1969) Contraction, deletion, and inherent variability of the English copula. *Language* **45**: 715–762.

Labov, William, Paul Cohen, Clarence Robins, & John Lewis (1968) A study of the nonstandard English of Black and Puerto Rican speakers in New York City. (Cooperative Research Report no. 3288). Tech. rep., United States Office of Education, Washington DC.

Liljencrants, J. & B. Lindblom (1972) Numerical Simulation of Vowel Quality Systems: The role of Perceptual Contrast. *Language* **48**: 839–861.

Littlestone, Nick (1988) Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm. *Mach. Learn.* **2**(4): 285–318.

Littlestone, Nick (1989) *Mistake Bounds and Logarithmic Linear-threshold Learning Algorithm*. Ph.D. thesis, UC Santa Cruz.

Maass, Wolfgang (1991) On-line learning with an oblivious environment and the power of randomization. In *COLT '91: Proceedings of the fourth annual workshop on Computational learning theory*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 167–178.

Maddieson, Ian (1984) *Patterns of Sound*. Cambridge: CUP.

Maddieson, Ian & Kristin Precoda (2002) UPSID and PHONEME (Version 1.1). Online database, UCLA.

Mitchell, Tom (1982) Generalization as search. *Artificial intelligence* **18**: 203–226.

Moreton, Elliott (2008) Analytic bias and phonological typology. *Phonology* **52**: 83–127.

Nagy, N. & B. Reynolds (1997) Optimality theory and variable word-final deletion in Faetar. *Language Variation and Change* **9**(1): 37–56.

Neu, Helene (1980) Ranking of constraints on /t,d/ deletion in American English: A Statistical Analysis. In *Locating Language in Time and Space*, William Labov, ed., New York: Academic, 37–54.

Patrick, Peter L (1992) Creoles at the intersection of variable processes: -t, d deletion and past-marking in the Jamaican mesolect. *Language Variation and Change* **3**: 171–189.

Patrick, Peter L (1999) *Urban Jamaican Creole: Variation in the mesolect.* Varieties of English Around the World, G17, Philadelphia & Amsterdam: John Benjamins.

Paunonen, Heikki (1995) Suomen kieli Helsingissä [The Finnish Language in Helsinki]. Helsingin yliopiston suomen kielen laitos, Helsinki.

Prince, Alan (2002) Entailed Ranking Arguments. *Rutgers Optimality Archive* **500**: 1–117, rOA-500. `roa.rutgers.edu`.

Prince, Alan & Adrian Brasoveanu (2005) Ranking and Necessity. *Rutgers Optimality Theory Archive* **(ROA) 794-1205**, roa.rutgers.edu.

Prince, Alan & Paul Smolensky (1993/2004) *Optimality Theory: Constraint Interaction in Generative Grammar*. MIT Press.

Prince, Alan & Bruce Tesar (1999) Learning Phonotactic Distributions, Ms. ROA 535.

Prince, Alan & Bruce Tesar (2004) Learning Phonotactic Distributions. In *Fixing Priorities: Constraints in Phonological Acquisition*, Joe Pater Kager, Rene & Wim Zonneveld, eds., Cambridge University Press.

Resnik, Philip & Eric Hardisty (2009) Gibbs Sampling for the Uninitiated. Online manuscript: http://www.umiacs.umd.edu/ resnik/pubs/gibbs.pdf.

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

Riggle, Jason (2009a) The Complexity of Ranking Hypotheses in Optimality Theory. *Computational Linguistics* **35**(1): 47–59.

Riggle, Jason (2009b) Generating Contenders. ROA 1044-0809.

Santa Ana, Otto (1991) *Phonetic Simplification Processes in the English of the Barrio: A Cross-Generational Sociolinguistic Study of the Chicanos of Los Angeles*. Ph.D. thesis, University of Pennsylvania.

Santa Ana, Otto (1992) Chicano English evidence for the exponential hypothesis: A variable rule pervades lexical phonology. *Language Variation and Change* **4**: 275–288.

Santa Ana, Otto (1996) Sonority and syllable structure in Chicano English. *Language Variation and Change* **8**: 63–89.

Selkirk, Elisabeth (1981) *Epenthesis and degenerate syllables in Cairene Arabic*. Cambridge MA: MIT, 111–140.

Steriade, Donca (1997) Phonetics in Phonology : the case of laryngeal neutralisations. UCLA Ms, http://www.linguistics.ucla.edu/people/steriade/papers/phoneticsinphonology.pdf.

Stevens, K., S. J. Keyser, & H. Kawasaki (1986) *Toward a phonetic and phonological theory of redundant features*. Hillsdale, NJ: Lawrence Erlbaum, 426–449.

Tagliamonte, Sali & Rosalind Temple (2005) New perspectives on an ol' variable: (t,d) in British English. *Language Variation and Change* **17**: 281–302.

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

Tesar, Bruce (1997) Multi-Recursive Constraint Demotion, ms., Rutgers University.

Tesar, Bruce (1998) An iterative strategy for language learning. *Lingua* **104**: 131–145.

Tesar, Bruce & Paul Smolensky (1993) The Learnability of Optimality Theory: An Algorithm and Some Basic Complexity Results. Tech. Rep. CU-CS-678-93, University of Colorado at Boulder Department of Computer Science.

Tesar, Bruce & Paul Smolensky (1996) Learnability in Optimality Theory (long version). Tech. rep., Deparment of Cognitive Science, Johns Hopkins University.

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

Tesar, Bruce & Paul Smolensky (2000) *Learnability in Optimalty Theory*. Cambridge: MIT Press.

Valiant, Leslie G. (1979) The Complexity of Computing the Permanent. *Theoretical Computer Science* **8**: 189–201.

Vapnik, V. N. & A. Cervonenkis (1971) On the uniform convergence of relative frequencies of events to their probabilities. *tpa* **16**: 264–280.

Wilson, Colin (2006) Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* **30**: 945–982.

Wolfram, Walt & Donna Christian (1976) Appalachian Speech. Washington, DC: Center for Applied Linguistics.

Wolfram, Walter A. (1969) *A Sociolinguistic Description of Detroit Negro Speech*. Washington, D.C.: Center for Applied Linguistics.

Zoll, Cheryl (1998) *Parsing below the Segment in A Constraint-Based Framework*. Stanford: CSLI Publications.