Robust Identification in the Limit from Incomplete Positive Data

Philip Kaelbling, Dakotah Lambert, Jeffrey Heinz

Robust Identification in the Limit from Incomplete Positive Data

Philip Kaelbling[1], Dakotah Lambert[2], and Jeffrey Heinz[3]

[1] pkaelbling@wesleyan.edu Department of Computer Science, Wesleyan University
[2] dakotahlambert@acm.org Université Jean Monnet Saint-Étienne, CNRS, Institut d Optique Graduate School, Laboratoire Hubert Curien UMR 5516
[3] jeffrey.heinz@stonybrook.edu Department of Linguistics and Institute for Advanced Computational Science, Stony Brook University

Abstract. Intuitively, a learning algorithm is robust if it can succeed despite adverse conditions. We examine conditions under which learning algorithms for classes of formal languages are able to succeed when the data presentations are systematically incomplete; that is, when certain kinds of examples are systematically absent. One motivation comes from linguistics, where the phonotactic pattern of a language may be understood as the intersection of formal languages, each of which formalizes a distinct linguistic generalization. We examine under what conditions these generalizations can be learned when the only data available to a learner belongs to their intersection. In particular, we provide three formal definitions of robustness in the identification in the limit from positive data paradigm, and several theorems which describe the kinds of classes of formal languages which are, and are not, robustly learnable in the relevant sense. We relate these results to classes relevant to natural language phonology.

Keywords: identification in the limit, grammatical inference, regular languages, model theory, locally testable, piecewise testable

1 Introduction

This paper presents an analysis of Gold-style learning [8] of formal languages from systematically deficient data and the conclusions one can draw from three different definitions of correctness. For our purposes, the omissions in the data arise from other constraints. We specifically consider data presentations which are the intersection of two languages, one of which is the target of learning.

The analysis is illustrated with, and motivated by, classes of formal languages that are both computationally natural and of particular interest to natural language phonology [11]. These classes are well-studied subregular classes which often have multiple characterizations, including language-theoretic, automata-theoretic, logical, and algebraic. The classes used to exemplify this work include the Strictly Local languages [20], the Strictly Piecewise languages [23], and the Tier-Based Strictly Local languages [13, 18].

As an example, suppose we are interested in learning the formal language L containing all strings which do not contain bb as a substring. As explained in more detail in section 2, a positive data presentation for this language would eventually include strings like *babaaca* (because it does not contain the bb substring). Now suppose the observable sequences are also subject to a constraint that words must not contain a b preceding c at any distance. In this case, the word *babaaca* would *not* be part of the data presentation. Is it still possible to learn L if such words are never presented?

We provide three formal definitions of robustness in the identification in the limit from positive data learning paradigm, and several theorems which describe the kinds of classes of formal languages which are, and are not, robustly learnable in the relevant sense.[4]

We opt to explore a modification of Gold-style instead of the Probably Approximately Correct learning framework (PAC 26) in order to avoid the issue of defining a distance between formal languages. In the PAC framework, data is drawn from a stationary distribution, and a learner is required to be reasonably correct based on the data presented to it. This data could be considered "deficient" if the distribution poorly represents the target concept. As discussed by Eyraud et al. [5], PAC is not necessarily well-suited for the problem of learning formal languages. The approximate nature of correctness in PAC requires a notion of distance between formal languages to judge the quality of a proposed solution. There are many feasible metrics [22, 4, 25], and the PAC results are expected to be sensitive to the chosen metric. We choose to study robust learning in a model where this is not a concern.

Generally, research on identification in the limit from positive data in the presence of data inaccuracies have identified the following three types [16, chap. 8].

1. Noisy data. A data presentation for a formal language L includes intrusions from the complement of L.
2. Incomplete data. A data presentation for a formal language L omits examples from L. That is, if E represents the set of omitted examples, the presentation is actually a text for $L - E$ rather than for L itself.
3. Imperfect data. Data presentations for a formal language L which both includes intrusions from the complement of L and omits examples from L.

In this work we only study the identification in the limit from incomplete positive data. Fulk and Jain [7] study the problem of learning from incomplete data when there are finitely many data points omitted, which is unlike the case we consider where there can be infinitely many omitted examples. On the other hand, Jain [14] considers cases where there are infinitely many omissions. This work, like that of Fulk and Jain [7], establishes hierarchies of classes that are exactly learnable or not in the presence of inaccurate data. While our first theorem

[4] The notion of robustness studied here is different from the one studied by Case et al. [3]. There, a class is "robustly learnable" if and only if its effective transformations are learnable too. As such, their primary interest is classes "outside the world of the recursively enumerable classes." This paper uses the term "robustly learnable" to mean learnable despite the absence of some positive evidence.

regards exact identification, our other theorems relax that requirement. Additionally, Freivalds et al. [6] and Jain et al. [15] consider learning from a finite set of examples which contains at least all the *good examples*, which intuitively are well-chosen illustrations of the language. As learners must succeed with finitely many examples, this scenario potentially omits infinitely many. The scenario we consider, however, does not make provision for good examples.

As just mentioned, our strongest definition of correctness requires a learner to recover exactly the target language on a text drawn from the intersection of two languages. A key result under this definition is that of its strength, namely that very few classes of languages are independent of interference under it.

Our main result comes under our second notion of correctness, strong robustness, which requires a learner only to recover a language compatible with the target grammar when restricted to the intersection. Under this definition, we show that classes of languages identifiable with string extension learners [10, 12] are strongly robust in the presence of interference from all other classes of languages.

Finally, we present our weakest correctness definition, weak robustness, removing the prior requirement of learning a language in the correct concept class. Under this definition the class of Tier-Based Strictly Local languages is robustly learnable, specifically by the algorithm presented by Lambert [17].

More generally, the results here are related to the question of whether two learnable classes of languages C and D imply a successful learning algorithm for the class of languages formed by their pointwise intersection $\{L_C \cap L_D : L_C \in C, L_D \in D\}$. In the case of identification in the limit from positive data, the answer in the general case is negative.[5] However, the results above – in particular strong robustness – help us understand the conditions sufficient for this situation to occur. In this way, this work helps take a step towards a compositional theory of language learning.

2 Background

2.1 Identification in the Limit

Gold [8] introduced a number of different definitions of what it means to learn a formal language. In this work, we concern ourselves only with the notion of learnability in the limit from positive data (ILPD), which is also called explanatory learning from text [16].

Let Σ denote a fixed finite set of symbols and Σ^* the set of all strings of finite length greater than or equal to zero. A formal language L (a constraint) is a subset of Σ^*.

A **text** t can be thought of as a function from the natural numbers to Σ^*. Let $\overrightarrow{t_n}$ represent the sequence $\langle t_0, t_1, \ldots, t_{n-1} \rangle$, the length-$n$ initial segment of a

[5] Alexander Clark (personal communication) provides a counterexample. Let $C = \{L_\infty, L_1, \ldots\}$ where $L_n = \{a^m : 0 < m < n\} \cup \{b^{n+1}\}$ and $L_\infty = a^+ \cup \{b\}$. Let $D = \{a^*\}$. Both classes are ILPD-learnable but $\{L_C \cap L_D : L_C \in C, L_D \in D\}$ is not.

text t. Note $\vec{t}_n$ is always of finite size. Let $\mathbb{T}$ denote all texts and $\vec{\mathbb{T}}$ represent the collection of finite initial segments of these texts. Using the notation for sequences instead of functions, we write t_i instead of $t(i)$. For a text t, let $\text{CT}(t) = \{w : \exists n \in \mathbb{N}, t_n = w\}$, and similarly $\text{CT}(\vec{t}_n) = \{w : 0 \leq i < n, t_i = w\}$. For a given language L, we say t is a text for L if and only if $\text{CT}(t) = L$. The set of all texts for a language L is denoted $\mathbb{T}_L$.

Osherson et al. [21] discuss a modification that allows a text to exist for the empty language: t_n may be either an element of L or a distinct symbol $\circledcirc$ representing a lack of data. Consequently, the empty language has exactly one text; for each $n \in \mathbb{N}, t_n = \circledcirc$. We denote this text with $t^{\circledcirc}$.

We denote with $\mathbb{G}$ a collection of grammars, by which we mean a set of finitely-sized representations, each of which is associated with a formal language in a well-defined way. If a grammar $G \in \mathbb{G}$ is associated with a formal language L, we write $\mathcal{L}(G) = L$, and say that G recognizes, accepts, or generates L.

An ILPD-learner is a function $\varphi\colon \vec{\mathbb{T}} \to \mathbb{G}$. The learner **converges on** t iff there is some grammar $G \in \mathbb{G}$ and some $i \in \mathbb{N}$ such that for all $j > i$, $\mathcal{L}(\varphi(\vec{t}_j)) = \mathcal{L}(G)$. If, for all texts t for a language L, it holds that φ converges on t to a grammar G such that $\mathcal{L}(G) = L$, then φ **identifies** L **in the limit**. If this holds for all languages of a class C, then φ identifies C in the limit and can be called a C-ILPD-learner.

Angluin [1] proved the following theorem.

Theorem 1. *Let C be a collection of languages which is indexed by some computbale function. C is* ILPD*-learnable iff there exists a computably enumerable family of finite sets S such that for each $L_i \in C$, there exists a finite $S_i \subseteq L_i$ such that for any $L_j \in C$ which contains S_i, $L_j \not\subset L_i$.*

The finite set S_i is called a **telltale set** for L_i with respect to C.

2.2 String Extension Learning

Heinz [10] introduced string extension learning, a general type of set-based ILPD-learning algorithm based solely on the information contained in strings. A generalization of this technique is discussed here; for another generalization, see Heinz et al. [12]. The learner is defined as follows:

$$\varphi(\vec{t}_i) = \begin{cases} \varnothing & \text{if } i = 0,\ t_i = \circledcirc \\ \varnothing \oplus f(t_i) & \text{if } i = 0,\ t_i \neq \circledcirc \\ \varphi(\vec{t}_{i-1}) & \text{if } i \neq 0,\ t_i = \circledcirc \\ \varphi(\vec{t}_{i-1}) \oplus f(t_i) & \text{otherwise,} \end{cases}$$

where f is a function that extracts information from a string and $\oplus$ is an operation for inserting information into a grammar.

Finally there is an interpretation relation $\models$, describing the language represented by the grammar. The statement $w \models G$ means that w satisfies the interpretation of G given by this relation. The language of the grammar G then

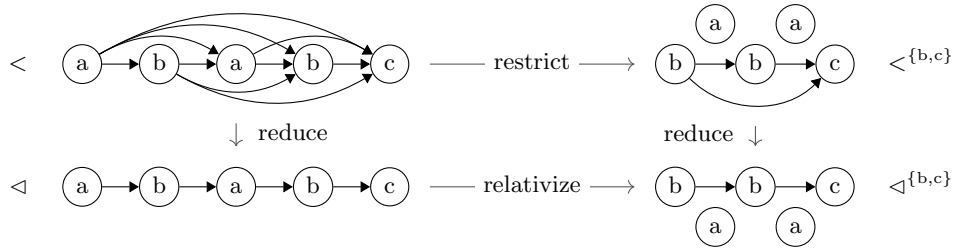

Fig. 1. Piecewise, local, and tier-based factors [18].

is $\mathcal{L}(G) = \{w : w \models G\}$. If it is the case that $w \models (G \oplus f(w))$ for all G and w, then φ is **consistent**. Often, a learner is defined with the following constraints: f extracts a set of factors of some sort, grammars are sets of the same type, $\oplus$ is set union, and $w \models G$ iff $f(w) \subseteq G$. Such a learner is consistent.

This may appear to simply refer to any incremental learner, but we add one further restriction: the class of string extension learners is defined to be the subfamily of these learners that are guaranteed to converge on any text. Every learner defined by Heinz et al. [12] satisfies this property, as does the incremental learner defined by Lambert [17] that will be explored further in section 3.3.

2.3 Model-theoretic Factors and Related Formal Language Classes

Lambert et al. [19] discuss a model-theoretic notion of factors. A simplification, sufficient for the present discussion, involves only a collection of symbol-labeled domain elements along with a binary relation between them. The relation induces a graph. A k-factor of a model is a collection of k nodes connected by the transitive closure of the relevant binary relation. Grammars and formal languages can be defined in terms of such k-factors.

Different binary relations give rise to different k-factors. Figure 1 shows some examples for the word *ababc*. The precedence relation ($<$) (upper left) yields several 3-factors: *aab*, *aac*, *aba*, *abb*, *abc*, *bab*, *bac*. The sucessor relation ($\lhd$) (lower left) yields fewer 3-factors: *aba*, *bab*, *abc*. The precedence relation can be restricted to a tier $T \subseteq \Sigma$ of salient symbols. In Figure 1, $T = \{\mathrm{b}, \mathrm{c}\}$. It follows that the tier-precedence relation ($<^{\{\mathrm{b},\mathrm{c}\}}$) (upper right) yields only the 2-factors: *bb*, *bc*. The tier-successor relation ($\lhd^{\{\mathrm{b},\mathrm{c}\}}$) (lower right), is a binary relation relating positions on the tier to the positions that are "next" on the tier. Hence, it yields the 2-factors *bb*, *bc*.

Consider grammars G which are sets of k-factors and say $w \models G$ only if the k-factors in w are a subset of G. Such a definition for the relations $\lhd, <, \lhd^T$ yields the classes of formal languages that are testable in the strict sense: strictly k-local (SL) [20], strictly k-piecewise (SP) [23, 9], or tier-based strictly k-local (TSL)

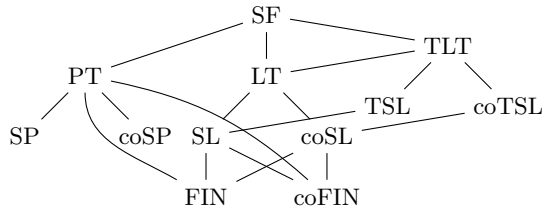

Fig. 2. A hierarchy by subclass of subregular classes.

[13, 18], respectively.[6] Such classes are string extension learnable where f maps w to its k-factors, and $\oplus$ is set union [10].

Next consider grammars G which are sets of sets of k-factors and and say $w \models G$ only if the k-factors in w are an element of G. Such a definition for the $\lhd, <$, and $<^T$ relations yields the testable classes: locally k-testable (LT) [20], piecewise k-testable (PT) [24], or tier-based locally k, T-testable (TLT) [18], respectively. These classes are string extension learnable where f maps w to its k-factors, and $\oplus$ is set insertion [10].

The model-theoretic perspective combined provides a uniform way to characterize these well-studied classes. It also fits well into the generalized string extension scheme above because both the functions f and the operation $\oplus$ are understood simply in terms of k-factors.

The aforementioned results hold for classes where the parameters k, T are fixed. It is of special interest in linguistics to learn the family of k, T-TSL languages when k is fixed but T is not. Lambert [17] provides an incremental learning algorithm for this class of languages, and in section 3.3, it is shown that this class is robustly learnable in a weak sense.

The aforementioned language classes and others are shown in Figure 2 including some complement classes indicated with the prefix 'co'. Many other subregular classes exist, but only these few will be discussed in this work. For more details, readers are referred to Lambert et al. [19]. SF is star-free [20].

3 Robustness

When two or more constraints interact, the intersection of their licensed sets may no longer provide enough data to learn the constraints. Formally, we are interested in whether languages in C can be ILPD-learned from texts that are systematically deficient in some way.

A rare sort of robustness is when the individual constraints are retrievable exactly from the intersection, entirely unaffected by the increased sparsity of data. Consider any $L \in C$ and any other language M. This sort of robustness would mean that L is ILPD-learnable on all texts for $L \cap M$. In this case, there is a learner φ which can still exactly learn L despite interference from M.

[6] Technically, local classes need to be augmented with symbols marking word edges.

If L might be unrecoverable, there could still be a guarantee that one can recover a grammar whose language produces the same intersection. This sort of robustness would mean that a learner φ, on any text for $L \cap M$, need only converge to a grammar G such that $\mathcal{L}(G) \cap M = L \cap M$. In this case, there is a learner φ which may fail to exactly learn L, but learns another language which is 'good enough' up to M.

We study two types of this latter form of robustness. If this equivalent constraint $\mathcal{L}(G)$ always belongs to the same class C as the original constraint L, then that class is strongly robust in the presence of the M; else the robustness is only weak. This section discusses these notions in order of decreasing strength.

3.1 Unaffectedness

The strongest form of robustness is that in which constraints are guaranteed to be extractable without loss of information from the interacting pattern.

Definition 1. *A class C is **unaffected by** another class D iff there is a learning function φ such that for all languages $L \in C$ and all languages $M \in D$, φ converges to a grammar for L on all texts for $L \cap M$.*

C is **affected by** D iff it is not unaffected by D. The strictness of this criterion is suggested by the following theorem.

Theorem 2. *Every* ILPD*-learnable class which includes two languages is affected by any class D where $\varnothing \in D$.*

Proof. Let C be a class which contains distinct languages L_1, L_2 and let D be a class containing $\varnothing$. There is only one text for $L_1 \cap \varnothing = L_2 \cap \varnothing = \varnothing$, which is $t^{\circledcirc}$. If a learner exists which correctly converges to L_1 on $t^{\circledcirc}$, it would not correctly converge to L_2 on this same text and vice versa. □

Nearly every class in Figure 2 contains the empty set. Over a non-empty alphabet, the sole exception is the class of cofinite languages coFIN, which may exclude only finitely many strings. However, even the cofinite languages can be shown to affect classes with general properties.

Theorem 3. *Every* ILPD*-learnable class which includes two distinct finite languages is affected by co*FIN*.*

Proof. Let C be a class which contains two distinct finite language L_1 and L_2. Because they are finite, their complements $\complement L_1$ and $\complement L_2$ belong to coFIN. The intersection $L_1 \cap \complement L_1 = L_2 \cap \complement L_2 = \varnothing$ has but a single text: $t^{\circledcirc}$. If a learner exists which correctly converges to L_1 on $t^{\circledcirc}$, it would not correctly converge to L_2 on this same text and vice versa. □

Because both FIN and SP contain both the empty language and at least one nonempty finite language (Σ^k for nonzero k), neither they nor any of their superclasses can be unaffected by either the FIN or coFIN classes.

The SL and SP classes are not saved from being affected by COFIN even if restricted to their subclasses containing only infinite languages. Let L be the language of all and only those words over Σ which, if longer than n symbols, do not contain a specific symbol $a \in \Sigma$. In other words, a appears only in words shorter than n symbols. Further let M be the cofinite language containing all and only those words over Σ of length at least n. The intersection of L and M is $(\Sigma - \{a\})^{\geq n}$, an $(n+2)$-SL proper subset of L. and which contains all and only those piecewise factors (subsequences) over $\Sigma - \{a\}$. For $k < n + 2$, all and only those local factors (substrings) over this same alphabet are present in the language. In any case, because a does not appear in the data, it will be forbidden. Therefore for any parameters, the SL and SP learners will converge on some superset of this intersection which contains no instances of a, rather than on L itself.

In general, ILPD-learnable classes are affected by certain overlapping classes.

Theorem 4. *If L is a language in an* ILPD*-learnable class C, and $M \subset L$ belongs to $C \cap D$ for some class D, then C is affected by D.*

Proof. Let C and D be language classes such that C is ILPD-learnable and contains a language L and D overlaps with C such that $C \cap D$ contains a language $M \subset L$. Then $L \cap M = M$ and, since C is ILPD-learnable and M is in C, the C-learner must converge to M on a text for $L \cap M$. □

Corollary 1. *An* ILPD*-learnable class that contains a language L is affected by all subclasses of itself that contain any smaller language $M \subset L$.*

Our main result on unaffectedness is a characterization of which classes C are not affected by which classes D. We prove this result by adapting Angluin's [1980] characterization of the ILPD-learnable classes. Following Osherson et al. [21], we obtain this result via an adaptation of Blum and Blum's [1975].

Theorem 5. *Let $L, M \subseteq \Sigma^*$ and suppose φ is a learning function which identifies L on all texts for $L \cap M$. Letting $\mathbb{T}_{L \cap M}$ denote all texts for $L \cap M$, then there is some $\sigma \in \overrightarrow{\mathbb{T}}_{L \cap M}$ such that*

1. $\text{CT}(\sigma) \subseteq L \cap M$
2. *$\varphi(\sigma) = G$ where $\mathcal{L}(G) = L$.*
3. $\forall \tau \in \overrightarrow{\mathbb{T}}_{L \cap M}[\text{CT}(\tau) \subseteq L \cap M \rightarrow \varphi(\sigma\tau) = \varphi(\sigma)]$.

In other words, if a learner identifies a language L in the limit on texts from $L \cap M$, then there is some point in each text from which the learner is 'locked' into a particular grammatical hypothesis.

Proof. The proof is by contradiction. If the theorem is not true, it must be the case that for every $\sigma \in \overrightarrow{\mathbb{T}}_{L \cap M}$ such that (1) and (2) above are true, there is a $\tau \in \overrightarrow{\mathbb{T}}_{L \cap M}$ such that $\text{CT}(\tau) \subseteq L \cap M$, but $\varphi(\sigma\tau) \neq \varphi(\sigma)$.

If this is true, then it is possible to construct a positive text for $L \cap M$ with which φ fails to converge, thus contradicting the initial assumption that φ

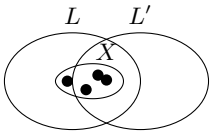

Fig. 3. Dots represent a telltale set for L, distinguishing it from L', despite interference from some third language M whose intersection with L is X.

identifies L in the limit on all texts for $L \cap M$. It will be helpful to consider some text t for $L \cap M$. Construct the new text q recursively as follows. Let $q^{(0)} = t_0$. Note that $\text{CT}(q^{(0)})$ is a subset of $L \cap M$. $q^{(n)}$ is determined by the following cases:

Case 1. $\varphi(q^{(n-1)}) = G$ where $\mathcal{L}(G) = L$. Then by the reductio assumption we know that there exists some τ_n such that $\text{CT}(\tau_n) \subseteq L \cap M$ and $\varphi(q^{(n-1)}\tau_n) \neq G$. Let $q^{(n)} = q^{(n-1)}\tau_n t_n$, and note that $\text{CT}(q^{(n)})$ is a subset of $L \cap M$.

Case 2. $\varphi(q^{(n-1)}) = G$ where $\mathcal{L}(G) \neq L$. Then let $q^{(n)} = q^{(n-1)}t_n$. As in the other case, $\text{CT}(q^{(n)})$ is a subset of $L \cap M$.

Observe that $\text{CT}(q) = L \cap M$ and thus q is a text for $L \cap M$. This is because t is a text for $L \cap M$ and an element of t is added to q at every step in its construction. However, φ fails to converge on q because for every $i \in \mathbb{N}$ such that $\varphi(q^{(i)}) = G$ where $L = \mathcal{L}(G)$, there is a later point $q^{(i+1)}$, where $\varphi(q^{(i+1)})$ does not equal G by the construction above (Case 1). Therefore, we contradict the original assumption that φ identifies L on all texts for $L \cap M$ and the reductio assumption is false, proving the theorem. □

Now one can state a property of all classes C which are unaffected by another class D. A crucial concept is the **telltale set despite interfence** of a language in some class, defined below and demonstrated in Figure 3.

Definition 2. *Any finite $S \subset \Sigma^*$ is a **telltale set** of a language $L \in C$ **despite interference** from $M \in D$ iff $S \subseteq L \cap M$ and for any $L' \in C$ such that $L' \cap M$ contains S, it holds that $L' \not\subset L$.*

If a learner guesses language L upon observing a telltale set for L despite interference from M, then it is guaranteed that the learner has guessed the smallest language in C which contains the sample. Thus the learner has not overgeneralized as no other language in the class of languages which includes the sample is strictly contained within L.

Theorem 6. *Let C, D be collections of languages which are both indexed by some computable functions. C is unaffected by D iff there exists a computably enumerable family of finite sets S such that for each $L_i \in C$ and $M_j \in D$, there exists a finite $S_{i,j} \subseteq L_i \cap M_j$ such that $S_{i,j}$ is a telltale set for L_i despite interference from M_j.*

Proof. ($\Rightarrow$) Suppose C is unaffected by D. Then there exists φ which for all $L \in C$ and $M \in D$ identifies L despite interference from M. By Theorem 5, there is a locking sequence σ for L where $\text{CT}(\sigma) \subseteq L \cap M$. We show that the $\text{CT}(\sigma)$ is

a telltale set for L despite interference from M. First, as locking sequences are finite, $\text{CT}(\sigma)$ is finite too. Now for contradiction assume that there is some $L' \in C$ such that $\text{CT}(\sigma) \subseteq L'$, and $L' \subset L$. Then, per Theorem 5, φ fails to identify L' on a text t for L' where t begins with $\sigma\tau$, as $\varphi(\sigma) = G$ where $\mathcal{L}(G) = L$.

($\Leftarrow$) Assume that for every $L \in C$ and $M \in D$, L has a telltale set S despite interference from M, and further assume some enumeration of grammars and of these sets. Let X be the first (only) telltale set such that $X \subseteq \text{CT}(\overrightarrow{t_i})$ and let $G = \varphi(\overrightarrow{t_i})$ be the first grammar in the enumeration such that $X \subseteq \text{CT}(\overrightarrow{t_i}) \subseteq \mathcal{L}(G)$ if such objects exist, otherwise let X and G be the first set and grammar in their respective enumerations.

Now consider any $L \in C, M \in D$, any text t for $L \cap M$ and let G be the n-th grammar in the enumeration, but the first such that $\mathcal{L}(G) = L$. As S is finite, there is an i_1 such that $S \subseteq \text{CT}(\overrightarrow{t_{i_1}}) \subseteq \mathcal{L}(G)$. Thus for all $j \geq i_1$, $\varphi(\overrightarrow{t_j})$ returns G unless there is some G' earlier in the enumeration such that $\mathcal{L}(G') \in C$, and S' is a telltale set for $\mathcal{L}(G')$ despite interference from M and $S' \subseteq \text{CT}(\overrightarrow{t_{i_1}}) \subseteq \mathcal{L}(G')$.

However, we can find $i_2 \geq i_1$ which ensures that no such G' exists. Suppose there is some G' earlier in the enumeration such that $S' \subseteq \text{CT}(\overrightarrow{t_{i_1}}) \subseteq \mathcal{L}(G')$. Then $\mathcal{L}(G')$ cannot properly include L because S' is a telltale set for $\mathcal{L}(G')$ and both $L, \mathcal{L}(G') \in C$. Thus there must be some sentence s in $L \cap M$ that is not in $\mathcal{L}(G') \cap M$. As t is a text for $L \cap M$, there is a k such that $s \in \text{CT}(\overrightarrow{t_k})$.

Thus for any $j \geq k$, $\varphi(\overrightarrow{t_j}) \neq G'$ since $s \notin \mathcal{L}(G')$ and thus $\text{CT}(\overrightarrow{t_j}) \not\subseteq \mathcal{L}(G')$. It follows that for each G_m (such that $\mathcal{L}(G_m) \in C$) which occurs earlier in the enumeration than G (i.e. $m < n$), there is some k_m such that $\text{CT}(\overrightarrow{t_{k_m}}) \not\subseteq \mathcal{L}(G_m)$. There are only finitely many grammars before G in the enumeration and so by letting i_2 be the largest element of $\{i_1\} \cup \{k_m : 0 \leq m < n\}$, we guarantee that for any $j \geq i_2$, $\varphi(\overrightarrow{t_j}) = G$. □

In short, for each $L \in C$ and for each $M \in D$, there must be a telltale set for L contained with $L \cap M$. This highlights the difficulty of this paradigm. The only classes unaffected by others to our knowledge are the singleton language classes $\{L\}$, which are unaffected by every class D. Future work involves identification of non-trivial C, D of linguistic interest such that C that is unaffected by D.

3.2 Strong Robustness

There are few cases of classes being unaffected by another. Yet this raises a question: should we care if the learned constraint is incorrect only on data that it cannot encounter? Learning a language consistent with the data should suffice.

Definition 3. *A class C is **strongly robust in the presence of** another class D iff there exists a learning function φ such that for all languages $L \in C$ and $M \in D$, there exists a grammar G such that $\mathcal{L}(G) \in C$, $\mathcal{L}(G) \cap M = L \cap M$, and φ converges to G on all texts for $L \cap M$.*

Theorem 7. *If a class C is intersection-closed (i.e. closed under finitary intersection) and string extension learnable by a learner φ which for any initial*

segment of a text $\overrightarrow{t_i}$ guarantees as output a unique minimum grammar $G = \varphi(\overrightarrow{t_i})$ where $\mathcal{L}(G) \in C$ such that $\text{CT}(\overrightarrow{t_i}) \subseteq \mathcal{L}(\varphi(\overrightarrow{t_i}))$[7], then C is strongly robust in the presence of any class D.

Proof. Let C be an intersection-closed, string extension learnable class whose associated learner φ guarantees a unique minimum grammar whose language is in C and compatible with the received text. That is, given any text t it holds that for any initial segment $\overrightarrow{t_i}$ of t we have $\text{CT}(\overrightarrow{t_i}) \subseteq \mathcal{L}(\varphi(\overrightarrow{t_i}))$ and there is no grammar $X \neq \varphi(\overrightarrow{t_i})$ such that $\mathcal{L}(X) \in C$ and $\text{CT}(\overrightarrow{t_i}) \subseteq \mathcal{L}(X) \subseteq \mathcal{L}(\varphi(\overrightarrow{t_i}))$ Further, let $L \in C$, let M be any language, and let G be the grammar obtained by applying φ to a text drawn from $L \cap M$.

If $L \subset \mathcal{L}(G)$ then $\mathcal{L}(G)$ is not the minimal language in C compatible with the data, contradicting the assumption.

Suppose by way of contradiction that $\mathcal{L}(G) \cap M \neq L \cap M$. If $\mathcal{L}(G) \cap M \subset L \cap M$ then there exists some $w \in L \cap M$ such that $w \not\models G$. But w is in the text, violating the assumption that φ is compatible with the data it receives. Then it must be that there is some $v \models G$ such that $v \in M - L$, and notably v cannot appear in the text. The language $\mathcal{L}(G) \cap L$ is in C by intersection-closure, is a subset of $\mathcal{L}(G)$ by definition, and does not contain v; this violates the assumption that φ returns a grammar for the smallest language compatible with the text.

The only remaining option is that $\mathcal{L}(G) \cap M = L \cap M$. As M was unrestricted, it follows that C is strongly robust in the presence of any class D. □

Each of the FIN, SL, LT, SP, and PT classes are intersection-closed and, when appropriately parameterized, string extension learnable in a way that guarantees a unique minimum language consistent with the text [10]. Therefore each of these classes is strongly robust in the presence of any class.

Corollary 2. *A intersection-closed class of* ILPD*-learnable languages C is strongly robust in the presence of any of its subclasses $C' \subseteq C$.*

Proof. Let C and D be classes of languages such that $D \subseteq C$, where C is ILPD-learnable and intersection-closed. Let $L \in C$ and $M \in D$. Then the intersection $L \cap M$ is in C. As C is ILPD-learnable, the intersection is learned exactly. □

If two classes, A and B, are string-extension learnable by φ_A and φ_B, respectively, then one can define a string-extension learner for their pointwise intersection, $A \Cap B = \{a \cap b \colon a \in A, b \in B\}$, as follows:

$$f(w) = \langle f_A(w), f_B(w) \rangle$$
$$\langle G_A, G_B \rangle \oplus \langle x, y \rangle = \langle G_A \oplus_A x, G_B \oplus_B y \rangle$$
$$w \models \langle G_A, G_B \rangle \iff w \models G_A \wedge w \models G_B.$$

The learner thus defined is a **pointwise string extension learner** for $A \Cap B$.

For example, the intersection closure of the TSL class, MTSL, is pointwise string extension learnable. Given the alphabet Σ over which the text is drawn,

[7] Note that this is a stronger guarantee than consistency.

construct $2^{|\Sigma|}$ k-SL learners in parallel, one for each subset of Σ. Each of these learners will be responsible for learning the constraints over its associated tier, by first projecting to that subset of Σ the words it encounters, then extracting the local factors of the result. Such a learner is not particularly efficient; for an alphabet of ten unique symbols, this results in 1,024 parallel SL learners.

Theorem 8. *If A and B are intersection-closed and string extension learnable, and both A and B are strongly robust in the presence of the other pointwise intersected with some third class C, then the class $A \Cap B$ is strongly robust in the presence of C.*

Proof. Let A and B be string extension learnable classes such that A is strongly robust in the presence of $B \Cap C$ and B is strongly robust in the presence of $A \Cap C$. Let φ_A and φ_B be the learners for A and B, respectively. Finally, let $L = L_A \cap L_B$ be some language in $A \Cap B$ and let $L' \in C$. Given some text t drawn from $L \cap L'$, $\mathcal{L}(\varphi_A(t)) \cap L_B \cap L' = L_A \cap L_B \cap L' = L \cap L'$, and $\mathcal{L}(\varphi_B(t)) \cap L_A \cap L' = L_A \cap L_B \cap L' = L \cap L'$ by strong robustness. The pointwise string extension learner φ for $A \Cap B$ exists such that $\mathcal{L}(\varphi(t)) = \mathcal{L}(\varphi_A(t)) \cap \mathcal{L}(\varphi_B(t)) = L \cap L'$. Therefore $A \Cap B$ is strongly robust in the presence of C. □

Suppose that A and B are classes that satisfy the conditions of Theorem 7. That is, they are string extension learnable in a way that guarantees as output a unique minimum language in the respective class, compatible with the text they were given. Then they are strongly robust in the face of any interactions, by that theorem. It then follows immediately from Theorem 8 that their pointwise intersection $A \Cap B$ is similarly strongly robust in the presence of any interactions. However, we cannot turn this around and make strong claims about A or B based on the learnability of $A \Cap B$. Consider the case where A contains the empty language and B is the singleton class containing only the empty language. Then $A \Cap B = B$, which as a singleton class is unaffected by any other class C, no matter what A is. Furthermore, suppose A and C are identical and intersection-closed, but not ILPD-learnable. Concretely, suppose $A = C = \text{Reg}$, the class of all regular languages. Then A cannot be strongly robustly learnable in the presence of C, because it is not learnable in the first place.

3.3 Weak Robustness

An ILPD-learner only guarantees convergence to a language in its target class when presented with a text for such a language. When given a text from a language not in the target class, the result can be anything, even a lack of convergence. A weaker form of robustness might then be a guarantee that the learner will necessarily converge to some language consistent with the data, even if that language is not in the target class C.

Definition 4. *A class C is **weakly robust in the presence of** another class D iff there exists a learning function φ such that for all languages $L \in C$ and $M \in D$, there exists a grammar G such that $\mathcal{L}(G) \cap M = L \cap M$, and φ converges to G on all texts for $L \cap M$.*

This suggests the existence of a third class X, a superclass of C, where X is strongly robust in the presence of D.

As a concrete example of weak robustness, we shall consider $C = D =$ TSL. Membership in TSL is closure under suffix substitution on some tier T, and under insertion and deletion of elements not on that tier. That is, for $x \in T^k$, if $u_1xu_2 \in L$, $v_1xv_2 \in L$, then $u_1xv_2 \in L$, and if $u_1au_2 \in L$ for $a \notin T$ then $u_1u_2 \in L$ and vice versa. Let $\Sigma = \{a, b, c\}$, L be the language forbidding ab on the $\{a, b\}$ tier, M be that forbidding bc on the $\{b, c\}$ tier. The intersection $L \cap M$ is not TSL for any tier T. No letter is freely insertable or deletable in $L \cap M$, so $T = \{a, b, c\}$. Notice that $b(a^k)a \in L \cap M$ and $a(a^k)c \in L \cap M$. If it were TSL, then we would expect by suffix-substitution that $b(a^k)c \in L \cap M$, but it is not, as $b(a^k)c \notin M$. It follows that $L \cap M \notin$ TSL.

Recall the earlier discussion on pointwise intersections. Suppose that A and B are classes such that $A \Cap B$ is ILPD-learnable. We have already noted that this provides no guarantees regarding the robustness or even learnability of A or B in the presence of some third class C. However, we can state that A and B are weakly robust in the presence of one another, as the learner for $A \Cap B$ is by definition guaranteed to converge exactly on texts from the intersection. In fact, this example is just such a case. For $A = B =$ TSL, their intersection is (a subclass of) MTSL and therefore learnable by the algorithm which uses $2^{|\Sigma|}$ k-SL learners operating in parallel mentioned earlier. We have thus shown that TSL is weakly robust in the presence of itself.

4 Conclusions

We motivated and discussed the notion of learning from data systematically lacking in completeness. The result is four categories of learnability. The strongest, unaffectedness, provides a guarantee that a telltale set for the target language remains present despite interference. This requires that the correct generalizations be made even for data that can never appear. Strong robustness, while weaker than unaffectedness, makes a more reasonable guarantee: the learned language is only necessarily consistent with the target on data that can naturally occur in the face of the other constraints. Weak robustness keeps this more reasonable guarantee, but allows the learner to use grammars outside the target class. Finally, if none of these hold, the class is not robust.

We showed that each of the FIN, SL, LT, SP, and PT classes are strongly robust in the presence of any other class. On the other hand, we showed that the tier-based strictly local class of constraints, while quite natural for descriptive phonology, fails to be even strongly robust in the case where the relevant tier is unknown. Yet it has superclasses that are strongly robust in the presence of some types of interference. Such a quality makes this class weakly robust: one might fail to learn the target grammar, but learn instead a compatible grammar from the superclass. In the case of TSL, the relevant superclass was MTSL.

Each of these robustness categories is parameterized by the class from which interfering constraints are drawn. A class may be strongly robust in the presence

of one class, yet not robust at all when faced with another. Open questions include characterizing the strongly and weakly robust learning paradigms. It would also be interesting to consider the effects of interference from other constraints when learning from good examples [6, 15], as well as the problem of learning from data presentations which misrepresent the target language by including examples that do not belong to it.

Finally, the strongly robust learning paradigm provides a sufficient condition for when the pointwise intersection of two learnable classes of languages C and D is itself also learnable. The fact that each of the FIN, SL, LT, SP, and PT classes are strongly robust in the presence of any other class implies that classes of languages which must satisfy constraints from more than one of these classes are also learnable. To put it another way, some learnable classes of languages can be factored into simpler learnable classes. We hope this work helps lead to a more fully developed *compositional* theory of language learning.

Acknowledgements

We acknowledge support from the Data + Computing = Discovery summer REU program at the Institute for Advanced Computational Science at Stony Brook University, supported by the NSF under award 1950052.

References

[1] Angluin, D.: Inductive inference of formal languages from positive data. Information and Control 45(2), 117–135 (May 1980)
[2] Blum, L., Blum, M.: Toward a mathematical theory of inductive inference. Information and Control 28(2), 125–155 (June 1975)
[3] Case, J., Jain, S., Stephan, F., Wiehagen, R.: Robust learning–rich and poor. J. Comput. Syst. Sci. 69(2), 123–165 (2004)
[4] Clark, A., Lappin, S.: Linguistic Nativism and the Poverty of the Stimulus. Wiley-Blackwell (2011)
[5] Eyraud, R., Heinz, J., Yoshinaka, R.: Efficiency in the identification in the limit learning paradigm. In: Heinz, J., Sempere, J. (eds.) Topics in Grammatical Inference, chap. 2, pp. 25–46. Springer, Berlin, Heidelberg (2016)
[6] Freivalds, R., Kinber, E., Wiehagen, R.: On the power of inductive inference from good examples. Theoretical Computer Science 110(1), 131–144 (1993)
[7] Fulk, M., Jain, S.: Learning in the presence of inaccurate information. Theoretical Computer Science 161, 235–261 (1996)
[8] Gold, E.M.: Language identification in the limit. Information and Control 10(5), 447–474 (May 1967)
[9] Haines, L.H.: On free monoids partially ordered by embedding. Journal of Combinatorial Theory 6(1), 94–98 (1969)
[10] Heinz, J.: String extension learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 897–906. Association for Computational Linguistics, Uppsala, Sweden (July 2010)

[11] Heinz, J.: The computational nature of phonological generalizations. In: Hyman, L., Plank, F. (eds.) Phonological Typology, Phonetics and Phonology, vol. 23, chap. 5, pp. 126–195. Mouton de Gruyter (2018)
[12] Heinz, J., Kasprzik, A., Kötzing, T.: Learning in the limit with lattice-structured hypothesis spaces. Theoretical Computer Science 457, 111–127 (October 2012)
[13] Heinz, J., Rawal, C., Tanner, H.G.: Tier-based strictly local constraints for phonology. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers. vol. 2, pp. 58–64. Association for Computational Linguistics, Portland, Oregon (2011)
[14] Jain, S.: Program synthesis in the presence of infinite number of inaccuracies. Journal of Computer and System Sciences 53, 583–591 (1996)
[15] Jain, S., Lange, S., Nessel, J.: On the learnability of recursively enumerable languages from good examples. Theoretical Computer Science 261, 3–29 (2001)
[16] Jain, S., Osherson, D., Royer, J.S., Sharma, A.: Systems That Learn: An Introduction to Learning Theory. The MIT Press, 2nd edn. (1999)
[17] Lambert, D.: Grammar interpretations and learning TSL online. In: Proceedings of the Fifteenth International Conference on Grammatical Inference. Proceedings of Machine Learning Research, vol. 153, pp. 81–91 (August 2021)
[18] Lambert, D.: Relativized adjacency. Journal of Logic, Language and Information (May 2023)
[19] Lambert, D., Rawski, J., Heinz, J.: Typology emerges from simplicity in representations and learning. Journal of Language Modelling 9(1), 151–194 (August 2021)
[20] McNaughton, R., Papert, S.A.: Counter-Free Automata. MIT Press (1971)
[21] Osherson, D.N., Stob, M., Weinstein, S.: Systems That Learn. MIT Press, Cambridge, MA (1986)
[22] Pin, J.E.: Profinite methods in automata theory. 26th International Symposium on Theoretical Aspects of Computer Science STACS 2009 (Feb 2009)
[23] Rogers, J., Heinz, J., Bailey, G., Edlefsen, M., Visscher, M., Wellcome, D., Wibel, S.: On languages piecewise testable in the strict sense. In: Ebert, C., Jäger, G., Michaelis, J. (eds.) The Mathematics of Language: Revised Selected Papers from the 10th and 11th Biennial Conference on the Mathematics of Language, LNCS/LNAI, vol. 6149, pp. 255–265. FoLLI/Springer (2010)
[24] Simon, I.: Piecewise testable events. In: Brakhage, H. (ed.) Automata Theory and Formal Languages, Lecture Notes in Computer Science, vol. 33, pp. 214–222. Springer-Verlag, Berlin (1975)
[25] Smetsers, R., Volpato, M., Vaandrager, F., Verwer, S.: Bigger is not always better: on the quality of hypotheses in active automata learning. In: Clark, A., Kanazawa, M., Yoshinaka, R. (eds.) The 12th International Conference on Grammatical Inference. Proceedings of Machine Learning Research, vol. 34, pp. 167–181. PMLR, Kyoto, Japan (17–19 Sep 2014)
[26] Valiant, L.G.: A theory of the learnable. Communications of the ACM 27(11), 1134–1142 (November 1984)