

Learnability of Grammars with Arbitrary Finite Exceptions *

Jacob Rowen

Stony Brook University

JACOB.ROWEN@STONYBROOK.EDU

Abstract

Inspired by natural language phenomena, we provide a characterization of the classes of formal languages that can be learned even if their grammars include finitely many exceptions. Specifically, we consider a setup where the learner must faithfully acquire any arbitrary exceptional strings present in the input presented to them, in addition to any general patterns present. We use the identifiability in the limit framework to show that classes of languages with finitely many arbitrarily omitted strings cannot be identified except in trivial cases, while languages with finitely many arbitrarily included strings can be identified non-trivially. We then present strategies for identification of these arbitrarily augmented languages.

Keywords: identifiability in the limit, finite augmentations, finite omissions, formal languages

1. Introduction

As documented in [Jain et al. \(1999, pp. 168-180 and 189-194\)](#), a number of studies have been done on identification in the limit from noisy and incomplete sources, with research continuing after that book’s publication, like [Tantini et al. \(2006\)](#), or even more recently, [Wu and Heinz \(2023\)](#) and [Kaelbling et al. \(2023\)](#). [Jain et al. \(1999, pp. 138-139 and 145-150\)](#) also discusses work from [Osherson and Weinstein \(1982\)](#) and [Case and Lynes \(1982\)](#) on more generous learning paradigms that allow the learner to learn languages that are at worst some finite set of modifications away from the target language.

What has not been directly considered, though, are classes where the target languages themselves have exceptions to what would otherwise be generalizable in the usual grammatical sense, either through non-rule-following additions or non-rule-following omissions. This is a problem that comes up with some regularity in the study of natural languages and their acquisition, in various guises. For example, one might consider a language to generally obey a particular set of phonological rules, but not for some finite set of loan words. As [Dresher and Kaye \(1990\)](#) put it: “It is useful to think of the grammar of a language as having a core ... and a periphery, which departs in various ways from the principles of the core, and which consists of more or less idiosyncratic rules and exceptions.”

For a concrete case, consider the English calque “long time no see.” This particular phrase is deemed grammatical by many English speakers, yet, it does not function like most syntactically valid phrases of English; it permits no lexical substitutions. “Big time no see,” “long moment no see”, “long time not see”, and “long time no observe” are all

* With thanks to Timothy Carlin-Burns, my friends and colleagues in and from the Stony Brook Linguistics Department, and most especially Jeffrey Heinz, who originally posed the question this paper answers, and whose feedback and support has been invaluable. Support from NSF Grant #2416183 is also gratefully acknowledged.

quite clearly ungrammatical. If a scientist proposed a grammar of English that generated any phrases similar to, but not exactly the same as “long time no see”, that would be an error. The status of such a phrase should, in some meaningful sense, be viewed as a non-rule-following addition to a grammar of English. This is not a novel view; all research using the Tolerance Principle (Yang, 2016), for example, relies entirely on the idea that some language structures are memorized, and not rule-following, in this way.

In this article, we investigate the problem of identifying families of languages with exceptions in the limit from positive data (Gold, 1967). The primary results are presented in Theorems 4, 14 and 18. We provide and prove an exact characterization of the class of language families that have the potential for identification in the limit from positive data even when they have these finite internal inconsistencies. Theorem 4 covers the identifiability of families with arbitrary finite omissions, which turn out to be quite limited and uninteresting, and Theorem 14 covers the identifiability of families with arbitrary finite augmentations, which are much richer. Theorem 18 provides a strategy for identifying a non-trivial subclass of the latter variety of families. Along the way, we introduce a handful of novel properties and constructions useful for the analysis of language families with exceptions. Finally, in Section 6, we discuss where future research could improve upon the work presented here.

2. Preliminaries

This section draws heavily from the Preliminaries section of Motoki et al. (1991).

Let Σ be a nonempty finite set of symbols and Σ^* be the countably infinite set of finite-length strings constructable from the symbols in Σ . A *language*¹ is a subset of Σ^* , and a *family of languages* is a potentially infinite set of subsets of Σ^* . A family of languages $\mathcal{L}$ is an *indexed family of languages* if and only if there exists an enumeration $L_1, L_2, \dots = \mathcal{L}$ such that a subset of the natural numbers can index the family. We interchangeably refer to indexed families of languages by their element languages or by their indices, as convenient. An indexed family of languages $\mathcal{L}$ is *recursive* if and only for each $L \in \mathcal{L}$, there exists a computable *membership function* of the form $f : \Sigma^* \rightarrow \{0, 1\}$ such that $f(w) = 1$ if $w \in L$, and $f(w) = 0$ if $w \notin L$.

We use $\mathcal{FIN}$ to refer to the indexed family of recursive languages that contains all of and only the finite subsets of Σ^* . We use the symbol ‘ $-$ ’ for relative set complement and the symbol ‘ Δ ’ for symmetric difference, i.e. $L_a \Delta L_b = (L_a - L_b) \cup (L_b - L_a)$. If $L_a \Delta L_b \in \mathcal{FIN}$, we say that L_a and L_b are *finitely distinct*; otherwise, they are *infinitely distinct*. Given any set of sets A , we define $\text{GCS}(A)$ to be the greatest common subset of A , i.e. $\text{GCS}(A) = \bigcap_{E \in A} E$. We refer to the greatest common subset of a family of languages (which itself may or may not be a language in that family) as that family’s *core*.

Next, we introduce the identification in the limit from positive data paradigm. A *text* for a language L is a countably infinite sequence that contains all of and only the strings in L . Consider a procedure that sequentially requests inputs from a text and, upon receiving each new input, outputs a single index for a language in a particular indexed family of languages. Given an indexed family of languages $\mathcal{L}$, such a procedure *identifies* $\mathcal{L}$ in the

1. At times in this article, we will define languages as sets of natural numbers. These are compatible with $\Sigma = \{0, 1, \dots, 8, 9\}$, or they can be interpreted equally well as sets of indices for strings in an enumeration of any arbitrary Σ^* . None of the results will hinge on this distinction.

limit from positive data if and only if, given inputs from any text for any $L \in \mathcal{L}$, only a finite number of the procedure's outputs are not indices for languages equal to L . If an effective procedure of this kind exists for a family, that family is *identifiable in the limit from positive data*.

Given an indexed family of languages $\mathcal{L}$, a set T is said to be a *telltale set* for a language $L \in \mathcal{L}$ if and only if $T \in \mathcal{FIN}$, $T \subseteq L$, and for every $L' \in \mathcal{L}$ such that $L' \subsetneq L$, L' fails to contain T . Angluin (1980, Theorem 1) showed that telltale sets can be used to characterize exactly those families that are identifiable in the limit from positive data.

Theorem 1 *An indexed family of nonempty recursive languages $\mathcal{L}$ is identifiable in the limit from positive data if and only if for all $L \in \mathcal{L}$, there exists an effective procedure that, given an index for L , enumerates all strings in its telltale set.*

Due to the generality of the classes of families under consideration, we will often prove only the mere existence of such telltale sets, with or without a computable procedure to enumerate them, which is a strictly weaker desideratum than identifiability in the limit from positive data. In other words, for a given family $\mathcal{L}$, we may only show that for all $L \in \mathcal{L}$, there exists a set T that is a telltale set for L . This is a necessary condition for identifiability in the limit from positive data, but, strictly speaking, not a sufficient one. Due to this weakening, we needn't worry about whether any of the languages are empty, and we needn't be concerned with whether basic set operations are universally effectively computable.

3. Formalizing Finite Exceptions

First, we must provide a particular formal definition of “finite exceptions” that properly captures their arbitrariness. Since, *prima facie*, a learner has no way to know what number of kind of exceptions to expect, we investigate the problem where, as far as the learner knows, the target language could have any finite number of exceptional inclusions or exclusions, and any individual exceptional string need not obey any rules or generalizations. As such, we provide the following definitions:

Definition 2 (Finite Augmentations) *Given an indexed family of recursive languages $\mathcal{L}$, let $\mathcal{L}^+ = \{L \cup F \mid L \in \mathcal{L}, F \in \mathcal{FIN}\}$, the family of arbitrary finite augmentations to $\mathcal{L}$.*

Definition 3 (Finite Omissions) *Given an indexed family of recursive languages $\mathcal{L}$, let $\mathcal{L}^- = \{L - F \mid L \in \mathcal{L}, F \in \mathcal{FIN}\}$, the family of arbitrary finite omissions from $\mathcal{L}$.*

It is important to note that both $\mathcal{L}^+$ and $\mathcal{L}^-$ families are indexed families of recursive languages. They can each be indexed by dovetailing an enumeration of their respective base-family $\mathcal{L}$ with an enumeration of $\mathcal{FIN}$, and their language membership functions can be defined by putting together the membership functions of the languages in $\mathcal{L}$ with the trivial one for $\mathcal{FIN}$. Suppose $L \in \mathcal{L}$ has a membership function $f_L(w)$, and $F \in \mathcal{FIN}$ has a membership function $f_F(w)$. Let $L' \in \mathcal{L}^+$ such that $L' = L \cup F$ and $L'' \in \mathcal{L}^-$ such that $L'' = L - F$. A membership function for L' can be defined $f_L(x) \vee f_F(x)$, and for L'' can be $f_L(x) \wedge \neg f_F(x)$.

Families with arbitrary finite omissions are not of much interest with regard to identifiability. If an indexed family of recursive languages $\mathcal{L}$ is a subset of $\mathcal{FIN}$, then both $\mathcal{L}^+$ and $\mathcal{L}^-$ are as well, leaving them identifiable by the trivial algorithm of simply memorizing each input and guessing the resulting finite set. However, the presence of a single infinite language renders $\mathcal{L}^-$ unidentifiable.

Theorem 4 *If a family with arbitrary finite omissions $\mathcal{L}^-$ is not a subset of $\mathcal{FIN}$, it is not identifiable in the limit from positive data.*

Proof If $\mathcal{L}^- \not\subseteq \mathcal{FIN}$, then $\mathcal{L} \not\subseteq \mathcal{FIN}$. Take $L \in \mathcal{L}^-, L' \in \mathcal{L}, F \in \mathcal{FIN}$ such that $L = L' - F$ and $L \notin \mathcal{FIN}$. For any finite set S where $S \subsetneq L$, there must exist at least one $e \in L$ such that $e \notin S$. Let $F' = F \cup \{e\}$. Since F is a finite set, by Definition 3, there must exist a language $L'' \in \mathcal{L}^-$ such that $L'' = L' - F'$, meaning $S \subsetneq L'' \subsetneq L$. Thus, no finite subset of L is a suitable telltale set, and by Theorem 1, the family is not identifiable. ■

On the other hand, families with finite augmentations provide much more room for study. Let us consider a concrete example.

Example 1 (The Primes Family) Let $\mathcal{P} = \{P_2, P_3, P_5, P_7, P_{11}, \dots\}$ such that for any prime natural number n , $P_n = \{n^1, n^2, n^3, \dots\}$. So, in ascending order of index, the first three languages in $\mathcal{P}$ are:

$$\begin{aligned} P_2 &= \{2, 4, 8, 16, \dots\} \\ P_3 &= \{3, 9, 27, 81, \dots\} \\ P_5 &= \{5, 25, 125, 625, \dots\} \end{aligned}$$

It should be clear to see that not only are all of the languages in $\mathcal{P}$ infinitely distinct from each other, but in fact, none shares a single string with any other. This family is trivial to identify in the limit from positive data, as determining the prime factor of the very first input will fully determine the target language.

Now consider $\mathcal{P}^+$, the augmentation of $\mathcal{P}$. Specifically, imagine a learner aware of the construction of $\mathcal{P}^+$ trying to identify $\mathcal{P}^+$ in the limit from positive data. Suppose they are given 6 as their first input. If they are clever, they will immediately recognize that 6 must be from the finite augmentations, as no language in $\mathcal{P}$ contains 6; it has two different prime factors. Now suppose the next input is 9; the learner cannot, at this point, be certain that the target language is an augmentation of P_3 , as 9 could simply be another exception. Even if the next inputs are 81, then 27, then 3, then 2187, the learner still cannot be certain. The input after that may be 125, and the target language may be an augmentation of P_5 with only finitely many powers of 3. But all is not lost; in any given language in $\mathcal{P}^+$, there are infinitely many powers of one particular prime number, and only finitely many of any other. If, at any given time step, the learner simply guesses that the prime they have seen the most of is the infinite one, and that any other inputs seen make up the full set of exceptions, then eventually, they will be correct. The infinitely many non-exceptions will eventually permanently outnumber the finitely many exceptions. A procedure that makes guesses in exactly this way will, indeed, identify $\mathcal{P}^+$ in the limit from positive data.

The Primes Family was deliberately constructed to make it as easy as possible to see how one could identify a family with arbitrary finite augmentations in the limit from positive data, but within it lies a very useful kernel of wisdom. If a learner’s target language is a union of some infinite language with some finite set of augmentations, identifying the infinite language correctly can be sufficient to identify the full augmented language, as any “exceptions” can simply be “memorized” in the way a learner of a finite language would. However, this strategy does not generalize perfectly to all identifiable families, nor does it provide clear criteria for which families are identifiable with arbitrary finite augmentations. These criteria are the topic of the following section.

4. MF Distinctness and FD Nesting

In this section, we give an exact characterization of which families still admit telltale sets even with arbitrary finite augmentations. We show that two restrictions on languages, relative to the rest of the languages in the family, are needed to prevent finite augmentations from creating infinitely descending chains of subsets. We call these properties *Mutually-Finitudinal (MF) Distinctness* and *Finitely Distinct (FD) Nesting*. As it happens, MF Distinctness and FD Nesting are restrictive enough together that, with or without finite augmentations, all families that have both properties admit telltale sets.

4.1. Mutually-Finitudinal Distinctness

The first property we introduce is MF Distinctness. If there were two languages L_a and L_b in the same family $\mathcal{L}$, where $L_b \subsetneq L_a$, but $L_a - L_b \notin \mathcal{FIN}$, then there would be an infinite number of possible finite sets that one could augment L_b with while still keeping it a subset of L_a . These finite sets would entirely stymie the construction of a telltale set for L_a upon augmentation. A restriction must be made to prevent this. Yet, prohibiting these infinite subset relationships turns out to be insufficient; if L_b had a few elements that L_a did not have, then the problem would simply shift to the finite augmentation of L_a that does have these missing elements. Thus, we arrive precisely at MF Distinctness:

Definition 5 (MF Distinct) *An indexed family of languages $\mathcal{L}$ is Mutually-Finitudinally (MF) Distinct if and only if $\forall L, L' \in \mathcal{L}, L - L' \notin \mathcal{FIN} \implies L' - L \notin \mathcal{FIN}$.*

In an MF Distinct family, any two languages either each have infinitely many elements the other does not, or each have only finitely many elements the other does not. Upon arbitrary finite augmentation, infinite differences remain infinite, and finite differences can disappear entirely when the right material is added. Consider $\mathcal{FIN}$ itself, for example, which is trivially MF Distinct. Every finite language is only a finite union away from being a superset of any other. Where there are languages in subset and superset relations in an MF Distinct family, they must behave similarly to those in $\mathcal{FIN}$, as the subset language will, by definition, have an empty set of elements that the superset language does not have, and the empty set is most certainly finite.

With these intuitions built, we can now begin to prove some important properties of MF Distinct families, starting with Lemma 6.

Lemma 6 *If a suitable telltale set exists for each language in a family with arbitrary finite augmentations $\mathcal{L}^+$, $\mathcal{L}^+$ is MF Distinct.*

Proof Suppose for the sake of contradiction that a suitable telltale set exists for each language in $\mathcal{L}^+$, but it is not MF Distinct. By Definition 5, there must exist $L_a, L_b \in \mathcal{L}^+$ such that $L_a - L_b \notin \mathcal{FIN}$, but $L_b - L_a \in \mathcal{FIN}$. It is clear to see that this cannot be the case unless $L_a \notin \mathcal{FIN}$. By Definition 2, there must be a language $L_c \in \mathcal{L}^+$ such that $L_c = L_a \cup (L_b - L_a) = L_a \cup L_b$. For any finite set S where $S \subsetneq L_c$, there is some $L_d \in \mathcal{L}^+$ such that $L_d = L_b \cup S$, which must be a proper subset of L_c , since there is no union of finite elements to L_b that results in a superset of L_a . Thus, no suitable telltale set for L_c exists, which is a contradiction. ■

Our current goal is to provide a characterization of the families that admit telltale sets when finitely augmented, but one of the key properties of MF Distinctness itself is that it is entirely independent of augmentation.

Lemma 7 *A family with arbitrary finite augmentations $\mathcal{L}^+$ is MF Distinct if and only if $\mathcal{L}$ is MF Distinct.*

Proof If $\mathcal{L}^+$ is MF Distinct, then $\mathcal{L}$ must be as well, since $\mathcal{L} \subseteq \mathcal{L}^+$. The implication must also be true the other way, since the finitude of the differences between any pair of languages is not changed by augmenting either or both with finitely many elements. ■

As alluded to above, MF Distinctness alone is not enough for a family with arbitrary finite augmentations to admit telltale sets. Consider, for example, the following very simple Co-Finite family from Jain et al. (1999, pp. 44):

Example 2 (The Missing String Family) *Let $\mathcal{L} = \{\Sigma^* - \{w\} \mid w \in \Sigma^*\}$. This family is quite trivially identifiable in the limit from positive data, as no language is a subset of any other, and quite trivially MF Distinct, as the symmetric difference between any two languages is exactly two elements. A procedure that simply iterates over an enumeration of $\mathcal{L}$, guessing each language until counter-evidence is provided, will eventually converge and identify $\mathcal{L}$ in the limit. However, augmentation adds Σ^* into the family, making $\mathcal{L}^+$ entirely unidentifiable, as Σ^* will admit no suitable telltale set.*

To rule out families like these, we will need a property like FD Nesting.

4.2. Finitely Distinct Nesting

Before properly introducing FD Nesting we consider the following lemma:

Lemma 8 *If a suitable telltale set exists for each language in a family with arbitrary finite augmentations $\mathcal{L}^+$, then for each $L \in \mathcal{L}^+$, $\{L' \in \mathcal{L}^+ \mid L' \subsetneq L\}$ is finite.*

Proof Suppose for the sake of contradiction that a suitable telltale set exists for each language in $\mathcal{L}^+$, but there exists a language $L_a \in \mathcal{L}^+$ such that $\{L' \in \mathcal{L}^+ \mid L' \subsetneq L_a\}$ is infinite. Let S be an arbitrary finite set where $S \subseteq L_a$. Because S is finite, $\{L' \in \mathcal{L}^+ \mid L_a - S \subseteq L' \subseteq L_a\}$ is finite (in fact its cardinality is no greater than $2^{|S|}$, as it is simply the set of languages that can be created by removing subsets of S from L_a). Thus, there must exist some $L_b \in \mathcal{L}^+$

such that $L_b \cup S \subsetneq L_a$. By Definition 2, there must exist $L_c \in \mathcal{L}^+$ such that $L_c = L_b \cup S$, and thus $S \subseteq L_c \subsetneq L_a$. Since such a language exists for any finite subset of L_a , no finite subset of L_a is a suitable telltale set, which is a contradiction. ■

Lemma 8 diagnoses the exact problem that the Missing String Family from Example 2 had. In fact, all MF Distinct families that do not admit telltale sets run into the very same problem. However, much like how MF Distinctness restricted more than just subsets and supersets to preserve its key properties over augmentation, the condition in Lemma 8 is not quite sufficient, as we are interested in describing the properties that the families must have before augmentation. Thus, we introduce Definition 9:

Definition 9 (FD Nested) *An indexed family of languages $\mathcal{L}$ is Finite-Distinctly (FD) Nested if and only if $\forall L \in \mathcal{L}, \forall F \in \mathcal{FIN}$, the set $\{L' \in \mathcal{L} \mid L' \subsetneq L \cup F\}$ is finite.*

In an FD Nested family, no language can even be a finite union away from being a superset of infinitely many others. Thus, upon arbitrary finite augmentation, no language will be a superset of infinitely many others. Once again, we may consider $\mathcal{FIN}$, which is trivially FD Nested. Even though every language in $\mathcal{FIN}$ is a finite union away from being a superset of any of the infinite others, a particular language with a particular finite union is a superset of only finitely many other languages. Once again, $\mathcal{FIN}$ provides a model of what families must look like, as all FD Nested families work similarly.

Now we can begin to prove properties of FD Nested families, starting with Lemma 10 and Lemma 11:

Lemma 10 *If a family with arbitrary finite augmentations $\mathcal{L}^+$ is such that for all $L \in \mathcal{L}^+$, $\{L' \in \mathcal{L}^+ \mid L' \subsetneq L\}$ is finite, then $\mathcal{L}^+$ is FD Nested.*

Proof Suppose for the sake of contradiction that for all $L \in \mathcal{L}^+$, $\{L' \in \mathcal{L}^+ \mid L' \subsetneq L\}$ is finite, but $\mathcal{L}^+$ is FD not Nested. By Definition 9, there must be some $L_a \in \mathcal{L}^+$ such that there exists $F \in \mathcal{FIN}$ where $\{L' \in \mathcal{L}^+ \mid L' \subsetneq L_a \cup F\}$ is not finite. But by Definition 2, $L_a \cup F \in \mathcal{L}^+$, so its corresponding set must be finite, which is a contradiction. ■

Lemma 11 *If a family with arbitrary finite augmentations $\mathcal{L}^+$ is FD Nested, $\mathcal{L}$ is FD Nested too.*

Proof $\mathcal{L}$ is a subset of $\mathcal{L}^+$. ■

Note that unlike what Lemma 7 showed for MF Distinct families, Lemma 11 only shows a one-way implication. This poses no serious problem, since, as we have already proven the biconditional for MF Distinctness, we can prove the other direction of the implication for FD Nesting alongside MF Distinctness:

Lemma 12 *If a family $\mathcal{L}$ is both MF Distinct and FD Nested, $\mathcal{L}^+$ is too.*

Proof By Lemma 7, if $\mathcal{L}$ is MF Distinct, then $\mathcal{L}^+$ is too. As for FD Nesting, suppose for the sake of contradiction that $\mathcal{L}$ is FD Nested, but $\mathcal{L}^+$ is not. By Definition 9, there must be some $L_a \in \mathcal{L}^+$ such that there exists $F \in \mathcal{FIN}$ where $\{L' \in \mathcal{L}^+ \mid L' \subsetneq L_a \cup F\}$ is not

finite. Due to the infinitude of that set, there must be some $L_b \in \mathcal{L}$ such that $L_b \subsetneq L_a \cup F$. Because $L_b \subsetneq L_a \cup F$, $L_b - L_a$ is finite. By Definition 5, $L_a - L_b$ must also be finite. Thus, $\{F' \in \mathcal{FIN} \mid L_b \cup F' \subsetneq L_a\}$ is finite, so $\{L_n \in \mathcal{L} \mid L_n \subsetneq L_a \cup F\}$ must be infinite. But note that there must be some $L_c \in \mathcal{L}, F' \in \mathcal{FIN}$ such that $L_c \cup F' = L_a$, and thus some $F'' \in \mathcal{FIN}$ such that $L_c \cup F'' = L_a \cup F$, meaning $\{L_n \in \mathcal{L} \mid L_n \subsetneq L_c \cup F''\}$ is infinite, precisely the inverse of $\mathcal{L}$ being FD Nested, and thus a contradiction. ■

One of the other useful properties that arises from a family being both MF Distinct and FD Nested is that every language in the family has a finite subset that distinguishes it from all of its subset languages. This subset makes for a suitable telltale set:

Lemma 13 *If an indexed family of recursive languages $\mathcal{L}$ is MF Distinct and FD Nested, then a suitable telltale set exists for each language in the family.*

Proof Let $L_a \in \mathcal{L}$, and construct $U = \bigcup_{\{L' \in \mathcal{L} \mid L' \subsetneq L_a\}} [L_a - L']$. Suppose for the sake of contradiction that U is not a suitable telltale set for L_a . By construction, $U \subseteq L_a$, and for every $L' \in \mathcal{L}$ such that $L' \subsetneq L_a$, L' fails to contain U , so it must be the case that $U \notin \mathcal{FIN}$. Note that $U \subseteq L_a$, so it must be that $L_a \notin \mathcal{FIN}$ as well. Because $\mathcal{L}$ is MF Distinct, by Definition 5, for any one $L' \in \mathcal{L}$ such that $L' \subsetneq L_a$, it must be the case that $L_a - L' \in \mathcal{FIN}$, so there must be an infinite number of such L' . But this would mean that $\mathcal{L}$ is not FD Nested, which is a contradiction. Thus, U is a suitable telltale set. ■

With that, we can now simply assemble all of these lemmas into our desired Theorem 14:

Theorem 14 *A suitable telltale set exists for each language in a family with arbitrary finite augmentations $\mathcal{L}^+$ if and only if $\mathcal{L}$ is MF Distinct and FD Nested.*

Proof If $\mathcal{L}$ is MF Distinct and FD Nested, then by Lemma 12, $\mathcal{L}^+$ is MF Distinct and FD Nested, so by Lemma 13, the suitable telltale sets exist.

If $\mathcal{L}$ is not MF Distinct, then by Lemma 7, $\mathcal{L}^+$ is not MF Distinct, so by Lemma 6, the telltale sets do not exist. If $\mathcal{L}$ is not FD Nested, then by Lemma 11, $\mathcal{L}^+$ is not FD Nested, so by Lemma 10, there is a language in $\mathcal{L}^+$ with infinitely many proper subset languages in $\mathcal{L}^+$. This means that by Lemma 8, the suitable telltale sets do not exist. ■

5. Learning Strategies

This section provides strategies for the construction of learning algorithms for particular families of languages. The previous section provides a way to determine if a family of languages with arbitrary finite augmentations is potentially identifiable, but not through any constructive means sufficient to define an identifying procedure. As restricted as it is in some ways, the class of MF Distinct and FD Nested families is broad enough to make a universal identification strategy difficult to formulate. For that reason, we define the Exceptional Core property to describe one usefully generalized learning approach. Then, we show an example of a family without an Exceptional Core and how to identify it in the limit from positive data with arbitrary finite augmentations regardless.

5.1. Decomposition

To identify MF Distinct and FD Nested families, it is very useful to break them down with a method we call *Finitely Distinct (FD) Decomposition*.

Definition 15 (FD Decomposition) *The Finitely-Distinct (FD) Decomposition of an indexed family of languages $\mathcal{L}$ is $FDD(\mathcal{L}) = \{\mathcal{L}_1, \mathcal{L}_2, \dots\}$ such that $\mathcal{L} = \bigcup_{i \in \mathbb{N}} \mathcal{L}_i$ and for any pair of $\mathcal{L}_i, \mathcal{L}_j$ such that $\mathcal{L}_i \neq \mathcal{L}_j$, for any $L_a, L_b \in \mathcal{L}_i$ and $L_c \in \mathcal{L}_j$, $L_a \triangle L_b \in \mathcal{FIN}$, and $L_a \triangle L_c \notin \mathcal{FIN}$.*

Lemma 16 *Given an indexed family of recursive languages $\mathcal{L} = \{L_1, L_2, \dots\}$, there exists an FD Decomposition for it.*

Proof Construct $\mathcal{L}_1 = \{L_i \in \mathcal{L} \mid L_i \triangle L_1 \in \mathcal{FIN}\}$. Note that if $A \triangle B \in \mathcal{FIN}$ and $B \triangle C \in \mathcal{FIN}$, then $A \triangle C \subseteq (A - B) \cup (C - B) \in \mathcal{FIN}$. Construct an indexed family of recursive languages $\mathcal{L}' = \{L'_1, L'_2, \dots\} = \mathcal{L} - \mathcal{L}_1$. Construct $\mathcal{L}_2 = \{L_i \in \mathcal{L}' \mid L_i \triangle L'_1 \in \mathcal{FIN}\}$. Construct an indexed family of recursive languages $\mathcal{L}'' = \{L''_1, L''_2, \dots\} = \mathcal{L}' - \mathcal{L}_2$. Repeat this process indefinitely. The resulting indexed set $\{\mathcal{L}_1, \mathcal{L}_2, \dots\}$ will be a sufficient FD Decomposition for $\mathcal{L}$. $\blacksquare$

Crucially, the FD Decomposition always groups together a language with its finite augmentations. Indeed, it groups together all languages that are finitely distinct from each other, hence the name. If a strategy is known to identify every subfamily in the FD Decomposition of a family, and a strategy is known to identify which subfamily a text belongs to, then these strategies compose into a strategy to identify the whole family. This is relatively easy to do with families composed entirely of subfamilies with Exceptional Cores.

5.2. Exceptional Cores

Definition 17 (Exceptional Core) *An indexed family of languages has an Exceptional Core if and only if for all $L \in \mathcal{L}$, $GCS(\mathcal{L}) \triangle L \in \mathcal{FIN}$.*

Intuitively, we call a family's core "Exceptional" when every language in the family can be effectively described as a finite set of "exceptions" to that core (making them "exceptionally" well-behaved). Note that $\mathcal{FIN}$ has a trivial Exceptional Core, the empty set, and for much the same reason that $\mathcal{FIN}$ is MF Distinct and FD Nested, a family with an Exceptional Core must be as well. If a family of languages has an Exceptional Core, then finding a telltale set for each language is very straightforward; the relative difference of the language and the core is sufficient. This strategy can be generalized using the FD Decomposition.

Theorem 18 *If an indexed family of recursive languages $\mathcal{L}$ is MF Distinct, and each of the subfamilies in $FDD(\mathcal{L})$ has an Exceptional Core, then given a language $L \in \mathcal{L}$ that belongs to a subfamily with Exceptional Core C , $L - C$ is a sufficient telltale set for L .*

Proof Recall that given some language $L \in \mathcal{L}$ and some set T , there are three things we must show to prove that T is a suitable telltale set for L : $T \in \mathcal{FIN}$, $T \subseteq L$, and for every $L' \in \mathcal{L}$ such that $L' \subsetneq L$, L' fails to contain T . Let $T = L - C$. By Definition 17, $T \in \mathcal{FIN}$, and it is trivial to see that $T \subseteq L$. Because $\mathcal{L}$ is MF Distinct, by Definition 5 and

Definition 15, any two languages where one is a subset of the other must belong to the same subfamily in $FDD(\mathcal{L})$. Thus, if there is some $L' \in \mathcal{L}$ such that $L' \subsetneq L$, then there is some $F' \in \{F \in \mathcal{FIN} \mid C \cup F \subsetneq L\}$ such that $L' = C \cup F'$. If $L - C$ failed to be a telltale set, we would obtain $L - C = T \subseteq L' = C \cup F' \subsetneq C \cup T = L$, which implies $T \subseteq C \cup F' \subsetneq C \cup T$, but $T \cap C = \emptyset$, so this would mean $T \subseteq F' \subsetneq T$, a naked contradiction. Thus, $L - C$ is a sufficient telltale set for L . $\blacksquare$

In the cases where the telltale sets can be effectively computed, the algorithm in Angluin (1980) identifies the family. This can be demonstrated quite nicely for the $\mathcal{P}^+$ family from Example 1:

Example 3 (The Primes Family, Redux) First, it will be useful to consider $FDD(\mathcal{P})$, which in fact is little different from considering $\mathcal{P}$ itself. Every pair of languages in $\mathcal{P}$ is infinitely distinct, so every language has a subfamily to itself. In $FDD(\mathcal{P}^+)$, the subfamily for some $P \in \mathcal{P}$ grows to contain all augmentations to P . The empty set is a sufficient telltale set for any language in $\mathcal{P}$, as none of these languages have any subsets in the family. Thus, quite neatly, the telltale set for some $P' \in \mathcal{P}^+$ where $P' = P \cup F$ for some $P \in \mathcal{P}, F \in \mathcal{FIN}$ is $P' - P = (P \cup F) - P = F - P$. In other words, for any language in $\mathcal{P}^+$, its set of augmentations alone is a sufficient telltale set, enumerable simply with an enumeration of $\mathcal{FIN}$.

It is not the case that all families that exclusively decompose into subfamilies with Exceptional Cores can have their telltale sets enumerated as simply as $\mathcal{P}^+$, but it is the case that if a procedure is found for enumerating the cores, an enumeration of $\mathcal{FIN}$ will do the rest. If the former enumeration is computable from indices for the cores, then this may be a path to an identification algorithm. Yet, as stated above, not all MF Distinct and FD Nested families have the Exceptional Cores that make this possible. Entirely in the arena of human intuition, we found it difficult to produce such counterexamples, so they may be meaningfully uncommon in some way. Regardless, we now consider the families without Exceptional Cores.

5.3. Non-Exceptional Cores

The investigation presented here would not be complete without a demonstration of an MF Distinct and FD Nested family that does not decompose into subfamilies with Exceptional Cores. We consider in detail the following example of a single finitely distinct family of infinite languages with an empty core:

Example 4 (The Un-Counting Family) Let $\Sigma = \{a, b\}, \mathcal{L} = \{L_1, L_2, \dots\}$ such that for each $n \in \mathbb{N}, L_n = \{a^n, b^n, b^{n+1}, b^{n+2}, \dots\}$, where x^k for some string x and some natural number k represents the string x concatenated with itself k times. So, in ascending order of index, the first three languages in $\mathcal{L}$ are:

$$\begin{aligned} L_1 &= \{a, b, bb, bbb, \dots\} \\ L_2 &= \{aa, bb, bbb, bbbb, \dots\} \\ L_3 &= \{aaa, bbb, bbbb, bbbbbb, \dots\} \end{aligned}$$

For any two $i, j \in \mathbb{N}$ where $1 \leq i < j$, we can see that $L_i - L_j = \{a^i, b^i, b^{i+1}, \dots, b^{j-1}\}$ and $L_j - L_i = \{a^j\}$. Thus, every pair of languages is finitely distinct, and the family is MF Distinct. Every language has a unique element, so no language is a subset of another. Moreover, a finite union to any language will contain only a finite number of strings of a 's, so the resulting language will only have a finite number of subset languages in $\mathcal{L}$. Thus, $\mathcal{L}$ is FD Nested.

Yet, for any given $w = b^k$ where $k \in \mathbb{N}$, the cardinality of the set $\{L \in \mathcal{L} \mid w \in L\}$ is exactly k . The core of $\mathcal{L}$ is empty. Since none of the languages are subsets of each other, telltale sets for languages in $\mathcal{L}$ are trivial; a valid procedure to identify $\mathcal{L}$ is to simply wait for a string of a 's to appear, the length of which will exactly identify the target language. But what of $\mathcal{L}^+$?

In many cases, telltale sets can be constructed the same way as the families with Exceptional Cores. For example, any string consisting of both a 's and b 's will be part of its language's telltale set, as these are obviously the result of augmentation. But consider the language $L_i \cup \{a^j\}$, where $i, j \in \mathbb{N}$ and $1 \leq i < j$. This language, built from a union of just one element with a language in $\mathcal{L}$, has a potentially quite large number of subset languages in $\mathcal{L}^+$: $L_i, L_j, L_j \cup \{a^i\}, L_j \cup \{a^i, b^i\}, L_j \cup \{a^i, b^i, b^{i+1}\}, \dots, L_j \cup \{a^i, b^i, b^{i+1}, \dots, b^{j-2}, b^{j-1}\}$. The minimal telltale set for this language is, evidently, $\{a^i, a^j, b^i, b^{i+1}, \dots, b^{j-2}, b^{j-1}\}$. To provide a full generalization, a sufficient telltale set of a language $L \in \mathcal{L}^+$ formed from the union of $L' \in \mathcal{L}$ and $F \in \mathcal{FIN}$ can be computed by taking the union of F with the union of symmetric differences of each language in $\mathcal{L}$ whose a -string is in L .

We conjecture that the Un-Counting Family may be in some way structurally representative of the finitely distinct families that fail to have Exceptional Cores. If so, there may be some useful way to generalize a strategy for constructing their telltale sets, and perhaps unify that strategy with the one for Exceptional Cores, but this merits further investigation than we present here.

6. Discussion

This concludes our investigation of families of languages with arbitrary finite exceptions. We showed in Theorem 4 that all of and only subsets of $\mathcal{FIN}$ are identifiable in the limit from positive data with arbitrary omissions. We showed in Theorem 14 that all of and only MF Distinct and FD Nested indexed families of recursive languages admit suitable telltale sets for identifiability. We showed in Theorem 18 how to use FD Decomposition to construct telltale sets for the subclass of MF Distinct and FD Nested families with only Exceptional Cores, and we showed in Example 4 what families outside this subclass may look like.

Now that we have surveyed the landscape of identifiable families of languages with arbitrary finite exceptions, we present some key takeaways to consider and future directions to explore.

6.1. Omission vs. Augmentation

As stated about somewhat different problems in Systems That Learn, we have once again shown that “missing data is strictly more disruptive than noise” (Jain et al., 1999, p. 175).

With the exception of subsets of $\mathcal{FIN}$, families with arbitrary finite omissions are uniquely a lost cause. Many exception-like phenomena, e.g. in natural language, afford analyses either in the form of a small list of exclusions, or a large, but finite, list of inclusions. To allow for identifiability in the limit from positive data, in the general case, it is the list of inclusions, not exclusions, that must be chosen. Furthermore, a family with arbitrary finite omissions and augmentations is no better than one with just omissions; arbitrary omissions immediately destroy FD Nesting.

6.2. Applications to Well-Known Families

For a family with arbitrary finite augmentations to be identifiable in the limit from positive data, either all of its languages must be finite, or all must be infinite. This can be thought of as analogous to Gold’s original insight about the unidentifiability of Superfinite classes (Gold, 1967). For a representative example, this generalization rules out the Strictly k -Local families (McNaughton and Papert, 1971) for any $k > 0$, as these families all include both finite and infinite languages. Of course, in some sense, Strictly k -Local families have a grammar-internal way to handle “exceptional” strings of length less than k , but such exceptions are not truly arbitrary in the sense defined here. Many well-studied families would have to be considerably more constrained to be both MF Distinct and FD Nested.

6.3. Practical and Linguistic Applications

One very natural way to build an MF Distinct and FD Nested family with Exceptional Cores is to define it through parameterization. Parameterization can make it relatively simple to ensure subclasses are infinitely different from each other, to satisfy MF Distinctness, and to avoid languages that are subsets of each other, to satisfy FD Nesting and ensure Exceptional Cores. For a linguistic example, consider systems of phonological stress, which can be parameterized neatly by properties like metrical headedness. There may be a path to substantial new insights by reconsidering so-called Principles and Parameters work like that of the aforementioned Dresher and Kaye (1990), now with a new perspective on parameters that comes with a new learning strategy. It would also be of great use to see how the results covered here intersect with traditional linguistic analyses that appeal to exceptionality, like irregular morphology.

6.4. Further Theoretical Work

There is still much theoretical grammatical inference work to be done in this area. As stated in Section 5.3, further research could possibly find a strategy capable of handling non-Exceptional Cores, rather than the partial result we provide. Furthermore, perhaps there is some even cleverer way to consider the problem of arbitrary finite augmentations so as to capture computable identifiability, not just the weaker property we pursued. However, we have only considered one framework, identification in the limit from positive data, and only looked at languages, as opposed to, for example, string-to-string functions. It is entirely unknown to us whether the landscape looks different for different kinds of inference problems when exceptionality is defined as it is here. Nearly-universal patterns are an entirely natural phenomenon to attempt to infer, and if differing approaches result in differing learning criteria, that could be a significant factor in choosing the right approach for a particular

domain. On the other hand, if differing approaches yield similar outcomes, then we have come upon something fundamental about inferring the nature of somewhat inconsistent data. In either case, we are excited to see what further insights can reveal about the learning of patterns alongside their exceptions.

References

- Dana Angluin. Inductive inference of formal languages from positive data. *Information and control*, 45(2):117–135, 1980.
- John Case and Christopher Lynes. Machine inductive inference and language identification. In *International Colloquium on Automata, Languages, and Programming*, pages 107–115. Springer, 1982.
- B Elan Dresher and Jonathan D Kaye. A computational learning model for metrical phonology. *Cognition*, 34(2):137–195, 1990.
- E Mark Gold. Language identification in the limit. *Information and control*, 10(5):447–474, 1967.
- Sanjay Jain, Daniel Osherson, James Royer, and Arun Sharma. *Systems that learn: an introduction to learning theory, second edition*. MIT press, 1999.
- Philip Kaelbling, Dakotah Lambert, and Jeffrey Heinz. Robust identification in the limit from incomplete positive data. In *International Symposium on Fundamentals of Computation Theory*, pages 276–290. Springer, 2023.
- Robert McNaughton and Seymour A Papert. *Counter-Free Automata (MIT research monograph no. 65)*. The MIT Press, 1971.
- Tatsuya Motoki, Takeshi Shinohara, and Keith Wright. The correct definition of finite elasticity: corrigendum to identification of unions. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, page 375, 1991.
- Daniel N Osherson and Scott Weinstein. Criteria of language learning. *Information and Control*, 52(2):123–138, 1982.
- Frédéric Tantini, Colin de La Higuera, and Jean-Christophe Janodet. Identification in the limit of systematic-noisy languages. In *International Colloquium on Grammatical Inference*, pages 19–31. Springer, 2006.
- Katherine Wu and Jeffrey Heinz. String extension learning despite noisy intrusions. In *International Conference on Grammatical Inference*, pages 80–95. PMLR, 2023.
- Charles Yang. *The price of linguistic productivity: How children learn to break the rules of language*. MIT press, 2016.