# Finite identification with positive and with complete data

Dick de Jongh[1] and Ana Lucia Vargas Sandoval[1]

Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands
d.h.j.dejongh@uva.nl
ana.varsa@gmail.com

**Abstract**

We study the differences between finite identifiability of recursive languages with positive and with complete data. We show that in finite families the difference lies exactly in the fact that for positive identification the families need to be antichains, while in in the infinite case it is less simple. We also show that with complete data there are no maximal learnable families, whereas with positive data there usually are, and we conjecture they always exist.

## 1 Introduction

The groundbreaking work of Gold (3) from 1967 started a new era for developing mathematical and computational frameworks for studying the formal process of *learning*. Gold's model, *identification in the limit*, has been studied for learning recursive functions, recursively enumerable languages, and recursive languages with *positive data* and with *complete data*. The learning task consists of identifying languages as members of a family of languages, the learning function can output infinitely many conjectures but they need to stabilize in one permanent conjecture. In Gold's model, a huge difference in power between learning with positive data and with complete data (i.e. positive plus negative data) is exposed. With positive data no family of languages containing all finite languages and at least one infinite one can be learnable. With complete data the learning task becomes almost trivial.
Based on Gold's model and results, Angluin's work (1) focuses on indexed families of recursive languages, i.e., families of languages with a uniform decision procedure for membership. Such families are of interest because of their naturalness for languages generated by types of grammars. In particular, Angluin (1) gave a characterization when Gold's learning task can be executed. Her work shows that many non-trivial families of recursive languages can be learned by means of positive data only.

A few years later, Mukouchi (6) (and simultaneously Lange and Zeugmann (5)) introduced the framework of *finite identification* "in Angluin's style" for both positive and complete data. The learning task is as in Gold's model with the difference that the learning function can only guess once. Mukouchi presents an Angluin style characterization theorem for positive and complete finite identification. As expected, finite identification with complete data is more powerful than with positive data only. However, the distinction is much less marked than in Gold's framework. His work didn't draw much attention until recently Gierasimczuk and de Jongh (2) further developed the theory of finite identification.

It is often believed that children do not use negative data when they learn their native language. In opposition to that, a large amount of theoretical and experimental work in computational linguistics has been conducted to analyze and test the intuition in the powerful contribution of "negative" data for improving and speeding up children's language acquisition (see Hiller and Fernandez (4)).

In this work, we focus on a more fine grained theoretical analysis of the distinction between finite identification with positive and with complete data in Angluin-style. Our aim is to formally study the concrete difference: what can we do more with complete information for families of recursive languages than with only positive information.
We start with finite identification of finite families, in which the distinction between positive and complete data comes out very clearly: the difference is exactly described by the fact that with positive data families can only be identified if they are antichains w.r.t. $\subseteq$. Then, we question whether any finitely identifiable family is contained in a maximal finitely identifiable one. First we address this in the positive data setting. Maximal learnable families are of special interest because a learner for a maximal learnable family is a learner for all of its subfamilies. We provide a mildly positive result for families concerning any number of finite languages and give some hints about the obstacles to a more general result. Then, surprisingly, we provide a negative result concerning maximal learnable families for finite identification

with complete data: any finitely identifiable family can be extended to a larger one which is also finitely identifiable and is therefore not maximal.

We then return to families which are antichains. We show that infinite antichains of infinite languages exist which can be identified with complete information but not with positive information only. For infinite antichains of finite languages we show that such an example cannot exist if the indexing of the languages is by canonical indexes. The case of arbitrary indexing is investigated but not fully solved.

# 2 Preliminaries

We use standard notions from recursion theory and learning theory (see e.g., Osherson and Weinstien (7)), and for "Angluin style" identification in the limit (see (1), (6)).

Since we can represent strings of symbols by natural numbers, we will always refer to $\mathbb{N}$ as our universal set. Thus *languages* are sets of natural numbers, i.e. $L \subseteq \mathbb{N}$. A *family* $\mathcal{L} = \{L_i | i \in \mathbb{N}\}$ will be *an indexed family of recursive languages*, i.e. the two-place predicate $y \in L_i$ is recursive. In case all languages are finite and there is a recursive function $F$ such that for each $i$, $F(i)$ is a canonical index for $L_i$, then we call $\mathcal{L}$ a *canonical family*. In finite identification a *learner* will be a total recursive function that takes its values in $\mathbb{N} \cup \{\uparrow\}$ where $\uparrow$ stands for *undefined*.

A *positive data presentation* of a language $L$ is an infinite sequence $\sigma^+ := x_1, x_2, \ldots$ of elements of $\mathbb{N}$ such that $\{x_1, x_2, \ldots\} = L$. A *complete data presentation* of a language $L$ is an infinite sequence of pairs $\sigma := (x_1, t_1), (x_2, t_2), \ldots$ of $\mathbb{N} \times \{0, 1\}$ such that $\{x_n | t_n = 1, n \geq 0\} = L$ and $\{x_m | t_m = 0, m \geq 0\} = \mathbb{N} \setminus L$. An initial segment of length $n$ of $\sigma$ is indicated by $\sigma[n]$. A family $\mathcal{L}$ of languages is said to be *finitely identifiable from positive data (p.f.i.)*, or *finitely identifiable from complete data (c.f.i.)*, if there exists a recursive learner $\varphi$ which satisfies the following: for any language $L_i$ of $\mathcal{L}$ and for any positive data sequence $\sigma^+$ (or complete data sequence $\sigma$) of $L_i$ as input to $\varphi$, $\varphi$ produces on exactly one initial segment $\sigma^+[n]$ a conjecture $\varphi(\sigma^+[n]) = j$ such that $L_j = L_i$, and stops.

Let $\mathcal{L}$ be a family of languages, and let $L$ be a language in $\mathcal{L}$. A finite set $D_L$ is a *definite tell-tale set* (DFTT) for $L$ if $D_L \subseteq L$ and $\forall L' \in \mathcal{L}, (D_L \subseteq L' \rightarrow L' = L)$.

A language $L'$ is said to be *consistent* with a pair of finite sets $(B, C)$ if $B \subseteq L'$ and $C \subseteq \mathbb{N} \setminus L'$. A pair of finite sets $D_L$, $\overline{D}_L$ is a *definite,co-definite pair of tell-tale sets* (DFTT, co-DFTT) for $L$ if $L$ is consistent with $(D_L, \overline{D}_L)$, and $\forall L' \in \mathcal{L}$, if $L'$ is consistent with $(D_L, \overline{D}_L)$ then, $L' = L$.

**Theorem 1.** (Mukouchi's Characterization Theorem)(6)(5)

*A family $\mathcal{L}$ of languages is finitely identifiable from positive data (p.f.i.) iff for every $L \in \mathcal{L}$ there is a uniformly computable DFTT set $D_L$, that is, there exists an effective procedure that on input i, index of L, produces the canonical index of some definite finite tell-tale of L and then halts.*

*A family $\mathcal{L}$ of languages is finitely identifiable from complete data (c.f.i.) iff there is, by an effective procedure, for every $L \in \mathcal{L}$ a uniformly computable pair of DFTT,co-DFTT sets $(D_L, \overline{D}_L)$.*

**Corollary 1.** (6) *If a family $\mathcal{L}$ has two languages such that $L_i \subset L_j$, then $\mathcal{L}$ is not p.f.i..*

# 3 Finite families of languages

This section is dedicated to finite families of languages. A pair of simple but striking results already provides a good insight on a feature underlying the difference between finite identification on positive and on complete data.

**Theorem 2.** *A finite family of languages $\mathcal{L}$ is finitely identifiable from positive data iff no language $L \in \mathcal{L}$ is a proper subset of another $L' \in \mathcal{L}$.*

**Theorem 3.** *Any finite collection of languages $\mathcal{L} = \{L_1, \ldots, L_n\}$ is finitely identifiable with complete data.*

# 4 Looking for maximal learnable families

## 4.1 Finding maximal p.f.i. families

In this section we study maximal p.f.i. families. We address the follow up question: Is each p.f.i. family contained in a maximal p.f.i. family?

**Theorem 4.** *Every recursive family of finite languages which is p.f.i. is contained in a maximal family of languages which is (non-effectively) p.f.i..*

Proving theorem 4 is by a classical Zorn lemma construction. If infinite languages are present in the family, such a Zorn lemma construction cannot be applied since not every family of incomparable languages is non-effectively p.f.i..

Conjecture 1: Every p.f.i. family can be effectively extended into a maximal effective p.f.i. family.

How many maximal extensions can a p.f.i. family have? Consider the following example: Let $\mathcal{L}^s$ be the family of all singletons. Clearly it is maximal with respect to p.f.i.. However if we take out one of the singletons, say $\{0\}$, we obtain a p.f.i. subfamily $\mathcal{L}^s_0$ which is no longer maximal and its only p.f.i. extension is $\mathcal{L}^s$. If we remove $\{1\}$ from $\mathcal{L}^s_0$, we can maximally extend this family in two different ways, either adding $\{0,1\}$ or adding $\{0\}$ and $\{1\}$. Thus we have two independent maximal p.f.i. extensions for $\mathcal{L}^s_1$. We can repeat this effective deletion-procedure finitely many times and still obtain finitely many extensions. For regaining maximality, we are indeed "restricted" in the structural sense. The following lemma illustrates this.

**Lemma 1.** *Let $\mathcal{L}$ be a maximal p.f.i. family and $\mathcal{L} \setminus \{x\}$ where $x \in \mathbb{N}$ and $\{x\} \in \mathcal{L}$. If $\mathcal{L}'$ is a maximal p.f.i. extension of $\mathcal{L} \setminus \{x\}$, then for all $L \in \mathcal{L}'$ which is not in $\mathcal{L} \setminus \{x\}$ we have that $L$ is of the form $\{x\} \cup A$ for some $A \subseteq L_i \in \mathcal{L} \setminus \{x\}$.*

In the following example we see that even when the languages are all finite, we can still regain uncountably many maximal p.f.i. extensions. Let $\mathcal{L} = \{\{0\} \cup \mathcal{L}_3\}$ where $\mathcal{L}_3 = \{\{i,j,k\} : i,j,k \in \mathbb{N} \setminus \{0\}\}$. Clearly $\mathcal{L}$ is maximal p.f.i. family. Consider $\mathcal{L}_3 = \mathcal{L} \setminus \{0\}$, by lemma 1 in order to regain maximality, the languages to add must be of the form $\{0\} \cup A$ for some $A \subseteq L_i$ for some $L_i \in \mathcal{L}_3\}$. Therefore we have the following procedure for achieving maximal p.f.i. extensions of $\mathcal{L}_3$: For each $B \subseteq \mathbb{N} \setminus \{0\}$ add the triplets of the form $\{0,n,m\}$ with $n \neq m$ and $n,m \in B$ and all the pairs of the form $\{0,c\}$ with $c \notin B$. This construction applies to all $B \subseteq \mathbb{N} \setminus \{0\}$, thus $\mathcal{L}_3$ has uncountable many maximal p.f.i. extensions.

Conjecture 2: Every p.f.i. family has either finitely many maximal p.f.i. extensions or uncountably many.

## 4.2 Do maximal c.f.i. families exist?

In this section we address the question whether every c.f.i. family is contained in a maximal one. Or in other words, if we can always find c.f.i. extensions for c.f.i. families. Surprisingly, we show that the latter is indeed always possible, the question whether maximal c.f.i. families exist is answered negatively.

**Theorem 5.** *Take $\mathcal{L}$ an indexed c.f.i. family and $L \in \mathcal{L}$. For any co-DFTT $\overline{D}_L$ of $L$, if $D_L \cup \{n\}$ is such that $n \notin \overline{D}_L \cup L$ then $\mathcal{L} \cup \{D_L \cup \{n\}\}$ is c.f.i..*

**Corollary 2.** *Maximal c.f.i. extensions do not exist for any c.f.i. family $\mathcal{L}$.*

There may be other ways of extending a c.f.i. family than the one described in Theorem 5 as the following example shows.

**Example 1.** *Take the family $\mathcal{L} = \{\{0\},\{0,1\},\{0,1,2\},\ldots,\{0,1,2,3,\ldots,n\},\ldots\}$. This family is c.f.i.. Note that for $L = \{0\}$ we can extend $\mathcal{L}$ with $L \cup \{2\}$ and preserve c.f.i. even though a co-DFTT is $\{1,2\}$. Moreover we can extend it with $L \cup \{3\}$, $L \cup \{4\}$ and so on, and preserve c.f.i..*

# 5 Infinite antichains

Contrary to the results of Section 3, c.f.i. identification is on infinite families more powerful on antichains than p.f.i. identification. The class of all co-singletons, $\{\mathbb{N} \setminus \{i\} \mid i \in \mathbb{N}\}$, is easily seen to be c.f.i. but not p.f.i. The case of infinite families of finite languages is less clear. It is a trivial fact that canonical families which are antichains are always p.f.i. By following a diagonalization strategy, we can construct a non-canonical family which is an antichain but not p.f.i.

**Theorem 6.** *There is a family $\mathcal{L}$ of finite languages which is an antichain and for which there is no canonically indexed $\{D_{f(n)} : n \in \omega\}$ such that $D_i \subseteq L_i$ for all $i \in \mathbb{N}$ and $D_i \not\subseteq L_j$ for all $j \neq i$, i.e, this family is not p.f.i.*

This example happens to be not c.f.i. either. The question remains open, whether there exists such a family which is c.f.i. but not p.f.i.

Conjecture 3: If $\mathcal{L}$ is an antichain of finite languages which is c.f.i., then it is p.f.i.

**Acknowledgement.** We thank S. Terwijn for introducing us to the diagonalization methods used in the proof of Theorem 6.

# References

[1] Dana Angluin. Inductive inference of formal languages from positive data. *Information and control*, 45(2):117–135, 1980.

[2] Nina Gierasimczuk and Dick de Jongh. On the complexity of conclusive update. *The Computer Journal*, pages 56(3):365–377, 2012.

[3] Mark E Gold. Language identification in the limit. *Information and control*, 10(5):447–474, 1967.

[4] Sarah Hiller and Raquel Fernández. A data-driven investigation of corrective feedback on subject omission errors in first language acquisition. *CoNLL 2016*, page 105, 2016.

[5] Steffen Lange and Thomas Zeugmann. Set-driven and rearrangement-independent learning of recursive languages. *Theory of Computing Systems*, 29(6):599–634, 1996.

[6] Yasuhito Mukouchi. Characterization of finite identification. In *Analogical and Inductive Inference*, pages 260–267. Springer, 1992.

[7] Daniel N Osherson, Michael Stob, and Scott Weinstein. *Systems that learn: An introduction to learning theory for cognitive and computer scientists.* The MIT Press, 1986.