Co-Learning of Recursive Languages from Positive Data

Freivalds, Rusins
Institute of Mathematics and Computer Science University of Latvia

Zeugmann, Thomas
Research Institute of Fundamental Information Science Kyushu University

出版情報：RIFIS Technical Report. 110, 1995-04. Research Institute of Fundamental Information Science, Kyushu University
バージョン：
権利関係：

Co–Learning of Recursive Languages from Positive Data

Rusins Freivalds*
Institute of Mathematics and
Computer Science
University of Latvia
Raina bulv. 29, Riga, Latvia
rusins@mii.lu.lv

Thomas Zeugmann
Research Institute of
Fundamental Information Science
Kyushu University 33
Fukuoka 812, Japan
thomas@rifis.kyushu-u.ac.jp

Abstract

The present paper deals with the co-learnability of enumerable families $\mathcal{L}$ of uniformly recursive languages from positive data. This refers to the following scenario. A family $\mathcal{L}$ of target languages as well as hypothesis space for it are specified. The co-learner is fed eventually all positive examples of an unknown target language L chosen from $\mathcal{L}$. The target language L is successfully co-learned if and only if the co-learner can definitely delete all but one possible hypotheses, and the remaining one has to correctly describe L.

We investigate the capabilities of co-learning in dependence on the choice of the hypothesis space, and compare it to language learning in the limit from positive data. We distinguish between *class preserving* learning ($\mathcal{L}$ has to be co-learned with respect to some suitably chosen enumeration of all and only the languages from $\mathcal{L}$), *class comprising* learning ($\mathcal{L}$ has to be co-learned with respect to some hypothesis space containing at least all the languages from $\mathcal{L}$), and *absolute co-learning* ($\mathcal{L}$ has to be co-learned with respect to all class preserving hypothesis spaces for $\mathcal{L}$).

Our results are manyfold. First, it is shown that co-learning is exactly as powerful as learning in the limit provided the hypothesis space is appropriately chosen. However, while learning in the limit is insensitive to the particular choice of the hypothesis space, the power of co-learning crucially depends on it. Therefore we study properties a hypothesis space should have in order to be suitable for co-learning. Finally, we derive sufficient conditions for absolute co-learnabilty, and separate it from finite learning.

*The first author was supported by the grant No. 93.599 form the Latvian Science Council.

1. Introduction

The present paper deals with the co-learnability of enumerable families $\mathcal{L}$ of uniformly recursive languages from positive data. This refers to the following scenario introduced by Freivalds, Karpinski and Smith (1994) in the setting of inductive inference of recursive functions. A family $\mathcal{L}$ of target languages as well as hypothesis space for it are specified. The co-learner is fed eventually all positive examples of an unknown target language L chosen from $\mathcal{L}$. The target language L is successfully co-learned if and only if the co-learner can definitely delete all but one possible hypotheses, and the remaining one has to correctly describe L. This approach derives its motivation from machine learning, where learning algorithms rather often start from a large finite set of possible guesses. Then, all but one are refuted during the learning process. Hence, our model is just the recursion theoretic counterpart of that approach.

We investigate the capabilities of co-learning in dependence on the choice of the hypothesis space, and compare it to learning in the limit, conservative learning, and finite learning. We distinguish between *class preserving* learning ($\mathcal{L}$ has to be co-learned with respect to some suitably chosen enumeration of all and only the languages from $\mathcal{L}$), *class comprising* learning ($\mathcal{L}$ has to be co-learned with respect to some hypothesis space containing at least all the languages from $\mathcal{L}$), and *absolute co-learning* ($\mathcal{L}$ has to be co-learned with respect to all class preserving hypothesis spaces for $\mathcal{L}$).

Our results are manyfold. First, it is shown that co-learning is exactly as powerful as learning in the limit provided the hypothesis space is appropriately chosen. However, while learning in the limit is insensitive to the particular choice of the hypothesis space, the power of co-learning crucially depends on it. The latter result is obtained while studying the co-learnability of the *pattern languages*. Moreover, proving that the pattern languages are not absolutely co-learnable but absolute conservatively inferable allows some deeper insight into the strength to refute *some* and *all but one* hypothesis.

Furthermore, we study properties a hypothesis space should have in order to be suitable for co-learning. Finally, we derive sufficient conditions for absolute co-learnabilty, and separate it from finite learning.

2. Notations and Definitions

Unspecified notations follow Rogers (1967). Let $\mathbb{N} = \{0, 1, 2, ...\}$ be the set of natural numbers. We set $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$. By $\langle \cdot, \cdot \rangle : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ we denote ***Cantor's pairing function***, i.e., $\langle x, y \rangle = ((x + y)^2 + 3x + y)/2$ for all $x, y \in \mathbb{N}$. We use $\mathcal{P}^n$ and $\mathcal{R}^n$ to denote the set of all n-ary partial recursive and total recursive functions over $\mathbb{N}$, respectively. The class of all $\{0, 1\}$ valued functions $f \in \mathcal{R}^n$ is denoted by $\mathcal{R}^n_{0,1}$. For $n = 1$ we omit the upper index, i.e., we set $\mathcal{P} = \mathcal{P}^1$, and $\mathcal{R} = \mathcal{R}^1$ as well as $\mathcal{R}_{0,1} = \mathcal{R}^1_{0,1}$.

Every function $\psi \in \mathcal{P}^2$ is called a numbering. Moreover, let $\psi \in \mathcal{P}^2$, then we write ψ_j instead of $\lambda x \psi(i, x)$. Furthermore, let $\psi \in \mathcal{R}^2_{0,1}$, then by $L(\psi_j)$ we denote the language generated or described by ψ_j, i.e., $L(\psi_j) = \{x \mid \psi_j(x) = 1,\ x \in \mathbb{N}\}$.

Moreover, we call $\mathcal{L} = (L(\psi_j))_{j \in \mathbb{N}}$ an ***indexed family*** (cf. Angluin (1980b)). For the sake of presentation, we restrict ourselves to consider exclusively indexed families of non-empty languages. Let $\mathcal{L}$ be an indexed family. Every numbering $\psi \in \mathcal{R}^2_{0,1}$ is called ***hypothesis space***. A hypothesis space $\psi \in \mathcal{R}^2_{0,1}$ is said to be ***class comprising*** for an indexed family $\mathcal{L}$ iff $range(\mathcal{L}) \subseteq \{L(\psi_j) | \ j \in \mathbb{N}\}$. Furthermore, we call a hypothesis space $\psi \in \mathcal{R}^2_{0,1}$ ***class preserving*** for $\mathcal{L}$ iff $range(\mathcal{L}) = \{L(\psi_j) | \ j \in \mathbb{N}\}$.

Let L be a language and let $t = s_0, s_1, s_2, ...$ be an infinite sequence of natural numbers such that $range(t) = \{s_k | k \in \mathbb{N}\} = L$. Then t is said to be a ***text*** for L or, synonymously, a ***positive presentation***. By $text(L)$ we denote the set of all positive presentations of L. Moreover, let t be a text, and let y be a number. Then t_y denotes the initial segment of t of length $y + 1$, i.e., $t_y = s_0, ..., s_y$. Finally, t_y^+ denotes the ***content*** of t_y, i.e., $t_y^+ = \{s_z | \ z \leq y\}$.

As in Gold (1967), we define an ***inductive inference machine*** (abbr. IIM) to be an algorithmic device which works as follows: The IIM takes as its input incrementally increasing initial segments of a text t and it either requests the next input, or it first outputs a hypothesis, i.e., a number, and then it requests the next input.

We interpret the hypotheses output by an IIM with respect to some suitably chosen hypothesis space $\psi \in \mathcal{R}^2_{0,1}$. When an IIM outputs a number j, we interpret it to mean that the machine is hypothesizing the language $L(\psi_j)$.

Furthermore, we define a ***co-learning machine*** (abbr. CLM) to be an algorithmic device working as follows: The CLM takes as its input incrementally increasing initial segments of a text t (as an IIM does) and it either requests the next input, or it first outputs a number, and then it requests the next input.

However, there is a *major* difference in the *semantics* of the output of an IIM and CLM, respectively. Let $\psi \in \mathcal{R}^2_{0,1}$ be any hypothesis space. Suppose a CLM M has been successively fed an initial segment t_y of a text t, and it has output numbers $j_0, ..., j_z$, $z \leq y$. Then we interpret $j = \min(\mathbb{N} \setminus \{j_0, ..., j_z\})$ as M's actual guess. Intuitively speaking, if a CLM outputs a number j, then it *definitely deletes* j from its list of potential hypotheses.

Let M be an IIM or a CLM, let t be a text, and $y \in \mathbb{N}$. Then we use $M(t_y)$ to denote the last number that has been output by M when successively fed t_y. We define convergence of IIMs as usual. Let t be a text, and let M be an IIM. The sequence $(M(t_y))_{y \in \mathbb{N}}$ is said to ***converge in the limit*** to the number j if and only if either $(M(t_y))_{y \in \mathbb{N}}$ is infinite and all but finitely many terms of it are equal to j, or $(M(t_y))_{y \in \mathbb{N}}$ is non-empty and finite, and its last term is j.

A CLM M is said to ***stabilize*** on a text t to a number j if and only if $\{j\} = \mathbb{N} \setminus \{M(t_y) | \ y \in \mathbb{N}\}$. Intuitively, a CLM M stabilizes itself on a number j if it outputs all but the natural number j when fed a text. Now we are ready to define learning and co-learning.

Definition 1. (Gold, 1967) *Let $\mathcal{L}$ be an indexed family, let L be a language, and let $\psi \in \mathcal{R}^2_{0,1}$ be a hypothesis space.* ***An IIM M CLIM–identifies L from text with respect to ψ*** *iff for every text t for L, there exists a $j \in \mathbb{N}$ such that the sequence $(M(t_y))_{y \in \mathbb{N}}$ converges in the limit to j and $L = L(\psi_j)$.*

Furthermore, M CLIM–identifies $\mathcal{L}$ with respect to ψ if and only if, for each $L \in$

range($\mathcal{L}$), *M CLIM–identifies L with respect to* ψ.

Finally, let CLIM denote the collection of all indexed families $\mathcal{L}$ *for which there are an IIM M and a hypothesis space* ψ *such that M CLIM–identifies* $\mathcal{L}$ *with respect to* ψ.

Since, by the definition of convergence, only finitely many data of L were seen by the IIM upto the (unknown) point of convergence, whenever an IIM identifies the language L, some form of learning must have taken place. For this reason, hereinafter the terms *infer*, *learn*, and *identify* are used interchangeably.

In Definition 1, *LIM* stands for "limit." Furthermore, the prefix C is used to indicate ***class comprising*** learning, i.e., the fact that $\mathcal{L}$ may be learned with respect to some class comprising hypothesis space ψ for $\mathcal{L}$. The restriction of *CLIM* to class preserving hypothesis spaces is denoted by *LIM* and referred to as ***class preserving*** inference. Moreover, we use the prefix A to express the fact that an indexed family $\mathcal{L}$ may be inferred with respect to ***all*** class preserving hypothesis spaces for $\mathcal{L}$, and we refer to this learning model as to ***absolute*** learning. We adopt this convention in the definitions of the learning types below.

The following proposition clarifies the relations between absolute, class preserving and class comprising learning in the limit.

Proposition 1. (Lange and Zeugmann, 1993c)

$ALIM = LIM = CLIM$

Note that, in general, it is not decidable whether or not an IIM M has already converged on a text t for the target language L. With the next definition, we consider a special case where it has to be decidable whether or not an IIM has successfully finished the learning task.

Definition 2. (Gold, 1967; Trakhtenbrot and Barzdin, 1970) *Let* $\mathcal{L}$ *be an indexed family, let L be a language, and let* $\psi \in \mathcal{R}^2_{0,1}$ *be a hypothesis space.* ***An IIM M CFIN–identifies L from text with respect to*** ψ *iff for every text t for L, there exists a* $j \in \mathbb{N}$ *such that M, when successively fed t, outputs the single hypothesis* j, $L = L(\psi_j)$, *and stops thereafter.*

Furthermore, M CFIN–identifies L with respect to ψ *if and only if, for each* $L \in$ *range*($\mathcal{L}$), *M CFIN–identifies L with respect to* ψ.

The resulting learning type is denoted by CFIN.

The following proposition states that, if an indexed family $\mathcal{L}$ can be $CFIN$–learned with respect to some hypothesis space ψ for it, then it can be finitely inferred with respect to every class preserving hypothesis space for $\mathcal{L}$.

Proposition 2. (Zeugmann, Lange and Kapur, 1995)

$AFIN = FIN = CFIN$

Next we adapt the definition of co-learnability introduced by Freivalds, Karpinski and Smith (1994) to language learning from positive data.

Definition 3. *Let* $\mathcal{L}$ *be an indexed family, let L be a language, and let* $\psi \in \mathcal{R}^2_{0,1}$ *be a hypothesis space.* ***A CLM M co–CFIN–identifies L from text with respect to*** ψ *iff for every text t for L, there exists a* $j \in \mathbb{N}$ *such that M on t stabilizes to j*

and $L = L(\psi_j)$.

Furthermore, M $co-CFIN$*–identifies* $\mathcal{L}$ *with respect to* ψ *if and only if, for each* $L \in range(\mathcal{L})$, M $co-CFIN$*–identifies* L *with respect to* ψ.

Finally, let $co-CFIN$ *denote the collection of all indexed families* $\mathcal{L}$ *for which there are a CLM* M *and a hypothesis space* ψ *such that* M $co-CFIN$*–identifies* $\mathcal{L}$ *with respect to* ψ.

Next we define ***conservative*** IIMs. Intuitively speaking, conservative IIMs maintain their actual hypothesis at least as long as the have received data that "provably misclassify" it. Hence, whenever a conservative IIM performs a mind change it is because it has perceived a clear contradiction between its hypothesis and the actual input.

Definition 4. (Angluin, 1980b) *Let* $\mathcal{L}$ *be an indexed family, let* L *be a language, and let* $\psi \in \mathcal{R}^2_{0,1}$ *be a hypothesis space.* ***An IIM*** M ***CCONSV–identifies*** L ***from text with respect to*** ψ *iff*

(1) M *CLIM–identifies* L *with respect to* ψ,

(2) *for every text* t *for* L *the following condition is satisfied:*
whenever M *on input* t_y *makes the guess* j_y *and then makes the guess* $j_{y+k} \neq j_y$ *at some subsequent step, then* $L(\psi_{j_y})$ *must fail to contain some string from* t^+_{y+k}.

Finally, M *CCONSV–identifies* $\mathcal{L}$ *with respect to* ψ *if and only if, for each* $L \in range(\mathcal{L})$, M *CCONSV–identifies* L *with respect to* ψ.

The resulting collection of sets CCONSV is defined in a manner analogous to that above.

The following proposition shows that conservative learning is sensitive to the particular choice of the hypothesis space.

Proposition 3. (Lange and Zeugmann, 1993b)

$ACONSV \subset CONSV \subset CCONSV \subset ALIM$

3. Results

As already mentioned in the Introduction, Freivalds, Karpinski and Smith (1994) recently studied co-learnability of recursive functions. On the other hand, in inductive inference functions and languages are usually very different from each other (cf., e.g., Osherson, Stob and Weinstein (1986) and the references therein). Hence, it is only natural to ask whether or not there are major differences between the co-learnability of recursive functions and recursive languages, too. In this section we provide both similarities and distinctions. However, the overall goal is far-reaching, and results presented in the following subsection will guide us to central questions concerning the co-inferability of recursive languages.

3.1. Basic Results

We start our investigations by clarifying whether or not the capabilities of co-learning do depend on the class of admissible hypothesis spaces. Clearly, $co-AFIN \subseteq$

$co-FIN \subseteq co-CFIN$. First we ask whether these inclusions are proper, and what are the lower and upper bounds of this hierarchy. The first theorem provides a first lower bound.

Theorem 1. *Let $\mathcal{L}$ be an indexed family. Then $\mathcal{L} \in FIN$ implies $\mathcal{L} \in co\text{–}AFIN$.*

Proof. Let $\psi \in \mathcal{R}^2_{0,1}$ be any class preserving hypothesis space for $\mathcal{L}$. By Proposition 2 there exists an IIM M that finitely infers $\mathcal{L}$ with respect to ψ. The desired CLM $\hat{M}$ can be defined as follows. Let $L \in range(\mathcal{L})$, $t \in text(L)$, and $y \in \mathbb{N}$. The CLM $\hat{M}$ simulates M on input t_y. Now, two cases are possible. First, M outputs nothing and request the next input. In this case $\hat{M}$ also requests the next input and does not output any hypothesis. Second, M outputs a hypothesis j and stops. Due to the definition of FIN we know that $L = L(\psi_j)$. Then $\hat{M}$ outputs, one at a time, all natural numbers but j. Clearly, $\hat{M}$ stabilizes on j, and hence it indeed $co-AFIN$–infers L. q.e.d.

Next we deal with the desired upper bound.

Theorem 2. *Let $\mathcal{L}$ be an indexed family, and let $\psi \in \mathcal{R}^2_{0,1}$ be any class comprising hypothesis space for $\mathcal{L}$. Then, $\mathcal{L} \in co-CFIN$ with respect to ψ implies $\mathcal{L} \in CLIM$ with respect to ψ.*

Proof. Let M be a CLM that witnesses $\mathcal{L} \in co-CFIN$ with respect to ψ. The desired IIM $\hat{M}$ is defined as follows. Let $L \in range(\mathcal{L})$, $t \in text(L)$, and $y \in \mathbb{N}$. $\hat{M}$ simulates M on input t_y. If M does not produce an output, then $\hat{M}$ requests the next input and outputs nothing. Otherwise, it outputs the least number not yet definitely deleted by M and requests the next input. Obviously, $\hat{M}$ $CLIM$–learns L. q.e.d.

As the latter theorem shows, co-learning is at most as powerful as learning in the limit. With the next theorem we establish the equality of $CLIM$ and $co-CFIN$.

Theorem 3. *Let $\mathcal{L}$ be an indexed family. If $\mathcal{L} \in CLIM$ then there exists a class preserving hypothesis space $\tau \in \mathcal{R}^2_{0,1}$ such that $\mathcal{L} \in co-FIN$ with respect to τ.*

Proof. Let $\mathcal{L}$ be an indexed family such that $\mathcal{L} \in CLIM$. By Proposition 1 we may assume, without loss of generality, that there are a class preserving hypothesis space $\psi \in \mathcal{R}^2_{0,1}$ for $\mathcal{L}$ and an IIM M such that M LIM–identifies $\mathcal{L}$ with respect to ψ. We define the desired class preserving hypothesis space τ as follows. For all $j, x, z \in \mathbb{N}$ we set $\tau_{\langle j,x \rangle}(z) = \psi_j(z)$. Hence, the hypothesis space τ contains for every language $L \in range(\mathcal{L})$ infinitely many descriptions. Moreover, given any description $\langle j, x \rangle$ one can easily compute infinitely many descriptions generating the same language $L(\tau_{\langle j,x \rangle})$. Applying the same technique as in Freivalds, Karpinski and Smith (1994) one directly obtains a CLM $\hat{M}$ that $co-FIN$–infers $\mathcal{L}$ with respect to τ. q.e.d.

Theorem 2 and 3 as well as Proposition 1 directly allow the following corollary.

Corollary 4.

(1) $ALIM = co-CFIN$,

(2) $co-FIN = co-CFIN$.

The latter corollary yields some insight into the potential capabilities of co-learning. In particular, we already know that every LIM–inferable indexed family is also co-

learnable provided the hypothesis space is appropriately chosen. Hence, in order to decide whether or not a particular indexed family can be co-learned one can apply any of the known criteria for LIM–inferability (cf., e.g., Angluin (1980b), Sato and Umayahara (1992)). On the other hand, if an indexed family $\mathcal{L}$ is $CLIM$–identifiable at all then it can be learned in the limit with respect to any class comprising hypothesis space for $\mathcal{L}$. That is, the principle capabilities of learning in the limit are insensitive to the particular choice of the hypothesis space. Therefore, it is only natural to ask whether or not the power of co-learnability does depend on the choice of the possible hypothesis spaces. We answer this question by clarifying the relations between absolute and class preserving co-learning. We achieve this goal by studying the co-learnability of the pattern languages introduced by Angluin (1980a). Nix (1983) outlined interesting applications of inference algorithms for pattern languages. Shinohara (1982), Kearns and Pitt (1989), Schapire (1991), Lange and Wiehagen (1991) as well as Wiehagen and Zeugmann (1994) studied polynomial time learnability of pattern languages. Furthermore, Zeugmann, Lange and Kapur (1995) investigated the inferability of pattern languages under various constraints of monotonicity. So let us define what are a pattern and a pattern language. Let $\Sigma = \{a, b, ..\}$ be any non–empty finite alphabet containing at least two elements. Furthermore, let $X = \{x_i \mid i \in \mathbb{N}\}$ be an infinite set of variables such that $\Sigma \cap X = \emptyset$. *Patterns* are non-empty strings from $\Sigma \cup X$, e.g., ab, ax_1ccc, $bx_1x_1cx_2x_2$ are patterns. $L(p)$, the language generated by pattern p is the set of strings which can be obtained by substituting non–null strings from Σ^* for the variables of the pattern p. Thus $aabbb$ is generable from pattern ax_1x_2b, while $aabba$ is not. Pat and PAT denote the set of all patterns and of all pattern languages over Σ, respectively. From a practical point of view it is highly desirable to choose the hypothesis space as small as possible. For that purpose we use the canonical form of patterns (cf. Angluin (1980a)). A pattern p is in *canonical form* provided that if k is the number of variables in p, then the variables occurring in p are precisely $x_0, ..., x_{k-1}$. Moreover, for every j with $0 \leq j < k-1$, the leftmost occurrence of x_j in p is left to the leftmost occurrence of x_{j+1} in p. If a pattern p is in canonical form then we refer to p as a canonical pattern. Let $Patc$ denote the set of all canonical patterns. Clearly, for every pattern p there exists a unique $q \in Patc$ such that $L(p) = L(q)$. Finally, choose any repetition free effective enumeration $p_0, p_1, ...$ of $Patc$ and define $PAT = (L(p_j))_{j \in \mathbb{N}}$. Since membership for pattern languages is uniformly decidable, there is a $\psi \in \mathcal{R}^2_{0,1}$ such that $L(p_j) = L(\psi_j)$ for all $j \in \mathbb{N}$ (cf. Angluin (1980a)). Note that ψ enumerates every pattern language exactly ones. By Angluin (1980b) we also know that $PAT \in CONSV$ with respect to ψ. However, PAT cannot be $co-FIN$–inferred with respect to ψ as our next theorem shows.

Theorem 5. *Let PAT and ψ be defined as above. Then, $PAT \notin co-FIN$ with respect to ψ.*

Proof. Suppose the converse, i.e., there is a CLM M that $co-FIN$–learns PAT with respect to ψ. Now, let k be the index of $L(x_1)$ in the hypothesis space ψ, i.e., $L(x_1) = L(\psi_k)$. We proceed in showing that there is a text $\hat{t} \in text(L(\psi_k))$ from which M fails to $co-FIN$–identify $L(x_1)$. For that purpose let $p \in Patc$ be any pattern with $L(p) \neq L(x_1)$, and let $t \in text(L(p))$. Since M $co-FIN$–infers $L(p)$, there exists a y such that $k = M(t_y)$, since otherwise M cannot stabilize on a correct hypothesis for $L(p)$. But now we observe that t_y is an initial segment of some text $\hat{t} \in text(L(\psi_k)) = text(L(x_1))$, since $L(x_1) = \Sigma^+$. Therefore, $(M(\hat{t}_z))_{z \in \mathbb{N}}$ cannot

stabilize on k, a contradiction. q.e.d.

The latter theorem directly implies the wanted separation of absolute and class preserving co-learnability.

Corollary 6. $co-AFIN \subset co-FIN$

Additionally, Theorem 5 provides a main ingredient to show that absolute conservative learning does not imply absolute co-inferability.

Corollary 7.

(1) $ACONSV \setminus co-AFIN \neq \emptyset$

(2) $AFIN \subset ACONSV$

Proof. First, we prove Assertion (1). In accordance with Theorem 5 it suffices to show that $PAT \in ACONSV$. Let M and ψ be chosen as in the proof of Theorem 5, i.e., M witnesses $PAT \in CONSV$ with respect to ψ. Now, let τ be any class preserving hypothesis space for PAT. We have to show that there exists an IIM $\hat{M}$ conservatively inferring PAT with respect to τ.

The main ingredient to the definition of $\hat{M}$ is the fact that PAT can be finitely inferred from positive and negative data with respect to ψ (cf. Lange and Zeugmann (1993a)). Therefore, we can define $\hat{M}$ as follows. Let $L \in PAT$, let $t \in text(L)$ and let $y \in \mathbb{N}$.

$\hat{M}(t_y) =$ "Compute $M(t_y)$. If M when successively fed t_y does not produce any hypothesis, then output nothing and request the next input.
Otherwise, let $j = M(t_y)$. Compute $\psi_j(0), ..., \psi_j(z)$ and search for the least index k such that $\tau_k(x) = \psi_j(x)$ for all $x \leq z$, where z is the least number such that all shortest strings in $L(\psi_j)$ are classified. Output k and request the next input."

We show that $\hat{M}$ conservatively infers PAT with respect to τ. Remember that ψ and τ are class preserving hypothesis spaces for PAT. Hence, if $j = M(t_y)$ then $L(\psi_j)$ is a pattern language. As shown in Lange and Zeugmann (1993a), if all shortest strings in $L(\psi_j)$ are classified, then $L(\psi_j) = L(\tau_k)$ provided $\tau_k(x) = \psi_j(x)$ for all $x \leq z$. Therefore, $\hat{M}$ is conservative and it learns PAT with respect to τ. Consequently, $PAT \in ACONSV$, and (1) is proved.

Now, Assertion (2) is an immediate consequence of $PAT \notin FIN$ (cf. Lange and Zeugmann (1993a)). q.e.d.

Furthermore, as we have seen, one-to-one hypothesis spaces ψ do not guarantee the co-inferability of the corresponding indexed families $(L(\psi_j))_{j \in \mathbb{N}}$. This nicely contrasts a result for the co-learnability of recursive functions (cf. Freivalds, Karpinski and Smith (1994), Theorem 3). Moreover, the proof technique applied in the demonstration of Theorem 5 allows the following generalizations.

Theorem 8. *Let $\mathcal{L} = (L_j)_{j \in \mathbb{N}}$ be any indexed family such that $L = \bigcup_{j \in \mathbb{N}} L_j \in range(\mathcal{L})$. Furthermore, let $\psi \in \mathcal{R}^2_{0,1}$ be any hypothesis space for $\mathcal{L}$ such that* card$(\{k |$

$k \in \mathbb{N},\ L(\psi_k) = L\}) < \infty$. *Then,* $\mathcal{L}$ *cannot be* $co-CFIN$*–identified with respect to* ψ.

Proof. Suppose the converse, i.e., there is a CLM M that $co-FIN$–learns $\mathcal{L}$ with respect to ψ. Let $L = \bigcup_{j \in \mathbb{N}} L_j$, and let $\{k_1, ..., k_m\}$ be the set of all ψ–indices that generate L. Furthermore, let $\hat{L} \in range(\mathcal{L})$ be any fixed language such that $\hat{L} \neq L$, and let $\hat{t}$ be any text for $\hat{L}$. Consequently, there has to be a $y \in \mathbb{N}$ such that M when successively fed $\hat{t}_y$ outputs at least the numbers from $\{k_1, ..., k_m\}$. Again, we observe that $\hat{t}_y$ constitutes an initial segment of some text t for L. Consequently, $(M(t_z))_{z \in \mathbb{N}}$ cannot stabilize on a correct index for L. q.e.d.

Theorem 9. *Let* $\mathcal{L} = (L_j)_{j \in \mathbb{N}}$ *be any indexed family containing at least two languages* L_k, L_z *such that* $L_k \subset L_z$. *Then, for any hypothesis space* $\psi \in \mathcal{R}^2_{0,1}$ *satisfying* $\text{card}(\{m \mid L(\psi_m) = L_z\}) < \infty$ *we have that* $\mathcal{L} \notin co-CFIN$ *with respect to* ψ.

Proof. Again, the same argument as above applies *mutatis mutandis.* q.e.d.

As we have seen, the power of co-learnability may heavily depend on the particular choice of the hypothesis space. However, Theorems 8 and 9 might suggest that inclusion of some languages in the target indexed family causes the sensitivity of co-learning with respect to the choice of the hypothesis space. Nevertheless, the situation is more complex as our next theorem shows.

Theorem 10. *There is an indexed family* $\mathcal{L}$ *such that*

(1) $L \not\subset \hat{L}$ *for all* $L, \hat{L} \in range(\mathcal{L})$,

(2) *there exists a class preserving hypothesis space* τ *for* $\mathcal{L}$ *with respect to which* $\mathcal{L}$ *cannot be* $co-FIN$ *learned.*

Proof. Let $M_0, M_1, M_2, ...$ be the canonical enumeration of all CLMs. We construct the desired indexed family by defining the numbering $\tau \in \mathcal{R}^2_{0,1}$. As we shall see, all languages are finite ones and they either contain one or two numbers. This is done as follows. By p_j we denote the jth prime number.

We define $\tau_{2j}(p_j) = \tau_{2j+1}(p_j) = 1$ for all $j \in \mathbb{N}$. Hence, $L(\tau_{2j})$ as well as $L(\tau_{2j+1})$ contain p_j. In order to complete the definition of τ let t^j_x be the finite sequence of length $x + 1$ with $content(t^j_x) = \{p_j\}$.

Then, for $x = 0, 1, ..., p_j - 1, p_j + 1, ...$ we successively define $\tau_{2j}(x)$ and $\tau_{2j+1}(x)$ as follows. Simulate the computation of M_j on input t^j_x. If M_j when fed t^j_x does not output a hypothesis or $n = M_j(t^j_x)$ satisfies $n \notin \{2j, 2j+1\}$ then set $\tau_{2j}(x) = \tau_{2j+1}(x) = 0$. Otherwise, define $\tau_{2j}(p_j^{x+2}) = 1$ and $\tau_{2j+1}(p_j^{x+3}) = 1$ and set $\tau_{2j}(x) = \tau_{2j+1}(x) = 0$ for all $x \in \mathbb{N}$ for which τ_{2j} and τ_{2j+1} are not defined yet.

Obviously, $\tau \in \mathcal{R}^2_{0,1}$. Moreover, Assertion (1) is an immediate consequence of our definition, since any two languages are either equal or incomparable.

We proceed with Assertion (2). Suppose the converse, i.e., $(L(\tau_z))_{z \in \mathbb{N}} \in co-FIN$ with respect to τ. Hence, there must be a CLM M witnessing the co-learnability of $(L(\tau_z))_{z \in \mathbb{N}}$ with respect to τ. Moreover, this CLM has to appear in the canonical enumeration of all CLMs. Thus, there is a j such that $M = M_j$. Now, consider M_j's behavior when successively fed t^j_x. We distinguish the following cases.

Case 1. M_j when successively fed t_x^j, $x \in \mathbb{N}$, does never output a number $n \in \{2j, 2j+1\}$.

By construction, $L(\tau_{2j}) = L(\tau_{2j+1}) = \{p_j\}$, and therefore $t^j = (t_x^j)_{x \in \mathbb{N}}$ constitutes a text for $L(\tau_{2j})$ as well as for $L(\tau_{2j+1})$. But M_j on input t^j does neither output $2j$ nor does it output $2j+1$. Thus, it cannot stabilize on input t^j, a contradiction.

Case 2. M_j when successively fed t_x^j, $x \in \mathbb{N}$, does output a number $n \in \{2j, 2j+1\}$.

Then, in accordance with the definition of τ we know that $\{p_j\} \neq L(\tau_{2j}) \neq L(\tau_{2j+1}) \neq \{p_j\}$, and that both languages contain p_j. Assume M_j outputs $2j$, say on input t_x^j. Then, t_x^j is an initial segment of a text t for $L(\tau_{2j})$ but M_j has definitely deleted the only correct hypothesis for $L(\tau_{2j})$ when fed t_x. Hence, it cannot co-learn $L(\tau_{2j})$ from text t, a contradiction. The remaining case that M_j outputs $2j+1$ can be analogously handled. q.e.d.

Next, we are interested in learning under what conditions hypothesis spaces are appropriate for co-inferability. This is done in the next subsection.

3.2. Main Results

This subsection is devoted to the problem why an indexed family that is co-learnable with respect to some hypothesis space ψ might become $co-FIN$–non-inferable with respect to other hypothesis spaces τ. First of all, we want to point to another difference between learning in the limit and co-inference. Gold (1967) proved that every IIM which learns an indexed family $\mathcal{L}$ with respect to some hypothesis space ψ can be effectively transformed into an IIM $\hat{M}$ inferring $\mathcal{L}$ with respect to some other hypothesis space τ provided that there is a limiting recursive compiler from ψ into τ. For co-learning, the situation is much more subtle. To see this, we introduce the following notation.

Definition 5. *Let $\psi, \tau \in \mathcal{R}_{0,1}^2$ be two hypothesis spaces. τ is said to be* ***reducible*** *to ψ (abbr. $\tau \leq_c \psi$) iff there exists a recursive compiler $c \in \mathcal{R}$ such that $\tau_j = \psi_{c(j)}$ for all $j \in \mathbb{N}$.*

Clearly, if $\mathcal{L}$ is an indexed family and $\psi, \tau \in \mathcal{R}_{0,1}^2$ are hypothesis spaces for $\mathcal{L}$ satisfying $\tau \leq_c \psi$, then $\mathcal{L} \in CLIM$ with respect to τ implies $\mathcal{L} \in CLIM$ with respect to ψ. In contrast, for $co-FIN$ we have the following theorem.

Theorem 11. *There are an indexed family $\mathcal{L}$ and class preserving hypothesis spaces ψ, τ for $\mathcal{L}$ such that*

(1) $\tau \leq_c \psi$,

(2) $\mathcal{L} \in co-FIN$ *with respect to* τ *but* $\mathcal{L} \notin co-FIN$ *with respect to* ψ.

Proof. We set $\mathcal{L} = PAT$. Since $PAT \in LIM$, by Theorem 3 we may conclude that there exists a class preserving hypothesis space τ such that $PAT \in co-FIN$ with respect to τ. Furthermore, let ψ be the class preserving hypothesis space for PAT from Theorem 5. Hence, $PAT \notin co-FIN$ with respect to ψ. It remains to show that there is a recursive compiler c such that $\tau \leq_c \psi$. But this has been implicitly done in the proof of Corollary 7. q.e.d.

Consequently, it is only natural to ask under what circumstances reducibility of hypothesis spaces does preserve co-learnability. Our next theorem provides a partial answer to this question.

Theorem 12. *Let $\mathcal{L}$ be an indexed family. Furthermore, let τ be any class preserving hypothesis space for $\mathcal{L}$ that contains precisely one index for every $L \in range(\mathcal{L})$. Then we have:*

$\mathcal{L} \in co-FIN$ with respect to τ implies $\mathcal{L} \in co-FIN$ with respect to any class preserving hypothesis space ψ provided $\psi \leq_c \tau$.

Proof. By assumption, there exists a CLM M witnessing $\mathcal{L} \in co-FIN$ with respect to τ. Let ψ be any class preserving hypothesis space for $\mathcal{L}$ with $\psi \leq_c \tau$. We have to construct a CLM $\hat{M}$ that $co-FIN$–infers $\mathcal{L}$ with respect to ψ. The desired CLM $\hat{M}$ may be defined as follows. Let $L \in range(\mathcal{L})$, and let $t \in text(L)$. Then, $\hat{M}$ when successively fed $(t_y)_{y \in \mathbb{N}}$ works as follows:

$\hat{M}$ simulates M when successively fed $(t_y)_{y \in \mathbb{N}}$ and keeps track of the following sets $I(\tau, y)$, $C(\psi, y)$, and $G(\psi, y)$ of τ and ψ indices. Let $I(\tau, y) = \{M(t_z) \mid z \leq y\}$. That is, $I(\tau, y)$ is the set of all τ indices that M has definitely deleted when successively fed t_y. Since τ is a one-to-one hypothesis space, we know that none of the indices $j \in I(\tau, y)$ may satisfy $L = L(\tau_j)$. However, we have to ensure that $\hat{M}$ definitely deletes all indices i in the hypothesis space ψ that are equivalent to one of the τ–indices in $I(\tau, y)$. Therefore, $\hat{M}$ additionally computes $C(\psi, y) = \{c(i) \mid 0 \leq i \leq y,\ c(i) \in I(\tau, y)\}$, and by dovetailing, it successively outputs all elements in $C(\psi, y)$. Moreover, let $a_y = \min(\mathbb{N} \setminus I(\tau, y))$, i.e., a_y is M's actual guess. The CLM $\hat{M}$ seeks the least index i_y such that $c(i_y) = a_y$, and computes $G(\psi, y) = \{i \mid i_y < i \leq i_y + y,\ c(i) = a_y\}$. Note that the *unbounded* search for i_y has to terminate, since ψ and τ are class preserving hypothesis spaces and $\psi \leq_c \tau$. Again, by dovetailing it successively outputs all elements from $G(\psi, y)$.

It remains to show that $\hat{M}$ witnesses $\mathcal{L} \in co-FIN$ with respect to ψ. Let $a = \min(\mathbb{N} \setminus \{M(t_y) \mid y \in \mathbb{N}\})$, i.e., a is the index the CLM M stabilizes to. We have to argue that $\hat{M}$ outputs all natural numbers except i, where i is the least number satisfying $c(i) = a$. In accordance with $\hat{M}$'s definition it is obvious that $\hat{M}$ does not output i. Moreover, by the definition of the sets $I(\tau, y)$ and $C(\psi, y)$ one straightforwardly obtains that $\hat{M}$ sometimes outputs all ψ–indices j with $L \neq L(\psi_i)$. Hence, it remains to argue that all but the ψ–index i of L are output, too. But this is ensured by the definition of the set $G(\psi, y)$ in which $\hat{M}$ successively keeps track of all other possible ψ–indices. Finally, if M changes its actual guess, say from a_y to a_{y+1} then any number in $G(\psi, y)$ which has not already been output has to appear in $I(\tau, y + r)$ for some $r \in \mathbb{N}$. Hence, $\hat{M}$ $co-FIN$–learns $\mathcal{L}$ with respect to ψ. q.e.d.

Note that the latter theorem establishes a certain type of "co-reducibility," i.e., instead of requiring $\tau \leq_c \psi$, as for "traditional" learning types, we demand $\psi \leq_c \tau$. This is, in general, a stronger requirement, since $\psi \leq_c \tau$ implies $\tau \leq_{\tilde{c}} \psi$. The latter implication easily follows, since τ is a one-to-one hypothesis space.

Moreover, the latter theorem can be successfully applied to solve the intriguing problem whether or not $AFIN \subset co-AFIN$. The affirmative answer is provided by our next theorem.

Theorem 13. $co-AFIN \setminus AFIN \neq \emptyset$

Proof. First, we define the desired indexed family $\mathcal{L}$ witnessing the announced separation. For the sake of presentation, we describe $\mathcal{L}$ as a family of languages over an alphabet Σ. As we shall see, $AFIN$ and $co-AFIN$ may be even separated over a one letter alphabet. We set $\Sigma = \{a\}$, and define $L_j = \{a^n \mid n \in \mathbb{N}^+, n \neq j\}$ for all $j \in \mathbb{N}^+$. Clearly, $\mathcal{L} = (L_j)_{j \in \mathbb{N}^+}$ is an indexed family.

Claim 1. $\mathcal{L} \notin AFIN$

It suffices to show that $\mathcal{L}$ cannot be finitely learned with respect to the hypothesis space $\mathcal{L}$. Suppose the converse, i.e., there is an IIM witnessing $\mathcal{L} \in FIN$ with respect to $\mathcal{L}$. We consider M's behavior on the following text t_{fool}. M is fed $a^2, a^3, ...$ until it outputs the hypothesis 1. In case it does not, we are already done, since then M does not finitely learn L_1 from its lexicographically ordered text. But if it does, say on input $a^2, a^3, ..., a^x$, we may define t_{fool} as follows. We set $t_{fool} = a^2, a^3, ..., a^x, a, a^{x+2}, a^{x+3}, ...$, i.e., t_{fool} is a text for L_{x+1}. However, when successively fed t_{fool} the IIM M converges to 1, and $L_1 \neq L_{x+1}$, a contradiction.

The remaining part of the proof, i.e., the demonstration of $\mathcal{L} \in co-AFIN$ is divided into two parts. First, we show that $\mathcal{L} \in co-FIN$ with respect to $\mathcal{L}$. Next, we apply Theorem 5 to prove that $\mathcal{L} \in co-FIN$ with respect to every class preserving hypothesis space ψ for $\mathcal{L}$.

Claim 2. $\mathcal{L} \in co-FIN$ with respect to $\mathcal{L}$.

The desired CLM M can be defined as follows. Let $L \in range(\mathcal{L})$, $t \in text(L)$, and let $y \in \mathbb{N}$. We define:

$M(t_y) =$ "If $y = 0$, then compute the unique number j such that $t_0 = a^j$. Output j, and request the next input.
For $y \geq 1$ proceed inductively as follows. Let $I(y)$ be the set of all numbers n such that $t_y^+ = \{a^n \mid n \in I(y)\}$. If $I(y) \setminus I(y-1) \neq \emptyset$, then output $j = \min(I(y) \setminus I(y-1))$, and request the next input.
Otherwise, output nothing, and request the next input."

Since $L \in range(\mathcal{L})$, there is a unique number k such that $L = L_k$. It remains to show that M stabilizes on t to k. In accordance with the definition of $\mathcal{L}$ we know that $a^n \in L_k$ for all $n \in \mathbb{N}^+ \setminus \{k\}$. Hence, k is never output by M. Furthermore, since t is a text for L_k, all numbers $n \in \mathbb{N}^+ \setminus \{k\}$ must be sometimes output by M. Thus, M stabilizes to k.

Claim 3. $\mathcal{L} \in co-AFIN$

Let $\psi \in \mathcal{R}_{0,1}^2$ be any class preserving hypothesis space for $\mathcal{L}$. By Theorem 5 it suffices to show that there is a recursive compiler $c \in \mathcal{R}$ such that $\psi \leq_c \mathcal{L}$. For the sake of presentation we suppress all the technicalities dealing with the relevant encoding, i.e., the isomorphism between the string over the alphabet $\{a\}$ and the natural numbers. The desired compiler c can be defined as follows. Let $i \in \mathbb{N}$. Compute $\psi_i(0), \psi_i(1), ...$ until the least $x \in \mathbb{N}$ with $\psi_i(x) = 0$ is found. Since ψ is a class preserving hypothesis space, this unbounded search has to terminate. Moreover, the number x encodes the unique missing string, say a^k, over the alphabet $\{a\}$ that characterizes $L(\psi_i)$. Thus, we can define $c(i) = k$. Obviously, $L(\psi_i) = L_k$, and hence c is a compiler from ψ to $\mathcal{L}$. q.e.d.

4. References

ANGLUIN, D. (1980a), Finding patterns common to a set of strings, *Journal of Computer and System Sciences*, **21**, 46 – 62.

ANGLUIN, D. (1980b), Inductive inference of formal languages from positive data, *Information and Control* **45**, 117 – 135.

ANGLUIN, D., AND SMITH, C.H. (1983), Inductive inference: theory and methods, *Computing Surveys* **15**, 237 – 269.

ANGLUIN, D., AND SMITH, C.H. (1987), Formal inductive inference, *in* "Encyclopedia of Artificial Intelligence" (St.C. Shapiro, Ed.), Vol. 1, pp. 409 – 418, Wiley-Interscience Publication, New York.

FREIVALDS, R., KARPINSKI, M., AND SMITH, C.H. (1994), Co-learning of total recursive functions, *in* "Proceedings 7th Annual ACM Conference on Computational Learning Theory," New Brunswick, July 1994, pp. 190 – 197, ACM Press, New York.

FREIVALDS, R., GOBLEJA, D., KARPINSKI, M., AND SMITH, C.H. (1994), Colearnability and FIN-identifiability of enumerable classes of total recursive functions, *in* "Proceedings 4th International Workshop on Analogical and Inductive Inference, AII'94," (S. Arikawa and K.P. Jantke, Eds.), Lecture Notes in Artificial Intelligence Vol. 872, pp. 100 – 105, Springer-Verlag, Berlin.

GOLD, E.M. (1967), Language identification in the limit, *Information and Control* **10**, 447 – 474.

HOPCROFT, J.E., AND ULLMAN, J.D (1969), "Formal Languages and their Relation to Automata," Addison-Wesley, Reading, Massachusetts.

KEARNS, M., AND PITT, L. (1989), A polynomial-time algorithm for learning k–variable pattern languages from examples, *in* "Proceedings 2nd Annual Workshop on Computational Learning Theory, August 1988, Boston," (D. Haussler and L. Pitt, Eds.), pp. 57 – 71, Morgan Kaufmann Publishers Inc., San Mateo.

LANGE, S., AND WIEHAGEN, R. (1991), Polynomial-time inference of arbitrary pattern languages, *New Generation Computing* **8**, 361 – 370.

LANGE, S., AND ZEUGMANN, T. (1993a), Monotonic versus non-monotonic language learning, *in* "Proceedings 2nd International Workshop on Nonmonotonic and Inductive Logic, December 1991, Reinhardsbrunn," (G. Brewka, K.P. Jantke and P.H. Schmitt, Eds.), Lecture Notes in Artificial Intelligence Vol. 659, pp. 254 – 269, Springer-Verlag, Berlin.

LANGE, S., AND ZEUGMANN, T. (1993b), Language learning in dependence on the space of hypotheses, *in* "Proceedings 6th Annual ACM Conference on Computational Learning Theory," Santa Cruz, July 1993, pp. 127 – 136, ACM Press, New York.

LANGE, S., AND ZEUGMANN, T. (1993c), Learning recursive languages with bounded mind changes, *International Journal of Foundations of Computer Science* **4**, 157 – 178.

MACHTEY, M., AND YOUNG, P. (1978) "An Introduction to the General Theory of Algorithms," North-Holland, New York.

Nix, R.P. (1983), Editing by examples, Yale University, Dept. Computer Science, Technical Report 280.

Osherson, D., Stob, M., and Weinstein, S. (1986), "Systems that Learn, An Introduction to Learning Theory for Cognitive and Computer Scientists," MIT-Press, Cambridge, Massachusetts.

Rogers, H.Jr. (1967), "Theory of Recursive Functions and Effective Computability", McGraw–Hill, New York.

Sato, M., and Umayahara, K. (1992), Inductive Inferability for formal languages from positive data, *IEICE Transactions on Information and Systems* **E–75D**, 415 – 419.

Schapire, R.E. (1990), Pattern languages are not learnable, *in* "Proceedings 3rd Annual Workshop on Computational Learning Theory", (M.A. Fulk and J. Case, Eds.), pp. 122 – 129, Morgan Kaufmann Publishers, Inc., San Mateo.

Shinohara, T. (1982), Polynomial time inference of extended regular pattern languages, *in* "Proceedings RIMS Symposia on Software Science and Engineering," Kyoto, Lecture Notes in Computer Science 147, pp. 115 – 127, Springer-Verlag, Berlin.

Trakhtenbrot, B.A., and Barzdin, J. (1970) "Конечные Автоматы (Поведение и Синтез)", Наука, Москва,
English translation: "Finite Automata–Behavior and Synthesis, Fundamental Studies in Computer Science 1", North-Holland, Amsterdam, 1973.

Wiehagen, R., and Zeugmann, T. (1994), Ignoring data may be the only way to learn efficiently, *Journal of Theoretical and Experimental Artificial Intelligence* **6**, 131 – 144.

Zeugmann, T., Lange, S., and Kapur, S. (1995), Characterizations of monotonic and dual monotonic language learning, *Information and Computation* **120**, No. 2, 155 – 173.