Learning by Erasing *

Steffen Lange	Rolf Wiehagen	Thomas Zeugmann
FB Math. & Informatik	FB Informatik	Department of
HTWK Leipzig	Universität Kaiserslautern	Informatics
PF 30066	PF 3049	Kyushu University 33
D-04251 Leipzig	D-67653 Kaiserslautern	Fukuoka 812-81
Germany	Germany	Japan

Abstract

Learning by erasing means the process of eliminating potential hypotheses from further consideration thereby converging to the least hypothesis never eliminated and this one must be a solution to the actual learning problem.

The present paper deals with learnability by erasing of indexed families of languages from both positive data as well as positive and negative data. This refers to the following scenario. A family $\mathcal{L}$ of target languages and a hypothesis space for it are specified. The learner is fed eventually all positive examples (all labeled examples) of an unknown target language L chosen from $\mathcal{L}$. The target language L is learned by erasing if the learner erases some set of possible hypotheses and the least hypothesis never erased correctly describes L.

The capabilities of learning by erasing are investigated in dependence on the requirement of what sets of hypotheses have to be or may be erased, and in dependence of the choice of the hypothesis space.

Class preserving learning by erasing ($\mathcal{L}$ has to be learned w.r.t. some suitably chosen enumeration of all and only the languages from $\mathcal{L}$), *class comprising* learning by erasing ($\mathcal{L}$ has to be learned w.r.t. some hypothesis space containing at least all the languages from $\mathcal{L}$), and *absolute* learning by erasing ($\mathcal{L}$ has to be learned w.r.t. all class preserving hypothesis spaces for $\mathcal{L}$) are distinguished.

For all these models of learning by erasing necessary and sufficient conditions for learnability are presented. A complete picture of all separations and coincidences of the learning by erasing models is derived. Learning by erasing is compared with standard models of language learning such as learning in the limit, finite learning and conservative learning The exact location of these types within the hierarchy of the learning by erasing models is established.

*A full version of this paper appeared as *Learning by Erasing*, RIFIS Technical Report RIFIS-TR-CS-122, RIFIS, Kyushu University 33, February 13, 1996; http://www.i.kyushu-u.ac.jp/ thomas/treport.html

1. Introduction

Learning by erasing means the process of eliminating potential hypotheses from further consideration thereby stabilizing to a correct hypothesis for the actual learning problem which will be never eliminated. This approach is motivated by similarities to human learning and human problem solving. In solving a problem we often find out several "non-solutions" to that problem first, contradicting the data we have or explaining them unsatisfactorily. We exclude these non-solutions from further consideration and keep only a more or less explicitly given remaining set of potential solutions. Often, at any time of the solving process, we have an actual "favored candidate" among all the remaining candidates which, though, up to now cannot be proved to be really a solution and which may change from time to time, too. Then the following can happen. Eventually we find a solution to the problem, can prove its correctness and hence successfully stop the solving process. Or, our "favored candidate" will be stable from some point on, it is really a solution, but we are not absolutely sure of that. The latter case is a version of successful learning in the limit, which is what we do in building theories or in writing computer programs.

Our main intention is a rigorous study of learning by erasing *in the limit*. A special case of our approach, the so-called co-learning, was introduced by Freivalds *et al.* [4], and further studied in [5] for learning r.e. classes of recursive functions. In that case the learner has to eliminate all but one correct hypothesis. Kummer [9] used this approach for showing that an r.e. class of recursive functions is co-learnable with respect to all of its numberings iff all of these numberings are equivalent, thus giving a learning-theoretic solution to a longstanding problem of recursion-theoretic numbering theory. We relax the all-but-one approach by giving the learner more freedom on which sets are allowed to erase eventually. We consider the following possibilities: $e\text{-}ARB$ – an arbitrary set of hypotheses may be erased, $e\text{-}MIN$ – exactly all hypotheses less than the least correct one have to be erased, $e\text{-}SUB$ – only incorrect hypotheses may be erased, $e\text{-}EQ$ – exactly all incorrect hypotheses have to be erased, $e\text{-}SUPER$ – all incorrect hypotheses have to be erased and an arbitrary set of correct hypotheses may be erased, too, $e\text{-}ALL$ – all but one hypotheses have to be erased.

Our objects to be learned are indexed families of recursive languages. We consider learning from text (positive data, only) and learning from informant (positive and negative data). We distinguish between *class preserving* learning (the hypothesis spaces exactly enumerate the family to be learned), *class comprising* learning (the hypothesis spaces enumerate a superset of the family to be learned), and *absolute* learning (the families have to be learned with respect to all class preserving hypothesis spaces). Note that the $e\text{-}ALL$-case above was already studied in [6]. Our results can be classified along the lines of characterizations, comparisons inside, and comparisons with known types of language learning.

Characterizations. For all types of learning by erasing we present char-

acterizations. In several cases the corresponding characteristic condition is a purely structural one, i.e., the language family may not contain any language together with a proper sublanguage. In other cases, for example for $e\text{-}SUB$, the characterization comes to the "granularity" of deriving characteristic learnability conditions for any given pair of a language family and a hypothesis space for it. Such granularity results were already derived in language learning theory (cf., e.g., [1], [2], [8], [16], [17]). However, our characterizations do work without the explicit use of so-called "telltales" which were commonly used in all previous characterizations in language learning. Surprisingly, up to now no such granularity results are known in Gold's [7] paradigm of learning recursive functions.

Comparisons inside. We derive a complete picture containing all separations and coincidences of the learning by erasing models defined. This picture is of a pretty regular structure. Several of these results follow from the characterizations above.

Comparisons with known types of language learning. We compare all models of learning by erasing with "standard" types of learning indexed families such as learning in the limit, finite learning and conservative learning, and determine the exact location of these learning types in the hierarchy of the learning by erasing models.

2. Notations and Definitions

Unspecified notations follow [13]. 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***. 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.

Every function $\psi \in \mathcal{P}^2$ is called a numbering. Let $\psi \in \mathcal{P}^2$, then we write ψ_j instead of $\lambda x \psi(j, x)$. By $\varphi \in \mathcal{P}^2$ we denote any fixed ***Gödel numbering*** of $\mathcal{P}$, and by $\Phi \in \mathcal{P}^2$ any associated ***complexity measure*** (cf. [3]). Let $\psi \in \mathcal{R}^2_{0,1}$, then by $L(\psi_j)$ we denote the language generated by ψ_j, i.e., $L(\psi_j) = \{x \mid \psi_j(x) = 1,\ x \in \mathbb{N}\}$, and by $co\text{-}L(\psi_j)$ its complement, i.e., $\mathbb{N} \setminus L(\psi_j)$. We call $\mathcal{L} = (L(\psi_j))_{j \in \mathbb{N}}$ an ***indexed family*** (cf. [1]). We restrict ourselves to consider exclusively indexed families of non-empty languages. An indexed family $\mathcal{L}$ is said to be ***inclusion-free*** iff $L \not\subset \hat{L}$ for all L, $\hat{L} \in range(\mathcal{L})$. Every numbering $\psi \in \mathcal{R}^2_{0,1}$ is called ***hypothesis space***. A hypothesis space ψ is said to be ***class comprising*** for $\mathcal{L}$ iff $range(\mathcal{L}) \subseteq \{L(\psi_j) \mid j \in \mathbb{N}\}$. We call a hypothesis space $\psi \in \mathcal{R}^2_{0,1}$ ***class preserving*** for $\mathcal{L}$ iff $range(\mathcal{L}) = \{L(\psi_j) \mid j \in \mathbb{N}\}$. Now, let $L \in range(\mathcal{L})$, and let ψ be any class comprising hypothesis space for $\mathcal{L}$. Then we set $min_\psi(L) = \min\{j \mid L(\psi_j) = L\}$.

Let L be a language and let $t = (s_j)_{j \in \mathbb{N}}$ be an infinite sequence of natural numbers with $content(t) = \{s_k \mid k \in \mathbb{N}\} = L$; then t is said to be a ***text*** for L.

By $text(L)$ we denote the set of all texts for L. Let t be a text, and $y \in \mathbb{N}$; then t_y denotes the initial segment of t of length $y+1$. Finally, t_y^+ denotes the ***content*** of t_y, i.e., $t_y^+ = \{s_z \mid z \leq y\}$.

We define an ***inductive inference machine*** (abbr. IIM) to be an algorithmic device working as follows: The IIM takes as its input incrementally increasing initial segments of a text and it either requests the next input, or it first outputs a hypothesis, i.e., a number, and then it requests the next input (cf. [7]). We interpret the hypotheses output by an IIM with respect to some hypothesis space $\psi \in \mathcal{R}_{0,1}^2$. If an IIM outputs j, we interpret it to mean that the IIM is hypothesizing the language $L(\psi_j)$.

Next we define an ***erasing learning machine*** (abbr. ELM) to be an algorithmic device working exactly as an IIM. However, there is a major difference in the *semantics* of the output of an IIM and an ELM, respectively. Let $\psi \in \mathcal{R}_{0,1}^2$ be any hypothesis space, and let t be a text. Suppose an ELM M has been successively fed t_y and it has output numbers $j_0, \ldots, j_z$. Then we interpret $j = \min(\mathbb{N} \setminus \{j_0, \ldots, j_z\})$ as M's ***actual guess.*** Intuitively, if an ELM outputs a number i, then it *definitely deletes* i from its list of potential hypotheses.

Let M be an IIM or an ELM, 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. The sequence $(M(t_y))_{y \in \mathbb{N}}$ is said to ***converge*** to a number j iff 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. An ELM M is said to ***stabilize*** to a number j on a text t iff its sequence of actual guesses converges to j, i.e., $j = \min(\mathbb{N} \setminus \{M(t_y) \mid y \in \mathbb{N}\})$.

Now we are ready to define learning and learning by erasing.

Definition 1 ([7]). *Let $\mathcal{L}$ be an indexed family, let L be a language, and let $\psi \in \mathcal{R}_{0,1}^2$ be a hypothesis space.* ***An IIM M CLIM–infers L from text w.r.t. 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 to j and $L = L(\psi_j)$.*

M CLIM–infers $\mathcal{L}$ w.r.t. ψ iff, for each $L \in range(\mathcal{L})$, M CLIM–infers L w.r.t. 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–infers $\mathcal{L}$ w.r.t. ψ.

By the definition of convergence, only finitely many data of L were seen by the IIM up to the (unknown) point of convergence, whenever it infers L. Hence, some form of learning must have taken place. Thus, the terms *infer*, *learn*, and *identify* are used interchangeably.

In Definition 1, *LIM* stands for "limit." 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 $\mathcal{L}$. Restricting *CLIM* to class preserving hypothesis spaces results in ***class preserving*** inference and is denoted by *LIM*. Finally, we use the prefix A to express the fact that $\mathcal{L}$ has to be inferred with

respect to *all* class preserving hypothesis spaces for it, 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 ([16]). *ALIM = LIM = CLIM.*

In general, it is not decidable whether or not an IIM M has already converged on a text t for the target language L. Adding this requirement to Definition 1 results in ***finite learning***. We denote the resulting learning type by *CFIN*. 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 $\mathcal{L}$ (cf. Proposition 2).

Proposition 2 ([17]). *AFIN = FIN = CFIN.*

Now, we define ***conservative*** IIMs. Intuitively, conservative IIMs maintain their actual hypothesis at least as long as they have received data that "provably misclassify" it.

Definition 2 ([1]). *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–infers L from text w.r.t. ψ*** *iff*

(1) *M CLIM–infers L w.r.t. ψ,*

(2) *for every text $t \in text(L)$ and for all $y, k \in \mathbb{N}$, if $M(t_y) \neq M(t_{y+k})$ then $t^+_{y+k} \not\subseteq L(\psi_{M(t_y)})$.*

M CCONSV–infers $\mathcal{L}$ w.r.t. ψ iff, for each $L \in range(\mathcal{L})$, M CCONSV–infers L w.r.t. ψ. The resulting learning type CCONSV is defined analogously as above.

Conservative learning is sensitive to the particular choice of the hypothesis space.

Proposition 3 ([16]). *ACONSV $\subset$ CONSV $\subset$ CCONSV $\subset$ ALIM.*

Next, we define learning by erasing.

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.* ***An ELM M e-CARB–infers L from text w.r.t. ψ*** *iff for every text t for L, there exists a $j \in \mathbb{N}$ with $L = L(\psi_j)$ such that M on t stabilizes to j.*

M e-CARB–infers $\mathcal{L}$ w.r.t. ψ iff, for each $L \in range(\mathcal{L})$, M e-CARB–infers L w.r.t. ψ.

Finally, let e-CARB denote the collection of all indexed families $\mathcal{L}$ for which there are an ELM M and a hypothesis space ψ such that M e-CARB–infers $\mathcal{L}$ w.r.t. ψ.

Definition 4. *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 ELM M is said to***

(A) ***e-CSUB–identify L from text w.r.t. ψ***

(B) ***e-CEQ–identify L from text w.r.t. ψ***

(C) ***e-CSUPER–identify L from text w.r.t. ψ***

(D) ***e-CALL–identify L from text w.r.t. ψ***

(E) ***e-CMIN–identify L from text w.r.t. ψ***

iff

M e-CARB–infers L from text w.r.t. ψ and the following conditions are satisfied

(A) $\{M(t_y) \mid y \in \mathbb{N}\} \subseteq \{j \mid L(\psi_j) \neq L,\ j \in \mathbb{N}\}$, *i.e., M is only allowed to erase hypotheses that are incorrect for L;*

(B) $\{M(t_y) \mid y \in \mathbb{N}\} = \{j \mid L(\psi_j) \neq L,\ j \in \mathbb{N}\}$, *i.e., M has to erase exactly all hypotheses that are incorrect for L;*

(C) $\{M(t_y) \mid y \in \mathbb{N}\} \supseteq \{j \mid L(\psi_j) \neq L,\ j \in \mathbb{N}\}$, *i.e., M has to erase all hypotheses that are incorrect for L but it may additionally erase correct hypotheses for L;*

(D) $\mathrm{card}(\mathbb{N} \setminus \{M(t_y) \mid y \in \mathbb{N}\}) = 1$, *i.e., M has to erase all but one hypothesis;*

(E) $\{M(t_y) \mid y \in \mathbb{N}\} = \{j \mid j < min_\psi(L),\ j \in \mathbb{N}\}$, *i.e., M has to erase exactly all hypotheses prior to the least correct index for L.*

We denote by *e-CSUB*, *e-CEQ*, *e-CSUPER*, *e-CALL*, and *e-CMIN* the collection of all those indexed families $\mathcal{L}$ for which there are a hypothesis space ψ and an ELM M inferring every language of it in the sense of *e-CSUB*, *e-CEQ*, *e-CSUPER*, *e-CALL*, and *e-CMIN* w.r.t. ψ, respectively.

All the types above have in common that at any step of the learning process the "favored candidate" will always be the *least* hypothesis not yet eliminated. This approach is justified by the following observations. First, by the principle of Occam's razor simple hypotheses should be "favored." Next, in case that even in the limit "many" hypotheses remain uncancelled, we get a distinguished final hypothesis, and thus one can decide from outside whether or not the learning process was successful. And finally, in case the learning machine eventually finds a provably correct hypothesis, then it can eliminate all the other hypotheses up to that one (or even all but that one) thereby making that hypothesis the least uncancelled one. Note that *e-ALL* coincides with co-learning from positive data as defined in [6]. Thus, all our definitions may be regarded as natural variations of this learning type.

3. Learning from Text

In this section, we compare the learning capabilities of all learning by erasing models from positive data to one another as well as to finite inference, learning in the limit and conservative identification from text. We analyze the power of learning by erasing in dependence on the set of admissible hypothesis spaces. Figure 1 displays the achieved separations and coincidences of the learning by

erasing models and the ordinary learning types defined. Each learning type is represented as a vertex in a directed graph. A directed edge (or path) from vertex A to vertex B indicates that A is a proper subset of B, and no edge (or path) between these vertices imply that A and B are incomparable. Finally, LT stands for ARB, SUB, EQ and $SUPER$, respectively.

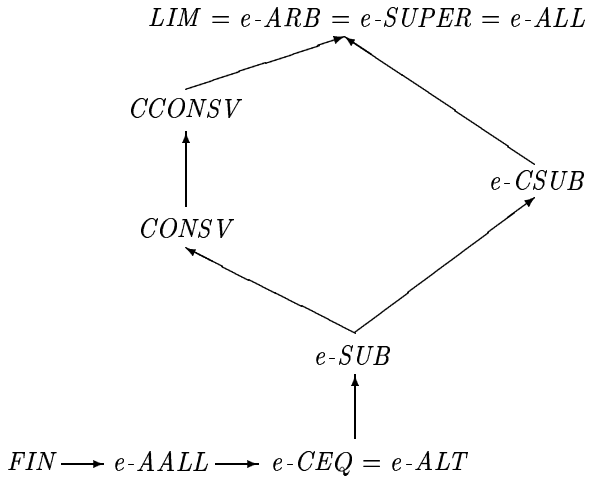

Figure 1. ***Learning by erasing versus ordinary inference***

The results displayed above are obtained via the following theorems. First, we consider class preserving hypothesis spaces. Our first theorem points to similarities and differences of the learning by erasing models defined above.

Theorem 1. (1) $FIN \subset e\text{-}EQ \subset e\text{-}SUB \subset CONSV$,

(2) *For all* $LT \in \{ARB,\ SUPER,\ ALL\}$, $e\text{-}LT = LIM$.

Theorem 1 and Proposition 3 together allow the following corollary summarizing the inclusions and equalities known so far.

Corollary 2. (1) $e\text{-}EQ \subset e\text{-}SUB \subset e\text{-}ALL = e\text{-}SUPER = e\text{-}ARB$.

(2) *For all* $LT \in \{SUPER,\ ALL,\ ARB\}$, $e\text{-}LT = e\text{-}CLT = LIM$.

Next, we study the relations between class preserving and class comprising learning for the remaining learning types. By Corollary 2, it remains to investigate the learning power of $e\text{-}CSUB$ and $e\text{-}CEQ$. The following theorem provides the desired complete picture.

Theorem 3. (1) $e\text{-}EQ = e\text{-}CEQ$,

(2) $e\text{-}SUB \subset e\text{-}CSUB \subset LIM$,

(3) $e\text{-}CSUB \ \#\ CCONSV$.

A closer look at the proof of the latter theorem clarifies that every inclusion-free indexed family is $e\text{-}EQ$–identifiable with respect to every class preserving

hypothesis space. Consequently, by Assertion (1) of Theorem 3 we immediately arrive at the following corollary.

Corollary 4. *e-AEQ = e-EQ = e-CEQ.*

Finally, we clarify the power as well as the limitations of absolute learning for the remaining learning models. Putting it all together we thus obtain all the relations displayed in Figure 1.

Theorem 5. (1) *For all* $LT \in \{ARB, SUB, SUPER\}$, $e\text{-}ALT = e\text{-}EQ$,
(2) $FIN \subset e\text{-}AALL \subset e\text{-}EQ$.

4. Learning from Informant

Now, we study learning by erasing from informant. Thus, we have to introduce some more notations and definitions. Let L be a language, and let $i = ((s_j, b_j))_{j \in \mathbb{N}}$ be any sequence of elements of $\mathbb{N} \times \{+, -\}$ such that $content(i) = \{s_k \mid k \in \mathbb{N}\} = \mathbb{N}$, $i^+ = \{s_k \mid (s_k, b_k) = (s_k, +), k \in \mathbb{N}\} = L$ and $i^- = \{s_k \mid (s_k, b_k) = (s_k, -), k \in \mathbb{N}\} = co\text{-}L$. Then we refer to i as an ***informant***. By $info(L)$ we denote the set of all informants for L. We use i_x to denote the initial segment of i of length $x + 1$, and define $i_x^+ = \{s_k \mid (s_k, +) \in i, k \leq x\}$ and $i_x^- = \{s_k \mid (s_k, -) \in i, k \leq x\}$. $CLIM.INF$ and $FIN.INF$ are defined analogously as their text counterparts by replacing everywhere text by informant. Finally, we extend all definitions of learning by erasing in the same way, and denote the resulting learning types by $e\text{-}CLT.INF$ for all $LT \in \{ARB, SUB, EQ, SUPER, MIN\}$.

Freivalds *et al.* [4] originally introduced both the learning types *e-ALL.INF* and implicitely *e-AALL.INF*, and referred to them as to *co–learning* (abbr. *co-FIN*). Furthermore, they considered the co–learnability of arbitrary recursively enumerable classes of total recursive functions. This contrasts our scenario, since we exclusively study the learnability of $\{0, 1\}$ valued functions. Nevertheless, their results easily translate into our setting. The following proposition displays the results obtained.

Proposition 4 ([4]). $FIN.INF \subseteq e\text{-}AALL.INF \subset e\text{-}ALL.INF = LIM.INF$.

Taking into account that $CLIM.INF = ALIM.INF$, one easily verifies that $e\text{-}ALL.INF = e\text{-}CALL.INF$. Moreover, Freivalds *et al.* [5] could improve Proposition 4 to $FIN.INF \subset co\text{-}FIN$ by using a deep result by Selivanov [15]. Note, however, that the separating function class is *not* $\{0, 1\}$ valued. The latter result raises two questions. First, which indexed families belong to *e-AALL.INF*, and second, whether or not $e\text{-}AALL.INF \setminus FIN.INF \neq \emptyset$, too.

The first question has been completely answered in [9] as the next proposition shows.

Proposition 5 ([9]). *Let $\mathcal{L}$ be any indexed family. Then $\mathcal{L} \in e\text{-}AALL.INF$ if and only if every class preserving hypothesis space for $\mathcal{L}$ has a recursive equality problem.*

Kummer [9] proved that every $\mathcal{L} \in e\text{-}AALL.INF$ must be discrete. An indexed family $\mathcal{L} = (L(\psi_j))_{j \in \mathbb{N}}$ is called ***discrete*** iff for every $k \in \mathbb{N}$, there is a finite function $\delta_k \subseteq \psi_k$ such that for all $j \in \mathbb{N}$, if $\delta_k \subseteq \psi_j$ then $\psi_k = \psi_j$. We refer to δ_k as to a ***separating function*** for ψ_k. $\mathcal{L} = (L(\psi_j))_{j \in \mathbb{N}}$ is said to be ***effectively discrete*** iff there exists an algorithm computing for every $k \in \mathbb{N}$ a separating function δ_k for ψ_k.

Our next theorem completely answers the second question posed above. Again, the proof is based on Selivanov's [15] result.

Theorem 6. $e\text{-}AALL \setminus FIN.INF \neq \emptyset$.

Thus, it remains to clarify the relations with respect to inclusion between the remaining learning by erasing models. This is done by the following theorem.

Theorem 7. *For all* $LT \in \{ARB, SUB, EQ, SUPER, MIN\}$,

$e\text{-}ALT.INF = e\text{-}LT.INF = e\text{-}CLT.INF = LIM.INF$.

So far we have studied separately learning from text and from informant. Now we focus our attention to the interplay between information presentation and learnability constraints. The first known result along this line of research is provided by the next proposition.

Proposition 6 ([16]). $FIN.INF \subset CONSV$.

Since $FIN.INF \subseteq e\text{-}AALL.INF$, the question arises whether or not Proposition 6 generalizes to $e\text{-}AALL.INF \subset CONSV$ or at least to $e\text{-}AALL.INF \subset LIM$. For answering it, we establish a consequence of Kummer's [9] characterization of $e\text{-}AALL.INF$.

Theorem 8. *Let* $\mathcal{L}$ *be any indexed family. If* $\mathcal{L}$ *is discrete, then* $\mathcal{L} \in LIM$.

Corollary 9. $e\text{-}AALL.INF \subset LIM$.

By Corollary 2 we may easily conclude:

Corollary 10. *For all* $LT \in \{ARB, SUPER, ALL\}$, $e\text{-}AALL.INF \subset e\text{-}LT$.

The following theorem enables us to clarify the relation between the remaining models of learning by erasing from text and informant, respectively.

Theorem 11. (1) $FIN.INF \setminus e\text{-}CSUB \neq \emptyset$,

(2) $e\text{-}EQ \setminus e\text{-}AALL.INF \neq \emptyset$.

Taking into account that $LIM \subset LIM.INF$ (cf. [7]), we directly arrive at the following corollary displaying the consequences of the latter theorem.

Corollary 12. *For all* $LT \in \{ARB, SUB, EQ, SUPER, ALL\}$ *and for all* $\lambda \in \{A, \varepsilon, C\}$, $e\text{-}\lambda LT \subset e\text{-}\lambda LT.INF$.

Putting it all together, we obtain the following Figure 2 summarizing the established relations of learning by erasing from text and informant, respectively. The semantics of Figure 2 is analogous to that of Figure 1. Let $LT \in \{ARB, SUPER, SUB, EQ, MIN\}$.

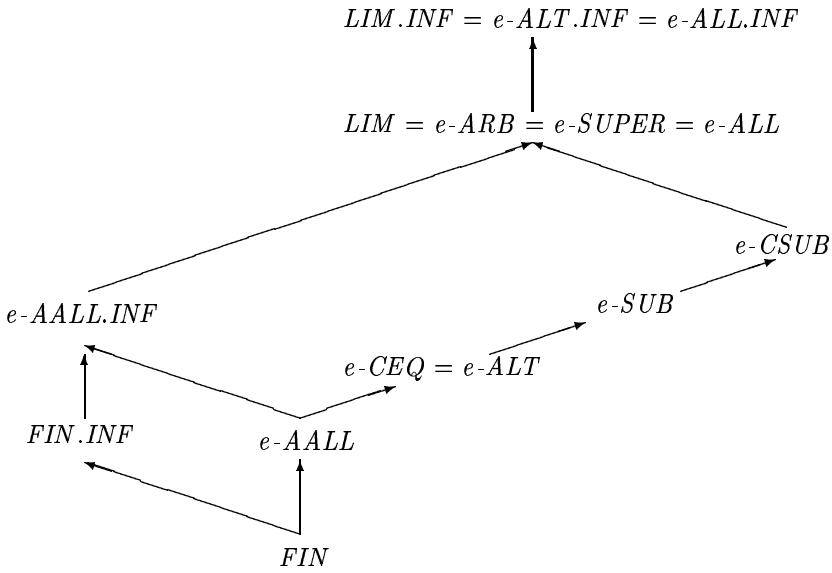

Figure 2. ***Relations between learning by erasing from text and informant***

5. Characterizations

In this section we present characterizations of all the learning by erasing models. These characterizations may help to gain a better understanding of what the defined learning models have in *common* and what their *differences* are. Our first result characterizes $e\text{-}ARB$, $e\text{-}EQ$, $e\text{-}SUB$ and $e\text{-}SUPER$ in purely structural terms.

Theorem 13. *Let $\mathcal{L}$ be any indexed family. $\mathcal{L} \in e\text{-}EQ$ if and only if $\mathcal{L}$ is inclusion-free.*

Taking Corollary 4 and Theorem 5 into consideration we may easily conclude:

Corollary 14. *Let $LT \in \{ARB, SUB, EQ, SUPER\}$, and let $\mathcal{L}$ be an indexed family. $\mathcal{L} \in e\text{-}ALT$ if and only if $\mathcal{C}$ is inclusion-free.*

For characterizing $e\text{-}AALL$ we have to combine the structural approach with the numbering theoretical one used by Kummer [9].

Theorem 15. *Let $\mathcal{L}$ be any indexed family. $\mathcal{L} \in e\text{-}AALL$ if and only if $\mathcal{L}$ is inclusion-free, and every class preserving hypothesis space for $\mathcal{L}$ has a recursive equality problem.*

Next, we characterize $e\text{-}CSUB$ and $e\text{-}SUB$. Now we derive necessary and sufficient conditions for any given pair of an indexed family and a hypothesis space for it. Again, the characterization is mainly based on the structural properties of the relevant hypothesis spaces. However, we have to add a recursive

component to these structural properties. Within the next definition we provide the necessary framework for establishing the desired characterization theorems.

Definition 5. *Let $\mathcal{L}$ be any indexed family, let $L \in range(\mathcal{L})$, and let ψ be any class comprising hypothesis space for $\mathcal{L}$. Then we set:*

(1) $Bad(\mathcal{L},\psi) = \{j \mid L \subset L(\psi_j) \text{ and } j < min_\psi(L) \text{ for some } L \in range(\mathcal{L})\}$,

(2) $Comp(\mathcal{L},\psi) = \{j \mid L(\psi_j) \notin range(\mathcal{L})\}$.

Theorem 16. *Let $\mathcal{L}$ be an indexed family. $\mathcal{L} \in e\text{-}CSUB$ if and only if there are a class comprising hypothesis space ψ for $\mathcal{L}$ and a recursively enumerable set W such that $Bad(\mathcal{L},\psi) \subseteq W \subseteq Comp(\mathcal{L},\psi)$.*

Proof. Necessity: Let $\mathcal{L} \in e\text{-}CSUB$. Hence, there are a class comprising hypothesis space $\hat{\psi}$ for $\mathcal{L}$ and an ELM $\hat{M}$ that witnesses $\mathcal{L} \in e\text{-}CSUB$ with respect to $\hat{\psi}$. We use M to define $f \in \mathcal{P}$ such that $W = range(f)$. For every $k \in \mathbb{N}$, let t^k denote the canonical text for the language $L(\psi_k)$ (cf. [17]). For every $k,\ x \in \mathbb{N}$, we set:

$$f(\langle k,x\rangle) = \begin{cases} M(t^k_x), & \text{if } content(t^k_x) \subseteq L(\psi_{M(t^k_x)}), \\ \text{not defined}, & \text{otherwise.} \end{cases}$$

Using the convention that, if M on input t^k_x does not output any hypothesis then $f(\langle k,x\rangle)$ is also not defined, we obviously have $f \in \mathcal{P}$. Next we show $Bad(\mathcal{L},\psi) \subseteq W \subseteq Comp(\mathcal{L},\psi)$.

Claim A. $W \subseteq Comp(\mathcal{L},\psi)$.

If $W = \emptyset$, we are done. Now, let $z = f(\langle k,x\rangle)$ for some $k,\ x \in \mathbb{N}$. By definition of f, we have $M(t^k_x) = z$ and $content(t^k_x) \subseteq L(\psi_z)$. Suppose, $L(\psi_z) \in range(\mathcal{L})$. Since $content(t^k_x) \subseteq L(\psi_z)$, t^k_x is an initial segment of some text $\hat{t}$ for $L(\psi_z)$. Thus M, when fed the text $\hat{t}$ for $L(\psi_z) \in range(\mathcal{L})$, outputs the correct ψ–index z for $L(\psi_z)$, a contradiction.

Claim B. $Bad(\mathcal{L},\psi) \subseteq W$.

Suppose the converse, i.e., there is a $z \in Bad(\mathcal{L},\psi) \setminus W$. Hence, $z < min_\psi(L)$ for some $L \in range(\mathcal{L})$ with $L \subset L(\psi_z)$. We distinguish the following cases.

Case 1. $L(\psi_z) \in range(\mathcal{L})$: Consider M when fed any text t for L. Because of $z < min_\psi(L)$, M eventually outputs z, say on input t_x, since otherwise it would stabilize on t to some $z' \leq z$ with $L(\psi_{z'}) \neq L$. However, since $L \subset L(\psi_z)$, the initial segment t_x may be extended to a text for $L(\psi_z)$ on which M outputs z. This is a contradiction to M $e\text{-}CSUB$–infers $\mathcal{L}$ with respect to ψ.

Case 2. $L(\psi_z) \notin range(\mathcal{L})$: Let k be any ψ–index for L. Consider M when fed the canonical text t^k for L. Since M $e\text{-}CSUB$–infers L from text, M must stabilize on t^k to $min_\psi(L)$. Because of $z < min_\psi(L)$, there has to be an $x \in \mathbb{N}$ with $M(t^k_x) = z$. Thus, $f(\langle k,x\rangle) = z$, and hence $z \in W$, again a contradiction. Claim B follows, and we are done.

Sufficiency: Let $\mathcal{L}$ be any indexed family, let ψ be a class comprising hypothesis space for $\mathcal{L}$, and let W be a recursively enumerable set with $Bad(\mathcal{L},\psi) \subseteq W \subseteq Comp(\mathcal{L},\psi)$. Let $\ell \in \mathbb{N}$ be such that $W = range(\varphi_\ell)$, and let W^z be the

elements of W, if any, enumerated after z steps of computation of φ_ℓ. We define an ELM M that witnesses $\mathcal{L} \in e\text{-}CSUB$ with respect to ψ. So, let $L \in range(\mathcal{L})$, $t \in text(L)$, and let $x \in \mathbb{N}$.

***ELM M*:** "On input t_x proceed as follows:

If $x = 0$, initialize $ToErase_0 = \emptyset$. Output nothing and request the next input.

Otherwise, test whether or not $ToErase_{x-1} = \emptyset$. In case it is, goto $(A1)$. Otherwise, goto $(A2)$.

$(A1)$ Determine the least index k that satisfies both $t_x^+ \subseteq L(\psi_k)$ and $k \notin W^x$. Update $ToErase_x = \{j \mid j < k\}$, output nothing, and request the next input.

$(A2)$ Determine $j = \min(ToErase_{x-1})$. Update $ToErase_x = ToErase_{x-1} \setminus \{j\}$, output j, and request the next input."

Since $t \in text(L)$ for some $L \in range(\mathcal{L})$, the unbounded search performed within Instruction $(A1)$ terminates for every $x \in \mathbb{N}$, and thus, M is an ELM. Let $\hat{k} = min_\psi(L)$. We show that M eventually outputs all natural numbers but $\hat{k}$.

Claim A. $\hat{k} \notin ToErase_x$ *for all* $x \in \mathbb{N}$.

Suppose the converse. Hence, there exists a least $x \in \mathbb{N}$ such that $\hat{k} \in ToErase_x$. By definition, M includes $\hat{k}$ into $ToErase_x$ iff either $t_x^+ \not\subseteq L(\psi_{\hat{k}})$ or an index $k < \hat{k}$ has been found that meets $p(k, \hat{k}) = 1$. Since $L(\psi_{\hat{k}}) = L$, $t_x^+ \not\subseteq L(\psi_{\hat{k}})$ cannot be observed. Moreover, $p(k, \hat{k}) = 1$ implies $L(\psi_k) = L$ contradicting that $\hat{k}$ is the least ψ–index for L. The claim follows.

Since M is exclusively outputting numbers $j \in ToErase_x$ for some $x \in \mathbb{N}$, the index $\hat{k}$ is never deleted. It remains to show that M eventually outputs all ψ–indices that are different from $\hat{k}$.

Claim B. *M, when successively fed t, outputs all* $k \in \mathbb{N} \setminus \{\hat{k}\}$.

We distinguish the following cases.

Case 1. $L(\psi_k) \neq L$: Since ψ is class preserving and $\mathcal{L}$ is inclusion-free, we know that $L \setminus L(\psi_k) \neq \emptyset$. Because of $t \in text(L)$, there must be a minimal y such that $t_{y+\ell}^+ \not\subseteq L(\psi_k)$ for all $\ell \in \mathbb{N}$. Thus, if $ToErase_{y-1} = \emptyset$ then $k \in ToErase_y$. Otherwise, there must be an $r \in \mathbb{N}$ such that $k \in ToErase_{y+r}$. Consequently, k is output eventually.

Case 2. $L(\psi_k) = L$: Since $\hat{k}$ is the least ψ–index of L, the inclusion-freeness of $\mathcal{L}$ implies $L(\psi_{\hat{k}}) \setminus L(\psi_j) \neq \emptyset$ for all $j < \hat{k}$. Hence, there exists a $y \in \mathbb{N}$ such that $t_y^+ \not\subseteq L(\psi_j)$ for all $j < \hat{k}$. Therefore, for all $x \geq y$ the unbounded search in Instruction (A1) terminates at $\hat{k}$. Moreover, since $L(\psi_k) = L$ we have $p(\hat{k}, k) = 1$. Consequently, there must be an $x \geq \max\{y, k\}$ such that $k \in ToErase_x$. Consequently, k is again eventually output.

Thus, Claim B follows, and the theorem is proved. q.e.d.

Theorem 17. *Let $\mathcal{L}$ be any indexed family. $\mathcal{L} \in e\text{-}SUB$ if and only if there is a class preserving hypothesis space ψ for $\mathcal{L}$ such that $Bad(\mathcal{L}, \psi) = \emptyset$.*

A closer look at the proofs above shows that, given any text t for any language L in the target class, the ELMs precisely erase all hypotheses less than the least correct one for L. Thus, we directly obtain all the remaining characterizations for learning by erasing.

Theorem 18. *For all $\lambda \in \{\varepsilon,\ A,\ C\}$, $e\text{-}\lambda MIN = e\text{-}\lambda SUB$.*

Finally, since $e\text{-}EQ = e\text{-}ASUB = e\text{-}AMIN$, we obtain the missing characterizations for the learning types $e\text{-}AMIN$, $e\text{-}MIN$ and $e\text{-}CMIN$ by Theorems 13, 17 and 16, respectively.

6. Conclusions

We defined various models of learning by erasing and related their learning power to one another as well as to learning in the limit, conservative identification, and finite inference. All but the $e\text{-}EQ$ learning model are sensitive with respect to the particular choice of the hypothesis space, thus nicely contrasting learning in the limit and finite learning. Moreover, the $e\text{-}SUB$ model is even very dependent on the set of admissible hypothesis spaces.

A further interesting aspect is provided by Theorem 1 and Corollary 2. These results show that the process of elimination cannot be restricted to *incorrect* hypotheses for achieving its full learning power. On the other hand, all learning by erasing models that are allowed to erase *correct* hypotheses, too, are as powerful as learning in the limit provided the hypothesis space is appropriately chosen (cf. Theorem 1). Thus, in order to decide whether or not a particular indexed family can be $e\text{-}LT$–learned, $LT \in \{ARB,\ SUPER,\ ALL\}$, one can apply any of the known criteria for LIM–inferability (cf., e.g., [1], [14]). These differences almost vanish if *absolute* learning is considered. Then we have a somehow opposite effect, i.e., erasing all but one guess is most restrictive with respect to the resulting learning capabilities.

The phenomena described above find their natural explanation in our characterization theorems. All models $e\text{-}ALT$ of absolute learning by erasing are constraint by the structural properties of the indexed families to be learned, i.e., they must be inclusion-free ($LT \in \{ARB,\ SUB,\ EQ,\ SUPER,\ ALL\}$), and in case of $e\text{-}AALL$, additionally, all hypothesis spaces must be equivalent with respect to reducibility.

Finally, in Section 4 we studied the problem whether or not information presentation may be traded versus learnability. The results obtained put the strength of $e\text{-}AALL.INF$ learning into the right perspective (cf. Figure 2). However, it remained open whether or not $e\text{-}AALL.INF \subset LIM$ can be sharpened to $e\text{-}AALL.INF \subset CCONSV$. Note that Theorem 8 cannot be sharpened to discreteness implies conservative learnability (cf. [10]).

7. References

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

[2] Baliga, G., Case, J., and Jain, S. (1996), Synthesizing enumeration techniques for language learning, *eCOLT*, eC-TR-96-003.

[3] Blum, M. (1967), A machine-independent theory of the complexity of recursive functions, *Journal of the ACM* **14**, 322 – 336.

[4] Freivalds, R., Karpinski, M., and Smith, C.H. (1994), Co-learning of total recursive functions, *in* "Proc. 7th Annual ACM Conference on Computational Learning Theory," pp. 190 – 197, ACM Press, New York.

[5] Freivalds, R., Gobleja, D., Karpinski, M., and Smith, C.H. (1994), Co-learnability and FIN-identifiability of enumerable classes of total recursive functions, *in* "Proc. 4th Int. Workshop on Analogical and Inductive Inference, AII'94," LNAI Vol. 872, pp. 100 – 105, Springer-Verlag, Berlin.

[6] Freivalds, R., and Zeugmann, T. (1995), Co–learning of recursive languages from positive data, RIFIS-TR-CS-110, RIFIS, Kyushu University 33.

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

[8] Kapur, S., and Bilardi, G. (1995), Language learning without overgeneralization, *Theoretical Computer Science* **141**, 151 – 162.

[9] Kummer, M. (1995), A learning-theoretic characterization of classes of recursive functions, *Information Processing Letters* **54**, 205 – 211.

[10] Lange, S., Wiehagen, R., and Zeugmann, T. (1996), Learning by erasing, RIFIS-TR-CS-122, RIFIS, Kyushu University 33.

[11] Lange, S., and Zeugmann, T. (1994), Characterization of language learning from informant under various monotonicity constraints, *Journal of Experimental & Theoretical Artificial Intelligence* **6**, 73 – 94.

[12] 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.

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

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

[15] Selivanov, V.L. (1976), Enumerations of families of general recursive functions, *Algebra and Logic* **15**, 128 – 141.

[16] Zeugmann, T., and Lange, S. (1995), A guided tour across the boundaries of learning recursive languages, *in* "Algorithmic Learning for Knowledge-Based Systems," LNAI Vol. 961, pp. 190 – 258, Springer-Verlag, Berlin.

[17] Zeugmann, T., Lange, S., and Kapur, S. (1995), Characterizations of monotonic and dual monotonic language learning, *Information and Computation* **120**, 155 – 173.