Learning from Informants: Relations between Learning Success Criteria

Martin Aschenbach, Timo Kötzing, and Karen Seidel

Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
D-14482 Potsdam

Abstract. Learning from positive and negative information, so-called *informants*, being one of the models for human and machine learning introduced by Gold [1967], is investigated. Particularly, naturally arising questions about this learning setting, originating in results on learning from solely positive information, are answered.

By a carefully arranged argument learners can be assumed to only change their hypothesis in case it is inconsistent with the data (such a learning behavior is called *conservative*). The deduced main theorem states the relations between the most important delayable learning success criteria, being the ones not ruined by a delayed in time hypothesis output.

Additionally, our investigations concerning the non-delayable requirement of consistent learning underpin the claim for *delayability* being the right structural property to gain a deeper understanding concerning the nature of learning success criteria.

In contrast to the vacillatory hierarchy for learning from solely positive information, we observe a *duality* depending on whether infinitely many *vacillations* between different (almost) correct hypotheses are still considered a successful learning behavior.

Keywords: language identification and approximation in the limit, informant learning, positive and negative data, consistent learning, delayable learning restrictions

1 Introduction

Research in the area of *inductive inference* aims at investigating the learning of formal languages and has connections to computability theory, complexity theory, cognitive science, machine learning, and more generally artificial intelligence. Setting up a classification program for deciding whether a given word belongs to a certain language can be seen as a problem in supervised machine learning, where the machine experiences labeled data about the target language. The label is 1 if the datum is contained in the language and 0 otherwise. The machine's task is to infer some rule in order to generate words in the language of interest and thereby generalize from the training samples.

According to Gold [1967] the learner is modelled by a computable function, successively receiving sequences incorporating more and more data. The source of labeled data is called an *informant*, which is supposed to be *complete in the limit*, i.e., every word in the language must occur at least once. Thereby, the learner possibly updates the current description of the target language (its hypothesis). Learning is considered successful, if after some finite time the learners' hypotheses yield good enough approximations to the target language. The original and most common learning success criterion is called **Ex**-*learning* and additionally requires that the learner eventually settles on exactly one correct hypothesis, which precisely captures the words in the language to be learned. As a single language can easily be learned, the interesting question is whether there is a learner successful on all languages in a fixed collection of languages.

Example. Consider $\mathcal{L} = \{ \mathbb{N}\backslash X \mid X \subseteq \mathbb{N} \text{ finite} \}$, the collection of all co-finite sets of natural numbers. Clearly, there is a computable function p mapping finite subsets $X \subseteq \mathbb{N}$ to $p(X)$, such that $p(X)$ encodes a program which stops if and only if the input is not in X. We call $p(X)$ an *index* for $\mathbb{N}\backslash X$. The learner is successful if for every finite $X \subseteq \mathbb{N}$ it infers $p(X)$ from a possibly very large but finite number of samples labeled according to $\mathbb{N}\backslash X$.
Regarding this example, let us assume the first two samples are $(60, 1)$ and $(2, 0)$. The first datum still leaves all options with $60 \notin X$. As the second datum tells us that $2 \in X$, we may make the learner guess $p(\{2\})$ until possibly more negative data is available. Thus, the collection of all co-finite sets of natural numbers is **Ex**-learnable from informants, simply by making the learner guess the complement of all negative information obtained so far. Since after finitely many steps all elements of the finite complement of the target language have been observed, the learner will be correct from that point onward.
It is well-known that this collection of languages cannot be learned from purely positive information. Intuitively, at any time the learner cannot distinguish the whole set of natural numbers from all other co-finite sets which contain all natural numbers presented to the learner until this point.

Learning from solely positive information, so-called *texts*, has been studied extensively, including many learning success criteria and other variations. Some results are summed up in Jain, Osherson, Royer, and Sharma [1999] and Case [2016]. We address the naturally arising question what difference it makes to learn from positive and negative information.

1.1 Our Contributions

For learning from texts there are entire maps displaying the pairwise relations of different well-known learning success criteria, see Kötzing and Palenta [2014], Kötzing and Schirneck [2016] and Jain, Kötzing, Ma, and Stephan [2016]. We give an equally informative map for **Ex**-learning from informants.
The most important requirements on the learning process when learning from informants are *conservativeness* (**Conv**), where only inconsistent hypotheses are allowed to be changed; *strong decisiveness* (**SDec**), forbidding to ever return

semantically to a withdrawn hypothesis; *strong monotonicity* (**SMon**), requiring that in every step the hypothesis incorporates the former one; *monotonicity* (**Mon**), fulfilled if in every step the set of correctly inferred words incorporates the formerly correctly guessed; *cautiousness* (**Caut**), for which never a strict subset of earlier conjectures is guessed. Lange, Zeugmann, and Kapur [1996] observed that requiring monotonicity is restrictive and that under the assumption of strong monotonicity even fewer collections of languages can be learned from informants. We complete the picture by answering the following questions regarding **Ex**-learning from informants positively:

1. Is every learnable collection of languages also learnable in a conservatively and strongly decisively way?
2. Are monotonic and cautious learning incomparable?

From positively answering the second question follow the above mentioned observations by Lange, Zeugmann, and Kapur [1996].

A diagram incorporating the resulting map is depicted in Figure 1. The complete map can be found in Figure 2.

Answering the first question builds on providing the two *normal forms* of (1) requiring learning success only on the information presented in the canonical order and (2) assuming the learner to be defined on all input sequences. Further, a regularity property borrowed from text learning plays a crucial role in the proof.

Requiring all of the learners guesses to be *consistent* with the positive and the negative information being presented to it so far makes learning harder. Next to this we also observe that the above normal forms cannot be assumed when the learner is required to act consistently. On the one hand, it is easier to find a learner for a collection of languages that consistently learns each of them only from the canonical presentation than finding one consistently learning them from arbitrary informants. On the other hand finding a total learner consistently **Ex**-learning a collection of languages is harder than finding a partial one.
We further transfer the concept of a learning success criterion to be invariant under time-delayed outputs of the hypotheses, introduced for learning from text in Kötzing and Palenta [2016] and generalized in Kötzing, Schirneck, and Seidel [2017], to the setting of learning from informants. Consistency is not *delayable* since a hypothesis which is consistent now might be inconsistent later due to new data. As this is the only requirement not being delayable, the results mentioned in the last paragraph justify the conjecture of delayability being the right property to proof more results that at once apply to all learning success criteria but consistency.

While the work of Lange and Zeugmann [1994] considers variously restricted learning of collections of recursive languages with a uniform decision procedure, the above mentioned results also apply to arbitrary collections of recursively enumerable sets. Further, our results are as strong as possible, meaning that negative results are stated for indexable families, if possible, and positive results for all collections of languages.

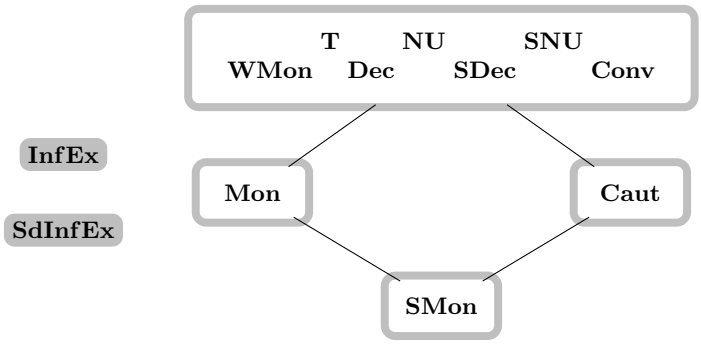

Fig. 1. Relations between delayable learning restrictions in **Ex**-learning from informants. Implications are represented as black lines from bottom to top. Two learning settings are equivalent if and only if they lie in the same grey outlined zone.

In this spirit we add to a careful investigation on how informant and text learning relate to each other by Lange and Zeugmann [1993]. We show that even for the most restrictive delayable learning success criterion when **Ex**-learning from informants there is a collection of recursive languages learnable in this setting that is not **Ex**-learnable from texts.

Case [1999] observed the vacillatory hierarchy for learning from texts. Thereby in the limit a vacillation between b many (almost) correct descriptions is allowed, where $b \in \mathbb{N}_{>0} \cup \{\infty\}$. In contrast we observe a duality by showing that, when learning from informants, requiring the learner to eventually output exactly one correct enumeration procedure is as powerful as allowing any finite number of correct descriptions in the limit. Furthermore, even facing all appropriate learning restrictions at hand gives us more learning power for $b = \infty$, known as behaviorally correct (**Bc**) learning. In particular, we obtain for all $b \in \mathbb{N}_{>0}$

$$[\mathbf{InfEx}] = \ldots = [\mathbf{InfEx}_b] = [\mathbf{InfEx}_{b+1}] = \ldots \subsetneq [\mathbf{InfBc}].$$

1.2 More Connections to Prior Research

In contrast to our observations, Angluin [1980] showed that requiring a conservative learning process is a restriction when learning from texts. Further, this is equivalent to cautious learning by Kötzing and Palenta [2016]. That monotonic learning is restrictive and incomparable to both of them in the text learning setting follows from Lange, Zeugmann, and Kapur [1996], Kinber and Stephan [1995], Jain and Sharma [1998] and Kötzing and Palenta [2016]. Further, when learning from texts, strong monotonicity is again the most restrictive assumption by Lange, Zeugmann, and Kapur [1996]. Strong decisiveness is restrictive by Baliga, Case, Merkle, Stephan, and Wiehagen [2008] and further is restricted by cautiousness/conservativeness on the one hand and monotonicity on the other

hand by Kötzing and Palenta [2016]. By the latter visualizations and a detailed discussion are provided.
When the learner does not have access to the order of presentation but knows the number of samples, the map remains the same as observed by Kötzing and Schirneck [2016].
In case the learner makes its decisions only based on the set of presented samples and ignores any information about the way it is presented, it is called *set-driven* (**Sd**). For such set-driven learners, when learning from texts, conservative, strongly decisive and cautious learning are no longer restrictive and the situation with monotonic and strong monotonic learning remains unchanged by Kinber and Stephan [1995] and Kötzing and Palenta [2016].
We observe that for delayable informant learning all three kinds of learners yield the same map. Thus, our results imply that negative information compensates for the lack of information set-driven learners have to deal with.

Gold [1967] was already interested in the above mentioned normal forms and proved that they can be assumed without loss of generality in the basic setting of pure **Ex**-learning, whereas our results apply to all delayable learning success criteria.

The name "delayability" refers to tricks in order to delay mind changes of the learner which were used to obtain polynomial computation times for the learners hypothesis updates as discussed by Pitt [1989] and Case and Kötzing [2009]. Moreover, it should not be confused with the notion of δ-delay by Akama and Zeugmann [2008], which allows satisfaction of the considered learning restriction δ steps later than in the un-δ-delayed version.

Osherson, Stob, and Weinstein [1986] analyze several restrictions for learning from informants and mention that cautious learning is a restriction to learning power; we extend this statement with our Proposition 22 in which we give one half of the answer to the second question above by providing a family of languages not cautiously but monotonically **Ex**-learnable from informants.

Furthermore, Osherson, Stob, and Weinstein [1986] consider a version of *conservativeness* where mind changes are only allowed if there is *positive* data contradicting the current hypothesis, which they claim to restrict learning power. In this paper, we stick to the more common definition of Blum and Blum [1975] and Bārzdiņš [1977], according to which mind changes are allowed also when there is negative data contradicting the current hypothesis.

1.3 Outline

In Section 2 the setting of learning from informants is formally introduced by transferring fundamental definitions and —as far as possible— observations from the setting of learning from texts. In Section 3 in order to derive the entire map of pairwise relations between delayable **Ex**-learning success criteria, normal forms

and a regularity property for such learning from informants are provided. Further, consistent learning is being investigated. In Section 4 we answer the questions above and present all pairwise relations of learning criteria in Theorem 24. In Section 5 we generalize the result by Gold [1967], we already gave a proof-sketch for, namely **Ex**-learning from texts to be harder than **Ex**-learning from informants. In Section 6 we provide the aforementioned anomalous hierarchy and vacillatory duality.

We kept every section as self-contained as possible. Unavoidably, all sections build on Section 2. Additionally, Section 4 builds on Section 3.

2 Informant Learning

We formally introduce the notion of an informant and transfer concepts and fundamental results from the setting of learning from text to learning from informant. This includes the learner itself, convergence criteria, locking sequences, learning restrictions and success criteria as well as a compact notation for comparing different learning settings. In the last subsection delayability as the central property of learning restrictions and learning success criteria is formally introduced.

As far as possible, notation and terminology on the learning theoretic side follow Jain, Osherson, Royer, and Sharma [1999], whereas on the computability theoretic side we refer to Odifreddi [1999].

We let $\mathbb{N}$ denote the *natural numbers* including 0 and write ∞ for an *infinite cardinality*. Moreover, for a function f we write $\mathrm{dom}(f)$ for its *domain* and $\mathrm{ran}(f)$ for its *range*. If we deal with (a subset of) a cartesian product, we are going to refer to the *projection functions* to the first or second coordinate by pr_1 and pr_2, respectively. For sets X, Y and $a \in \mathbb{N}$ we write $X =^a Y$, if X *equals* Y *with* a *anomalies*, i.e., $|(X \setminus Y) \cup (Y \setminus X)| \leqslant a$, where $|.|$ denotes the *cardinality function*. In this spirit we write $X =^* Y$, if there exists some $a \in \mathbb{N}$ such that $X =^a Y$. Further, $X^{<\omega}$ denotes the *finite sequence*s over X and X^ω stands for the *countably infinite sequence*s over X. Additionally, $X^{\leqslant\omega} := X^{<\omega} \cup X^\omega$ denotes the set of all *countably finite or infinite sequence*s over X. For every $f \in X^{\leqslant\omega}$ and $t \in \mathbb{N}$, we let $f[t] := \{(s, f(s)) \mid s < t\}$ denote the *restriction of* f *to* t. Finally, for sequences $\sigma, \tau \in X^{<\omega}$ their concatenation is denoted by $\sigma^\frown\tau$ and we write $\sigma \sqsubseteq \tau$, if σ is an initial segment of τ, i.e., there is some $t \in \mathbb{N}$ such that $\sigma = \tau[t]$. In our setting, we typically have $X = \mathbb{N} \times \{0,1\}$. We denote by $\mathfrak{P}$ and $\mathfrak{R}$ the set of all partial functions $f : \mathrm{dom}(f) \subseteq \mathbb{N} \times \{0,1\}^{<\omega} \to \mathbb{N}$ and total functions $f : \mathbb{N} \times \{0,1\}^{<\omega} \to \mathbb{N}$, respectively.

Let $L \subseteq \mathbb{N}$. If L is recursively enumerable, we call L a *language*. In case its characteristic function is computable, we say it is a *recursive language*. Moreover, we call $\mathcal{L} \subseteq \mathrm{Pow}(\mathbb{N})$ a *collection of (recursive) languages*, if every $L \in \mathcal{L}$ is a (recursive) language. In case there exists an enumeration $\{L_\xi \mid \xi \in \Xi\}$ of $\mathcal{L}$, where $\Xi \subseteq \mathbb{N}$ is recursive and a computable function f with $\mathrm{ran}(f) \subseteq \{0,1\}$ such

that $x \in L_\xi \Leftrightarrow f(x,\xi) = 1$ for all $\xi \in \Xi$ and $x \in \mathbb{N}$, we say $\mathcal{L}$ is an *indexable family of recursive languages*. By definition indexable families are collections of recursive languages with a uniform decision procedure.
Further, we fix a programming system φ as introduced in Royer and Case [1994]. Briefly, in the φ-system, for a natural number p, we denote by φ_p the partial computable function with program code p. We call p an *index* for $W_p := \mathrm{dom}(\varphi_p)$. For a finite set $X \subseteq \mathbb{N}$ we denote by $\mathrm{ind}(X)$ a canonical index for X. In reference to a Blum complexity measure, for all $p, t \in \mathbb{N}$, we denote by $W_p^t \subseteq W_p$ the recursive set of all natural numbers less or equal to t, on which the machine executing p halts in at most t steps. Moreover, by s-m-n we refer to a well-known recursion theoretic observation, which gives nice finite and infinite recursion theorems, like Case's Operator Recursion Theorem ORT. Finally, we let $H = \{\, p \in \mathbb{N} \mid \varphi_p(p)\downarrow \,\}$ denote the halting problem.

2.1 Informants and Learners

Intuitively, for any natural number x an *informant for a language* L answers the question whether $x \in L$ in finite time. More precisely, for every natural number x the informant I has either $(x,1)$ or $(x,0)$ in its range, where the first is interpreted as $x \in L$ and the second as $x \notin L$, respectively.

Definition 1. *(i) Let $f \in (\mathbb{N}\times\{0,1\})^{\leqslant\omega}$. We denote by*

$$\mathrm{pos}(f) := \{y \in \mathbb{N} \mid \exists x \in \mathbb{N}\colon \mathrm{pr}_1(f(x)) = y \wedge \mathrm{pr}_2(f(x)) = 1\},$$
$$\mathrm{neg}(f) := \{y \in \mathbb{N} \mid \exists x \in \mathbb{N}\colon \mathrm{pr}_1(f(x)) = y \wedge \mathrm{pr}_2(f(x)) = 0\}$$

the sets of all natural numbers, about which f gives some positive or negative information, respectively.

(ii) Let L be a language. We call every function $I : \mathbb{N} \to \mathbb{N}\times\{0,1\}$ such that $\mathrm{pos}(I) \cup \mathrm{neg}(I) = \mathbb{N}$ and $\mathrm{pos}(I) \cap \mathrm{neg}(I) = \varnothing$ an informant. *Further, we denote by* **Inf** *the set of all informants and the* set of all informants for the language L *is defined as*

$$\mathbf{Inf}(L) := \{I \in \mathbf{Inf} \mid \mathrm{pos}(I) = L\}.$$

(iii) Let I be an informant. If for every time $t \in \mathbb{N}$ reveals information about t itself, for short $\mathrm{pr}_1(I(t)) = t$, we call I a canonical informant.

It is immediate, that $\mathrm{neg}(I) = \mathbb{N}\backslash L$ for every $I \in \mathbf{Inf}(L)$. Gold [1967] referred to a canonical informant as *methodical informant*.

We employ Turings model for human computers which is the foundation of all modern computers to model the processes in human and machine learning.

Definition 2. *A* learner *is a (partial) computable function*

$$M : dom(M) \subseteq (\mathbb{N}\times\{0,1\})^{<\omega} \to \mathbb{N}.$$

The set of all partial computable functions $M : \mathrm{dom}(M) \subseteq \mathbb{N}\times\{0,1\}^{<\omega} \to \mathbb{N}$ and total computable functions $M : \mathbb{N}\times\{0,1\}^{<\omega} \to \mathbb{N}$ are denoted by $\mathcal{P}$ and $\mathcal{R}$, respectively.

2.2 Convergence Criteria and Locking Sequences

Convergence criteria tell us what quality of the approximation and syntactic accuracy of the learners' eventual hypotheses are necessary to call learning successful. Further, we proof that learning success implies the existence of sequences on which the learner is locked in a way corresponding to the convergence criterion. We will use locking sequences to show that a collection of languages cannot be learned in a certain way.

Definition 3. *Let M be a learner and $\mathcal{L}$ a collection of languages. Further, let $a \in \mathbb{N} \cup \{*\}$ and $b \in \mathbb{N}_{>0} \cup \{*, \infty\}$.*

(i) Let $L \in \mathcal{L}$ be a language and $I \in \mathbf{Inf}(L)$ an informant for L presented to M.
  (a) We call $h = (h_t)_{t \in \mathbb{N}} \in \mathbb{N}^\omega$, where $h_t := M(I[t])$ for all $t \in \mathbb{N}$, the learning sequence of M on I.
  (b) M learns L from I with a anomalies and vacillation number b in the limit*, for short M $\mathbf{Ex}_b^a$-learns L from I or $\mathbf{Ex}_b^a(M, I)$, if there is a time $t_0 \in \mathbb{N}$ such that $|\{ h_t \mid t \geqslant t_0 \}| \leqslant b$ and for all $t \geqslant t_0$ we have $W_{h_t} =^a L$.*
(ii) M learns $\mathcal{L}$ with a anomalies and vacillation number b in the limit*, for short M $\mathbf{Ex}_b^a$-learns $\mathcal{L}$, if $\mathbf{Ex}_b^a(M, I)$ for every $L \in \mathcal{L}$ and every $I \in \mathbf{Inf}(L)$.*

The intuition behind *(i)(b)* is that, sensing I, M eventually only vacillates between at most b-many hypotheses, where the case $b = *$ stands for eventually finitely many different hypotheses. In convenience with the literature, we omit the superscript 0 and the subscript 1.

Ex-learning, also known as *explanatory learning*, is the most common definition for successful learning and corresponds to the notion of identifiability in the limit by Gold [1967], where the learner eventually decides on one correct hypothesis. On the other end of the hierarchy of convergence criteria is *behaviorally correct learning*, for short **Bc**- or $\mathbf{Ex}_\infty$-learning, which only requires the learner to be eventually correct, but allows infinitely many syntactically different hypotheses in the limit. Behaviorally correct learning was introduced by Osherson and Weinstein [1982]. The general definition of $\mathbf{Ex}_b^a$-learning for $a \in \mathbb{N} \cup \{*\}$ and $b \in \mathbb{N}_{>0} \cup \{*\}$ was first mentioned by Case [1999].
In our setting, we also allow $b = \infty$ and subsume all $\mathbf{Ex}_b^a$ under the notion of a *convergence criterion*, since they determine in which semi-topological sense the learning sequence needs to have L as its limit, in order to succeed in learning L.

In the following we transfer an often employed observation by Blum and Blum [1975] to the setting of learning from informants and generalize it to all convergence criteria introduced in Definition 3.

Definition 4. *Let M be a learner, L a language and $a \in \mathbb{N} \cup \{*\}$ as well as $b \in \mathbb{N}_{>0} \cup \{*, \infty\}$. We call $\sigma \in (\mathbb{N} \times \{0,1\})^{<\omega}$ a $\mathbf{Ex}_b^a$-*locking sequence for M on L, *if* $\mathrm{Cons}(\sigma, L)$ *and*

$$\exists D \subseteq \mathbb{N} \; \big(\, |D| \leqslant b \; \wedge \; \forall \tau \in (\mathbb{N} \times \{0,1\})^{<\omega} \\ \big(\mathrm{Cons}(\tau, L) \Rightarrow \big(M(\sigma^\frown\tau){\downarrow} \wedge W_{M(\sigma^\frown\tau)} =^a L \wedge M(\sigma^\frown\tau) \in D \big) \big) \big)$$

Further, a locking sequence for M on L *is a* **Ex***-locking sequence for M on L.*

Intuitively, the learner M is locked by the sequence σ onto the language L in the sense that no presentation consistent with L can circumvent M guessing admissible approximations to L and additionally all guesses based on an extension of σ are captured by a finite set of size at most b.

Note that the definition implies $M(\sigma)\!\downarrow$, $W_{M(\sigma)} =^a L$ and $M(\sigma) \in D$.

Lemma 5. *Let M be a learner, $a \in \mathbb{N} \cup \{*\}$, $b \in \mathbb{N}_{>0} \cup \{*, \infty\}$ and L a language $\mathbf{Ex}^a_b$-identified by M. Then there is a $\mathbf{Ex}^a_b$-locking sequence for M on L.*

Proof. This is a contradictory argument. Without loss of generality M is defined on $\varnothing$. Assume towards a contradiction for every σ with $\mathrm{Cons}(\sigma, L)$, $M(\sigma)\!\downarrow$ and $W_{M(\sigma)} =^a L$ and for every finite $D \subseteq \mathbb{N}$ with at most b elements there exists a sequence $\tau^D_\sigma \in (\mathbb{N} \times \{0,1\})^{<\omega}$ with

$$\mathrm{Cons}(\tau^D_\sigma, L) \wedge \big(M(\sigma^\frown \tau^D_\sigma)\!\uparrow \vee \neg W_{M(\sigma^\frown \tau^D_\sigma)} =^a L \vee M(\sigma^\frown \tau^D_\sigma) \notin D \big).$$

Let I_L denote the canonical informant for L. We obtain an informant for L on which M does not $\mathbf{Ex}^a_b$-converge by letting

$$\begin{aligned} I &:= \bigcup_{n \in \mathbb{N}} \sigma_n, \text{ with} \\ \sigma_0 &:= I_L[1], \\ \sigma_{n+1} &:= \sigma_n^\frown \tau^{D_n}_{\sigma_n}{}^\frown I_L(n+1) \end{aligned}$$

for all $n \in \mathbb{N}$, where in $D_n := \{ M(\sigma_i^-) \mid \max\{0, n-b+1\} \leqslant i \leqslant n \}$ we collect M's at most b-many last relevant hypotheses. Since I is an informant for L by having interlaced the canonical informant for L, the learner M $\mathbf{Ex}^a_b$-converges on I. Therefore, let n_0 be such that for all t with $\sigma^-_{n_0} \sqsubseteq I[t]$ we have $h_t\!\downarrow$ and $W_{h_t} =^a L$. Then certainly $\{ M(\sigma_i^-) \mid n_0 \leqslant i \leqslant n_0 + b \}$ has cardinality $b+1$, a contradiction. ∎

Obviously, an appropriate version also holds when learning from text is considered.

2.3 Learning Success Criteria

We list the most common requirements that combined with a convergence criterion define when a learning process is considered successful. For this we first recall the notion of consistency of a sequence with a set according to Blum and Blum [1975] and Bārzdiņš [1977].

Definition 6. *Let $f \in (\mathbb{N} \times \{0,1\})^{\leqslant\omega}$ and $A \subseteq \mathbb{N}$. We define*

$$\mathrm{Cons}(f, A) \quad :\Leftrightarrow \quad \mathrm{pos}(f) \subseteq A \ \wedge \ \mathrm{neg}(f) \subseteq \mathbb{N} \backslash A$$

and say f is consistent with *A.*

The choice of learning restrictions in the following definition is justified by prior investigations of the corresponding criteria, when learning from texts, by Kötzing and Palenta [2016], Kötzing and Schirneck [2016] and Jain, Kötzing, Ma, and Stephan [2016].

Definition 7. *Let M be a learner, $I \in \mathbf{Inf}$ an informant and $h = (h_t)_{t\in\mathbb{N}} \in \mathbb{N}^\omega$ the learning sequence of M on I. We write*

(i) $\mathbf{Cons}(M, I)$ *(Angluin [1980]), if M is* consistent on I*, i.e., for all t*

$$\mathrm{Cons}(I[t], W_{h_t}).$$

(ii) $\mathbf{Conv}(M, I)$ *(Angluin [1980]), if M is* conservative on I*, i.e., for all s, t with $s \leqslant t$*

$$\mathrm{Cons}(I[t], W_{h_s}) \;\Rightarrow\; h_s = h_t.$$

(iii) $\mathbf{Dec}(M, I)$ *(Osherson, Stob, and Weinstein [1982]), if M is* decisive on I*, i.e., for all r, s, t with $r \leqslant s \leqslant t$*

$$W_{h_r} = W_{h_t} \;\Rightarrow\; W_{h_r} = W_{h_s}.$$

(iv) $\mathbf{Caut}(M, I)$ *(Osherson, Stob, and Weinstein [1986]), if M is* cautious on I*, i.e., for all s, t with $s \leqslant t$*

$$\neg W_{h_t} \subsetneq W_{h_s}.$$

(v) $\mathbf{WMon}(M, I)$ *(Jantke [1991],Wiehagen [1991]), if M is* weakly monotonic on I*, i.e., for all s, t with $s \leqslant t$*

$$\mathrm{Cons}(I[t], W_{h_s}) \;\Rightarrow\; W_{h_s} \subseteq W_{h_t}.$$

(vi) $\mathbf{Mon}(M, I)$ *(Jantke [1991],Wiehagen [1991]), if M is* monotonic on I*, i.e., for all s, t with $s \leqslant t$*

$$W_{h_s} \cap \mathrm{pos}(I) \subseteq W_{h_t} \cap \mathrm{pos}(I).$$

(vii) $\mathbf{SMon}(M, I)$ *(Jantke [1991],Wiehagen [1991]), if M is* strongly monotonic on I*, i.e., for all s, t with $s \leqslant t$*

$$W_{h_s} \subseteq W_{h_t}.$$

(viii) $\mathbf{NU}(M, I)$ *(Baliga, Case, Merkle, Stephan, and Wiehagen [2008]), if M is* non-U-shaped on I*, i.e., for all r, s, t with $r \leqslant s \leqslant t$*

$$W_{h_r} = W_{h_t} = \mathrm{pos}(I) \;\Rightarrow\; W_{h_r} = W_{h_s}.$$

(ix) $\mathbf{SNU}(M, I)$ *(Case and Moelius [2011]), if M is* strongly non-U-shaped on I*, i.e., for all r, s, t with $r \leqslant s \leqslant t$*

$$W_{h_r} = W_{h_t} = \mathrm{pos}(I) \;\Rightarrow\; h_r = h_s.$$

(x) $\mathbf{SDec}(M, I)$ *(Kötzing and Palenta [2016]), if* M *is* strongly decisive on I*, i.e., for all* r, s, t *with* $r \leqslant s \leqslant t$

$$W_{h_r} = W_{h_t} \;\Rightarrow\; h_r = h_s.$$

The following lemma states the implications between almost all of the above defined learning restrictions, which form the foundation of our research. Figure 2 includes the resulting backbone, which is slightly different from the one for learning from texts, since **WMon** does not necessarily imply **NU** in the context of learning from informants.

Lemma 8. *Let* M *be a learner and* $I \in \mathbf{Inf}$ *an informant. Then*

(i) $\mathbf{Conv}(M, I)$ *implies* $\mathbf{SNU}(M, I)$ *and* $\mathbf{WMon}(M, I)$.
(ii) $\mathbf{SDec}(M, I)$ *implies* $\mathbf{Dec}(M, I)$ *and* $\mathbf{SNU}(M, I)$.
(iii) $\mathbf{SMon}(M, I)$ *implies* $\mathbf{Caut}(M, I), \mathbf{Dec}(M, I), \mathbf{Mon}(M, I), \mathbf{WMon}(M, I)$.
(iv) $\mathbf{Dec}(M, I)$ *and* $\mathbf{SNU}(M, I)$ *imply* $\mathbf{NU}(M, I)$.
(v) $\mathbf{WMon}(M, I)$ *does* not *imply* $\mathbf{NU}(M, I)$.

Proof. Verifying the claimed implications is straightforward. In order to verify *(v)*, consider $L = 2\mathbb{N}$. Fix $p, q \in \mathbb{N}$ such that $W_p = 2\mathbb{N} \cup \{1\}$ and $W_q = 2\mathbb{N}$ and define the learner M for all $\sigma \in \mathbb{N} \times \{0,1\}^{<\omega}$ by

$$M(\sigma) = \begin{cases} p, & \text{if } 1 \in \operatorname{neg}(\sigma) \wedge 2 \notin \operatorname{pos}(\sigma); \\ q, & \text{otherwise.} \end{cases}$$

In order to prove $\mathbf{WMon}(M, I)$ for every $I \in \mathbf{Inf}(L)$, let I be an informant for L and $\mathfrak{s}_I(x) := \min\{t \in \mathbb{N} \mid \operatorname{pr}_1(I(t)) = x\}$, i.e., $\mathfrak{s}_I(1)$ and $\mathfrak{s}_I(2)$ denote the first occurance of $(1,0)$ and $(2,1)$ in $\operatorname{ran}(I)$, respectively. Then we have for all $t \in \mathbb{N}$

$$W_{h_t} = \begin{cases} 2\mathbb{N} \cup \{1\}, & \text{if } \mathfrak{s}_I(1) < t \leqslant \mathfrak{s}_I(2); \\ 2\mathbb{N}, & \text{otherwise.} \end{cases}$$

We have $W_{h_s} = W_{M(I[s])} = 2\mathbb{N} \cup \{1\}$ as well as $1 \in \operatorname{neg}(I[t])$ for all $s, t \in \mathbb{N}$ with $\mathfrak{s}_I(1) < s \leqslant \mathfrak{s}_I(2)$ and $t > \mathfrak{s}_I(2)$. Therefore, $\neg\operatorname{Cons}(I[t], W_{h_s})$ because of $\operatorname{neg}(I[t]) \nsubseteq \mathbb{N} \backslash W_{h_s}$. We obtain $\mathbf{WMon}(M, I)$ since whenever $s \leqslant t$ in $\mathbb{N}$ are such that $\operatorname{Cons}(I[t], W_{h_s})$, we know that $W_{h_s} = 2\mathbb{N} \cup \{1\}$ can only hold if likewise $\mathfrak{s}_I(1) < t \leqslant \mathfrak{s}_I(2)$ and hence $W_{h_t} = 2\mathbb{N} \cup \{1\}$, which yields $W_{h_s} \subseteq W_{h_t}$. Furthermore, if $W_{h_s} = 2\mathbb{N}$ all options for W_{h_t} satisfy $W_{h_s} \subseteq W_{h_t}$. Otherwise, in case M observes the canonical informant I for L, we have $W_{h_0} = W_{h_1} = 2\mathbb{N}$, $W_{h_2} = 2\mathbb{N} \cup \{1\}$ and $W_{h_t} = 2\mathbb{N}$ for all $t > 2$, which shows $\neg\mathbf{NU}(M, I)$. ∎

By the next definition, in order to characterize what successful learning means, we choose a convergence criterion from Definition 3 and may pose additional learning restrictions from Definition 7.

Definition 9. *Let* $\mathbf{T} := \mathfrak{P} \times \mathbf{Inf}$ *denote the whole set of pairs of possible learners and informants. We denote by*

$$\Delta := \{\, \mathbf{Caut}, \mathbf{Cons}, \mathbf{Conv}, \mathbf{Dec}, \mathbf{SDec}, \mathbf{WMon}, \mathbf{Mon}, \mathbf{SMon}, \mathbf{NU}, \mathbf{SNU}, \mathbf{T} \,\}$$

the set of admissible learning restrictions *and by*

$$\Gamma := \{\, \mathbf{Ex}_b^a \mid a \in \mathbb{N} \cup \{*\} \ \wedge \ b \in \mathbb{N}_{>0} \cup \{*, \infty\} \,\}$$

the set of convergence criteria*. Further, if*

$$\beta \in \{\, \bigcap_{i=0}^{n} \delta_i \cap \gamma \mid n \in \mathbb{N}, \forall i \leqslant n \, (\delta_i \in \Delta) \text{ and } \gamma \in \Gamma \,\} \subseteq \mathfrak{P} \times \mathbf{Inf},$$

we say that β *is a* learning success criterion*.*

Note that every convergence criterion is indeed a learning success criterion by letting $n = 0$ and $\delta_0 = \mathbf{T}$, where the latter stands for no restriction.

We refer to all $\delta \in \{\mathbf{Caut}, \mathbf{Cons}, \mathbf{Dec}, \mathbf{Mon}, \mathbf{SMon}, \mathbf{WMon}, \mathbf{NU}, \mathbf{T}\}$ also as *semantic* learning restrictions, as they allow for proper semantic convergence.

2.4 Comparing the Learning Power of Learning Settings

In order to state observations about how two ways of defining learning success relate to each other, the learning power of the different settings is encapsulated in notions $[\alpha\mathbf{Inf}\beta]$ defined as follows.

Definition 10. *Let* $\alpha \subseteq \mathcal{P}$ *be a property of partial computable functions from the set* $(\mathbb{N} \times \{0,1\})^{<\omega}$ *to* $\mathbb{N}$ *and* β *a learning success criterion. We denote by* $[\alpha\mathbf{Inf}\beta]$ *the set of all collections of languages that are* β*-learnable from informants by a learner* M *with the property* α*.*
In case the learner only needs to succeed on canonical informants, we denote the corresponding set of collections of languages by $[\alpha\mathbf{Inf}_{\text{can}}\beta]$*.*

In the learning success criterion at position β, the learning restrictions to meet are denoted in alphabetic order, followed by a convergence criterion.
At position α, we restrict the set of admissible learners by requiring for example totality. The properties stated at position α are *independent of learning success.*

For example, a collection of languages $\mathcal{L}$ lies in $[\mathcal{R}\mathbf{Inf}_{\text{can}}\mathbf{ConvSDecEx}]$ if and only if there is a total learner M conservatively, strongly decisively **Ex**-learning every $L \in \mathcal{L}$ from canonical informants. The latter means that for every canonical informant I for some $L \in \mathcal{L}$ we have $\mathbf{Conv}(M, I)$, $\mathbf{SDec}(M, I)$ and $\mathbf{Ex}(M, I)$.

Note that it is also conventional to require M's hypothesis sequence to fulfill certain learning restrictions, not asking for the success of the learning process. For instance, we are going to show that there is a collection of languages $\mathcal{L}$ such that:

- there is a learner which behaves consistently on all $L \in \mathcal{L}$ and **Ex**-learns all of them, for short $\mathcal{L} \in [\mathbf{InfConsEx}]$.
- there is no learner which **Ex**-learns every $L \in \mathcal{L}$ and behaves consistently on all languages, for short $\mathcal{L} \notin [\mathbf{ConsInfEx}]$.

The existence of $\mathcal{L}$ is implicit when writing $[\mathbf{ConsInfEx}] \subsetneq [\mathbf{InfConsEx}]$.

This notation makes it also possible to distinguish the mode of information presentation. If the learner observes the language as solely positive information, we write $[\alpha\mathbf{Txt}\beta]$ for the collections of languages β-learnable by a learner with property α from texts. Of course for α and β the original definitions for the setting of learning from texts have to be used.

2.5 Delayability

We now introduce a property of learning restrictions and learning success criteria, which allows general observations, not bound to the setting of **Ex**-learning, since it applies to all of the learning restrictions introduced in Definition 7 except consistency.

Definition 11. *Denote the set of all unbounded and non-decreasing functions by* $\mathfrak{S}$*, i.e.,* $\mathfrak{S} := \{ \mathfrak{s} : \mathbb{N} \to \mathbb{N} \mid \forall x \in \mathbb{N}\, \exists t \in \mathbb{N} \colon \mathfrak{s}(t) \geqslant x \text{ and } \forall t \in \mathbb{N} \colon \mathfrak{s}(t+1) \geqslant \mathfrak{s}(t) \}$*. Then every* $\mathfrak{s} \in \mathfrak{S}$ *is a so called* admissible simulating function*.*

A predicate $\beta \subseteq \mathfrak{P} \times \mathbf{Inf}$ *is* delayable*, if for all* $\mathfrak{s} \in \mathfrak{S}$*, all* $I, I' \in \mathbf{Inf}$ *and all partial functions* $M, M' \in \mathfrak{P}$ *holds: Whenever we have* $\operatorname{pos}(I'[t]) \supseteq \operatorname{pos}(I[\mathfrak{s}(t)])$*,* $\operatorname{neg}(I'[t]) \supseteq \operatorname{neg}(I[\mathfrak{s}(t)])$ *and* $M'(I'[t]) = M(I[\mathfrak{s}(t)])$ *for all* $t \in \mathbb{N}$*, from* $\beta(M, I)$ *we can conclude* $\beta(M', I')$*.*

The unboundedness of the simulating function guarantees $\operatorname{pos}(I) = \operatorname{pos}(I')$ and $\operatorname{neg}(I) = \operatorname{neg}(I')$.

In order to give an intuition for delayability, think of β as a learning restriction or learning success criterion and imagine M to be a learner. Then β is delayable if and only if it carries over from M together with an informant I to all learners M' and informants I' representing a delayed version of M on I. More concretely, as long as the learner M' conjectures $h_{\mathfrak{s}(t)} = M(I[\mathfrak{s}(t)])$ at time t and has, in form of $I'[t]$, at least as much data available as was used by M for this hypothesis, M' with I' is considered a delayed version of M with I.

The next result guarantees that arguing with the just defined properties covers all of the considered learning restrictions but consistency.

Lemma 12. *(i) Let* $\delta \in \Delta$ *be a learning restriction. Then* δ *is delayable if and only if* $\delta \neq \mathbf{Cons}$*.*

(ii) Every convergence criterion $\gamma \in \Gamma$ *is delayable.*

(iii) The intersection of finitely many delayable predicates on $\mathfrak{P} \times \mathbf{Inf}$ *is again delayable. Especially, every learning success criterion* $\beta = \bigcap_{i=0}^{n} \delta_i \cap \gamma$ *with* $\delta_i \in \Delta \backslash \{\mathbf{Cons}\}$ *for all* $i \leqslant n$ *and* $\gamma \in \Gamma$*,* β *is delayable.*

Proof. We approach (i) by showing, that **Cons** is not delayable. To do so, consider $\mathfrak{s} \in \mathfrak{S}$ with $\mathfrak{s}(t) := \lfloor \frac{t}{2} \rfloor$, $I, I' \in \mathbf{Inf}$ defined by $I(x) := (\lfloor \frac{x}{2} \rfloor, \mathbb{1}_{2\mathbb{N}}(\lfloor \frac{x}{2} \rfloor))$ and $I'(x) := (x, \mathbb{1}_{2\mathbb{N}}(x))$, where $\mathbb{1}_{2\mathbb{N}}$ stands for the characteristic function of all even natural numbers. By s-m-n there are learners M and M' such that for all $\sigma \in (\mathbb{N} \times \{0,1\})^{<\omega}$

$$W_{M(\sigma)} = \{x \in \mathbb{N} \mid (x \text{ even } \wedge x \leqslant \lfloor \frac{|\sigma|}{2} \rfloor) \vee (x \text{ odd } \wedge x > \lfloor \frac{|\sigma|}{2} \rfloor)\}$$
$$W_{M'(\sigma)} = \{x \in \mathbb{N} \mid (x \text{ even } \wedge x \leqslant \lfloor \frac{|\sigma|}{4} \rfloor) \vee (x \text{ odd } \wedge x > \lfloor \frac{|\sigma|}{4} \rfloor)\}.$$

Further, $\mathbf{Cons}(M, I)$ is easily verified since for all $t \in \mathbb{N}$

$$\mathrm{pos}(I[t]) = \{x \in \mathbb{N} \mid x \text{ even } \wedge x \leqslant \lfloor \frac{t-1}{2} \rfloor\} \subseteq W_{M(I[t])}$$
$$\mathrm{neg}(I[t]) = \{x \in \mathbb{N} \mid x \text{ odd } \wedge x \leqslant \lfloor \frac{t-1}{2} \rfloor)\} \subseteq \mathbb{N} \backslash W_{M(I[t])}$$

but on the other hand $\neg\mathbf{Cons}(M', I')$ since for all $t > 2$

$$\mathrm{pos}(I'[t]) = \{x \in \mathbb{N} \mid x \text{ even } \wedge x < t\}$$
$$\not\subseteq \{x \in \mathbb{N} \mid (x \text{ even } \wedge x \leqslant \lfloor \frac{t}{4} \rfloor) \vee (x \text{ odd } \wedge x > \lfloor \frac{t}{4} \rfloor)\} = W_{M'(I'[t])}.$$

The remaining proofs for (i) and (ii) are straightforward. Basically, for **Dec**, **SDec**, **SMon** and **Caut**, the simulating function $\mathfrak{s}$ being non-decreasing and $M'(I'[t]) = M(I[\mathfrak{s}(t)])$ for all $t \in \mathbb{N}$ would suffice, while for $\mathbf{NU}, \mathbf{SNU}$ and **Mon** one further needs that the informants I and I' satisfy $\mathrm{pos}(I) = \mathrm{pos}(I')$. The proof for **WMon** and **Conv** to be delayable, requires all assumptions, but $\mathfrak{s}$'s unboundedness. Last but not least, in order to prove that every convergence criterion $\gamma = \mathbf{Ex}^a_b$, for some $a \in \mathbb{N} \cup \{*\}$ and $b \in \mathbb{N}_{>0} \cup \{*, \infty\}$, carries over to delayed variants, one essentially needs both characterizing properties of $\mathfrak{s}$ and of course $M'(I'[t]) = M(I[\mathfrak{s}(t)])$. Finally, (iii) is obvious. ∎

3 Delayability vs. Consistency: Canonical Informants and Totality

In order to facilitate smooth proofs later on, we discuss normal forms for learning from informants. First, we consider the notion of set-drivenness. In Lemma 14 we show for delayable learning success criteria, that every collection of languages that is learnable from canonical informants is also learnable by a set-driven learner from arbitrary informants. By Proposition 15 this does not hold for consistent **Ex**-learning. This also implies that consistency is a restriction when learning from informants. Moreover, in Lemma 17 we observe that only considering total learners does not alter the learnability of a collection of languages in case of a delayable learning success criterion. This does not hold for consistent **Ex**-learning by Proposition 18.

3.1 Set-driven Learners and Canonical Informants

We start by formally capturing the intuition for a learner being set-driven, given in the introduction.

Definition 13 (Wexler and Culicover [1980]). *A learner M is* set-driven, *for short* $\mathbf{Sd}(M)$, *if for all* $\sigma, \tau \in \mathbb{N} \times \{0,1\}^{<\omega}$

$$(\,\mathrm{pos}(\sigma) = \mathrm{pos}(\tau) \wedge \mathrm{neg}(\sigma) = \mathrm{neg}(\tau)\,) \;\Rightarrow\; M(\sigma) = M(\tau).$$

Schäfer-Richter [1984] and Fulk [1985] showed that set-drivenness is a restriction when learning only from positive information and also the relation between the learning restrictions differ as observed by Kötzing and Palenta [2016].

In the next Lemma we observe that, by contrast, set-drivenness is not a restriction in the setting of learning from informants. Concurrently, we generalize Gold [1967]'s observation, stating that considering solely canonical informants to determine learning success does not give more learning power, to arbitrary delayable learning success criteria.

Lemma 14. *Let β be a delayable learning success criterion. Then*

$$[\mathbf{Inf}_{\mathrm{can}}\beta] = [\mathbf{SdInf}\beta].$$

Proof. Clearly, we have $[\mathbf{Inf}_{\mathrm{can}}\beta] \supseteq [\mathbf{SdInf}\beta]$. For the other inclusion, let $\mathcal{L}$ be β-learnable by a learner M from canonical informants. We proceed by formally showing that rearranging the input on the initial segment of $\mathbb{N}$, we already have complete information about at that time, is an admissible simulation in the sense of Definition 11. Let $L \in \mathcal{L}$ and $I' \in \mathbf{Inf}(L)$. For every $f \in (\mathbb{N} \times \{0,1\})^{\leqslant\omega}$, thus especially for I' and all its initial segments, we define $\mathfrak{s}_f \in \mathfrak{S}$ for all t for which $f[t]$ is defined, by

$$\mathfrak{s}_f(t) = \sup\{x \in \mathbb{N} \mid \forall w < x \colon w \in \mathrm{pos}(f[t]) \cup \mathrm{neg}(f[t])\},$$

i.e., the largest natural number x such that for all $w < x$ we know, whether $w \in \mathrm{pos}(f)$. In the following f will either be I' or one of its initial segments, which in any case ensures $\mathrm{pos}(f[t]) \subseteq L$ for all appropriate t. By construction, $\mathfrak{s}_f$ is non-decreasing and if we consider an informant I, since $\mathrm{pos}(I) \cup \mathrm{neg}(I) = \mathbb{N}$, $\mathfrak{s}_I$ is also unbounded. In order to employ the delayability of β, we define an operator $\Sigma \colon (\mathbb{N} \times \{0,1\})^{\leqslant\omega} \to (\mathbb{N} \times \{0,1\})^{\leqslant\omega}$ such that for every $f \in (\mathbb{N} \times \{0,1\})^{\leqslant\omega}$ in form of $\Sigma(f)$ we obtain a canonically sound version of f. $\Sigma(f)$ is defined on all $t < \mathfrak{s}_f(|f|)$ in case f is finite and on every $t \in \mathbb{N}$ otherwise by

$$\Sigma(f)(t) := \begin{cases} (t,0), & \text{if } (t,0) \in \mathrm{ran}(f); \\ (t,1), & \text{otherwise.} \end{cases}$$

Intuitively, in $\Sigma(f)$ we sortedly and without repetitions sum up all information contained in f up to the largest initial segment of $\mathbb{N}$, f without interruption

informs us about. For a finite sequence σ the canonical version $\Sigma(\sigma)$ has length $\mathfrak{s}_\sigma(|\sigma|)$. Now consider the learner M' defined by

$$M'(\sigma) = M(\Sigma(\sigma)).$$

Since $I := \Sigma(I')$ is a canonical informant for L, we have $\beta(M, I)$. Moreover, for all $t \in \mathbb{N}$ holds $\mathrm{pos}(I[\mathfrak{s}_{I'}(t)]) \subseteq \mathrm{pos}(I'[t])$ and $\mathrm{neg}(I[\mathfrak{s}_{I'}(t)]) \subseteq \mathrm{neg}(I'[t])$ by the definitions of $\mathfrak{s}_{I'}$ and of I using Σ. Finally,

$$M'(I'[t]) = M(\Sigma(I'[t])) = M(\Sigma(I')[\mathfrak{s}_{I'}(t)]) = M(I[\mathfrak{s}_{I'}(t)])$$

and the delayability of β yields $\beta(M', I')$. ∎

Therefore, while considering delayable learning from informants, looking only at canonical informants already yields the full picture also for set-driven learners. Clearly, the picture is also the same for so-called *partially set-driven learners* that base their hypotheses only on the set and the number of samples.

The next proposition answers the arising question, whether Lemma 14 also holds, when requiring the non-delayable learning restriction of consistency, negatively.

H denotes the halting problem.

Proposition 15. *For* $\mathcal{L} := \{2H \cup 2(H \cup \{x\}) + 1 \mid x \in \mathbb{N}\}$ *holds*

$$\mathcal{L} \in [\mathcal{R}\mathbf{Inf}_{\mathrm{can}}\mathbf{ConsConvSDecSMonEx}] \backslash [\mathbf{InfConsEx}].$$

Particularly, $[\mathbf{InfConsEx}] \subsetneq [\mathbf{Inf}_{\mathrm{can}}\mathbf{ConsEx}]$.

Proof. Let $p : \mathbb{N} \to \mathbb{N}$ be computable such that $W_{p(x)} = 2H \cup 2(H \cup \{x\}) + 1$ for every $x \in \mathbb{N}$ and let h be an index for $2H \cup 2H + 1$. Consider the total learner M defined by

$$M(\sigma) = \begin{cases} p(x), & \text{if } x \text{ with } 2x \in \mathrm{neg}(\sigma) \text{ and } 2x+1 \in \mathrm{pos}(\sigma) \text{ exists;} \\ h, & \text{otherwise} \end{cases}$$

for every $\sigma \in (\mathbb{N} \times \{0,1\})^{<\omega}$. Clearly, M conservatively, strongly decisively and strongly monotonically **Ex**-learns $\mathcal{L}$ from informants and on canonical informants for languages in $\mathcal{L}$ it is consistent.

Now, assume there is a learner M such that $\mathcal{L} \in \mathbf{InfConsEx}(M)$. By Lemma 5 there is a locking sequence σ for $2H \cup 2H + 1$. By s-m-n there is a computable function

$$\chi(x) = \begin{cases} 1, & \text{if } M(\sigma) = M(\sigma^\frown(2x+1, 1)); \\ 0, & \text{otherwise.} \end{cases}$$

By the consistency of M on $\mathcal{L}$, we immediately obtain that χ is the characteristic function for H, a contradiction. ∎

Note, that there must not be an indexable family witnessing the difference stated in the previous proposition, since every indexable family is consistently and conservatively **Ex**-learnable by enumeration.

Gold [1967] further introduces *request informants for M and L*. As the name already suggests, there is an interaction between the learner and the informant in the sense that the learner decides, about which natural number the informant should inform it next. His observation $[\mathbf{InfEx}] = [\mathbf{Inf}_{\mathrm{can}}\mathbf{Ex}] = [\mathbf{Inf}_{\mathrm{req}}\mathbf{Ex}]$ seems to hold true when facing arbitrary delayable learning success criteria, but fails in the context of the non-delayable learning restriction of consistency.

Since $\mathcal{L}$ in Proposition 15 lies in $[\mathbf{Inf}_{\mathrm{can}}\mathbf{Ex}]$, which by Lemma 14 equals $[\mathbf{InfEx}]$, we gain that for learning from informants consistent **Ex**-learning is weaker than **Ex**-learning, i.e., $[\mathbf{InfConsEx}] \subsetneq [\mathbf{InfEx}]$.

We now show that, as observed for learning from texts by Jain, Osherson, Royer, and Sharma [1999], a consistent behavior regardless learning success cannot be assumed in general, when learning from informants.

Proposition 16. *For $\mathcal{L} := \{\, \mathbb{N}, H \,\}$ holds*

$$\mathcal{L} \in [\mathcal{R}\mathbf{InfConsConvSDecEx}] \backslash [\mathbf{ConsInfEx}].$$

In particular, $[\mathbf{ConsInfEx}] \subsetneq [\mathbf{InfConsEx}]$.

Proof. Fix an index h for H and an index p for $\mathbb{N}$. The total learner M with

$$M(\sigma) = \begin{cases} p, & \text{if } \mathrm{neg}(\sigma) = \varnothing; \\ h, & \text{otherwise} \end{cases}$$

for every $\sigma \in (\mathbb{N} \times \{0,1\})^{<\omega}$ clearly consistently, conservatively and strongly decisively **Ex**-learns $\mathcal{L}$.

Aiming at the claimed proper inclusion, assume there is a consistent learner M for $\mathcal{L}$ from informants. Since M learns H, by Lemma 5, we gain a locking sequence $\sigma \in (\mathbb{N} \times \{0,1\})^{<\omega}$ for M on H, which means $\mathrm{Cons}(\sigma, H)$, $W_{M(\sigma)} = H$ and for all $\tau \in (\mathbb{N} \times \{0,1\})^{<\omega}$ with $\mathrm{Cons}(\tau, H)$ holds $M(\sigma^\frown\tau)\downarrow = M(\sigma)$. By letting

$$\chi(x) := \begin{cases} 1, & \text{if } M(\sigma^\frown(x,1)) = M(\sigma); \\ 0, & \text{otherwise} \end{cases}$$

for all $x \in \mathbb{N}$, we can decide H by the global consistency of M, a contradiction. ∎

3.2 Total Learners

Similar to full-information learning from text we show that for delayable learning restrictions totality is not a restrictive assumption.

Lemma 17. *Let β be a delayable learning success criterion. Then*

$$[\mathbf{Inf}\beta] = [\mathcal{R}\mathbf{Inf}\beta].$$

Proof. Let $\mathcal{L} \in [\mathbf{Inf}\beta]$ and M be a learner witnessing this. Without loss of generality we may assume that $\varnothing \in \mathrm{dom}(M)$. We define the total learner M' by letting $\mathfrak{s}_M : (\mathbb{N} \times \{0,1\})^{<\omega} \to \mathbb{N}$,

$$\sigma \mapsto \sup\{s \in \mathbb{N} \mid s \leqslant |\sigma| \text{ and } M \text{ halts on } \sigma[\![s]\!] \text{ after at most } |\sigma| \text{ steps}\}$$

and

$$M'(\sigma) := M(\sigma[\![\mathfrak{s}_M(\sigma)]\!]).$$

The convention $\sup(\varnothing) = 0$ yields that $\mathfrak{s}_M$ is total and it is computable, since for M only the first $|\sigma|$-many steps have to be evaluated on σ's finitely many initial segments. One could also employ a Blum complexity measure here. Hence, M' is a total computable function.
In order to observe that M' $\mathbf{Inf}\beta$-learns $\mathcal{L}$, let $L \in \mathcal{L}$ and I be an informant for L. By letting $\mathfrak{s}(t) := \mathfrak{s}_M(I[\![t]\!])$, we clearly obtain an unbounded non-decreasing function, hence $\mathfrak{s} \in \mathfrak{S}$. Moreover, for all $t \in \mathbb{N}$ from $\mathfrak{s}(t) \leqslant t$ immediately follows

$$\begin{aligned} &\mathrm{pos}(I[\![\mathfrak{s}(t)]\!]) \subseteq \mathrm{pos}(I[\![t]\!]),\ \mathrm{neg}(I[\![\mathfrak{s}(t)]\!]) \subseteq \mathrm{neg}(I[\![t]\!]) \text{ as well as} \\ &\quad M'(I[\![t]\!]) = M(I[\![\mathfrak{s}_M(I[\![t]\!])]\!]) = M(I[\![\mathfrak{s}(t)]\!]). \end{aligned}$$

By the delayability of β and with $I' = I$, we finally obtain $\beta(M', I)$. ∎

By the next proposition also for learning from informants requiring the learner to be total is a restrictive assumption for the non-delayable learning restriction of consistency. For learning from texts this was observed by Wiehagen and Zeugmann [1995] and generalized to δ-delayed consistent learning from texts by Akama and Zeugmann [2008].

Proposition 18. *There is a* collection of decidable languages *witnessing*

$$[\mathcal{R}\mathbf{InfConsEx}] \subsetneq [\mathbf{InfConsEx}].$$

Proof. Let o be an index for $\varnothing$ and define for all $\sigma \in (\mathbb{N} \times \{0,1\})^{<\omega}$ the learner M by

$$M(\sigma) := \begin{cases} o, & \text{if } \mathrm{pos}(\sigma) = \varnothing; \\ \varphi_{\max(\mathrm{pos}(\sigma))}(\langle \sigma \rangle), & \text{otherwise.} \end{cases}$$

We argue that $\mathcal{L} := \{ L \subseteq \mathbb{N} \mid L \text{ is decidable and } L \in \mathbf{InfConsEx}(M) \}$ is not consistently learnable by a total learner from informants. Assume towards a contradiction M' is such a learner. For a sequence σ of natural numbers we denote by $\overline{\sigma}$ the corresponding canonical finite informant sequence, ending with the highest value σ takes. Further, for a natural number x we denote by $\mathrm{seq}(x)$

the unique element of $(\mathbb{N}\times\{0,1\})^{<\omega}$ with $\langle \mathrm{seq}(x)\rangle = x$. Then by padded ORT there are $e, z \in \mathbb{N}$ and functions $a, b : \mathbb{N}^{<\omega} \to \mathbb{N}$, such that

$$\forall \sigma, \tau \in \mathbb{N}^{<\omega} \, (\, \sigma \sqsubsetneq \tau \Rightarrow \max\{a(\sigma), b(\sigma)\} < \min\{a(\tau), b(\tau)\} \,), \tag{1}$$

with the property that for all $\sigma \in \mathbb{N}^{<\omega}$ and all $i \in \mathbb{N}$

$$\begin{aligned}
\sigma_0 &= \varnothing; \\
\sigma_{i+1} &= \sigma_i \, ^\frown \begin{cases} a(\sigma_i), & \text{if } M'(\overline{\sigma_i{}^\frown a(\sigma_i)}) \neq M'(\overline{\sigma_i}); \\ b(\sigma_i), & \text{otherwise;} \end{cases} \\
\varphi_z(y) &= \begin{cases} 1, & \text{if } y \in \mathrm{pos}(\overline{\sigma_y}); \\ 0, & \text{otherwise;} \end{cases} \\
W_e &= \bigcup_{i\in\mathbb{N}} \mathrm{pos}(\overline{\sigma_i}); \\
\varphi_{a(\sigma)}(x) &= \begin{cases} e, & \text{if } M'(\overline{\sigma{}^\frown a(\sigma)}) \neq M'(\overline{\sigma}) \text{ and} \\ & \forall y \in \mathrm{pos}(\mathrm{seq}(x)) \, \varphi_z(y) = 1 \wedge \\ & \forall y \in \mathrm{neg}(\mathrm{seq}(x)) \, \varphi_z(y) = 0; \\ \mathrm{ind}(\mathrm{pos}(\mathrm{seq}(x))), & \text{otherwise;} \end{cases} \\
\varphi_{b(\sigma)}(x) &= \begin{cases} e, & \text{if } \forall y \in \mathrm{pos}(\mathrm{seq}(x)) \, \varphi_z(y) = 1 \wedge \\ & \forall y \in \mathrm{neg}(\mathrm{seq}(x)) \, \varphi_z(y) = 0; \\ \mathrm{ind}(\mathrm{pos}(\mathrm{seq}(x))), & \text{otherwise;} \end{cases}
\end{aligned} \tag{2}$$

Note that φ_z witnesses W_e's decidability by (1) and with this whether $\varphi_{a(\sigma)}$ and $\varphi_{b(\sigma)}$ output e or stick to p depends on $\mathrm{Cons}(\mathrm{seq}(x), W_e)$. Clearly, we have $W_e \in \mathcal{L}$ and thus M' also **InfConsEx**-learns W_e. By the **Ex**-convergence there are $e', j \in \mathbb{N}$, where j is minimal, such that $W_{e'} = W_e$ and for all $i \geqslant j$ we have $M'(\overline{\sigma_i}) = e'$ and hence $M'(\overline{\sigma_i{}^\frown a(\sigma_i)}) = M'(\overline{\sigma_i})$ by (2).
We now argue that $L := \mathrm{pos}(\overline{\sigma_j}) \cup \{a(\sigma_j)\} \in \mathcal{L}$. Let I be an informant for L and $t \in \mathbb{N}$. By (2) we observe that M is consistent on I as

$$M(I[t]) = \varphi_{\max(\mathrm{pos}(I[t]))}(\langle I[t]\rangle) = \begin{cases} e, & \text{if } \mathrm{Cons}(I[t], W_e); \\ \mathrm{ind}(\mathrm{pos}(I[t])), & \text{otherwise.} \end{cases}$$

Further, by the choice of j as well as (1) and (2) we have

$$a(\sigma_j) \notin W_e = W_{e'}, \tag{3}$$

and with this $W_{M(I[t])} = L$, if $\mathrm{pos}(I[t]) = L$.
On the other hand M' does not consistently learn L as by the choice of j we obtain $M'(\overline{\sigma_j{}^\frown a(\sigma_j)}) = M'(\overline{\sigma_j}) = e'$ and $\neg\mathrm{Cons}(\overline{\sigma_j{}^\frown a(\sigma_j)}, W_{e'})$ by (3), a contradiction. ∎

4 Relations between Delayable Learning Success Criteria

In order to reveal the relations between the delayable learning restrictions in **Ex**-learning from informants, we provide a regularity property of learners, called *syntactic decisiveness*, for **Ex**-learning in Lemma 20.
Most importantly, in Proposition 21 we acquire that conservativeness and strongly decisiveness do not restrict informant learning. After this, Propositions 23 and 22 provide that cautious and monotonic learning are incomparable, implying that both these learning settings are strictly stronger than strongly monotonic learning and strictly weaker than unrestricted learning. The overall picture is summarized in Figure 2 and stated in Theorem 24.

4.1 Syntactically Decisive Learning

A further beneficial property, requiring a learner never to *syntactically* return to an abandoned hypothesis, is supplied.

Definition 19 (Kötzing and Palenta [2016]). *Let M be a learner, L a language and I an informant for L. We write*

$\mathbf{SynDec}(M, I)$*, if M is* syntactically decisive *on I, i.e.,*

$$\forall r, s, t\colon (r \leqslant s \leqslant t \wedge h_r = h_t) \Rightarrow h_r = h_s.$$

The following easy observation shows that this variant of decisiveness can always be assumed in the setting of **Ex**-learning from informants. This is employed in the proof of our essential Proposition 21, showing that conservativeness and strong decisiveness do not restrict **Ex**-learning from informants.

Lemma 20. *We have* $[\mathbf{InfEx}] = [\mathbf{SynDecInfEx}]$.

Proof. Since obviously $[\mathbf{SynDecInfEx}] \subseteq [\mathbf{InfEx}]$, it suffices to show that every **InfEx**-learnable collection of languages is also **SynDecInfEx**-learnable. For, let $\mathcal{L} \in [\mathbf{InfEx}]$ and M witnessing this. In the definition of the learner M', we make use of a one-one computable padding function $\mathrm{pad} : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ such that $W_p = \mathrm{dom}(\varphi_p) = \mathrm{dom}(\varphi_{\mathrm{pad}(p,x)}) = W_{\mathrm{pad}(p,x)}$ for all $p, x \in \mathbb{N}$. Now, consider M' defined by

$$M'(\sigma) := \begin{cases} \mathrm{pad}(M(\sigma), |\sigma|), & \text{if } M(\sigma^-) \neq M(\sigma); \\ M'(\sigma), & \text{otherwise.} \end{cases}$$

M' behaves almost like M with the crucial difference, that whenever M performs a mind change, M' semantically guesses the same language as M did, but syntactically its hypothesis is different from all former ones. The padding function's defining property and the assumption that M **InfEx**-learns $\mathcal{L}$ immediately yield the **SynDecInfEx**-learnability of $\mathcal{L}$ by M'. ∎

Note that **SDec** implies **SynDec**, which is again a delayable learning restriction. Thus by Lemma 14, in the proof of Lemma 20 we could have also restricted our attention to canonical informants. It is further easy to see that Lemma 20 also holds for all other convergence criteria introduced and the simulation does not destroy any of the learning restrictions introduced in Definition 7.

4.2 Conservative and Strongly Decisive Learning

The following proof for **ConvSDecEx**-learning being equivalent to **Ex**-learning from informants builds on the normal forms of canonical presentations and totality provided in Section 3 as well as the regularity property introduced in the last subsection.

Proposition 21. *We have* $[\mathbf{InfEx}] = [\mathbf{InfConvSDecEx}]$.

Proof. Obviously $[\mathbf{InfEx}] \supseteq [\mathbf{InfConvSDecEx}]$ and by the Lemmas 14, 17 and 20 it suffices to show $[\mathcal{R}\mathbf{SynDecInfEx}] \subseteq [\mathbf{Inf}_{\mathrm{can}}\mathbf{ConvSDecEx}]$.
In the following for every set X and $t \in \mathbb{N}$, let $X[t]$ denote the canonical informant sequence of the first t elements of X.

Now, let $\mathcal{L} \in [\mathcal{R}\mathbf{SynDecInfEx}]$ and M a learner witnessing this. In particular, M is total and on informants for languages in $\mathcal{L}$ we have that M never returns to a withdrawn hypothesis. We want to define a learner M' which mimics the behavior of M, but modified such that, if σ is a locking sequence, then the hypothesis of M' codes the same language as the guess of M. However, if σ is not a locking sequence, then the language guessed by M' should not include data that M changes its mind on in the future. Thus, carefully in form of a recursively defined $\subseteq$-increasing sequence $(A^t_\sigma)_{t\in\mathbb{N}}$ in the guess of M' we only include the elements of the hypothesis of M that do not cause a mind change of M when looking more and more computation steps ahead. The following formal definitions make sure, this can be done in a computable way.
For every $\sigma \in (\mathbb{N}\times\{0,1\})^{<\omega}$, $t \in \mathbb{N}$ with $t \geqslant |\sigma|$ and $D \subseteq W^t_{M(\sigma)}$, we let

$$\mathbb{r}^t_\sigma(D) = \min\{|\sigma| \leqslant r \leqslant t \mid D \subseteq W^r_{M(\sigma)}\}.$$

Moreover, we define*

$$\mathcal{X}^t_\sigma(D) = \{\, X \subseteq W^t_{M(\sigma)} \mid \max(X) < \inf(W^t_{M(\sigma)}\backslash X),\, D \subsetneq X \text{ and } M(\sigma) = M(\, W^t_{M(\sigma)}[\, \mathbb{r}^t_\sigma(X)+1\,]\,)\,\}.$$

In the following we abbreviate $X \subseteq W^t_{M(\sigma)}$ and $\max(X) < \inf(W^t_{M(\sigma)}\backslash X)$ by $X \sqsubseteq W^t_{M(\sigma)}$ and say that X is an initial subset of $W^t_{M(\sigma)}$.

Aiming at providing suitable hypotheses $p(\sigma)$ for the conservative strongly decisive learner M', given σ, we carefully enumerate more and more elements included in $W_{M(\sigma)}$. We are going to start with the positive information provided

* We suppose $\inf(\varnothing) = \infty$ for convenience.

by σ. Having obtained A^t_σ with $\mathcal{X}^t_\sigma(A^t_\sigma)$ we have a set at hand that contains all initial subsets X of $W^t_{M(\sigma)}$ strictly incorporating A^t_σ, for which M does not differentiate between σ and the appropriate initial segment $W^t_{M(\sigma)}[\,\mathbb{r}^t_\sigma(X)+1\,]$ of the canonical informant of M's guess on σ. Thus $\mathcal{X}^t_\sigma(A^t_\sigma)$ contains our candidate sets for extending A^t_σ. The length $\mathbb{r}^t_\sigma(X)+1$ of the initial segment is minimal such that X is a subset of $W^{\mathbb{r}^t_\sigma(X)}_{M(\sigma)}$ and at least $|\sigma|$ to assure **Ex**-convergence of the new learner.
For an arbitrary $\sigma \in (\mathbb{N}\times\{0,1\})^{<\omega}$ this reads as follows

$$A^0_\sigma = \mathrm{pos}(\sigma);$$
$$\forall t \in \mathbb{N} : A^{t+1}_\sigma = \begin{cases} W^t_{M(\sigma)}, & \text{if } \mathrm{neg}(\sigma) \cap A^t_\sigma \neq \emptyset; \\ \max_\subseteq \mathcal{X}^t_\sigma(A^t_\sigma), & \text{else if } \mathcal{X}^t_\sigma(A^t_\sigma) \neq \emptyset; \\ A^t_\sigma, & \text{otherwise.} \end{cases}$$

Furthermore, using s-m-n, we define $p : (\mathbb{N}\times\{0,1\})^{<\omega} \to \mathbb{N}$ as a one-one function, such that for all $\sigma \in (\mathbb{N}\times\{0,1\})^{<\omega}$

$$W_{p(\sigma)} = \bigcup_{t\in\mathbb{N}} A^t_\sigma. \tag{4}$$

In the following, for all $\tau \in (\mathbb{N}\times\{0,1\})^{<\omega}$ we denote by τ' the largest initial segment of τ for which $M'(\tau') = M'(\tau)$, i.e., the last time M' performed a mind change. Finally, we define our new learner M' by

$$M'(\sigma) = \begin{cases} p(\sigma), & \text{if } |\sigma| = 0; \\ p(\sigma), & \text{else if } M((\sigma^-)') \neq M(\sigma) \wedge \neg\mathrm{Cons}(\sigma, A^{|\sigma|}_{(\sigma^-)'}); \\ M'(\sigma^-), & \text{otherwise.} \end{cases}$$

That is, M' follows the mind changes of M once a suitably inconsistent hypothesis has been seen. All hypotheses of M are poisoned in a way to ensure that we can decide inconsistency.

Let us first observe that M' **Ex**-learns every $L \in \mathbf{InfEx}(M)$ from informants. For, let t_0 be minimal such that, for all $t \geqslant t_0$, $M(L[t]) = M(L[t_0])$. Thus, $e := M(L[t_0])$ is a correct hypothesis for L.

If M' does not make a mind change in or after t_0, then M' converged already before that mind change of M. Thus, let $s_0 < t_0$ be minimal such that for all $t \geqslant s_0$, $e' := M'(L[s_0]) = M'(L[t])$. As p is one-one and M learns syntactically decisive, we have $M(L[s_0]) \neq M(L[t])$ for all $t \geqslant t_0$. From $(L[t-1])' = L[s_0]$ and the definition of M' we get $\mathrm{Cons}(L[t], A^t_{L[s_0]})$ for all $t \geqslant t_0$. Thus, $W_{e'} = L$, because the final hypothesis $W_{e'}$ of M' contains all elements of L and no other by Equation (4).

In case M' makes a mind change in or after t_0, let $t_1 \geqslant t_0$ be the time of that mind change. As M does not perform mind changes after t_0, the learner M' cannot make further mind changes and therefore converges to $e' := p(L[t_1])$.

By construction we have $A^t_{L[t_1]} \subseteq W_e = L$ for all $t \in \mathbb{N}$ and with it $W_{e'} \subseteq L$ by Equation 4. Towards a contradiction, suppose $W_{e'} \subsetneq L$ and let $x \in L \backslash W_{e'}$ be minimal. By letting s_0 such that $\mathrm{pos}(L[x]) \subseteq A^{s_0}_{L[t_1]}$ and $x \in W^{s_0}_e$, every initial subset of $W^{s_0}_e$ extending $A^{s_0}_{L[t_1]}$ would necessarily contain x. Therefore we have $A^s_{L[t_1]} = A^{s_0}_{L[t_1]}$ and $\mathcal{X}^s_{L[t_1]}(A^{s_0}_{L[t_1]}) = \varnothing$ for all $s \geqslant s_0$. We obtain the **Ex**-convergence of M' by constructing $s_2 \geqslant s_0$ with $\mathcal{X}^{s_2}_{L[t_1]}(A^{s_0}_{L[t_1]}) \neq \varnothing$. For this, let $y := \max(A^{s_0}_{L[t_1]} \cup \{x\})$ which implies $A^{s_0}_{L[t_1]} \subsetneq \mathrm{pos}(L[y+1])$. Moreover, let $s_1 \geqslant t_1$ be large enough such that $L[y+1] = W^{s_1}_e[y+1]$. Thus, by letting $r := \mathbb{r}^{s_1}_{L[t_1]}(\mathrm{pos}(L[y+1])) + 1$ we gain $r = \mathbb{r}^s_{L[t_1]}(\mathrm{pos}(L[y+1])) + 1$ for all $s \geqslant s_1$, where the latter denotes the time window considered in the third requirement for $\mathrm{pos}(L[y+1]) \in \mathcal{X}^s_{L[t_1]}(A^{s_0}_{L[t_1]})$. Furthermore, let $s_2 \geqslant s_1$ with $L[r] = W^{s_2}_e[r]$. By the definition of r we have $r > t_1 \geqslant t_0$ and gain

$$\mathrm{pos}(L[y+1]) \sqsubseteq W^{s_2}_e, \quad A^{s_2}_{L[t_1]} = A^{s_0}_{L[t_1]} \subsetneq \mathrm{pos}(L[y+1]) \quad \text{and}$$
$$M(L[t_1]) = e = M(L[r]) = M(\, W^{s_2}_e[r]\,),$$

for short $\mathrm{pos}(L[y+1]) \in \mathcal{X}^{s_2}_{L[t_1]}(A^{s_0}_{L[t_1]})$, implying $\mathcal{X}^{s_2}_{L[t_1]}(A^{s_0}_{L[t_1]}) \neq \varnothing$.

Now we come to prove that M' is conservative on every $L \in \mathbf{InfEx}(M)$. For, let t be such that $M'(L[t]) \neq M'(L[t+1])$. Let $e' := M'(L[t])$ and let $t' \leqslant t$ be minimal such that $M'(L[t']) = e'$. From the mind change of M' we get $\neg\mathrm{Cons}(L[t+1], A^{t+1}_{L[t']})$. In case it holds $\mathrm{neg}(L[t+1]) \cap A^{t+1}_{L[t']} \neq \varnothing$, since $A^{t+1}_{L[t']} \subseteq W_{e'}$, we would immediately observe $\neg\mathrm{Cons}(L[t+1], W_{e'})$. Therefore, we may assume $\mathrm{pos}(L[t+1]) \backslash A^{t+1}_{L[t']} \neq \varnothing$. Suppose, by way of contradiction, $W_{e'}$ is consistent with $L[t+1]$, i.e., $\mathrm{pos}(L[t+1]) \subseteq W_{e'}$ and $\mathrm{neg}(L[t+1]) \cap W_{e'} = \varnothing$. Then we have $\mathrm{neg}(L[t+1]) \cap A^s_{L[t']} = \varnothing$ for all $s \in \mathbb{N}$. Since $\mathrm{pos}(L[t+1]) \subseteq W_{e'}$, there is t_0 minimal such that

$$L[t+1] = A^{t_0+1}_{L[t']}[t+1]. \tag{5}$$

We have $\mathrm{neg}(L[t']) \cap A^{t_0}_{L[t']} = \varnothing$ as otherwise $\neg\mathrm{Cons}(L[t+1], W_{e'})$. Because t_0 was minimal, we have $A^{t_0}_{L[t']} \subsetneq A^{t_0+1}_{L[t']}$ and with this $A^{t_0+1}_{L[t']} \in \mathcal{X}^{t_0}_{L[t']}(A^{t_0}_{L[t']})$ by the definition of $A^{t_0+1}_{L[t']}$. In particular, this tells us

$$A^{t_0+1}_{L[t']} \sqsubseteq W^{t_0}_{M(L[t'])} \quad \text{and} \tag{6}$$
$$M(L[t']) = M(\, W^{t_0}_{M(L[t'])}[\, \mathbb{r}^{t_0}_{L[t']}(A^{t_0+1}_{L[t']}) + 1\,]\,). \tag{7}$$

and therefore with

$$L[t'] \sqsubseteq L[t+1] \overset{(5)}{\sqsubseteq} A^{t_0+1}_{L[t']} \overset{(6)}{\sqsubseteq} W^{t_0}_{M(L[t'])}[\, \mathbb{r}^{t_0}_{L[t']}(A^{t_0+1}_{L[t']}) + 1\,]$$

by Equation (7) and M's syntactic decisivenes we get $M(L[t']) = M(L[t+1])$. Therefore, M' did not make a mind change in $t+1$, a contradiction. ∎

4.3 Completing the Picture of Delayable Learning

The next two propositions show that monotonic and cautious **Ex**-learning are incomparable on the level of indexable families. With Proposition 21 this yields all relations between delayable **Ex**-learning success criteria as stated in Theorem 24.

We extend the observation of Osherson, Stob, and Weinstein [1986] for cautious learning to restrict learning power with the following result. The positive part has already been discussed in the example in the introduction.

Proposition 22. *For the* indexable family $\mathcal{L} := \{\mathbb{N}\backslash X \mid X \subseteq \mathbb{N} \text{ finite}\}$ *holds*

$$\mathcal{L} \in [\mathbf{InfMonEx}]\backslash[\mathbf{InfCautBc}].$$

Particularly, $[\mathbf{InfCautEx}] \subsetneq [\mathbf{InfEx}]$.

Proof. In order to approach $\mathcal{L} \notin [\mathbf{InfCautBc}]$, let M be a **InfBc**-learner for $\mathcal{L}$ and I_0 the canonical informant for $\mathbb{N}$. Moreover, let t_0 be such that $W_{M(I_0[t_0])} = \mathbb{N}$. Let I_1 be the canonical informant for $L_1 := \mathbb{N}\backslash\{t_0\}$. Since M learns L_1, there is $t_1 > t_0$ such that $W_{M(I_1[t_1])} = L_1$. We have $I_1[t_0] = I_0[t_0]$ and hence M is not cautiously learning L_1 from I_1.

We now show the **MonEx**-learnability. By s-m-n there is a computable function $p : \mathbb{N} \to \mathbb{N}$ such that for all finite sets X holds $W_{p(\langle X\rangle)} = \mathbb{N}\backslash X$, where $\langle X\rangle$ denotes a canonical code for X as already employed in the proof of Proposition 23. We define the learner M by letting for all $\sigma \in \mathbb{N}\times\{0,1\}^{<\omega}$

$$M(\sigma) = p(\langle \mathrm{neg}(\sigma)\rangle).$$

The corresponding intuition is that M includes every natural number in its guess, not explicitly excluded by σ. Clearly, M learns $\mathcal{L}$ and behaves monotonically on $\mathcal{L}$, since for every $X \subseteq \mathbb{N}$ finite, every informant I for $\mathbb{N}\backslash X$ and every $t \in \mathbb{N}$, we have $W_{M(I[t])} \supseteq \mathbb{N}\backslash X$ and therefore $W_{M(I[t])} \cap \mathbb{N}\backslash X = \mathbb{N}\backslash X$. ∎

This reproves $[\mathbf{InfSMonEx}] \subsetneq [\mathbf{InfMonEx}]$ observed by Lange, Zeugmann, and Kapur [1996] also on the level of indexable families.

In the next proposition the learner can even be assumed cautious on languages it does not identify. Thus, according to Definition 10 we write this success independent property of the learner on the left side of the mode of presentation.

Proposition 23. *For the* indexable family

$$\mathcal{L} := \{2X \cup (2(\mathbb{N}\backslash X) + 1) \mid X \subseteq \mathbb{N} \text{ finite or } X = \mathbb{N}\}$$

holds $\mathcal{L} \in [\mathbf{CautInfEx}]\backslash[\mathbf{InfMonBc}]$.
Particularly, $[\mathbf{InfMonEx}] \subsetneq [\mathbf{InfEx}]$.

Proof. We first show $\mathcal{L} \notin [\mathbf{InfMonBc}]$. Let M be a $\mathbf{InfBc}$-learner for $\mathcal{L}$. Further, let I_0 be the canonical informant for $L_0 := 2\mathbb{N} \in \mathcal{L}$. Then there exists t_0 such that $W_{M(I_0[2t_0])} = 2\mathbb{N}$. Moreover, consider the canonical informant I_1 for

$$L_1 := 2\{0, \ldots, t_0\} \cup (2(\mathbb{N} \backslash \{0, \ldots, t_0\}) + 1) \in \mathcal{L}$$

and let $t_1 > t_0$ such that $W_{M(I_1[2t_1])} = L_1$. Similarly, we let I_2 be the canonical informant for

$$L_2 := 2\{0, \ldots, t_0, t_1 + 1\} \cup (2(\mathbb{N} \backslash \{0, \ldots, t_0, t_1 + 1\}) + 1) \in \mathcal{L}$$

and choose $t_2 > t_1$ with $W_{M(I_2[2t_2])} = L_2$. Since $2(t_1 + 1) \in (L_0 \cap L_2) \backslash L_1$ and by construction $I_2[2t_0] = I_0[2t_0]$ as well as $I_2[2t_1] = I_1[2t_1]$, we obtain

$$2(t_1 + 1) \in W_{M(I_2[2t_0])} \cap L_2 \quad \text{and} \quad 2(t_1 + 1) \notin W_{M(I_2[2t_1])} \cap L_2$$

and therefore M does not learn L_2 monotonically from I_2.

Let us now adress $\mathcal{L} \in [\mathbf{CautInfEx}]$. Fix $p \in \mathbb{N}$ such that $W_p = 2\mathbb{N}$. Further, by s-m-n there is a computable function $q : \mathbb{N} \to \mathbb{N}$ with $W_{q(\langle X \rangle)} = X \cup (2\mathbb{N} \backslash X) + 1$, where $\langle X \rangle$ stands for a canonical code of the finite set X. We define the learner M for all $\sigma \in \mathbb{N} \times \{0, 1\}^{<\omega}$ by

$$M(\sigma) = \begin{cases} p, & \text{if } \mathrm{pos}(\sigma) \subseteq 2\mathbb{N}; \\ q(\langle \mathrm{pos}(\sigma) \cap 2\mathbb{N} \rangle), & \text{otherwise.} \end{cases}$$

Intuitively, M guesses $2\mathbb{N}$ as long as no odd number is known to be in the language L to be learned. If for sure $L \neq 2\mathbb{N}$, then M assumes that all even numbers known to be in L so far are the only even numbers therein.
It is easy to verify that M is computable and by construction it learns $\mathcal{L}$. For establishing the cautiousness, let L be any language, I an informant for L and $s \leqslant t$. Furthermore, assume $W_{M(I[s])} \neq W_{M(I[t])}$. In case $\mathrm{pos}(I[s]) \nsubseteq 2\mathbb{N}$, we have $x \in (\mathrm{pos}(I[t]) \cap 2\mathbb{N})$ with $x \notin (\mathrm{pos}(I[s]) \cap 2\mathbb{N})$ and therefore as desired $W_{M(I[t])} \backslash W_{M(I[s])} \neq \varnothing$. Then $\mathrm{pos}(I[s]) \subseteq 2\mathbb{N}$ implies $W_{M(I[s])} = 2\mathbb{N}$ and thus again $W_{M(I[t])} \backslash W_{M(I[s])} \neq \varnothing$. ∎

We sum up the preceding results in the next theorem and also represent them in Figure 2.

Theorem 24. *We have*

(i) $\forall \delta \in \{\mathbf{Conv}, \mathbf{Dec}, \mathbf{SDec}, \mathbf{WMon}, \mathbf{NU}, \mathbf{SNU}\} : [\mathbf{Inf}\delta\mathbf{Ex}] = [\mathbf{InfEx}]$.
(ii) $[\mathbf{InfMonEx}] \perp [\mathbf{InfCautEx}]$.

Proof. The first part is an immediate consequence of Proposition 21 and so is the second part of the Propositions 22 and 23. ∎

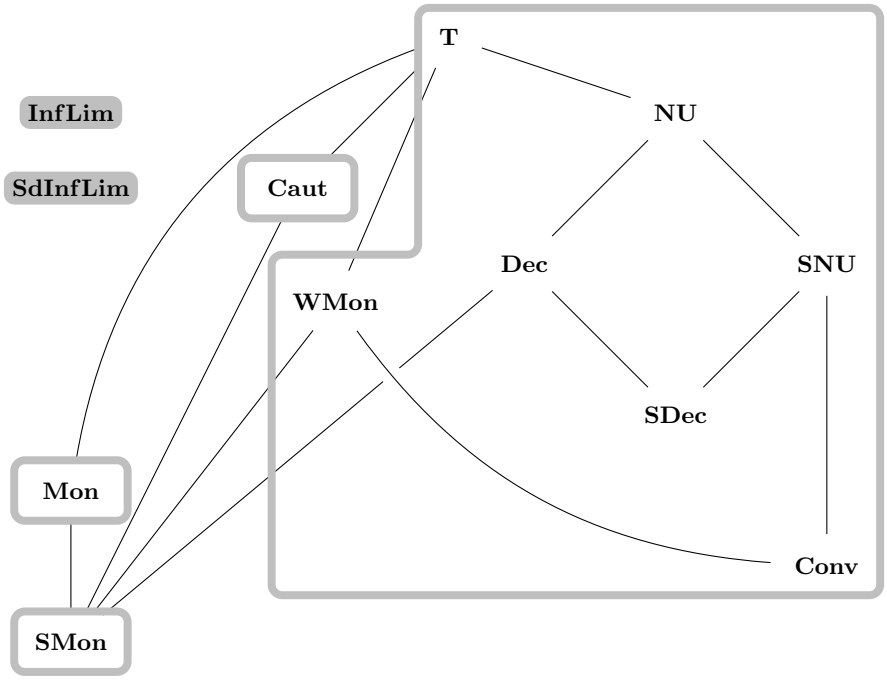

Fig. 2. Relations between delayable learning restrictions in full-information (explanatory) **Ex**-learning of languages from informants. The implications according to Lemma 8 are represented as black lines from bottom to top. Two learning settings are equivalent if and only if they lie in the same grey outlined zone as stated in Theorem 24.

5 Outperforming Learning from Texts

Already Gold [1967] observed [**TxtEx**] $\subsetneq$ [**InfEx**] and later on Lange and Zeugmann [1993] further investigated the interdependencies when considering the different monotonicity learning restrictions. For instance, they showed that there exists an indexable family $\mathcal{L} \in$ [**InfMonEx**]\[**TxtEx**] $\neq \varnothing$ and in contrast that for indexable families **InfSMonEx**-learnability implies **TxtEx**-learnability. We show that this inclusion fails on the level of families of recursive languages even with all learning restrictions at hand.

Proposition 25. *For the class of recursive languages*

$$\mathcal{L} := \{2(L \cup \{x\}) \cup 2L + 1 \mid L \text{ is recursive} \wedge W_{\min(L)} = L \wedge x \geqslant \min(L)\}$$

holds $\mathcal{L} \in$ [**InfConvSDecSMonEx**]\[**TxtEx**].

Proof. Let p_m denote an index for $2W_m \cup 2W_m + 1$ and $p_{m,x}$ an index for $2(W_m \cup \{x\}) \cup 2W_m + 1$. The learner M will look for the minimum of the presented set and moreover try to detect the exception x, in case it exists. Thus,

it checks for all m such that $2m \in \mathrm{pos}(\sigma)$ or $2m+1 \in \mathrm{pos}(\sigma)$ whether for all $k < m$ holds $2k \in \mathrm{neg}(\sigma)$ or $2k+1 \in \mathrm{neg}(\sigma)$. In case m has this property relative to σ, we write $\min_L(m, \sigma)$ as m is the minimum of the language presented. Further, M tries to find x such that $2x \in \mathrm{pos}(\sigma)$ and $2x+1 \in \mathrm{neg}(\sigma)$ and we abbreviate by $\mathrm{exc}_L(x, \sigma)$ that x is such an exception. Consider the learner M for all $\sigma \in (\mathbb{N} \times \{0,1\})^{<\omega}$ defined by

$$M(\sigma) = \begin{cases} \mathrm{ind}(\varnothing), & \text{if there is no } m \text{ with } \min_L(m, \sigma); \\ p_m, & \text{if } \min_L(m, \sigma) \text{ and there is no } x \text{ with } \mathrm{exc}_L(x, \sigma); \\ p_{m,x}, & \text{if } \min_L(m, \sigma) \text{ and } x \text{ is minimal with } \mathrm{exc}_L(x, \sigma). \end{cases}$$

Clearly, M conservatively, strongly decisively and strongly monotonically **Ex**-learns $\mathcal{L}$.
To observe $\mathcal{L} \notin [\mathbf{TxtEx}]$, assume there exists M such that $\mathcal{L} \in \mathbf{TxtEx}(M)$. By s-m-n there exists $e \in \mathbb{N}$ such that for all $i \in \mathbb{N}$

$$A_\sigma(i) = \{\, k \in \mathbb{N} \mid M(\sigma) \neq M(\sigma^\frown(2e+4i)^k) \,\};$$
$$B_\sigma(i) = \{\, k \in \mathbb{N} \mid M(\sigma) \neq M(\sigma^\frown(2e+4i+2)^k) \,\};$$
$$\sigma_0 = (2e, 2e+1);$$
$$\sigma_{i+1} = \begin{cases} \sigma_i, & \text{if } A_{\sigma_i}(i) = B_{\sigma_i}(i) = \varnothing \\ & \text{or } i > 0 \wedge \sigma_{i-1} = \sigma_i; \\ \sigma_i^\frown(2e+4i)^{\inf(A_{\sigma_i}(i))\frown}(2e+4i+1), & \text{if } A_{\sigma_i}(i) \neq \varnothing \\ & \wedge \inf(A_{\sigma_i}(i)) \leqslant \inf(B_{\sigma_i}(i))); \\ \sigma_i^\frown(2e+4i+2)^{\inf(B_{\sigma_i}(i))\frown}(2e+4i+3), & \text{if } B_{\sigma_i}(i) \neq \varnothing \\ & \wedge \inf(B_{\sigma_i}(i)) < \inf(A_{\sigma_i}(i)); \end{cases}$$
$$W_e = \bigcup_{i \in \mathbb{N}} \{n \mid 2n+1 \in \mathrm{ran}(\sigma_i)\}.$$

W_e is recursive, because it is either finite or we can decide it along the construction of the σ_i. Thus, $2W_e \cup 2W_e + 1 \in \mathcal{L}$. If for some index i holds $\sigma_{i+1} = \sigma_i$, then M fails to learn $2(W_e \cup \{e+2i\}) \cup 2W_e + 1$ or $2(W_e \cup \{e+2i+1\}) \cup 2W_e + 1$. On the other hand, if there is no such i, by letting $T := \bigcup_{i \in \mathbb{N}} \sigma_i$ we obtain a text for $2W_e \cup 2W_e + 1$, on which M performs infinitely many mindchanges. ∎

6 Vacillatory Duality

We compare the convergence criteria $\mathbf{Ex}_b^a$ from Definition 3 for different parameters $a \in \mathbb{N} \cup \{*\}$ and $b \in \mathbb{N}_{>0} \cup \{*, \infty\}$. The duality depending on whether $b = \infty$ for fixed a follow from the Propositions 26 and 27.

We separate **InfEx**- and **InfBc**-learning at the level of families of recursive languages, even when requiring the **Bc**-learning sequence to meet all introduced delayable semantic learning restrictions. As every indexable family of recursive languages is **Ex**-learnable from informants by enumeration, the result is optimal.

Proposition 26. *For the collection of recursive languages*

$$\mathcal{L} = \{L \cup \{x\} \mid L \subseteq \mathbb{N} \text{ is recursive} \wedge W_{\min(L)} = L \wedge x \geqslant \min(L)\}$$

holds $\mathcal{L} \in [\mathbf{InfSMonBc}] \backslash [\mathbf{InfEx}]$.

Proof. By Lemma 14 it suffices to show

$$\mathcal{L} \in [\mathbf{Inf}_{\mathrm{can}}\mathbf{SMonBc}] \backslash [\mathbf{Inf}_{\mathrm{can}}\mathbf{Ex}].$$

By s-m-n there are $p : \mathbb{N} \times \{0,1\}^{<\omega} \times \mathbb{N} \to \mathbb{N}$ and a learner M such that for all $\sigma \in \mathbb{N} \times \{0,1\}^{<\omega}$ and $x \in \mathbb{N}$

$$W_{p(\sigma,x)} = W_{\min(\mathrm{pos}(\sigma))} \cup \{x\} \text{ and}$$

$$M(\sigma) = \begin{cases} o, & \text{if } \mathrm{pos}(\sigma) = \varnothing; \\ \min(\mathrm{pos}(\sigma)), & \text{else if } \mathrm{pos}(\sigma) \backslash W^{|\sigma|}_{\min(\mathrm{pos}(\sigma))} = \varnothing; \\ p(\sigma, x), & \text{else if } x = \min(\mathrm{pos}(\sigma) \backslash W^{|\sigma|}_{\min(\mathrm{pos}(\sigma))}); \end{cases}$$

where o refers to the canonical index for the empty set. Let $L \cup \{x\} \in \mathcal{L}$ with $L \subseteq \mathbb{N}$ recursive, $W_{\min(L)} = L$ and $x \geqslant \min(L)$ and let I be the canonical informant for $L \cup \{x\}$. Then for all $t > \min(L)$ we have $W_{\min(\mathrm{pos}(I[t]))} = W_{\min(L)} = L$. Further, let m be minimal such that $\{y \in L \mid y < x\} \subseteq W^m_{\min(L)}$. Since $x \geqslant \min(L)$ the construction yields for all $t \in \mathbb{N}$

$$W_{h_t} = \begin{cases} \varnothing, & \text{if } t \leqslant \min(L); \\ L, & \text{else if } \min(L) \leqslant t < \max\{x+1, m\}; \\ L \cup \{x\}, & \text{otherwise.} \end{cases}$$

This can be easily verified, since in case $y \in L$ we have $L = L \cup \{y\}$ and establishes the $\mathbf{Inf}_{\mathrm{can}}\mathbf{SMonBc}$-learnability of $\mathcal{L}$ by M.

In order to approach $\mathcal{L} \notin [\mathbf{Inf}_{\mathrm{can}}\mathbf{Ex}]$, assume to the contrary that there is a learner M that $\mathbf{Inf}_{\mathrm{can}}\mathbf{Ex}$-learns $\mathcal{L}$. By Lemma 17 M can be assumed total. We are going to define a recursive language L with $W_{\min(L)} = L$ helpful for showing that not all of $\mathcal{L}$ is $\mathbf{Inf}_{\mathrm{can}}\mathbf{Ex}$-learned by M. In order to do so, for every canonical $\sigma \in \mathbb{N} \times \{0,1\}^{<\omega}$ we define sets $A^0_\sigma, A^1_\sigma \subseteq \mathbb{N}$. For this let I^0_σ stand for the canonical informant of $\mathrm{pos}(\sigma)$, whereas I^1_σ denotes the canonical informant of $\mathrm{pos}(\sigma) \cup \{|\sigma|\}$. In A^0_σ we collect all $t > |\sigma|$ for which M's hypothesis on $I^0_\sigma[t]$ is different from $M(\sigma)$. Similarly, in A^1_σ we capture all $t > |\sigma|$ such that M on $I^1_\sigma[t]$ makes a guess different from $M(\sigma)$. This reads as follows

$$A^0_\sigma := \{\, t \in \mathbb{N} \mid t > |\sigma| \wedge M(I^0_\sigma[t]) \neq M(\sigma) \,\},$$
$$A^1_\sigma := \{\, t \in \mathbb{N} \mid t > |\sigma| \wedge M(I^1_\sigma[t]) \neq M(\sigma) \,\}.$$

Note that for every $t > |\sigma|$

$$I^0_\sigma[t] = \sigma^\frown(\,(|\sigma|, 0), (|\sigma|+1, 0), \ldots, (t-1, 0)\,),$$
$$I^1_\sigma[t] = \sigma^\frown(\,(|\sigma|, 1), (|\sigma|+1, 0), \ldots, (t-1, 0)\,).$$

By s-m-n there exists $p \in \mathbb{N}$ such that[⋆⋆]

$$\sigma_0 = (\,(0,0),\ldots,(p-1,0),(p,1)\,),$$
$$\forall i \in \mathbb{N}\colon \sigma_{i+1} = \begin{cases} \sigma_i, & \text{if } A^0_{\sigma_i} = A^1_{\sigma_i} = \varnothing; \\ I^0_{\sigma_i}[\min(A^0_{\sigma_i})], & \text{if } \inf(A^0_{\sigma_i}) \leqslant \inf(A^1_{\sigma_i}); \\ I^1_{\sigma_i}[\min(A^1_{\sigma_i})], & \text{otherwise;} \end{cases}$$
$$W_p = \bigcup_{i\in\mathbb{N}} \operatorname{pos}(\sigma_i).$$

By construction $p = \min(W_p)$ and W_p is recursive, which immediately yields $L := W_p \in \mathcal{L}$. Further, for every $i \in \mathbb{N}$ from $\sigma_i \neq \sigma_{i+1}$ follows $M(\sigma_i) \neq M(\sigma_{i+1})$. Aiming at a contradiction, let I be the canonical informant for L, which implies $\bigcup_{i\in\mathbb{N}} \sigma_i \sqsubseteq I$. Since M **Ex**-learns L and thus does not make infinitely many mind changes on I, there exists $i_0 \in \mathbb{N}$ such that for all $i \geqslant i_0$ we have $\sigma_i = \sigma_{i_0}$. But then for all $t > |\sigma_{i_0}|$ holds

$$M(I^0_{\sigma_{i_0}}[t]) = M(\sigma_{i_0}) = M(I^1_{\sigma_{i_0}}[t]),$$

thus M does not learn at least one of $L = \operatorname{pos}(\sigma_{i_0})$ and $L \cup \{|\sigma_{i_0}|\}$ from their canonical informants. On the other hand both of them lie in $\mathcal{L}$ and therefore, M had not existed in the beginning. ▪

Since allowing infinitely many different correct hypotheses in the limit gives more learning power, the question arises, whether finitely many hypotheses already allow to learn more collections of languages. The following proposition shows that, as observed by Bārzdiņš and Podnieks [1973] and Case and Smith [1983] for function learning, the hierarchy of vacillatory learning collapses when learning languages from informants.

Proposition 27. *Let $a \in \mathbb{N} \cup \{*\}$. Then* $[\mathbf{InfEx}^a] = [\mathbf{InfEx}^a_*]$.

Proof. Clearly, $[\mathbf{InfEx}^a] \subseteq [\mathbf{InfEx}^a_*]$. For the other inclusion let $\mathcal{L}$ be in $[\mathbf{InfEx}^a_*]$ and M a learner witnessing this. By Lemma 17 we assume that M is total. In the construction of the $\mathbf{Ex}^a$-learner M', we employ the recursive function $\Xi : (\mathbb{N}\times\{0,1\})^{<\omega} \times \mathbb{N} \to \mathbb{N}$, which given $\sigma \in (\mathbb{N}\times\{0,1\})^{<\omega}$ and $p \in \mathbb{N}$ alters p such that $W^{|\sigma|}_{\Xi(\sigma,p)} \cap \operatorname{neg}(\sigma) = \varnothing$ and moreover, if $\sigma \sqsubseteq \tau$ are such that $W^{|\sigma|}_p \cap \operatorname{neg}(\sigma) = W^{|\tau|}_p \cap \operatorname{neg}(\tau)$, then $\Xi(\sigma,p) = \Xi(\tau,p)$. One way to do this is by letting $\Xi(\sigma,p)$ denote the unique program, which given x successively checks, whether $x = y_i$, where $(y_i)_{i<|\operatorname{neg}(\sigma)|}$ is the increasing enumeration of $\operatorname{neg}(\sigma)$. As soon as the answer is positive, the program goes into a loop. Otherwise it executes the program encoded in p on x, which yields

$$\varphi_{\Xi(\sigma,p)}(x) = \begin{cases} \uparrow, & \text{if } x \in \operatorname{neg}(\sigma); \\ \varphi_p(x), & \text{otherwise.} \end{cases}$$

[⋆⋆] Again we use the convention $\inf(\varnothing) = \infty$.

Now, M' works as follows:

I. Compute $p_i := M(\sigma[i])$ for all $i \leqslant |\sigma|$.
II. Withdraw all p_i with the property $|\mathrm{neg}(\sigma) \cap W_{p_i}^{|\sigma|}| > a$.
III. Define $M'(\sigma)$ to be a code for the program coresponding to the union vote of all $\Xi(\sigma, p_i)$, for which p_i was not withdrawn in the previous step:
Given input x, for n from 0 till ∞ do the following: If $i := \pi_1(n) \leqslant |\sigma|$, $|\mathrm{neg}(\sigma) \cap W_{p_i}^{|\sigma|}| \leqslant a$ and $\Phi_{\Xi(\sigma,p_i)}(x) \leqslant \pi_2(n)$, then return 0; otherwise increment n.
This guarantees

$$\varphi_{M'(\sigma)}(x) = \begin{cases} 0, & \text{if } \exists\, i \leqslant |\sigma| \,(\, |\mathrm{neg}(\sigma) \cap W_{p_i}^{|\sigma|}| \leqslant a \;\wedge\; \varphi_{\Xi(\sigma,p_i)}(x){\downarrow}\,); \\ \uparrow, & \text{otherwise.} \end{cases}$$

Intuitively, $M'(\sigma)$ eliminates all commission errors in guesses of M on initial segments of σ, not immediately violating the allowed number of anomalies, and then asks whether one of them converges on the input, which implies

$$W_{M'(\sigma)} = \bigcup_{i \leqslant |\sigma|, |\mathrm{neg}(\sigma) \cap W_{p_i}^{|\sigma|}| \leqslant a} W_{\Xi(\sigma, M(\sigma[i]))}.$$

In order to show $\mathcal{L} \in \mathbf{InfEx}^a(M)$, let $L \in \mathcal{L}$ and $I \in \mathbf{Inf}(L)$. As $\mathcal{L} \in \mathbf{Ex}_*^a(M)$, there is t_0 such that all of M's hypotheses are in $\{h_s \mid s \leqslant t_0\}$ and additionally $|\, W_{h_s}^{t_0} \cap \mathbb{N}\backslash L\,| > a$ for all $s \leqslant t_0$ with $|\, W_{h_s} \cap \mathbb{N}\backslash L| > a$. Moreover, we can assume that for all $s \leqslant t_0$ with $|\, W_{h_s} \cap \mathbb{N}\backslash L\,| \leqslant a$ we have observed all commission errors in at most t_0 steps, which formally reads as $W_{h_s} \cap \mathbb{N}\backslash L = W_{h_s}^{t_0} \cap \mathbb{N}\backslash L$.
Then for all $t \geqslant t_0$ we obtain the same set of indices

$$A := \{\, \Xi(I[t], p_i) \mid i \leqslant t \wedge |\mathrm{neg}(I[t]) \cap W_{p_i}^t| \leqslant a \,\}$$

and therefore M' will return syntactically the same hypothesis, namely, h'_{t_0}.
It remains to argue for $W_{h'_{t_0}} =^a L$. By construction and the choice of t_0 there are no commission errors, i.e., $W_{h'_{t_0}} \cap \mathbb{N}\backslash L = \varnothing$. Further, since $\varphi_{h'_{t_0}}(x)$ exists in case there is at least one $p \in A$ such that $\varphi_p(x)$ exists, there are at most a arguments, on which $\varphi_{h'_{t_0}}$ is undefined. ∎

This contrasts the results in language learning from texts by Case [1999], observing for every $a \in \mathbb{N} \cup \{*\}$ a hierarchy

$$\begin{aligned}[\mathbf{TxtEx}^a] \subsetneq \ldots \subsetneq [\mathbf{TxtEx}_b^a] \subsetneq [\mathbf{TxtEx}_{b+1}^a] \subsetneq \ldots \\ \subsetneq \bigcup_{b \in \mathbb{N}_{>0}} [\mathbf{TxtEx}_b^a] \subsetneq [\mathbf{TxtEx}_*^a] \subseteq [\mathbf{TxtBc}^a].\end{aligned}$$

For learning from informants we gain for every $a \in \mathbb{N} \cup \{*\}$ a duality

$$\begin{aligned}[\mathbf{InfEx}^a] = \ldots = [\mathbf{InfEx}_b^a] = [\mathbf{InfEx}_{b+1}^a] = \ldots \\ = \bigcup_{b \in \mathbb{N}_{>0}} [\mathbf{InfEx}_b^a] = [\mathbf{InfEx}_*^a] \subsetneq [\mathbf{InfBc}^a].\end{aligned}$$

7 Learning Characteristic Functions of Collections of Recursive Languages

We now turn to the setting in which we want to learn a set of Boolean classifiers. In Machine Learning the input is usually considered a labeled element of $\mathbb{R}^d$. It is reasonable to consider only the countably many d-tuples x of computable reals $\mathbb{R}^d_{\text{comp}}$. By fixing a (non-computable) enumeration $\mathbb{R}^d_{\text{comp}} = \langle x_i \mid i < \mathbb{N} \rangle$, we might as a first attempt identify i with x_i. Then our hypothesis space is the set of all Boolean functions. We will later restrict ourselves to total computable Boolean functions.

Definitions 1 for informant and 2 for the learner are independent of the interpretation of the hypothesis. The Definition 3 of convergence criteria has to be slightly modified as follows.

Definition 28. *Let M be a learner and $\mathcal{L}$ a collection of recursive languages. Further, let $a \in \mathbb{N} \cup \{*\}$ and $b \in \mathbb{N}_{>0} \cup \{*, \infty\}$.*

(i) Let $L \in \mathcal{L}$ be a language and $I \in \mathbf{Inf}(L)$ an informant for L presented to M.
  (a) We call $h = (h_t)_{t \in \mathbb{N}} \in \mathbb{N}^\omega$, where $h_t := M(I[t])$ for all $t \in \mathbb{N}$, the learning sequence of M on I.
  (b) M learns L from I with a anomalies and vacillation number b in the limit*, for short M $\mathbf{Ex}^a_{Cb}$-learns L from I or $\mathbf{Ex}^a_{Cb}(M, I)$, if there is a time $t_0 \in \mathbb{N}$ such that $|\{ h_t \mid t \geqslant t_0 \}| \leqslant b$ and for all $t \geqslant t_0$ we have $\mathrm{Diff}_L(h_t) = \{x \in \mathbb{N} \mid \varphi_{h_t}(x) \neq \chi_L(x)\}$ has at most size a.*
(ii) M learns $\mathcal{L}$ with a anomalies and vacillation number b in the limit*, for short M $\mathbf{Ex}^a_{Cb}$-learns $\mathcal{L}$, if $\mathbf{Ex}^a_{Cb}(M, I)$ for every $L \in \mathcal{L}$ and every $I \in \mathbf{Inf}(L)$.*

We also have to adjust the Definition 4 of locking sequences.

Definition 29. *Let M be a learner, L a language and $a \in \mathbb{N} \cup \{*\}$ as well as $b \in \mathbb{N}_{>0} \cup \{*, \infty\}$. We call $\sigma \in (\mathbb{N} \times \{0,1\})^{<\omega}$ an* $\mathbf{Ex}^a_{Cb}$-locking sequence for M on L, *if* $\mathrm{Cons}(\sigma, L)$ *and*

$$\exists D \subseteq \mathbb{N} \; \big(\; |D| \leqslant b \; \wedge \; \forall \tau \in (\mathbb{N} \times \{0,1\})^{<\omega} \\ \big(\mathrm{Cons}(\tau, L) \Rightarrow \big(M(\sigma^\frown\tau)\!\downarrow \wedge |\mathrm{Diff}_L(M(\sigma^\frown\tau))| \leqslant a \wedge M(\sigma^\frown\tau) \in D \big) \big) \big)$$

Then the proof of Lemma 5 immediately transfers and we obtain the following lemma.

Lemma 30. *Let M be a learner, $a \in \mathbb{N} \cup \{*\}$, $b \in \mathbb{N}_{>0} \cup \{*, \infty\}$ and L a language $\mathbf{Ex}^a_{Cb}$-identified by M. Then there is a $\mathbf{Ex}^a_{Cb}$-locking sequence for M on L.*

We also have to adjust the Definition 6 of consistency in the following way.

Definition 31. *Let φ be a Boolean computable function. We define*

$$\operatorname{pos}(\varphi) = \{x \in \mathbb{N} \mid \varphi(x)\downarrow = 1\};$$
$$\operatorname{neg}(\varphi) = \{x \in \mathbb{N} \mid \varphi(x)\downarrow = 0\}.$$

Let $f \in (\mathbb{N} \times \{0,1\})^{\leqslant \omega}$. We say f is consistent with φ*, for short* $\operatorname{Cons}(f, \varphi)$*, if*

$$\operatorname{pos}(f) \subseteq \operatorname{pos}(\varphi) \ \wedge \ \operatorname{neg}(f) \subseteq \operatorname{neg}(\varphi).$$

Let C_{h_t} denote $\operatorname{pos}(\varphi_{h_t})$. By replacing W_{h_t} by C_{h_t}, the definitions of the learning restrictions in Definition 7, learning success criteria in Definition 9 and learning criteria in Definition 10 remain the same. The implications (independent of the learning success criterion at hand) between the delayable learning restrictions as stated in Lemma 8 hold accordingly.

Moreover, the Definition 11 and basic Lemma 12 concerning delayability remain unchanged. Also Lemma 14 about considering canonical informants being sufficient and Lemma 17 about totality being no restriction for delayable learning success criteria still hold as the proofs only refer to the abstract concept of delayability.

To our knowledge Machine Learning algorithms only hypothesize total classifiers. Denote the set of encoded programs for total Boolean functions on $\mathbb{N}$ by $C\mathbf{Ind}$. Then we will from now on only allow the learner M to hypothesize elements of $C\mathbf{Ind}$ on data consistent with some classifier to be learned. We denote by $[\mathbf{Inf}C\mathbf{IndEx}_C]$ the collection of all recursive languages $\mathbf{Ex}_C$-learnable by such a learner M from informants. In Definition 9 in the learning success criterion at position β, we write $C\mathbf{Ind}$ between the learning restrictions to be met and the convergence criterion.
With $[C\mathbf{IndInfEx}_C]$ we refer to the collection of all recursive languages $\mathbf{Ex}_C$-learnable by a learner with range contained in $C\mathbf{Ind}$. These learners only output hypotheses for total computable Boolean functions and in Definition 9 we write $C\mathbf{Ind}$ as part of α.
Later we might consider appropriately chosen subsets of $C\mathbf{Ind}$ as hypothesis space.

In this setting we can assume the learner to output only hypotheses consistent with the input on relevant data. This is done by patching the hypothesis according to the finitely many training data points the learner has received so far.

Proposition 32. *We have*

$$[\mathbf{Inf}_{\text{can}}C\mathbf{IndEx}] = [\mathbf{SdInfCons}C\mathbf{IndEx}].$$

Proof. We use the idea from Lemma 14. Thus, the new learner outputs M's hypothesis h on the largest complete canonical informant with information only from the current input σ. As h is an index for a total function, we can, in a uniformly computable way, obtain a hypothesis h_σ from h such that

(i) φ_{h_σ} is consistent with all data in σ and
(ii) $h_\sigma = h$ if σ is consistent with φ_h.

More precisely, the computable operator maps an index h of a computable function $\varphi_h : \mathbb{N} \to \{0,1\}$ and a finite informant sequence σ to an index h_σ of a computable function φ_{h_σ} with

$$\varphi_{h_\sigma}(x) = \begin{cases} 1, & \text{if } x \in \operatorname{pos}(\sigma); \\ 0, & \text{else if } x \in \operatorname{cnt}(\sigma); \\ \varphi_h(x), & \text{otherwise.} \end{cases}$$

The simulation only requires information about $\operatorname{cnt}(\sigma) = \operatorname{pos}(\sigma) \cup \operatorname{neg}(\sigma)$ and thus the learner is set-driven. Further, $h_\sigma = h$ whenever φ_h is consistent with σ. As M converges on the canonical informant and we only alter h in case at least one datum in σ is inconsistent with φ_h, we obtain the convergence of the new learner. Clearly, it is consistent by construction. ∎

Summing up, as consistency of the input data with a hypothesized total computable Boolean functions is computable, C**Ind**-learners can be assumed consistent while learning. By the same argument $\tau(C\mathbf{Ind})$-learners can be assumed $\tau(\mathbf{Cons})$.

It is easy to see that **Ex** can be replaced by every convergence criterion (and also **Mon**).

On the other hand, it is easy to adapt the proof of Proposition 18 as follows.

Proposition 33. *There is a* collection of decidable languages *witnessing*

$$[\mathcal{R}\mathbf{InfCons}C\mathbf{IndEx}_C] \subsetneq [\mathbf{InfCons}C\mathbf{IndEx}_C].$$

Proof. Let o be an index for the everywhere 0-function. Further, define for all $\sigma \in (\mathbb{N} \times \{0,1\})^{<\omega}$ the learner M by

$$M(\sigma) := \begin{cases} o, & \text{if } \operatorname{pos}(\sigma) = \varnothing; \\ \varphi_{\max(\operatorname{pos}(\sigma))}(\langle \sigma \rangle), & \text{otherwise.} \end{cases}$$

We argue that $\mathcal{L} := \{ L \subseteq \mathbb{N} \mid L \text{ is decidable and } L \in \mathbf{InfConsEx}_C(M) \}$ is not consistently learnable by a total learner from informants. Assume towards a contradiction M' is such a learner. For a sequence σ of natural numbers we denote by $\overline{\sigma}$ the corresponding canonical finite informant sequence, ending with the highest value σ takes. Further, for a natural number x we denote by $\tau(x)$ the unique element of $\mathbb{N}^{<\omega}$ with $\langle \tau(x) \rangle = x$. Then by padded ORT there are $e, z \in \mathbb{N}$ and functions $a, b : \mathbb{N}^{<\omega} \to \mathbb{N}$, such that

$$\forall \sigma, \tau \in \mathbb{N}^{<\omega} \, (\, \sigma \subsetneq \tau \Rightarrow \max\{a(\sigma), b(\sigma)\} < \min\{a(\tau), b(\tau)\} \,), \tag{8}$$

with the property that for all $\sigma \in \mathbb{N}^{<\omega}$ and all $i \in \mathbb{N}$

$$\begin{aligned}
\sigma_0 &= \varnothing;\\
\sigma_{i+1} &= \sigma_i {}^\frown \begin{cases} a(\sigma_i), & \text{if } M'(\overline{\sigma_i {}^\frown a(\sigma_i)}) \neq M'(\overline{\sigma_i});\\ b(\sigma_i), & \text{otherwise;} \end{cases} \qquad (9)\\
\varphi_e(y) &= \begin{cases} 1, & \text{if } y \in \operatorname{pos}(\overline{\sigma_y});\\ 0, & \text{otherwise;} \end{cases}\\
\varphi_{a(\sigma)}(x) &= \begin{cases} e, & \text{if } \operatorname{Cons}(\tau(x), \varphi_e) \text{ and } M'(\overline{\sigma {}^\frown a(\sigma)}) \neq M'(\overline{\sigma});\\ \operatorname{ind}(\operatorname{pos}(\tau(x))), & \text{otherwise;} \end{cases}\\
\varphi_{b(\sigma)}(x) &= \begin{cases} e, & \text{if } \operatorname{Cons}(\tau(x), \varphi_e);\\ \operatorname{ind}(\operatorname{pos}(\tau(x))), & \text{otherwise;} \end{cases}
\end{aligned}$$

Consider the decidable language $L_e = \operatorname{pos}(\varphi_e)$. Clearly, we have $L_e \in \mathcal{L}$ and thus M' also $\mathbf{InfConsEx}_C$-learns L_e. By the $\mathbf{Ex}_C$-convergence there are $e', j \in \mathbb{N}$, where j is minimal, such that $\varphi_{e'} = \varphi_e$ and for all $i \geqslant j$ we have $M'(\overline{\sigma_i}) = e'$ and hence $M'(\overline{\sigma_i {}^\frown a(\sigma_i)}) = M'(\overline{\sigma_i})$ by (9).
We now argue that $L := \operatorname{pos}(\overline{\sigma_j}) \cup \{a(\sigma_j)\} \in \mathcal{L}$. Let I be an informant for L and $t \in \mathbb{N}$. By (9) we observe that M is consistent on I as

$$M(I[t]) = \varphi_{\max(\operatorname{pos}(I[t]))}(\langle I[t] \rangle) = \begin{cases} e, & \text{if } \operatorname{Cons}(I[t], \varphi_e);\\ \operatorname{ind}(\operatorname{pos}(I[t])), & \text{otherwise.} \end{cases}$$

Further, by the choice of j we have $\neg\mathbf{Cons}(\,(a(\sigma_j), 1), \varphi_e\,)$. If $\operatorname{pos}(I[t]) = L$, we obtain $\varphi_{M(I[t])} = \operatorname{ind}_L$. On the other hand M' does not consistently learn L as by the choice of j we obtain $M'(\overline{\sigma_j {}^\frown a(\sigma_j)}) = M'(\overline{\sigma_j}) = e'$ and $\neg\operatorname{Cons}(\overline{\sigma_j {}^\frown a(\sigma_j)}, L_e)$, a contradiction. ∎

Thus, learning algorithms not defined on all inputs have strictly more learning power.

As we clearly can do a padding-trick for C-indices, similar to Lemma 20, we might assume the learner to be syntactically decisive. Furthermore, the separations of **Caut**, **Mon** and **SMon** are still valid as they are witnessed by indexable families. Thus, the interesting question is whether **Conv** and **SDec** are also not restrictive for binary classifiers. We now observe that this still holds true but the proof is much simpler than for W-indices, because the consistency of data with hypotheses is decidable.

Theorem 34. *For $\delta \in \{\mathbf{T}, \mathbf{Mon}\}$ holds*

$$[\mathbf{Inf}\delta C\mathbf{IndBc}_C] = [\mathbf{InfConvSDec}\delta C\mathbf{IndEx}_C].$$

Proof. By the comment after Proposition 32 we assume $\delta \subseteq \mathbf{Cons}$. Let $\mathcal{L} \in [\mathbf{Inf}\delta\mathbf{Bc}]$ and the learner M witnessing this. It is an easy exercise to check that

the following learner acts as required, where σ is a finite informant sequence and $\xi \in \mathbb{N} \times \{0,1\}$:

$$M'(\varnothing) = M(\varnothing);$$
$$M'(\sigma^\frown\xi) = \begin{cases} M(\sigma^\frown\xi) & \text{if } \neg\mathrm{Cons}(\sigma^\frown\xi, M'(\sigma)); \\ M'(\sigma) & \text{otherwise.} \end{cases}$$

Note that the consistency of M on $\mathcal{L}$ is only employed to obtain **SDec**. ∎

Corollary 35. $[\mathbf{Inf}_{\mathrm{can}}C\mathbf{IndBc}_C] = [\mathbf{InfConsConv}C\mathbf{IndEx}_C]$.

For $\tau(C\mathbf{Ind})$-learners the simulation in Theorem 34 preserves totality.

In a nutshell for learners only outputting C-indices, we obtain the same map as for W-indices. In contrast, **Cons** is not a restriction anymore.

Moreover, $\mathbf{Bc}_C$-learning is not weaker than explanatory learning and thus the vacillatory hierarchy collapses.

8 Further Research

Future investigations could address the relationships between the different delayable learning restrictions for other convergence criteria, where the general results in Section 3 may be helpful.

According to Osherson, Stob, and Weinstein [1986] requiring the learner to base its hypothesis only on the previous one and the current datum, makes **Ex**-learning harder. While the relations between the delayable learning restrictions for these so called *iterative learners* in the presentation mode of solely positive information has been investigated by Jain, Kötzing, Ma, and Stephan [2016], so far this has not been done when learning from informants. For indexable families, this was already of interest to Lange and Zeugmann [1992], Lange and Grieser [2003] and Jain, Lange, and Zilles [2006]. In the corresponding map each of **Caut**, **Mon** and **SMon** is separated from all other learning restrictions. Moreover, **Conv** restricts iterative learning from informant and we are sure that also **SNU** does. It remains open, whether all syntactic learning criteria have the same learning power. Further, it seems like settling **NU**, **Dec** and **WMon** requires completely new techniques. This model is of special interest as it models the behavior of neural networks. Further improvements to the model would be a more appropriate hypothesis space, a probabilistic presentation of the data and other convergence criteria. For C-indices the incomparability of **Caut** and **Mon**, as well as the separation of **Conv** are still valid.

For automatic structures as alternative approach to model a learner, there have been investigations on how different types of text effect the **Ex**-learnability, see Jain, Luo, and Stephan [2010] and Hölzl, Jain, Schlicht, Seidel, and Stephan

[2017]. The latter started investigating how learning from canonical informants and learning from text relate to one another in the automatic setting. A natural question seems to be what effect other kinds of informants and learning success criteria have.

Last but not least, rating the models value for other research aiming at understanding the capability of human and machine learning is another very challenging task to tackle.

Acknowledgments and Funding

This work was supported by the German Research Foundation (DFG) under Grant KO 4635/1-1 (SCL).

Bibliography

Y. Akama and T. Zeugmann. Consistent and coherent learning with δ-delay. *Information and Computation*, 206(11):1362–1374, 2008.

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

G. Baliga, J. Case, W. Merkle, F. Stephan, and R. Wiehagen. When unlearning helps. *Information and Computation*, 206:694–709, 2008.

J. Bārzdiņš. Inductive inference of automata, functions and programs. In *Amer. Math. Soc. Transl.*, pages 107–122, 1977.

J. Bārzdiņš and K. Podnieks. The theory of inductive inference. In *Mathematical Foundations of Computer Science*, 1973.

L. Blum and M. Blum. Toward a mathematical theory of inductive inference. *Information and Control*, 28:125–155, 1975.

J. Case. The power of vacillation in language learning. *SIAM Journal on Computing*, 28(6):1941–1969, 1999.

J. Case. Gold-style learning theory. In *Topics in Grammatical Inference*, pages 1–23. 2016.

J. Case and T. Kötzing. Difficulties in forcing fairness of polynomial time inductive inference. In *Proc. of Algorithmic Learning Theory*, pages 263–277, 2009.

J. Case and S. Moelius. Optimal language learning from positive data. *Information and Computation*, 209:1293–1311, 2011.

J. Case and C. Smith. Comparison of identification criteria for machine inductive inference. *Theoretical Computer Science*, 25(2):193–220, 1983.

M. Fulk. *A Study of Inductive Inference Machines*. PhD thesis, SUNY at Buffalo, 1985.

E. Gold. Language identification in the limit. *Information and Control*, 10: 447–474, 1967.

R. Hölzl, S. Jain, P. Schlicht, K. Seidel, and F. Stephan. Automatic learning from repetitive texts. In *Proc. of Algorithmic Learning Theory*, pages 129–150, 2017.

S. Jain and A. Sharma. Generalization and specialization strategies for learning r.e. languages. *Annals of Mathematics and Artificial Intelligence*, 23(1-2):1–26, 1998.

S. Jain, D. Osherson, J. Royer, and A. Sharma. *Systems that Learn: An Introduction to Learning Theory*. MIT Press, Cambridge, Massachusetts, second edition, 1999.

S. Jain, S. Lange, and S. Zilles. Towards a better understanding of incremental learning. In *ALT*, volume 4264 of *Lecture Notes in Computer Science*, pages 169–183, 2006.

S. Jain, Q. Luo, and F. Stephan. Learnability of automatic classes. In *LATA*, pages 321–332, 2010.

S. Jain, T. Kötzing, J. Ma, and F. Stephan. On the role of update constraints and text-types in iterative learning. *Information and Computation*, 247:152–168, 2016.

K. P. Jantke. Monotonic and nonmonotonic inductive inference of functions and patterns. In *Nonmonotonic and Inductive Logic, 1st International Workshop, Proc.*, pages 161–177, 1991.

E. Kinber and F. Stephan. Language learning from texts: mindchanges, limited memory, and monotonicity. *Information and Computation*, 123(2):224–241, 1995.

T. Kötzing and R. Palenta. A map of update constraints in inductive inference. In *Algorithmic Learning Theory*, pages 40–54, 2014.

T. Kötzing and R. Palenta. A map of update constraints in inductive inference. *Theoretical Computer Science*, 650:4–24, 2016.

T. Kötzing and M. Schirneck. Towards an atlas of computational learning theory. In *33rd Symposium on Theoretical Aspects of Computer Science*, 2016.

T. Kötzing, M. Schirneck, and K. Seidel. Normal forms in semantic language identification. In *Proc. of Algorithmic Learning Theory*, pages 493–516. PMLR, 2017.

S. Lange and G. Grieser. Variants of iterative learning. *Theoretical computer science*, 292(2):359–376, 2003.

S. Lange and T. Zeugmann. Types of monotonic language learning and their characterization. In *Proc. 5th Annual ACM Workshop on Comput. Learning Theory*, pages 377–390, New York, NY, 1992. ACM Press.

S. Lange and T. Zeugmann. Monotonic versus non-monotonic language learning. In *Proc. of Nonmonotonic and Inductive Logic*, pages 254–269, 1993.

S. Lange and T. Zeugmann. Characterization of language learning from informant under various monotonicity constraints. *Journal of Experimental & Theoretical Artificial Intelligence*, 6(1):73–94, 1994.

S. Lange, T. Zeugmann, and S. Kapur. Monotonic and dual monotonic language learning. *Theoretical Computer Science*, 155(2):365–410, 1996.

P. Odifreddi. *Classical Recursion Theory*, volume II. Elsivier, Amsterdam, 1999.

D. Osherson and S. Weinstein. Criteria of language learning. *Information and Control*, 52:123–138, 1982.

D. Osherson, M. Stob, and S. Weinstein. Learning strategies. *Information and Control*, 53:32–51, 1982.

D. Osherson, M. Stob, and S. Weinstein. *Systems that Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists*. MIT Press, Cambridge, Mass., 1986.

L. Pitt. Inductive inference, DFAs, and computational complexity. In *Proc. of AII (Analogical and Inductive Inference)*, pages 18–44, 1989.

J. Royer and J. Case. *Subrecursive Programming Systems: Complexity and Succinctness*. Research monograph in *Progress in Theoretical Computer Science*. Birkhäuser Boston, 1994.

G. Schäfer-Richter. Über Eingabeabhängigkeit und Komplexität von Inferenzstrategien, 1984. Dissertation, RWTH Aachen.

K. Wexler and P. Culicover. *Formal Principles of Language Acquisition*. MIT Press, Cambridge, Massachusetts, 1980.

R. Wiehagen. A thesis in inductive inference. In *Nonmonotonic and Inductive Logic, 1st International Workshop, Proc.*, pages 184–207, 1991.

R. Wiehagen and T. Zeugmann. Learning and consistency. In *Algorithmic Learning for Knowledge-Based Systems*, pages 1–24. 1995.