LEARNING LANGUAGES IN THE LIMIT FROM POSITIVE INFORMATION WITH FINITELY MANY MEMORY CHANGES

TIMO KÖTZING, KAREN SEIDEL

Abstract. We investigate learning collections of languages from texts by an inductive inference machine with access to the current datum and its memory in form of states. The bounded memory states (**BMS**) learner is considered successful in case it eventually settles on a correct hypothesis while exploiting only finitely many different states.

We give the complete map of all pairwise relations for an established collection of learning success restrictions. Most prominently, we show that non-U-shapedness is not restrictive, while conservativeness and (strong) monotonicity are. Some results carry over from iterative learning by a general lemma showing that, for a wealth of restrictions (the *semantic* restrictions), iterative and bounded memory states learning are equivalent. We also give an example of a non-semantic restriction (strongly non-U-shapedness) where the two settings differ.

1. Introduction

We are interested in the problem of algorithmically learning a description for a formal language (a computably enumerable subset of the set of natural numbers) when presented successively all and only the elements of that language; this is sometimes called *inductive inference*, a branch of (algorithmic) learning theory. For example, a learner h might be presented more and more even numbers. After each new number, h outputs a description for a language as its conjecture. The learner h might decide to output a program for the set of all multiples of 4, as long as all numbers presented are divisible by 4. Later, when h sees an even number not divisible by 4, it might change this guess to a program for the set of all multiples of 2.

Many criteria for deciding whether a learner h is *successful* on a language L have been proposed in the literature. Gold, in his seminal paper [Gol67], gave a first, simple learning criterion, **TxtEx**-*learning*[1], where a learner is *successful* iff, on every *text* for L (listing of all and only the elements of L) it eventually stops changing its conjectures, and its final conjecture is a correct description for the input sequence. Trivially, each single, describable language L has a suitable constant function as an **TxtEx**-learner (this learner constantly outputs a description for L). Thus, we are interested in analyzing for which *classes of languages* $\mathcal{L}$ is there a *single learner* h learning *each* member of $\mathcal{L}$. This framework is also sometimes known as *language learning in the limit* and has been studied extensively, using a wide range of learning criteria similar to **TxtEx**-learning (see, for example, the textbook [JORS99]).

One major criticism of the model suggested by Gold is it's excessive use of memory: for each new hypothesis the entire history of past data is available. Iterative learning is the most common variant of learning in the limit which addresses memory constraints: the memory of the learner on past data is just its current hypothesis. Due to the padding lemma, this memory is still not void, but finitely many data can be memorized in the hypothesis. Another way of restricting the memory is in analogy to the computation of finite automata: a learner can pass on not it's current hypothesis, but a *state* which can be used in the computation of the next hypothesis (and next state). This was introduced in [CCJS07] and called *bounded memory states (BMS)* learning.

There is already a quite comprehensive body of work on iterative learning [CK10, CM08b, JKMS16, JMZ13, JORS99]. In contrast, the rather natural setting of storing just a single state is not analyzed much at all.

It is a reasonable assumption to have a countable reservoir of states. Hence, we use natural numbers as such. Note that allowing arbitrary use of all natural numbers as states would effectively allow a learner to

[1] **Txt** stands for learning from a *text* of positive examples; **Ex** stands for *explanatory*.

store all seen data in the state, thus giving the same mode as Gold's original setting. Probably the minimal way to restrict the use of states is to demand that a learner must stop using new states in order to be considered successfully learning (but may still traverse among the finitely many states produced so far, and may use infinitely many states on garbage data). It was claimed that this setting is equivalent to iterative learning [CCJS07, Remark 38] (this restriction is here called *ClassBMS*, we will call it $\mathbf{TxtBMS_*Ex}$). However, this was only remarked for the plain setting of explanatory learning; for further restrictions, the setting is completely unknown, only for explicit constant state bounds a few scattered results are known [CCJS07, CK13].

In this paper, we consider a wealth of restrictions, described in detail in Section 2 (after an introduction to the general notation of this paper). Following the approach of giving *maps* of pairwise relations suggested in [KS16], we give a complete map in Figure 1. We note that this map is the same as the map for iterative learning given in [JKMS16], but partially for different reasons.

Based on carefully arranged definitions suitable for the general result, in Lemma 3.1 we show that, for many restrictions (the so-called *semantic* restrictions, where only the semantics of hypotheses are restricted) the learning setting with bounded memory states is equivalent to learning iteratively. This proves and generalizes the aforementioned remark in [CCJS07] to a wide class of restrictions. The iterative learner uses the hypotheses of the $\mathbf{BMS_*}$-learner on an equivalent text and additionally pads a subgraph of the translation diagram of the **BMS**-learner to it. It keeps track of all states visited so far together with the datum which caused the first transfer to the respective state. This way we can reconstruct the last first-time-visited state while observing the equivalent sequence. Moreover, the equivalent text prevents the iterative learner from returning to a previously visited state but the last one and hence enables the **Ex**-convergence.

However, if restrictions are not semantic, then iterative and bounded memory states learning can differ. We show this concretely for the case of so-called *strongly non-U-shaped* learning in Theorem 4.5; the proof uses an intricate ORT-argument, indicating that the two settings, while different, are very similar nonetheless. It is based on the proof that strong non-U-shapedness restricts $\mathbf{BMS_*Ex}$-learning. The proof of the latter result combines the techniques for showing that strong non-U-shapedness restricts iterative learning, as stated in [CK13, Theorem 5.7], and that not every by an iterative learner strongly monotonically learnable set of languages is strongly non-U-shapedly learnable by an iterative learner, see [JKMS16, Theorem 5]. Moreover, it relies on showing that state decisiveness can be assumed in Lemma 4.1.

The remainder of Section 4 completes the map given in Figure 1 for the case of syntactic restrictions (since these do not carry over from the setting of iterative learning). All syntactic learning requirements are closely related to strongly locking learners. The fundamental concept of a locking sequence was introduced by [BB75]. For a similar purpose than ours [JKMS16] introduced strongly locking learners. We generalize their construction for certain syntactically restricted iterative learners from a strongly locking iterative learner. Finally, we obtain that all non-semantic learning restrictions also coincide for $\mathbf{BMS_*}$-learning.

2. Learners, Success Criteria and other Terminology

As far as possible, we follow [JORS99] on the learning theoretic side and [Odi99] for computability theory. We recall the most essential notation and definitions.

We let $\mathbb{N}$ denote the *natural numbers* including 0. 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.

Further, $X^{<\omega}$ denotes the *finite sequence*s over the set X and X^{ω} stands for the *countably infinite sequence*s over X. For every $\sigma \in X^{<\omega}$ and $t \leqslant |\sigma|$, $t \in \mathbb{N}$, we let $\sigma[t] := \{(s, \sigma(s)) \mid s < t\}$ denote the *restriction of* σ *to* t. Moreover, for sequences $\sigma, \tau \in X^{<\omega}$ their concatenation is denoted by $\sigma^\frown\tau$. Finally, we write $\mathrm{last}(\sigma)$ for the last element of σ , $\sigma(|\sigma|-1)$, and σ^- for the initial segment of σ without $\mathrm{last}(\sigma)$, i.e. $\sigma[|\sigma|-1]$. Clearly, $\sigma = \sigma^{-\frown}\mathrm{last}(\sigma)$.

For a finite set $D \subseteq \mathbb{N}$ and a finite sequence $\sigma \in X^{<\omega}$, we denote by $\langle D\rangle$ and $\langle\sigma\rangle$ a canonical index for D or σ, respectively. Further, we fix a Gödel pairing function $\langle .,. \rangle$ with two arguments.

Let $L \subseteq \mathbb{N}$. We interpret every $n \in \mathbb{N}$ as a code for a word. If L is recursively enumerable, we call L a *language*.

We fix a programming system φ as introduced in [RC94]. 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 defined as $\mathrm{dom}(\varphi_p)$.

In reference to a Blum complexity measure Φ_p, 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, i.e.

$$W_p^t = \{x \mid x \leqslant t \ \wedge \ \Phi_p(x) \leqslant t\}.$$

Moreover, the well-known s-m-n theorem gives finite and infinite recursion theorems. We will refer to Case's Operator Recursion Theorem ORT in its 1-1-form.

Throughout the paper, we let $\Sigma = \mathbb{N} \cup \{\#\}$ be the input alphabet with $n \in \mathbb{N}$ interpreted as code for a word in the language and # interpreted as pause symbol, i.e. no new information. Further, let $\Omega = \mathbb{N} \cup \{?\}$ be the output alphabet with $p \in \mathbb{N}$ interpreted as φ-index and ? as no hypothesis or repetition of the last hypothesis, if existent.

A *learner* is always a (partial) computable function

$$M : \mathrm{dom}(M) \subseteq \Sigma^{<\omega} \to \Omega.$$

The set of all total computable functions $M : \Sigma^{<\omega} \to \Omega$ is denoted by $\mathcal{R}$.

Let $f \in \Sigma^{<\omega} \cup \Sigma^{\omega}$, then the *content of* f, defined as $\mathrm{content}(f) := \mathrm{ran}(f) \backslash \{\#\}$, is the set of all natural numbers, about which f gives some positive information. The *set of all texts for the language* L is defined as

$$\mathbf{Txt}(L) := \{T \in \Sigma^{\omega} \mid \mathrm{content}(T) = L\}.$$

Definition 2.1. Let M be a learner. M is an *iterative learner*, for short $M \in \mathbf{It}$, if there is a computable (partial) hypothesis generating function $h_M : \Omega \times \Sigma \to \Omega$ such that $M = h_M^{\ddagger}$ where $h_M^{\ddagger}$ is defined on finite sequences by

$$\begin{aligned} h_M^{\ddagger}(\epsilon) &= ?; \\ h_M^{\ddagger}(\sigma^\frown x) &= h_M(h_M^{\ddagger}(\sigma), x). \end{aligned}$$

Definition 2.2. Let M be a learner. M is a *bounded memory states learner*, for short $M \in \mathbf{BMS}$, if there are a computable (partial) hypothesis generating function $h_M : \mathbb{N} \times \Sigma \to \Omega$ and a computable (partial) state transition function $s_M : \mathbb{N} \times \Sigma \to \mathbb{N}$ such that $\mathrm{dom}(h_M) = \mathrm{dom}(s_M)$ and $M = h_M^*$ where h_M^* and s_M^* are defined on finite sequences by

$$\begin{aligned} s_M^*(\epsilon) &= 0; \\ h_M^*(\sigma^\frown x) &= h_M(s_M^*(\sigma), x); \\ s_M^*(\sigma^\frown x) &= s_M(s_M^*(\sigma), x). \end{aligned}$$

Note that every iterative learner gives a **BMS**-learner by identifying the hypothesis space Ω with the set of states via a computable bijection between $\mathbb{N}$ and Ω. The resulting **BMS**-learner will succeed on the same languages the iterative learner does learn. Further, as the set of visited states contains exactly all hypotheses the learner puts out, the **BMS**-learner only uses finitely many states on all texts for languages it explanatory learns. In [CCJS07, Rem. 38] the equality $[\mathbf{TxtBMS}_*\mathbf{Ex}] = [\mathbf{ItTxtEx}]$ is claimed. This also follows from our more general Lemma 3.1.

Definition 2.2 may be stated more generally for arbitrary finite or infinite sets of states Q, instead of $\mathbb{N}$. Moreover, s_M^* and h_M^* can easily be generalized to functions taking also a starting state s as input by

$$\begin{aligned} s_M^*(s, \epsilon) &= s; \\ h_M^*(s, \sigma^\frown x) &= h_M(s_M^*(s, \sigma), x); \\ s_M^*(s, \sigma^\frown x) &= s_M(s_M^*(s, \sigma), x). \end{aligned}$$

We now clarify what we mean by succesful learning.

Definition 2.3. Let M be a learner and $\mathcal{L}$ a collection of languages.

(1) Let $L \in \mathcal{L}$ be a language and $T \in \mathbf{Txt}(L)$ a text for L presented to M.
   (a) We call $h = (h_t)_{t\in\mathbb{N}} \in \Omega^\omega$, where $h_t := M(T[t])$ for all $t \in \mathbb{N}$, the *learning sequence of M on T*.
   (b) *M learns L from T in the limit*, for short M **Ex**-learns L from T or $\mathbf{Ex}(M,T)$, if there exitsts t_0 such that $W_{h_{t_0}} = \text{content}(T)$ and $\forall t \geqslant t_0$ $(\ h_t \neq ? \ \Rightarrow\ h_t = h_{t_0}\ )$.

(2) *M learns $\mathcal{L}$ in the limit*, for short M **Ex**-learns $\mathcal{L}$, if $\mathbf{Ex}(M,T)$ for every $L \in \mathcal{L}$ and every $T \in \mathbf{Txt}(L)$.

Ex-learning is the most common definition for successful learning in inductive inference and corresponds to the notion of identifiability in the limit by [Gol67], where the learner eventually decides on one correct hypotheses.

In our investigations, the most important additional requirement on a successful learning process is for a **BMS**-learner to use finitely many states only, as stated in the following definition.

Definition 2.4. Let M be a **BMS**-learner and $T \in \mathbf{Txt}$. We say that M *uses finitely many memory states on T*, for short $\mathbf{BMS}_*(M,T)$, if $\{\, s^*_M(T[t]) \mid t \in \mathbb{N}\,\}$ is finite.

We list the most common additional requirements regarding the learning sequence, which may tag a learning process. For this we first recall the notion of consistency of a sequence with a set.

Definition 2.5. Let $f \in \Sigma^{<\omega} \cup \Sigma^\omega$ and $A \subseteq \Sigma$. We define

$$\mathbf{Cons}(f,A) \quad :\Leftrightarrow \quad \text{content}(f) \subseteq A$$

and say *f is consistent with A*.

The listed properties of the learning sequence have been in the center of different investigations. Studying how they relate to one another did begin quite recently in [KP16], [KS16], [JKMS16] and [AKS18].

Definition 2.6. Let M be a learner, $T \in \mathbf{Txt}$ and $h = (h_t)_{t\in\mathbb{N}} \in \Omega^\omega$ the learning sequence of M on T. We write

(1) $\mathbf{Conv}(M,T)$ ([Ang80]), if M is *conservative on T*, i.e., for all s,t with $s \leqslant t$ holds $\mathbf{Cons}(T[t], W_{h_s}) \ \Rightarrow\ h_s = h_t$.
(2) $\mathbf{Dec}(M,T)$ ([OSW82]), if M is *decisive on T*, i.e., for all r,s,t with $r \leqslant s \leqslant t$ holds $W_{h_r} = W_{h_t} \ \Rightarrow\ W_{h_r} = W_{h_s}$.
(3) $\mathbf{Caut}(M,T)$ ([OSW86]), if M is *cautious on T*, i.e., for all s,t with $s \leqslant t$ holds $\neg W_{h_t} \subsetneq W_{h_s}$.
(4) $\mathbf{WMon}(M,T)$ ([Jan91],[Wie91]), if M is *weakly monotonic on T*, i.e., for all s,t with $s \leqslant t$ holds $\mathbf{Cons}(T[t], W_{h_s}) \ \Rightarrow\ W_{h_s} \subseteq W_{h_t}$.
(5) $\mathbf{Mon}(M,T)$ ([Jan91],[Wie91]), if M is *monotonic on T*, i.e., for all s,t with $s \leqslant t$ holds $W_{h_s} \cap \text{content}(T) \subseteq W_{h_t} \cap \text{pos}(T)$.
(6) $\mathbf{SMon}(M,T)$ ([Jan91],[Wie91]), if M is *strongly monotonic on T*, i.e., for all s,t with $s \leqslant t$ holds $W_{h_s} \subseteq W_{h_t}$.
(7) $\mathbf{NU}(M,T)$ ([BCM$^+$08]), if M is *non-U-shaped on T*, i.e., for all r,s,t with $r \leqslant s \leqslant t$ holds $W_{h_r} = W_{h_t} = \text{content}(T) \ \Rightarrow\ W_{h_r} = W_{h_s}$.
(8) $\mathbf{SNU}(M,T)$ ([CM11]), if M is *strongly non-U-shaped on T*, i.e., for all r,s,t with $r \leqslant s \leqslant t$ holds $W_{h_r} = W_{h_t} = \text{content}(T) \ \Rightarrow\ h_r = h_s$.
(9) $\mathbf{SDec}(M,T)$ ([KP16]), if M is *strongly decisive on T*, i.e., for all r,s,t with $r \leqslant s \leqslant t$ holds $W_{h_r} = W_{h_t} \ \Rightarrow\ h_r = h_s$.
(10) $\mathbf{Wb}(M,T)$ ([KS16]), if M is *witness-based on T*, i.e., for all r,t such that for some s with $r < s \leqslant t$ holds holds $h_r \neq h_s$ we obtain $\text{content}(T[s]) \cap (W_{h_t} \backslash W_{h_r}) \neq \varnothing$.

It has been observed that $\mathbf{Conv}(M,T)$ implies $\mathbf{SNU}(M,T)$ and $\mathbf{WMon}(M,T)$; $\mathbf{SDec}(M,T)$ implies $\mathbf{Dec}(M,T)$ and $\mathbf{SNU}(M,T)$; $\mathbf{SMon}(M,T)$ implies $\mathbf{Caut}(M,T), \mathbf{Dec}(M,T), \mathbf{Mon}(M,T)$, $\mathbf{WMon}(M,T)$ and finally $\mathbf{Dec}(M,T)$, $\mathbf{WMon}(M,T)$ and $\mathbf{SNU}(M,T)$ imply $\mathbf{NU}(M,T)$. Figure 1 includes the resulting backbone with arrows indicating the aforementioned implications. Further, $\mathbf{Wb}(M,T)$ implies $\mathbf{Conv}(M,T)$, $\mathbf{SDec}(M,T)$ and $\mathbf{Caut}(M,T)$.

In order to characterize what successful learning means, these predicates may be combined with the explanatory convergence criterion. For this, we let $\Delta := \{\,\mathbf{Caut}, \mathbf{Conv}, \mathbf{Dec}, \mathbf{SDec}, \mathbf{WMon}, \mathbf{Mon}, \mathbf{SMon}, \mathbf{NU}, \mathbf{SNU}, \mathbf{T}\,\}$ denote the set of *admissible learning restrictions*, with $\mathbf{T}$ standing for no restriction. Further, a *learning success criterion* is an element of

$$\{\bigcap_{i=0}^{n} \delta_i \cap \mathbf{Ex} \mid n \in \mathbb{N}, \forall i \leqslant n(\delta_i \in \Delta)\}.$$

Note that plain explanatory convergence is a learning success criterion by letting $n = 0$ and $\delta_0 = \mathbf{T}$.

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 do not require the learner to settle on exactly one hypothesis.

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 $[\alpha\mathbf{Txt}\beta]$ defined as follows.

Definition 2.7. Let α be a property of partial computable functions from the set $\Sigma^{<\omega}$ to $\mathbb{N}$ and β a learning success criterion. We denote by $[\alpha\mathbf{Txt}\beta]$ the set of all collections of languages that are β-learnable from texts by a learner M with the property α.

At position α, we restrict the set of admissible learners for example by requiring them to be iterative or finite bounded memory states learners. The properties stated at position α are *independent of learning success.*

In contrast, at position β, the required learning behavior and convergence criterion are specified.

For example, a collection of languages $\mathcal{L}$ lies in $[\mathbf{BMSTxtBMS_*ConvEx}]$ if and only if there is a bounded memory states learner M conservatively explanatory learning every $L \in \mathcal{L}$ from texts while using only finite memory. More concretely, for all $L \in \mathcal{L}$ and for every text $T \in \mathbf{Txt}(L)$ we have $\mathbf{Conv}(M,T)$, $\mathbf{BMS}_*(M,T)$ and $\mathbf{Ex}(M,T)$.

The proof of our general observation employs a property of learning requirements and learning success criteria, that applies to all such considered in this paper.

Definition 2.8. 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 β on pairs of learners and texts *allows for simulation on equivalent text*, if for all simulating functions $\mathfrak{s} \in \mathfrak{S}$, all texts $T, T' \in \mathbf{Txt}$ and all learners M, M' holds: Whenever we have $\mathrm{content}(T'[t]) = \mathrm{content}(T[\mathfrak{s}(t)])$ and $M'(T'[t]) = M(T[\mathfrak{s}(t)])$ for all $t \in \mathbb{N}$, from $\beta(M,T)$ we can conclude $\beta(M',T')$.

Intuitively, as long as the learner M' conjectures $h_{\mathfrak{s}(t)} = M(T[\mathfrak{s}(t)])$ at time t and has, in form of $T'[t]$, the same data available as was used by M for this hypothesis, M' on T' is considered to be a simulation of M on T.

It is easy to see that all learning success criteria considered in this paper allow for simulation on equivalent text.

3. Relations between Semantic Learning Requirements

We show that bounded memory states learners and iterative learners have equal learning power, when a semantic learning requirement is added to the standard convergence criterion. With this the results from iterative learning are transferred to this setting.

The following lemma formally establishes the equal learning power of iterative and $\mathbf{BMS}_*$-learning for all learning success criteria but **Conv**, **SDec** and **SNU**. We are going to prove in Section 4 that even for the three aforementioned non-semantic additional requirements we obtain the same behavior.

Lemma 3.1. *Let δ allow for simulation on equivalent text.*

(1) We have $[\mathbf{TxtBMS}_*\delta\mathbf{Ex}] \supseteq [\mathbf{ItTxt}\delta\mathbf{Ex}]$.

(2) If δ is semantic then $[\mathbf{TxtBMS}_*\delta\mathbf{Ex}] = [\mathbf{ItTxt}\delta\mathbf{Ex}]$.

Proof. (1) and "$\supseteq$" of (2). Let M be an iterative learner, i.e. there is a computable function $h_M : \Omega \times \Sigma \to \Omega$ with $M = h_M^\ddagger$ where $h_M^\ddagger(\epsilon) = ?$ and $h_M^\ddagger(\sigma^\frown x) = h_M(h_M^\ddagger(\sigma), x)$ for all $\sigma \in \Sigma^{<\omega}$ and $x \in \Sigma$. We show that M can be obtained as a state driven learner by using the hypotheses also as states. For this, we fix the computable bijection $\pi : Q \to \Omega$ with computable inverse, defined by $\pi(0) = ?$ and $\pi(i) = i - 1$ for all $i > 0$. Then the learner $N = h_N^*$ with $\langle s_N, h_N\rangle(q, x) = (\pi^{-1}(h_M(\pi(q), x)), h_M(\pi(q), x))$ is as wished because the state corresponds via π directly to the last hypothesis of M and so the learners M and N act identically.

Formally, this follows by an induction showing for every $\tau \in \Sigma^{<\omega}$ that $s_N^*(\tau) = \pi^{-1}(M(\tau))$ and moreover if $|\tau| > 0$ we have $N(\tau) = M(\tau)$. The claim holds for $\tau = \epsilon$, because of $s_N^*(\epsilon) = 0 = \pi^{-1}(M(\epsilon))$. In case there are $\sigma \in \Sigma^{<\omega}$ and $x \in \Sigma$ such that $\tau = \sigma^\frown x$, we may assume $s_N^*(\sigma) = \pi^{-1}(M(\sigma))$ and obtain

$$\begin{aligned} s_N^*(\tau) &\overset{\text{Def. } s_N^*}{=} s_N(s_N^*(\sigma), x) \overset{s_N^*(\sigma)=\pi^{-1}(M(\sigma))}{=} s_N(\pi^{-1}(M(\sigma)), x) \overset{\text{Def. } s_N}{=} \pi^{-1}(h_M(M(\sigma), x)) \overset{M=h_M^\ddagger}{=} \pi^{-1}(M(\tau)),\\ N(\tau) &\overset{N=h_N^*}{=} h_N(s_N^*(\sigma), x) \overset{s_N^*(\sigma)=\pi^{-1}(M(\sigma))}{=} h_N(\pi^{-1}(M(\sigma)), x) \overset{\text{Def. } h_N}{=} h_M(M(\sigma), x)) \overset{M=h_M^\ddagger}{=} M(\tau). \end{aligned}$$

That M in case of learning success uses only finitely many states follows immediately from the **Ex**-convergence, implying to output only finitely many pairwise distinct hypotheses.

"$\subseteq$" of (2). Let $\mathcal{L} \in [\mathbf{TxtBMS}_*\delta\mathbf{Ex}]$ be witnessed by the learner M, i.e., there is $\langle s_M, h_M\rangle : Q \times \Sigma \to Q \times \Omega$ such that $M = h_M^*$. Further, we may assume that for all $L \in \mathcal{L}$ and $T \in \mathbf{Txt}(L)$ the set of visited states $s_M^*[\{T[t] \mid t \in \mathbb{N}\}]$ is finite and M $\delta\mathbf{Ex}$-learns L from T.

Intuitively, the iterative learner $M_{\mathbf{It}}$ uses the hypotheses of M on an equivalent text $\hat{T}$ and additionally pads a subgraph $V(\sigma)$ of the translation diagram of the **BMS**-learner M to it. In $V(\sigma)$, which is being build after having observed σ, we keep track of all states visited so far together with the datum which caused the first transfer to the respective state. In order to assure **Ex**-convergence, we do not change the subgraph in case the new state had already been visited after some proper initial segment of σ was observed. From $V(\sigma)$ we can reconstruct the last first-time-visited state $s_{M_{\mathbf{It}}}^*(\sigma)$ of M while observing the equivalent sequence corresponding to σ. Moreover, we build the equivalent text $\hat{T}$ by inserting a path of already observed data leading to state $s_{M_{\mathbf{It}}}^*(\sigma)$, in case this is necessary to prevent the learner $M_{\mathbf{It}}$ from returning to a previously visited state but the last one. With this strategy we make sure that the last state is the one we are currently in, as keeping track of the current state while observing the original text may destroy the **Ex**-convergence.

Formally, we define functions pump : $\Sigma^{<\omega}\backslash\{\epsilon\} \times \mathbb{N} \to \Sigma^{<\omega}$ and $V : \Sigma^{<\omega} \to \Sigma^{<\omega}$ by

$$\begin{aligned} \mathrm{pump}(V(\sigma), x) &= \begin{cases} x, & \text{if } s_M(s_{M_{\mathbf{It}}}^*(\sigma), x) \notin \mathrm{pr}_1[V(\sigma)];\\ x^\frown \mathrm{path}(s_M(s_{M_{\mathbf{It}}}^*(\sigma), x), s_{M_{\mathbf{It}}}^*(\sigma)), & \text{otherwise;} \end{cases}\\ V(\epsilon) &= \epsilon;\\ V(\sigma^\frown x) &= \begin{cases} V(\sigma)^\frown\langle s_M(s_{M_{\mathbf{It}}}^*(\sigma), x), x\rangle, & \text{if } s_M(s_{M_{\mathbf{It}}}^*(\sigma), x) \notin \mathrm{pr}_1[V(\sigma)];\\ V(\sigma), & \text{otherwise;} \end{cases} \end{aligned}$$

with the application of the projection to the first coordinate extracting the set of visited states. Moreover, for states $s_0, s_1 \in S$ with $\mathrm{path}(s_0, s_1)$ we refer to the unique sequence $(\sigma(i), \sigma(i+1), \ldots, \sigma(j))$ of second coordinates in $V(\sigma)$ such that $(s_0, \sigma(i))^\frown \ldots ^\frown(s_1, \sigma(j))$ is an intermediate sequence in $V(\sigma)$. The learner $M_{\mathbf{It}}$ is now defined by

$$M_{\mathbf{It}}(\sigma^\frown x) = \mathrm{pad}(h_M^*(s_{M_{\mathbf{It}}}^*(\sigma), \mathrm{pump}(V(\sigma), x)), V(\sigma^\frown x)).$$

By construction $s_{M_{\mathbf{It}}}^*(\sigma) = \mathrm{last}(\mathrm{pr}_1(V(\sigma)))$ and therefore the hypothesis of $M_{\mathbf{It}}$ on some sequence $\sigma^\frown x$ is always only based on $V(\sigma)$ and x, which makes $M_{\mathbf{It}}$ iterative.

The text $\hat{T} = \bigcup_{t\in\mathbb{N}} \tau_t$ with $\tau_0 = \epsilon$ and $\tau_{t+1} = \tau_t{}^\frown\mathrm{pump}(V(T[t]), T(t))$ is a text for L. Let $\mathfrak{s} : \mathbb{N} \to \mathbb{N}, t \mapsto |\tau_t|$ be the corresponding simulating function. As for all $t \in \mathbb{N}$ holds $\mathrm{content}(T[t]) = \mathrm{content}(\hat{T}[\mathfrak{s}(t)])$ and

$M_{\mathbf{It}}(T[t]) = \mathrm{pad}(M(\hat{T}[\mathfrak{s}(t)]), V(T[t]))$, we obtain $W_{M_{\mathbf{It}}(T[t])} = W_{M(\hat{T}[\mathfrak{s}(t)])}$ and because δ is semantic and afsoet, we conclude the semantic δ-convergence of $M_{\mathbf{It}}$ on T. Having in mind that M uses only finitely many pairwise distinct states $V(T[t])$ stabilizes. Paired with the **Ex**-convergence of M on $\hat{T}$ we conclude the **Ex**-convergence of $M_{\mathbf{It}}$ on T. □

Note that obviously the proof is identical for learning from positive and negative information, introduced by [Gol67]. In this learning model the information the learner receives is labeled, like in binary classification, and has to be complete in the limit. See [AKS18] for a formal definition, a summary of results on this model and the complete map.

With Lemma 3.1 the following results transfer from learning with iterative learners and it remains to investigate the relations to and between the non-semantic requirements **Conv**, **SDec** and **SNU**.

Theorem 3.2. *(1)* $[\mathbf{TxtBMS_*NUEx}] = [\mathbf{TxtBMS_*Ex}]$
(2) $[\mathbf{TxtBMS_*DecEx}] = [\mathbf{TxtBMS_*WMonEx}] = [\mathbf{TxtBMS_*CautEx}] = [\mathbf{TxtBMS_*Ex}]$
(3) $[\mathbf{TxtBMS_*MonEx}] \subsetneq [\mathbf{TxtBMS_*Ex}]$
(4) $[\mathbf{TxtBMS_*SMonEx}] \subsetneq [\mathbf{TxtBMS_*MonEx}]$

Proof. The respective results for iterative learners can be found in [CM08a, Theorem 2], [JKMS16, Theorem 10], [JKMS16, Theorem 3] and [JKMS16, Theorem 2]. □

4. Relations to and between Syntactic Learning Requirements

The following lemma establishes that we may assume $\mathbf{BMS}_*$-learners to never go back to withdrawn states. We are going to employ this property in almost all of the following proofs.

Lemma 4.1. *Let β be a learning success criterion allowing for simulation on equivalent text and $\mathcal{L} \in [\mathbf{TxtBMS}_*\beta]$. Then there is a* **BMS***-learner N such that N never returns to a withdrawn state and* $\mathbf{BMS}_*\beta$*-learns $\mathcal{L}$ from texts.*

Proof. Let M be a **BMS**-learner with $\mathcal{L} \in \mathbf{TxtBMS}_*\beta(M)$. We employ a construction similar to the one in the proof of Theorem 3.1. Again for $V \in ({}^{<\omega}Q \times \Sigma)$ with pairwise distinct first coordinates and $s' \in \mathrm{pr}_1[V]$ by $\mathrm{path}(V, s')$ we denote the unique sequence of second coordinates $x_0{}^\frown \dots {}^\frown x_\xi$ of V such that $(s', x_0)^\frown \dots {}^\frown(\mathrm{last}(\mathrm{pr}_1[V]), x_\xi)$ is a final segment of V. The **BMS** learner N is initialized with state $\mathrm{pad}(0, (0, \#))$ and for every $s \in Q$, $V \in ({}^{<\omega}Q \times \Sigma)$ and $x \in \Sigma$ defined by

$$s_N(\langle s, V\rangle, x) = \begin{cases} \langle s, V\rangle, & \text{if } s_M(s,x) \in \mathrm{pr}_1[V]; \\ \langle s_M(s,x), V^\frown(s_M(s,x), x)\rangle, & \text{otherwise;} \end{cases}$$

$$h_N(\langle s, V\rangle, x) = \begin{cases} h_M^*(s, x^\frown \mathrm{path}(V, s_M(s,x))), & \text{if } s_M(s,x) \in \mathrm{pr}_1[V]; \\ h_M(s,x), & \text{otherwise.} \end{cases}$$

By construction N is a $\mathbf{BMS}_*$-learner, as it only uses states $\langle s, V\rangle$ where $s = \mathrm{pr}_1(\mathrm{last}(V))$ is a state used by M and for every $s \in Q$, visited by M, there is exactly one sequence $V \in ({}^{<\omega}Q \times \Sigma)$ such that $\langle s, V\rangle$ is used by N. The learner N simulates M on an equivalent text just as in the proof of Theorem 3.1. □

We show that strongly monotonically $\mathbf{BMS}_*$-learnability does not imply strongly non-U-shapedly $\mathbf{BMS}_*$-learnability.

Theorem 4.2. $[\mathbf{TxtBMS_*SMonEx}] \nsubseteq [\mathbf{TxtBMS_*SNUEx}]$

Proof. Consider the **BMS**-learner M initialized with state $\langle\, ?, \langle\varnothing\rangle\rangle$ and h_M and s_M for every $e \in \Omega$, $D \subseteq \mathbb{N}$ finite and $x \in \Sigma$ defined by:

$$s_M(\langle e, \langle D\rangle\rangle, x) = \begin{cases} \langle e, \langle D\rangle\rangle, & \text{if } x \in D \cup \{\#\} \ \vee \ \varphi_x(e) = e; \\ \langle \varphi_x(e), \langle D \cup \{x\}\rangle\rangle, & \text{else if } \varphi_x(e) \neq e; \\ \uparrow, & \text{otherwise.} \end{cases}$$

$$h_M(\langle e, \langle D\rangle\rangle, x) = \begin{cases} e, & \text{if } x \in D \cup \{\#\} \ \vee\ \varphi_x(e) = e; \\ \varphi_x(e), & \text{else if } \varphi_x(e) \neq e; \\ \uparrow, & \text{otherwise.} \end{cases}$$

Thus, M is self-learning by interpreting the datum x as a program and the conjectures are generated by applying this program to the last hypothesis. (We identify φ_x with the function obtained by using a bijection from $\mathbb{N}$ to Ω.) Further, in form of the states, the last hypothesis as well as exactly the data that already lead to a mind change of M is stored.

Let $\mathcal{L} = \mathbf{TxtBMS_*SMonEx}(M)$.

Assume there is a $\mathbf{BMS}_*$-learner N with hypothesis generating function h_N and state transition function s_N, such that $\mathcal{L} \subseteq \mathbf{TxtBMS_*SNUEx}(N)$. By Lemma 4.1 we assume that N does not return to withdrawn states.

We are going to obtain a language $L \in \mathcal{L}$ not strongly non-U-shapedly learned by N by applying 1-1 ORT and thereby refering to the c.e. predicates MC and NoMC defined for fixed $a, b \in \mathcal{R}$, all $k \in \mathbb{N}$ and $\sigma \in \Sigma^{<\omega}$ with the help of the formulas $\psi_k(\ell)$, expressing that the $\mathbf{BMS}_*$-learner N does not perform a mind- or state-change on the text $a[k]^\frown b(k)^\frown \#^\infty$ after having observed $a[k]^\frown b(k)^\frown \#^\ell$. The predicates state that N does converge and (not) make a mind-change when observing σ after having observed $a[k]^\frown a(k)^\frown \#^{\ell_k}$, with ℓ_k being the least ℓ with $\psi_k(\ell)$.

$$\begin{aligned} \psi_k(\ell) &\Leftrightarrow N(a[k]^\frown b(k)^\frown \#^\ell) = N(a[k]^\frown b(k)^\frown \#^{\ell+1}) \ \wedge\ s^*_N(a[k]^\frown b(k)^\frown \#^\ell) = s^*_N(a[k]^\frown b(k)^\frown \#^{\ell+1}); \\ \mathrm{NoMC}(k, \sigma) &\Leftrightarrow \exists \ell_k \in \mathbb{N}\,(\, \psi_k(\ell_k) \ \wedge\ \forall \ell < \ell_k\, \neg\psi_k(\ell) \ \wedge\ N(a[k]^\frown b(k)^\frown \#^{\ell_k \frown}\sigma)\downarrow = N(a[k]^\frown b(k)^\frown \#^{\ell_k})\,); \\ \mathrm{MC}(k, \sigma) &\Leftrightarrow \exists \ell_k \in \mathbb{N}\,(\, \psi_k(\ell_k) \ \wedge\ \forall \ell < \ell_k\, \neg\psi_k(\ell) \ \wedge\ N(a[k]^\frown b(k)^\frown \#^{\ell_k \frown}\sigma)\downarrow \neq N(a[k]^\frown b(k)^\frown \#^{\ell_k})\,). \end{aligned}$$

Now, let p be an index for the program which on inputs $k \in \mathbb{N}$ and $\sigma \in \Sigma^{<\omega}$ searches for ℓ_k. In case ℓ_k exists, the program encoded in p runs N on $a[k]^\frown b(k)^\frown \#^{\ell_k \frown}\sigma$. Hence, $\Phi_p(k, \sigma)$ stands for the number of computation steps the program just described needs on input k, σ. By the definition of p we have $\Phi_p(k, \sigma)\uparrow$ if and only if $\ell_k\uparrow$ or $N(a[k]^\frown b(k)^\frown \#^{\ell_k \frown}\sigma)\uparrow$.

We abbreviate with $({}^{<\omega}a, i) = {}^{<\omega}\leqslant_i(\mathrm{ran}(a[i]) \cup \{\#\})$ the set of all finite sequences over $\mathrm{ran}(a[i]) \cup \{\#\}$ with length at most i. Moreover, we employ a well-order $<_a$ on $({}^{<\omega}\mathrm{ran}(a))$ by letting $\rho <_a \sigma$ if and only if for the unique i_ρ such that $\rho \in ({}^{<\omega}a, i_\rho + 1) \backslash ({}^{<\omega}a, i_\rho)$ holds $\sigma \notin ({}^{<\omega}a, i_\rho + 1)$ or else $\sigma \notin ({}^{<\omega}a, i_\rho)$ and at the same time $\langle\rho\rangle < \langle\sigma\rangle$. For constructing L we will also make use of the c.e. sets

$$E_k = \{\, a(i) \mid \forall \sigma \in ({}^{<\omega}a, i)\, \mathrm{NoMC}(k, \sigma) \ \vee\ (\exists \sigma \forall \rho <_a \sigma\, \mathrm{NoMC}(k, \rho) \ \wedge\ \Phi_p(k, \sigma) > i\,)\,\}.$$

It is easy to see that E_k is finite and equals $\{\, a(i) \mid i < \max(\{i_{\sigma_0}\} \cup \{\Phi_p(k, \sigma) \mid \sigma \leqslant_a \sigma_0\})\,\}$ if and only if for $\sigma_0 \in ({}^{<\omega}\mathrm{ran}(a))$ holds $\mathrm{MC}(k, \sigma_0)$ and $\mathrm{NoMC}(k, \sigma)$ for all $\sigma <_a \sigma_0$. Otherwise $E_k = \mathrm{ran}(a)$.

By 1-1 ORT there are $a, b, e_1, e_2 \in \mathcal{R}$ with pairwise disjoint ranges and $e_0 \in \mathbb{N}$, such that

$$\varphi_{a(i)}(e) = \begin{cases} e_0 & \text{if } e \in \{?, e_0\}; \\ e_2(k) & \text{else if } e = e_1(k) \text{ for some } k \leqslant i; \\ e, & \text{otherwise;} \end{cases}$$

$$\varphi_{b(k)}(e) = \begin{cases} e_1(k) & \text{if } e \in \{?, e_0\}; \\ e, & \text{otherwise;} \end{cases}$$

$$W_{e_0} = \begin{cases} \mathrm{ran}(a[t_0]) & \text{if } t_0 \text{ is minimal with } \forall t \geqslant t_0\,(\, N(a[t]) = N(a[t_0]) \wedge s^*_N(a[t]) = s^*_N(a[t_0])\,); \\ \mathrm{ran}(a), & \text{no such } t_0 \text{ exists.}; \end{cases}$$

$$W_{e_1(k)} = \mathrm{content}(a[k]) \cup \{b(k)\} \cup \begin{cases} E_k & \text{if } \exists \sigma_0\,(\,\mathrm{MC}(k, \sigma_0) \ \wedge\ \forall \sigma <_a \sigma_0\, \mathrm{NoMC}(k, \sigma)\,); \\ \varnothing, & \text{otherwise;} \end{cases}$$

$$W_{e_2(k)} = \mathrm{content}(a[k]) \cup \{b(k)\} \cup E_k.$$

As $W_{e_0} \in \mathcal{L}$ by construction, N has to learn it and hence t_0 exists.

We first observe that there exists σ_0 such that $\mathrm{MC}(t_0, \sigma_0)$ and $\mathrm{NoMC}(t_0, \sigma)$ for all $\sigma <_a \sigma_0$. Assume otherwise, then either $\ell_{t_0}\uparrow$ or for all $\sigma \in ({}^{<\omega}\mathrm{ran}(a))$ holds $\mathrm{NoMC}(t_0, \sigma)$ or for σ_0 minimal with $\neg\mathrm{NoMC}(t_0, \sigma_0)$ we have $N(a[t_0]^\frown b(t_0)^\frown \#^{\ell_{t_0}\frown}\sigma_0)\uparrow$. Anyhow, this would mean $E_{t_0} = \mathrm{ran}(a)$. By the definition of e_1, e_2 and our converse assumption we obtain $W_{e_1(t_0)} = \mathrm{content}(a[t_0]) \cup \{b(t_0)\}$ and $W_{e_2(t_0)} = \mathrm{ran}(a) \cup \{b(t_0)\}$. It can be easily checked that $W_{e_1(t_0)}$ and $W_{e_2(t_0)}$ are strongly monotonically learned by M and hence lie in $\mathcal{L}$. As N has to learn $W_{e_1(t_0)}$ from the text $a[t_0]^\frown b(t_0)^\frown \#^\infty$, we know $\ell_{t_0}\downarrow$ and moreover $W_{N(a[t_0]^\frown b(t_0)^\frown \#^\ell)} = W_{e_1(t_0)}$ holds for all $\ell \geqslant \ell_{t_0}$. Moreover, N has to learn $W_{e_2(t_0)}$ from all the texts $a[t_0]^\frown b(t_0)^\frown \#^{\ell_{t_0}\frown}\sigma^\frown a$ with $\sigma \in ({}^{<\omega}\mathrm{ran}(a))$. Thus, $N(a[t_0]^\frown b(t_0)^\frown \#^{\ell_{t_0}\frown}\sigma)\downarrow$ for all $\sigma \in ({}^{<\omega}\mathrm{ran}(a))$. Because of our converse assumption, the only option left is $\mathrm{NoMC}(t_0, \sigma)$ for all $\sigma \in ({}^{<\omega}\mathrm{ran}(a))$. Since this is equivalent to $N(a[t_0]^\frown b(t_0)^\frown \#^{\ell_{t_0}\frown}\sigma) = N(a[t_0]^\frown b(t_0)^\frown \#^{\ell_{t_0}})$ for all $\sigma \in ({}^{<\omega}\mathrm{ran}(a))$, N cannot learn both $W_{e_1(t_0)}$ and $W_{e_2(t_0)}$. Hence σ_0 exists.

By the choice of t_0 and σ_0 we obtain $E_{t_0} = \mathrm{content}(a[t_1])$ for $t_1 = \max(\{i_{\sigma_0}\} \cup \{\Phi_p(k, \sigma) \mid \sigma \leqslant_a \sigma_0\}) \in \mathbb{N}$. Let $\hat{t} = \max\{t_0, t_1\}$ and $L = \mathrm{content}(a[\hat{t}]) \cup \{b(t_0)\}$. Then $W_{e_1(t_0)} = W_{e_2(t_0)} = L \in \mathcal{L}$ and by construction of E_{t_0} we have $\mathbf{Cons}(\sigma_0, L)$. Because of $\hat{t} \geqslant t_0$, we obtain $s^*_N(a[\hat{t}]) = s^*_N(a[t_0])$. With this and the choice of t_0 we conclude $N(a[\hat{t}]^\frown b(t_0)^\frown \#^\ell) = N(a[t_0]^\frown b(t_0)^\frown \#^\ell)$ for all $\ell \in \mathbb{N}$. Further, as N learns L from the text $a[\hat{t}]^\frown b(t_0)^\frown \#^\infty$ we have $W_{N(a[\hat{t}]^\frown b(t_0)^\frown \#^{\ell_{t_0}})} = L$. On the other hand by $\mathrm{MC}(t_0, \sigma_0)$ we obtain $N(a[\hat{t}]^\frown b(t_0)^\frown \#^{\ell_{t_0}}) \neq N(a[\hat{t}]^\frown b(t_0)^\frown \#^{\ell_{t_0}\frown}\sigma_0)$, which forces N to perform a syntactic U-shape on the text $a[\hat{t}]^\frown b(t_0)^\frown \#^{\ell_{t_0}\frown}\sigma_0{}^\frown \#^\infty$ for L. □

For inferring the relations between the syntactic learning requirements **SNU**, **SDec** and **Conv**, we refer to **Wb**. All these criteria are closely related to strongly locking learners, which we define in the following.

It was observed by [BB75] that the learnability of every language L by a learner M is witnessed by a sequence σ, consistent with L, such that $M(\sigma)$ is an index for L and no extension of σ consistent with L will lead to a mind-change of M. Such a sequence σ is called *locking sequence for M on L*. For a similar purpose as ours [JKMS16] introduced strongly locking learners. A learner M acts strongly locking on a language L, if for every text T for L there is an initial segment σ of T that is a locking sequence for M on L.

The proof of the following proposition generalizes the construction of a conservative and strongly decisive iterative learner from a strongly locking iterative learner in [JKMS16, Theorem 8]. With it we obtain in the Corollary thereafter, that all non-semantic learning restrictions coincide.

Theorem 4.3. *Let $\mathcal{L}$ be a set of languages* $\mathbf{BMS}_*\mathbf{Ex}$*-learned by a strongly locking* $\mathbf{BMS}$*-learner. Then*

$$\mathcal{L} \in [\mathbf{TxtBMS}_*\mathbf{WbEx}].$$

Proof. Let $\mathcal{L} \in [\mathbf{TxtBMS}_*\mathbf{Ex}]$ be learned by the strongly locking learner M. By Lemma 4.1 we may assume that M does not return to withdrawn states.

We proceed in two steps. First we construct a learner M' conservatively $\mathbf{BMS}_*\mathbf{Ex}$-learning at least $\mathcal{L}$ in a strong sense, i.e.,

$$\forall \sigma \in \Sigma^{<\omega}\ \forall x \in \Sigma\ (\, M'(\sigma^\frown x) \neq M'(\sigma) \ \Rightarrow\ x \notin W_{M'(\sigma)}\,). \tag{1}$$

That we require the last datum to violate consistency with the former hypothesis fits the setting of **BMS**-learners and is also called locally conservative by [JLZ06]. Second, with such a learner at hand, we are going to construct a learner N which $\mathbf{BMS}_*\mathbf{Ex}$-learns $\mathcal{L}$ in a witness-based fashion.

For defining the strongly conservative learner M', we employ a one-one function $f : \mathbb{N} \times Q \to \Omega$ satisfying

$$W_{f(e,s)} = \bigcup_{t \in \mathbb{N}} \begin{cases} W_e^t, & \text{if } \forall x \in W_e^t (\, h_M(s, x) = e \ \wedge\ s_M(s, x) = s\,); \\ \varnothing, & \text{otherwise} \end{cases}$$

for every hypothesis $e \in \mathbb{N} \subseteq \Omega$ and state $s \in Q$. The existence of f is granted by the smn theorem. Thus, f takes into account only the initial part of W_e not necessary to possibly justify a mind-change or state-change later on. Now define for all $\sigma \in \Sigma^{<\omega}$

$$M'(\sigma) = f(M(\sigma), s^*_M(\sigma)).$$

As M never returns to withdrawn states and behaves strongly locking while $\mathbf{BMS_*Ex}$-learning $\mathcal{L}$, M' also $\mathbf{Ex}$-learns $\mathcal{L}$. For $\sigma \neq \epsilon$ the values of $M(\sigma)$ and $s^*_M(\sigma)$ only depend on $s^*_M(\sigma^-)$ and $\mathrm{last}(\sigma)$ and hence M' is a $\mathbf{BMS_*}$-learner with $s_{M'} = s_M$. Moreover, by construction it is conservative in the strong sense defined in (1).

We now define the witness-based learner N. In addition to thinning out the hypotheses of M', as we did with the hypotheses of M when constructing M' from M, we patch all data causing mind-changes to it. This data is stored in the states used by N. Further, we only alter our old hypothesis in case we can guarantee the existence of a witness justifying the possible mind-change. To do this in a computable way, we need to store also the last hypothesis of M' in the states of N.

For every datum $x \in \Sigma$, data-sequence $\sigma \in \Sigma^{<\omega}$, hypothesis $e \in \mathbb{N} \subseteq \Omega$ and every finite sequence MC of natural numbers, interpreted as pairs of hypotheses and data, we define a state transition function s_N, auxiliary hypothesis generating function h, recursive function $g : \mathbb{N}^2 \to \Omega$ and the learner N by

$$h(\langle s, \langle \mathrm{MC} \rangle\rangle, x) = \begin{cases} h_{M'}(s, \#), & \text{if } x \in \mathrm{pr}_2[\mathrm{MC}]; \\ h_{M'}(s, x), & \text{otherwise;} \end{cases}$$

$$s_N(\langle s, \langle \mathrm{MC} \rangle\rangle, x) = \begin{cases} \langle s_{M'}(s, \#), \langle \mathrm{MC} \rangle\rangle, & \text{if } x \in \mathrm{pr}_2[\mathrm{MC}] \ \wedge \ h_{M'}(s, \#) = \mathrm{pr}_1(\mathrm{last}(\mathrm{MC})); \\ \langle s_{M'}(s, \#), \langle \mathrm{MC}^\frown \langle h_{M'}(s, \#), \# \rangle\rangle\rangle, & \text{if } x \in \mathrm{pr}_2[\mathrm{MC}] \ \wedge \ h_{M'}(s, \#) \neq \mathrm{pr}_1(\mathrm{last}(\mathrm{MC})); \\ \langle s_{M'}(s, x), \langle \mathrm{MC} \rangle\rangle, & \text{else if } h_{M'}(s, x) = \mathrm{pr}_1(\mathrm{last}(\mathrm{MC})); \\ \langle s_{M'}(s, x), \langle \mathrm{MC}^\frown \langle h_{M'}(s, x), x \rangle\rangle\rangle, & \text{otherwise;} \end{cases}$$

$$W_{g(e, \langle s, \langle \mathrm{MC} \rangle\rangle)} = \mathrm{pr}_2[\mathrm{MC}] \cup W_e;$$

$$N(\sigma^\frown x) = \begin{cases} ?, & \text{if } h^*(\sigma^\frown x) = ?; \\ g(h^*(\sigma^\frown x), s^*_N(\sigma^\frown x)), & \text{else if } h^*(\sigma^\frown x) \neq \mathrm{pr}_1(\mathrm{last}(\mathrm{decode}(\mathrm{pr}_2(s^*_N(\sigma))))); \\ N(\sigma), & \text{otherwise.} \end{cases}$$

Thus with the help of g the data stored in the second coordinates of MC is patched to the language encoded in e. Further, N only makes a mind-change if h^* does, as $h^*(\sigma) = \mathrm{pr}_1(\mathrm{last}(\mathrm{decode}(\mathrm{pr}_2(s^*_N(\sigma)))))$. The learner h^* behaves like M' on the text, in which every datum repeatedly causing a mind-change is replaced by the pause symbol.

Let $L \in \mathcal{L}$ and $T \in \mathbf{Txt}(L)$. It is easy to see that for the text T' recursively defined by

$$T'(t) = \begin{cases} \#, & \text{if } \exists s < t \, (\, T(s) = T(t) \ \wedge \ M'(T'[s]^\frown T(s)) \neq M'(T'[s]) \,); \\ T(t), & \text{otherwise,} \end{cases}$$

holds $h^*(T[t]) = M'(T'[t])$ for all $t \in \mathbb{N}$. This follows with a simultaneous induction also showing $\mathrm{pr}_1(s^*_N(T[t])) = s^*_{M'}(T'[t])$. Hence h^* on T behaves like M' on $T' \in \mathbf{Txt}(L)$.

Because M' $\mathbf{Ex}$-converges on T', it makes only finitely many mind-changes and uses only finitely many states, which implies that N also only uses finitely many states. Let $e = M'(T'[t_0])$ be the final correct hypothesis of M' on T' with $t_0 \in \mathbb{N}$ chosen appropriately. Because M' never returns to withdrawn states, the states of N also stabilize. Moreover, $N(T[t_0])$ has to be correct since $\mathrm{pr}_2[\mathrm{MC}] \subseteq W_e$.

As already mentioned, N learns every $L \in \mathcal{L}$ witness-based because M' is strongly conservative. Every time N performs a mind-change on T, so does M' on T'. Therefore, there is a responsible datum x which was not in the former hypothesis of M' and also has not occured so far, as no datum in T' causes more than one mind-change. This datum x will be contained in all languages hypothesized by N in the future. □

With the former theorem it is straightforward to observe that in the $\mathbf{BMS_*Ex}$-setting conservative, strongly decisive and strongly non-U-shaped $\mathbf{Ex}$-learning are equivalent.

Corollary 4.4. *We have* $[\mathbf{TxtBMS_*ConvEx}] = [\mathbf{TxtBMS_*SDecEx}] = [\mathbf{TxtBMS_*SNUEx}]$.

Proof. On the one hand a conservative or strongly decisive learning behavior is also a strongly non-U-shaped learning behavior. On the other hand, a learner behaving strongly non-U-shaped proceeds strongly locking and, by Theorem 4.3, from a strongly locking learner we may construct a learner with at least equal learning power, acting witness-based and hence also conservatively and strongly decisively. □

By [JKMS16, Theorem 2] and Lemma 3.1 (1) we obtain

$$[\mathbf{TxtBMS_*ConvEx}] \nsubseteq [\mathbf{TxtBMS_*SMonEx}].$$

From this we conclude with Theorem 4.2 and Corollary 4.4 the following incomparability

$$[\mathbf{TxtBMS_*ConvEx}] \perp [\mathbf{TxtBMS_*SMonEx}].$$

Similarly, with [JKMS16, Theorem 3] and again Lemma 3.1 (1) we obtain $[\mathbf{TxtBMS_*ConvEx}] \nsubseteq [\mathbf{TxtBMS_*MonEx}]$. As Theorem 4.2 implies $[\mathbf{TxtBMS_*MonEx}] \nsubseteq [\mathbf{TxtBMS_*SNUEx}]$, with Corollary 4.4 follows

$$[\mathbf{TxtBMS_*ConvEx}] \perp [\mathbf{TxtBMS_*MonEx}].$$

Because Theorem 4.2 also reproves $[\mathbf{TxtBMS_*SNUEx}] \subsetneq [\mathbf{TxtBMS_*Ex}]$, first observed in [CK13, Th. 3.10], we completed the map for $\mathbf{BMS_*Ex}$-learning from texts. An overview is depicted in Figure 1.

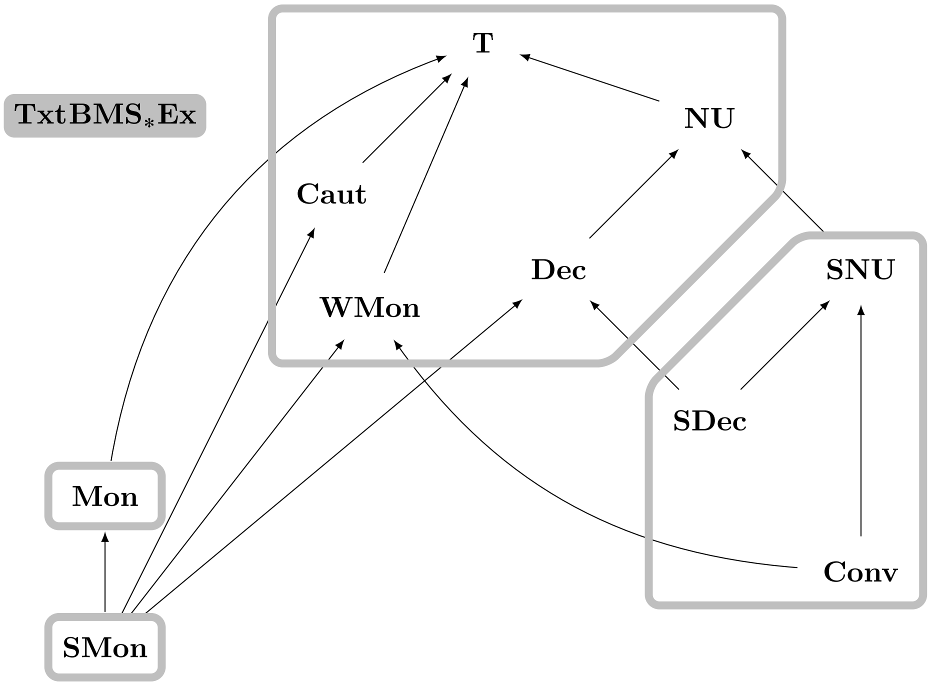

FIGURE 1. Relations between delayable learning restrictions in explanatory finitely bounded memory states learning of languages from informants. The arrows represent implications independent of the model. The outlined areas stand for equivalence classes with respect to learning power, when the underlying model is $\mathbf{TxtBMS_*Ex}$.

As this map equals the one for **It**-learning, naturally the question arises, whether a result similar to Lemma 3.1 can be observed for the syntactic learning criteria. In the following we show that this is not the case.

Theorem 4.5. $[\mathbf{ItTxtSNUEx}] \subsetneq [\mathbf{TxtBMS_*SNUEx}]$

Proof. Clearly, the inclusion holds. Similar to the proof of Lemma 4.2, we consider the **BMS**-learner M initialized with state $\langle\langle\,?,0\rangle,\langle\varnothing\rangle\rangle$ and h_M and s_M for every $\langle e,\xi\rangle \in \Omega$, $D \subseteq \mathbb{N}$ finite and $x \in \Sigma$ defined by:

$$s_M(\langle\langle e,\xi\rangle,\langle D\rangle\rangle,x) = \begin{cases} \langle\langle e,\xi\rangle,\langle D\rangle\rangle, & \text{if } x \in D \cup \{\#\} \ \vee\ \pi_1(\,\varphi_x(\langle e,\xi\rangle)\,) = e; \\ \langle\varphi_x(\langle e,\xi\rangle),\langle D \cup \{x\}\rangle\rangle, & \text{else if } \pi_1(\,\varphi_x(\langle e,\xi\rangle)\,) \neq e; \\ \uparrow, & \text{otherwise.} \end{cases}$$

$$h_M(\langle\langle e,\xi\rangle,\langle D\rangle\rangle,x) = \begin{cases} e, & \text{if } x \in D \cup \{\#\} \ \vee\ \pi_1(\,\varphi_x(\langle e,\xi\rangle)\,) = e; \\ \pi_1(\,\varphi_x(\langle e,\xi\rangle)\,), & \text{else if } \pi_1(\,\varphi_x(\langle e,\xi\rangle)\,) \neq e; \\ \uparrow, & \text{otherwise.} \end{cases}$$

Additionally to the last hypothesis as well as exactly the data that already lead to a mind change of M, some parameter $\xi \in \{0,1,2\}$ is stored, indicating whether a further mind-change may cause a syntactic U-shape.

Let $\mathcal{L} = \mathbf{TxtBMS_*SNUEx}(M)$.

Assume there is an iterative learner N with hypothesis generating function h_N and $\mathcal{L} \subseteq \mathbf{ItTxtSNUEx}(N)$.

We obtain $L \in \mathcal{L}\backslash\mathbf{ItTxtSNUEx}(N)$ by applying 1-1 ORT referring to the Σ_1-predicates MC and NoMC, expressing that N does (not) perform a mind-change on a text built from parameters $a, b \in \mathcal{R}$. More specifically, the predicates state that N does converge and (not) make a mind-change when observing $\sigma \in \Sigma^{<\omega}$ after having observed $a[k]^\frown b(k)^\frown\#^{\ell_k}$, with $k \in \mathbb{N}$.

$$\begin{aligned} \psi_k(\ell) &\Leftrightarrow N(a[k]^\frown b(k)^\frown\#^{\ell}) = N(a[k]^\frown b(k)^\frown\#^{\ell+1}); \\ \mathrm{NoMC}(k,\sigma) &\Leftrightarrow \exists \ell_k \in \mathbb{N}\,(\,\psi_k(\ell_k)\ \wedge\ \forall \ell < \ell_k\ \neg\psi_k(\ell)\ \wedge\ N(a[k]^\frown b(k)^\frown\#^{\ell_k\frown}\sigma)\!\downarrow\, = N(a[k]^\frown b(k)^\frown\#^{\ell_k})\,); \\ \mathrm{MC}(k,\sigma) &\Leftrightarrow \exists \ell_k \in \mathbb{N}\,(\,\psi_k(\ell_k)\ \wedge\ \forall \ell < \ell_k\ \neg\psi_k(\ell)\ \wedge\ N(a[k]^\frown b(k)^\frown\#^{\ell_k\frown}\sigma)\!\downarrow\, \neq N(a[k]^\frown b(k)^\frown\#^{\ell_k})\,). \end{aligned}$$

By 1-1 ORT there are $a, b, e_1, e_2 \in \mathcal{R}$ with pairwise disjoint ranges and $e_0 \in \mathbb{N}$, such that

$$\varphi_{a(i)}(\langle e,\xi\rangle) = \begin{cases} \langle e_0,\xi\rangle, & \text{if } e \in \{?, e_0\}; \\ \langle e_1(k),1\rangle, & \text{else if } e = e_1(k) \text{ for some } k \leqslant i \text{ and } \xi = 0 \text{ and } i \text{ even;} \\ \langle e_1(k),2\rangle, & \text{else if } e = e_1(k) \text{ for some } k \leqslant i \text{ and } \xi = 0 \text{ and } i \text{ odd;} \\ \langle e_2(k),0\rangle, & \text{else if } e = e_1(k) \text{ for some } k \leqslant i \text{ and } \xi = 1 \text{ and } i \text{ odd;} \\ \langle e_2(k),0\rangle, & \text{else if } e = e_1(k) \text{ for some } k \leqslant i \text{ and } \xi = 2 \text{ and } i \text{ even;} \\ \langle e,\xi\rangle, & \text{otherwise;} \end{cases}$$

$$\varphi_{b(k)}(\langle e,\xi\rangle) = \begin{cases} \langle e_1(k),\xi\rangle, & \text{if } e \in \{?, e_0\}; \\ \langle e,\xi\rangle, & \text{otherwise;} \end{cases}$$

$$W_{e_0} = \begin{cases} \mathrm{ran}(a[t_0]), & \text{if } t_0 \text{ is minimal with } \forall t \geqslant t_0\, N(a[t]) = N(a[t_0]); \\ \mathrm{ran}(a), & \text{no such } t_0 \text{ exists;} \end{cases}$$

$$W_{e_1(k)} = \mathrm{content}(a[k]) \cup \{b(k)\} \cup \begin{cases} \{a(i_k)\} & \text{if } \exists i_k \geqslant k \text{ first found } (\mathrm{MC}(k, a(i_k)); \\ \varnothing, & \text{otherwise;} \end{cases}$$

$$W_{e_2(k)} = \mathrm{ran}(a) \cup \{b(k)\}.$$

As the learner constantly puts out e_0 on every text for W_{e_0}, we have $W_{e_0} \in \mathcal{L}$. Thus, also N has to learn the finite language W_{e_0} and t_0 exists. Note that by the iterativeness of N we obtain $N(a[t_0]) = N(a[t_0]^\frown a(i))$ for all $i \geqslant t_0$ and with this

$$N(a[t_0]^\frown b(t_0)^\frown\#^{\ell_{t_0}}) = N(a[t_0]^\frown a(i)^\frown b(t_0)^\frown\#^{\ell_{t_0}}) \text{ for all } i \geqslant t_0.$$

$W_{e_1(t_0)}$ and $W_{e_2(t_0)}$ also lie in $\mathcal{L}$. To see that M explanatory learns both of them, note that, after having observed $b(t_0)$, M only changes its mind from $e_1(t_0)$ to $e_2(t_0)$ after having seen $a(i)$ and $a(j)$ with $i, j \geqslant t_0$ and $i \in 2\mathbb{N}$ as well as $j \in 2\mathbb{N}+1$. This clearly happens for every text for the infinite language $W_{e_2(t_0)}$. As $|W_{e_1(t_0)}\backslash(\mathrm{content}(a[t_0]) \cup \{b(t_0)\})\,| \leqslant 1$, this mind change never occurs for any text for $W_{e_1(t_0)}$. The syntactic non-U-shapedness of the learning processes can be easily seen as for all $k, l \in \mathbb{N}$ the languages W_{e_0}, $W_{e_1(k)}$ and $W_{e_2(l)}$ are pairwise distinct and the learner never returns to an abandoned hypothesis.

Next, we show the existence of $i_{t_0} \geqslant t_0$ with $\mathrm{MC}(t_0, a(i_{t_0}))$. Assume towards a contradiction that i_{t_0} does not exist. Hence, it holds $W_{e_1(t_0)} = \mathrm{content}(a[t_0]) \cup \{b(t_0)\}$. As M learns this language from the text $a[t_0]^\frown b(t_0)^\frown\#^\infty$, so does N. The convergence of N implies the existence of ℓ_{t_0}. Thus, for every $i \in \mathbb{N}$ we have $N(a[t_0]^\frown b(t_0)^\frown\#^{\ell_{t_0}\frown}a(i))\uparrow$ or $N(a[t_0]^\frown b(t_0)^\frown\#^{\ell_{t_0}\frown}a(i)) = N(a[t_0]^\frown b(t_0)^\frown\#^{\ell_{t_0}})$. Because N is iterative and learns $W_{e_2(t_0)}$, it may not be undefined and therefore always the latter is the case. But then N will not learn

$W_{e_1(t_0)}$ and $W_{e_2(t_0)}$ as they are different but it does not make a mind-change on the text $a[t_0]^\frown b(t_0)^\frown \#^{\ell_{t_0}}{}^\frown a$ after having observed the initial segment $a[t_0]^\frown b(t_0)^\frown \#^{\ell_{t_0}}$, due to its iterativeness. Hence i_{t_0} exists.

By the choice of i_{t_0}, the learner N does perform a syntactic U-shape on the following text for $W_{e_1(t_0)}$

$$a[t_0]^\frown a(i_{t_0})^\frown b(t_0)^\frown \#^{\ell_{t_0}}{}^\frown a(i_{t_0})^\frown \#^\infty.$$

More precisely, t_0 and ℓ_{t_0} were chosen such that $N(a[t_0]^\frown a(i_{t_0})^\frown b(t_0)^\frown \#^{\ell_{t_0}})$ has to be correct and the characterizing property of i_{t_0} assures $N(a[t_0]^\frown a(i_{t_0})^\frown b(t_0)^\frown \#^{\ell_{t_0}}) \neq N(a[t_0]^\frown a(i_{t_0})^\frown b(t_0)^\frown \#^{\ell_{t_0}}{}^\frown a(i_{t_0}))$. This is a contradiction to $W_{e_1(t_0)} \in \mathbf{ItTxtSNUEx}(N)$. Thus no iterative learner can explanatory syntactically non-U-shapedly learn $\mathcal{L}$. □

Note that by Corollary 4.4 we also obtain $[\mathbf{ItTxtSDecEx}] \subsetneq [\mathbf{TxtBMS_*SDecEx}]$ and $[\mathbf{ItTxtConvEx}] \subsetneq [\mathbf{TxtBMS_*ConvEx}]$.

5. Related Open Problems

We have given a complete map for learning with bounded memory states, where, on the way to success, the learner must use only finitely many states. Future work can address the complete maps for learning with an a priori bounded number of memory states, which needs very different combinatorial arguments. Results in this regard can be found in [CCJS07] and [CK13]. We expect to see trade-offs, for example allowing for more states may make it possible to add various learning restrictions (just as non-deterministic finite automata can be made deterministic at the cost of an exponential state explosion).

Also memory-restricted learning from positive and negative data (so-called informant) has only partially been investigated for iterative learners and to our knowledge not at all for other models of memory-restricted learning. Very interesting also in regard of 1-1 hypothesis spaces that prevent coding tricks is the **Bem**-hierarchy, see [FJO94], [LZ96] and [CJLZ99].

In the spirit of grammatical inference, we encourage to investigate the learnability of carefully chosen indexable families arising from applied machine learning or cognitive science research.

Acknowledgements. This work was supported by DFG Grant Number KO 4635/1-1. We are grateful to the people supporting us.

References

[AKS18] M. Aschenbach, T. Kötzing, and K. Seidel. Learning from informants: Relations between learning success criteria. *arXiv preprint arXiv:1801.10502*, 2018.

[Ang80] D. Angluin. Inductive inference of formal languages from positive data. *Information and control*, 45(2):117–135, 1980.

[BB75] L. Blum and M. Blum. Toward a mathematical theory of inductive inference. *Information and Control*, 28:125–155, 1975.

[BCM+08] G. Baliga, J. Case, W. Merkle, F. Stephan, and R. Wiehagen. When unlearning helps. *Information and Computation*, 206:694–709, 2008.

[CCJS07] L. Carlucci, J Case, S. Jain, and F. Stephan. Results on memory-limited U-shaped learning. *Information and Computation*, 205:1551–1573, 2007.

[CJLZ99] J. Case, S. Jain, S. Lange, and T. Zeugmann. Incremental concept learning for bounded data mining. *Information and Computation*, 152:74–110, 1999.

[CK10] J. Case and T. Kötzing. Strongly non-U-shaped learning results by general techniques. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 181–193. Omnipress, 2010.

[CK13] John Case and Timo Kötzing. Memory-limited non-u-shaped learning with solved open problems. *Theoretical Computer Science*, 473:100–123, 2013.

[CM08a] J. Case and S. Moelius. U-shaped, iterative, and iterative-with-counter learning. *Machine Learning*, 72:63–88, 2008.

[CM08b] J. Case and S. E. Moelius. U-shaped, iterative, and iterative-with-counter learning. *Machine Learning*, 72:63–88, 2008.

[CM11] J. Case and S. Moelius. Optimal language learning from positive data. *Information and Computation*, 209:1293–1311, 2011.

[FJO94] M. Fulk, S. Jain, and D. Osherson. Open problems in Systems That Learn. *Journal of Computer and System Sciences*, 49(3):589–604, December 1994.

[Gol67] E. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[Jan91] 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.

[JKMS16] 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.

[JLZ06] 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.

[JMZ13] S. Jain, S. Moelius, and S. Zilles. Learning without coding. *Theoretical Computer Science*, 473:124–148, 2013.

[JORS99] S. Jain, D. Osherson, J. Royer, and A. Sharma. *Systems that Learn: An Introduction to Learning Theory.* MIT Press, Cambridge, Massachusetts, second edition, 1999.

[KP16] T. Kötzing and R. Palenta. A map of update constraints in inductive inference. *Theoretical Computer Science*, 650:4–24, 2016.

[KS16] T. Kötzing and M. Schirneck. Towards an atlas of computational learning theory. In *33rd Symposium on Theoretical Aspects of Computer Science*, 2016.

[LZ96] S. Lange and T. Zeugmann. Incremental learning from positive data. *Journal of Computer and System Sciences*, 53:88–103, 1996.

[Odi99] P. Odifreddi. *Classical Recursion Theory*, volume II. Elsivier, Amsterdam, 1999.

[OSW82] D. Osherson, M. Stob, and S. Weinstein. Learning strategies. *Information and Control*, 53:32–51, 1982.

[OSW86] 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.

[RC94] J. Royer and J. Case. *Subrecursive Programming Systems: Complexity and Succinctness.* Research monograph in *Progress in Theoretical Computer Science.* Birkhäuser Boston, 1994.

[Wie91] R. Wiehagen. A thesis in inductive inference. In *Nonmonotonic and Inductive Logic, 1st International Workshop, Proc.*, pages 184–207, 1991.