Normal Forms in Semantic Language Identification

Timo Kötzing TIMO.KOETZING@HPI.DE
Martin Schirneck MARTIN.SCHIRNECK@HPI.DE
Karen Seidel KAREN.SEIDEL@HPI.DE
Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3
D-14482 Potsdam

Editor: Steve Hanneke and Lev Reyzin

Abstract

We consider language learning in the limit from text where all learning restrictions are semantic, that is, where any conjecture may be replaced by a semantically equivalent conjecture. For different such learning criteria, starting with the well-known **TxtGBc**-learning, we consider three different normal forms: strongly locking learning, consistent learning and (partially) set-driven learning. These normal forms support and simplify proofs and give insight into what behaviors are necessary for successful learning (for example when consistency in conservative learning implies cautiousness and strong decisiveness).

We show that strongly locking learning can be assumed for partially set-driven learners, even when learning restrictions apply. We give a very general proof relying only on a natural property of the learning restriction, namely, allowing for simulation on equivalent text. Furthermore, when no restrictions apply, also the converse is true: every strongly locking learner can be made partially set-driven. For several semantic learning criteria we show that learning can be done consistently. Finally, we deduce for which learning restrictions partial set-drivenness and set-drivenness coincide, including a general statement about classes of infinite languages. The latter again relies on a simulation argument.

Keywords: behaviorally correct learning, language identification in the limit, learning restriction, normal form, semantic learning

1. Introduction

Gold (1967) introduced a framework for the learning of languages as a branch of algorithmic learning theory. It has since been called *inductive inference* or *learning in the limit.* This branch analyzes the problem of algorithmically learning a description for a formal language (a computably enumerable subset of the natural numbers) when presented successively all and only the elements of that language. 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 description 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 himself gave a first, simple learning criterion, **TxtGEx**-

learning[1], where a learner is successful iff on every text for L it will eventually pick a conjecture which it keeps forever, and this final conjecture is a correct description for L. We will use W-indices as conjectures, that is, any conjecture is a natural number interpreted as a program which describes the set of numbers on which the program terminates. Trivially, each single, describable language L has a constant function as a **TxtGEx**-learner (this learner constantly outputs a description for L). Thus, we are interested in analyzing for which *classes of languages* $\mathcal{L}$ there is a *single learner* h learning *each* member of $\mathcal{L}$. This framework has been studied extensively, using a wide range of learning criteria similar to **TxtGEx**-learning (see, for example, the textbook by Jain et al., 1999).

Note that Gold required for successful learning that the learner stops changing the conjecture eventually, keeping a final correct conjecture. This imposes a *syntactic* restriction on the learning process: not only must the hypotheses eventually be correct, but it also must not change any more. There is also a purely *semantic* version of this restriction called *behaviorally correct* (**Bc**) learning of languages by Case and Lynes (1982) and Osherson and Weinstein (1982), where a learner h is successful on a language L iff on every text for L it eventually only outputs conjectures which are correct descriptions for L. This is an important distinction when target languages can have multiple descriptions, and deciding the equivalence of two descriptions is not computable (as is the case with W-indices).

There are a number of desirable properties one might want to have in a learner. For example, the learner can be required to be *consistent* with the data seen so far, that is, any conjecture contains the data it is based on (introduced by Angluin, 1980). Consistency is another example for a semantic restriction: any specific conjecture is just as good as any semantically equivalent conjecture. An example for an extensively studied learning restriction which is *not* semantic is *conservativeness* (also by Angluin, 1980), the restriction of not *syntactically* changing the conjecture while it is still consistent with the data. In this paper we will only consider semantic restrictions, of which we give more below.

When working with learners, it is often desirable that they behave in some normalized way, that they fulfill some simple assumptions. The most famous normal form for learners is the *Fulk Normal Form* (Fulk, 1990). He showed that every **TxtGEx**-learner can be assumed to have a number of additional properties. Among other things, the Fulk Normal Form requires the learner to be *strongly locking*, that is, on any text there is a point such that enough data was presented and the learner will not change its mind regardless of what data from the target language is presented. Such a learner would truly converge and would not be ready to change the conjecture any more for a certain continuation of the data. For **TxtGBc**-learning, the corresponding notion is that of *strongly* **Bc**-*locking*, where syntactic mind changes are still allowed as long as they are not semantic. See Jain et al. (1999) for the generalized concept of locking sequences in other contexts than **Ex**-style convergence.

Another part of the Fulk Normal Form is *partial set-drivenness* or *rearrangement independence*. This describes the situation in which a learner is insensitive to the exact sequence of the input data and merely depends on the set of data presented and the length of the input sequence. We speak of *set-driven* learners, introduced by Wexler and Culicover (1980), when their output depends *only* on the data. This second regularity property is desirable in that the exact order of presentation does not have an impact on the learning process.

[1] **Txt** stands for learning from a *text* of positive examples; **G** stands for Gold, who introduced this model, and is used to indicate full-information learning; **Ex** stands for *explanatory*.

For **TxtGEx**-learning, it is well-known that partial set-drivenness is not a restriction, but interestingly it cannot be assumed to be set-driven (see Schäfer-Richter, 1984; Fulk, 1985).

We mentioned consistency above. It would support the learning process if the learner only makes conjectures that at least contain the data seen so far. However, for **TxtGEx**-learning, consistency is a restriction of the learning power (Bārzdiņš, 1977); this is called the *inconsistency phenomenon* and follows from **TxtGEx**-learning requiring syntactic convergence. If we relax our criterion to **TxtGBc**, we can trivially see that every **TxtGBc**-learnable class can be **TxtGBc**-learned consistently (**TxtGConsBc**-learned for short). Every new hypothesis is generated by *patching in* the current data. This requires a change of the conjecture in every step, but for correct conjectures this change is only syntactic.

Our goal is to study the mentioned three normal forms, (1) strongly **Bc**-locking, (2) (partial) set-drivenness and (3) consistency in the area of **TxtGBc**-learning. For several semantic restrictions δ, we discuss δ-restricted **TxtGBc**-learning (**TxtG**$\boldsymbol{\delta}$**Bc**-learning).

Before we consider the different normal forms, we introduce notation and learning criteria in Section 2, followed by a general presentation of semantic learning in Section 3. In Section 4 we see that, surprisingly, strongly **Bc**-locking is equivalent to partial set-drivenness. Furthermore, for many learning restrictions δ, we can assume δ-restricted partially set-driven learning to be strongly locking, and any set-driven learner is always strongly **Bc**-locking. These last two statements are very general: They apply to all learning restrictions δ we consider in this paper and a wealth of further restrictions. Finally, we also see that every non-U-shaped (**NU**) learner is automatically strongly locking.

In Section 5 we consider consistency. While it is trivial that pure **TxtGBc**-learning (without any further learning restriction) allows for consistent learning, the situation is much less clear when an additional learning restriction δ is considered. We were able to show that many important semantic learning restrictions δ allow for consistent learning. Our results in this section are specific to the respective restrictions and employ proofs tailored to the particular structure of the different δ, since most restrictions do not allow for simple patching of the hypotheses as in the case of **TxtGBc**.

In Section 6 we reveal when a partially set-driven learner can be assumed to be set-driven. Here consistency serves as a normal form supporting our proofs. It is known that, for classes $\mathcal{L}$ of infinite languages, **TxtGEx**-learning can be made set-driven (Osherson et al., 1986). We pick up the idea of this proof to show, for a wide range of learning criteria, partially set-driven learning can be made set-driven for classes of infinite languages.

Wherever possible we give our results not just for concrete learning criteria, but for any learning criteria fulfilling some natural axiom. This makes the theorem more versatile and applicable also to learning criteria not yet invented, as long as they fulfill the axiom.

We conclude in Section 7. Note that, due to space considerations, some proofs are omitted in the main part of the paper and can be found in the appendix.

2. Language Learning in the Limit

In this section we give a brief overview over mathematical details used in this paper, focusing on less standard notation. While full mathematical prelims can be found in Appendix A, most notation is standard and follows the textbooks by Rogers (1967) and Jain et al. (1999), except for the notation on learning criteria, which follows Kötzing (2009).

A *text* for a language L is an infinite sequence of all and only the elements of L, with possible additional listings of the *pause symbol* #. A *learner* is any partial computable function $h \in \mathcal{P}$. An *interaction operator* is an operator β taking as arguments a function $h \in \mathcal{P}$ (the learner) and a text $T \in \mathbf{Txt}$, and outputs a (possibly partial) function $p \in \mathfrak{P}$, the sequence of conjectures. The most common interaction operator is $\mathbf{G}$, which is defined such that, for all learners h, texts T and indices i, $\mathbf{G}(h, T)(i) = h(T[i])$. Partial set-driven ($\mathbf{Psd}$), set-driven ($\mathbf{Sd}$), and iterative ($\mathbf{It}$) learning can be defined analogously. A learner h is said to be *confluently iterative* ($\mathbf{CflIt}$) just in case it is an iterative learner which gives the same output on any two sequences of inputs that contain the exact same data. In this work, whenever we refer to *all interaction operators* we mean those we just defined.

One can establish a hierarchy among the interaction operators by noticing that some can be simulated by others (Case and Kötzing, 2016). For two interaction operators β, β', we say *β-learners can be translated into β'-learners*, written $\beta \preccurlyeq \beta'$, if, for every learner h, there is some learner h' such that on arbitrary texts T the resulting sequence of hypotheses of h working on T is the same as that of h'. That is, $\forall T \in \mathbf{Txt}\colon \beta(h, T) = \beta'(h', T)$. For example, an $\mathbf{Sd}$-learner can be translated into an $\mathbf{Psd}$-learner by simply ignoring the additional information of the number of the current iteration.

A *learning restriction* is a predicate δ on a total learning sequence $p \in \mathfrak{R}$ and a text $T \in \mathbf{Txt}$. For example, for all pairs (p, T),

$$\mathbf{Ex}(p, T) \Leftrightarrow p \in \mathfrak{R} \wedge (\exists n_0\colon \forall n \geq n_0 : p(n) = p(n_0) \wedge W_{p(n_0)} = \mathrm{content}(T)).$$

We consider a number of further restrictions in this work.

$$\begin{aligned}
\mathbf{Bc}(p, T) &\Leftrightarrow \exists n_0\, \forall n \geq n_0\colon W_{p(n)} = \mathrm{content}(T);\\
\mathbf{Conv}(p, T) &\Leftrightarrow \forall i, j\colon (i \leq j \wedge \mathrm{content}(T[j]) \subseteq W_{p(i)}) \Rightarrow p(i) = p(j);\\
\mathbf{Caut}(p, T) &\Leftrightarrow \forall i, j\colon W_{p(i)} \subset W_{p(j)} \Rightarrow i \leq j;\\
\mathbf{Dec}(p, T) &\Leftrightarrow \forall i, j, k\colon (i \leq j \leq k \wedge W_{p(i)} = W_{p(k)}) \Rightarrow W_{p(i)} = W_{p(j)};\\
\mathbf{SDec}(p, T) &\Leftrightarrow \forall i, j, k\colon (i \leq j \leq k \wedge W_{p(i)} = W_{p(k)}) \Rightarrow p(i) = p(j);\\
\mathbf{Mon}(p, T) &\Leftrightarrow \forall i, j\colon i \leq j \Rightarrow W_{p(i)} \cap \mathrm{content}(T) \subseteq W_{p(j)} \cap \mathrm{content}(T);\\
\mathbf{SMon}(p, T) &\Leftrightarrow \forall i, j\colon i \leq j \Rightarrow W_{p(i)} \subseteq W_{p(j)};\\
\mathbf{WMon}(p, T) &\Leftrightarrow \forall i, j\colon (i \leq j \wedge \mathrm{content}(T[j]) \subseteq W_{p(i)}) \Rightarrow W_{p(i)} \subseteq W_{p(j)}.\\
\mathbf{NU}(p, T) &\Leftrightarrow \forall i, j, k\colon (i \leq j \leq k \wedge W_{p(i)} = W_{p(k)} = \mathrm{content}(T)) \Rightarrow W_{p(i)} = W_{p(j)};\\
\mathbf{SNU}(p, T) &\Leftrightarrow \forall i, j, k\colon (i \leq j \leq k \wedge W_{p(i)} = W_{p(k)} = \mathrm{content}(T)) \Rightarrow p(i) = p(j);
\end{aligned}$$

Finally, we let $\mathbf{T}$ denote the absence of any restriction. We combine restrictions δ and δ' by intersecting them and denote this by juxtaposition. A *learning criterion* is a tuple $(\alpha, \mathcal{C}, \beta, \delta)$, where $\mathcal{C}$ is a set of admissible learners (typically $\mathcal{P}$ or $\mathcal{R}$), β is an interaction operator, and α, δ are learning restrictions. We write $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta$ to denote this learning criterion, omitting $\mathcal{C}$ in case of $\mathcal{C} = \mathcal{P}$ and the restrictions if they equal $\mathbf{T}$. Let $h \in \mathcal{C}$ be an admissible learner. We say that learner h *$\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta$-learns* a language L iff on *arbitrary* texts $T \in \mathbf{Txt}$, $\alpha(\beta(h, T), T)$ holds, and for all texts $T \in \mathbf{Txt}(L)$, $\delta(\beta(h, T), T)$ holds. The class of languages $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta$-learned by h is denoted by $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta(h)$. Finally, we write $[\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta]$ to denote the set of all $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta$-learnable classes of languages.

3. Semantic Learning

The main idea of semantic language learning is to consider only the conjectured languages during the learning process, not their syntactic representation. The learner may be subject to restrictions regarding the languages it proposes, but it is free to encode its guesses by hypotheses it sees fit. Of course, a common hypothesis space is a necessary foundation for intelligible conjectures.

In this paper we consider semantic learning restrictions. Intuitively, a semantic restriction allows for replacement of a hypothesis by another that describes the same language. A pseudo-semantic restriction has the additional constraint that no new syntactic mind change may be introduced by this replacement. The following definitions were first given by Kötzing (2014). For a partial function $p \in \mathfrak{P}$ we fix the two sets

$$\mathbf{Sem}(p) = \{p' \in \mathfrak{P} \mid \forall n\colon (p(n)\!\downarrow \Leftrightarrow p'(n)\!\downarrow) \wedge (p(n)\!\downarrow \Rightarrow W_{p'(n)} = W_{p(n)})\};$$
$$\mathbf{Mc}(p) = \{p' \in \mathfrak{P} \mid \forall n\colon p(n)\!\downarrow = p(n+1)\!\downarrow \Rightarrow p'(n)\!\downarrow = p'(n+1)\!\downarrow\}.$$

A learning restriction δ is said to be *semantic* if for any sequence p and text T, $(p,T) \in \delta$ and $p' \in \mathbf{Sem}(p)$ implies $(p',T) \in \delta$. A restriction is said to be *pseudo-semantic* if this implication holds for all $p' \in \mathbf{Sem}(p) \cap \mathbf{Mc}(p)$. All restrictions defined above are pseudo-semantic and all but **Conv**, **SDec**, **SNU**, and **Ex** are also semantic.

The natural success criterion for semantic learning is behaviorally correct identification (**Bc**), where we require the learner to eventually only output correct hypotheses for the language it infers. The reason we disregard restrictions that are only pseudo-semantic but not semantic is that those cannot utilize the full potential of behaviorally correct learning. In fact, even when paired with **Bc** these restrictions fall back to **Ex**-convergence.

Proposition 1 *If $\delta \in \{\mathbf{Conv}, \mathbf{SDec}, \mathbf{SNU}, \mathbf{Ex}\}$, then $\delta\mathbf{Bc} = \delta\mathbf{Ex}$.*

As stated in Section 2, we require a successful learner to be always defined on texts for languages it is able to identify. However, it might as well diverge on texts for languages it cannot learn in the first place. The next theorem shows that in semantic learning it is sufficient to focus on total functions as admissible learners.

Theorem 2 *For any interaction operator β and semantic restriction δ holds*

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

3.1. Semantic Closure

We define the *semantic closure* of a learning restriction δ as the set

$$\mathbf{SemCl}(\delta) = \{(p',T) \mid p' \in \mathbf{Sem}(p) \wedge (p,T) \in \delta\}.$$

We augment the collection of sequences observing δ (with respect to some text) with all sequences that are semantically equivalent. It is immediate that $\delta \subseteq \mathbf{SemCl}(\delta)$ and that δ

is semantic iff we have equality. The closures of **SDec**, **SNU**, and **Ex** are already among the criteria mentioned in Section 2; namely, they are **Dec**, **NU**, and **Bc**, respectively. Regarding the last pseudo-semantic restriction, **Conv**, we introduce *semantic conservativeness* as the semantic closure **SemCl**(**Conv**). In more detail, we define for any $p \in \mathfrak{R}$ and $T \in \mathbf{Txt}$,

$$\mathbf{SemConv}(p, T) \Leftrightarrow \forall i, j \colon (i \leq j \wedge \mathrm{content}(T[j]) \subseteq W_{p(i)}) \Rightarrow W_{p(i)} = W_{p(j)}.$$

Already on the level of predicates every semantically conservative sequence is also weakly monotone, $\mathbf{SemConv} \subseteq \mathbf{WMon}$.

It was shown by Kötzing and Palenta (2016) that, for **Ex**-learning, cautiousness (**Caut**) and **WMon** are equivalent on the level of learnable classes. To study these restrictions also in the semantic setting, we adopt their approach of defining new intermediate criteria as common upper and lower bounds. **Bc**-learning requires that the learner eventually conjectures the target language. Consequently, a cautious **Bc**-learner never guesses a proper superset of the target. We follow Kötzing and Palenta (2016) in calling this behavior *target cautiousness*,

$$\mathbf{Caut_{Tar}}(p, T) \Leftrightarrow \forall i \colon \neg(W_{p(i)} \supset \mathrm{content}(T)).$$

We have already seen the inclusion $\mathbf{Caut} \cap \mathbf{Bc} \subseteq \mathbf{Caut_{Tar}}$. Observe that we also have $\mathbf{WMon} \subseteq \mathbf{Caut_{Tar}}$. As a common lower bound for **Caut** and **SemConv**, we define a learning sequence to be *semantically witness-based* by

$$\mathbf{SemWb}(p, T) \Leftrightarrow \forall i, k \colon (\exists j \colon (i \leq j \leq k \wedge W_{p(i)} \neq W_{p(j)})) \Rightarrow (\mathrm{content}(T[k]) \cap W_{p(k)}) \setminus W_{p(i)} \neq \emptyset.$$

Intuitively, if there has been a semantic mind change between the positions i and k, then there must be an observation in content$(T[k])$, which is now explained by the hypothesis $p(k)$, but has not been explained prior by $p(i)$. This "witness" justifies the mind change at position j. $\mathbf{SemWb} \subseteq \mathbf{Caut} \cap \mathbf{SemConv}$ is easily verified. As a side note, **SemWb** is in fact the semantic closure of what Kötzing and Schirneck (2016) defined as *witness-based*.

3.2. Semantic Equivalence of Interaction Operators

We weaken the notion of translating one learner into another (see Section 2) to adjust to the needs of semantic learning. For two interaction operators β, β', we say *β-learners can be translated semantically into β'-learners*, written $\beta \preccurlyeq_{\text{sem}} \beta'$, if for every β-learner h, there is a β'-learner h' such that for all texts T and positions n

$$W_{\beta(h,T)(n)} = W_{\beta'(h',T)(n)}.$$

Note that this entails the output $\beta'(h', T)(n)$ being defined if and only if $\beta(h, T)(n)$ is. Learner h' working on some text may give different hypotheses than h, but they describe the same sequence of recursively enumerable sets, the same semantics. If the translation is possible in both ways, we write $\beta \cong_{\text{sem}} \beta'$ and say that β and β' are *semantically equivalent*. The importance of $\preccurlyeq_{\text{sem}}$ stems from the fact that, if restrictions α and δ are semantic, $\beta \preccurlyeq_{\text{sem}} \beta'$ implies $[\tau(\alpha)\mathbf{Txt}\beta\delta] \subseteq [\tau(\alpha)\mathbf{Txt}\beta'\delta]$.

Theorem 3 $\mathbf{G} \cong_{\text{sem}} \mathbf{It}$*,* $\mathbf{Sd} \cong_{\text{sem}} \mathbf{CflIt}$*.*

The result follows from a padding argument. The output of the learner is used not only to describe a conjectured language, but also to encode information about the input seen so far. As a consequence of Theorem 2 and Theorem 3 we only need to consider full-information, partially set-driven, and set-driven learners which are defined on all inputs. All three interaction operators have in common that the learner at least has access to the content of the input sequence, the set of all data points presented so far.

4. Strongly Locking Learners

A locking sequence for a learner on some language encapsulates sufficient information for the learner to uniquely identify this language. The following definition for **Bc**-learning was given by Jain et al. (1999). Let $h \in \mathcal{P}$ be a **G**-learner. A sequence $\sigma \in \mathbb{S}\mathrm{eq}(L)$ is a **Bc**-*locking sequence* for h on L if for every sequence $\tau \in \mathbb{S}\mathrm{eq}(L)$, $W_{h(\sigma \diamond \tau)} = L$. Not only does h infer a correct description of the target language on σ itself (choose $\tau = \varepsilon$ as the empty sequence); moreover, no extension of σ with data from L makes the learner change its mind semantically. Of course, every such extension of σ is a locking sequence as well. The transfer to (partially) set-driven learning is immediate. A finite set $D \subseteq L$ is a **Bc**-*locking set* for h on L if for all $D \subseteq D' \subseteq L$, $W_{h(D')} = L$. A pair (D, t) with $t \geq |D|$ is a **Bc**-*locking information* if for all D' as before and numbers t' such that $t' - t \geq |D' \backslash D|$, $W_{h(D',t')} = L$. We also use the term locking information to subsume all three concepts.

A central observation of Blum and Blum (1975) is that every learner has a locking sequence for every language it **Ex**-learns. The same holds for **Bc**-convergence (the proof is identical). However, it is well-known that there are learners such that no initial sequence of a text serves as a locking sequence. We show in this section that in some cases we can forgo learners demonstrating this undesired behavior. Following Kötzing and Palenta (2016), we call a learner h *strongly* **Bc**-*locking* on some language L, if for every text $T \in \mathbf{Txt}(L)$ there is a position n_0 such that the sequence $T[n_0]$, the set $\mathrm{content}(T[n_0])$ or the pair $(\mathrm{content}(T[n_0], n_0)$, respectively, is a **Bc**-locking information for h on L. We say h is *strongly* **Bc**-*locking* if it is strongly **Bc**-locking on every language it learns.

Another very useful result carries over from **Ex**-learning (see Case and Kötzing, 2016).

Theorem 4 *Every* **Sd**-*learner is strongly* **Bc**-*locking.*

Partially set-driven learners are not strongly locking in general. Unfortunately, many known techniques to make them so are incompatible with **Bc**-convergence or violate other semantic restrictions (compare Fulk, 1990; Jain et al., 1999). We use an alternative approach to characterize a large class of learning criteria for which strongly locking learners suffice.

Let $\boldsymbol{R}$ denote the set of all unbounded non-decreasing functions $r \colon \mathbb{N} \to \mathbb{N}$. Note that such functions have an unbounded lower limit, that is, for every $m \in \mathbb{N}$, there are only finitely many n such that $r(n) < m$. A learning restriction δ allows for *simulation on equivalent text* if, for all $T, T' \in \mathbf{Txt}$ with $\mathrm{content}(T) = \mathrm{content}(T')$, partial functions $p \in \mathfrak{P}$ and $r \in \boldsymbol{R}$, the following holds: Whenever $\delta(p, T')$ and $\forall n \colon \mathrm{content}(T[n]) = \mathrm{content}(T'[r(n)])$, we have $\delta(p \circ r, T)$.

The intuition behind this definition lies in the name: a learner h' seeing one text T may simulate a learner h on a different text T', provided that T and T' are texts for the same language, and h' on T uses later and later conjectures given by h on T', but always only uses hypotheses of h that are based on all the data already available to h. Thus, the texts T and T' are not just equivalent in that they are for the same language, but also they are used in a data-synchronous way.

Note that this is a generalization of the *delayable* restrictions of Kötzing and Palenta (2016). Since all restrictions defined in Section 2 are delayable, they also allow for simulation on equivalent text. In fact, the strength of this definition lies in the fact that most of the studied learning restrictions in the literature allow for simulation on equivalent text.

Theorem 5 *Let restriction δ allow for simulation on equivalent text. Then, every class of languages that is* $\mathbf{TxtPsd}\delta\mathbf{Bc}$*-learnable is so learnable by a strongly* $\mathbf{Bc}$*-locking learner.*

Proof Let h be a total $\mathbf{Psd}$-learner (Theorem 2) such that $\mathcal{L} \subseteq \mathbf{TxtPsd}\delta\mathbf{Bc}(h)$ and consider the learner h' defined on finite sets D and numbers $t \geq |D|$ as

$$h'(D, n) = h(D, 2n).$$

Let us first check $\mathcal{L} \subseteq \mathbf{TxtPsd}\delta\mathbf{Bc}(h')$. To this end, let $L \in \mathcal{L}$ and $T \in \mathbf{Txt}(L)$. $T'(n) = T(\lfloor \frac{n}{2} \rfloor)$ denotes the text in which all data from T is repeated once, before the next data point occurs. Let $p = \mathbf{Psd}(h, T')$ denote the learning sequence in case T' is presented to h, namely $n \mapsto h(\mathrm{content}(T'[n]), n)$. Because of $T' \in \mathbf{Txt}(L)$ and $L \in \mathbf{TxtPsd}\delta\mathbf{Bc}(h)$, we obtain $\delta\mathbf{Bc}(p, T')$. Furthermore, the function $r(n) = 2n$ is non-decreasing and has unbounded lower limit, $r \in \boldsymbol{R}$ and, for all n, we have $\mathrm{content}(T'[r(n)]) = \mathrm{content}(T[n])$. Let $p' = \mathbf{Psd}(h', T)$ be the learning sequence resulting from T being presented to h', that is, $n \mapsto h(\mathrm{content}(T[n]), 2n)$. Observe that $p' = p \circ r$ is the composition of the learning sequence p with function r. From this we conclude $\delta\mathbf{Bc}(p', T)$ since the learning restriction $\delta\mathbf{Bc}$ allows for learning on equivalent texts. Hence, h' learns L from T.

It remains to verify that h' is strongly $\mathbf{Bc}$-locking. As h learns L there is a $\mathbf{Bc}$-locking information for h on L. In other words, there is some (D_0, n_0) such that $D_0 \subseteq L$ and $W_{h(D_0,n_0)} = L$, and for all possible extensions (D, n) compatible with L ($n \geq n_0$, $D_0 \subseteq D$, $D \subseteq L$, and $|D \setminus D_0| \leq n - n_0$), we have $W_{h(D,n)} = L$. Let again T be a text for L and n_1 large enough such that $D_0 \subseteq \mathrm{content}(T[n_1])$. Since every extension of a locking information for h on L is again a locking information, it suffices to find an n' such that $(\mathrm{content}(T[n']), 2n')$ is an extension of (D_0, n_0). By letting $n' = n_0 + n_1$ we immediately get $n_0 \leq 2n'$, $D_0 \subseteq \mathrm{content}(T[n'])$, $|\mathrm{content}(T[n']) \setminus D_0| \leq n' \leq 2n' - n_0$ and, of course, $\mathrm{content}(T[n']) \subseteq L$. Therefore, $(\mathrm{content}(T[n']), 2n')$ is a locking information for h' on L. Learner h' is strongly $\mathbf{Bc}$-locking by the arbitrary choice of T. ■

The proof of the last theorem gives a simple construction to make a learner strongly locking while *simultaneously* preserving the restriction δ. It does so for a rich class of learning restrictions. It remains an open question whether such a general statement also holds for the case of full-information learning, i.e., for the interaction operator $\mathbf{G}$. The next two theorems are partial results in this direction, already covering a large portion of the learning restrictions in question.

Theorem 6 *A class of languages is* $\mathbf{TxtGBc}$*-learnable by a strongly* $\mathbf{Bc}$*-locking learner if and only if it is* $\mathbf{TxtPsdBc}$*-learnable.*

Theorem 7 *Let β be an interaction operator and $\delta \subseteq \mathbf{NU}$ a restriction. Every class of languages that can be $\mathbf{Txt}\beta\delta\mathbf{Bc}$-learned is in fact so learned by a strongly $\mathbf{Bc}$-locking learner.*

5. Consistency

A natural requirement in language learning is to ask the learner to always include all its current knowledge in its hypothesis. That is, the learner's output has to be consistent with the content of the input sequence (Angluin, 1980). More formally, *consistency* is defined as the following predicate on infinite sequences $p \in \mathfrak{P}$ and texts $T \in \mathbf{Txt}$,

$$\mathbf{Cons}(p, T) \Leftrightarrow \forall i \colon \mathrm{content}(T[i]) \subseteq W_{p(i)}.$$

Consistency is a semantic restriction and obviously allows for simulation on equivalent text. It is known to severely limit the capabilities of $\mathbf{Ex}$-learners (see Jain et al., 1999). In the context of $\mathbf{Bc}$-convergence, however, the additional requirement of consistency does not reduce the learning power of many restrictions. One can even extend this requirement to texts for languages the learner cannot identify.

Theorem 8 *Let β be any interaction operator and $\delta \in \{\mathbf{T}, \mathbf{Caut_{Tar}}, \mathbf{Mon}, \mathbf{SMon}, \mathbf{WMon}, \mathbf{SemConv}, \mathbf{SemWb}\}$. Then, $[\tau(\mathbf{Cons})\mathbf{Txt}\beta\delta\mathbf{Bc}] = [\mathbf{Txt}\beta\delta\mathbf{Bc}]$.*

The inclusion $[\tau(\mathbf{Cons})\mathbf{Txt}\beta\delta\mathbf{Bc}] \subseteq [\mathbf{Txt}\beta\delta\mathbf{Bc}]$ is trivial, the other direction will follow from Lemma 9 through 12 below. We distinguish them by the technique used to ensure consistency. For their proofs we fix some notation. Let a text T and a finite initial sequence $\sigma \sqsubset T$ be given. We use $|\sigma|$ to denote the length of σ. For some learner h, we abbreviate the output $\beta(h, T)(|\sigma|)$ of h on σ to $h(\sigma)$. This does not mean that h has access to the order or even the number of data points presented thus far. However, by Theorem 3 we can assume that at least $\mathrm{content}(\sigma)$ is known to h. Using the *S-m-n Theorem* (compare Rogers, 1967) we construct mappings $\mathcal{P} \to \mathcal{P}$; $h \mapsto g$ such that the resulting learner g is consistent. We say such a transformation *preserves* some learning restriction δ if $(\beta(g, T), T) \in \delta$ whenever $(\beta(h, T), T) \in \delta$. In the proofs we repeatedly exploit the following simple observation: For any two indices $i \leq j$, we have $\mathrm{content}(T[i]) \subseteq \mathrm{content}(T[j]) \subseteq \mathrm{content}(T)$.

The first lemma follows just from patching the set of seen data points into the output.

Lemma 9 *The restrictions $\mathbf{T}$, $\mathbf{Mon}$, and $\mathbf{SMon}$ allow for consistent $\mathbf{Bc}$-learning.*

The W-hypothesis space prohibits an effective test for consistency. Given a universal enumeration procedure for all languages, one can at least determine consistency in the limit. We fix such a procedure in the following. For any program e, let $\Phi_e \colon \mathbb{N} \to \mathbb{N}$ be the *complexity measure* associated with the recursive function φ_e (Blum, 1967). We set $W_e^s = \{x \leq s \mid \Phi_e(x) \leq s\}$ as the collection of natural numbers at most s whose membership to W_e can be verified within s steps. A complete description of the finite set W_e^s can be computed from e and s, and $\bigcup_{s \in \mathbb{N}} W_e^s = W_e$. Furthermore, for any finite set D, we have $D \subseteq W_e$ if and only if there is an s such that $D \subseteq W_e^s$. This is the key technique to ensure consistency for the learning restrictions discussed in the following. Lemma 11 is a simple equality on the level of predicates.

Lemma 10 *The restrictions* **WMon** *and* $\mathbf{Caut_{Tar}}$ *allow for consistent* **Bc**-*learning.*

Lemma 11 $\mathbf{Cons} \cap \mathbf{SemWb} = \mathbf{Cons} \cap \mathbf{SemConv}$.

In the proof of the next lemma we employ *one-one texts.* Let $T \in \mathbf{Txt}$ be an arbitrary text. We construct the one-one sequence $T_{\text{1-1}}$ by deleting every occurrence of # as well as all duplicate data points from T. Note that for every i we find a unique $r(i) \leq i$ such that $\text{content}(T_{\text{1-1}}[r(i)]) = \text{content}(T[i])$. If content$(T)$ is finite, we amend $T_{\text{1-1}}$ with infinitely many pause symbols to make it a text.

Lemma 12 *The restrictions* **SemWb** *and* **SemConv** *allow for consistent* **Bc**-*learning.*

Proof By Lemma 11 it is sufficient to show this for **SemConv**. We give separate transformations for three relevant interaction operators β, namely, **G**, **Psd**, and **Sd**.

Case 1 $\beta = \mathbf{G}$.

For a finite sequence $\sigma \in \mathbb{S}\text{eq}$, we let $\sigma^- = \sigma[|\sigma|-1]$ denote the prefix of σ containing all but the last entry, or $\sigma^- = \varepsilon$ the empty sequence if $\sigma = \varepsilon$ is empty. Let h be a **G**-learner, there are recursive functions f and g such that

$$W_{f(\sigma)} = \text{content}(\sigma) \cup \bigcup_{s \in \mathbb{N}} \begin{cases} \emptyset, & \text{if content}(\sigma) \not\subseteq W^s_{h(\sigma)}; \\ W^s_{h(\sigma)}, & \text{otherwise;} \end{cases}$$

$$W_{g(\sigma)} = \begin{cases} W_{g(\sigma^-)}, & \text{if content}(\sigma) = \text{content}(\sigma^-); \\ W_{f(\sigma)}, & \text{otherwise.} \end{cases}$$

As long as the content of the input sequence does not change, the output of g also stays the same. If a new data point is shown, the respective hypothesis of the initial learner h is tested for consistency (in the limit), depending on the outcome either only the updated content is conjectured or the guess of h is adopted. The learner g is globally consistent.

Regarding **Bc**, let $L \in \mathbf{TxtGSemConvBc}(h)$ a language that h learns and $T \in \mathbf{Txt}(L)$ a text for L. If L is infinite, T shows new data points on infinitely many positions n, on all but finitely many of them, we have $W_{h(T[n])} = L$. The guesses are consistent and are hence adopted as the output of g. Also, g does not change its mind between these positions. If L is finite, there is a latest new point in T at position n_0. Since **SemConv** implies $\mathbf{Caut_{Tar}}$ (as a predicate), the conjecture $h(T[n_0])$ cannot describe a proper superset of L. Either $W_{h(T[n_0])}$ is inconsistent or equal to L, in both cases we have $W_{g(T[n])} = L$ for all $n \geq n_0$.

Suppose $i \leq j$ are indices with $\text{content}(T[j]) \subseteq W_{g(T[i])}$, we show that $W_{g(T[i])} = W_{g(T[j])}$. If the content does not change between the positions, this is obvious. We thus assume that i and j are such that $W_{g(T[i])} = W_{f(T[i])}$ and $W_{g(T[j])} = W_{f(T[j])}$. This means, $T(j-1)$ is a new data point contained in $W_{g(T[i])}$. The only way to achieve this is by hypothesis $h(T[i])$ being consistent. We get $\text{content}(T[j]) \subseteq W_{g(T[i])} = W_{h(T[i])}$ and arrive at $W_{h(T[i])} = W_{h(T[j])}$ via **SemConv**. Finally, $h(T[j])$ must also be consistent with the later observation $T[j]$,

$$W_{g(T[i])} = W_{h(T[i])} = W_{h(T[j])} = W_{g(T[j])}.$$

Case 2 $\beta = \mathbf{Psd}$.

Recall that partially set-driven learners have access to the length of the input sequence, they are, however, oblivious to the order in which the data is presented. We employ a construction that uses even less information, the original learner h is queried on inputs that only depend on the content seen so far. There is a **Psd**-learner g such that on finite sets $D \subset \mathbb{N}$ and numbers $t \geq |D|$,

$$W_{g(D,t)} = D \cup \bigcup_{s \in \mathbb{N}} \begin{cases} \emptyset, & \text{if } D \nsubseteq W^s_{h(D,|D|)}; \\ W^s_{h(D,|D|)}, & \text{otherwise.} \end{cases}$$

Let $L \in \mathbf{TxtPsdSemConvBc}(h)$ and $T \in \mathbf{Txt}(L)$. By transitioning to one-one texts we can interpret the output $h(\mathrm{content}(T[i]), |\mathrm{content}(T[i])|)$ of h on T as the hypothesis $h(\mathrm{content}(T_{\text{1-1}}[r(i)]), r(i))$ of h on $T_{\text{1-1}}$. Note that h learns L from $T_{\text{1-1}}$. Consequently, if L is infinite, almost all queried conjectures of h are correct. If L is finite, then hypothesis $h(L, |L|)$ is either correct or inconsistent. In all cases g converges to a semantically correct description of L on T, which yields the preservation of **Bc**. Regarding **SemConv**, let again $i \leq j$ be such that $\mathrm{content}(T[j]) \subseteq W_{g(\mathrm{content}(T[i]),i)}$. In case $h(\mathrm{content}(T_{\text{1-1}}[r(i)]), r(i))$ has been inconsistent, we have

$$W_{g(\mathrm{content}(T[i]),i)} = \mathrm{content}(T[i]) = \mathrm{content}(T[j]) = W_{g(\mathrm{content}(T[j]),j)}.$$

The last equality is due to the fact that in the computation of $g(\mathrm{content}(T[j]), j)$ learner h is *again* queried on input $(\mathrm{content}(T_{\text{1-1}}[r(i)]), r(i))$ since the content has not changed. If h initially has been consistent, we get

$$W_{g(\mathrm{content}(T[i]),i)} = W_{h(\mathrm{content}(T_{\text{1-1}}[r(i)]),r(i))} = W_{h(\mathrm{content}(T_{\text{1-1}}[r(j)]),r(j))} = W_{g(T[j])}$$

from the semantic conservatism of h on text $T_{\text{1-1}}$.

Case 3 $\beta = \mathbf{Sd}$.

This case follows from the fact that the construction of g in the **Psd**-case does not make use of the length t of the input sequence. Although the transformed learner is now based on a *set-driven* h, the preservation of **SemConv** can be seen exactly as above. ■

As seen in Lemma 11, the additional requirement of consistency can also reveal new connections among the different learning restrictions and, subsequently, the collections of identifiable classes. Another example is given in the next theorem.

Theorem 13 $\mathbf{Cons} \cap \mathbf{WMon} \subseteq \mathbf{Caut} \cap \mathbf{Dec}$.

Corollary 14 *The following relations hold for any interaction operator* β.

(i) $[\mathbf{Txt}\beta\mathbf{SemWbBc}] = [\mathbf{Txt}\beta\mathbf{SemConvBc}]$;

(ii) $[\mathbf{Txt}\beta\mathbf{WMonBc}] \subseteq [\mathbf{Txt}\beta\mathbf{CautDecBc}]$.

6. Set-Drivenness

We have seen in Section 3 that, in semantic learning, we can confine ourselves to total learners interacting with the text via the operators **G**, **Psd**, and **Sd**. For (partially) set-driven learning we even can assume the learner to be strongly locking. This raises the question whether we can further reduce the class of relevant learners. For example, it would simplify the learning process tremendously if the learner only needs to know the content of the input sequence to identify any learnable class; that is, if set-driven learners suffice. In this section we study the relation between partially set-driven learners and their set-driven counterparts in our semantic setting. As it turns out, in general they are unequal in learning power. There are, however, restrictions δ for which $[\mathbf{TxtPsd}\delta\mathbf{Bc}] = [\mathbf{TxtSd}\delta\mathbf{Bc}]$ holds.

We start with the separation result. The technique used in its proof is based on a result by Kötzing and Palenta (2016) which is, in turn, drawing on a proposition given by Schäfer-Richter (1984) and Fulk (1985) regarding set-driven **Ex**-learning. We extend the theorem to show that semantic learning is not covered completely by set-driven learners.

Theorem 15 $[\mathbf{TxtPsdConsMonSDecEx}] \setminus [\mathbf{TxtSdBc}] \neq \emptyset$.

Corollary 16 *If* $\delta \in \{\mathbf{T}, \mathbf{Cons}, \mathbf{Dec}, \mathbf{Mon}, \mathbf{NU}\}$*, we have* $[\mathbf{TxtSd}\delta\mathbf{Bc}] \subset [\mathbf{TxtPsd}\delta\mathbf{Bc}]$.

We now present learning restrictions for which **Psd**-learners and **Sd**-learners are equally powerful. In the upcoming proofs we employ several tools introduced in the earlier parts of this paper. Recall from Section 4 that a learning restriction allows for simulation on equivalent text if it permits the reasonable reuse of hypotheses gained from a different text showing the same content. The one-one texts defined in Section 5 are natural examples for such equivalent texts. The transition from an arbitrary text $T \in \mathbf{Txt}$ to the equivalent one-one text $T_{\text{1-1}}$ consists of deleting all pause symbols and repetitions. Regarding the associated non-decreasing functions $r \in \boldsymbol{R}$ with unbounded lower limit, the following observation is useful: For all n such that $\mathrm{content}(T[n]) \subset \mathrm{content}(T)$, there is exactly one assignment $n \mapsto r(n)$ satisfying $\mathrm{content}(T_{\text{1-1}}[(r(n))]) = \mathrm{content}(T[n])$, namely, $r(n) = |\mathrm{content}(T[n])|$. If $\mathrm{content}(T)$ is infinite, this uniquely defines a total function $r \in \boldsymbol{R}$. If the content is finite, we can choose $r(n) = n$ for the larger n.

Theorem 17 *Let* $\mathcal{L}$ *be a class of* infinite *languages and* δ *be a restriction that allows for simulation on equivalent text. Then,* $\mathcal{L} \in [\mathbf{TxtPsd}\delta]$ *if and only if* $\mathcal{L} \in [\mathbf{TxtSd}\delta]$.

Proof One direction is trivial. For the other one, let h be a **Psd**-learner that δ-learns $\mathcal{L}$. Suppose $L \in \mathcal{L}$ is a language and $T \in \mathbf{Txt}(L)$ a text for L. Construct an **Sd**-learner g derived from h by defining $g(D) = h(D, |D|)$ on all finite sets D. We claim that g also δ-learns L from T. To this end, let function $r \in \boldsymbol{R}$ be as above. The construction of g gives

$$\begin{aligned}\mathbf{Sd}(g,T)(n) &= g(\mathrm{content}(T[n])) = h(\mathrm{content}(T[n]), |\mathrm{content}(T[n])|)\\ &= h(\mathrm{content}(T_{\text{1-1}}[r(n)]), r(n)) = \mathbf{Psd}(h, T_{\text{1-1}})(r(n)).\end{aligned}$$

By our assumption h learns L from the text $T_{\text{1-1}}$, meaning $\delta(\mathbf{Psd}(h, T_{\text{1-1}}), T_{\text{1-1}})$. As δ allows for simulation on equivalent text, $\delta(\mathbf{Psd}(h, T_{\text{1-1}}) \circ r, T)$ follows. By the equality shown above, $\delta(\mathbf{Sd}(g,T), T)$ and g learns L from T. ■

It was observed by Case (1999) that, for infinite languages, set-driven learning is not a restriction in non-semantic settings without further learning restrictions.

Under some conditions we can apply the ideas from Theorem 17 also to classes of finite languages.

Theorem 18 *Let $\delta \subseteq \mathbf{Caut_{Tar}Cons}$ be a semantic restriction that allows for simulation on equivalent text. Then, $[\mathbf{TxtPsd}\delta\mathbf{Bc}] = [\mathbf{TxtSd}\delta\mathbf{Bc}]$.*

Proof We only need to prove $[\mathbf{TxtPsd}\delta\mathbf{Bc}] \subseteq [\mathbf{TxtSd}\delta\mathbf{Bc}]$. Let again h be a learner, $L \in \mathbf{TxtPsd}\delta\mathbf{Bc}(h)$ a language it identifies, and $g(D) = h(D, |D|)$ the derived learner. If L is infinite we get $L \in \mathbf{TxtSd}\delta\mathbf{Bc}(g)$ as in the proof of Theorem 17. Assume L is finite and let $T \in \mathbf{Txt}(L)$ be a text for L. In the first part of this proof we exploit the semanticity of δ to show that learner g identifies L from the one-one text $T_{\text{1-1}}$. In the second part we show that g learns L from text T using that δ allows for simulation on equivalent text.

We compare the learning sequences $\mathbf{Psd}(h, T_{\text{1-1}})$ and $\mathbf{Sd}(g, T_{\text{1-1}})$ of the two learners h and g working on the same text $T_{\text{1-1}}$. At positions $n < |L|$, we have

$$g(\text{content}(T_{\text{1-1}}[n])) = h(\text{content}(T_{\text{1-1}}[n]), n),$$

thus $\mathbf{Psd}(h, T_{\text{1-1}})(n) = \mathbf{Sd}(g, T_{\text{1-1}})(n)$. Since h learns L from $T_{\text{1-1}}$ by assumption, the learning sequence $(\mathbf{Psd}(h, T_{\text{1-1}}), T_{\text{1-1}})$ satisfies δ and hence also $\mathbf{Caut_{Tar}Cons}$. If $n \geq |L|$, consistency gives $L \subseteq W_{h(\text{content}(T_{\text{1-1}}[n]),n)}$ and the target cautiousness enforces equality. On the other hand

$$g(\text{content}(T_{\text{1-1}}[n])) = g(L) = h(L, |L|).$$

This hypothesis is in general syntactically different from $h(\text{content}(T_{\text{1-1}}[n]), n)$, but both are for the target language L. In total, $\mathbf{Psd}(h, T_{\text{1-1}})$ and $\mathbf{Sd}(g, T_{\text{1-1}})$ describe the same sequence of sets. Now $\delta(\mathbf{Sd}(g, T_{\text{1-1}}), T_{\text{1-1}})$ follows from learning restriction δ being semantic.

For the second part, we turn to the general text T for L. Let n_0 be the minimal n such that $\text{content}(T[n]) = L$. Recall that we have $r(n) = n$ for all $n \geq n_0$ to meet the requirement of an unbounded lower limit. The equality $\text{content}(T_{\text{1-1}}[r(n)]) = \text{content}(T[n])$ still holds for those n. It is now easy to see that the sequences $\mathbf{Sd}(g, T_{\text{1-1}})$ and $\mathbf{Sd}(g, T)$ of the learner g on the texts T and $T_{\text{1-1}}$, respectively, differ only by a composition with r,

$$\mathbf{Sd}(g, T)(n) = g(\text{content}(T[n])) = g(\text{content}(T_{\text{1-1}}[r(n)])) = \mathbf{Sd}(g, T_{\text{1-1}})(r(n)).$$

Since δ allows for simulation on equivalent text, $(\mathbf{Sd}(g, T), T) \in \delta$. By the discussion above the learning sequence also observes $\mathbf{Bc}$, learner g identifies L from T. ■

Together with Theorem 8 we get the following equalities on the level of learnable classes. In the case of **SMon**, the result was known before (Kötzing and Schirneck, 2016).

Corollary 19

(i) *If* $\delta \in \{\mathbf{Caut_{Tar}}, \mathbf{SemWb}, \mathbf{SemConv}, \mathbf{SMon}, \mathbf{WMon}\}$, $[\mathbf{TxtPsd}\delta\mathbf{Bc}] = [\mathbf{TxtSd}\delta\mathbf{Bc}]$.

(ii) $[\mathbf{TxtPsdConsCautBc}] = [\mathbf{TxtSdConsCautBc}]$.

It remains open whether **Caut** also allows for consistent **Bc**-learning. For the opposite statement it would suffice to show $[\mathbf{TxtSdCautBc}] \subset [\mathbf{TxtPsdCautBc}]$.

7. Conclusion and Future Work

This work lays the foundation for future analyses of semantic learning restrictions by providing versatile tools. We considered three different normal forms for learners and showed for many learning criteria that some or all of the normal forms can be assumed. These normal forms already supported proofs in this work and can be even more beneficial in future work when analyzing the relations *between* different semantic learning criteria.

We did not give any results for when a full-information learner can be assumed to be partially set-driven, which is a natural extension of our work. This would also complement our theorems for when partial set-driven learning can be assumed set-driven.

For consistent learning, natural future steps would be to (a) give constructions to establish this normal form for other learning restrictions, such as *cautious*, *decisive*, and *non-U-shaped* learning and (b) to give a more general, unifying proof, establishing the normal form for a range of learning restrictions simultaneously.

Acknowledgments

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

References

Dana Angluin. Inductive Inference of Formal Languages from Positive Data. *Information and Control*, 45:117–135, 1980.

Ganesh Baliga, John Case, Wolfgang Merkle, Frank Stephan, and Rolf Wiehagen. When Unlearning Helps. *Information and Computation*, 206:694 – 709, 2008.

Jānis Bārzdiņš. Inductive Inference of Automata, Functions and Programs. In *American Mathematical Society Translations*, pages 107–122, 1977.

Leonore Blum and Manuel Blum. Toward a Mathematical Theory of Inductive Inference. *Information and Control*, 28:125–155, 1975.

Manuel Blum. A Machine Independent Theory of the Complexity of Recursive Functions. *Journal of the ACM*, 14:322–336, 1967.

John Case. Periodicity in Generations of Automata. *Mathematical Systems Theory*, 8: 15–32, 1974.

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

John Case and Timo Kötzing. Strongly Non-U-Shaped Learning Results by General Techniques. *Information and Computation*, 251:1–15, 2016.

John Case and Christopher Lynes. Machine Inductive Inference and Language Identification. In *Proceedings of the 9th International Colloquium on Automata, Languages and Programming (ICALP)*, pages 107–115, 1982.

John Case and Samuel Moelius. Optimal Language Learning from Positive Data. *Information and Computation*, 209:1293–1311, 2011.

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

Mark Fulk. Prudence and Other Conditions on Formal Language Learning. *Information and Computation*, 85:1–11, 1990.

Mark Gold. Language Identification in the Limit. *Information and Control*, 10:447–474, 1967.

Sanjay Jain, Daniel Osherson, James Royer, and Arun Sharma. *Systems that Learn: An Introduction to Learning Theory.* MIT Press, Cambridge (MA), 2nd edition, 1999.

Klaus Jantke. Monotonic and Non-monotonic Inductive Inference. *New Generation Computing*, 8:349–360, 1991.

Timo Kötzing. *Abstraction and Complexity in Computational Learning in the Limit.* PhD thesis, University of Delaware, 2009.

Timo Kötzing. A Solution to Wiehagen's Thesis. In *Proceedings of the 31st Symposium on Theoretical Aspects of Computer Science (STACS)*, pages 494–505, 2014.

Timo Kötzing and Raphaela Palenta. A Map of Update Constraints in Inductive Inference. *Theoretical Computer Science*, 650:4–24, 2016.

Timo Kötzing and Martin Schirneck. Towards an Atlas of Computational Learning Theory. In *Proceedings of the 33rd Symposium on Theoretical Aspects of Computer Science (STACS)*, pages 47:1–47:13, 2016.

Steffen Lange and Thomas Zeugmann. Monotonic versus Non-monotonic Language Learning. In *Proceedings of the 2nd International Workshop on Nonmonotonic and Inductive Logic (NIL)*, pages 254–269, 1993.

Daniel Osherson and Scott Weinstein. Criteria of Language Learning. *Information and Control*, 52:123–138, 1982.

Daniel Osherson, Michael Stob, and Scott Weinstein. Learning Strategies. *Information and Control*, 53:32–51, 1982.

Daniel Osherson, Michael Stob, and Scott Weinstein. *Systems that Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists.* MIT Press, Cambridge (MA), 1986.

Hartley Rogers. *Theory of Recursive Functions and Effective Computability.* McGraw-Hill, New York, 1967. Reprinted by MIT Press, Cambridge (MA), 1987.

Gisela Schäfer-Richter. *Über Eingabeabhängigkeit und Komplexität von Inferenzstrategien.* PhD thesis, RWTH Aachen, 1984.

Kenneth Wexler and Peter Culicover. *Formal Principles of Language Acquisition.* MIT Press, Cambridge (MA), 1980.

Rolf Wiehagen. A Thesis in Inductive Inference. In *Proceedings of the 1st International Workshop on Nonmonotonic and Inductive Logic (NIL)*, pages 184–207, 1991.

Appendix A. Language Learning in the Limit (Full Length)

In this section we formally define our setting of learning in the limit and the associated learning criteria. We follow the system given by Kötzing (2009). For a background on computability see the textbook of Rogers (1967). We use symbols $\subset$ and $\subseteq$ to distinguish the proper subset and subset relation between sets. $\mathbb{N} = \{0, 1, 2, \dots\}$ denotes the set of natural numbers and $c, e, i, j, k, n, s, t \in \mathbb{N}$ elements thereof. The set of all partial and total functions $\mathbb{N} \to \mathbb{N}$ is denoted by $\mathfrak{P}$ and $\mathfrak{R}$, respectively. The subset of all partial (total) computable functions is denoted by $\mathcal{P}$ ($\mathcal{R}$). If a partial function $p \in \mathfrak{P}$ is defined at some position n, we mark this fact with $p(n)\downarrow$; otherwise, we write $p(n)\uparrow$. We fix an effective numbering $\{\varphi_e\}_{e\in\mathbb{N}}$ of $\mathcal{P}$ and let $W_e = \mathrm{dom}(\varphi_e)$ denote the e-th recursively enumerable set. This way, we interpret the natural number e as a *hypothesis* for the set W_e.

A *learner* is a partial computable function $h \in \mathcal{P}$. A *language* is a recursively enumerable set $L \subseteq \mathbb{N}$ of natural numbers. The symbol $\#$ shall be read as *pause*. Any total function $T \colon \mathbb{N} \to \mathbb{N} \cup \{\#\}$ is called a *text*, the collection of all texts is $\mathbf{Txt}$. For any text (or other sequence) T, we let the *content* of T be the set $\mathrm{content}(T) = \mathrm{range}(T)\setminus\{\#\}$. For any given language L, a *text for* L is a text T such that $\mathrm{content}(T) = L$, the collection of all texts for L is $\mathbf{Txt}(L)$. For any n, we use $T[n]$ to denote the sequence $(T(0), \dots, T(n-1))$ of length n (the empty sequence ε when $n = 0$). Such initial parts of texts is what learners usually get as information. For some language L, let $\mathbb{S}\mathrm{eq}(L)$ denote the set of finite sequences of elements of $L \cup \{\#\}$. We abbreviate $\mathbb{S}\mathrm{eq}(\mathbb{N})$ to $\mathbb{S}\mathrm{eq}$, variables σ, τ range over $\mathbb{S}\mathrm{eq}$. The relation $\tau \sqsubseteq \sigma$ means that τ is a prefix of σ. For some $x \in \mathbb{N} \cup \{\#\}$, $\sigma \diamond x$ denotes the result of appending x to σ.

An *interaction operator* is an operator β taking as arguments a function $h \in \mathcal{P}$ (the learner) and a text $T \in \mathbf{Txt}$, and outputs a (possibly partial) function $p \in \mathfrak{P}$. Intuitively, β defines how a learner can interact with a given text to produce a sequence of hypotheses. We define the interaction operators $\mathbf{G}$ (*Gold-style* or *full-information learning*, Gold, 1967), $\mathbf{Psd}$ (*partially set-driven learning*, Schäfer-Richter, 1984), $\mathbf{Sd}$ (*set-driven learning*, Wexler and Culicover, 1980), and $\mathbf{It}$ (*iterative learning*, Wexler and Culicover, 1980) as follows. For all learners h, texts T, and indices i,

$$\begin{aligned}
\mathbf{G}(h,T)(i) &= h(T[i]);\\
\mathbf{Psd}(h,T)(i) &= h(\mathrm{content}(T[i]), i);\\
\mathbf{Sd}(h,T)(i) &= h(\mathrm{content}(T[i]));\\
\mathbf{It}(h,T)(i) &= \begin{cases} h(\varepsilon), & \text{if } i = 0;\\ h(\mathbf{It}(h,T)(i-1), T(i-1)), & \text{otherwise.}\end{cases}
\end{aligned}$$

In set-driven learning, the learner has access to the set of all previous data, but not to the full sequence as in $\mathbf{G}$-learning. In partially set-driven learning, the learner has the set of data and the current iteration number. $\mathbf{Psd}$-learning is sometimes also called *rearrangement-independent learning* (Blum and Blum, 1975). In iterative learning, the learner can access its last hypothesis as well as the most recent data point. Hereby, $h(\varepsilon)$ denotes the initial hypothesis of learner h. A learner h is said to be *confluently iterative* just in case it is both set-driven and iterative. The associated interaction operator is denoted $\mathbf{CflIt}$. In this work, whenever we refer to *all interaction operators* we mean those we just defined.

One can establish a hierarchy among the interaction operators by noticing that some can be simulated by others (Case and Kötzing, 2016). For two interaction operators β, β', we say *β-learners can be translated into β'-learners*, written $\beta \preccurlyeq \beta'$, if, for every learner h, there is some learner h' such that on arbitrary texts T the resulting sequence of hypotheses of h working on T is the same as that of h'. That is, $\forall T \in \mathbf{Txt}\colon \beta(h,T) = \beta'(h',T)$. For example, an **Sd**-learner can be translated into an **Psd**-learner by simply ignoring the additional information of the number of the current iteration. Clearly, all learners investigated in this paper can be translated into **G**-learners.

Successful learning requires the learner to observe certain restrictions, for example convergence to a correct index. A *learning restriction* is a predicate δ on a learning sequence $p \in \mathfrak{P}$ and a text $T \in \mathbf{Txt}$. We give the important example of *explanatory learning* (**Ex**, Gold, 1967) defined such that, for all pairs (p,T)

$$\mathbf{Ex}(p,T) \Leftrightarrow p \in \mathfrak{R} \wedge (\exists n_0\colon \forall n \geq n_0 : p(n) = p(n_0) \wedge W_{p(n_0)} = \text{content}(T)).$$

There are several other learning restrictions under investigation in this work. We would like to point out that we tacitly assume successful learning to be total, as we did for **Ex**.

$$\begin{aligned}
\mathbf{Bc}(p,T) &\Leftrightarrow \exists n_0\ \forall n \geq n_0\colon W_{p(n)} = \text{content}(T);\\
\mathbf{Conv}(p,T) &\Leftrightarrow \forall i,j\colon (i \leq j \wedge \text{content}(T[j]) \subseteq W_{p(i)}) \Rightarrow p(i) = p(j);\\
\mathbf{Caut}(p,T) &\Leftrightarrow \forall i,j\colon W_{p(i)} \subset W_{p(j)} \Rightarrow i \leq j;\\
\mathbf{Dec}(p,T) &\Leftrightarrow \forall i,j,k\colon (i \leq j \leq k \wedge W_{p(i)} = W_{p(k)}) \Rightarrow W_{p(i)} = W_{p(j)};\\
\mathbf{SDec}(p,T) &\Leftrightarrow \forall i,j,k\colon (i \leq j \leq k \wedge W_{p(i)} = W_{p(k)}) \Rightarrow p(i) = p(j);\\
\mathbf{Mon}(p,T) &\Leftrightarrow \forall i,j\colon i \leq j \Rightarrow W_{p(i)} \cap \text{content}(T) \subseteq W_{p(j)} \cap \text{content}(T);\\
\mathbf{SMon}(p,T) &\Leftrightarrow \forall i,j\colon i \leq j \Rightarrow W_{p(i)} \subseteq W_{p(j)};\\
\mathbf{WMon}(p,T) &\Leftrightarrow \forall i,j\colon (i \leq j \wedge \text{content}(T[j]) \subseteq W_{p(i)}) \Rightarrow W_{p(i)} \subseteq W_{p(j)}.\\
\mathbf{NU}(p,T) &\Leftrightarrow \forall i,j,k\colon (i \leq j \leq k \wedge W_{p(i)} = W_{p(k)} = \text{content}(T)) \Rightarrow W_{p(i)} = W_{p(j)};\\
\mathbf{SNU}(p,T) &\Leftrightarrow \forall i,j,k\colon (i \leq j \leq k \wedge W_{p(i)} = W_{p(k)} = \text{content}(T)) \Rightarrow p(i) = p(j);
\end{aligned}$$

One could require that a conjecture which is consistent with the data must not be changed; this is known as *conservative* learning (**Conv**, Angluin, 1980). In *cautious* learning (**Caut**, Osherson et al., 1982) the learner is not allowed to ever give a conjecture for a strict subset of a previously conjectured set. In *non-U-shaped* learning (**NU**, Baliga et al., 2008) a learner may never *semantically* abandon a correct conjecture; in *strongly non-U-shaped* learning (**SNU**, Case and Moelius, 2011) not even syntactic changes are allowed after giving a correct conjecture. In *decisive* learning (**Dec**, Osherson et al., 1982), a learner may never return to a *semantically* abandoned conjecture; in *strongly decisive* learning (**SDec**, Kötzing, 2014) the learner may not even return to *syntactically* abandoned conjectures. Finally, a number of monotonicity requirements are studied (Jantke, 1991; Wiehagen, 1991; Lange and Zeugmann, 1993): in *strongly monotone* learning (**SMon**) the conjectured sets may only grow; in *monotone* learning (**Mon**) only incorrect data may be removed; and in *weakly monotone* learning (**WMon**) the conjectured set may only grow while it is consistent. Finally, we let $\mathbf{T}$ denote the absence of any restriction (even totality). We combine any two learning restrictions δ and δ' by intersecting them; we denote this by juxtaposition.

Now a *learning criterion* is a tuple $(\alpha, \mathcal{C}, \beta, \delta)$, where $\mathcal{C}$ is a set of admissible learners (typically $\mathcal{P}$ or $\mathcal{R}$), β is an interaction operator, and α, δ are learning restrictions. We write $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta$ to denote this learning criterion, omitting $\mathcal{C}$ in case of $\mathcal{C} = \mathcal{P}$ and the restrictions if they equal $\mathbf{T}$. Let $h \in \mathcal{C}$ be an admissible learner. We say that learner h $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta$*-learns* a language L iff on *arbitrary* texts $T \in \mathbf{Txt}$, $\alpha(\beta(h, T), T)$ holds, and for all texts $T \in \mathbf{Txt}(L)$, $\delta(\beta(h, T), T)$ holds. The class of languages $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta$-learned by h is denoted by $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta(h)$. Finally, we write $[\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta]$ to denote the set of all $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta$-learnable classes of languages.

Splitting a learning criteria into its particles enables one to study the influence of the different factors independently of the others. The relations between different learners, interaction operators, and learning restrictions then translate back to the collections of learnable classes. This is made formal in the next lemma by Case and Kötzing (2016).

Lemma 20 *Let $\alpha \subseteq \alpha', \delta \subseteq \delta'$ be learning restrictions, $\mathcal{C} \subseteq \mathcal{C}'$ classes of admissible learners and $\beta \preccurlyeq \beta'$ two interaction operators. Then, we have $[\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta] \subseteq [\tau(\alpha')\mathcal{C}'\mathbf{Txt}\beta'\delta']$.*

Appendix B. Omitted Proofs

In this appendix we present section-wise the proofs omitted in the main part.

B.1. Proofs of Section 3

Proposition 1 *If $\delta \in \{\mathbf{Conv}, \mathbf{SDec}, \mathbf{SNU}, \mathbf{Ex}\}$, then $\delta\mathbf{Bc} = \delta\mathbf{Ex}$.*

Proof The case $\delta = \mathbf{Ex}$ is trivial, of the remaining restrictions we only prove $\delta = \mathbf{Conv}$. The reasoning for **SDec** and **SNU** is very similar. The inclusion $\mathbf{ConvEx} \subseteq \mathbf{ConvBc}$ is obvious as **Ex** (as a predicate) implies **Bc**. Now suppose $p \in \mathfrak{P}$ is an infinite sequence and $T \in \mathbf{Txt}$ a text such that $\mathbf{ConvBc}(p, T)$ holds. There is an index n_0 such that all $n \geq n_0$ satisfy $W_{p(n)} = \mathrm{content}(T)$. In particular, $\mathrm{content}(T[n + 1]) \subseteq W_{p(n)}$. Conservatism now enforces that hypothesis $p(n) = p(n + 1)$ are equal, implying $\mathbf{ConvEx}(p, T)$. ■

Theorem 2 *For any interaction operator β and semantic restriction δ holds*

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

Proof We use symbol c to denote (the code number of) the input to some learner $h \in \mathcal{P}$. For example, if $\beta = \mathbf{G}$, c stands for the initial sequence $\sigma \sqsubseteq T$ of some text T; if $\beta = \mathbf{It}$, c encodes the ordered pair of the last hypothesis and the new data point.

The *S-m-n Theorem* (see Rogers, 1967) implies the existence of a total recursive function $h' \in \mathcal{R}$ such that on any c

$$W_{h'(c)} = \begin{cases} \emptyset, & \text{if } h(c)\uparrow; \\ W_{h(c)}, & \text{otherwise.} \end{cases}$$

Whenever h is defined, the output of h' describes the same set. Since the learning restriction δ is semantic, learner h' identifies all languages in $\mathbf{Txt}\beta\delta(h)$. The converse inclusion $[\mathcal{R}\mathbf{Txt}\beta\delta] \subseteq [\mathbf{Txt}\beta\delta]$ follows from Lemma 20. ■

Theorem 3 $\mathbf{G} \cong_{\text{sem}} \mathbf{It}$*,* $\mathbf{Sd} \cong_{\text{sem}} \mathbf{CflIt}$*.*

Proof Basic computability theory tells us that there is a total recursive padding-function $\text{pad} \in \mathcal{R}$ strongly monotonically increasing with $W_{\text{pad}(e,\sigma)} = W_e$ for any hypothesis e and finite sequence σ (compare Rogers, 1967). By the strong monotonicity of pad its inverse $\text{unpad}_2 \colon \text{pad}(e,\sigma) \mapsto \sigma$ is also computable. We abbreviate $\text{unpad}_2(c)$ as σ_c.

For the first part we show $\mathbf{It} \preccurlyeq_{\text{sem}} \mathbf{G}$ and $\mathbf{G} \preccurlyeq_{\text{sem}} \mathbf{It}$. The possibility to translate $\mathbf{It}$-learners into $\mathbf{G}$-learners *syntactically*, $\mathbf{It} \preccurlyeq \mathbf{G}$, implies $\mathbf{It} \preccurlyeq_{\text{sem}} \mathbf{G}$. To establish the converse, let a $\mathbf{G}$-learner h be given and consider the following $\mathbf{It}$-learner h',

$$\begin{aligned} h'(\varepsilon) &= \text{pad}(h(\varepsilon), \varepsilon); \\ h'(c, x) &= \text{pad}(h(\sigma_c \diamond x), \sigma_c \diamond x). \end{aligned}$$

Learner h' working on some text T imitates h in the following way. It pads the initial sequence $\sigma \sqsubseteq T$ seen so far into its hypothesis together with the original guess $h(\sigma)$. In the next step it regains this information by unpadding its own prior hypothesis c. It then appends the new data point x, simulates h on the new sequence $\sigma \diamond x$, and finally pads the result again for the next iteration. While this alters h's guesses syntactically, it preserves them semantically, we have $W_{\mathbf{It}(h',T)(n)} = W_{\mathbf{G}(h,T)(n)}$ for all n.

The argument showing the semantic equivalence of $\mathbf{Sd}$ and $\mathbf{CflIt}$ is very similar. By definition $\mathbf{CflIt} \preccurlyeq \mathbf{Sd}$, thus $\mathbf{CflIt} \preccurlyeq_{\text{sem}} \mathbf{Sd}$. Starting from an $\mathbf{Sd}$-learner h, it is straightforward to construct a semantically equivalent $\mathbf{It}$-learner,

$$\begin{aligned} h'(\varepsilon) &= \text{pad}(h(\emptyset), \varepsilon); \\ h'(c, x) &= \text{pad}(h(\text{content}(\sigma_c \diamond x)), \sigma_c \diamond x). \end{aligned}$$

Observe that h' satisfies the additional requirement that on any two sequences that have the same content the resulting hypothesis is the same. Hence, h' is in fact a $\mathbf{CflIt}$-learner. ■

B.2. Proofs of Section 4

Theorem 4 *Every* $\mathbf{Sd}$*-learner is strongly* $\mathbf{Bc}$*-locking.*

Proof Let h be an $\mathbf{Sd}$-learner and $L \in \mathbf{TxtSdBc}(h)$ a language it identifies. There is a $\mathbf{Bc}$-locking set for h on L, say D_0. Let $T \in \mathbf{Txt}(L)$ be a text for L and n_0 such that $\text{content}(T[n_0]) \supseteq D_0$. Then, $\text{content}(T[n_0])$ is a $\mathbf{Bc}$-locking set for h on L. ■

Theorem 6 *A class of languages is* $\mathbf{TxtGBc}$*-learnable by a strongly* $\mathbf{Bc}$*-locking learner if and only if it is* $\mathbf{TxtPsdBc}$*-learnable.*

Proof If a concept class is **TxtPsdBc**-learnable, it is so learnable by some strongly **Bc**-locking learner by Theorem 5. The learner can be translated into an equivalent **G**-learner (compare Section 2). This translation completely preserves the learning behavior, in particular the property of being strongly locking.

For the opposite implication, assume the **G**-learner h is strongly **Bc**-locking and identifies a class $\mathcal{L} \subseteq \mathbf{TxtGBc}(h)$. For any finite set D and number $t \geq |D|$, σ_D^t shall denote the *canonical sequence* which lists the members of D in ascending order and is then padded with pause symbols to length t (if needed). We define a **Psd**-learner g on such D and t by

$$g(D,t) = h(\sigma_D^t).$$

Let $L \in \mathcal{L}$ be a language. We denote by T_L the *canonical text* listing L in ascending order (possibly padded with # if L is finite). Since h is strongly locking, there is an index n_0 such that the initial part $T_L[n_0]$ is a **Bc**-locking sequence for h on L. Let $T \in \mathbf{Txt}(L)$ be any text for L and $n_1 \geq n_0$ large enough such that $\mathrm{content}(T_L[n_0]) \subseteq \mathrm{content}(T[n_1])$. Then, the canonical sequence $\sigma^{n_1}_{\mathrm{content}(T[n_1])} \sqsupseteq T_L[n_0]$ is an extension of the **Bc**-locking sequence. This shows $W_{g(\mathrm{content}(T[n]),n)} = L$ for $n \geq n_1$. ■

Theorem 7 *Let β be an interaction operator and $\delta \subseteq$ **NU** a restriction. Every class of languages that can be **Txt**$\beta\delta$**Bc**-learned is in fact so learned by a strongly **Bc**-locking learner.*

Proof Let h be a β-learner and $\mathcal{L} \subseteq \mathbf{Txt}\beta\delta\mathbf{Bc}(h)$ a class of languages it identifies. To ease notation, we use $h(\sigma)$ to denote the hypothesis the h outputs after seeing sequence σ. This shall not indicate that h, in its computation, can rely on the order or even the length of σ.

Let $L \in \mathcal{L}$ be a language and $T \in \mathbf{Txt}(L)$ a text for L. As h learns L from T, there is an n_0 such that $W_{h(T[n_0])} = L$. For any finite extension $T[n_0] \diamond \tau$ with $\tau \in \mathbb{S}\mathrm{eq}(L)$, there is a text $T' \in \mathbf{Txt}(L)$ such that the extension $T[n_0] \diamond \tau \sqsubset T'$ is an initial part. Learner h observes restriction δ on T' and thus also **NU**. As $h(T[n_0])$ is already a conjecture for the content of T', namely for L, h never abandons this guess (semantically) on T', implying $W_{h(T[n_0]\diamond\tau)} = L$. So, $T[n_0]$ is a **Bc**-locking sequence and h is strongly **Bc**-locking by the arbitrary choice of T.

■

B.3. Proofs of Section 5

Lemma 9 *The restrictions* **T**, **Mon**, *and* **SMon** *allow for consistent* **Bc**-*learning.*

Proof Let $h \in \mathcal{R}$ be some total learner (compare Theorem 2). By the S-m-n Theorem there is a total recursive function g such that

$$W_{g(\sigma)} = \mathrm{content}(\sigma) \cup W_{h(\sigma)}.$$

By construction learner g is consistent on arbitrary texts. It is also not hard to see that this transformation preserves **Bc**. Let T be a text and n_0 an index such that for all $n \geq n_0$, $W_{h(T[n])} = \mathrm{content}(T)$, then $W_{g(T[n])} = \mathrm{content}(T[n]) \cup W_{h(T[n])} = \mathrm{content}(T)$.

It is left to prove that the transformation preserves **T**, **SMon**, **Mon**, and $\mathbf{Caut}_{\mathbf{Tar}}$, respectively. The case $\delta = \mathbf{T}$ is trivial. For **SMon**, note that for all $i \leq j$

$$W_{h(T[i])} \subseteq W_{h(T[j])} \Rightarrow \text{content}(T[i]) \cup W_{h(T[i])} \subseteq \text{content}(T[j]) \cup W_{h(T[j])},$$

so g observes **SMon** whenever h does. In the same spirit, we have

$$\begin{aligned}&\text{content}(T) \cap W_{h(T[i])} \subseteq \text{content}(T) \cap W_{h(T[j])} \Rightarrow \\ &\qquad \text{content}(T) \cap (\text{content}(T[i]) \cup W_{h(T[i])}) \subseteq \text{content}(T) \cap (\text{content}(T[j]) \cup W_{h(T[j])}).\end{aligned}$$

The preservation of **Mon** follows from that. ■

Lemma 10 *The restrictions* **WMon** *and* $\mathbf{Caut}_{\mathbf{Tar}}$ *allow for consistent* **Bc**-*learning.*

Proof Given h there is a total recursive function g such that

$$W_{g(\sigma)} = \text{content}(\sigma) \cup \bigcup_{s \in \mathbb{N}} \begin{cases} \emptyset, & \text{if content}(\sigma) \nsubseteq W^s_{h(\sigma)}; \\ W^s_{h(\sigma)} & \text{otherwise.} \end{cases}$$

Learner g is consistent and if $W_{h(\sigma)} = \text{content}(T)$, then $W_{g(\sigma)} = W_{h(\sigma)}$, preserving **Bc**.

Let T be a text on which the original learner h observes **WMon** and indices $i \leq j$ such that $\text{content}(T[j]) \subseteq W_{g(T[i])}$. We claim that $W_{g(T[i])} \subseteq W_{g(T[j])}$. If the conjecture of h at position i has not been consistent, $\text{content}(T[i]) \nsubseteq W_{h(T[i])}$, we get $W_{g(T[i])} = \text{content}(T[i])$. By assumption $W_{g(T[i])}$ also contains the content of the longer sequence $T[j]$, hence, the extension has not shown new data. The consistency of g implies

$$W_{g(T[i])} = \text{content}(T[i]) = \text{content}(T[j]) \subseteq W_{g(T[j])}.$$

If the original guess $h(T[i])$ has been consistent, we have $\text{content}(T[j]) \subseteq W_{g(T[i])} = W_{h(T[i])}$. As h is weakly monotonic on T, it only grows its conjectured sets between i and j and the later guess $h(T[j])$ must also be consistent,

$$W_{g(T[i])} = W_{h(T[i])} \subseteq W_{h(T[j])} = W_{g(T[j])}.$$

It is easy to see that the transformation also preserves $\mathbf{Caut}_{\mathbf{Tar}}$. In neither of the cases, $W_{g(\sigma)} = \text{content}(\sigma)$ or $W_{g(\sigma)} = W_{h(\sigma)}$, the conjecture of g can be proper supersets of the text's content as long as the original learner h is target cautious. ■

Lemma 11 $\mathbf{Cons} \cap \mathbf{SemWb} = \mathbf{Cons} \cap \mathbf{SemConv}$.

Proof **SemWb** implies **SemConv** without any additional assumptions. Conversely, let a function $p \in \mathfrak{P}$ and text $T \in \mathbf{Txt}$ be such that $(p, T) \in \mathbf{Cons} \cap \mathbf{SemConv}$. If there are positions $i \leq j$ such that there is a semantic mind change $W_{p(i)} \neq W_{p(j)}$, **SemConv** implies that hypothesis $p(i)$ must have been inconsistent with the observation at position j,

$$\text{content}(T[j]) \nsubseteq W_{p(i)}.$$

Due to **Cons**, $\mathrm{content}(T[k]) = \mathrm{content}(T[k]) \cap W_{p(k)}$ for any position k. So, if $k \geq j$,

$$\emptyset \neq \mathrm{content}(T[j]) \backslash W_{p(i)} \subseteq (\mathrm{content}(T[k]) \cap W_{p(k)}) \backslash W_{p(i)}$$

and p observes **SemWb** with respect to text T. ■

Theorem 13 $\mathbf{Cons} \cap \mathbf{WMon} \subseteq \mathbf{Caut} \cap \mathbf{Dec}$.

Proof First, we prove $\mathbf{Cons} \cap \mathbf{WMon} \subseteq \mathbf{Caut}$. In order to reach a contradiction, assume there is an infinite sequence p and a text T such that $(p, T) \in \mathbf{Cons} \cap \mathbf{WMon}$, but $\mathbf{Caut}(p, T)$ does not hold. There must be indices $i \leq j$ such that

$$W_{p(i)} \supset W_{p(j)}.$$

Sequence p is consistent on T, thus, $\mathrm{content}(T[j]) \subseteq W_{p(j)} \subseteq W_{p(i)}$. The earlier conjecture $p(i)$ is consistent with the later observation $\mathrm{content}(T[j])$, **WMon** yields the contradiction

$$W_{p(i)} \subseteq W_{p(j)}.$$

For the other inclusion, assume $\mathbf{Dec}(p, T)$ does not hold. There are $i \leq j \leq k$ such that

$$W_{p(i)} = W_{p(k)} \wedge W_{p(i)} \neq W_{p(j)}.$$

Consistency gives $\mathrm{content}(T[j]) \subseteq W_{p(k)} = W_{p(i)}$, using the weak monotonicity of h (twice) we arrive at $W_{p(i)} \subset W_{p(j)} \subset W_{p(k)}$, a contradiction. ■

B.4. Proofs of Section 6

Theorem 15 $[\mathbf{TxtPsdConsMonSDecEx}] \backslash [\mathbf{TxtSdBc}] \neq \emptyset$.

Proof There is a total recursive function ind $\in \mathcal{R}$ such that for all finite sets $D \subset \mathbb{N}$, $W_{\mathrm{ind}(D)} = D$ holds (see Rogers, 1967). We define a **Psd**-learner h on any such D and numbers $t \geq |D|$ as

$$h(D, t) = \begin{cases} \mathrm{ind}(\emptyset), & \text{if } D = \emptyset; \\ \varphi_{\max(D)}(t), & \text{otherwise.} \end{cases}$$

Consider the class of languages $\mathcal{L} = \mathbf{TxtPsdConsMonSDecEx}(h)$ that h identifies. We claim that $\mathcal{L}$ cannot be learned set-drivenly. Assume by way of contradiction there is an **Sd**-learner h' for $\mathcal{L}$. Let the predicate Q, defined on pairs (D, t), be such that

$$Q(D, t) \Leftrightarrow D \subset W^t_{h'(D)}.$$

As h' is recursive, so is Q. The *Operator Recursion Theorem* (Case, 1974) yields a program e and a total recursive function $f \in \mathcal{R}$ strictly monotone increasing such that

$$W_e = \mathrm{range}(f)$$
$$\varphi_{f(n)}(t) = \begin{cases} \mathrm{ind}(\mathrm{content}(f[n])), & \text{if } Q(\mathrm{content}(f[n]), t); \\ e, & \text{otherwise.} \end{cases}$$

Case 1 $Q(\text{content}(f[n]), t)$ does not hold for any n or t.

Then, $W_e \in \mathcal{L}$ because function $\varphi_{f(n)}$ is constant and for any $n \in \mathbb{N}$, $T \in \mathbf{Txt}(W_e)$,

$$h(\text{content}(T[n]), n) = \varphi_{\max(\text{content}(T[n]))}(n) = e.$$

Learner h' must hence identify W_e from text f. There is a natural number n_0 such that $W_{h'(\text{content}(f[n]))} = W_e$ for all $n \geq n_0$. This gives

$$\text{content}(f[n]) \subseteq W^t_{h'(\text{content}(f[n]))}$$

for all t sufficiently large. As $Q(\text{content}(f[n]), t)$ does not hold, we have $\text{content}(f[n]) = W_e$, which contradicts that W_e is infinite.

Case 2 There are n and t such that $Q(\text{content}(f[n]), t)$ holds.

Let (n_0, t_0) be the lexicographically minimal pair with $Q(\text{content}(f[n_0]), t_0)$. In the following we abbreviate $\text{content}(f[n_0])$ as L_0. The predicate Q witnesses that

$$L_0 \neq W_{h'(L_0)}$$

So, learner h' cannot identify L_0. Contrariwise, on any text $T \in \mathbf{Txt}(L_0)$ the **Psd**-learner h changes its mind exactly twice. From $\text{ind}(\emptyset)$ to e once T shows its first non-pause symbol and subsequently to $\text{ind}(L_0)$ as soon as the pair $(\max(\text{content}(T[n]), n)$ is (lexicographically) larger than $(f(n_0{-}1), t_0)$. This learning sequence observes the restrictions **Cons**, **Mon**, **SDec**, and **Ex**. We conclude that L_0 is in $\mathcal{L}$, a contradiction. ■