

Generic Results for Concatenation Hierarchies

Thomas Place and Marc Zeitoun

the date of receipt and acceptance should be inserted later

Abstract In the theory of formal languages, the understanding of concatenation hierarchies of regular languages is one of the most fundamental and challenging topics. In this paper, we survey progress made on this problem since 1971. We also establish new generic statements regarding this problem.

1 Introduction

This paper has a dual vocation. The first is to outline progress seen during the last 50 years about *concatenation hierarchies* of regular languages. The second is to provide *generic statements* and *elementary proofs* of some of the core results on this topic, which were obtained previously in restricted cases. In this introduction, we present the historical background, first highlighting the motivations and the key ideas that emerged since the mid 60s. In a second part, we describe the contributions of the paper, which are either new proofs of existing results or generalizations thereof.

Historical background: a short survey of 50 years of research. Concatenation hierarchies were introduced in order to understand the interplay between two basic constructs used to build *regular languages*: Boolean operations and concatenation. The story started in 1956 with Kleene's theorem [18], one of the core results in automata theory. It states that languages of finite words recognized by finite automata are exactly the ones that can be described by regular expressions, *i.e.*, are built from the singleton languages and the empty set using a finite number of times the basic operations of union, concatenation, and iteration (a.k.a. Kleene star).

As Kleene's theorem provides another syntax for regular languages, it makes it possible to classify them according to the hardness of their description by such an expression. The notion of star-height was designed for this purpose. The *star-height* of a regular expression is its maximum number of nested Kleene stars. The *star-height* of a regular language is the minimum star-height of a regular expression defining the language. As there are languages of arbitrary star-height [12, 11], this notion is an

Funded by the DeLTA project (ANR-16-CE40-0007)

LaBRI, Bordeaux University, France

appropriate complexity measure. Eggan [12] raised in 1963 the natural question of computing the star-height of a regular language (see also Brzozowski [7]): “Given a regular language and a natural number n , is there an expression of star-height n defining the language?”.

This question, called the *star-height problem*, is an instance of the *membership problem*. Given a class \mathcal{C} of regular languages, the membership problem for \mathcal{C} asks whether \mathcal{C} is a decidable class, that is:

Input: A regular language L .
Output: Does L belong to \mathcal{C} ?

Thus, the star-height problem asks whether membership is decidable for the class \mathcal{H}_n of languages of star-height n , for each n . It was first solved in 1988 by Hashiguchi [16], but it took 17 more years to obtain simpler proofs, see [17, 42, 5].

Kleene’s theorem also implies that adding *complement* to our set of basic operations does not make it possible to define more languages. Thus, instead of just considering regular expressions, one may consider *generalized* regular expressions, where complement is allowed (in addition to union, concatenation and Kleene star). This yields the notion of *generalized star-height*, which is defined as the star-height, but replacing “regular expression” by “generalized regular expression”. One may then ask the same question: is there an algorithm to compute the *generalized star-height* of a regular language? Despite its simple statement, this question, raised in 1980 by Brzozowski [7, 6], is still open. Even more, one does not know whether there exists a regular language of generalized star-height greater than 1. In other terms, membership is open for the class of languages of generalized star-height 1 (see [27] for a historical presentation).

This makes it relevant to already focus on languages of generalized star height 0, *i.e.*, that can be described using only union, concatenation and Boolean operations (including complement), *without* the Kleene star. Such languages are called *star-free*. It turns out that even for this restricted class, the membership problem is difficult. It was solved in 1965 by Schützenberger in a seminal paper.

Theorem 1 (Schützenberger [43]) *Membership is decidable for the class of star-free languages.*

Star-free languages rose to prominence due to their robustness: they enjoy several equivalent characterizations. The most important, of logical flavor, was discovered in 1971 by McNaughton and Papert. The key idea is that one may describe languages with logical sentences: any word may be viewed as a logical structure made of a linearly ordered sequence of positions, each carrying a label. In first-order logic over words (denoted by $\text{FO}(<)$), one may quantify these positions, compare them with a predicate “ $<$ ” interpreted as the (strict) linear order, and check their labels (for any letter a , a unary “label” predicate selecting positions with label a is available). Therefore, each $\text{FO}(<)$ sentence states a property over words and defines the language of all words that satisfy it.

Theorem 2 (McNaughton & Papert [21]) *For any regular language L , the following properties are equivalent:*

- L is star-free.
- L can be defined by an $\text{FO}(<)$ sentence.

Let us point out that this connection between star-free and first-order definable languages is rather intuitive. Indeed, there is a clear correspondence between union, intersection and complement for star-free languages with the Boolean connectives in FO($<$) sentences. Moreover, concatenation corresponds to existential quantification.

Just as the star-height measures how complex a regular language is, a natural complexity for star-free languages is the required number of *alternations* between concatenation and complement operations for building a given star-free language from basic ones. This led Brzozowski and Cohen [8] to introduce in 1971 a hierarchy of classes of regular languages, called the *dot-depth hierarchy*. It classifies all star-free languages into full levels, indexed by natural numbers: 0, 1, 2, . . . , and half levels, indexed by half natural numbers: $\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$, etc. Roughly speaking, levels count the number of alternations between concatenation and Boolean operations (including complement) that are necessary to express a given star-free language.

More formally, the hierarchy is built by using, alternately, two closure operations starting from level 0: *Boolean* and *polynomial* closures. Given a class of languages \mathcal{C} , its *Boolean closure*, denoted $Bool(\mathcal{C})$, is the least Boolean algebra containing \mathcal{C} . Polynomial closure is slightly more complicated as it involves *marked concatenation*. Given two languages K and L , a marked concatenation of K with L is a language of the form KaL for some $a \in A$. The *polynomial closure* of \mathcal{C} , denoted $Pol(\mathcal{C})$, is the least class of languages containing \mathcal{C} and closed under union, intersection and marked concatenation (*i.e.*, $K \cup L$, $K \cap L$ and KaL belong to \mathcal{C} for any $K, L \in \mathcal{C}$ and $a \in A$).

The dot-depth hierarchy is now defined as follows (n denotes a positive integer):

- Level 0 is the class $\{\emptyset, \{\varepsilon\}, A^+, A^*\}$ (where A is the working alphabet).
- The *half level* $n + \frac{1}{2}$ is the *polynomial closure* of the full level n .
- The *full level* $n + 1$ is the *Boolean closure* of the half level $n + \frac{1}{2}$.

A side remark is that this is not the original definition. First, the historical definition of the dot-depth started from another class at level 0. However, both definitions coincide at level 1 and above. Second, the polynomial closure of a class \mathcal{C} was defined as the least class containing \mathcal{C} and closed under *union* and concatenation. This definition is seemingly weaker, as it does not explicitly insist for $Pol(\mathcal{C})$ to be closed under intersection. However, Arfi [2, 3] (assuming that \mathcal{C} satisfies some mild closure properties) and Pin [25] (assuming slightly more general properties on \mathcal{C}) showed that the two definitions are equivalent.

The union of all levels in the dot-depth hierarchy is the class of all star-free languages. Moreover, Brzozowski and Knast proved in 1978 that the dot-depth hierarchy is strict: every level contains strictly more languages than the previous one.

Theorem 3 (Brzozowski & Knast [9]) *The dot-depth hierarchy is strict when the alphabet contains at least two letters.*

This shows in particular that classes built using Boolean and polynomial closure do not satisfy the same closure properties, in general. Typically, when \mathcal{C} is a class of languages, $Pol(\mathcal{C})$ is closed under marked concatenation but *not* under complement, while $Bool(\mathcal{C})$ is closed under complement but *not* under marked concatenation. The fact that the hierarchy is strict motivates the investigation of the membership problem for each level.

Problem 4 (Membership for the dot-depth hierarchy) Given a level in the dot-depth hierarchy, is membership decidable for this level?

Using the framework developed by Schützenberger in his proof for deciding whether a language is star-free, Knast proved in 1983 that level 1 enjoys decidable membership, via a combinatorially intricate proof.

Theorem 5 (Knast [19]) *Level 1 in the dot-depth has decidable membership.*

The case of half levels required to adapt Schützenberger’s approach, which was designed to deal with Boolean algebras only (recall that half levels are *not* Boolean algebras since the hierarchy is infinite). In 1995, Pin [24] modified the framework to handle half levels. Membership was then solved for level $\frac{1}{2}$ by Pin and Weil in 2002, as well as for level $\frac{3}{2}$ by Glaßer and Schmitz in 2007.

Theorem 6 (Pin & Weil [29, 30, 31], Glaßer & Schmitz [14]) *Levels $\frac{1}{2}$ and $\frac{3}{2}$ in the dot-depth hierarchy have decidable membership.*

One may now wonder why level 0 in the dot-depth hierarchy is $\{\emptyset, \{\varepsilon\}, A^+, A^*\}$. It would be natural to start from $\{\emptyset, A^*\}$, and to apply the very same construction scheme for higher levels. This is exactly the definition of the Straubing-Thérien hierarchy, introduced independently in 1981 by Straubing [48] and Thérien [50]. Its definition follows the same pattern as that of the dot-depth, except that level 0 is $\{\emptyset, A^*\}$.

Like the dot-depth, the Straubing-Thérien hierarchy is strict and spans the whole class of star-free languages. One can show this by proving that level n in the dot-depth hierarchy sits between levels n and $n + 1$ in the Straubing-Thérien hierarchy. This makes membership a relevant problem for each level in this hierarchy as well.

Problem 7 (Membership for the Straubing-Thérien hierarchy) Given a level in the Straubing-Thérien hierarchy, is membership decidable for this level?

Just as for the dot-depth hierarchy, level 1 in the Straubing-Thérien hierarchy was shown to be decidable by Simon in 1972 (actually before the formal definition of the hierarchy itself). The first half levels were solved in 1987 by Arfi who relied, for level $\frac{3}{2}$, on a difficult result of Hashiguchi [15]. In 1995, Pin and Weil presented a self-contained proof using the adaptation [24] of the framework of Schützenberger to classes that are not closed under complement.

Theorem 8 (Simon [44, 45]) *Level 1 in the Straubing-Thérien hierarchy has decidable membership.*

Theorem 9 (Arfi [2, 3], Pin & Weil [29, 30]) *Levels $\frac{1}{2}$ and $\frac{3}{2}$ in the Straubing-Thérien hierarchy have decidable membership.*

In fact, the dot-depth and the Straubing-Thérien hierarchies are closely related. First, as already stated, they are interleaved. More importantly, Straubing proved in 1985 an effective reduction between the membership problems associated to their full levels, which Pin and Weil adapted to half levels in 2002.

Theorem 10 (Straubing [49], Pin & Weil [31]) *Membership for a level in the dot-depth reduces to membership for the same level in the Straubing-Thérien hierarchy.*

This theorem is crucial. Indeed, from a combinatorial point of view, the Straubing-Thérien hierarchy is much simpler to deal with than the dot-depth. This is evidenced by all recent publications on the topic: most results for the dot-depth are obtained indirectly as corollaries of results for the Straubing-Thérien hierarchy via Theorem 10. This is the case for the last result about membership that we state, which dates back to 2014, and concludes the state of the art about membership for both hierarchies.

Theorem 11 (Place & Zeitoun [34, 40], Place [32, 33]) *Membership is decidable for levels 2 , $\frac{5}{2}$ and $\frac{7}{2}$ in both the dot-depth and the Straubing-Thérien hierarchies.*

Note that there is a gap between levels $\frac{5}{2}$ and $\frac{7}{2}$: it is unknown whether level 3 has decidable membership. This is because full levels are actually harder to cope with than half levels. Indeed, the framework that was developed recently to solve instances of membership relies on more general problems, called separation and covering [36], which we shall discuss in Section 5. The crucial point is that closure under concatenation product—which holds for half but not for full levels—is an essential ingredient in the methodology elaborated for solving separation and covering.

Now that we have surveyed the most prominent results regarding membership for two concatenation hierarchies, let us explain an extra but important motivation for investigating this problem. Recall that the initial incentive was to understand the interplay between Boolean operations and concatenation, two operations at the heart of language theory. Additionally, Thomas discovered in 1982 a tight connection between the dot-depth hierarchy and first-order logic, which suffices by itself to motivate an in-depth investigation of this hierarchy. The core idea is the following: since star-free languages are exactly those that one can define in first-order logic, it is desirable to refine this correspondence level by level, in each of the hierarchies considered so far. The beautiful result of Thomas establishes such a correspondence.

To present it, we first slightly extend the standard signature used in first-order logic over words. In addition to the linear order and the label predicates, we add:

- The binary *successor* “+1”, which holds for pairs of the form $(i, i + 1)$.
- The unary *minimum* “min”, that selects the leftmost position of the word.
- The unary *maximum* “max”, that selects the rightmost position of the word.
- The nullary *empty* “ ε ” predicate, which holds for the empty word only.

We denote by $\text{FO}(<, +1, \min, \max, \varepsilon)$ the resulting logic. Since these predicates are all definable in $\text{FO}(<)$, adding them in the signature does not increase the overall expressive power of first-order logic. In other words, $\text{FO}(<)$ and $\text{FO}(<, +1, \min, \max, \varepsilon)$ are equally expressive. However, this enriched signature makes it possible to define fragments of first-order logic that correspond to levels of the dot-depth hierarchy.

To this end, we classify $\text{FO}(<, +1, \min, \max, \varepsilon)$ sentences by counting their number of quantifier *alternations*. Given an integer $n \geq 0$, a sentence is “ $\Sigma_n(<, +1, \min, \max, \varepsilon)$ ” (resp. “ $\Pi_n(<, +1, \min, \max, \varepsilon)$ ”) when it is a formula from $\text{FO}(<, +1, \min, \max, \varepsilon)$ whose prenex normal form has either:

- *exactly* n blocks of quantifiers, the leftmost being an “ \exists ” (resp. a “ \forall ”) block,
- or *strictly less* than n blocks of quantifiers.

For example, a formula of $\text{FO}(<, +1, \min, \max, \varepsilon)$ whose prenex normal form is:

$$\exists x_1 \exists x_2 \forall x_3 \exists x_4 \varphi(x_1, x_2, x_3, x_4) \quad (\text{with } \varphi \text{ quantifier-free}),$$

is Σ_3 . Observe that while $\text{FO}(<)$ and $\text{FO}(<, +1, \min, \max, \varepsilon)$ have the same expressiveness, the enriched signature increases the expressive power of individual levels.

The negation of a $\Sigma_n(<, +1, \min, \max, \varepsilon)$ sentence is not $\Sigma_n(<, +1, \min, \max, \varepsilon)$ in general (it is $\Pi_n(<, +1, \min, \max, \varepsilon)$). Thus, the corresponding classes of languages are not closed under complement and it makes sense to define $\mathcal{B}\Sigma_n(<, +1, \min, \max, \varepsilon)$ sentences as finite Boolean combinations of $\Sigma_n(<, +1, \min, \max, \varepsilon)$ sentences. This yields a

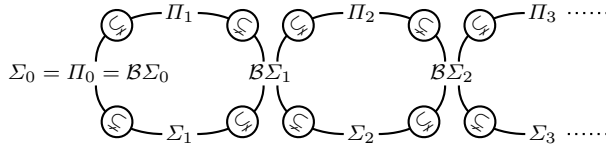


Fig. 1 The quantifier alternation hierarchy of first-order logic

strict hierarchy of classes of languages depicted in Fig. 1 (abusing notation, we denote a class of languages by the set of sentences defining it).

The correspondence discovered by Thomas relates levels of the dot-depth hierarchy with levels in the quantifier alternation hierarchy of enriched first-order logic.

Theorem 12 (Thomas [52]) *For any alphabet A , any integer $n \in \mathbb{N}$ and any language $L \subseteq A^*$, the two following properties hold:*

1. L has dot-depth n if and only if L belongs to $\mathcal{B}\Sigma_n(<, +1, \min, \max, \varepsilon)$.
2. L has dot-depth $n + \frac{1}{2}$ if and only if L belongs to $\Sigma_{n+1}(<, +1, \min, \max, \varepsilon)$.

Some years later in 1986, a similar correspondence was established between levels in the Straubing-Thérien hierarchy and in the quantifier alternation hierarchy over the signature consisting of the linear order and the label predicates. Such levels, denoted by $\mathcal{B}\Sigma_n(<)$ and $\Sigma_n(<)$, are defined analogously as for the enriched signature.

Theorem 13 (Perrin & Pin [23]) *For any alphabet A , any integer $n \in \mathbb{N}$ and any language $L \subseteq A^*$, the two following properties hold:*

1. L has level n in the Straubing-Thérien hierarchy if and only if L belongs to $\mathcal{B}\Sigma_n(<)$.
2. L has level $n + \frac{1}{2}$ in the Straubing-Thérien hierarchy if and only if L belongs to $\Sigma_{n+1}(<)$.

Content of the paper. The line of research that we surveyed spans over 45 years. This is why results are disseminated in the literature, and often tailored to one or the other of the two hierarchies. Moreover, their proofs often rely on involved algebraic or topological tools, and sometimes use other hard results as black boxes. In this paper, we present a unified framework that captures all of them. In the following, we call *finitely based* a concatenation hierarchy whose level 0 is a *finite* Boolean algebra closed under quotient (see Section 2). We present six theorems that suffice to recover **all** known results that we presented so far, providing new proofs for four of them:

1. We give a new proof that the polynomial closure of a lattice of regular languages closed under quotient is also closed under intersection.
2. We prove a generalization of Theorem 3: *any* finitely based hierarchy is *strict*.
3. We state that levels $\frac{1}{2}$, 1 and $\frac{3}{2}$ of *any* finitely based hierarchy have decidable separation, hence also decidable membership.
4. We prove the following transfer result (even for non-finitely based hierarchies): if level $n - \frac{1}{2}$ has decidable separation, then level $n + \frac{1}{2}$ has decidable membership.
5. We generalize Theorem 10 to separation, with a language-theoretic formulation.
6. We generalize Theorems 12 and 13 to *any* hierarchy, by showing that one can describe any concatenation hierarchy by an associated logical fragment.

We provide new proofs for Items 1, 2, 4 and 6 (Theorems 29, 42, 50 and 73). For Items 3 and 5 (Theorems 52 and 65), see [36, 35, 40] or the full papers [41, 37, 39].

Related work. This paper is a *survey* of results that are scattered in the literature and that mostly rely on different techniques as ours. Actually, several results in the theory of regular languages have been obtained indirectly, through the lens of finite semigroup theory, which was developed around algebraic and topological machinery. Our objective is to obtain direct, accessible, self-contained proofs of more generic statements. On the way, we drop unnecessary hypotheses and we underline the key ones. As an example, consider Theorem 65, which states a reduction from separation for the dot-depth hierarchy to separation for the Straubing-Thérien hierarchy. A full and self-contained proof requires only a few pages [37, 35]. Part of this statement can be recovered by combining already known results, but this is far from being direct as this relies on several papers: Almeida [1] related separation with an algebraic problem. Steinberg [47] then obtained a “delay theorem” for this problem, which, combined with results of Straubing [49] and with the correspondences between the hierarchies of languages and their algebraic counterparts, finally gives the result (for integer levels only). Along the years, results were obtained in the algebraic setup. However, relying on it requires much expertise in order to get an overall intuition. Finally, it often misses constructiveness of a separator in separation algorithms, contrary to our framework.

Organization. In Section 2, we set up the notation and present basic notions. We define Boolean and polynomial closures in Section 3, where we also prove that closure under intersection for polynomial closure is implied by closure under union and marked concatenation, if the class we start from satisfies mild closure properties. In Section 4, we define generic concatenation hierarchies and we state their basic properties. We also prove that concatenation hierarchies with a finite basis are strict. In Section 5, we state the main results about separation and membership for concatenation hierarchies, and we prove a characteristic property of $Pol(\mathcal{C})$, which yields a reduction from \mathcal{C} -separation to $Pol(\mathcal{C})$ -membership. Section 6 investigates the two historical hierarchies: the dot-depth and the Straubing-Thérien hierarchies. Finally, Section 7 presents the generic logical definition of concatenation hierarchies, thus generalizing the result of Thomas [52]. This paper is the full version of [38].

2 Preliminary definitions and tools

In this section, we set up the main definitions about classes of languages. We then present the problems we are interested in: “membership” and “separation”.

2.1 Finite words and classes of languages

Throughout the paper, we fix a finite alphabet A . We let ε be the empty word. The set of all finite words over A is denoted by A^* , and the set $A^* \setminus \{\varepsilon\}$ of all nonempty words over A is denoted by A^+ . If w is a word and $w = a_1 \cdots a_n$ with $a_i \in A$, the set of *positions* of w is $\{1, \dots, n\}$. Moreover, the length of w , denoted by $|w|$ is its number of positions. For two positions i, j we define $w]i, j[$ as the word $a_{i+1} \cdots a_{j-1}$ if $i + 1 \leq j - 1$, and as ε if $i + 1 > j - 1$. We define similarly $w[i, j[$ as $a_i \cdots a_{j-1}$ if $i \leq j - 1$ and as ε otherwise. We define $w]i, j[$ symmetrically. Given $w \in A^*$, we let $\text{alph}(w)$ be the set of letters appearing in w , that is, the smallest set $B \subseteq A$ such that $w \in B^*$. We say that $\text{alph}(w)$ is the *alphabet* of w . A *language (over A)* is a subset of A^* . Finally, a *class of languages* is a set of languages over A .

Remark 14 Our definition of a class of languages is simpler than the usual one. When dealing with several alphabets, a class of languages is often defined as a *correspondence* mapping any finite alphabet A to a set of languages $\mathcal{C}(A)$ over A . Our simpler definition is justified by the fact that we mainly use one fixed alphabet in the paper.

We shall consider several fundamental operations on languages:

- Boolean operations (union, intersection and complement),
- Left and right quotients. If w is a word and L is a language, then the left (resp. right) quotient $w^{-1}L$ (resp. Lw^{-1}) of L by w is the following language:

$$w^{-1}L \stackrel{\text{def}}{=} \{v \in A^* \mid wv \in L\}, \quad Lw^{-1} \stackrel{\text{def}}{=} \{v \in A^* \mid vw \in L\}.$$

Note that for any language L , any $a \in A$ and any $w \in A^*$, we have $(wa)^{-1}L = a^{-1}(w^{-1}L)$ and $L(aw)^{-1} = (Lw^{-1})a^{-1}$. Therefore, a class is closed under taking quotients if and only if it is closed under taking quotients by any *letter*. We shall freely use this fact throughout the paper. Another basic fact that we shall use without further reference is that quotients commute with Boolean operations. For instance, $w^{-1}(K \cup L) = w^{-1}K \cup w^{-1}L$ and $w^{-1}(A^* \setminus L) = A^* \setminus (w^{-1}L)$.

- The concatenation of two languages $K, L \subseteq A^*$ is defined as:

$$KL \stackrel{\text{def}}{=} \{uv \mid u \in K \text{ and } v \in L\}.$$

- Finally, given a letter $a \in A$, the *marked concatenation* of K and L by a is the language $K\{a\}L$, also written KaL .

All classes considered in this paper satisfy robust properties, which we present now.

- A class of languages is a *lattice* if it contains \emptyset and A^* and it is closed under union and intersection.
- A *Boolean algebra* is a lattice closed under complement.
- A class of languages is *quotienting* when it is closed under taking (left and right) quotients by words of A^* .

Example 15 Let AT be the class of languages over A consisting of all finite Boolean combinations of languages A^*aA^* , for $a \in A$. The name “AT” stands for “alphabet testable”: a language belongs to AT when membership of a word in this language depends only on the set of letters occurring in the word. It is straightforward to verify that AT is a *finite* quotienting Boolean algebra.

We denote by REG the class of all regular languages over A . All classes that we consider consist of regular languages only, *i.e.*, are sub-classes of REG . Recall that regular languages can be equivalently defined by nondeterministic finite automata (NFA), regular expressions, finite monoids or monadic second-order logic. In this paper, we assume a basic knowledge in automata theory, but we shall rather work with the definition based on monoids, which we recall in Section 5. Moreover, we rely on the following characterization of regular languages, due to Myhill and Nerode.

Theorem 16 (Myhill and Nerode [22]) *Let $L \subseteq A^*$ be a language. The following properties are equivalent:*

1. L is regular,
2. L has finitely many left quotients,

3. L has finitely many right quotients.

Membership and separation. For a class \mathcal{C} of languages, the most basic question is whether one can test membership of an input regular language in the class \mathcal{C} . In other words, we want to design an algorithm deciding whether an input language belongs to \mathcal{C} , or to prove that no such algorithm exists. The corresponding decision problem is called \mathcal{C} -membership (or membership for \mathcal{C}).

Membership problem for \mathcal{C} :

Input: A regular language L .
Question: Does L belong to \mathcal{C} ?

Recent solutions to the membership problem for specific classes actually consider a more general problem, the \mathcal{C} -separation problem (or separation problem for \mathcal{C}), which is stated as follows:

Separation problem for \mathcal{C} :

Input: Two regular languages L_1, L_2 .
Question: Is there a language K from \mathcal{C} such that $L_1 \subseteq K$ and $K \cap L_2 = \emptyset$?

We say that a language K such that $L_1 \subseteq K$ and $K \cap L_2 = \emptyset$ is a *separator* of (L_1, L_2) . Observe that since regular languages are closed under complement, there is a straightforward reduction from membership to separation. Indeed, an input language L belongs to \mathcal{C} exactly when it can be \mathcal{C} -separated from its complement.

2.2 Finite quotienting lattices and canonical preorders

Here, we present simple mathematical tools associated to *finite* quotienting lattices. Of course, most finite quotienting lattices are not very interesting. The only example which is featured prominently in this paper is the class AT of alphabet testable languages (see Example 15). However, when manipulating arbitrary quotienting lattices in proofs, we shall often reduce ourselves to finite ones using the following lemma.

Lemma 17 *Let \mathcal{C} be a quotienting lattice of regular languages and let \mathcal{C}' be a finite subclass of \mathcal{C} . Then, there exists a finite quotienting lattice \mathcal{D} such that $\mathcal{C}' \subseteq \mathcal{D} \subseteq \mathcal{C}$.*

Proof We define \mathcal{D} as the least quotienting lattice containing \mathcal{C}' . Since \mathcal{C} is a quotienting lattice itself, it is immediate \mathcal{D} is a quotienting lattice such that $\mathcal{C}' \subseteq \mathcal{D} \subseteq \mathcal{C}$. We prove that \mathcal{D} is finite. By definition, any language in \mathcal{D} is built from languages in \mathcal{C}' using unions, intersections and quotients. Hence, since quotients commute with Boolean operations, it follows that any language in \mathcal{D} is built from quotients of languages in \mathcal{C}' using unions and intersections. By hypothesis \mathcal{C}' is a finite set of regular languages. Hence, by Theorem 16, there are finitely many quotients of languages in \mathcal{C}' and we obtain that \mathcal{D} is finite. \square

Canonical preorders. Given an arbitrary class \mathcal{C} , we associate a *canonical preorder relation over A^** as follows. Given $w, w' \in A^*$, we write $w \leq_{\mathcal{C}} w'$ if and only if:

$$\text{For all } L \in \mathcal{C}, \quad w \in L \Rightarrow w' \in L.$$

It is immediate from the definition that $\leq_{\mathcal{C}}$ is indeed transitive and reflexive. In the context of formal language theory, such preorders have been introduced in [13]. In

our context, while the above definition makes sense for *arbitrary* classes, it is meant to be used for *finite* lattices: all results that we present require this hypothesis.

We say that a language $L \subseteq A^*$ is an *upper set* for $\leq_{\mathcal{C}}$ when for any two words $u, v \in A^*$, if $u \leq_{\mathcal{C}} v$ and $u \in L$, then also $v \in L$. In other words, L is an upper set when it coincides with its upwards closure $\uparrow_{\mathcal{C}} L$ with respect to the preorder $\leq_{\mathcal{C}}$, where $\uparrow_{\mathcal{C}} L$ is defined as the set of words that sit above some word of L :

$$\uparrow_{\mathcal{C}} L = \{u \in A^* \mid \exists w \in L, w \leq_{\mathcal{C}} u\}.$$

Lemma 18 *Let \mathcal{C} be a lattice. Then, for any word $w \in A^*$, we have:*

$$\uparrow_{\mathcal{C}} w = \bigcap_{L \in \mathcal{C} \text{ and } w \in L} L.$$

In particular, if \mathcal{C} is finite, the canonical preorder $\leq_{\mathcal{C}}$ has finitely many upper sets.

Proof The equality follows directly from the definition of $\leq_{\mathcal{C}}$. Let us prove the second assertion. By definition, an upper set is a union of languages of the form $\uparrow_{\mathcal{C}} w$. Hence, it suffices to prove that there are finitely many languages $\uparrow_{\mathcal{C}} w$, which follows immediately from the equality of the lemma and the finiteness of \mathcal{C} . \square

We now prove the second important property of the preorder $\leq_{\mathcal{C}}$: we use the fact that \mathcal{C} is a finite lattice to characterize the languages belonging to \mathcal{C} . They are exactly the upper sets for $\leq_{\mathcal{C}}$.

Lemma 19 *Let \mathcal{C} be a finite lattice. Then, \mathcal{C} is exactly the set of upper sets of $\leq_{\mathcal{C}}$.*

Proof Assume first that $L \in \mathcal{C}$. Then for all $w \in L$ and all w' such that $w \leq_{\mathcal{C}} w'$, we have $w' \in L$ by definition of $\leq_{\mathcal{C}}$. Hence, L is an upper set.

Assume now that L is an upper set. Observe that since \mathcal{C} is finite and closed under intersection, for any word w , the upper set $\uparrow_{\mathcal{C}} w$ belongs to \mathcal{C} by Lemma 18: it is the intersection of all languages in \mathcal{C} containing w . Furthermore, $L = \bigcup_{w \in L} \uparrow_{\mathcal{C}} w$. By Lemma 18, there are only finitely many sets of the form $\uparrow_{\mathcal{C}} w$. Since \mathcal{C} is closed under finite union, L belongs to \mathcal{C} . \square

While Lemma 19 states an equivalence, we mainly use the left to right implication (or rather its contrapositive). It is useful for proving that a given language L does *not* belong to \mathcal{C} , or that two languages K, L are *not* \mathcal{C} -separable. We describe this application in the following corollary.

Corollary 20 *Let \mathcal{C} be a finite lattice and let $K, L \subseteq A^*$ be two languages. Then, the two following properties hold:*

1. L does **not** belong to \mathcal{C} if and only if there exist $w \in L$ and $w' \notin L$ such that $w \leq_{\mathcal{C}} w'$.
2. L is **not** \mathcal{C} -separable from K if and only if there exist $w \in L$ and $w' \in K$ such that $w \leq_{\mathcal{C}} w'$.

Proof The first assertion is immediate from Lemma 19. We prove the second. Assume that there exist $w \in L$ and $w' \in K$ such that $w \leq_{\mathcal{C}} w'$. Hence, for any language H separating L from K , we have $w \in H$ and $w' \notin H$. Since $w \leq_{\mathcal{C}} w'$, this means that $H \notin \mathcal{C}$ by definition of $\leq_{\mathcal{C}}$. Therefore, L is not \mathcal{C} -separable from K .

Conversely, assume that L is not \mathcal{C} -separable from K . For any $w \in A^*$, note that $\uparrow_{\mathcal{C}} w$ belongs to \mathcal{C} by Lemma 19, since $\uparrow_{\mathcal{C}} w$ is an upper set for $\leq_{\mathcal{C}}$. We define:

$$H = \bigcup_{w \in L} \uparrow_{\mathcal{C}} w.$$

This union is finite since $\leq_{\mathcal{C}}$ has finitely many upper sets. Hence, $H \in \mathcal{C}$. Moreover, $L \subseteq H$ by definition. Hence, since L is not \mathcal{C} -separable from K , we have $H \cap K \neq \emptyset$. Let $w' \in H \cap K$. By definition $w' \in \uparrow_{\mathcal{C}} w$ for some $w \in L$ which means that $w \leq_{\mathcal{C}} w'$. Therefore, we have found $w \in L$ and $w' \in K$ such that $w \leq_{\mathcal{C}} w'$. \square

Example 21 Let $L = a^*b^*$. Then $L \notin \text{AT}$. Indeed, we have $ab \in L$ and $ba \notin L$ while $\text{alph}(ab) = \text{alph}(ba)$ which means that $ab \leq_{\text{AT}} ba$. For the same reason, a^+b^+ is not AT-separable from b^+a^+ . \square

Canonical preorders for quotienting lattices. We now present additional properties of the canonical $\leq_{\mathcal{C}}$ preorder that hold when the finite lattice \mathcal{C} is closed under quotients. The key property is given in the following lemma: closure under quotients for \mathcal{C} corresponds to compatibility with word concatenation for $\leq_{\mathcal{C}}$.

Lemma 22 *A finite lattice \mathcal{C} is closed under quotient if and only if its associated canonical preorder $\leq_{\mathcal{C}}$ is a precongruence for word concatenation. That is, for any words u, v, u', v' ,*

$$u \leq_{\mathcal{C}} u' \quad \text{and} \quad v \leq_{\mathcal{C}} v' \quad \Rightarrow \quad uv \leq_{\mathcal{C}} u'v'.$$

Proof We do the proof for lattices. First assume that \mathcal{C} is closed under quotients and let u, u', v, v' be four words such that $u \leq_{\mathcal{C}} u'$ and $v \leq_{\mathcal{C}} v'$. We have to prove that $uv \leq_{\mathcal{C}} u'v'$. Let $L \in \mathcal{C}$ and assume that $uv \in L$. This means that $v \in u^{-1} \cdot L$. By closure under left quotient, we have $u^{-1}L \in \mathcal{C}$, hence, since $v \leq_{\mathcal{C}} v'$, we obtain that $v' \in u^{-1} \cdot L$ and therefore that $uv' \in L$. It now follows that $u \in L(v')^{-1}$. Using closure under right quotient, we obtain that $L(v')^{-1} \in \mathcal{C}$. Therefore, since $u \leq_{\mathcal{C}} u'$, we conclude that $u' \in L(v')^{-1}$ which means that $u'v' \in L$, as desired.

Conversely, assume that $\leq_{\mathcal{C}}$ is a precongruence. Let $L \in \mathcal{C}$ and $w \in A^*$, we prove that $w^{-1}L \in \mathcal{C}$ (the proof for right quotients is symmetrical). By Lemma 19, we have to prove that $w^{-1}L$ is an upper set. Let $u \in w^{-1}L$ and $u' \in A^*$ such that $u \leq_{\mathcal{C}} u'$. Since $\leq_{\mathcal{C}}$ is a precongruence, we have $wu \leq_{\mathcal{C}} wu'$. Hence, since L is an upper set (it belongs to \mathcal{C}) and $wu \in L$, we have $wu' \in L$. We conclude that $u' \in w^{-1}L$, which completes the proof. \square

3 Boolean and polynomial closures

Classes of languages are often built from simpler classes by using *closure operators*. We are interested in two such operators: Boolean and polynomial closure. In this section, we define these operators and describe some of their important properties.

3.1 Definition

Boolean closure. Given a class of languages \mathcal{C} , the *Boolean closure of \mathcal{C}* , denoted by $Bool(\mathcal{C})$, is the least Boolean algebra containing \mathcal{C} , *i.e.*, the least class of languages containing \mathcal{C} closed under union, intersection and complement. Observe that by definition, the Boolean closure of a class is a Boolean algebra. In particular, it follows that Boolean closure is an idempotent operation: for any class \mathcal{C} , we have $Bool(Bool(\mathcal{C})) = Bool(\mathcal{C})$. Furthermore, Boolean closure preserves many properties of the input class \mathcal{C} . In particular, it preserves closure under quotient.

Proposition 23 *For any quotienting class of languages \mathcal{C} , the Boolean closure of \mathcal{C} is a quotienting Boolean algebra.*

Proof Immediate since quotients commute with Boolean operations. \square

Remark 24 Not all closure properties are preserved under Boolean closure. The most significant example is closure under concatenation. In fact, all classes that we build with Boolean closure will *not* be closed under concatenation. This is a problem, since our techniques for solving membership and separation rely on concatenation.

Polynomial closure. We turn to the second operation: *polynomial closure*. Let \mathcal{C} be a class of languages. We say that a language $L \subseteq A^*$ is a *\mathcal{C} -monomial* when there exists a natural number $n \in \mathbb{N}$, $L_0, \dots, L_n \in \mathcal{C}$ and $a_1, \dots, a_n \in A$ such that,

$$L = L_0 a_1 L_1 a_2 L_2 \cdots a_n L_n.$$

The least integer n for which this property holds is called the *degree* of L . Finally, a *\mathcal{C} -polynomial* is a finite union of \mathcal{C} -monomials. The degree of a \mathcal{C} -polynomial L is the least integer n such that L is a union of \mathcal{C} -monomials having degree at most n .

We may now define polynomial closure. For any class of languages \mathcal{C} , we call *polynomial closure of \mathcal{C}* the class of all \mathcal{C} -polynomials. We denote it by $Pol(\mathcal{C})$. Note that $\mathcal{C} \subseteq Pol(\mathcal{C})$, since the languages in \mathcal{C} are the \mathcal{C} -monomials of degree 0.

Example 25 Consider $\mathcal{C} = \{\emptyset, A^*\}$. Then $Pol(\mathcal{C})$ consists of all finite unions of languages of the form $A^* a_1 A^* \cdots a_n A^*$ for $n \geq 0$ and $a_1, \dots, a_n \in A$. \square

Example 26 Consider the class AT of alphabet testable languages from Example 15. One may check that $Pol(AT)$ consists of all finite unions of languages of the form

$$B_0^* a_1 B_1^* a_2 B_2^* \cdots a_n B_n^* \quad \text{with } a_1, \dots, a_n \in A \text{ and } B_0, \dots, B_n \subseteq A. \quad \square$$

It is not immediate from the definition that classes built with polynomial closure have much structure. First, $Pol(\mathcal{C})$ is closed under union, by definition. Moreover, we have the following result, which holds when \mathcal{C} is quotienting.

Lemma 27 *For any quotienting class \mathcal{C} , its polynomial closure $Pol(\mathcal{C})$ is quotienting and closed under union, marked concatenation and classical concatenation.*

Proof We first show that $Pol(\mathcal{C})$ is closed under quotients. Given any $w \in A^*$ and any \mathcal{C} -polynomial L over A , we have to prove that $w^{-1}L$ and Lw^{-1} are \mathcal{C} -polynomials as well. We present a proof for $w^{-1}L$ (the argument for Lw^{-1} is symmetrical). We may assume without loss of generality that $w = a \in A$, since $(ua)^{-1}L = a^{-1}(u^{-1}L)$.

Moreover, since quotients commute with union, it suffices to consider the case when L is a \mathcal{C} -monomial. Hence, there exist $L_0, \dots, L_n \in \mathcal{C}$ and $a_1, \dots, a_n \in A$ such that,

$$L = L_0 a_1 L_1 a_2 L_2 \cdots a_n L_n.$$

There are two cases depending on whether $\varepsilon \in L_0$ and $a_1 = a$, or not. We have:

$$a^{-1}L = \begin{cases} (a^{-1}L_0)a_1 L_1 a_2 L_2 \cdots a_n L_n \cup L_1 a_2 L_2 \cdots a_n L_n & \text{if } \varepsilon \in L_0 \text{ and } a_1 = a, \\ (a^{-1}L_0)a_1 L_1 a_2 L_2 \cdots a_n L_n & \text{otherwise.} \end{cases}$$

Since \mathcal{C} is closed under quotient by hypothesis, $w^{-1}L$ is a union of \mathcal{C} -monomials and is itself a \mathcal{C} -polynomial. This terminates the proof for closure under quotient.

That $Pol(\mathcal{C})$ is closed under union and marked concatenation is immediate by definition. Let us show closure under classical concatenation. Let K and L be two languages in $Pol(\mathcal{C})$. One may verify that:

$$KL = \begin{cases} \bigcup_{a \in A} Ka(a^{-1}L) \cup K & \text{if } \varepsilon \in L, \\ \bigcup_{a \in A} Ka(a^{-1}L) & \text{if } \varepsilon \notin L. \end{cases}$$

In either case, we obtain a language from $Pol(\mathcal{C})$ by closure under quotients and marked concatenation. This concludes the proof. \square

When \mathcal{C} contains the singleton language $\{\varepsilon\}$, one may prove a stronger variant of Lemma 27. While simple, this observation is important as we will rely on it later.

Lemma 28 *Let \mathcal{C} be a class of languages. Then, $\{\varepsilon\} \in \mathcal{C}$ if and only if $\{\varepsilon\} \in Pol(\mathcal{C})$. Moreover, in that case, for any $w \in A^+$ and $K, L \in Pol(\mathcal{C})$, we have:*

$$KwL \in Pol(\mathcal{C}) \quad wL \in Pol(\mathcal{C}) \quad Kw \in Pol(\mathcal{C}) \quad \{w\} \in Pol(\mathcal{C}).$$

Proof Clearly, if $\{\varepsilon\} \in \mathcal{C}$, then $\{\varepsilon\} \in Pol(\mathcal{C})$ since $\mathcal{C} \subseteq Pol(\mathcal{C})$. Conversely, it suffices to observe that a \mathcal{C} -polynomial of degree $n \geq 1$ cannot be equal to $\{\varepsilon\}$. Hence, if $\{\varepsilon\} \in Pol(\mathcal{C})$, then $\{\varepsilon\}$ is a \mathcal{C} -polynomial of degree 0, *i.e.*, an element of \mathcal{C} .

We now assume that $\{\varepsilon\} \in Pol(\mathcal{C})$. Consider $w \in A^+$ and $K, L \in Pol(\mathcal{C})$. We have to prove that $KwL, wL, Kw, \{w\} \in Pol(\mathcal{C})$. We present a proof for KwL (the other cases follow by choosing K or L to be $\{\varepsilon\}$). Let $w = a_1 \cdots a_\ell$ with $a_1, \dots, a_\ell \in A$. It is immediate that $KwL = Ka_1\{\varepsilon\}a_2\{\varepsilon\} \cdots \{\varepsilon\}a_\ell L$. Hence, $KwL \in Pol(\mathcal{C})$ by closure under marked concatenation. \square

Without more hypothesis on the input class \mathcal{C} , these are the only properties of $Pol(\mathcal{C})$ that we are able to prove. However, it was proved by Arfi that when \mathcal{C} is a quotienting lattice consisting only of regular languages, $Pol(\mathcal{C})$ is a quotienting lattice as well. In other words, in that case, polynomial closure preserves closure under intersection. The proof is not immediate and depends on all hypotheses: the fact that \mathcal{C} is closed under quotient and contains only regular languages is crucial.

Theorem 29 (Arfi [2, 3], Pin [25]) *Let \mathcal{C} be a quotienting lattice of regular languages. Then $Pol(\mathcal{C})$ is also a quotienting lattice of regular languages. In particular, $Pol(\mathcal{C})$ is the least lattice containing \mathcal{C} and closed under marked concatenation.*

Proof Let us fix a quotienting lattice \mathcal{C} of regular languages. We prove that $Pol(\mathcal{C})$ is a quotienting lattice of regular languages. It is closed under union and quotients by Lemma 27. Hence, we may concentrate on closure under intersection. Let K, L be two \mathcal{C} -polynomials and let $m, n \in \mathbb{N}$ be their degrees. We have to prove that $K \cap L$ is a \mathcal{C} -polynomial as well. Note that since intersection distributes over union, we may assume without loss of generality that K and L are both \mathcal{C} -monomials. We begin by treating the special case when either K or L has degree 0 using induction on the degree of the other one. We then use this special case to prove the general one.

Special Case: Either K or L has degree 0. By symmetry, we may assume that K has degree 0, which means that $K \in \mathcal{C}$. We use induction on the degree n of the \mathcal{C} -monomial L . If $n = 0$, then K and L both belong to \mathcal{C} , which is closed under intersection by hypothesis. Hence we conclude that $K \cap L \in \mathcal{C} \subseteq Pol(\mathcal{C})$. Otherwise, L has degree $n \geq 1$ and by definition, it can be decomposed as follows: $L = L_1 b L_2$ where $L_1 \in \mathcal{C}$ and L_2 is a \mathcal{C} -monomial of degree at most $n - 1$.

Observe that a word w belongs to $K \cap L$ when it belongs to K and can be decomposed as $w = w_1 b w_2$ with $w_1 \in L_1$ and $w_2 \in L_2$. Given any word $u \in A^*$, we let Q_u be the set of all words $x \in A^*$ such that $u \in K(bx)^{-1}$. We claim that the following equality holds:

$$K \cap L = \bigcup_{u \in A^*} \left(L_1 \cap \bigcap_{x \in Q_u} K(bx)^{-1} \right) \cdot b \cdot (L_2 \cap (ub)^{-1} K). \quad (1)$$

Before proving this claim, let us explain why it helps conclude the proof that $K \cap L \in Pol(\mathcal{C})$. First observe that since $\mathcal{C} \subseteq REG$, we know from the Myhill-Nerode theorem (Theorem 16) that there are finitely many quotients of K . Hence, there are finitely many languages of the form $\bigcap_{x \in Q_u} K(bx)^{-1}$ and $(ub)^{-1} K$, which means that the union over all $u \in A^*$ in (1) actually ranges over finitely many distinct languages. Therefore, since $Pol(\mathcal{C})$ is closed under finite union, it suffices to prove that for any $u \in A^*$,

$$\left(L_1 \cap \bigcap_{x \in Q_u} K(bx)^{-1} \right) \cdot b \cdot (L_2 \cap (ub)^{-1} K) \text{ belongs to } Pol(\mathcal{C}).$$

Since $Pol(\mathcal{C})$ is closed under marked concatenation (see Lemma 27), it suffices to prove that the two following properties hold:

$$L_1 \cap \bigcap_{x \in Q_u} K(bx)^{-1} \in Pol(\mathcal{C}) \quad \text{and} \quad L_2 \cap (ub)^{-1} K \in Pol(\mathcal{C}).$$

For the first property, we know by hypothesis that $L_1 \in \mathcal{C}$. Moreover, for any $x \in Q_u$, $K(bx)^{-1} \in \mathcal{C}$ since it is a quotient of $K \in \mathcal{C}$. Furthermore, since K is regular (as $K \in \mathcal{C} \subseteq REG$), it has finitely many quotients, which means that $L_1 \cap \bigcap_{x \in Q_u} K(bx)^{-1}$ is a finite intersection of languages in \mathcal{C} and belongs to \mathcal{C} as well. Since $\mathcal{C} \subseteq Pol(\mathcal{C})$, it belongs also to $Pol(\mathcal{C})$. For the second property, note that $(ub)^{-1} K \in \mathcal{C}$ since \mathcal{C} is closed under quotient. Therefore, since L_2 is a \mathcal{C} -polynomial of degree at most $n - 1$ by hypothesis, we may use induction to conclude that $L_2 \cap (ub)^{-1} K \in Pol(\mathcal{C})$.

It remains to prove that (1) holds. Assume first that $w \in K \cap L$. Then $w \in K$ and it can be decomposed as $w = w_1 b w_2$ with $w_1 \in L_1$ and $w_2 \in L_2$. It follows that

$w_2 \in L_2 \cap (w_1 b)^{-1} K$ and by definition of Q_{w_1} , we have $w_1 \in L_1 \cap \bigcap_{x \in Q_{w_1}} K(bx)^{-1}$. This concludes the proof of the left to right inclusion.

Conversely, assume that w belongs to the right hand side of (1). We have to prove that $w \in K \cap L$. It is immediate from the definition that $w \in L_1 b L_2 = L$. It remains to prove that $w \in K$. By definition, there exists $u \in A^*$ such that w can be decomposed as $w = w_1 b w_2$ with $w_1 \in \bigcap_{x \in Q_u} K(bx)^{-1}$ and $w_2 \in (ub)^{-1} K$. Since $w_2 \in (ub)^{-1} K$, we have $u \in K(bw_2)^{-1}$ and therefore $w_2 \in Q_u$ by definition. We conclude that $w_1 \in K(bw_2)^{-1}$, which exactly means that $w = w_1 b w_2 \in K$.

General case. We now assume that K and L both have arbitrary degrees m and n . To prove that $K \cap L \in \text{Pol}(\mathcal{C})$, we proceed by induction on the sum $m + n$ of the degrees. The argument is similar to the one above. If $m = 0$ or $n = 0$, this is exactly the special case. Hence, we assume that $m, n \geq 1$: K and L may be decomposed as:

$$K = K_1 a K_2 \quad \text{and} \quad L = L_1 b L_2,$$

where $K_1, K_2 \in \mathcal{C}$ and L_1, L_2 are \mathcal{C} -monomials of degree at most $m - 1$ and $n - 1$. Observe that a word w belongs to $K \cap L$ if and only if it admits two decompositions witnessing its membership in $K_1 a K_2$ and $L_1 b L_2$, respectively. We use this observation to break down $K \cap L$ as the union of two languages (or three depending on whether $a = b$ or not). Consider the three following languages:

$$\begin{aligned} H_\ell &= \{w_1 a w_2 b w_3 \mid w_1 \in K_1, w_2 b w_3 \in K_2, w_1 a w_2 \in L_1 \text{ and } w_3 \in L_2\}, \\ H_r &= \{w_1 b w_2 a w_3 \mid w_1 \in L_1, w_2 a w_3 \in L_2, w_1 b w_2 \in K_1 \text{ and } w_3 \in K_2\}, \\ H_c &= \{w_1 a w_2 \mid w_1 \in L_1 \cap K_1 \text{ and } w_2 \in L_2 \cap K_2\}. \end{aligned}$$

It is simple to verify that $K \cap L = H_\ell \cup H_r$ when $a \neq b$ and $K \cap L = H_\ell \cup H_r \cup H_c$ when $a = b$. Hence, it suffices to prove that H_ℓ, H_r and H_c are \mathcal{C} -polynomials. Since the proof is similar for all three cases, we concentrate on H_ℓ .

Given any word $u \in A^*$, we write P_u for the set of all words $x \in A^*$ such that $u \in L_1(ax)^{-1}$, i.e., $P_u = (ua)^{-1} L_1$. We claim that the following equality holds:

$$H_\ell = \bigcup_{u \in A^*} \left(K_1 \cap \bigcap_{x \in P_u} L_1(ax)^{-1} \right) \cdot a \cdot (K_2 \cap ((ua)^{-1} L_1) b L_2). \quad (2)$$

Before establishing (2), let us use it to prove that $H_\ell \in \text{Pol}(\mathcal{C})$. Since $\mathcal{C} \subseteq \text{REG}$, we know from the Myhill-Nerode Theorem (Theorem 16) that there are finitely many quotients of L_1 . Hence, there are finitely many languages of the form $\bigcap_{x \in P_u} L_1(ax)^{-1}$ and $((ua)^{-1} L_1) b L_2$ for $u \in A^*$. Since $\text{Pol}(\mathcal{C})$ is closed under finite union, we obtain from (2) that in order to show $H_\ell \in \text{Pol}(\mathcal{C})$, it suffices to prove that for all $u \in A^*$,

$$\left(K_1 \cap \bigcap_{x \in P_u} L_1(ax)^{-1} \right) \cdot a \cdot (K_2 \cap ((ua)^{-1} L_1) b L_2) \text{ belongs to } \text{Pol}(\mathcal{C}).$$

Let $u \in A^*$. Since $\text{Pol}(\mathcal{C})$ is closed under marked concatenation (see Lemma 27) it suffices to prove that the two following properties hold:

$$K_1 \cap \bigcap_{x \in P_u} L_1(ax)^{-1} \in \text{Pol}(\mathcal{C}) \quad \text{and} \quad K_2 \cap ((ua)^{-1} L_1) b L_2 \in \text{Pol}(\mathcal{C}).$$

The argument for the first property is identical to the special case that we treated above. Indeed, we know that $K_1 \in \mathcal{C}$ by hypothesis. Moreover, for any $x \in P_u$, $L_1(ax)^{-1} \in \mathcal{C}$ (it is a quotient of $L_1 \in \mathcal{C}$). Finally, $L_1 \in \mathcal{C} \subseteq \text{REG}$ has finitely many quotients and $K_1 \cap \bigcap_{x \in P_u} L_1(ax)^{-1}$ is a finite intersection of languages in \mathcal{C} and belongs to \mathcal{C} as well, whence to $\text{Pol}(\mathcal{C})$. For the other property, we have $(ua)^{-1}L_1 \in \mathcal{C}$ since \mathcal{C} is closed under quotient. Moreover, since L_2 is a \mathcal{C} -monomial of degree at most $n-1$, this entails that $((ua)^{-1}L_1)bL_2$ is a \mathcal{C} -monomial of degree at most n . Finally, K_2 is a \mathcal{C} -monomial of degree at most $m-1$. Therefore, we may use induction on the sum of the degrees to conclude that $K_2 \cap ((ua)^{-1}L_1)bL_2 \in \text{Pol}(\mathcal{C})$.

It remains to prove that (2) holds. Assume first that $w \in H_\ell$. Then $w = w_1aw_2bw_3$ with $w_1 \in K_1$, $w_2bw_3 \in K_2$, $w_1aw_2 \in L_1$ and $w_3 \in L_2$. It follows that $w_2 \in (w_1a)^{-1}L_1$ and therefore, $w_2bw_3 \in K_2 \cap ((w_1a)^{-1}L_1)bL_2$. Finally, by definition of P_{w_1} , we have $w_1 \in K_1 \cap \bigcap_{x \in P_{w_1}} L_1(ax)^{-1}$. Therefore, we conclude that,

$$w \in \left(K_1 \cap \bigcap_{x \in P_{w_1}} L_1(ax)^{-1} \right) \cdot a \cdot \left(K_2 \cap ((w_1a)^{-1}L_1)bL_2 \right).$$

This concludes the proof of the left to right inclusion. Conversely, assume that w belongs to the right hand side of (2). We have to prove that $w \in H_\ell$. By definition, there exists $u \in A^*$ such that $w = w_1ay$ with,

$$w_1 \in K_1 \cap \bigcap_{x \in P_u} L_1(ax)^{-1} \quad \text{and} \quad y \in K_2 \cap ((ua)^{-1}L_1)bL_2.$$

Therefore, y can be decomposed as $y = w_2bw_3$ with $w_2 \in (ua)^{-1}L_1$ and $w_3 \in L_2$. Hence, we have $w = w_1aw_2bw_3$ with $w_1 \in K_1$, $w_2bw_3 \in K_2$ and $w_3 \in L_2$. To prove that $w \in H_\ell$, it remains to prove that $w_1aw_2 \in L_1$. Since $w_2 \in (ua)^{-1}L_1$, we have $u \in L_1(aw_2)^{-1}$ and therefore $w_2 \in P_u$ by definition. We conclude that $w_1 \in L_1(aw_2)^{-1}$, which exactly means that $w = w_1aw_2 \in L_1$, as desired. \square

Additional operations. We end this section by looking at two additional operations that are built from Boolean and polynomial closures. The first one is simply the composition of the two: given any class \mathcal{C} , we write $B\text{Pol}(\mathcal{C})$ for the class $\text{Bool}(\text{Pol}(\mathcal{C}))$. Combining the results of the previous subsections, we obtain the following result.

Proposition 30 *If \mathcal{C} is a quotienting class, then the class $B\text{Pol}(\mathcal{C})$ is a quotienting Boolean algebra.*

The second operation is motivated by a simple observation about polynomial closure. We proved in Theorem 29 that $\text{Pol}(\mathcal{C})$ is a lattice provided that \mathcal{C} is a quotienting lattice of regular languages. However, it is simple to verify that in general, $\text{Pol}(\mathcal{C})$ is *not* closed under complement, even when \mathcal{C} is.

Example 31 Consider the class $\mathcal{C} = \{\emptyset, A^*\}$. It turns out that $\text{Pol}(\mathcal{C})$ is not closed under complement. Indeed, it is clear that $A^+ = \bigcup_{a \in A} A^*aA^* \in \text{Pol}(\mathcal{C})$. However, it follows from Lemma 28 that its complement $\{\varepsilon\}$ does not belong to $\text{Pol}(\mathcal{C})$. \square

When dealing with a lattice \mathcal{D} which is not closed under complement, it makes sense to consider the *complement class* which we denote by $\text{co-}\mathcal{D}$. More precisely, $\text{co-}\mathcal{D}$ contains all complements of languages in \mathcal{D} : for any language $L \subseteq A^*$, we have $L \in \text{co-}\mathcal{D}$

if and only if $A^* \setminus L \in \mathcal{D}$. Note that since \mathcal{D} is a lattice, so is $co\mathcal{D}$ by De Morgan's laws. We shall often consider the complement of classes which are polynomial closures: given a class \mathcal{C} , we consider $co\text{-}Pol(\mathcal{C})$.

It is simple to transfer most of the closure properties of some class \mathcal{D} to its complement class $co\mathcal{D}$. This yields the following proposition.

Proposition 32 *If \mathcal{D} is a quotienting lattice, then so is $co\mathcal{D}$.*

Proof The property of being a quotienting lattice is transferred from \mathcal{D} to $co\mathcal{D}$ using De Morgan's laws and the fact that quotients commute with Boolean operations. \square

However, not all properties from \mathcal{D} are inherited by $co\mathcal{D}$. A notable example is closure under concatenation and marked concatenation. While $Pol(\mathcal{C})$ is closed under marked concatenation (and usual concatenation when \mathcal{C} is quotienting), this is not the case for $co\text{-}Pol(\mathcal{C})$ in general, even when \mathcal{C} is a quotienting Boolean algebra. This is easy to check for $\mathcal{C} = \{A^*, \emptyset\}$, since A^* belongs to $co\text{-}Pol(\mathcal{C})$ while A^*aA^* does not.

By definition, $co\mathcal{D}$ and \mathcal{D} are dual and the associated membership and separation problems are inter-reducible. Indeed, $L \in co\mathcal{D}$ if and only if $A^* \setminus L \in \mathcal{D}$ and L_1 is $co\mathcal{D}$ -separable from L_2 if and only if L_2 is \mathcal{D} -separable from L_1 . Hence, when considering these problems, one may simply work with \mathcal{D} instead of $co\mathcal{D}$.

In view of these observations, one might wonder about our motivation for considering the complement of polynomial closure. Indeed, we just explained that $co\text{-}Pol(\mathcal{C})$ is less robust than $Pol(\mathcal{C})$, while the associated decision problems are symmetrical to the ones associated to $Pol(\mathcal{C})$. Our motivation is explained by the next lemma, which can be used to bypass Boolean closure in some cases.

Lemma 33 *Let \mathcal{C} be a quotienting Boolean algebra of regular languages. Then,*

$$Pol(BPol(\mathcal{C})) = Pol(co\text{-}Pol(\mathcal{C})).$$

Proof It is clear that $co\text{-}Pol(\mathcal{C}) \subseteq BPol(\mathcal{C})$, whence $Pol(co\text{-}Pol(\mathcal{C})) \subseteq Pol(BPol(\mathcal{C}))$. We show that $BPol(\mathcal{C}) \subseteq Pol(co\text{-}Pol(\mathcal{C}))$. It will follow that,

$$Pol(BPol(\mathcal{C})) \subseteq Pol(Pol(co\text{-}Pol(\mathcal{C}))) = Pol(co\text{-}Pol(\mathcal{C})).$$

Let $L \in BPol(\mathcal{C})$. By definition, L is a Boolean combination of \mathcal{C} -polynomials. De Morgan's laws show that L is built by applying unions and intersections to languages that are either \mathcal{C} -polynomials (*i.e.*, elements of $Pol(\mathcal{C})$) or complements of \mathcal{C} -polynomials (*i.e.*, elements of $co\text{-}Pol(\mathcal{C})$). It follows that $L \in Pol(co\text{-}Pol(\mathcal{C}))$, since $Pol(\mathcal{C}) \subseteq Pol(co\text{-}Pol(\mathcal{C}))$, $co\text{-}Pol(\mathcal{C}) \subseteq Pol(co\text{-}Pol(\mathcal{C}))$ and $Pol(co\text{-}Pol(\mathcal{C}))$ is closed under union and intersection by Theorem 29. \square

Let us finish the section with a recap of the closure properties that we proved for the four operations on classes that we defined. We present it in Fig. 2.

3.2 A characteristic property of $Pol(\mathcal{C})$

We complete the definitions by presenting a characteristic property applying to the polynomial closure of every **finite** quotienting lattice. Its main application will be to show that specific languages do not belong to $Pol(\mathcal{C})$. Recall that when \mathcal{C} is a finite quotienting lattice, we associate a canonical preorder $\leq_{\mathcal{C}}$ over A^* : $w \leq_{\mathcal{C}} w'$ if and only if every $L \in \mathcal{C}$ containing w contains w' as well. Moreover, since \mathcal{C} is closed under quotients, $\leq_{\mathcal{C}}$ is compatible with word concatenation by Lemma 22.

	Intersection and Union	Complement	Quotient	Concat.	Marked concat.
$BPol(\mathcal{C})$	Y	Y	Y	N	N
$Bool(\mathcal{C})$	Y	Y	Y	N	N
$Pol(\mathcal{C})$	Y	N	Y	Y	Y
$co-Pol(\mathcal{C})$	Y	N	Y	N	N

Fig. 2 Closure properties satisfied by any quotienting lattice of regular languages \mathcal{C}

Proposition 34 *Let \mathcal{C} be a finite quotienting lattice and $L \in Pol(\mathcal{C})$. There exist numbers $h, p \geq 1$ such that for every $\ell \geq h$ and $u, v, x, y \in A^*$ satisfying $u \leq_{\mathcal{C}} v$,*

$$xu^{p\ell+1}y \in L \quad \Rightarrow \quad xu^{p\ell}vu^{p\ell}y \in L.$$

Before we prove Proposition 34, let us state an important corollary which we shall use to prove that particular languages are not $Pol(\mathcal{C})$ -separable.

Corollary 35 *Let \mathcal{C} be a finite quotienting lattice and two languages $H, L \subseteq A^*$. Assume that there exist $u, v, x, y \in A^*$ such that $u \leq_{\mathcal{C}} v$ and for every $n \geq 1$, $xu^n y \in H$ and $xu^n vu^n y \in L$. Then H is **not** $Pol(\mathcal{C})$ -separable from L .*

Proof By contradiction assume that we have $K \in Pol(\mathcal{C})$ which separates H from L . By hypothesis on $u, v, x, y \in A^*$, Proposition 34 yields $h, p \geq 1$ such that:

$$xu^{ph+1}y \in K \quad \Rightarrow \quad xu^{ph}vu^{ph}y \in K.$$

Moreover, $xu^{ph+1}y \in H$ and $xu^{ph}vu^{ph}y \in L$ by hypothesis and since K separates H from L , this implies $xu^{ph+1}y \in K$ and $xu^{ph}vu^{ph}y \notin K$. This contradicts the above implication, finishing off the proof. \square

The remainder of the section is devoted to proving Proposition 34. We fix the finite quotienting lattice \mathcal{C} and $L \in Pol(\mathcal{C})$ for the proof. We first need to choose $h, p \geq 1$. We start with p whose choice is based on the following fact.

Fact 36 *There exists $p \geq 1$ such that for any $m, m' \geq 1$ and $w \in A^*$, $w^{pm} \leq_{\mathcal{C}} w^{pm'}$.*

Proof Let \sim be the equivalence on A^* generated by $\leq_{\mathcal{C}}$. Since $\leq_{\mathcal{C}}$ is a preorder with finitely many upper sets which is compatible with concatenation (see Lemma 19 and 22), \sim must be a congruence of finite index. Therefore, the set A^*/\sim of \sim -classes is a finite monoid. It is standard that for any finite monoid M , there exists a natural number $\omega(M)$ such that for any $s \in M$ $s^{\omega(M)}$ is idempotent ($s^{\omega(M)} = s^{\omega(M)}s^{\omega(M)}$). It suffices to choose $p = \omega(A^*/\sim)$. \square

We now define $h \geq 1$. Since L belongs to $Pol(\mathcal{C})$, it is \mathcal{C} -polynomial. We let n as its degree and define $h = n + 1$. It remains to show that h satisfies the property described in Proposition 34. Let $\ell \geq h$ and $u, v, x, y \in A^*$ satisfying $u \leq_{\mathcal{C}} v$. We show that:

$$xu^{p\ell+1}y \in L \quad \Rightarrow \quad xu^{p\ell}vu^{p\ell}y \in L.$$

Consequently, assume that $xu^{p\ell+1}y \in L$. Since L is a \mathcal{C} -polynomial of degree n , there exists a \mathcal{C} -monomial $L_0a_1L_1 \cdots a_mL_m$ of degree $m \leq n$ with $w \in L_0a_1L_1 \cdots a_mL_m \subseteq L$. It follows that $xu^{p\ell+1}y$ admits a unique decomposition:

$$xu^{p\ell+1}y = w_0a_1w_1 \cdots a_mw_m.$$

such that $w_i \in L_i$ for all $i \leq m$. Recall that by definition $\ell \geq h = n + 1 \geq m + 1$. Therefore, it follows from the pigeonhole principle that some infix u^p of $xu^{p\ell+1}y$ must be contained within one of the infixes w_i . Let us formalize this result.

Lemma 37 *There exist $i \leq m$, $j_1, j_2 < \ell$ such that $j_1 + 1 + j_2 = \ell$ and $x_1, x_2 \in A^*$ satisfying,*

- $w_i = x_1u^px_2$.
- $w_0a_1w_1 \cdots a_ix_1 = xu^{pj_1}$.
- $x_2a_{i+1} \cdots a_mw_m = u^{pj_2+1}y$.

We now use Lemma 37 to finish the proof. By Fact 36, we have the following inequality:

$$u^p \leq_{\mathcal{C}} u^{p(\ell+1)} = u^{p(j_1+1+j_2+1)} = u^{p(j_2+1)}uu^{p(j_1+1)-1}.$$

Moreover, since $u \leq_{\mathcal{C}} v$ and $\leq_{\mathcal{C}}$ is compatible with concatenation this yields that,

$$u^p \leq_{\mathcal{C}} u^{p(j_2+1)}uu^{p(j_1+1)-1} \leq_{\mathcal{C}} u^{p(j_2+1)}vu^{p(j_1+1)-1}.$$

Using again compatibility with concatenation we obtain,

$$w_i = x_1u^px_2 \leq_{\mathcal{C}} x_1u^{p(j_2+1)}vu^{p(j_1+1)-1}x_2.$$

Therefore, since $w_i \in L_i$ which is a language of \mathcal{C} , it follows from the definition of $\leq_{\mathcal{C}}$ that $x_1u^{p(j_2+1)}vu^{p(j_1+1)-1}x_2 \in L_i$. Therefore, since $w_j \in L_j$ for all j ,

$$w_0a_1w_1 \cdots a_ix_1u^{p(j_2+1)}vu^{p(j_1+1)-1}x_2a_{i+1} \cdots a_mw_m \in L_0a_1L_1 \cdots a_mL_m.$$

By the last two assertions in Lemma 37, we get $xu^{p\ell}vu^{p\ell}y \in L_0a_1L_1 \cdots a_mL_m$. Since we have $L_0a_1L_1 \cdots a_mL_m \subseteq L$ by definition, this implies that $xu^{p\ell}vu^{p\ell}y \in L$, finishing the proof.

4 Concatenation hierarchies: definition and generic results

We may now move to the main topic of this paper: concatenation hierarchies. As explained in the introduction, a natural complexity measure for star-free languages is the number of alternations between concatenation and complement that are required to build a given language from basic star-free languages. Generalizing this idea leads to the notion of concatenation hierarchy, which is meant to classify languages according to such a complexity measure. We first define concatenation hierarchies. Then, we prove that finitely based hierarchies are strict.

4.1 Definition

A particular hierarchy is built from a starting class of languages \mathcal{C} , which is called its *basis*. In order to get nice properties, we restrict \mathcal{C} to be a quotienting Boolean algebra of regular languages. This is the only parameter of the construction, meaning that once the basis has been chosen, the construction is entirely generic: each new level is built from the previous one by applying a closure operation: either *Boolean closure*, or *polynomial closure*. We speak of the “(concatenation) hierarchy of basis \mathcal{C} ”.

In the concatenation hierarchy of basis \mathcal{C} , languages are classified into levels of two distinct kinds: full levels (denoted by $0, 1, 2, 3, \dots$) and half levels (denoted by $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \dots$). The definition is as follows:

- Level 0 is the basis (*i.e.*, our parameter class \mathcal{C}). We denote it by $\mathcal{C}[0]$.
- *Half levels* are the *polynomial closure* of the previous full level: for any $n \in \mathbb{N}$, level $n + \frac{1}{2}$ is the polynomial closure of level n . We denote it by $\mathcal{C}[n + \frac{1}{2}]$.
- *Integer levels* are the *Boolean closure* of the previous half level: for any $n \in \mathbb{N}$, level $n + 1$ is the Boolean closure of level $n + \frac{1}{2}$. We denote it by $\mathcal{C}[n + 1]$.

We give a graphical representation of the construction process of a concatenation hierarchy in Fig. 3 below.

$$0 \xrightarrow{\text{Pol}} \frac{1}{2} \xrightarrow{\text{Bool}} 1 \xrightarrow{\text{Pol}} \frac{3}{2} \xrightarrow{\text{Bool}} 2 \xrightarrow{\text{Pol}} \frac{5}{2} \xrightarrow{\text{Bool}} 3 \xrightarrow{\text{Pol}} \frac{7}{2} \dots\dots$$

Fig. 3 A concatenation hierarchy

Observe that by definition, for any $n \in \mathbb{N}$, we have $\mathcal{C}[n] \subseteq \mathcal{C}[n + \frac{1}{2}] \subseteq \mathcal{C}[n + 1]$. However, these inclusions need not be strict. For instance, if the basis is closed under Boolean operations and marked concatenation (such as the class of star-free languages, or the whole class REG), the associated hierarchy collapses at level 0. Of course, the interesting hierarchies are the ones that *are* strict.

What we gain by imposing that the basis must be a quotienting Boolean algebra of regular languages are the following properties, obtained as an immediate corollary of Theorem 29 and Proposition 23.

Proposition 38 *Let \mathcal{C} be a quotienting Boolean algebra of regular languages and consider the concatenation hierarchy of basis \mathcal{C} . Then, every half level is a quotienting lattice of regular languages and every full level is a quotienting Boolean algebra of regular languages.*

In view of Lemma 27, the half levels are even more robust.

Proposition 39 *Let \mathcal{C} be a quotienting Boolean algebra of regular languages. Then, all half levels in the concatenation hierarchy of basis \mathcal{C} are closed under concatenation and marked concatenation.*

It is important to note that for a hierarchy to be strict, half levels must not be closed under complement and full levels must not be closed under marked concatenation. Indeed, a half level that is closed under complement would be equal to its Boolean

closure and the hierarchy would collapse at this level. Similarly, a full level that is closed under marked concatenation would be equal to its polynomial closure and the hierarchy would collapse as well.

Proposition 40 *Let \mathcal{C} be a quotienting Boolean algebra of regular languages. The following properties are equivalent:*

1. *The concatenation hierarchy of basis \mathcal{C} is strict.*
2. *No half level in the hierarchy of basis \mathcal{C} is closed under complement.*
3. *No full level in the hierarchy of basis \mathcal{C} is closed under marked concatenation.*

In practice, techniques for solving membership or separation problems rely heavily on the concatenation operation. Hence, we prefer to consider classes that are closed under concatenation, which excludes full levels: they cannot be closed under concatenation in a strict hierarchy. This is apparent in the proofs applying to concatenation hierarchies: most results—even those that apply to full levels—are based on the investigation of a half level. This is possible because of the following observation. When only considering the half levels, one may bypass the full levels in the definition by applying polynomial closure to the complements of half levels. This trick is based on Lemma 33. Let \mathcal{C} be a quotienting Boolean algebra of regular languages. Observe that by definition, for any $n \geq 1$, level $\mathcal{C}[n + \frac{1}{2}]$ is defined as:

$$\mathcal{C}[n + \frac{1}{2}] = \text{Pol}(\mathcal{C}[n]) = \text{Pol}(\text{BPol}(\mathcal{C}[n - 1])).$$

Applying Lemma 33, we obtain the following alternate definition of level $\mathcal{C}[n + \frac{1}{2}]$:

$$\mathcal{C}[n + \frac{1}{2}] = \text{Pol}(\text{BPol}(\mathcal{C}[n - 1])) = \text{Pol}(\text{co-Pol}(\mathcal{C}[n - 1])) = \text{Pol}(\text{co-}\mathcal{C}[n - \frac{1}{2}]).$$

The important point here is that the class $\text{co-}\mathcal{C}[n - \frac{1}{2}]$ is much simpler to manipulate than $\mathcal{C}[n]$. Indeed, the associated membership and separation problems are dual with the ones for the class $\mathcal{C}[n - \frac{1}{2}]$, which is closed under concatenation and marked concatenation. Altogether, we obtain the following proposition.

Proposition 41 *Let \mathcal{C} be a quotienting Boolean algebra of regular languages and consider the associated concatenation hierarchy. Then, for any natural number $n \geq 1$,*

$$\mathcal{C}[n + \frac{1}{2}] = \text{Pol}(\text{co-}\mathcal{C}[n - \frac{1}{2}]).$$

In view of Proposition 41, we update the construction process of a concatenation hierarchy in Fig. 4 (the new arrows are horizontal).

4.2 Strictness of finitely based hierarchies

As explained in the introduction, concatenation hierarchies first appeared in the literature with two specific hierarchies: the dot-depth was introduced in 1971 [8] and the Straubing-Thérien hierarchy ten years later [48, 50]. Although both of them were investigated intensively, their understanding is still far from being satisfactory. For instance, membership algorithms are known only for the lower levels in both hierarchies. A common feature to these two hierarchies is that their bases are *finite*.

Here, we look at *finitely based* hierarchies. We prove that *any* such hierarchy is strict for alphabets of size 2 or more, meaning that any half or full level contains strictly more languages than the preceding ones.

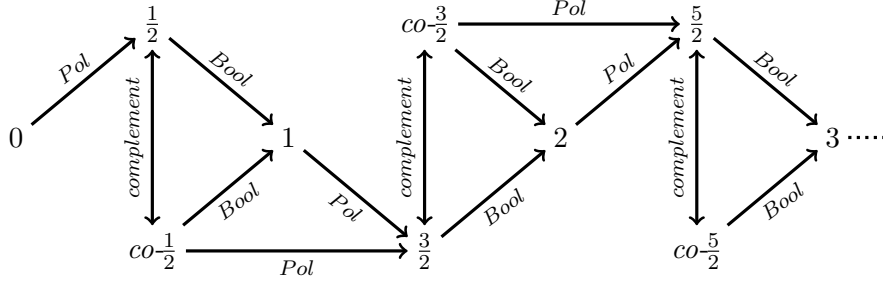


Fig. 4 A concatenation hierarchy with complement levels

As we just explained, we will first prove that any finitely based hierarchy is *strict*. To prove strictness, one must show that no two consecutive levels are identical. The condition that the basis is finite may seem to be restrictive, but it is already challenging and it covers the two classical cases (namely, the dot-depth and the Straubing-Thérien hierarchies).

Theorem 42 *Let \mathcal{C} be a finite quotienting Boolean algebra of languages. Then, the concatenation hierarchy of basis \mathcal{C} is strict for any alphabet of size at least two.*

The remainder of the section is devoted to proving Theorem 42. We assume that the working alphabet A contains at least two distinct letters a and b . Our objective is to prove that for any finite quotienting Boolean algebra of (regular) languages \mathcal{C} , the associated concatenation hierarchy is strict, that is, for any $n \in \mathbb{N}$:

$$\mathcal{C}[n] \subsetneq \mathcal{C}[n + \frac{1}{2}] \subsetneq \mathcal{C}[n + 1].$$

We prove this as a corollary of a more general result. Let us first introduce some terminology. We call *unambiguous family* an infinite language $U \subseteq A^+$ satisfying the two following conditions:

1. For any $u \in U^+$, the decomposition $u = u_1 \cdots u_n$ with $u_1, \dots, u_n \in U$ witnessing membership in U^+ is *unique*. In the literature, such a set U is usually called a *variable-length code*.
2. Moreover, if $v \in U$ is an infix of u , then $v = u_i$ for some $i \leq n$. In the terminology of [4], U is a comma-free code.

Example 43 A typical example of unambiguous family is $U = ab^+a$. In fact, this is exactly the family that we use below to prove Theorem 42. \square

Consider an arbitrary (possibly infinite) quotienting Boolean algebra of regular languages \mathcal{C} and an unambiguous family U . We say that \mathcal{C} is *non-separating for U* when there exist a language L and $V \subseteq U$ satisfying the four following conditions:

$$\begin{cases} U \setminus V \text{ is infinite.} \\ A^*LA^* = L. \\ L \in \text{Pol}(\mathcal{C}). \\ (A^* \setminus L) \cap V^+ \text{ is \textbf{not} } \text{Pol}(\mathcal{C})\text{-separable from } L \cap V^+. \end{cases} \quad (3)$$

We may now state our general result. Any concatenation hierarchy (even with an infinite basis) which is non-separating for some unambiguous family U , is strict.

Proposition 44 *Let \mathcal{C} be a quotienting Boolean algebra of regular languages. Assume that there exists an unambiguous family $U \subseteq A^*$ such that \mathcal{C} is non-separating for U . Then, the concatenation hierarchy of basis \mathcal{C} over A is strict.*

We divide the proof in two parts: first, we explain how Proposition 44 can be used to prove Theorem 42. Then, we present the proof of Proposition 44 itself.

Proof (of Theorem 42, assuming Proposition 44) Our objective is to show that the concatenation hierarchy of basis \mathcal{C} is strict for A . We first reduce the problem to the case where $\{\varepsilon\} \in \mathcal{C}$.

Lemma 45 *Let \mathcal{C} be a finite quotienting Boolean algebra. Then, there exists a finite quotienting Boolean algebra \mathcal{D} such that $\{\varepsilon\} \in \mathcal{D}$ and the concatenation hierarchy of basis \mathcal{C} is strict if and only if that of basis \mathcal{D} is strict.*

Proof We define \mathcal{D} as the least Boolean algebra containing \mathcal{C} and such that $\{\varepsilon\} \in \mathcal{D}$. By definition, \mathcal{D} is also finite and it is a Boolean algebra. Moreover, since quotients commute with Boolean operations, since \mathcal{C} is closed under quotient and since the only quotients of $\{\varepsilon\}$ are $\{\varepsilon\}$ and \emptyset , \mathcal{D} is closed under quotient as well.

We prove that for any $n \in \mathbb{N}$, $\mathcal{C}[n] \subseteq \mathcal{D}[n] \subseteq \mathcal{C}[n+1]$. This implies immediately that the concatenation hierarchy of basis \mathcal{C} is strict if and only if that based on \mathcal{D} is. By definition of concatenation hierarchies, it suffices to verify that these inclusions hold for $n = 0$, i.e., $\mathcal{C} \subseteq \mathcal{D} \subseteq \mathcal{C}[1]$. Clearly, we have $\mathcal{C} \subseteq \mathcal{D}$. For the other inclusion, we have $\mathcal{C} \subseteq \mathcal{C}[1]$ and $\mathcal{C}[1]$ is a Boolean algebra. Hence, by definition of \mathcal{D} , it suffices to prove that $\{\varepsilon\} \in \mathcal{C}[1]$ to conclude that $\mathcal{D} \subseteq \mathcal{C}[1]$. This is immediate, since $A^+ = \bigcup_{a \in A} A^* a A^* \in \mathcal{C}[\frac{1}{2}]$. Therefore, $\{\varepsilon\} = A^* \setminus A^+ \in \mathcal{C}[1]$, finishing the proof. \square

In view of Lemma 45, we now assume that $\{\varepsilon\} \in \mathcal{C}$. Let $U = ab^+a$. Clearly, U is unambiguous. If we prove that \mathcal{C} is non-separating for U , it will follow from Proposition 44 that the concatenation hierarchy of basis \mathcal{C} is strict. Our objective is therefore to exhibit $L \subseteq \{a, b\}^*$ and $V \subseteq U$ satisfying (3).

Since \mathcal{C} is a **finite** Boolean algebra, there exist two natural numbers $1 \leq k \leq \ell$ such that $b^k \leq_{\mathcal{C}} b^\ell$. We let $u = ab^k a$ and $v = ab^\ell a$. Clearly, $u, v \in U$. Moreover, since \mathcal{C} is a finite quotienting Boolean algebra, we get from Lemma 22 that $u \leq_{\mathcal{C}} v$. We define L as follows:

$$L = A^* v A^* = A^* ab^\ell a A^*.$$

Note that since $A^* \in \mathcal{C}$ (as \mathcal{C} is a Boolean algebra) and $\{\varepsilon\} \in \mathcal{C}$ by hypothesis, it is immediate from Lemma 28 that $L \in \text{Pol}(\mathcal{C})$. Moreover, $A^* L A^* = L$ by definition. Finally, we define $V = \{u, v\} \subseteq U$. Since U is infinite, so is $U \setminus V$. We show that $(A^* \setminus L) \cap V^+$ is not $\text{Pol}(\mathcal{C})$ -separable from $L \cap V^+$.

We use Corollary 35. Recall that $u \leq_{\mathcal{C}} v$. Moreover, since $L = A^* ab^\ell a A^*$, $u = ab^k a$ and $v = ab^\ell a$ with $k \neq \ell$, it is immediate that for every $n \geq 1$, we have $u^n \in (A^* \setminus L) \cap V^+$ and $u^n v u^n \in L \cap V^+$. Thus, Corollary 35 (in the case when $x = y = \varepsilon$) yields that $(A^* \setminus L) \cap V^+$ is not $\text{Pol}(\mathcal{C})$ -separable from $L \cap V^+$. \square

We now prove Proposition 44, as a consequence of two lemmas that we present now.

Lemma 46 *Let \mathcal{C} be a quotienting Boolean algebra of regular languages. Assume that there exists an unambiguous family $U \subseteq A^*$ such that \mathcal{C} is non-separating for U . Then, $\text{Pol}(\mathcal{C})$ is not closed under complement.*

Lemma 47 *Let \mathcal{C} be a quotienting Boolean algebra of regular languages. Assume that there exists an unambiguous family $U \subseteq A^*$ such that \mathcal{C} is non-separating for U . Then, $BPol(\mathcal{C})$ is non-separating for U .*

Proof (of Proposition 44 assuming Lemmas 46 and 47) Combining Lemmas 46 and 47 yields that for any quotienting Boolean algebra of regular languages \mathcal{C} which is non-separating for some unambiguous family, every half-level in the associated concatenation hierarchy is not closed under complement. Proposition 40 entails that the concatenation hierarchy of basis \mathcal{C} is strict over A . Thus, Proposition 44 is proved. \square

To conclude the proof of Theorem 42, it remains to prove Lemmas 46 and 47.

Proof (of Lemma 46) It follows from our hypothesis that there exist $L \in Pol(\mathcal{C})$ and $V \subseteq U$ such that $(A^* \setminus L) \cap V^+$ is **not** $Pol(\mathcal{C})$ -separable from $L \cap V^+$. Observe that $A^* \setminus L$ clearly separates from $(A^* \setminus L) \cap V^+$ from $L \cap V^+$. Hence, $A^* \setminus L \notin Pol(\mathcal{C})$ by hypothesis. Since $L \in Pol(\mathcal{C})$ by (3), we conclude that $Pol(\mathcal{C})$ is not closed under complement, which terminates the proof of Lemma 46. \square

We turn to Lemma 47 whose proof is more involved.

Proof (of Lemma 47) By hypothesis, \mathcal{C} is non-separating for U , *i.e.*, we have L and $V \subseteq U$ satisfying (3). We need to prove that $BPol(\mathcal{C})$ is non-separating for U as well. By definition, this amounts to finding $K \in Pol(BPol(\mathcal{C}))$ and $W \subseteq U$ satisfying the appropriate properties. We first build K and W . Let w be a word in the (infinite) set $U \setminus V$ and let,

$$\begin{cases} K = A^*w(A^+ \setminus L)wA^*, \\ W = V \cup \{w\}. \end{cases}$$

It is immediate that $U \setminus W$ is infinite and that $A^*KA^* = K$. Let us verify that $K \in Pol(BPol(\mathcal{C}))$. First, $A^+ = \bigcup_{a \in A} A^*aA^* \in Pol(\mathcal{C})$, and since $L \in Pol(\mathcal{C})$ as well, we have $A^+ \setminus L \in BPol(\mathcal{C})$. Moreover, observe that $\{\varepsilon\} = A^* \setminus A^+ \in BPol(\mathcal{C})$. Hence, Lemma 28 shows that K belongs to $Pol(BPol(\mathcal{C}))$.

What remains to show is that $(A^* \setminus K) \cap W^+$ is **not** $Pol(BPol(\mathcal{C}))$ -separable from $K \cap W^+$ (see (3)). We proceed by contradiction. Assume that $(A^* \setminus K) \cap W^+$ is $Pol(BPol(\mathcal{C}))$ -separable from $K \cap W^+$. We have the following fact.

Fact 48 *There exists some finite quotienting lattice \mathcal{D} such that $\mathcal{D} \subseteq Pol(\mathcal{C})$ and $(A^* \setminus K) \cap W^+$ is $Pol(co\mathcal{D})$ -separable from $K \cap W^+$.*

Proof By hypothesis, we have $H \in Pol(BPol(\mathcal{C}))$ separating $(A^* \setminus K) \cap W^+$ from $K \cap W^+$. By Lemma 33, $Pol(BPol(\mathcal{C})) = Pol(co\text{-}Pol(\mathcal{C}))$. Hence, $H \in Pol(co\text{-}Pol(\mathcal{C}))$ and there exists some finite class $\mathcal{C}' \subseteq Pol(\mathcal{C})$ such that $H \in Pol(co\mathcal{C}')$. We may then use Lemma 17 to get a finite quotienting lattice \mathcal{D} such that $\mathcal{C}' \subseteq \mathcal{D} \subseteq Pol(\mathcal{C})$. Clearly, $H \in Pol(co\mathcal{D})$ which concludes the proof. \square

Since \mathcal{D} is a finite quotienting lattice, it is immediate that $co\mathcal{D}$ is one as well. Recall that by hypothesis, $(A^* \setminus L) \cap V^+$ is not $Pol(\mathcal{C})$ -separable from $L \cap V^+$. Hence, since $\mathcal{D} \subseteq Pol(\mathcal{C})$, $(A^* \setminus L) \cap V^+$ is not \mathcal{D} -separable from $L \cap V^+$. Moreover, by definition of $co\mathcal{D}$, this also implies that $L \cap V^+$ is not $co\mathcal{D}$ -separable from $(A^* \setminus L) \cap V^+$. Therefore, Corollary 20 yields $u \in L \cap V^+$ and $v \in (A^* \setminus L) \cap V^+$ such that $u \leq_{co\mathcal{D}} v$. We have the following fact.

Fact 49 *Let $n \geq 1$. Then, $(www)^n \in (A^* \setminus K) \cap W^+$ and $(www)^n wvw(www)^n \in K \cap W^+$.*

Let us first use Fact 49 to finish the main argument. Since $co\mathcal{D}$ is a finite quotienting lattice, Lemma 22 states that $\leq_{co\mathcal{D}}$ is compatible with word concatenation. Together with $u \leq_{co\mathcal{D}} v$, this implies that $www \leq_{co\mathcal{D}} wvw$. Consequently, Corollary 35 (in the case $x = y = \varepsilon$) and Fact 49 yield that $(A^* \setminus K) \cap W^+$ is **not** $Pol(co\mathcal{D})$ -separable from $K \cap W^+$. This contradicts Fact 48, which completes the main proof of Lemma 47.

It remains to prove Fact 49. Let us start with $(www)^n \in (A^* \setminus K) \cap W^+$. By definition, $w \in W$ and $u \in L \cap V^+$. Since $V^+ \subseteq W^+$ by definition, it follows that $(www)^n \in W^+$. It remains to show that $(www)^n \in A^* \setminus K$, i.e., $(www)^n \notin K$. Since $K = A^*w(A^+ \setminus L)wA^*$, this amounts to proving that for any infix of the form wzw in $(www)^n$ with $z \neq \varepsilon$, we have $z \in L$. Consider such an infix. By definition, $w \in U$ and $u \in L \cap V^+ \subseteq U^+$. Therefore, since U is unambiguous, z must contain $u \in L$ as an infix. Since $L = A^*LA^*$ by hypothesis, it follows that $z \in L$. We get $(www)^n \notin K$.

It remains to show that $(www)^n wvw(www)^n \in K \cap W^+$. Since $w \in W$ and $u, v \in V^+ \subseteq W^+$, it is immediate that $(www)^n wvw(www)^n \in W^+$. Furthermore, we have $v \in (A^* \setminus L) \cap V^+ \subseteq A^+ \setminus L$ by definition. This yields that $(www)^n wvw(www)^n \in K = A^*w(A^+ \setminus L)wA^*$. We conclude that $(www)^n wvw(www)^n \in K \cap W^+$, as desired. \square

5 Membership and separation

Now that we know that finitely based concatenation hierarchies are strict, we focus on membership and separation for each individual level in such a hierarchy. We present an exhaustive and *generic* state of the art regarding these problems. However, presenting the algorithms themselves would require introducing too much material. For this reason, we only detail one of these results. For the others, we simply state the problems which are known to be decidable, without describing the algorithms.

5.1 State of the art

For levels in finitely based concatenation hierarchies, membership and separation are both unexpectedly hard, despite their straightforward formulations. The overall state of the art can be formulated with only three theorems. We state two of them in this section and the last one in Section 6.

The first result (which we shall detail and prove) connects $Pol(\mathcal{C})$ -membership to \mathcal{C} -separation: the former reduces to the latter.

Theorem 50 (Place & Zeitoun [34]) *For every quotienting lattice of regular languages \mathcal{C} , the membership problem for $Pol(\mathcal{C})$ reduces to the separation problem for \mathcal{C} .*

Remark 51 This statement is slightly more general than the original one of [34] which was specific to classes \mathcal{C} within the Straubing-Thérien hierarchy. However, it turns out that the proof arguments of [34] suffice to obtain Theorem 50.

Among the known membership and separation results for concatenation hierarchies, Theorem 50 is the simplest one. However, it is central as it motivates the investigation of separation for the levels of concatenation hierarchies, even when one is only interested in

membership. Of course, Theorem 50 only states a reduction. Getting actual algorithms requires the following theorem, which summarizes the known separation results for finitely based hierarchies: the problem is decidable up to level $\frac{3}{2}$.

Theorem 52 (Place & Zeitoun [40, 32, 33]) *If \mathcal{C} is a finite quotienting Boolean algebra, separation is decidable for the following classes:*

1. $Pol(\mathcal{C})$,
2. $BPol(\mathcal{C})$,
3. $Pol(BPol(\mathcal{C}))$.

Hence, in any finitely based hierarchy, levels $\frac{1}{2}$, 1 and $\frac{3}{2}$ have decidable separation.

Let us point out that Theorem 50 summarizes several distinct results which are all difficult. When combined with Theorem 50, it yields the following corollary for the membership problem.

Corollary 53 *Let \mathcal{C} be a finite quotienting Boolean algebra. Then, the levels $\frac{1}{2}$, 1, $\frac{3}{2}$ and $\frac{5}{2}$ in the concatenation hierarchy of basis \mathcal{C} have decidable membership.*

Proof The result for levels $\frac{1}{2}$, 1 and $\frac{3}{2}$ is immediate from Theorem 52 since membership reduces to separation. For level $\frac{5}{2}$, we know from Lemma 33 that $\mathcal{C}[\frac{5}{2}] = Pol(co\mathcal{C}[\frac{3}{2}])$. By Proposition 23 and Theorem 29, $co\mathcal{C}[\frac{3}{2}]$ is a quotienting lattice which has decidable separation by Theorem 52 (the problem is symmetric to $\mathcal{C}[\frac{3}{2}]$ -separation). Hence, Theorem 50 yields that $\mathcal{C}[\frac{5}{2}]$ has decidable membership. \square

Theorem 52 and Corollary 53 apply in particular to the Straubing-Thérien and dot-depth hierarchies, since their bases are finite. Therefore, several classical results that we presented in Section 1, namely Theorems 5, 6, 8, 9 and 10, are immediate corollaries of Theorem 52. Note however that we do not recover Theorem 11 yet. Nevertheless, we will see in the next section that this result is also a corollary of Theorem 52.

5.2 A generic characterization of $Pol(\mathcal{C})$

We now focus on proving Theorem 50. It is based on a generic algebraic characterization of $Pol(\mathcal{C})$. To present it, we need the algebraic definition of regular languages with finite monoids. We briefly recall this standard notion (see [26] for more details).

Recognition by monoids. A *monoid* is a set M endowed with an associative multiplication $(s, t) \mapsto s \cdot t$ (we often write st for $s \cdot t$) having a neutral element 1_M , *i.e.*, such that $1_M \cdot s = s \cdot 1_M = s$ for every $s \in M$. An *idempotent* of a monoid M is an element $e \in M$ such that $ee = e$. It is folklore that for every *finite* monoid M , there exists a natural number $\omega(M)$ (denoted by ω when M is understood) such that for any $s \in M$, the element s^ω is an idempotent.

Observe that A^* is a monoid whose multiplication is concatenation (the neutral element is ε). Thus, given an arbitrary finite monoid M , we may consider monoid morphisms $\alpha : A^* \rightarrow M$. Given such a morphism and some language $L \subseteq A^*$, we say that L is *recognized* by α when there exists an *accepting set* $F \subseteq M$ such that $L = \alpha^{-1}(F)$. It is well-known that a language is regular if and only if it can be recognized by a morphism into a *finite* monoid. Furthermore, given any language L , one may define a canonical monoid morphism $\alpha_L : A^* \rightarrow M_L$ (computable when L is

regular) recognizing L . We briefly recall its definition. One may associate an equivalence \equiv_L over A^* to the language L : the *syntactic congruence of L* . Given $u, v \in A^*$, $u \equiv_L v$ if and only if $xuy \in L \Leftrightarrow xvy \in L$ for all $x, y \in A^*$. It is known and simple to verify that “ \equiv_L ” is a congruence on A^* . Thus, the set of equivalence classes $M_L = A^*/\equiv_L$ is a monoid (called the *syntactic monoid of L*) and the map $\alpha_L : A^* \rightarrow M_L$, which maps any word to its equivalence class is a morphism (called syntactic morphism of L). It is well-known that L is regular if and only if M_L is finite (*i.e.*, \equiv_L has finite index). In that case, one may compute the syntactic morphism $\alpha_L : A^* \rightarrow M_L$ from any representation of L (such as a finite automaton).

Finally, since $Pol(\mathcal{C})$ is not closed under complement in general, characterizing the languages in $Pol(\mathcal{C})$ requires slightly more information than just their syntactic morphism. We need a canonical partial order defined on their syntactic monoid. Consider a language L . Let $\alpha_L : A^* \rightarrow M_L$ its syntactic morphism and $F \subseteq M_L$ the corresponding accepting set ($L = \alpha_L^{-1}(F)$). Given $s, t \in M_L$, we write $s \leq_L t$ if and only if $qsr \in F \Rightarrow qtr \in F$ for all $q, r \in M_L$. It is standard and simple to verify that \leq_L satisfies the two following properties:

- \leq_L is a partial order on M_L compatible with multiplication: if $s_1, s_2, t_1, t_2 \in M_L$, $s_1 \leq_L t_1$ and $s_2 \leq_L t_2$, then $s_1 s_2 \leq_L t_1 t_2$.
- The accepting set F is an upper set for \leq_L : for every $s \in F$ and $t \in M_L$, if $s \leq_L t$, then $t \in F$.

Characterization. We may now present the generic characterization of $Pol(\mathcal{C})$. The statement is parametrized by a relation depending on \mathcal{C} . Consider a class \mathcal{C} , a finite monoid M and a morphism $\alpha : A^* \rightarrow M$. For any pair $(s, t) \in M^2$, we say that,

$$(s, t) \text{ is a } \mathcal{C}\text{-pair for } \alpha \text{ when } \alpha^{-1}(s) \text{ is \textbf{not} } \mathcal{C}\text{-separable from } \alpha^{-1}(t). \quad (4)$$

For the sake of improved readability, we shall often abuse terminology and simply say that (s, t) is a \mathcal{C} -pair, omitting the morphism α . By definition, the set of \mathcal{C} -pairs for α is finite: it is a subset of $M \times M$. Moreover, having a \mathcal{C} -separation algorithm in hand is clearly enough to compute all \mathcal{C} -pairs for any input morphism α .

We now characterize of $Pol(\mathcal{C})$ when \mathcal{C} is an arbitrary quotienting lattice by a property of the syntactic morphism of the languages in $Pol(\mathcal{C})$ and the canonical order defined on the syntactic monoid.

Theorem 54 *Let \mathcal{C} be a quotienting lattice of regular languages and let L be a regular language. The two following properties are equivalent:*

1. $L \in Pol(\mathcal{C})$.
2. The syntactic morphism $\alpha_L : A^* \rightarrow M_L$ of L satisfies the following property:

$$s^{\omega+1} \leq_L s^\omega t s^\omega \quad \text{for all } \mathcal{C}\text{-pairs } (s, t) \in M_L^2. \quad (5)$$

Clearly, Theorem 54 states a reduction from $Pol(\mathcal{C})$ -membership to \mathcal{C} -separation. Indeed, the syntactic morphism of a regular language can be computed and Equation (5) can be decided as soon as one is able to compute all \mathcal{C} -pairs (this is possible with an algorithm for \mathcal{C} -separation). Hence, Theorem 50 is an immediate corollary.

This terminates the presentation of the algebraic characterization of $Pol(\mathcal{C})$. We now turn to its proof. Let \mathcal{C} be a quotienting lattice of regular languages, and let L be some regular language L . We start with 1) \Rightarrow 2): when $L \in Pol(\mathcal{C})$, α_L satisfies Equation (5).

Direction 1) \Rightarrow 2)

Assume that $L \in \text{Pol}(\mathcal{C})$. We have to show that the syntactic morphism α_L satisfies Equation (5): for every \mathcal{C} -pair $(s, t) \in M_L^2$, $s^{\omega+1} \leq_L s^\omega t s^\omega$.

Fact 55 *There exists a finite quotienting lattice $\mathcal{D} \subseteq \mathcal{C}$ such that $L \in \text{Pol}(\mathcal{D})$.*

Proof Since $L \in \text{Pol}(\mathcal{C})$, we have $L \in \text{Pol}(\mathcal{C}')$ for some finite class $\mathcal{C}' \subseteq \mathcal{C}$. It then follows from Lemma 17 that there exists a finite quotienting lattice \mathcal{D} such that $\mathcal{C}' \subseteq \mathcal{D} \subseteq \mathcal{C}$. Clearly, $L \in \text{Pol}(\mathcal{D})$. \square

Since (s, t) is a \mathcal{C} -pair, we know that $\alpha^{-1}(s)$ is not \mathcal{C} -separable from $\alpha^{-1}(t)$. Therefore, since $\mathcal{D} \subseteq \mathcal{C}$, it follows that $\alpha^{-1}(s)$ is not \mathcal{D} -separable from $\alpha^{-1}(t)$. By Corollary 20 this yields $u \in \alpha^{-1}(s)$ and $v \in \alpha^{-1}(t)$ such that $u \leq_{\mathcal{D}} v$. Since $L \in \text{Pol}(\mathcal{D})$, we may apply Proposition 34 which yields $h, p \geq 1$ such that for every $x, y \in A^*$,

$$xu^{ph\omega+1}y \in L \quad \Rightarrow \quad xu^{ph\omega}vu^{ph\omega}y \in L$$

By definition of the syntactic order \leq_L on M_L , it then follows that,

$$s^{\omega+1} = \alpha(u^{ph\omega+1}) \leq_L \alpha(u^{ph\omega}vu^{ph\omega}) = s^\omega t s^\omega$$

This concludes the proof for this direction.

Direction 2) \Rightarrow 1)

It now remains to prove the harder “2) \Rightarrow 1)” direction of Theorem 54. Assuming that the syntactic morphism of L satisfies (5), we show that $L \in \text{Pol}(\mathcal{C})$.

The proof is based on the factorization forest theorem of Simon which we briefly recall now. We refer the reader to [20, 10] for more details and a proof. Consider a finite monoid M and a morphism $\alpha : A^* \rightarrow M$. An α -factorization forest is an ordered unranked tree whose nodes are labeled by words in A^* . For any inner node x with label $w \in A^*$, if $w_1, \dots, w_n \in A^*$ are the labels of its children listed from left to right, then $w = w_1 \cdots w_n$. Moreover, all nodes x in the forest must be of the three following kinds:

- *Leaves*, which are labeled by either a single letter or the empty word.
- *Binary inner nodes*, which have exactly two children.
- *Idempotent inner nodes*, which may have an arbitrary number of children. However, the labels w_1, \dots, w_n of these children must satisfy $\alpha(w_1) = \cdots = \alpha(w_n) = e$ where e is an idempotent element of M .

Given a word $w \in A^*$, an α -factorization forest for w is an α -factorization forest whose root is labeled by w . The *height* of a factorization forest is the largest $h \in \mathbb{N}$ such that it contains a branch with h inner nodes (a single leaf has height 0). We turn to the factorization forest theorem of Simon: there exists a bound depending only on M such that every word admits an α -factorization forest of height at most this bound.

Theorem 56 ([46, 20]) *Consider a morphism $\alpha : A^* \rightarrow M$. For all words $w \in A^*$, there exists an α -factorization forest for w of height at most $3|M| - 1$.*

We may now start the proof of the direction “2) \Rightarrow 1)” in Theorem 54. We assume that the syntactic morphism $\alpha_L : A^* \rightarrow M_L$ satisfies (5). We use the following proposition to show that $L \in \text{Pol}(\mathcal{C})$.

Proposition 57 *For all $h \in \mathbb{N}$ and all $s \in M_L$, there exists $H_{s,h} \in \text{Pol}(\mathcal{C})$ such that for all $w \in A^*$:*

- *If $w \in H_{s,h}$ then $s \leq_L \alpha_L(w)$.*
- *If $\alpha_L(w) = s$ and w admits an α_L -factorization forest of height at most h , then $w \in H_{s,h}$.*

Assume for now that Proposition 57 holds. Given $h = 3|M_L| - 1$, for all $s \in M_L$, consider the language $H_{s,h} \in \text{Pol}(\mathcal{C})$ associated to s and h by Proposition 57. We know from Theorem 56 that all words in A^* admit an α_L -factorization forest of height at most $3|M_L| - 1$. Therefore, for all $w \in A^*$ we have,

1. If $w \in H_{s,h}$ then $s \leq_L \alpha_L(w)$.
2. If $\alpha_L(w) = s$, then $w \in H_{s,h}$.

Recall that $L = \alpha_L^{-1}(F)$ for an accepting set $F \subseteq M_L$ which is an upper set for \leq_L . We show that $L = \bigcup_{s \in F} H_{s,h}$ which concludes the proof: $L \in \text{Pol}(\mathcal{C})$ since it is a union of languages $H_{s,h} \in \text{Pol}(\mathcal{C})$. By Item 2 above, we have $L \subseteq \bigcup_{s \in F} H_{s,h}$. Moreover, since F is an upper set, we know that for if $s \in F$ and $s \leq_L t$, then $t \in F$. Hence, Item 1 above implies that $\bigcup_{s \in F} H_{s,h} \subseteq L$. This concludes the proof of Theorem 54.

It now remains to prove Proposition 57. We begin with a lemma which defines the basic languages in \mathcal{C} that we will use in the construction of our languages in $\text{Pol}(\mathcal{C})$. Note that this is also where we use the hypothesis that (5) holds.

Lemma 58 *For every idempotent $e \in M_L$, there exists a language K_e belonging to \mathcal{C} (and therefore to $\text{Pol}(\mathcal{C})$) which satisfies the two following properties,*

1. *For all $u \in K_e$, we have $e \leq_L e \cdot \alpha_L(u) \cdot e$.*
2. *$\alpha^{-1}(e) \subseteq K_e$.*

Proof Let $T \subseteq M_L$ be the set of all elements $t \in M_L$ such that (e, t) is **not** a \mathcal{C} -pair (i.e., $\alpha^{-1}(e)$ is \mathcal{C} -separable from $\alpha^{-1}(t)$). By definition, for all $t \in T$, there exists a language $G_t \in \mathcal{C}$ which separates $\alpha^{-1}(e)$ from $\alpha^{-1}(t)$. We let $K_e = \bigcap_{t \in T} G_t$. Clearly, $K_e \in \mathcal{C}$ since \mathcal{C} is a lattice. Moreover, $\alpha^{-1}(e) \subseteq K_e$ since the inclusion holds for all languages G_t . Finally, given $u \in K_e$, it is immediate from the definition that $\alpha_L(u)$ does not belong to T which means that $(e, \alpha_L(u))$ is a \mathcal{C} -pair. The first item is now immediate from (5) since e is idempotent. \square

We may now start the proof of Proposition 57. Let $h \in \mathbb{N}$ and $s \in M_L$. We construct $H_{s,h} \in \text{Pol}(\mathcal{C})$ by induction on h . Assume first that $h = 0$. Note that the nonempty words having an α -factorization forest of height at most 0 are all single letters. We let $B = \{b \in A \mid \alpha_L(b) = s\}$. Moreover, we use the language $K_{1_{M_L}}$ as defined in Lemma 58 for the neutral element 1_{M_L} (which is an idempotent). There are two cases depending on whether $s = 1_{M_L}$ or not. If $s \neq 1_{M_L}$, we let,

$$H_{s,0} = \bigcup_{b \in B} K_{1_{M_L}} b K_{1_{M_L}}.$$

Otherwise, when $s = 1_{M_L}$, we let,

$$H_{s,0} = K_{1_{M_L}} \cup \bigcup_{b \in B} K_{1_{M_L}} b K_{1_{M_L}}.$$

Note that $H_{s,0} \in \text{Pol}(\mathcal{C})$ by Lemma 27 as we only applied marked concatenation and unions to $K_{1_{M_L}} \in \mathcal{C} \subseteq \text{Pol}(\mathcal{C})$. We now prove that this definition satisfies the two conditions in Proposition 57. We do the proof for the case when $s \neq 1_{M_L}$ (the other case is similar).

Assume first that $w \in H_{s,0}$, we have to prove that $s \leq_L \alpha_L(w)$. By definition $w = ubu'$ with $u, u' \in K_{1_{M_L}}$ and $b \in B$. Hence, $\alpha_L(w) = \alpha_L(u)s\alpha_L(u')$. Since $u, u' \in K_{1_{M_L}}$, we obtain from the first assertion in Lemma 58 that $1_{M_L} \leq_L \alpha_L(u)$ and $1_{M_L} \leq_L \alpha_L(u')$. It follows that $s \leq_L \alpha_L(u)s\alpha_L(u') = \alpha_L(w)$.

We turn to the second item. Let $w \in A^*$ such that $\alpha_L(w) = s$ and w admits an α_L -factorization forest of height at most 0. Since we assumed that $s \neq 1_{M_L}$, w cannot be empty. We have to prove that $w \in H_{s,0}$. By hypothesis, w is a one letter word $b \in B$. Hence, $w \in K_{1_{M_L}}bK_{1_{M_L}}$ since $\varepsilon \in K_{1_{M_L}}$ by the second assertion in Lemma 58.

Assume now that $h > 0$. There are two cases depending on whether s is idempotent or not. We treat the idempotent case (the other case is left to the reader: this is essentially a simpler version of the same proof). Hence, we assume that s is an idempotent, which we denote by e . We begin by constructing $H_{e,h}$ and then prove that it satisfies the conditions in the proposition. For all $t \in M_L$, one may use induction to construct $H_{t,h-1} \in \text{Pol}(\mathcal{C})$ such that for all $w \in A^*$:

- If $w \in H_{t,h-1}$ then $t \leq_L \alpha_L(w)$.
- If $\alpha_L(w) = t$ and w admits an α_L -factorization forest of height at most $h-1$, then $w \in H_{t,h-1}$.

We define $H_{e,h}$ as the union of three languages. Intuitively, the first one contains the words which have an α_L -factorization forest of height at most $h-1$, the second one, words having an α_L -factorization forest of height h and whose root is a binary node, and the third one, words with an α_L -factorization forest of height h and whose root is an idempotent node. Let K_e be as defined in Lemma 58.

$$H_{e,h} = H_{e,h-1} \cup \bigcup_{t_1 t_2 = e} (H_{t_1,h-1} H_{t_2,h-1}) \cup H_{e,h-1} K_e H_{e,h-1}$$

By definition, $H_{e,h}$ is a union of concatenations of languages in $\text{Pol}(\mathcal{C})$ and therefore belongs to $\text{Pol}(\mathcal{C})$ itself by Lemma 27. We need to prove that it satisfies the conditions of the proposition. Choose some $w \in A^*$ and assume first that $w \in H_{e,h}$. We need to prove that $e \leq_L \alpha_L(w)$.

- If $w \in H_{e,h-1}$, then this is by definition of $H_{e,h-1}$.
- If $w \in H_{t_1,h-1} H_{t_2,h-1}$ for $t_1, t_2 \in M_L$ such that $t_1 t_2 = e$, then by definition, $w = w_1 w_2$ with $t_1 \leq_L \alpha_L(w_1)$ and $t_2 \leq_L \alpha_L(w_2)$. It follows that $e = t_1 t_2 \leq_L \alpha_L(w_1 w_2) = \alpha_L(w)$.
- Finally, if $w \in H_{e,h-1} K_e H_{e,h-1}$, we obtain that $w = w_1 u w_2$ with $e \leq_L \alpha_L(w_1)$, $u \in K_e$ and $e \leq_L \alpha_L(w_2)$. The first assertion in Lemma 58 yields $e \leq_L e \alpha_L(u) e$. Hence, since $e \alpha_L(u) e \leq_L \alpha_L(w_1) \alpha_L(u) \alpha_L(w_2) = \alpha_L(w)$, we get $e \leq_L \alpha_L(w)$.

For the second condition, assume that $\alpha_L(w) = e$ and w admits an α_L -factorization forest of height at most h . We prove that $w \in H_{e,h}$. There are again three cases.

- First, if w admits an α_L -factorization forest of height at most $h-1$, then $w \in H_{e,h-1}$ by definition.
- Second, if w admits an α_L -factorization forest of height h whose root is a binary node, then $w = w_1 w_2$ with w_1, w_2 admitting forests of height at most $h-1$. Let

$t_1 = \alpha_L(w_1)$ and $t_2 = \alpha_L(w_2)$. Observe that $t_1 t_2 = \alpha_L(w) = e$. By definition, we have $w_1 \in H_{t_1, h-1}$ and $w_2 \in H_{t_2, h-1}$. Hence, $w \in H_{t_1, h-1} H_{t_2, h-1} \subseteq H_{e, h}$, as desired.

- Finally, if w admits an α_L -factorization forest of height h whose root is an idempotent node, then $w = w_1 u w_2$ with $\alpha_L(w_1) = \alpha_L(u) = \alpha_L(w_2) = e$ and w_1, w_2 admit forests of height at most $h - 1$. It follows that $w_1, w_2 \in H_{e, h-1}$ and since $\alpha_L(u) = e$, it is immediate that $u \in K_e$ by second item in Lemma 58. We conclude that $w \in H_{e, h-1} K_e H_{e, h-1} \subseteq H_{e, h}$.

This concludes the proof of Proposition 57.

6 Two fundamental concatenation hierarchies

This section is devoted to the dot-depth and Straubing-Thérien hierarchies. The dot-depth is the concatenation hierarchy whose basis is:

$$\text{DD}[0] = \{\emptyset, \{\varepsilon\}, A^+, A^*\},$$

while the Straubing-Thérien hierarchy is the concatenation hierarchy whose basis is:

$$\text{ST}[0] = \{\emptyset, A^*\}.$$

For $q \in \mathbb{N}$ or $q \in \frac{1}{2} + \mathbb{N}$, we denote by $\text{DD}[q]$ level q in the dot-depth hierarchy, and by $\text{ST}[q]$ level q in the Straubing-Thérien hierarchy. It is easy to see both hierarchies classify the star-free languages:

$$\text{SF} = \bigcup_{n \geq 0} \text{DD}[n] = \bigcup_{n \geq 0} \text{ST}[n].$$

This was the original motivation of Brzozowski and Cohen for considering the dot-depth hierarchy: for each language, one counts the number of alternations between complement and concatenation that are required to define it.

The Straubing-Thérien hierarchy quickly gained attention in the mid 80s after it was observed to be “more fundamental” than the dot-depth. This informal claim is motivated by the two following properties:

1. Straubing [49] showed that any full level in the dot-depth hierarchy is obtained by applying a uniform operation to the corresponding level in the Straubing-Thérien hierarchy. This result makes it possible to reduce membership for a level in the dot-depth to the same problem for corresponding level the Straubing-Thérien hierarchy. This result was pushed later to half-levels [31] and to separation [35, 37].
2. An important result that we already stated in Section 5 is that separation is decidable up to level $\frac{3}{2}$ in any hierarchy with a finite basis. In the particular case of the Straubing-Thérien hierarchy, it follows from a theorem of Pin and Straubing [28] that for every q , level $q + 1$ is also level q in another finitely based hierarchy. While simple, this result is crucial, as it allows us to lift the decidability results from Section 5 up to level $\frac{5}{2}$ in the Straubing-Thérien hierarchy (and therefore in the dot-depth as well by the first item above).

Thanks to the generic analysis carried out in previous sections, we know that levels of both hierarchies satisfy robust properties: since their bases are quotienting Boolean algebras, it follows from Proposition 38 that all half-levels are quotienting lattices and all full levels are quotienting Boolean algebras. Moreover, it follows from Proposition 39 that all half-levels are closed under concatenation and marked concatenation.

Additionally, Theorem 42 shows that over an alphabet of size at least 2, both hierarchies are strict (as their bases $ST[0]$ and $DD[0]$ are finite). In particular, this means that Theorem 3 is a simple corollary of Theorem 42.

Remark 59 We proved Theorem 3 for the dot-depth as the corollary of a more general theorem. However, there exist many specific proofs. This includes the original proof by Brzozowski and Knast [9] who exhibit languages L_n for all $n \geq 1$ such that L_n has dot-depth n but not dot-depth $n - 1$. The definition is as follows: $L_1 = (ab)^*$, and for $n \geq 2$, $L_n = (aL_{n-1}b)^*$. As expected, the difficulty is to prove that L_n does not have dot-depth $n - 1$. This proof has often been revisited. Let us mention the game theoretic proofs of Thomas [53, 54] or that by Thérien [51].

Let us also mention the proof of Straubing [48], which relies on a different approach based on algebra. Instead of working with the classes $DD[n]$, this argument proves strict inclusions between associated algebraic classes (namely, semigroup varieties). \square

Examples. Let us present some typical examples of languages that do or do not belong to the first levels of the dot-depth hierarchy.

Example 60 (Languages of dot-depth 1/2) Let $A = \{a, b\}$ and consider $DD[\frac{1}{2}] = Pol(\mathcal{C})$. We show that the language $L = a^*b^*$ does not belong to $DD[\frac{1}{2}]$. Indeed, by contradiction, assume that $L \in DD[\frac{1}{2}]$. Since $DD[0] = \{\emptyset, \{\varepsilon\}, A^+, A^*\}$, it is immediate that $a \leq_{DD[0]} b$. Moreover, for all $n \geq 1$, we have $a^n b \in L$ and $a^n b a^n b \notin L$. Hence, Corollary 34 (for $u = a$, $v = b$, $x = \varepsilon$ and $y = b$) yields that L is not $DD[\frac{1}{2}]$ -separable from $A^* \setminus L$ (i.e., $L \notin DD[\frac{1}{2}]$). \square

Example 61 (Languages of dot-depth one) Consider the alphabet $A = \{a, b\}$. The typical example of language having dot-depth one is $(ab)^*$. Indeed, we have $(ab)^* = A^* \setminus (bA^* \cup A^*a \cup A^*aaA^* \cup A^*bbA^*)$. Moreover, a similar argument as that of Example 60 shows that $(ab)^* \notin DD[\frac{1}{2}]$. Hence, $(ab)^*$ has dot-depth one. \square

Example 62 (Languages of dot-depth 3/2) We prove that the language $K = (a(ab)^*b)^*$ does not belong to $DD[\frac{3}{2}]$. Let us start with a preliminary result. We show that K is not $co\text{-}DD[\frac{1}{2}]$ -separable from $L = a(ab)^*a(ab)^*b$.

Since $DD[0] = \{\emptyset, \{\varepsilon\}, A^+, A^*\}$, it is immediate that $ab \leq_{DD[0]} b$. Moreover, for any $n \geq 1$, $a(ab)^n aabb \in L$ and $a(ab)^n b(ab)^n aabb \in K$. Hence, Corollary 35 (for $u = ab$, $v = b$, $x = a$ and $y = aabb$) yields that L is not $DD[\frac{1}{2}]$ -separable from K . By definition of $co\text{-}DD[\frac{1}{2}]$, this yields that K is not $co\text{-}DD[\frac{1}{2}]$ -separable from L .

We may now show that $K = (a(ab)^*b)^*$ does not belong to $DD[\frac{3}{2}]$. By contradiction, assume that $K \in DD[\frac{3}{2}]$. By Proposition 41, we know that $DD[\frac{3}{2}] = Pol(co\text{-}DD[\frac{1}{2}])$. It is then simple to verify using Lemma 17 that there exists some finite quotienting lattice $\mathcal{D} \subseteq co\text{-}DD[\frac{1}{2}]$ such that $K \in Pol(\mathcal{D})$. Since $\mathcal{D} \subseteq co\text{-}DD[\frac{1}{2}]$, we know that K is not \mathcal{D} -separable from L . Hence, Corollary 20 yields $u \in K$ and $v \in L$ such that $u \leq_{\mathcal{D}} v$. Furthermore, it is clear that for any $n \geq 1$, $u^n \in K$ (as $u \in K = (a(ab)^*b)^*$) and $u^n v u^n \notin K$ (as $u^n v u^n \in K L K$ and it is clear that $K L K \cap K = \emptyset$). Hence, Corollary 35 (for $x = y = \varepsilon$) yields that K is not $Pol(\mathcal{D})$ -separable from its complement which is a contradiction since $K \in Pol(\mathcal{D})$ by hypothesis. \square

Example 63 Let $A = \{a, b\}$. One can verify that $(ab)^*$ does not belong to $\text{ST}[1]$. However, it belongs to $\text{ST}[2]$. Indeed, observe that the singleton language $\{\varepsilon\}$ belongs to $\text{ST}[1]$ ($\{\varepsilon\} = A^* \setminus (\bigcup_{a \in A} A^* a A^*)$). Thus, bA^* , $A^* a$, $A^* aa A^*$ and $A^* bb A^*$ belong to $\text{ST}[\frac{3}{2}]$ and we may use the usual approach: $(ab)^*$ is the complement of $bA^* \cup A^* a \cup A^* aa A^* \cup A^* bb A^*$ and therefore belongs to $\text{ST}[2]$. \square

Example 64 (Languages of dot-depth two) Consider the alphabet $A = \{a, b\}$. The language $(a(ab)^* b)^*$ has dot-depth two. Indeed, one may verify that it is the complement of the following language:

$$(ab)^* b A^* + A^* aa (ba)^* a A^* + A^* b (ba)^* bb A^* + A^* a (ab)^*.$$

Clearly, the above language has dot-depth $\frac{3}{2}$ since $(ab)^*$ and $(ba)^*$ have dot-depth one by the previous example. Hence, $(a(ab)^* b)^*$ has dot-depth two. \square

Membership and Separation. By Theorem 10, if membership is decidable for some level in the Straubing-Thérien hierarchy, then it is also decidable for the corresponding level in the dot-depth hierarchy. Actually, the state of the art with respect to membership is the same for both hierarchies. In [35, 37], we generalized Theorem 10 to cope with separation, by defining a generic operation on classes of languages, which maps each level of the Straubing-Thérien hierarchy to the corresponding level in the dot-depth hierarchy, and which preserves decidability of separation (and of the more general “covering problem” [36, 41]). We outline the state of the art in Fig. 5.

Theorem 65 (Place & Zeitoun [35, 37]) *For any level $q \in \mathbb{N}$ or $q \in \frac{1}{2} + \mathbb{N}$, $\text{DD}[q]$ has decidable separation if and only if so does $\text{ST}[q]$.*

The alphabet trick. We now connect the Straubing-Thérien hierarchy with the concatenation hierarchy whose basis is the class AT of alphabet testable languages. While simple, this theorem has important consequences.

Let us briefly recall the definition of the alphabet testable languages. We shall connect two classes to the Straubing-Thérien hierarchy: AT itself and a weaker one, which we denote by WAT . For any alphabet A , recall that AT consists of all Boolean combinations of languages of the form $A^* a A^*$, for $a \in A$. Moreover, WAT contains all unions of languages B^* for $B \subseteq A$. Since $B^* = A^* \setminus \bigcup_{b \notin B} A^* b A^*$, we have $\text{WAT} \subseteq \text{AT}$. We already know that AT is a quotienting Boolean algebra of regular languages and one may verify that WAT is a quotienting lattice (closure under intersection follows from the fact that $B^* \cap C^* = (B \cap C)^*$).

It was proved by Pin and Straubing [28] that level $\frac{3}{2}$ in the Straubing-Thérien hierarchy¹ is also the class $\text{Pol}(\text{WAT})$.

Remark 66 Another formulation (which is essentially the original one by Pin and Straubing) is to say that $\text{ST}[\frac{3}{2}]$ contains exactly the unions of languages of the form,

$$B_0^* a_1 B_1^* a_2 B_2^* \cdots a_n B_n^* \quad \text{with } B_0, \dots, B_n \subseteq A.$$

We reformulate this result in the following crucial theorem.

¹ In fact, the original formulation of Pin and Straubing considers level 2 in the Straubing-Thérien hierarchy and not level $\frac{3}{2}$.

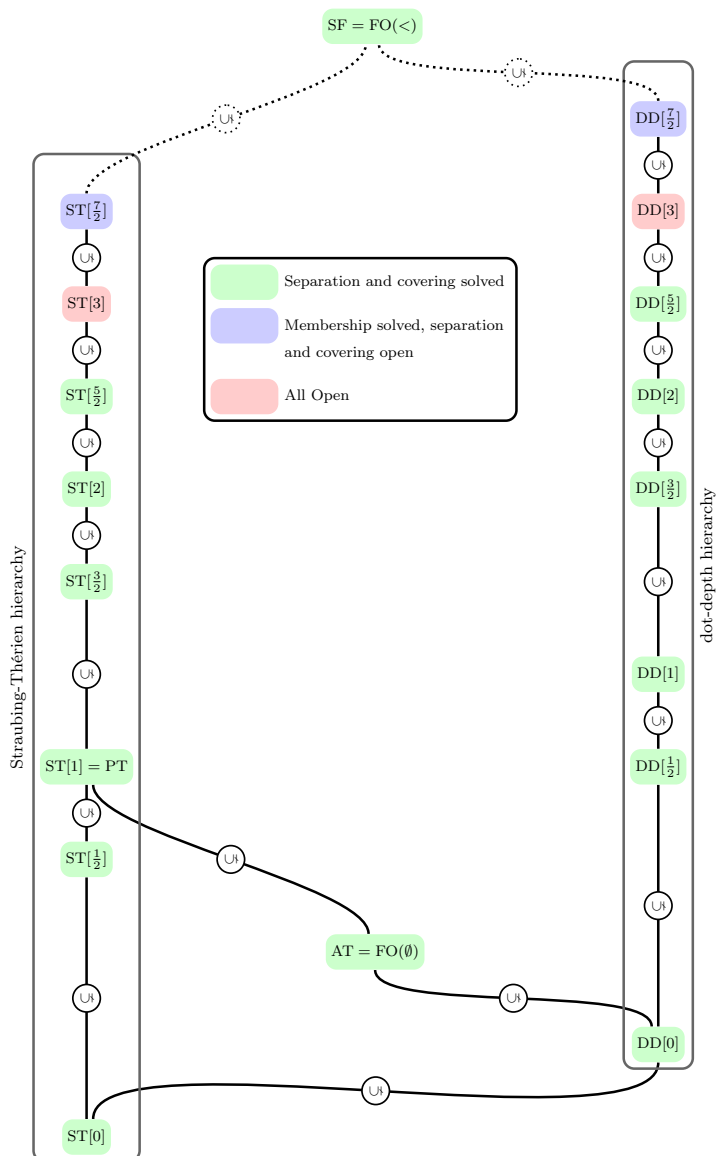


Fig. 5 Overview of classes. For the sake of avoiding clutter, inclusions between levels in the Straubing-Thérien and dot-depth hierarchies are omitted.

Theorem 67 (Pin and Straubing [28]) *Level $\frac{3}{2}$ in the Straubing-Thérien hierarchy satisfies the following property:*

$$\text{ST}[\frac{3}{2}] = \text{Pol}(\text{WAT}) = \text{Pol}(\text{AT}).$$

In particular, any level $q \geq \frac{3}{2}$ (half or full) in the Straubing-Thérien hierarchy corresponds exactly to level $q - 1$ in the concatenation hierarchy of basis AT.

The important point here is that while AT is more involved than $ST[0]$ as a basis, it remains *finite*. Therefore, Theorem 67 states that any level $q \geq \frac{3}{2}$ in the Straubing-Thérien hierarchy is also the level $q - 1$ in another hierarchy whose basis is finite. This saves us one level for the decidability results, therefore yielding Theorem 11.

Indeed, we know that separation is decidable up to level $\frac{3}{2}$ in *any* concatenation hierarchy whose basis is finite. This of course applies to the Straubing-Thérien hierarchy since $ST[0]$ is clearly finite. However, Theorem 67 allows us to go one step further and to lift these results to levels 2 and $\frac{5}{2}$ in the particular case of the Straubing-Thérien hierarchy. Indeed, they are also levels 1 and $\frac{3}{2}$ in the hierarchy of finite basis AT.

We now prove Theorem 67. Since $WAT \subseteq AT$, the inclusion $Pol(WAT) \subseteq Pol(AT)$ is trivial. We show that $Pol(AT) \subseteq ST[\frac{3}{2}] \subseteq Pol(WAT)$.

For the inclusion $Pol(AT) \subseteq ST[\frac{3}{2}]$, observe that $AT \subseteq ST[1]$. Indeed, Consider $L \in AT$. By definition, L is the Boolean combination of languages A^*aA^* for some $a \in A$. Clearly, $A^*aA^* \in ST[\frac{1}{2}]$ for any $a \in A$. Hence, $L \in ST[1] = Bool(ST[\frac{1}{2}])$. Consequently, $Pol(AT) \subseteq ST[\frac{3}{2}]$ by definition.

The inclusion $ST[\frac{3}{2}] \subseteq Pol(WAT)$ is more involved. We first reduce the proof to that of a simpler inclusion. Recall that we showed in Proposition 41 that,

$$ST[\frac{3}{2}] = Pol(co-ST[\frac{1}{2}]).$$

Therefore, in order to prove that $ST[\frac{3}{2}] \subseteq Pol(WAT)$, it suffices to show the following inclusion:

$$co-ST[\frac{1}{2}] \subseteq Pol(WAT) \tag{6}$$

Indeed, it will follow that $Pol(co-ST[\frac{1}{2}]) \subseteq Pol(Pol(WAT)) = Pol(WAT)$. We now concentrate on proving (6). This is a consequence of the following lemma.

Lemma 68 *For any $a_1, \dots, a_n \in A$, we have $A^* \setminus A^*a_1A^* \cdots a_nA^* \in Pol(WAT)$.*

Before proving Lemma 68, let us use it to show that the inclusion 6 holds. By definition, any language $L \in co-ST[\frac{1}{2}]$ is the complement of another language in the class $ST[\frac{1}{2}] = Pol(ST[0])$. Hence, by definition of polynomial closure, there exist $ST[0]$ -monomials K_1, \dots, K_m such that,

$$L = A^* \setminus \left(\bigcup_{i \leq m} K_i \right) = \bigcap_{i \leq m} A^* \setminus K_i$$

As $ST[0] = \{\emptyset, A^*\}$ all $ST[0]$ -monomials K_1, \dots, K_m have the form $A^*a_1A^* \cdots a_nA^*$. It follows from Lemma 68 that $A^* \setminus K_i \in Pol(WAT)$ for all $i \leq m$. Finally, since $Pol(WAT)$ is closed under intersection, we conclude that $L \in Pol(WAT)$, as desired.

Remark 69 While the above argument may seem simple, let us point out that we implicitly used Theorem 29 which is an involved result. On one hand, we used the original definition of polynomial closures for $Pol(ST[0])$ (*i.e.*, it contains the unions of $ST[0]$ -monomials). On the other hand, we used the fact that $Pol(WAT)$ is closed under intersection which is not immediate from the definition: this is where we need Theorem 29.

It remains to prove Lemma 68. Consider n letters $a_1, \dots, a_n \in A$. By a *piece* of a word, we mean a scattered subword. Our goal is to show that $A^* \setminus A^* a_1 A^* \cdots a_n A^* \in \text{Pol}(\text{WAT})$. For all $k \leq n$, we let $L_k = A^* \setminus A^* a_1 A^* \cdots a_k A^*$. Observe that by definition, for any $k \leq n$, the language L_k consists of all words w such that $a_1 \cdots a_k$ is **not** a piece of w . We prove by induction on k that $L_k \in \text{Pol}(\text{WAT})$ for all $k \leq n$.

When $k = 1$, this is immediate since $L_1 = (A \setminus \{a_1\})^*$, which belongs to WAT (and therefore to $\text{Pol}(\text{WAT})$) by definition. We now assume that $k \geq 2$. Consider the following language H_k :

$$H_k = (A \setminus \{a_k\})^* \cup L_{k-1} a_k (A \setminus \{a_k\})^*$$

By induction hypothesis, we have $L_{k-1} \in \text{Pol}(\text{WAT})$. Moreover, it is immediate from the definition of WAT that $(A \setminus \{a_k\})^* \in \text{WAT} \subseteq \text{Pol}(\text{WAT})$. Hence, we conclude that $H_k \in \text{Pol}(\text{WAT})$ using closure under marked concatenation and union. We now show that $L_k = H_k$, which terminates the proof.

We begin with $L_k \subseteq H_k$. Let $w \in L_k$. We consider two cases depending on whether w contains the letter a_k or not. If $a_k \notin \text{alph}(w)$, then $w \in (A \setminus \{a_k\})^*$ which is a subset of H_k by definition. Hence, $w \in H_k$. Otherwise, $a_k \in \text{alph}(w)$. We decompose w by looking at the rightmost copy of the letter a_k : $w = ua_k v$ with $u \in A^*$ and $v \in (A \setminus \{a_k\})^*$. It is now immediate that $u \in L_{k-1}$. Indeed, otherwise the word $a_1 \cdots a_{k-1}$ would be a piece of u and therefore $a_1 \cdots a_k$ would be a piece of w which is not possible since $w \in L_k$ by hypothesis. We conclude that $w = ua_k v \in L_{k-1} a_k (A \setminus \{a_k\})^*$ which is a subset of H_k by definition. Thus, $w \in H_k$.

We turn to the second inclusion: $H_k \subseteq L_k$. Let $w \in H_k$. If $w \in (A \setminus \{a_k\})^*$, then it is clear that $a_1 \cdots a_k$ is not a piece of w which means that $w \in L_k$. Otherwise, $w \in L_{k-1} a_k (A \setminus \{a_k\})^*$. Thus, $w = ua_k v$ with $u \in L_{k-1}$ and $a_k \notin \text{alph}(v)$. By contradiction, assume that $a_1 \cdots a_k$ is a piece of w . Since $a_k \notin \text{alph}(v)$, it follows that $a_1 \cdots a_{k-1}$ must be a piece of u , which is impossible since $u \in L_{k-1}$. We conclude that $a_1 \cdots a_k$ is not a piece of w , and therefore that $w \in L_k$. This terminates the proof of Lemma 68. \square

7 The link with logic

In this section, we present *quantifier alternations hierarchies*, whose levels are defined by fragments of first-order logic. Such hierarchies classify languages according to the type of sentences needed to define them: the classifying parameter is the number of alternations between \exists and \forall quantifiers that are necessary to define a language.

The main theorem is a *generic* correspondence between concatenation and quantifier alternation hierarchies. For every basis \mathcal{C} , the concatenation hierarchy of basis \mathcal{C} corresponds exactly to the quantifier alternation hierarchy within a well-chosen variant of first-order logic. This generic connection was originally observed by Thomas [52], who obtained it in a specific case: he showed that the dot-depth hierarchy corresponds exactly to the quantifier alternation hierarchy within the variant $\text{FO}(<, +1, \min, \max, \varepsilon)$ of first-order logic.

We organize the section as follows: we first present first-order logic and quantifier alternation hierarchies. Then, we state and prove the main theorem of the section: the correspondence between concatenation and quantifier alternation hierarchies. Finally, we instantiate this result on the dot-depth and Straubing-Thérien hierarchies.

7.1 Quantifier alternation hierarchies

For defining languages with first-order logic, we view words as relational structures: a word of length n is a sequence of positions $\{1, \dots, n\}$ labeled over alphabet A . A *signature* is a set of predicate symbols, each of them having an arity. Given a word of length n , a predicate of arity k is interpreted as a k -ary relation on the set $\{1, \dots, n\}$ of positions of the word. Important examples of predicates are the following:

- ε , the nullary “empty” predicate, which holds when the word is empty. That is, given a word w , the predicate ε holds when $w = \varepsilon$.
- For each $a \in A$, a unary “label” predicate, also denoted by a . Given a word w and a position i in w , $a(i)$ holds when position i in w carries letter a .
- $\min(x)$, the unary “minimum” predicate, which selects the first position of a word.
- $\max(x)$, the unary “maximum” predicate, which selects the last position.
- $<$, the binary “order” predicate, interpreted as the linear order on positions.
- $+1$, the binary “successor” predicate, interpreted as the successor relation.

Each signature \mathcal{S} defines a variant of first-order logic, which we denote by $\text{FO}(\mathcal{S})$. For concrete signatures, we will not write the label predicates, *i.e.*, they will be always understood. For instance, $\text{FO}(<)$ denotes the variant of first-order logic for the signature consisting of the order predicate *and* all label predicates.

For a given signature \mathcal{S} , we define the semantics of $\text{FO}(\mathcal{S})$ of first-order logic as follows: one may quantify over positions of a word, use Boolean connectives as well as the \top (true) and \perp (false) formulas, and test properties of the quantified positions using the predicate symbols from \mathcal{S} . Each first-order sentence of $\text{FO}(\mathcal{S})$ therefore defines a language over A^* .

More formally, let $w = b_1 \dots b_n \in A^*$ be a word and let \mathcal{X} be some finite set of first-order variables. An *assignment of \mathcal{X} in w* is a map μ from \mathcal{X} to the set of positions of w (*i.e.*, $\mu : \mathcal{X} \rightarrow \{1, \dots, n\}$). In particular, if μ is an assignment of \mathcal{X} in w , x a variable (not necessarily in \mathcal{X}) and i a position in w , we will denote by $\mu[x \mapsto i] : (\mathcal{X} \cup \{x\}) \rightarrow \{1, \dots, n\}$, the assignment of $\mathcal{X} \cup \{x\}$ in w that is identical to μ except that it maps x to i . We can now define the semantic of a first-order formula.

Let φ be a first-order formula and assume that \mathcal{X} contains all free variables of φ . Then, for every word $w = b_1 \dots b_n \in A^*$ and every assignment μ of \mathcal{X} in w , we say that w *satisfies* φ under μ , written $w, \mu \models \varphi$, when one the following properties hold:

- $\varphi := \top$.
- $\varphi := P(x_1, \dots, x_k)$ for some predicate $P \in \mathcal{S}$ and $P(\mu(x_1), \dots, \mu(x_k))$ holds.
- $\varphi := \exists x \Psi$ and there exists a position $i \in \{1, \dots, n\}$ such that $w, \mu[x \mapsto i] \models \Psi$.
- $\varphi := \forall x \Psi$ and for every position $i \in \{1, \dots, n\}$, we have $w, \mu[x \mapsto i] \models \Psi$.
- $\varphi := \Psi \vee \Gamma$ and $w, \mu \models \Psi$ or $w, \mu \models \Gamma$.
- $\varphi := \Psi \wedge \Gamma$ and $w, \mu \models \Psi$ and $w, \mu \models \Gamma$.
- $\varphi := \neg \Psi$ and $w, \mu \not\models \Psi$ (w does not satisfy Ψ under μ).

When φ is a sentence, whether $w, \mu \models \varphi$ does not depend on μ . In that case, we simply write $w \models \varphi$. Every sentence φ defines the language $\{w \in A^* \mid w \models \varphi\}$.

We now define a hierarchy of fragments within $\text{FO}(\mathcal{S})$ by classifying all $\text{FO}(\mathcal{S})$ sentences according to the number of quantifier alternations within their parse trees. For $i \in \mathbb{N}$, a formula is $\Sigma_i(\mathcal{S})$ (resp. $\Pi_i(\mathcal{S})$) when it is a $\text{FO}(\mathcal{S})$ formula whose prenex normal form has either:

- exactly n blocks of quantifiers, the leftmost being an “ \exists ” (resp. a “ \forall ”) block,
- or strictly less than n blocks of quantifiers.

For example, a formula whose prenex normal form is

$$\forall x_1 \forall x_2 \exists x_3 \forall x_4 \varphi(x_1, x_2, x_3, x_4) \quad (\text{with } \varphi \text{ quantifier-free})$$

is $\Pi_3(\mathcal{S})$. Observe that a $\Pi_i(\mathcal{S})$ formula is the negation of a $\Sigma_i(\mathcal{S})$ formula. Finally, a $\mathcal{B}\Sigma_i(\mathcal{S})$ formula is a Boolean combination of $\Sigma_i(\mathcal{S})$ formulas. Note that by definition, we have $\mathcal{B}\Sigma_{i-1}(\mathcal{S}) \subseteq \Sigma_i(\mathcal{S}) \subseteq \mathcal{B}\Sigma_i(\mathcal{S})$ and $\mathcal{B}\Sigma_{i-1}(\mathcal{S}) \subseteq \Pi_i(\mathcal{S}) \subseteq \mathcal{B}\Sigma_i(\mathcal{S})$ for every $i \geq 1$. It is also clear that every $\text{FO}(\mathcal{S})$ formula belongs to some of these classes.

We lift this syntactic definition to the semantic level: for $X = \text{FO}(\mathcal{S})$, $\Sigma_i(\mathcal{S})$, $\Pi_i(\mathcal{S})$ or $\mathcal{B}\Sigma_i(\mathcal{S})$, we say that a language L is X -definable if it can be defined by an X -formula. Abusing notation, we also denote by X the class of X -definable languages. This gives us a hierarchy of languages depicted in Fig. 6.

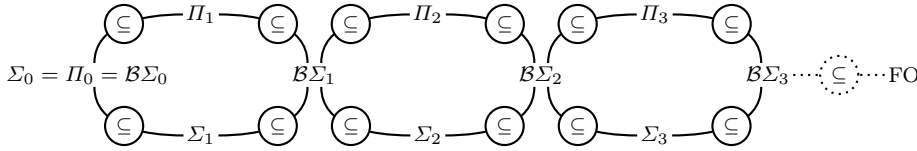


Fig. 6 Quantifier alternation hierarchy of first-order logic

Whether a particular hierarchy is strict depends on its signature \mathcal{S} . The two most prominent hierarchies in the literature are known to be strict. These are:

- The one associated to the logic $\text{FO}(<)$.
- The one associated to the logic $\text{FO}(<, +1, \min, \max, \varepsilon)$.

Remark 70 It is a classical result that $\text{FO}(<)$ and $\text{FO}(<, +1, \min, \max, \varepsilon)$ have the same expressive power: all predicates available in $\text{FO}(<, +1, \min, \max, \varepsilon)$ can be defined from the linear order. However, this is not the case for levels in their respective quantifier alternation hierarchies. Intuitively, the reason is that defining the predicates “+1”, “min” and “max” from “<” costs quantifier alternations.

Finally, a last useful lemma is that we can bypass $\mathcal{B}\Sigma_n(\mathcal{S})$ formulas in the definition of quantifier alternation hierarchies.

Lemma 71 *For any $n \geq 0$, any $\Sigma_{n+1}(\mathcal{S})$ formula is equivalent to a formula of the form $\psi \vee \forall x \perp$ or $\psi \wedge \exists x \top$, where ψ belongs to the closure of $\Pi_n(\mathcal{S})$ under existential quantification.*

Proof On nonempty words, any formula from $\Sigma_{n+1}(\mathcal{S})$ is equivalent to its prenex normal form, which by definition either belongs itself to $\Pi_n(\mathcal{S})$, or is of the form $\exists x_1 \dots \exists x_k \psi$ where ψ is a $\Pi_n(\mathcal{S})$ formula. The disjunction with $\forall x \perp$ and the conjunction with $\exists x \top$ are used to add or remove the empty word from the language of the formula. This concludes the proof. \square

7.2 Main theorem

We are ready to present and prove the generic correspondence existing between quantifier alternation and concatenation hierarchies. More precisely, we show that for any basis \mathcal{C} , one may define an appropriate signature (also denoted by \mathcal{C}) such that the concatenation hierarchy of basis \mathcal{C} and the quantifier alternation hierarchy within $\text{FO}(\mathcal{C})$ are identical.

Consider an arbitrary basis \mathcal{C} . We associate a signature to \mathcal{C} and consider the variant of first-order logic equipped with this signature. As usual, the signature associated to \mathcal{C} contains all label predicates: for every $a \in A$, we have a unary predicate (also denoted by “ a ”) which is interpreted as the unary relation selecting all positions whose label is a . Moreover, for every language $L \in \mathcal{C}$, we add four predicates:

- A binary predicate I_L interpreted as follows: given a word w and two positions i, j in w , $I_L(i, j)$ holds when $i < j$ and the infix $w[i, j[$ belongs to L .
- A unary predicate P_L interpreted as follows: given a word w and a position i in w , $P_L(i)$ holds when the prefix $w[1, i[$ belongs to L .
- A unary predicate S_L interpreted as follows: given a word w and a position i in w , $S_L(i)$ holds when the suffix $w]i, |w|]$ belongs to L .
- A nullary predicate N_L interpreted as follows: given a word w , N_L holds when w belongs to L .

Abusing notation, we denote by $\text{FO}(\mathcal{C})$ the associated variant of first-order logic.

Remark 72 Observe that these signatures always contain the label predicates and the linear order “ $<$ ”. Indeed, by definition, “ $<$ ” is the binary predicate I_{A^*} , and A^* belongs to \mathcal{C} since it is a quotienting Boolean algebra. Thus, all variants of first-order logic that we consider here are at least as expressive as $\text{FO}(<)$. In fact, $\text{FO}(<) = \text{FO}(\mathcal{C})$ when $\mathcal{C} = \{\emptyset, A^*\}$. We shall detail this point in the next section.

We now state the theorem establishing an exact correspondence between the concatenation hierarchy of basis \mathcal{C} and the quantifier alternation hierarchy within $\text{FO}(\mathcal{C})$.

Theorem 73 *Let \mathcal{C} be a quotienting Boolean algebra. For every $n \in \mathbb{N}$ and every language $L \subseteq A^*$, the two following properties hold:*

1. $L \in \mathcal{C}[n]$ if and only if L can be defined by a $\mathcal{B}\Sigma_n(\mathcal{C})$ sentence.
2. $L \in \mathcal{C}[n + \frac{1}{2}]$ if and only if L can be defined by a $\Sigma_{n+1}(\mathcal{C})$ sentence.

The remainder of the section is devoted to proving Theorem 73. A first observation is that we may concentrate on the second item as the first one is a simple corollary. Indeed, for $n = 0$, a $\mathcal{B}\Sigma_0(\mathcal{C})$ sentence is by definition a Boolean combination of atomic formulas which do not involve variables. This includes \perp, \top and the nullary predicates N_L for $L \in \mathcal{C}$. By definition of the predicates N_L and since \mathcal{C} is a Boolean algebra, it follows that $\mathcal{B}\Sigma_0(\mathcal{C}) = \mathcal{C} = \mathcal{C}[0]$. A second observation is that for $n \geq 1$, we have $\mathcal{B}\Sigma_n(\mathcal{C}) = \text{Bool}(\Sigma_n(\mathcal{C}))$. Hence, the equality $\Sigma_n(\mathcal{C}) = \mathcal{C}[n - \frac{1}{2}]$ immediately yields $\mathcal{B}\Sigma_n(\mathcal{C}) = \mathcal{C}[n]$.

We now concentrate on proving the second item in Theorem 73. The proof is divided in two steps, one for each inclusion. We first show the easier one, namely,

$$\mathcal{C}[n + \frac{1}{2}] \subseteq \Sigma_{n+1}(\mathcal{C}).$$

The proof is by induction on n . The key ingredient is the following lemma which states that for every $n \in \mathbb{N}$, $\Sigma_n(\mathcal{C})$ is closed under marked concatenation.

Lemma 74 *Let $n \in \mathbb{N}$ and $L_1, L_2 \subseteq A^*$ be two languages in $\Sigma_n(\mathcal{C})$. Then, for every $a \in A$, the marked concatenation $L_1 a L_2$ also belong to $\Sigma_n(\mathcal{C})$.*

Proof Let L_1 and L_2 be languages defined by two $\Sigma_n(\mathcal{C})$ sentences φ_1 and φ_2 . We have to construct a third sentence ψ that defines $L_1 a L_2$. Let x be a fresh variable with respect to both φ_1 and φ_2 . We build two formulas $\varphi'_1(x)$ and $\varphi'_2(x)$ (each with x as a single free variable) with the following semantics. Given $w \in A^*$ and μ an assignment for w with domain $\{x\}$:

- $w, \mu \models \varphi'_1(x)$ if and only if the prefix $w[1, \mu(x)[$ belongs to L_1 (that is, if and only if $w[1, \mu(x)[\models \varphi_1$). Observe that this prefix may be empty when $\mu(x) = 1$.
- $w, \mu \models \varphi'_2(x)$ if and only if the suffix $w]\mu(x), |w|]$ belongs to L_2 (that is, if and only if $w]\mu(x), |w|] \models \varphi_2$). Observe that this suffix may be empty when $\mu(x) = |w|$.

The constructions of $\varphi'_1(x)$ and $\varphi'_2(x)$ are symmetrical. Let us describe that of $\varphi'_1(x)$. We build it from φ_1 as follows:

1. We relativize quantifications to positions that are to the left of x . That is, we replace every sub-formula of the form $\exists y \xi$ (resp. $\forall y \xi$) by $\exists y y < x \wedge \xi$ (resp. $\forall y \neg(y < x) \vee \xi$).
2. We replace atomic formulas of the form N_L for some $L \in \mathcal{C}$ by $P_L(x)$.
3. We replace atomic formulas of the form $S_L(y)$ for some $L \in \mathcal{C}$ by $I_L(y, x)$.

Clearly, $\varphi'_1(x)$ is also a $\Sigma_n(\mathcal{C})$ formula and one may verify that it satisfies the above property. We can now define ψ for $L_1 a L_2$ as follows:

$$\psi = \exists x a(x) \wedge \varphi'_1(x) \wedge \varphi'_2(x).$$

It is obvious that ψ is a $\Sigma_n(\mathcal{C})$ sentence defining the language $L_1 a L_2$.

We may now prove that $\mathcal{C}[n + \frac{1}{2}] \subseteq \Sigma_{n+1}(\mathcal{C})$ for every $n \in \mathbb{N}$. We proceed by induction on n . When $n = 0$, we first note that $\mathcal{C} \subseteq \Sigma_1(\mathcal{C})$. Indeed, every language L of \mathcal{C} is defined by the atomic sentence N_L . Therefore, $\mathcal{C}[\frac{1}{2}] = \text{Pol}(\mathcal{C}) \subseteq \Sigma_1(\mathcal{C})$, since $\Sigma_1(\mathcal{C})$ is closed under union and marked concatenation.

When $n \geq 1$, we know that $\mathcal{C}[n + \frac{1}{2}] = \text{Pol}(\text{co}\mathcal{C}[n - \frac{1}{2}])$ by Proposition 41. By induction hypothesis, we have $\mathcal{C}[n - \frac{1}{2}] \subseteq \Sigma_n(\mathcal{C})$ and therefore,

$$\text{co}\mathcal{C}[n - \frac{1}{2}] \subseteq \mathcal{B}\Sigma_n(\mathcal{C}) \subseteq \Sigma_{n+1}(\mathcal{C}).$$

Hence, since $\Sigma_{n+1}(\mathcal{C})$ is closed under union and marked concatenation, we obtain as desired that $\mathcal{C}[n + \frac{1}{2}] \subseteq \Sigma_{n+1}(\mathcal{C})$, finishing the proof for this direction.

It remains to establish the converse inclusion, *i.e.*, that for every $n \in \mathbb{N}$:

$$\Sigma_{n+1}(\mathcal{C}) \subseteq \mathcal{C}[n + \frac{1}{2}]. \quad (7)$$

Since the proof works inductively on the formulas, we have to explain how we handle free variables. We do this using Büchi's classical idea, *i.e.*, by encoding a word and a assignment of first-order variables as a single word over an extended alphabet.

Let $\mathcal{X} = \{x_1, x_2, x_3, \dots\}$ be an infinite linearly ordered set of first-order variables. One may assume that all FO(\mathcal{C}) formulas that we consider only use variables from \mathcal{X} . Given $\ell \in \mathbb{N}$, we use the alphabet $A_\ell = \{0, 1\}^\ell \times A$ to represent pairs (w, μ) with $w \in A^*$ and μ an assignment of $\{x_1, \dots, x_\ell\}$ in the positions of w .

For every $h \leq \ell$, we denote by $\pi_h : A_\ell \rightarrow \{0, 1\}$ the projection on component h . Similarly, we denote by $\pi_A : A_\ell \rightarrow A$ the projection on the rightmost component (component $\ell + 1$). Note that there are actually several mappings π_A , one for each value of ℓ , and similarly for π_h . Which mapping we use will be clear from the context.

We can now present the encoding. Let $w = a_1 \cdots a_n \in A^*$ and let μ be an assignment of $\{x_1, \dots, x_\ell\}$ in w . We encode the pair (w, μ) by the word $[w]_\mu = \bar{b}_1 \cdots \bar{b}_n$ of $(A_\ell)^*$ where for all $i \leq n$, the letter $\bar{b}_i \in A_\ell$ is defined as follows:

- $\pi_A(\bar{b}_i) = a_i$.
- For all $h \leq \ell$,
 - If $i = \mu(x_h)$, then $\pi_h(\bar{b}_i) = 1$.
 - If $i \neq \mu(x_h)$, then $\pi_h(\bar{b}_i) = 0$.

Note that when $\ell = 0$, we have $A_0 = A$ and $[w]_\mu = w$ (μ is the empty assignment). Clearly, the map $(w, \mu) \mapsto [w]_\mu$ is injective (however, it is not surjective since for every $h \leq \ell$, there is exactly one position i such that $\pi_h(\bar{b}_i) = 1$). For $\ell \geq 1$, we define the following class of languages over the alphabet A_ℓ :

$$\mathcal{C}_\ell \stackrel{\text{def}}{=} \{\pi_A^{-1}(L) \subseteq A_\ell^* \mid L \in \mathcal{C}\}.$$

It is straightforward to verify that \mathcal{C}_ℓ is a quotienting Boolean algebra of regular languages. Moreover, for any $\ell \in \mathbb{N}$, we define a morphism $\alpha_\ell : A_\ell^* \rightarrow A_{\ell+1}^*$ as follows: given $(i_1, \dots, i_\ell, a) \in A_\ell$, we let $\alpha_\ell(i_1, \dots, i_\ell, a) = (i_1, \dots, i_\ell, 0, a) \in A_{\ell+1}$. We now state a connection between the concatenation hierarchies of bases \mathcal{C}_ℓ and $\mathcal{C}_{\ell+1}$.

Fact 75 *For any $\ell, n \in \mathbb{N}$ and any $K \in \mathcal{C}_{\ell+1}[n + \frac{1}{2}]$, we have $\alpha_\ell^{-1}(K) \in \mathcal{C}_\ell[n + \frac{1}{2}]$.*

Proof This is immediate by induction on n and the definition of concatenation hierarchies using the following properties. For any $K_1, K_2 \subseteq A_{\ell+1}^*$, we have:

1. By definition, when $K_1 \in \mathcal{C}_{\ell+1}$, we have $\alpha_\ell^{-1}(K_1) \in \mathcal{C}_\ell$.
2. $\alpha_\ell^{-1}(K_1 \cup K_2) = \alpha_\ell^{-1}(K_1) \cup \alpha_\ell^{-1}(K_2)$.
3. $\alpha_\ell^{-1}(A_{\ell+1}^* \setminus K_1) = A_\ell^* \setminus \alpha_\ell^{-1}(K_1)$.
4. For every $\bar{b} \in A_{\ell+1}$, we have $\alpha_\ell^{-1}(K_1 \bar{b} K_2) = \alpha_\ell^{-1}(K_1) \alpha_\ell^{-1}(\bar{b}) \alpha_\ell^{-1}(K_2)$.

This concludes the proof of Fact 75. \square

The proof of the remaining inclusion (7) relies on the following proposition.

Proposition 76 *Let $\ell, n \in \mathbb{N}$ and let φ be a $\Sigma_{n+1}(\mathcal{C})$ formula whose free variables belong to $\{x_1, \dots, x_\ell\}$. Then, there exists a language $L_{\ell, \varphi} \in \mathcal{C}_\ell[n + \frac{1}{2}]$ such that for every $w \in A^*$ and every assignment μ of $\{x_1, \dots, x_\ell\}$ in w , we have,*

$$[w]_\mu \in L_{\ell, \varphi} \quad \text{if and only if} \quad w, \mu \models \varphi. \quad (8)$$

Note that the special case $\ell = 0$ of Proposition 76 yields the following corollary.

Corollary 77 *Let $n \in \mathbb{N}$ and let φ be a $\Sigma_{n+1}(\mathcal{C})$ sentence. Then, there exists a language $L \in \mathcal{C}[n + \frac{1}{2}]$ such that for every $w \in A^*$, we have,*

$$w \in L \quad \text{if and only if} \quad w \models \varphi.$$

Corollary 77 implies that for all $n \in \mathbb{N}$, we have $\Sigma_{n+1}(\mathcal{C}) \subseteq \mathcal{C}[n + \frac{1}{2}]$, which is the inclusion (7) that remained to be proved, concluding the proof of Theorem 73.

It remains to prove Proposition 76. Let $\ell, n \in \mathbb{N}$ and let φ be a $\Sigma_{n+1}(\mathcal{C})$ formula whose free variables belong to $\{x_1, \dots, x_\ell\}$. We construct by induction on n a language $L_{\ell, \varphi} \in \mathcal{C}_\ell[n + \frac{1}{2}]$ satisfying Condition (8) of Proposition 76.

Recall that we showed in Lemma 71 that we may assume without loss of generality that φ is built from negations of $\Sigma_n(\mathcal{C})$ formulas using existential quantifications. We use a sub-induction on this construction. We start with the base case, which we treat differently for $n = 0$ and $n \geq 1$ (essentially, the former amounts to treating atomic formulas, while the later is immediate by induction on n).

Base case. φ is the negation $\varphi = \neg\psi$ of some $\Sigma_n(\mathcal{C})$ formula ψ .

We first treat the case $n \geq 1$, which is where we use induction on n . Indeed, induction yields $L_{\ell, \psi} \in \mathcal{C}_\ell[n - \frac{1}{2}]$ such that for every $w \in A^*$ and every assignment μ of $\{x_1, \dots, x_\ell\}$ in w , we have,

$$[w]_\mu \in L_{\ell, \psi} \quad \text{if and only if} \quad w, \mu \models \psi.$$

It suffices to choose $L_{\ell, \varphi} = A_\ell^* \setminus L_{\ell, \psi} \in \mathcal{C}_\ell[n] \subseteq \mathcal{C}_\ell[n + \frac{1}{2}]$, which clearly meets the conditions in Proposition 76.

It remains to treat the case $n = 0$. By definition, the $\Sigma_0(\mathcal{C})$ formulas are the quantifier-free formulas. Thus, $\varphi = \neg\psi$ is itself a $\Sigma_0(\mathcal{C})$ formula. In other words φ is a Boolean combination of atomic formulas. Moreover, if we allow the equality predicate in the signature, we may eliminate all negations in φ . Indeed, using De Morgan's laws, one may first push all negations to atomic formulas. Furthermore, given any atomic formula, its negation is equivalent to a $\Sigma_0(\mathcal{C})$ formula without negation (this is where we need equality). Indeed, given $a \in A$, $\neg a(x)$ is equivalent to $\bigvee_{c \neq a} c(x)$. Finally, for any $K \in \mathcal{C}$, we have the following (recall that since \mathcal{C} is a quotienting Boolean algebra, $A^* \setminus H$ belongs to \mathcal{C} as well):

- $\neg I_K(x, y)$ is equivalent to $I_{A^*}(y, x) \vee x = y \vee I_{A^* \setminus K}(x, y)$.
- $\neg P_K(x)$ is equivalent to $P_{A^* \setminus K}(x)$.
- $\neg S_K(x)$ is equivalent to $S_{A^* \setminus K}(x)$.
- $\neg N_K$ is equivalent to $N_{A^* \setminus K}$.

Hence, we may assume without loss of generality that there is no negation in φ , which is therefore in $\Sigma_0(\mathcal{C})$. Hence, φ is built from atomic formulas using conjunctions and disjunctions. We may handle disjunctions and conjunctions in the obvious way. Hence, it suffices to treat the cases when φ is atomic.

There are two kinds of atomic formulas: those involving the label predicates and those which are specific to \mathcal{C} . Moreover, we also need to treat equality since we used it above to eliminate negations. Let us first assume that $\varphi = a(x_h)$ for some $h \leq \ell$ and some $a \in A$. Consider the set B of all letters in A_ℓ whose component h is equal to 1 and whose component $\ell + 1$ is equal to a :

$$B = \{\bar{b} \in A_\ell \mid \pi_h(\bar{b}) = 1 \text{ and } \pi_A(\bar{b}) = a\}.$$

It now suffices to define $L_{\ell, \varphi} = A_\ell^* B A_\ell^* \in \text{Pol}(\mathcal{C}_\ell) = \mathcal{C}_\ell[\frac{1}{2}]$. It is again immediate from the definitions that $L_{\ell, \varphi}$ satisfies the conditions in Proposition 76.

We now assume that $\varphi = "x_g = x_h"$ for some $g, h \leq \ell$. We now let B as the set of all letters in A_ℓ whose components g and h are both equal to 1.

$$B = \{\bar{b} \in A_\ell \mid \pi_g(\bar{b}) = 1 \text{ and } \pi_h(\bar{b}) = 1\}.$$

It now suffices to define $L_{\ell,\varphi} = A_\ell^* B A_\ell^* \in \text{Pol}(\mathcal{C}_\ell) = \mathcal{C}_\ell[\frac{1}{2}]$. It is then immediate from the definitions that $L_{\ell,\varphi}$ satisfies the conditions in Proposition 76.

It remains to treat the predicates given by \mathcal{C} . Since the argument is the same for all four kinds, we only treat the case when $\varphi = I_K(x_i, x_j)$, for some $K \in \mathcal{C}$ and $i, j \leq \ell$. We may assume that $i \neq j$, since $I_K(x_i, x_i)$ is equivalent to \perp . By symmetry, we may then assume that $i < j$. Let B_i and B_j be the following sub-alphabets of A_ℓ :

$$\begin{cases} B_i = \{\bar{b} \in A_\ell \mid \pi_i(\bar{b}) = 1\}, \\ B_j = \{\bar{b} \in A_\ell \mid \pi_j(\bar{b}) = 1\}. \end{cases}$$

We define $L_{\ell,\varphi} = A_\ell^* B_i \pi_A^{-1}(K) B_j A_\ell^*$. Recall that the language $\pi_A^{-1}(K) \subseteq A_\ell^*$ belongs to \mathcal{C}_ℓ (by definition of \mathcal{C}_ℓ). Hence, we have $L_{\ell,\varphi} \in \text{Pol}(\mathcal{C}) = \mathcal{C}[\frac{1}{2}]$. One may then verify that $L_{\ell,\varphi}$ satisfies the conditions in Proposition 76.

This concludes the base case of our structural induction on the formula φ . We now consider the inductive case which is handled uniformly for $n = 0$ and $n \geq 1$.

Inductive case: First-order quantification. Assume that φ is of the form $\exists x \psi$. Since variables can be renamed, we may assume without loss of generality that $x = x_{\ell+1}$, i.e., $\varphi = \exists x_{\ell+1} \psi$. Thus, all free variables of ψ belong to $\{x_1, \dots, x_{\ell+1}\}$. Applying induction to ψ yields a language $L_{\ell+1,\psi} \in \mathcal{C}_{\ell+1}[n + \frac{1}{2}]$ such that for every $w \in A^*$ and every assignment γ of $\{x_1, \dots, x_{\ell+1}\}$ in w , we have,

$$[w]_\gamma \in L_{\ell+1,\psi} \quad \text{if and only if} \quad w, \gamma \models \psi.$$

We first define $L_{\ell,\varphi} \in \mathcal{C}_\ell[n + \frac{1}{2}]$ and then prove that it satisfies (8). Given any word $u \in A_{\ell+1}^*$, we say that u is *good* when there exists exactly one position in u whose label \bar{b} satisfies $\pi_{\ell+1}(\bar{b}) = 1$ (which implies that the labels \bar{c} of all other positions satisfy $\pi_{\ell+1}(\bar{c}) = 0$). Let $\pi_{1,\dots,\ell,A} : A_{\ell+1}^* \rightarrow A_\ell^*$ be the projection which discards component $\ell + 1$ in words belonging to $A_{\ell+1}^*$. More precisely,

$$\pi_{1,\dots,\ell,A}(i_1, \dots, i_{\ell+1}, a) = (i_1, \dots, i_\ell, a).$$

We now define $L_{\ell,\varphi} \subseteq A_\ell^*$ as the following language:

$$L_{\ell,\varphi} = \{\pi_{1,\dots,\ell,A}(u) \mid u \in L_{\ell+1,\psi} \text{ and } u \text{ is good}\}. \quad (9)$$

It remains to prove that $L_{\ell,\varphi} \in \mathcal{C}_\ell[n + \frac{1}{2}]$ and that it satisfies Property (8) from Proposition 76. We first deal with Property (8).

Lemma 78 *Let $w \in A^*$ and let μ be an assignment of $\{x_1, \dots, x_\ell\}$ in the positions of w . Then, we have,*

$$[w]_\mu \in L_{\ell,\varphi} \quad \text{if and only if} \quad w, \mu \models \varphi.$$

Proof Assume first that $[w]_\mu \in L_{\ell,\varphi}$. By definition, there exists $u \in L_{\ell+1,\psi}$ which is good and such that $\pi_{1,\dots,\ell,A}(u) = [w]_\mu$. Since u is good, there exists exactly one position in u whose label \bar{b} satisfies $\pi_{\ell+1}(\bar{b}) = 1$. Let i be this position and let γ be the assignment $\mu[x_{\ell+1} \mapsto i]$ of $\{x_1, \dots, x_{\ell+1}\}$ in w . It follows immediately from the definitions that $u = [w]_\gamma$. Since $u \in L_{\ell+1,\psi}$, it follows that $w, \gamma \models \psi$, which exactly says that $w, \mu \models \varphi$ since $\varphi = \exists x_{\ell+1} \psi$ and $\gamma = \mu[x_{\ell+1} \mapsto i]$.

Conversely, assume that $w, \mu \models \varphi$. It follows that there exists a position i in w such that $w, \mu[x_{\ell+1} \mapsto i] \models \psi$. Let $\gamma = \mu[x_{\ell+1} \mapsto i]$. By definition of $L_{\ell+1,\psi}$, we have $[w]_\gamma \in L_{\ell+1,\psi}$. Clearly, $[w]_\gamma$ is good and therefore, we have $[w]_\mu = \pi_{1,\dots,\ell,A}([w]_\gamma) \in L_{\ell,\varphi}$. This concludes the proof. \square

It remains to prove that $L_{\ell,\varphi} \in \mathcal{C}_\ell[n + \frac{1}{2}]$. The argument is based on the next lemma.

Lemma 79 (Splitting lemma) *Let \mathcal{C} be a quotienting lattice of regular languages over A and let $B \subseteq A$. Consider a language $L \in \mathcal{C}$. Then, $L \cap A^*BA^*$ is a finite union of languages of the form PbS where $b \in B$ and $P, S \in \mathcal{C}$.*

Proof First observe that we may assume without loss of generality that B is a singleton. Indeed, we have $L \cap A^*BA^* = \bigcup_{b \in B} L \cap A^*bA^*$. Hence, it suffices to apply the lemma in the singleton case for each language $L \cap A^*bA^*$. Therefore, we now assume that B is a singleton $\{b\}$.

For any $u \in A^*$, let $Q_u = (ub)^{-1}L = \{v \in A^* \mid ubv \in L\}$. Consider the following language L' :

$$L' = \bigcup_{u \in A^*} \left(\bigcap_{v \in Q_u} L(bv)^{-1} \right) \cdot b \cdot (ub)^{-1}L \quad (10)$$

We claim that $L \cap A^*bA^* = L'$. Before we prove this equality, let us explain why it allows us to conclude the proof. Since $L \in \mathcal{C}$, we know by hypothesis on \mathcal{C} that L is regular. Hence, it follows from the Myhill-Nerode Theorem (Theorem 16) that there are finitely many quotients of L . In particular, this means that in (10), the union over all $u \in A^*$ and the intersections over all $v \in Q_u$ are actually finite. Moreover, since \mathcal{C} is a quotienting lattice, we obtain that for every $u \in A^*$,

$$\bigcap_{v \in Q_u} L(bv)^{-1} \in \mathcal{C} \quad \text{and} \quad (ub)^{-1}L \in \mathcal{C}.$$

Hence, this concludes the proof of Lemma 79: $L \cap A^*bA^*$ is a finite union of languages of the form PbS where $P, S \in \mathcal{C}$. It remains to prove that $L \cap A^*bA^* = L'$.

To prove that $L \cap A^*bA^* = L'$, assume first that $w \in L \cap A^*bA^*$. It follows that $w = ubv' \in L$ for some $u, v' \in A^*$. Hence, $v' \in (ub)^{-1}L$. Moreover, $u \in L(bv)^{-1}$ for any $v \in Q_u$ by definition. Hence, $u \in \bigcap_{v \in Q_u} L(bv)^{-1}$. We now conclude that,

$$w \in \left(\bigcap_{v \in Q_u} L(bv)^{-1} \right) \cdot b \cdot (ub)^{-1}L.$$

Therefore, $w \in L'$. We have proved that $L \cap A^*bA^* \subseteq L'$.

Conversely, assume that $w \in L'$. We obtain $u \in A^*$ such that w admits a decomposition $w = u'bv'$ with $u' \in \bigcap_{v \in Q_u} L(bv)^{-1}$ and $v' \in (ub)^{-1}L$. In particular, since $v' \in (ub)^{-1}L$, we have $ubv' \in L$ which means that $v' \in Q_u$ by definition. Combined with the fact that $u' \in \bigcap_{v \in Q_u} L(bv)^{-1}$, this yields $u' \in L(bv')^{-1}$, which exactly says that $w = u'bv' \in L \cap A^*bA^*$. \square

Recall that in order to finish off the proof of Proposition 76, it remains to show that the language $L_{\ell,\varphi}$, defined in Equation (9), belongs to $\mathcal{C}_\ell[n + \frac{1}{2}]$. Let $B \subseteq A_{\ell+1}$ be the set of all letters in $A_{\ell+1}$ whose component $\ell + 1$ is 1:

$$B = \{\bar{b} \in A_{\ell+1} \mid \pi_{\ell+1}(\bar{b}) = 1\}.$$

Note that by definition, every *good* word $u \in A_{\ell+1}^*$ belongs to $A_{\ell+1}^*BA_{\ell+1}^*$. Since by Proposition 38, $\mathcal{C}[n + \frac{1}{2}]$ is a quotienting lattice of regular languages, we may apply Lemma 79 to $L_{\ell+1,\psi} \in \mathcal{C}_{\ell+1}[n + \frac{1}{2}]$:

$$L_{\ell+1,\psi} \cap A_{\ell+1}^*BA_{\ell+1}^* = \bigcup_{j \leq m} P_j \bar{b}_j S_j \quad (11)$$

where for all $j \leq m$, $\bar{b}_j \in B$ and $P_j, S_j \in \mathcal{C}_{\ell+1}[n + \frac{1}{2}]$. For all $j \leq m$, let $\bar{c}_j = \pi_{1,\dots,\ell,A}(\bar{b}_j) \in A_\ell$. Recall that $\alpha_\ell : A_\ell^* \rightarrow A_{\ell+1}^*$ is defined as the following morphism. For every letter $(i_1, \dots, i_\ell, a) \in A_\ell$, we have $\alpha_\ell(i_1, \dots, i_\ell, a) = (i_1, \dots, i_\ell, 0, a) \in A_{\ell+1}$. We have the following fact.

Fact 80 $L_{\ell,\varphi} = \bigcup_{j \leq m} \alpha_\ell^{-1}(P_j) \bar{c}_j \alpha_\ell^{-1}(S_j)$.

Proof We first consider $v \in L_{\ell,\varphi}$. We exhibit $j \leq m$ such that $v \in \alpha_\ell^{-1}(P_j) \bar{c}_j \alpha_\ell^{-1}(S_j)$. By definition of $L_{\ell,\varphi}$, we get $u \in L_{\ell+1,\psi}$ which is good and such that $v = \pi_{1,\dots,\ell,A}(u)$. Since u is good, we have,

$$u \in L_{\ell+1,\psi} \cap A_{\ell+1}^*BA_{\ell+1}^*.$$

It then follows from (11) that we have $u \in P_j \bar{b}_j S_j$ for some $j \leq m$. Hence, we may decompose u as $u = u_1 \bar{b}_j u_2$ with $u_1 \in P_j$ and $u_2 \in S_j$. Therefore, we have,

$$v = \pi_{1,\dots,\ell,A}(u) = \pi_{1,\dots,\ell,A}(u_1 \bar{b}_j u_2) = \pi_{1,\dots,\ell,A}(u_1) \bar{c}_j \pi_{1,\dots,\ell,A}(u_2).$$

Finally, since $u = u_1 \bar{b}_j u_2$ is good and $\bar{b}_j \in B$, we know that \bar{b}_j is the only letter in u whose component $\ell + 1$ is equal to 1. Hence, the component $\ell + 1$ of any letter in u_1 or u_2 is 0. By definition of α_ℓ , it follows that $\alpha_\ell(\pi_{1,\dots,\ell,A}(u_1)) = u_1$ and $\alpha_\ell(\pi_{1,\dots,\ell,A}(u_2)) = u_2$. Thus, since $u_1 \in P_j$ and $u_2 \in S_j$, we get $\pi_{1,\dots,\ell,A}(u_1) \in \alpha_\ell^{-1}(P_j)$ and $\pi_{1,\dots,\ell,A}(u_2) \in \alpha_\ell^{-1}(S_j)$. Finally, this yields $v \in \alpha_\ell^{-1}(P_j) \bar{c}_j \alpha_\ell^{-1}(S_j)$ which concludes this direction of the proof.

Conversely, assume that $v \in \alpha_\ell^{-1}(P_j) \bar{c}_j \alpha_\ell^{-1}(S_j)$ for some $j \leq m$. We have to prove that $v \in L_{\ell,\varphi}$. By hypothesis, we have $v = v_1 \bar{c}_j v_2$ with $v_1 \in \alpha_\ell^{-1}(P_j)$ and $v_2 \in \alpha_\ell^{-1}(S_j)$. Consider the following word $u \in A_{\ell+1}^*$:

$$u = \alpha_\ell(v_1) \bar{b}_j \alpha_\ell(v_2) \in P_j \bar{b}_j S_j.$$

Observe that by definition, u is good and $v = \pi_{1,\dots,\ell,A}(u)$. Moreover, it follows from (11) that $u \in L_{\ell+1,\psi}$. Thus, $v \in L_{\ell,\varphi}$ by definition of $L_{\ell,\varphi}$. \square

Fact 80 concludes the proof since it is immediate from Fact 75 that for all $j \leq m$, $\alpha_\ell^{-1}(P_j)$ and $\alpha_\ell^{-1}(S_j)$ both belong to $\mathcal{C}_\ell[n + \frac{1}{2}]$. Thus, we obtain Fact 80 that the language $L_{\ell,\varphi}$ is a finite union of marked concatenations of languages in $\mathcal{C}_\ell[n + \frac{1}{2}]$ and therefore belongs to $\mathcal{C}_\ell[n + \frac{1}{2}]$ itself. \square

7.3 Back to the dot-depth and Straubing-Thérien hierarchies

We now apply Theorem 73 to the two classical examples: the dot-depth and Straubing-Thérien hierarchies. Note that the logical characterization of the dot-depth hierarchy is historically the first result of this kind which was discovered by Thomas [52]. Therefore, Theorem 73 is a generalization of this original result.

Recall that the basis of the dot-depth hierarchy is $DD[0] = \{\emptyset, \{\varepsilon\}, A^+, A^*\}$. It turns out that the associated variant of first-order logic, $FO(DD[0])$, is exactly $FO(<, +1, \min, \max, \varepsilon)$. Indeed, according to Theorem 73 the predicates available in $FO(DD[0])$ are as follows:

1. The label predicates.
2. The predicate I_{A^*} , which by definition is equivalent to the order predicate $<$.
3. The predicates $P_{A^*}, S_{A^*}, N_{A^*}$ which always hold, hence they are equivalent to \top .
4. The predicates $I_\emptyset, P_\emptyset, S_\emptyset, N_\emptyset$ which never hold, hence they are equivalent to \perp .
5. The predicate I_ε , which by definition is equivalent to the successor predicate $+1$.
6. The predicate I_{A^+} , and by definition I_{A^+} is equivalent to $\neg I_\varepsilon$.

Thus, the only useful predicates in $FO(DD[0])$ are exactly those that are available in $FO(<, +1, \min, \max, \varepsilon)$: the label predicates, the linear order predicate, and the successor predicate. Therefore, we re-obtain Theorem 12 as a corollary of Theorem 73.

On the other hand, the basis of the Straubing-Thérien hierarchy is $ST[0] = \{\emptyset, A^*\}$, so that we miss the predicates I_ε and I_{A^+} . Therefore, Theorem 13 is also a corollary of Theorem 73.

8 Conclusion

In this paper, we surveyed 50 years of progress in the understanding of concatenation hierarchies. We presented a new proof that closure under intersection is implied by polynomial closure if the class we start from is a quotienting lattice. We then established that if level 0 is a finite quotienting Boolean algebra, then the corresponding hierarchy is strict and we stated that its levels $\frac{1}{2}, 1, \frac{3}{2}$ have decidable separation. We proved a result transferring decidability of separation for some level to decidability of membership for the next half level, entailing that level $\frac{5}{2}$ has decidable membership for finitely based hierarchies. We also observed that in the Straubing-Thérien hierarchy, level $q \geq 1$ ($q \in \mathbb{N}$ or $q \in \frac{1}{2} + \mathbb{N}$) coincides with level $q - 1$ in a concatenation hierarchy whose basis is finite, hence we got decidability of separation for levels $\frac{1}{2}, 1, \frac{3}{2}, 2$ and $\frac{5}{2}$ in the Straubing-Thérien hierarchy, and decidability of membership for level $\frac{7}{2}$. We then transferred all these results to the dot-depth hierarchy via a generic construction. Finally, we proved a generic logical definition of concatenation hierarchies encompassing the ones established for the dot-depth and the Straubing-Thérien hierarchies.

Some of the research directions following this work are obvious: it is desirable to generalize this approach to capture all levels of such a concatenation hierarchy. This seems however to be difficult. We would also want to test such techniques for other structures than words, for instance, trees. Another short-term interesting topic is to reprove and generalize results that were obtained in the particular case of the Straubing-Thérien hierarchy regarding *unambiguous polynomial closure*. We leave this question for future work.

References

- [1] J. Almeida. “Some algorithmic problems for pseudovarieties”. In: *Publicaciones Mathematicae Debrecen* 54 (1999). Proc. of Automata and Formal Languages, VIII, pp. 531–552.
- [2] M. Arfi. “Polynomial Operations on Rational Languages”. In: *STACS’87. Lect. Notes Comp. Sci.* Springer, 1987, pp. 198–206.
- [3] M. Arfi. “Opérations Polynomiales et Hiérarchies de Concaténation”. In: *Theoret. Comp. Sci.* 91.1 (1991), pp. 71–84.
- [4] J. Berstel, D. Perrin, and Ch. Reutenauer. *Codes and Automata*. Vol. 129. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2009.
- [5] M. Bojanczyk. “Star Height via Games”. In: *LICS’15*. IEEE Computer Society, 2015, pp. 214–219.
- [6] J. A. Brzozowski. “Developments in the Theory of regular Languages”. In: *IFIP Congress*. 1980, pp. 29–40.
- [7] J. A. Brzozowski. “Open problems about regular languages”. In: *Formal Language Theory*. Ed. by R. V. Book. Academic Press, 1980, pp. 23–47.
- [8] J. A. Brzozowski and R. S. Cohen. “Dot-Depth of Star-Free Events”. In: *J. Comput. System Sci.* 5.1 (1971), pp. 1–16.
- [9] J. A. Brzozowski and R. Knast. “The Dot-Depth Hierarchy of Star-Free Languages is Infinite”. In: *J. Comput. System Sci.* 16.1 (1978), pp. 37–55.
- [10] T. Colcombet. “Factorization Forests for Infinite Words and Applications to Countable Scattered Linear Orderings”. In: *Theoretical Computer Science* 411.4-5 (2010), pp. 751–764.
- [11] F. Dejean and M. P. Schützenberger. “On a Question of Eggan”. In: *Information and Control* 9.1 (1966), pp. 23–25.
- [12] L. C. Eggan. “Transition Graphs and the Star-Height of Regular Events.” In: *Michigan Math. J.* 10.4 (1963), pp. 385–397.
- [13] M. Gehrke, S. Grigorieff, and J.-E. Pin. “A topological approach to recognition”. In: *Proc. of ICALP’10*. Lect. Notes Comp. Sci. Springer, 2010.
- [14] C. Glaßer and H. Schmitz. “Languages of Dot-Depth $3/2$ ”. In: *Theory of Computing Systems* 42.2 (2007), pp. 256–286.
- [15] K. Hashiguchi. “Representation theorems on regular languages”. In: *J. Comput. System Sci.* 27.1 (1983), pp. 101–115.
- [16] K. Hashiguchi. “Algorithms for Determining Relative Star Height and Star Height”. In: *Information and Computation* 78.2 (1988), pp. 124–169.
- [17] D. Kirsten. “Distance Desert Automata and the Star Height Problem”. In: *RAIRO-Theor. Inf. Appl.* 39.3 (2005), pp. 455–509.
- [18] S. C. Kleene. “Representation of Events in Nerve Nets and Finite Automata”. In: *Annals of Mathematics Studies* 34. Ed. by C. Shannon and J. McCarthy. New Jersey: Princeton University Press, 1956, pp. 3–41.
- [19] R. Knast. “A Semigroup Characterization of Dot-Depth One Languages”. In: *RAIRO - Theoretical Informatics and Applications* 17.4 (1983), pp. 321–330.
- [20] M. Kufleitner. “The Height of Factorization Forests”. In: *Proceedings of the 33rd International Symposium on Mathematical Foundations of Computer Science*. MFCS’08. 2008, pp. 443–454.
- [21] R. McNaughton and S. A. Papert. *Counter-Free Automata*. MIT Press, 1971.

- [22] A. Nerode. “Linear Automaton Transformations”. In: *Proceedings of the American Mathematical Society* 9.4 (1958), pp. 541–544.
- [23] D. Perrin and J.-É. Pin. “First-Order Logic and Star-Free Sets”. In: *J. Comput. System Sci.* 32.3 (1986), pp. 393–406.
- [24] J.-É. Pin. “A variety theorem without complementation”. In: *Russian Mathem. (Iz. VUZ)* 39 (1995), pp. 74–83.
- [25] J.-É. Pin. “An Explicit Formula for the Intersection of two Polynomials of Regular Languages”. In: *DLT 2013*. Vol. 7907. Lect. Notes Comp. Sci. Springer, 2013, pp. 31–45.
- [26] J.-É. Pin. “Mathematical Foundations of Automata Theory”. In preparation. 2016. URL: <https://www.irif.fr/~jep/MPRI/MPRI.html>.
- [27] J.-É. Pin. “Open problems about regular languages, 35 years later”. In: *The Role of Theory in Computer Science. Essays Dedicated to Janusz Brzozowski*. World Scientific, 2017.
- [28] J.-É. Pin and H. Straubing. “Monoids of Upper Triangular Boolean Matrices”. In: *Semigroups. Structure and Universal Algebraic Problems*. Vol. 39. North-Holland, 1985, pp. 259–272.
- [29] J.-É. Pin and P. Weil. “Polynomial Closure and Unambiguous Product”. In: *ICALP’95*. Lect. Notes Comp. Sci. Springer, 1995, pp. 348–359.
- [30] J.-É. Pin and P. Weil. “Polynomial Closure and Unambiguous Product”. In: *Theory of Computing Systems* 30.4 (1997), pp. 383–422.
- [31] J.-É. Pin and P. Weil. “The Wreath Product Principle for Ordered Semigroups”. In: *Communications in Algebra* 30.12 (2002), pp. 5677–5713.
- [32] T. Place. “Separating Regular Languages with Two Quantifier Alternations”. In: *LICS’15*. 2015, pp. 202–213.
- [33] T. Place. “Separating Regular Languages with Two Quantifiers Alternations”. To appear. For a preliminary version, see <https://arxiv.org/abs/1707.03295>. 2018.
- [34] T. Place and M. Zeitoun. “Going Higher in the First-Order Quantifier Alternation Hierarchy on Words”. In: *ICALP’14*. Lect. Notes Comp. Sci. Springer, 2014, pp. 342–353.
- [35] T. Place and M. Zeitoun. “Separation and the Successor Relation”. In: *STACS’15*. Lipics, Leibniz-Zentrum fuer Informatik, 2015, pp. 662–675.
- [36] T. Place and M. Zeitoun. “The Covering Problem: A Unified Approach for Investigating the Expressive Power of Logics”. In: *MFCS’16*. Lipics, Leibniz-Zentrum fuer Informatik, 2016, 77:1–77:15.
- [37] T. Place and M. Zeitoun. “Adding successor: A transfer theorem for separation and covering”. In: *CoRR* abs/1709.10052 (2017). URL: <http://arxiv.org/abs/1709.10052>.
- [38] T. Place and M. Zeitoun. “Concatenation Hierarchies: New Bottle, Old Wine”. In: *CSR 2017*. Ed. by P. Weil. Vol. 10304. Lect. Notes Comp. Sci. Springer, 2017, pp. 25–37.
- [39] T. Place and M. Zeitoun. “Going Higher in First-Order Quantifier Alternation Hierarchies on Words”. In: *CoRR* abs/1707.05696 (2017). URL: <http://arxiv.org/abs/1707.05696>.
- [40] T. Place and M. Zeitoun. “Separation for dot-depth two”. In: *LICS 2017*. IEEE Computer Society, 2017, pp. 1–12.
- [41] T. Place and M. Zeitoun. “The Covering Problem”. In: *CoRR* abs/1707.03370 (2017). URL: <http://arxiv.org/abs/1707.03370>.
- [42] J. Sakarovitch. *Elements of Automata Theory*. Cambridge University Press, 2009.

-
- [43] M. P. Schützenberger. “On Finite Monoids Having Only Trivial Subgroups”. In: *Information and Control* 8.2 (1965), pp. 190–194.
- [44] I. Simon. “Hierarchies of Events of Dot-Depth One”. PhD thesis. University of Waterloo, Canada, 1972.
- [45] I. Simon. “Piecewise Testable Events”. In: *Proceedings of the 2nd GI Conference on Automata Theory and Formal Languages*. Springer, 1975, pp. 214–222.
- [46] I. Simon. “Factorization Forests of Finite Height”. In: *Theoret. Comp. Sci.* 72.1 (1990), pp. 65–94.
- [47] B. Steinberg. “A delay theorem for pointlikes”. In: 63.3 (2001), pp. 281–304.
- [48] H. Straubing. “A Generalization of the Schützenberger Product of Finite Monoids”. In: *Theoret. Comp. Sci.* 13.2 (1981), pp. 137–150.
- [49] H. Straubing. “Finite Semigroup Varieties of the Form $V * D$ ”. In: *Journal of Pure and Applied Algebra* 36 (1985), pp. 53–94.
- [50] D. Thérien. “Classification of Finite Monoids: The Language Approach”. In: *Theoret. Comp. Sci.* 14.2 (1981), pp. 195–208.
- [51] D. Thérien. “The Power of Diversity”. In: *Descriptive Complexity of Formal Systems*. Ed. by M. Holzer, M. Kutrib, and G. Pighizzini. Vol. 6808. Lect. Notes Comp. Sci. Springer, 2011, pp. 43–54.
- [52] W. Thomas. “Classifying Regular Events in Symbolic Logic”. In: *J. Comput. System Sci.* 25.3 (1982), pp. 360–376.
- [53] W. Thomas. “An application of the Ehrenfeucht-Fraïssé game in formal language theory”. In: *Mémoires de la Société Mathématique de France* 16 (1984), pp. 11–21.
- [54] W. Thomas. “A concatenation game and the dot-depth hierarchy”. In: *Computation Theory and Logic*. Berlin, Heidelberg: Springer, 1987, pp. 415–426.