Classes of Languages Generated by the Kleene Star of a Word

Laure Daviaud[1] and Charles Paperman[2(✉)]

[1] LIF, UMR7279, CNRS, Aix-Marseille Université, Marseille, France
[2] Warsaw University, Warsaw, Poland
charles.paperman@gmail.com

Abstract. In this paper, we study the lattice and the Boolean algebra, possibly closed under quotient, generated by the languages of the form u^*, where u is a word. We provide effective equational characterisations of these classes, i.e. one can decide using our descriptions whether a given regular language belongs or not to each of them.

1 Introduction

Equational descriptions of regular languages is a successful and long-standing approach to obtain characterisations of classes of regular languages. One of the first results about equational descriptions is Schützenberger's theorem [10] on star-free languages. In the case of a variety of regular languages, Reiterman's theorem [9] guarantees the existence of a characteristic set of profinite equations. This theorem has been extended to several kinds of classes of languages, including lattices and Boolean algebras. The reader could refer to [3,6] for a more detailed presentation. Let $\mathcal{U}$ be the class of all languages of the form u^*, where u is a word. The aim of this paper is to study the four classes of regular languages $\mathcal{L}$, $\mathcal{B}$, $\mathcal{L}q$ and $\mathcal{B}q$ obtained respectively as the closure of $\mathcal{U}$ under the following operations: finite union and finite intersection (lattice operations) for $\mathcal{L}$, finite union, finite intersection and complement (Boolean operations) for $\mathcal{B}$, lattice operations and quotients for $\mathcal{L}q$ and Boolean operations and quotients for $\mathcal{B}q$.

Our main result is an equational characterisation for each of these four classes. These equational characterisations being effective, they give as a counterpart the decidability of the membership problem: One can decide whether a given regular language belongs to $\mathcal{L}$, $\mathcal{B}$, $\mathcal{L}q$ and $\mathcal{B}q$ respectively. In addition to describing $\mathcal{L}$, $\mathcal{B}$, $\mathcal{L}q$ and $\mathcal{B}q$ in terms of equations, our results also provide a general form for the languages belonging to each of these classes.

Motivations. Our motivation for the study of these classes are threefold. First, Restivo suggested a few years ago to characterise the variety of languages generated by the languages of the form u^*, where u is a word. Given that a variety of languages is a class of regular languages closed under Boolean operations,

The second author is supported by WCMCS.

G.F. Italiano et al. (Eds.): MFCS 2015, Part I, LNCS 9234, pp. 167–178, 2015.
DOI: 10.1007/978-3-662-48057-1_13

quotients and inverses of morphisms, our result can be viewed as a first step towards the solution of Restivo's problem.

Our second reason for studying these classes was to provide non trivial applications of the equational theory of regular languages as defined by Gehrke, Grigorieff and Pin in [3,6]. There are indeed plenty of examples of known equational characterisations of varieties of languages, but not so much of classes of languages that are not closed under inverses of morphisms or under quotients.

Our third motivation is rather a long term perspective since it has to do with the (generalised) star-height problem, a long standing open problem on regular languages [7]. It appears that a key step towards this problem would be to characterise the Boolean algebra generated by the languages of the form F^*, where F is a finite language. The case $F = \{u\}$ studied in this paper is certainly a very special case, but it gives an insight into the difficulty of the general problem.

Related Work. A related class is the class of *slender languages* [4,11], which can be written as a finite union of languages of the form xu^*y, where $x, u, y \in A^*$. The class of *slender or full languages* is a lattice closed under quotients that is therefore characterised by a set of equations. These equations correspond in fact to patterns that cannot be found in any minimal automaton that computes a slender language. In our case, equations provided to characterise classes $\mathcal{L}$, $\mathcal{B}$, $\mathcal{L}q$ and $\mathcal{B}q$ can also be seen as forbidden patterns in automata. Then, we deduce normal forms for the languages in $\mathcal{L}$, $\mathcal{B}$, $\mathcal{L}q$ and $\mathcal{B}q$.

Organization of the Paper. Section 2 gives classical definitions and properties about the algebraic automata theory and profinite semigroups. Section 3 is dedicated to the study of the syntactic monoid of u^* for a given word u. In particular, we present useful algebraic properties of the syntactic monoid of u^*. Section 4 presents equational theory of regular languages: it first gives classical results, then presents the equations satisfied by u^*, and finally gives the characterisations of $\mathcal{L}$, $\mathcal{L}q$ and $\mathcal{B}q$. The study of $\mathcal{B}$ is much more intricate and involves specific tools that are given in Sect. 5. Finally, Sect. 6 presents decidability issues. Sections 2 to 6 deal with alphabet with at least two letters. The case of a unary alphabet is simpler and derives from the two-letter case. It is treated in Sect. 7.

Notations. We denote by A a finite alphabet with at least two letters, by A^* the set of words on A, by 1 the empty word and by $|u|$ the length of a word u.

2 Recognisability and the Profinite Monoid

In this section, we introduce the definitions of recognisability by monoids and of profinite monoid. For more details, the reader could refer to [2].

Monoids and Recognisability. A monoid M is a set equipped with a binary associative operation with a neutral element denoted by 1. The product of x and y is denoted by xy. An element e of M is idempotent if $e^2 = e$. An element

$0 \in M$ is a zero of M if for all $x \in M$, $0x = x0 = 0$. Given two monoids M and N, $\varphi : M \to N$ is a morphism if for all $x, y \in M$, $\varphi(xy) = \varphi(x)\varphi(y)$ and $\varphi(1) = 1$.

In a finite monoid, every element has an idempotent power: for all $x \in M$, there is $n_x \in \mathbb{N} - \{0\}$ such that x^{n_x} is idempotent. The smallest n_x satisfying this property is called the *index* of x. Moreover, there is an integer $n \neq 0$ such that for all $x \in M$, x^n is idempotent. For instance, one could take the product of the n_x. The smallest integer satisfying this property is called the *index* of the monoid and is denoted by ω. Thus, x^ω is the unique idempotent in the subsemigroup generated by x.

Given a monoid M and a morphism $\varphi : A^* \to M$, a language L is said to be *recognised* by (M, φ) if there is $P \subseteq M$ such that $L = \varphi^{-1}(P)$. The language L is said to be recognised by M if there is φ such that (M, φ) recognises L. A language is regular if and only if it is recognised by a finite monoid. Moreover, the smallest monoid that recognises a regular language L is unique up to isomorphism and is called the *syntactic monoid* of L. The associated morphism φ is called the *syntactic morphism* and $\varphi(L)$ is called the *syntactic image* of L. Furthermore, for each word u, we call $\varphi(u)$ the syntactic image of u with respect to L. The syntactic monoid of a regular language can be computed as it is the transition monoid of the minimal (deterministic) automaton of L.

Free Profinite Monoid. Given two words u and v, a monoid M separates u and v if there is a morphism $\varphi : A^* \to M$ such that $\varphi(u) \neq \varphi(v)$. If $u \neq v$, there is a finite monoid that separates u and v. A distance d can be defined on A^* as follows: $d(u, u) = 0$ and if $u \neq v$, $d(u, v) = 2^{-n}$ where n is the smallest size of a monoid that separates u and v. Moreover this distance is ultrametric.

Every finite monoid is seen as a metric space equipped with the distance $d(x, y) = 1$ if $x \neq y$ and $d(x, y) = 0$ otherwise. This implies that every morphism from A^* to a finite monoid is a uniformly continuous function.

We briefly recall some useful definitions and results on the *free profinite monoid*. We refer to [2] for an extended presentation of this subject. The free profinite monoid of A^*, denoted by $\widehat{A^*}$ can be defined as the completion for the distance d of A^*. It is a compact space such that A^* is a dense subset of $\widehat{A^*}$. Its elements are called *profinite words*. It is known that a language L is regular if and only if $\overline{L}$ is open and closed in $\widehat{A^*}$, where $\overline{L}$ is the topological closure of L in $\widehat{A^*}$.

Finally, every morphism from A^* to some finite monoid M can be uniquely extended to a uniformly continuous morphism from $\widehat{A^*}$ to M. By abuse of notation, a morphism and its extension will be denoted by the same symbol.

The two following examples are profinite words that are not finite words and that will be intensively used in the remainder of the paper.

Example 1. (Idempotent power). Given a word $u \in A^*$, the sequence $(u^{n!})_n$ converges in $\widehat{A^*}$. Its limit is denoted by u^ω. Given a finite monoid M and a morphism $\varphi : \widehat{A^*} \to M$, $\varphi(u^\omega) = \varphi(u)^\omega$.

Example 2. (Zero [1,8]*).* Let A be an alphabet with at least two letters and fix a total order on it. Let $(u_n)_n$ be the sequence of all words ordered by the induced shortlex order. We set: $v_0 = u_0$ and for all $n \in \mathbb{N}$, $v_{n+1} = (v_n u_{n+1} v_n)^{(n+1)!}$. The sequence $(v_n)_n$ converges in $\widehat{A^*}$ and we denote by ρ_A its limit. Given a finite monoid M and a morphism $\varphi : \widehat{A^*} \to M$, if M has a zero then $\varphi(\rho_A) = 0$.

3 The Languages u^*

As mentioned in the introduction, our goal is to describe classes generated by the languages u^*. We will see in Sect. 4 that proving the correctness of such characterisations requires a precise description of the structure of the syntactic monoid of a given language u^* and particularly of its idempotents.

Therefore, this section addresses this study by exhibiting some properties of the syntactic monoid of u^*. Let us introduce two notions useful to study the languages of the form u^*. A word u is said to be *primitive* if for all words v and all integers n, the condition $u = v^n$ implies $n = 1$ and $v = u$. A word v is said to be a *conjugate* of u if there are words u_1, u_2 such that $u = u_1u_2$ and $v = u_2u_1$. The study of the syntactic monoid of u^* highly depends on the fact whether u is primitive or not. Without loss of generality, we consider now the studied language to be of the form $(u^m)^*$ for u a primitive word and m a positive integer.

In the syntactic monoid of $(u^m)^*$, there is a zero that is the syntactic image of words that cannot be completed into a word of $(u^m)^*$. Idempotent elements are exactly this zero, the neutral element and the syntactic images of the conjugates of u^m. Thus there are $|u|+2$ idempotents. Moreover, if the idempotent power of a syntactic image of a word is not zero, then this word has to be a power of a conjugate of u. Finally, the index of the syntactic images of u and of its conjugates is m. All these properties are instanciated in Example 3.

Example 3. We show in Fig. 1 the minimal deterministic automaton and monoid representation of the language $(aab)^*$. The elements in boxes are the elements

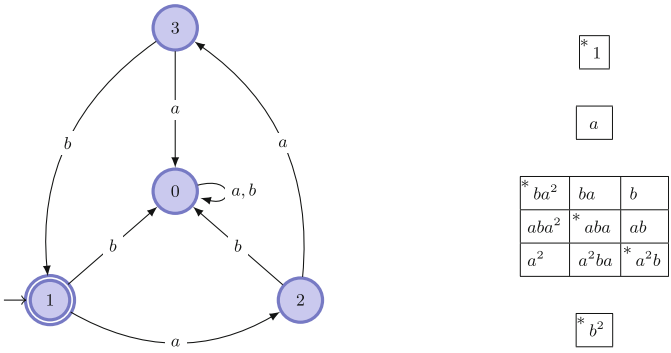

Fig. 1. Minimal deterministic automaton and monoid representation of $(aab)^*$

of the syntactic monoid of $(aab)^*$. An element has a star in its box if it is idempotent. The conjugates of aab are aab, aba and baa. Finally, the syntactic image of b^2 is a zero of the monoid.

4 Equational Characterisations of $\mathcal{L}$, $\mathcal{L}q$ and $\mathcal{B}q$

This section covers the equational theory of regular languages. First, Sect. 4.1 presents known results about equations. Then Sect. 4.2 applies this theory to the study of $\mathcal{L}$, $\mathcal{L}q$ and $\mathcal{B}q$, by giving equations that characterise them.

4.1 Equational Characterisations of Algebraic Structures of Regular Languages

A *lattice* (resp. a *Boolean algebra*) of languages of A^* is a class of languages containing the empty language $\emptyset$, the full language A^* and which is closed under finite union and finite intersection (resp. finite union, finite intersection and complement). A class of languages $\mathcal{L}$ is *closed under quotients* if for all $L \in \mathcal{L}$, for all $u \in A^*$, $u^{-1}L$ and Lu^{-1} belong to $\mathcal{L}$. Recall that $u^{-1}L = \{v \mid uv \in L\}$ and $Lu^{-1} = \{v \mid vu \in L\}$. Let u and v be two profinite words. A language $L \subseteq A^*$ satisfies the equation $u \to v$ if the condition $u \in \overline{L}$ implies $v \in \overline{L}$. It satisfies $u \leqslant v$ if for all words x, y, $xuy \in \overline{L}$ implies $xvy \in \overline{L}$. The notation $u \leftrightarrow v$ is a shortcut for $u \to v$ and $v \to u$ and similarly $u = v$ is a shortcut for $u \leqslant v$ and $v \leqslant u$. Observe that given a regular language L and its syntactic morphism $\varphi : A^* \to M$, the language L satisfies $u = v$ if and only if $\varphi(u) = \varphi(v)$ in M. A class of languages $\mathcal{L}$ is defined by a set of equations E if the following equivalence holds: $L \in \mathcal{L}$ if and only if L satisfies all the equations in E.

The kind of equations used to describe a class of languages is strongly related to its closure operations. The two following propositions formalise this statement.

Proposition 1. (Theorem 5.2 [3]). *A class of regular languages is defined by a set of equations of the form $u \to v$ (resp. $u \leftrightarrow v$) if and only if it is a lattice (resp. a Boolean algebra) of regular languages.*

Proposition 2. (Theorem 7.2 [3]). *A class of regular languages is defined by a set of equations of the form $u \leqslant v$ (resp. $u = v$) if and only if it is a lattice (resp. a Boolean algebra) of regular languages closed under quotients.*

Equations with zero. The existence of a zero in a syntactic monoid is given by the equations:

$$\rho_A x = x\rho_A = \rho_A$$

If these equations are satisfied, we will use the notation 0 instead of ρ_A. For example the set of equations:

$$\begin{cases} \rho_A \leqslant x \\ \rho_A x = x\rho_A = \rho_A \end{cases} \quad \text{is replaced by } 0 \leqslant x$$

The zero has been used to describe several classes of languages. For instance, the equations $0 \leqslant x$ for $x \in A^*$ describe exactly the so called *nondense* languages. Another example is the class of slender or full languages defined in the introduction [4,11]. This class of languages is a lattice closed under quotients; it is described by the following equations:

$$0 \leqslant x \text{ for } x \in A^*$$
$$x^\omega uy^\omega = 0 \text{ for } x, y \in A^+,\ u \in A^* \text{ and } i(uy) \neq i(x)$$

where $i(v)$ is the first letter of v for any $v \in A^+$ [5].

4.2 Characterisations of $\mathcal{L}$, $\mathcal{L}q$ and $\mathcal{B}q$

We give here a list of equations used in the study of $\mathcal{L}$, $\mathcal{L}q$, $\mathcal{B}q$ and $\mathcal{B}$. The proofs of the characterisations of these classes by some sets of equations are made in two steps. We first verify that the equations are correct and then check for their completeness. For the first step, it is sufficient to prove that for all words u, the language u^* satisfies the set of equations. From the nature of the equations $(\to, \leftrightarrow, \leqslant, =)$, we then obtain directly that the whole lattice, Boolean algebra and their closure under quotients satisfy the given set of equations. This step of correctness can be derived from the structure of the languages of the form u^*, presented in Sect. 3. The second step is to prove that only the languages in the desired structures satisfy the set of equations. This step is more intricate since it requires a full understanding of the combinatorics of the classes we consider. First we define the two following languages:

$$P_u = \bigcup_{p \text{ prefix of } u} u^*p \quad \text{and} \quad S_u = \bigcup_{s \text{ suffix of } u} su^*$$

The equations:

$$x^\omega y^\omega = 0 \text{ for } x, y \in A^* \text{ such that } xy \neq yx \qquad (E_1)$$
$$x^\omega y = 0 \text{ for } x, y \in A^* \text{ such that } y \notin P_x \qquad (E_2)$$
$$yx^\omega = 0 \text{ for } x, y \in A^* \text{ such that } y \notin S_x \qquad (E_3)$$
$$x^\omega \leqslant 1 \text{ for } x \in A^* \qquad (E_4)$$
$$0 \leqslant 1 \qquad (E_5)$$
$$x^\ell \leftrightarrow x^{\omega+\ell} \text{ for } x \in A^*,\ \ell > 0 \qquad (E_6)$$
$$x^\omega \to 1 \text{ for } x \in A^* \qquad (E_7)$$
$$x \to x^\ell \text{ for } x \in A^*,\ \ell > 0 \qquad (E_8)$$

Some equations are clearly satisfied by u^* such as equations (E_8) and (E_7). Indeed, if $v \in u^*$ then for all ℓ, v^ℓ is also a power of u and belongs to u^* (E_8). Similarly, 1 always belongs to u^* (E_7). Proving that u^* satisfies the other equation is more difficult and requires to analyse the structure of its syntactic monoid. In particular, the role of the idempotents is important. The following theorem gives the equational characterisations of $\mathcal{B}q$, $\mathcal{L}q$ and $\mathcal{L}$.

Theorem 1. *Over a finite alphabet with at least two letters:*

1. *The class $\mathcal{B}q$ is defined by equations (E_1), $(E_2)$and (E_3).*
2. *The class $\mathcal{L}q$ is defined by equations (E_1), (E_2), (E_3) and (E_4).*
3. *The class $\mathcal{L}$ is defined by equations (E_1), (E_4) and (E_8).*

To prove these characterisations we introduce a normal form for the languages in $\mathcal{B}q$, $\mathcal{L}q$ and $\mathcal{L}$. More precisely, we prove that a language that satisfies the sets of equations can be written in a normal form. Finally, normal forms imply membership in the classes $\mathcal{B}q$, $\mathcal{L}q$ or $\mathcal{L}$. We now sketch briefly the proofs.

We start with the most general class $\mathcal{B}q$ and then we restrict to the classes $\mathcal{L}q$ and $\mathcal{L}$ by adding sets of equations in the equational characterisation. Hence, let us start with $\mathcal{B}q$. First, we remark that the finite languages are in $\mathcal{B}q$, as for instance, the language $\{aab\}$. Indeed, $\{aab\} = a^{-1}(aaab)^* \cap (aab)^*$. Given a word u, and a non-negative integer r, we denote by $u^{\geqslant r}$ the language u^*u^r. Since this language can be rewritten as $u^* - \{1, u, \ldots, u^{r-1}\}$, it belongs to $\mathcal{B}q$. Similarly, by using the closure by quotient we capture the languages $u^{\geqslant r}p$ and $su^{\geqslant r}$ where p (resp. s) is a prefix (resp. a suffix) of u. Finally, the following normal form fully characterises the class $\mathcal{B}q$: if L is a nonfull language in $\mathcal{B}q$, then L can be written as

$$\left(\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i\right) \cup F \text{ or } \left(\left(\bigcup_{i=1}^{k} u_i^{\geqslant r_i} p_i\right) \cup F\right)^c$$

where $(u_i)_{i=1\ldots k}$ and F are finite sets of words, p_i is a prefix of u_i and $(r_i)_{i=1\ldots k}$ are integers. We have sketched the proof that all the languages that can be written in this normal form are in $\mathcal{B}q$. The difficult part is to prove that every regular language that satisfies the equations can be written in the normal form.

We can achieve the reduction from $\mathcal{B}q$ to $\mathcal{L}q$, that is removing the closure by complement, by adding the set of equations (E_4). in the equational characterisation. Furthermore, we obtain that the normal form is a restriction of the previous one: if $L \in \mathcal{L}q$ is nonfull, then

$$L = \left(\bigcup_{i=1}^{k} u_i^* p_i\right) \cup F$$

Remark 1. The proof is constructive: assuming that a language L satisfies the set of equations, one can compute the words and the integers giving the normal form.

Example 4. The language A^*aaA^* is not in $\mathcal{B}q$. Indeed, the first equation is not satisfied since the syntactic image of the words ab and b are idempotents, but the syntactic image of abb is not syntactically equal to 0. However, the language $A^*(aa + bb)A^*$ satisfies the three sets of equations and is therefore in $\mathcal{B}q$ but not in $\mathcal{L}$ since the set of equations (E_4) is not satisfied: the syntactic image of aa is 0, and by equation (E_4), $0 \leqslant 1$, so 1 should be in the language but that is not the case. We can even give the normal form of this language:

$$A^*(aa + bb)A^* = ((ab)^* \cup (ab)^*a \cup (ba)^* \cup (ba)^*b)^c$$

In order to study $\mathcal{L}$ and $\mathcal{B}$, we have to remove the "closure under quotients" from the characterisations above. We deal with these cases by introducing an intermediate Boolean algebra (resp. lattice) denoted by $\widetilde{\mathcal{B}}$ (resp. $\widetilde{\mathcal{L}}$). The latter classes are generated by the following languages, which correspond to a certain form of quotients:

$$\widetilde{\mathcal{U}} = \{(u^m)^* u^r \mid u \in A^*,\ m > 0,\ 0 \leqslant r < m\}$$

The study of these two classes is an intermediate step since:

$$\mathcal{B} \subseteq \widetilde{\mathcal{B}} \subseteq \mathcal{B}q \quad \text{and} \quad \mathcal{L} \subseteq \widetilde{\mathcal{L}} \subseteq \mathcal{L}q$$

Proposition 3. *Over a finite alphabet with at least two letters:*

1. *The class $\widetilde{\mathcal{B}}$ is defined by equations (E_1) and (E_6).*
2. *The class $\widetilde{\mathcal{L}}$ is defined by equations (E_1), (E_6), (E_5) and (E_7).*

From this proposition, we can see that the language presented in Example 4 $A^*(aa + bb)A^*$ is not in $\widetilde{\mathcal{B}}$, and therefore it is neither in $\mathcal{L}$ nor in $\mathcal{B}$, since the equation (E_6) is not satisfied. Indeed, it is sufficient to consider the word aba, and to remark that $(aba)^2aba \in A^*(aa+bb)A^*$ but $aba \notin A^*(aa+bb)A^*$. As for the preceding cases, the languages in $\widetilde{\mathcal{B}}$ and $\widetilde{\mathcal{L}}$ can be written in a normal form: if L is a nonfull language in $\widetilde{\mathcal{B}}$, then

$$L \cup \{1\} = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i} \quad \text{or} \quad L - \{1\} = \left(\bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i} \right)^c$$

Similarly, if L is a nonfull language in $\widetilde{\mathcal{L}}$, then $L = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i}$ where $(u_i)_{i=1,\dots,k}$ are words and m, $(r_i)_{i=1,\dots,k}$ are integers. Finally, we can characterise the classes $\mathcal{L}$ and $\mathcal{B}$ by restricting the set of integers r_i that can be obtained in the normal form of $\widetilde{\mathcal{L}}$ and $\widetilde{\mathcal{B}}$. Regarding $\mathcal{L}$, one can prove that the only possible choice for r_i is 0. Thus, a nonfull language L in $\mathcal{L}$ is of the form $L = \bigcup_{i=1}^{k} u_i^*$. Unlike the class $\mathcal{L}$, the case of $\mathcal{B}$ can not be deduced directly from the case of $\widetilde{\mathcal{B}}$ and it is much more complicated. It is the subject of the next section.

5 The Case of the Boolean Algebra $\mathcal{B}$

We enter here the most intricate part of the description of the classes generated by the languages of the form u^*. The idea is to restrict the possible integers r_i we can obtain in the description of $\widetilde{\mathcal{B}}$. For that, we will define equivalence relations over the integers. Once this will be done, the main difficulty will be to translate properties over integers into profinite equations. In order to do that, we will introduce profinite numbers. This issue is addressed in Sect. 5.1 that first defines which sets of integers are allowed for the r_i and then translates it into equations. Finally, Sect. 5.2 aggregates all these notions to give the characterisation of $\mathcal{B}$.

5.1 Equivalence Classes Over $\mathbb{N}$ and Profinite Numbers

Let m be an integer, and r and s be in $\{0, \ldots, m-1\}$, let us define $r \equiv_m s$ if and only if $\gcd(r, m) = \gcd(s, m)$. Remark that $\equiv_m$ is an equivalence relation. Intuitively, a language in $\mathcal{B}$ with m as the index of its syntactic monoid, will not be able to separate two integers that are equivalent with respect to $\equiv_m$. More precisely, let L be a language in $\mathcal{B}$ with m as the index of its syntactic monoid and $r \equiv_m s$. Then for all words u and for all k, k', we have $u^{km+r} \in L$ if and only if $u^{k'm+s} \in L$.

Example 5. We introduce the language $L = (a^2)^* - (a^6)^*$. This language is, by definition, in $\mathcal{B}$. The index of its syntactic monoid is 6. Classes for $\equiv_6$ are $\{1, 5\}$, $\{2, 4\}$ and $\{3\}$. Thus, L cannot separate a word in $(a^6)^*a^2$ from a word in $(a^6)^*a^4$. Therefore, since $(a^6)^*a^2$ is in L, $(a^6)^*a^4$ is also in L. Since L belongs to $\mathcal{B}$, it also belongs to $\widetilde{\mathcal{B}}$ and we have a convenient normal form given by Proposition 3:

$$L = (a^2)^* - (a^6)^* = (a^6)^*a^2 \cup (a^6)^*a^4.$$

The equivalence relation $\equiv_m$ allows to give the form of the languages in $\mathcal{B}$. The next step is to translate it in terms of equations. The difficulty comes from the fact that $\equiv_m$ depends on the parameter m that represents the index of the syntactic monoid of a given language. So, this cannot be directly translated into a set of equations that are supposed to not depend on a specific language.

Profinite Numbers. Consider a one-letter alphabet $B = \{a\}$ and the profinite monoid $\widehat{B^*}$. There is an isomorphism from B^* to $\mathbb{N}$ that associates a word to its length. Then there is a unique set $\widehat{\mathbb{N}}$ and a unique isomorphism $\psi : \widehat{B^*} \to \widehat{\mathbb{N}}$ such that $\mathbb{N} \subseteq \widehat{\mathbb{N}}$ and $\widehat{\psi}$ coincides with ψ on $\mathbb{N}$. Elements of $\widehat{\mathbb{N}}$ are called *profinite numbers*. They are limits of sequences of integers, in the sense of the topology of the set of words on a one-letter alphabet. Given a word u, and a profinite number α, u^α corresponds to the profinite word that is the limit of the words u^{α_n} where $(\alpha_n)_n$ is a sequence of integers converging to α.

Let $\mathcal{P} = \{p_1 < p_2 < \ldots < p_n < \ldots\}$ be a cofinite sequence of prime numbers. That is, a sequence of prime numbers such that only a finite number of prime numbers are not used in the sequence. Consider the sequence defined by $z_n^{\mathcal{P}} = (p_1 \cdots p_n)^{n!}$. The sequence $(z_n^{\mathcal{P}})_{n>0}$ is converging in $\widehat{\mathbb{N}}$ and we denote by $z^{\mathcal{P}}$ its limit.

We can give now the last set of equations needed to characterise $\mathcal{B}$ and that conveys the notion of equivalence over $\mathbb{N}$ defined above. Denote by Γ the set of pairs of profinite numbers $(dz^{\mathcal{P}}, dpz^{\mathcal{P}})$ satisfying the three following conditions:

- $\mathcal{P}$ is a cofinite sequence of prime numbers,
- $p \in \mathcal{P}$,
- if q divides d then $q \notin \mathcal{P}$.

Let us define the set of equations (E_9) by:

$$x^\alpha \leftrightarrow x^\beta \text{ for all } (\alpha, \beta) \in \Gamma \qquad (E_9)$$

5.2 Characterisation of $\mathcal{B}$

The following result combines the notions given in Sect. 5.1 and characterises the class $\mathcal{B}$.

Theorem 2. *Over a finite alphabet with at least two letters, the class $\mathcal{B}$ is defined by equations (E_1), (E_6) and (E_9).*

Sketch of the Proof. Firstly, we prove that u^* satisfies (E_1), (E_6) and (E_9). For (E_9), essentially, $dpz^{\mathcal{P}}$ is a multiple of $dz^{\mathcal{P}}$ so u^* satisfies $x^{dz^{\mathcal{P}}} \to x^{dpz^{\mathcal{P}}}$. Conversely, thanks to the definition of Γ, for n large enough, $dz^{\mathcal{P}}_{n+1}$ is a multiple of $dpz^{\mathcal{P}}_n$ and thus u^* satisfies $x^{dpz^{\mathcal{P}}} \to x^{dz^{\mathcal{P}}}$.

The reverse implication is proved in two steps. First, we prove that if a nonfull language L satisfies (E_1), (E_6) and (E_9), then just like for the other classes, it has a normal form:

$$L \cup \{1\} = \bigcup_{i=1}^{k} \bigcup_{r \in S_i} (u_i^m)^* u_i^r \quad \text{or} \quad \big(L - \{1\}\big)^c = \bigcup_{i=1}^{k} \bigcup_{r \in S_i} (u_i^m)^* u_i^r$$

where m is an integer, $(u_i)_{i=1,\dots,k}$ is a finite set of words, and S_i is an equivalence class of $\equiv_{m,\cdot}$. We start by using the first part of Proposition 3 to prove that L belongs to $\widetilde{\mathcal{B}}$. So L can be written as:

$$L \cup \{1\} = \bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i} \quad \text{or} \quad L - \{1\} = \Big(\bigcup_{i=1}^{k} (u_i^m)^* u_i^{r_i}\Big)^c$$

We prove that for all $t \equiv_m r$, u^r belongs to L if and only if u^t belongs to L. The idea is the following: Let φ be the syntactic morphism of L, consider any cofinite sequence of prime numbers $\mathcal{P}$. If all the prime divisors of m are in $\mathcal{P}$, then for all n large enough, m divides $z^{\mathcal{P}}_n$ and thus for all words x, $\varphi(x^{dz^{\mathcal{P}}}) = \varphi(x^{\omega}) = \varphi(x^{dpz^{\mathcal{P}}})$. If none of the prime divisors of m is in $\mathcal{P}$, then for all n large enough, $z^{\mathcal{P}}_n$ is of the form $km + 1$. Then $\varphi(x^{dz^{\mathcal{P}}}) = \varphi(x^{\omega+d})$ and $\varphi(x^{dpz^{\mathcal{P}}}) = \varphi(x^{\omega+dp})$. Finally, d and dp under the conditions that define the set Γ, represent integers in the same equivalence class with respect to m that are then linked by (E_9). Other situations are combinations of these two.

Once we have the normal form for L, what is left is to prove that a language that can be written in this normal form belongs to $\mathcal{B}$. This is done by proving that:

$$\bigcup_{p \in \overline{r}^m} (u^m)^* u^p = (u^d)^* - \bigcup_{\substack{k \text{ s.t.} \\ 0 \leqslant k \leqslant m \\ gcd(k, \frac{m}{d}) \neq 1}} (u^{kd})^*$$

where $\overline{r}^m$ is the equivalence class of r for $\equiv_m$ and $d = \gcd(m, r)$.

6 Decidability

The characterisations that are given in Theorems 1 and 2 yield as a counterpart the decidability of the classes $\mathcal{B}q$, $\mathcal{L}q$, $\mathcal{L}$ and $\mathcal{B}$: given a regular language L, one can decide if L belongs to said classes. Every single equation is effectively testable. The main issue is to test an infinite set of equations in finite time. The idea is to test the equations in the syntactic monoid of L that is finite and thus test a finite number of equations. The first step is to compute M, the syntactic monoid of L, m its index, φ the syntactic morphism and P, the syntactic image of L. They are all computable from the minimal automaton of L. Then, it is sufficient to check if the sets of equations are satisfied directly in M and P, which are finite. More precisely:

(E_4): For all x, y, $z \in M$, $yx^m z \in P \Rightarrow yz \in P$
(E_5): Particular case of (E_4)
(E_6): For all $x \in M$, for all $0 < \ell < m$, $x^\ell \in P \Leftrightarrow x^{m+\ell} \in P$
(E_7): Particular case of (E_4)
(E_8): For all $x \in M$, for all $0 < \ell \leqslant 2m$, $x \in P \Rightarrow x^\ell \in P$
(E_9): Thanks to the notion of equivalence classes given in Sect. 5.1, testing equations in (E_9) is the same as testing that for all $x \in M$, for all $0 \leqslant r, s < m$ such that $r \equiv_m s$, $x^r \in P \Leftrightarrow x^s \in P$.

It is much more difficult to translate sets of equations (E_1), (E_2) and (E_3) in M since conditions"$xy \neq yx$","$y \notin P_x$"and"$y \notin S_x$" cannot be translated directly in M.

(E_1): Consider $x, y \in M$ such that $x^m y^m \neq 0$. One has to check that for all words $u \in \varphi^{-1}(x)$, $v \in \varphi^{-1}(y)$, $uv = vu$. This problem is decidable.
(E_2): Consider $x, y \in M$ such that $x^m y \neq 0$. One has to check that for all words $u \in \varphi^{-1}(x)$, $v \in \varphi^{-1}(y)$, $v \in P_u$. This problem is decidable.
(E_3): Same as (E_2)

7 The Case of a Unary Alphabet

This section summarises results for a unary alphabet. In this case, the syntactic monoid of a language of the form $(a^k)^*$ has no zero and even more the construction of ρ_A, given in [1,8] for larger alphabets, does not make sense for a singleton alphabet. But using the fact that a regular language on the alphabet $A = \{a\}$ is a finite union of languages of the form $(a^q)^* a^p$ for non negative integers p and q, we can derive from proofs made for the general case that $\mathcal{B}q$ is the set of all regular languages. The set $\mathcal{L}q$ is the set of the languages that are finite unions of languages of the form $(a^q)^* a^p$ with $p < q$ and is characterised by (E_4). The set $\mathcal{L}$ is the set of the languages that are finite unions of languages of the form $(a^q)^*$ and is characterised by (E_4) and (E_8). Finally, $\mathcal{B}$ is characterised by (E_6) and (E_9).

8 Conclusion

This paper offers an equational description of the lattice, Boolean algebra and their closure under quotients generated by the languages of the form u^*. These descriptions illustrate the power of the topological framework introduced by [3]. In particular, it gives us tools to describe in an effective way these classes of languages.

A lot of combinatorial phenomena have been understood and analysed to obtain these results. The next step could be to investigate either the case of the classes of languages generated by F^* where F is a finite set of words, or the case of the classes generated by $u_1^* u_2^* \dots u_k^*$ with $u_1, \dots, u_k$ some finite words. Each of these questions are interesting to have a better understanding of the phenomena that appear in the study of the variety generated by the languages u^* and of the generalised star-height problem.

References

1. Almeida, J., Volkov, M.V.: Profinite identities for finite semigroups whose subgroups belong to a given pseudovariety. J. Algebra Appl. **2**(2), 137–163 (2003)
2. Almeida, J., Weil, P.: Relatively free profinite monoids: an introduction and examples. In: Semigroups, Formal Languages and Groups (York, 1993), of NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., vol. 466, pp. 73–117. Kluwer Acad. Publ., Dordrecht (1995)
3. Gehrke, M., Grigorieff, S., Pin, J.É.: Duality and equational theory of regular languages. In: Aceto, L., Damgård, I., Goldberg, L.A., Halldórsson, M.M., Ingólfsdóttir, A., Walukiewicz, I. (eds.) ICALP 2008, Part II. LNCS, vol. 5126, pp. 246–257. Springer, Heidelberg (2008)
4. Honkala, J.: On slender languages. In: Paun, B., Rozenberg, G., Salomaa, A. (eds.)Current Trends in Theoretical Computer Science, pp. 708–716. World Scientific Publishing, River Edge (2001)
5. Pin, J.É.: Mathematical foundations of automata theory. http://www.liafa.jussieu.fr/~jep/PDF/MPRI/MPRI.pdf
6. Pin, J.É.: Profinite methods in automata theory. In: Albers, S., Marion, J.-Y. (eds.) 26th International Symposium on Theoretical Aspects of Computer Science (STACS 2009), pp. 31–50. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany (2009)
7. Pin, J.É., Straubing, H., Thérien, D.: Some results on the generalized star-height problem. Inf. Comput. **101**, 219–250 (1992)
8. Reilly, N.R., Zhang, S.: Decomposition of the lattice of pseudovarieties of finite semigroups induced by bands. Algebra Univers. **44**(3–4), 217–239 (2000)
9. Reiterman, J.: The Birkhoff theorem for finite algebras. Algebra Univers. **14**(1), 1–10 (1982)
10. Schützenberger, M.P.: On finite monoids having only trivial subgroups. Inf. Control **8**, 190–194 (1965)
11. Yu, S.: Regular languages. In: Rozenberg, G., Salomaa, A. (eds.) Handbook of Language Theory, vol. 1, chap. 2, pp. 679–746. Springer, Heidelberg (1997)