Languages of profinite words and the limitedness problem

Szymon Toruńczyk[*]

University of Warsaw

Abstract. We present a new, self-contained proof of the limitedness problem. The key novelty is a description using profinite words, which unifies and simplifies the previous approaches, and seamlessly extends the theory of regular languages. We also define a logic over profinite words, called MSO+inf and show that the satisfiability problem of MSO+$\mathbb{B}$ reduces to the satisfiability problem of our logic.

1 Introduction

This paper is an attempt to establish a natural framework for problems related to the limitedness problem. A notable example of such a problem is the decidability of the logic MSO+$\mathbb{B}$.

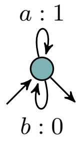


Fig. 1. A distance automaton over the input alphabet $\{a, b\}$.

The *limitedness problem* was introduced by Hashiguchi [8] on his way to solving the famous star height problem. In its basic form, it concerns *distance automata*, i.e. nondeterministic automata, whose transitions are additionally labeled by nonnegative, integer weights, such as the one depicted in Figure 1. A distance automaton is *limited* if there exists a bound n such that every accepted word has some accepting run whose sum of weights is bounded by n. Thus the *limitedness problem* is a decision problem which asks whether a given automaton is limited. The automaton in the example is not limited: the words $a, a^2, a^3, \ldots$ require accepting runs of ever larger weights.

The *logic MSO*+$\mathbb{B}$ was introduced by Bojańczyk in his dissertation (see also [2]) in relation with a problem concerning modal μ-calculus. It is an extension of the usual MSO logic – over infinite trees or words – by the quantifier $\mathbb{B}$, defined so that the formula $\mathbb{B}X.\varphi(X)$ holds if and only if all the sets of positions X satisfying the formula φ in the given model have a commonly bounded size. A typical language of infinite words defined in this logic is:

$$L_B = \{a^{n_1}ba^{n_2}b\ldots : \quad \text{the sequence } n_1, n_2, \ldots \text{ is bounded}\}.$$

Note that this language is not ω-regular, as its complement does not contain any ultimately periodic word. As a far-reaching project (see [3] for a survey),

[*] Author supported by ERC Starting Grant “Sosna”.

Bojańczyk posed the question of decidability of satisfiability of the logic MSO+$\mathbb{B}$ over infinite trees. Still, it is not even known to be decidable over infinite words.

A syntactic fragment of the logic MSO+$\mathbb{B}$ has been shown decidable in [4]. The key tool used in this paper is a model of automata called *ωB-automata*. Later, the authors discovered that limitedness of distance automata can be easily decided using their results concerning ωB-automata. The link with the limitedness problem has been exploited in [6], where Colcombet defined *B-automata* and developed his theory of regular cost functions and stabilization semigroups. B-automata directly generalize distance automata, by allowing more than one counter which, moreover, can be reset.

Our contribution is a theory which we believe to be the appropriate setting for considering limitedness of B-automata, and related problems. As a starting point, we see that B-automata naturally define languages of *profinite* words. The set of profinite words has a rich algebraic and topological structure, which we find very useful in the context of limitedness.

For instance, consider the distance automaton from Figure 1. There is a *profinite* word, denoted a^ω (not to be confused with the infinite word) which witnesses the fact that the automaton is *not* limited – this word can be defined as the limit of the sequence of finite words $(a^{n!})_{n=1}^{\infty}$. We say that this profinite word does *not* belong to the language of this automaton; the language of this automaton consists of profinite words which only have finitely many a's, such as b or $b^\omega a$.

We call the class of languages of profinite words defined by B-automata *B-regular languages*. Our main result states that this class can be characterized in terms of logic, regular expressions and semigroups. The result generalizes the main results of the papers [11, 13, 9, 1, 4], and implies the main result of [6, 7]. The description in terms of semigroups immediately implies decidability of the limitedness problem for B-automata, which, in our framework is simply the question of language universality. In particular, together with Kirsten's elegant reduction of the star height problem to the limitedness problem, our result gives yet another proof of decidability of the star height problem. The result also implies decidability of a more general problem – limitedness of Boolean combinations of B-automata. The remaining characterizations are primarily of conceptual value, as they manifest both that our framework is appropriate, and that the class of B-regular languages is robust. Note that most of these characterizations are also available in the framework of Colcombet. One exception is a new, finite-index characterization of B-regular languages, à la the Myhill-Nerode theorem; it seems that this result cannot be even phrased in the other frameworks.

Lastly, we show that our framework is suited for dealing with the satisfiability problem for MSO+$\mathbb{B}$ over infinite words – we prove that this problem can be reduced to the satisfiability problem of a new logic MSO+inf over profinite words, which we introduce here. This seems impossible in the other frameworks. In fact, our reduction is very general, and works for very many logics. The proof extends Büchi's ideas, and consists of two key ingredients: convergent Ramsey factorizations of infinite words, and a model of deterministic automata over infinite words with a profinite acceptance condition.

Related work. Several proofs of decidability of the limitedness problem exist [8, 11, 13, 9, 1, 6]. Our proof builds on ideas from all of these papers, and simplifies them greatly. Hashigushi's #-expressions acquire a new, concrete meaning in our framework, as simply defining profinite words. We extend Leung's insight of considering the compact topological semigroup of all matrices over the tropical semiring, to considering the profinite semigroup. Also, Leung introduced finite versions of his topological semigroups, which are predecessors of stabilization semigroups of Colcombet. The factorization forests of Simon play a key role in the main technical part of our proof. The proof of Kirsten applies to a model very similar to B-automata, but with a hierarchical constraint on the counter operations. Kirsten generalized Leung's proof, providing further instances of stabilization semigroups; however, the topological insights of Leung disappeared, as he no longer considered compact topological semigroups.

Colcombet used ideas from [4] and of Kirsten in [7], where he developed his theory of regular cost functions. In his theory, a B-automaton defines a *B-regular cost function* – an equivalence class of number-valued functions. These cost functions also have equivalent descriptions in terms of regular expressions, logic and semigroups. The crucial discovery of that paper is the tight two-way correspondence between stabilization semigroups (defined there) and B-automata. Still, the topological insights of Leung remained missing.

On a general level, and also on the level of proof structure, our approach resembles the approach of Colcombet. We outline the key differences. As we deal with languages which are subsets of a topological semigroup, many classical notions naturally lift to our setting – such as recognizable subsets, Myhill-Nerode equivalence, homomorphisms. In Colcombet's framework, cost functions are not sets, and have no apparent algebraic nor topological structure (they only have a lattice structure, corresponding to the lattice ordering of languages). Because of this, the natural notions mentioned above do not exist, or have non-obvious definitions – an example is the complex notion of *compatible mapping* [6], which corresponds to our ∞-*homomorphism.* Even the notion of a Boolean combination of cost functions is meaningless. As a result, cost functions are not well-suited for the study of the full logic MSO+$\mathbb{B}$. On a technical level, the proofs in [6, 7] deal with the relative notions of "big" vs. "small" values, and this relativity needs to be carefully controlled in the calculations and proofs. In our more abstract setting, we deal with the absolute notions of infinite vs. finite, and computations involve usual set-theoretic equalities.

Outline of the paper. First, we recall the definitions of B- and S-automata, and of profinite words. Next, we show how languages of profinite words can be defined using automata, regular expressions and logic. Then we present our main technical tool – recognition by homomorphisms. In Section 5, we state the central result. Finally, we show a link between languages of infinite words and of profinite words. Due to space limitations, many details are deferred to the appendix.

Acknowledgements. I would like to thank Mikołaj Bojańczyk and Thomas Colcombet for many useful comments.

2 Preliminaries

Let us fix a finite alphabet A; finite words are assumed to be elements of A. In the examples, we will more concretely assume the alphabet $A = \{a, b\}$. By $\mathbb{N}$ we denote $\{0, 1, 2, \ldots\}$, and by $\overline{\mathbb{N}}$ we denote $\mathbb{N} \cup \{\omega\}$. We treat $\overline{\mathbb{N}}$ as a compact metric space, in which $d(m, n) = |2^{-m} - 2^{-n}|$ (where $2^{-\omega} = 0$).

B-automata and S-automata (implicit in [4], defined in [6]) are nondeterministic automata over finite words, equipped with a finite number of counters. There are two counter operations available for each counter: *inc* increases the current value of the counter by 1 and *reset* sets the value to 0. A transition of a B- or S-automaton may trigger any sequence of operations on its counters. If the operation *reset* is performed in a run ρ on a counter which currently stores a value n, then we say that *n is a reset value* in the considered run ρ. The two models – B- and S-automata – differ in the semantics of the functions they define.

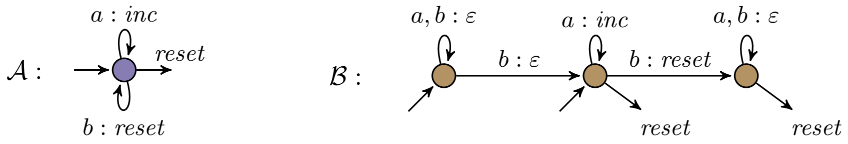
First, consider a B-automaton $\mathcal{A}$. Since $\mathcal{A}$ is nondeterministic, there might be many runs over a single word. For a particular run ρ, we define the *value* of ρ as its maximal reset value. Next, the *valuation* $f_{\mathcal{A}}(w)$ of an input word w under the automaton $\mathcal{A}$ is the minimum of the values of all accepting runs ρ over w:

$$f_{\mathcal{A}}(w) = \min_{\rho} \max\{n : \textit{ in the run } \rho\textit{, the value } n \textit{ is a reset value}\}.$$

Note that min ranges only over the accepting runs ρ of $\mathcal{A}$. We assume $\max(\emptyset) = 0$ and $\min(\emptyset) = \omega$, so if $\mathcal{A}$ has no accepting run over w, then $f_{\mathcal{A}}(w) = \omega$.

If $\mathcal{A}$ is an S-automaton, the definition of a valuation $f_{\mathcal{A}}(w)$ of an input word w is completely dual – simply swap min with max in the formula above.

Example 1 (The running example). Let $\mathcal{A}$ be the B-automaton with one counter which is depicted in the left-hand side of the figure below.



We declare that the automaton resets its counter after reading the entire word – this extra feature can be easily eliminated using nondeterminism. Then,

$$f_{\mathcal{A}}(w) = \max\{n_1, n_2, \ldots, n_k\} \qquad \textit{for } w = a^{n_1} b a^{n_2} \ldots b a^{n_k}.$$

Now consider the S-automaton $\mathcal{B}$ depicted in the right-hand side of the figure. It has one counter, which is also assumed to be reset at the end of the run. The reader can check that each accepting run of $\mathcal{B}$ over an input word w corresponds to a block of a's in w, and that $f_{\mathcal{B}}(w)$ is the length of the largest block of a's in w. Therefore, $f_{\mathcal{B}}$ and $f_{\mathcal{A}}$ are precisely the same function from A^+ to $\overline{\mathbb{N}}$.

Example 2. Let $\mathcal{A}$ be a finite nondeterministic automaton. If we view $\mathcal{A}$ as a B-automaton with no counters, the induced function assigns 0 to any word accepted by $\mathcal{A}$ and ω to any rejected word. Dually, if we treat $\mathcal{A}$ as an S-automaton, the induced function assigns ω to any accepted word, and 0 to any rejected word.

A B- or S-automaton is said to be *limited* if the function $f_{\mathcal{A}}$ has finite range (it may nevertheless contain the value ω). The *limitedness problem* for B- or S-automata is then to decide whether a given B- or S-automaton is limited. The automata in the example are not limited, since $f_{\mathcal{A}}(a^n) = n$ for any $n \in \mathbb{N}$.

Profinite words should be thought of as limits of sequences of finite words, with respect to *all* regular languages. A formal definition follows (see e.g. [12] for more details). We say that an infinite sequence $w_1, w_2, \ldots \in A^+$ of finite (nonempty) words *ultimately* belongs to the regular language $L \subseteq A^+$ if almost all the words $w_1, w_2, \ldots$ belong to L. We say that a sequence of words is *convergent*, if for any regular language L, the sequence ultimately belongs to L or ultimately belongs to the complement of L. Every constant sequence is convergent. The sequence $a, a^{2!}, a^{3!}, \ldots$ is also convergent, as follows from a pumping argument for regular languages. However, the sequence $a, a^2, a^3, \ldots$ is not convergent, since the regular language $(aa)^+$ only contains every other of its elements. Two convergent sequences are *equivalent* if they belong ultimately to precisely the same regular languages. In other words, interleaving one sequence with the other yields a convergent sequence. An equivalence class of convergent sequences is a *profinite word*. A profinite word is uniquely specified by the set of regular languages to which it ultimately belongs. For example, the equivalence class of the convergent sequence $a, a^{2!}, a^{3!}, \ldots$, which is a profinite word denoted a^ω, ultimately belongs to the languages $a^+, (aa)^+, (aaa)^+, \ldots$, and does not ultimately belong to the languages $a^* \cdot b \cdot a^*$ nor $a \cdot (aa)^+$. We denote profinite words by $x, y, \ldots$, and the set of all profinite words by $\widehat{A^+}$. We define $\widehat{A^*} = \widehat{A^+} \cup \{\varepsilon\}$, where ε is the *empty word*. Note that the set of finite words A^+ naturally embeds into the set of profinite words $\widehat{A^+}$, via constant convergent sequences. We call subsets of $\widehat{A^+}$ or of $\widehat{A^*}$ *languages of profinite words*.

The set of profinite words forms a semigroup: if $w_1, w_2, \ldots$ and $v_1, v_2, \ldots$ are two convergent sequences, then the sequence $w_1v_1, w_2v_2, \ldots$ is also convergent. There is another important operation on profinite words, called the *ω-power*. The ω-power of a convergent sequence $w_1, w_2, w_3, \ldots$ is the sequence $w_1^1, w_2^{2!}, w_3^{3!}, \ldots$, which also turns out to be convergent. This operation induces an operation $x \mapsto x^\omega$ defined over profinite words.

The set of profinite words carries a compact metric: the distance between two profinite words x, y is $\frac{1}{n}$, where n is the smallest size – measured as size of the minimal automaton – of a regular language L such that x ultimately belongs to L and y does not. This metric is compatible with the notion of convergence defined above. In particular, the set A^+ of finite words is dense in the set of profinite words, $\widehat{A^+}$. Multiplication and the ω-power are continuous mappings over $\widehat{A^+}$. One can prove that $x^\omega = \lim_{n\to\infty} x^{n!}$ for any $x \in \widehat{A^+}$.

The closure $\overline{L}$ in $\widehat{A^+}$ of any regular language $L \subseteq A^+$ turns out to be both closed and open, i.e. *clopen* in $\widehat{A^+}$. Conversely, any clopen subset of $\widehat{A^+}$ is of the form $\overline{L}$ for some regular language L, so clopen sets correspond precisely to regular languages. Any open set in $\widehat{A^+}$ is a (possibly infinite) union of clopen sets.

3 Languages of profinite words

In this section we discuss several ways of describing languages of profinite words – via automata, regular expressions and logic.

B- and S-regular languages. The essential idea underlying our theory is to consider B- and S-automata as processing not only finite words, but also profinite words. Let $\mathcal{A}$ be a B- or S-automaton. The following, simple observation relies on the fact that for each $n \in \mathbb{N}$, the language $\{w \in A^+ : \ f_{\mathcal{A}}(w) < n\}$ is regular.

Fact 1. *Let $w_1, w_2, \ldots$ be a convergent sequence of finite words. Then, the sequence $f_{\mathcal{A}}(w_1), f_{\mathcal{A}}(w_2), \ldots$ is convergent in $\overline{\mathbb{N}} = \mathbb{N} \cup \{\omega\}$.*

Therefore, it makes sense to define, for any $x \in \widehat{A^+}$,

$$\widehat{f_{\mathcal{A}}}(x) \quad \overset{def}{=} \quad \lim_{n\to\infty} f_{\mathcal{A}}(w_n),$$

where $w_1, w_2, \ldots$ is any sequence of finite words which converges to x. This value may happen to be ω. It is straightforward to show that $\widehat{f_{\mathcal{A}}}$ is a well-defined continuous function from $\widehat{A^+}$ to $\overline{\mathbb{N}}$. Moreover, by density of A^+ in $\widehat{A^+}$, the continuous extension of $f_{\mathcal{A}}$ to $\widehat{A^+}$ is unique, so we will further identify $f_{\mathcal{A}}$ with the continuous mapping $\widehat{f_{\mathcal{A}}} \colon \widehat{A^+} \to \overline{\mathbb{N}}$.

Similarly to the idea underlying cost functions [6], we do not care about the exact values of the function $f_{\mathcal{A}}$ (this would quickly lead to undecidability, as demonstrated by Krob [10]). What we care about is over which sequences of words, $f_{\mathcal{A}}$ grows indefinitely. By continuity of $f_{\mathcal{A}}$ and compactness of $\widehat{A^+}$, this is encoded in the set

$$\{x \in \widehat{A^+} : \ f_{\mathcal{A}}(x) = \omega\}.$$

This is a closed set as the inverse image of a point under a continuous mapping.

This motivates the following definitions. For an S-automaton $\mathcal{A}$, we define the set $L(\mathcal{A})$ consisting of all profinite words x such that $f_{\mathcal{A}}(x) = \omega$. For a B-automaton $\mathcal{A}$, we define $L(\mathcal{A})$ dually, as the language of all profinite words x such that $f_{\mathcal{A}}(x) < \omega$. In either case, we call $L(\mathcal{A})$ the *language recognized by $\mathcal{A}$*. The reason why the definitions differ is that S-automata try to maximize, while B-automata try to minimize the value of a run. We call a language $L \subseteq \widehat{A^+}$ *B-regular* (respectively, *S-regular*), if it is recognized by a B-automaton (respectively, S-automaton). Note that S-regular languages are closed, and B-regular languages are open subsets of $\widehat{A^+}$. In particular, a language is both B- and S-regular if and only if it is clopen.

Example 3. Let $\mathcal{A}$ be the B-automaton from Example 1, computing the largest block of a's. Then $L(\mathcal{A})$ is the language of all profinite words for which every block of a's has uniformly bounded length:

$$L(\mathcal{A}) = \{x \in \widehat{A^+} : \ f_{\mathcal{A}}(x) < \omega\} = \bigcup_{n\in\mathbb{N}} \{x \in \widehat{A^+} : \ x \text{ has no infix } a^n\}.$$

It is not difficult to show (using compactness and continuity of multiplication) that a profinite word has arbitrarily long blocks of a's if and only if it contains

a^ω as an infix. (We say that u is an *infix* of v if $v = v_1 \cdot u \cdot v_2$ for some, potentially empty, profinite words v_1, v_2.) Therefore, if $\mathcal{B}$ is the S-automaton from Example 1 (recall that $f_\mathcal{A} = f_\mathcal{B}$), we deduce that

$$L(\mathcal{B}) = \{x \in \widehat{A^+} : \ f_\mathcal{B}(x) = \omega\} = \widehat{A^+} - L(\mathcal{A}) = \{x_1 \cdot a^\omega \cdot x_2 : \ x_1, x_2 \in \widehat{A^+}\}.$$

Limitedness. Assume that we want to test for limitedness of a B-automaton $\mathcal{A}$. It is easy to reduce the general case to the case when the underlying finite automaton accepts all finite words (to do this, it suffices to consider the disjoint union of $\mathcal{A}$ and $\mathcal{A}'$, where $\mathcal{A}'$ is a B-automaton which maps all words accepted by $\mathcal{A}$ to ω, and the rest to 0). Then, an immediate compactness argument shows:

Fact 2. *A B-automaton $\mathcal{A}$ which accepts all finite words is limited iff $L(\mathcal{A}) = \widehat{A^+}$.*

Closure properties. As usual with nondeterministic automata, both classes – of B- and S-regular languages – are closed under language projection, and also under union and intersection.They are not, however, closed under complements: the complement of the B-regular language $L(\mathcal{A})$ from the previous example is not B-regular, since it is not an open set. However, this complement is an S-regular language, as it is equal to $L(\mathcal{B})$. More generally, we will prove the difficult result that complements of B-regular languages are S-regular, and vice versa.

The logic MSO+inf. We introduce the logic MSO+inf over profinite words. First, we define its base fragment, the logic MSO. A formula of this logic describes a set of profinite words. Usually, in the case of finite or infinite words, one sees such a word as a model whose elements are positions of the word, and so a formula of MSO speaks about sets of positions of the word. However, in profinite words, “positions” are not well-defined. To define the logic MSO over profinite words, we view the constructs of MSO as operations on languages of profinite words. We describe how to interpret the second-order existential quantifier $\exists$; for the other constructs, the idea is even simpler. We view the quantifier $\exists$ as language projection. What language do we project? A formula $\varphi(X)$ beneath a quantifier $\exists$ defines a language L_φ over the extended alphabet $A \times \{0,1\}$. For example, $\varphi(X) = a(X) \wedge \text{singleton}(X)$ defines the language L_φ of those profinite words over $A \times \{0,1\}$, which contain precisely one symbol $(a,1)$ and no other symbols with a 1 on the second coordinate. We define the language of the formula $\exists X.\varphi(X)$ as the projection of the language L_φ, forgetting about the second coordinate. Therefore, $\exists X.a(X) \wedge \text{singleton}(X)$ describes the set of profinite words which have precisely one letter a.

With similar ideas, it is easy to interpret all the usual constructs of MSO as language operations: the Boolean connectives $\wedge, \vee, \neg$, the binary predicates $<, \in$ and the unary predicates $a(X)$, per each letter $a \in A$. This way, we define the semantic of the MSO logic over profinite words. This logic describes precisely the class of clopen sets. To go beyond that, we add a predicate $\inf(X)$ which holds in a profinite word over $A \times \{0,1\}$ if it has infinitely many 1's on the second coordinate. This is a closed, but not open property of profinite words over the alphabet $A \times \{0,1\}$, so it is not definable in MSO. We denote the logic

MSO extended by the quantifier inf by MSO+inf and distinguish the syntactic fragment MSO+inf$^+$ (resp., MSO+inf$^-$) where the predicate inf appears only under an even (resp. odd) number of negations.

Example 4. Consider the S-regular language $L(\mathcal{B})$ from Example 3: "there is an infinite block of a's". It can be described by the following formula of MSO+inf$^+$:

$$\exists X. \inf(X) \;\wedge\; \forall x,y,z.\big(x \in X \;\wedge\; z \in X \;\wedge\; (x < y < z) \implies (y \in X \wedge a(y))\big).$$

This example can be easily extended, yielding the following.

Proposition 3. *B-regular languages are definable in* MSO+inf$^-$*, and S-regular languages are definable in* MSO+inf$^+$*. The translations are effective.*

B- and S-regular expressions. We consider the usual syntax of regular expressions, except that apart from the usual Kleene star, which corresponds to *unrestricted* iteration, there are two new iteration operations: *finite* iteration, denoted $L^{<\infty}$, and *infinite* iteration, denoted L^{∞}. Formally, we define *profinite sequences* of profinite words, as profinite words over the alphabet A with an additional *separator* symbol †. A profinite word $x \in \widehat{A^+}$ is an *element* of a profinite sequence $\hat{x}$ if $\dagger x \dagger$ is an infix of $\dagger \hat{x} \dagger$. The *concatenation* of $\hat{x}$ is obtained by removing the symbols †. We define L^{∞} (resp. $L^{<\infty}$ and L^*) as concatenations of profinite sequences containing infinitely (resp. finitely, arbitrarily) many separators, and whose elements belong to L. *B-regular expressions* can only use the exponents $<\!\infty$ and $*$, while *S-regular expressions* can only use the exponents ∞ and $*$.

Example 5. The B-regular expression $(a^{<\infty}\, b)^*\, a^{<\infty}$ describes precisely the language accepted by the B-automaton $\mathcal{A}$ from Example 3 – "every block of a's has a finite length". The S-regular expression $(a+b)^*\, a^{\infty}\, (a+b)^*$ describes precisely the complement of $L(\mathcal{A})$, i.e. the language accepted by the S-automaton $\mathcal{B}$.

Mimicking the standard translation from regular expressions to automata we get:

Proposition 4. *A language defined by a B-/S-regular expression is B-/S-regular.*

4 Recognizable languages

Syntactic congruence. Just as multiplication is intimately related with regular languages, multiplication together with the ω-power over $\widehat{A^+}$ turn out to be of central importance for B- and S-regular languages. For notational reasons, we view $(\widehat{A^+}, \cdot, \omega)$ as an algebra over the signature $\langle \,\cdot\, , \# \rangle$, where the ω-power of $\widehat{A^+}$ plays the role of the operation $\#$ of the signature. Let $L \subseteq \widehat{A^+}$. Its $\langle \,\cdot\, , \# \rangle$*-syntactic congruence* $\simeq_L$ is the coarsest equivalence relation over $\widehat{A^+}$ which preserves multiplication, the ω-power, and membership in L.

Example 6. Let $L = (a^{<\infty}\, b)^*\, a^{<\infty}$ be the language of the B-automaton which computes the maximal length of a block of a's. It is easy to see that the equivalence classes of $\simeq_L$ (and also of $\simeq_K$, for $K = \widehat{A^+} - L$) are:

$$a^{<\infty}, \qquad (a^{<\infty}\, b)^+\, a^{<\infty}, \qquad (a+b)^*\, a^{\infty}\, (a+b)^*.$$

Stabilization semigroups. We consider languages $L \subseteq \widehat{A^+}$ whose $\langle \cdot, \# \rangle$-syntactic congruence has a finite index. Such a set yields a finite $\langle \cdot, \# \rangle$-*syntactic algebra*, i.e. the quotient $S_L = \widehat{A^+}/{\simeq_L}$. Since $\simeq_L$ is a congruence, the syntactic algebra is equipped with two operations – the usual multiplication, and *stabilization*, denoted $\#$, which stems from the ω-power in the profinite semigroup. The syntactic algebra also naturally inherits the *quotient* topology from $\widehat{A^+}$, which is usually non-Hausdorff, i.e. there might be singleton sets which are not closed. (However, if L is a closed or open language, then the quotient topology is T_0, i.e. if $x \in \overline{\{y\}}$ and $y \in \overline{\{x\}}$ for $x, y \in S_L$, then $x = y$.) Multiplication and stabilization in S_L are continuous with respect to the topology, and also satisfy several properties which are easily derived from the properties of multiplication and the ω-power over $\widehat{A^+}$. Namely, for $s, t, e \in S$:

$$\begin{aligned} s \cdot (t \cdot s)^\# &= (s \cdot t)^\# \cdot s & s^\# \cdot s^\# &= s^\# \\ (s^\#)^\# &= s^\# & e \cdot e^\# &= e^\# \qquad \text{for idemptent } e \\ (s^n)^\# &= s^\# \qquad \text{for } n = 1, 2, 3 \ldots & s^\# &\in \overline{\{s^n : \ n \in \mathbb{N}\}}. \end{aligned}$$

A *stabilization semigroup* is a T_0 topological space S equipped with two continuous operations $\cdot$ and $\#$ satisfying the above axioms, apart from associativity of $\cdot$.

Example 7. Let S_L denote the quotient set induced by the language L from Example 6. As noted there, S_L consists of three equivalence classes, which we denote by $[a]$, $[b]$ and $[a^\omega]$, respectively. Multiplication, stabilization and topology over S_L flow from the properties of the three equivalence classes: multiplication is commutative and each element is idempotent, $[a^\omega]$ is the zero element and $[a]$ is the neutral element; stabilization maps $[a]$ to $[a^\omega]$ and s to s otherwise; $[a^\omega]$ is contained in the closure of $[a]$ and in the closure of $[b]$.

Recognizability. We consider an analogue of the notion of recognizability by semigroups in the classical theory. Recall that a subset $L \subseteq \widehat{A^+}$ is *recognizable* if there is a mapping $\alpha \colon A \to S$ to a finite discrete semigroup such that for the induced homomorphism $\hat{\alpha} \colon \widehat{A^+} \to S$ we have $L = \hat{\alpha}^{-1}(F)$ for some $F \subseteq S$.

Instead of semigroups, we deal with finite stabilization semigroups. A *homomorphism* $\hat{\alpha}$ from $\widehat{A^+}$ to a stabilization semigroup S is required to preserve multiplication and map the ω-power in $\widehat{A^+}$ to stabilization in S. We use a notion of invariance of $\hat{\alpha}$ under *infinite substitutions*, which intuitively means that if a profinite word x is factorized into a profinite sequence of factors, and each factor x_i is replaced by some other factor y_i with $\hat{\alpha}(x_i) = \hat{\alpha}(y_i)$, then, for the resulting concatenation y of the factors y_i, $\hat{\alpha}(x) = \hat{\alpha}(y)$. We say that such a homomorphism $\hat{\alpha} \colon \widehat{A^+} \to S$ is an ∞-*homomorphism.* The following result plays a pivotal role in the theory, and its proof is difficult comparing to the classical case.

Theorem 5. *Let $\alpha \colon A \to S$ be any mapping from a finite alphabet A to a finite stabilization semigroup S. Then there exists a unique ∞-homomorphism $\hat{\alpha} \colon \widehat{A^+} \to S$ extending α. The mapping $\hat{\alpha}$ is continuous. Its image is the subset of S generated from $\alpha(A)$ by the operations $\langle \cdot, \# \rangle$.*

Note that the extension $\hat{\alpha}$ is not necessarily the *unique* continuous homomorphic extension of α. We call $\hat{\alpha}$ the ∞-homomorphism *induced* by α. We say that a language $L \subseteq \widehat{A^+}$ is *recognized* by $\hat{\alpha}\colon \widehat{A^+} \to S$ if $L = \hat{\alpha}^{-1}(F)$ for some $F \subseteq S$; if additionally F is closed (resp. open) in S, we say that L is $\downarrow$*-recognizable* (resp. $\uparrow$*-recognizable*). Note that a recognizable set is described in a finite manner by $\alpha\colon A \to S$ and $F \subseteq S$. It is crucial that the image of $\hat{\alpha}$ can be computed from α.

Example 8. Let S be the stabilization semigroup $\widehat{A^+}/\!\simeq_L$ from the previous example, whose elements are $[a], [b], [a^\omega]$. Let $\alpha\colon A \to S$ map a to $[a]$ and b to $[b]$. We will check that the quotient mapping $\alpha_L\colon \widehat{A^+} \to S$ is the ∞-homomorphism induced by α. We argue that α_L is invariant under infinite substitutions. Consider a profinite word x, and choose some factorization of x. Replace each factor by some other factor, with the same image under α_L. Schematically:

$$\begin{array}{rcccccc} x = & aaa & aaba & aaa & \cdots & ab^\omega a & baaab \\ & \downarrow & \downarrow & \downarrow & \cdots & \downarrow & \downarrow \\ y = & aaaaa & (ab)^\omega & aaaaaaa & \cdots & aaaaabaaaa & aaaaabaaa \end{array}$$

Intuitively, it is clear that if the original word x contains no infinite block of a's, then no such block can appear in the resulting word y either. Hence, $\alpha_L(y) = \alpha_L(x)$.

The proof of Theorem 5 extends the idea of Simon's factorization trees to profinite words and stabilization semigroups, which we shortly describe. Start with any profinite word x. We want to determine the *type* of x, i.e. $\hat{\alpha}(x)$. If x is a single letter a, then its type is $\alpha(a)$. If not, we try to factorize x into a profinite sequence of factors, for which the type can be determined. We use three rules:

- If $x = x_1 \cdot x_2$, and $\hat{\alpha}(x_1) = s_1$, $\hat{\alpha}(x_2) = s_2$, then $\hat{\alpha}(x) = s_1 \cdot s_2$,
- If x factorizes into finitely many factors, each of idempotent type e, then $\hat{\alpha}(x) = e$,
- If x factorizes into infinitely many factors, each of idempotent type e, then $\hat{\alpha}(x) = e^\#$.

We prove by induction on $|S|$ that in a finite number of steps, depending only on $|S|$, using the above three rules, any profinite word x can be iteratively split into single letters. Moreover, we prove that the resulting type does not depend on the chosen "factorization tree". The proof of existence of factorization trees is similar to the proof of Simon's theorem, and proceeds by induction on the size of S. The proof of uniqueness requires the use of the axioms of stabilization semigroups. It is similar to a proof of analogous statement in [7]. An important difference is that there, only finite words have factorization trees, and their output is unique only in an asymptotic way.

The standard Cartesian-product construction yields several closure properties for recognizable languages. For closure under projection, we use two enhanced variants of the powerset construction, similar to constructions from [7].

Proposition 6. *Recognizable languages are closed under Boolean combinations.* $\downarrow$*-recognizable (resp.* $\uparrow$*-recognizable) languages are closed under unions and intersections. Complements of* $\downarrow$*-recognizable languages are* $\uparrow$*-recognizable and vice versa.* $\downarrow$*-recognizable and* $\uparrow$*-recognizable languages are closed under projections.*

By inductively applying the above to formulas of MSO+inf, we get:

Corollary 1. *Languages definable in* MSO+inf$^-$ *are* $\downarrow$*-recognizable, and languages definable in* MSO+inf$^+$ *are* $\uparrow$*-recognizable. The translations are effective.*

5 The main results

The main theorem collects the notions and results listed above, proving the equivalence of several characterizations. The last one is a finite-index characterization of B-automata. Up to our knowledge, such a characterization has not been – and perhaps cannot be – phrased in the remaining frameworks.

Theorem 7. *Let $L \subseteq \widehat{A^+}$ and $K = \widehat{A^+} - L$ be its complement. The following conditions 1-9 are equivalent:*

1. L is defined by a B-regular expression,	*5. K is defined by an S-regular expression,*
2. $L = L(\mathcal{A})$ for some B-automaton $\mathcal{A}$,	*6. $K = L(\mathcal{B})$ for some S-automaton $\mathcal{B}$,*
3. L is definable in MSO+inf$^-$*,*	*7. K is definable in* MSO+inf$^+$*,*
4. L is $\uparrow$-recognizable,	*8. K is $\downarrow$-recognizable,*

9. The $\langle\cdot, \#\rangle$-syntactic congruence of K has finite index and $K = \overline{K \cap A^{\langle\cdot,\omega\rangle}}$.

In the last characterization, $A^{\langle\cdot,\omega\rangle}$ is the set of profinite words which can be generated from A by applying multiplication and the ω-power – they are analogues of ultimately periodic words in the theory of ω-regular languages. It follows that a B- or S-regular language is determined by its elements contained in $A^{\langle\cdot,\omega\rangle}$, similarly as an ω-regular language is determined by its ultimately periodic words.

By the last part of Theorem 5, the image of an ∞-homomorphism to a finite stabilization semigroup can be computed using a fixed point calculation. Hence, emptiness of recognizable languages is decidable. This proves the following.

Theorem 8. *Emptiness of Boolean combinations of B-regular languages is decidable. In particular, the limitedness problem is decidable for B-automata.*

The above result extends the decidability results of Hashiguchi and Kirsten. As emptiness of Boolean combinations reduces to inclusion testing, it is equivalent to the main result of [7] – that the domination relation is decidable for B-automata.

6 From infinite words to profinite words

We describe a connection between ω-words (i.e. mappings from $\mathbb{N}$ to A) and profinite words. Recall that any ω-regular language can be presented as a finite union of languages of the form $U \cdot V^\omega$, where $U, V \subseteq A^+$ are regular languages of finite words. We generalize this observation, and provide a meta-reduction between the satisfiability problems for logics over ω-words to corresponding logics over profinite words. The proof resembles Büchi's original proof of decidability of MSO. Instead of the usual Ramsey lemma, we use the following observation (originating from [5]): For any ω-word $w \in A^\omega$ there is a factorization $w = u_0 \cdot u_1 \cdot u_2 \cdots$ such that the sequence $u_0, u_1, u_2, \ldots$ is convergent to some $u_\infty \in \widehat{A^+}$. The proof is an easy, repeated application of the usual Ramsey lemma.

Let $V \subseteq \widehat{A^+}$ be a language of profinite words, and $\varepsilon > 0$ a real number. Consider the following language of infinite words $V^\omega_\varepsilon \subseteq A^\omega$:

$$V^\omega_\varepsilon \overset{def}{=} \{v_1 \cdot v_2 \cdot v_3 \cdots : \exists v_\infty \in V^* : \lim_{n\to\infty} v_n = v_\infty \textit{ and } \forall_n d(v_n, v_\infty) < \varepsilon\}.$$

For a regular language $U \subseteq A^+$ of finite words, we say that the expression $U \cdot V^\omega$ is *well-formed* if the language $U \cdot V_\varepsilon^\omega$ does not depend on the choice of $0 < \varepsilon \leq 1/n$, where n is the size of the minimal automaton recognizing U. In this case, we define the language $U \cdot V^\omega$ as $U \cdot V_\varepsilon^\omega$. For example, the expression $(a+b)^* \cdot (a^{<\infty}b)^\omega$ is well-formed and describes the language L_B from the introduction. For a class $\mathcal{L}$ of languages of profinite words, let $\omega\mathcal{L}$ denote the class of all finite unions of languages defined by well-formed expressions $U \cdot V^\omega$ with $U \subseteq A^+$ regular and $V \in \mathcal{L}$.

In the following theorem, by REGULAR, B-REGULAR, S-REGULAR, MSO+inf, we denote the corresponding classes of languages of profinite words, and to each we apply the map $\mathcal{L} \mapsto \omega\mathcal{L}$ as described above, yielding classes of languages of infinite words. The proof of the theorem is very general. It generalizes Büchi's proof of decidability of MSO over infinite words.

Theorem 9. *Every ω-regular language is in ω*REGULAR*. Every ωB-regular language is in ω*B-REGULAR*. Every ωS-regular language is in ω*S-REGULAR*. Every* MSO+$\mathbb{B}$ *definable language is in ω*MSO+inf*. The translations are effective.*

The reduction described above allows to transfer results from profinite words to ω-words. For instance, the main results of [4] (concerning ωB- and ωS-regular languages) follow from the results in our paper. More importantly, we get:

Corollary 2. *The satisfiability problem for the logic* MSO+$\mathbb{B}$ *over ω-words reduces to the satisfiability problem for the logic* MSO+inf *over profinite words.*

We mention that by refining our Theorem 9, Skrzypczak [14] proved that a language of infinite words which is both ωB-regular and ωS-regular must in fact be ω-regular – reflecting the immediate, analogous fact for profinite words.

Conclusion. We presented a new proof and framework for the limitedness problem. We rise the question of decidability of the logic MSO+inf over profinite words.

References

1. P. A. Abdulla, P. Krcál, and W. Yi. R-automata. In *CONCUR*, pages 67–81, 2008.
2. Mikołaj Bojańczyk. A bounding quantifier. In *Computer Science Logic*, volume 3210 of *Lecture Notes in Computer Science*, pages 41–55, 2004.
3. Mikołaj Bojańczyk. Beyond ω-regular languages. In *STACS*, pages 11–16, 2010.
4. Mikołaj Bojańczyk and Thomas Colcombet. Bounds in ω-regularity. In *Logic in Computer Science*, pages 285–296, 2006.
5. Mikołaj Bojańczyk and Eryk Kopczyński. Ramsey's theorem for colors from a metric space. Submitted, 2010.
6. Thomas Colcombet. The theory of stabilisation monoids and regular cost functions. In Susanne Albers, Alberto Marchetti-Spaccamela, Yossi Matias, Sotiris E. Nikoletseas, and Wolfgang Thomas, editors, *ICALP (2)*, volume 5556 of *Lecture Notes in Computer Science*, pages 139–150. Springer, 2009.
7. Thomas Colcombet. Regular cost functions, part i: Logic and algebra over words. submitted, 2011.
8. Kosaburo Hashiguchi. Limitedness theorem on finite automata with distance functions. *Journal of Computer and System Sciences*, 24:233–244, 1982.
9. Daniel Kirsten. Distance desert automata and the star height problem. *Theoretical Informatics and Applications*, 39(3):455–511, 2005.
10. Daniel Krob. The equality problem for rational series with multiplicities in the tropical semiring is undecidable. In Werner Kuich, editor, *ICALP*, volume 623 of *Lecture Notes in Computer Science*, pages 101–112. Springer, 1992.
11. Hing Leung. On the topological structure of a finitely generated semigroup of matrices. *Semigroup Forum*, 37:273–278, 1988.
12. Jean-Éric Pin. Profinite methods in automata theory. In Susanne Albers and Jean-Yves Marion, editors, *STACS*, volume 3 of *LIPIcs*, pages 31–50. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany, 2009.
13. Imre Simon. On semigroups of matrices over the tropical semiring. *ITA*, 28(3-4):277–294, 1994.
14. Michał Skrzypczak. Separation property for ωB- and ωS-regular languages. Submitted, January 2012.