Cellular automata, \( \omega \omega \)-regular sets, and sofic systems

Karel Culik II

Department of Computer Science, University of South Carolina, Columbia, SC 29208, USA

Sheng Yu

Department of Mathematical Sciences, Kent State University, Kent, OH 44242, USA

Received 15 May 1988

Revised 15 September 1989

Abstract


We study and compare several mechanisms for defining sets of biinfinite words (\( \omega \omega \)-languages), namely \( \omega \omega \)-finite automata, adherences of regular languages and sofic systems. We show that a sofic system is a topologically closed \( \omega \omega \)-regular set. We also show that there is a one-to-one correspondence between sofic systems and the adherences of regular languages. We give a complete proof of the closure of the \( \omega \omega \)-regular sets under \( \omega \omega \)-rational relations and under Boolean operations. Finally, we disprove Hurd's conjecture [15] on bi-extensible subsets of languages, and show that the conjecture would hold if a different definition were used.

1. Introduction

Since sofic systems were introduced by Weiss [24] in 1973, they have been widely studied in relation to cellular automata and dynamical systems. A sofic system is defined to be the image of a subshift of finite type under a continuous mapping [24,15]. A sofic system can also be characterized by the properties of the set of its finite substrings. In [24], it was mentioned that “the sofic systems are essentially those that can be defined by finite automata”. But there were no further statements.

* Research supported by the Natural Sciences and Engineering Research Council of Canada grants OGP1N007 and A7403 and the National Sciences Foundation Grant No. CCR-8702752.
and proofs besides this only statement. Hurd later gave a detailed proof [15] of the fact that \( K \subseteq S^\mathbb{Z} \) is a sofic system if and only if the set of all finite subwords of \( K \) is a regular language. (Note that he neglected to mention that \( K \) should be a subshift.)

In [19], Nivat and Perrin defined "a biinfinite word as the equivalence class under the shift of a two-sided infinite sequence". They studied \( \omega \omega \)-regular sets and other properties of biinfinite words based on results on one-way infinite words, for which many deep results have already been available.

Both \( \omega \omega \)-regular sets and sofic systems are naturally related to linear cellular automata. However, it appears that no effort has been made to study them together. Many questions need to be answered. For example, what are the differences and relationship between sofic systems and \( \omega \omega \)-regular sets (i.e. recognizable sets of biinfinite words)? What are the algebraic limits and adherence sets of regular languages and their relation to sofic systems? We try to answer these and related questions in this paper.

In the next section, we formally define biinfinite words, \( \omega \omega \)-regular sets, and \( \omega \omega \)-rational relations. We show that \( \omega \omega \)-regular sets are closed under \( \omega \omega \)-finite transduction. We also give a complete proof of the closure of the family of \( \omega \omega \)-regular sets under complementation. The same result has been given in [19].

In Section 3, we define and study the basic properties of algebraic limit and adherence sets of languages, in particular, regular languages. In Section 4, the relationship between sofic systems and \( \omega \omega \)-regular sets, algebraic limits, and adherences of regular languages are discussed.

In the final section, we study the bi-extensible subsets of languages. We disprove Hurd's conjecture [15] and show that Hurd's conjecture would be true if a different definition of extensible subset was used. We also show that the adherences, but not the algebraic limits, of context-free languages have the property that the sets of their finite subwords are context-free. This property of context-free languages parallels the relation of sofic systems to regular languages.

2. Biinfinite words, \( \omega \omega \)-regular sets, and \( \omega \omega \)-rational relations

Let \( \mathbb{L} \) be the set of integers, and \( S \) be a finite alphabet. A configuration \( c \) is a function \( c: \mathbb{L} \rightarrow S \), i.e. \( c \in S^{\mathbb{L}} \). The term configuration is used for cellular automata. The configuration space \( S^\mathbb{L} \) of a linear cellular automaton is the product of infinitely many finite sets \( S \). The product topology on \( S^\mathbb{L} \) with \( S \) endowed with the discrete topology is compact by Tychonoff's theorem [16, Theorem 5.13]. A subbasis of open sets for this topology consists of all sets of the form \( \{c \in S^\mathbb{L} \mid c_i = a \} \) where \( i \in \mathbb{L} \) and \( a \in S \). A set \( C \subseteq S^\mathbb{L} \) is open if it is a union of finite intersections of sets in the subbasis.

A right shift \( \sigma \) is a mapping \( \sigma: S^{\mathbb{L}} \rightarrow S^{\mathbb{L}} \) such that \( (\sigma(c))_i = c_{i+1} \), \( -\infty < i < \infty \). A left shift mapping can be similarly defined. A set \( X \subseteq S^{\mathbb{L}} \) is said to be shift-
invariant if $\sigma(X) = X$. Let $c$ be a configuration. The biinfinite word ($\omega\omega$-word) $v$ generated by $c$ is a subset of $S^\mathbb{Z}$ such that (1) $c \in v$; (2) $v$ is shift-invariant; and (3) for any shift-invariant subset $u$ that contains $c$, $v \subseteq u$. It is clear that $v = \{\sigma^i(c) | -\infty < i < \infty\}$. A subshift is a shift-invariant closed set.

Finite automata that recognize sets of biinfinite words have been defined by Nivat and Perrin [19] and studied in, e.g., [2,10,19]. Here we use a slightly different but equivalent definition which is more convenient for our purpose.

An $\omega\omega$-finite automaton ($\omega\omega$-FA) $M$ is a quintuple $(Q, S, \delta, Q_L, Q_R)$, where

- $Q$ is the finite set of state;
- $S$ is the input alphabet;
- $\delta$ is the transition function;
- $Q_L \subseteq Q$ is the set of left (accepting) states; and
- $Q_R \subseteq Q$ is the set of right (accepting) states.

A biinfinite word $v$ is said to be recognized by $M$ if there is a mapping $\mathbb{Z} \to Q$, i.e. a biinfinite sequence of states

$$\ldots, q_{-2}, q_{-1}, q_0, q_1, q_2, \ldots$$

and a configuration $c$ in $v$ such that, for all $j \in \mathbb{Z}$,

1. $\delta(q_j, c_j) = q_{j+1}$; and
2. there exist $m, n \in \mathbb{Z}$, $m \leq j \leq n$, such that $q_m \in Q_L$ and $q_n \in Q_R$.

In other words, $v$ is said to be recognized by $M$ if there is a biinfinite computation of $M$ on a configuration $c$ in $v$ such that there is a left state appearing arbitrarily early, and there is a right state appearing arbitrarily late in the computation. Such a computation is called an accepting computation.

The set of biinfinite words recognized by $M$ is denoted $B(M)$. We call $B(M)$ an $\omega\omega$-regular set. Clearly, every $\omega\omega$-regular set is shift-invariant.

**Example 2.1.** Let $M = (Q, S, \delta, Q_L, Q_R)$ be an $\omega\omega$-FA, where $Q = \{0, 1\}$, $S = \{a, b\}$, $Q_L = \{0\}$, $Q_R = \{1\}$, and $\delta$ is given in Fig. 1. The set of biinfinite words recognized by $M$ is the set of all words which have infinitely many $a$'s followed by infinitely many $b$'s, i.e. $a^*ab^\omega$. □

Finite or one-way infinite words can be considered as special cases of biinfinite words in the following sense: A special quiescent symbol is specified such that a one-
way infinite word (ω-word) is a biinfinite word with infinitely many quiescent symbols on the left end, and a finite word is a biinfinite word with a finite consecutive nonquiescent subword.

In an ωω-FA, a left (right) state that is not in a cycle can be changed into a non-left (non-right) state without affecting the set of biinfinite words recognized by the ωω-FA. A state which cannot be reached from a left state or from which no right state can be reached is useless—it does not contribute to the recognition of any biinfinite word. We say that an ωω-FA is reduced if it satisfies the following conditions:

(i) Every left state is in a cycle.
(ii) Every right state is in a cycle.
(iii) Every state can be reached from some left state.
(iv) From every state some right state can be reached.

Obviously, for any given ωω-FA we can construct a reduced one that recognizes the same set of biinfinite words.

An ωω-finite transducer $T$ is a 6-tuple $(P, S, S', \varrho, P_L, P_R)$ where

- $P$ is the finite set of states,
- $S$ is the input alphabet,
- $S'$ is the output alphabet,
- $\varrho : P \times (S \cup \{\lambda\}) \to P \times S'^*$ is the transition function,
- $P_L \subseteq P$ is the set of left (accepting) states, and
- $P_R \subseteq P$ is the set of right (accepting) states.

A biinfinite word $d$ is an output on input $c$ under the ωω-finite transducer $T$ if there are a biinfinite sequence of states

$$..., P_{-2}, P_{-1}, P_0, P_1, ...$$

and a biinfinite sequence of strings $x_j \in S'^*$ such that

$$... x_{-2} x_{-1} = ... d(-2)d(-1),$$

and for all $j \in \mathbb{Z}$

1. $\varrho(P_j, c(j)) = (P_{j+1}, x_j)$, and
2. there exist $m, n \in \mathbb{Z}, m \leq j \leq n$, such that $p_m \in P_L$ and $p_n \in P_R$.

The relation defined by an ωω-finite transducer is called an ωω-rational relation. Clearly, the Nivat theorem [4] can be generalized to ωω-rational relations.

**Theorem 2.2.** The family of ωω-regular sets is closed under ωω-rational relations.

**Proof.** Let $C = B(M)$ for some ωω-FA $M = (Q, S, \delta, Q_L, Q_R)$. Let $T = (P, S, S', \varrho, P_L, P_R)$ be an ωω-finite transducer. We shall construct an ωω-FA $M'$ such that $T(C) = B(M')$. 

Define \( \pi_L : \{0, 1, 2\} \times Q \times P \to \{0, 1, 2\} \) by

\[
\pi_L(2, q, p) = 0;
\]

\[
\pi_L(0, q, p) = \begin{cases} 
0, & \text{if } q \not\in Q_L, \\
1, & \text{if } q \in Q_L;
\end{cases}
\]

\[
\pi_L(1, q, p) = \begin{cases} 
1, & \text{if } p \not\in P_L, \\
2, & \text{if } p \in P_L.
\end{cases}
\]

Similarly, \( \pi_R : \{0, 1, 2\} \times Q \times P \to \{0, 1, 2\} \) is defined by

\[
\pi_R(2, q, p) = 0;
\]

\[
\pi_R(0, q, p) = \begin{cases} 
0, & \text{if } q \not\in Q_R, \\
1, & \text{if } q \in Q_R;
\end{cases}
\]

\[
\pi_R(1, q, p) = \begin{cases} 
1, & \text{if } p \not\in P_R, \\
2, & \text{if } p \in P_R.
\end{cases}
\]

The \( \omega \omega \)-FA \( M' \) simulates simultaneous execution of \( M \) and \( T \); the states of \( M' \) have also two additional components whose purpose is to remember the passage through left and right states of \( M \) and \( T \). Define \( M' = (Q', S', S', Q^L, Q^R) \), where

\[
Q' = \{0, 1, 2\} \times \{0, 1, 2\} \times Q \times P,
\]

\[
Q^L = \{2\} \times \{0, 1, 2\} \times Q \times P,
\]

\[
Q^R = \{0, 1, 2\} \times \{2\} \times Q \times P,
\]

and

\[
\delta'((i_1, j_1, q_1, p_1), x') = (i_2, j_2, q_2, p_2),
\]

if \( x' \in S'^* \) and there exists \( a \in S \cup \{\lambda\} \) such that

\[
\delta(q_1, a) = q_2, \quad \delta(p_1, a) = (p_2, x'),
\]

\[
\pi_L(i_1, q_1, p_1) = i_2, \quad \pi_R(j_1, q_1, p_1) = j_2.
\]

Obviously, we could add finitely many additional states to \( Q' \) and replace the definition of \( \delta'(q', x') \) by one written in terms of \( \delta'(q', a') \), for \( a' \in S' \). Accepting computations of \( M' \) are in one-to-one correspondence with simultaneous accepting computations of \( M \) and \( T \). Thus it is easy to check that \( T(C) = B(M') \).

In [19], it was stated that the family of \( \omega \omega \)-regular sets is closed under Boolean operations. Here we give a different proof. Note that it is not obvious that the closure properties of \( \omega \)-regular sets imply that \( \omega \omega \)-regular sets are closed under complementation. For example, \( \omega a b^+a^\omega \) can be represented by the pair of one-way infinite words \((a^\omega, b^+a^\omega)\). It is not obvious how the complement of such a set of biinfinite words is represented by the Boolean operations of the two \( \omega \)-regular sets.
In the following, we give a relatively detailed proof for the closure property of \( \omega \omega \)-regular sets under complementation. But first, we need to introduce the notion of canonical expressions of \( \omega \omega \)-regular sets. Let \( C \) be an \( \omega \omega \)-regular set accepted by an \( \omega \omega \)-FA \( A = (Q, S, \delta, Q_L, Q_R) \). We assume that \( A \) is reduced. For each state \( q \) in \( Q \), let \( A(q) \) be a FA \( (Q, S, \delta, q, Q_R) \) where \( \delta \) is defined by \( p' \in \delta(p, a) \) if and only if \( p \in \delta(p', a) \), and \( A(q) = (Q, S, \delta, q, Q_R) \). Let \( |A| \) denote the set of one-way infinite words recognized by \( A \), i.e. the set of one-way infinite words that have infinitely many prefixes in \( L(A) \). Then it is obvious that \( C = \bigcup_{q \in Q} |A(q)| \). Note that by \( D \) and \( E \) are sets of one-way infinite words, we mean the set of biinfinite words that is the shift-closure of the set of all configurations formed by concatenation of the reversal of a word in \( D \) and a word in \( E \). It is not difficult to see that the expression \( C = \bigcup_{q \in Q} |A(q)|^R |A(q)| \) has the property that, for each biinfinite word \( w \) in \( C \), no matter how \( w \) is splitted into two one-way infinite words \( x \) and \( y \) such that \( w = xy \) there exists a \( q \in Q \) such that \( x \in |A(q)| \) and \( y \in |A(q)| \). We call this expression a canonical expression of \( C \). We state this formally as follows:

**Definition 2.3.** Let \( C \) be a set of biinfinite words in \( S^\mathbb{Z} \). The expression \( \bigcup_{i=1}^{n} D_i^R E_i \), where \( n \geq 0 \) and \( D_i \) and \( E_i, 1 \leq i \leq n \), are \( \omega \)-regular sets, is called a canonical expression of \( C \) if \( C = \bigcup_{i=1}^{n} D_i^R E_i \) and for any configuration \( c \in C \) and \( t \in \mathbb{Z} \) there exists an integer \( i, 1 \leq i \leq n \), such that \( c_{t-i} c_{t-i+1} \ldots \in D_i \) and \( c_{t} c_{t+1} \ldots \in E_i \).

**Lemma 2.4** (Representation Lemma). A set of biinfinite words is \( \omega \omega \)-regular if and only if it can be presented by \( D_1^R E_1 \cup D_2^R E_2 \cup \cdots \cup D_n^R E_n \) where \( D_1, \ldots, D_n, E_1, \ldots, E_n \) are \( \omega \)-regular sets and \( D \) denotes the reversal of \( D \). Every \( \omega \omega \)-regular set has a canonical expression.

The first part of the lemma has been stated in [19]. The second part is clear from the above argument.

**Theorem 2.5.** The family of \( \omega \omega \)-regular sets is closed under complementation.

**Proof.** Let \( C \) be an \( \omega \omega \)-regular set and \( \bigcup_{i=1}^{n} D_i^R E_i \) a canonical expression of \( C \). We construct a new canonical expression

\[
C = \bigcup_{j=1}^{m} D_j^R E_j' \quad \text{such that} \quad \bigcup_{j=1}^{m} D_j^R \bigcup_{i=1}^{n} D_i \quad \text{and} \quad D_j \cap D_t' = \emptyset \quad \text{for all} \quad 1 \leq s, t \leq m \quad \text{and} \quad s \neq t.
\]

(2.1)

This can be done because \( \omega \)-regular sets of one-way infinite words are closed under Boolean operations and results are constructible. See [7,8,18,22,23,20] for details. Later we will refer this expression as (2.1). Similarly, we construct another canonical expression.
Cellular automata, \( \omega \)-regular sets, and sofic systems

\[ C = \bigcup_{k=1}^{m} D_k^r E_k^u \] such that \( \bigcup_{k=1}^{m} E_k^u = \bigcup_{i=1}^{n} E_i \) and \( E_s^u \cap E_t^u = \emptyset \)
for all \( 1 \leq s, t \leq l \) and \( s \neq t \). \hfill (2.2)

We call this expression (2.2). The following properties of (2.1) and (2.2) are obvious: note that \( \overline{X} \) is used to denote that complement of \( X \), where \( X \) is either a set of one-way infinite words or a set of biinfinite words according to the context of the notation.

1. If \( c \in C \), then for any \( t \in \mathbb{Z} \) there is exactly one integer \( i \), \( 1 \leq i \leq m \), such that \( c_{t-i} c_{t-i-1} \ldots \in D_i' \) and \( c_{t-i+1} \ldots \in E_i' \). Note that \( c_{t-i} c_{t-i+1} \ldots \) may also be in \( E_i' \) for some \( 1 \leq j \leq m \) and \( j \neq i \).

2. If \( c \in C \), then for any \( t \in \mathbb{Z} \) there is exactly one integer \( j \), \( 1 \leq j \leq l \), such that \( c_{t-j} c_{t-j-1} \ldots \in D_j' \) and \( c_{t-j+1} \ldots \in E_j' \). Note that \( c_{t-j} c_{t-j+1} \ldots \) may also be in \( D_j' \) for some \( 1 \leq i \leq l \) and \( i \neq j \).

3. If \( c \in C \) and \( c \in D_i^r E_i^u \) for some \( 1 \leq i \leq m \), then (i) there does not exist an integer \( t \) such that \( c_{t-i} c_{t-i-1} \ldots \in D_i' \) and, therefore, (ii) \( c \notin D_i^r E_i^u \).

4. If \( c \in C \) and \( c \in D_j^r E_j^u \) for some \( 1 \leq j \leq l \), then (i) there does not exist an integer \( t \) such that \( c_{t-j} c_{t-j-1} \ldots \in D_j' \) and, therefore, (ii) \( c \notin D_j^r E_j^u \).

Let \( D' = \bigcup_{i=1}^{m} D_i' \) and \( E'' = \bigcup_{j=1}^{l} E_j'' \). Now, we claim that
\[ C = \left( \bigcup_{j=1}^{m} D_j^r E_j^u \right) \cup \left( \bigcup_{k=1}^{l} D_k^r E_k^u \right) \cup D^r E'' \]

For convenience, we use \( C' \) to denote the right-hand side of the above equation.

First, we prove that \( C \subseteq C' \). Let \( c \in C \), i.e. \( c \notin C \). Then there are three cases for \( c \):

1. Case 1: \( c_{t-i} c_{t-i-1} \ldots \in D' \). Then \( c_{t-i} c_{t-i-1} \ldots \in D_i' \) for some \( 1 \leq i \leq m \) and \( c_i c_{i+1} \ldots \in E_i' \). Then \( c \notin D_i^r E_i^u \).

2. Case 2: \( c_{t-j} c_{t-j-1} \ldots \in E'' \). Then \( c_{t-j} c_{t-j-1} \ldots \in E_j'' \) for some \( 1 \leq j \leq l \) and \( c_{t-j+1} \ldots \notin D_j' \). Then \( c \notin D_j^r E_j^u \).

3. Case 3: \( c_{t-i} c_{t-i-1} \ldots \notin D' \) and \( c_i c_{i+1} \ldots \notin E'' \). Then \( c \notin D^r E'' \).

The above three cases are not mutually exclusive but they cover all the possibilities. Hence, \( c \in C' \).

Now, we prove that \( C' \subseteq C \), i.e. \( C \subseteq C' \). Let \( c \in C \). Consider (2.1). For each \( 1 \leq i \leq m \), if \( c \in D_i^r E_i^u \) then \( c \notin D_i^r E_i^u \). Similarly, for each \( 1 \leq j \leq l \), if \( c \in D_j^r E_j^u \) then, by the property (3) of these expressions above, \( c \notin D_j^r E_j^u \). So, \( c \notin \bigcup_{i=1}^{m} D_i^r E_i^u \). Similar consideration applies to (2.2). So, \( c \notin \bigcup_{j=1}^{l} D_j^r E_j^u \). Since, for any \( t \in \mathbb{Z} \), \( c_{t-i} c_{t-i-1} \ldots \in D' \) and \( c_i c_{i+1} \ldots \in E'' \), \( c \notin D^r E'' \). Therefore, \( c \in C' \).

By now, we have proved \( C' = \overline{C} \). Since \( C' \) is an \( \omega \omega \) regular set by Lemma 2.4, \( \overline{C} \) is an \( \omega \omega \)-regular set.

**Corollary 2.6.** The family of \( \omega \omega \)-regular sets is closed under union, complementation and intersection.
We conclude this section with two results on the connection between the \( \omega \omega \)-regularity of a set \( C \) of biinfinite words and the regularity of the set \( L[C] \) defined as follows: For a biinfinite word \( c \in S^\infty \),

\[
L[c] = \{ w \in S^* \mid \text{w is a finite subword of } c \},
\]

and, for \( C \subseteq S^\infty \),

\[
L[C] = \bigcup_{c \in C} L[c].
\]

**Theorem 2.7** Let \( C \) be a set of biinfinite words. If \( C \) is \( \omega \omega \)-regular, then \( L[C] \) is a regular language.

**Proof.** Let \( M=(Q,S,\delta,q_0,F) \) be a reduced \( \omega \omega \)-FA such that \( C=B(M) \). We modify \( M \) to produce an FA \( M' \) accepting \( L[C] \), as follows. We add a new start state that has a \( \lambda \)-transition to every state of \( M \), and make every state of \( M \) a final state. Since \( M \) is reduced, every word accepted by \( M' \) can be extended to a biinfinite word in \( C \). By the construction of \( M' \), every subword of a biinfinite word recognized by \( M \) is accepted by \( M' \). Therefore, \( L[C] \) is the language accepted by \( M' \), and thus is regular. \( \square \)

**Theorem 2.8.** If \( R \subseteq S^* \) is a regular set, then the set \( \{ c \in S^\infty \mid L[c] \subseteq R \} \) is \( \omega \omega \)-regular.

**Proof.** Let \( C=\{ c \in S^\infty \mid L[c] \subseteq R \} \). We can assume, without loss of generality, that every subword of every word in \( R \) is in \( R \). Thus there is a finite automaton \( M \) that accepts \( R \) and such that every state in \( M \) is final. Let \( M=(Q,S,\delta,q_0,F) \), where \( q_0 \) is the start state and \( F=Q \) is the set of final states. Assume that every state in \( Q \) can be reached from the start state. Define an \( \omega \omega \)-FA \( M' \) by \( M'=(Q,S,\delta,Q_L,Q_R) \) where \( Q_L=Q_R=Q \). We are going to show that \( C=B(M') \).

To show that \( B(M') \subseteq C \), choose any \( c \in B(M') \) and any finite subword \( w \) of \( c \). Then there exist states \( q_1,q_2 \in Q \) such that \( q_1w \xrightarrow{\delta} q_2 \). Since every state in \( Q \) can be reached from \( q_0 \), we have \( q_0x \xrightarrow{\delta} q_1 \) for some \( x \in S^* \). It follows that \( xw \in R \), and therefore also \( w \in R \), because \( R \) contains every subword of every word in \( R \). We conclude that \( L[c] \subseteq R \), and \( c \in C \).

To show that \( C \subseteq B(M') \), choose any \( c \in C \). We use the infinity lemma [17, p. 383] to prove that there is a biinfinite path in \( M' \) labeled by \( c \). We form an oriented tree in which all the finite paths in \( M' \) labeled by the words \( c(-j)\ldots c(j) \), \( j \geq 0 \), are vertices. A path \( \pi \) (of length \( 2j+3 \)) labeled by \( c(-j)\ldots c(j+1) \) is a son of a path \( \pi' \) (of length \( 2j+1 \)) labeled by \( c(-j)\ldots c(j) \) if \( \pi \) is a concatenation of one transition in \( M' \) followed by \( \pi' \) followed by one transition in \( M' \). All paths (of length 1) labeled by \( c(0) \) are sons of a special root element. The oriented tree is infinite, and every vertex has finite degree. By the infinity lemma there is an infinite path from the root in the tree. Thus there is an infinite sequence of finite paths in \( M' \) labeled by finite
Example 2.9. Let $X$ be the set of all the biinfinite words over $\{a, b\}$ that have a prime number of $a$'s. Let $Y$ be the set of all biinfinite words over $\{a, b\}$. Then $L[X] = L[Y] = \{a, b\}^*$. 

In the above example, $X$ and $Y$ have the same set of finite subwords although $X \neq Y$. Clearly, $Y$ can be recognized by an $\omega\omega$-FA, but $X$ cannot. This example shows that the finite subwords of biinfinite words do not always capture the characteristics of the biinfinite words themselves. This suggests that it is useful to study directly the properties of biinfinite words as well as their relations with finite subwords.

Theorem 2.10. If $C \subseteq S^\mathbb{Z}$ is shift-invariant then the set $\{c \in S^\mathbb{Z} \mid L[c] \subseteq L[C]\}$ is the closure of $C$ in the product topology.

Proof. Let $D = \{c \in S^\mathbb{Z} \mid L[c] \subseteq L[C]\}$. The complement of $D$ in $S^\mathbb{Z}$ is open. Indeed, if $c' \notin D$ then $c'(i) \ldots c'(j) \notin L[C]$ for some $i, j \in \mathbb{Z}$, $i \leq j$. In that case the set 
\[
\{c \in S^\mathbb{Z} \mid c(i) \ldots c(j) = c'(i) \ldots c'(j)\},
\]
which is a neighborhood of $c'$ in the product topology, does not intersect $D$.

Since $D$ is closed and $C \subseteq D$, it follows that the closure $\bar{C}$ of $C$ is a subset of $D$. To prove that $D \subseteq \bar{C}$, choose any $d \in D$. Then for every $j \geq 0$ the word $d(-j) \ldots d(j)$ is a subword of some $c_j \in C$. Since $C$ is shift-invariant, we can choose $c_j$ so that $d(-j) \ldots d(j) = c_j(-j) \ldots c_j(j)$. But then $d$ is the limit of the sequence $(c_j | j = 0, 1, \ldots)$ in the product topology, which proves that $d \in \bar{C}$. 

Corollary 2.11. Let $C \subseteq S^\mathbb{Z}$ be a shift-invariant closed set (a subshift). Then $C$ is $\omega\omega$-regular if and only if $L[C]$ is regular.

Proof. By Theorems 2.10, 2.7 and 2.8.

Corollary 2.12. Let $C_1, C_2 \subseteq S^\mathbb{Z}$ be shift-invariant closed sets (subshifts). Then $C_1 = C_2$ if and only if $L[C_1] = L[C_2]$.

Proof. By Theorem 2.10.

3. Algebraic limits and adherences of languages

The concept of algebraic limits, in terms of biinfinite words, of languages have
been defined in [19,2,10]. Our definition of the adherence of a language, which
defines a set of biinfinite words, is a natural extension of the definition in [5,12],
which defines a set of one-way infinite words.

Let $x, y \in \Sigma^*$. If there exist $u, v \in \Sigma^+$ such that $uxv = y$, then $y$ is called an extension of $x$, denoted $y > x$ or $x < y$. A biinfinite word $\tau$ is called the algebraic limit of an infinite sequence of finite words $x_1, x_2, \ldots, x_n, \ldots$ if $x_1 < x_2 < \cdots < x_n < \cdots$ and $x_i$ is a subword of $\tau$ for all $i \geq 1$.

Let $L$ be an arbitrarily given language. A biinfinite word $\tau$ is called an algebraic limit of the language $L$ if there is an infinite sequence of words $x_1, x_2, \ldots, x_n, \ldots$ in $L$ such that $\tau$ is the algebraic limit of the sequence. The set of all algebraic limits of $L$, denoted $\text{limit}(L)$, is called the algebraic limit set of $L$.

A biinfinite word $\tau$ is called an adherence of a language $L$ if there is an infinite sequence of words $y_1, y_2, \ldots, y_n, \ldots$ such that each $y_i$, $i \geq 1$, is a subword of a word in $L$ and $\tau$ is the algebraic limit of the sequence. The set of all adherences of $L$, denoted $\text{adherence}(L)$, is called the adherence set of $L$.

The next theorem has been proved in [19].

**Theorem 3.1.** The family of algebraic limit sets of regular languages is a proper subset of the family of $\omega\omega$-regular sets.

The following example shows that an $\omega\omega$-regular set is not necessarily the algebraic limit of a regular language. Let $C$ be an $\omega\omega$-regular set accepted by the $\omega\omega$-FA $A = (\{p, q, r\}, \{0, 1\}, \delta, Q_L, Q_R)$ where $Q_L = \{p\}$ and $Q_R = \{r\}$ and $\delta$ is defined by the transition diagram in Fig. 2. Clearly, this $\omega\omega$-regular set is not the algebraic limit of any regular set.

For a language $L \subseteq \Sigma^*$, we define

$$\mathcal{P}(L) = \{x \mid x \text{ is a subword of } w \text{ for some } w \in L\}.$$ 

The next result is obvious.

**Theorem 3.2.** For any language $L$, $\text{adherence}(L) = \text{adherence}(\mathcal{P}(L)) = \text{limit}(\mathcal{P}(L))$.

**Corollary 3.3.** The adherence of any regular language is $\omega\omega$-regular.

**Theorem 3.4.** For any language $L$, adherence($L$) is a topologically closed set.

![Fig. 2. An $\omega\omega$-FA $A$.](image-url)
Proof. Let $D =$ adherence$(L)$. Then $D =$ adherence$(\mathcal{P}(L))$. Let $c \notin D$. Then there is a finite subword $c_i \ldots c_j$, $i < j$, of $c$ such that $c_i \ldots c_j \notin \mathcal{P}(L)$. Clearly, $c$ is in the open set $O = \{ o \mid o_i = c_i, \ldots, o_j = c_j \}$ and adherence$(\mathcal{P}(L)) \cap O = \emptyset$. Hence the complement of adherence$(L)$ is open and adherence$(L)$ is closed. □.

Note that the algebraic limit set of a regular language is not necessarily topologically closed. For example, limit$(a^* b^* c^+)$ is $\omega\omega$-regular but not topologically closed.

Corollary 3.5. The family of adherence sets of regular languages is a proper subset of the family of algebraic limit sets of regular languages.

4. Sofic systems

A set $C \subseteq S^Z$ is called a subshift if it is a shift-invariant and topologically closed set. A subshift $C \subseteq S^Z$ is said to be of finite type if $S^* - L[C]$ is finite [15,24]. A sofic system is the image of a subshift of finite type under a shift-invariant continuous map. The following theorem has been proved by Hurd in [15].

Theorem 4.1. $C \subseteq S^Z$ is a sofic system if and only if $C$ is a subshift and $L[C]$ is a regular language.

The next result is obtained by using the result from the previous section and the above theorem.

Theorem 4.2. The following statements are equivalent:

1. $C \subseteq S^Z$ is a sofic system.
2. $C =$ adherence$(R)$ for some regular language $R$.
3. $C$ is a topologically closed $\omega\omega$-regular set.
4. $C$ is the topologically closure of limit$(R)$ for some regular language $R$.

Proof. (1) ⇒ (3): Let $C$ be a sofic system. By Theorem 4.1, $C$ is a subshift and $L[C]$ is a regular language. Then, by Corollary 2.11, $C$ is $\omega\omega$-regular.

(3) ⇒ (1): Since $C$ is $\omega\omega$-regular, $L[C]$ is a regular language by Theorem 2.7. Then, by Theorem 4.1, $C$ is a sofic system.

(2) ⇒ (4): Obvious.

(4) ⇒ (3): By Theorem 3.1.

(3) ⇒ (2): Let $C' =$ adherence$(L[C])$. It is easy to verify that $L[C'] = L[C]$. Since both $C$ and $C'$ are closed, $C = C'$ by Corollary 2.12. Therefore, $C$ is the adherence of the regular language $L[C]$. □
Theorem 4.2 tells that the family of all sofic systems is exactly the family of adherence sets of regular languages.

5. Bi-extensible subsets of languages

First we introduce three definitions of extensible subsets, and then we show the relationship among these definitions and some properties of these sets. We disprove Hurd's conjecture which states that context-free languages are closed under operation. However, we prove that context-free languages are indeed closed under the operation $E_{\text{sub}}$. At the end, we will show how the bi-extensible sets are related to the limits and adherences of languages.

Given a language $L \subseteq \Sigma^*$, the bi-extensible subset of $L$, denoted $E(L)$, is defined as follows:

$$E(L) = \{ w \in L \mid \text{for any } N > 0, \text{ there exist } u, v \in \Sigma^* \text{ and } |u|, |v| \geq N \text{ such that } uwv \in L \}.$$ 

The serially bi-extensible subset of $L$, denoted $E_{\text{se}}(L)$, is defined as follows:

$$E_{\text{se}}(L) = \{ w \in L \mid \text{there are infinite sequences of nonempty words } u_1, u_2, \ldots \text{ and } v_1, v_2, \ldots \text{ such that } u_1, u_2, \ldots w v_1, v_2, \ldots \in L \text{ for all } t \geq 1 \}.$$ 

The set of extensible subwords of $L$, denoted $E_{\text{sub}}(L)$, is the set of all subwords of words in $L$ that are infinitely bi-extensible. Formally,

$$E_{\text{sub}} = \{ w \in \Sigma^* \mid \text{for any } N > 0, \text{ there exist } u, v \in \Sigma^* \text{ and } |u|, |v| > N \text{ such that } uwv \in L \}.$$ 

The differences among the three definitions can be seen from the next example.

**Example 5.1.** Let $L = \{ ca^i b' c \mid i > 0 \} \cup \{ a^i b \mid j > 0 \}$. Then $E(L) = \{ a^i b \mid j > 0 \}$, $E_{\text{se}}(L) = \emptyset$, and $E_{\text{sub}}(L) = \{ a^i b^j \mid i, j > 0 \}$.

The definition of $E(L)$ has been given in [15], where the sets $E_{\text{se}}(L)$ and $E_{\text{sub}}(L)$ have also been studied but never been formally defined.

For a given $L$, the three sets $E(L)$, $E_{\text{se}}(L)$, and $E_{\text{sub}}(L)$ have the following relations and properties:

$$E_{\text{se}}(L) \subseteq E(L) \subseteq E_{\text{sub}}(L); \quad (A1)$$

$$E_{\text{sub}}(L) \cap L = E(L); \quad (A2)$$

$$\mathcal{P}(E_{\text{sub}}(L)) = E_{\text{sub}}(L); \quad (A3)$$

$$E_{\text{sub}} = \{ w \in \Sigma^* \mid \text{for any } N > 0, \text{ there exist } u, v \in \Sigma^* \text{ and } |u|, |v| > N \text{ such that } uwv \in L \}.$$
Cellular automata, \( \omega \)-regular sets, and sofic systems

\[ \mathcal{E}_{se}(\mathcal{E}_{se}(L)) = \mathcal{E}_{se}(L); \]
\[ \mathcal{E}(\mathcal{F}(L)) = \mathcal{E}_{se}(\mathcal{F}(L)) = \mathcal{E}_{sub}(\mathcal{F}(L)) = \mathcal{E}_{sub}(L). \]

From (A5) we see that if \( \mathcal{F}(L) = L \) then the three definitions are equivalent.

Let \( L_1, L_2 \subseteq \Sigma^* \). Then

\[ \mathcal{E}(L_1 \cup L_2) \supseteq \mathcal{E}(L_1) \cup \mathcal{E}(L_2), \]
\[ \mathcal{E}_{se}(L_1 \cup L_2) \supseteq \mathcal{E}_{se}(L_1) \cup \mathcal{E}_{se}(L_2), \]
\[ \mathcal{E}_{sub}(L_1 \cup L_2) \supseteq \mathcal{E}_{sub}(L_1) \cup \mathcal{E}_{sub}(L_2); \]
\[ \mathcal{E}(L_1 \cap L_2) \subseteq \mathcal{E}(L_1) \cap \mathcal{E}(L_2), \]
\[ \mathcal{E}_{se}(L_1 \cap L_2) \subseteq \mathcal{E}_{se}(L_1) \cap \mathcal{E}_{se}(L_2), \]
\[ \mathcal{E}_{sub}(L_1 \cap L_2) = \mathcal{E}_{sub}(L_1) \cap \mathcal{E}_{sub}(L_2). \]

Similar properties have been stated in [15]. However, in [15] it is incorrectly stated that \( \mathcal{E}(L_1 \cup L_2) = \mathcal{E}(L_1) \cup \mathcal{E}(L_2) \) and \( \mathcal{E}(L_1 \cap L_2) = \mathcal{E}(L_1) \cap \mathcal{E}(L_2) \). The following examples show that the equalities do not necessarily hold.

**Example 5.2.** Let \( L_1 = \{ a^ib^i \mid i \geq 2 \} \) and \( L_2 = \{ ab \} \). Then \( \mathcal{E}(L_1 \cup L_2) = \{ a^ib^i \mid i \geq 1 \} \), but \( \mathcal{E}(L_1) \cup \mathcal{E}(L_2) = \{ a^ib^i \mid i \geq 2 \} \). So, \( \mathcal{E}(L_1 \cup L_2) \neq \mathcal{E}(L_1) \cup \mathcal{E}(L_2) \). For intersection, consider \( L_1 = c^*ab^+ \) and \( L_2 = a^+bc^* \).

Note that, for any integer \( N > 0 \), there exists a regular language \( R \) such that \( \mathcal{E}^N(R) \supset \mathcal{E}_{se}(R) \) but \( \mathcal{E}^N(R) \neq \mathcal{E}_{se}(R) \). However, reader can verify the next result by using the pigeonhole principle.

**Theorem 5.3.** Let \( R \) be a regular language. Then there exists an integer \( N \) for \( R \) such that \( \mathcal{E}_{se}(R) = \mathcal{E}^N(R) \).

In [15], it is proved neatly that the family of regular languages is closed under the \( \mathcal{E} \) operation. Therefore, regular languages are closed under \( \mathcal{E}_{se} \), by the above theorem, and \( \mathcal{E}_{sub} \) as well. Hurd also showed that the family of context-sensitive languages is not closed under \( \mathcal{E} \). For context-free languages, the question has been open, and Hurd conjectured that they are closed under \( \mathcal{E} \) [15]. Our next result shows that Hurd’s conjecture is not true in general. We also show that, however, the conjecture would be true if it was the operator \( \mathcal{E}_{sub} \) rather than \( \mathcal{E} \) that was considered.

**Lemma 5.4.** There exists a context-free language \( L \) such that \( \mathcal{E}(L), \mathcal{E}_{se}(L), \mathcal{F}(\mathcal{E}(L)), \) and \( \mathcal{F}(\mathcal{E}_{se}(L)) \) are not context-free languages.

**Proof.** Let \( L = L(G) \) and \( G = (\Sigma, N, P, S) \) be a context-free grammar where \( \Sigma = \{ a, b, c, \# \}, N = \{ S, D, S_1, S_2, A, B \}, \) and \( P: \)
Let \( R = \# a'b'c' \# \# aabc \# \# abcc \# \). We claim that
\[
\mathcal{E}(L) \cap R = \{ \# a^ib^ic^i \# \# aabc \# \# abcc \# | i > 0 \}
\]
which is not a context-free language. Let \( L_R \) denote the language on the right side of the equation. First we show that \( L_R \subseteq \mathcal{E}(L) \cap R \). It is clear that \( L_R \subseteq L \cap R \). So, it suffices to show that every word \( w \in L_R \) is bi-extensible in \( L \), i.e. for any integer \( t > 0 \) there exist \( u, v \in \Sigma^* \), \( |u|, |v| \geq t \), such that \( uvw \in L \). Let \( w \) be an arbitrary word in \( L_R \). Then for any given \( t > 0 \) we can choose \( u = \# a'b'c# \# a'b'c# \) and \( v = \# abc' \), and it is clear that \( uvw \in L \). So, \( w \in \mathcal{E}(L) \). Now we show that \( \mathcal{E}(L) \cap R \subseteq L_R \), i.e. \( w \in L_R \) implies \( w \in \mathcal{E}(L) \cap R \). We can restrict our attention to the words of the form \( w = a'b^ic^j aabcabcc \) where \( i \neq j \) or \( j \neq k \). If \( i \neq j \), then \( w \notin L \) and therefore \( w \notin \mathcal{E}(L) \). If \( j \neq k \), then \( w \) is not left extensible and, therefore, \( w \notin \mathcal{E}(L) \). So, we can conclude that \( \mathcal{E}(L) \cap R = L_R \). It is clear that \( L_R \) is not context-free. By the closure property of context-free languages, \( \mathcal{E}(L) \) is not a context-free language. It is easy to verify that \( \mathcal{E}_n(L) \cap R = \mathcal{E}(\mathcal{E}(L)) \cap R = \mathcal{E}(\mathcal{E}_n(L)) \cap R = L_R \). So, \( \mathcal{E}_n(L) \), \( \mathcal{E}(\mathcal{E}(L)) \) and \( \mathcal{E}(\mathcal{E}_n(L)) \) are not context-free languages.  

**Theorem 5.5.** The family of context-free languages is not closed under the following operations: \( \mathcal{E}, \mathcal{E}_n \), \( \mathcal{J}, \mathcal{E} \) and \( \mathcal{J}, \mathcal{E}_n \).

The following is an extension of the iteration theorem for context-free languages [11]. It will be used in the proof of our next theorem.

**Lemma 5.6.** Let \( G = (\Sigma, N, P, S) \) be a context-free grammar and \( L = L(G) \). Then for any integer \( t > 0 \), there exists an integer \( p(t) \) such that for each \( z \in L \) and any set \( D \) of distinguished positions in \( z \), if \( |D| \geq p(t) \), then there is a decomposition

\[
z = u x_1 \ldots x_l w y_1 \ldots y_j v
\]
such that:

1. There exists \( A \in N \) such that
\[
A \Rightarrow \ast u Av \Rightarrow \ast u x_1 A y_1 v \Rightarrow \cdots \Rightarrow u x_1 \ldots x_l A y_1 \ldots y_j v \Rightarrow u x_1 \ldots x_l w y_1 \ldots y_j v.
\]
2. For any \( i_1, \ldots, i_l \geq 0 \),
\[
ux_1^{i_1} \ldots x_l^{i_l} w y_1^{j_1} \ldots y_l^{j_l} v \in L.
\]
3. Let \( K(x) \) denote the distinguished positions of \( K \) in \( x \). Then
Theorem 5.7. The family of context-free languages is closed under the \( \epsilon \) _sub_ operation.

Proof. Let \( G = (\Sigma, N, P, S) \) be a context-free grammar and \( L = L(G) \). Without loss of generality, we assume that \( G \) is reduced and \( \epsilon \)-free. We define a grammar \( G' = (\Sigma \cup \{\$\}, N, P', S) \), which is an augmented grammar of \( G \), where

\[
P' = P \cup \{ A \rightarrow \$A \mid A = *uAv \text{ in } G \text{ and } u, v \in \Sigma^+ \}
\]

\[
\cup \{ A \rightarrow A \mid A = *uA \text{ in } G \text{ and } u \in \Sigma^+ \}
\]

\[
\cup \{ A \rightarrow AS \mid A = *uAv \text{ in } G \text{ and } v \in \Sigma^+ \}
\]

\[
\cup \{ A \rightarrow S \mid A = *uAv \text{ in } G \text{ and } uv \in \Sigma^+ \}
\]

Let \( L' = L(G') \) and \( L'' = \{ w \mid u_1$u_2$w_3$u_4 \in L' \} \). Then we claim that \( \epsilon _{\text{sub}}(L) = L'' \) and \( L'' \) is context-free. First, we prove that \( \epsilon _{\text{sub}}(L) = L'' \).

Let \( w \in L'' \). Then there exist \( u_1, u_2, u_3, u_4 \in \Sigma^* \) such that \( u_1$u_2$w_3$u_4 \in L(G') \). Consider the following two cases:

(i) The two \( \$'s \) are derived by a production \( A \rightarrow \$A \) for some \( A \in N \). Then there exist \( x, y \in \Sigma^* \) such that \( u_1xx$u_2$w_3y$u_4 \in L \) for all \( i \geq 0 \). Therefore, \( w \in \epsilon _{\text{sub}}(L) \).

(ii) The two \( \$'s \) are derived separately by the productions of the form \( A \rightarrow \$A \), \( A \rightarrow A\$, or \( A \rightarrow \$. Let \( A \) and \( B \) be the two nonterminals that derive the two \( \$'s \), respectively. If \( A \rightarrow \$A \) or \( A \rightarrow A\$, then the \( \$ \) can be replaced by \( x' \) for some \( x \in \Sigma^+ \) and for all \( i \geq 0 \). If \( A \rightarrow \$, then for any integer \( n > 0 \) there exists a word \( x \in \Sigma^+ \) such that \( |x| \geq n \) and \( A \Rightarrow *x \). The same argument applies on the second \( \$ \) as well. Then \( w \in \epsilon _{\text{sub}}(L) \).

Above are the only possible two cases. Therefore, \( w \in \epsilon _{\text{sub}}(L) \). Since \( w \) is arbitrary, \( L'' \subseteq \epsilon _{\text{sub}}(L) \).

Let \( z \in \epsilon _{\text{sub}}(L) \) and let \( |z| = t \). Choose \( z_1, z_2 \in \Sigma^+ \) such that \( |z_1|, |z_2| \geq p(t) \) and \( z' = z_1zz_2 \in L \), where \( p(t) \) is the number in Lemma 5.6. Consider all the positions in \( z_1 \). By Lemma 5.6, we have

\[
S = *uAv = \ldots = ux_1 \ldots x_iAy_1 \ldots y_iu \Rightarrow ^+
\]

\[
ux_1 \ldots x_iwy_1 \ldots y_iu = z'.
\]

In the following, we use \( x \preceq y \) to denote that \( x \) is a subword (not necessarily a proper subword) of \( y \). There are two cases:

Case 1: \( K(u), K(x_1), \ldots, K(x_i), K(w) \neq \emptyset \).

(a) If \( z \preceq wy_1 \ldots y_i \) and \( A \rightarrow \$A \) in \( G' \), then \( u\$x_1 \ldots x_iwy_1 \ldots y_iu \) in \( L' \) and therefore \( z \) is in \( L'' \).
Fig. 3. Finite transducer $M$. (Note: "a" denotes any symbol in $\Sigma$ and $\$ \notin \Sigma$.)

(b) If $z \leq wy_1 \ldots y_1$ and $A \not\to \$A\$, then $y_1 \ldots y_1 = \varepsilon$ and $A = \$A$ in $G'$. So, we can insert a $\$ between $u$ and $x_1$ and the word is still in $L(G')$.

(c) If $z \leq y_1 \ldots y_1 \nu$, then $w$ can be replaced by a $\$.

(d) Otherwise, $z \leq wy_1 \ldots y_1 \nu$ and $z$ crosses $w, y_1 \ldots y_1$, and $\nu$. Then $|y_1 \ldots y_1| < |z| = t$, and then there is $1 \leq i \leq t$ such that $i = \varepsilon$. So, we can replace $x_i$ by a $\$.

**Case 2:** $K(w), K(y_1), \ldots, K(y_1), K(\nu) \neq \emptyset$. Then $z \leq \nu$. We can replace $w$ by a $\$.

From the above argument on the positions in $z_1$, we know that there can be either two $\$’s on both sides of $z$ or one $\$ on the left side of $z$. We apply a symmetric argument on the positions in $z_2$. So, for any $z \in \mathcal{E}_{\text{sub}}(L)$ there exists a word of the form $u_1 \$ u_2 z u_3 \$ u_4 \in L'$. Therefore, $z \in L''$ and $\mathcal{E}_{\text{sub}}(L) \subseteq L''$.

We should still show that $L''$ is context-free. We construct a finite transducer $M$ as shown in Fig. 3. Obviously, $L'' = M(L')$. Since context-free languages are closed under finite transduction, $L''$ is context-free. Thus, $\mathcal{E}_{\text{sub}}(L)$ is context-free. $\square$

The next theorem shows the relationship between the bi-extensible subsets of languages and the limit and adherence sets of languages.

**Theorem 5.8.** For any language $L_0$, the following relations hold:

1. $L[\text{limit}(L_0)] = \mathcal{P}(\mathcal{E}_{\text{sub}}(L_0))$.
2. $L[\text{adherence}(L_0)] = \mathcal{P}(\mathcal{E}_{\text{sub}}(L_0))$.

By the above result, we can again show that if $L_0$ is a regular language, then both $L[\text{limit}(L_0)]$ and $L[\text{adherence}(L_0)]$ are regular languages. For an arbitrary context-free language $L_1$, $L[\text{limit}(L_1)]$ may not be context-free. The context-free language given in Lemma 5.4 is a counterexample. However, $L[\text{adherence}(L_1)]$ is context-free for every context-free language $L_1$. Since sofic systems are adherence sets of regular languages, it is natural to define another hierarchy of sets of biinfinite words related to context-free languages. We call it $CF$ systems.

**Definition 5.9.** A set $C \subseteq \Sigma^\mathbb{Z}$ is called a $CF$ system if $C = \text{adherence}(L)$ for some context-free language $L$.

**Theorem 5.10.** A set $C \subseteq \Sigma^\mathbb{Z}$ is a $CF$ system if and only if $L[C]$ is context-free and $C$ is a subshift.
References