Decision Problems For Convex Languages

Janusz Brzozowski, Jeffrey Shallit, and Zhi Xu

David R. Cheriton School of Computer Science,
University of Waterloo, Waterloo, ON, Canada N2L 3G1
{brzozo,shallit,z5xu}@uwaterloo.ca

Abstract. We examine decision problems for various classes of convex languages, previously studied by Ang and Brzozowski under the name "continuous languages". We can decide whether a language L is prefix-, suffix-, factor-, or subword-convex in polynomial time if L is represented by a DFA, but the problem is PSPACE-hard if L is represented by an NFA. If a regular language is not convex, we prove tight upper bounds on the length of the shortest words demonstrating this fact, in terms of the number of states of an accepting DFA. Similar results are proved for some subclasses of convex languages: the prefix-, suffix-, factor-, and subword-closed languages, and the prefix-, suffix-, factor-, and subword-free languages.

1 Introduction

A word x is a *factor* of a word w if $w = uxv$ for some words u and v. A word x is a *subword* of w if x is a subsequence of w. Thierrin [1] introduced convex languages with respect to the subword relation, and Ang and Brzozowski [2] generalized this concept to arbitrary relations.

A language L is *prefix-convex* if $u, w \in L$ with u a prefix of w implies that any word v must also be in L if u is a prefix of v and v is a prefix of w. L is *prefix-free* if $w \in L$ implies that no proper prefix of w is in L. L *prefix-closed* if $w \in L$ implies that every prefix of w is also in L.

Similar definitions hold for suffix-, factor-, and subword-convex languages, and suffix-, factor-, and subword-free and closed languages. Prefix-free languages (prefix codes) were studied by Berstel and Perrin [3]. Han has recently considered X-free languages for various values of X, such as prefix, suffix, factor and subword [4]. A factor-closed language is often called *factorial*.

We consider the computational complexity of testing whether a given language is prefix-convex, suffix-convex, etc., prefix-closed, suffix-closed, etc., for a total of 12 different problems. The computational complexity of these decision problems depends on how the language is represented. If it is specified by a DFA, the decision problem is solvable in polynomial time. If it is represented as a regular expression or an NFA, the decision problem is PSPACE-complete. We also consider the following question: given that a language is *not* prefix-convex, suffix-convex, etc., what is a good upper bound on the length of the shortest words (*witnesses*) demonstrating this fact?

In Section 2 we study the complexity of testing for convexity for languages represented by DFA's, and include testing for closure and freeness as special cases. In Section 3 we exhibit shortest witnesses to the lack of convexity. Convex languages specified by NFA's and context-free grammars are briefly studied in Section 4. Section 5 concludes the paper. Owing to the space constraints, we have had to omit many results and proofs; they can be found in the full version of our paper [5].

2 Decision problems for languages specified by DFA's

We will show that, if a regular language L is represented by a DFA M with n states, it is possible to test the property of prefix-, suffix-, factor-, and subword-convexity efficiently, in fact, in $O(n^3)$ time.

Let $\trianglelefteq$ be one of the four relations *prefix, suffix, factor,* or *subword.* The basic idea is as follows: L is *not* $\trianglelefteq$-convex if and only if there exist words $u, w \in L$, $v \notin L$, such that $u \trianglelefteq v \trianglelefteq w$. Given M, we create an NFA-ϵ M' with $O(n^3)$ states and transitions that accepts the language $\{w \in L(M) \ : \$ there exist $u \in L(M), v \notin L(M)$ such that $u \trianglelefteq v \trianglelefteq w\}$. Then $L(M') = \emptyset$ if and only if $L(M)$ is $\trianglelefteq$-convex. We can test the emptiness of $L(M')$ using depth-first search in time linear in the size of M'. This gives an $O(n^3)$ algorithm for testing the $\trianglelefteq$-convexity.

Since the constructions for all four properties are similar, we handle the hardest case (factor-convexity) in detail, and refer the reader to [5] for the rest.

Factor-convex languages Suppose $M = (Q, \Sigma, \delta, q_0, F)$ is a DFA accepting the language $L = L(M)$, and suppose M has n states. We construct an NFA-ϵ M' such that $L(M')$ is the set of words $w \in \Sigma^*$ such that there exist $u, v \in \Sigma^*$ such that u is a factor of v, v is a factor of w, and $u, w \in L, v \notin L$. Clearly $L(M') = \emptyset$ if and only if $L(M)$ is factor-convex.

States of M' are quadruples, where components 1, 2, and 3 keep track of where M is upon processing w, v, and u (respectively). The last component is a flag indicating the present *mode* of the simulation process. Formally, $M' = (Q', \Sigma, \delta', q_0', F')$, where $Q' = Q \times Q \times Q \times \{1, 2, 3, 4, 5\}$, $q_0' = [q_0, q_0, q_0, 1]$, $F' = F \times (Q - F) \times F \times \{5\}$, and

1. $\delta'([p, q_0, q_0, 1], a) = \{[\delta(p, a), q_0, q_0, 1]\}$, for all $p \in Q, a \in \Sigma$;
2. $\delta'([p, q_0, q_0, 1], \epsilon) = \{[p, q_0, q_0, 2]\}$, for all $p \in Q$;
3. $\delta'([p, q, q_0, 2], a) = \{[\delta(p, a), \delta(q, a), q_0, 2]\}$, for all $p, q \in Q, a \in \Sigma$;
4. $\delta'([p, q, q_0, 2], \epsilon) = \{[p, q, q_0, 3]\}$, for all $p, q \in Q$;
5. $\delta'([p, q, r, 3], a) = \{[\delta(p, a), \delta(q, a), \delta(r, a), 3]\}$, for all $p, q, r \in Q, a \in \Sigma$;
6. $\delta'([p, q, r, 3], \epsilon) = \{[p, q, r, 4]\}$, for all $p, q, r \in Q$;
7. $\delta'([p, q, r, 4], a) = \{[\delta(p, a), \delta(q, a), r, 4]\}$, for all $p, q, r \in Q, a \in \Sigma$;
8. $\delta'([p, q, r, 4], \epsilon) = \{[p, q, r, 5]\}$, for all $p, q, r \in Q$;
9. $\delta'([p, q, r, 5], a) = \{[\delta(p, a), q, r, 5]\}$, for all $p, q, r \in Q, a \in \Sigma$.

One can verify that the contruction is correct, and that the NFA-ϵ M' has $3n^3 + n^2 + n$ states and $(3|\Sigma| + 2)n^3 + (|\Sigma| + 1)(n^2 + n)$ transitions, where $|\Sigma|$ is the cardinality of Σ [5]. In other words, the following theorem holds:

Theorem 1. *If M is a DFA with n states, there exists an NFA-ϵ M' with $O(n^3)$ states and transitions such that M' accepts the language $L(M') = \{w \in \Sigma^* :$ there exist $u, v \in \Sigma^*$ such that u is a factor of v, v is a factor of w, and $u, w \in L, v \notin L\}$.*

Corollary 1. *We can decide if a given regular language L accepted by a DFA with n states is factor-convex in $O(n^3)$ time.*

Factor-closed languages The language L is *not* factor-closed if and only if there exist words v, w such that v is a factor of w, and $w \in L$, while $v \notin L$. Given a DFA M accepting L, we construct an NFA-ϵ M' such that $L(M') = \{w \in \Sigma^* :$ there exists $v \in \Sigma^*$ such that v is a factor of w, and $w \in L, v \notin L\}$. Then $L(M') = \emptyset$ if and only if $L(M)$ is factor-closed. The size of M' is $O(n^2)$.

States of M' are triples, where components 1 and 2 keep track of where M is upon processing w and v (respectively). The last component is a flag as before. Formally, $M' = (Q', \Sigma, \delta', q_0', F')$, where $Q' = Q \times Q \times \{1, 2, 3\}$; $q_0' = [q_0, q_0, 1]$; $F' = F \times (Q - F) \times \{3\}$; and

1. $\delta'([p, q_0, 1], a) = \{[\delta(p, a), q_0, 1]\}$ for $p \in Q$, $a \in \Sigma$.
2. $\delta'([p, q_0, 1], \epsilon) = \{[p, q_0, 2]\}$, for all $p \in Q$;
3. $\delta'([p, q, 2], a) = \{[\delta(p, a), \delta(q, a), 2]\}$, for all $p, q \in Q$;
4. $\delta'([p, q, 2], \epsilon) = \{[p, q, 3]\}$, for all $p, q \in Q$;
5. $\delta'([p, q, 3], a) = \{[\delta(p, a), q, 3]\}$, for $p, q \in Q$, $a \in \Sigma$.

M' has $2n^2 + n$ states and $(2|\Sigma| + 1)n^2 + (|\Sigma| + 1)$ transitions. Thus we have Theorem 2. (This result was previously obtained by Béal et al. [6, Prop. 5.1, p. 13] through a slightly different approach.)

Theorem 2. *We can decide if a given regular language L accepted by a DFA with n states is factor-closed in $O(n^2)$ time.*

The converse of the relation "u is a factor of v" is "v contains u as a factor". This relation and similar converse relations derived from the prefix, suffix, and subword relations, lead to "converse-closed languages" [2]. Subword-closed and converse-subword-closed languages were characterized by Thierrin [1]. It has been shown by de Luca and Varricchio [7] that a language L is factor-closed (factorial, in their terminology) if and only if it is a complement of an ideal, that is, if and only if $L = \overline{\Sigma^* K \Sigma^*}$ for some $K \subseteq \Sigma^*$. Ang and Brzozowski [2] noted that a language is an ideal if and only if it is converse-factor-closed, that is, if, for every $u \in L$, each word of the form $v = xuy$ is also in L. Thus, to test whether L is converse-factor-closed, we must check that there is no pair (u, v) such that $u \in L$, $v \notin L$, and u is a factor of v. This is equivalent to testing whether $\overline{L}$ is factor-closed. Then the following is an immediate consequence of Theorem 1:

Corollary 2. *We can decide if a given regular language L accepted by a DFA with n states is an ideal in $O(n^2)$ time.*

Factor-free languages Factor-free (also known as infix-free) languages have been studied recently by Han et al. [8], who gave efficient algorithms for determining if the language accepted by an NFA is prefix-, suffix-, or factor-free. We can decide whether a DFA language is factor-free in $O(n^2)$ time with the automaton we used for testing factor-closure, except that the set of accepting states is now $F' = F \times F \times \{3\}$.

3 Minimal witnesses

Let $\trianglelefteq$ represent one of the four relations: factor, prefix, suffix, or subword. A necessary and sufficient condition that a language L be *not* $\trianglelefteq$-convex is the existence of a triple (u, v, w) of words, where $u, w \in L$, $v \notin L$, $u \trianglelefteq v$, and $v \trianglelefteq w$. We call such a triple a *witness* to the lack of $\trianglelefteq$-convexity. A witness (u, v, w) is *minimal* if every other witness (u', v', w') satisfies $|w| < |w'|$, or $|w| = |w'|$ and $|v| < |v'|$, or $|w| = |w'|$, $|v| = |v'|$, and $|u| < |u'|$. The *size* of a witness is $|w|$.

Similarly, if L is not $\trianglelefteq$-closed, then (v, w) is a *witness* if $w \in L$, $v \notin L$, and $v \trianglelefteq w$. A witness (v, w) is *minimal* if there exists no witness (v', w') such that $|w'| < |w|$, or $|w'| = |w|$ and $|v'| < |v|$. The *size* is again $|w|$. For $\trianglelefteq$-freeness, *witness, minimal witness,* and *size* are defined as for $\trianglelefteq$-closure, except that both words are in L.

Suppose we are given a regular language L specified by an n-state DFA M, and we know that L is not $\trianglelefteq$-convex (respectively, $\trianglelefteq$-closed or $\trianglelefteq$-free). A natural question then is, what is a good upper bound on the size of the shortest witness that demonstrates the lack of this property?

3.1 Factor-convexity

From Theorem 1, we deduce Corollary 3, which gives an $O(n^3)$ upper bound for the length of a witness to the lack of factor-convexity. This bound is best possible, as is shown in Theorem 3, whose proof appears in Section 3.3.

Corollary 3. *Suppose L is accepted by a DFA with n states and L is not factor-convex. Then there exists a witness (u, v, w) such that $|w| \leq 3n^3 + n^2 + n - 1$.*

Theorem 3. *There is a class of non-factor-convex regular languages L_n, accepted by DFA's with $O(n)$ states, such the size of the minimal witness is $\Omega(n^3)$.*

Factor-closure Theorem 2 gives us a $O(n^2)$ upper bound on the length of a witness to the failure of the factor-closed property:

Corollary 4. *If L is accepted by a DFA with n states and L is not factor-closed, then there exists a witness (v, w) such that $|w| \leq 2n^2 + n - 1$.*

This $O(n^2)$ upper bound is best possible. Let $M = (Q, \Sigma, \delta, q_0, F)$ be a DFA, where $Q = \{q_0, q_1, \cdots, q_n, q_{n+1}, p_0, p_1, \cdots, p_n, p_{n+1}\}$, $\Sigma = \{\mathtt{0}, \mathtt{1}\}$, and $F = Q \setminus \{q_{n+1}\}$. The transition function is

$\delta(q_0, \mathtt{0}) = q_0$, $\delta(q_0, \mathtt{1}) = q_1$, $\delta(q_{n+1}, \mathtt{0}) = q_{n+1}$, $\delta(q_{n+1}, \mathtt{1}) = q_{n+1}$,

$$\delta(q_i, \mathtt{0}) = \begin{cases} q_{i+1}, & \text{if } 0 < i < n; \\ q_1, & \text{if } 0 < i = n, \end{cases}$$

$$\delta(q_i, \mathtt{1}) = \begin{cases} q_1, & \text{if } 0 < i < n-1; \\ p_0, & \text{if } 0 < i = n-1; \\ q_{n+1}, & \text{if } 0 < i = n; \end{cases}$$

$$\delta(p_j, \mathtt{0}) = \begin{cases} p_{j+1}, & \text{if } 0 \leq j < n; \\ q_0, & \text{if } 0 \leq j = n; \end{cases}$$

$$\delta(p_j, \mathtt{1}) = \begin{cases} q_{n+1}, & \text{if } 0 \leq j < n; \\ p_{n+1}, & \text{if } 0 \leq j = n; \end{cases}$$

and $\delta(p_{n+1}, \mathtt{0}) = q_{n+1}$, $\delta(p_{n+1}, \mathtt{1}) = q_{n+1}$. The DFA M has $2n + 4$ states. The following theorem holds [5]:

Theorem 4. *For the DFA M above, let $L = L(M)$. For any witness (u, v) to the lack of factor-closure we have $|v| \geq (n+1)^2 - 1$, and this bound is achievable.*

Factor-freeness From the remarks at the end of Section 2, we get

Corollary 5. *If L is accepted by a DFA with n states and L is not factor-free, then there exists a witness (v, w) such that $|w| \leq 2n^2 + n - 1$.*

Up to a constant, Corollary 5 is best possible, as the following theorem shows.

Theorem 5. *There is a class of languages accepted by DFA's with $O(n)$ states, such that the smallest witness to the lack of factor-freeness is of size $\Omega(n^2)$.*

Proof. Let $L = \mathtt{bb}(\mathtt{a}^n)^+\mathtt{b} \cup \mathtt{b}(\mathtt{a}^{n+1})^+\mathtt{b}$. This language can be accepted by a DFA with $2n + 6$ states. However, the shortest witness to lack of factor-freeness is $(\mathtt{ba}^{n(n+1)}\mathtt{b}, \mathtt{bba}^{n(n+1)}\mathtt{b})$, which has size $n^2 + n + 3$. □

3.2 Prefix-convexity

For prefix-convexity, we have the following theorem.

Theorem 6. *Let M be a DFA with n states. If $L(M)$ is not prefix-convex, there is a witness (u, v, w) with $|w| \leq 2n - 1$. Furthermore, this bound is best possible, as for all $n \geq 2$, there exists a unary DFA with n states that achieves this bound.*

Proof. If $L(M)$ is not prefix-convex, then such a witness (u, v, w) exists. Without loss of generality, assume that (u, v, w) is minimal. Now write $w = uyz$, where $v = uy$ and $w = vz$.

Let $\delta(q_0, u) = p$, $\delta(p, y) = q$, and $\delta(q, z) = r$. Let P be the path from q_0 to r traversed by uvw, and let P_1 be the states from q_0 to p (not including p), P_2 be the states from p to q (not including q), and P_3 be the states from q to r (not including r); see Figure 1. Since (u, v, w) is minimal, we know that every state of P_3 is rejecting, since we could have found a shorter w if there were an accepting state among them. Similarly, every state of P_2 must be accepting, for, if there were a rejecting state among them, we could have found a shorter y and hence a shorter v. Finally, every state of P_1 must be rejecting, since, if there were an accepting state, we could have found a shorter u.

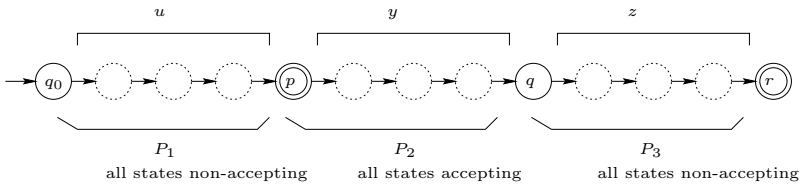

Fig. 1. The acceptance path for w

Let $r_i = |P_i|$ for $i = 1, 2, 3$. There are no repeated states in P_3, for if there were, we could cut out the loop to get a shorter w; the same holds for P_2 and P_1. Thus $r_i \leq n - 1$ for $i = 1, 2, 3$. Now P_1 and P_2 are disjoint, since all the states of P_1 are rejecting, while all the states of P_2 are accepting. Similarly, the states of P_3 are disjoint from P_2. So $r_1 + r_2 \leq n$ and $r_2 + r_3 \leq n$. It follows that $r_1 + r_2 + r_3 \leq 2n - r_3$. Since $r_3 \geq 1$, it follows that $|w| \leq 2n - 1$.

To see that $2n - 1$ is optimal, consider the DFA of n states accepting the unary language $L = \mathtt{a}^{n-1}(\mathtt{a}^n)^*$. Then L is not prefix-convex, and the shortest witness is $(\mathtt{a}^{n-1}, \mathtt{a}^n, \mathtt{a}^{2n-1})$. □

Prefix-closure For prefix-closed languages we can get an even better bound.

Theorem 7. *Let M be an n-state DFA, and suppose $L = L(M)$ is not prefix-closed. Then the minimal witness (v, w) showing L is not prefix-closed has $|w| \leq n$, and this is best possible.*

Proof. Assume that (v, w) is a minimal witness. Consider the path P from q_0 to $q = \delta(q_0, w)$, passing through $p = \delta(q_0, v)$. Let P_1 denote the part of the path P from q_0 to p (not including p) and P_2, the part of the path from p to q (not including q). Then all the states traversed in P_2 must be rejecting; otherwise, we would get a shorter w. Similarly, all the states traversed in P_1 must be accepting, because otherwise we could get a shorter v. Neither P_1 nor P_2 contains a repeated

state, because if they did, we could "cut out the loop" to get a shorter v or w. Furthermore, the states in P_1 are disjoint from P_2. So the total number of states in the path to w (not counting q) is at most n. Thus $|w| \leq n$.

The result is best possible, as the example of the unary language $L = (\mathtt{a}^n)^*$ shows. This language is not prefix-closed, can be accepted by a DFA with n states, and the smallest witness is $(\mathtt{a}, \mathtt{a}^n)$. □

Prefix-freeness For the prefix-free property we have:

Theorem 8. *If L is accepted by a DFA with n states and is not prefix-free, then there exists a witness (v, w) with $|w| \leq 2n - 1$. The bound is best possible.*

Proof. The proof is similar to that of Theorem 6. The bound is achieved by a unary DFA accepting $\mathtt{a}^{n-1}(\mathtt{a}^n)^*$. □

3.3 Suffix-convexity

For the suffix-convex property, the cubic upper bound implied by Corollary 3 is best possible, up to a constant factor.

Theorem 9. *There is a class of non-suffix-convex regular languages L_n, accepted by DFA's with $O(n)$ states, such the size of the minimal witness is $\Omega(n^3)$.*

Proof. Let $L = \mathtt{bbb}(\mathtt{a}^{n-1})^+ \cup \mathtt{bb}(\mathtt{a} + \mathtt{aa} + \cdots + \mathtt{a}^{n-1})(\mathtt{a}^n)^* \cup \mathtt{b}(\mathtt{a}^{n+1})^+$. Then L can be accepted by a DFA with $3n + 5$ states, as illustrated in Figure 2.

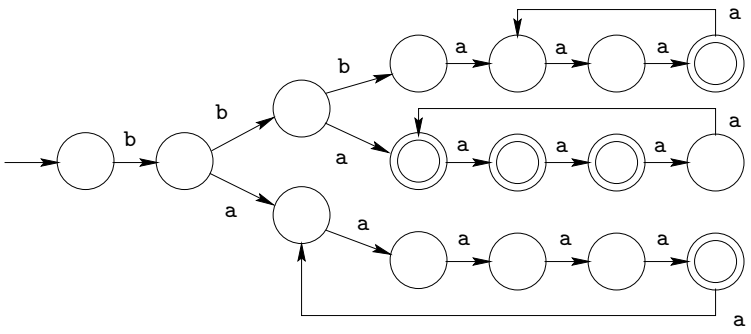

Fig. 2. Example of the construction in Theorem 9 for $n = 4$. All unspecified transitions go to a rejecting "dead state" d that cycles on all inputs.

It can be verified that L is not suffix-convex and the shortest witness is $(\mathtt{ba}^i, \mathtt{bba}^i, \mathtt{bbba}^i)$, where $i = \mathrm{lcm}(n-1, n, n+1) \geq (n-1)n(n+1)/2$. □

A similar technique can be used for non-factor-convex languages. This allows us to prove Theorem 3 in the same way Theorem 9, except we use the language $L\mathtt{b}$ instead.

Suffix-closure Obviously, a witness to the failure of suffix-closure is also a witness to the failure of factor-closure. So the proof of Theorem 4 shows that the bound $(n+1)^2 - 1$ also holds for suffix-closed languages.

Ang and Brzozowski pointed out [2] that a language L is factor-closed if and only if L is both prefix-closed and suffix-closed. The next result [5] shows that a long minimal witness for factor-closure must also be a witness for suffix-closure.

Proposition 1. *Let M be a DFA of n states, and $L = L(M)$. Let v be the shortest word such that there is $u \notin L, v \in L, |v| > n$ and u is a factor of v. Then u is a suffix of v.*

Suffix-freeness

Theorem 10. *There exists a class of languages accepted by DFA's with $O(n)$ states, such that the smallest witness to the lack of suffix-freeness is of size $\Omega(n^2)$.*

Proof. Let $L = \mathtt{bb}(\mathtt{a}^n)^+ \cup \mathtt{b}(\mathtt{a}^{n+1})^+$. This language is accepted by a DFA with $2n + 5$ states. However, the shortest witness to the lack of suffix-freeness $(\mathtt{ba}^{n(n+1)}, \mathtt{bba}^{n(n+1)})$ has size $n^2 + n + 2$. □

3.4 Subword-convexity

We now turn to subword properties. First, we recall some facts about the pumping lemma. If $w = a_1 \cdots a_m$ with $a_i \in \Sigma$ for $1 \leq i \leq m$, we write $w[i, j]$ for the factor $a_i \cdots a_j$. Assume that $M = (Q, \Sigma, \delta, q_0, F)$ is an n-state DFA, $m \geq n$, let $q \in Q$, and consider the state sequence $S(q, w) = (\delta(q, w[1, 0]), \ldots, \delta(q, w[1, m]))$. We know that some state in $S(q, w)$ must appear more than once, because there are only n distinct states in M. Let $\delta(q, w[1, i])$ be the first state that appears more than once in S, and let $x = w[1, i]$. Moreover, let $\delta(q, w[1, j])$ be the first state in $S(q, w)$ equal to $\delta(q, w[1, i])$, and let $y = w[i + 1, j]$. Finally, let $z = w[j + 1, m]$. Then $w = xyz$, where $|xy| \leq n$, $|y| > 0$, and $|z| \geq m - n$, and $\delta(q, x) = \delta(q, xy)$. By the pumping lemma, $xy^*z \subseteq L$. By the definition of x and y, all the states in the sequence $S(q, w[1, j - 1])$ are distinct. For a word w with $|w| = m \geq n$, we refer to the factorization $w = xyz$ as the *canonical factorization of w with respect to q*.

Subword-closure Here $v \trianglelefteq w$ means v is a subword of w. If $L = L(M)$ is not subword-closed, then (v, w) is a *witness* if $w \in L$, $v \notin L$, and $v \trianglelefteq w$.

Lemma 1. *Let M be a DFA with $n \geq 2$ states such that $L(M)$ is not subword-closed. For any witness (v, w), there exists a witness (v', w') with $|w'| \leq n$ and $w' \trianglelefteq w$.*

Proof. We will show that, for any witness (v, w) with $|w| \geq n+1$, we can find a witness (v', w') with $|w'| < |w|$ and $w' \trianglelefteq w$. The lemma then follows.

Suppose that (v, w) is a minimal witness, and $|w| = m \geq n+1$. Then the canonical factorization of w is $w = xyz$, where $|xy| \leq n$, $|y| > 0$, and $|z| \geq m - n > 0$.

If there is a z' such that $z' \trianglelefteq z$ and $xyz' \notin L$, then $xz' \notin L$, since xyz' and xz' lead to the same state in M. Then (xz', xz) is a witness with $|xz| < |w|$ and $xz \trianglelefteq w$. Thus we can assume that

$$z' \trianglelefteq z \text{ implies } xyz' \in L. \tag{1}$$

Since $v \trianglelefteq w = xyz$, we can write $v = v_x v_y v_z$, where $v_x \trianglelefteq x$, $v_y \trianglelefteq y$, and $v_z \trianglelefteq z$. Clearly, $v \trianglelefteq xyv_z$. If $v_z \neq z$, then by (1), we have $xyv_z \in L$, and (v, xyv_z) is a witness with $|xyv_z| < |w|$ and $xyv_z \trianglelefteq w$. Thus we may assume that our witness has the form $(v_x v_y z, xyz)$.

In the particular case that $z' = \epsilon$, (1) implies that $xy \in L$. If $y' \trianglelefteq y$ and $xy' \notin L$, then (xy', xy) is a witness with $|xy| < |w|$ and $xy \trianglelefteq w$. Thus $y' \trianglelefteq y$ implies $xy' \in L$.

Finally, if $x' \trianglelefteq x$ and $x' \notin L$, then (x', x) is a witness with $|x| < |w|$ and $x \trianglelefteq w$. Thus $x' \trianglelefteq x$ implies $x' \in L$. Altogether, we may assume that all the states along the path spelling w in M are accepting. We know that the states in $S = (\delta(q_0, w[1,0]), \ldots, \delta(q_0, w[1, |xy|-1]))$ are all distinct. Also, the states in $S' = (\delta(q_0, v_x v_y z[1,1]), \ldots, \delta(q_0, v_x v_y z[1, |z|-1]))$ are all accepting and distinct; otherwise, v would not be shortest.

We now claim that no state can be in both S and S'. For suppose that $\delta(q_0, w[1,i]) = \delta(q_0, v_x v_y z[1,k])$, for some $0 \leq i \leq |x|$, $0 < k < |z|$. Then $(w[1,i]z[k+1,|z|], xz)$ is a witness with $|xz| < |w|$ and $xz \trianglelefteq w$, since $w[1,i] = x[1,i]$, and $x[1,i]z[k+1,|z|] \trianglelefteq xz$. Next, if $\delta(q_0, xy[1,j]) = \delta(q_0, v_x v_y z[1,k])$, for some $0 < j < |y|$, $0 < k < |z|$, then $(xy[1,j]z[k+1,|z|], xyz[k+1,|z|])$ is a witness with $|xyz[k+1,|z|]| < |w|$ and $xyz[k+1,|z|] \trianglelefteq w$, since $xy[1,j]z[k+1,|z|] \trianglelefteq xyz[k+1,|z|]$, and $xyz[k+1,|z|] \in L$ by (1).

Under these conditions M must have $|xy| + (|z|-1) = |xyz| - 1$ distinct accepting states, and at least one rejecting state. Hence $|xyz| = |w| \leq n$ and we have found a witness with the required properties. □

Corollary 6. *Let M be a DFA with $n \geq 2$ states. If $L(M)$ is not subword-closed, there exists a witness (v, w) with $|w| \leq n$. Furthermore, this is the best possible bound, as there exists a unary DFA with n states that achieves this bound.*

For $n = 1$, L is either $\emptyset$ or Σ^*, and these languages are subword-closed.

Subword-freeness

Lemma 2. *Let M be a DFA with $n \geq 2$ states such that $L(M)$ is not subword-free. For any witness (u, w), there exists a witness (u', w') with $|w'| \leq 2n-1$, and $w' \trianglelefteq w$.*

Corollary 7. *Let M be a DFA with $n \geq 2$ states. If $L(M)$ is not subword-free, there exists a witness (u, w) with $|w| \leq 2n-1$. This is the best possible bound, as there exists a unary DFA with $2n-1$ states that achieves this bound.*

Subword-convexity

Lemma 3. *Let M be a DFA with $n \geq 2$ states such that $L(M)$ is not subword-convex. For any witness (u, v, w), there exists a witness (u', v', w') with $w' \trianglelefteq w$, and $|w'| \leq 3n - 2$.*

Proof. We will show that, for any witness (u, v, w) with $|w| \geq 3n - 1$, we can find a witness (u', v', w') with $|w'| < |w|$ and $w' \trianglelefteq w$. The lemma then follows.

We assume without loss of generality that v is a shortest possible word corresponding to the given w, and u is a shortest word corresponding to v and w.

First, consider the witness (u, v) to the lack of subword-closure of the language $\overline{L}$. By Lemma 1, there exists a witness (u', v') to the failure of subword-closure of $\overline{L}$ such that $v' \trianglelefteq v$ and $|v'| \leq n$. Therefore we can assume that we have a witness (u, v, w) to the failure of subword-convexity such that $|v| \leq n$.

Suppose that (u, v, w) is a minimal witness, and $|w| \geq 3n - 1$. Then the canonical factorization of w is $w = x_1 y_1 z_1$, where $|x_1 y_1| \leq n$, $|y_1| > 0$, and $|z_1| \geq 2n - 1 \geq n > 0$. Consider the states

$$p_0 = \delta(q_0, x_1 y_1),\ p_1 = \delta(q_0, x_1 y_1 z_1[1,1]),\ \cdots,\ p_{|z_1|} = \delta(q_0, x_1 y_1 z_1).$$

Since $|z_1| \geq n$, there must be at least one pair (p_i, p_j) of states such that $p_i = p_j$. If p_0 is the state that is repeated, let i be the greatest index such that $p_0 = p_i$, and let $x_2 = \epsilon$, $y_2 = z_1[1, i]$, and $z_2 = z_1[i + 1, |z_1|]$. If p_i is the first state that is repeated, let j be the greatest index such that $p_i = p_j$, and let $x_2 = z_1[1, i]$, $y_2 = z_1[i+1, j]$, and $z_2 = z_1[j+1, |z_1|]$. If $\delta(q_0, x_1 y_1 x_2 y_2), \delta(q_0, x_1 y_1 x_2 y_2 z_2[1,1])$, $\ldots, \delta(q_0, x_1 y_1 x_2 y_2 z_2)$ has no repeated states, we stop. Otherwise, we apply the same procedure to z_2, and so on. In any case, eventually we reach a z_k for which no repeated states exist. Then we have the factorization $w = x_1 y_1 x_2 y_2 \cdots x_k y_k z_k$, where $x_1 y_1^* x_2 y_2^* \cdots x_k y_k^* z_k \subseteq L$, $|x_2 \cdots x_k z_k| < n$ (otherwise, there would be repeated states), $|y_i| > 0$, for $i = 1, \ldots, k$, and $k \geq 2$.

For any $y_2' \trianglelefteq y_2, \cdots, y_k' \trianglelefteq y_k$, we have $x_1 y_1 x_2 y_2' \cdots x_k y_k' z_k \in L$. Otherwise, the triple $(x_1 x_2 \cdots x_k z_k,\ x_1 x_2 y_2' \cdots x_k y_k' z_k,\ x_1 x_2 y_2 \cdots x_k y_k z_k)$ is a witness with $|x_1 x_2 y_2 \cdots x_k y_k z_k| < |w|$, and $x_1 x_2 y_2 \cdots x_k y_k z_k \trianglelefteq w$.

Since $v \trianglelefteq w$, we can now write $v = v_{x_1} v_{y_1} v_{x_2} v_{y_2} \cdots v_{x_k} v_{y_k} v_{z_k}$, where $v_{x_1} \trianglelefteq x_1$, etc. If there is a y_i with $i \geq 2$, such that $v_{y_i} = \epsilon$, then we can replace that y_i by ϵ in w and obtain a smaller witness. Hence each v_{y_i} must be nonempty. By the same argument, if there is a letter in y_i, for $i \geq 2$, that is not used in v_{y_i}, then that letter can be removed, yielding a smaller witness. Therefore $y_i = v_{y_i}$ for $i = 2, \ldots, k$. We claim that $|y_2 \cdots y_k| < |v|$; otherwise $v = v_{y_2} \cdots v_{y_k} = y_2 \cdots y_k$ and $(u, v, x_1 x_2 y_2 \cdots x_k y_k z_k)$ is a witness with $|x_1 x_2 y_2 \cdots x_k y_k z_k| < |w|$. Thus $|y_2 \cdots y_k| < |v| \leq n$, and $|w| = |x_1 y_1| + |x_2 \cdots x_k z_k| + |y_2 \cdots y_k| \leq n + (n - 1) + (n - 1) = 3n - 2$. □

Corollary 8. *Let M be a DFA with $n \geq 2$ states. If $L(M)$ is not subword-convex, there exists a witness (u, v, w) with $|w| \leq 3n - 2$.*

We do not know whether $3n - 2$ is the best bound. The unary language $\mathtt{a}^{n-1}(\mathtt{a}^n)^*$ is accepted by a DFA with n states and has a minimal witness $(\mathtt{a}^{n-1}, \mathtt{a}^n, \mathtt{a}^{2n-1})$, showing that $2n - 1$ is achievable.

4 Languages specified by other means

4.1 Languages specified by NFA's

Some of our decision problems become PSPACE-complete if M is represented by an NFA. Our fundamental tool is the following classical lemma [9]:

Lemma 4. *Let T be a one-tape deterministic Turing machine and $p(n)$ a polynomial such that T never uses more than $p(|x|)$ space on input x. Then there is a finite alphabet Δ and a polynomial $q(n)$ such that we can construct a regular expression r_x in $q(|x|)$ steps, such that $L(r_x) = \Delta^*$ if T doesn't accept x, and $L(r_x) = \Delta^* - \{w\}$ for some nonempty w (depending on x) otherwise. Similarly, we can construct an NFA M_x in $q(|x|)$ steps, such that $L(M_x) = \Delta^*$ if T doesn't accept x, and $L(M_x) = \Delta^* - \{w\}$ for some nonempty w (depending on x) otherwise.*

Theorem 11. *The problem of deciding whether a given regular language L, represented by an NFA or regular expression, is prefix-convex (resp., suffix-, factor-, subword-convex), or prefix-closed (resp., suffix-, factor-, subword-closed) is PSPACE-complete.*

For the properties of prefix-, suffix, and factor-closed properties, this result was essentially already proved by Hunt and Rosenkrantz [10, Thm. 3.4].

The situation is different for deciding the property of prefix-freeness, suffix-freeness, etc., for languages represented by NFA's, as the following theorem shows. This was proved by Han et al. [8] through a different approach.

Theorem 12. *Let M be an NFA with n states and t transitions. Then we can decide in $O(n^2 + t^2)$ time whether $L(M)$ is prefix-free (resp., suffix-free, factor-free, subword-free).*

Minimal witnesses for NFA's We have already seen that the length of the minimal witness for the lack of convexity or closure is polynomial in the size of the DFA. For the case of NFA's, however, this bound no longer holds.

Theorem 13. *There is a class of NFA's with $O(n)$ states such that the shortest witness to the lack prefix-convexity (resp., suffix-, factor-, subword-convexity) or prefix-closure (resp., suffix-, factor-, subword-closure) is of length $2^{\Omega(n)}$.*

Theorem 14. *There exists a class of languages, accepted by NFA's with $O(n)$ states and $O(n)$ transitions, such that the minimal witness for the lack of prefix-freeness is of length $\Omega(n^2)$.*

For the lack of subword-freeness, we cannot improve the bound we obtained for DFA's in Corollary 7, as the proof we presented there also works for NFA's.

4.2 Languages specified by context-free grammars

If L is represented by a context-free grammar, then the decision problems corresponding to convex and closed languages become undecidable. This follows easily from a well-known result that the set of invalid computations of a Turing machine is a CFL [11, Lemma 8.7, p. 203]. Similarly, the decision problems corresponding to the properties of prefix-free, suffix-free, and factor-free become undecidable for CFL's, as shown by Jürgensen and Konstantinidis [12, Thm. 9.5, p. 581]. However, testing subword-freeness is still decidable for CFL's:

Theorem 15. *There is an algorithm that, given a context-free grammar G, will decide if $L(G)$ is subword-free.*

5 Conclusions

We have shown that we can decide in $O(n^3)$ time whether a language specified by a DFA is prefix-, suffix-, factor-, or subword-convex, and that the corresponding closure and freeness properties can be tested in $O(n^2)$ time. If L is specified by an NFA or a regular expression, these problems are PSPACE-complete.

Acknowledgments: This research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

1. Thierrin, G.: Convex languages. In Nivat, M., ed.: Automata, Languages, and Programming. North-Holland (1973), 481–492.
2. Ang, T., Brzozowski, J.: Continuous languages. In Csuhaj-Varjú, E., Ésik, Z., eds.: Proc. 12th International Conference on Automata and Formal Languages. Computer and Automation Research Institute, Hungarian Academy of Sciences (2008), 74–85.
3. Berstel, J., Perrin, D.: Theory of Codes. Academic Press, New York (1985).
4. Han, Y.S.: Decision algorithms for subfamilies of regular languages using state-pair graphs. Bull. European Assoc. Theor. Comput. Sci. (93) (October 2007) 118–133.
5. Brzozowski, J.A., Shallit, J., Xu, Z.: Decision problems for convex languages. Preprint, `http://arxiv.org/abs/0808.1928`.
6. Béal, M.P., Crochemore, M., Mignosi, F., Restivo, A., Sciortino, M.: Computing forbidden words of regular languages. Fund. Inform. **56** (2003), 121–135.
7. de Luca, A., Varricchio, S.: Some combinatorial properties of factorial languages. In Capocelli, R., ed.: Sequences. Springer (1990), 258–266.
8. Han, Y.S., Wang, Y., Wood, D.: Infix-free regular expressions and languages. Internat. J. Found. Comp. Sci. **17** (2006), 379–393.
9. Aho, A., Hopcroft, J., Ullman, J.: The Design and Analysis of Computer Algorithms. Addison-Wesley (1974).
10. Hunt, III, H.B., Rosenkrantz, D.J.: Computational parallels between the regular and context-free languages. SIAM J. Comput. **7** (1978), 99–114.
11. Hopcroft, J.E., Ullman, J.D.: Introduction to Automata Theory, Languages, and Computation. Addison-Wesley (1979).
12. Jürgensen, H., Konstantinidis, S.: Codes. In Rozenberg, G., Salomaa, A., eds.: Handbook of Formal Languages, Vol. 1. Springer-Verlag (1997), 511–607.