

# Ordering Regular Languages: A Danger Zone

Giovanna D’Agostino, Davide Martincigh, and Alberto Policriti

University of Udine, Italy

**Abstract.** Ordering the collection of states of a given automaton starting from an order of the underlying alphabet is a natural move towards a computational treatment of the language accepted by the automaton. Along this path, Wheeler *graphs* have been recently introduced as an extension/adaptation of the Burrows-Wheeler Transform (the now famous BWT, originally defined on strings) to graphs. These graphs constitute an important data-structure for languages, since they allow a very efficient storage mechanism for the transition function of an automaton, while providing a fast support to all sorts of substring queries. This is possible as a consequence of a property—the so-called *path coherence*—valid on Wheeler graphs and consisting in an ordering on nodes that “propagates” to (collections of) strings. By looking at a Wheeler graph as an automaton, the ordering on strings corresponds to the co-lexicographic order of the words entering each state. This leads naturally to consider the class of regular languages accepted by Wheeler automata, i.e. the Wheeler languages.

It has been shown that, as opposed to the general case, the classic determination by powerset construction is polynomial on Wheeler languages. As a consequence, most of the classical problems turn out to be “easy”—that is, solvable in polynomial time—on Wheeler languages. Moreover, deciding whether a DFA is Wheeler and deciding whether a DFA accepts a Wheeler language is polynomial.

Our contribution here is to put an upper bound to easy problems. For instance, whenever we generalize by switching to general NFAs or by not fixing an order of the underlying alphabet, the above mentioned problems become “hard”—that is NP-complete or even PSPACE-complete.

## 1 Introduction

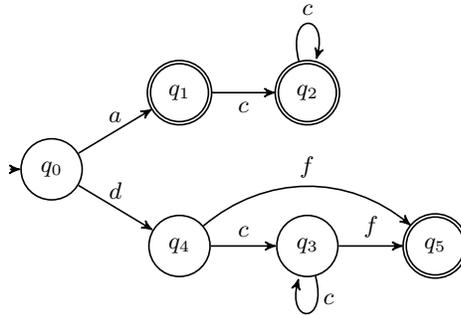
Adding an order to a class of structures (graphs, groups, monoids...) is a natural move. Moreover, ordering is a basic data-structuring mechanism, strongly favoring computational manipulations of the considered class.

Over the class of finite automata, order has been added e.g. in [12], where the order must propagate along equally labeled transitions. Much more recently,

---

Copyright © 2021 for the individual papers by the papers’ authors. Copyright © 2021 for the volume as a collection by its editors. This volume and its papers are published under the Creative Commons License Attribution 4.0 International (CC BY 4.0).

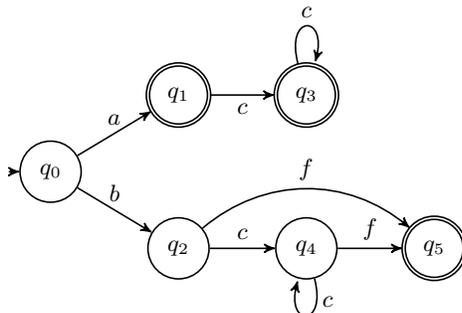
in an effort to find a common denominator to a number of different algorithmic techniques, Gagie et al. proposed (in [13]) a new simple strategy for enforcing and using order on a given automaton. Starting from an underlying order of the alphabet, the order on the states (formally given in Definition 1) must: i) agree with the order of the labels of their incoming edges, and ii) be coherent on target/source nodes, for pairs of arcs with equal labels. It turns out that this kind of automata, called Wheeler automata, (a) admit an efficient index data structure for searching subpaths labeled with a given query pattern, and (b) enable a representation of the graph in a space proportional to that of the edges’ labels since the topology can be encoded with just  $O(1)$  bits per node [13] (as well as enabling more advanced compression mechanisms, see [3,11]). This is in contrast with the fact that general graphs require a logarithmic (in the graph’s size) number of bits per edge to be represented, as well as with recent results showing that in general, the subpath search problem can not be solved in subquadratic time, unless the strong exponential time hypothesis is false [4,5,6,7,10].



**Fig. 1.** A WDFAs  $\mathcal{A}$  recognizing the language  $\mathcal{L}_d = ac^* + dc^*f$ . Condition (i) of Definition 1 implies input consistency and induces the partial order  $q_1 < q_2, q_3 < q_4 < q_5$ . From condition (ii) it follows that  $\delta(q_1, c) \leq \delta(q_4, c)$ , thus  $q_2 < q_3$ . Therefore, the only order that could make  $\mathcal{A}$  Wheeler is  $q_0 < q_1 < q_2 < q_3 < q_4 < q_5$ . The reader can verify that condition (ii) holds for each pair of equally labeled edges.

In Figure 1 is depicted an example of a Wheeler automaton (see also Definition 1). Notice that the minimum DFA recognizing  $\mathcal{L}$  is not input consistent, hence not Wheeler; we will return on this point later. One is naturally led to consider the class of Wheeler languages, i.e. the regular languages recognized by some Wheeler automaton, as well as to raise the question of whether it is possible to decide, just by looking at a (generic, possibly not Wheeler) DFA, whether the language it recognizes is Wheeler. Wheeler languages have been studied extensively in [2], where it is shown that we can decide in polynomial time whether a DFA recognizes a Wheeler language. Unfortunately, trying to

use the same strategy on a NFA language does not work. In fact, in this paper we show that the problem for NFAs becomes PSPACE-complete.



**Fig. 2.** A failed attempt to build a W DFA recognizing the language  $\mathcal{L}_b = ac^* + bc^*f$ . Condition (i) of Definition 1 induces the partial order  $q_1 < q_2 < q_3, q_4 < q_5$ , and condition (ii) implies  $\delta(q_1, c) \leq \delta(q_2, c)$ , thus  $q_3 < q_4$ . But if we apply condition (ii) again, we obtain  $\delta(q_2, c) \leq \delta(q_3, c)$ , a contradiction.

Consider the example in Figure 2 and confront it with Figure 1. Even though apparently we made an insignificant change to the automaton, that is we have simply swapped character-labels  $d$  and  $b$ , the new language recognized by the automaton turns out to be *not* Wheeler. This example illustrates that the order of the underlying alphabet plays a significant role when determining whether an automaton, and even a language, is Wheeler. Notice that swapping  $d$  and  $b$  can be seen as considering a different order on  $\Sigma$ : from  $a < c < d < f$  to  $a < d < c < f$ . Again a natural question is raised: when presented with an automaton (or a language) that is *not* Wheeler, is there a different way to order the alphabet so that the automaton (language) becomes Wheeler? Being able to answer in polynomial time to this question would be very helpful for applications. Unfortunately, as shown in Section 3, this problem turns out to be NP-complete even for DFAs.

WNFAs have the following useful property: turning a WNFA into a DFA using the (classic) powerset construction actually results in a W DFA with at most twice the number of states of the original WNFA. In other words, the blow-up of states that we can observe when converting NFAs into DFAs, does not occur for Wheeler non-deterministic automata. Nevertheless, it is known (see [1]) that a blow-up of states can occur even when we switch from the *minimum* DFA recognizing a language  $\mathcal{L}$  to the minimum W DFA recognizing  $\mathcal{L}$ . As a last contribution, in this paper we give an answer to a question put forward in [1]. We provide an algorithm to compute the minimum W DFA starting from the minimum DFA. Our algorithm works in exponential time in the worst case and,

since the dimension of the output might be exponential in the dimension of the input, is optimal for the task.

Due to space constraints, we present only the sketches of the proofs of our results. The full proofs, complete with all the missing details, can be found in the Appendix.

## 2 Wheeler languages

First of all, we fix some notation. Let  $\Sigma$  denote a finite alphabet endowed with a total order  $(\Sigma, <)$ . We denote by  $\Sigma^*$  the set of finite words over  $\Sigma$ , with  $\varepsilon$  being the empty word, and we extend the order  $<$  over  $\Sigma$  to the *co-lexicographic* order  $(\Sigma^*, <)$ , where  $\alpha < \beta$  if and only the reverse of  $\alpha$ , i.e.  $\alpha$  read from the right to the left, precedes lexicographically the reverse of  $\beta$ . Given two words  $\alpha, \beta \in \Sigma^*$ , we denote by  $\alpha \dashv \beta$  the property that  $\alpha$  is a suffix of  $\beta$ . For a language  $\mathcal{L} \subseteq \Sigma^*$ , we denote by  $\text{Pref}(\mathcal{L})$  the set of prefixes of strings in  $\mathcal{L}$ . We denote by  $\mathcal{A} = (Q, q_0, \delta, F, \Sigma)$  a finite automaton (NFA), with  $Q$  as set of states,  $q_0$  initial state,  $\delta : Q \times \Sigma \rightarrow 2^Q$  transition function, and  $F \subseteq Q$  final states. The dimension of  $\mathcal{A}$ , denoted by  $|\mathcal{A}|$ , is defined as the number of states of  $\mathcal{A}$ . An automaton is deterministic (DFA) if  $|\delta(q, a)| \leq 1$ , for all  $q \in Q$  and  $a \in \Sigma$ . As customary, we extend  $\delta$  to operate on strings as follows: for all  $q \in Q$ ,  $a \in \Sigma$  and  $\alpha \in \Sigma^*$

$$\delta(q, \varepsilon) = \{q\}, \quad \delta(q, \alpha a) = \bigcup_{v \in \delta(q, \alpha)} \delta(v, a).$$

We denote by  $\mathcal{L}(\mathcal{A}) = \{\alpha \in \Sigma^* : \delta(q_0, \alpha) \cap F \neq \emptyset\}$  the language accepted by the automaton  $\mathcal{A}$ . We make the assumption that every automaton is basic, that is, every state is reachable from the initial state and every state can reach at least one final state. Notice that this assumption is not restrictive, since removing from a NFA every state not reachable from  $q_0$  and every state from which is impossible to reach a final state can be done in linear time and does not change the accepted language. It immediately follows that:

- there might be only one state without incoming edges, namely  $q_0$ ;
- every word that can be read starting from  $q_0$  belongs to  $\text{Pref}(\mathcal{L})$ .

Lastly, given a finite set  $Q$ , totally ordered by the relation  $<$ , we say that  $I \subseteq Q$  is an interval if and only if for all  $q, q', q'' \in Q$  with  $q \leq q' \leq q''$ , if  $q, q'' \in I$  then  $q' \in I$ . We denote by  $I := [q_{\min}, q_{\max}]$  a generic interval, where  $q_{\min}$  ( $q_{\max}$ ) is the minimum (maximum) element of  $I$  with respect to  $<$ .

The class of Wheeler automata has been recently introduced in [13]. An automaton in this class has the property that there exists a total order on its states that is propagated along equally labeled transition. Moreover, the order must be compatible with the underlying order of the alphabet:

**Definition 1 (Wheeler Automaton).** *A Wheeler NFA (WNFA)  $\mathcal{A}$  is a NFA  $(Q, q_0, \delta, F, \Sigma)$  endowed with a binary relation  $<$ , such that:  $(Q, <)$  is a linear order having the initial state  $q_0$  as minimum,  $q_0$  has no in-going edges, and the*

following two (Wheeler) properties are satisfied. Let  $v_1 \in \delta(u_1, a_1)$  and  $v_2 \in \delta(u_2, a_2)$ :

- (i)  $a_1 \prec a_2 \rightarrow v_1 < v_2$
- (ii)  $(a_1 = a_2 \wedge u_1 < u_2) \rightarrow v_1 \leq v_2$ .

A Wheeler DFA (W DFA) is a WNFA in which the cardinality of  $\delta(u, a)$  is always less than or equal to one.

In Figure 1 is depicted an example of a W DFA.

*Remark 1.* A consequence of Wheeler property (i) is that  $A$  is *input-consistent*, that is all transitions entering a given state  $u \in Q$  have the same label: if  $u \in \delta(v, a)$  and  $u \in \delta(w, b)$ , then  $a = b$ . Therefore the function  $\lambda : Q \rightarrow \Sigma$  that associate to each state the unique label of its incoming edges is well defined. For the state  $q_0$ , the only one without incoming edges, we set  $\lambda(q_0) := \#$ .

In [13] it is shown that W DFAs have a property called *path coherence*: let  $\mathcal{A} = (Q, q_0, \delta, F, \Sigma)$  be a W DFA according to the order  $(Q, <)$ . Then for every interval of states  $I = [q_i, q_j]$  and for all  $\alpha \in \Sigma^*$ , the set  $J$  of states reachable starting from any state of  $I$  by reading  $\alpha$  is also an interval. *Path coherence* allows us to transfer the order  $<$  over the states of  $Q$  to the co-lexicographic order  $\prec$  over the words entering the states: for each state  $q$  define the set  $I_q = \{\alpha : \delta(q_0, \alpha) = q\}$ . With abuse of notation, we extend  $\preceq$  to subsets of  $\Sigma^*$  as follows:  $U \preceq V$ , with  $U, V \subseteq \Sigma^*$ , if and only if  $\alpha \prec \beta$  for each  $\alpha \in U$ ,  $\beta \in V$  such that  $\{\alpha, \beta\} \not\subseteq U \cap V$ . Then, two states  $q$  and  $p$  with  $I_q \neq I_p$  satisfy  $q < p$  if and only if  $I_q \preceq I_p$  (again proved in [1]). An immediate consequence of this fact is that a W DFA admits an unique order of its states that makes it Wheeler and this order is univocally determined by the co-lexicographic order of any word entering its states. This result is important for two different reasons. First of all, it makes possible to decide in polynomial time whether a DFA is Wheeler: for each state  $q$ , pick a word  $\alpha_q$  entering it and order the states reflecting the co-lexicographic order of the words  $\{\alpha_q : q \in Q\}$ ; then check if the order satisfies the Wheeler conditions. Secondly, it is the key to adapt Myhill-Nerode Theorem to Wheeler automata. We recall the following definition.

**Definition 2 (Myhill-Nerode equivalence).** Let  $\mathcal{L} \subseteq \Sigma^*$  be a language. Given a word  $\alpha \in \Sigma^*$ , we define the right context of  $\alpha$  as

$$\alpha^{-1}\mathcal{L} := \{\gamma \in \Sigma^* : \alpha\gamma \in \mathcal{L}\},$$

and we denote by  $\equiv_{\mathcal{L}}$  the Myhill-Nerode equivalence on  $\text{Pref}(\mathcal{L})$  defined as

$$\alpha \equiv_{\mathcal{L}} \beta \iff \alpha^{-1}\mathcal{L} = \beta^{-1}\mathcal{L}.$$

The (classic) Myhill-Nerode Theorem, among many other things, establishes a bijection between equivalence classes of  $\equiv_{\mathcal{L}}$  and the states of the minimum DFA recognizing  $\mathcal{L}$ . This minimum automaton is also unique up to isomorphism and a similar result, fully proved in [2], holds for Wheeler languages as well.

In order to state such an analogous of Myhill-Nerode Theorem for Wheeler languages, the equivalence  $\equiv_{\mathcal{L}}$  is replaced by the equivalence  $\equiv_{\mathcal{L}}^c$  defined below.

**Definition 3.** *The input consistent, convex refinement  $\equiv_{\mathcal{L}}^c$  of  $\equiv_{\mathcal{L}}$  is defined as follows.  $\alpha \equiv_{\mathcal{L}}^c \beta$  if and only if*

- $\alpha \equiv_{\mathcal{L}} \beta$ ,
- $\alpha$  and  $\beta$  end with the same character, and 2
- for all  $\gamma \in \text{Pref}(\mathcal{L})$ , if  $\min(\alpha, \beta) \preceq \gamma \preceq \max(\alpha, \beta)$ , then  $\alpha \equiv_{\mathcal{L}} \gamma \equiv_{\mathcal{L}} \beta$ .

The Myhill-Nerode Theorem for Wheeler languages proves that there exists a minimum (in the number of states) W DFA recognizing  $\mathcal{L}$ . As in the classic case, states of the minimum automaton are, in fact,  $\equiv_{\mathcal{L}}^c$ -equivalence classes, this time consisting of *intervals* of words. Also such W DFA is unique up to isomorphism.

A further important consequence, especially for testing Wheelerness, is stated in the following Lemma (again proved in [2]).

**Lemma 1.** *A regular language  $\mathcal{L}$  is Wheeler if and only if all monotone sequences in  $(\text{Pref}(\mathcal{L}), \prec)$  become eventually constant modulo  $\equiv_{\mathcal{L}}$ . In other words, for all sequences  $(\alpha_i)_{i \geq 0}$  in  $\text{Pref}(\mathcal{L})$  with*

$$\alpha_1 \preceq \alpha_2 \preceq \dots \alpha_i \preceq \dots \quad \text{or} \quad \alpha_1 \succeq \alpha_2 \succeq \dots \alpha_i \succeq \dots$$

*there exists an  $n$  such that  $\alpha_h \equiv_{\mathcal{L}} \alpha_k$ , for all  $h, k \geq n$ .*

Lemma 1 shows how it is possible to recognize whether a language  $\mathcal{L}$  is Wheeler simply by verifying a property on the words of  $\text{Pref}(\mathcal{L})$ : trying to find a W DFA that recognizes  $\mathcal{L}$  is no longer needed to decide the Wheelerness of  $\mathcal{L}$ . As it turns out, we can verify whether the property depicted in Lemma 1 is satisfied just by looking at *any* DFA recognizing  $\mathcal{L}$ , as shown in Theorem 1 (see [2]).

**Theorem 1.** *Let  $\mathcal{A}$  be a DFA such that  $\mathcal{L} = \mathcal{L}(\mathcal{A})$ , with initial state  $q_0$  and dimension  $n = |\mathcal{A}|$ .*

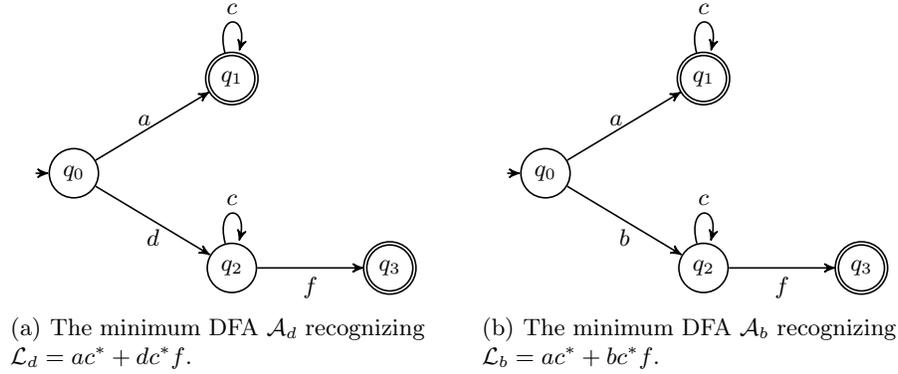
*$\mathcal{L}$  is not Wheeler if and only if there exist  $\mu, \nu$  and  $\gamma$  in  $\Sigma^*$ , with  $\gamma \not\prec \mu, \nu$ , such that:*

1.  $\mu \not\equiv_{\mathcal{L}} \nu$  and they label paths from  $q_0$  to states  $u$  and  $v$ , respectively;
2.  $\gamma$  labels two cycles, one starting from  $u$  and one starting from  $v$ ;
3.  $\mu, \nu \prec \gamma$  or  $\gamma \prec \mu, \nu$ .

*The length of the words  $\mu, \nu$  and  $\gamma$  satisfying the above can be bounded:*

4.  $|\mu|, |\nu| \leq |\gamma| \leq n^3 + 2n^2 + n + 2$ .

Since in this work we make an extensive use of Theorem 1, here is a simple example on how and why it works. Consider the automata depicted in Figure 3. As shown in Figure 1 and 2, the language  $\mathcal{L}_d$  is Wheeler whereas  $\mathcal{L}_b$  is not, as can be easily proved using Theorem 1. In fact, consider the automaton  $\mathcal{A}_b$ . By setting  $\mu := a$ ,  $\nu := b$  and  $\gamma := c$ , one can verify conditions 1-3 of the theorem are satisfied. Notice that condition  $a \not\equiv_{\mathcal{L}_b} b$  follows immediately from the fact that  $\delta(q_0, a) \neq \delta(q_0, b)$  in the minimum DFA  $\mathcal{A}_b$ .



**Fig. 3.** The minimum DFAs recognizing the languages  $\mathcal{L}_d$  (Wheeler) and  $\mathcal{L}_b$  (not Wheeler).

If we try to transpose the same reasoning to the automaton  $\mathcal{A}_d$  by setting  $\mu = a$ ,  $\nu = d$  and  $\gamma = c$ , condition 3 of Theorem 1 is no longer satisfied. We can not find 3 words satisfying conditions 1-3 of Theorem 1, therefore  $\mathcal{L}_b$  is Wheeler.

The polynomial bound given by condition 4 of Theorem 1 allows us to design an algorithm that decides whether a given DFA recognizes a Wheeler language: using dynamic programming (see [1]), it is possible to keep track of all the relevant paths and cycles inside the DFA and check, in polynomial time, whether there exists three words satisfying the conditions of the theorem.

Things change if, instead of a DFA, we are given a NFA. Trying to exploit the same idea used for DFAs does not work: the problem of deciding whether two words  $\mu$  and  $\nu$  read by a NFA are Myhill-Nerode equivalent is PSAPCE-complete, whereas is polynomial on DFAs (simply compute the minimum DFA using Hopcroft's algorithm and check whether the two states reached by  $\mu$  and  $\nu$  coincide). Even worse, the straightforward attempt of building the minimum DFA recognizing the NFA's language might lead to a blow-up of the states, resulting in an exponential time (and exponential space) algorithm.

We show that the problem of deciding whether a NFA recognizes a Wheeler language is indeed hard, but doesn't require exponential time to be solved: the problem turns out to be PSPACE-complete. To show this, we first need to adapt Theorem 1 to work on NFAs, as described in the following corollary.

**Corollary 1.** *Let  $\mathcal{A} = (Q, q_0, \delta, F, \Sigma)$  be a NFA of dimension  $n := |\mathcal{A}|$ . Then  $\mathcal{L} := \mathcal{L}(\mathcal{A})$  is not Wheeler if and only if there exist three words  $\mu, \nu, \gamma$ , with  $\gamma \mathcal{A} \mu, \nu$ , such that*

1.  $\mu\gamma^i \notin_{\mathcal{L}} \nu\gamma^j$  for all  $0 \leq i, j \leq 2^n$ ;
2.  $\gamma$  labels two cycles, one starting from a state  $p \in \delta(q_0, \mu)$  and one from a state  $r \in \delta(q_0, \nu)$ ;

3.  $\mu, \nu \prec \gamma$  or  $\gamma \prec \mu, \nu$ .

The length of the words  $\mu, \nu$  and  $\gamma$  satisfying the above can be bounded:

4.  $|\mu|, |\nu| < |\gamma| \in O(n^3 \cdot 2^{3n})$ .

*Proof (Sketch).* ( $\implies$ ) Let  $\mathcal{D}$  be the minimum DFA recognizing the language  $\mathcal{L}$ , with initial state  $\hat{q}_0$ . Recall that  $\mathcal{D}$  might have up to  $2^n$  states. From Theorem 1 we know that  $\mathcal{L}$  is not Wheeler if and only if there exist  $\hat{\mu}, \hat{\nu}$  and  $\hat{\gamma}$  satisfying conditions 1-4 of Theorem 1. Although there is not a perfect correspondence between cycles in  $\mathcal{A}$  and cycles in  $\mathcal{D}$ , whenever we find in  $\mathcal{D}$  a path  $\hat{\mu}$  (respectively,  $\hat{\nu}$ ) and a cycle  $\hat{\gamma}$  both ending in the same state  $\hat{u}$  ( $\hat{v}$ ), we can find in  $\mathcal{A}$  a path  $\mu = \hat{\mu}\hat{\gamma}^i$  ( $\nu = \hat{\nu}\hat{\gamma}^j$ ) and a cycle  $\gamma_1 = \hat{\gamma}^{k_1}$  ( $\gamma_2 = \hat{\gamma}^{k_2}$ ) both ending in the same state  $u$  ( $v$ ), for some  $i, j, k_1, k_2 \leq n$ . Note that the word  $\hat{\gamma}^{(n+1)k_1k_2}$  labels two cycles starting from  $u$  and  $v$ . Moreover, we have

$$|\hat{\gamma}^{(n+1)k_1k_2}| \geq (n+1)|\hat{\gamma}^{k_1k_2}| \geq (n+1)|\hat{\gamma}| \geq (i+1)|\hat{\gamma}| > |\hat{\mu}\hat{\gamma}^i|,$$

and similarly  $|\hat{\gamma}^{(n+1)k_1k_2}| > |\hat{\nu}\hat{\gamma}^j|$ . Since  $|\hat{\gamma}| \in O(2^{3n})$ , we have  $|\hat{\gamma}^{(n+1)k_1k_2}| \in O(n^3 \cdot 2^{3n})$  and therefore the words  $\mu, \nu$  and  $\gamma := \hat{\gamma}^{(n+1)k_1k_2}$  satisfy conditions 2-4 of this corollary. Finally, condition 1 is satisfied for all  $i, j \geq 0$  because when reading  $\mu\gamma^i$  and  $\nu\gamma^j$  on  $\mathcal{D}$ , they reach different states.

( $\impliedby$ ) As we did while proving the opposite direction, given three words  $\mu, \nu, \gamma$  in  $\mathcal{A}$  satisfying condition 1-3, we can always find, for some  $i, j, k \leq 2^n$ , three words  $\hat{\mu} = \mu\gamma^i$ ,  $\hat{\nu} = \nu\gamma^j$  and  $\hat{\gamma} = \gamma^k$  in  $\mathcal{D}$  that almost satisfy condition 1-3 of Theorem 1. The only requirement that might be missing is that  $\hat{\mu} \not\equiv_{\mathcal{L}} \hat{\nu}$ . The upper bound  $2^n$  in condition 1 ensures that this requirement is fulfilled, hence we can apply Theorem 1 to conclude that  $\mathcal{L}$  is not Wheeler.

Despite the fact the the bound in condition 4 has become exponential by switching to NFAs, it is still possible to check in polynomial space (but exponential time) whether there are three words  $\mu, \nu$  and  $\gamma$  satisfying the conditions of Corollary 1. Thus we can prove the following:

**Proposition 1.** *Given a NFA  $\mathcal{A}$ , deciding whether the language  $\mathcal{L} := \mathcal{L}(\mathcal{A})$  is Wheeler is PSPACE-complete.*

*Proof (Sketch).* To prove that the problem is in PSPACE (=NPSPACE), we guess (using non-determinism) three words  $\mu, \nu$  and  $\gamma$  satisfying the conditions of Corollary 1. Since this words are too long to be stored, we first guess their length bit by bit and then we guess their characters one by one, starting from the last one and proceeding from right to left. Meanwhile, for all  $\alpha \in \{\mu, \nu, \gamma\}$  and for all state  $t$  of  $\mathcal{A}$ , we follow backwards the edges of  $\mathcal{A}$  labeled as the character of  $\alpha$  that we are guessing. This way, we can compute the set of states from which is possible to reach  $t$  by reading  $\alpha$ . This allows us to find, for all  $i, j \leq 2^n$ , the sets  $Q_\mu^i := \delta(q_0, \mu\gamma^i)$  and  $Q_\nu^j := \delta(q_0, \nu\gamma^j)$ . We can then test whether  $\mu\gamma^i \not\equiv_{\mathcal{L}} \nu\gamma^j$  by confronting the language accepted by the automaton  $\mathcal{A}$  with set of initial states  $Q_\mu^i$  and the language accepted by the automaton  $\mathcal{A}$  with set of initial states  $Q_\nu^j$ .

This can be done in polynomial space since the problem of deciding whether two NFAs recognize the same language is known to be PSPACE-complete.

To prove the hardness of the problem, we show a polynomial reduction from the universality problem for NFA, i.e. the problem of deciding whether the language accepted by a NFA  $\mathcal{A}$  over the alphabet  $\Sigma$  is such that  $\mathcal{L}(\mathcal{A}) = \Sigma^*$ . Let  $\mathcal{A} = (Q, q_0, \delta, F, \Sigma)$  be a NFA and let  $\mathcal{L} := \mathcal{L}(\mathcal{A})$ . We can assume, without loss of generality, that  $q_0 \in F$ , otherwise  $\mathcal{A}$  would not accept the empty word and we could immediately derive that  $\mathcal{L} \neq \Sigma^*$ . Starting from  $\mathcal{A}$ , we build in constant time a NFA  $\mathcal{A}''$  over the alphabet  $\Sigma \cup \{a, b, c\}$  that recognizes the language

$$\mathcal{L}'' := a \cdot (\mathcal{L}c)^* \cdot \mathcal{L} + b \cdot (\Sigma + c)^*,$$

where  $a, b, c$  are three characters not in  $\Sigma$  and such that  $a \prec b \prec c$  (the order of the characters of  $\Sigma$  is irrelevant in this proof). If we prove that  $\mathcal{L} = \Sigma^*$  if and only if  $\mathcal{L}''$  is Wheeler, the reduction is complete and the thesis follows. Notice that, since  $\varepsilon \in \mathcal{L}$ , the following property holds:

$$\mathcal{L} = \Sigma^* \iff (\mathcal{L}c)^* \cdot \mathcal{L} = (\Sigma + c)^*. \quad (1)$$

Let us prove that  $\mathcal{L} = \Sigma^*$  implies that  $\mathcal{L}''$  is not Wheeler. If  $\mathcal{L} = \Sigma^*$ , then by (1) we have  $\mathcal{L}'' = (a + b)(\Sigma + c)^*$ . The minimum DFA recognizing  $\mathcal{L}''$  has only one cycle, therefore from Theorem 1 it follows that  $\mathcal{L}''$  is Wheeler.

We next prove that  $\mathcal{L}''$  not Wheeler implies  $\mathcal{L} = \Sigma^*$ . Notice that  $a^{-1}\mathcal{L}'' = (\mathcal{L}c)^* \cdot \mathcal{L}$  so that, by (1) if  $\mathcal{L} \neq \Sigma^*$ , then  $a^{-1}\mathcal{L}'' = (\mathcal{L}c)^* \cdot \mathcal{L} \neq (\Sigma + c)^*$ . However we have  $b^{-1}\mathcal{L}'' = (\Sigma + c)^*$ , thus  $a \not\equiv_{\mathcal{L}''} b$ . Moreover, it can be proved that, for all  $n$ , both  $a \equiv_{\mathcal{L}''} ac^n$  and  $b \equiv_{\mathcal{L}''} bc^n$  hold. Therefore the monotone sequence

$$ac \prec bc \prec ac^2 \prec bc^2 \prec \dots \prec ac^n \prec bc^n \prec \dots$$

is not constant modulo  $\equiv_{\mathcal{L}''}$  and from Lemma 1 it follows that  $\mathcal{L}(\mathcal{A}'')$  is not Wheeler.

### 3 Generalized Wheelerness

As we have already pointed out in the introduction, changing the underlying order of the alphabet might turn a Wheeler language into a not Wheeler one and vice versa. For instance, consider again the Wheeler languages  $\mathcal{L}_d$  and the regular (but not Wheeler) language  $\mathcal{L}_b$  depicted in Figure 3. If we change the order of  $\Sigma$  from  $a \prec c \prec d \prec f$  to  $a \prec d \prec c \prec f$ , the Wheeler language  $\mathcal{L}_d$  turns into a non-Wheeler language (isomorphic to  $\mathcal{L}_b$  under the isomorphism  $\varphi$  between alphabets that fixes characters  $a, c, f$  and sends  $d$  into  $\varphi(d) = b$ ). Hence, by not fixing an apriori order of the alphabet  $\Sigma$  we enlarge the class of languages.

**Definition 4 (Generalized Wheelerness).** *A NFA  $\mathcal{A}$  over the alphabet  $\Sigma$  is called a Generalized Wheeler Automaton (GWNFA) if and only if there exists*

an ordering of the elements of  $\Sigma$  that makes  $\mathcal{A}$  Wheeler.

A language  $\mathcal{L}$  is called *generalized Wheeler* (for short *GW*) if and only if there exists a *GW*NFA that recognizes  $\mathcal{L}$ .

Let  $\mathcal{A}$  be a WDFA. Then, every word  $\alpha$  that labels a cycle in  $\mathcal{A}$  is *primitive* (see [2]), that is there exists no  $\beta \neq \varepsilon$  and  $i > 1$  such that  $\alpha = \beta^i$ . A direct consequence of this property is that Wheeler languages form a subclass of star-free languages, i.e. the class of languages that can be defined by a regular expression not containing the Kleene star. Since star-free expressions, and thus star-free languages, are closed under permutations of the alphabet, even GW languages must be a subclass of star-free languages. Here we show that the inclusion is strict, therefore GW languages must be studied separately.

**Proposition 2.** *If  $|\Sigma| \geq 2$ , then the set  $\{\mathcal{L} \subseteq \Sigma^* : \mathcal{L} \text{ is GW}\}$  is a proper subset of  $\{\mathcal{L} \subseteq \Sigma^* : \mathcal{L} \text{ is star-free}\}$ .*

*Proof.* Let  $a, b$  be two distinct characters of  $\Sigma$ , and consider the language  $\mathcal{L} = a(aba)^*a + ba(aba)^*b$ . It is possible to prove that  $\mathcal{L}$  is star-free, but  $\mathcal{L}$  is not GW: consider the sequence  $(\alpha_i)_{i \geq 2}$  with  $\alpha_{2n} = a(aba)^n$  and  $\alpha_{2n+1} = ba(aba)^n$ . Since, for all  $i$ , the word  $\alpha_i$  is a prefix of  $\alpha_{i+1}$ , independently from how  $a$  and  $b$  are ordered we have

$$aaba \prec baaba \prec aabaaba \prec baabaaba \prec \dots \prec a(aba)^i \prec ba(aba)^i \prec \dots$$

Moreover, for all  $i$  we have  $a(aba)^i \not\equiv_{\mathcal{L}} ba(aba)^i$ , since  $a(aba)^i \cdot a$  belongs to  $\mathcal{L}$  but  $ba(aba)^i \cdot a$  does not. We can then apply Lemma 1 to conclude that  $\mathcal{L}$  is not Wheeler. Since this result does not depend on the order of the alphabet,  $\mathcal{L}$  is not GW.

We mentioned that we can decide in polynomial time whether a DFA is Wheeler. On the contrary, deciding whether a NFA is Wheeler is NP-complete, even when we bound the outdegree of each state of the NFA to be at most 5 (see [8]). As one may expect, deciding whether a NFA is a GWNFA is not easier than deciding whether a NFA is Wheeler. In fact, we show in Proposition 3 that the problem is NP-complete. We actually prove a stronger result in Proposition 4: even deciding whether a DFA is a GWNFA is NP-complete. Since the proof of Proposition 4 is cumbersome, we decided to present also its weaker version, i.e. Proposition 3, which shows a more natural reduction from the problem of deciding whether a NFA is Wheeler. Both proofs can be found in the Appendix. It is worth noticing that the proof of Proposition 3 can be adapted to work even on DFAs, hence giving an alternative way to prove that deciding whether a DFA is a GWNFA is NP-complete. Nonetheless, Proposition 4 is still stronger, since it also proves that deciding whether a DFA recognizes a GW language is NP-complete.

**Proposition 3 (GWNFA hardness).** *Let  $\mathcal{A}$  be a NFA. Deciding whether  $\mathcal{A}$  is a GWNFA is NP-complete.*

**Proposition 4 (GW DFA and GW languages hardness).** *Let  $\mathcal{L} \subseteq \Sigma^*$  be a language and  $\mathcal{A}$  be a DFA. Both the problems of deciding whether  $\mathcal{A}$  is a GWNFA and deciding whether  $\mathcal{L}$  is GW are NP-complete.*

## 4 DFA to W DFA

As discussed in Section 1, a Wheeler automaton can be represented more compactly than a generic finite automaton. Therefore, if we are given a Wheeler language  $\mathcal{L}$  represented as a DFA  $\mathcal{A}$  that recognizes  $\mathcal{L}$ , we may be tempted to look for a W DFA  $\mathcal{A}'$  that recognizes the same language to achieve a better, i.e. more compact, representation. Unfortunately, this approach might not work in our favor: in [1], a family  $(\mathcal{L}_m)_{m \geq 1}$  of Wheeler languages with the property that the dimension of the minimum W DFA recognizing  $\mathcal{L}_m$  is exponential in the dimension of the minimum DFA recognizing  $\mathcal{L}_m$  is presented. Note that we can always assume that, whenever we are given a DFA or a W DFA, the automaton is minimum. In fact, both tasks of minimizing a DFA and minimizing a W DFA can be done in polynomial time (see [9] for DFAs and [1] for W DFAs). Here we answer (positively) to the open question put forward in [1] whether there exists an algorithm to compute the minimum W DFA starting from the minimum DFA. Our algorithm works in exponential time in the worst case and, since the dimension of the output might be exponential in the dimension of the input, is optimal for the task.

Let  $\mathcal{L} = \mathcal{L}(\mathcal{A})$  be the language recognized by the given (minimum) DFA  $\mathcal{A}$ . Recall that there is a 1 to 1 correspondence between the  $\equiv_{\mathcal{L}}^c$ -classes and the states of the minimum W DFA recognizing  $\mathcal{L}$ . First of all, our algorithm identifies a representative for each  $\equiv_{\mathcal{L}}^c$ -class; to be able to do this in exponential time, we first need to put a bound on the length of such representatives.

**Lemma 2.** *Let  $\mathcal{A}$  be the minimum DFA recognizing the Wheeler language  $\mathcal{L} = \mathcal{L}(\mathcal{A})$  over the alphabet  $\Sigma$ , and let  $C_1, \dots, C_m$  be the pairwise distinct equivalence classes of  $\equiv_{\mathcal{L}}^c$ . Then, for each  $1 \leq i \leq m$ , there exists a word  $\alpha_i \in C_i$  such that  $|\alpha_i| < n + n^2$ , where  $n := |\mathcal{A}|$ .*

*Proof (Sketch).* Suppose by contradiction that there exists a class  $C_i$  such that for all  $\alpha \in C_i$  it holds  $|\alpha| \geq n + n^2$ , and let  $\alpha \in C_i$  be a word of minimum length. Since  $\mathcal{A}$  has  $n$  states, there must exist a factor of  $\alpha$  that corresponds to a cycle in the run of  $\alpha$  on  $\mathcal{A}$ . Let  $\beta$  be the word obtained by erasing such factor from  $\alpha$ . Since  $\alpha$  and  $\beta$  ends in the same state of  $\mathcal{A}$ , we have  $\alpha \equiv_{\mathcal{L}} \beta$ . From the minimality of  $\alpha$  it follows that  $\alpha \not\equiv_{\mathcal{L}}^c \beta$ , therefore there must exist a word  $\eta$  that is not Myhill-Nerode equivalent to  $\alpha$  and that is included (co-lexicographically) between  $\alpha$  and  $\beta$ . We can further assume that the length of  $\eta$  is greater than  $n^2$ , hence we can identify a common factor of  $\alpha$  and  $\eta$  that labels two cycles in the runs of  $\alpha$  and  $\eta$  on  $\mathcal{A}$ . We can then apply Theorem 1 to conclude that  $\mathcal{L}$  is not Wheeler, a contradiction.

The algorithm works as follow: first of all, it generates the list containing every word of length less than  $n + n^2$  that can be read on  $\mathcal{A}$ . This can be done in exponential time since the list contains at most  $|\Sigma|^{n+n^2}$  words. Then, we order the list co-lexicographically and we apply the following proposition.

**Proposition 5 (Minimum DFA to minimum W DFA).** *Let  $\mathcal{A}$  be the minimum automaton recognizing  $\mathcal{L} = \mathcal{L}(\mathcal{A})$  with  $|\mathcal{A}| = n$ , over the alphabet  $\Sigma$  with*

$|\Sigma| = \sigma$ . Let  $d := n + n^2$  and define

$$\text{Pref}(\mathcal{L})^{\leq d} := \{\alpha \in \text{Pref}(\mathcal{L}) : |\alpha| \leq d\}.$$

Assume that we are given the elements of  $\text{Pref}(\mathcal{L})^{\leq d}$  in co-lexicographic order, i.e.  $\text{Pref}(\mathcal{L})^{\leq d} = \{\alpha_1, \dots, \alpha_k\}$  with  $\alpha_i \prec \alpha_{i+1}$ . Then it is possible to build the minimum W DFA recognizing  $\mathcal{L}$  in  $O(k + n^2 \cdot \sigma \cdot m \log m)$  time, where  $m$  is the dimension of the output.

*Proof (Sketch).* From Lemma 2 we know that each  $\equiv_{\mathcal{L}}^c$ -class has at least one representative in  $\text{Pref}(\mathcal{L})^{\leq d}$ . By considering the list  $[\alpha_1]_{\mathcal{L}}, \dots, [\alpha_k]_{\mathcal{L}}$  we can spot such representatives:  $\alpha_i$  represents a  $\equiv_{\mathcal{L}}^c$ -class if and only if  $\alpha_i$  and  $\alpha_{i+1}$  belong to different  $\equiv_{\mathcal{L}}^c$ -classes or they belong to the same class but their last character is different.

Each representative of a  $\equiv_{\mathcal{L}}^c$ -class become a state of our W DFA. To compute the edges of the automaton, we simply check, for all representative  $\beta$  and for all  $c \in \Sigma$ , the  $\equiv_{\mathcal{L}}^c$ -class of  $\beta \cdot c$ .

## 5 Conclusions

Having an order on the set of states of an automaton that is consistent with a given order of the underlying alphabet has important implication from a practical point of view. In fact, deterministic or even non-deterministic finite state automata enjoying this property—Wheeler automata—can be used to efficiently analyse the language they accept by standard tools.

In this work we established a few limitations along the directions that one can imagine to take in order to extend the nice properties of Wheeler languages.

More specifically, we proved that adding as a degree of freedom the possibility to re-order the underlying alphabet produces a significantly more complex (NP-complete) class of languages. In addition, the polynomial test we have for testing whether a language is Wheeler when the latter is presented by a *deterministic* automaton, turns out much more complex (PSPACE-complete) if the language is presented by a non-deterministic one. The picture is completed by explicitly giving an algorithm that turns a DFA into a Wheeler automaton, whenever this is possible.

Proving the above limitations should clarify the role played by apparently secondary aspects of the definition of Wheeler language. This kind of study should help to better understand the nature of Wheeler languages, with the ultimate goal of singling out any feature that may admit some sort of extension.

Consider, in more general terms, the following open problem: is there a class of automata  $\mathcal{A}$  properly extending the class of Wheeler automata and such that their deterministic equivalent  $\mathcal{D}_{\mathcal{A}}$  is such that  $|\mathcal{D}_{\mathcal{A}}| \in \text{poly}(|\mathcal{A}|)$ ?

## 6 Appendix

**Proof of Corollary 1.** Let  $\mathcal{D} = (\hat{Q}, \hat{q}_0, \hat{\delta}, \hat{F}, \Sigma)$  be the minimum DFA recognizing  $\mathcal{L}$ . Clearly  $\mathcal{D}$  has at most  $2^n$  states.

( $\Leftarrow$ ) From condition 2 it follows that  $\mu\gamma^* \subseteq \text{Pref}(\mathcal{L})$ , so consider the following list of  $2^n + 1$  states of  $\mathcal{D}$ :

$$\hat{\delta}(\hat{q}_0, \mu\gamma^0), \hat{\delta}(\hat{q}_0, \mu\gamma^1), \dots, \hat{\delta}(\hat{q}_0, \mu\gamma^{2^n}).$$

Since  $\mathcal{D}$  has at most  $2^n$  states, there must exist two integers  $0 \leq h < k \leq 2^n$  such that  $\hat{\delta}(\hat{q}_0, \mu\gamma^h) = \hat{\delta}(\hat{q}_0, \mu\gamma^k)$ . Therefore  $\gamma^{k-h}$  labels a cycle starting from  $\hat{\delta}(\hat{q}_0, \mu\gamma^h)$ . Similarly, there exist  $0 \leq h' < k' \leq 2^n$  such that  $\gamma^{k'-h'}$  labels a cycle starting from  $\hat{\delta}(\hat{q}_0, \nu\gamma^{h'})$ . The words

$$\begin{aligned} \hat{\mu} &:= \mu\gamma^h \\ \hat{\nu} &:= \nu\gamma^{h'} \\ \hat{\gamma} &:= \gamma^{\text{lcm}(k-h, k'-h') \cdot 2^n}, \end{aligned}$$

where the factor  $2^n$  in the definition of  $\hat{\gamma}$  ensures that  $|\hat{\mu}|, |\hat{\nu}| \leq |\hat{\gamma}|$ , satisfy condition 2 of Theorem 1. Condition 1 of Theorem 1 follows automatically from conditions 1 of this corollary. Lastly, condition 3 of Theorem 1 follows from conditions 3 of this corollary and the fact that  $\gamma \not\mathcal{A} \mu, \nu$ . Thus we can apply Theorem 1 to conclude that  $\mathcal{L}$  is not Wheeler.

( $\Rightarrow$ ) Since  $\mathcal{L} = \mathcal{L}(\mathcal{D})$  is not Wheeler, let  $\hat{\mu}, \hat{\nu}, \hat{\gamma}$  be as in Theorem 1. The DFA  $\mathcal{D}$  has at most  $2^n$  states, hence the length of  $\hat{\gamma}$  is bounded by the constant  $2^{3n} + 2 \cdot 2^{2n} + 2^n + 2$ . We have  $\hat{\mu}\hat{\gamma}^* \subseteq \text{Pref}(\mathcal{L})$ , so let  $t_0 = q_0, t_1, \dots, t_m$  be a run of  $\hat{\mu}\hat{\gamma}^n$  over  $\mathcal{A}$ . We set  $u := |\hat{\mu}|$  and  $g := |\hat{\gamma}|$ , and consider the list of  $n+1$  states

$$t_u, t_{u+g}, t_{u+2g}, \dots, t_{u+ng} = t_m$$

Since  $\mathcal{A}$  has  $n$  states, there must exist two integers  $0 \leq h < k \leq n$  such that  $t_{u+hg} = t_{u+kg}$ . That is, there exists a state  $p := t_{u+hg}$  such that  $p \in \delta(q_0, \hat{\mu}\hat{\gamma}^h)$  and  $\hat{\gamma}^{k-h}$  labels a cycle starting from  $p$ . We can repeat the same argument for a run of  $\hat{\nu}\hat{\gamma}^n$  over  $\mathcal{A}$  to find a state  $r$  and two integers  $h', k'$  such that  $r \in \delta(q_0, \hat{\nu}\hat{\gamma}^{h'})$  and  $\hat{\gamma}^{k'-h'}$  labels a cycle starting from  $r$ . We can then define the words

$$\begin{aligned} \mu &:= \hat{\mu}\hat{\gamma}^h \\ \nu &:= \hat{\nu}\hat{\gamma}^{h'} \\ \gamma &:= \hat{\gamma}^{\text{lcm}(k-h, k'-h') \cdot n} \end{aligned}$$

which satisfy the conditions 2 and 3.

Condition 4 is satisfied since  $|\hat{\gamma}| \leq 2^{3n} + 2 \cdot 2^{2n} + 2^n + 2$  and  $\text{lcm}(k-h, k'-h') < n^2$ . Finally, condition 1 is satisfied for all  $i, j \geq 0$ . Indeed, for all  $l$  the words  $\hat{\mu}$  and  $\hat{\mu}\hat{\gamma}^l$  lead to the same state of  $\mathcal{D}$ , thus  $\hat{\mu} \equiv_{\mathcal{L}} \hat{\mu}\hat{\gamma}^l$ . Similarly, for all  $l$  we also have  $\hat{\nu} \equiv_{\mathcal{L}} \hat{\nu}\hat{\gamma}^l$ . Since  $\forall i \exists s_i$  such that  $\mu\gamma^i = \hat{\mu}\hat{\gamma}^{s_i}$ , and similarly,  $\forall j \exists s_j$  such that  $\nu\gamma^j = \hat{\nu}\hat{\gamma}^{s_j}$ , the thesis follows from  $\hat{\mu} \not\equiv_{\mathcal{L}} \hat{\nu}$ .  $\square$

**Proof of Proposition 1.** First of all we need to prove that the problem is in PSPACE. We will show instead that its complement is in NPSPACE, then the thesis follows from Savitch's Theorem, which states that NPSPACE = PSPACE, and the fact that PSPACE is closed under complementation. We prove that checking the conditions in Corollary 1 is in NPSPACE. We can use non-determinism to guess, bit by bit, the length of  $\mu, \nu$  and  $\gamma$  and store this guessed information in three counters  $u, v, g$  respectively, using  $O(n)$  space for each. We also need to guess the ending states  $p, r \in Q$  of  $\mu, \nu$ . Then we start guessing the characters of  $\mu, \nu$  and  $\gamma$  starting from their last one and proceeding backwards toward their first one, checking condition 3 of Corollary 1 in constant space. Whenever we guess a character of  $\mu$  (respectively,  $\nu, \gamma$ ) we decrease by one the counter  $u$  ( $v, g$ ), so we know when the guessing stops. If condition 3 is satisfied, we calculate (we will show later how) the sets  $\delta(q_0, \mu)$ ,  $\delta(q_0, \nu)$ ,  $\delta(p, \gamma)$ , and  $\delta(r, \gamma)$  and check condition 2, that is, whether  $p \in \delta(q_0, \mu)$ ,  $r \in \delta(q_0, \nu)$ ,  $p \in \delta(p, \gamma)$ , and  $r \in \delta(r, \gamma)$ . If condition 2 is satisfied, we consider condition 1: for fixed  $i, j \leq 2^n$  consider the automata  $A_{\mu\gamma}^i$  and  $A_{\nu\gamma}^j$  obtained from the NFA  $\mathcal{L}(\mathcal{A})$  by considering as initial states the sets  $\delta(q_0, \mu\gamma^i)$ ,  $\delta(q_0, \nu\gamma^j)$ , respectively. Notice that we have  $\mu\gamma^i \not\equiv_{\mathcal{L}} \nu\gamma^j$  if and only if  $\mathcal{L}(A_{\mu\gamma}^i) \neq \mathcal{L}(A_{\nu\gamma}^j)$  and checking whether  $\mathcal{L}(A_{\mu\gamma}^i) = \mathcal{L}(A_{\nu\gamma}^j)$  can be done in polynomial space, since deciding whether two NFAs recognize the same language is a well-known PSPACE-complete problem.

To conclude the proof, we claim that we are able to calculate in polynomial space, for all  $q \in Q$  and for all  $\beta \in \{\mu, \nu, \gamma\}^*$ , the set  $\delta(q, \beta)$ . While guessing  $\mu$  character by character, we can compute, for each state  $q$ , the set  $Q_{\mu,q}$  of the states from which is possible to reach  $q$  reading  $\mu$ . To build  $Q_{\mu,q}$  we start from the set  $Q^0 := \{q\}$  and we follow backwards the edges entering  $q$  and labeled as the last character of  $\mu$ . We call  $Q^1$  this new set of states and we repeat the process by following backwards the edges entering each state of  $Q^1$  and labeled as the second to last character of  $\mu$ . Proceeding inductively, we compute the sets  $Q^0, Q^1, \dots, Q^u = Q_{\mu,q}$ ; notice that to calculate  $Q^{k+1}$  we only need  $Q^k$  and the  $k$ -th to last character of  $\mu$ , thus we can update (instead of storing) the set  $Q^k$ . We can do the same for  $\nu$  and  $\gamma$  to compute, for each  $q \in Q$ , the sets  $Q_{\nu,q}$  and  $Q_{\gamma,q}$ . Once we have stored, for all  $q \in Q$  and for all  $\alpha \in \{\mu, \nu, \gamma\}$ , the sets  $Q_{\alpha,q}$ , our claim follows easily. For instance, to compute  $\delta(q_0, \mu\gamma)$  we build, for all  $t \in Q$ , the set

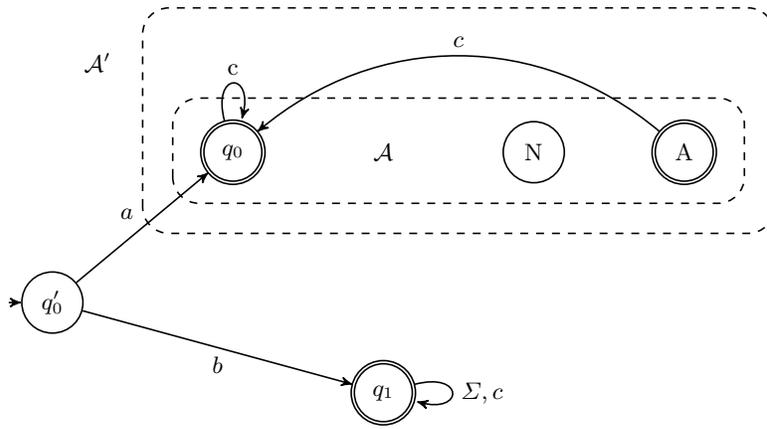
$$Q(t) := \{q \in Q : t \in \delta(q, \mu\gamma)\} = \bigcup_{p \in Q_{\gamma,t}} Q_{\mu,p}.$$

Then we have  $t \in \delta(q_0, \mu\gamma)$  if and only if  $q_0 \in Q(t)$ .

To prove the completeness of the problem, we will show a polynomial reduction from the universality problem for NFA, i.e. the problem of deciding whether the language accepted by a NFA  $\mathcal{A}$  over the alphabet  $\Sigma$  is such that  $\mathcal{L}(\mathcal{A}) = \Sigma^*$ .

Let  $\mathcal{A} = (Q, q_0, \delta, F, \Sigma)$  be a NFA and let  $\mathcal{L} = \mathcal{L}(\mathcal{A})$ . We can assume without loss of generality that  $q_0 \in F$ , otherwise  $\mathcal{A}$  would not accept the empty word and we could immediately derive that  $\mathcal{L} \neq \Sigma^*$ . Let  $a, b, c$  be three characters

not in  $\Sigma$  and such that  $a \prec b \prec c$  with respect to the lexicographical order (the order of the characters of  $\Sigma$  is irrelevant in this proof). First, we build the automaton  $\mathcal{A}'$  starting from  $\mathcal{A}$  by adding an edge  $(q_f, q_0, c)$  for each final state  $q_f \in F$ , see the top part of Figure 1. Notice that  $\mathcal{A}'$  recognizes the language  $\mathcal{L}' = \mathcal{L}(\mathcal{A}') = (\mathcal{L}c)^* \cdot \mathcal{L}$ , and it is straightforward to prove that  $\mathcal{L} = \Sigma^*$  if and only if  $\mathcal{L}' = (\Sigma + c)^*$ : if  $\mathcal{L} = \Sigma^*$ , let  $\alpha$  be a word in  $(\Sigma + c)^*$  containing  $n$  occurrences of  $c$ . Then  $\alpha = \alpha_0 c \alpha_2 c \dots \alpha_{n-1} c \alpha_n$  for some  $\alpha_1, \dots, \alpha_n \in \Sigma^*$ . Hence  $\alpha \in (\Sigma^*c)^* \cdot \Sigma^* = \mathcal{L}'$ . On the other hand, if  $\mathcal{L} \neq \Sigma^*$  let  $\alpha$  be a word in  $\Sigma^* \setminus \mathcal{L}$ . Then  $\alpha \cdot c \notin \mathcal{L}'$ .



**Fig. 4.** The automaton  $\mathcal{A}''$ . Every accepting state of  $\mathcal{A}$ , labeled A in the figure, has a back edge labeled  $c$  connecting it to  $q_0$ . Conversely, non-accepting states of  $\mathcal{A}$ , labeled N in the figure, do not have such back edges.

We build a second automaton  $\mathcal{A}''$  as depicted in Figure 4. Let  $\mathcal{L}'' = \mathcal{L}(\mathcal{A}'')$  be the language recognized by  $\mathcal{A}''$ . We claim that  $\mathcal{L} = \Sigma^*$  if and only if  $\mathcal{L}''$  is Wheeler.

( $\implies$ ) If  $\mathcal{L} = \Sigma^*$ , we have already proved that  $\mathcal{L}' = (\Sigma + c)^*$ . Hence we have  $\mathcal{L}'' = (a + b) \cdot (\Sigma + c)^*$ . The minimum DFA recognizing  $\mathcal{L}''$  has only one loop, therefore by Theorem 1  $\mathcal{L}''$  is Wheeler.

( $\impliedby$ ) If  $\mathcal{L} \neq \Sigma^*$ , let  $\alpha$  be a word in  $\Sigma^* \setminus \mathcal{L}$ . Notice that  $\alpha \neq \varepsilon$  since we assumed that  $\varepsilon \in \mathcal{L}$ . Every possible run of  $\alpha$  over  $\mathcal{A}$  must lead to a non-accepting state, hence  $\alpha \cdot c \notin \mathcal{L}'$ . This implies that for all  $i \geq 0$  we have  $a \cdot c^i \cdot \alpha \cdot c \notin \mathcal{L}''$  (notice that the only edge labeled  $c$  leaving  $q_0$  ends in  $q_0$ ). On the other hand, for all  $j \geq 0$  we have  $bc^j \cdot \alpha \cdot c \in \mathcal{L}''$ , hence for all  $i, j \geq 0$  we have  $ac^i \not\equiv_{\mathcal{L}''} bc^j$ . Thus the following monotone sequence in  $\text{Pref}(\mathcal{L}'')$

$$ac \prec bc \prec acc \prec bcc \prec \dots \prec ac^n \prec bc^n \prec \dots$$

is not eventually constant modulo  $\equiv_{\mathcal{L}''}$ . From Lemma 1 it follows that  $\mathcal{L}''$  is not Wheeler.  $\square$

**Proof of Lemma 2.** Suppose by contradiction that there exists a class  $C_i$  such that for all  $\alpha \in C_i$  it holds  $|\alpha| \geq n + n^2$ , and let  $\alpha \in C_i$  be a word of minimum length. Consider the first  $n + 1$  states  $q_0 = t_0, \dots, t_n$  of  $\mathcal{A}$  visited by reading the first  $n$  characters of  $\alpha$ . Since  $\mathcal{A}$  has only  $n$  states, there must exist  $0 \leq i, j \leq n$  with  $i < j$  such that  $t_i = t_j$ . Let  $\alpha'$  be the prefix of  $\alpha$  of length  $i$  (if  $i = 0$  then  $\alpha' = \varepsilon$ ), let  $\delta$  be the factor of  $\alpha$  of length  $j - i$  labeling the path  $t_i, \dots, t_j$ , and let  $\zeta$  be the suffix of  $\alpha$  such that  $\alpha = \alpha'\delta\zeta$ . By construction, the words  $\alpha$  and  $\beta := \alpha'\zeta$  end in the same state, hence  $\alpha \equiv_{\mathcal{L}} \beta$ . Moreover, from  $|\beta| < |\alpha|$  and the minimality of  $\alpha$  it follows that  $\alpha \not\equiv_{\mathcal{L}}^c \beta$ .

Suppose that  $\alpha \prec \beta$ , the other case being completely symmetrical. Since  $\alpha$  and  $\beta$  share the same suffix  $\zeta$ , they end with the same character. This means that the words  $\alpha$  and  $\beta$ , which are Myhill-Nerode equivalent but not  $\equiv_{\mathcal{L}}^c$  equivalent, were not split into two distinct  $\equiv_{\mathcal{L}}^c$ -classes due to input-consistency, therefore there must exist a word  $\eta$  such that  $\alpha \prec \eta \prec \beta$  and  $\eta \not\equiv_{\mathcal{L}} \alpha$ . Formally, assume by contradiction that for all words  $\eta$  such that  $\alpha \prec \eta \prec \beta$  it holds  $\eta \equiv_{\mathcal{L}} \alpha$ . Then, by definition of  $\equiv_{\mathcal{L}}^c$ , it would follow  $\alpha \equiv_{\mathcal{L}}^c \beta$ , a contradiction.

Let  $\eta$  be a word such that  $\alpha \prec \eta \prec \beta$  and  $\eta \not\equiv_{\mathcal{L}} \alpha$ . From  $\zeta \dashv \alpha, \beta$  it follows that  $\zeta \dashv \eta$ , so we can write  $\eta = \eta'\zeta$  for some  $\eta' \in \Sigma^*$ . Recall that by construction  $\alpha = \alpha'\delta\zeta$  with  $|\alpha'\delta| \leq n$ , hence  $|\zeta| \geq n^2$ . Consider the last  $n^2 + 1$  states  $r_0, \dots, r_{n^2}$  of  $\mathcal{A}$  visited by reading the word  $\alpha$ , and the last  $n^2 + 1$  states  $p_0, \dots, p_{n^2}$  visited by reading the word  $\eta$ . Since  $\mathcal{A}$  has only  $n$  states, there must exist  $0 \leq i, j \leq n^2$  with  $i < j$  such that  $(r_i, p_i) = (r_j, p_j)$ . Notice that it can't be  $r_i = p_i$ , otherwise from the determinism of  $\mathcal{A}$  it would follow  $r_{n^2} = p_{n^2}$ ; from the minimality of  $\mathcal{A}$  it would then follow  $\alpha \equiv_{\mathcal{L}} \eta$ , a contradiction.

Let  $\zeta''$  be the suffix of  $\zeta$  of length  $n^2 - j$ , and let  $\gamma$  be the factor of  $\zeta$  of length  $j - i$  labeling the path  $r_i, \dots, r_j$ . Since  $|\zeta| \geq n^2$ , there exists  $\zeta' \in \Sigma^*$  such that  $\zeta = \zeta'\gamma\zeta''$ . We can then rewrite  $\alpha, \eta$  and  $\beta$  as

$$\begin{aligned}\alpha &= \alpha'\delta\zeta = \alpha'\delta\zeta'\gamma\zeta'' \\ \eta &= \eta'\zeta = \eta'\zeta'\gamma\zeta'' \\ \beta &= \alpha'\zeta = \alpha'\zeta'\gamma\zeta''.\end{aligned}$$

Let  $k$  be an integer such that  $|\gamma^k|$  is greater than  $|\alpha'\delta\zeta'|$  and  $|\eta'\zeta'|$ . Set  $\mu := \eta'\zeta'$ ; from  $\alpha \prec \eta \prec \beta$  it follows that  $\alpha'\delta\zeta' \prec \mu \prec \alpha'\zeta'$ . If  $\gamma^k \prec \mu$  set  $\nu := \alpha'\zeta'$ , otherwise set  $\nu := \alpha'\delta\zeta'$ . In both cases, the hypothesis of Theorem 1 are satisfied, since  $\gamma^k$  labels two cycles starting from the states  $r_i$  and  $p_i$ , that we have proved to be distinct. We can conclude that  $\mathcal{L}$  is not Wheeler, a contradiction, and the thesis follows.  $\square$

**Proof of Proposition 5.** Consider the pairwise distinct equivalence classes  $C_1, \dots, C_m$  of the equivalence  $\equiv_{\mathcal{L}}^c$ . Clearly, the minimum Wheeler automaton recognizing  $\mathcal{L}$  has  $m$  states. We can assume without loss of generality that the

equivalence classes are co-lexicographically ordered, i.e.  $C_i \prec C_{i+1}$  for all  $i < m$ . For sake of simplicity, given a word  $\alpha \in \Sigma^*$  we will write  $[\alpha]$  to indicate its equivalence class modulo  $\equiv_{\mathcal{L}}$ , and we will write  $[\alpha]_W$  to indicate its equivalence class modulo  $\equiv_{\mathcal{L}}^c$ .

Let  $K := \{1, \dots, k\}$  and, for all  $i$ , let  $K_i := \{j \in K : [\alpha_j]_W = C_i\}$ . Since each  $C_i$  is convex in  $\text{Pref}(\mathcal{L})$ , each  $K_i$  must be convex in  $K$ , that is each  $K_i$  is an interval. Therefore the list of equivalence classes  $[\alpha_1]_W, \dots, [\alpha_k]_W$  must be partitioned in consecutive runs of the same class, each class appearing in one and only one run. From Lemma 2 we know that each equivalence class has at least one representative in  $\text{Pref}(\mathcal{L})^{\leq d}$ , hence the list  $[\alpha_1]_W, \dots, [\alpha_k]_W$  must contain exactly  $m$  runs.

For all  $1 \leq j < k$ , we have  $[\alpha_j]_W \neq [\alpha_{j+1}]_W$  if and only if

$$\left([\alpha_j] \neq [\alpha_{j+1}]\right) \vee \left([\alpha_j] = [\alpha_{j+1}] \wedge \text{last}(\alpha_j) \neq \text{last}(\alpha_{j+1})\right),$$

where  $\text{last}(\alpha)$  denotes the last character of  $\alpha$ . This means that we can identify the  $m$  runs of the equivalence  $\equiv_{\mathcal{L}}^c$  just by looking at the two lists  $\alpha_1, \dots, \alpha_k$  and  $[\alpha_1], \dots, [\alpha_k]$ : whenever  $\text{last}(\alpha_j) \neq \text{last}(\alpha_{j+1})$  or  $[\alpha_j] \neq [\alpha_{j+1}]$ , we know that a new  $\equiv_{\mathcal{L}}^c$  run must start at  $\alpha_{j+1}$ .

In  $O(k)$  we are able to determine the  $m$  runs and to pick a representative for each of them, i.e. we can find  $m$  indexes  $i_1, \dots, i_m$  such that for all  $1 \leq j \leq m$  it holds  $\alpha_{i_j} \in C_j$ . We call the set  $\{a_{i_1}, \dots, a_{i_m}\}$  a *fingerprint* of the language  $L$ , i.e. a set of words that has cardinality  $m$  such that distinct elements of the set belong to distinct  $\equiv_{\mathcal{L}}^c$ -classes.

We show how to build the minimum Wheeler DFA recognizing  $\mathcal{L}$ , starting from any fingerprint of  $\mathcal{L}$  and the standard minimum DFA recognizing  $\mathcal{L}$ . Let  $\{\beta_1, \dots, \beta_m\}$  be a fingerprint of  $\mathcal{L}$  and let  $\mathcal{A}$  be the minimum DFA recognizing  $\mathcal{L}$ . We can assume without loss of generality that  $\beta_1 \prec \dots \prec \beta_m$ . We build the automaton  $\mathcal{A}^W = (Q, \beta_1, \delta, F, \Sigma)$ , where the set of states is  $Q = \{\beta_1, \dots, \beta_m\}$  and the set of final states is  $F = \{\beta_j : \beta_j \in \mathcal{L}\}$ . The transition function  $\delta$  can be computed as follow. For all  $1 \leq j \leq m$  and for all  $c \in \Sigma$ , check whether  $\beta_j \cdot c \in \text{Pref}(\mathcal{L})$ . If  $\beta_j \cdot c \notin \text{Pref}(\mathcal{L})$ , there are no edges labeled  $c$  that exit from  $\beta_j$ . If instead  $\beta_j \cdot c \in \text{Pref}(\mathcal{L})$ , locate  $\beta_j \cdot c$  using a binary search. There are three possible cases.

1.  $\beta_j \cdot c \prec \beta_1$ . Then  $\delta(\beta_j, c) = \beta_1$ .
2.  $\beta_m \prec \beta_j \cdot c$ . Then  $\delta(\beta_j, c) = \beta_m$ .
3. There exists  $s$  such that  $\beta_s \preceq \beta_j \cdot c \preceq \beta_{s+1}$ . It can not be the case that both  $\beta_j c \not\equiv_{\mathcal{L}} \beta_s$  and  $\beta_j c \not\equiv_{\mathcal{L}} \beta_{s+1}$ , since  $\{\beta_1, \dots, \beta_m\}$  is a fingerprint of  $\mathcal{L}$ . Hence we distinguish three cases.
  - (a)  $\beta_s \equiv_{\mathcal{L}} \beta_j c \not\equiv_{\mathcal{L}} \beta_{s+1}$ . Then  $\delta(\beta_j, c) = \beta_s$ .
  - (b)  $\beta_s \not\equiv_{\mathcal{L}} \beta_j c \equiv_{\mathcal{L}} \beta_{s+1}$ . Then  $\delta(\beta_j, c) = \beta_{s+1}$ .
  - (c)  $\beta_s \equiv_{\mathcal{L}} \beta_j c \equiv_{\mathcal{L}} \beta_{s+1}$ . Since  $\{\beta_1, \dots, \beta_m\}$  is a fingerprint of  $\mathcal{L}$ , it is either  $c = \text{last}(\beta_j c) = \text{last}(\beta_s)$ , in which case  $\delta(\beta_j, c) = \beta_s$ , or  $c = \text{last}(\beta_{s+1})$ , in which case  $\delta(\beta_j, c) = \beta_{s+1}$ .

□

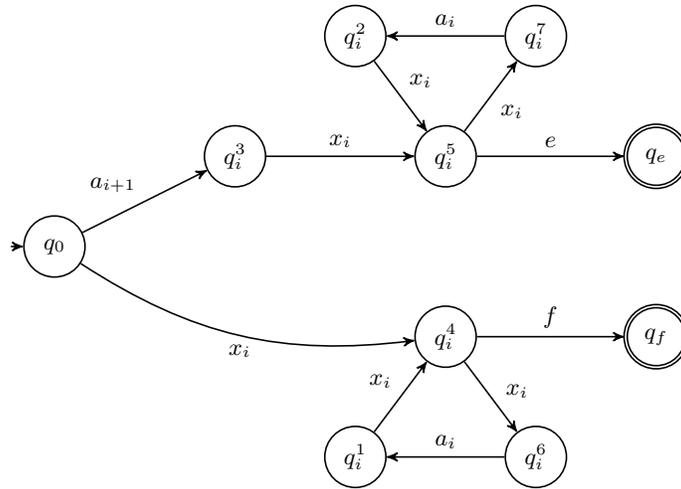
**Proof of Proposition 3.** The problem is in NP, since we can use non-determinism to guess the order of the alphabet and then check whether such order makes the NFA Wheeler.

To prove the hardness, we show a polynomial reduction from the problem of deciding whether a NFA is Wheeler. Let  $\mathcal{A}$  be a NFA with initial state  $q_0$ , over the alphabet  $\Sigma = \{a_1 \dots, a_\sigma\}$  ordered by the relation  $a_1 \prec \dots \prec a_\sigma$ . We want to build a new automaton  $\mathcal{A}'$  such that  $\mathcal{A}'$  is a GWNFA if and only if  $\mathcal{A}$  is Wheeler.  $\mathcal{A}'$  will be an automaton of size  $|\mathcal{A}| + O(\sigma)$  over the alphabet  $\Sigma'$  of size  $O(\sigma)$ .

The automaton  $\mathcal{A}'$  will be built starting from  $\mathcal{A}$  and adding extra states and transitions. We define the new alphabet as

$$\Sigma' = \{a_1, \dots, a_\sigma, x_1, \dots, x_{\sigma-1}, e, f\},$$

with  $x_i, e, f \notin \Sigma$ , and we add two final states  $q_e$  and  $q_f$ . We then build  $\sigma - 1$  gadgets, one for each pair of consecutive characters  $(a_i, a_{i+1})$  of  $\Sigma$ , each one connected to  $\mathcal{A} \cup \{q_e, q_f\}$  as depicted in Figure 5. This completes the construction of the automaton  $\mathcal{A}'$ . Notice that  $\mathcal{A}$  is input-consistent, but in general it is not deterministic.



**Fig. 5.** The gadget  $G_i$ , connected to  $q_0$  and to the sinks  $q_e$  and  $q_f$ .

We want to show that  $\mathcal{A}$  is Wheeler according to the order  $(\Sigma, \prec)$  if and only if  $\mathcal{A}'$  is a GWNFA.

( $\implies$ ) Define the order  $\prec'$  over  $\Sigma'$  by setting

$$a_1 \prec' \dots \prec' a_\sigma \prec' x_1 \prec' \dots \prec' x_{\sigma-1} \prec' e \prec' f.$$

We show that  $\mathcal{A}'$  is Wheeler according to  $(\Sigma', \prec')$  by ordering its states. Since  $\mathcal{A}$  is Wheeler, there already exists an order of its states that makes  $\mathcal{A}$  Wheeler. Therefore, we simply need to extend this order to the states of  $\mathcal{A}'$ . Recall that in [1] it has been proved that a WNFA is Wheeler if and only if, for each pair of states  $q, p$  such that  $I_q \neq I_p$ , either  $I_q \preceq I_p$  or  $I_p \preceq I_q$  holds, where by definition  $I_q \preceq I_p$  if and only if, for all  $\alpha \in I_q$  and for all  $\beta \in I_p$  such that  $\{\alpha, \beta\} \not\subseteq I_q \cap I_p$ , we have  $\alpha \prec \beta$ . Therefore we check, for each pair of states  $q$  and  $p$  in  $\mathcal{A}'$ , that either  $I_q \preceq I_p$  (implying  $q < p$ ) or  $I_p \preceq I_q$  (implying  $p < q$ ) holds. Note that when  $I_q$  and  $I_p$  are disjoint, the condition  $I_q \preceq I_p$  translates to the following: for all  $\alpha \in I_q$  and for all  $\beta \in I_p$ ,  $\alpha \prec \beta$ . In the discussion that follows we never compare two states that belong to  $\mathcal{A}$ , since the order between them is already established.

First of all, we will order, for each  $i$ , the states with incoming edge  $x_i$ . Since  $x_i \notin \Sigma$ , the only states to compare are the one belonging to the gadget  $G_i$ , i.e.  $q_i^4, q_i^5, q_i^6, q_i^7$ . Consider the languages

$$\begin{aligned} I_4 &:= I_{q_i^4} = x_i(x_i a_i x_i)^* \\ I_5 &:= I_{q_i^5} = a_{i+1} x_i(x_i a_i x_i)^* \\ I_6 &:= I_{q_i^6} = x_i x_i(a_i x_i x_i)^* \\ I_7 &:= I_{q_i^7} = a_{i+1} x_i x_i(a_i x_i x_i)^*. \end{aligned}$$

Since  $a_{i+1} \prec' x_i$  we have  $I_4, I_5 \prec I_6, I_7$ . Moreover, consider any two words  $\alpha \in I_4$  and  $\beta \in I_5$ . If  $|\alpha| < |\beta|$ , we have  $\alpha \prec \beta$  hence  $\alpha \prec \beta$ . If  $|\alpha| \geq |\beta|$  instead, then the last  $|\beta|$  characters of  $\alpha$  must be  $a_i x_i(x_i a_i x_i)^m$ , where  $m$  is such that  $\beta = a_{i+1} x_i(x_i a_i x_i)^m$ . From  $a_i \prec' a_{i+1}$  we still get  $\alpha \prec \beta$ , hence  $I_4 \prec I_5$ . Similarly we can prove that  $I_6 \prec I_7$ . It immediately follows that we need to order the states as follows:  $q_i^4 < q_i^5 < q_i^6 < q_i^7$ .

Secondly, for each  $1 \leq i \leq \sigma$  we need to sort the states with incoming edges labeled  $a_i$ . Note that the automaton  $\mathcal{A}$  might contain states with incoming label  $a_i$  and we need to consider such states as well. For  $i = \sigma$ , the task is easy. The only gadget with a state labeled  $a_\sigma$  is  $G_{\sigma-1}$ , and such state is  $q_{\sigma-1}^3$ . Notice that  $I_{q_{\sigma-1}^3} = \{a_\sigma\}$ . Let  $q$  be a state of  $\mathcal{A}$  with  $\lambda(q) = a_\sigma$ . Since every word in  $I_q$  ends with  $a_\sigma$ , it trivially follows that  $I_{q_{\sigma-1}^3} \preceq I_q$ . If  $I_q \neq \{a_\sigma\}$ , we are forced to set  $q_{\sigma-1}^3 < q$ . If instead  $I_q = \{a_\sigma\}$ , the order of  $q$  and  $q_{\sigma-1}^3$  does not matter. For sake of consistency, we set  $q_{\sigma-1}^3 < q$ .

For  $i = 1$ , the only gadget with states labeled  $a_1$  is  $G_1$ , and such states are  $q_1^1$  and  $q_1^2$ . Let  $q$  be a state of  $\mathcal{A}$  with  $\lambda(q) = a_1$  and consider the languages

$$\begin{aligned} I_1 &:= I_{q_1^1} = x_1 x_1 a_1 (x_1 x_1 a_1)^* \\ I_2 &:= I_{q_1^2} = a_2 x_1 x_1 a_1 (x_1 x_1 a_1)^*. \end{aligned}$$

For all  $\alpha \in I_q$ , we have either  $\alpha = a_1$  or  $\alpha = \alpha' a_j a_1$  for some  $\alpha' \in \Sigma^*$  and some  $1 \leq j \leq \sigma$ . In both cases,  $\alpha$  must precede co-lexicographically every word of  $I_1$  and  $I_2$ , thus  $I_q \prec I_1, I_2$ . To compare  $I_1$  and  $I_2$ , consider any two words  $\alpha \in I_1$  and  $\beta \in I_2$ . If  $|\alpha| < |\beta|$ , we have  $\alpha \dashv \beta$  hence  $\alpha \prec \beta$ . If  $|\alpha| \geq |\beta|$  instead, then the last  $|\beta|$  characters of  $\alpha$  must be  $a_1 x_1 x_1 a_1 (x_1 x_1 a_1)^m$ , where  $m$  is such that  $\beta = a_2 x_1 x_1 a_1 (x_1 x_1 a_1)^m$ . From  $a_1 \prec' a_2$  we still get  $\alpha \prec \beta$ , hence  $I_1 \prec I_2$ . It follows that  $q_1^1$  and  $q_1^2$  must follow  $q$  (and all the other states of  $\mathcal{A}$  with label  $a_1$ ), and we must set  $q_1^1 < q_1^2$ .

Lastly, if  $1 < i < \sigma$  we should compare the states  $q_{i-1}^3$ ,  $q_i^1$  and  $q_i^2$  with the states in  $Q_i := \{q \in \mathcal{A} : \lambda(q) = a_i\}$ . Applying both of the reasoning discussed in the cases  $i = 1$  and  $i = \sigma$ , we can conclude that the state  $q_{i-1}^3$  precedes all the states in  $Q_i$  and that the states  $q_i^1$  and  $q_i^2$  follow all the states in  $Q_i$  and must be ordered as  $q_i^1 < q_i^2$ .

The order of the states of  $\mathcal{A}'$  that we described makes  $\mathcal{A}'$  Wheeler with respect to  $(\Sigma', \prec')$ , hence making it a GWNFA.

( $\Leftarrow$ ) If  $\mathcal{A}'$  is a GWNFA, then there exists an order  $\prec'$  over  $\Sigma'$  that makes  $\mathcal{A}'$  Wheeler. Since  $\mathcal{A}$  is a sub-automaton of  $\mathcal{A}'$ , it follows that even  $\mathcal{A}$  is Wheeler according to  $\prec'$ . Let  $\tilde{\prec}$  be the restriction of  $\prec'$  over the alphabet  $\Sigma$ ; we want to show that  $\tilde{\prec}$  is the same order as  $\prec$ . Assume by contradiction that  $\tilde{\prec} \neq \prec$ . If for all  $1 \leq i < \sigma$  we have  $a_i \tilde{\prec} a_{i+1}$ , then  $\tilde{\prec} = \prec$ , a contradiction. Hence there exists  $1 \leq i < \sigma$  such that  $a_{i+1} \tilde{\prec} a_i$ . Since  $\prec'$  extends  $\tilde{\prec}$ , this implies that  $a_{i+1} \prec' a_i$ . We will show that  $\mathcal{A}'$  is not Wheeler according to  $\prec'$ , a contradiction. Define the words  $\mu := x_i$ ,  $\nu := a_{i+1} x_i$  and  $\gamma := x_i a_i x_i$ . From  $\mu \dashv \gamma$  and  $a_{i+1} \prec' a_i$  we have  $\mu, \nu \prec \gamma$ . The word  $\gamma$  labels two cycles in  $\mathcal{A}'$  starting from two distinct states, i.e.  $q_i^4$  and  $q_i^5$ . Moreover,  $\mu$  and  $\nu$  label two paths that start from the initial state  $q_0$  and end in  $q_i^4$  and  $q_i^5$  respectively. Since  $q_i^4$  and  $q_i^5$  are not Myhill-Nerode equivalent, we can apply Theorem 1 to conclude that  $\mathcal{A}'$  is not Wheeler according to  $\prec'$ , a contradiction. Thus  $\tilde{\prec}$  and  $\prec$  coincide.

We have shown that  $\mathcal{A}$  is Wheeler according to  $\prec'$ , and that  $\prec'$  extends  $\prec$ . Therefore we can conclude that  $\mathcal{A}$  is Wheeler according to  $\prec$ .  $\square$

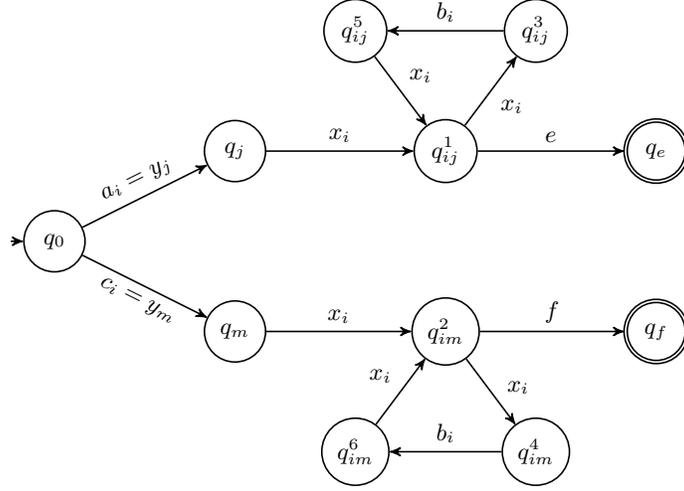
**Definition 5 (Betweenness).** *Input: a list  $Y$  of  $n$  distinct elements  $Y = y_1, \dots, y_n$  and  $k < n^3$  ordered triples  $(a_1, b_1, c_1), \dots, (a_k, b_k, c_k)$ , where each element of each triple belongs to  $Y$ . Elements belonging to the same triple are distinct.*

*Output: yes/no answer. The answer is “yes” if and only if there exists a total order  $<$  of  $Y$  such that, for each  $k$ , either  $a_k < b_k < c_k$  or  $a_k > b_k > c_k$ .*

**Proof of Proposition 4.** We can prove that both problems are in NP using an argument similar to the one employed in the proof of Proposition 3.

To prove the hardness, we show a polynomial reduction from the betweenness problem to both of the problems described; we will use exactly the same reduction for both problems. We start from an instance  $I = (Y, K)$  of the betweenness problem, where  $Y$  is the set  $Y = \{y_1 \dots, y_n\}$  and  $K \subseteq \mathcal{P}(T^3)$  is a collection of  $k$  triples  $(a_1, b_1, c_1), \dots, (a_k, b_k, c_k)$ , for some  $1 < k < n^3$ . We build a DFA  $\mathcal{A}$  of size  $O(n + k)$ , over an alphabet of size  $O(n + k)$ . The alphabet is

$\Sigma = Y \cup \{x_1, \dots, x_k, e, f\}$ , where we introduce a new character  $x_i$  for each triple  $(a_i, b_i, c_i) \in K$  and two extra “ending” characters  $e$  and  $f$ . To build  $G$ , we start with the initial state  $q_0$  connected with  $n$  states  $q_1, \dots, q_n$  through the edges  $(q_0, y_j, q_j)$  for each  $1 \leq j \leq n$ . We also add two sinks  $q_e$  and  $q_f$ , the only final states. We add the states  $q_{ij}^1, q_{ij}^3, q_{ij}^5$  (see Figure 6) and the transitions



**Fig. 6.** The gadget  $G_i$  related to the  $i$ -th triple.

$$\delta(q_j, x_i) = q_{ij}^1, \quad \delta(q_{ij}^1, x_i) = q_{ij}^3, \quad \delta(q_{ij}^3, b_i) = q_{ij}^5, \quad \delta(q_{ij}^5, x_i) = q_{ij}^1$$

and  $\delta(q_{ij}^1, e) = q_e$ . We repeat the same process with  $c_i$ : given the integer  $m$  such that  $c_i = y_m$ , we add the states  $q_{im}^2, q_{im}^4, q_{im}^6$  and the transitions

$$\delta(q_m, x_i) = q_{im}^2, \quad \delta(q_{im}^2, x_i) = q_{im}^4, \quad \delta(q_{im}^4, b_i) = q_{im}^6, \quad \delta(q_{im}^6, x_i) = q_{im}^2$$

and  $\delta(q_{im}^2, f) = q_f$ . Lastly, we remove the states among  $q_1, \dots, q_n$  that don't have outgoing edges. More formally, we define the sets  $A := \{a_1, \dots, a_k\}$  and  $C := \{c_1, \dots, c_k\}$  and we remove from  $G$  all the states  $q_j$  such that  $y_j \notin A \cup C$ . We show that the instance  $I = (Y, K)$  of the betweenness problem is satisfiable if and only if  $\mathcal{A}$  is a GWNFA, if and only if  $\mathcal{L}(\mathcal{A})$  is GW.

$(\implies)$  Since  $I = (Y, K)$  is satisfiable, there exists an ordering  $\pi : Y \rightarrow \{1, \dots, n\}$  of the elements of  $Y$  satisfying  $I$ . We order  $\Sigma$  as follows:

$$\pi^{-1}(1) \prec \dots \prec \pi^{-1}(n) \prec x_1 \prec \dots \prec x_k \prec e \prec f.$$

This ordering induce a partial order on the states of  $\mathcal{A}$ , where states with different incoming labels are ordered by such labels. Therefore we only need to order

the states of  $\mathcal{A}$  with the same incoming label.

For each  $1 \leq i \leq k$ , the only states of  $\mathcal{A}$  with incoming label  $x_i$  are  $q_{ij}^1, q_{ij}^3, q_{im}^2, q_{im}^4$ , where  $j$  and  $m$  are integers such that  $a_i = y_j$  and  $c_i = y_m$ . Since, by construction, the order  $\pi$  satisfies the instance  $I$ , then only two cases can occur: either  $\pi(a_i) < \pi(b_i) < \pi(c_i)$ , or  $\pi(c_i) < \pi(b_i) < \pi(a_i)$ . In the first case, we set  $q_{ij}^1 < q_{im}^2 < q_{ij}^3 < q_{im}^4$ . To realize that this is in fact the correct order of the states, consider the following languages:

$$\begin{aligned} I_1 &:= I_{q_{ij}^1} = \{\alpha \in \Sigma^* : \delta(q_0, \alpha) = q_{ij}^1\} = a_i x_i (x_i b_i x_i)^* \\ I_2 &:= I_{q_{im}^2} = c_i x_i (x_i b_i x_i)^* \\ I_3 &:= I_{q_{ij}^3} = a_i x_i x_i (b_i x_i x_i)^* \\ I_4 &:= I_{q_{im}^4} = c_i x_i x_i (b_i x_i x_i)^*. \end{aligned}$$

Since, by construction, we have  $a_i, b_i, c_i \prec x_i$ , it follows that  $I_1, I_2 \prec I_3, I_4$ . Moreover, from  $\pi(a_i) < \pi(b_i) < \pi(c_i)$  we also have that  $I_1 \prec I_2$  and  $I_3 \prec I_4$ , which completes the ordering. Symmetrically, if  $\pi(c_i) < \pi(b_i) < \pi(a_i)$  then we set  $q_{im}^2 < q_{ij}^1 < q_{im}^4 < q_{ij}^3$ .

We still need to order the states of  $\mathcal{A}$  whose incoming labels belong to  $Y$ . For each  $1 \leq p \leq n$ , the states with incoming label  $y_p$  belong to the sets  $V_5 := \{q_{ij}^5 : b_i = y_p\}$  or  $V_6 := \{q_{im}^6 : b_i = y_p\}$  or  $V_p$ , where  $V_p = \{q_p\}$  if  $q_p$  is a state of  $\mathcal{A}$  (i.e. if  $y_p \in A \cup C$ ) and  $V_p = \emptyset$  otherwise. If  $V_p = \{q_p\}$ , we set  $q_p$  as the smallest state. We then sort the states of  $V_5$  and  $V_6$  by their first subscript; when the first subscript is equal, i.e. the two states that we want to confront are  $q_{ij}^5$  and  $q_{im}^6$  (with  $a_i = y_j$  and  $c_i = y_m$ ), then we set  $q_{ij}^5 < q_{im}^6$  if  $\pi(a_i) < \pi(b_i) < \pi(c_i)$  and we set  $q_{im}^6 < q_{ij}^5$  if  $\pi(c_i) < \pi(b_i) < \pi(a_i)$ . This can be deduced by confronting the following languages:

$$\begin{aligned} I_5 &:= I_{q_{ij}^5} = a_i x_i x_i b_i (x_i x_i b_i)^* \\ I_6 &:= I_{q_{i'm}^6} = c_{i'} x_{i'} x_{i'} b_{i'} (x_{i'} x_{i'} b_{i'})^*. \end{aligned}$$

If  $i < i'$  then, by construction, we have  $x_i \prec x_{i'}$ , hence  $I_5 \prec I_6$ . Symmetrically, if  $i' < i$  then we have  $I_6 \prec I_5$ . Lastly, if  $i = i'$  then the order between  $I_5$  and  $I_6$  is determined solely by  $a_i$  and  $c_{i'} = c_i$ .

Since there is only one state with incoming label  $e$  and only one state with incoming label  $f$ , we have finished. The order described makes  $\mathcal{A}$  Wheeler, thus  $\mathcal{A}$  is a GWNFA and  $\mathcal{L}(\mathcal{A})$  is GW.

( $\Leftarrow$ ) Assume that the instance  $I = (Y, K)$  of the betweenness problem is unsatisfiable. We prove that  $\mathcal{L}(\mathcal{A})$  is not GW. Assume by contradiction that  $\mathcal{L}(\mathcal{A})$  is GW, then there exists an ordering  $\pi'$  of the elements of  $\Sigma$  such that  $\mathcal{L}(\mathcal{A})$  is Wheeler according to said order. Recall that  $Y \subseteq \Sigma$  and consider the order  $\pi := \pi'|_Y$ . Since  $(Y, K)$  is unsatisfiable,  $\pi$  must violate one of the constraints, i.e. there exists an  $1 \leq i \leq k$  such that either  $\pi(a_i), \pi(c_i) < \pi(b_i)$  or  $\pi(b_i) < \pi(a_i), \pi(c_i)$ . Define the words  $\mu := a_i x_i$ ,  $\nu := c_i x_i$  and  $\gamma := x_i b_i x_i$ ; then it is either  $\mu, \nu \prec \gamma$  or  $\gamma \prec \mu, \nu$  (here the co-lexicographic order  $\prec$  is calculated

with respect to  $\pi'$ ). By construction,  $\gamma$  labels two cycles in  $\mathcal{A}$  starting from two distinct states,  $q_{ij}^1$  and  $q_{im}^2$ , which are not Myhill-Nerode equivalent. Moreover,  $\mu$  and  $\nu$  label two paths that start from the initial state  $q_0$  and end in  $q_{ij}^1$  and  $q_{im}^2$  respectively. We can then apply the Theorem 1 to conclude that  $\mathcal{L}(\mathcal{A})$  is not Wheeler according to  $\pi'$ , a contradiction. Therefore  $\mathcal{L}(\mathcal{A})$  is not GW, which automatically implies that  $\mathcal{A}$  is not a GWNFA.  $\square$

## References

1. Alanko, J., D'Agostino, G., Policriti, A., Prezza, N.: Regular languages meet prefix sorting. In: Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms. pp. 911–930 (2020). <https://doi.org/10.1137/1.9781611975994.55>, <https://epubs.siam.org/doi/abs/10.1137/1.9781611975994.55>
2. Alanko, J., D'Agostino, G., Policriti, A., Prezza, N.: Wheeler Languages. CoRR arXiv:2002.10303 (Feb 2020)
3. Alanko, J., Gagie, T., Navarro, G., Seelbach Benkner, L.: Tunneling on wheeler graphs. In: 2019 Data Compression Conference (DCC). pp. 122–131 (2019). <https://doi.org/10.1109/DCC.2019.00020>
4. Backurs, A., Indyk, P.: Which regular expression patterns are hard to match? In: 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS). pp. 457–466 (2016). <https://doi.org/10.1109/FOCS.2016.56>
5. Equi, M., Grossi, R., Mäkinen, V.: On the Complexity of Exact Pattern Matching in Graphs: Binary Strings and Bounded Degree. In: ICALP 2019 - 46th International Colloquium on Automata, Languages and Programming. pp. 1–15. Patras, Greece (Jul 2019), <https://hal.inria.fr/hal-02338498>
6. Equi, M., Mäkinen, V., Tomescu, A.I.: Graphs cannot be indexed in polynomial time for sub-quadratic time string matching, unless seth fails. In: Bureš, T., Dondi, R., Gamper, J., Guerrini, G., Jurdziński, T., Pahl, C., Sikora, F., Wong, P.W. (eds.) SOFSEM 2021: Theory and Practice of Computer Science. pp. 608–622. Springer International Publishing, Cham (2021)
7. Gibney, D., Hoppenworth, G., Thankachan, S.V.: Simple reductions from formula-sat to pattern matching on labeled graphs and subtree isomorphism. In: Le, H.V., King, V. (eds.) 4th Symposium on Simplicity in Algorithms, SOSA 2021, Virtual Conference, January 11-12, 2021. pp. 232–242. SIAM (2021). <https://doi.org/10.1137/1.9781611976496.26>, <https://doi.org/10.1137/1.9781611976496.26>
8. Gibney, D., Thankachan, S.V.: On the hardness and inapproximability of recognizing wheeler graphs. In: 27th Annual European Symposium on Algorithms, ESA 2019, September 9-11, 2019, Munich/Garching, Germany. pp. 51:1–51:16 (2019). <https://doi.org/10.4230/LIPIcs.ESA.2019.51>, <https://doi.org/10.4230/LIPIcs.ESA.2019.51>
9. Hopcroft, J.E.: An  $n \log n$  algorithm for minimizing states in a finite automaton. Tech. rep., Stanford University (January 1971)
10. Potchin, A., Shallit, J.: Lengths of words accepted by nondeterministic finite automata. Information Processing Letters **162**, 105993 (2020). <https://doi.org/https://doi.org/10.1016/j.ipl.2020.105993>, <https://www.sciencedirect.com/science/article/pii/S0020019020300806>

11. Prezza, N.: On locating paths in compressed tries. In: Proceedings of the Thirty-Second Annual ACM-SIAM Symposium on Discrete Algorithms. p. 744–760. Society for Industrial and Applied Mathematics, USA (2021)
12. Shyr, H., Thierrin, G.: Ordered automata and associated languages. *Tamkang J. Math* (5), 9–20 (1974)
13. Travis Gagie, Giovanni Manzini e Sirén, J.: Wheeler graphs: A framework for bwt-based data structures. *Theoretical computer science* **698**, 67–78 (2017)