Probabilistic and Nondeterministic Unary Automata

Gregor Gramlich[⋆]

Institut für Informatik
Johann Wolfgang Goethe–Universität Frankfurt
Robert-Mayer-Straße 11-15
60054 Frankfurt am Main, Germany
gramlich@thi.informatik.uni-frankfurt.de
Fax: +49 - 69 - 798-28814

Published in
Proceedings of Mathematical Foundations of Computer Science 2003, pp. 460–469.
Lecture Notes in Computer Science 2747

Abstract. We investigate unary regular languages and compare deterministic finite automata (DFA's), nondeterministic finite automata (NFA's) and probabilistic finite automata (PFA's) with respect to their size.
Given a unary PFA with n states and an ϵ-isolated cutpoint, we show that the minimal equivalent DFA has at most $n^{\frac{1}{2\epsilon}}$ states in its cycle. This result is almost optimal, since for any $\alpha < 1$ a family of PFA's can be constructed such that every equivalent DFA has at least $n^{\frac{\alpha}{2\epsilon}}$ states. Thus we show that for the model of probabilistic automata with a constant error bound, there is only a polynomial blowup for cyclic languages.
Given a unary NFA with n states, we show that efficiently approximating the size of a minimal equivalent NFA within the factor $\frac{\sqrt{n}}{\ln n}$ is impossible unless $P = NP$. This result even holds under the promise that the accepted language is cyclic. On the other hand we show that we can approximate a minimal NFA within the factor $\ln n$, if we are given a cyclic unary n-state DFA.

1 Introduction

Regular languages and finite state automata as their acceptance devices, are well studied objects. We consider DFA's, NFA's and PFA's with isolated cutpoint and compare their sizes.

For an n-state PFA with ϵ-isolated cutpoint, the equivalent DFA needs at most $(1 + \frac{1}{2\epsilon})^{n-1}$ states [10]. For a unary alphabet, Milani and Pighizzini [9] show the tight bound[1] of $e^{\Theta(\sqrt{n \ln n})}$ for the number of states in the cycle of the minimal DFA. This result does not depend on the size of the isolation ϵ and the proof of the lower bound actually relies on an isolation that tends to zero. We show that the isolation ϵ plays a crucial role, namely that L can be accepted by a DFA with at most $n^{\frac{1}{2\epsilon}}$ states in its cycle. Thus, for constant isolation ϵ, we improve the upper bound of Milani and Pighizzini to be a polynomial in n.

The minimization problem for DFA's can be efficiently solved. But for a given DFA, the problem of determining the minimal number of states of an equivalent **NFA** is $PSPACE$-complete [6]. A result of Stockmeyer and Meyer [11] shows that the problem of minimizing a given NFA is $PSPACE$-complete for a binary alphabet and NP-complete for a unary alphabet.

We show that, given an n-state NFA accepting L, it is impossible to efficiently approximate the number of states of a minimal NFA accepting L within a factor of $\frac{\sqrt{n}}{\ln n}$ unless $P = NP$. This result holds even under the promise that L is a unary cyclic language and can be extended to PFA's with isolated cutpoint. On the other hand we show that if we are given a unary cyclic n-state DFA accepting L, then we can efficiently construct an equivalent NFA with at most $k \cdot (1 + \ln n)$ states, where k is the number of states of a minimal NFA accepting L. This contrasts with a result of

⋆ partially supported by DFG project SCHN503/2-1

1 The bound of $\Theta(e^{\sqrt{n \ln n}})$ stated in the article had to be corrected, due to [2].

Jiang et al. [5] who show that the number of states of a minimal NFA, equivalent to a given unary DFA, cannot be computed in polynomial time, unless $NP \subseteq DTIME(n^{O(\ln n)})$. This result even holds, if we restrict the DFA to accept only cyclic languages.

The next section gives a short introduction into unary NFA's and unary PFA's. Unary PFA's with ϵ-isolated cutpoint, resp. unary NFA's, are investigated in sections 3 and 4 respectively.

2 Preliminaries

We consider unary languages $L \subseteq \{a\}^*$. A unary regular language is recognized by a DFA that starts with a possibly empty path and ends in a non-empty cycle.

A language L is *ultimately d-cyclic*, if there is a $\mu \in \mathbb{N}_0$, so that $(a^j \in L \Leftrightarrow a^{j+d} \in L)$ holds for any $j \geq \mu$ and we say that d is an *ultimate period* of L. A smallest ultimate period is called the *minimal ultimate period* $c(L)$ and any ultimate period is a multiple of the minimal ultimate period. L is called *cyclic*, if the path of the minimal DFA for L is empty. For cyclic languages we use the term *period* instead of ultimate period and *d-cyclic* (resp. minimally d-cyclic) instead of ultimately d-cyclic (resp. minimally utimately d-cyclic).

The *size* of an automaton A is the number of states of A. For a given regular language L, we use $\mathrm{nsize}(L)$ as the minimal size of an NFA accepting L.

A normal form for unary NFA's is established by Chrobak in [1]. His construction converts a given NFA N with n states into an equivalent NFA N' consisting of a deterministic path and several deterministic cycles. Only the last state of the path branches nondeterministically into one state of each cycle. The path of N' has length $O(n^2)$, and the number of all states in the cycles is bounded by n. Chrobak proves, that $L(N')$ is ultimately d-cyclic, where d is the least common multiple of the length of the cycles in N'. For cyclic languages we introduce union automata as automata in Chrobak normal form with an empty path.

Definition 1. *A union automaton U is described by a collection $(A_1, \ldots, A_k)$ of cyclic DFA's. U accepts an input w iff there is an A_i, such that A_i accepts w. The size of U is defined as $\sum_{i=1}^{k} s_i$, where s_i is the number of states of A_i.*

To convert a union automaton U into an NFA with a single inital state, we simply add one state q_0 and transitions from q_0 to each state that succeeds an initial state of the deterministic automata that U consists of.

Jiang, McDowell and Ravikumar [5] show a structural result about minimal unary NFA's accepting cyclic languages.

Fact 1. *[5] Let L be a minimally D-cyclic unary language.*

Every minimal NFA accepting L can be obtained by converting some minimal union automaton U accepting L into an NFA. Moreover D is the least common multiple of the cycle lengths of U.

Consider the prime factorization of $D = p_1^{\alpha_1} \cdot \ldots \cdot p_r^{\alpha_r}$, where the p_i are distinct and $\alpha_i \in \mathbb{N}$, then every NFA accepting L has at least $p_1^{\alpha_1} + \ldots + p_r^{\alpha_r}$ states.

This result offers some clues about the composition of the (ultimate) period of a unary language which also apply to probabilistic finite automata which we define as follows. A unary PFA M with a set Q of n states is described by a stochastic $n \times n$ matrix A, a stochastic row vector π representing the initial distribution, and a column vector $\eta \in \{0,1\}^n$ indicating the final states. Observe that $\pi A^j \eta$ is the acceptance probability for input a^j. The language accepted by M with respect to a cutpoint $\lambda \in [0,1]$ is $L(M, \lambda) = \{a^j | \pi A^j \eta > \lambda\}$. We call cutpoint λ ϵ-isolated, if for any $j \in \mathbb{N}_0$: $|\pi A^j \eta - \lambda| \geq \epsilon$. We call a cutpoint isolated, if there is an $\epsilon > 0$, so that it is ϵ-isolated.

We regard A as the stochastic matrix of a finite Markov chain $\mathcal{M}$, with rows and columns indexed by states, and consider the representation of $\mathcal{M}$ as a directed graph $G_A = (V, E)$ with $V = Q$. An arc from state q to state p exists in G_A, if $A_{p,q} > 0$. We call a strongly connected component $B \subseteq Q$ in G_A ergodic[2], if starting in any state $q \in B$, we cannot reach any state outside of B. States within an ergodic component are called ergodic states, non-ergodic states are called transient. For an ergodic component B, the period of $q \in B$ is defined as

[2] Unlike some authors we do not require an ergodic component to be aperiodic.

$$d_q = \gcd\{j| \text{ starting in } q \text{ one can reach } q \text{ with exactly } j \text{ steps}\}.$$

All states $q \in B$ have the same period $d = d_q$, which we call the period of B.

Factorization and primality play an important role for (ultimate) periods. To estimate the size of the i-th prime number we use the following fact.

Fact 2. *[4] If p_i is the i-th prime number, then $i \ln i \leq p_i \leq 2i \ln i$ for $i \geq 3$.*

3 Unary PFA's with ϵ-Isolated Cutpoint

In [9] Milani and Pighizzini show, that the ergodic components of a unary PFA with isolated cutpoint basically play the same role as the cycles of an NFA in Chrobak normal form. The least common multiple D of the periods of these components is an ultimate period of the language $L(M, \lambda)$ accepted by the PFA.

This result does not take the isolation into account and yields an exponential upper bound for the ultimate period, namely $c(L(M, \lambda)) = e^{O(\sqrt{n \ln n})}$ where n is the number of states in the PFA. We show that the ultimate period $c(L(M, \lambda))$ decreases significantly with increasing isolation and this results in a polynomial upper bound for $c(L(M, \lambda))$, if ϵ is a constant.

As a first step, Lemma 1 shows that the period d_i of an ergodic component B_i with absorption probability $r_i < 2\epsilon$, where

$$r_i := \lim_{t\to\infty} \sum_{p \in B_i} (\pi A^t)_p = \text{prob(a random walk is eventually absorbed into } B_i),$$

does not play a role for $c(L(M, \lambda))$, neither do periods of collections of ergodic components with small combined absorption probability.

Lemma 1. *Let $B_1, \ldots, B_m$ be the ergodic components of a Markov chain with periods d_i and absorption probabilities r_i, respectively. If the corresponding PFA M accepts $L := L(M, \lambda)$ with ϵ-isolated cutpoint, then for any $I \subseteq \{1, \ldots, m\}$ with $\sum_{i \in I} r_i > 1 - 2\epsilon$, $D(I) := \mathrm{lcm}\{d_i | i \in I\}$ is an ultimate period of L and thus is a multiple of $c(L)$.*

Proof (Sketch). For an ultimate period D of L the limit $A^\infty := \lim_{t\to\infty}(A^D)^t$ exists, where we require convergence in each entry of the matrix. This can be shown by bringing the matrix A into a normal form (see Gantmacher [3]), so that the stochastic submatrix A_i for each ergodic component B_i forms a block within A. If B_i has period d_i, then $\lim_{t\to\infty}(A_i^{d_i})^t$ exists. Since D is a multiple of every d_i, the limit of $(A^D)^t$ exists. As a consequence from [9] and from the existence of this limit, for every δ there must be a $\mu_\delta \in \mathbb{N}$, such that for every $j \geq \mu_\delta$, $a^j \in L \Leftrightarrow a^{j+D} \in L$ and

$$\sum_{q \in Q} |(\pi A^j)_q - (\pi A^{(j \bmod D)} A^\infty)_q| < \delta.$$

Let $I \subseteq \{1, \ldots, m\}$ be a set of indices with $\sum_{i \in I} r_i > 1 - 2\epsilon$. Assume that $D(I)$ is not an ultimate period of L. Then there is some $j > \mu_\delta$ with $a^j \in L$ and $a^{j+D(I)} \notin L$. So $\pi A^j \eta \geq \lambda + \epsilon$ and $\pi A^{j+D(I)} \eta \leq \lambda - \epsilon$, and thus $\pi \left(A^j - A^{j+D(I)}\right) \eta \geq 2\epsilon$. Let $(x)^+ = x$ if $x > 0$, and let $(x)^+ = 0$ otherwise. Remember, that $\eta \in \{0, 1\}^n$. Then we have with $Q_I := \bigcup_{i \in I} B_i \cup \{q | q \text{ transient}\}$

$$\begin{aligned} 2\epsilon &\leq \sum_{q \in Q} (\pi(A^j - A^{j+D(I)}))_q \\ &\leq \sum_{q \in Q_I} (\pi(A^j - A^{j+D(I)}))_q^+ + \sum_{q \notin Q_I} (\pi(A^j - A^{j+D(I)}))_q^+. \end{aligned} \tag{1}$$

The proof of the existence of A^∞ also shows that if we restrict the matrix A to all the states in Q_I and call the resulting substochastic matrix A_I, then the limit $\lim_{t\to\infty}(A_I^{D(I)})^t$ exists as well. And so, for $\delta = 2\epsilon - \sum_{i \notin I} r_i$ and for any $j \geq \mu_\delta$, we get

$$\sum_{q \in Q_I} (\pi(A^j - A^{j+D(I)}))_q^+ < \delta. \tag{2}$$

But on the other hand, for any $j \geq 0$

$$\sum_{q \notin Q_I} (\pi(A^j - A^{j+D(I)}))_q^+ \leq \sum_{q \notin Q_I} (\pi A^j)_q \leq \sum_{i \notin I} r_i = 2\epsilon - \delta. \tag{3}$$

The second inequality follows, since the absorption probability is the limit of a monotonically increasing sequence. So we have reached a contradiction, since the sum of (3) and (2) does not satisfy (1). $\square$

We can now exclude some prime powers as potential divisors of $c(L(M, \lambda))$.

Definition 2. *Let M be a PFA with ergodic periods d_i and absorption probabilities r_i. We call a prime power $q = p^s$ ϵ-essential (for M), if*

$$\sum_{i:\ q \text{ divides } d_i} r_i \geq 2\epsilon \qquad \text{and} \qquad \sum_{i:\ q \cdot p \text{ divides } d_i} r_i < 2\epsilon.$$

Lemma 2. *If λ is ϵ-isolated for a PFA M, then*

$$D = \prod_{q \text{ is } \epsilon-\text{essential}} q.$$

is an ultimate period of $L = L(M, \lambda)$. Hence D is a multiple of $c(L)$.

Proof. Assume that $c(L)$ is a multiple of a prime power p^k which does not divide any ϵ-essential prime power. Let $J = \{i | p^k \text{ divides } d_i\}$, and let $I = \{1, \ldots, m\} \setminus J$ be the complement of J. Then p^k does not divide any d_i with $i \in I$ and thus p^k does not divide $D(I) = \text{lcm}\{d_i | i \in I\}$.Since p^k does not divide any ϵ-essential prime power, we have that $\sum_{i \in J} r_i < 2\epsilon$, and so $\sum_{i \in I} r_i > 1 - 2\epsilon$. According to Lemma 1, $D(I)$ is a multiple of $c(L)$. But on the other hand $D(I)$ is not a multiple of p^k. This is a contradiction, since p^k was assumed to divide $c(L)$. $\square$

Now we show the tight upper bound for the minimal ultimate period of a language accepted by an ϵ-isolated PFA.

Theorem 1. a) *For any unary PFA M with n states and ϵ-isolated cutpoint λ*

$$c(L(M, \lambda)) \leq n^{\frac{1}{2\epsilon}}.$$

b) *For any $0 \leq \alpha < 1$ and any $\epsilon = \frac{1}{2m}$ with $m \in \mathbb{N}$, there is a PFA M with n states and ϵ-isolated cutpoint λ, such that $c(L(M, \lambda)) > n^{\frac{\alpha}{2\epsilon}}$.*

Proof. **a)** Let M have m ergodic components with periods $d_1, \ldots, d_m$. Set $D := \prod_{q \text{ is } \epsilon-\text{essential}} q$ and remember, that $\sum_{i:\ q \text{ divides } d_i} r_i \geq 2\epsilon$ for any ϵ-essential q, then

$$\begin{aligned} D^{2\epsilon} = \prod_{q \text{ is } \epsilon-\text{essential}} q^{2\epsilon} &\leq \prod_{q \text{ is } \epsilon-\text{essential}} q^{\sum_{i:\ q \text{ divides } d_i} r_i} \\ &= \prod_{i=1}^{m} \prod_{\substack{q \text{ is } \epsilon-\text{essential}, \\ q \text{ divides } d_i}} q^{r_i} \leq \prod_{i=1}^{m} d_i^{r_i}. \end{aligned}$$

Now, since $\sum_{i=1}^{m} r_i = 1$, the weighted arithmetic mean is at least as large as the geometric mean, and thus

$$\sum_{i=1}^{m} r_i d_i \geq \prod_{i=1}^{m} d_i^{r_i}.$$

Since $D \geq c(L(M, \lambda))$ with Lemma 2, we obtain

$$n \geq \sum_{i=1}^{m} d_i \geq \sum_{i=1}^{m} r_i d_i \geq \prod_{i=1}^{m} d_i^{r_i} \geq D^{2\epsilon} \geq c(L(M, \lambda))^{2\epsilon}.$$

And the claim follows.

b) Let $p_1, p_2, \ldots$ be the sequence of prime numbers. We define the languages

$$L_{k,m} = \left\{ a^j \,\middle|\, j \equiv 0 \left(\bmod \prod_{i=k}^{k+m-1} p_i \right) \right\}$$

for $k, m \geq 1$. Obviously $c(L_{k,m}) = \prod_{i=k}^{k+m-1} p_i \geq p_k^m \geq (k \ln k)^m$.

On the other hand $L_{k,m}$ can be accepted by a PFA with isolation $\epsilon = \frac{1}{2m}$ and cutpoint $\lambda = 1 - \frac{1}{2m}$ as follows. We define a "union automaton with an initial distribution" by setting up m disjoint cycles of length $p_k, p_{k+1}, \ldots, p_{k+m-1}$, respectively. The transition probability from one state to the next in a cycle is 1. There is exactly one final state in each cycle and the initial distribution places probability $\frac{1}{m}$ on each final state. For every word $a^z \in L_{k,m}$ we have $z \equiv 0 (\bmod\ p_i)$ for every $k \leq i \leq k+m-1$ and for every word $a^z \notin L_{k,m}$ there is at least one i with $z \not\equiv 0 (\bmod\ p_i)$. Thus a word is either accepted with probability 1, or it can reach acceptance probability at most $1 - \frac{1}{m}$.

Applying Fact 2, the number of states in the PFA is

$$\begin{aligned} n_{k,m} = \sum_{i=k}^{k+m-1} p_i &\leq 2 \sum_{i=k}^{k+m-1} i \ln i \leq 2 \int_k^{k+m} x \ln x \, dx \\ &= 2 \left[\frac{x^2}{2} \ln x - \frac{x^2}{4} \right]_{x=k}^{x=k+m} \\ &\leq (k^2 + 2km + m^2) \ln(k+m) - k^2 \ln k \\ &= k^2 \ln \left(1 + \frac{m}{k} \right) + (2km + m^2) \ln(k+m). \end{aligned}$$

But since $k \ln \left(1 + \frac{m}{k}\right) = \ln \left(1 + \frac{m}{k}\right)^k \leq \ln e^m = m$,

$$n_{k,m} \leq km + (2km + m^2) \ln(k+m) \leq (3km + m^2) \ln(k+m).$$

Thus for any $0 \leq \alpha < 1$, any constant $m = \frac{1}{2\epsilon}$ and a sufficiently large k, we have

$$c(L_{k,m}) \geq (k \ln k)^m > ((3km + m^2) \ln(k+m))^{\alpha m} \geq n_{k,m}^{\frac{\alpha}{2\epsilon}},$$

and the claim follows. □

Our result shows that for a fixed isolation ϵ, the ultimate period of the language accepted by the PFA M with n states is only polynomial in n.

4 Approximating the Size of a Minimal NFA

Stockmeyer and Meyer [11] show, that the universe problem $L(N) \neq \Sigma^*$ is NP-complete for regular expressions and NFA's N, even if we consider only unary languages. Since our argument is based on their construction, we show the proof.

Fact 3. *[11] For a unary NFA N, it is NP-hard to decide, if $L(N) \neq \{a\}^*$.*

Proof. We reduce 3SAT to the universe problem for unary NFA's. Let Φ be a 3CNF-formula over n variables with m clauses. Let $p_1, \ldots, p_n$ be the first n primes and set $D := \prod_{i=1}^n p_i$. According to the chinese remainder theorem, the function $\mu : \mathbb{N}_0 \to \mathbb{N}_0^n$ with $\mu(x) = (x \bmod p_1, \ldots, x \bmod p_n)$ is injective, if we restrict the domain to $\{0, \ldots, D-1\}$. We call x a code (for an assignment), if $\mu(x) \in \{0,1\}^n$.

We construct a union automaton N_Φ that accepts $\{a\}^*$ iff Φ is not satisfiable. We first make sure, that $L_{0,\Phi} = \{a^k | k \text{ is not a code}\}$ is accepted. Therefore, for every prime p_i ($p_i > 2$) we construct a cycle that accepts the words a^j with $j \not\equiv 0 (\bmod\ p_i) \wedge j \not\equiv 1 (\bmod\ p_i)$. So there are 2 non-final states and $(p_i - 2)$ final states in the cycle. For every clause C of Φ with variables $x_{i_1}, x_{i_2}, x_{i_3}$ we construct a cycle C^* of length $p_{i_1} p_{i_2} p_{i_3}$. C^* will accept

$$\{a^k | \text{ the assignment } k \bmod p_{i_j} \text{ for } x_{i_j}\ (j = 1,2,3) \text{ does not satisfy } C\}.$$

Since the falsifying assignment is unique for the three variables in question, exactly one state is accepting in C^*.

The construction can be done in time polynomial in the length of Φ. If there is a word $a^j \notin L(N_\Phi)$, then j is a code for a satisfying assignment. On the other hand every satisfying assignment has a code j and a^j is not accepted by N_Φ. □

We set $L_\Phi = L(N_\Phi)$ for the automaton N_Φ constructed above. Observe that L_Φ is a union of cyclic languages and hence itself cyclic. Obviously if $\Phi \notin 3SAT$, then the minimal NFA for L_Φ has size 1. We will show, that for $\Phi \in 3SAT$ every NFA accepting L_Φ must have at least $\sum_{i=2}^n p_i$ states, which implies Theorem 2.

Theorem 2. *Given an NFA N with n states, it is impossible to efficiently approximate* $\text{nsize}(L(N))$ *within a factor of $\frac{\sqrt{n}}{\ln n}$ unless $P = NP$.*

We first determine a lower bound for the period of L_Φ.

Lemma 3. *For any given 3CNF-formula $\Phi \in 3SAT$ the minimal period of L_Φ is either $D := \prod_{i=2}^n p_i$ or $2D$.*

Proof. L_Φ is $2D$-cyclic, since $2D$ is the least common multiple of the cycle lengths of N_Φ. Assume that neither D nor $2D$ is the minimal period of L_Φ. Then there is $i \geq 2$, such that $d = \frac{D}{p_i}$ is a period of L_Φ. We know that $a^{qp_i+2} \in L_{0,\Phi}$ for every $q \in \mathbb{N}$, because $qp_i + 2$ does not represent a code. Since $L_{0,\Phi} \subseteq L_\Phi$ and we assume that L_Φ is d-cyclic, a^{qp_i+2+rd} belongs to L_Φ for every $r \in \mathbb{N}$ as well.

On the other hand, since $L_\Phi \neq \{a\}^*$, there is an $a^l \notin L_\Phi$, and so $a^{l+td} \notin L_\Phi$ for every $t \in \mathbb{N}$. It is a contradiction, if we find $q, r, t \in \mathbb{N}_0$, so that $qp_i + 2 + rd = l + td$, since the corresponding word has to be in L_Φ because of the left-hand side of the equation and cannot be in L_Φ because of the right-hand side.

$$\begin{aligned}
\exists q, r, t : qp_i + 2 + rd = l + td &\Leftrightarrow \exists q, r, t : qp_i = l - 2 + (t - r)d \\
&\Leftrightarrow \quad \exists q : qp_i \equiv l - 2 \qquad \pmod d \\
&\Leftrightarrow \quad \exists q : \quad q \equiv (l - 2)p_i^{-1} \qquad \pmod d
\end{aligned}$$

The multiplicative inverse of p_i modulo d exists, since $\gcd(p_i, d) = 1$, and we have obtained the desired contradiction. □

We will need a linear relation between the number of clauses and variables in the CNF-formula.

Fact 4. *Let $E3SAT - E5$ be the satisfiability problem for formulae with exactly 3 literals in every clause and every variable appearing in exactly 5 distinct clauses, then $E3SAT - E5$ is NP-complete.*

The following Lemma determines a lower bound for the size of an NFA equivalent to N_Φ, if Φ is satisfiable.

Lemma 4. *Let $\Phi \in E3SAT - E5$ and assume that Φ consists of m clauses. Then* $\text{nsize}(L(N_\Phi)) \geq cm^2 \ln m$ *for some constant c.*

Proof. We know from Lemma 3, that $L(N_\Phi)$ is either minimally D-cyclic or $2D$-cyclic with $D = \prod_{i=2}^n p_i$ where n is the number of variables in Φ. Applying Fact 1 the size of a minimal NFA accepting L_Φ is at least $\sum_{i=2}^n p_i$. We observe that

$$\sum_{i=2}^{n} p_i \geq \sum_{i=1}^{n} i \ln i \geq \int_1^n x \ln x \, dx \geq \frac{n^2}{4} \ln n$$

We have $5n = 3m$ and thus $\text{nsize}(L_\Phi) \geq cm^2 \ln m$ for some constant c. □

Finally we determine an upper bound for the size of the NFA N_Φ.

Lemma 5. *Let Φ be a 3CNF formula with m clauses and exactly 5 appearances of every variable. Then the NFA N_Φ has size at most $O\left(m^4 (\ln m)^3\right)$ and at least $\Omega\left(m^2 \ln m\right)$.*

Proof. The number of states in a cycle for a clause is a product of three primes. So there are at most $m \cdot p_n^3 = O(m\,(m \ln m)^3)$ states in all of these cycles. The cycles recognizing $L_{0,\Phi}$ have $\sum_{i=2}^{n} p_i = \Theta(n^2 \ln n)$ states, where n is the number of variables of Φ. Since $n = \Theta(m)$ the claim follows. □

Proof (of Theorem 2). Assume that the polynomial time deterministic algorithm A approximates $\text{nsize}(L(N))$ within the factor $\frac{\sqrt{s}}{\ln s}$ for an NFA N with s states. We show that the satisfiablity problem can be decided in polynomial time.

Let Φ be the given input for the $E3SAT-E5$ problem, where we assume that Φ has n variables and m clauses. We construct the NFA N_Φ as in fact 3. If Φ is not satisfiable, then $\text{nsize}(L_\Phi) = 1$, and according to Lemma 5 the algorithm A claims that an equivalent NFA with at most

$$\frac{\sqrt{s}}{\ln s} = \frac{\sqrt{(O(m^4(\ln m)^3))}}{\ln(\Omega(m^2(\ln m)))} = o(m^2 \ln m)$$

states exists. Since $\sum_{i=2}^{n} p_i = \Theta(m^2 \ln m)$, the claimed number of states is asymptotically smaller than $\text{nsize}(L_\Psi)$ for any satisfiable formula Ψ with the same number of clauses as Φ. Hence with the help of A, we can decide if Φ is satisfiable within polynomial time. □

Remark 1. For every $0 < \epsilon \leq 1$ the same construction as in the proof of Theorem 2 can be used to show that it is not possible to approximate the size of a minimal PFA with isolation ϵ equivalent to a given n-state PFA with isolation $c \cdot n^{-\frac{1}{4}}$ within the factor $\frac{\sqrt{n}}{\ln n}$.

For a given formula Φ with m clauses we construct the PFA M_Φ with m cycles[3] and uniform initial distribution for the initial states of each cycle. We define the cutpoint as $\lambda = \frac{1}{2m}$. Hence a word is accepted by M_Φ iff it is accepted by at least one cycle. Thus the cutpoint λ is δ-isolated with $\delta = \frac{1}{2m} \geq c \cdot n^{-\frac{1}{4}}$ for some appropriate c, and M_Φ behaves like a union automaton. Since $L(M_\Phi, \lambda)$ is the same language as considered before, it is 1-cyclic if Φ is not satisfiable and has period $D = \prod_{i=2}^{m} p_i$ or $2D$ if Φ is satisfiable. Every PFA with isolated cutpoint that accepts a language with period $\prod_{i=2}^{m} p_i$ has at least $\sum_{i=2}^{m} p_i$ states [8], independent of the actual isolation.

The approximation complexity changes if a unary cyclic language is specified by a **DFA** M, although the decision problem, namely to decide whether there is a k-state NFA accepting the cyclic language $L(M)$, is not efficiently solvable unless $NP \subseteq DTIME(n^{O(\ln n)})$ [5].

Theorem 3. *Given a unary cyclic DFA accepting L with D states, an NFA for L with at most* $\text{nsize}(L) \cdot (1 + \ln D)$ *states can be computed in polynomial time. Observe that* $\text{nsize}(L) \cdot (1 + \ln D) = O\left(\text{nsize}(L)^{\frac{3}{2}} \sqrt{\ln \text{nsize}(L)}\right)$.

Proof. We reduce the optimization problem for a given cyclic DFA M to an instance of the weighted set cover problem. We can assume M to be a minimal cyclic D-state DFA with the set of states $Q = \{0, \ldots, D-1\}$, 0 as the initial state, and final states $F \subseteq Q$. Then $L(M) = \{a^{j+kD} | j \in F, k \in \mathbb{N}_0\}$.

For every d_l that divides D we construct a deterministic cycle C_l with period d_l. The union automaton consisting of these cycles will accept $L(M)$, if we choose the final states of C_l as follows: For each $a^j \in L$ with $0 \leq j < d_l$, we let C_l accept a^j, iff $a^{j+k \cdot d_l} \in L(M)$ for any $0 \leq k < \frac{D}{d_l}$. Remember, that we don't have to check for a^x with $x \geq D$, since $L(M)$ is D-cyclic and d_l divides D.

At this stage the union automaton will have a lot of unnecessary cycles. Therefore we define an instance of the set cover problem, where we introduce a set

$$T_l := \{j | 0 \leq j < D, a^j \text{ is accepted by } C_l\}$$

of weight $w_l := d_l$ for every cycle C_l. The universe is $\{j | 0 \leq j < D, a^j \in L(M)\}$. The instance can be constructed in polynomial time, since the number of divisors of D is less than D and thus the set cover problem consists of at most D sets with at most D elements.

[3] To check the validity of a code we can also use the clause cycles.

If N is a minimal NFA accepting $L(M)$, then we know from Fact 1 that N is a union automaton (with an additional initial state) that consists of cycles with periods that divide D. Every cycle C^* of N corresponds to a set T_l and the accepted words of C^* up to length $D-1$ are contained in T_l.

So a minimal union automaton with n states can be expressed by a set cover of weight n. On the other hand, every set cover can be considered to be a union automaton. Thus a minimal set cover corresponds to a minimal NFA.

The greedy algorithm for the weighted set cover problem approximates the optimal set cover within the factor $H(k) = \sum_{i=1}^{k} \frac{1}{k} \leq 1 + \ln k$, where k is the size of the largest set [7]. For an n-state NFA N Chrobak [2] bounds the size of $c(L(N))$ by the Landau function and receives $D = e^{O(\sqrt{n \ln n})}$. □

5 Conclusions and Open Problems

In Theorem 1 we have shown that PFA's with constant isolation lead to only polynomially smaller automata in comparison to cyclic unary DFA's. It is not hard to observe that PFA's with constant isolation are negatively exponentially smaller than DFA's for non-cyclic unary languages. The size relation between minimal PFA's and minimal DFA's for non-cyclic unary languages is to be further explored.

The hardness result of Theorem 2 for minimizing unary NFA's is tight within a square, since size $\frac{\sqrt{n}}{\ln n}$ is excluded for a given NFA of size n. Is Theorem 2 "essentially" optimal?

Jiang and Ravikumar [6] state the open problem of approximating a minimal NFA given a DFA. Specifically to determine the complexity of designing an NFA accepting $L(M)$ with at most $\text{nsize}(L(M))^k$ states for a given DFA M and a given k. We have answered the question for the case of unary cyclic DFA's and $k > \frac{3}{2}$ in Theorem 3.

References

1. Chrobak, M.: *Finite automata and unary languages*, Theoretical Computer Science 47, 1986, pp. 149-158.
2. Chrobak, M.: *Errata to: "Finite automata and unary languages"*, Theoretical Computer Science 302, 2003, pp. 497-498.
3. Gantmacher, F.R.: *Theory of Matrices*, Vol. II, Chelsea, New York, 1959.
4. Graham, R., Knuth, D., Patashnik, O.: *Concrete Mathematics*, Addison Wesley, Reading, Massachusetts, 1989.
5. Jiang, T., McDowell, E., Ravikumar, B.: *The structure and complexity of minimal NFA's over a unary alphabet*, Int. J. Found. of Comp. Sci., 2, 1991, pp. 163-182.
6. Jiang, T., Ravikumar, B.: *Minimal NFA problems are hard*, SIAM Journal on Computing, 22 (1), 1993, pp. 1117-1141.
7. Hochbaum, D. (editor): *Approximation algorithms for NP-hard problems*, PWS Publishing Company, Boston, 1997.
8. Mereghetti, C., Palano, B., Pighizzini, G.: *On the succinctness of deterministic, nondeterministic, probabilistic and quantum finite automata*, DCAGRS 2001.
9. Milani, M., Pighizzini, G.: *Tight bounds on the simulation of unary probabilistic automata by deterministic automata*, DCAGRS 2000.
10. Rabin, M.: *Probabilistic automata*, Information and Control, 1963, pp. 230-245.
11. Stockmeyer, L., Meyer, A.: *Word Problems Requiring Exponential Time*, Proc. of the 5th Ann. ACM Symposium on Theory of Computing, New York, 1973, pp. 1-9.