A Small NFA for $\{a^i : i \neq n\}$

Exposition by William Gasarch

1 Credit where Credit is Due

These notes are based on Jeff Shallit's slides on the Frobenius Problem [3] and some emails I had with him. None of this is my work.

2 Introduction

Consider the following language: $L_n = \{a^i : i \neq n\}$.

There is a $n+2$ state DFA for L_n (we prove this later, though it's easy). Can we do better? How about with an NFA?

We show:

1. The $n+2$ state DFA for L_n is optimal.
2. There is a $\sqrt{n} + O((\log n)^2(\log\log n))$ state NFA for L_n. a $\sqrt{n} + O((\log n)^2(\log\log n))$ state NFA for L_n for some $c < 2$.
3. Any NFA for L_n has $> \sqrt{n}$ states.

There is an appendix which has some needed lemmas from Number Theory.

3 A DFA For L_n With $n+2$ States

Theorem 3.1 *There is a DFA for L_n with $n+2$ states; however, there is no DFA for L_n with $n+1$ states.*

Proof:

The DFA for L_n has states for how many a's have been seen up to n, and then a state for 'I have seen $\geq n+1$ states'. Formally:

There are states $\{0, 1, 2, \ldots, n+1\}$. 0 is the start state. For $0 \leq i \leq n$ state i means that i a's have been seen so far. State $n+1$ means $\geq n+1$ a's have been seen. All states are accepting EXCEPT n.

For $0 \le s \le n$ $\delta(s, a) = s + 1$.

$\delta(n + 1, a) = n + 1$.

Let M be a DFA for L_n. We show that M has $\ge n + 2$ states. Let 0 be the start state.

Look at states:

$\delta(0, a^0)$

$\delta(0, a^1)$

$\delta(0, a^2)$

$\delta(0, a^3)$

$\vdots$

$\delta(0, a^{n-1})$

These are all accepting states.

I claim they are all DIFFERENT states. Assume, by way of contradiction, that $1 \le i < j \le n-1$ but

$$\delta(0, a^i) = \delta(0, a^j).$$

Then

$$\delta(0, a^i \cdot a^{n-j}) = \delta(0, a^j \cdot a^{n-j})$$

Hence

$$\delta(0, a^{n+(i-j)}) = \delta(0, a^n)$$

Since $n + (i - j) < n$, the LHS is an ACCEPT state. But the RHS is clearly a REJECT state. This is a contradiction. Hence there are n states listed above. They are all accept states. There is also at least one reject state. Hence there are at least $n + 1$ states. But there's more! Let r be the reject state. Hence $\delta(0, a^n) = r$. Look at $\delta(0, a^{n+1})$. We leave it to the reader to show that it cannot be any of the states mentioned. Hence it is another state. Total number of states: $n + 2$.

$\blacksquare$

4 An NFA for L_{107} With 23 States

Theorem 4.1 *There exists an NFA for L_{107} with 23 States.*

Proof:

What is the smallest NFA for L_{107}? Let us rephrase the question: How can a number i PROVE that its NOT 107? The next lemma will yield a small helpful NFA.

Claim 1:

1. There DO NOT exist $c, d \in \mathsf{N}$ such that $107 = 10c + 13d$.

2. $(\forall i \geq 108)(\exists c, d \in \mathsf{N})[i = 10c + 13d]$.

Proof of Claim 1:

1) We narrow down what c, d must be.

$$107 = 10c + 13d$$

take this equation mod 10.

$$7 \equiv 3d \pmod{10}$$

Multiply both sides by 7 (the inverse of 3 mod 10)

$$49 \equiv 21d \pmod{10}$$

$$9 \equiv d \pmod{10}$$

Hence $d \geq 9$ so

$$107 = 10c + 13d \geq 10c + 13 \times 9 = 13d + 117$$

This cannot happen.

2) We prove this by induction on n.

We view this as expressing n in terms of 10-cent coins and 13-cent coins.

Base Case: $108 = 13 \times 6 + 10 \times 3$

Ind. Hyp. Assume that $n \geq 108$ and that $(\exists c, d \in \mathsf{N})[n = 10c + 13d]$.

We prove that $(\exists c', d' \in \mathsf{N})[n + 1 = 10c' + 13d']$

Case 1: $c \geq 9$. Intuitively we can remove nine 10-cent coins and add in seven 13-cent coins to end up +1. Formally

$$10(c - 9) + 13(d + 7) = 10c + 13d + 1 = n + 1$$

Case 2: $d \geq 3$. Intuitively we can remove three 13-cent coins and add in four 10-cent coins to end up +1. Formally

$$10(c + 4) + 13(d - 3) = 10c + 13d + 1 = n + 1$$

Case 3: $c \leq 8$ and $d \leq 2$. Then $n = 10c + 13d \leq 80 + 26 = 106 < 108$. Hence this case cannot occur.

End of Proof of Claim 1:

We describe the NFA for L_{107}

1. There is a start state s that has many e-transitions out of it which we describe.

2. One of the e transitions is to a state q that is accepting and has a loop of size 13 (of non-accept states) but with one shortcut- there is an transition on a from the 9th element in the cycle to q. Hence one can go from q to q with either a^{10} or a^{13}. This branch will accept all strings of the form $\{a^i : i \geq 108\}$ and will NOT accept a^{107}. This part has 13 states.

3. For each $m \in \{4, 5, 7\}$ (1) let $107 \equiv a_m \pmod{m}$, (2) create DFA M_p that accepts

$$\{a^i : i \not\equiv a_m \pmod{m}\}$$

(3) put a transition between s and the start state of M_m. Clearly none of these loops accept a^{107}. This part has $4+5+7=16$ states.

Let a^i be a string that is rejected. Since a^i is not accepted by the first branch, $i \le 107$. Since they are not accepted by ANY other branch, for all $m \in \{4,5,7\}$, $i \equiv a_m \pmod{m}$. Since $4 \times 5 \times 7 = 140 > 107$, by Lemma A.1 there is at most one such i. Since $i = 107$ does work, $a^{107}i$ is the only string thats accepted.

The total number of states is $13 + 16 = 23$. ∎

5 Rel Prime Convention AND Loop Notation

In the description of the NFA in the proof of Theorem 4.1 we needed a set of rel prime numbers with product ≥ 107 and (we hope) a small sum. We will use this technique in this paper many times. Rather than repeat the details, we will just give the rel prime numbers.

We will need the Loop-and-shortcut from the proof of Theorem 4.1 later.

Def 5.1 Let $x < y \in \mathsf{N}$. Then LOOP(y,x) is the NFA that has (1) a start state s which is also the only accept state, (2) a loop of size y around s, and (3) a shortcut– a transition on a from the $x-1$'s state in the cycle to s. Note that LOOP(y,x) accepts $\{a^i : (\exists c,d \in \mathsf{N})[i = cx + dy]\}$ and has y states.

We will later need a generalization of LOOP(y,x).

Def 5.2 Let $x < y \in \mathsf{N}$ and let $m \in \mathsf{N}$. Then LOOP(y,x,m) is the NFA that has (1) has a chain of accept states from the start to a state s' which is also an accept state, (2) a loop of size y around s', and (3) a shortcut– a transition on a from the $x-1$'s state in the cycle to s. Note that LOOP(y,x) accepts $\{a^i : (\exists c,d \in \mathsf{N})[i = cx + dy + m]\}$ and has y states.

We will later need a generalization of LOOP(y,x).

6 The Inverse Frobenius Problem

What was special about 107 that made the NFA for L_{107} small? The key was (1) any $i \ge 108$ can be written as a sum of 10's and 13's, (2) 107 CANNOT be written as a sum of 10's and 13's.

Given a number, n, I want to find two numbers x_1, x_2 such that

- n cannot be written as a sum of x_1's and x_2's
- $(\forall i \ge n+1)(\exists c, d)[i = cx_1 + dx_2]$.

This is the inverse the Frobenius problem:

Frobenius problem: Given coins of denominations $(x_1, \ldots, x_m)$ find n such that n cannot be formed with those coins but all numbers $\ge n+1$ can.

The following lemma solves the $m = 2$ case of the Frobenius problem and will give us an infinite number of n such that L_n has an NFA with $\le \sqrt{n} + O((\log n)^2(\log\log n))$ states.

Lemma 6.1 *Let $x, y \in \mathsf{N}$, relatively prime. Let $n = xy - x - y$.*

1. *There DO NOT exist $c, d \in \mathsf{N}$ such that $n = xc + yd$.*
2. $(\forall i \ge n+1)(\exists c, d \in \mathsf{N})[i = xc + yd]$.
3. *Assume $y > x$.* LOOP(y, x) *(1) does not accept a^n, (2) accepts all of the strings in $\{a^i : i \ge n+1\}$, (3) we not care what else it accepts. This follows from (1) and (2).*

Proof:

1) Assume, by way of contradiction, that there exists c, d such that

$$xy - x - y = xc + yd$$

Take this mod x

$$-y \equiv yd \pmod{x}$$

Since x and y are rel prime y has an inverse so we get

$$b \equiv -1 \pmod{x}.$$

Since $b \ge 0$ we get $b \ge x - 1$.

Similarly we get $a \geq y-1$. Hence

$$xy - x - y = xc + yd \geq x(y-1) + y(x-1) = 2xy - x - y$$

$$xy \geq 2xy$$

Since $x, y \geq 1$ we get

$$1 \geq 2$$

which is a contradiction.

2) Omitted for now but the proof is on Shallit's Slides [3].

∎

We show one example.

Theorem 6.2 *There exists an NFA for L_{2069} with 75 States.*

Proof: Since 46 and 47 are relatively prime and $46 \times 47 - 46 - 47 = 2069$, by Lemma 6.1,

1. There DO NOT exist $c, d \in \mathsf{N}$ such that $2069 = 46c + 47d$.

2. $(\forall i \geq 2070)(\exists c, d \in \mathsf{N})[i = 46c + 47d]$.

We can now present the NFA for L_{2069}.

1. There is a start state s that has many e-transitions out of it which we describe.

2. One of the e transitions is to LOOP(47, 46). This branch will accept all strings of the form $\{a^i : i \geq 2070\}$ and will NOT accept a^{2069}. This part has 47 states.

3. Use the set of rel prime numbers $\{2, 3, 5, 7, 11\}$. Note that $2 \times 3 \times 5 \times 7 \times 11 = 2310 > 2069$ and $2 + 3 + 5 + 7 + 11 = 28$.

The total number of states is $47 + 28 = 75$. ∎

7 For Infinitely Many n There is a $\sqrt{n}+O((\log n)^2(\log\log n))$ State NFA for L_n

Theorem 7.1 *Let $x \in \mathsf{N}$, $x \ge 2$. Let $n = x^2 - x - 1 \in \mathsf{N}$. (Note that $x = \sqrt{n} + O(1)$.) There is a $\sqrt{n}+O((\log n)^2(\log\log n))$ state NFA for L_n.*

Proof:

We describe the NFA for L_n:

1. There is a start state s. There will be many e-transitions from it.

2. One of the e transitions is to LOOP$(x+1, x)$. This branch (1) does not accept a^n, (2) accepts $\{a^i : i \ge n+1\}$, (3) we don't care what else it accepts. The number of states is $x+1 \le \sqrt{n}+O(1)$.

3. Let ℓ be the least number such that the product of the first ℓ primes is $\ge n$. Use the set of rel prime numbers $\{p_1, \ldots, p_\ell\}$ (p_i is the ith prime). By Lemma B.1 $\sum_{i=1}^{\ell} p_i = O(\ell^2 \log \ell) = O((\log n)^2 \log\log n)$.

The total number of states is:

$$\sqrt{n}+O((\log n)^2(\log\log n))$$

∎

8 A $\sqrt{n}+O((\log n)^2(\log\log n))$ State NFA for L_n and Some Tips on Getting Less States

Is there always a small NFA for L_n? Yes. We show three ways of obtaining a small NFA for L_{1000}. After the first way we have a general theorem. We then give two smaller NFA's and some non-rigorous advice on how to get a smaller NFAs in general.

8.1 An NFA for L_{1000} With 68 States

Theorem 8.1 *There exists an NFA for L_{1000} with 68 States.*

Proof: Let $x = \lfloor\sqrt{1000}\rfloor = 32$ and $y = x + 1 = 33$. Note that $xy - x - y = 991$. By an easy variant of Lemma 6.1 (1) there does not exist c, d such that $1000 = 32c+33d+9$, (2) for all $i \geq 1001$ there does exist c, d such that $n = 32c + 33d + 9$.

Note that LOOP$(33, 32, 9)$ (1) does not accept a^{1000}, (2) accepts $\{a^i : i \geq 1001\}$ (3) we don't care what else it accepts.

We describe the NFA for L_{1000}

1. There is a start state s that will have many transitions out of it.

2. (This does not need an e-transition.) LOOP$(33, 32, 9)$ comes out of the start state. The number of states on this branch is $33 + 9 = 42$ (this includes the start state).

3. We use the set of rel prime numbers $\{3, 5, 7, 11\}$. Note that $3 \times 5 \times 7 \times 11 = 1155 > 1000$ and that $3 + 5 + 7 + 11 = 26$.

The total number of states is and has $42 + 26 = 68$ states. ∎

The proof of Theorem 8.1 generalizes.

Theorem 8.2 *Let $n \in \mathsf{N}$. There exists a $\sqrt{n} + O((\log n)^2(\log\log n))$ state NFA for L_n.*

Proof:

Let $x = \lfloor\sqrt{n}\rfloor$ and $y = \lfloor\sqrt{n}\rfloor + 1$. Note that

$$xy - x - y = (\sqrt{n})(\sqrt{n} + 1) - 2\sqrt{n} + O(1) = n - \sqrt{n} + O(1) = n - m$$

where m is within $O(1)$ of $\sqrt{n}$.

We describe the NFA for L_n.

1. There is a start state s that will have many transitions out of it.

2. (This does not need an e-transition.) From the start state have LOOP(y, x, m). This takes $m + y = \sqrt{n} + O(1)$ states.

3. This part of the NFA is identical to that in Theorem 7.1. The number of states is $O((\log n)^2 \log\log n)$.

The total number of states is $\sqrt{n} + O((\log n)^2(\log\log n))$. ∎

8.2 NFA for L_{1000} With 65 States

Theorem 8.3 *There exists an NFA for L_{1000} with 65 states.*

Proof: Let $x = 34$, $y = 39$, and $n = 39 \times 34 - 39 - 34 = 1253$. Hence LOOP$(39, 34)$ (1) does not accept a^{1253} (this does not help us), and (2) accepts $\{a^i : i \geq 1253\}$.

We need to NOT get 1000.

We show that there is NO c, d such that $34c + 39d = 1000$. Assme, by way of contradiction, that

$$1000 = 34c + 39d$$

Mod out by 34

$$14 \equiv 5d \pmod{34}$$

Multiply both sides by 7 since $5 \times 7 = 35 \equiv 1 \pmod{34}$.

$$14 \times 7 \equiv d \pmod{34}$$

$$d \equiv 14 \times 7 \equiv 98 \equiv 30 \pmod{34}$$

SO $d \equiv 30 \pmod{34}$. Hence $d \geq 30$. But then

$34c + 39d \geq 34c + 39 \times 30 = 1170 > 1000$.

Hence LOOP$(39, 34)$ does not accept 1000.

We describe the NFA for L_{1000}.

1. There is a start state s that will have many transitions out of it.

2. From the start state there is an e-transition to LOOP$(39, 34)$. This takes 39 states.

3. We use the set of rel prime numbers $\{3, 5, 7, 11\}$. Note that $3 \times 5 \times 7 \times 11 = 1155 > 1000$ and that $3 + 5 + 7 + 11 = 26$.

The total number of states is $39 + 26 = 65$. ∎

8.3 One More Potential Tip for Reducing the Number of States

In the proof of Theorem 8.1 we constructed an NFA M_2 that used the set of rel primes numbers $\{3, 5, 7, 11\}$ since $3 \times 5 \times 7 \times 11 = 1155 \geq 1000$. We noted that M_2 has $3 + 5 + 7 + 11 = 26$ states Could we have picked a set of rel primes numbers with product ≥ 1000 but sum ≤ 26? One can show NO. But for L_n there may be a clever way to pick the set which leads to some savings. We suspect the savings is not much since this is part of the log-term.

Another possible savings: We have been ignoring what the big loop part accepts that is under n. It is plausible that the big loop part ends up accepting all $i \leq n - 1$ with n having the correct equivalence classes mod some prime. This may enable you to use less primes.

8.4 Finding a Small NFA for L_n

Given n we want to find a small NFA for L_n. Here is a procedure.

1) Find $x < y$ such that $xy - x - y$ is closer to n and y is small. There are several cases.

1. $n = xy - x - y$. Build an NFA with loops of size y with a shortcut to create an x-loop. This NFA has y states.

2. $xy - x - y < n$. Use a chain of size $n - (xy - x - y)$ from the initial state to the state where you the loop of size y. This NFA has $y + (n - xy + x + y) = x + 2y + n - xy$ states.

3. $xy - x - y > n$. We also need that n cannot be written as $cx + dy$. Then can use a loop of y. This NFA has y states.

Take the smallest of these three NFA's and call it M_1. If case 1 happens that will surely be the smallest.

2) Find a set or rel primes numbers A such that $\prod_{i\in A} i \geq n$ and $\sum_{i\in A} i$ is minimized. Use this to build part of the NFA as in Theorem 7.1.

3) The final NFA is an OR of M_1 and M_2.

9 Every NFA for L_n has $\geq \sqrt{n}$ States

Chroback [2] proved the following.

Theorem 9.1 *Let L be a co-finite unary regular language. If there is an NFA for L with n states then there is an NFA for L of the following form:*

- *There is a sequence of $\leq n^2$ states from the start state to a state we will call X. Note that there is no nondeterminism involved yet.*
- *From X there are e-transitions to $X_1, \ldots, X_m$. (This is nondeterministic.)*
- *Each X_i is part of a cycle C_i. All of the C_i are disjoint.*

The following theorem is due to Jeff Shallit and was communicated to me by email.

Theorem 9.2 *Let L be a cofinite unary language where the shortest string that is not in L is of length n. Any NFA for L requires $\Omega(\sqrt{n})$ states*

Proof:

Assume there was an NFA with $< \sqrt{n}$ states for L_n. Then by Theorem 9.1 there would be an NFA for L with a path from the start state to a state X of length $< n$ and then from X a branch to many cycles. Let X_i and cycle's C_i as described in Theorem 9.1.

Run a^n through the NFA and try out all paths. For each i there will be a point in C_i that you end up at. Let n_i be the length of C_i. For every i there is a state on C_i that rejects. Hence the strings $a^{n+Kn_1n_2\cdots n_m}$ are all rejected. This is an infinite number of strings. This is a contradiction.

∎

10 Open Problems

For every n, (1) there is an NFA for L_n with $\sqrt{n}$ states (omitting some log terms), but (2) there is no NFA for L_n with $\sqrt{n}$ states. We would like to close this gap. The upper bound might be improved with some lemmas from number theory. The lower bound might be improved by a more in depth study of Theorem 9.1. And, of course, its possible either or both require new techniques.

A A Lemma from Easy Number Theory

We use the following well known lemma. We include the proof for completeness.

Lemma A.1

1. *Let m_1, m_2 be relatively prime. Let $0 \le a_1 \le m_1 - 1$ and Let $0 \le a_2 \le m_2 - 1$. Let A be the set*

$$A = \{i : i \equiv a_1 \pmod{m_1}\} \cap \{i : i \equiv a_2 \pmod{m_2}\} \cap \{i : i \le m_1 m_2\}$$

   Then $|A| \le 1$.

2. *Let $m_1, \ldots, m_\ell$ be relatively prime. Let $a_1, \ldots, a_\ell$ be such that, for all $1 \le i \le \ell$, $0 \le a_i \le m_i - 1$, and $n \equiv a_i \pmod{m_i}$. Let A be the set*

$$\left(\bigcap_{i=1}^{\ell} \{i : i \equiv a_i \pmod{m_i}\}\right) \cap \{i : i \le m_1 m_2 \cdots m_\ell\}.$$

   Then $|A| \le 1$. (This follows from part 1 and induction so we omit the proof of this part.)

Proof:

Assume $x, y \in A$ and $x < y$. Then $x \equiv y \pmod{m_1}$ and $x \equiv y \pmod{m_2}$.

Since $x - y$ is a multiple of both m_1 and m_2, and m_1, m_2 are rel prime, $x - y$ is a multiple of $m_1 m_2$. But then $y = x + k m_1 m_2 > m_1 m_2$. This is a contradiction. ▮

B A Lemma from Hard Number Theory

We use the following lemma. We do not include the proof; however, see [1] for both references and more precise estimates.

Lemma B.1 *Let $\ell \in \mathsf{N}$. Let $p_1, \ldots, p_\ell$ be the first ℓ primes. Then $\sum_{p \leq \ell} p = O(\ell^2 \log \ell)$.*

References

[1] C. Axler. On the sum of the first n primes, 2014. `https://arxiv.org/pdf/1409.1777.pdf`.

[2] M. Chrobak. Finite automata and unary languages. *TCS*, 47:149–158, 1986. `http://www.sciencedirect.com/science/article/pii/0304397586901428`.

[3] J. Shallit. The Frobenius problem and its generalization. `slides:https://cs.uwaterloo.ca/~shallit/Talks/frob14.pdf`.