Powers of Regular Languages

Szilárd Zsolt Fazekas

Research Group in Mathematical Linguistics
Universidad Rovira i Virgili
szilard.zsolt@estudiants.urv.cat

Abstract. In this paper we prove that it is decidable whether the set $\mathrm{pow}(L)$, which we get by taking all the powers of all the words in some regular language L, is regular or not. The problem was originally posed by Calbrix and Nivat in 1995. Partial solutions have been given by Cachat for unary languages and by Horváth et al. for various kinds of exponent sets for the powers and regular languages which have primitive roots satisfying certain properties. We show that the regular languages which have a regular power are the ones which are 'almost' equal to their Kleene-closure.

1 Introduction

Calbrix and Nivat defined the power $\mathrm{pow}(L)$ of a language L in a paper about prefix and period languages of rational ω-languages [3]:

$$\mathrm{pow}(L) = \{w^i \mid w \in L, i \geq 1\}.$$

It is easy to see that there are examples of regular languages L for which $\mathrm{pow}(L)$ is regular, and examples for which $\mathrm{pow}(L)$ is not regular. Take, for instance, the regular language ab^* whose power is not even context-free. Calbrix and Nivat posed the problem of characterizing those regular languages whose powers are also regular, and to decide whether a given regular language has this property. They conjectured that "rational languages such that their power is also rational are 'almost' a union of rational subsemigroups of Σ^* and the point is to give the right sense to this almost". Cachat [2] gave a partial solution to this problem showing that for a regular language L over a one-letter alphabet, it is decidable whether $\mathrm{pow}(L)$ is regular. Unfortunately Cachat's result cannot be extended to arbitrary alphabets since he translated unary languages into sets of natural numbers to reach his solution. Horváth et al. [5] provided more partial results. They looked at arbitrary alphabets and did a case analysis based on the primitive roots of the regular languages in question. However, the case when $L \cap \sqrt{L}$ is not regular and $\sqrt{\mathrm{pow}(L) \setminus L}$ is finite was left open, and thus a complete solution to the original problem was not given. They also considered other exponent sets instead of the whole set of natural numbers and an algorithm based on [2] to decide whether the power of a regular language with finite primitive root is regular or not. We will use the results of [5] (and [2] resp.) as part of the

V. Diekert and D. Nowotka (Eds.): DLT 2009, LNCS 5583, pp. 221–227, 2009.

decision procedure. There exist several other papers that study regular languages containing powers of their words or consisting solely of non-primitive words, see [4,6]. Recently Anderson et al. [1] presented a characterization of regular languages consisting only of powers.
In this paper we characterize the regular languages whose power is also regular. First we present a short overview of the notions and results needed to proceed with the paper, and then we go on to solve the decidability problem posed by Calbrix and Nivat by reducing it to decidable subproblems.

2 Preliminaries

In this section we briefly recall some definitions and known results needed throughout the rest of the paper. Let Σ be a fixed finite nonempty alphabet. By Σ^* we mean the free monoid generated by Σ, that is the set of all words over Σ. The empty word we denote by λ, and $\Sigma^+ = \Sigma^* \setminus \{\lambda\}$. A language over Σ is a subset L of Σ^*. For a word $p \in \Sigma^*$, $|p|$ denotes the length of p, and for a set M, $|M|$ denotes the cardinality of M. For a natural number k, p^k denotes the concatenation of k copies of the word p, and $p^0 = \lambda$. As usual, p^* denotes the set $\{p^k : k \geq 0\}$, and p^*q the set $\{p^kq : k \geq 0\}$. For two words $u, v \in \Sigma^*$ and a language $L \subseteq \Sigma^*$, by saying that $u \equiv v(P_L)$ we mean the following

$$xuy \in L \text{ if and only if } xvy \in L \text{ for all } x, y \in \Sigma^*.$$

For a word $w \in \Sigma^*$ the congruence class $[w]_{P_L}$ consists of all words congruent with w according to P_L, that is $[w]_{P_L} = \{v \in \Sigma^* \mid w \equiv v(P_L)\}$. Since P_L is a congruence relation, $\Sigma^*/P_L = \{[w] \mid w \in \Sigma^*\}$ is a monoid, which is called the *syntactic monoid* of L and denoted by $\mathrm{Synt}(L)$.
A *non-deterministic finite automaton* (NFA) is a quintuple $\mathcal{A} = \{\Sigma, Q, I, F, \sigma\}$ with the usual conventions, i.e. Σ is the input alphabet, Q is the set of states, I is the set of initial states, F is the set of final states and $\sigma : Q \times \Sigma \to 2^Q$ is the transition function. We will use σ as an extended transition function taking words as second arguments instead of only letters, interpreted as follows $\sigma(p, ab) = \{q \in \sigma(q_1, b) \mid q_1 \in \sigma(p, a)\}$. A finite automaton is *deterministic* (DFA) if I is a singleton and $p, q \in \sigma(r, w)$ implies $p = q$.
When talking about regular languages we will often mean languages accepted by some finite automaton or languages which are unions of some of the equivalence classes of a congruence relation of finite index. As it is well known from Rabin and Scott [8] these language classes are in fact the same. We note here that the size of the syntactic monoid for a regular language L is at most $2^{|Q|^2}$, if Q is the set of states of some NFA accepting L.

Definition 1. *A word p is primitive if there is no word $q \neq p$ and no positive integer n such that $p = q^n$. We denote the language of all primitive words over a given alphabet Σ by Q.*

Definition 2. *The (primitive) root of a word $p \in \Sigma^+$ is the unique primitive word q such that $p = q^n$ for some $n \geq 1$. $\sqrt{p}$ denotes the root of p. For a language L, $\sqrt{L} = \{\sqrt{p} : p \in L \wedge p \neq \lambda\}$ is the root of L.*

The next result is known as the theorem of Fine and Wilf. Intuitively it tells us how far two periodic events (strings) have to match in order to guarantee a common period, that is to guarantee that the two sequences are ultimately the same.

Theorem 1. *Let $x, y \in \Sigma^*$, $n = |x|$, $m = |y|$, $d = \gcd(n, m)$. If two powers x^p and y^q of x and y have a common left factor of length at least equal to $n+m-d$, then x and y are powers of the same word.*

The following is a well known theorem by Lyndon and Schützenberger.

Theorem 2. *[7] If $u^m v^n = w^k$ for non-empty words u, v, w and natural numbers $m, n, k \geq 2$ then $\sqrt{u} = \sqrt{v} = \sqrt{w}$.*

This is a basic property of words that will prove useful to us later on through a theorem by Shyr and Yu, which we present here with a short proof to better illustrate our argument.

Corollary 1. *[9] Let u, v be primitive words such that $u \neq v$. Then there is at most one non-primitive word in u^+v^+.*

Proof. Let $w = u^m v^n$ be non-primitive. Then either $m = 1$ or $n = 1$ by Theorem 2. So, by symmetry let $uv^n = w^i$ for some primitive word w and $i \geq 2$. We may choose n to be minimal with that property. It is enough to show that all uv^{n+k} are primitive. By contradiction, suppose that uv^{n+k} is not primitive for some $k \geq 1$, that is there exists some $z \in Q$ and $j \geq 2$ such that $z^j = uv^{n+k}$. It follows that $w^i v^k = z^j$.

First consider the case $k \geq 2$. Since $i, j, k \geq 2$, we can apply the Lyndon-Schützenberger theorem and get that $\sqrt{w} = \sqrt{v} = \sqrt{z}$, but then $u = v$, a contradiction.

Now let us see the case $k = 1$. Non-primitivity is invariant to cyclic shifts, so w^i and z^j being non-primitive gives us that $v^n u$ and $v^n uv$ are non-primitive too. Hence there are words $w_1, z_1 \in Q$ such that $w_1^i = v^n u$ and $z_1^j = v^n uv$, moreover $|w_1| = |w|$ and $|z_1| = |z|$. From here $z_1^j = w_1^i v$. As v is a prefix of w_1^i, we have that z_1^j has both periods $|z_1|$ and $|w_1|$. Now we can apply the theorem of Fine and Wilf and get that $z_1 = w_1 = v$. This means $w = z = v$ and then $u = v$, a contradiction again. □

Corollary 2. *For all words $x, y, z \in \Sigma^*$ with $y \neq \lambda$ with $|\sqrt{xyz}| \neq |\sqrt{y}|$, there is at most one non-primitive word in the language xy^+z.*

Proof. Suppose there exist numbers $i, j \geq 1$ with $i < j$ such that both xy^iz and xy^jz are non-primitive. Non-primitivity is invariant to cyclic shifts, so zxy^i and zxy^j are non-primitive too.

If zx is non-primitive then we can apply the Lyndon-Schützenberger theorem on zxy^j and get that $\sqrt{zx} = \sqrt{y}$. This would mean $\sqrt{zxy} = \sqrt{y}$ and from here $|\sqrt{xyz}| = |\sqrt{y}|$, a contradiction.

If zx is primitive then from Theorem 1 we have that only one of the words zxy^i and zxy^j is non-primitive, contradicting our original assumption. □

3 The Power of a Regular Language

There are easy examples for non-trivial regular languages that do not have a regular power. Besides the one mentioned in the introduction one could take $aaa(aa)^*$ whose power $\{a^k : k \text{ is not a power of } 2\}$ is not even context-free (in particular, powers of regular languages are not semi-linear, in general).

In fact, as it turns out, it is quite difficult to come up with examples of regular languages L with regular power other than the ones for which $L = L^*$ or $L = L^* \setminus K$, where either K is finite or $K = \bigcup_{w \in S} w^*$ for some finite set of words S. This seems to justify the conjecture formulated by Calbrix and Nivat cited before. Hence, rather than trying to solve the case left open in [5] one might try a new approach.

Indeed, as we will shortly see, we can give an equivalent criterion for a regular language to have a regular power, i.e., we can now give sense to that 'almost'.

Theorem 3. *Let L be a regular language. Then $pow(L)$ is regular if and only if $\mathrm{pow}(L) \setminus L$ is a regular language such that its primitive root is a finite language.*

Proof. The class of regular languages is closed under union and taking the difference of two sets, therefore if $\mathrm{pow}(L) \setminus L$ is a regular language then so is $(\mathrm{pow}(L) \setminus L) \cup L = pow(L)$.

Now let us look at the "only if" part. If $\mathrm{pow}(L)$ is regular then so is $L_{\mathrm{diff}} = \mathrm{pow}(L) \setminus L$. Note that L_{diff} consists solely of non-primitive words. Let n be the number of states of the minimal deterministic automaton accepting L_{diff}. Now suppose that $\sqrt{L_{\mathrm{diff}}}$ is infinite. In this case there must be some $w \in L_{\mathrm{diff}}$ such that $|\sqrt{w}| > n$. On the other hand the pumping property of regular languages tells us that $w = xyz$ for some $xz, y \notin \{\lambda\}$ with $|y| \leq n$ such that $xy^iz \in L_{\mathrm{diff}}$ for all $i \geq 0$, so xy^iz is non-primitive for all i. Corollary 2 says that in this case $|\sqrt{xyz}| = |\sqrt{y}| \leq |y| \leq n$, contradicting the assumption $|\sqrt{w}| > n$. □

Lemma 1. *Let L be a regular language given by an NFA having n states. If $pow(L)$ is regular, then we have*

$$pow(L) \subseteq L \cup \{\sqrt{u}^i \mid u \in L \wedge |u| \leq \max(n^2, m) \wedge i \geq 1\},$$

where m is the size of $\mathrm{Synt}(L)$.

Proof. We have seen in Theorem 3 that $pow(L)$ being regular means that it has to be a subset of $L \cup \bigcup_{u \in U} u^+$ for some finite set U of words. We need to prove that for every $w \in L_{\mathrm{diff}}$ there is a $u \in L$ such that $w \in \sqrt{u}^+$ and $|u| \leq \max(n^2, m)$.

Take the shortest $u \in L$ such that w is a power of u. If $|u| > \max(n^2, m)$ then according to the pumping property of regular languages u can be written as xyz for some $y \neq \lambda \neq xz$ such that $xy^jz \in L$ for all $j \geq 0$. Here we can distinguish two cases.

1. If $|y|$ is a multiple of $|\sqrt{u}|$ then $|\sqrt{u}| \leq n$. As $u \in \sqrt{u}^+ \cap L$ we can apply the pumping argument on powers of $\sqrt{u}$ as if it was a unary language. If k is the smallest number for which $\sqrt{u}^k \in L$ then $k \leq n$, or else there would be some numbers p, q with $p < q < k$ such that from the initial state we reach the same state by reading $\sqrt{u}^p$ or $\sqrt{u}^q$, and we could cut out $\sqrt{u}^{q-p}$ from the word. From here we get that there is a word $\sqrt{u}^k \leq n^2$ having the same root as w.
2. We are left with the case when in any decomposition $u = xyz$, $|y|$ is not a multiple of $|\sqrt{xyz}|$ and $|u| > m$. Then we find a decomposition $u = xyz$ with $0 < |y| \leq m$ and $[xy] = [x]$ in $\mathrm{Synt}(L)$. This way we know that $xy^j z \in L$, for all $j \geq 1$. As a consequence of Corollary 2 we also know that at most one of these $xy^j z$ can be a non-primitive word. At the same time L_{diff} has finite root, hence for all but finitely many values of j, $(xy^j z)^+ \subseteq L$, so we find some j such that $xy^j z \in L$ and $xy^j z$ is primitive and at the same time $(xy^j z)^+ \subseteq L$. Due to $[xy] = [x]$ in $\mathrm{Synt}(L)$ we can conclude $(xyz)^+ \subseteq L$. However, we supposed that $w \in L_{\mathrm{diff}}$ is some power of xyz, a contradiction.

So for every $w \in L_{\mathrm{diff}}$ there is some $u \in L$, with $|u| \leq \max(n^2, m)$ such that $\sqrt{w} = \sqrt{u}$ and this concludes the proof. □

To make it easy to see why the latter half of the previous theorem can be checked effectively, we should replace $\max(n^2, m)$ with a bound depending only on the number of states n of the automaton accepting L.

Remark 1. Let L be a regular language given by an NFA having n states. If $pow(L)$ is regular, then we have

$$pow(L) \subseteq L \cup \{u^i \mid u \in L \wedge |u| \leq 2^{n^2} \wedge i \geq 1\},$$

where m is the size of $\mathrm{Synt}(L)$.

Proof. This is clear because $n^2 < 2^{n^2}$ and the syntactic monoid is a divisor of the monoid of Boolean $n \times n$ matrices, so $\mathrm{Synt}(L)$ has size at most 2^{n^2}. □

Let us recall the following result from the paper by Calbrix and Nivat about languages which are equal to their power.

Lemma 2. *[3] Let L be a regular language of Σ^*. Then $\mathrm{pow}(L) = L$ if and only if there are regular languages $(L_i)_{1 \leq i \leq n}$ such that $L = \bigcup_{i=1}^{n} L_i^+$.*

The statement above is useful for testing if a language is equal to its power or not, we only need to specify the languages L_i for an effective construction. Using the syntactic monoid of L gives us the tool we need. We can translate $pow(L) = L$ into the following statement involving the congruence classes of P_L:

$$\bigcup_{u \in L} [u]^+ \subseteq L = \bigcup_{u \in L} [u] \subseteq \bigcup [u]^+.$$

Given an automaton accepting L we can effectively construct its syntactic monoid and from here we can effectively define the set of words in the class

$[u]$ for all $u \in L$. In the case of a regular language P_L induces a finite number of classes, so we can decide whether the equality holds or not. Hence, we can state the following.

Proposition 1. *For a regular language L it is decidable whether $pow(L) = L$ holds or not.*

Now we are ready to proceed with the algorithm. Theorem 3 reduces the original problem to an equivalent one of deciding whether the language, in some sense, lacks only a "few" words to be equal to its power. Lemma 1 provides the means to find those "few" missing words and after adding them to our starting language Proposition 1 will tell us whether the result is a power or not, that is whether the power of the original language is regular or not.

Theorem 4. *For a regular language L it is decidable whether* $\mathrm{pow}(L)$ *is regular.*

Proof. We propose the following algorithm:

1. Input: an NFA $\mathcal{A} = \{\Sigma, Q, I, F, \sigma\}$.
2. Output: "YES", if $\mathrm{pow}(L(\mathcal{A}))$ is regular, and "NO" otherwise.
3. $U = \emptyset$
4. FOR all words $w \in L(\mathcal{A})$ shorter than $2^{|Q|^2}$:
5. —IF $w^* \setminus L(\mathcal{A}) \neq \emptyset$ THEN:
6. ——IF $\mathrm{pow}((\sqrt{w})^* \cap L(\mathcal{A}))$ is regular THEN add w to U
7. ——ELSE output "NO"
8. compute the syntactic monoid for $L' = L(\mathcal{A}) \setminus \bigcup_{u \in U} (\sqrt{u})^*$
9. IF $L' = \mathrm{pow}(L')$ then output "YES"
10. ELSE output "NO"

The enumeration of words in $L(\mathcal{A})$ shorter than $2^{|Q|^2}$ can be done in finite time due to the length limit. The condition in line 5 can be checked effectively too. First we have to perform the difference of two regular languages, then check whether the result is empty or not. As it is stated in [5], the condition in line 6 can be verified using Cachat's algorithm ([2]), because $(\sqrt{w})^* \cap L(\mathcal{A})$ is isomorphic to a unary language, which can be computed effectively. In step 8 we have to compute the syntactic monoid for a regular language, which is the difference of a regular language and the finite union of some regular languages, all effectively presented. If a regular language L is equal to $M \cup N$ for some regular languages M and N, such that $\sqrt{M} \cap \sqrt{N} = \emptyset$, then from the closure properties of the regular class we get that $\mathrm{pow}(L)$ is regular if and only if both $\mathrm{pow}(M)$ and $\mathrm{pow}(N)$ are regular. Moreover, L and M being powers implies N being a power as well. Therefore, in step 9 we only need to check whether a regular language is equal to its power or not; by Proposition 1 this is decidable too. Hence, the algorithm terminates after finitely many steps; however, the complexity is at least exponential due to both Cachat's algorithm and the exponential length bound on the words we need to check in step 4. □

4 Conclusion

We managed to characterize regular languages that have regular power following the conjecture of Calbrix and Nivat formulated in [3] and we gave an effective albeit inefficient procedure to decide this property. Although the decision procedure is not a direct extension of previous results ([2,5]), Cachat's algorithm is needed in an essential step, which identifies those "few words" in $\mathrm{pow}(L)$ missing from L.

Acknowledgements

The author would like to thank the referees and, in particular, Volker Diekert for their suggestions on correcting the proofs and improving the overall presentation of the paper.

References

1. Anderson, T., Rampersad, N., Santean, N., Shallit, J.: Finite Automata, Palindromes, Powers, and Patterns. In: Martín-Vide, C., Otto, F., Fernau, H. (eds.) LATA 2008. LNCS, vol. 5196, pp. 52–63. Springer, Heidelberg (2008)
2. Cachat, T.: The Power of One-Letter Rational Languages. In: Kuich, W., Rozenberg, G., Salomaa, A. (eds.) DLT 2001. LNCS, vol. 2295, pp. 145–154. Springer, Heidelberg (2002)
3. Calbrix, H., Nivat, M.: Prefix and Period Languages of Rational omega-Languages. In: Dassow, J., Rozenberg, G., Salomaa, A. (eds.) Developments in Language Theory 1995, pp. 341–349. World Scientific, Singapore (1996)
4. Dömösi, P., Horváth, G., Ito, M.: A small hierarchy of languages consisting of non-primitive words. Publ. Math. Debrecen 64(3-4), 261–267 (2004)
5. Horváth, S., Leupold, P., Lischke, G.: Roots and Powers of Regular Languages. In: Ito, M., Toyama, M. (eds.) DLT 2002. LNCS, vol. 2450, pp. 220–230. Springer, Heidelberg (2003)
6. Lischke, G.: The root of a language and its complexity. In: Kuich, W., Rozenberg, G., Salomaa, A. (eds.) DLT 2001. LNCS, vol. 2295, pp. 272–280. Springer, Heidelberg (2002)
7. Lyndon, R.C., Schützenberger, M.P.: On the equation $a^M = b^N c^P$ in a free group. Michigan Math. Journ. 9, 289–298 (1962)
8. Rabin, M.O., Scott, D.: Finite automata and their decision problems. IBM J. Res. Develop. 3, 114–125 (1959)
9. Shyr, H.J., Yu, S.S.: Non-primitive words in the language p^+q^+. Soochow J. Math. 20, 535–546 (1994)