Lie Complexity of words

Jason Bell (Joint with Jeffrey Shallit)

July 26, 2021

Motivating Problem

Let Σ be a finite alphabet and let w be a right-infinite word over the alphabet Σ. We're interested in the set of factors $\mathrm{Fac}(\mathrm{w})$ of w. This is the set of words of the form $w[i, \ldots, i+n]$ with $i, n \geq 0$.

In particular, we're interested in the following question:

Can we determine the set of primitive factors v of w with the property that v^n is a factor of w for every n?

In the study of this problem, we'll introduce some new invariants for a word, which have not really been looked at and could be interesting objects of study.

Equivalence of right-infinite words

Given two right-infinite words w, w' over Σ, we declare that w is *equivalent* to w' if

$$\mathrm{Fac}(\mathrm{w}) = \mathrm{Fac}(\mathrm{w}').$$

Then this induces an equivalence relation $\sim$ on the set of right-infinite words over Σ, and we let $[\mathrm{w}]$ denote the equivalence class of w. Although equivalent words need not be the same, many natural combinatorial properties are preserved under this equivalence.

We recall that a right-infinite word w is *periodic* if there is some finite word v such that

$$\mathrm{w} = vvv \cdots = v^{\omega}.$$

The word w is *recurrent* if every factor v occurs infinitely often in w, and w is *uniformly recurrent* if every factor v occurs infinitely often *uniformly*; i.e., there is some $N = N(v)$ such that every factor of w of length N contains a copy of v.

If w and w' are equivalent right-infinite words over Σ, then the following hold:

1. w is recurrent if and only if w' is recurrent;
2. w is uniformly recurrent if and only if w' is uniformly recurrent;
3. w is periodic if and only if w' is periodic.

We put a partial order $\prec$ on the equivalence classes of right-infinite words over Σ by declaring that $[w] \preceq [w']$ if $\mathrm{Fac}(w) \supseteq \mathrm{Fac}(w')$. Then $[w] \preceq [w']$ and $[w'] \preceq [w]$ if and only if $[w] = [w']$.

We now introduce three sets one can associate with a right-infinite word.

Definition

Let w be a right-infinite word over Σ. We define the following sets:

1. the *total spectrum* of w is the set

$$\mathrm{Tot}(\mathsf{w}) = \{[\mathsf{v}] \colon [\mathsf{w}] \preceq [\mathsf{v}]\};$$

2. the *recurrent spectrum* of w is the set

$$\mathrm{Rec}(\mathsf{w}) = \{[\mathsf{v}] \colon \mathsf{v} \text{ recurrent}, [\mathsf{w}] \preceq [\mathsf{v}]\};$$

3. the *uniformly recurrent spectrum* of w is the set

$$\mathrm{URec}(\mathsf{w}) = \{[\mathsf{v}] \colon \mathsf{v} \text{ unif.recurrent}, [\mathsf{w}] \preceq [\mathsf{v}]\};$$

4. the *periodic spectrum* of w is the set

$$\mathrm{Per}(\mathsf{w}) = \{[\mathsf{v}] \colon \mathsf{v} \text{ periodic}, [\mathsf{w}] \preceq [\mathsf{v}]\};$$

Then we have the containments

$$\mathrm{Tot}(\mathsf{w}) \supseteq \mathrm{Rec}(\mathsf{w}) \supseteq \mathrm{URec}(\mathsf{w}) \supseteq \mathrm{Per}(\mathsf{w}). \tag{1}$$

Furthermore, $\mathrm{Tot}(\mathsf{w})$, $\mathrm{Rec}(\mathsf{w})$, $\mathrm{URec}(\mathsf{w})$, and $\mathrm{Per}(\mathsf{w})$ are all posets under the ordering $\prec$, and all elements of $\mathrm{URec}(\mathsf{w})$ and $\mathrm{Per}(\mathsf{w})$ are maximal.

Examples

Let $\mathsf{w} = xy^\omega$. Then $\mathrm{Per}(\mathsf{w}) = \mathrm{URec}(\mathsf{w}) = \mathrm{Rec}(\mathsf{w}) = [y^\omega]$, $\mathrm{Tot}(\mathsf{w}) = \{[y^\omega], [\mathsf{w}]\}$.

Let w' be the Thue-Morse word $xyyxyxxy\cdots$. Then the recurrent, uniformly recurrent, and total spectra are all just the singleton $\{[\mathsf{w}']\}$; the periodic spectrum is empty!

While $\mathrm{Per}(w)$ can be empty, $\mathrm{URec}(w)$ (and hence $\mathrm{Rec}(w)$) is never empty.

Why? A theorem of Furstenberg says that if w is a right-infinite word, then there is always a uniformly right-infinite word w' such that $\mathrm{Fac}(w') \subseteq \mathrm{Fac}(w)$. Then $[w'] \in \mathrm{URec}(w)$.

Zariski Topology

We recall that a topology on a set X is a subset $\mathcal{C}$ of the power set of X, call the *closed subsets of* X, with the properties: $\emptyset, X \in \mathcal{C}$; $\mathcal{C}$ is closed under finite unions and under arbitrary intersections.

Let w be a right-infinite word over a finite alphabet Σ. For each subset S of $\mathrm{Fac}(\mathrm{w})$ that is closed under the process of taking factors, we define a set

$$\mathcal{C}(S) = \{[u] \in \mathrm{Rec}(\mathrm{w})\colon \mathrm{Fac}(u) \subseteq S\}.$$

Then in addition to being posets, $\mathrm{Rec}(\mathrm{w})$, $\mathrm{URec}(\mathrm{w})$, and $\mathrm{Per}(\mathrm{w})$ also have a structure as a topological space in which the closed sets are the sets of the form $\mathcal{C}(S)$.

Then an interesting question becomes: *What are these topological spaces/posets for a given right-infinite word* w*?* In particular, these are interesting invariants for automatic words w, in which case one gets finiteness results. In fact, one has more general results of this flavour for words of *linear factor complexity*.

We recall that if w is a right-infinite word, then we have a *factor complexity function*, $p_{\mathrm{w}} : \mathbb{N} \to \mathbb{N}$, whose value at n is the number of distinct factors of length n of w.

(A useful fact is that when we look at a set X of the form $\mathrm{Per}(w)$, $\mathrm{Rec}(w)$, $\mathrm{URec}(w)$ as topological spaces then the map $p : X \to \mathbb{N}^{\mathbb{N}}$ given by $p([v]) = p_v : \mathbb{N} \to \mathbb{N}$ is continuous when we give $\mathbb{N}^{\mathbb{N}}$ the *Alexandrov product* topology!)

Theorem

Let w *be a right-infinite word and suppose that* $p_{\mathrm{w}}(n) = \mathrm{O}(n)$ *(e.g.,* w *is k-automatic). Then* $\mathrm{Per}(\mathrm{w})$, $\mathrm{URec}(\mathrm{w})$, *and* $\mathrm{Rec}(\mathrm{w})$ *are finite sets.*

In fact, if $p_{\mathrm{w}}(n) \leq Cn$ for n sufficiently large with $C > 1$, then we have

$$\#\mathrm{Per}(\mathrm{w}) \leq \limsup \kappa := p_{\mathrm{w}}(n) - p_{\mathrm{w}}(n-1) + 1 < \infty,$$

$$\#\mathrm{URec}(\mathrm{w}) \leq \kappa + C + 1,$$

and for w recurrent,

$$\#\mathrm{Rec}(\mathrm{w}) \leq \kappa + (1 + C)^{C+1}.$$

In particular, if w is a *k*-automatic sequence, the numbers

$$\#\mathrm{Per}(\mathsf{w}),$$

$$\#\mathrm{URec}(\mathsf{w}),$$

and

$$\#\mathrm{Rec}(\mathsf{w})$$

are quantities that we can associate with w that measure its complexity in some sense. It would be interesting to find an algorithm to determine the poset

$$\mathrm{Rec}(\mathsf{w}).$$

Example

Let $\mathrm{w} = 01101000100000001\cdots$; i.e., $w[n] = 1$ if n is a power of 2 and $w[n] = 0$ otherwise. Then $\mathrm{Tot}(\mathrm{w})$ is infinite and

$$\mathrm{Rec}(\mathrm{w}) = \mathrm{URec}(\mathrm{w}) = \mathrm{Per}(\mathrm{w}) = \{[0^\omega]\}.$$

For the purposes of this talk, we're just going to look at the inequality

$$\#\mathrm{Per}(\mathsf{w}) \leq \kappa,$$

for words w with $p_{\mathsf{w}}(n) = O(n)$. In this setting it has a more concrete reformulation.

It says that the number of primitive factors v of w such that v^n is a factor of w for every $n \geq 1$ is a finite set, and if we work up to cyclic equivalence then there are at most κ primitive such primitive factors v (up to cyclic equivalence).

Lie Complexity

To obtain this first bound, we'll introduce a new complexity function $L_{\mathrm{w}} : \mathbb{N} \to \mathbb{N}$, which we call the *Lie complexity function*.

We again let w be a right-infinite word over Σ. We say that two words v, v' over Σ^* are *cyclically equivalent*, which we write $v \sim_C v'$ if v and v' are cyclic permutations of one another.

E.g.,

$$abcdab \sim_C cdabab \not\sim_C bacdab.$$

We define the *Lie complexity* of w to be the function

$$L_w(n) := \#\{v \in \mathrm{Fac}(w) \colon |v| = n, v' \in \mathrm{Fac}(w)\ \forall v' \sim_C v\}.$$

That is, $L_w(n)$ counts the number of cyclic equivalence classes of length n with the property that every word in the equivalence class is a factor of w.

Note this is different from the cyclic complexity defined by Cassaigne, Fici, Sciortino, Zamboni. In particular, $L_w(n) \leq c_w(n)$, where $c_w(n)$ is the cyclic complexity of w, as defined by CFSZ.

Example

If $\mathrm{w} = 01101001\cdots$ is the Thue-Morse word, then $L_{\mathrm{w}}(n) \in \{0, 1, 2, 3\}$ for all n and:

- $L_{\mathrm{w}}(n) = 3$ if and only if $n = 2$;
- $L_{\mathrm{w}}(n) = 2$ if and only if $n = 1, 4$ or a number of the form $3 \cdot 2^k$ with $k \geq 0$;
- $L_{\mathrm{w}}(n) = 1$ if $n = 0$ or $n = 2^k$ for some $k \geq 3$;
- otherwise, $L_{\mathrm{w}}(n) = 0$.

Relevant remark: $L_{\mathrm{w}}(n)$ is a 2-automatic!

Why is Lie complexity relevant?

Notice that if w is a right-infinite word and if v is a factor such that v^2 is also a factor of w then every cyclic permutation of v is a factor of w and so $L_{\mathrm{w}}(|v|) \geq 1$, and so the Lie complexity function at n is bounding the number of square factors of w of the form vv with $|v| = n$ up to cyclic equivalence (on v).

Our main result about the Lie complexity function is the following.

Theorem

(B-Shallit) Let w *be a right-infinite word. Then*

$$L_{\mathrm{w}}(n) \le p_{\mathrm{w}}(n) - p_{\mathrm{w}}(n-1) + 1.$$

Cassaigne showed that if $p_{\mathrm{w}}(n) = \mathrm{O}(n)$, then $p_{\mathrm{w}}(n) - p_{\mathrm{w}}(n-1) = \mathrm{O}(1)$, so in this case $L_{\mathrm{w}}(n)$ is uniformly bounded.

Corollaries

Theorem

(B-Shallit) Let w *be a right-infinite word with* $p_{\mathrm{w}}(n) = \mathrm{O}(n)$*. Then up to cyclic equivalence there are at most*

$$\kappa := \limsup(p_{\mathrm{w}}(n) - p_{\mathrm{w}}(n-1) + 1)$$

primitive words v *with the property that* v^n *is a factor of* w *for every* $n \geq 1$*.*

We note that Klouda and Starosta proved a similar result for pure morphic words.

Theorem

(B-Shallit) If w *is* k*-automatic then* $L_{\mathrm{w}}(n)$ *is a* k*-automatic sequence.*

I will focus on these two corollaries today, since the proof of the inequality

$$L_{\mathrm{w}}(n) \leq p_{\mathrm{w}}(n) - p_{\mathrm{w}}(n-1) + 1$$

is a bit technical.

Technical note: the proof of this inequality is what motivates the term *Lie complexity*. The proof goes by associating a graded Lie algebra

$$\mathcal{L} = \bigoplus_{n=0}^{\infty} \mathcal{L}_n$$

and an element $u \in \mathcal{L}_1$ such that

$$L_{\mathrm{w}}(n) \leq \dim(\mathrm{coker}(\mathrm{ad}_u|_{\mathcal{L}_{n-1}})).$$

First Corollary

Let's now explain why the inequality

$$L_{\mathrm{w}}(n) \leq p_{\mathrm{w}}(n) - p_{\mathrm{w}}(n-1) + 1$$

gives that there are at most

$$\kappa := \limsup(p_{\mathrm{w}}(n) - p_{\mathrm{w}}(n-1) + 1)$$

primitive words v (up to cyclic equivalence) with the property that v^n is a factor of w for every $n \geq 1$.

Recall this is the same as showing $\mathrm{Per}(\mathrm{w}) \leq \kappa$.

Proof via contradiction....

Suppose that there exist distinct equivalence classes

$$[u_1^\omega], \ldots, [u_s^\omega]$$

in $\mathrm{Per}(\mathrm{w})$ with $s > \kappa$. Pick D such that u_i^D is not a factor of u_j^ω whenever $i \neq j$. Let

$$b := D|u_1| \cdot |u_2| \cdots |u_s|.$$

Then by construction, for each $n \geq 1$, the words $u_1^{nb/|u_1|}, \ldots, u_s^{nb/|u_s|}$ are cyclically inequivalent words of length nb with the property that every cyclic permutation occurs as a factor of w. Hence $L_{\mathrm{w}}(bn) \geq s > \kappa$ for every $n \geq 1$, which is a contradiction.

Second Corollary: Automaticity

Now we'll look at the proof that if w is k-automatic then $L_{\mathrm{w}}(n)$ is also k-automatic.

Remark. Since automatic words have linearly bounded factor complexity, we know $L_{\mathrm{w}}(n)$ takes only finitely many values, which plays a key role in the proof.

To do this, we use a result of Charlier, Rampersad, and Shallit:

Theorem

Let s *be a* k*-automatic sequence.*

(a) *There is an algorithm that, given a well-formed first-order logical formula* φ *in* $FO(\mathbb{N}, +, 0, 1, n \to \mathsf{s}[n])$ *having no free variables, decides if* φ *is true or false.*

(b) *Furthermore, if* φ *has free variables, then the algorithm constructs an automaton recognizing the representation of the values of those variables for which* φ *evaluates to true.*

In case one doesn't know what this means, we look at an example. Suppose that s is the Thue-Morse sequence (which is 2-automatic). It's well-known that s does not contain factors of the form v^3. Notice this can be phrased as a first-order statement as follows:

$$\exists n, i \geq 0, \text{such that } s[n, \ldots, n+i] = s[n+i+1, \ldots, n+2i]$$

$$\text{and } s[n, \ldots, n+i] = s[n+2i+1, \ldots, n+3i].$$

Now technically, we're not allowed to use multiplication, but things such as $2i+1$ and $3i$ can be written as $i+i+1$ and $i+i+i$.

Then the theorem says we can decide whether this is true or false (well, we know it's false, but we can decide algorithmically).

The theorem says even more: it says that for a given automatic sequence, the set of (n, i) such that

$$s[n, \ldots, n+i] = s[n+i+1, \ldots, n+2i]$$

$$\text{and } s[n, \ldots, n+i] = s[n+2i+1, \ldots, n+3i]$$

are both true is automatic and one can explicitly build an automaton that accepts these pairs.

Now if w is an automatic word, the formula $L_{\mathrm{w}}(n) \geq i$ is expressible within this first-order framework as follows:

- We can find all factors of w with the property that all cyclic permutations are again factors of w by the first
- We can also decide when two length n factors are cyclicly equivalent.
- Finally, for a nonnegative integer i, we can determine the set of n for which there are at least i cyclicly inequivalent factors of w such that every cyclic permutation of the words is also a factor of w.

This allows us to show that the set of n for which $L_{\mathrm{w}}(n) \geq i$ is k-automatic for every $i \geq 0$ when w is k-automatic. Since $L_{\mathrm{w}}(n)$ is uniformly bounded, we get the result.

Example

Let us look at an example in a different base, and where there are factors of unbounded exponent. Let $\mathrm{c} = 101000101\cdots$ be the *Cantor sequence*, which is the fixed point of the morphism $1 \to 101$ and $0 \to 000$. Then

$$L_{\mathrm{c}}(n) = \begin{cases} 3, & \text{if } n = 4; \\ 2, & \text{if } n = 0, 1, 3 \text{ or } 2 \cdot 3^k \text{ for } k \geq 0; \\ 1, & \text{otherwise.} \end{cases}$$

Example

Let f be the Fibonacci word, the fixed point of the morphism sending 0 to 01 and 1 to 0. Define the Fibonacci numbers by $F_0 = 0$, $F_1 = 1$, and $F_n = F_{n-1} + F_{n-2}$ for $n \geq 2$. Then

$$L_{\mathrm{f}}(n) = \begin{cases} 1, & \text{if } n = 0 \text{ or } n = F_k \text{ for } k \geq 4 \text{ or} \\ & \quad n = F_k + F_{k-3} \text{ for } k \geq 4 \text{ ;} \\ 2, & \text{if } n = 1, 2; \\ 0, & \text{otherwise.} \end{cases}$$

Example

Let TR be the Tribonacci word, the fixed point of the morphism sending 0 to 01, 1 to 02, and 2 to 0. Define the Tribonacci numbers by $T_0 = 0$, $T_1 = 1$, $T_2 = 1$, and $T_n = T_{n-1} + T_{n-2} + T_{n-3}$ for $n \geq 3$. Then

$$L_{\mathrm{TR}}(n) = \begin{cases} 1, & \text{if } n = 0 \text{ or } n = T_k \text{ for } k \geq 5 \text{ or} \\ & \ n = T_k + T_{k-1} \text{ for } k \geq 3 \text{ or} \\ & n = T_k + T_{k-4} \text{ for } k \geq 5\ ; \\ 2, & \text{if } n = 4; \\ 3, & \text{if } n = 1, 2; \\ 0, & \text{otherwise.} \end{cases}$$

Finally, we give an example where $L_{\mathrm{w}}(n) = 0$ for $n \geq 2$:

Example

Let $\Sigma = \{x_1, \ldots, x_6, y_1, \ldots, y_6\}$ and let $\Phi : \Sigma^* \to \Sigma^*$ be the morphism given by

$$\begin{array}{lll} x_1 \mapsto x_1x_2y_1y_2 & x_2 \mapsto x_1x_3y_1y_3 & x_3 \mapsto x_1x_4y_1y_4 \\ x_4 \mapsto x_1x_5y_1y_5 & x_5 \mapsto x_1x_6y_1y_6 & x_6 \mapsto x_2x_3y_2y_3 \\ y_1 \mapsto x_2x_4y_2y_5 & y_2 \mapsto x_2x_5y_3y_4 & y_3 \mapsto x_2x_6y_2y_6 \\ y_4 \mapsto x_3x_4y_3y_5 & y_5 \mapsto x_3x_5y_3y_6 & y_6 \mapsto x_3x_6y_4y_5. \end{array}$$

and let $\mathrm{w} = \Phi^{\omega}(x_1)$. Then w is 2-automatic and work of B-Madill shows that $L_w(n) = 0$ for $n \geq 2$.

The preceding example gives a word in which every factor of length at least two has some cyclic conjugate that is not a factor. Badkobeh and Ochem give an example of such a word over a 5-letter alphabet. In general, the property that $L_{\mathrm{w}}(n) = 0$ for $n \geq i$ has been studied over various alphabets.

One loose end!

We've shown that $L_{\mathrm{w}}(n)$ is uniformly bounded whenever $L_{\mathrm{w}}(n)$ has linearly bounded factor complexity. But it still leaves the question of what happens when the factor complexity grows slightly faster than linearly (e.g., $\leq n \log n$): is $L_{\mathrm{w}}(n)$ still uniformly bounded?

In fact we can show this fails.

Theorem

(B-Shallit) Let $f : \mathbb{N} \to \mathbb{N}$ *be a function that tends to infinity as* $n \to \infty$ *and let* Σ *be a finite alphabet. Then there is a right-infinite recurrent word* w *over* Σ *such that* $p_{\mathrm{w}}(n) \leq nf(n)$ *for* n *sufficiently large and such that* w *has infinitely many distinct primitive factors* y *with the property that* y^n *is a factor of* w *for every* n. *In particular* $\limsup L_{\mathrm{w}}(n) = \infty$.

One last question!

One thing that is intriguing is that for automatic words w, the set $\mathrm{Rec}(w)$ is a finite partially ordered set.

Question: Can one characterize the finite posets which occur as some $\mathrm{Rec}(w)$ with w a k-automatic word?

Thanks!