Lyndon words and de Bruijn sequences

Narad Rampersad

June 1, 2014

Let Σ be an ordered alphabet. We denote the *lexicographic order* on Σ^* by $u < v$ if either

- u is a proper prefix of v; or,
- u has a prefix xa and v has a prefix xb, where a and b are letters and $a < b$.

A primitive word is a *Lyndon word* if it is lexicographically smaller than all of its proper, non-empty suffixes.

Example 1. *Over the alphabet* $\{0, 1\}$ *with the usual order* $0 < 1$, *the word* 0001 *is a Lyndon word, since it is lexicographically smaller than its suffixes* 001, 01, *and* 1.

The word 0010 *is not a Lyndon word, since the suffix* 0 *is less than* 0010.

The word 0101 *is not a Lyndon word because it is not primitive.*

Lemma 2. *If u and v are Lyndon words and $u < v$, then uv is a Lyndon word.*

Proof. Let s be a proper suffix of uv. If $s = xv$, where x is a non-empty suffix of u, then $u < x$, and since u is not a prefix of x, we have $uv < xv = s$, as required.

Now suppose that s is a suffix of v. Then $v \leq s$. If u is not a prefix of v, then from $u < v$ we have $uv < v \leq s$, as required. If u is a prefix of v, then write $v = uv'$. We then have $v < v'$ and therefore $uv < uv' = v \leq s$, as required. □

Theorem 3. *Let w be a non-empty word. Then w has a unique factorization*

$$w = w_1 w_2 \cdots w_t,$$

where each w_i is a Lyndon word and

$$w_1 \geq w_2 \geq \cdots \geq w_t.$$

This factorization is called the standard factorization *of w. Furthermore, the factor w_1 is the longest Lyndon prefix of w.*

Proof. To show that such a factorization exists, consider any factorization of $w = u_1 u_2 \cdots u_r$ as a product of Lyndon words. This can always be done since individual letters are Lyndon words. If there is a pair of consecutive factors $u_i u_{i+1}$ such that $u_i < u_{i+1}$, then these two factors can be merged into a single Lyndon factor by Lemma 2. Repeatedly merge all such pairs of factors until the resulting factorization is a product of non-increasing factors.

The uniqueness of the factorization will follow from the claim that w_1 is the longest Lyndon prefix of w. To show this, suppose to the contrary that $p = w_1 \cdots w'_j$ is a Lyndon prefix of w, where $j > 1$ and w'_j is a prefix of w_j. Then since w_j is Lyndon, we have $w'_j \leq w_j \leq \cdots \leq w_1 < p$. Thus $p > w'_j$, contradicting the hypothesis that p is Lyndon.

The uniqueness of the factorization now follows by repeating the previous argument to $w_2 \cdots w_t$, $w_3 \cdots w_t$, etc. $\square$

We now give a characterization of the non-empty prefixes of Lyndon words.

Lemma 4. *Let w be a prefix of a Lyndon word and let a and b be letters with $a < b$. If $w = ua$, then ub is a Lyndon word.*

Proof. Write $ua = xy$ and $ub = xy'$, where x is non-empty. We must show that $ub < y'$. Let z be the prefix of u of length $|y|$. Since ua is a prefix of a Lyndon word, let uat be such a Lyndon word. Then $z < uat < yt$. Since $|z| = |y|$, this implies $z < y$. However, $y < y'$, so $z < y'$ and consequently $ub < y'$. $\square$

Lemma 5. *Let w be a non-empty prefix of a Lyndon word. Then $w = w_1^k w'_1$ for some $k \geq 1$, where w_1 is the longest Lyndon prefix of w and w'_1 is a prefix of w_1.*

Proof. Suppose that w does not have the claimed form. Then there exist letters $a \neq b$ and a prefix pa of w_1 such that $w = w_1^k pbv$ for some $k \geq 1$ and some word v. Let wx be an extension of w to a Lyndon word; i.e.,

$$wx = w_1^k pbvx$$

is a Lyndon word. If $a < b$, then Lemma 4 implies that $w_1^k pb$ is a Lyndon prefix of w, contradicting the maximality of w_1. If $a > b$, then $pbvx < w_1^k pbvx = wx$, since wx begins with pa and $pbvx$ begins with pb. This contradicts the fact that wx is a Lyndon word. $\square$

Lemma 6. *Let v be a Lyndon word. Then for any $k \geq 1$ and any prefix v' of v, the word $w = v^k v'$ is either a prefix of a Lyndon word or equals c^j, where c is the largest letter of the alphabet and $j \geq 2$ is some integer.*

Proof. Suppose first that v is a single letter. If $v < c$, then $v^k c$ is a Lyndon word and so w is a prefix of a Lyndon word. If $v = c$, then either $w = c$, which is a Lyndon word, or $w = c^j$ for some $j \geq 2$.

So now suppose that $|v| > 1$. Let a be the last letter of v. Since v is a Lyndon word we have $v < a$. The single letter a is also a Lyndon word, so by Lemma 2 the word va is a Lyndon word. Indeed, by repeated application of Lemma 2, we see that $v^{k+1}a$ is a Lyndon word. Since w is a prefix of $v^{k+1}a$, the result follows. $\square$

Lemmas 5 and 6 establish a correspondence between Lyndon words and prefixes of Lyndon words. This can be used as the basis for an algorithm generating all Lyndon words of length at most n in increasing lexicographic order. The following algorithm considers all words $a_1 \cdots a_n$ that are prefixes of Lyndon words. By Lemma 5 there is a corresponding index $j \leq n$ such that $a_1 \cdots a_j$ is a Lyndon word. The algorithm outputs each $a_1 \cdots a_j$ discovered in this manner. In this way it outputs all Lyndon words of length at most n in lexicographic order.

```
 1: procedure GenerateLyndonWords(n, k)            ▷ Alphabet: {0, ..., k − 1}
 2:     [a_1, ..., a_n] ← [0, ..., 0]
 3:     j ← 1
 4:     a_0 ← −1
 5:     Output a_1 ··· a_j.
 6:     j ← n
 7:     while a_j = k − 1 do
 8:         j ← j − 1
 9:     end while
10:     if j = 0 then
11:         Terminate.
12:     else
13:         a_j ← a_j + 1                          ▷ a_1 ··· a_j is now Lyndon by Lemma 4
14:     end if
15:     for k ← j + 1, ..., n do
16:         a_k ← a_{k−j}          ▷ a_1 ··· a_n is now the periodic extension of a_1 ··· a_j
17:     end for
18:     Goto 5.
19: end procedure
```

The remarkable thing about this algorithm is that if it is modified to only output Lyndon words $a_1 \cdots a_j$ when $j|n$, then the resulting concatenation of Lyndon words is a circular de Bruijn word.

Theorem 7. *Let $w_1 < w_2 < \cdots < w_t$ be the Lyndon words whose lengths divide n that are output by* GenerateLyndonWords(n, k). *Then $w_1 w_2 \cdots w_t$ is a circular de Bruijn word of order n.*

Proof. We first observe that the word $w_1 w_2 \cdots w_t$ has length k^n. To see this, let $\lambda_k(d)$ (resp. $\pi_k(d)$) denote the number of Lyndon (resp. primitive) words of length d over a k-letter alphabet. Then we have

$$|w_1 w_2 \cdots w_t| = \sum_{d|n} d\lambda_k(d) = \sum_{d|n} \pi_k(d) = k^n.$$

It suffices then to show that every word of length n appears in $w_1 w_2 \cdots w_t w_1 w_2$.

For ease of exposition, let us now suppose that $k = 10$, so that we are working over the alphabet $0, \ldots, 9$. Of course, the argument remains valid for arbitrary k.

Every word of length n has the form $(uv)^d$, where $d|n$ and vu is a Lyndon word (i.e., one of the w_i). We consider two cases:

Case 1: u contains at least one letter different from 9. Let $w_i = vu$. The next Lyndon word of length at most n after w_i in lexicographic order is generated by first deleting any trailing 9's of $(vu)^d$ and then increasing the last letter of the resulting word. It follows that w_{i+1} begins with $(vu)^{d-1}v$ and so $w_i w_{i+1}$ contains $(uv)^d$, as required.

Case 2: $u = 9^j$ for some $j \geq 1$. We consider two subcases:

Subcase 2A: $d = 1$. If v is all 0's, i.e. $uv = 9^j 0^{n-j}$, then uv appears in $w_{t-1}w_t w_1 w_2 = 89^n 0^n 1$. We suppose then that v is not all 0's. Let i be the least index such that w_i begins with v. The word v is the prefix of a Lyndon word and therefore is the periodic extension of its longest Lyndon prefix v'.

We first establish that $w_{i-1} \leq v' \leq w_i$. Suppose to the contrary that $v' < w_{i-1}$. Let $\hat{v}$ be the next Lyndon word of length at most n after v' in the lexicographic order. The word $\hat{v}$ is obtained by deleting any trailing 9's from $(v')^{n/|v'|}$ and then incrementing the last letter. Note that v is a prefix of $\hat{v}$: if this were not the case, then, since v is a fractional power of v', we would have $\hat{v} > v$, which is not possible, since v is a prefix of w_i and $\hat{v} < w_i$. Now, since $v' < w_{i-1}$, we have $\hat{v} \leq w_{i-1}$. Therefore there exists $w_{i'} \leq w_{i-1}$ such that $\hat{v}$ is a prefix of $w_{i'}$. Then v is a prefix of $w_{i'}$, contradicting the minimality of i. So $w_{i-1} \leq v' \leq w_i$, as claimed.

The word of length n considered by the algorithm just prior to generating v' is $(v' - 1)9^{n-|v'|}$, where $v' - 1$ denotes the word obtained by decreasing the last letter of v' by 1. The word $(v' - 1)9^{n-|v'|}$ is the periodic extension of its longest Lyndon prefix v'', which is the Lyndon word immediately preceding v' in the lexicographic order. Note that v'' ends with at least $n - |v'| \geq n - |v| = |u|$ 9's. It follows that either v'' equals $(v' - 1)9^{n-|v'|}$ or $|v''| < |v'|$. Now if $v' > w_{i-1}$, then we repeat this argument, if necessary, with v'', and then with its Lyndon predecessor v''', and so on, until we reach w_{i-1}. Each Lyndon word in this decreasing sequence ends with at least $|u|$ 9's, and consequently, w_{i-1} ends with u. Therefore $w_{i-1}w_i$ contains uv, as required.

Now if $v' = w_{i-1}$, the argument of the previous paragraph shows that w_{i-2} ends with u. Now $w_{i-1}w_i$ begins with $v'v$; however, the word v is itself a fractional power of v', so $v'v$ also begins with v. Consequently, $w_{i-2}w_{i-1}w_i$ contains uv, as required.

Subcase 2B: $d > 1$. Let $w_i = vu$. The next Lyndon word of length at most n after w_i in lexicographic order is generated by first deleting any trailing 9's of $(vu)^d$ and then increasing the last letter of the resulting word. It follows that w_{i+1} begins with $(vu)^{d-1}$. Now the word of length n considered by the algorithm just prior to outputting w_i is the word $w' = (vu - 1)9^{(d-1)|vu|}$, where $vu - 1$ denotes the word obtained by decreasing the last letter of vu by 1. We claim that w' is itself a Lyndon word, and hence equal to w_{i-1}. The word w' is the periodic extension of its longest Lyndon prefix. This Lyndon prefix must end with $9^{(d-1)|vu|}$, but now this forces it to equal w'. Thus $w' = w_{i-1}$ ends with u and so $w_{i-1}w_i w_{i+1}$ contains $(uv)^d$, as required. □

Example 8. *Concatenating the Lyndon words of lengths* 1, 2, 3, *and* 6 *gives the following circular binary de Bruijn word of order* 6*:*

```
0|000001|000011|000101|000111|001|001011|001101|001111|01|010111|011|011111|1
```

Notes

Theorem 3 is due to Chen, Fox, and Lyndon (1958). The algorithm GenerateLyndonWords is due to Duval (1983). Theorem 7 is due to Fredriksen and Maiorana (1978). Our treatment is based largely on Knuth TACP 4A.