Shuffled Lyndon Words

L.J. Cummings
University of Waterloo Waterloo, Ontario
Canada N2L 3G1

M.E. Mays
West Virginia University
Morgantown, West Virginia
U.S.A. 26506

Abstract. The set of Lyndon words of length N is obtained by choosing those strings of length n over a finite alphabet which are lexicographically least in the aperiodic equivalence classes determined by cyclic permutation. We prove that interleaving two Lyndon words of length n produces a Lyndon word of length $2n$. For the binary alphabet $\{0,1\}$ we represent the set of Lyndon words of length n as vertices of the n-cube. It is known that the set of Lyndon words of length n form a connected subset of the n-cube. A path of vertices in the n-cube is a list of strings of length n in which adjacent strings differ in a single bit. Using paths of Lyndon words in the n-cube we construct longer paths of Lyndon words in higher order cubes by shuffling and concatenation.

Introduction

Let A be a finite alphabet equipped with a total order $<$. A **string**, w, of length n over A is a mapping $w\colon \{1,\ldots,n\} \to A$. Usually, a string w is denoted by $w = a_1 \cdots a_n$ an with $a_1,\ldots,a_n \in A$ or by $w = a[1]\cdots a[n]$. We let A^n denote the set of all strings of length n and set

$$A^* = \bigcup_{0\le n} A^n$$

where $A^0 = \{\lambda\}$, if λ is the empty string, and $A^1 = A$. We set $A^+ = A^* - \lambda$. If $w \in A^+$ then we denote the concatenation of w with itself k times by w^k.

Consider the action of the full cycle permutation $\pi = (12\cdots n)$ on A^n given by

$$w^k = a_{\pi(1)}\cdots a_{\pi(n)}$$

for each $w = a_1\cdots a_n \in A^n$. The relation $u \sim v$ if $v = u^{\pi^m}$ for some non-negative integer m, is an equivalence relation on A^n. The resulting equivalence classes are often called circular strings. We are especially interested in aperiodic circular strings. A string w is **aperiodic** if $w \neq u^m$ for any nonempty substring u and positive integer m. A circular string is aperiodic if every word in the equivalence class determined by the string is aperiodic. A string for which the corresponding circular string is aperiodic is called **primitive**. A standard argument by Möbius

inversion shows that the number of primitive strings of length n constructed from an alphabet A of cardinality σ is given by

$$\sum_{d|n} \mu(n/d)\sigma^d$$

and hence the number of aperiodic circular strings is

$$s(n,\sigma) = \frac{1}{n}\sum_{d|n} \mu(n/d)\sigma^d. \tag{1}$$

From (1) we see that the number of aperiodic circular strings over an alphabet of cardinality σ is asymptotic to σ^n/n as either σ or n approach infinity.

Lexicographical Ordering and Lyndon Words

The set A^+ can be ordered **lexicographically** as follows:

Definition 1. *Strings u and v in A^+ satisfy $u < v$ if*

$$v = uv', \text{ for some } v' \in A^+ \tag{2}$$

or

$$u = ras, v = rbt \text{ and } a < b \text{ for some } a, b \in A, r, s, t \in A^*. \tag{3}$$

Lemma 1 contains two well-known properties of the lexicographical ordering.

Lemma 1. *If $u, v, w,$ and z are strings in A^+ then*

$$u < v \text{ if and only if } wu < wv, \text{ for all } w \in A^* \tag{4}$$

and

$$\text{if } u < v \text{ and } v \neq uv' \text{ for any } v' \in A^* \text{ then } uw < vz. \tag{5}$$

Another well-known consequence of the lexicographical ordering of strings is the following:

Proposition 1. *If $w, w', w'' \in A^+$ and $w' < w$ but $w \neq w'u$ for any $u \in A^+$ then*

$$w'w < w < ww''.$$

We will need the following special case of proposition 1:

If p and q are positive integers and w is a non-constant string in A^+ then

$$a^p w < w < wb^q \tag{6}$$

where a is any element of A less than the first entry of w with respect to the given order $<$ on A and b is an arbitrary element of A. ∎

Definition 2. *We denote by L_n the set of primitive strings of length n in A^n which are lexicographically least in the equivalence classes determined by cyclic permutation. The strings in L_n are called Lyndon words. By convention we take $L_1 = A$. We further define $L = \bigcup_{0 \leq n} L_n$.*

Lyndon words were first introduced by R.C. Lyndon to define bases for the quotients of the lower central series of a free group, or equivalently, a basis of the free Lie algebra [1], [5]. For other applications see [6] and [7].

We list the Lyndon words in L through $n = 6$:

0, 1, 01, 001, 011, 0001, 0011, 0111, 00001, 00011, 00101,
00111, 01011, 01111, 000001, 000011, 000101, 000111, 001011,
001101, 001111, 010111, 011111.

Proofs of the following two lemmas about Lyndon words are contained in [4].

Lemma 2. *A string $w \in L$ if and only if $w = uv$ for some $u, v \in L$ such that $u < v$ in the lexicographical order determined by a total ordering of the underlying alphabet A.*

If $w = uv \in A^*$ and u and v are non-empty then v is called a **proper right factor** of w.

Lemma 3. *A string $w \in L$ if and only if w is strictly less than each of its proper right factor in the lexicographical order of A^+ determined by a total ordering of the underlying alphabet A.*

Lemma 2 yields a recursive algorithm to generate all the Lyndon words in L_n from shorter words, but the same string may be generated in different ways, requiring "look-ups" to avoid repetitions. The first example of this sort occurs in L_4:

$$0011 = (001)1 = 0(011).$$

Viewed as an algorithm, Lemma 3 would require testing each of σ^n strings of length n to determine L_n. Repetitions may be avoided by putting further restrictions on the words to be concatenated. For instance, distinct Lyndon words of length $2n$ could be generated by requiring that u and v both have length n, but not all Lyndon words of length $2n$ would be generated in this way.

J.-P. Duval [3] has given an excellent algorithm which generates all Lyndon words of length n in lexicographic order. The algorithm requires only the previous word of the list to be held in memory.

Shuffled Strings

Lemma 2 may also be viewed as providing a "weak" binary operation on L: given words $u, v \in L$, $u \neq v$, then exactly one of uv or $vu \in L$. We study a different operation which "shuffles" the words of L.

Definition 3. *If* $u = a_1 \cdots a_n$ *and* $v = b_1 \cdots b_n$ *in* A^n *then the string*

$$(u \mid v) = a_1 b_1 a_2 b_2 \cdots a_n b_n$$

in A^{2n} *is called "*u *shuffle* v*".*

We note that Definition 3 is a special case of the shuffle product as defined in **[4; p. 108]**.

The interleaving described in the definition above can be viewed as an addition if the alphabet symbols appearing in u and v are interpreted as digits of numbers which are first appropriately padded with zeros. A convenient way to do this padding is to change the radix from m to m^2 and to multiply the digits of u by m to shift them to the left one position. This leads to an arithmetical characterization of shuffling.

Proposition 2. *Let* $u = a_1 \cdots a_n$ *and* $v = b_1 \cdots b_n$ *be strings over the alphabet* $\{0, 1, \ldots, m-1\}$. *If*

$$(u)_m = \sum_{i=1}^{n} a_i m^{n-i}$$

and

$$(u)_{m^2} = \sum_{i=1}^{n} a_i m^{2(n-i)}$$

then

$$(u \mid v)_m = m \cdot (u)_{m^2} + (v)_{m^2}. \tag{7}$$

We now turn to some basic facts about shuffled strings and the lexicographical ordering.

Lemma 4. *If* $w_1 < w_2$ *and* $w_3 < w_4$ *then*

$$(w_1 \mid w_3) < (w_2 \mid w_4), \tag{8}$$

Conversely, if (8) holds then at least one of $w_1 < w_2$ *or* $w_3 < w_4$ *is true.*

Proof: We compare strings of the same length one alphabet letter a time. Let $w_1 = a_1 \cdots a_n$, $w_2 = b_1 \cdots b_n$, $w_3 = c_1 \cdots c_n$, $w_4 = d_1 \cdots d_n$. Since $w_1 < w_2$, either $a_1 < b_1$, in which case (8) is true, or $a_1 = b_1$. In the latter case, we compare the second letters of the shuffled words. Here, since $w_3 < w_4$, either $c_1 < d_1$, in which case (8) is true, or $c_1 = d_1$. If $a_1 = b_1$ and $c_1 = d_1$ we shift right one letter and repeat the argument. Since the inequalities postulated were strict, (3) of Definition 1 ensures that (8) eventually holds.

Conversely, suppose (8) holds. In Definition 1, (2) is not applicable since $(w_1 \mid w_3)$ and $(w_2 \mid w_4)$ have the same length. From (3) there exists $i \in \{1, \ldots, n\}$ such that either

$$a_1 c_1 \cdots a_i c_i = b_1 d_1 \cdots b_i d_i \text{ and } a_{i+1} < b_{i+1} \tag{9}$$

or

$$a_1 c_1 \cdots a_i = b_1 d_1 \cdots b_i \text{ and } c_{i+1} < d_{i+1}. \tag{10}$$

Now, if (9) holds then

$$a_1 = b_1, \ldots, a_i = b_i \text{ and } a_{i+1} < b_{i+1}.$$

In this case $w_1 < w_2$. In the same way, (10) implies $w_3 < w_4$. ∎

Lemma 5. *If $u, v \in A^n$ and $u \leq v$ then $(u \mid v) \leq (v \mid u)$.*

Proof: We proceed by induction on n. First note for $n = 1$ that if $a, b \in A$ and $a \leq b$ then $ab \leq ba$. Now suppose the result is true for strings of lengths $1, \ldots, n-1$. If $u = a_1 \cdots a_n < v = b_1 \cdots b_n$ then we compare the strings

$$(u \mid v) = a_1 b_1 a_2 b_2 \cdots a_{n-1} b_{n-1} (a_n b_n)$$

and

$$(v \mid u) = b_1 a_1 b_2 a_2 \cdots b_{n-1} a_{n-1} (b_n a_n).$$

Either $a_1 \cdots a_{n-1} < b_1 \cdots b_{n-1}$ in which case the induction hypothesis establishes the required inequality independently of a_n and b_n or the initial segments of length $n-1$ are equal, in which case $a_n \leq b_n$, since by hypothesis, $u \leq v$. ∎

We observe that $u \neq v$ implies $(u \mid v) \neq (v \mid u)$ so that Lemma 5 may be stated with strict inequalities as well.

We next study the periodicity of the shuffle of two arbitrary strings of length n. For our purposes, we say a string $w \in A^*$ has **period** d if $w = u^m$ for some $u \in A^+$ with $u \in A^d$ and $m > 1$ is chosen as large as possible. Let $u = a_1 \cdots a_n$ and $v = b_1 \cdots b_n$ and assume $(u \mid v)$ has period d. If $d = 1$ then $(u \mid v) = a_1^{2n}$ which implies that $u = v = a_1^n$. If $d = 2$ then

$$a_1 b_1 = a_2 b_2 = \cdots = a_n b_n$$

which implies $u = a_1^n$ and $v = b_1^n$.

Lemma 6. *If $u, v \in A^n$ and $(u \mid v)$ has period d, then d even implies both u and v are periodic with period $d/2$. If d is odd then v is a cyclic permutation of u.*

Proof: We have already noted the cases $d = 1$ and $d = 2$. We change notation slightly for convenience in this proof only. Set $u = a[1] \cdots a[n]$ and $v = b[1] \cdots b[n]$ where $a[i], b[j] \in A$. Note that since $(u \mid v)$ has length $2n$, $2n = dm$ for some $m > 1$.

If d is even then

$$\begin{aligned}(u \mid v) =& a[1]b[1] \cdots a[d/2]b[d/2]a[(d/2)+1]b[(d/2)+1] \\ & \cdots a[d]b[d] \cdots a[n-(d/2)+1]b[n-(d/2)+1] \cdots a[n]b[n].\end{aligned}$$

Hence,

$$\begin{aligned} a[1] = a[(d/2)+1] &= \cdots = a[n-(d/2)+1] \\ b[1] = b[(d/2)+1] &= \cdots = b[n-(d/2)+1] \\ &\vdots \\ a[d/2] = a[d] &= \cdots = a[n] \\ b[d/2] = b[d] &= \cdots = b[n]. \end{aligned}$$

Therefore,

$$\begin{aligned} u &= (a[1] \cdots a[d/2])^m \\ v &= (b[1] \cdots b[d/2])^m. \end{aligned}$$

If d is odd then

$$\begin{aligned}(u \mid v) =& a[1]b[1] \cdots b[(d-1)/2]a[(d+1)/2]b[(d+1)/2]a[(d+3)/2] \\ & \cdots a[d]b[d] \cdots b[n-(n/2)+1]a[n-(d/2)+2] \cdots a[n]b[n].\end{aligned}$$

Hence,

$$\begin{aligned} a[1] = b[(d+1)/2] &= \cdots = b[n-(d/2)+1] \\ b[1] = a[(d+3)/2] &= \cdots = a[n-(d/2)+2] \\ &\vdots \\ b[(d-1)/2] = a[d] &= \cdots = a[n] \\ a[(d+1)/2] = b[d] &= \cdots = b[n]. \end{aligned}$$

We conclude that

$$v = b[1] \cdots b[n] = a[(d+3)/2] \cdots a[n]a[1] \cdots a[(d+1)/2],$$

showing that v is a cyclic permutation of u. ∎

Conversely, if u has odd period d then there is precisely one cyclic permutation of u, namely, $v = u\left[\frac{n+1}{2}\right] \ldots u[n]u[1] \ldots u\left[\frac{n-1}{2}\right]$ such that $(u \mid v)$ has period d. In all other cases, $(u \mid v)$ has period $2d$.

Theorem 1. *If $n \geq 2$ and $u, v \leq L_n$ with $u \leq v$ then $(u \mid v) \in L_{2n}$.*

Proof: We establish first that $(u \mid v)$ is a primitive string. Suppose a cyclic permutation w of $(u \mid v)$ is periodic with period d. Then $w = (u' \mid v')$ for some strings $u, v \in A^*$. It follows that if w is an even cyclic permutation of $(u \mid v)$ then u' is a cyclic permutation of u and v' is a cyclic permutation of v. If w is an odd cyclic permutation of $(u \mid v)$ then u' is a cyclic permutation of v and v' is a cyclic permutation of u. If d is even then u', v' are periodic by Lemma 6 and therefore so are u and v, contradicting the assumption that $u, v \in L_n$. If d is odd then v' is a cyclic permutation of u'. Therefore, v is a cyclic permutation of u, again contradicting the assumption that $u, v \in L_n$.

To complete the proof, note that if $(u \mid v) \notin L_{2n}$ then some cyclic permutation of $(u \mid v)$, say, $w = (u' \mid v') \in L_{2n}$, and $w < (u \mid v)$ in the lexicographical ordering.

As before, if w is an even cyclic permutation of $(u \mid v)$ then u' is a cyclic permutation of u and v is a cyclic permutation of v. But $w = (u' \mid v') < (u \mid v)$ implies by the converse of Lemma 4 that at least one of $u' < u$ or $v' < v$ holds, contradicting either the assumption that $u \in L_n$ or that $v \in L_n$. If w is an odd cyclic permutation of $(u \mid v)$ then u' is a cyclic permutation v^{π^m} of v for some non-negative integer m and v' is a cyclic permutation u^{π^p} of u for some non-negative integer p. By Lemma 5, our hypothesis $u \leq v$ implies $(u \mid v) < (v \mid u)$. Thus, $w = (u' \mid v') < (u \mid v)$ implies $(u' \mid v') < (v \mid u)$. Again by the converse of Lemma 4 we conclude that at least one of $u' < u$ or $v' < v$ holds. But then either

$$u' = v^{\pi^m} < v$$

or

$$v' = u^{\pi^p} < u.$$

From Definition 2 we now conclude that either $v \notin L_n$ or $u \notin L_n$ completing the proof. ∎

One can compare the operations of concatenation and shuffling by noting that concatenation produces a Lyndon word uv from Lyndon words u and v under the restriction $u < v$. On the other hand, shuffling produces a Lyndon word $(u \mid v)$ under the weaker restriction $u \leq v$ but requires that u and v have the same length. In particular, if u is any Lyndon word of length n then $(u \mid u)$ is a Lyndon word if $n > 1$ but $u \cdot u = u^2$ is never a Lyndon word. The following proposition states an obvious relationship between concatenation and shuffling.

Proposition 3. *For any strings w_1, w_2, w_3 and $w_4 \in A^n$,*

$$(w_1 \mid w_2)(w_3 \mid w_4) = ((w_1 w_3) \mid (w_2 w_4)).$$

Paths of Lyndon Words in the N-Cube

We now restrict our attention to the binary alphabet $A = \{0, 1\}$.

It was shown in [2] that the set of binary Lyndon words in L_n when viewed as vertices of the n-cube, form a connected subset. It was conjectured there that for $n > 2$ the strings in L_n can be listed so that there is only one bit change between successive strings. Equivalently, the conjecture states that the subgraph of the n-cube determined by L_n for $n > 2$ has a Hamilton path; i.e., a path containing every vertex of the subgraph once and only once. If the conjecture is true then the storage and generation of binary Lyndon words becomes computationally very efficient because, for fixed n, only one string need be stored together with a list of $S(n,2) - 1$ change digits. We note that Duval's Algorithm [3] does not provide a solution to this problem because the lexicographical ordering of L_n has many adjacent words which are not at distance 1 in the n-cube.

Definition 4. *The* n-**cube** *is the graph whose* **vertices** *are the* 2^n *strings of* $\{0,1\}^n$. *The* **edges** *of the* n*-cube are the pairs* (α,β) *with* $\alpha,\beta \in : \alpha,\beta \in \{0,1\}^n$ *and* $d(\alpha,\beta) = 1$, *where* $d(\alpha,\beta)$ *denotes the Hamming distance between* α *and* β*; i.e., the number of bits in which* α *and* β *differ. Any* **path** *in the* n*-cube is an ordered list of vertices* $w_1,\ldots,w_m$ *where* $d(w_i,w_{i+1}) = 1$ *for* $i = 1,\ldots,m-1$. *A* **path** *in* L_n *is a path of the* n*-cube whose vertices are Lyndon words.*

Theorem 2. *If there is a path in* L_n *with* m *distinct vertices then there is a path with* $2m - 1$ *distinct vertices in* L_{2n}.

Proof: The proof is constructive. If $w_1,\ldots,w_m$ is a path in L_n then $d(w_i,w_{i+1}) = 1$ for $i = 1,\ldots,m-1$. Therefore, $d((w_i \mid w_i),(w_{i+1} \mid w_{i+1})) = 2$. By Theorem 1 $(w_i \mid w_i)$ and $(w_{i+1} \mid w_{i+1})$ are in L_n. We construct a path in L_{2n} by interpolating between each pair $(w_j \mid w_j)$ and $(w_{j+1} \mid w_{i+1})$ the vertex $(w_i \mid w_{i+1}) \in L_{2n}$ if $w_i < w_{i+1}$ or the vertex $(w_{i+1} \mid w_i) \in L_{2n}$ if $w_i > w_{i+1}$. Theorem 1 ensures that the appropriate choice in each case is a Lyndon word. If, for example, $w_i < w_{i+1}$ then

$$d((w_i \mid w_i),(2_i \mid w_{i+1})) = d((w_i \mid w_{i+1}),(w_{i+1} \mid w_{i+1})) = 1.$$

Note, further, that all vertices of the constructed path in L_{2n} are distinct since the initial path of vertices in L_n was assumed to contain only distinct vertices. ∎

Repeated application of Theorem 2 leads to longer paths in higher order cubes but this is not as useful as might be hoped since the number of Lyndon words increases at the same rate.

Corollary 2.1. *Let* k *be a positive integer. If there exists a path of* m *distinct vertices in* L_n *then there exists a path of* $2^k(m-1)+1$ *distinct vertices in* $L_{2^k n}$.

Proof: The kth iteration of the construction of Theorem 3 yields Lyndon words of length $2^k n$. An inductive argument shows that the number of words produced is $2^k m - (2^k - 1) = 2^k(m-1) + 1$. ∎

In the following theorem we combine shuffling and concatenation to obtain longer paths of distinct vertices of Lyndon words.

Theorem 3. *Let r and s be positive integers. If there is a path of m distinct vertices in L_n then there is a path of $2r(m-1)+1$ distinct vertices in L_{2rn+s}.*

Proof: Let $w_1, \ldots, w_m$ be a path of m distinct vertices in L_n. With these vertices, construct a path $y_1, \ldots, y_{2m-1}$ of distinct vertices in L_{2n} as in Theorem 2. If $y_i < y_{i+1}$ in the lexicographical ordering then construct by concatenation the r strings

$$\begin{aligned} z_0^{(i)} &= 0^p y_i^r 1^q \\ z_1^{(i)} &= 0^p y_i^{r-1} y_{i+1} 1^q \\ &\vdots \\ z_{r-1}^{(i)} &= 0^p y_i y_{i+1}^{r-1} 1^q \\ z_r^{(i)} &= 0^p y_{i+1}^r 1^q, \end{aligned} \tag{11}$$

where p and q are positive integers such that $p+q=s$.

If $y_i > y_{i+1}$ then we construct instead a set of strings similar to those of (11) by interchanging the roles of y_i and y_{i+1}.

The strings thus constructed are in L_{2rn+s}:

$$z_0^{(1)}, \ldots, z_{r-1}^{(1)}, z_0^{(2)}, \ldots, z_0^{(2m-1)}, \ldots, z_{r-1}^{(2m-1)}, z_r^{(2m-1)}.$$

They are seen to be distinct by direct comparison since $w_1, \ldots, w_m$ are assumed distinct. Since we start with $2m-1$ distinct shuffles of $w_1, \ldots, w_m$ and we interpolate $r-1$ strings between each $z_0^{(i)} = 0^p y_i^r 1^q$; $i = 1, \ldots, 2m-1$ there are $2m-1+(2m-2)(r-1) = 2r(m-1)+1$ distinct strings thus constructed.

We now show that each of the constructed strings $z_t^{(i)}$ is in L_{2rn+s}. The $z_t^{(i)}$ come in various forms depending on the choices made in the construction of $y_1, \ldots, y_{2m-1}$. Suppose that $z_t^{(i)} = 0^p(w_i \mid w_i)^{r-t}(w_i \mid w_{i+1})^t 1^q$. Then, $y_i = (w_i \mid w_i)$ and $y_{i+1} = (w_i \mid w_{i+1})$ and $y_i < y_{i+1}$. Now y_i is not a prefix of y_{i+1}, simply because they have the same length. From (4) we obtain $y_i^{r-t} < y_{i+1}^t$ when we take $w = y_i^{r-t-1}$ and $z = y_{i+1}^{t-1}$. Now from (6) we obtain

$$0^p y_i^{r-t} \leq y_i^{r-t} < y_{i+1}^t \leq y_{i+1}^t 1^q. \tag{12}$$

It now follows from Lemma 2 that $z_t^{(i)}$ is a Lyndon word. The argument is similar in the other cases.

Note that we have assumed that r and s are positive integers. Thus, for $s = p+q$ either of p or q may be zero, but at least one is non-zero. This is required because the construction would otherwise contain strings y_i^r which are periodic and so not Lyndon words.

References

1. K.T. Chen, R.H. Fox, R.C. Lyndon, *Free differential calculus IV—the quotient groups of the lower central series*, Annals of Mathematics **68** (1958), 81–95.
2. L.J. Cummings, *Connectivity of Lyndon words in the N-cube*, The Journal of Combinatorial Mathematics and Combinatorial Computing **3** (1988), 93–96.
3. J.-P. Duval, *Génération d'une section des classes de conjugation et arbre des mots de Lyndon de longeur bornée.* report 88-20, March 1988, LITP, Paris
4. M. Lothaire, "Combinatorics on Words", Addison-Wesley, Massachusetts, 1983.
5. R.C. Lyndon, *On Burnside's problem I*, Transactions of the American Mathematical Society **77** (1954), 202–215.
6. N. Metropoulis, G.-C. Rota, *Witt vectors and the algebra of necklaces*, Advances in Mathematics **50** (1983), 95–125.
7. C. Reutenauer, *Mots circulaires et polynômes irréductibles.* report 87-39, June 1987, LITP, Paris