Shorter strings containing all $k$-element permutations

Eugen Zălinescu

Computer Science Department, ETH Zurich, Switzerland

Abstract

We consider the problem of finding short strings that contain all permutations of order $k$ over an alphabet of size $n$, with $k \leq n$. We show constructively that $k(n - 2) + 3$ is an upper bound on the length of shortest such strings, for $n \geq k \geq 10$. Consequently, for $n \geq 10$, the shortest strings that contain all permutations of order $n$ have length at most $n^2 - 2n + 3$. These two new upper bounds improve with one unit the previous known upper bounds.

Keywords: Combinatorial problems, Permutations, Subsequences

1. Introduction

The problem of finding (shortest) strings that contain, as subsequences, all permutations of order $n$ was identified by R. Karp in 1971, and stated in [3]. In the sequel, it has been investigated in [7, 1, 5, 4, 6], which show constructively that $n^2 - 2n + 4$ is an upper bound on the length of shortest such strings, each work presenting a different construction. Moreover, M. Newey [7] showed that this bound is tight for $n \leq 7$. However, while it is tempting to conjecture tightness also for $n \geq 8$, counterexamples of length 83 and 102 have been recently and anonymously identified [2] for $n = 10$ and respectively $n = 11$. Motivated by these counterexamples, we present a construction of strings of length $n^2 - 2n + 3$ that contain all permutations of order $n$, for any $n \geq 10$.

A generalization of the above problem consists of finding (shortest) strings that contain all permutations of order $k$ over an alphabet of size $n$. (For $k = n$ we find the original problem.) M. Newey [7] and C. Savage [8] present different constructions which show that $k(n - 2) + 4$ is an upper bound on the length of shortest such strings. Following C. Savage’s approach, we use the construction from the particular case to build strings of length $k(n - 2) + 3$ that contain all permutations of order $k$ over an alphabet of size $n$. We thus obtain the new upper bound $k(n - 2) + 3$ in the general case, for $n \geq k \geq 10$.

2. Preliminaries

We mainly use L. Adleman’s [1] notations.

When a string $\sigma$ is a subsequence of a string $\gamma$, we say that $\gamma$ contains $\sigma$, or $\sigma$ is contained in $\gamma$. For instance, $abc$ contains $ac$. Given a set of strings $Q$, 

the string $\gamma$ is $Q$-complete if $\gamma$ contains each string in $Q$. For example, $aba$ is $\{ab, ba\}$-complete.

For an alphabet $\Sigma$, we let $P_\Sigma$ denote the set of all permutations of $\Sigma$, and $L_k^\Sigma$ denote the set of all strings in $\Sigma^*$ of length $k$ in which no letter occurs twice.

For any $n \geq 1$, $\Sigma_n$ denotes an alphabet of size $n$. We assume without loss of generality that $\Sigma_n = \{1, \ldots, n\}$. $P_n$ abbreviates $P_{\Sigma_n}$ and $L_k^n$ abbreviates $L_k^{\Sigma_n}$. For instance, $P_3 = \{123, 132, 213, 231, 312, 321\}$, $L_2^3 = \{12, 13, 21, 23, 31, 32\}$, the string $1213121$ is $P_3$-complete, and the string $1213211$ is $L_2^3$-complete.

Note that $L_1^n = \Sigma_n$, $L_n^n = P_n$, and any element of $P_n$ is $L_1^n$-complete. For instance, $P_3 = \{123, 132, 213, 231, 312, 321\}$, $L_2^3 = \{12, 13, 21, 23, 31, 32\}$, the string $1213121$ is $P_3$-complete, and the string $1213211$ is $L_2^3$-complete.

Concatenation of strings is written as juxtaposition. The notation is extended to sets of strings as expected. For example, $\gamma Q := \{\gamma \sigma \mid \sigma \in Q\}$.

A string $\sigma = s_1s_2\ldots s_k \in \Sigma_n^*$ with $s_i \in \Sigma_n$ is an $R_n$-string if $s_{i+1} = (s_i \mod n) + 1$ for all $1 \leq i \leq k - 1$. We denote by $R^\alpha_n(a)$ the $R_n$-string of length $\ell$ that starts with the letter $a$. For instance, $R^3_5(2) = 23123$.

3. Two previous constructions

Before presenting the details of our construction, we sketch the ones of [1] and [8].

L. Adleman’s construction [1] starts from the string $R^{(n-2)(n-1)+2}_{n-1}(1)$. Next, the string is split into $n - 1$ strings ($n - 3$ strings of length $n - 2$, first and last string of length $n - 1$); we call these strings $\alpha$-blocks. Finally, the letter $n$ is inserted between the blocks, before the first block, and after the last block. The string thus obtained is $P_n$-complete. The proof is based on the following key property: any $R_n$-string of length $k(n - 1) + 1$ is $L_n^k$-complete.

C. Savage [8] generalizes L. Adleman’s construction by starting instead from the string $R^{(k-2)(k-1)+2}_{k-1}(1)$, splitting it into $k - 1$ blocks, and inserting the string $\{k(k+1)\ldots n\}$ between the blocks, before the first block, and after the last block. The obtained string is $L_n^k$-complete and the proof is based on the same key property.

Our approach is similar. We start directly from $k - 1$ so-called $\beta$-blocks, which are build from $\alpha$-blocks by inserting the letter $(k - 1)$ in appropriate positions. We then apply C. Savage’s construction on the $\beta$-blocks, inserting the string $k(k+1)\ldots n$ between, before, and after the $\beta$-blocks, to obtain an $L_n^k$-complete string. However, as we do not have an analogue of the key property for $\beta$-blocks, our proof is completely different from the ones in [1] and [8].

4. A construction of $L_n^k$-complete strings of length $k(n - 2) + 3$

From now on, we fix two integers $n$ and $k$, with $n \geq k \geq 10$. Also, we denote the letter $(k - 1)$ by $A$.

\footnote{Actually, instead of the string $k(k+1)\ldots n$, any permutation of $\{k, k+1, \ldots, n\}$ would do.}
The construction makes use of the so-called $\alpha$-blocks and $\beta$-blocks. The $\alpha$-blocks $\alpha_1, \ldots, \alpha_{k-1}$ are the following strings:

$$\begin{align*}
\alpha_1 &:= \alpha_2 := R_{k-2}^k(1) = 12 \ldots (k-2), \\
\alpha_j &:= R_{k-2}^j(k-j+2), \text{ for } 3 \leq j \leq k-3, \\
\alpha_{k-1} &:= \alpha_{k-2} := R_{k-2}^k(4) = 45 \ldots (k-2)123.
\end{align*}$$

The $\beta$-blocks $\beta_1, \ldots, \beta_{k-1}$ are the following strings:

$$\begin{align*}
\beta_1 &:= \alpha_1 A, \quad \beta_2 := \alpha_2, \quad \beta_3 := A\alpha_3, \\
\beta_j &:= \alpha_1 a_2 \ldots a_{k-4} A a_{k-3}, \text{ for } 4 \leq j \leq \lfloor \frac{k}{2} \rfloor - 1, \\
\beta_j &:= \alpha_j, \text{ for } \lfloor \frac{k}{2} \rfloor \leq j \leq \lceil \frac{k}{2} \rceil \\
\beta_j &:= \alpha_1 A a_2 \ldots a_{k-4} a_{k-3}, \text{ for } \lceil \frac{k}{2} \rceil + 1 \leq j \leq k-4, \\
\beta_{k-3} &:= \alpha_{k-3} A, \quad \beta_{k-2} := \alpha_{k-2}, \quad \beta_{k-1} := A\alpha_{k-1},
\end{align*}$$

where $a_j = a_1 \ldots a_{k-4}$ for $4 \leq j \leq k-4$. Note that $\lfloor \frac{k}{2} \rfloor \leq j \leq \lceil \frac{k}{2} \rceil$ is equivalent with $j = \ell$ when $k = 2\ell$, and with $j \in \{\ell, \ell+1\}$ when $k = 2\ell + 1$.

Table 1 details the $\beta$-blocks in all cases. A letter $b \in \Sigma_{k-1}$ is missing from $\beta_j$ if $b$ does not occur in $\beta_j$.

Let $\Sigma_{k,n} := \{k, \ldots, n\}$ and $\Sigma_{A,n} := \Sigma_{k,n} \cup \{A\}$. The set of strings $T_n^k$ is

$$\{ \tau_1 \beta_1 \tau_2 \ldots \tau_{k-1} \beta_{k-1} \tau_k \mid \tau_j \in P_{\Sigma_{A,n}} \text{ for } k \text{ odd and } j = \lceil \frac{k}{2} \rceil, \\
\tau_j \in P_{\Sigma_{k,n}} \text{ otherwise} \}. $$

That is, strings in $T_n^k$ are obtained by concatenating the $\beta$-blocks and inserting between them, before the first block, and after the last block, arbitrary permutations of the set $\Sigma_{k,n}$; with one exception: for $k$ odd ($k = 2\ell + 1$), the block inserted in the middle (that is, $\tau_{\ell+1}$) is an arbitrary permutation of $\Sigma_{A,n}$.

The set we have just built is the focus of this paper. As examples, the following two strings are in $T_{10}^k$ and respectively in $T_{13}^k$:

- $a \cdot 12345678A \cdot a \cdot 12345678 \cdot a \cdot A1234567 \cdot a \cdot 812345A6 \cdot a \cdot 7812345 \cdot a \cdot 64781234 \cdot a \cdot 56781234 \cdot a \cdot 45678123 \cdot a \cdot a$ and
- $bcd \cdot 123456789A \cdot cdb \cdot 123456789 \cdot dbc \cdot A12345678 \cdot cdb \cdot 9123456A7 \cdot cdb \cdot 89123456 \cdot bAdc \cdot 78912345 \cdot dcb \cdot 6A7891234 \cdot dcb \cdot 56789123A \cdot dcb \cdot A56789123 \cdot bdc \cdot A456789123 \cdot bdc$, $bc$.

where $\cdot$ denotes concatenation and delimits the $\beta$-blocks, and $a$, $b$, $c$, and $d$ denote the letters 10, 11, 12, and respectively 13. We recall that $A = 9$ in the first string, and $A = 10$ in the second.

5. Main result

This section is devoted to the proof of the following theorem.

**Theorem 1.** Any string in $T_n^k$ is $L_n^k$-complete and has length $k(n-2)+3$. 

3
In what follows, \( \gamma \) denotes an arbitrary element of \( S_j \).

**Lemma 1.** The following statements hold:

1. For any \( 1 \leq j \leq k - 1 \), any \( \gamma \in S_j \), \( \gamma \) is \( L_{2n}^{(k)} \)-complete.
2. For any \( 1 \leq j \leq \left\lfloor \frac{k}{2} \right\rfloor - 1 \), any \( \gamma \in S_j \), \( \gamma^{-2} \) is \( L_{2n}^{(k)}_{\Sigma - \{a_1, \ldots, a_j\}} \)-complete.

**Proof.** We will often use the following property: if \( \gamma = a_m a_{m-1} \ldots a_1 \in \Sigma^* \) is \( L_{2n}^{(k)} \)-complete, then \( \gamma^{-1} \) is \( L_{2n}^{(k)}_{\Sigma - \{a_1, \ldots, a_j\}} \)-complete, where \( i < m \) and \( j \leq |\Sigma| - i \).

In what follows, \( \gamma_j \) denotes an arbitrary element of \( S_j \).
We proceed by induction on \( j \).

**Base cases.** Clearly, \( \gamma_1 \) is \( \Sigma^L_{2n} \)-complete and \( \gamma_1^{-2} \) is \( \Sigma^L_{2n-(k-2,A)} \)-complete. We also easily check that \( \gamma_2 \) is \( \Sigma^L_{2n} \)-complete and \( \gamma_2^{-2} \) is \( \Sigma^L_{2n-(k-3,A)} \)-complete.

**Inductive case.** Suppose now that \( j \geq 3 \). We let \( \gamma_j = \gamma_{j-1} \tau_j \beta_j \) with \( \gamma_{j-1} \in S_j \) and \( \gamma_j \in \mathcal{P}_{2k,n} \), and \( \gamma_{j-1} = \gamma_{j-2} \tau_j \beta_{j-1} \), with \( \gamma_{j-2} \in S_j \) and \( \tau_j \in \mathcal{P}_{2k,n} \). By induction hypothesis, \( \gamma_{j-1} \) is \( \Sigma^L_{2n} \)-complete, \( \gamma_{j-2} \) is \( \Sigma^L_{2n} \)-complete, and, if \( j \leq \lfloor \frac{k}{2} \rfloor - 1 \) then \( \gamma_j^{-2} \) is \( \Sigma^L_{2n-(k-3,A)} \)-complete and \( \gamma_j^{-2} \) is \( \Sigma^L_{2n-(k-j+1,A)} \)-complete.

We consider the two statements in turn.

(1) Let \( \sigma = s_1 \ldots s_j \in \mathcal{L}^j_{\Sigma_n} \).

Suppose first that \( s_j \) occurs in \( \tau_j \beta_j \). As \( \gamma_{j-1} \) is \( \Sigma^L_{2n} \)-complete, \( \gamma_{j-1} \) contains \( s_1 \ldots s_j \). Thus \( \gamma_j \) contains \( \sigma \).

Suppose now that \( s_j \) does not occur in \( \tau_j \beta_j \). Then \( j \leq k - 2 \). Note that for \( k \) odd and \( j = \lfloor \frac{k}{2} \rfloor \), the letter \( A \) is missing from \( \beta_j \), but it is not missing from \( \tau_j \beta_j \). Depending on the value of \( j \), we can have one of the following cases:

(a) \( s_j = k - j + 1 \), for \( 3 \leq j \leq k - 3 \). We have that the last letter of \( \gamma_{j-1} \) is \( (k - j + 1) \). Then \( \gamma_j^{-1} \) is \( \Sigma^L_{2n-(k-j+1)} \)-complete. Hence \( s_1 \ldots s_j \) is contained in \( \gamma_j^{-1} \), and \( s_j \) equals the last letter of \( \gamma_{j-1} \). Thus, \( \sigma \) is contained in \( \gamma_j^{-1} \), and moreover in \( \gamma_j \).

(b) \( s_j = A \), for \( j = k - 2 \). As \( \gamma_{j-1} \) is \( \Sigma^L_{2n} \)-complete and the last letter of \( \gamma_{j-1} \) is \( A \), we have that \( \gamma_j^{-1} \) is \( \Sigma^L_{2n-(A)} \)-complete. Hence \( \sigma \) is contained in \( \gamma_j^{-1} \).

(c) \( s_j = A \), for \( j = \lfloor \frac{k}{2} \rfloor \), \( k \) even, and for \( j = \lceil \frac{k}{2} \rceil \), \( k \) odd. Note that \( j \geq 5 \) and that \( s_j \) equals the second last letter of \( \beta_{j-1} \). We can have:

(i) \( s_{j-1} \notin \{k - j + 1, k - j + 2\} \). That is, \( s_{j-1} \) is neither the missing letter from \( \beta_{j-1} \), nor the letter after \( A \) in \( \beta_{j-1} \) (i.e. the last letter of \( \beta_{j-1} \)). Then \( s_{j-1} \) is contained in \( \tau_{j-1} \beta_{j-2} \). As \( \gamma_{j-2} \) is \( \Sigma^L_{2n} \)-complete, \( s_1 \ldots s_{j-2} \) is contained in \( \gamma_{j-2} \). Hence \( \sigma \) is contained in \( \gamma_{j-1} \).

(ii) \( s_{j-1} = k - j + 2 \) (the missing letter from \( \beta_{j-1} \)). As \( \gamma_{j-2} \) is \( \Sigma^L_{2n} \)-complete and the last letter of \( \beta_{j-2} \) is \( (k - j + 2) \), we have that \( \gamma_{j-1}^{-1} \) is \( \Sigma^L_{2n-(k-j+2)} \)-complete, hence also \( \Sigma^L_{2n-(k-j+2,A)} \)-complete. Hence \( s_1 \ldots s_{j-2} \) is contained in \( \gamma_{j-2}^{-1} \), and \( s_{j-1} \) equals the last letter of \( \gamma_{j-2} \). Hence \( \sigma \) is contained in \( \gamma_{j-1} \).

(iii) \( s_{j-1} = k - j + 1 \) (the last letter of \( \beta_{j-1} \)). We distinguish two cases:

- For \( j - 2 \geq 4 \), as \( \gamma_{j-2}^{-2} \) is \( \Sigma^L_{2n-(k-j+1,A)} \)-complete and the last letter of \( \gamma_{j-2}^{-2} \) is \( (k - j + 1) \), it follows that \( \gamma_{j-2}^{-3} \) is \( \Sigma^L_{2n-(k-j+1,A)} \)-complete. Thus \( s_1 \ldots s_{j-2} \) is contained in \( \gamma_{j-2}^{-3} \) and \( s_{j-1} \) equals the third last letter of \( \gamma_{j-2} \). Hence \( \sigma \) is contained in \( \gamma_{j-2} \).
If \( j - 2 < 4 \) then \( j - 2 = 3 \) (since \( j \geq 5 \)). The last letter of \( \gamma_{j-2}^{-1} \) is \((k-j+1)\). It follows that \( \gamma_{j-2}^{-1} \) is \( L_{\Sigma_n-(\{k-j+1,A\})}^{j-2} \)-complete.

Thus \( s_1 \ldots s_{j-2} \) is contained in \( \gamma_{j-2}^{-1} \) and \( s_{j-1} \) equals the second last letter of \( \gamma_{j-2} \). Hence \( \sigma \) is contained in \( \gamma_{j-1} \).

(2) Let \( \sigma = s_1 \ldots s_j \in L_{\Sigma_n-(\{k-j-1,A\})}^j \). We can have:

(a) \( s_j \notin \{k-j+1,k-j\} \). That is, \( s_j \) is neither the missing letter of \( \beta_j \) nor the last letter of \( \beta_j \). Then \( s_j \) is contained in \( \tau_j \beta_j^{-2} \). And as \( \gamma_{j-1} \) is \( L_{\Sigma_n}^{j-1} \)-complete, it follows that \( \sigma \) is contained in \( \gamma_{j-2}^{-1} \).

(b) \( s_j = k-j+1 \). As \( \gamma_{j-1} \) is \( L_{\Sigma_n}^{j-1} \)-complete and \((k-j+1)\) is the last letter of \( \gamma_{j-1} \) we have that \( \gamma_{j-1}^{-1} \) is \( L_{\Sigma_n}^{j-1-(k-j+1)} \)-complete. It follows that \( \gamma_{j-1}^{-1} \) contains \( s_1 \ldots s_{j-1} \), and thus \( \sigma \) is contained in \( \gamma_{j-1} \), thus in \( \gamma_{j-2}^{-1} \).

(c) \( s_j = k-j \). As \( \gamma_j^{-2} \) is \( L_{\Sigma_n-(\{k-j,A\})}^{j-1} \)-complete, we have that \( \gamma_j^{-2} \) contains \( s_1 \ldots s_{j-1} \). For \( j-1 \leq 3 \), \((k-j)\) is the second last letter of \( \gamma_{j-1} \), hence \( \gamma_{j-1}^{-1} \) contains \( \sigma \). For \( j-1 > 3 \), \((k-j)\) is the third last letter of \( \gamma_{j-1} \), hence \( \gamma_{j-1}^{-3} \) is \( L_{\Sigma_n-(\{k-j,A\})}^{j-1} \)-complete, thus \( \gamma_{j-1}^{-1} \) contains \( \sigma \).

The following lemma implies that the reverse of any string in \( T_n^k \) is isomorphic with a string in \( T_n^k \). Let us denote by \( \overline{\sigma} \) the reverse of a string \( \omega \). Also, for \( \gamma = \tau_1 \beta_1 \tau_2 \ldots \tau_{k-1} \beta_k-1 \tau_k \in T_n^k \), let \( \gamma_j := \beta_j \tau_{j+1} \ldots \beta_{k-1} \tau_k \) for any \( 1 \leq j \leq k-1 \).

Let \( g \) be the bijection on \( \Sigma_n \) given by \( g(a) := 1 + (k + 1 - a) \pmod{(k - 2)} \) for \( 1 \leq a \leq k-2 \), and \( g(a) := a \) for \( k-1 \leq a \leq n \).

Lemma 2. For any \( \gamma \in T_n^k \), for any \( 1 \leq j \leq k-1 \), we have \( \overline{\gamma_j} \in g(S_{k-j}) \).

Proof. Clearly, \( \tau \in P_{\Sigma_k,n} \) for any \( \tau \in P_{\Sigma_k,n} \).

We have \( \overline{\gamma_j} := g(\alpha_{k-j}) \) for any \( 1 \leq j \leq k-1 \). Hence also \( \overline{\beta_j} := g(\beta_{k-j}) \) for any \( 1 \leq j \leq k-1 \). It follows that \( \overline{\gamma_j} = \beta_j \tau_{j+1} \ldots \beta_{k-1} \tau_k \) is \( \overline{\gamma_j} = \tau_k \beta_{k-1} \ldots \tau_{j+1} \beta_j = \tau_k g(\beta_1) \ldots \tau_{j+1} g(\beta_{k-j}) = g(\tau_k \beta_1 \ldots \tau_{j+1} \beta_{k-j}) \). As \( \overline{\gamma_j} \beta_{j+1} \beta_{j-k} \in S_{k-j} \), we have \( \overline{\gamma_j} \in g(S_{k-j}) \). \( \square \)

We can now conclude the proof of Theorem 1.

Let \( \gamma \) be an arbitrary string in \( T_n^k \) and \( \sigma = s_1 \ldots s_k \) be an arbitrary string in \( L_k \). Let \( j \) be such that \( s_j \in \Sigma_{k,n} \), with \( 1 \leq j \leq k \). There is such a \( j \) by the pigeonhole principle, as \( |\Sigma_n - \Sigma_{k,n}| = k - 1 \) and the \( k \) letters of \( \sigma \) are distinct.

Let \( \gamma' = \tau_1 \beta_1 \tau_2 \ldots \tau_{j-1} \beta_{j-1} \) and \( \gamma'' = \beta_j \tau_{j+1} \beta_{j+1} \ldots \beta_{k-1} \tau_k \). Then \( \gamma = \gamma' \gamma'' \). (If \( j = 1 \) then \( \gamma' = \varepsilon \) and \( \gamma = \tau_1 \gamma'' \), while if \( j = k \) then \( \gamma'' = \varepsilon \) and \( \gamma = \gamma' \tau_k \), where \( \varepsilon \) is the empty string.)

As \( \gamma' \in S_{j-1} \), by Lemma 1(1), we have that \( \gamma' \) contains \( s_1 \ldots s_{j-1} \). By Lemma 2, we have that \( \overline{\gamma'} \in g(S_{k-j}) \). Thus, again by Lemma 1(1), we have that \( g^{-1}(\overline{\gamma'}) \) contains \( g^{-1}(\tau_k \ldots s_{j+1}) \), and hence \( \gamma'' \) contains \( s_{j+1} \ldots s_k \). Clearly, \( \tau_j \) contains \( s_j \). Putting the three pieces together, we obtain that \( \gamma \) contains \( \sigma \).
6. Conclusions

In this paper we have built a set $T^k_n$ of strings, each string being of length $k(n-2)+3$ and containing all permutations of order $k$ over an alphabet of size $n$, for $n \geq k \geq 10$. We thus improve by one unit the previous known upper bound on the length of the shortest such strings, which was $k(n-2)+4$. It remains open if further improvements are possible. We also do not know whether the set $T^k_n$ is complete, that is, whether there exist other such strings, not isomorphic with the ones in $T^k_n$.

In the particular case when $k = n$, our construction shows that $f(n) \leq n^2-2n+3$, for $n \geq 10$, where $f(n)$ denotes the length of the shortest strings that contain all permutations of order $n$ (over an alphabet of size $n$). As M. Newey [7] proved that $f(n) = n^2-2n+4$, for $3 \leq n \leq 7$, it is compelling to find the values of $f(8)$ and $f(9)$.

Acknowledgments. Mohammad Torabi Dashti introduced us to the problem. Lăcrămioara Aștefănoaei made useful comments. We warmly thank them both. We also thank the anonymous referees for their helpful suggestions.


