3-2-2015

Twins in words and long common subsequences in permutations

Boris Bukh
Carnegie Mellon University, bbukh@math.cmu.edu

Lidong Zhou

Published In
Israel Journal of Mathematics, accepted.

Twins in words and long common subsequences in permutations

Boris Bukh Lidong Zhou

Abstract

A large family of words must contain two words that are similar. We investigate several problems where the measure of similarity is the length of a common subsequence.

We construct a family of $n^{1/3}$ permutations on n letters, such that LCS of any two of them is only $cn^{1/3}$, improving a construction of Beame, Blais, and Huynh-Ngoc. We relate the problem of constructing many permutations with small LCS to the twin word problem of Axenovich, Person and Puzynina. In particular, we show that every word of length n over a k-letter alphabet contains two disjoint equal subsequences of length $cnk^{-2/3}$. Connections to other extremal questions on common subsequences are exhibited.

Many problems are left open.

1 Introduction

This paper grew out of attempts to solve the twin word problem of Axenovich, Person and Puzynina [1]. These attempts gave rise to several problems on common subsequences in words, some of which appear to us even more appealing. However, we begin with the twin word problem, as it is the core that the other problems link to.

A *word* is a sequence of letters from some fixed finite alphabet. Since the nature of the alphabet is not important to us, we will usually use $[k] \stackrel{\text{def}}{=} \{1, 2, \dots, k\}$ as a canonical k-letter alphabet. The set of all words of length n is thus denoted $[k]^n$. The i'th letter of a word w is denoted $w[i]$. A *subword* of a word w is a word consisting of several consecutive letters from w. In contrast, a *subsequence* in w is a word consisting of letters from w that are in order, but not necessarily consecutive. For example, radar is a subsequence, but not a subword of abracadabra.

Two subsequences w_1, w_2 of a word w are said to be *twins* if they are equal as words, and no letter from w is in both w_1 and w_2. For example, the word 0110010010101101 contains twins of length 7, as seen from $\overline{01}1\underline{0}\overline{0}\underline{10}\overline{0101}\underline{01}\underline{101}$, where overlines and underlines indicate which of the two subsequences (if any) the letter is in. Let $\mathrm{LT}(w)$ be the length of the longest twins contained in a word w. We also define $\mathrm{LT}(k, n) \stackrel{\text{def}}{=} \min_w \mathrm{LT}(w)$, where the minimum is taken over all words $w \in [k]^n$. The twin problem of Axenovich–Person–Puzynina is to determine how large $\mathrm{LT}(k, n)$ is. The following is a remarkable result:

Theorem 1 (Axenovich–Person–Puzynina [1], Theorem 1)**.** *For words over a binary alphabet we have*

$$\mathrm{LT}(2, n) \geq \tfrac{1}{2}n - o(n).$$

Since the upper bound of $n/2$ is trivial, the result is tight. A very interesting problem, which we do not address in this paper, is to determine the exact behavior of the $o(n)$ term. We refer the reader

to [1] for the best known bounds. Instead we consider what happens as the alphabet size grows. First, we recall the results of Axenovich–Person–Puzynina:

$$\begin{aligned}
&\mathrm{LT}(k,n) \geq \tfrac{1}{k}n - o(n) && \text{for all } k \geq 2, && (1)\\
&\mathrm{LT}(5,n) \leq 0.49n, \\
&\mathrm{LT}(k,n) \leq \left(\frac{e}{\sqrt{k}} - \frac{e^2}{k} + O(k^{-3/2})\right) n && \text{for all large } k. && (2)
\end{aligned}$$

(The upper bound on $\mathrm{LT}(k,n)$ appearing in [1] is different; we have provided its asymptotic form.)

The upper bounds on LT from [1] use the union bound, which for $k = 2, 3, 4$ yields only the trivial bound $\mathrm{LT}(k,n) \leq n/2$. In particular, prior to the present work the best bounds on $\mathrm{LT}(3,n)$ and $\mathrm{LT}(4,n)$ were $1/3 \leq \lim_n \mathrm{LT}(3,n)/n \leq 1/2$, and $1/4 \leq \lim_n \mathrm{LT}(4,n)/n \leq 1/2$. We improve both lower and upper bounds.

The first two results are improved lower bounds.

Theorem 2. *(Proof is in section 3) Let $k \geq 3$. Then we have $\mathrm{LT}(k,n) \geq \frac{1.02}{k}n - o(n)$.*

Theorem 3. *(Proof is in section 2) For the twin problem over a k-letter alphabet we have*

$$\mathrm{LT}(k,n) \geq 3^{-4/3}k^{-2/3}n - 3^{-1/3}k^{1/3}.$$

While theorem 3 is stronger than theorem 2 for large k, together they constitute an improvement over (1) for all $k \geq 3$.

The next result is a new upper bound for $\mathrm{LT}(k,n)$. It is a tiny improvement over (2) for large k, but importantly it shows that $\mathrm{LT}(4,n)$ is not asymptotic to $n/2 + o(n)$. (The question of whether or not $\mathrm{LT}(3,n)$ is asymptotic to $n/2 + o(n)$ remains open.)

Theorem 4. *(Proof is in section 4) Suppose $k \geq 2$, and suppose $\alpha \in [1/k, 1/2]$ is a constant such that the real number*

$$(1-2\alpha)\log\frac{1}{1-2\alpha} - \alpha\log(\alpha^2 k) - 2\alpha\log\left(\frac{2}{1+\sqrt{1-1/k}}\right) + (1-2\alpha)\log(1-1/k)$$

is negative. Then $\mathrm{LT}(k,n) \leq \alpha n$ for all sufficiently large n. In particular, $\mathrm{LT}(4,n) \leq 0.4932n$, $\mathrm{LT}(5,n) \leq 0.48n$, and $\mathrm{LT}(k,n) \leq \left(\frac{e}{\sqrt{k}} - \frac{e^2+1/2}{k} + O(k^{-3/2})\right) n + o(n)$ for all large k.

In the next section we give a short proof of theorem 3. The proof will illustrate a connection with several extremal problems on the longest common subsequences, which we discuss there as well.

2 Common subsequences

A *common subsequence* of two words w and w' is a word that is a subsequence of both w and w'. A common subsequence of more than two words is defined analogously. We denote the length of the longest common subsequence of a set $\mathcal{W}$ of words by $\mathrm{LCS}(\mathcal{W})$. For notational sanity we write $\mathrm{LCS}(w,w')$ in place of $\mathrm{LCS}(\{w,w'\})$.

A special, but important class of words are permutations. A *permutation* is a word in which each letter of the alphabet appears exactly once. We denote by $\mathcal{P}_k$ the set of all permutations on k letters. Note that the group structure of permutations is of no concern in this paper; permutations are just words. A slight generalization of permutations are multipermutations. Given a vector $\vec{s} = (s_1, \ldots, s_k) \in \mathbb{Z}_+^k$, an *$\vec{s}$-multipermutation* is a word in which the letter $l \in [k]$ appears exactly s_l times. Length of an $\vec{s}$-multipermutation is $\|\vec{s}\| \stackrel{\text{def}}{=} \sum_l s_l$. We denote the set of all $\vec{s}$-multipermutations by $\mathcal{P}_{\vec{s}}$.

The basis for our proof of theorem 3 is a result of Beame and Huynh-Ngoc [4, Lemma 5.9] which asserts that if $\pi_1, \pi_2, \pi_3 \in \mathcal{P}_k$ are three permutations, then for some pair $i < j$ we have $\mathrm{LCS}(\pi_i, \pi_j) \geq k^{1/3}$. We require a slight generalization of that result:

Lemma 5. *Let $\pi_1, \pi_2, \pi_3 \in \mathcal{P}_{\vec{s}}$ be three $\vec{s}$-multipermutations of length $\|\vec{s}\| = n$. Then there is a pair $i < j$ such that $\mathrm{LCS}(\pi_i, \pi_j) \geq n^{1/3}$.*

Proof. For each of π_1, π_2, π_3, replace the j'th occurrence of letter l by the letter l_j. An example: 212331 becomes $2_1 1_1 2_2 3_1 3_2 1_2$. This turns π_i into a permutation π_i' over an alphabet of size n. Since $\mathrm{LCS}(\pi_i, \pi_j) \geq \mathrm{LCS}(\pi_i', \pi_j')$, the lemma follows from the result of Beame–Huynh-Ngoc stated above. □

Proof of theorem 3. It suffices to show that every word of length $3k$ contains twins of length $(k/3)^{1/3}$, for then the general result would follow by partitioning a word of length n into intervals of length $3k$. So assume that w is of length $3k$.

For each $l \in [k]$ let C_l be the number of copies of the letter l in w. Define a vector $\vec{s}$ by $s_l = \lfloor C_l/3 \rfloor$. Then w contains three disjoint subsequences w_1, w_2, w_3, each of which is a $\vec{s}$-multipermutation. Thus, by the preceding lemma there are $i < j$ such that

$$\mathrm{LCS}(w_i, w_j) \geq \Big(\sum_{l \in [k]} \lfloor C_l/3 \rfloor\Big)^{1/3} \geq \Big(\sum_{l \in [k]} (C_l - 2)/3\Big)^{1/3} = \big((3k - 2k)/3\big)^{1/3} = (k/3)^{1/3}.$$

Since $\mathrm{LT}(w) \geq \mathrm{LCS}(w_i, w_j)$, the proof is complete. □

There are two natural approaches to improve upon this proof. First, one might hope that using more than three subsequences would increase the $k^{1/3}$ bound in the Beame and Huynh-Ngoc result. Second, instead of replacing copies of a same symbol by different symbols in lemma 5, we might hope to take advantage of the repetitions. We examine these approaches in order.

Common subsequences in large sets of permutations. For a set $\mathcal{W}$ of words let $\mathrm{LCS}_T(\mathcal{W}) \stackrel{\text{def}}{=} \max \mathrm{LCS}(W')$ where the maximum is taken over all sets $\mathcal{W}' \subset \mathcal{W}$ of size T. For a family of words $\mathcal{F}$ let

$$\mathrm{LCS}_T(t, \mathcal{F}) \stackrel{\text{def}}{=} \min_{\mathcal{W} \in \binom{\mathcal{F}}{t}} \mathrm{LCS}_T(\mathcal{W}).$$

With this notation, the Beame–Huynh-Ngoc theorem asserts that $\mathrm{LCS}_2(3, \mathcal{P}_k) \geq k^{1/3}$. For a constant number of permutations, a matching upper bound was proved by Beame, Blais and Huynh-Ngoc [3, Theorem 2]:

$$\mathrm{LCS}_2(t, \mathcal{P}_k) \leq 32(tk)^{1/3} \qquad \text{for all } t \leq k^{1/2}.$$

We tweak the construction from [3] to show optimality for $t \leq k^{1/3}$:

Theorem 6. *(Proof is in section 7) For every t and k satisfying $3 \le t \le k^{1/2}$, we have*

$$\mathrm{LCS}_2(t, \mathcal{P}_k) \le 4k^{1/3} + O(k^{7/40}) \qquad \text{if } t \le k^{1/3}, \tag{3}$$

$$\mathrm{LCS}_2(t, \mathcal{P}_k) \le 4t + O(t^{21/40}) \qquad \text{if } k^{1/3} \le t \le k^{1/2}. \tag{4}$$

We do not know if the constant of 4 in theorem 6 is sharp. However, the next theorem shows that, for many permutations, the constant is not 1, as the result of Beame–Huynh-Ngoc might have suggested.

Theorem 7. *(Proof is in section 8) For every t we have* $\mathrm{LCS}_2(t, \mathcal{P}_k) \ge (\frac{7}{4} - \frac{8}{t})^{1/6} k^{1/3} - 2$.

The connections between LCS_2 and LT go both ways: The next result, theorem 8 translates upper bounds on LCS_2 into upper bounds on LT. In the opposite direction, we posit a conjecture asserting a lower bound on LCS_2 "on average", and show in theorem 10 that the conjecture implies non-trivial lower bounds on LT.

Theorem 8. *(Proof is in section 5) If there exists a constant C such that for all k we have* $\mathrm{LCS}_2(2k, \mathcal{P}_k) \le Ck^{1/3}$, *then for all k we have* $\mathrm{LT}(k, n) \le 6Ck^{-2/3}n + o(n)$.

The following conjecture is a generalization of a well-known fact that the expected length of the longest increasing subsequence in a permutation sampled uniformly from $\mathcal{P}_k$ has length at least $\sqrt{k}$.

Conjecture 9. *Consider an arbitrary probability distribution on $\mathcal{P}_k$. Let π_1, π_2 be two permutations sampled independently from $\mathcal{P}_k$. Then* $\mathbb{E}\big[\mathrm{LCS}(\pi_1, \pi_2)\big] \ge \sqrt{k}$.

It might even be true that $\mathbb{E}\big[\mathrm{LCS}(\pi_1, \pi_2)\big]$ is minimized for the uniform measure on $\mathcal{P}_k$; in that case, the bound would be asymptotic to $2\sqrt{k}$ by the work of Logan–Shepp [8] and Veršik–Kerov [9]. It is straightforward to verify this stronger form of the conjecture for $k \le 3$. The authors have no idea how to approach the conjecture (or its negation).

The truth of the conjecture implies a lower bound for $\mathrm{LCS}_2(t, \mathcal{P}_k)$ for large t. Indeed, given a set of t permutations, one can apply the conjecture to the uniform probability distribution on that set. In that case, the probability that $\pi_1 = \pi_2$ is $1/t$, and so we deduce that $\frac{k}{t} + \mathrm{LCS}_2(t, \mathcal{P}_k) \ge \sqrt{k}$.

As promised, the conjecture implies an improvement on theorem 3:

Theorem 10. *(Proof is in section 6) Conjecture 9 implies that* $\mathrm{LT}(k, n)/n \ge \frac{1}{100} k^{-12/19} \log^{-8/19} k - o(1)$ *for any fixed k and large n.*

Common subsequences in small sets of multipermutations. As expected, we can improve the bound in lemma 5 by using letter repetitions:

Theorem 11. *(Proof is in section 8) Let $\pi_1, \pi_2, \pi_3 \in \mathcal{P}_{\vec{s}}$ be three $\vec{s}$-multipermutations of length $\|\vec{s}\| = n$. Then there is a pair $i < j$ such that* $\mathrm{LCS}(w_i, w_j) \ge \left(\frac{1}{6}\sum_{l\in[k]} s_l^2\right)^{1/3}$. *In other words,* $\mathrm{LCS}_2(3, \mathcal{P}_{\vec{s}}) \ge \left(\frac{1}{6}\sum_{l\in[k]} s_l^2\right)^{1/3}$.

Theorem 12. *(Proof is in section 7) The previous bound is sharp: If $\vec{s} = (s, \dots, s) \in \mathbb{Z}_+^k$ and $s \le \frac{1}{5}k$, then* $\mathrm{LCS}_2(4, \mathcal{P}_{\vec{s}}) \le (2s^2k)^{1/3} + \frac{5}{3}s + s^{4/3}k^{-1/3}$.

Surprisingly, this improvement does not hold in the setting of the classical Erdős–Szekeres theorem!

To explain the meaning of the preceding exclamation we must first recall the statement of the Erdős–Szekeres theorem. The *reverse* of a word w, denoted $\operatorname{rev} w$, is the word obtained by writing w backward. For example, $\operatorname{rev} \mathsf{abracadabra} = \mathsf{arbadacarba}$. Let

$$\mathrm{LCS}^{\mathrm{r}}(w, w') \stackrel{\text{def}}{=} \max\big(\mathrm{LCS}(w, w'), \mathrm{LCS}(w, \operatorname{rev} w')\big).$$

The Erdős–Szekeres theorem [6, p. 467] asserts that if $\pi = \mathsf{12}\ldots\mathsf{k}$, then $\mathrm{LCS}^{\mathrm{r}}(\pi, \pi') \geq \sqrt{k}$ for every $\pi' \in \mathcal{P}_k$. Since $\mathrm{LCS}^{\mathrm{r}}(\pi, \pi')$ is unchanged by relabelling the alphabet, the inequality $\mathrm{LCS}^{\mathrm{r}}(\pi, \pi') \geq \sqrt{k}$ holds for every $\pi, \pi' \in \mathcal{P}_k$. Hence, by the same reasoning as in the proof of lemma 5 it follows that $\mathrm{LCS}^{\mathrm{r}}(\pi, \pi') \geq \sqrt{n}$ for every two $\vec{s}$-multipermutations π, π' of length n. In view of theorem 11, it is quite a surprise then that this bound is sharp!

Theorem 13. *(Proof is in section 7) If $\vec{s} = (s, \ldots, s) \in \mathbb{Z}_+^k$ and $n = \|\vec{s}\|$, then there exist two $\vec{s}$-multipermutations π, π' such that $\mathrm{LCS}^{\mathrm{r}}(\pi, \pi') \leq \sqrt{n} + s$.*

Common subsequences in large sets of multipermutations. In an attempt to improve theorem 3 it is natural to combine the two approaches, and consider many multipermutations. Alas, we have been unable to extend theorem 12 or to prove better lower bounds on $\mathrm{LCS}_2(t, \mathcal{P}_{\vec{s}})$.

3 Regularity lemma and proof theorem 2

The key ingredient in the proof of theorem 1 was a regularity lemma for words. We state a version of the lemma that we need.

For a word $w \in [k]^n$ and another word u of length l that is smaller than n, we define *frequency of u in w* to be the probability that a randomly chosen l-letter-long subword of w is a copy of u. We denote the frequency by $f_w(u)$. A word $w \in [k]^n$ is (ε, L)-*regular* if whenever w' is a subword of w of length at least εn, then

$$\big|f_w(u) - f_{w'}(u)\big| < \varepsilon \qquad \text{for every } u \text{ of length at most } L.$$

Lemma 14 (Regularity lemma for words)**.** *For every $\varepsilon > 0$ and every L there is a number $M = M(\varepsilon, k, L)$ such that the following holds: Every sufficiently long word $w \in [k]^n$ can be partitioned into at most M subwords such that the total length of the subwords that are not (ε, L)-regular is at most εn.*

This lemma is slightly different than what appears in [1, Theorem 6]. The difference is that in [1] the result was asserted only for $L = 1$. However, the general case can be deduced from this special case:

Proof that the special case of $L = 1$ implies the case of general L. Assume k, ε and L are fixed. Let $\mathcal{W}$ be the set of all words in $[k]$ of length at most L. We claim that we may take $M(\varepsilon, k, L) = M(\varepsilon/2, 2^{|\mathcal{W}|}, 1)$. Indeed, given a word $w \in [k]^n$ we define a word W in the alphabet $2^{\mathcal{W}}$ via

$$W[i] \stackrel{\text{def}}{=} \{u \in \mathcal{W} : \text{the subword starting from } w[i] \text{ of length } \operatorname{len}(u) \text{ is equal to } u\}.$$

Since a subword of W that is $(\varepsilon/2, 1)$-regular corresponds to a (ε, L)-regular subword of w, the result follows from the $L = 1$ case of the regularity lemma applied to W. (The reason for $\varepsilon/2$ deteriorating into ε is the edge effect — the words that start too close to a subword boundary are miscounted.) □

For brevity, we call a word w simply *ε-regular* if it is $(\varepsilon, 1/\varepsilon)$-regular.

In context of proving the lower bounds on the twin word problem, the regularity lemma allows us to assume that the word under consideration is ε-regular, for any fixed ε. Indeed, suppose the bound of $\mathrm{LT}(u) \geq \alpha \operatorname{len}(u)$ is valid for all ε-regular words u of length exceeding n_0, and $w = w_1 \cdots w_m$ is the partition into $m \leq M$ parts described in the lemma. Then the bound $\mathrm{LT}(u) \geq \alpha\big(\operatorname{len}(u) - n_0\big)$ is valid for all ε-regular words, and so $\mathrm{LT}(w) \geq \alpha(n - mn_0) - \varepsilon n$.

We shall prove theorem 2 for $k = 3$. For a general k theorem 2 follows from the $k = 3$ case by stripping all but the three most popular letters from a word. So, we assume that $w \in [3]^n$ is ε^2-regular for some fixed, but arbitrarily small $\varepsilon > 0$. Set

$$\beta \stackrel{\text{def}}{=} \min\big(f(1), f(2), f(3)\big).$$

We may also assume that $\beta > 0.02$, for otherwise $\mathrm{LT}(w) \geq 0.49n - o(n)$ follows from theorem 1 applied to the two most-frequent letters in w.

For brevity, we write $f(u)$ in place of $f_w(u)$ throughout the remainder of this section, and also $f(u_1 + u_2 + \cdots)$ in place of $f(u_1) + f(u_2) + \cdots$.

Lemma 15. *We have* $\mathrm{LT}(w)/n \geq \frac{1}{2} f(11 + 22 + 33) - O(1/n)$.

Proof. Write $w = w_1 w_2 \cdots w_t$ be a partition of w into subwords that consist of a single letter of the alphabet. For example, 2223312222111 would be partitioned as 222 33 1 2222 111. The number of subwords is equal to

$$1 + (n-1) f(12 + 13 + 21 + 23 + 31 + 32) = n\big(1 - f(11 + 22 + 33)\big) + O(1).$$

Since $\mathrm{LT}(w_i) \geq \operatorname{len}(w_i)/2 - 1/2$, and $\mathrm{LT}(w) \geq \sum_i \mathrm{LT}(w_i)$, the lemma follows from $\sum_i \operatorname{len}(w_i) = n$. □

Lemma 16. *Define* $\alpha_1 = f(21 + 31)$, $\alpha_2 = f(12 + 32)$, $\alpha_3 = f(13 + 23)$. *For each* $l \in [3]$ *we then have*

$$\mathrm{LT}(w)/n \geq \tfrac{1}{2}\Big(1 - f(l) + \frac{\alpha_l^2}{1 - f(l)}\Big) - O(\varepsilon).$$

Proof. In view of the symmetry it suffices to prove the case $l = 3$. Let $t \stackrel{\text{def}}{=} \lfloor 1/\varepsilon \rfloor$. Partition w into t subwords of length at least εn each; say, $w = w_1 w_2 \cdots w_t$. Note that since w is ε^2-regular, each of w_i is ε-regular.

We will find twins of the following form:

	w_1	w_2	w_3	$\cdots$	w_{t-1}	w_t
First twin	1's and 3's	2's and 3's	1's and 3's	$\cdots$	2's and 3's	
Second twin		1's and 3's	2's and 3's	$\cdots$	1's and 3's	2's and 3's

So, the first twin will contain 1's and 3's, but no 2's from w_1, etc. To assure that no letter is in both twins, we adopt the following rule: Only a 3 that immediately follows a 1 (for $w_1, w_3, \dots$) or a 2 (for $w_2, w_4, \dots$) in w can appear in the first twin. Similarly, only a 3 that immediately follows a 2 (for $w_3, w_5, \dots$) or a 1 (for $w_2, w_4, \dots$) in w can appear in the second twin.

Next, we show how to find a long common subsequence between w_1 and w_2 consisting only of 1's and 3's that satisfies the restriction on 3's specified above. We will do so by first finding subsequences u_1 and u_2 of w_1 and w_2 respectively that consist only of 1's, and then adding 3's where possible.

Let r_1 and r_2 be the number of 1's in w_1 and w_2, respectively. Set $r \stackrel{\text{def}}{=} \min(r_1, r_2) - \varepsilon^2 n$. Note that $r = \big(f(1) - O(\varepsilon)\big)\varepsilon n$ by the regularity of w. Let u_1 be the subsequence of w_1 that consists of the first r occurrences of 1 in w_1. Suppose that the 1's in w_2 are at positions $i_1, i_2, \dots$. Pick an integer m uniformly at random from 0 to $\varepsilon^2 n$, and let u_2 be the subsequence $w_2[i_m], w_2[i_{m+1}], \dots, w_2[i_{m+r-1}]$.

As words, u_1 and u_2 are both words of length r that contain only 1's. As subsequences of w, they are more interesting. Of r letters that u_1 contains, $\big(f(13) - O(\varepsilon)\big)\varepsilon n$ are followed by a 3 in w_1. Say, $u_1[i]$ is followed by a 3 in w_1. Consider the letter $u_2[i]$. Due to the choice of m, the $u_2[i]$ is chosen uniformly from all 1's in an interval of length at least $\varepsilon^2 n$. So, crucially, since w_2 is ε-regular, the probability that $u_2[i]$ is followed by a 3 is $\frac{f(13)-O(\varepsilon)}{f(1)-O(\varepsilon)} = f(13)/f(1) - O(\varepsilon)$. By linearity of expectation, this implies that there is an m such that for at least

$$\frac{f(13) - O(\varepsilon)}{f(1)}\big(f(13) - O(\varepsilon)\big)\varepsilon n = \frac{f(13)^2 - O(\varepsilon)}{f(1)}\varepsilon n$$

values of i both $u_1[i]$ and $u_2[i]$ are followed by a 3. Hence, we can extend u_1 and u_2 to subsequences u_1' and u_2' of w_1 and w_2, respectively, of length at least

$$\left(f(1) + \frac{f(13)^2 - O(\varepsilon)}{f(1)}\right)\varepsilon n.$$

Similar matches can be found between w_2 and w_3, between w_3 and w_4, etc. Concatenation of these matches yields a pair of twins that are large:

$$\begin{aligned}\mathrm{LT}(w) &\geq \lfloor t/2 \rfloor \left(f(1) + \frac{f(13)^2 - O(\varepsilon)}{f(1)}\right)\varepsilon n + \lfloor (t-1)/2 \rfloor \left(f(2) + \frac{f(23)^2 - O(\varepsilon)}{f(2)}\right)\varepsilon n \\ &= \tfrac{1}{2}\left(f(1) + f(2) + \frac{f(13)^2}{f(1)} + \frac{f(23)^2}{f(2)} - O(\varepsilon)\right) n,\end{aligned}$$

which by the Cauchy–Schwarz inequality applied to the vectors $\big(\frac{f(13)}{\sqrt{f(1)}}, \frac{f(23)}{\sqrt{f(2)}}\big)$ and $\big(\sqrt{f(1)}, \sqrt{f(2)}\big)$ implies that

$$2\,\mathrm{LT}(w)/n \geq f(1) + f(2) + \frac{\alpha_3^2}{f(1) + f(2)} - O\big(\varepsilon\big).$$

Since $f(1) + f(2) = 1 - f(3)$, the proof of the lemma is complete. □

The preceding two lemmas are enough to deduce theorem 2. Indeed, applying lemma 16 for

$l = 1, 2, 3$ and adding the resulting bounds we obtain

$$\begin{aligned} 6\,\mathrm{LT}(w)/n &\geq 3 - f(1+2+3) + \frac{\alpha_1^2}{1-f(1)} + \frac{\alpha_2^2}{1-f(2)} + \frac{\alpha_3^2}{1-f(3)} - O(\varepsilon) \\ &= 2 + \frac{\alpha_1^2}{1-f(1)} + \frac{\alpha_2^2}{1-f(2)} + \frac{\alpha_3^2}{1-f(3)} - O(\varepsilon) \\ &\geq 2 + \tfrac{1}{2}(\alpha_1+\alpha_2+\alpha_3)^2 - O(\varepsilon), \end{aligned}$$

where the last line follows from applying Cauchy–Schwarz inequality to vectors $\left(\frac{\alpha_1}{\sqrt{1-f(1)}}, \frac{\alpha_2}{\sqrt{1-f(2)}}, \frac{\alpha_3}{\sqrt{1-f(3)}}\right)$ and $\left(\sqrt{1-f(1)}, \sqrt{1-f(2)}, \sqrt{1-f(3)}\right)$, and then using $f(1+2+3) = 1$ to simplify the resulting expression. Since $\alpha_1+\alpha_2+\alpha_3 = 1-f(11+22+33)$, in view of the bound from lemma 15 we conclude that

$$\mathrm{LT}(w)/n \geq \min_x \max(\tfrac{1}{3} + \tfrac{1}{12}x^2, \tfrac{1}{2} - \tfrac{1}{2}x) - O(\varepsilon) = \frac{4-\sqrt{11}}{2} - O(\varepsilon).$$

Since $\frac{4-\sqrt{11}}{2} > \frac{1}{3} \cdot 1.02$ and ε is arbitrary, theorem 2 follows.

4 Proof of theorem 4

We will show that with high probability a random word of length n satisfies the conclusion of theorem 4. Recall that $w[i]$ denotes the i'th letter of the word w.

Twins w_1 and w_2 in $w \in [k]^n$ are said to be *monotone* if $w_1[i]$ precedes $w_2[i]$ in w for all i.

Lemma 17. *If* $\mathrm{LT}(w) \geq m$, *then* w *contains monotone twins of length* m.

Proof. The condition implies that w contains twins of length m. However, if $w[p_1]\cdots w[p_m]$ and $w[p'_1]\cdots w[p'_m]$ are twins, then so are $w[\bar{p}_1]\cdots w[\bar{p}_m]$ and $w[\bar{p}'_1]\cdots w[\bar{p}'_m]$, where $\bar{p}_i = \min(p_i, p'_i)$ and $\bar{p}'_i = \max(p_i, p'_i)$. □

To each pair (w_1, w_2) of monotone twins in w we associate a word $R(w_1, w_2, w) \in \{0,1,2\}^n$ by the rule

$$R(w_1,w_2,w)[i] = \begin{cases} 0 & \text{if } w[i] \text{ is neither in } w_1 \text{ nor in } w_2, \\ 1 & \text{if } w[i] \text{ is in } w_1, \\ 2 & \text{if } w[i] \text{ is in } w_2. \end{cases}$$

The word $R(w_1, w_2, w)$ records the *roles* of letters of w in the pair (w_1, w_2). For example, consider the monotone pair $1\overline{0}\underline{0}\overline{1}\overline{1}\underline{1}0\underline{1}\overline{1}\underline{1}01$, where the overlines indicate the letters in w_1 and the underlines indicate the letters in w_2. For this pair, $R = 012112021200$.

A monotone pair (w_1, w_2) in w is said to be *regular* if the following two conditions hold:

a) There exist no two numbers $i < j$ satisfying the following: $R[i] = 2$ and $R[j] = 1$, and $R[k] = 0$ for all $i < k < j$, and $w[i] = w[j]$.

b) There exist no two numbers $i < j$ satisfying the following: $R[i] \in \{1,2\}$ and $R[j] = 0$, and $R[k] = 0$ for all $i < k < j$, and $w[i] = w[j]$.

We can express these conditions in the overline/underline notation: The condition (a) forbids the pattern $\underline{\mathsf{x}}???\overline{\mathsf{x}}$, whereas the condition (b) forbids the patterns $\overline{\mathsf{x}}???\mathsf{x}$ and $\underline{\mathsf{x}}???\mathsf{x}$, where the question marks denote letters that are not in the twins.

Lemma 18. *If* $\mathrm{LT}(w) \geq m$*, then* w *contains regular twins of length* m*.*

Proof. Pick monotone twins (w_1, w_2) in w such that $R(w_1, w_2, w)$ is lexicographically minimal. Then (w_1, w_2) is regular. Indeed, if (w_1, w_2) were not regular, then swapping the roles of $w[i]$ and $w[j]$ would lead to monotone twins with a lexicographically smaller value of R. □

Let $\mathcal{R}^n_m$ consist of all words $R \in \{\mathsf{0}, \mathsf{1}, \mathsf{2}\}^n$ in which letter $\mathsf{1}$ and letter $\mathsf{2}$ occur m times each. For $R \in \{\mathsf{0}, \mathsf{1}, \mathsf{2}\}^n$ let $p(R)$ be the number of occurrences of the pattern $\mathsf{20^*1}$, i.e., a $\mathsf{2}$ followed by zero or more $\mathsf{0}$'s, and then followed by a $\mathsf{1}$. Also, let $z(R)$ be the length of the longest prefix of R that contains only $\mathsf{0}$'s. Let $\mathcal{R}^n_{m,p,z} \stackrel{\text{def}}{=} \{R \in \mathcal{R}^n_m : p(R) = p,\ z(R) = z\}$. For example, $\mathsf{012112021200} \in \mathcal{R}^{12}_{4,2,1}$.

Let $\mathcal{M} \subset \{\mathsf{0}, \mathsf{1}, \mathsf{2}\}^n$ be the set of all words in which every prefix contains at least as many $\mathsf{1}$'s as $\mathsf{2}$'s. Note $R(w_1, w_2, w) \in \mathcal{M}$ for every pair of monotone twins. For $R \in \mathcal{M}$ let B_R be the event that a word w chosen uniformly at random from $[k]^n$ contains a regular pair (w_1, w_2) satisfying $R(w_1, w_2, w) = R$.

Lemma 19. *Suppose* $R \in \mathcal{R}^n_{m,p,z} \cap \mathcal{M}$*, then* $\Pr[B_R] = (1/k)^m (1 - 1/k)^{p+n-2m-z}$*.*

Proof. Imagine that each letter $w[i]$ of w is in its own box, which is labeled $R[i]$. Call a box labeled with $\mathsf{0}$ a *0-box*, etc. Imagine also that boxes are connected by red and blue wires according to the following rules:

- The i'th box labeled $\mathsf{1}$ is connected by a blue wire to the i'th box labeled $\mathsf{2}$.
- If the first non-zero-labeled box that precedes a $\mathsf{1}$-box is a $\mathsf{2}$-box, then the $\mathsf{1}$-box and the $\mathsf{2}$-box are connected by a red wire.
- Each $\mathsf{0}$-box is connected with the preceding non-zero-labeled box by a red wire.

The following condition is clear from the definition of a regular pair:

> A word w contains a regular pair (w_1, w_2) satisfying $R(w_1, w_2, w) = R$ if and only if the blue wires connect the boxes containing the same letters, and red wires connect the boxes containing different letters.

The boxes start closed, and we open them one by one. Each time we open a box we look at the letter that it contains, and at the letters of all previously-opened boxes that this box is connected to. If the condition above is violated, we abort.

The order for opening the boxes is not arbitrary: we first open all non-zero boxes from left to right, and only then open the $\mathsf{0}$-boxes. This order ensures that each time we open a box, there is at most one wire that connects it to a previously-opened box. Thus, the probability that we abort is $1 - 1/k$ if that the wire is blue, and $1/k$ if the wire is red.

Since the number of blue wires is m, and the number of red wires is $p + (n - 2m - z)$, the lemma follows. □

Lemma 20. *The size of* $\mathcal{R}^n_{m,p,z}$ *is* $|\mathcal{R}^n_{m,p,z}| = \binom{n-z}{2m}\binom{m}{p}^2$.

Proof. To each word $R \in \mathcal{R}^n_{m,p,z}$ associate the word $R' \in \mathcal{R}^{2m}_{m,p,0}$ obtained by removing all 0's from R. Since R starts with a prefix of z 0's, one can recover R from R' by specifying the positions of the remaining 0's. So, $|\mathcal{R}^n_{m,p,z}| = \binom{n-z}{2m}|\mathcal{R}^{2m}_{m,p,0}|$, and to complete the proof it suffices to compute $|\mathcal{R}^{2m}_{m,p,0}|$.

Every $R \in \mathcal{R}^{2m}_{m,p,0}$ is necessarily of the form

$$\square 21 \square 21 \square 21 \cdots 21 \square 21 \square,$$

where there are $p+1$ bins separated by the p occurrences of 21, and each bin contains a subword of the form 1^*2^* (the subword might be empty). The bins contain a total of $m-p$ 1's and the same number of 2's. There are $\binom{(m-p)+(p+1)-1}{(p+1)-1} = \binom{m}{p}$ ways of placing $m-p$ identical objects into $p+1$ labeled bins. In particular, there are $\binom{m}{p}$ ways to place $m-p$ 0's into $p+1$ bins, and the same number of ways to place $m-p$ 1's into the bins. Since the choices for placement of 0's and 1's are independent, we conclude that $|\mathcal{R}^{2m}_{m,p,0}| = \binom{m}{p}^2$ as promised. □

Let $m = \alpha n$ for the constant α from theorem 4. The union bound and the three preceding lemmas imply that

$$\begin{aligned}
\Pr[\mathrm{LT}(w) \geq m] &\leq \sum_{p,z} \sum_{R \in \mathcal{R}^n_{m,p,z} \cap \mathcal{M}} \Pr[B_R] \leq \sum_{p,z \geq 0} \binom{n-z}{2m} \binom{m}{p}^2 (1/k)^m (1-1/k)^{p+n-2m-z} \\
&= (1/k)^m (1-1/k)^{n-2m} \sum_p \binom{m}{p}^2 (1-1/k)^p \sum_{z \geq 0} \binom{n-z}{2m} (1-1/k)^{-z} \\
&\leq (1/k)^m (1-1/k)^{n-2m} n^2 \max_p \binom{m}{p}^2 (1-1/k)^p \max_{z \geq 0} \binom{n-z}{2m} (1-1/k)^{-z}.
\end{aligned}$$

Let $f(p) = \binom{m}{p}^2 (1-1/k)^p$ and $g(z) = \binom{n-z}{2m}(1-1/k)^{-z}$. Since $g(z+1)/g(z) = \frac{n-z-2m}{n-z} \cdot \frac{k}{k-1}$ and $m \geq n/2k$ the maximum of $g(z)$ is attained at $z = 0$. Similarly, $f(p+1)/f(p) = \left(\frac{m-p}{p+1}\right)^2 (1-1/k)$ implies that the maximum of $f(p)$ is attained when $p = \frac{m}{1+(1-1/k)^{-1/2}} + O(1)$. We plug these into the displayed formula above, use the asymptotic formula $\frac{\log \binom{n}{\beta n}}{n} = \beta \log \frac{1}{\beta} + (1-\beta) \log \frac{1}{1-\beta} + o(1)$, and simplify to obtain:

$$\frac{\log \Pr[\mathrm{LT}(w) \geq \alpha n]}{n} \leq (1-2\alpha) \log \frac{1}{1-2\alpha} - \alpha \log(\alpha^2 k) - 2\alpha \log \left(\frac{2}{1+\sqrt{1-1/k}} \right) + (1-2\alpha) \log(1-1/k) + o(1).$$

Whenever the expression on the right is negative, we have $\Pr[\mathrm{LT}(w) \geq \alpha n] < 1$ for large enough n, and so $\mathrm{LT}(n) < \alpha n$ for those n.

We note that the first two terms in the inequality above are the same as in the bound obtained in [1]. The last two terms are new to our analysis.

The bounds on $\mathrm{LT}(4, n)$, $\mathrm{LT}(5, n)$ and the asymptotic bound on $\mathrm{LT}(k, n)$ for large k in theorem 4 were obtained using the MATHEMATICA software package. The code that we used is available at http://www.borisbukh.org/code/twins_lcs13.html.

5 Proof of theorem 8

Lemma 21. *Suppose k, K, A, B are natural numbers that satisfy $\mathrm{LT}(K, n) \le A$ and $\mathrm{LCS}_2(K, \mathcal{P}_k) \le B$. Then $\mathrm{LT}(k, kn) \le (2n-1)B + kA$.*

Proof. Let $w \in [K]^n$ be a word such that $\mathrm{LT}(w) \le A$. Let $\pi_1, \dots, \pi_K \in \mathcal{P}_k$ be a family of K permutations such that $\mathrm{LCS}(\pi_i, \pi_j) \le B$ for all distinct i, j.

Replace each letter l in w by the permutation π_l to obtain a word $w' \in [k]^{kn}$. We claim that $\mathrm{LT}(w') \le (2n-1)B + kA$. Indeed, suppose w'_1 and w'_2 are twins in w'. Let $\mathcal{T}$ be the set of all pairs (i', j') such that $w'[i']$ is a letter in w'_1 that is matched to $w'[j']$, which is in w'_2. Let $\mathcal{H} \stackrel{\text{def}}{=} \{(\lfloor i'/k \rfloor, \lfloor j'/k \rfloor) : (i', j') \in \mathcal{T}\}$. Let $\mathcal{M} \stackrel{\text{def}}{=} \{(i, j) \in \mathcal{H} : w[i] = w[j]\}$. We may also suppose that the twins w'_1, w'_2 were chosen among all twins of their length so that $\sum_{i,j \in \mathcal{H}} (i + j)$ is minimized.

A simple picture explains the meaning of $\mathcal{T}$ and $\mathcal{H}$ just defined. Imagine a kn-by-kn square that is subdivided into n^2 k-by-k squares, and imagine that the kn letters of w' are laid along each of the axes. The intervals of length k on the axes correspond to the original letters of w. For each pair of matching letters of w'_1 and w'_2 draw a corresponding point — these will be points of $\mathcal{T}$. The set of k-by-k squares that are hit by $\mathcal{T}$ is the set $\mathcal{H}$. The squares in $\mathcal{H} \setminus \mathcal{M}$ cannot contain more than B points each since they correspond to different permutations. The points of $\mathcal{T}$ form a graph of a monotone function, and so $|\mathcal{H}| \le 2n - 1$. Hence, at most $(2n-1)B$ points of $\mathcal{T}$ fall into $\mathcal{H} \setminus \mathcal{M}$.

It remains to bound $|\mathcal{M}|$. We claim that no two squares in $\mathcal{M}$ share an x-coordinate, and no two squares share a y-coordinate. Indeed, suppose $(i, j_1), (i, j_2) \in \mathcal{M}$ with $j_1 < j_2$. Then by moving all the points from the square indexed by (i, j_2) to the one indexed by (i, j_1) we would obtain a pair of twins of same length, but with the smaller value of $\sum_{i,j \in \mathcal{H}} (i + j)$. Since this contradicts the minimality assumption, the claim follows. Put $w_1 = \{w[i] : (i, j) \in \mathcal{M}\}$ and $w_2 = \{w[j] : (i, j) \in \mathcal{M}\}$. The preceding claim implies that w_1 and w_2 are twins in w, and so $|\mathcal{M}| \le A$. Hence, the number of points of $\mathcal{T}$ in $\mathcal{M}$ is at most kA. Together with the estimate on points $\mathcal{T}$ in $\mathcal{H} \setminus M$, this implies that $|\mathcal{T}| \le (2n-1)B + kA$. Since $\mathrm{LT}(w') = |\mathcal{T}|$, the proof is complete. □

Suppose the condition of theorem 8 holds, and C is a constant as in the theorem. Let α_k be the least real number such that $\mathrm{LT}(k, n) \le \alpha_k n + o(n)$. The preceding lemma implies that $\mathrm{LT}(k, kn) \le 2Cnk^{1/3} + k\,\mathrm{LT}(2k, n)$, and so $\alpha_k \le 2Ck^{-2/3} + \alpha_{2k}$. Thus,

$$\alpha_k \le 2Ck^{-2/3} + 2C(2k)^{-2/3} + \cdots + 2C(2^{t-1}k)^{-2/3} + \alpha_{2^t k}.$$

As $t \to \infty$, the last term tends to 0 by inequality (2), and the sum of the remaining terms tends to $\frac{2C}{1-2^{-2/3}} k^{-2/3} \le 6Ck^{-2/3}$.

6 Proof of theorem 10

Throughout this section we employ the following notation. For a word w and a letter l we write $l \in w$ if the letter l occurs in w.

Lemma 22. *Suppose $\bar{u} = u_1 \cdots u_r$ and $\bar{u}' = u'_1 \cdots u'_r$ are words that are the concatenation of r other words. Then $\mathrm{LCS}(\bar{u}, \bar{u}') \le \sum_{i,j} \mathrm{LCS}(u_i, u'_j)$.*

Proof. A common sequence of $\bar{u}$ and $\bar{u}'$ can be broken up as a concatenation of common sequences of u_i and u_j over various i, j. □

Proof of theorem 10. In view of the discussion following the regularity lemma in section 3, it suffices to prove $\mathrm{LT}(w)/n \geq \frac{1}{100}k^{-12/19}\log^{-8/9} k - o(1)$ for all ε-regular words $w \in [k]^n$. So, we assume that $w \in [k]^n$ is given, and that it is ε-regular.

Let $m = n/12k$. Pick an integer r from the interval $[0, n/4]$ uniformly at random, and pick another integer r' from the interval $[n/2, 3n/4]$ uniformly at random independently from r. Starting from the position r partition w into m intervals of length $3k$ each. Note that the total length of these intervals is $n/4$, and so these intervals are completely contained in the first half of w. Let $w_1, \ldots, w_m \in [k]^{3k}$ be the subwords in these intervals. Similarly, starting from the position r' define m subwords $w'_1, \ldots, w'_m \in [k]^{3k}$ that are completely contained in the second half of w.

Since w is ε-regular for every word $u \in [k]^{3k}$ and every $i \in [m]$ we have $\Pr[w_i = u] = \mu(\{u\}) + O(\varepsilon)$, and $\Pr[w'_i = u] = \mu(\{u\}) + O(\varepsilon)$. Here, μ is a probability measure on $[k]^{3k}$. Furthermore, w_i and w'_i are independent. Note that

$$\mathrm{LT}(w) \geq \sum_{i=1}^{m} \mathrm{LT}(w_i) \qquad \text{and} \qquad \mathrm{LT}(w) \geq \sum_{i=1}^{m} \mathrm{LCS}(w_i, w'_i). \tag{5}$$

We shall show that at least one of these two bounds is large.

Let $0 < \alpha < 1/4$ be a parameter to be chosen later. Put

$$S \stackrel{\text{def}}{=} \{u \in [k]^{3k} : \text{ fewer than } \alpha k \text{ distinct letters occur in } u\}.$$

We distinguish two cases, depending on $\mu(S)$.

a) Suppose $\mu(S) \geq \frac{1}{2}$. If $u \in S$ is arbitrary, then treating u as a word over an αk-letter alphabet, and using theorem 3 we conclude that

$$\mathrm{LT}(u) \geq 3^{-4/3}(\alpha k)^{-2/3}3k - 3^{-1/3}(\alpha k)^{1/3} = 3^{-1/3}\alpha^{-2/3}(1-\alpha)k^{1/3} \geq \tfrac{1}{2}\alpha^{-2/3}k^{1/3}.$$

Thus, $\mathrm{LT}(w) \geq \mathbb{E}\big[\sum_i \mathrm{LT}(w_i)\big] \geq m \cdot \big(\mu(S) - O(\varepsilon)\big)\frac{1}{2}\alpha^{-2/3}k^{1/3} \geq \frac{1}{48}\alpha^{-2/3}k^{-2/3} - O(\varepsilon n)$.

b) Suppose $\mu(\bar{S}) \geq \frac{1}{2}$. Let u and u' be two words sampled independently from $[k]^{3k}$ according to measure μ. Let $L = \{l \in [k] : \Pr[l \in u] \geq \alpha/4\}$. Since $\mu(\bar{S}) \geq \frac{1}{2}$, for random $l \in [k]$ we have $\Pr_{l,u}[l \in u] \geq \alpha/2$ and so it follows that $|L| \geq \alpha k/4$. For a word v that contains every letter of L at least once denote by $\pi_L(v)$ the permutation on the alphabet L obtained by taking in v the first occurrence of each letter. If there is a letter of L that does not occur in v, we let $\pi_L(v) = \emptyset$, where $\emptyset$ is the empty word.

Let $r \stackrel{\text{def}}{=} 4\alpha^{-1}\log(4k)$. Let $u_1, \ldots, u_r$ and $u'_1, \ldots, u'_r$ be $2r$ words sampled independently from $[k]^{3k}$ according to measure μ. Let $\bar{u} = u_1 \cdots u_r$ and $\bar{u}' = u'_1 \cdots u'_r$. Put $\pi = \pi_L(\bar{u})$ and $\pi' = \pi_L(\bar{u}')$. For each fixed $l \in L$, we have $\Pr[l \notin \bar{u}] \leq (1-\alpha/4)^r \leq 1/4k$. Hence, $\Pr[\pi = \emptyset] \leq 1/4$. Similarly $\Pr[\pi' = \emptyset] \leq 1/4$. By conjecture 9 it follows that

$$\begin{aligned}\mathbb{E}\big[\mathrm{LCS}(\pi, \pi')\big] &= \mathbb{E}\big[\mathrm{LCS}(\pi, \pi') | \pi \neq \emptyset \wedge \pi' \neq \emptyset\big] \Pr[\pi \neq \emptyset \wedge \pi' \neq \emptyset] \\ &\geq \tfrac{1}{2}\sqrt{|L|} \geq \tfrac{1}{4}\sqrt{\alpha k}.\end{aligned}$$

By lemma 22 and the linearity of expectation we have

$$r^2\mathbb{E}\big[\mathrm{LCS}(u,u')\big] = \sum_{i,j\in[r]} \mathbb{E}\big[\mathrm{LCS}(u_i,u'_j)\big] \geq \mathbb{E}\big[\mathrm{LCS}(\bar{u},\bar{u}')\big].$$

Since $\mathrm{LCS}(\bar{u},\bar{u}') \geq \mathrm{LCS}(\pi,\pi')$, we can combine the two preceding inequalities with (5) to obtain

$$\mathrm{LT}(w) \geq \mathbb{E}\Big[\sum_{i=1}^{m} \mathrm{LCS}(w_i,w'_i)\Big] \geq m\cdot\big(\tfrac{1}{r^2}\cdot\tfrac{1}{4}\sqrt{\alpha k} - O(\varepsilon)\big) \geq \tfrac{1}{1000}\alpha^{5/2}\frac{k^{-1/2}}{\log^2 k}n - O(\varepsilon n).$$

Thus no matter which of the two cases holds, we have

$$\frac{\mathrm{LT}(w)}{n} \geq \min\big(\tfrac{1}{48}\alpha^{-2/3}k^{-2/3}, \tfrac{1}{1000}\alpha^{5/2}\tfrac{k^{-1/2}}{\log^2 k}\big) - O(\varepsilon).$$

Since $\varepsilon > 0$ is arbitrary, setting $\alpha = 3k^{-1/19}\log^{12/19} k$ yields the desired result. □

7 Constructions

In this section we prove theorems 6, 12 and 13. All of these results are proved by exhibiting explicit constructions of families of (multi)permutations with required properties. The constructions are very similar to one another, and so we start by describing their commonality.

In all the constructions the alphabet is $X\times Y\times\cdots$ for some sets $X, Y, \dots$. This way each letter l can be thought of as a vector $(x, y, \dots)$. Given an injective map $f\colon X\times Y\times\cdots\to\mathbb{Z}\times\cdots\times\mathbb{Z}$ we define π_f to be the permutation in which letter l precedes letter l' if $f(l)$ is lexicographically smaller than $f(l')$. As a notation, we write $[\![f_1\ f_2\ \dots]\!]$ for the permutation associated to the function $f = (f_1, f_2, \dots)$. For example, $\mathrm{rev}[\![f_1\ f_2\ \dots]\!] = [\![-f_1\ -f_2\ \dots]\!]$.

As another example, if $X = Y = Z = [n]$ the four permutations

$$\begin{aligned}
\pi_1 &= [\![\ \ x\ \ \ y\ \ \ z]\!],\\
\pi_2 &= [\![-x\ \ -y\ \ \ z]\!],\\
\pi_3 &= [\![-x\ \ \ y\ \ -z]\!],\\
\pi_4 &= [\![\ \ x\ \ -y\ \ -z]\!].
\end{aligned}$$

satisfy $\mathrm{LCS}(\pi_i,\pi_j) \leq n$ for $i\neq j$. Indeed, in any sequence $(x_1,y_1,z_1),(x_2,y_2,z_2),\dots$ that is increasing in both π_i and π_j only one of the coordinates can change. (This example is from [4]).

Proof of theorem 6. For $x\in\mathbb{Z}/p\mathbb{Z}$ let $\overline{x}$ be the element of $\{0,1,\dots,p-1\}$ that is congruent to x. Let $X = Y = Z = \mathbb{Z}/p\mathbb{Z}$, and for each $i\in\mathbb{Z}/p\mathbb{Z}$ define the permutation

$$\pi_i = \Big[\!\!\Big[\overline{i^2x+iy+z}\quad \overline{2ix+y}\quad \overline{x}\Big]\!\!\Big].$$

We claim that $\mathrm{LCS}(\pi_i,\pi_j)\leq 4p-2$ for all $i\neq j$. Indeed, suppose w is a common subsequence of π_i and π_j. Say, w is the sequence $(x_1,y_1,z_1),(x_2,y_2,z_2),\dots\in(\mathbb{Z}/p\mathbb{Z})^3$. For $(I,J)\in\{0,1,\dots,p-1\}^2$ put

$$B(I,J) \stackrel{\text{def}}{=} \big\{(x,y,z)\in(\mathbb{Z}/p\mathbb{Z})^3 : \overline{i^2x+iy+z} = I,\ \overline{j^2x+jy+z} = J\big\}.$$

If $(x, y, z) \in B(I, J)$ we say that (x, y, z) is in the *bin* (I, J). Let $(I_1, J_1), (I_2, J_2), \dots$ be the sequence of bins for $(x_1, y_1, z_1), (x_2, y_2, z_2), \dots$. It is clear that $I_1 \le I_2 \le \cdots$ and $J_1 \le J_2 \le \cdots$, and so at most $2p - 1$ bins are occupied. Hence, it suffices to prove that no more than 2 letters of w fall into a same bin.

Since $i \neq j$, the two defining equations of $B(I, J)$ are linearly independent, and so the set $B(I, J)$ is a line in $(\mathbb{Z}/p\mathbb{Z})^3$. A computation yields that, for fixed (I, J), the set $B(I, J)$ is of the form $\{(x_0, y_0, z_0) + t(1, -i - j, ij) : t \in \mathbb{Z}/p\mathbb{Z}\}$ for some (x_0, y_0, z_0). Thus, the set

$$\Big\{\big(2ix + y, 2jx + y\big) : (x, y, z) \in B(I, J)\Big\}$$

is the line $\Big\{\big(2ix_0 + y_0 + (i - j)t, 2jx_0 + y_0 + (j - i)t\big) : t \in \mathbb{Z}/p\mathbb{Z}\Big\}$. As this line has slope -1, the set

$$\Big\{\big(\overline{2ix + y}, \overline{2jx + y}\big) : (x, y, z) \in B(I, J)\Big\}$$

is a union of at most two line segments of slope -1. Since the sequence w is increasing in both π_i and π_j, it follows that w can contain at most two points from $B(I, J)$. Hence, $\mathrm{LCS}_2(p, \mathcal{P}_{p^3}) \le 4p - 2$. The inequality (3) follows by choosing p to be the smallest prime exceeding $k^{1/3}$, and noting that $p \le k^{1/3} + O(k^{7/40})$ by [2].

To derive the inequality (3) we start with a family of t permutations over a t^3-letter alphabet such that LCS of any two is at most $4t + O(t^{21/40})$. We then select some k letters of the alphabet, and delete all the remaining letters from all the permutations. □

We extend the $[\![\cdots]\!]$ notation to the multipermutations. The alphabet is still $X \times Y \times \cdots$ for some sets $X, Y, \dots$. There is now an additional set R, and to each injective function $f\colon X \times Y \times \cdots \times R \to \mathbb{Z} \times \cdots \times \mathbb{Z}$ we associate a multipermutation in which each letter occurs $|R|$ times, as follows. The set R indexes copies of the same letter in π_f. The occurrence of letter l indexed by $r \in R$ precedes the occurrence of letter l' indexed by r' if $f(l, r)$ is lexicographically smaller than $f(l', r')$.

Proof of theorem 13. Let $X = [k_1]$, $Y = [k_2]$ and $R = [s]$. Consider the following two multipermutations:

$$\begin{aligned} \pi &= [\![x \quad y \quad\ \ r]\!], \\ \pi' &= [\![x \quad r \quad -y]\!]. \end{aligned}$$

We first bound the $\mathrm{LCS}(\pi, \pi')$. Suppose $(x_1, y_1, r_1), (x_2, y_2, r_2), \dots$ and $(x_1, y_1, r'_1), (x_2, y_2, r'_2), \dots$ are subsequences of π and π', respectively, that are equal as words. Then, for each i, either $x_{i+1} > x_i$, or $x_{i+1} = x_i$ and $r_{i+1} > r_i$. Hence $\mathrm{LCS}(\pi, \pi') \le k_1 s$.

We next bound $\mathrm{LCS}(\pi, \mathrm{rev}\,\pi')$. Consider a pair of sequences as above. For each i we must have $x_{i+1} = x_i$, and also $y_{i+1} \ge y_i$ and $r'_{i+1} \le r'_i$ with at least one of the inequalities being strict. Hence $\mathrm{LCS}(\pi, \mathrm{rev}\,\pi') \le k_2 + s - 1$.

Given any $n = ks$, let k_1 be the closest integer to $\sqrt{k/s} + \frac{1}{2}$, and let $k_2 = \lceil k/k_1 \rceil$. One can then verify that $k_1 s \le \sqrt{ks} + s$ and $k_2 + s \le \lceil \sqrt{ks} \rceil + s \le \sqrt{ks} + s + 1$. □

Proof of theorem 12. With hindsight let $X = [k_1]$, $Y = [k_2]$ and $Z = [k_3]$, where k_2 is the closest integer to $(k/4s)^{1/3} + \frac{1}{3}$, and $k_1 = 2k_2$, and $k_3 = \lceil k/k_1k_2 \rceil$. Put $R = [s]$. Consider the following four multipermutations:

$$\begin{aligned}
\pi_1 &= [\![\ \ x \quad\ \ y \quad\ \ z \quad\ \ r]\!],\\
\pi_2 &= [\![-x \quad -y \quad\ \ r \quad\ \ z]\!],\\
\pi_3 &= [\![-x \quad\ \ y \quad -z \quad\ \ r]\!],\\
\pi_4 &= [\![\ \ x \quad -y \quad\ \ r \quad -z]\!].
\end{aligned}$$

To speed up the computation of LCS we observe that $\mathrm{LCS}\left(\begin{smallmatrix}[\![x & f]\!],\\ [\![x & g]\!]\end{smallmatrix}\right) = |X|\,\mathrm{LCS}\left(\begin{smallmatrix}[\![f]\!],\\ [\![g]\!]\end{smallmatrix}\right)$ and that $\mathrm{LCS}\left(\begin{smallmatrix}[\![\ x & f]\!],\\ [\![-x & g]\!]\end{smallmatrix}\right) = \mathrm{LCS}\left(\begin{smallmatrix}[\![f]\!],\\ [\![g]\!]\end{smallmatrix}\right)$. We thus have

$$\begin{aligned}
\mathrm{LCS}(\pi_1,\pi_2) &= \mathrm{LCS}\left(\begin{smallmatrix}[\![z & r]\!],\\ [\![r & z]\!]\end{smallmatrix}\right) = |Z| + s - 1,\\
\mathrm{LCS}(\pi_1,\pi_4) &= |X|\,\mathrm{LCS}\left(\begin{smallmatrix}[\![z & \ \ r]\!],\\ [\![r & -z]\!]\end{smallmatrix}\right) = |X|s,\\
\mathrm{LCS}(\pi_1,\pi_3) &= |Y|\,\mathrm{LCS}([\![r]\!],[\![r]\!]) = |Y|s,\\
\mathrm{LCS}(\pi_2,\pi_4) &= |Y|\,\mathrm{LCS}\left(\begin{smallmatrix}[\![r & \ \ z]\!],\\ [\![r & -z]\!]\end{smallmatrix}\right) = |Y|(2s-1).
\end{aligned}$$

The derivation of the equality $\mathrm{LCS}(\pi_3,\pi_4) = |Z| + s - 1$ is same as of $\mathrm{LCS}(\pi_1,\pi_2)$, whereas the proof of $\mathrm{LCS}(\pi_2,\pi_3) = |X|s$ is analogous to that of $\mathrm{LCS}(\pi_1,\pi_4)$.

With the choice of k_1, k_2, k_3 made above we have $(2s-1)|Y| \le |X|s \le (2s^2k)^{1/3} + \frac{5}{3}s$, and $|Z| + s - 1 \le k/k_1k_2 + s \le k/2\big((k/4s)^{1/3} - 1/6\big)^2 = (2s^2k)^{1/3} + \frac{5}{3}s + \frac{(1458k)^{1/3} - 2s^{1/3}}{((54k)^{1/3} - s^{1/3})^2} \cdot \frac{s^{4/3}}{3}$, which is at most $(2s^2k)^{1/3} + \frac{5}{3}s + s^{4/3}k^{-1/3}/3$ for $s \le \frac{1}{5}k$. □

8 Lower bounds

In this section we prove theorem 7 which gives the lower bound on $\mathrm{LCS}_2(t, \mathcal{P}_k)$ for large t, and theorem 11 that shows how to take advantage of repetitions in multipermutations.

Proof of theorem 7. Suppose that $\{\pi_1, \ldots, \pi_t\}$ is any set of t permutations, and put $L_{i,j} \stackrel{\text{def}}{=} \mathrm{LCS}(\pi_i, \pi_j)$. For a letter $l \in [k]$ and $\pi \in \mathcal{P}_k$ let $\pi\{l\}$ be the prefix of π that ends with l. For $i < j$ define a function $f_{i,j}\colon [k] \to [L_{i,j}]$ by $f_{i,j}(l) = \mathrm{LCS}\big(\pi_i\{l\}, \pi_j\{l\}\big)$.

We say that a pair of letters $\{l, l'\}$ is *nice* to the triple $i_1 < i_2 < i_3$ if the three differences $f_{i_1,i_2}(l) - f_{i_1,i_2}(l')$, $f_{i_1,i_3}(l) - f_{i_1,i_3}(l')$, and $f_{i_2,i_3}(l) - f_{i_2,i_3}(l')$ are either all negative, or all non-negative.

Observation. If l and l' are distinct, then there are at least $\big(\frac{7}{16} - \frac{2}{t}\big)\binom{t}{3}$ triples $i_1 < i_2 < i_3$ for which $\{l, l'\}$ are nice.

Proof of the observation. Consider the complete graph on the vertex set $[t]$. Color its edge $i < j$ red if $f_{i,j}(l) < f_{i,j}(l')$ and blue if $f_{i,j}(l) \ge f_{i,j}(l')$. A triple $i_1 < i_2 < i_3$ is nice to $\{l, l'\}$ if $i_1i_2i_3$ is a monochromatic triangle in the coloring. Let $\mathcal{A} = \{i : l \text{ precedes } l' \text{ in } \pi_i\}$ and $\mathcal{B} = \{i : l' \text{ precedes } l \text{ in } \pi_i\}$. Note that all the edges in $\mathcal{A}$ are red, whereas all the edges in $\mathcal{B}$ are blue.

For an $i \in \mathcal{A}$ let $d(i)$ be the number of blue edges connecting i to $\mathcal{B}$. For a $j \in \mathcal{B}$ let $d(j)$ be the number of red edges connecting j to $\mathcal{A}$. Finally let M be the number of monochromatic triangles in our complete graph. Since $|\mathcal{A}||\mathcal{B}| = \sum_{i\in\mathcal{A}} d(i) + \sum_{j\in\mathcal{B}} d(j)$ and $x \mapsto \binom{x}{2}$ is convex, it follows that

$$\begin{aligned} M &= \binom{|\mathcal{A}|}{3} + \binom{|\mathcal{B}|}{3} + \sum_{i\in\mathcal{A}} \binom{d(i)}{2} + \sum_{j\in\mathcal{B}} \binom{d(j)}{2}. \\ &\geq \binom{|\mathcal{A}|}{3} + \binom{|\mathcal{B}|}{3} + t\binom{|\mathcal{A}||\mathcal{B}|/t}{2}. \end{aligned}$$

Let $x = |\mathcal{A}|$. Then $|\mathcal{B}| = t - x$. With a bit of calculus we can compute the derivative of the right-hand side of the inequality above with respect to x to be $(x/t - 1/2)\big(t(t-1) - 2xt + 2x^2\big)$, from which it follows that the minimum is at $x = n/2$, and so

$$M \geq 2\binom{t/2}{3} + t\binom{t/4}{2} \geq \left(\frac{7}{16} - \frac{2}{t}\right)\binom{t}{3}. \qquad \square$$

The observation implies that there is a triple $i_1 < i_2 < i_3$ that is nice to $(\frac{7}{16} - \frac{2}{t})\binom{k}{2}$ pairs $\{l, l'\} \in \binom{[k]}{2}$. Note, however, that the number of unordered pairs $\{(x,y,z),(x',y',z')\}$ such that $x \leq x', y \leq y', z \leq z'$ in a box $[L_{i,j}] \times [L_{i,k}] \times [L_{j,k}]$ is less than $\binom{L_{i_1,i_2}+1}{2}\binom{L_{i_1,i_3}+1}{2}\binom{L_{i_2,i_3}+1}{2}$. Hence,

$$\left(\frac{7}{16} - \frac{2}{t}\right)\binom{k}{2} \leq \binom{L_{i_1,i_2}+1}{2}\binom{L_{i_1,i_3}+1}{2}\binom{L_{i_2,i_3}+1}{2}.$$

Since $(k-1)^2/2 \leq \binom{k}{2}$ and $\binom{L+1}{2} \leq (L+1)^2/2$, we have $\max(L_{i_1,i_2}, L_{i_1,i_3}, L_{i_2,i_3}) \geq \big(\frac{7}{4} - \frac{8}{t}\big)^{1/6} k^{1/3} - 2$. $\square$

For a proof of theorem 11 we need a lemma about monotone functions that is of independent interest. A function $f(x,y)$ of two variables is said to be *strongly monotone* if the inequalities $f(x,y) \leq f(x',y)$, $f(x,y) \leq f(x,y')$ and $f(x,y) < f(x',y')$ hold whenever $x < x'$ and $y < y'$.

Lemma 23. *Suppose $f_1, f_2, f_3 \colon [s]^2 \to \mathbb{Z}$ are strongly monotone functions, and*

$$f(x,y,z) = \big(f_1(x,y), f_2(x,z), f_3(y,z)\big).$$

Then f takes at least $s^2/6$ distinct values on $[s]^3$.

Proof. The key observation is that if $f(x,y,z) = f(x',y',z')$, then (x,y,z) and (x',y',z') agree in at least one coordinate. Indeed, if it were not true, then by swapping p with p' and renaming the coordinates if necessary, we could have arranged that $x < x'$ and $y < y'$, which would have contradicted the strong monotonicity of f_1.

Let $(x_0,y_0,z_0) \in [s]^3$ be arbitrary, and consider the set $E \stackrel{\text{def}}{=} \{(x,y,z) : f(x,y,z) = f(x_0,y_0,z_0)\}$. The observation tells us that the set E is contained in a union of three coordinate hyperplanes, namely $H_x = \{(x_0,y,z) : y,z \in [s]^2\}$, $H_y = \{(x,y_0,z) : x,z \in [s]^2\}$, and $H_z = \{(x,y,z_0) : x,y \in [s]^2\}$. We claim that each of these hyperplanes contains at most $2s-1$ points of E. Indeed, $H_x \cap E$ cannot contain two points (x_0,y,z) and (x_0,y',z') such that $y < y'$ and $z < z'$, for that would have contradicted strong monotonicity of the function f_3. In particular $y - z$ is distinct as (x_0,y,z) runs over points of $H_x \cap E$. Thus, $|E| \leq 3(2s-1) \leq 6s$. Since (x_0,y_0,z_0) was arbitrary, the image of f must be of size at least $s^3/6s$, which completes the proof. $\square$

Proof of theorem 11. Put $L_{i,j} \stackrel{\text{def}}{=} \mathrm{LCS}(\pi_i, \pi_j)$. Let $\mathcal{I}$ be the set of all pairs (l,p) consisting of a letter $l \in [k]$ and an integer $1 \le p \le s_l$. For $(l,p) \in \mathcal{I}$ denote by $\pi_i\{l,p\}$ the prefix of the multipermutation π_i that ends with the p'th copy of the letter l. Put $\mathcal{I}_3 \stackrel{\text{def}}{=} \big\{(l,p_1,p_2,p_3) : (l,p_i) \in \mathcal{I}\big\}$. Define a function $f\colon \mathcal{I}_3 \to [L_{1,2}] \times [L_{1,3}] \times [L_{2,3}]$ by

$$f(l,p_1,p_2,p_3) = \big(\mathrm{LCS}(\pi_1\{l,p_1\},\pi_2\{l,p_2\}),\ \mathrm{LCS}(\pi_1\{l,p_1\},\pi_3\{l,p_3\}),\ \mathrm{LCS}(\pi_2\{l,p_2\},\pi_3\{l,p_3\})\big).$$

If $l,l' \in [k]$ are two different letters, then $f(l,p) \ne f(l',p')$ for all $p \in [s_l]^3$ and $p' \in [s_{l'}]^3$. Indeed, interchanging the roles of (l,p) and (l',p') and renaming the multipermutations if needed, we may assume that $\pi_1\{l,p_1\}$ and $\pi_2\{l,p_2\}$ are longer than $\pi_1\{l',p'_1\}$ and $\pi_2\{l',p'_2\}$ respectively, and so $f_1(l,p) > f_1(l',p')$ because the longest common subsequence between $\pi_1\{l',p'_1\}$ and $\pi_2\{l',p'_2\}$ can be extended to a longer common subsequence of $\pi_1\{l,p_1\}$ and $\pi_2\{l,p_2\}$.

Hence, $f(l,[s_l]^3) \cap f(l',[s'_l]^3) = \emptyset$ for distinct l,l'. However, the preceding lemma shows that for any fixed letter l we have $|f(l,[s_l]^3)| \ge s_l^2/6$. Thus $\frac{1}{6}\sum s_l^2 \le L_{1,2}L_{1,3}L_{2,3}$ and the theorem follows. □

9 Concluding remarks

- The upper bounds on $\mathrm{LT}(k,n)$ both in this paper and in [1] come from random words. Estimating $\mathrm{LT}(w)$ for a random $w \in [k]^n$ is an interesting problem on its own. A result of Kiwi, Loebl and Matoušek [7] asserts that if w_1, w_2 are two random words of length $n/2$, then $\mathbb{E}[\mathrm{LCS}(w_1,w_2)] \sim \frac{1}{\sqrt{k}}n$ for large k. Hence $1 \le \lim_n \frac{\mathbb{E}[\mathrm{LT}(w)]}{n\sqrt{k}} \le e$ for large k.
- In [1] a more general problem has been considered: Instead of looking for twins in w, one can seek T-tuplets, which are T-tuples of disjoint subsequences that are equal as words. One then defines the quantity $\mathrm{LT}_T(w)$ to be the length of longest T-tuplets in w, and defines $\mathrm{LT}_T(k,n)$ in a manner analogous to $\mathrm{LT}(k,n)$. Theorem 3 extends easily to $\mathrm{LT}_T(k,n) \ge C_T k^{-1+1/\binom{2T-1}{T}} n$ with the help of the following estimate:

  Theorem 24. *For every T and k we have $\mathrm{LCS}_T(2T-1,\mathcal{P}_k) \ge k^{1/\binom{2T-1}{T}}$.*

  Proof. For a letter $l \in [k]$ and any set of permutations $P \subset \mathcal{P}_k$ we denote by $P\{l\}$ the prefixes of all $\pi \in P$ that end with the letter l. Suppose $\Pi \subset \mathcal{P}_k$ is any set of $2T-1$ permutations. For each $I \subset \binom{\Pi}{T}$ and $l \in [k]$ put $f_I(l) \stackrel{\text{def}}{=} \mathrm{LCS}(I\{l\})$. Define the function $f\colon [k] \to \mathbb{N}^{\binom{\Pi}{T}}$ by $f(l)_I = f_I(l)$. If $l,l' \in [k]$ are distinct letters, then by the pigeonhole principle there are at least T permutations in which the relative order of l and l' is the same, and so $f(l) \ne f(l')$. So, f is injective, and thus its image cannot be contained in a box with of length smaller than $k^{1/\binom{2T-1}{T}}$. □

  The preceding theorem is sharp. First, $2T-1$ cannot be reduced to $2T-2$ because of a family that consists of $T-1$ copies of $\mathsf{123\cdots k}$ and $T-1$ copies of $\mathsf{k\cdots 321}$. Second, the bound itself cannot be improved, for $\mathrm{LCS}_T(2T,\mathcal{P}_k) \le k^{1/\binom{2T-1}{T}}$ as the following example shows: For each $I \subset \binom{[2T-1]}{T}$ associate a set $X_I = [k]$, and define $f_{i,I}\colon X_I \to \mathbb{Z}$ by $f_{i,I}(x_I) = x_I$ if $i \in I$ and $f_{i,I}(x_I) = -x_I$ if $i \notin I$. Also define $f_{0,I}$ by $f_{0,I}(x_I) = x_I$. Define $f_i = (f_{i,I})_I$, which is a function from $\prod_I X_I$ to $\mathbb{Z}^{\binom{2T-1}{T}}$, and put $\pi_i = [\![f_i]\!]$ as in section 7. It is then easy to check that $\Pi = \{\pi_0,\pi_1,\ldots,\pi_{2T-1}\}$ satisfies $\mathrm{LCS}_T(\Pi) = k$.

- In this paper we neglected to address the most basic extremal problem on LCS in sets of words, the estimation of $\mathrm{LCS}_2(t,[k]^n)$. Denote by l^n the word made of n copies of the letter l. The family $\{1^n, 2^n, \ldots, \mathsf{k}^n\}$ shows that $\mathrm{LCS}_2(k,[k]^n) = 0$. By labeling a word with the most popular letter that occurs in it, and applying the pigeonhole principle, one derives $\mathrm{LCS}_2(k+1,[k]^n) \geq n/k$. The bound is sharp, as the family of $k+2$ words $\{1^n, 2^n, \ldots, \mathsf{k}^n, 1^{n/k}2^{n/k}\cdots\mathsf{k}^{n/k}, \mathsf{k}^{n/k}\cdots 2^{n/k}1^{n/k}\}$ shows. For fixed $t \geq k+3$ it is possible to prove that $\mathrm{LCS}_2(t,[k]^n) \leq \big(1+o(1)\big)n/k$ by considering a family of words of the form $w_m = \big(1^m\cdots\mathsf{k}^m\big)^{n/mk}$ for a quickly growing sequence of values of m. Indeed, if $m_1 < m_2$, then a common subsequence of w_{m_1} and w_{m_2} is necessarily of the form $l_1^{p_1}l_2^{p_2}\ldots l_r^{p_r}$ for some $r \leq n/m_2$, but each subsequence of form l^p in w_1 spans an interval of length at least $\lfloor\frac{p-1}{m_1}\rfloor m_1 k \geq (p-m_1)k$ in w_1. Since $\sum(p_i-m_1)k \leq n$, we obtain that $\mathrm{LCS}_2(w_{m_1},w_{m_2}) \leq \big(\frac{1}{k}+\frac{m_1}{m_2}\big)n$. In a recent work of the first author with Jie Ma [5], we showed that this construction is essentially optimal when the number of words is constant.

- It is hard to resist the conjecture that the correct bound in lemma 23 is s^2 in place of $s^2/6$. If true, it would be sharp in view of the function $f(x,y,z) = (x,x,y)$.

Acknowledgments. The first author gratefully acknowledges helpful discussions with Yury Person and Sevak Mkrtchyan. We thank the referees for suggestions that helped to improve the exposition. All the errors remain ours.

References

[1] Maria Axenovich, Yury Person, and Svetlana Puzynina. A regularity lemma and twins in words. *J. Combin. Theory Ser. A*, 120(4):733–743, 2013. arXiv:1204.2180.

[2] Roger C. Baker, Glyn Harman, and János Pintz. The difference between consecutive primes. II. *Proc. London Math. Soc. (3)*, 83(3):532–562, 2001.

[3] Paul Beame, Eric Blaise, and Dang-Trinh Huynh-Ngoc. Longest common subsequences in sets of permutations. arXiv:0904.1615, April 2009.

[4] Paul Beame and Dang-Trinh Huynh-Ngoc. On the value of multiple read/write streams for approximating frequency moments. *Electronic Colloquium on Computational Complexity (ECCC)*, May 2008. Technical Report TR08-024.

[5] Boris Bukh and Jie Ma. Longest common subsequences in sets of words. *SIAM Disc. Math.*, 28(4):2042–2049, 2014. arXiv:1406.7017.

[6] Paul Erdős and George Szekeres. A combinatorial problem in geometry. *Compositio Math.*, 2:463–470, 1935.

[7] Marcos Kiwi, Martin Loebl, and Jiří Matoušek. Expected length of the longest common subsequence for large alphabets. *Adv. Math.*, 197(2):480–498, 2005. arXiv:math/0308234.

[8] B. F. Logan and L. A. Shepp. A variational problem for random Young tableaux. *Advances in Math.*, 26(2):206–222, 1977.

[9] A. M. Veršik and S. V. Kerov. Asymptotic behavior of the Plancherel measure of the symmetric group and the limit form of Young tableaux. *Dokl. Akad. Nauk SSSR*, 233(6):1024–1027, 1977.