Longest common subsequences in sets of words

Boris Bukh*
Jie Ma†

Abstract

Given a set of $t \geq k+2$ words of length n over a k-letter alphabet, it is proved that there exists a common subsequence among two of them of length at least $\frac{n}{k} + cn^{1-1/(t-k-2)}$, for some $c > 0$ depending on k and t. This is sharp up to the value of c.

1 Introduction

A *word* is a sequence of symbols from some fixed finite alphabet. For the problems in this paper only the size of the alphabet is important. So we will use $[k] \stackrel{\text{def}}{=} \{1, 2, \ldots, k\}$ for a canonical k-letter alphabet. The family of all words of length n over a k-letter alphabet is thus denoted by $[k]^n$. For a word w, a *subsequence* is any word obtained by deleting zero or more symbols from w. By a *subword* of w, we mean a subsequence of w consisting of consecutive symbols. For example, 1334 is a subsequence but not a subword of 12341234. A *common subsequence* of w and w' is a word that is a subsequence of both w and w'.

A general principle asserts that every sufficiently large collection of objects necessarily contains a pair of similar objects. In this paper, we treat the case when the objects are words, and similarity is measured by length of a common subsequence. We use $\mathrm{LCS}(w, w')$ to denote the length of the longest common subsequence of words w and w'. For a set $\mathcal{W}$ of words, let $\mathrm{LCS}(\mathcal{W}) \stackrel{\text{def}}{=} \max \mathrm{LCS}(w, w')$ where the maximum is taken over all pairs $\{w, w'\}$ in $\mathcal{W}$. We also allow $\mathcal{W}$ to be a multiset, so $\mathcal{W}$ might contain some elements multiple times.

For an integer $t \geq 2$ and a family $\mathcal{F}$ of words, let

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

A *permutation* of length k is a word over $[k]$ in which every symbol appears exactly once. Let $\mathcal{P}_k$ be the set of all permutations of length k. Much of the inspiration for our work comes

*Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Email: bbukh@math.cmu.edu. Supported in part by U.S. taxpayers through NSF grant DMS-1201380.

†School of Mathematical Sciences, University of Science and Technology of China, Hefei, Anhui 230026, China.

from the results of Beame–Huynh-Ngoc [4], Beame–Blais–Huynh-Ngoc [3] and Bukh–Zhou [2] that can be summarized as

$$\begin{aligned} \mathrm{LCS}(3, \mathcal{P}_k) &= k^{1/3} + O(1), \\ \mathrm{LCS}(4, \mathcal{P}_k) &= k^{1/3} + O(1), \\ 1.001k^{1/3} \le \mathrm{LCS}(t, \mathcal{P}_k) &\le 4k^{1/3} + O(k^{7/40}) \qquad \text{for } 5 \le t \le k^{1/3}. \end{aligned}$$

The problem of bounding $\mathrm{LCS}(t, \mathcal{P}_k)$ is closely related to the longest twin problem of Axenovich, Person and Puzynina [1]. Here, two subsequences w_1, w_2 of the same word w are *twins* if they are equal as words but the sets of positions of symbols from w retained in w_1 and in w_2 are disjoint. It was shown in [2] that if $\mathrm{LCS}(t, \mathcal{P}_k)$ is small for $t \ge 2k$, then there are words that contain no long twins, and that a converse (which is more technical to state) also holds.

In this paper, we consider $\mathrm{LCS}(t, [k]^n)$. We let w^m be the concatenation of m copies of w, e.g., $(\mathsf{343})^2 = \mathsf{343343}$. For $t \le k$ we have $\mathrm{LCS}(t, [k]^n) = 0$ as the family $\{\mathsf{1}^n, \mathsf{2}^n, \dots, \mathsf{t}^n\}$ shows. For $t = k + 1$ we have[1]

$$\mathrm{LCS}(k+1, [k]^n) = \tfrac{n}{k}.$$

The upper bound is attained by the family $\{\mathsf{1}^n, \mathsf{2}^n, \dots, \mathsf{k}^n, (\mathsf{12...k})^{\frac{n}{k}}\}$ of $k+1$ words. The lower bound is a consequence of two simple facts: the most popular letter in a word occurs at least $\frac{n}{k}$ times, and the most popular letter is the same in two out of $k+1$ words. The following is our main theorem, which determines the asymptotic magnitude of $\mathrm{LCS}(t, [k]^n)$ for all $t \ge k + 2$.

Theorem 1. *For nonnegative integers k, r and n such that $k \ge 2$ and $n \ge k(10r)^{9r}$, there exists $c = \Theta(r^{-9}k^{1/r-2})$ such that*

$$\mathrm{LCS}(r + k + 2, [k]^n) \ge \frac{n}{k} + cn^{1-\frac{1}{r}}.$$

This theorem is sharp up to the value of c. For $0 \le i \le r$, put

$$m_i \stackrel{\text{def}}{=} (n/k)^{i/r},$$

and define words

$$\begin{aligned} w_i &\stackrel{\text{def}}{=} (\mathsf{1}^{m_i}\mathsf{2}^{m_i} \dots \mathsf{k}^{m_i})^{\frac{n}{km_i}}, \\ \operatorname{rev} w_i &\stackrel{\text{def}}{=} (\mathsf{k}^{m_i} \dots \mathsf{2}^{m_i}\mathsf{1}^{m_i})^{\frac{n}{km_i}}. \end{aligned}$$

Note that $\operatorname{rev} w_i$ is the word obtained by reversing the symbols of w_i. We claim that for the family

$$\mathcal{W} \stackrel{\text{def}}{=} \{w_0, w_1, \dots, w_r, \operatorname{rev} w_r, \mathsf{1}^n, \mathsf{2}^n, \dots, \mathsf{k}^n\} \tag{1}$$

we have

$$\mathrm{LCS}(\mathcal{W}) \le \frac{n}{k} + k^{1/r}n^{1-1/r}.$$

[1]For readability, we omit the floor and ceiling signs throughout the paper.

Indeed, $\mathrm{LCS}(w_i, \mathsf{j}^n) = \frac{n}{k}$ is clear, and every common subsequence of w_r and $\operatorname{rev} w_r$ is of the form i^m, and so is of length at most n/k. Hence, it suffices to bound $\mathrm{LCS}(w_i, w_j)$ and $\mathrm{LCS}(w_i, \operatorname{rev} w_j)$ for $i < j$ (though we need a bound on $\mathrm{LCS}(w_i, \operatorname{rev} w_j)$ only for $j = r$). The two cases are similar: Any common subsequence of w_i and w_j (or $\operatorname{rev} w_j$) must be of the form $k_1^{p_1} k_2^{p_2} \dots k_s^{p_s}$ for some $s \le \frac{n}{m_j}$, where $k_l \in [k]$ for $l = 1, 2, \dots, s$. Since each subsequence $k_l^{p_l}$ of w_i spans a subword of length at least $(\lceil \frac{p_l}{m_i} \rceil - 1)km_i \ge (p_l - m_i)k$ in w_i, it follows that $n \ge \sum_{l=1}^{s} (p_l - m_i)k$, implying that $\mathrm{LCS}(w_i, w_j) = \max \sum_l p_l \le \frac{n}{k} + \frac{m_i}{m_j} n \le \frac{n}{k} + k^{1/r} n^{1-1/r}$.

A word $w \in [k]^n$ is called *balanced* if it contains the same number of i's as j's for any $\mathsf{i}, \mathsf{j} \in [k]$. Let B_k^n be the family containing all balanced words in $[k]^n$. As we shall see, the assertion of Theorem 1 reduces to the following result on family B_k^n.

Theorem 2. *For nonnegative integers k, r and n such that $k \ge 2$ and $n \ge k(10r)^{9r}/2$, there exists $c' = \Theta(r^{-9} k^{1/r-1})$ such that*

$$\mathrm{LCS}(r + 2, B_k^n) \ge \frac{n}{k} + c' n^{1-\frac{1}{r}}.$$

For large n, Theorem 2 is sharp up to the value of c', as witnessed by the family of words $\{w_0, w_1, \dots, w_r, \operatorname{rev} w_r\}$, where w_i's are as in (1). For values of n that are comparable to k, the rates of growth as $k \to \infty$ for $\mathrm{LCS}(3, B_k^n)$ and for $\mathrm{LCS}(4, B_k^n)$ have been determined in [2, Theorems 11 and 12].

The rest of the paper is organized as follows. In the next section, we reduce Theorems 1 and 2 to the LCS for balanced binary words (see Theorem 3). The proof of Theorem 3 will be completed in section 3. The last section contains a couple of open problems. In this paper, we do not attempt to optimize the constants, and instead aim for simpler presentation.

2 Reductions

In this section we deduce Theorems 1 and 2 from the following special case $k = 2$ of Theorem 2, which we state separately.

Theorem 3. *Let r and n be nonnegative integers such that $n \ge (10r)^{9r}$. For any set $\mathcal{W}$ of $r + 2$ balanced words in $\{0, 1\}^n$, we have*

$$\mathrm{LCS}(\mathcal{W}) \ge \frac{n}{2} + \Omega\left(r^{-9}\right) \cdot n^{1-1/r}.$$

Proof of Theorem 2. (*Assume that Theorem 3 holds.*) Consider a multiset $\mathcal{W}$ of arbitrary $r + 2$ balanced words from $[k]^n$ and let

$$\mathcal{W}' \stackrel{\text{def}}{=} \{\text{the subsequence of } w \text{ consisting of all } 1\text{'s and } 2\text{'s for every } w \in \mathcal{W}\}.$$

Then $\mathcal{W}'$ is a multiset of $r+2$ balanced words from $\{1, 2\}^{n'}$ for $n' = \frac{2n}{k} \ge (10r)^{9r}$. By Theorem 3, we have $\mathrm{LCS}(\mathcal{W}) \ge \mathrm{LCS}(\mathcal{W}') \ge \frac{n'}{2} + \Omega(r^{-9})(n')^{1-1/r} = \frac{n}{k} + \Omega(r^{-9} k^{1/r-1}) \cdot n^{1-1/r}$. □

Proof of Theorem 1. (*Assume that Theorem 2 holds.*) Let c be a small constant. Consider an arbitrary set of $r+k+2$ words from $[k]^n$. Call $w \in [k]^n$ *unhinged* if some letter occurs in w at least $\frac{n}{k} + cn^{1-1/r}$ times and *hinged* otherwise. Observe that if there are $k+1$ unhinged words in the set, then some two of them have LCS of length at least $\frac{n}{k} + cn^{1-1/r}$. Thus we may assume that there are at least $r+2$ hinged words. Since each hinged word of length n contains a subsequence that is a balanced word of length $n - k^2 cn^{1-1/r}$, by Theorem 2, some two hinged words have LCS of length at least

$$\frac{n - k^2 cn^{1-\frac{1}{r}}}{k} + c' \left(n - k^2 cn^{1-\frac{1}{r}}\right)^{1-\frac{1}{r}} \geq \frac{n}{k} + cn^{1-\frac{1}{r}},$$

provided $c = \Theta\left(\frac{c'}{k}\right)$. This proves Theorem 1. □

3 The proof of Theorem 3

Throughout this proof, let $\alpha \stackrel{\text{def}}{=} 10^{-6} r^{-9}$ and $\beta \stackrel{\text{def}}{=} \frac{1}{40000} r^{-6}$, and let $\mathcal{W}$ be a set consisting of arbitrary balanced words $w^{(1)}, w^{(2)}, \ldots, w^{(r+2)}$ in $\{0,1\}^n$. Moreover, we assume that $r \geq 2$, as it is easy to see that $\mathrm{LCS}(\mathcal{W}) \geq \frac{n}{2}$ when $r = 0$ and $\mathrm{LCS}(\mathcal{W}) \geq \frac{n}{2} + 1$ when $r = 1$.

We give a very brief outline of the proof before proceeding. A crucial idea is to consider the scale on which 0's and 1's alternate in a word. For example, in word $(01)^{n/2}$ alternation (between 0's and 1's) happens on scale $\Theta(1)$, whereas in word $0^{n/2}1^{n/2}$ the alternation scale is about $\Theta(n)$. The proof will first find two words, say $w^{(1)}$ and $w^{(2)}$, of "comparable" alternation scale, and then show, in effect, that $\mathrm{LCS}(w^{(1)}, w^{(2)})$ is large.

We shall think of words as made of a sequence of *distinguishable* 0's and 1's. That means that if we say "let z be a 0 in word w", then the variable z refers to a particular 0. For example, if z is the 3'rd 0 in the word 1100101101, and w' is the word obtained from w by removing the 1'st and 4'th zeros, namely $w' = 11010111$, then z becomes the 2'nd zero in w'.

For two symbols a, b from w such that a is to the left of b, we denote by $w[a,b]$ the subword of w starting from a and ending with b. We also use $w(a,b)$ to denote the subword of w obtained from $w[a,b]$ by deleting a and b. The notations $w[a,b)$ and $w(a,b]$ are defined similarly.

If z is a 0 in a word w, its *position*, denoted $P_w(z)$, is the number of 1's to the left of z. When the word w is clear from the context, we will drop the subscript of $P_w(z)$ and write simply $P(z)$. Note that several 0's might have the same position. If z is the j'th 0 in w, we say that its *expected position* is j, because in a random word the expected value of $P(z)$ is j. We say that a 0 is *good* in w if its position differs from its expected position by at most $\alpha n^{1-1/r}$. If a 0 is not good, then its position is either to the left or to the right of its expected position. In these cases we call such a 0 *left-bad* and *right-bad* respectively. The following claim will be used frequently.

Claim 1. *If a subword of $w^{(i)}$ contains N 1's, then it contains at most $N + 2\alpha n^{1-1/r}$ good 0's.*

Proof. Let z_l and z_r be the leftmost and the rightmost good 0's in the subword. By definition, we have $P(z_r) - P(z_l) \leq N$. From the goodness of z_r, we see that its expected position differs

from $P(z_r)$ by at most $\alpha n^{1-1/r}$; similarly it holds for z_l. Therefore, the expected positions of z_r and z_l differ by at most $N + 2\alpha n^{1-1/r}$, implying this claim. □

We introduce a concept closely related to the alternation scale described in the outline. A subword is called a *0-rich interval of length L* if it contains exactly L good 0's and no more than $L/10$ 1's. The *type* of a good 0 is the largest integer t such that this 0 is contained in a 0-rich interval of length exactly $n^{t/r}$. Note that a type of a good 0 is well-defined since every good zero is contained in a 0-rich interval of length 1. Also note that a type cannot be $r-1$. Indeed, if there existed a 0-rich interval of length $n^{1-1/r}$, then Claim 1 would imply that $n^{1-1/r} \leq n^{1-1/r}/10 + 2\alpha n^{1-1/r}$, a contradiction. We define a *type of a bad 0* to be either left-bad or right-bad. Thus a type of each 0 is an element of $\{0, 1, \ldots, r-2, \text{left-bad}, \text{right-bad}\}$.

To be able to refer to individual 0's, we define $0_j^{(i)}$ as the j'th 0 in word $w^{(i)}$. As our proof does not treat 0's and 1's symmetrically, we do not need a similar notation to refer to individual 1's.

Fix an integer j and consider $0_j^{(1)}, 0_j^{(2)}, \ldots, 0_j^{(r+2)}$. We may assume that at most one of these zeros is left-bad, and at most one of them is right-bad. Suppose, on the contrary, that both $0_j^{(1)}$ and $0_j^{(2)}$ are left-bad. Then we can obtain a common subsequence of $w^{(1)}$ and $w^{(2)}$ with length at least $n/2+\alpha n^{1-1/r}$ by matching up the first j 0's and then 1's to the right of $0_j^{(1)}$ and $0_j^{(2)}$. Hence, in this case $\mathrm{LCS}(\mathcal{W}) \geq \mathrm{LCS}(w^{(1)}, w^{(2)}) \geq n/2 + \alpha n^{1-1/r} = n/2 + \Omega(r^{-9}) \cdot n^{1-1/r}$. The case of two right-bad 0's is similar.

Hence, for any integer j, two of $0_j^{(1)}, 0_j^{(2)}, \ldots, 0_j^{(r+2)}$ are of the same type, and that type is one of $0, 1, \ldots, r-2$. By the pigeonhole principle, there are two words, say $w^{(1)}$ and $w^{(2)}$, and some $t \in \{0, 1, \ldots, r-2\}$ such that the set

$$\mathcal{T} \stackrel{\text{def}}{=} \{j : \text{ both } 0_j^{(1)} \text{ and } 0_j^{(2)} \text{ have type } t\}$$

has size at least $\frac{n/2}{\binom{r+2}{2}(r-1)} \geq \frac{n}{2r^3}$. We will show that $w^{(1)}$ and $w^{(2)}$ contain a common subsequence of length $n/2 + \Omega(n^{1-1/r})$.

We partition each of $w^{(1)}$ and $w^{(2)}$ into *blocks* that contain exactly $\beta n^{1-1/r}$ many 1's. To be more precise, for each $i \in \{1, 2\}$, the k'th block (denoted by $B_k^{(i)}$) of word $w^{(i)}$ is defined to be the subword $w^{(i)}[a_{k-1}, a_k)$, where a_k denotes the $(k \cdot \beta n^{1-1/r} + 1)$'th 1 in word $w^{(i)}$.

For each $i \in \{1, 2\}$ and each $j \in \mathcal{T}$, choose a 0-rich interval of length $n^{t/r}$ containing $0_j^{(i)}$ and call this interval $I_j^{(i)}$. By shrinking $I_j^{(i)}$ if necessary, we may assume that both leftmost and rightmost symbols in $I_j^{(i)}$ are good 0's. An integer $j \in \mathcal{T}$ is *consistent* if $I_j^{(1)} \subset B_k^{(1)}$ and $I_j^{(2)} \subset B_k^{(2)}$ for some k. Let $\mathcal{S} = \{j \in \mathcal{T} : j \text{ is consistent}\}$.

Claim 2. $|\mathcal{S}| \geq \frac{n}{4r^3}$.

Proof. For each $i \in \{1, 2\}$, let $L_k^{(i)}$ be the subword of $w^{(i)}$ spanning the last $2\alpha n^{1-1/r}$ 1's in the block $B_k^{(i)}$ and the first $2\alpha n^{1-1/r}$ 1's in the block $B_{k+1}^{(i)}$. By Claim 1, we see that $L_k^{(i)}$ contains

at most $6\alpha n^{1-1/r}$ good $\mathsf{0}$'s, and hence the set $L^{(i)} \stackrel{\text{def}}{=} \{\text{all good } \mathsf{0}\text{'s contained in } \cup_k L_k^{(i)}\}$ is of size at most $6\alpha n^{1-1/r} \cdot \frac{n/2}{\beta n^{1-1/r}} = \frac{3\alpha n}{\beta}$. Let

$$\mathcal{T}' \stackrel{\text{def}}{=} \{j \in \mathcal{T} : \mathsf{0}_j^{(i)} \notin L^{(i)} \text{ for each } i = 1, 2\}.$$

It is clear that $|\mathcal{T}'| \geq |\mathcal{T}| - |L^{(1)}| - |L^{(2)}| \geq \frac{n}{2r^3} - \frac{6\alpha n}{\beta} \geq \frac{n}{4r^3}$.

Now it suffices to show that $\mathcal{T}' \subseteq \mathcal{S}$. Consider an arbitrary integer $j \in \mathcal{T}'$. Assume that $\mathsf{0}_j^{(1)} \in B_k^{(1)}$ for some k. By the definition of $\mathcal{T}'$, it holds that

$$(k-1)\beta n^{1-1/r} + 2\alpha n^{1-1/r} < P(\mathsf{0}_j^{(1)}) \leq k\beta n^{1-1/r} - 2\alpha n^{1-1/r}.$$

As $t \leq r-2$ and $n \geq (10r)^{9r}$, the $\mathsf{0}$-rich interval $I_j^{(1)}$ has at most $n^{t/r}/10 \leq 2\alpha n^{1-1/r}$ $\mathsf{1}$'s, implying that $I_j^{(1)} \subset B_k^{(1)}$. By the goodness of $\mathsf{0}_j^{(1)}$ and $\mathsf{0}_j^{(2)}$, we obtain that $|P(\mathsf{0}_j^{(1)}) - P(\mathsf{0}_j^{(2)})| \leq 2\alpha n^{1-1/r}$, which implies that $\mathsf{0}_j^{(2)} \in B_k^{(2)}$. By the definition of $\mathcal{T}'$ again, in fact we have $(k-1)\beta n^{1-1/r} + 2\alpha n^{1-1/r} < P(\mathsf{0}_j^{(2)}) \leq k\beta n^{1-1/r} - 2\alpha n^{1-1/r}$. Repeating the same argument, we see $I_j^{(2)} \subset B_k^{(2)}$. So j is consistent and hence $j \in \mathcal{S}$, finishing the proof of Claim 2. □

With slight abuse of notation, let $\mathcal{S} \cap B_k^{(i)} \stackrel{\text{def}}{=} \{\mathsf{0}_j^{(i)} \in B_k^{(i)} : j \in \mathcal{S}\}$. Clearly, $\mathcal{S} \cap B_k^{(1)}$ and $\mathcal{S} \cap B_k^{(2)}$ are of the same size, say s_k. Then s_k satisfy

$$0 \leq s_k \leq (\beta + 2\alpha)n^{1-1/r} \quad \text{and} \quad \sum_k s_k = |\mathcal{S}| \geq \frac{n}{4r^3}, \tag{2}$$

where the first inequality follows by Claim 1. For fixed k and $i \in \{1, 2\}$, consider the family of all $\mathsf{0}$-rich intervals $I_j^{(i)}$ that belong to $B_k^{(i)}$ as j ranges over $\mathcal{S}$. It is clear that the union of $I_j^{(i)}$'s from this family contains all $\mathsf{0}$'s in $\mathcal{S} \cap B_k^{(i)}$. By the Vitali covering lemma, there is a subfamily, denoted by $\mathcal{I}_k^{(i)}$, consisting of pairwise disjoint intervals $I_j^{(i)}$ whose union contains at least one third of the $\mathsf{0}$'s in $\mathcal{S} \cap B_k^{(i)}$. Since each $I_j^{(i)}$ contains at most $n^{t/r}$ $\mathsf{0}$'s from $\mathcal{S} \cap B_k^{(i)}$, we derive

$$|\mathcal{I}_k^{(i)}| \geq \frac{s_k}{3n^{t/r}}. \tag{3}$$

Let $\mathcal{I}^{(i)} \stackrel{\text{def}}{=} \cup_k \mathcal{I}_k^{(i)}$. The intervals in $\mathcal{I}^{(i)}$ are disjoint, for intervals in $\mathcal{I}_k^{(i)}$ and $\mathcal{I}_{k'}^{(i)}$ for $k \neq k'$ are contained in non-overlapping blocks.

We shall pick an integer Q in the interval $(-\beta n^{1-1/r}, \beta n^{1-1/r})$ uniformly at random, and define words $\dot{w}^{(1)}$ and $\dot{w}^{(2)}$ as follows. If $Q \geq 0$, let $\dot{w}^{(1)} \stackrel{\text{def}}{=} w^{(1)}$ and $\dot{w}^{(2)}$ be obtained from $w^{(2)}$ by removing the first Q $\mathsf{1}$'s; otherwise, let $\dot{w}^{(2)} \stackrel{\text{def}}{=} w^{(2)}$ and $\dot{w}^{(1)}$ be obtained from $w^{(1)}$ by removing the first $-Q$ $\mathsf{1}$'s.

For an interval $I \in \mathcal{I}^{(i)}$, its *left-position* (resp. *right-position*) in $w^{(i)}$ is the position of the leftmost (resp. rightmost) good $\mathsf{0}$ in w. We denote left- and right-positions by $LP(I)$ and

$RP(I)$. We define the left- and right-positions of an interval in $\dot{w}^{(i)}$ similarly, and denote them by $\dot{LP}(I)$ and $\dot{RP}(I)$. We note that for intervals $I_1 \in \mathcal{I}^{(1)}$ and $I_2 \in \mathcal{I}^{(2)}$

$$LP(I_2) - LP(I_1) - Q = \dot{LP}(I_2) - \dot{LP}(I_1). \tag{4}$$

We say that two intervals $I_1 \in \mathcal{I}^{(1)}$ and $I_2 \in \mathcal{I}^{(2)}$ are *close* or (I_1, I_2) is a *close pair*, if

$$|\dot{LP}(I_2) - \dot{LP}(I_1)| \leq \frac{1}{20} n^{t/r}.$$

Suppose that intervals $I_1 \in \mathcal{I}^{(1)}$ and $I_2 \in \mathcal{I}^{(2)}$ are close, then as $0 \leq \dot{RP}(I_i) - \dot{LP}(I_i) \leq \frac{1}{10} n^{t/r}$, we also have

$$|\dot{RP}(I_2) - \dot{RP}(I_1)| \leq \frac{3}{20} n^{t/r}. \tag{5}$$

Claim 3. *Each interval in $\mathcal{I}^{(1)}$ is close to at most $n^{1/r}$ intervals in $\mathcal{I}^{(2)}$. Similarly, each interval in $\mathcal{I}^{(2)}$ is close to at most $n^{1/r}$ intervals in $\mathcal{I}^{(1)}$.*

Proof. Suppose, on the contrary, that an interval $I \in \mathcal{I}^{(1)}$ is close to $J_1, J_2, \ldots, J_d \in \mathcal{I}^{(2)}$ with $\dot{LP}(J_1) < \dot{LP}(J_2) < \ldots < \dot{LP}(J_d)$, where $d \stackrel{\text{def}}{=} n^{1/r} + 1$. Let J be the subword of $w^{(2)}$ starting from the leftmost good $\mathsf{0}$ of J_1 and ending with the leftmost good $\mathsf{0}$ of J_d. By the closeness of (I, J_1) and of (I, J_d), we have $|\dot{LP}(J_1) - \dot{LP}(J_d)| \leq \frac{1}{10} n^{t/r}$, which implies that J has at most $\frac{1}{10} n^{t/r} \leq \frac{1}{10} n^{(t+1)/r}$ $\mathsf{1}$'s. Since J also contains at least $(d-1) \cdot n^{t/r} = n^{(t+1)/r}$ good $\mathsf{0}$'s, every $\mathsf{0}$ in J is contained in a $\mathsf{0}$-rich interval of length $n^{(t+1)/r}$. Hence the type of any $\mathsf{0}$ in J_1 is at least $t+1$. Yet from the construction of $\mathcal{I}^{(2)}$, it is evident that J_1 contains at least one $\mathsf{0}$ of type t. This contradiction finishes the proof of Claim 3. □

Some intervals in the first block, i.e., those in $\mathcal{I}_1^{(1)} \cup \mathcal{I}_1^{(2)}$, might be destroyed in the passage from $w^{(1)}$ and $w^{(2)}$ to their dotted counterparts. So let $k \geq 2$ and consider two arbitrary intervals $I_1 \in \mathcal{I}_k^{(1)}$ and $I_2 \in \mathcal{I}_k^{(2)}$. In view of (4), I_1 and I_2 are close if and only if

$$|LP(I_2) - LP(I_1) - Q| \leq \frac{1}{20} n^{t/r}. \tag{6}$$

Since I_1 and I_2 are in the same block, there exists an integer $q \in (-\beta n^{1-1/r}, \beta n^{1-1/r})$ such that $LP(I_2) = LP(I_1) + q$. Therefore there are at least $\frac{1}{20} n^{t/r}$ choices of Q's for which (6) holds, namely Q can be any integer in $[q - \frac{1}{20} n^{t/r}, q + \frac{1}{20} n^{t/r}] \cap (-\beta n^{1-1/r}, \beta n^{1-1/r})$. This shows that the probability that I_1 and I_2 are close is at least

$$p \stackrel{\text{def}}{=} \frac{\frac{1}{20} n^{t/r}}{2\beta n^{1-1/r}} = \frac{1}{40\beta} n^{1/r + t/r - 1}. \tag{7}$$

Let $E \subset \mathcal{I}^{(1)} \times \mathcal{I}^{(2)}$ be the set of close pairs (I_1, I_2). Then the expectation of $|E|$ is at least $p \cdot \left(\sum_{k \geq 2} |\mathcal{I}_k^{(1)}| |\mathcal{I}_k^{(2)}| \right)$. There must exist some $Q \in (-\beta n^{1-1/r}, \beta n^{1-1/r})$ such that the size of

E is at least its expectation. Fix such a Q. Note that this also fixes $\dot{w}^{(1)}, \dot{w}^{(2)}$ and the set E. By (2), (3), (7) and the Cauchy–Schwarz inequality, we derive

$$|E| \geq p \cdot \left(\sum_{k \geq 2} |\mathcal{I}_k^{(1)}||\mathcal{I}_k^{(2)}| \right) \geq \frac{n^{1-t/r}}{5000r^6}, \tag{8}$$

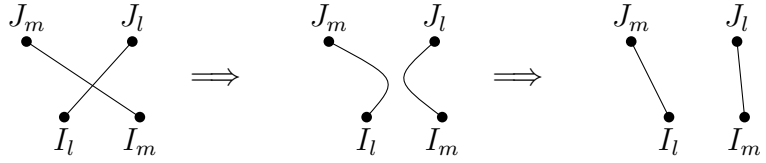
since the summation is over at most $\frac{n/2}{\beta n^{1-1/r}} = \frac{n^{1/r}}{2\beta}$ terms.

Claim 4. *There exist $\frac{|E|}{2n^{1/r}}$ close pairs (I_i, J_i) in E such that*

$$\begin{aligned} \dot{LP}(I_1) < \dot{LP}(I_2) < \ldots < \dot{LP}(I_{|E|/2n^{1/r}}), \\ \dot{LP}(J_1) < \dot{LP}(J_2) < \ldots < \dot{LP}(J_{|E|/2n^{1/r}}). \end{aligned} \tag{9}$$

Proof. We can view E as the edge set of a bipartite graph G with bipartition $(\mathcal{I}^{(1)}, \mathcal{I}^{(2)})$. We desire to find a large matching $I_1J_1, I_2J_2, \ldots$ satisfying (9). Identify $I \in \mathcal{I}^{(1)}$ with the point $(\dot{LP}(I), 0)$ in the Euclidean plane, and identify $J \in \mathcal{I}^{(2)}$ with the point $(\dot{LP}(J), 1)$. Edges will be represented by line segments. Among all the matchings of maximum size, pick one that minimizes the total Euclidean length of edges.

We claim that this matching satisfies (9). Suppose, on the contrary, that $\dot{LP}(I_l) < \dot{LP}(I_m)$ and $\dot{LP}(J_l) > \dot{LP}(J_m)$. Then the line segments $\overline{I_lJ_l}$ and $\overline{I_mJ_m}$ cross. That implies that line segments $\overline{I_lJ_m}$ and $\overline{I_mJ_l}$ are both shorter than $\max\big(\mathrm{dist}(I_l, J_l), \mathrm{dist}(I_l, J_l)\big)$, and so $I_lJ_m, I_mJ_l \in E$. This contradicts the choice of the matching, since replacing edges I_lJ_l and I_mJ_m with I_lJ_m and I_mJ_l decreases the total length of the matching as the following picture demonstrates.



The bound on the size of the matching follows from Claim 3, which in the present language asserts that the maximum degree in E is at most $n^{1/r}$. Indeed, for any matching M of size less than $|E|/2n^{1/r}$ there is an $e \in E$ not adjacent to any edge of M. Hence, a maximal matching has at least $|E|/2n^{1/r}$ edges. □

Finally, using the close pairs of Claim 4, we find a long common subsequence of $w^{(1)}$ and $w^{(2)}$. For convenience write $\lambda \stackrel{\text{def}}{=} |E|/2n^{1/r}$ and without loss assume that $Q \geq 0$. For $1 \leq i \leq \lambda - 1$, let A_i be the subword of $\dot{w}^{(1)}$ between the intervals I_i and I_{i+1} and B_i be the subword of $\dot{w}^{(2)}$ between the intervals J_i and J_{i+1}. In addition, let A_0 be the subword of $\dot{w}^{(1)}$ before the interval I_1, and A_λ be the subword of $\dot{w}^{(1)}$ after the interval I_λ; the definitions of B_0 and B_λ are similar. Let us consider the common subsequence w of $\dot{w}^{(1)}$ and $\dot{w}^{(2)}$, which consists of the common $\mathsf{0}$'s of I_i and J_i and the common $\mathsf{1}$'s of A_i and B_i for all $0 \leq i \leq \lambda$. By (5) and (6), we have

$$|\dot{RP}(J_i) - \dot{RP}(I_i)| \leq \frac{3}{20} n^{t/r} \quad \text{and} \quad |\dot{LP}(J_{i+1}) - \dot{LP}(I_{i+1})| \leq \frac{1}{20} n^{t/r},$$

which shows that for each $i < \lambda$, the counts of 1's in A_i and in B_i differ by at most $\frac{1}{5}n^{t/r}$. Also note that each I_i contains at most $\frac{1}{10}n^{t/r}$ 1's, thus the number of 1's in $\dot{w}^{(2)}$ but not in w is at most $\frac{\lambda}{2}n^{t/r}$. By (8) as well as the facts that $\lambda = |E|/2n^{1/r}$ and $|Q| < \beta n^{1-1/r}$, we derive that

$$\begin{aligned}\mathrm{LCS}(\mathcal{W}) &\geq \mathrm{LCS}(w^{(1)}, w^{(2)}) \geq \mathrm{LCS}(\dot{w}^{(1)}, \dot{w}^{(2)}) \geq |w| \geq \lambda \cdot n^{t/r} + \left(\frac{n}{2} - |Q| - \frac{\lambda}{2} \cdot n^{t/r}\right) \\ &\geq \frac{n}{2} - \beta n^{1-1/r} + \frac{n^{1-1/r}}{20000r^6} = \frac{n}{2} + \Omega(r^{-6}) \cdot n^{1-1/r}.\end{aligned}$$

This completes the proof of Theorem 3.

4 Two problems

In this paper we proved that $\mathrm{LCS}(r + k + 2, [k]^n) = \frac{n}{k} + \Theta_{r,k}(n^{1-1/r})$. It is possible that the coefficient in the big-theta notation need not depend on r, but we have been unable to prove so. In particular, what is the smallest r such that $\mathrm{LCS}(r + k + 2, [k]^n) \geq 1.01\frac{n}{k}$? Is it asymptotic to $\Theta(\log n)$?

Another worthy problem is the length of the longest common subsequence between two random words. A superadditivity argument shows that the expected length of such a subsequence is asymptotic to $\gamma_k n$ for some constant γ_k. Kiwi–Loebl–Matoušek [5] proved that $\gamma_k\sqrt{k} \to 2$ as $k \to \infty$, but the value of γ_k is not known for any $k \geq 2$ (including the case $k = 4$ that is natural for the problem of DNA comparison).

These two problems are connected. Azuma's inequality implies that $\mathrm{LCS}(w, w')$ for random $w, w' \in [k]^n$ is concentrated in an interval of length $\sqrt{n}$ with sub-Gaussian tails. It thus follows that for any $\epsilon > 0$ one can find a family $\mathcal{F}$ of exponentially many words from $[k]^n$ such that $\mathrm{LCS}(\mathcal{F}) \leq (\gamma_k + \epsilon + o(1)) \cdot n$.

References

[1] M. Axenovich, Y. Person and S. Puzynina, *A regularity lemma and twins in words*, J. Combin. Theory Ser. A, **120(4)** (2013), 733-743.

[2] B. Bukh and L. Zhou, *Twins in words and long common subsequences in permutations*, submitted.

[3] P. Beame, E. Blaise and D. Huynh-Ngoc, *Longest common subsequences in sets of permutations*, arXiv:0904.1615.

[4] P. Beame and D. Huynh-Ngoc, *On the value of multiple read/write streams for approximating frequency moments*, Electronic Colloquium on Computational Complexity (ECCC) (2008), 499-508.

[5] M. Kiwi, M. Loebl and J. Matoušek, *Expected length of the longest common subsequence for large alphabets*, Adv. Math., **197(2)** (2005), 480-498.