Longest Common Subsequence on Weighted Sequences.

Kipouridis, Evangelos; Tsichlas, K.

Published in:
Longest Common Subsequence on Weighted Sequences.

DOI:
10.4230/LIPIcs.CPM.2020.19

Publication date:
2020

Document version
Early version, also known as pre-print

Citation for published version (APA):
https://doi.org/10.4230/LIPIcs.CPM.2020.19
Abstract
We consider the general problem of the Longest Common Subsequence (LCS) on weighted sequences. Weighted sequences are an extension of classical strings, where in each position every letter of the alphabet may occur with some probability. In this paper we provide faster algorithms and prove a series of hardness results for more general variants of the problem. In particular, we provide an NP-Completeness result on the general variant of the problem instead of the log-probability version used in earlier papers, already for alphabets of size 2. Furthermore, we design an EPTAS for bounded alphabets, which is also an improved, compared to previous results, PTAS for unbounded alphabets. These are in a sense optimal, since it is known that there is no FPTAS for bounded alphabets, while we prove that there is no EPTAS for unbounded alphabets. Finally, we provide a matching conditional (under the Exponential Time Hypothesis) lower bound for any PTAS. As a side note, we prove that it is sufficient to work with only one threshold in the general variant of the problem.

1 Introduction
We consider the problem of determining the LCS (Longest Common Subsequence) on weighted sequences. Weighted sequences, also known as p-weighted sequences or Position Weighted Matrices (PWM) \cite{8, 23} are probabilistic sequences which extend the notion of strings, in the sense that in each position there is some probability for each letter of an alphabet $\Sigma$ to occur there.

Weighted sequences were introduced as a tool for motif discovery and local alignment and are extensively used in molecular biology \cite{16}. They have been studied both in the context of short sequences (binding sites, sequences resulting from multiple alignment, etc.) and on large sequences, such as complete chromosome sequences that have been obtained using a whole-genome shotgun strategy \cite{21, 24}. Weighted sequences are able to keep all the information produced by such strategies, while classical strings impose restrictions that oversimplify the original data.

Basic concepts concerning the combinatorics of weighted sequences (like pattern matching, repeats discovery and cover computation) were studied using weighted suffix trees \cite{18}, Crochemore’s partitioning \cite{8, 9, 13}, the Karp-Miller-Rabin algorithm \cite{13}, and other approaches \cite{30}. Other interesting results include approximate and gapped pattern matching \cite{8, 23}, property matching \cite{24}, swapped matching \cite{27}, the all-covers and all-seeds problem \cite{26, 29}, and extracting motifs \cite{20}. There are also some more practical results on mapping short weighted sequences to a reference genome \cite{6}, a problem that has also been studied.
The Longest Common Subsequence (LCS) problem is a well-known measure of similarity between two strings. Given two strings, the output should be the length of the longest subsequence common to both strings. Dynamic programming solutions [17, 25] for this problem are classical textbook algorithms in Computer Science. LCS has been applied in computational biology for measuring the commonality of DNA molecules or proteins which may yield similar functionality. A very interesting survey on algorithms for the LCS can be found in [10]. The current LCS algorithms are considered optimal, since matching lower bounds (under the Strong Exponential Time Hypothesis) were proven [1, 11].

Extensions of this problem on more general structures have also been investigated (trees and matrices [4], run-length encoded strings [7], and more). One interesting variant of the LCS is the Heaviest Common Subsequence (HCS) where the matching between different letters is assigned a different weight, and the goal is to maximize the weight of the common subsequence, rather than its length.

The problem studied in this paper is the weighted LCS (WLCs) problem. It was introduced by Amir et al. [3] as an extension of the classical LCS problem on weighted sequences. Given two weighted sequences, the goal is to find a string which has a high probability of appearing in both sequences. Amir et al. initially solved an easier version of this problem in polynomial time, but unfortunately its applications are limited. As far as the general problem is concerned, they gave an NP-Hardness result on a closely related problem, which they call the log-probability version of WLCs. In short, the problem is the same, but all products in its definition are replaced with sums. Their proof is based on a Turing reduction and only works for unbounded alphabets. Finally, Amir et al. provide an \( \frac{1}{5} \)-approximation algorithm for the WLCs problem.

Cygan et al. [14] provided an NP-Completeness result on the decision log-probability version of WLCs, already for alphabets of size 2 (for alphabets of size 1 the solution is trivial since there is no uncertainty). They also gave an \( \frac{1}{2} \)-approximation algorithm and a PTAS, while also noticing that no FPTAS can exist, unless \( P = NP \). Finally, they proved that every instance of the problem can be reduced to a more restricted class of instances. However, for this to be achieved their algorithm needs to perform exact computations of roots and logarithms that may make the algorithm to err.

In this paper, we provide an NP-Completeness result on the general problem instead of the log-probability version, already for alphabets of size 2. Furthermore, we design an EPTAS for bounded alphabets which is also an improvement when compared to the PTAS of [14] for unbounded alphabets. Then, we prove that there is no EPTAS for unbounded alphabets by proving that the problem is \( W[1] \)-hard, and provide a matching conditional (under the Exponential Time Hypothesis - ETH) lower bound on any PTAS for this problem. We also prove that every instance of WLCs can be reduced to a restricted class of instances without using roots and logarithms, thus being able to actually achieve exact computations without rounding errors that can make the algorithm to err.

As noted in the previous paragraph, apart from providing more hardness results and faster algorithms we also circumvent the need to work with roots and logarithms as the previous results did. In short, by taking advantage of the property that \((ab)^c = a^c b^c\) and setting \( c \) to be an appropriate logarithm, previous results managed to transform any instance to a more manageable form. However, this transformation introduces an error that can make the algorithm to err as shown in Section 3. Table 1 summarizes the above discussion. Table 2 summarizes our results depending on the alphabet-size.
In Section 2, we provide necessary definitions and discuss the model of computation. In Section 3, we reduce the problem to only one threshold while showing how the previous approach to make this reduction can make the algorithm err. In Section 4, we show that the WLCS is NP-Complete while in Section 5, we provide the EPTAS algorithm for bounded alphabets and the improved PTAS for unbounded alphabets. In Section 6, we show that there can be no EPTAS for unbounded alphabets by showing that this problem is W[1]-hard and in Section 7, we describe the matching conditional lower bound. Finally, we conclude in Section 8.

Table 1 Results on WLCS.

<table>
<thead>
<tr>
<th></th>
<th>Amir et al.</th>
<th>Cygan et al.</th>
<th>Our results</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP-Hardness result</td>
<td>Log-probability version</td>
<td>Log-probability version</td>
<td>WLCS</td>
</tr>
<tr>
<td>Type of Reduction</td>
<td>Turing</td>
<td>Karp</td>
<td>Karp</td>
</tr>
<tr>
<td>Size of Alphabet</td>
<td>Unbounded</td>
<td>Already from size 2</td>
<td>Already from size 2</td>
</tr>
<tr>
<td>Reduction to a restricted class of instances</td>
<td>No</td>
<td>Yes, by assuming exact computations of logarithms</td>
<td>Yes</td>
</tr>
<tr>
<td>Approximation Algorithms</td>
<td>(1 - \frac{1}{n})-Approximation</td>
<td>PTAS</td>
<td>EPTAS for bounded alphabets, Improved PTAS for unbounded alphabets</td>
</tr>
<tr>
<td>Proof that no EPTAS exists for unbounded alphabets</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Lower bound on any PTAS</td>
<td>No</td>
<td>No</td>
<td>Matching the upper bound, under ETH</td>
</tr>
</tbody>
</table>

Table 2 Results depending on the Alphabet Size

<table>
<thead>
<tr>
<th>Alphabet Size</th>
<th>Previous Results</th>
<th>Our results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Trivial</td>
<td>Trivial</td>
</tr>
<tr>
<td>2\ldots O(1)</td>
<td>No FPTAS possible (unless (P = NP))</td>
<td>Achieved EPTAS</td>
</tr>
<tr>
<td>(\omega(1))</td>
<td>Achieved PTAS</td>
<td>No EPTAS possible (unless (FPT = W[1])), Improved PTAS</td>
</tr>
</tbody>
</table>

2 Preliminaries

2.1 Basic Definitions

Let \(\Sigma\) be a finite alphabet, \(\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_K\}\). We deal both with bounded (\(K = O(1)\)) and unbounded alphabets. \(\Sigma^d\) denotes the set of all words of length \(d\) over \(\Sigma\). \(\Sigma^*\) denotes the set of all words over \(\Sigma\).

Definition 1 (Weighted Sequence). A weighted sequence \(X = x_1x_2\ldots x_n\) of length \(|X| = n\) over an alphabet \(\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_K\}\) is a sequence of sets of pairs of the form (letter, probability). More formally:

\[x_i = \{(\sigma_j, p_{i}^{(X)}(\sigma_j)) : 1 \leq j \leq K\}\]
where $K_{j=1}^{K} p_i^{(X)}(\sigma_j) = 1$ for any $i$.

By $WS(\Sigma)$ we denote the set of all weighted sequences over $\Sigma$. Let $X \in WS(\Sigma)$. Let $Seq_d^{[X]}$ be the set of all increasing sequences of $d$ positions in $X$. For a string $s \in \Sigma^d$ and $\pi \in Seq_d^{[X]}$, define $P_X(\pi, s)$ as the probability that the substring on positions corresponding to $\pi$ in $X$ equals $s$. More formally, if $\pi = (i_1, i_2, \ldots, i_d)$ and $s_k$ denotes the $k$-th letter of $s$,

$$P_X(\pi, s) = \prod_{k=1}^{d} p_i^{(X)}(s_k)$$

Denote

$$SUBS(X, a) = \{ s \in \Sigma^* : \exists \pi \in Seq_d^{[X]}(P_X(\pi, s) \geq a) \}$$

That is, $SUBS(X, a)$ is the set of deterministic strings which match a subsequence of $X$ with probability at least $a$. Every $s \in SUBS(X, a)$ is called an $a$-subsequence of $X$.

The decision problem we consider is the following:

► **Definition 2** ($(a_1, a_2) - WLCS$ decision problem). Given two weighted sequences $X, Y$, two cut-off probabilities $a_1, a_2$ and a number $k$, find if the longest string $s \in SUBS(X, a_1) \cap SUBS(Y, a_2)$ has at least $k$ letters.

The $WLCS$ problem is the $(a_1, a_2) - WLCS$ problem, where $a_1 = a_2$. We denote these (equal) probabilities by $a$ ($a = a_1 = a_2$) for concreteness.

Let us note that the problem is only interesting if $|\Sigma| \geq 2$. For $|\Sigma| = 1$ the problem is trivial since there is no uncertainty at all. The same letter appears in every position in both strings with probability 1, and thus the answer is simply the length of the shorter weighted sequence.

### 2.2 Model of Computation

Our model of computation is the word $RAM$, with word size $w = \Omega(\log I)$, where $I$ is the input size in bits. Thus, arithmetic operations between words take constant time. However, due to the nature of our problem, it is necessary to compute products of many numbers. This can produce numbers that are much larger than the word size. We even allow numbers in the input to be larger than $2^w$ (these numbers just need to use more than one word to be represented). We generally assume that each number in the input is represented by at most $B$ bits, but we do not pose any constraint on $B$. Of course, in cases where we deal with numbers that occupy many words, we no longer have unit-cost arithmetic operations.

### 3 One Threshold is Enough

We show that the $(a_1, a_2) - WLCS$ and $WLCS$ are equivalent. Before doing that, let us assume that it is possible to multiply two $B$-bit numbers from the input in (polynomial) $Mul_w(B)$ time, where $w$ is the word-size. For example, for integers there exists a multiplication algorithm by Fürer [15] with time complexity $Mul_w(B) = O(B \log B 2^{O((\log^*) B)})$ (generally the running time can also depend on $w$, although in this case it doesn’t). We establish the complexity of multiplying $O(n)$ $B$-bit numbers. Our divide and conquer algorithm splits the numbers into two (equal sized) groups, recursively multiply each, and multiply the results in $Mul_w \left( \frac{B}{2} \right)$ time.
Lemma 3. Multiplying $O(n)$ $B$-bit numbers costs $O(Mul_W(nB) \log(nB))$ time.

Proof. Follows from the Master Theorem.

Corollary 4. We first provide a sketch of the proof. Our goal is to use the same weighted sequences algorithm is $O(\cdot)$ to reduce it to an instance $w$ with the letter $a$, $w$ is the length of the weighted sequences $X$ and $Y$, and $B$ is the maximum number of bits for the representation of a number in the input (not to be confused with the word-size $w$). Since a number may need many words to be represented.

Proof. We first provide a sketch of the proof. Our goal is to use the same weighted sequences with one additional position at the end. We introduce a new letter (‘#’) which only appears in this position, and we make sure that any correct algorithm will pick it, by making its probability very appealing (high). Since we can’t assign a probability higher than one, increasing it is simulated by reducing all other probabilities, in all positions. Knowing that this specific letter will be picked at this specific position allows us to choose the two corresponding probabilities in a way that completes the proof. Additionally, in order for the probabilities to sum to 1 in every position, we introduce two auxiliary letters (‘#’ and ‘\$’). Let $a = m^k a_1$. The new sequences $X'$ and $Y'$ are constructed as follows:

$$p_i^{(X')}(\sigma) = mp_i^{(X)}(\sigma), 1 \leq i \leq |X|, \sigma \in \Sigma$$

$$p_i^{(X')}(\#') = 1$$

$$p_i^{(X')}(\#) = 1 - \sum_{\sigma \in \Sigma \setminus \{\#\}} p_i^{(X)}(\sigma), 1 \leq i \leq |X|$$

$$p_i^{(Y)}(\sigma) = mp_i^{(Y)}(\sigma), 1 \leq i \leq |Y|, \sigma \in \Sigma$$

$$p_i^{(Y)}(\#') = \frac{a_1}{a_2}$$

$$p_i^{(Y)}(\$) = 1 - \sum_{\sigma \in \Sigma \setminus \{\$\}} p_i^{(Y)}(\sigma), 1 \leq i \leq |Y| + 1$$

All unspecified probabilities are equal to 0.

Suppose there exists a solution to $(X, Y, a_1, a_2, k)$. Then, there exist two increasing subsequences $\pi_1 = (i_1, \ldots, i_k), \pi_2 = (j_1, \ldots, j_k)$ and a string $s$ such that $P_X(\pi_1, s) \geq a_1, P_Y(\pi_2, s) \geq a_2$. Let $\pi_1 = (i_1, \ldots, i_k, |X| + 1), \pi_2 = (j_1, \ldots, j_k, |Y| + 1)$ and $s'$ be equal to $s$ extended with the letter ‘\$’. It holds that $P_X(\pi_1, s') = m^k P_X(\pi_1, s) \geq m^k a_1 = a, P_Y(\pi_2, s') = m^k P_Y(\pi_2, s) \geq m^k a_2 = a$.

Conversely, suppose there exists a solution to $(X', Y', a, k + 1)$. Then, there exist increasing subsequences $\pi_1 = (i_1, \ldots, i_{k+1}), \pi_2 = (j_1, \ldots, j_{k+1})$ and a string $s$ such that $P_X(\pi_1, s) \geq a, P_Y(\pi_2, s) \geq a$. First of all, notice that, due to $p_i^{(X')}(\$) = p_i^{(Y')}(\#') = 0$
Longest Common Subsequence on Weighted Sequences

for all \( i, s \) doesn’t contain letters ‘$’ and ‘#’. In addition, the letter ‘%’ only appears at the last position, and it is the only possible option for this position. Finally, the last position shall be used on both subsequences, because differently \( P_X((\pi_1, s), P_Y((\pi_2, s) \leq n^{k+1} < a. \)

Thus, the last letter of \( s \) is ‘%’. If we denote by \( s’ \) the string \( s \) without its last letter, it holds that \( P_X((i_1, ..., i_k), s’ \geq a_1, P_Y((j_1, ..., j_k), s’ \geq a_2.

The computation of \( a \) requires \( O(Mul_w(nB) \log (nB)) \) time due to Lemma [3] and the \( n|\Sigma| \)-multiplications of two numbers with at most \( B \) bits each cost \( O(n|\Sigma|Mul_w(B)) \). All other computations take linear time.

We note that [14] proved the same result, but their reduction required computations with real numbers (raising to the log-probability version to \( WLC\, S \)). For this reason, we believe that although it is easier to prove \( NP\)-Completeness for the integer log-probability version of the problem, there is no easy way to use it for proving

\[
\begin{array}{cccc}
X & 1 & 2 & 3 \\
\text{a} & 1 & 1 & 1 \\
\text{b} & 0 & 0 & 0 \\
\# & 0 & 0 & 0 \\
\end{array}
\quad
\begin{array}{cccc}
Y & 1 & 2 & 3 \\
\text{a} & x & % & 1 \\
\text{b} & 1 - x & % & 0 \\
\end{array}
\]

where \( 0 \leq x \leq 1 \) is a constant to be specified later. For \( x = 1 \), the weighted \( LCS \) is \( aaaa \) and for \( x < 1 \) the weighted \( LCS \) is \( aaaa \). The transformation described in [14] would give \( a = \frac{1}{\pi}, \gamma = \frac{1}{2} \) and the new sequences would be:

\[
\begin{array}{cccc}
\text{X'} & 1 & 2 & 3 \\
\text{a} & x^1 & % & 1 \\
\text{b} & (1 - x)^1 & % & 0 \\
\# & 1 - x^1 - (1 - x)^1 & 1 - 2 * \frac{1}{2} & 0 \\
\end{array}
\quad
\begin{array}{cccc}
\text{Y'} & 1 & 2 & 3 \\
\text{a} & x^1 & % & 1 \\
\text{b} & (1 - x)^1 & % & 0 \\
\# & 1 - x^1 - (1 - x)^1 & 1 - 2 * \frac{1}{2} & 0 \\
\end{array}
\]

Since \( \frac{1}{2} \) is an irrational number, it is rounded to some number \( r = \lfloor \frac{1}{2} \rfloor \). Suppose \( r < \frac{1}{2} \). In this case, when \( x = 1 \), while the weighted \( LCS \) is \( aaaa \) the algorithm will return \( aaaa \) due to the rounding errors. On the other hand, if \( r > \frac{1}{2} \), we can always find an appropriate \( x < 1 \) such that the weighted \( LCS \) should have been \( aaaa \) but the algorithm returns \( aaaa \) due to the rounding errors. To show this, let \( x = \left( \frac{k-1}{k} \right)^2 \) for some integer \( k \). Then \( x^2 = \left( \frac{k-1}{k} \right)^3 \).

It holds that \( \left( \frac{k-1}{k} \right)^3 r^2 \) is an increasing function of \( k \) which converges to \( r^2 > \frac{1}{2} \). Thus, we can find a big enough \( k \) such that \( x^2 r^2 \geq \frac{1}{2} \) and err on this particular example.

4 NP-Completeness

An NP-Completeness proof for the integer log-probability version of the \( WLC\, S \) problem has been given in [14]. This is a closely related problem, with the main difference being that products are replaced with sums. We know of no way to reduce from this log-probability version to \( WLC\, S \) other than exponentiating. As stated in our explanation of our model of computation in Section 2, there is no limit on the number of bits needed to represent a single number (it will just occupy a lot of words). This means that, if the input consisted of \( I \) bits, and there was a number (probability) represented with \( \frac{I}{\log} \) bits, exponentiating would result in a number with \( 2^{\frac{I}{\log}} \) bits, meaning the reduction would not be a polynomial-time one. For this reason, we believe that although it is easier to prove \( NP\)-Completeness for the integer log-probability version of the problem, there is no easy way to use it for proving
NP-Completeness for the general version. We, thus, give a reduction from Subset Product which proves NP-Completeness directly for the general problem.

Notice that for alphabets consisting of one letter, the problem is trivial since there is no uncertainty at all. In the following, we prove that even for alphabets consisting of two letters, the problem is NP-Complete.

**Definition 6 (Subset Product).** Given a set $L$ of $n$ integers and an integer $P$, find if there exists a subset of the numbers in $L$ with product $P$.

**Lemma 7.** $WLCS$ is NP-Complete, even for alphabets of size 2.

**Proof.** Obviously $WLCS \in NP$ since the increasing subsequences $\pi_1, \pi_2$ and the string $s$ for which $P_X(\pi_1, s) \geq a, P_Y(\pi_2, s) \geq a$ are a certificate which, along with the input, can be used to verify in polynomial time that the problem has a solution.

Let $(L, P)$ be an instance of Subset Product and let $n = |L|$. By $L_i$ we denote the $i$-th number of the set $L$. We give a polynomial-time reduction to a $(X, Y, a, k)$ instance of $WLCS$, with alphabet size 2 (we call the letters $'a'$ and $'b'$).

The core idea is the following: The weighted sequences will have $n$ positions (plus 2 more for technical reasons related to the threshold $a$). The number $k$ will be equal to the length of the sequences, meaning that we pick every position, and the only question is whether we picked letter $'a'$ or letter $'b'$. Letter $'a'$ in position $i$ will correspond to picking the $i$-th number in the original Subset Product, while letter $'b'$ corresponds to not picking it. Finally, the letters $'a'$ picked in $X$ will form an inequality of the form: "some product is $\geq P$", while the same letters in $Y$ will form the inequality: "the same product is $\leq P$". For these two to hold simultaneously, it must be the case that we found some product equal to $P$, which is the goal of the original Subset Product.

More formally, the weighted sequences have size $n + 2$. In the following, let $c_i = \frac{1}{1 + L_i}$ and $d_i = \frac{1}{1 + \frac{1}{L_i}}$

$$
\begin{align*}
p_i^{(X)}('a') &= c_i L_i, 1 \leq i \leq n \\
p_i^{(X)}('b') &= c_i, 1 \leq i \leq n \\
p_{n+1}^{(X)}('a') &= 1 \\
p_{n+1}^{(X)}('b') &= 0 \\
p_{n+2}^{(X)}('b') &= \frac{1}{P^2} \\
p_{n+2}^{(X)}('b') &= 1 - \frac{1}{P^2} \\
p_i^{(Y)}('a') &= \frac{d_i}{L_i}, 1 \leq i \leq n \\
p_i^{(Y)}('b') &= d_i, 1 \leq i \leq n \\
p_{n+1}^{(Y)}('a') &= \frac{\prod_{i=1}^n c_i}{\prod_{j=1}^n d_i} \\
p_{n+1}^{(Y)}('b') &= 1 - p_{n+1}^{(Y)}('a') \\
p_{n+2}^{(Y)}('a') &= 1
\end{align*}
$$
Longest Common Subsequence on Weighted Sequences

\[ p_{n+2}(b') = 0 \]

where \( c_i \) and \( d_i \) have been chosen so that probabilities sum to 1. Finally, we set \( k = n + 2 \) and \( a = \prod_{i=1}^{n} c_i \).

First of all, notice that since we must find a string of length \( n + 2 \), we must choose a letter from every position. Thus, a choice of letter at some position on \( X \) corresponds to the same choice of letter at that position on \( Y \). The choice of letter on positions \( n + 1 \) and \( n + 2 \) is in both cases since

\[ p_X(n+1)(b') = p_Y(n+2)(b') = 0 \]

Suppose that the numbers at positions \( \{i_1, \ldots, i_\ell\} \) give product \( P \):

\[ \ell \prod_{j=1}^{\ell} L_{i_j} = P \]

Then, we form the string \( s \) by picking \( \prime a' \) at positions \( \{i_1, \ldots, i_\ell, n + 1, n + 2\} \) and \( \prime b' \) at all other positions. Thus

\[ P_X(\{1, 2, \ldots, n + 2\}, s) = \frac{\prod_{j=1}^{\ell} L_{i_j} \prod_{j=1}^{n} c_i}{P^2} \geq a \]

\[ P_Y(\{1, 2, \ldots, n + 2\}, s) = \frac{\prod_{j=1}^{n} d_i \prod_{j=1}^{\ell} c_i}{\prod_{j=1}^{\ell} L_{i_j} \prod_{j=1}^{n} d_i} = a \]

Conversely, suppose a solution for the WLCS problem, where the string \( s \) is formed by picking \( \prime a' \) at positions \( \{i_1, \ldots, i_\ell, n + 1, n + 2\} \) and \( \prime b' \) at all other positions. It holds that:

\[ P_X(\{1, 2, \ldots, n + 2\}, s) = \frac{\prod_{j=1}^{\ell} L_{i_j} \prod_{j=1}^{n} c_i}{P^2} \geq a \rightarrow \prod_{j=1}^{\ell} L_{i_j} \geq P \]

\[ P_Y(\{1, 2, \ldots, n + 2\}, s) = \frac{\prod_{j=1}^{n} d_i \prod_{j=1}^{\ell} c_i}{\prod_{j=1}^{\ell} L_{i_j} \prod_{j=1}^{n} d_i} \geq a \rightarrow \prod_{j=1}^{\ell} L_{i_j} \leq P \]

The above imply that \( \prod_{j=1}^{\ell} L_{i_j} = P \). Finally, notice that all computations are done in polynomial time, due to Lemma 3.

\[ \blacksquare \]

5 EPTAS for Bounded Alphabets, Improved PTAS for Unbounded Alphabets

We now give an Efficient Polynomial Time Approximation Scheme (EPTAS) for the case where our alphabet size is bounded (\(|\Sigma| = O(1)|\)). Let us notice that this is the case when working with DNA sequences (\(|\Sigma| = 4\), the most usual application of weighted sequences. The same algorithm is an improved (when compared to [14]) PTAS in the case of unbounded alphabets. This means that the WLCS problem is Fixed-Parameter Tractable for constant size alphabets and thus belongs to the corresponding complexity class FPT as shown in Theorem 10.

The authors in [14] first noted that there is no FPTAS unless \( P = NP \), and so we can only hope for an EPTAS. Our result relies on their following result:
Lemma 8. It is possible to find, in polynomial time, a solution of size \(d\) to the WLCS optimization problem such that the optimal value \(\text{OPT}\) is guaranteed to be either \(d\) or \(d + 1\) (however we do not know which one holds).

Their PTAS uses the above result and in case the approximation is guaranteed to be good enough \((d > (1 - \epsilon)(d + 1))\), which implies that \(d > (1 - \epsilon)\text{OPT}\), it stops. Else, it holds that \(\frac{1}{\epsilon} \geq d + 1 \geq \text{OPT}\), and the PTAS exhaustively searches all subsequences of \(X\), all subsequences of \(Y\), and all possible strings of length \(d + 1\), for a total complexity of

\[
O \left( Mul_w \left( \frac{B}{\epsilon} \right) \log \left( \frac{B}{\epsilon} \right) |\Sigma|^\frac{n}{\epsilon} \right)^2
\]

\(\text{Mul}_w(\frac{B}{\epsilon})\log(\frac{B}{\epsilon})\) is insignificant compared to the other terms, and it is the time needed to multiply \(d + 1\) numbers with at most \(B\)-bits each. Our EPTAS improves the exhaustive search to

\[
O \left( Mul_w \left( \frac{B}{\epsilon} \right) \frac{n}{\epsilon} |\Sigma|^\frac{1}{\epsilon} \right)
\]

which is polynomial in the input size, in case of bounded alphabets. The following lemma is needed.

Lemma 9. Given a weighted sequence \(X\) of \(n\) positions, and a string \(s\) of length \(d\), it is possible to find the maximum number \(a\) such that there exists an increasing subsequence \(\pi\) of length \(d\) for which \(P_X(\pi, s) = a\). The running time of the algorithm is \(O(\text{Mul}_w(dB)n^d)\), where \(B\) is the maximum number of bits needed to represent each probability in \(X\).

Proof. We use dynamic programming. Let \(s_j\) be the string formed by the first \(j\) letters of \(s\), \(c_j\) be the \(j\)-th letter of \(s\) and \(\text{opt}_X(i, j)\) be the maximum number such that there exists an increasing subsequence \(\pi_j\)'s of length \(j\) whose last term \(\pi_j^\prime\) is at most \(i\) and for which \(P_X(\pi', s_j) = \text{opt}_X(i, j)\). Since we choose whether \(c_j\) is picked from the \(i-th\) position of \(X\), it holds that:

\[
\text{opt}_X(i, j) = \max \{ \text{opt}_X(i - 1, j), \text{opt}_X(i - 1, j - 1) P_i^{(X)}(c_j) \}
\]

For the base cases, \(\text{opt}_X(i, 0) = 1\) for all \(i\) (we can always form the empty-string with certainty, by not picking anything), and \(\text{opt}_X(0, j) = 0\) for \(j > 0\) (not picking anything will never give us a non-empty string). We are interested in the value \(\text{opt}_X(|X|, |s|)\).

Theorem 10. WLCS \(\in\text{FPT}\) for bounded alphabets.

Proof. For bounded alphabets, the number of strings of length \(\ell\) is \(|\Sigma|^\ell\), which is a function of \(\ell\). To test whether a given string \(s\) is a solution to our problem, we use Lemma[8] on both \(X\) and \(Y\). Let our parameter be \(d\), the length of the solution. Then, all possible strings are examined, in non-decreasing length, until a length is found where no string is a solution. The time complexity is \(O \left( \sum_{k=1}^{d+1} \text{Mul}_w(kB)k|\Sigma|^k/n \right) = O(\text{Mul}_w(dB)d|\Sigma|^{d+2}/n)\).

Now we are ready to give our EPTAS.

Theorem 11. For any value \(\epsilon \in (0, 1]\) there exists an \((1 - \epsilon)\)-approximation algorithm for the WLCS problem which runs in \(O \left( \text{poly}(I) + \frac{n}{\epsilon^2} \text{Mul}_w \left( \frac{B}{\epsilon} \right) |\Sigma|^\frac{1}{\epsilon} \right)\) time and uses \(O(\text{poly}(I))\) space, where \(I\) is the input size, \(n = |X| + |Y|\) and \(B\) is the maximum number of bits needed to represent a probability in \(X\) and \(Y\). Consequently, the WLCS problem admits an EPTAS for bounded alphabets.
We begin by using Lemma 8 to find an $a$-subsequence of length $d$, such that the optimal solution is at most $d + 1$. If $d + 1 \geq \frac{1}{2}$, we are done, since in that case we have an \[ \frac{d + 1}{2} = 1 - \frac{d + 1}{2} \geq (1 - \epsilon) \] approximation. Otherwise, we try all possible strings $s \in |\Sigma|^{d+1}$, and use Lemma 9 to check if any one of them can appear in both weighted sequences with probability at least $a$.

\section{No EPTAS for Unbounded Alphabets}

We have already seen that there is no \textit{FPTAS} for \textit{WLCS}, even for alphabets of size 2, unless $P = NP$. We have also shown an \textit{EPTAS} for bounded alphabets and a \textit{PTAS} for unbounded alphabets. The natural question that arises is: Is it possible to give an \textit{EPTAS} for unbounded alphabets?

We answer this question negatively, by proving that \textit{WLCS} is $W[1]$-hard, meaning that it doesn’t admit an \textit{EPTAS} (and is in fact not even in \textit{FPT}) unless $FPT = W[1]$. To show this, we give a 2-step $FPT$-reduction from Perfect Code, which was shown to be $W[1]$-Complete in \cite{12}, to $k$-sized Subset Product and then to \textit{WLCS}. The $k$-sized Subset Product problem is the Subset Product problem with the additional constraint that the target subset must be of size $k$.

\begin{definition}[Perfect Code] Given an undirected graph $G$ and a positive integer $k$, find if $G$ has a $k$-element perfect code. A perfect code is a set of vertices $V' \subseteq V$ with the property that for each vertex $u \in V$ there is precisely one vertex in $N_G(u) \cap V'$, where $N_G(u)$ is the set of adjacent nodes of $u$ in $G$.
\end{definition}

Notice that the definition of a perfect code implies that there is a perfect code iff there is a set $V' \subseteq V$ for which $\bigcup_{u \in V'} N_G(u) = V$ and $N_G(u) \cap N_G(v) = \emptyset$ for all $u, v \in V', u \neq v$. First we show that $k$-sized Subset Product is $W[1]$-hard.

\begin{lemma} \label{lem:k-sproj} $k$-sized Subset Product is $W[1]$-hard.
\end{lemma}

\begin{proof}
Let $(G = (V, E), k)$ be an instance of Perfect Code. Suppose that the vertices are $V = \{1, \ldots, n\}$. First of all, we compute the first $n$ prime numbers using the Sieve of Eratosthenes. We denote the $i$-th prime number as $p_i$. The set of positive integers $L = \{L_1, L_2, \ldots, L_n\}$ as well as the positive integer $P$ are defined as follows:

\[ L_v = \prod_{u \in N_G(v)} p_u, \quad P = \prod_{v=1}^n p_v \]

Notice that due to the unique prime factorization theorem, a subset of $k$ numbers from the set $L$ have product $P$ iff $G$ has a $k$-element Perfect Code.

The size of our primes is $O(n \log n)$ due to the prime number theorem. Thus, they require $O(\log n)$ bits to be represented. Each integer in $L$, as well as $P$, are computed using Corollary 4 in $O(n \log^3(n) 2^{O(\log^* n)})$ time, for an overall $O(n^2 \log^3(n) 2^{O(\log^* n)})$ complexity for our reduction. Since the new parameter $k$ is the same as the old one (no dependence on $n$), our reduction is in fact an \textit{FPT}-reduction.

Our result for this section is the following.

\begin{theorem} \label{thm:nep} \textit{WLCS} is $W[1]$-hard.
\end{theorem}

\begin{proof}
To prove the theorem we create diagonal weighted sequences. That is, we require each letter to appear only in one position and vice-versa. In this way, the subsequences
picked for $X$ and $Y$ will be the same. The above rule will be broken by the addition of two auxiliary letters that are there to make the probabilities add up to 1 in each position. This creates no problem because we make sure that these letters are never picked. Finally, we force the product to be equal to our target, by forcing it to be at most our target and at least our target at the same time.

More formally, let $(L = \{L_1, L_2, \ldots, L_n\}, k, P)$ be an instance of the $k$-sized Subset Product problem and let $M = m^{k+1}$, where $m$ is the maximum number in set $L$. Notice that if $m^k \leq P$ then we only need to check the product of the highest $k$ numbers of $L$, which means the problem is solvable in polynomial time. Thus we can assume that $M \geq m^k > P$.

The alphabet of $X, Y$ is $\Sigma = \{1, 2, \ldots, n, n+1, n+2, n+3\}$ and we set $a = \frac{1}{PM^k}$.

$$p_i^{(X)}(i) = \frac{L_i}{M}, \quad p_i^{(Y)}(i) = \frac{1}{ML_i}, \quad 1 \leq i \leq n$$

$$p_{n+1}^{(X)}(n+1) = \frac{1}{P^2}, \quad p_{n+1}^{(Y)}(n+1) = 1$$

$$p_i^{(X)}(n+2) = 1 - p_i^{(X)}(i), \quad 1 \leq i \leq n+1$$

$$p_i^{(Y)}(n+3) = 1 - p_i^{(Y)}(i), \quad 1 \leq i \leq n+1$$

All non-specified probabilities are equal to 0. Notice that symbols $n+2$ and $n+3$ are used to guarantee that probabilities sum up to 1.

We show that the instance $(X, Y, a, k+1)$ has a solution iff $(L, k, P)$ has a solution. Suppose there exists a solution to $(L, k, P)$. Then, there exists an increasing subsequence $\pi = (i_1, \ldots, i_k)$ such that $\prod_{j=1}^{k} L_{i_j} = P$. Let $\pi'$ be $\pi$ extended by the number $n+1$ and $s$ be the string $i_1i_2\ldots i_{k+1}$. It holds that $P_X^{\pi'}(s) = P_Y^{\pi'}(s) = a$.

Conversely, suppose there exists a solution to $(X, Y, a, k+1)$. Then there exist increasing subsequences $\pi = (i_1, \ldots, i_{k+1}), \pi' = (j_1, \ldots, j_{k+1})$ and a string $s$ such that $P_X^{\pi}(s) \geq a, P_Y^{\pi'}(s) \geq a$. First of all, notice that, due to $p_i^{(X)}(n+3) = p_i^{(Y)}(n+2) = 0$ for all $i$, $s$ doesn’t contain letters $n+2$ and $n+3$, which leaves only one choice for every position. Also each letter appears only once in each sequence, and in the same position. Thus, $\pi = \pi'$, and due to our construction the $i-th$ letter of $s$ is the $i-th$ member of $\pi$. Finally, not picking position $n+1$ would result in $P_Y^{(\pi', s)} < a$ due to our definition of $M$. Thus, the last letter of $s$ is $n+1$. It holds that:

$$P_X^{\{i_1, \ldots, i_k\}, s} \geq a \Rightarrow \prod_{i=1}^{k} L_{\pi_i}/P^M \geq \frac{1}{PM^k} \Rightarrow \prod_{i=1}^{k} L_{\pi_i} \geq P$$

$$P_Y^{\{i_1, \ldots, i_k\}, s} \geq a \Rightarrow \frac{1}{M^k \prod_{i=1}^{k} L_{\pi_i}} \geq \frac{1}{PM^k} \Rightarrow \prod_{i=1}^{k} L_{\pi_i} \leq P$$

The above two inequalities imply a $k$-sized subset of $L$ with product equal to $P$.

The reduction is a polynomial-time one, due to Lemma $\ref{lemma:reduction}$. More than that, it is an FPT-reduction since the new parameter $k$ is equal to the old parameter incremented by one, and thus has no dependence on $n$.

### 7 Matching Conditional Lower Bound on any PTAS

Patrascu et al. $\cite{Patrascu}$ proved that any algorithm for solving the $d$-SUM problem requires $n^{\Omega(d)}$ time, unless the Exponential Time Hypothesis (ETH) fails. To show this, they proved the following.
Proposition 15. Let $f$ be an 1-in-3 SAT instance with $n$ variables and $O(n)$ clauses, where each variable appears only in a constant number of clauses. Under ETH, there is an (unknown) constant $s_3$ such that there exists no algorithm to solve $f$ in $O(2^{s_3n})$ time for $\delta < s_3$.

By assuming an $n^{o(d)}$ time algorithm for $d$-SUM they disproved the above fact, which can’t happen under ETH. We use the same technique for proving an $n^{O(d)}$ lower bound for $k$-sized Subset Product.

Lemma 16. Assuming the ETH, the problem of $k$-sized Subset Product can’t be solved in $O(n^{\frac{d}{k+1}}), k < n^{0.99},$ where $n$ is the number of integers each one of which has at most $O(\log n(\log k + \log \log n))$ bits, and $P$ is the target which can be arbitrarily big.

Proof. Let $f$ be an 1-in-3 SAT instance with $N$ variables and $M = O(N)$ clauses, where each variable appears only in a constant number of clauses, and $k > \frac{1}{\sqrt{N}}$. Conceptually, we split the variables of $f$ into $k$ blocks of equal size - apart from the last block that may have smaller size. Each block contains at most $\frac{N}{k}$ variables, and thus there are at most $2^N$ different assignments of values to the group-of-variables within a block. For each block and for each one of these assignments we generate a number which serves as an identifier of the corresponding block and assignment. Thus, there will be $n = k2^\frac{N}{k}$ different identifiers.

Let $p_i$ be the $i$-th prime number. In order to compute an identifier, we initialize it to $p_b$, where $b$ is the index of the identifier’s corresponding block. Then, we run through all of the $M = O(N)$ clauses and do the following: suppose we process the $i$-th clause and let $0 \leq j \leq 3$ be the number of variables of the identifier’s corresponding assignment that satisfy the clause. We update the identifier by multiplying it with $p_i^j$.

Since each variable appears only in a constant number of clauses, each identifier will be a product of at most $O(\frac{N}{k})$ numbers. The prime number theorem guarantees at most $O(\log N)$ bits to represent each factor, which means the identifiers will have at most $O(\frac{N}{k}\log N)$ bits. Using the fact that $n = k2^\frac{N}{k}$, each identifier is represented by $O(\log n(\log k + \log \log n))$ bits.

These $n$ identifiers, along with the target $P = \prod_{i=1}^{k+1} p_i$ (recall that $p_i$ is the $i$-th prime number), form a $k$-sized Subset Product instance. This preprocessing step costs $O(2^\frac{N}{k})$ time, ignoring polynomial terms, which is more efficient than $O(2^{s_3N})$.

Due to the unique prime factorization, a solution to the $k$-sized Subset Product corresponds to a solution in $f$ and vice-versa. If the running time of the $k$-sized Subset Product was $O(n^{\frac{d}{k+1}})$ then we could solve the above instance in $O((k2^\frac{N}{k})^{\frac{d}{k+1}})$ time.

Since $k = \frac{n^{0.99}}{2^\frac{1}{N}}$ and $k < n^{0.99}$, it follows that $\frac{n}{k} < n^{0.99} \implies n^{0.99} < 2^{99N}$. But $k < n^{0.99}$, which means $k < 2^{99N}$. Thus the previous running time becomes $O(2^{100N+s_3N})$. Both the preprocessing step and the solution of the $k$-sized Subset Product can be achieved in time $O(2^{6N})$, where $\delta < s_3$. However, this would violate Proposition 15.

Using the above, we are ready to prove our (matching) lower bound, conditional on ETH.

Theorem 17. Under ETH, there is no PTAS for $W LCS$ with running time $|I|^{o(\frac{1}{\delta})}$, where $|I|$ is the input size in bits.

Proof. Suppose that such an algorithm $A(I, \epsilon)$ existed. Let $R()$ be the polynomial time reduction from $k$-sized Subset Product to $W LCS$ given in the proof of Theorem 14. Then,
there is a solution to \( k \)-sized Subset Product iff there is a solution to \( WLCS \) of size \( k + 1 \), or, equivalently, iff the optimal solution to \( WLCS \) is at least \( k + 1 \).

Using the hypothetical \( A(I, \epsilon) \) with an appropriate value of \( \epsilon \), we will solve \( k \)-sized Subset Product more efficiently than possible, thus reaching a contradiction.

Consider the following algorithm for \( k \)-sized Subset Product, where there are \(|L|\) numbers in the input, each having \( O(\log |L|)(\log k + \log \log |L|) \) bits and \( k < |L|^{0.99} \). Given an instance \((L, k, P)\), we define the instance for the \( WLCS \) to be \( I = R(L, k, P) \). We run \( A(I, \frac{1}{2k+1}) \) and if the output is at least \( k + 1 \) we return that \((L, k, P)\) is satisfied, otherwise we return that it cannot be satisfied.

Note that if \( k \)-sized Subset Product is solvable, then \( OPT(I) \geq k + 1 \), and the value output by \( A \) is at least \((1 - \frac{1}{2k+1})(k+1) = k + \frac{1}{2} > k \). Thus, the value output by \( A \) is at least \( k + 1 \). On the other hand, if \( k \)-sized Subset Product is not solvable, then \( OPT(I) < k + 1 \), and obviously the value output by \( A \) is at most \( k \).

Thus we found an algorithm for \( k \)-sized Subset Product whose running time is \(|I|^{o(k)} \). Since \( I \) is obtained by a polynomial time reduction, its size is at most polynomially bigger than \(|(L, k, P)|\). Thus, the above running time becomes \(|(L, k, P)|^{o(k)} \). Under our assumptions, this becomes \(|L|^{o(k)} \), which is not feasible under \( ETH \), due to Lemma 16.

8 Conclusion

We managed to prove \( NP \)-Completeness for the \( WLCS \) decision problem, and give a \( PTAS \) along with a matching conditional lower bound for the optimization problem. In the most usual setting, where the alphabet size is constant, the above \( PTAS \) is in fact an \( EPTAS \), and it is known that no \( FPTAS \) can exist unless \( P = NP \). Finally we gave a transformation such that algorithms for the \( WLCS \) problem can be applied for the \((a_1, a_2) - WLCS \) problem.

In proving that \( WLCS \) doesn’t admit any \( EPTAS \), we proved that it is \( W[1] - hard \). It may be interesting to determine the exact complexity of \( WLCS \) in the \( W[\|] - hierarchy \).

References


