A Central Limit Theorem for the Length of the Longest Common Subsequence in Random Words

Christian Houdre, Ümit Islak

A Central Limit Theorem for the Length of the Longest Common Subsequence in Random Words

Christian Houdré[*‡] Ümit Işlak[†‡]

September 12, 2014

Abstract

Let $(X_k)_{k\geq 1}$ and $(Y_k)_{k\geq 1}$ be two independent sequences of independent identically distributed random variables having the same law and taking their values in a finite alphabet. Let LC_n be the length of longest common subsequences in the two random words $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$. Under assumptions on the distribution of X_1, LC_n is shown to satisfy a central limit theorem. This is in contrast to the limiting distribution of the length of longest common subsequences in two independent uniform random permutations of $\{1, \ldots, n\}$, which is shown to be the Tracy-Widom distribution.

[*]School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia, 30332-0160, houdre@math.gatech.edu. Research supported in part by a Simons Foundation Fellowship, grant #267336. Many thanks to both the LPMA of the Université Pierre et Marie Curie and to CIMAT for their hospitality while part of this research was carried out.

[†]Department of Mathematics, University of Southern California, Los Angeles, California, 90089-2532, islak@usc.edu. I am grateful to L. Goldstein for introducing me to Stein's method and, in particular, to Chatterjee's normal approximation theorem. Also, many thanks to the LPMA of the Université Pierre et Marie Curie for its hospitality while part of this research was carried out.

[‡]Both authors would like to thank the French Scientific Attachés Fabien Agenes and Nicolas Florsch for their consular help. Without them, this research might not have existed.

Keywords: Longest Common Subsequence, Central Limit Theorem, Optimal Alignment, Last Passage Percolation, Stein's Method, Tracy-Widom Distribution, Supersequences.

MSC 2010: 05A05, 60C05, 60F05.

1 Introduction

We study below the asymptotic behavior, in law, of the length of the longest common subsequence of two random words. Although it has been extensively studied from an algorithmic point of view in various disciplines such as, computer science, bio-informatics, or statistical physics, to name but a few of them, theoretical results on the longest common subsequence are rather sparse. To present our framework, let $X = (X_i)_{i\geq 1}$ and $Y = (Y_i)_{i\geq 1}$ be two infinite sequences whose coordinates take their values in $\mathcal{A}_m = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_m\}$, a finite alphabet of size m.

Next, LC_n, the length of the Longest Common Subsequences (LCS) of the random words $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$, is the maximal integer $k \in \{1, \ldots, n\}$, such that there exist $1 \leq i_1 < \cdots < i_k \leq n$ and $1 \leq j_1 < \cdots < j_k \leq n$, such that:

$$X_{i_s} = Y_{j_s}, \qquad \text{for} \quad \text{all} \qquad s = 1, 2, \ldots, k.$$

LC_n is a measure of the similarity/dissimilarity of the words which is often used in pattern matching and the asymptotic behavior of its law is the purpose of our study. In computer science, $2(n - LC_n)$ is the edit (or Levenshtein) distance which is the minimal number of indels (insertions/deletions) to transform one word into the other.

The study of LC_n has a long history starting with the well known result of Chvátal and Sankoff [7] asserting that

$$\lim_{n\to\infty} \frac{\mathbb{E}LC_n}{n} = \gamma_m^*. \tag{1.1}$$

However, to this day, the exact value of γ_m^* (which depends on the distribution of X_1 and on the size of the alphabet) is unknown, even in "simple cases", such as for uniform Bernoulli random variables. Nevertheless, its asymptotic behavior as the alphabet size grows is given, for X_1 uniformly distributed, by:

$$\lim_{m\to\infty} \sqrt{m}\gamma_m^* = 2, \tag{1.2}$$

as shown by Kiwi, Loebl and Matoušek ([14]).

Chvátal and Sankoff's first asymptotic result was sharpened by Alexander ([1]) who proved that

$$\gamma_m^* n - C_A\sqrt{n \log n} \leq \mathbb{E}LC_n \leq \gamma_m^* n, \tag{1.3}$$

where $C_A > 0$ is a universal constant (depending neither on n nor on the distribution of X_1). Next, Steele [20] was the first to obtain the upper-order of the variance proving, in particular, that $\operatorname{Var} LC_n \leq n$, but finding a lower-order bound is more illusive. For Bernoulli random variables and/or in various instances where there is a strong "bias" such as high asymmetry or mixed common and increasing subsequence problems, the lower bound is also shown to be of order n ([10], [12], [15]). In all these cases, the central r-th, $r \geq 1$, moment of LC_n can also be shown to be of order $n^{r/2}$ (see the concluding remarks in [11]). This strongly hints at the asymptotic normality of LC_n, in these contexts, although similar moments estimates can lead to a non-Gaussian limiting law in a related model ([3]). Here is our main result:

Theorem 1.1 *Let $(X_i)_{i\geq 1}$ and $(Y_i)_{i\geq 1}$ be two independent sequences of iid random variables with values in $\mathcal{A}_m = \{\alpha_1, \alpha_2, ..., \alpha_m\}$, and let*

$$\mathbb{P}(X_1 = \alpha_k) = \mathbb{P}(Y_1 = \alpha_k) = p_k, \qquad k = 1, 2, ..., m.$$

Let $p_{j_0} > 1/2$, for some $j_0 \in \{1, ..., m\}$, let $K = \min(2^{-4}10^{-2}e^{-67}, 1/800m)$, and let $\max_{j\neq j_0} p_j \leq \min\{2^{-2}e^{-5}K/m, K/2m^2\}$. Then for all $n \geq 1$,

$$d_W\left(\frac{LC_n - \mathbb{E}LC_n}{\sqrt{\operatorname{Var} LC_n}}, \mathcal{G}\right) \leq C\frac{1}{n^{1/8}}, \tag{1.4}$$

where d_W is the Monge-Kantorovich-Wasserstein distance, $\mathcal{G}$ a standard normal random variable and where $C > 0$ is a constant independent of n.

The above result is the first of its kind. It contrasts, in particular, with the related Bernoulli matching problem where the limiting law is the Tracy-Widom distribution ([16]). Both the LCS and Bernoulli matching models are last passage percolation models with respectively dependent and independent weights, possibly explaining the different limiting laws. In both cases, the expectation is linear in n, but the variance in Bernoulli matching is sublinear (of order $n^{2/3}$), while in our LCS case it is linear. Theorem 1.1 further contrasts with the corresponding limiting law for the length of the longest common subsequences in a pair of independent uniform random permutations of $\{1, \ldots, n\}$. In that problem, the emergence of the Tracy–Widom distribution has sometimes been speculated, and we show in the last section of the paper that this hypothesis is indeed true (the expectation there is of order $\sqrt{n}$ and the variance of order $n^{1/3}$).

As far as the content of the paper is concerned, the next section contains the proof of Theorem 1.1, and a remark discussing some elements of this proof. Then, in the last section, various extensions and generalizations as well as some related open questions are discussed. In particular, the proof, that the length of longest common subsequences in uniform random permutations converges to the Tracy-Widom distribution, is included there.

2 Proof of Theorem 1.1

The aim of this section is to provide a proof of our main theorem by a three step method. First making use of a recent result of Chatterjee ([4]) on Stein's method (see [6] for an overview of the method, including Chatterjee's normal approximation result), second using moment estimates for LC_n ([11]) and third developing correlation estimates based, in part, on short string-lengths genericity results obtained in [13]. We start by fixing notation and recalling some preliminaries.

Throughout the paper, $X = (X_i)_{i\geq 1}$ and $Y = (Y_i)_{i\geq 1}$ denote two independent infinite sequences whose coordinates are iid and take their values in $\mathcal{A}_m = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_m\}$, a finite alphabet of size m. Recall next that the Kolmogorov and Monge-Kantorovich-Wasserstein distances, d_K and d_W, between two probability distributions ν_1 and ν_2 on $\mathbb{R}$, are respectively defined as

$$d_K(\nu_1, \nu_2) = \sup_{h\in\mathcal{H}_1} \left| \int h d\nu_1 - \int h d\nu_2 \right|,$$

where $\mathcal{H}_1 = \{\mathbf{1}_{(-\infty,x]} : x \in \mathbb{R}\}$, and

$$d_W(\nu_1, \nu_2) = \sup_{h\in\mathcal{H}_2} \left| \int h d\nu_1 - \int h d\nu_2 \right|,$$

where $\mathcal{H}_2 = \{h : \mathbb{R} \to \mathbb{R} : |h(x) - h(y)| \leq |x - y|\}$. If ν_2 is absolutely continuous with its density bounded by C, then

$$d_K(\nu_1, \nu_2) \leq \sqrt{2C d_W(\nu_1, \nu_2)}, \tag{2.1}$$

as seen, for example, in [18] where a proof of (2.1) is also given. Thus, Theorem 1.1 implies via (2.1), that

$$d_K\left(\frac{LC_n - \mathbb{E}LC_n}{\sqrt{\operatorname{Var} LC_n}}, \mathcal{G}\right) \leq C\left(\frac{2}{\pi}\right)^{1/4} \frac{1}{n^{1/16}}. \tag{2.2}$$

Both (1.4) and (2.2) imply that, properly centered and normalized, LC_n converges in distribution to a standard normal random variable.

Let us continue by introducing some more notation following those of [4]. Let $W = (W_1, W_2, \ldots, W_n)$ and $W' = (W_1', W_2', ..., W_n')$ be two independent identically distributed $\mathbb{R}^n$-valued random vectors whose components are also independent. For $A \subset [n] := \{1, 2, \ldots, n\}$, define the random vector W^A by setting

$$W_i^A = \begin{cases} W_i' & \text{if } i \in A \\ W_i & \text{if } i \notin A, \end{cases}$$

with for $A = \{j\}$, and further ease of notation, we write W^j for $W^{\{j\}}$.

For a given Borel measurable function $f : \mathbb{R}^n \to \mathbb{R}$ and $A \subset [n]$, let

$$T_A := \sum_{j \notin A} \Delta_j f(W) \Delta_j f(W^A),$$

where

$$\Delta_j f(W) := f(W) - f(W^j).$$

Finally, let

$$T = \frac{1}{2} \sum_{A \subsetneq [n]} \frac{T_A}{\binom{n}{|A|}(n - |A|)},$$

where $|A|$ denotes the cardinality of A. Here is Chatterjee's normal approximation result.

Theorem 2.1 *[4] Let all the terms be defined as above, and let* $0 < \sigma^2 := \operatorname{Var} f(W) < \infty$. *Then,*

$$d_W \left(\frac{f(W) - \mathbb{E} f(W)}{\sqrt{\operatorname{Var} f(W)}}, \mathcal{G} \right) \leq \frac{\sqrt{\operatorname{Var} T}}{\sigma^2} + \frac{1}{2\sigma^3} \sum_{j=1}^{n} \mathbb{E} |\Delta_j f(W)|^3, \tag{2.3}$$

where $\mathcal{G}$ *is a standard normal random variable.*

Remark 2.1 *In [4], the variance term obtained in Theorem 2.1 is actually* $\operatorname{Var} \mathbb{E}(T | f(W))$ *but the above bound, with the larger* $\operatorname{Var} T$, *already present in [4], is enough for our purpose.*

Two notes before we begin the proof of Theorem 1.1.

(1) We do not keep track of constants in the proof since doing so would make the argument a lot lengthier. Therefore, a constant C may vary from an expression to another. Note, however, that C will always be independent of n.

(2) We do not worry about having quantities (e.g. length of longest common subsequences of two random words) like $\log n, n^{\alpha}$, *etc.* which should actually be $[n^{\alpha}]$, $[\log n]$, etc. This does not cause any problems as we are interested in asymptotic bounds. The proof can be revised with minor changes (and some notational burden) to make the statements more precise.

Let us start the proof of Theorem 1.1 and to do so, let

$$W = (X_1, \dots, X_n, Y_1, \dots, Y_n), \tag{2.4}$$

and let

$$f(W) = LC_n(X_1 \cdots X_n; Y_1 \cdots Y_n).$$

We begin by estimating the second term on the right-hand side of (2.3), and to do so, recall Theorem 1.1 of [11].

Theorem 2.2 *[11] Let the hypotheses of Theorem 1.1 hold, and let* $1 \leq r < \infty$. *Then, there exists a constant* $C > 0$ *depending on* r, m, p_{j_0} *and* $\max_{j \neq j_0} p_j$, *such that, for all* $n \geq 1$,

$$\mathbb{E}|LC - \mathbb{E}LC_n|^r \geq Cn^{r/2}. \tag{2.5}$$

Using the estimate in (2.5) with $r = 2$, we have

$$\sigma^2 = \operatorname{Var} LC_n \geq Cn, \qquad n \geq 1,$$

where C is a constant independent of n. Therefore,

$$\sigma^3 \geq Cn^{3/2}, \qquad n \geq 1, \tag{2.6}$$

yielding

$$\frac{1}{2\sigma^3} \sum_{j=1}^{2n} \mathbb{E}|\Delta_j f(W)|^3 \leq C \frac{1}{\sqrt{n}}, \tag{2.7}$$

since $|\Delta_j f(W)| \leq 1$.

Next we move to the estimation of the variance term in (2.3). Setting

$$\mathcal{S}_1 := \{(A, B, j, k) : A \subsetneqq [2n], B \subsetneqq [2n], j \notin A, k \notin B\}, \tag{2.8}$$

$\operatorname{Var} T$ can be expressed as

$$\begin{aligned}
\operatorname{Var} T &= \frac{1}{4} \operatorname{Var} \left(\sum_{A \subsetneqq [2n]} \sum_{j \notin A} \frac{\Delta_j f(W) \Delta_j f(W^A)}{\binom{2n}{|A|}(2n - |A|)} \right) \\
&= \frac{1}{4} \sum_{A \subsetneqq [2n], j \notin A} \sum_{B \subsetneqq [2n], k \notin B} \frac{Cov(\Delta_j f(W) \Delta_j f(W^A), \Delta_k f(W) \Delta_k f(W^B))}{\binom{2n}{|A|}(2n - |A|)\binom{2n}{|B|}(2n - |B|)} \\
&= \frac{1}{4} \sum_{(A,B,j,k) \in \mathcal{S}_1} \frac{Cov(\Delta_j f(W) \Delta_j f(W^A), \Delta_k f(W) \Delta_k f(W^B))}{\binom{2n}{|A|}(2n - |A|)\binom{2n}{|B|}(2n - |B|)}. \qquad (2.9)
\end{aligned}$$

Our strategy is now to divide $\mathcal{S}_1$ into several pieces and then to estimate the contributions of each piece separately. The following proposition, and a conditional version of it which easily follows from similar arguments, will be used repeatedly throughout the proof.

Proposition 2.1 *Let $\mathcal{R}$ be a subset of $[2n]^2$, and let*

$$\mathcal{S}^* = \{(A, B, j, k) : A \subsetneqq [2n], B \subsetneqq [2n], j \notin A, k \notin B, (j, k) \in \mathcal{R}\}.$$

Let $g : \mathcal{S}^* \to \mathbb{R}$ be such that $\|g\|_\infty \le C$, then

$$\sum_{(A,B,j,k) \in \mathcal{S}^*} \left| \frac{g(A, B, j, k)}{\binom{2n}{|A|}(2n - |A|)\binom{2n}{|B|}(2n - |B|)} \right| \le C|\mathcal{R}|.$$

Proof. First, observe that since $\|g\|_\infty \le C$,

$$\begin{aligned}
&\sum_{(A,B,j,k) \in \mathcal{S}^*} \left| \frac{g(A, B, j, k)}{\binom{2n}{|A|}(2n - |A|)\binom{2n}{|B|}(2n - |B|)} \right| \\
&\quad \le C \sum_{(A,B,j,k) \in \mathcal{S}^*} \left(\frac{1}{\binom{2n}{|A|}(2n - |A|)\binom{2n}{|B|}(2n - |B|)} \right).
\end{aligned}$$

Expressing $\sum_{(A,B,j,k) \in \mathcal{S}^*}$ in terms of $\mathcal{R}$, using basic results about binomial coefficients and performing some elementary manipulations lead to

$$\sum_{(A,B,j,k) \in \mathcal{S}^*} \frac{1}{\binom{2n}{|A|}(2n - |A|)\binom{2n}{|B|}(2n - |B|)}$$

$$
\begin{aligned}
&= \sum_{(j,k)\in\mathcal{R}} \sum_{\substack{A\subsetneq[2n]:A\not\ni j \\ B\subsetneq[2n]:B\not\ni k}} \frac{1}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\
&= \sum_{(j,k)\in\mathcal{R}} \left(\sum_{s,r=0}^{2n-1} \sum_{\substack{A\not\ni j,|A|=s \\ B\not\ni k,|B|=r}} \frac{1}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \right) \\
&= \sum_{(j,k)\in\mathcal{R}} \left(\sum_{s,r=0}^{2n-1} \sum_{\substack{A\not\ni j,|A|=s \\ B\not\ni k,|B|=r}} \frac{1}{\binom{2n}{s}(2n-s)\binom{2n}{r}(2n-r)} \right) \\
&= \sum_{(j,k)\in\mathcal{R}} \left(\sum_{s,r=0}^{2n-1} \frac{\binom{2n-1}{s}\binom{2n-1}{r}}{\binom{2n}{s}(2n-s)\binom{2n}{r}(2n-r)} \right) \\
&= \sum_{(j,k)\in\mathcal{R}} \left(\sum_{s,r=0}^{2n-1} \frac{1}{(2n)^2} \right) \\
&= |\mathcal{R}|,
\end{aligned}
$$

from which the result follows. □

Clearly, taking $\mathcal{R} = [2n]^2$, Proposition 2.1 yields the estimate

$$
\sum_{(A,B,j,k)\in\mathcal{S}_1} \left(\frac{Cov(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \right) \leq 8n^2,
$$

which is not good enough for our purposes, and we therefore begin a detailed estimation study to improve the order of the upper bound to $o(n^2)$.

To do so, we start by giving a slight variation of a result from [13] which can be viewed as a microscopic short-lengths genericity principle, and which will turn out to be an important tool in our proof. This principle, valid not only for common sequences but in much greater generality (see [13]), should prove useful in other contexts.

Assume that $n = vd$, and let the integers

$$
r_0 = 0 \leq r_1 \leq r_2 \leq r_3 \leq ... \leq r_{d-1} \leq r_d = n, \tag{2.10}
$$

be such that

$$
LC_n = \sum_{i=1}^{d} |LCS(X_{v(i-1)+1}X_{v(i-1)+2}\cdots X_{vi}; Y_{r_{i-1}+1}Y_{r_{i-1}+2}\cdots Y_{r_i})|, \tag{2.11}
$$

where $|LCS(X_{v(i-1)+1}X_{v(i-1)+2}\cdots X_{vi}; Y_{r_{i-1}+1}Y_{r_{i-1}+2}\cdots Y_{r_i})|$ is the length of the longest common subsequence of the words $X_{v(i-1)+1}X_{v(i-1)+2}\cdots X_{vi}$ and $Y_{r_{i-1}+1}Y_{r_{i-1}+2}\cdots Y_{r_i}$ (with the understanding that this length is zero if the X-part is aligned with gaps). Next, let $\epsilon > 0$ and let $0 < s_1 < 1 < s_2$, be two reals such that

$$\tilde{\gamma}(s_1) < \tilde{\gamma}(1) = \gamma_m^* \quad \text{and} \quad \tilde{\gamma}(s_2) < \tilde{\gamma}(1) = \gamma_m^*$$

where

$$\tilde{\gamma}(s) = \lim_{n\to\infty} \frac{\mathbb{E}LC_n(X_1\cdots X_n; Y_1\cdots Y_{sn})}{n(1+s)/2}, \quad s > 1.$$

(See [13] for the existence of, and estimates on, s_1 and s_2.) Finally, let E^n_{ϵ,s_1,s_2} be the event that for all integer vectors $(r_0, r_1, ..., r_d)$ satisfying (2.10) and (2.11), we have

$$|\{i \in [d] : vs_1 \le r_i - r_{i-1} \le vs_2\}| \ge (1-\epsilon)d. \tag{2.12}$$

Then, E^n_{ϵ,s_1,s_2} enjoys the following concentration property:

Theorem 2.3 *[13] Let $\epsilon > 0$. Let $0 < s_1 < 1 < s_2$ be such that $\tilde{\gamma}(s_1) < \tilde{\gamma}(1) = \gamma_m^*$ and $\tilde{\gamma}(s_2) < \tilde{\gamma}(1) = \gamma_m^*$, and let $\delta \in (0, \min(\gamma_m^* - \tilde{\gamma}(s_1), \gamma_m^* - \tilde{\gamma}(s_2)))$. Let the integer v be such that*

$$\frac{1+\log(v+1)}{v} \le \frac{\delta^2\epsilon^2}{16}. \tag{2.13}$$

Then,

$$\mathbb{P}(E^n_{\epsilon,s_1,s_2}) \ge 1 - \exp\left(-n\left(-\frac{1+\log(v+1)}{v} + \frac{\delta^2\epsilon^2}{16}\right)\right), \tag{2.14}$$

for all $n = n(v, \epsilon, \delta)$ large enough.

Remark 2.2 *Instead of (2.10), the corresponding condition in [13] is:*

$$r_0 = 0 < r_1 < r_2 < r_3 < ... < r_{d-1} < r_d = n. \tag{2.15}$$

However, in general, there is no guarantee that there exists an optimal alignment satisfying both (2.11) and (2.15). Indeed, for a simple counterexample, let $n = 4$, $\mathcal{A} = [2]$, $d = v = 2$, and let

$$X = (1, 1, 0, 0), \qquad Y = (0, 0, 1, 1).$$

Then, any optimal alignment satisfying (2.11) must have a cell with no terms in the Y-part which is clearly incompatible with (2.15). (This counterexample can easily be extended to $n = 6$, $\mathcal{A} = [2]$, $d = 3, v = 2$, letting $X = (1, 1, 0, 0, 1, 1)$, $Y = (0, 0, 1, 1, 0, 0)$, and so on.)

In general, there always exists an optimal alignment $(r_0, r_1, r_2, ..., r_d)$ satisfying both (2.10) and (2.11) with, say, $v = n^\alpha$ as above. (Consider any one of the longest common subsequences and choose the r_i's so that these two conditions are satisfied.) Therefore, we slightly changed the framework of [13] as the argument below requires the existence of an optimal alignment with (2.11) for any value of X and Y. However, the proof of Theorem 2.3 proceeds as the proof of the corresponding result in [13], and is therefore omitted. (The only difference is that counting the cases of equality, an upper estimate on the number of integer-vectors $(0 = r_0, r_1, \ldots, r_{d-1}, r_d = n)$ satisfying (2.10) is now given by

$$\binom{n+d}{d} \leq \frac{(n+d)^d}{d!} \leq \left(\frac{e(n+d)}{d}\right)^d = (e(v+1))^d, \tag{2.16}$$

leading to the terms involving $\log(v+1)$ rather than just $\log v$, when using (2.15) and an estimate on $\binom{n}{d}$.)

Remark 2.3 *In [13], the statement of Theorem 2.3 is given for "sufficiently large n". However, as indicated at the end of the proof there, it is possible to find a more quantitative estimate using Alexander's results (1.3). In fact, a lower bound valid for all n, in terms of v, ϵ and δ, holds true. Indeed, at first, from the end of the proof of the main theorem in [13], one can easily check that the following condition on n is sufficient for (2.14) to hold:*

$$\frac{4C_A^2}{(\delta^* - \delta)^2} \frac{\log n}{n} \leq \epsilon^2,$$

where $\delta^* - \delta$ is a fixed positive quantity and C_A is a positive constant such that $\gamma_m^* n - C_A\sqrt{n \log n} \leq \mathbb{E}LC_n$. (One can find explicit numerical estimates on C_A using Rhee's proof [17].)

In our context, here is how to choose ϵ so that the estimate in (2.14) holds true for all $n \geq 1$ and $v = n^\alpha$, $0 < \alpha < 1$. Let $c_1 > 0$ be a constant such that

$$c_1^2 \geq \frac{32}{\delta^2},$$

and

$$c_1^2\left(\frac{1+\log\left(n^\alpha+1\right)}{n^\alpha}\right)\geq\frac{4C_A^2}{(\delta^*-\delta)^2}\frac{\log n}{n},\qquad \textit{for all } n\geq 1.$$

Setting,

$$\epsilon^2=c_1^2\frac{1+\log\left(n^\alpha+1\right)}{n^\alpha},$$

(2.13) *holds for* $v=n^\alpha$ *and therefore,*

$$\mathbb{P}(E^n_{\epsilon,s_1,s_2})\geq 1-e^{-n^{1-\alpha}(1+\log\left(n^\alpha+1\right))},\tag{2.17}$$

for all $n\geq 1$.

Let us return to the proof of Theorem 1.1, and the estimation of (2.9). First, for notational convenience, below we write $\sum_1$ in place of $\Sigma_{(A,B,j,k)\in\mathcal{S}_1}$. Also, for random variables U,V and a random variable Z taking its values in $R\subset\mathbb{R}$, and with another abuse of notation, we write $Cov_{Z=z}(U,V)$ for $\mathbb{E}((U-\mathbb{E}U)(V-\mathbb{E}V)|Z=z)$, $z\in R$.

Let, now, the random variable Z be the indicator function of the event E^n_{ϵ,s_1,s_2}, where $\epsilon=c_1\sqrt{(1+\log\left(v+1\right))/v}$, i.e., $Z=\mathbf{1}_{E^n_{\epsilon,s_1,s_2}}$, with $v=n^\alpha$ and with c_1 as in Remark 2.3. Then,

$$\begin{aligned}\sum_1&\frac{Cov(\Delta_jf(W)\Delta_jf(W^A),\Delta_kf(W)\Delta_kf(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\&=\sum_1\frac{Cov_{Z=0}(\Delta_jf(W)\Delta_jf(W^A),\Delta_kf(W)\Delta_kf(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\mathbb{P}(Z=0)\\&+\sum_1\frac{Cov_{Z=1}(\Delta_jf(W)\Delta_jf(W^A),\Delta_kf(W)\Delta_kf(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\mathbb{P}(Z=1).\end{aligned}\tag{2.18}$$

To estimate the first term on the right-hand side of (2.18), first note that $Cov_{Z=0}(\Delta_jf(W)\Delta_jf(W^A),\Delta_kf(W)\Delta_kf(W^B))\leq 4$, which when combined with the estimate in (2.17) and Proposition 2.1, immediately lead to

$$\begin{aligned}\sum_1&\frac{Cov_{Z=0}(\Delta_jf(W)\Delta_jf(W^A),\Delta_kf(W)\Delta_kf(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\mathbb{P}(Z=0)\\&\leq 4n^2e^{-n^{1-\alpha}(1+\log\left(n^\alpha+1\right))}.\end{aligned}\tag{2.19}$$

For the second term on the right-hand side of (2.18), begin with the trivial bound on $\mathbb{P}(Z=1)$ to get

$$\sum_{1} \frac{Cov_{Z=1}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\mathbb{P}(Z=1)$$
$$\leq \sum_{1} \frac{Cov_{Z=1}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}. \qquad (2.20)$$

Finer decompositions are then needed to handle this last summation, and for this purpose, we specify an optimal alignment with certain properties.

Recall from Remark 2.2 that there always exists an optimal alignment $\mathbf{r} = (r_0, r_1, r_2, ..., r_d)$ satisfying both (2.10) and (2.11) with $v = n^\alpha$ as above. *In the sequel,* $\mathbf{r}$ *denotes a uniquely defined optimal alignment which also specifies the pairs, in the sequences* X *and* Y*, contributing to the longest common subsequence.* Such an alignment always exists, as just noted, and so we can define an injective map from (X, Y) to the set of alignments. This abstract construction is enough for our purposes, since the argument below is independent of the choice of the alignment. Note also that conditionally on the event $\{Z=1\}$, $\mathbf{r}$ satisfies (2.12).

To continue, we need another definition and some more notation.

Definition 2.1 *For the optimal alignment* $\mathbf{r}$*, each of the sets*

$$\{X_{v(i-1)+1}X_{v(i-1)+2}\cdots X_{vi}; Y_{r_{i-1}+1}Y_{r_{i-1}+2}\cdots Y_{r_i}\}, \qquad i = 1, ..., d,$$

is called a cell *of* $\mathbf{r}$*.*

In particular, and clearly, any optimal alignment with $v = n^\alpha$ has $d = n^{1-\alpha}$ cells.

Let us next introduce some more notation used below. For any given $j \in [2n]$, let P_j be the cell containing W_j where, again, $W = (W_1, \ldots, W_{2n}) = (X_1, \ldots, X_n, Y_1, \ldots, Y_n)$. We write $P_j = (P_j^1; P_j^2)$ where P_j^1 (resp. P_j^2) is the subword of X (resp. Y) corresponding to P_j. Note that, for each $j \in [2n]$, P_j^1 contains n^α letters but that P_j^2 might be empty.

Further, when P_j^2 is not empty, we define

$$a_j = \begin{cases} \min\{i : W_i \text{ is in } P_j^1\}, & \text{if } 1 \leq j \leq n \\ \min\{i : W_i \text{ is in } P_j^2\}, & \text{if } n+1 \leq j \leq 2n, \end{cases}$$

$$b_j = \begin{cases} \max\{i : W_i \text{ is in } P_j^1\}, & \text{if } 1 \le j \le n \\ \max\{i : W_i \text{ is in } P_j^2\}, & \text{if } n+1 \le j \le 2n, \end{cases}$$

$$a'_j = \begin{cases} \min\{i : W_i \text{ is in } \ P_j^2\}, & \text{if } 1 \le j \le n \\ \min\{i : W_i \text{ is in } \ P_j^1\}, & \text{if } n+1 \le j \le 2n, \end{cases}$$

and

$$b'_j = \begin{cases} \max\{i : W_i \text{ is in } \ P_j^2\}, & \text{if } 1 \le j \le n \\ \max\{i : W_i \text{ is in } \ P_j^1\}, & \text{if } n+1 \le j \le 2n. \end{cases}$$

(When P_j^2 is empty, the corresponding definitions do not make sense but this will not be an issue. Indeed, such cases are taken care of by the decompositions done later, in particular, see the definition of $S_{1,2,1}$ below.)

Let us illustrate our purpose on an example.

Example 2.1 *Take $n = 12$ and $\mathcal{A} = [3]$. Let*

$$X = (1, 1, 2, 1, 2, 1, 1, 2, 1, 1, 3, 1),$$

$$Y = (2, 1, 1, 3, 2, 3, 1, 2, 1, 1, 1, 1).$$

and $W = (X, Y)$. Then, $LC_{12} = 8$, and choosing $v = 3$, the number of cells in the optimal alignment is $d = 4$. One possible choice for these cells is

$$(X_1X_2X_3; Y_1Y_2Y_3Y_4Y_5) = (112; 21132),$$

$$(X_4X_5X_6; \emptyset) = (121; \emptyset),$$

$$(X_7X_8X_9; Y_6Y_7Y_8Y_9) = (121; 3121),$$

and

$$(X_{10}X_{11}X_{12}; Y_{10}Y_{11}Y_{12}) = (131; 111).$$

For example, focusing on $W_8 = X_8$, we have

$$P_8 = (P_8^1; P_8^2) = (121; 3121),$$

$a_8 = 7$, $b_8 = 9$, $a'_8 = 18$ and $b'_8 = 21$. If instead, we consider $W_{19} = Y_7$, then $a_{19} = 18$, $b_{19} = 21$, $a'_{19} = 7$, and $b'_{19} = 9$. Note that, in this example, the values of a, b and a', b' are interchanged since W_8 and W_{19} are in the same cell. Finally, for W_{23}, then $a_{23} = 22$, $b_{23} = 24$, $a'_{23} = 10$, and $b'_{23} = 12$

We return to the proof of Theorem 1.1 and define the following subsets of $\mathcal{S}_1$ with respect to the alignment $\mathbf{r}$:

$$\mathcal{S}_{1,1} = \{(A, B, j, k) \in \mathcal{S}_1 : W_j \text{ and } W_k \text{ are in the same cell of } \mathbf{r}\},$$

and

$$\mathcal{S}_{1,2} = \{(A, B, j, k) \in \mathcal{S}_1 : W_j \text{ and } W_k \text{ are in different cells of } \mathbf{r}\}.$$

Clearly, $\mathcal{S}_{1,1} \cap \mathcal{S}_{1,2} = \emptyset$ and $\mathcal{S}_1 = \mathcal{S}_{1,1} \cup \mathcal{S}_{1,2}$. Now, for a given subset $\mathcal{S}$ of $\mathcal{S}_1$, and for $(A, B, j, k) \in \mathcal{S}_1$, define $Cov_{Z=1,(A,B,j,k),\mathcal{S}}$ to be

$$Cov_{Z=1,(A,B,j,k),\mathcal{S}}(X, Y) = \mathbb{E}\left((X - \mathbb{E}X)(Y - \mathbb{E}Y)\mathbf{1}_{(A,B,j,k)\in\mathcal{S}}|Z = 1\right).$$

We write $Cov_{Z=1,\mathcal{S}}(X, Y)$ instead of $Cov_{Z=1,(A,B,j,k),\mathcal{S}}(X, Y)$ when the value of (A, B, j, k) is clear from the context.

We continue the decomposition of the right-hand side of (2.20) as

$$\begin{aligned}
&\sum_1 \frac{Cov_{Z=1}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n - |A|)\binom{2n}{|B|}(2n - |B|)} \\
&= \sum_1 \frac{Cov_{Z=1,\mathcal{S}_{1,1}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n - |A|)\binom{2n}{|B|}(2n - |B|)} \\
&\quad + \sum_1 \frac{Cov_{Z=1,\mathcal{S}_{1,2}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n - |A|)\binom{2n}{|B|}(2n - |B|)} \qquad (2.21)
\end{aligned}$$

To clarify the notation note that, for example,

$$\begin{aligned}
&\sum_1 \frac{Cov_{Z=1,\mathcal{S}_{1,1}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n - |A|)\binom{2n}{|B|}(2n - |B|)} \\
&\qquad = \sum_1 \mathbb{E}\left(\frac{g(A, B, j, k)\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}}{\binom{2n}{|A|}(2n - |A|)\binom{2n}{|B|}(2n - |B|)}\Big|Z = 1\right),
\end{aligned}$$

where

$$\begin{aligned}
g(A, B, j, k) \quad &= \left(\Delta_j f(W)\Delta_j f(W^A) - \mathbb{E}(\Delta_j f(W)\Delta_j f(W^A))\right) \\
&\quad \times\left(\Delta_k f(W)\Delta_k f(W^B) - \mathbb{E}(\Delta_k f(W)\Delta_k f(W^B))\right). \qquad (2.22)
\end{aligned}$$

To glimpse into the proof, let us stop for a moment to present some of its key steps. Our first intention is to show that, thanks to our conditioning

on the event E^n_{ϵ,s_1,s_2}, the number of terms contained in $\mathcal{S}_{1,1}$ is "small". To achieve this conclusion, a corollary to Theorem 2.3, see Theorem 2.4 below, is used. The next step will be based on estimations for the indices in $\mathcal{S}_{1,2}$. Here we will observe that we have enough independence (see the decomposition in (2.31)) to show that the contributions of the covariance terms from $\mathcal{S}_{1,2}$ are "small". This will require a lot more steps, as we shall see below.

Let us now focus on the first term on the right-hand side of (2.21). Letting g be as in (2.22), and using arguments similar to those used in the proof of Proposition 2.1, we have,

$$\begin{aligned}
&\sum_1 \frac{\left|Cov_{Z=1,\mathcal{S}_{1,1}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))\right|}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\
&\quad \le \mathbb{E}\left(\sum_1 \frac{|g(A,B,j,k)|\,\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\Big| Z=1\right) \\
&\quad \le 4\mathbb{E}\left(\sum_1 \frac{\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\Big| Z=1\right) \\
&\quad = 4\mathbb{E}\left(|\mathcal{R}||Z=1\right), \qquad (2.23)
\end{aligned}$$

where

$$\mathcal{R} = \{(j,k)\in[2n]^2 : W_j \text{ and } W_k \text{ are in the same cell of } \mathbf{r}\}.$$

To estimate (2.23), for each $i=1,\dots,d$, let $|\mathcal{R}_i|$ be the number of pairs of indices $(j,k)\in[2n]^2$ that are in the ith-cell, and let $\mathcal{G}_i$ be the event that $s_1 n^\alpha \le r_i - r_{i-1} \le s_2 n^\alpha$. Then,

$$\begin{aligned}
\mathbb{E}\left(|\mathcal{R}|\,|Z=1\right) &= \sum_{i=1}^{n^{1-\alpha}} \mathbb{E}(|\mathcal{R}_i|\,|Z=1) \\
&= \sum_{i=1}^{n^{1-\alpha}} \mathbb{E}(|\mathcal{R}_i|\mathbf{1}_{\mathcal{G}_i}\,|Z=1) + \sum_{i=1}^{n^{1-\alpha}} \mathbb{E}(|\mathcal{R}_i|\mathbf{1}_{\mathcal{G}_i^c}\,|Z=1). \qquad (2.24)
\end{aligned}$$

For the first term on the right-hand side of (2.24), note that, when $\mathcal{G}_i$ holds true, the i-th cell can contain at most $n^\alpha + s_2 n^\alpha = (1+s_2)n^\alpha$ letters (n^α is for the letters in X and $s_2 n^\alpha$ is for the letters in Y), and thus,

$$|\mathcal{R}_i|\mathbf{1}_{\mathcal{G}_i} \le (1+s_2)^2 n^{2\alpha}.$$

This gives

$$\sum_{i=1}^{n^{1-\alpha}} \mathbb{E}(|\mathcal{R}_i|\mathbf{1}_{\mathcal{G}_i}\big|Z=1) \leq (1+s_2)^2 n^{1+\alpha}. \tag{2.25}$$

For the estimation of the second term on the right-hand side of (2.24), we first recall a corollary to Theorem 2.3 stated in [13]. To do so, we need to introduce a different understanding of the LCS problem. Following [13], we consider alignments as subsets of $\mathbb{R}^2$, in the following way: If the i-th letter of X gets aligned with the j-th letter of Y, then the set representing the alignment is to contain (i, j).

Now, let H^n_{ϵ,s_2} be the event that all the points representing any optimal alignment of $X_1 \cdots X_n$ with $Y_1 \cdots Y_n$ are below the line $y = s_2 x + s_2 n\epsilon + s_2 n^\alpha$. Then,

Theorem 2.4 *[13] With the notation of Theorem 2.3,*

$$\mathbb{P}(H^n_{\epsilon,s_2}) \geq 1 - \exp\left(-n\left(-\frac{1+\log(v+1)}{v} + \frac{\delta^2\epsilon^2}{16}\right)\right),$$

for all $n = n(v, \epsilon, \delta)$, large enough.

Now, choosing ϵ as in Remark 2.3, the conclusion of Theorem 2.4 holds for any $n \geq 1$. The proof of Theorem 2.4 is based on the observation that

$$E^n_{\epsilon,s_1,s_2} \subset H^n_{\epsilon,s_2},$$

we refer the reader to [13] for details.

Returning to the estimation of the second term on the right-hand side of (2.24), we first write

$$\sum_{i=1}^{n^{1-\alpha}} \mathbb{E}(|\mathcal{R}_i|\mathbf{1}_{\mathcal{G}_i^c}|Z=1) = \sum_{i=1}^{n^{1-\alpha}} \mathbb{E}\left(|\mathcal{R}_i|\mathbf{1}_{\mathcal{G}_i^c}\mathbf{1}_{H^n_{\epsilon,s_2}}\Big|Z=1\right), \tag{2.26}$$

since, indeed, $E^n_{\epsilon,s_1,s_2} = \{Z=1\} \subset H^n_{\epsilon,s_2}$, so that $Z=1$ implies $\mathbf{1}_{H^n_{\epsilon,s_2}} = 1$.

Continuing, we start by focusing on estimating the first element of the sum, i.e., $\mathbb{E}(|\mathcal{R}_1|\mathbf{1}_{\mathcal{G}_1^c}\mathbf{1}_{H^n_{\epsilon,s_2}}|Z=1)$. To do so, let K^n_{ϵ,s_2} be the event that $X_1 \cdots X_{n^\alpha}$ is mapped to a subset of $Y_1 \cdots Y_{2s_2 n^\alpha + s_2 n\epsilon}$, for any alignment. Then, we have

$$H^n_{\epsilon,s_2} \subset K^n_{\epsilon,s_2},$$

where the inclusion follows from the definition of H^n_{ϵ,s_2}. Thus,

$$|\mathcal{R}_1|\mathbf{1}_{H^n_{\epsilon,s_2}} \leq \left((1+2s_2)\,n^\alpha + s_2 n\epsilon\right)^2,$$

yielding

$$\begin{aligned}\mathbb{E}(|\mathcal{R}_1|\mathbf{1}_{\mathcal{G}_1^c}\mathbf{1}_{H^n_{\epsilon,s_2}}\,|Z=1\,) &\leq \mathbb{E}(|\mathcal{R}_1|\mathbf{1}_{H^n_{\epsilon,s_2}}\,|Z=1\,)\\ &\leq \left((1+s_2)\,n^\alpha + s_2 n\epsilon\right)^2.\end{aligned}$$

In a similar way, we have

$$|\mathcal{R}_i|\mathbf{1}_{H^n_{\epsilon,s_2}} \leq \left((1+2s_2)\,n^\alpha + s_2 n\epsilon\right)^2, \qquad i=2,\ldots,n^\alpha,$$

since, again, when H^n_{ϵ,s_2} occurs, a cell must contain at most $s_2 n\epsilon + 2s_2 n^\alpha$ terms from the Y sequence. (To see that this is indeed the case, assume that the Y part of a cell contains more than $s_2 n\epsilon + 2s_2 n^\alpha$ terms while H^n_{ϵ,s_2} occurs. Then, just move the first $i-1$ cells to the end of the sequences to get an optimal alignment whose first cell has more than $s_2 n\epsilon + 2s_2 n^\alpha$ terms, giving a contradiction.)

Therefore, for any $i=1,\ldots,n^\alpha$,

$$\mathbb{E}(|\mathcal{R}_i|\mathbf{1}_{\mathcal{G}_i^c}\mathbf{1}_{H^n_{\epsilon,s_2}}\,|Z=1\,) \leq \left((1+2s_2)\,n^\alpha + s_2 n\epsilon\right)^2. \tag{2.27}$$

But, thanks to the $\mathbf{1}_{\mathcal{G}_i^c}$ terms and to the conditioning on $Z=1$, at most $\epsilon n^{1-\alpha}$ of the summands in (2.26) are nonzero and so, from (2.27),

$$\sum_{i=1}^{n^{1-\alpha}} \mathbb{E}\left(|\mathcal{R}_i|\mathbf{1}_{\mathcal{G}_i^c}\Big|Z=1\right) \leq \epsilon n^{1-\alpha}\left((1+2s_2)\,n^\alpha + s_2 n\epsilon\right)^2. \tag{2.28}$$

For $\epsilon = (c_1^2(1+\log(n^\alpha+1))/n^\alpha)^{1/2}$, the estimate (2.28) lead to:

$$\begin{aligned}&\sum_{i=1}^{n^{1-\alpha}} \mathbb{E}(|\mathcal{R}_i|\mathbf{1}_{\mathcal{G}_i^c}\,|Z=1\,)\\ &\quad\leq C_1 n^{1+\alpha/2}(\log n^\alpha)^{1/2} + C_2 n^{2-\alpha}\log n^\alpha + C_3 n^{3-5\alpha/2}(\log n^\alpha)^{3/2},\end{aligned} \tag{2.29}$$

where C_1, C_2, and C_3 are constants independent of n.

Hence, combining (2.25) and (2.29),

$$\mathbb{E}(|\mathcal{R}||Z=1\,)$$

$$\leq C(n^{1+\alpha} + n^{1+\alpha/2}(\log n^{\alpha})^{1/2} + n^{2-\alpha}\log n^{\alpha} + n^{3-5\alpha/2}(\log n^{\alpha})^{3/2}),$$

which, in turn, yields

$$\sum_{1} \frac{\left|Cov_{Z=1,\mathcal{S}_{1,1}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))\right|}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}$$
$$\leq C(n^{1+\alpha} + n^{1+\alpha/2}(\log n^{\alpha})^{1/2} + n^{2-\alpha}\log n^{\alpha} + n^{3-5\alpha/2}(\log n^{\alpha})^{3/2}) \quad (2.30)$$

and, this last estimate takes care of the first sum on the right-hand side of (2.21).

Next we move to the estimation of the second term on the right-hand side of (2.21), which is given by,

$$\sum_{1} \frac{Cov_{Z=1,\mathcal{S}_{1,2}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}.$$

To estimate the summands in the last expression, we need to decompose the covariance terms in such a way that independence of certain random variables occurs, simplifying te estimates. For this purpose, for each $i \in [2n]$, let $f(P_i) = LC(P_i)$ be the length of the longest common subsequence of P_i^1 and P_i^2, the coordinates of the cell $P_i = (P_i^1; P_i^2)$, and set

$$\tilde{\Delta}_i f(W) := f(P_i) - f(P_i'),$$

where P_i' is the same as P_i except that W_i is now replaced with the independent copy W_i'. In words, $\tilde{\Delta}_i f(W)$ is the difference between the length of the longest common subsequence restricted to P_i and its modified version at coordinate i. Now for $(A, B, j, k) \in \mathcal{S}_1$,

$$\begin{aligned}
&Cov_{Z=1,\mathcal{S}_{1,2}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B)) = \\
&\quad Cov_{Z=1,\mathcal{S}_{1,2}}((\Delta_j f(W) - \tilde{\Delta}_j f(W))\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B)) \\
&\quad + Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)(\Delta_j f(W^A) - \tilde{\Delta}_j f(W^A)), \Delta_k f(W)\Delta_k f(W^B)) \\
&\quad + Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), (\Delta_k f(W) - \tilde{\Delta}_k f(W))\Delta_k f(W^B)) \\
&\quad + Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)(\Delta_k f(W^B) - \tilde{\Delta}_k f(W^B)) \\
&\quad + Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B)). \qquad (2.31)
\end{aligned}$$

Above, we used the bilinearity of $Cov_{Z=1,\mathcal{S}_{1,2}}$ to express the left-hand side as a telescoping sum. (Except for the conditioning, this decomposition is akin to a decomposition developed in [7].)

Let us begin by estimating the last term on the right-hand side of (2.31). To do so, first observe that

$$Cov(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B)) = 0,$$

since the random variables $\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A)$ and $\tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B)$ belong to different cells and are therefore independent. Now, recalling that $\mathcal{S}_1 = \mathcal{S}_{1,1} \cup \mathcal{S}_{1,2}$, and conditioning on Z gives

$$\begin{aligned}
0 &= Cov(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B)) \\
&= Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B))\mathbb{P}(Z=1) \\
&\quad + Cov_{Z=1,\mathcal{S}_{1,1}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B))\mathbb{P}(Z=1) \\
&\quad + Cov_{Z=0,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B))\mathbb{P}(Z=0) \\
&\quad + Cov_{Z=0,\mathcal{S}_{1,1}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B))\mathbb{P}(Z=0) \quad (2.32)
\end{aligned}$$

Thus,

$$\begin{aligned}
&|Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B))| \leq \\
&\quad |Cov_{Z=1,\mathcal{S}_{1,1}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B)| \\
&\quad + \left|\frac{\mathbb{P}(Z=0)}{\mathbb{P}(Z=1)} Cov_{Z=0,\mathcal{S}_{1,1}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B))\right| \\
&\quad + \left|\frac{\mathbb{P}(Z=0)}{\mathbb{P}(Z=1)} Cov_{Z=0,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B))\right| \quad (2.33)
\end{aligned}$$

By making use of the estimate in (2.17), the last two terms on the right-hand side of (2.33) are clearly bounded by $Ce^{-n^{1-\alpha}(1+\log(n^\alpha+1))}$. Also, as in passing from (2.23) to (2.30), we have

$$\begin{aligned}
&\sum_1 \frac{|Cov_{Z=1,\mathcal{S}_{1,1}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B))|}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \leq \\
&\quad C(n^{1+\alpha} + n^{1+\alpha/2}(\log n^\alpha)^{1/2} + n^{2-\alpha}\log n^\alpha + n^{3-5\alpha/2}(\log n^\alpha)^{3/2}).
\end{aligned}$$

Combining these two observations, we arrive at

$$\sum_1 \frac{|Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\Delta}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\Delta}_k f(W^B))|}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \leq$$

$$C(n^{1+\alpha}+n^{1+\alpha/2}(\log n^{\alpha})^{1/2}+n^{2-\alpha}\log n^{\alpha}+n^{3-5\alpha/2}(\log n^{\alpha})^{3/2} \\ +n^{2}e^{-n^{1-\alpha}(1+\log(n^{\alpha}+1))}), \tag{2.34}$$

finishing the estimation of the last term in (2.31).

Next we obtain an upper bound for the first of the remaining four summands in (2.31), the other three terms can be estimated similarly and so, the details for these are omitted. To do so, let

$$U := (\Delta_j f(W) - \tilde{\Delta}_j f(W))\Delta_j f(W^A),$$

and

$$V := \Delta_k f(W)\Delta_k f(W^B),$$

so that we wish to estimate $Cov_{Z=1,\mathcal{S}_{1,2}}(U,V)$. But,

$$\begin{aligned}
&\left|Cov_{Z=1,\mathcal{S}_{1,2}}(U,V)\right| \\
&\quad = \left|\mathbb{E}((U-\mathbb{E}U)(V-\mathbb{E}V)\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}|\,Z=1)\right| \\
&\quad \le \mathbb{E}(|UV|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}|\,Z=1)+\mathbb{E}|V|\mathbb{E}(|U|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}|\,Z=1) \\
&\quad +\mathbb{E}|U|\mathbb{E}(|V|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}|\,Z=1)+\mathbb{E}|U|\mathbb{E}|V|\mathbb{E}(\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}|\,Z=1) \\
&\quad := T_1+T_2+T_3+T_4.
\end{aligned}$$

Note here that $T_i, i=1,2,3,4$ are functions of (A,B,j,k). Let us begin by estimating

$$T_1 = \mathbb{E}_{Z=1}|((\Delta_j f(W)-\tilde{\Delta}_j f(W))\Delta_j f(W^A))(\Delta_k f(W)\Delta_k f(W^B))\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}|.$$

Since $|\Delta_j f(W^A)(\Delta_k f(W)\Delta_k f(W^B))| \le 1$,

$$T_1 \le \mathbb{E}_{Z=1}\left(|\Delta_j f(W)-\tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right). \tag{2.35}$$

A similar estimate also reveals that

$$T_2 \le \mathbb{E}_{Z=1}\left(|\Delta_j f(W)-\tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right). \tag{2.36}$$

Next, for T_3 and T_4, since $|V| \le 1$,

$$\begin{aligned}
T_3+T_4 \le 2\mathbb{E}|U| &\le 2\mathbb{E}|\Delta_j f(W)-\tilde{\Delta}_j f(W)| \\
&= 2\mathbb{E}_{Z=1}\left(|\Delta_j f(W)-\tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)\mathbb{P}(Z=1)
\end{aligned}$$

$$
\begin{aligned}
&+2\mathbb{E}_{Z=1}\left(|\Delta_j f(W)-\tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}\right)\mathbb{P}(Z=1)\\
&+2\mathbb{E}_{Z=1}\left(|\Delta_j f(W)-\tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)\mathbb{P}(Z=0)\\
&+2\mathbb{E}_{Z=1}\left(|\Delta_j f(W)-\tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}\right)\mathbb{P}(Z=0)\\
\leq\ & 2\mathbb{E}_{Z=1}\left(|\Delta_j f(W)-\tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)\\
&+2\mathbb{E}_{Z=1}\left(|\Delta_j f(W)-\tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}\right)\\
&+Ce^{-n^{1-\alpha}(1+\log(n^\alpha+1))}, \qquad (2.37)
\end{aligned}
$$

where we used the trivial bound on $\mathbb{P}(Z=1)$, and also (2.17), for the last inequality.

Now, denote by $h(A,B,j,k)$ the sum of the first four terms on the right-hand side of (2.31). Then, performing estimations as in getting (2.35), (2.36) and (2.37), for the second to fourth term of this sum, and observing that $|\Delta_j f(W)-\tilde{\Delta}_j f(W)|$ is equal in distribution to $|\Delta_j f(W^A)-\tilde{\Delta}_j f(W^A)|$, while $|\Delta_k f(W)-\tilde{\Delta}_k f(W)|$ is equal in distribution to $|\Delta_k f(W^B)-\tilde{\Delta}_k f(W^B)|$, we obtain

$$
\begin{aligned}
&\sum_1\left|\frac{h(A,B,j,k)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\right|\\
&\quad\leq C\sum_1\frac{\mathbb{E}_{Z=1}\left(|\Delta_j f(W)-\tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\
&\qquad+C\sum_1\frac{\mathbb{E}_{Z=1}\left(|\Delta_j f(W)-\tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\
&\qquad+C\sum_1\frac{\mathbb{E}_{Z=1}\left(|\Delta_k f(W)-\tilde{\Delta}_k f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\
&\qquad+C\sum_1\frac{\mathbb{E}_{Z=1}\left(|\Delta_k f(W)-\tilde{\Delta}_k f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\
&\qquad+C\sum_1\frac{e^{-n^{1-\alpha}(1+\log(n^\alpha+1))}}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}.
\end{aligned}
$$

By making use of a symmetry argument, this gives

$$\begin{aligned}
&\sum_1 \left| \frac{h(A,B,j,k)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \right| \\
&\quad \leq C \sum_1 \frac{\mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)| \mathbf{1}_{(A,B,j,k)\in \mathcal{S}_{1,2}} \right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\
&\qquad + C \sum_1 \frac{\mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)| \mathbf{1}_{(A,B,j,k)\in \mathcal{S}_{1,1}} \right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\
&\qquad + C \sum_1 \frac{e^{-n^{1-\alpha}(1+\log(n^\alpha+1))}}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}.
\end{aligned}$$

As with previous computations, using (2.17), the third sum on the above right-hand side is bounded by

$$Cn^2 e^{-n^{1-\alpha}(1+\log(n^\alpha+1))}, \tag{2.38}$$

while the middle sum is bounded by

$$C(n^{1+\alpha} + n^{1+\alpha/2}(\log n^\alpha)^{1/2} + n^{2-\alpha}\log n^\alpha + n^{3-5\alpha/2}(\log n^\alpha)^{3/2}), \tag{2.39}$$

using (2.23). Hence, we are left with estimating

$$\sum_1 \mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)| \mathbf{1}_{(A,B,j,k)\in \mathcal{S}_{1,2}} \right). \tag{2.40}$$

To handle (2.40), we divide $\mathcal{S}_{1,2}$ into two more pieces; $\mathcal{S}_{1,2,1}$ and $\mathcal{S}_{1,2,2}$. This decomposition will turn out to be crucial in (2.47) where we take care of the covariance terms from $\mathcal{S}_{1,2,2}$. Recall that, for the optimal alignment $\mathbf{r}$, we have $d = n^{1-\alpha}$ cells. Also, for any given $j \in [n]$, we set $a_j = \min\{i : X_i \text{ is in } P_j^1\}$ and $b_j = \max\{i : X_i \text{ is in } P_j^1\}$ (with similar definitions, for $j \in \{n+1, \ldots, 2n\}$).

Let now,

$$\begin{aligned}
\mathcal{S}_{1,2,1} &= \{(A,B,j,k) \in \mathcal{S}_{1,2} : 1 \leq j \leq n\} \\
&\quad \cap \{(A,B,j,k) \in \mathcal{S}_{1,2} : |j - a_j| \leq n^{\alpha/2} \text{ or } |j - b_j| \leq n^{\alpha/2} \text{ or } P_j^2 = \emptyset\},
\end{aligned}$$

$$\widetilde{\mathcal{S}}_{1,2,1} = \{(A,B,j,k) \in \mathcal{S}_{1,2} : n+1 \leq j \leq 2n\}$$

$$\cap\{(A,B,j,k)\in\mathcal{S}_{1,2} : |j-a_j|\leq n^{\alpha/2} \text{ or } |j-b_j|\leq n^{\alpha/2}\},$$

and

$$\mathcal{S}_{1,2,2} = \mathcal{S}_{1,2} - (\mathcal{S}_{1,2,1}\cup\widetilde{\mathcal{S}}_{1,2,1}).$$

(Recall that $P_j^1\neq\emptyset$, for any $j\in[2n]$, explaining the difference between the definitions of $\mathcal{S}_{1,2,1}$ and $\widetilde{\mathcal{S}}_{1,2,1}$.) To estimate

$$\sum_1 \frac{\mathbb{E}_{Z=1}\left(|(\Delta_j f(W)-\tilde{\Delta}_j f(W))|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2,1}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)},$$

set

$$\begin{aligned}\mathcal{T}_1 &= \big\{(j,k)\in[2n]^2 : (A,B,j,k)\in\mathcal{S}_1, 1\leq j\leq n,\\ &\qquad\qquad |j-a_j|\leq n^{\alpha/2} \text{ or } |j-b_j|\leq n^{\alpha/2}\big\},\end{aligned}$$

and

$$\mathcal{T}_2 = \{(j,k)\in[2n]^2 : (A,B,j,k)\in\mathcal{S}_1, 1\leq j\leq n, P_j^2=\emptyset\},$$

so that, mimicking computations previously performed,

$$\sum_1 \frac{\mathbb{E}_{Z=1}|(\Delta_j f(W)-\tilde{\Delta}_j f(W))|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2,1}}}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \leq \mathbb{E}(|\mathcal{T}_1|+|\mathcal{T}_2|\,|\,Z=1). \tag{2.41}$$

Next, observe that

$$|\mathcal{T}_1|\leq Cn^{1-\alpha}n^{\alpha/2}n = Cn^{2-\alpha/2}. \tag{2.42}$$

($n^{1-\alpha}$ is for the number of cells, $2n^{\alpha/2}$ is for the number of coordinates in a given cell which are at most at distance $n^{\alpha/2}$ from their endpoints and n is for the number of possible values of k so that $(A,B,j,k)\in\mathcal{S}_{1,2,1}$.)

Also, by the very definition of Z,

$$\begin{aligned}\mathbb{E}(|\mathcal{T}_2|\big|Z=1) &\leq \epsilon n^{1-\alpha}n^{\alpha}n\\ &\leq Cn^{2-\alpha/2}(\log n^{\alpha})^{1/2}.\end{aligned} \tag{2.43}$$

(At most $\epsilon n^{1-\alpha}$ many cells may have empty Y-part, each of which containing n^{α} different j values. The n term is again for the number of possible values of k)

Combining (2.41), (2.42) and (2.43), lead to

$$\sum_{1} \frac{\mathbb{E}_{Z=1}\left(|(\Delta_j f(W) - \tilde{\Delta}_j f(W))|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2,1}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \leq C(n^{2-\alpha/2} + Cn^{2-\alpha/2}(\log n^{\alpha})^{1/2}). \quad (2.44)$$

Similar estimates also give

$$\sum_{1} \frac{\mathbb{E}_{Z=1}\left(|(\Delta_j f(W) - \tilde{\Delta}_j f(W))|\mathbf{1}_{(A,B,j,k)\in\tilde{\mathcal{S}}_{1,2,1}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \leq Cn^{2-\alpha/2}. \quad (2.45)$$

Let us now deal with the sum over $\mathcal{S}_{1,2,2}$. To do so, in the rest of the proof, $LCS(W)$ denotes the vector $(X_{i_1}\cdots X_{i_\ell}; Y_{j_1}\cdots Y_{j_\ell})$ obtained via the unique optimal alignment $\mathbf{r}$ we specified, so that $|LCS(W)| = \ell$. Further, we say that the pair $(X_i, Y_j) \in \{X_1, \ldots, X_n\} \times \{Y_1, \ldots, Y_n\}$ is in $LCS(W)$, if (X_i, Y_j) contributes to the length of the longest common subsequence (with respect to $\mathbf{r}$).

Now, write $\mathcal{S}_{1,2,2} = \mathcal{S}_{1,2,2,1} \cup \mathcal{S}_{1,2,2,2}$ where,

$$\begin{aligned}\mathcal{S}_{1,2,2,1} &= \{(A,B,j,k) \in \mathcal{S}_{1,2,2} : W_j \text{ and } W_{a'_j} \text{ are in } LCS(W)\} \\ &\quad \cup \{(A,B,j,k) \in \mathcal{S}_{1,2,2} : W_j \text{ and } W_{b'_j} \text{ are in } LCS(W)\},\end{aligned}$$

and where $\mathcal{S}_{1,2,2,2} = \mathcal{S}_{1,2,2} - \mathcal{S}_{1,2,2,1}$. The estimates for the covariance terms corresponding to $\mathcal{S}_{1,2,2,1}$ and $\mathcal{S}_{1,2,2,2}$ are, in a sense, similar to our previous computations. We will show that there are few terms (on average) in the index set $\mathcal{S}_{1,2,2,1}$, and that the covariance terms are themselves small when dealing with the indices in $\mathcal{S}_{1,2,2,2}$.

As in obtaining (2.44), we set

$$\begin{aligned}\mathcal{T}_3 = \{(j,k) \in [2n]^2 : (A,B,j,k) \in \mathcal{S}_1,\ W_j \text{ and } W_{a'_j} \text{ are in } LCS(W)\} \\ \cup \{(j,k) \in [2n]^2 : (A,B,j,k) \in \mathcal{S}_1,\ W_j \text{ and } W_{a'_j} \text{ are in } LCS(W)\}.\end{aligned}$$

We have $n^{1-\alpha}$ cells and each cell may contain at most four j values that contribute to the longest common subsequence by being matched with a'_j or b'_j. Hence the number of (j,k) pairs in $[2n]^2$ that are in $\mathcal{S}_{1,2,2,1}$ for some A, B is bounded by $4n^{1-\alpha}n = 4n^{2-\alpha}$, where the n term is for the possible choices of k values. That is,

$$|\mathcal{T}_3| \leq 4n^{2-\alpha},$$

and using Proposition 2.1,

$$\sum_{1} \frac{\mathbb{E}_{Z=1}\left(|(\Delta_j f(W) - \tilde{\Delta}_j f(W))|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2,2,1}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \leq Cn^{2-\alpha}. \tag{2.46}$$

Next, let

$$F_j := \left(\bigcup_{r=j-n^{\alpha/2}}^{j-1} \{W_r \text{ is in } LCS(W)\}\right) \bigcap \left(\bigcup_{r=j+1}^{j+n^{\alpha/2}} \{W_r \text{ is in } LCS(W)\}\right),$$

which is the event that the longest common subsequence of $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$, with respect to $\mathbf{r}$, contains terms which are close to W_j on both of its sides. Using the events F_j,

$$\begin{aligned}
&\sum_{1} \frac{\mathbb{E}_{Z=1}\left(|(\Delta_j f(W) - \tilde{\Delta}_j f(W))|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2,2,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\
&= \sum_{1} \left(\frac{\mathbb{E}_{Z=1}\left(|(\Delta_j f(W) - \tilde{\Delta}_j f(W))|\mathbf{1}_{F_j}\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2,2,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \right. \\
&\quad \left. + \frac{\mathbb{E}_{Z=1}\left(|(\Delta_j f(W) - \tilde{\Delta}_j f(W))|\mathbf{1}_{F_j^c}\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2,2,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \right) \\
&= \sum_{1} \frac{\mathbb{E}_{Z=1}\left(|(\Delta_j f(W) - \tilde{\Delta}_j f(W))|\mathbf{1}_{F_j^c}\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2,2,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\
&\leq C \sum_{1} \frac{\mathbb{E}_{Z=1}\left(\mathbf{1}_{F_j^c}\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2,2,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\
&\leq C \sum_{1} \frac{\mathbb{P}(F_j^c, (A,B,j,k) \in \mathcal{S}_{1,2,2,2})}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)},
\end{aligned}$$

where the second equality follows, observing that $(\Delta_j f(W) - \tilde{\Delta}_j f(W))\mathbf{1}_{F_j} = 0$, and the last one via (2.17).

Then,

$$F_j^c \cap \{(A,B,j,k) \in \mathcal{S}_{1,2,2,2})\}$$

$$= \left(\left(\bigcap_{r=j-n^{\alpha/2}}^{j-1} \{W_r \text{ is not in } LCS(W)\} \right) \bigcap \{(A,B,j,k) \in \mathcal{S}_{1,2,2,2})\} \right)$$
$$\bigcup \left(\left(\bigcap_{r=j+1}^{j+n^{\alpha/2}} \{W_r \text{ is not in } LCS(W)\} \right) \bigcap \{(A,B,j,k) \in \mathcal{S}_{1,2,2,2}\} \right),$$

and so

$$\begin{aligned}
&\mathbb{P}(F_j^c \cap \{(A,B,j,k) \in \mathcal{S}_{1,2,2,2})\}) \\
&\leq \mathbb{P}\left(\bigcap_{r=j-n^{\alpha/2}}^{j-1} \{W_r \text{ is not in } LCS(W)\} \cap \{(A,B,j,k) \in \mathcal{S}_{1,2,2,2}\} \right) \\
&\quad + \mathbb{P}\left(\bigcap_{r=j+1}^{j+n^{\alpha/2}} \{W_r \text{ is not in } LCS(W)\} \cap \{(A,B,j,k) \in \mathcal{S}_{1,2,2,2}\} \right).
\end{aligned}$$

Since for $(A,B,j,k) \in \mathcal{S}_{1,2,2,2}$, the pair $(W_j, W_{a'_j})$ is not included in the longest common subsequence, it follows that

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{r=j-n^{\alpha/2}}^{j-1} \{W_r \text{ is not in } LCS(W)\} \cap \{(A,B,j,k) \in \mathcal{S}_{1,2,2,2}\} \right) \\
&\leq \mathbb{P}\Big(\text{none of the terms in } \{W_{j-n^{\alpha/2}}, \ldots, W_{j-1}\} \text{ is equal to } W_{a'_j}\Big) \\
&= \mathbb{P}\left(\bigcap_{r=j-n^{\alpha/2}}^{j-1} \{W_r \neq W_{a'_j}\} \right) \\
&= \left(1 - \sum_{i=1}^{m} p_i^2\right)^{n^{\alpha/2}}.
\end{aligned}$$

Similarly, for $(A,B,j,k) \in \mathcal{S}_{1,2,2,2}$, the pair $(W_j, W_{b'_j})$ is not included in the longest common subsequence and so

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{r=j+1}^{j+n^{\alpha/2}} \{W_r \text{ is not in } LCS(W)\} \cap \{(A,B,j,k) \in \mathcal{S}_{1,2,2,2}\} \right) \\
&\leq \mathbb{P}\Big(\text{none of the terms in } \{W_{j+1}, \ldots, W_{j+n^{\alpha/2}}\} \text{ is equal to } W_{b'_j}\Big)
\end{aligned}$$

$$= \left(1 - \sum_{i=1}^{m} p_i^2\right)^{n^{\alpha/2}}.$$

Hence, from the above,

$$\mathbb{P}(F_j^c \cap \{(A,B,j,k) \in \mathcal{S}_{1,2,2,2}\}) \leq 2\left(1 - \sum_{i=1}^{m} p_i^2\right)^{n^{\alpha/2}}. \tag{2.47}$$

Thus, (2.47) and Proposition 2.1 lead to:

$$\begin{aligned}\sum_{1} \frac{\mathbb{P}(F_j^c, (A,B,j,k) \in \mathcal{S}_{1,2,2,2})}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} &\leq \sum_{1} \frac{2e^{n^{\alpha/2}\log\left(1-\sum_{i=1}^{m} p_i^2\right)}}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\ &\leq Cn^2 e^{n^{\alpha/2}\log\left(1-\sum_{i=1}^{m} p_i^2\right)}. \end{aligned} \tag{2.48}$$

Combining (2.19), (2.30), (2.34), (2.38), (2.39), (2.44), (2.45), (2.46) and (2.48), gives

$$\begin{aligned}\operatorname{Var} T \leq\ & C\Big(n^2 e^{-n^{1-\alpha}(1+\log(n^\alpha+1))} + n^{1+\alpha} + n^{1+\alpha/2}(\log n^\alpha)^{1/2} \\ & + n^{2-\alpha}\log n^\alpha + n^{3-5\alpha/2}(\log n^\alpha)^{3/2} + n^{2-\alpha/2} \\ & + n^{2-\alpha/2}(\log n^\alpha)^{1/2} + n^{2-\alpha} + n^2 e^{n^{\alpha/2}\log\left(1-\sum_{i=1}^{m} p_i^2\right)}\Big).\end{aligned}$$

Therefore, Theorem 2.3 and (2.7), as well as the choice $\alpha = 3/4$, above, ensure that:

$$d_W\left(\frac{LC_n - \mathbb{E}LC_n}{\sqrt{\operatorname{Var} LC_n}}, \mathcal{G}\right) \leq C\left(\frac{1}{n^{1/4}}\right)^{1/2} + C\frac{1}{n^{1/2}} \leq C\frac{1}{n^{1/8}},$$

holds for every $n \geq 1$, with $C > 0$ a constant independent of n. □

Remark 2.4 *(i) The arguments presented here will also prove a central limit theorem for the length of the longest common subsequence in the uniform setting, or for any distribution of X_1, as soon as the variance estimate*

$$\operatorname{Var} LC_n \geq Cn,$$

holds true for some constant C independent of n. In fact, even a sublinear lower bound for the variance, so as to compensate for our estimate on $\operatorname{Var} T$,

would do it, e.g., $n^{7/8}$ will do (although most likely, the variance of LC_n is linear in n).

(ii) The constant C in Theorem 1.1 is independent of n but depends on m, on s_1 and s_2 of Theorem 2.3, as well as on $\max_{j=1,\ldots,m} p_j$ and $\max_{j\neq j_0} p_j$ of Theorem 2.5.

(iii) Some arguments of the proof could, somehow, be simplified by making the events $\mathcal{S}_{1,2,1}$, $\widetilde{\mathcal{S}}_{1,2,1}$ and $\mathcal{S}_{1,2,2,1}$ part of $\mathcal{S}_{1,1}$. However, it is our belief that the current approach makes the arguments clearer.

(iv) Of course, there is no reason for our rate $n^{1/8}$ to be sharp. Already, instead of the choice $v = n^{\alpha}$, a choice such as $v = h(n)$, for some optimal function h would improve the rate.

(v) From a known duality between the length of longest common subsequence of two random words and the length of the shortest common supersequence (see Dančík [8]), our result also implies a central limit theorem for this latter case.

3 Concluding Remarks

We conclude the paper with a discussion on longest common subsequences in random permutations and in a final remark, present some potential extensions, perspectives and questions we believe are of interest.

Theorem 1.1 shows that the Gaussian distribution appears as the limiting law for the length in longest common subsequences of random words. However, the Tracy-Widom distribution has also been hypothesized as the limiting law in such contexts. It turns out, as shown next, that it is indeed the case for certain distributions on permutations.

First, it is folklore that, if $\pi = (\pi_1, \ldots, \pi_n)$ is any element of the symmetric group S_n, then

$$LI_n(\pi) = LC_n((1, 2, \ldots, n), (\pi_1, \pi_2, \ldots, \pi_n)), \tag{3.1}$$

where $LI_n(\pi)$ is the length of the longest increasing subsequence in $\pi = (\pi_1, \ldots, \pi_n)$, while $LC_n((1, 2, \ldots, n), (\pi_1, \pi_2, \ldots, \pi_n))$, is the length of the longest common subsequence of the identity permutation id and of the permutation π. In the equality (3.1), replacing id by an arbitrary permutation ρ and taking for π a uniform random permutation in S_n lead to:

Proposition 3.1 *(i) Let $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$ be a fixed permutation in S_n and let π be a uniform random permutation in S_n. Then,*

$$LI_n(\pi) =_d LC_n((\rho_1, \rho_2, \ldots, \rho_n), (\pi_1, \pi_2, \ldots, \pi_n)), \tag{3.2}$$

where $=_d$ denotes equality in distribution.

(ii) Let ρ and π be two independent uniform random permutations in S_n, and let $x \in \mathbb{R}$. Then,

$$\mathbb{P}(LC_n(\rho, \pi) \leq x) = \mathbb{P}(LI_n(\pi) \leq x). \tag{3.3}$$

Proof. To begin the proof of (i), let $\pi' \in S_n$ be such that $\pi'_i = \rho_i$. Then, $\pi'' := \pi\pi'$ is still a uniformly random permutation, and so

$$\begin{aligned} LC_n((\rho_1, \rho_2, \ldots, \rho_n), (\pi_1, \pi_2, \ldots, \pi_n)) &=_d LC_n((\rho_1, \rho_2, \ldots, \rho_n), (\pi''_1, \pi''_2, \ldots, \pi''_n)) \\ &= LC_n((\rho_1, \rho_2, \ldots, \rho_n), (\pi_{\rho_1}, \pi_{\rho_2}, \ldots, \pi_{\rho_n})), \end{aligned}$$

where for the second equality we used $\pi''_i = \pi\pi'_i = \pi_{\rho_i}$. Clearly,

$$LC_n((\rho_1, \rho_2, \ldots, \rho_n), (\pi_{\rho_1}, \pi_{\rho_2}, \ldots, \pi_{\rho_n})) =_d LC_n((1, 2, \ldots, n), (\pi_1, \pi_2, \ldots, \pi_n)),$$

and so (3.1) finishes the proof of (i).

Let us now prove (ii).

$$\begin{aligned} \mathbb{P}(LC_n(\rho, \pi) \leq x) &= \sum_{\gamma \in S_n} \mathbb{P}(LC_n(\gamma, \pi) \leq x | \rho = \gamma) \mathbb{P}(\rho = \gamma) \\ &= \frac{1}{n!} \sum_{\gamma \in S_n} \mathbb{P}(LC_n((\gamma_1, \ldots, \gamma_n), (\pi_1, \ldots, \pi_n)) \leq x) \\ &= \frac{1}{n!} \sum_{\gamma \in S_n} \mathbb{P}(LI_n(\pi) \leq x) \\ &= \mathbb{P}(LI_n(\pi) \leq x), \end{aligned}$$

where the third equality follows from (3.2). This proves (ii). □

Clearly, the identity (3.3), which in fact is easily seen to remain true if ρ is a random permutation in S_n with an arbitrary distribution, shows that the probabilistic behavior of $LC_n(\rho, \pi)$ is identical to the probabilistic

behavior of $LI_n(\pi)$. Among the many results presented in Romik [19], the mean asymptotic result of Logan-Shepp and Vershik-Kerov implies that:

$$\lim_{n\to+\infty} \frac{\mathbb{E}LC_n(\rho,\pi)}{2\sqrt{n}} = 1.$$

Moreover, the distributional asymptotic result of Baik-Deift and Johansson implies that, as $n \to +\infty$,

$$\frac{LC_n(\rho,\pi) - 2\sqrt{n}}{n^{1/6}} \longrightarrow F_2, \qquad \text{in distribution,}$$

where F_2 is the Tracy-Widom distribution whose cdf is given by

$$F_2(t) = \exp\left(-\int_t^\infty (x-t)u^2(x)dx\right),$$

where u is the solution to the Painlevé II equation:

$$u_{xx} = 2u^3 + xu \qquad \text{with} \qquad u(x) \sim -Ai(x) \quad \text{as} \quad x \to \infty.$$

To finish, let us list a few venues for future research that we find of potential interest.

Remark 3.1 *(i) First, the methods of the present paper can also be used to study sequence comparison with a general scoring functions S. Namely, $S : \mathcal{A}_m \times \mathcal{A}_m \to \mathbb{R}^+$ assigns a score to each pair of letters (the LCS corresponds to the special case where $S(a,b) = 1$ for $a = b$ and $S(a,b) = 0$ for $a \neq b$). This requires more work, but is possible, and will be presented in a separate publication (see [9]).*

(ii) Another important step would be to extend the central limit theorem result to three or more sequences. Such an attempt would require, at first, to use the variance estimates as stated in the concluding remarks of [11] and then generalize to higher dimensions the closeness to the diagonal results of [13].

(iii) As challenging is the the loss of independence, both between and inside the sequences, and the loss of identical distributions, both within and between the sequences. Results for this type of frameworks will be presented elsewhere.

(iv) It would also be of interest to study the random permutation versions of (i)–(iii) above.

References

[1] K. S. Alexander. *The rate of convergence of the mean length of the longest common subsequence.* Ann. Appl. Probab., 4(4), 1074-1082, 1994.

[2] Baik, J., Deift, P. and Johansson, K., *On the distribution of the length of the longest increasing subsequence of random permutations*, Journal of the American Mathematical Society 12 (4): 1119-1178, 1999.

[3] J.-C. Breton, C. Houdré. *On the limiting law of the longest common and increasing subsequence in random words.* In Preparation.

[4] S. Chatterjee. *A new method of normal approximation.* Ann. Probab. 36 no. 4, 1584-1610, 2008.

[5] S. Chatterjee, S. Sen. *Minimal spanning trees and Stein's method.* Preprint arXiv:math/1307.1661, 2013.

[6] L. H. Y. Chen, L. Goldstein, Q.-M. Shao, *Normal approximation by Stein's method.* Probability and its Applications. Springer, Heidelberg, 2011.

[7] V. Chvátal, D. Sankoff. *Longest common subsequences of two random sequences.* J. Appl. Probab. 12, 306-315, 1975.

[8] V. Dančík. *Common subsequences and supersequences and their expected length.* Combinatorics, Probability and Computing 7, 365-373, 1998.

[9] C. Houdré, Ü. Işlak. *A central limit theorem for sequence comparison with a general scoring function.* In Preparation.

[10] C. Houdré, J. Lember, H. Matzinger. *On the longest common increasing binary subsequence.* C.R. Acad. Sci. Paris Ser. I 343, 589–594, 2006.

[11] C. Houdré, J. Ma. *On the order of the central moments of the length of the longest common subsequence.* Preprint arXiv:1212.3265v2, 2012.

[12] C. Houdré, H. Matzinger. *On the variance of the optimal alignment score for an asymmetric scoring function.* Preprint arXiv:math/0702036, 2007.

[13] C. Houdré, H. Matzinger. *Closeness to the diagonal for longest common subsequences.* Preprint arxiv:math/0911.2031, 2011.

[14] M. Kiwi, M. Loebl, J. Matoušek. *Expected length of the longest common subsequence for large alphabets.* Adv. Math. 197 (2005), no. 2, 480-498.

[15] J. Lember, H. Matzinger. *Standard deviation of the longest common subsequence.* Ann. Probab. 37, no. 3, 1192-1235, 2009.

[16] S. N. Majumdar, S. Nechaev. *Exact asymptotic results for the Bernoulli matching model of sequence alignment.* Phys. Rev. E (3) 72, no. 2, 4 pp., 2005.

[17] W. Rhee. *On rates of convergence for common subsequences and first passage time.* Ann. Appl. Probab. 5, no. 1, 44-48, 1995.

[18] N. F. Ross. *Fundamentals of Stein's method*, Probability Surveys, 8, 210-293 (electronic), 2011.

[19] D. Romik. *The surprising mathematics of longest increasing subsequences.* Cambridge University Press, 2014.

[20] J. M. Steele. *An Efron-Stein inequality for nonsymmetric statistics.* Ann. Statist. 14, 753-758, 1986.