Electronic Journal of Probability

Electron. J. Probab. **28** (2023), article no. 3, 1–24.
ISSN: 1083-6489 https://doi.org/10.1214/22-EJP894

A central limit theorem for the length of the longest common subsequences in random words[*]

Christian Houdré[†] Ümit Işlak[‡]

Abstract

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

Keywords: longest common subsequences; random words; central limit theorem; optimal alignments; last passage percolation; Stein's method; Ulam's problem; random permutations; Tracy-Widom distribution; edit/Levenshtein distance; supersequences.
MSC2020 subject classifications: 05A05; 60C05; 60F05; 60F10.
Submitted to EJP on May 19, 2022, final version accepted on December 16, 2022.

1 Introduction

We explore here the asymptotic behavior, in law, of the length of the longest common subsequences of two random words. Although the study of this length is decade-old, and extensive from an algorithmic point of view, in various disciplines such as, computer science, bioinformatics, or statistical physics, its mathematically rigorous results are rather sparse. Below, we obtain the first result on the limiting law of this length, when properly centered and scaled.

To begin with, let us present our framework. Throughout, let $X = (X_i)_{i\geq 1}$ and $Y = (Y_i)_{i\geq 1}$ be two infinite sequences whose coordinates take their values in $\mathcal{A}_m = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_m\}$, a finite alphabet of size m. Next, let LC_n be the length of the longest

[*] CH's research was supported in part by a Simons Foundation Fellowship, grant #267336 and the grants #246283 and #524678 from the Simons Foundation.

[†] School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia, 30332-0160, USA.
E-mail: houdre@math.gatech.edu

[‡] Bogazici University, Faculty of Arts and Science, Department of Mathematics, 34342, Bebek-Istanbul, Turkey. E-mail: umit.islak1@boun.edu.tr

common subsequences (LCSs) of the random words $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$, i.e., LC_n is the maximal integer $k \in \{1, \dots, n\}$, such that there exist $1 \le i_1 < \cdots < i_k \le n$ and $1 \le j_1 < \cdots < j_k \le n$, for which:

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

As well known, LC_n is a measure of the similarity/dissimilarity of the two words/strings which is often used in pattern matching, e.g., in computer science the edit (or Levenshtein) distance is the minimal number of indels (insertions/deletions) to transform one string into the other and is therefore given by $2(n - LC_n)$. (The reader will find in [9], [32], [36] and [40] numerous examples of the relevance of longest common subsequences in various applications.)

The asymptotic study of LC_n began with the well known result of Chvátal and Sankoff [13] asserting, via a superadditivity argument, that

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

whenever, for example, $(X_i)_{i\ge1}$ and $(Y_i)_{i\ge1}$ are two independent sequences of independent identically distributed (iid) random variables having the same law.

However, to this day, the exact value of $\gamma_m^* = \sup_{n\ge1} \mathbb{E}LC_n/n$ (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 known (see Kiwi, Loebl and Matoušek ([26])) and given, for X_1 uniformly distributed, by:

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

Chvátal and Sankoff’s law of large numbers was further sharpened by Alexander ([2]) who proved that

$$\gamma_m^* n - K_A\sqrt{n\ln n} \le \mathbb{E}LC_n \le \gamma_m^* n, \tag{1.3}$$

where $K_A > 0$ is a universal constant (which depends neither on n nor on the distribution of X_1). Next, Steele [37] obtained via the Efron–Stein inequality the first upper bound on the variance of LC_n proving, in particular, that:

$$\operatorname{Var} LC_n \le \left(1 - \sum_{k=1}^{m} p_k^2\right) n, \tag{1.4}$$

where $p_k = \mathbb{P}(X_1 = \alpha_k), k = 1, \dots, m$. However, finding the order of the lower bound is much more illusive and remains unknown in many instances, in particular for iid uniform Bernoulli random variables. Some of the instances in which, and methods for which, a variance lower bound matching the linear upper bound have been obtained are further described below. Before doing so, let us state our main result:

Theorem 1.1. Let $(X_i)_{i\ge1}$ and $(Y_i)_{i\ge1}$ be two independent sequences of iid random variables with values in $\mathcal{A}_m = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_m\}$ and having the same law. Assume that $\operatorname{Var} LC_n \ge Kn$, for some positive constant K independent of $n \ge 1$. Let $0 < \eta < 1/10$, then for all $n \ge 1$,

$$d_W\left(\frac{LC_n - \mathbb{E}LC_n}{\sqrt{\operatorname{Var} LC_n}}, \mathcal{G}\right) \le C\frac{1}{n^{\frac{1}{10}-\eta}}, \tag{1.5}$$

where d_W is the Monge-Kantorovich-Wasserstein distance, where $\mathcal{G}$ a standard normal random variable and where $C > 0$ is a constant depending on K, on m, and on the distribution of X_1, but is independent of n.

Recall next that the Kolmogorov and Monge-Kantorovich-Wasserstein distances, d_K and d_W, between two probability distributions μ_1 and μ_2 on $\mathbb{R}$, are respectively defined as

$$d_K(\mu_1, \mu_2) = \sup_{h \in \mathcal{H}_1} \left| \int h d\mu_1 - \int h d\mu_2 \right|,$$

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

$$d_W(\mu_1, \mu_2) = \sup_{h \in \mathcal{H}_2} \left| \int h d\mu_1 - \int h d\mu_2 \right|,$$

where $\mathcal{H}_2 = \{h : \mathbb{R} \to \mathbb{R} : |h(x) - h(y)| \leq |x - y|\}$. Recall, further, that if μ_2 is absolutely continuous, with respect to the Lebesgue measure, and with density $\mu_2(dx)/dx$ essentially bounded, i.e., such that $\|\mu_2(dx)/dx\|_\infty < +\infty$, then,

$$d_K(\mu_1, \mu_2) \leq \sqrt{2\|\mu_2(dx)/dx\|_\infty d_W(\mu_1, \mu_2)}, \tag{1.6}$$

e.g., see Ross [35] or the Appendix in [3]. Thus, Theorem 1.1 implies via (1.6), that

$$d_K\left(\frac{LC_n - \mathbb{E}LC_n}{\sqrt{\operatorname{Var} LC_n}}, \mathcal{G}\right) \leq C^{1/2} \left(\frac{2}{\pi}\right)^{1/4} \frac{1}{n^{\frac{1}{20} - \frac{\eta}{2}}}, \tag{1.7}$$

and so, properly centered and normalized, LC_n converges in distribution to a standard normal random variable as long as $\operatorname{Var} LC_n$ is assumed to be of linear order.

Let us carefully review and discuss the assumption on the variance of LC_n present in the statement of our main theorem. As indicated in (1.4), $\operatorname{Var} LC_n \leq n$, however contradictory conjectures on the order of this variance have also appeared in the literature: A sub-linear conjecture (of order $o(n^{2/3})$) in [13] and a linear one in Waterman [39] (see also [2]). The linear order, which we believe to be the correct one, has been verified in a few situations that we briefly describe next:

• This linear lower bound is proved in [28] or [22] for iid random variables (Bernoulli or finite-alphabet ones) which are highly biased, in that a single letter is taken with very high (but fixed) probability. In that case, changing in any configuration, a low probability letters into the high probability one, is more likely to increase LC_n by one unit than to decrease it by one unit. This change (which clearly has no effect for uniformly distributed letters) reduces variability and the new longest common subsequences provide the variance lower bound.

• Beyond the strongly biased cases just mentioned, a linear order for the variance has been obtained in other situations closer to the iid uniform case. In particular, in a framework where either a letter is missing or long blocks are added within the iid uniform framework or in various other settings, as seen in the many references given in ([4], [6], [18], [21], [23], . . .).. Within these frameworks, modifications of the tools presented in our current approach would also lead to a central limit theorem, without any further assumption on the variance. In all these situations, the central r-th, $r \geq 1$, moments of LC_n can also be shown to be of order $n^{r/2}$ (see the concluding remarks in [22]). This last fact might hint at the asymptotic normality of LC_n, although similar moments estimates can lead to a non-Gaussian limiting law in a closely related problem, i.e., in the study of LCI_n, the length of the longest common and increasing subsequences of two random words, over a totally ordered finite alphabet (see [8], [16]).

• Early extensive simulations (with n of order 10^4) by Boutet de Monvel [7] seemed to indicate, in the uniform case, a variance of order at least $n^{2\omega'}$ with $\omega' \approx 0.418$ and even a normal asymptotic law. More recent extensive simulations (with n of order 10^6) (see [29]) seem to indicate (in both the uniform and non-uniform binary cases) that the variance is of order n as the lengths of the sequences are the larger to date, an order one-hundred times bigger than the ones in [7].

• As it will become clear from the proof of the theorem just stated, a mere sublinear lower bound on the variance will also lead to a normal limiting law, e.g., a lower bound of order at least $n^{9/10+\eta}$, $\eta > 0$ will do (although, and again, it is our belief that the variance of LC_n is linear in n, but nevertheless $9/10 > 2\omega'$). Note also that the proof of this theorem provides for α (to be defined) such that $4/5 < \alpha < 1$, a rate of $1/n^{(1-\alpha)/2}$, while for $2/3 < \alpha < 4/5$, a different rate, of order $1/n^{1-3(1-\alpha/2)/2}$, can be obtained in a similar way (see (2.40)), under a linear variance lower bound.

Remark 1.1. Theorem 1.1 is the first of its kind. It contrasts, in particular, with the corresponding result in the related Bernoulli matching problem where, as shown by Majumdar and Nechaev ([31]), the limiting law is the Tracy-Widom one. Both the LCS and Bernoulli matching models are directed last passage vertex/site 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 the Bernoulli matching problem is sublinear (of order $n^{2/3}$), while in our LCS case it is assumed linear. (The assumption on the order of the LCS-variance is additionally described at length in the next two paragraphs.) Let us describe how the LCS problem can be represented as a directed last passage percolation (LPP) problem with dependent weights. Indeed, let the set of vertices be

$$V := \{0, 1, 2, \ldots, n\} \times \{0, 1, 2, \ldots, n\},$$

and let the set of oriented edges $\mathcal{E} \subset V \times V$ contain horizontal, vertical and diagonal edges. The horizontal edges are oriented to the right, while the vertical edges are oriented upwards, both having unit length. The diagonal edges point up-right at a $\pi/4$-angle and have length $\sqrt{2}$. Hence,

$$\mathcal{E} := \{(v, v + e_1), (v, v + e_2), (v, v + e_3) : v \in V\},$$

where $e_1 := (1, 0)$, $e_2 := (0, 1)$ and $e_3 := (1, 1)$. With the horizontal and vertical edges, we associate a weight of 0. With the diagonal edge from (i, j) to $(i + 1, j + 1)$ we associate the weight 1 if $X_{i+1} = Y_{j+1}$ and 0 (or $-\infty$) otherwise. In this manner, we obtain that LC_n is equal to the total weight of the heaviest paths going from $(0, 0)$ to (n, n). (Another directed LPP representation can be obtained via $LC_n = \max_{\pi \in SI} \sum_{(i,j)\in\pi} \mathbf{1}_{\{X_i=Y_j\}}$, where SI refers to the set of all paths with *strictly* increasing steps, i.e., paths with *both* coordinates strictly increasing from a step to another, from $(0, 0)$ to the East, $x = n$, or North, $y = n$, boundary. A third representation would be as above but where now the paths going from $(0, 0)$ to (n, n) have either strictly increasing steps or North or East unit steps. Again to the strictly increasing steps the associated weight is $\mathbf{1}_{\{X_i=Y_j\}}$ while to the North as well as to the East unit steps is associated a weight value of 0. As a final representation one could still proceed with strictly increasing paths but with the requirement that one ends the paths with a 1.) Note that the weights in our percolation representations are not “truly 2-dimensional” and, in our opinion, this could be a further reason for the order of magnitude of the mean, variance as well as the limiting law in the LCS problem to be different from other first/last passage-related models.

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 $\{1, \ldots, n\}$. Indeed, in sequence comparison problems, the emergence of the Tracy–Widom distribution has sometimes been contemplated/speculated, e.g., see [1]. We show, in the last section of the present paper, that this is correct when analyzing the asymptotic behavior of the length of the longest common subsequences of two independent uniform random permutations of $\{1,\dots,n\}$ (the expectation there is of order $\sqrt{n}$ and the variance of order $n^{1/3}$).

Finally, let us remark that some of the ideas/techniques developed to prove lower bounds on $VarLC_n$ have been further developed in the context of first passage percolation, providing, to date, the best lower bound available on the variance of the passage time (see [14]).

As far as the content of the paper is concerned, the lengthy next section contains the proof of Theorem 1.1, which is preceded by a discussion of some elements of its proof. Then, in the third 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 two independent uniform random permutations of $\{1,\dots,n\}$ 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 the main theorem by a three-step method. The first step makes use of a relatively recent theorem of Chatterjee ([10]) on Stein's method (see [12] for an overview of the method, including Chatterjee's normal approximation results via exchangeable pairs); the second uses simple moment estimates for LC_n derived from our lower bound variance assumption; and the third develops lengthy correlation estimates based, in part, on short string-lengths genericity results obtained in [24]. We start by fixing notation and recalling some preliminaries.

Throughout this section, $X=(X_i)_{i\geq 1}$ and $Y=(Y_i)_{i\geq 1}$ are two independent sequences whose coordinates are iid, with a common law, and taking their values in $\mathcal{A}_m=\{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\dots,\boldsymbol{\alpha}_m\}$, a finite alphabet of size m.

Let us continue by introducing some more notation following those of [10]. Let $W=(W_1,W_2,\dots,W_n)$ and $W'=(W_1',W_2',...,W_n')$ be two iid $\mathbb{R}^n$-valued random vectors whose components are also independent. For $A\subset[n]:=\{1,2,\dots,n\}$, 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, W^j is short for $W^{\{j\}}$, while $W^{\emptyset}=W$.

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

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

and again, $T_{\emptyset}=\sum_{j=1}^n(\Delta_j f(W))^2$. 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, and where the sum above is taken over all the proper subsets (including $T_{\emptyset}$) of $[n]$. Here is Chatterjee's result.

Theorem 2.1. *[10] Let all the terms be defined as above, and let* $0<\sigma^2:=\operatorname{Var}f(W)<\infty$. *Then,*

$$d_W\left(\frac{f(W)-\mathbb{E}f(W)}{\sqrt{\operatorname{Var}f(W)}},\mathcal{G}\right)\leq\frac{\sqrt{\operatorname{Var}T}}{\sigma^2}+\frac{1}{2\sigma^3}\sum_{j=1}^n\mathbb{E}|\Delta_j f(W)|^3, \tag{2.1}$$

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

Remark 2.2. (i) In [10], the variance term as displayed in (2.1) is actually replaced by $\operatorname{Var}\mathbb{E}(T|f(W))$ but the above bound, with the larger $\operatorname{Var}T$, already presented in [10], is enough for our purpose.

(ii) Our proof bounds the right-hand side of (2.1) and next, using (1.6), bounds the corresponding Kolmogorov distance. An alternate way to obtain convergence in distribution would be to first use a more recent result of Lachièze-Rey and Peccati [27], directly bounding the Kolmogorov distance, which could then be estimated by adapting the techniques presented below.

Two small comments are in order before beginning the proof of Theorem 1.1.

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

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

Let us start with a sketch of proof Theorem 1.1 and to do so, set

$$W := (X_1, \dots, X_n, Y_1, \dots, Y_n), \tag{2.2}$$

and set

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

We begin by estimating the second term on the right-hand side of (2.1). To do so, recall our assumption:

$$\sigma^2 := \mathbb{E}(LC_n - \mathbb{E}LC_n)^2 \geq Kn. \tag{2.3}$$

Therefore,

$$\sigma^3 \geq Cn^{3/2}, \qquad n \geq 1, \tag{2.4}$$

yielding

$$\frac{1}{2\sigma^3} \sum_{j=1}^{2n} \mathbb{E}|\Delta_j f(W)|^3 \leq C \frac{1}{\sqrt{n}}, \tag{2.5}$$

since $|\Delta_j f(W)| \leq 1$. This last estimate takes care of the second term on the right-hand side of (2.1).

Next, let us move to the estimation of the variance term in (2.1). Setting

$$\mathcal{S}_1 := \{(A, B, j, k) : A \subsetneq [2n], B \subsetneq [2n], j \notin A, k \notin B\}, \tag{2.6}$$

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

$$\begin{aligned} \operatorname{Var}T &= \frac{1}{4} \operatorname{Var}\left(\sum_{A \subsetneq [2n]} \sum_{j \notin A} \frac{\Delta_j f(W) \Delta_j f(W^A)}{\binom{2n}{|A|}(2n - |A|)} \right) \\ &= \frac{1}{4} \sum_{A \subsetneq [2n], j \notin A} \sum_{B \subsetneq [2n], k \notin B} \frac{Cov(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W) \Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\ &= \frac{1}{4} \sum_{(A,B,j,k) \in \mathcal{S}_1} \frac{Cov(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W) \Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}. \end{aligned} \tag{2.7}$$

Our strategy is now to further divide $\mathcal{S}_1$ into two main pieces by conditioning on a, yet to be defined, high probability event E^n_{ϵ,s_1,s_2}, ensuring that LCSs are made of an accumulation of relatively short strings. More precisely,

Lemma 2.3. *Let $Z = \mathbf{1}_{E^n_{\epsilon,s_1,s_2}}$, then*

$$4\operatorname{Var} T = \sum_{(A,B,j,k)\in\mathcal{S}_1} \frac{Cov(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\mathbb{P}(Z=0) \tag{2.8}$$

$$+\sum_{(A,B,j,k)\in\mathcal{S}_1} \frac{Cov(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\mathbb{P}(Z=1). \tag{2.9}$$

To estimate each of the two terms in the above lemma, the following proposition, and a conditional version of it, which easily follows from similar arguments, will be used repeatedly throughout the proof.

Proposition 2.4. *Let $\mathcal{R}$ be a subset of $[2n]^2$, and let*

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

Let $g : \mathcal{S}^* \to \mathbb{R}$ with $\|g\|_\infty < +\infty$, then

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

Proof. First, since $\|g\|_\infty < +\infty$,

$$\sum_{(A,B,j,k)\in\mathcal{S}^*} \left| \frac{g(A,B,j,k)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \right|$$
$$\le \|g\|_\infty \sum_{(A,B,j,k)\in\mathcal{S}^*} \left(\frac{1}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \right).$$

Next, 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

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

$$
\begin{aligned}
&= \sum_{(j,k)\in\mathcal{R}} \left(\sum_{s,r=0}^{2n-1} \frac{\frac{(2n-1)!}{(2n-1-s)!s!}\,\frac{(2n-1)!}{(2n-1-r)!r!}}{\frac{(2n)!}{(2n-s)!s!}(2n-s)\frac{(2n)!}{(2n-r)!r!}(2n-r)} \right)\\
&= \sum_{(j,k)\in\mathcal{R}} \left(\sum_{s,r=0}^{2n-1} \frac{1}{(2n)^2} \right)\\
&= |\mathcal{R}|,
\end{aligned}
$$

from which the result follows. □

Taking $\mathcal{R} = [2n]^2$, $g(A,B,j,k) = Cov(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))$ which is such that $\|g\|_\infty \leq 1$, Proposition 2.4 yields the estimate

$$
\sum_{(A,B,j,k)\in\mathcal{S}_1} \left(\frac{Cov(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \right) \leq 4n^2. \tag{2.10}
$$

Hence, $\operatorname{Var} T \leq n^2$ giving a suboptimal result for our purposes, and we therefore begin a detailed estimation study to improve the variance upper bound to $o(n^2)$.

To do so, we start by giving a slight variation of a result from [24] 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 [24]), should prove useful in other contexts.

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

$$
r_0 = 0 \leq r_1 \leq r_2 \leq r_3 \leq \ldots \leq r_{d-1} \leq r_d = n, \tag{2.11}
$$

be such that

$$
LC_n = \sum_{i=1}^{d} |LCS(X_{v(i-1)+1}X_{v(i-1)+2}\cdots X_{vi}; Y_{r_{i-1}+1}Y_{r_{i-1}+2}\cdots Y_{r_i})|, \tag{2.12}
$$

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 subsequences of the words/strings $X_{v(i-1)+1}X_{v(i-1)+2}\cdots X_{vi}$ and $Y_{r_{i-1}+1}Y_{r_{i-1}+2}$ $\cdots Y_{r_i}$ (with the understanding that this length is zero if none of the letters of the X-part are aligned with letters of the Y-part, i.e., if the X-part is only 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

$$
\tilde{\gamma}(s) = \lim_{n\to\infty} \frac{\mathbb{E}LC_n(X_1\cdots X_n; Y_1\cdots Y_{sn})}{n(1+s)/2}, \quad s > 0.
$$

(See [24] 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 $(r_0, r_1, \ldots, r_d)$ satisfying (2.11) and (2.12), we have

$$
|\{i \in [d] : vs_1 \leq r_i - r_{i-1} \leq vs_2\}| \geq (1-\epsilon)d. \tag{2.13}
$$

In words, E^n_{ϵ,p_1,p_2} is the (random) set of optimal alignments of $X_1\cdots X_n$ and $Y_1\cdots Y_n$ satisfying (2.11), for which a proportion of at least $1-\epsilon$ of the integer intervals $[r_{i-1}+1, r_i]_{\mathbb{N}}$, $i = 1, 2, \ldots, d$, have their length between vs_1 and vs_2.

As stated next, E^n_{ϵ,s_1,s_2} holds with high probability. Broadly, our next theorem asserts that for any $\epsilon > 0$, there exists v large enough, but fixed, such that if X is divided into segments of length v then, typically (at least a fraction $1-\epsilon$ of segments), and with high probability, the LCSs match these segments to segments of similar length in Y.

Theorem 2.5. *[24] 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*

$$\frac{1 + \ln(1+v)}{v} \leq \frac{\delta^2 \epsilon^2}{16}. \tag{2.14}$$

Then,

$$\mathbb{P}(E^n_{\epsilon, s_1, s_2}) \geq 1 - \exp\left(-n\left(-\frac{1+\ln(1+v)}{v} + \frac{\delta^2\epsilon^2}{16}\right)\right), \tag{2.15}$$

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

Remark 2.6. In [24], instead of (2.11), the corresponding condition is:

$$r_0 = 0 < r_1 < r_2 < r_3 < ... < r_{d-1} < r_d = n. \tag{2.16}$$

which is made up of strict inequalities becoming weak inequalities in (2.11). The rationale for this difference is that, in general, there is no guarantee that there exists an optimal alignment, i.e., a longest common subsequence, satisfying both conditions (2.12) and (2.16). 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 (2.12) must have a piece (soon to be called a "cell") with no terms in the Y-part and this is clearly incompatible with (2.16). (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 (2.11) and (2.12) with, say, $v = n^\alpha$, $0 < \alpha < 1$, 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 change the framework of [24] as forthcoming arguments require the existence of an optimal alignment satisfying (2.12) for any value of X and Y. However, the proof of Theorem 2.5, above, proceeds as the proof of the corresponding result (Theorem 2.2!) in [24], 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 (2.11) is now given by

$$\binom{n+d}{d} \leq \frac{(n+d)^d}{d!} \leq \left(\frac{e(n+d)}{d}\right)^d = (e(1+v))^d, \tag{2.17}$$

leading to the terms involving $\ln(1+v)$ rather than just $\ln v$ [24], when using (2.16) and the estimate $n^d/d! \leq (ev)^d$ to upper-bound $\binom{n}{d}$.)

Remark 2.7. In [24], the statement corresponding to Theorem 2.5 is given for "all n large enough". However, as indicated at the end of the proof there, it is possible to find a more quantitative estimate using Alexander's result (1.3). In fact a lower bound, in terms of ϵ and δ, is valid for all $n \geq 1$. Indeed, at first, from the end of the proof of Lemma 3.1 there, preceding the main theorem in [24], one can easily verify that the following relation between n and ϵ is sufficient for (2.15) to hold:

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

where $0 < \delta < \delta^* := \min(\gamma_m^* - \tilde{\gamma}(s_1), \gamma_m^* - \tilde{\gamma}(s_2))$ is a fixed positive quantity and K_A is a positive constant such that $\gamma_m^* n - K_A\sqrt{n \ln n} \leq \mathbb{E}LC_n$. (One can find explicit numerical estimates on K_A using Rhee's [33] proof of (1.3).)

In our context, here is how to choose ϵ so that the estimate in (2.15) 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

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

Setting,

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

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

$$\mathbb{P}(E^n_{\epsilon,s_1,s_2}) \geq 1 - e^{-n^{1-\alpha}(1+\ln(1+n^\alpha))} \geq 1 - e^{-(1+\ln 2)}, \tag{2.18}$$

for all $n \geq 1$.

Let us return to the proof of Theorem 1.1, and the estimation of (2.7). First, for notational convenience, below we write $\sum_{\mathcal{S}_1}$ 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 $\epsilon = c_1\sqrt{(1+\ln(1+v))/v}$, i.e., let $Z = \mathbf{1}_{E^n_{\epsilon,s_1,s_2}}$, with $v = n^\alpha$ and with c_1 as in Remark 2.7. Then, we arrive at the decomposition of Lemma 2.3

$$\begin{aligned}\sum_{\mathcal{S}_1} & \frac{Cov(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\ &= \sum_{\mathcal{S}_1} \frac{Cov_{Z=0}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \mathbb{P}(Z=0) \\ &+ \sum_{\mathcal{S}_1} \frac{Cov_{Z=1}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \mathbb{P}(Z=1).\end{aligned} \tag{2.19}$$

To estimate the first term on the right-hand side of (2.19), first note that

$$Cov_{Z=0}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B)) \leq 1,$$

which when combined with the estimate in (2.18) and (2.10), immediately leads to

$$\begin{aligned}\sum_{\mathcal{S}_1} & \frac{Cov_{Z=0}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \mathbb{P}(Z=0) \\ &\leq 4n^2 e^{-n^{1-\alpha}(1+\ln(1+n^\alpha))}.\end{aligned} \tag{2.20}$$

For the second term on the right-hand side of (2.19), begin with the trivial bound on $\mathbb{P}(Z=1)$ to get

$$\begin{aligned}\sum_{\mathcal{S}_1} & \frac{Cov_{Z=1}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \mathbb{P}(Z=1) \\ &\leq \sum_{\mathcal{S}_1} \frac{Cov_{Z=1}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}.\end{aligned} \tag{2.21}$$

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 2.6 that there always exists an optimal alignment $\mathbf{r} = (r_0, r_1, r_2, ..., r_d)$ satisfying both (2.11) and (2.12) with $v = n^\alpha$, $0 < \alpha < 1$. In the sequel, $\mathbf{r}$ denotes such a (fixed) optimal alignment which also specifies the pairs, in the strings $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$, contributing to the longest common subsequence.[1] Such an alignment always exists, as just noted, and so we can define an injective map from $(X_1 \cdots X_n, Y_1 \cdots Y_n)$ to the set of alignments, making various definitions (such as the ones for $\mathcal{S}_{1,1}$ and $\mathcal{S}_{1,2}$ on page 12) below well defined. 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 (2.13).

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

Definition 2.8. *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, any optimal alignment with $v = n^\alpha$ has $d = n^{1-\alpha}$ cells.

Let us next introduce some more notation which will be 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, as the following example shows:

Example 2.9. Let $n = 12$ and $\mathcal{A} = [3]$, and let

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

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

and $W = (X, Y)$. Then, $LC_{12} = 8$, obtained for example through $(1, 1, 2, 1, 2, 1, 1, 1)$, while 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

$$(X_{10}X_{11}X_{12}; Y_{10}Y_{11}Y_{12}) = (131; 112).$$

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

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

Returning to the proof of Theorem 1.1, 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

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

[1] This alignment might change according to different realizations of the words $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$, i.e., it is random; but, for each realization and although there are possibly more than one choice of optimal alignments, we just fix one of those.

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

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

and, moreover, 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.

Continuing with the decomposition of the right-hand side of (2.21),

$$\begin{aligned}
&\sum_{\mathcal{S}_1} \frac{Cov_{Z=1}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\
= \quad &\sum_{\mathcal{S}_1} \frac{Cov_{Z=1,\mathcal{S}_{1,1}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\
&+\sum_{\mathcal{S}_1} \frac{Cov_{Z=1,\mathcal{S}_{1,2}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}, \qquad (2.22)
\end{aligned}$$

where to further clarify the notation note that, for example,

$$\begin{aligned}
&\sum_{\mathcal{S}_1} \frac{Cov_{Z=1,\mathcal{S}_{1,1}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\
&\qquad = \sum_{\mathcal{S}_1} \mathbb{E}\left(\frac{g(A,B,j,k)\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \Big| Z = 1 \right),
\end{aligned}$$

where

$$\begin{aligned}
g(A,B,j,k) \;=\; &\left(\Delta_j f(W)\Delta_j f(W^A) - \mathbb{E}(\Delta_j f(W)\Delta_j f(W^A))\right) \\
&\times\left(\Delta_k f(W)\Delta_k f(W^B) - \mathbb{E}(\Delta_k f(W)\Delta_k f(W^B))\right). \qquad (2.23)
\end{aligned}$$

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", while a further next step will be based on estimations for the indices in $\mathcal{S}_{1,2}$. Here we will observe that, as the letters are in different cells, we have enough independence (see the decomposition in (2.29)) to show that the contributions of the covariance terms from $\mathcal{S}_{1,2}$ are "small".

Let us now focus on the first term on the right-hand side of (2.22). Letting g be as in (2.23), and using arguments similar to those used in the proof of Proposition 2.4, we have,

$$\begin{aligned}
&\sum_{\mathcal{S}_1} \frac{\left|Cov_{Z=1,\mathcal{S}_{1,1}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))\right|}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\
&\qquad \le \mathbb{E}\left(\sum_{\mathcal{S}_1} \frac{|g(A,B,j,k)|\,\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \Big| Z = 1 \right) \\
&\qquad \le 4\mathbb{E}\left(\sum_{\mathcal{S}_1} \frac{\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \Big| Z = 1 \right) \\
&\qquad = 4\mathbb{E}\left(|\mathcal{R}||Z = 1\right), \qquad (2.24)
\end{aligned}$$

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 (2.24), 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 $\mathcal{T}_i$ be the event that $s_1 n^\alpha \le r_i - r_{i-1} \le s_2 n^\alpha$. Then,

$$\begin{aligned}\mathbb{E}\left(|\mathcal{R}|\,|Z=1\right) &= \sum_{i=1}^{n^{1-\alpha}} \mathbb{E}(|\mathcal{R}_i|\,|Z=1) \\ &= \sum_{i=1}^{n^{1-\alpha}} \mathbb{E}(|\mathcal{R}_i|\mathbf{1}_{\mathcal{T}_i}\,|Z=1) + \sum_{i=1}^{n^{1-\alpha}} \mathbb{E}(|\mathcal{R}_i|\mathbf{1}_{\mathcal{T}_i^c}\,|Z=1). \qquad (2.25)\end{aligned}$$

For the first term on the right-hand side of (2.25), note that, when $\mathcal{T}_i$ holds true, the X-part of the i-th cell can contain at most n^α letters while the Y-part can contain at most $s_2 n^\alpha$ ones. Thus,

$$|\mathcal{R}_i|\mathbf{1}_{\mathcal{T}_i} \le s_2 n^{2\alpha},$$

and this leads to:

$$\sum_{i=1}^{n^{1-\alpha}} \mathbb{E}(|\mathcal{R}_i|\mathbf{1}_{\mathcal{T}_i}\big|Z=1) \le s_2 n^{1+\alpha}. \qquad (2.26)$$

For the estimation of the second term on the right-hand side of (2.25), we first observe that letting $I := \{i \in [n^{1-\alpha}] : \mathcal{T}_i \text{ does not occur}\}$, we have

$$\sum_{i=1}^{n^{1-\alpha}} \mathbb{E}(|\mathcal{R}_i|\mathbf{1}_{\mathcal{T}_i^c}\big|Z=1) = \mathbb{E}\left(\sum_{i=1}^{n^{1-\alpha}} |\mathcal{R}_i|\mathbf{1}_{\mathcal{T}_i^c}|Z=1\right) = \mathbb{E}\left(\sum_{i\in I} |\mathcal{R}_i|\big|Z=1\right).$$

Noting that $|\mathcal{R}_i| \le 4n^2$, we have

$$\mathbb{E}\left(\sum_{i\in I} |\mathcal{R}_i|\big|Z=1\right) \le 4n^2 \mathbb{E}\left(\sum_{i\in I} 1\,\big|Z=1\right) \le 4n^2 \mathbb{E}\left(|I|\,\big|\,Z=1\right).$$

Next, by definition, given that $Z=1$, $|I| \le \epsilon n^{1-\alpha}$ and so

$$\mathbb{E}\left(|I|\,\big|\,Z=1\right) \le \epsilon n^{1-\alpha} = c_1 \left(\frac{1+\ln(1+n^\alpha)}{n^\alpha}\right)^{1/2} n^{1-\alpha} \le C n^{1-3\alpha/2} (\ln n^\alpha)^{1/2}.$$

Thus, we obtain

$$\sum_{i=1}^{n^{1-\alpha}} \mathbb{E}(|\mathcal{R}_i|\mathbf{1}_{\mathcal{T}_i^c}\big|Z=1) \le 4n^2 \mathbb{E}\left(|I|\,\big|\,Z=1\right) \le C n^{3-3\alpha/2} (\ln n^\alpha)^{1/2},$$

and when $\alpha > 2/3$, the above right-hand side is $o(n^2)$.

Hence, using also (2.26), it follows that:

$$\mathbb{E}(|\mathcal{R}||Z=1) \le C n^{1+\alpha} + C n^{3-3\alpha/2} (\ln n^\alpha)^{1/2}, \qquad (2.27)$$

which, in turn, yields via (2.24),

$$\begin{aligned}\sum_{\mathcal{S}_1} \frac{\left|Cov_{Z=1,\mathcal{S}_{1,1}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))\right|}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\ \le C n^{1+\alpha} + C n^{3-3\alpha/2} (\ln n^\alpha)^{1/2}. \qquad (2.28)\end{aligned}$$

This last estimate takes care of the first sum on the right-hand side of (2.22) as, again, this last right-hand side is $o(n^2)$, when $2/3 < \alpha < 1$.

Hence, from here on, we henceforth assume that α *is a real greater than* $2/3$ *and smaller than* 1.

We move next to estimating the second term on the right-hand side of (2.22), which is given by:

$$\sum_{\mathcal{S}_1} \frac{Cov_{Z=1,\mathcal{S}_{1,2}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B))}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}.$$

To estimate the summands in the above expression, we decompose the covariance terms in such a way that (conditional) independence of certain random variables occurs, therefore simplifying the estimates themselves. For this purpose, for each $i \in [2n]$, let $f(P_i) = LC(P_i)$ be the length of the longest common subsequences of P_i^1 and P_i^2, the coordinates of the cell $P_i = (P_i^1; P_i^2)$. Now, 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 subsequences of the two random words forming P_i, and the length of their modified versions at coordinate i, i.e., the words forming P_i'. Now for $(A, B, j, k) \in \mathcal{S}_1$,

$$\begin{aligned}
&Cov_{Z=1,\mathcal{S}_{1,2}}(\Delta_j f(W)\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B)) = \\
&\quad Cov_{Z=1,\mathcal{S}_{1,2}}((\Delta_j f(W) - \tilde{\Delta}_j f(W))\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B)) \\
&\quad + Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)(\Delta_j f(W^A) - \tilde{\tilde{\Delta}}_j f(W^A)), \Delta_k f(W)\Delta_k f(W^B)) \\
&\quad + Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\tilde{\Delta}}_j f(W^A), (\Delta_k f(W) - \tilde{\Delta}_k f(W))\Delta_k f(W^B)) \\
&\quad + Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\tilde{\Delta}}_j f(W^A), \tilde{\Delta}_k f(W)(\Delta_k f(W^B) - \tilde{\tilde{\Delta}}_k f(W^B))) \\
&\quad + Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\tilde{\Delta}}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\tilde{\Delta}}_k f(W^B)),
\end{aligned} \tag{2.29}$$

where, for any $i \notin A$, we also set $\tilde{\tilde{\Delta}}_i f(W^A) = f(W^A\big|_{P_i}) - f(W^{A\cup\{i\}}\big|_{P_i})$, with $W^A\big|_{P_i}$ and $W^{A\cup\{i\}}\big|_{P_i}$ being the restrictions of W^A and $W^{A\cup\{i\}}$ to the cell P_i, respectively. 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 step, this decomposition is akin to a decomposition developed in [11].)

Let us start by estimating the last term on the right-hand side of (2.29). Letting $\xi_j := \tilde{\Delta}_j f(W)\tilde{\tilde{\Delta}}_j f(W^A)$ and $\xi_k := \tilde{\Delta}_k f(W)\tilde{\tilde{\Delta}}_k f(W^B)$, with a slight abuse of notation as ξ_j depends on A, while ξ_k depends on B, we have

$$\begin{aligned}
&Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\tilde{\Delta}}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\tilde{\Delta}}_k f(W^B)) \\
&\qquad = \mathbb{E}((\xi_j - \mathbb{E}\xi_j)(\xi_k - \mathbb{E}\xi_k)\mathbf{1}((A, B, j, k) \in \mathcal{S}_{1,2})|Z = 1) \\
&\qquad = \frac{\mathbb{E}((\xi_j - \mathbb{E}\xi_j)(\xi_k - \mathbb{E}\xi_k)\mathbf{1}((A, B, j, k) \in \mathcal{S}_{1,2}, Z = 1))}{\mathbb{P}(Z = 1)} \\
&\qquad = \frac{\mathbb{E}((\xi_j - \mathbb{E}\xi_j)(\xi_k - \mathbb{E}\xi_k)\mathbf{1}((A, B, j, k) \in \mathcal{S}_{1,2}))}{\mathbb{P}(Z = 1)} \\
&\qquad\quad - \frac{\mathbb{E}((\xi_j - \mathbb{E}\xi_j)(\xi_k - \mathbb{E}\xi_k)\mathbf{1}((A, B, j, k) \in \mathcal{S}_{1,2}, Z = 0))}{\mathbb{P}(Z = 1)}
\end{aligned}$$

Note that the second term on the right-most equality is, in absolute value, at most

$$Ce^{-n^{1-\alpha}(1+\ln(1+n^\alpha))},$$

as from (2.18), $\mathbb{P}(Z = 1) \geq 1 - 1/(2e)$ and $\mathbb{P}(Z = 0) \leq e^{-n^{1-\alpha}(1+\ln(1+n^\alpha))}$.

So we focus on the other term and evaluate $\mathbb{E}((\xi_j - \mathbb{E}\xi_j)(\xi_k - \mathbb{E}\xi_k)\mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,2}))$. We have

$$\begin{aligned}&\mathbb{E}((\xi_j - \mathbb{E}\xi_j)(\xi_k - \mathbb{E}\xi_k)\mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,2}))\\ &\qquad = \mathbb{E}((\xi_j - \mathbb{E}\xi_j)(\xi_k - \mathbb{E}\xi_k) \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,2})\mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,2}),\end{aligned}$$

and, by conditional independence, the above right-hand side is equal to

$$\mathbb{E}((\xi_j - \mathbb{E}\xi_j) \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,2}))\mathbb{E}((\xi_k - \mathbb{E}\xi_k) \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,2}))\mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,2}).$$

Now,

$$\begin{aligned}&\mathbb{E}((\xi_j - \mathbb{E}\xi_j) \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,2}))\\ &\quad= \mathbb{E}((\xi_j \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,2})) - \mathbb{E}\xi_j\\ &\quad= \mathbb{E}((\xi_j \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,2}))\\ &\qquad - \mathbb{E}((\xi_j \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,2}))\mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,2})\\ &\qquad - \mathbb{E}((\xi_j \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,1}))\mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,1}).\end{aligned}$$

Using elementary manipulations, the last equality can be rewritten as

$$(\mathbb{E}((\xi_j \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,2})) - \mathbb{E}((\xi_j \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,1})))\mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,1}).$$

Following exactly the same steps, write $\mathbb{E}((\xi_k - \mathbb{E}\xi_k) \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,2}))$ as

$$(\mathbb{E}((\xi_k \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,2})) - \mathbb{E}((\xi_k \mid \mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,1})))\mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,1}).$$

Combining these observations, and using again (2.18), $\mathbb{P}(Z=1) \geq 1 - 1/(2e)$, $n \geq 1$, we obtain

$$\begin{aligned}&Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\tilde{\Delta}}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\tilde{\Delta}}_k f(W^B))\\ &\quad\leq \frac{C\mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,2})(\mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,1}))^2}{\mathbb{P}(Z=1)} + C\mathbb{P}(Z=0)\\ &\quad\leq C\mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,1}) + C\mathbb{P}(Z=0)\\ &\quad\leq C(\mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,1}, Z=1) + \mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,1}, Z=0)) + C\mathbb{P}(Z=0)\\ &\quad\leq C\mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,1} \mid Z=1)\mathbb{P}(Z=1) + C\mathbb{P}(Z=0).\end{aligned}$$

We therefore arrive at:

$$\begin{aligned}&\sum_{\mathcal{S}_1} \frac{\left|Cov_{Z=1,\mathcal{S}_{1,2}}(\tilde{\Delta}_j f(W)\tilde{\tilde{\Delta}}_j f(W^A), \tilde{\Delta}_k f(W)\tilde{\tilde{\Delta}}_k f(W^B))\right|}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\ &\qquad\leq \sum_{\mathcal{S}_1} \frac{C\mathbb{P}((A,B,j,k) \in \mathcal{S}_{1,1}|Z=1)\mathbb{P}(Z=1) + C\mathbb{P}(Z=0)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\ &\qquad\leq \sum_{\mathcal{S}_1} \frac{C\mathbb{E}(\mathbf{1}((A,B,j,k) \in \mathcal{S}_{1,1})|Z=1) + C\mathbb{P}(Z=0)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\ &\qquad\leq Cn^{1+\alpha} + Cn^2 e^{-n^{1-\alpha}(1+\log(1+n^\alpha))},\end{aligned} \tag{2.30}$$

where for the last step (2.27) is used, as well as the estimates in (2.10) and (2.18).

We continue by obtaining upper bounds for the first four summands in (2.29). We just focus on the estimation of the first of these four terms since the other three can be estimated in a similar way. Indeed, it will be clear from the discussion below that the third of these four terms can be estimated in exactly the same way as done for the first of the four. Also, with steps similar to the ones performed in estimating this first term, one can easily see that the estimation of the second and fourth of these terms reduces to the estimation of

$$\mathbb{E}_{Z=1,\mathcal{S}_{1,2}}|\Delta_j f(W^A) - \tilde{\tilde{\Delta}}_j f(W^A)|.$$

(Again, and throughout, $\mathbb{E}_{Z=1}$ is short for conditional expectation given $\{Z=1\}$, while $\mathbb{E}_{Z=1,\mathcal{S}_{1,2}}(\cdot) = \mathbb{E}_{Z=1}(\cdot \mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}})$.) Next,

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

Now, writing $\mathcal{S}^A_{1,2}$ in place of $\mathcal{S}_{1,2}$ when using the sequence W^A instead of W, the last inequality, just above, leads to:

$$\begin{aligned}
&\mathbb{E}_{Z=1,\mathcal{S}_{1,2}}|\Delta_j f(W^A) - \tilde{\tilde{\Delta}}_j f(W^A)| \\
&\quad \le \frac{\mathbb{E}\left(|\Delta_j f(W^A) - \tilde{\tilde{\Delta}}_j f(W^A)|\right)}{\mathbb{P}(Z=1)} \\
&\quad = \frac{\mathbb{E}\left(|\Delta_j f(W^A) - \tilde{\tilde{\Delta}}_j f(W^A)|\mathbf{1}((A,B,j,k)\in\mathcal{S}^A_{1,2})\right)}{\mathbb{P}(Z=1)} \\
&\qquad + \frac{\mathbb{E}\left(|\Delta_j f(W^A) - \tilde{\tilde{\Delta}}_j f(W^A)|\mathbf{1}((A,B,j,k)\notin\mathcal{S}^A_{1,2})\right)}{\mathbb{P}(Z=1)}.
\end{aligned} \tag{2.31}$$

Then, since

$$\begin{aligned}
&|\Delta_j f(W^A) - \tilde{\tilde{\Delta}}_j f(W^A)|\mathbf{1}((A,B,j,k)\in\mathcal{S}^A_{1,2}) \\
&\qquad =_d |\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}((A,B,j,k)\in\mathcal{S}_{1,2}),
\end{aligned}$$

the first term on the right-hand side of (2.31) is equal to

$$\frac{\mathbb{E}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}((A,B,j,k)\in\mathcal{S}_{1,2})\right)}{\mathbb{P}(Z=1)}, \tag{2.32}$$

which we will estimate further, below, when working out the estimation of the first term on the right-hand side of (2.31). Also, for the second term in (2.31), noting that $|\Delta_j f(W) - \tilde{\tilde{\Delta}}_j f(W)| \le 2$, and that by the iid assumption W and W^A are identically distributed, we have

$$\begin{aligned}
&\frac{\mathbb{E}\left(|\Delta_j f(W^A) - \tilde{\tilde{\Delta}}_j f(W^A)|\mathbf{1}((A,B,j,k)\notin\mathcal{S}^A_{1,2})\right)}{\mathbb{P}(Z=1)} \\
&\qquad = \frac{\mathbb{E}\left(|\Delta_j f(W) - \tilde{\tilde{\Delta}}_j f(W)|\mathbf{1}((A,B,j,k)\notin\mathcal{S}_{1,2})\right)}{\mathbb{P}(Z=1)} \\
&\qquad \le C\mathbb{E}(\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}} \mid Z=1),
\end{aligned}$$

and then

$$\sum_{\mathcal{S}_1} \frac{\mathbb{E}\left(|\Delta_j f(W^A) - \tilde{\tilde{\Delta}}_j f(W^A)|\mathbf{1}((A,B,j,k) \notin \mathcal{S}_{1,2}^A)\right)}{\mathbb{P}(Z=1)\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}$$
$$\leq C \sum_{\mathcal{S}_1} \frac{\mathbb{E}(\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}} \mid Z=1)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}.$$

But, this last term on the right-hand side was already shown to be bounded by $Cn^{1+\alpha} + Cn^{3-3\alpha/2}(\ln n^\alpha)^{1/2}$, while reaching out (2.28). Therefore, focusing on the estimation of $\mathbb{E}_{\mathcal{S}_{1,2}}|\Delta_j f(W) - \tilde{\Delta}_j f(W)|/\mathbb{P}(Z=1)$ or, indeed, merely on the estimation of $\mathbb{E}_{\mathcal{S}_{1,2}}|\Delta_j f(W) - \tilde{\Delta}_j f(W)|$, will suffice for our purposes for the second and the fourth of the terms in (2.29). This will be done while discussing the estimation of the first term below as noted earlier.

So, we can now focus on estimating the first term in (2.29) (and as already indicated, similar arguments will provide a similar estimate for the other three terms) which is given by:

$$Cov_{Z=1,\mathcal{S}_{1,2}}((\Delta_j f(W) - \tilde{\Delta}_j f(W))\Delta_j f(W^A), \Delta_k f(W)\Delta_k f(W^B)).$$

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,

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

and 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 $|\Delta_j f(W^A)(\Delta_k f(W)\Delta_k f(W^B))| \leq 1$,

$$T_1 \leq \mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right). \tag{2.33}$$

A similar estimate also reveals that

$$T_2 \leq \mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right). \tag{2.34}$$

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

$$\begin{aligned}
T_3 + T_4 \leq 2\mathbb{E}|U| &\leq 2\mathbb{E}|\Delta_j f(W) - \tilde{\Delta}_j f(W)| \\
&= 2\mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)\mathbb{P}(Z=1) \\
&\quad +2\mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}\right)\mathbb{P}(Z=1) \\
&\quad +2\mathbb{E}_{Z=0}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)\mathbb{P}(Z=0)
\end{aligned}$$

$$\begin{aligned} &+2\mathbb{E}_{Z=0}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}\right)\mathbb{P}(Z=0)\\ \leq\ \ & 2\mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)\\ &+2\mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}\right)\\ &+Ce^{-n^{1-\alpha}(1+\ln(1+n^{\alpha}))}, \end{aligned} \tag{2.35}$$

where $\mathbb{E}_{Z=0}$ (resp. $\mathbb{E}_{Z=1}$) is short for the conditional expectation given $\{Z=0\}$ (resp. given $\{Z=1\}$), and where we used the trivial bound on $\mathbb{P}(Z=1)$, and also (2.18), 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 (2.29). Then, performing estimations as in getting (2.33), (2.34) and (2.35), for the first and third term of this sum, and keeping in mind the discussion following (2.31), so that similar estimates also hold true for the second and fourth term of the sum, we obtain

$$\begin{aligned} &\sum_{\mathcal{S}_1}\left|\frac{h(A,B,j,k)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\right|\\ &\quad\leq C\sum_{\mathcal{S}_1}\frac{\mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\ &\quad\quad+C\sum_{\mathcal{S}_1}\frac{\mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\ &\quad\quad+C\sum_{\mathcal{S}_1}\frac{\mathbb{E}_{Z=1}\left(|\Delta_k f(W) - \tilde{\Delta}_k f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\ &\quad\quad+C\sum_{\mathcal{S}_1}\frac{\mathbb{E}_{Z=1}\left(|\Delta_k f(W) - \tilde{\Delta}_k f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\ &\quad\quad+C\sum_{\mathcal{S}_1}\frac{e^{-n^{1-\alpha}(1+\ln(1+n^{\alpha}))}}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}. \end{aligned}$$

Noting that the sums involving k's are identical to the sums involving j's, we rewrite this last upper bound as

$$\begin{aligned} &\sum_{\mathcal{S}_1}\left|\frac{h(A,B,j,k)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\right|\\ &\quad\leq C\sum_{\mathcal{S}_1}\frac{\mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\ &\quad\quad+C\sum_{\mathcal{S}_1}\frac{\mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,1}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}\\ &\quad\quad+C\sum_{\mathcal{S}_1}\frac{e^{-n^{1-\alpha}(1+\ln(1+n^{\alpha}))}}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}. \end{aligned}$$

As with previous computations, using (2.10) and (2.18), the third sum on the above right-hand side is itself upper-bounded by

$$Cn^2e^{-n^{1-\alpha}(1+\ln(1+n^{\alpha}))}, \tag{2.36}$$

while, using (2.24) and (2.27), the middle sum is upper-bounded by

$$Cn^{1+\alpha}. \tag{2.37}$$

Therefore, we are just left with estimating

$$\sum_{\mathcal{S}_1} \frac{\mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}.$$

Noting that

$$\begin{aligned}\sum_{\mathcal{S}_1} &\frac{\mathbb{E}_{Z=1}\left(|\Delta_j f(W) - \tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \\ &\leq \sum_{\mathcal{S}_1} \frac{\mathbb{E}_{Z=1}|\Delta_j f(W) - \tilde{\Delta}_j f(W)|}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)}, \end{aligned} \tag{2.38}$$

we can just focus on estimating $\mathbb{E}_{Z=1}|\Delta_j f(W) - \tilde{\Delta}_j f(W)|$. To do so, the following simple proposition will be useful.

Proposition 2.10. *For any $j \in [2n]$,*

$$\Delta_j f(W) \leq \widetilde{\Delta}_j f(W).$$

Proof. Assume not, and that $\Delta_j f(W) > \widetilde{\Delta}_j f(W)$. Then either $\Delta_j f(W) = 1$ and $\widetilde{\Delta}_j f(W) = 0$, or $\Delta_j f(W) = 0$ and $\widetilde{\Delta}_j f(W) = -1$. Consider the former. Then, changing the jth coordinate does not affect the length of the longest common subsequence of the cell containing j. Since the coordinates outside that particular cell have not been changed, the overall length of the longest common subsequence cannot decrease, that is, Δ_j cannot be 1. The other case is similar. □

Returning to the estimation of $\mathbb{E}_{Z=1}|\Delta_j f(W) - \tilde{\Delta}_j f(W)|$, using the domination property obtained in Proposition 2.10, we have

$$\mathbb{E}_{Z=1}|\Delta_j f(W) - \tilde{\Delta}_j f(W)| = \mathbb{E}_{Z=1}(\tilde{\Delta}_j f(W)) - \mathbb{E}_{Z=1}(\Delta_j f(W)).$$

We now claim that both terms on the right-hand side of the last expression are exponentially small in n. Let us first deal with $\mathbb{E}_{Z=1}(\Delta_j f(W))$, the other term, which is similar, is dealt with afterwards.

We have

$$\begin{aligned}\mathbb{E}_{Z=1}(\Delta_j f(W)) &= \mathbb{E}_{Z=1}(\Delta_j f(W)\mathbf{1}(Z^j = 1)) + \mathbb{E}_{Z=1}(\Delta_j f(W)\mathbf{1}(Z^j = 0)) \\ &= \frac{\mathbb{E}(\Delta_j f(W)\mathbf{1}(Z=1)\mathbf{1}(Z^j = 1))}{\mathbb{P}(Z=1)} \\ &\quad + \frac{\mathbb{E}(\Delta_j f(W)\mathbf{1}(Z=1)\mathbf{1}(Z^j = 0))}{\mathbb{P}(Z=1)},\end{aligned}$$

where Z^j is the indicator random variable defined in the same way as Z, except that the j^{th} coordinate of W is replaced by the independent copy W'_j. Note that, for any $j \in [2n]$, Z and Z^j are identically distributed but that they are certainly not independent.

Looking, first, at the second term in the last expression, we have, with the help of (2.18), and since Z and Z^j are identically distributed,

$$\frac{\left|\mathbb{E}(\Delta_j f(W)\mathbf{1}(Z=1)\mathbf{1}(Z^j=0))\right|}{\mathbb{P}(Z=1)} \leq \frac{\mathbb{P}(Z^j=0)}{\mathbb{P}(Z=1)} \leq Ce^{-n^{1-\alpha}(1+\ln(1+n^{\alpha}))}.$$

Also, writing

$$\begin{aligned}\mathbb{E}(\Delta_j f(W)\mathbf{1}(Z=1)\mathbf{1}(Z^j=1)) &= \mathbb{E}((f(W)-f(W^j))\mathbf{1}(Z=1)\mathbf{1}(Z^j=1))\\ &= \mathbb{E}(f(W)\mathbf{1}(Z=1)\mathbf{1}(Z^j=1))\\ &\quad - \mathbb{E}(f(W^j)\mathbf{1}(Z=1)\mathbf{1}(Z^j=1))\\ &= 0,\end{aligned}$$

since, again, Z and Z^j are identically distributed. These observations yield

$$|\mathbb{E}_{Z=1}(\Delta_j f(W))| \leq Ce^{-n^{1-\alpha}(1+\ln(1+n^{\alpha}))}.$$

Similarly, noting that the expectation is conditional on $Z=1$, replacing n by n^{α}, we have

$$|\mathbb{E}_{Z=1}(\tilde{\Delta}_j f(W)\mathbf{1}((A,B,j,k)\in\mathcal{S}_{1,2}))| \leq Ce^{-n^{(1-\alpha)\alpha}(1+\log(1+n^{\alpha^2}))}.$$

(The reason for this last inequality is the fact that the configurations belong to $\mathcal{S}_{1,2}$ and, in that case, we just deal with a scaled version of the LCS problem.)

Now, note that

$$\begin{aligned}\left|\mathbb{E}_{Z=1}(\tilde{\Delta}_j f(W))\right| &\leq |\mathbb{E}_{Z=1}(\tilde{\Delta}_j f(W)\mathbf{1}((A,B,j,k)\in\mathcal{S}_{1,1}))|,\\ &\quad + |\mathbb{E}_{Z=1}(\tilde{\Delta}_j f(W)\mathbf{1}((A,B,j,k)\in\mathcal{S}_{1,2}))|,\end{aligned}$$

and, via Proposition 2.4,

$$\sum_{\mathcal{S}_1}\frac{|\mathbb{E}_{Z=1}(\tilde{\Delta}_j f(W)\mathbf{1}((A,B,j,k)\in\mathcal{S}_{1,1}))|}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \leq Cn^{1+\alpha},$$

and

$$\sum_{\mathcal{S}_1}\frac{|\mathbb{E}_{Z=1}(\tilde{\Delta}_j f(W)\mathbf{1}((A,B,j,k)\in\mathcal{S}_{1,2}))|}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \leq Cn^2e^{-n^{(1-\alpha)\alpha}(1+\log(1+n^{\alpha^2}))},$$

which, when combined, yields

$$\sum_{\mathcal{S}_1}\frac{|\mathbb{E}_{Z=1}(\tilde{\Delta}_j f(W))|}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \leq Cn^{1+\alpha}.$$

Thus, from (2.38) and the above estimates,

$$\sum_{\mathcal{S}_1}\frac{\mathbb{E}_{Z=1}\left(|\Delta_j f(W)-\tilde{\Delta}_j f(W)|\mathbf{1}_{(A,B,j,k)\in\mathcal{S}_{1,2}}\right)}{\binom{2n}{|A|}(2n-|A|)\binom{2n}{|B|}(2n-|B|)} \leq Cn^{1+\alpha}. \tag{2.39}$$

Combining (2.20), (2.28), (2.30), (2.36), (2.37) and (2.39) finally gives

$$\operatorname{Var} T \leq C\left(n^2e^{-n^{1-\alpha}(1+\ln(1+n^{\alpha}))} + n^{1+\alpha} + n^{3-3\alpha/2}(\ln n^{\alpha})^{1/2}\right). \tag{2.40}$$

Therefore, Theorem 2.1 and (2.5), ensure that:

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

holds for every $n\geq 1$, with $C>0$ a constant independent of n, and for $\alpha>4/5$ as then $1+\alpha>3-3\alpha/2$. □

Remark 2.11. (i) The constant C in Theorem 1.1 is independent of n but depends on m, on α, on s_1 and s_2 of Theorem 2.5, which in turn depend on the distribution of X_1, as well as on the quantities involved in the constant K and C in (2.3)–(2.5).

(ii) Of course, there is no reason for our rate $1/n^{(1-\alpha)/2}$ to be sharp (as previously mentioned, for $2/3 < \alpha < 4/5$, the rate $1/n^{1-3(1-\alpha/2)/2}$ is possible). Also, instead of the choice $v = n^\alpha$, a choice such as $v = h(n)$, for some optimal function h would improve this rate. Can we conjecture that the optimal rate in Kolmogorov distance is $1/\sqrt{n}$?

(iii) From a known duality between the length of a longest common subsequence of two random words and the length of a shortest common supersequence (see Dančík [15]), our result also implies a central limit theorem for this latter case.

3 Concluding remarks

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

Theorem 1.1 shows that the Gaussian distribution appears as the limiting law for the length of the longest common subsequences of two random words. However, the Tracy-Widom distribution has also been hypothesized as the limiting law in sequence comparison problems, e.g., [1]. It turns out, as shown next, that it is indeed the case for certain distributions on permutations.

First, it is folklore that, if $\pi = (\pi_1, \dots, \pi_n)$ is any element of the symmetric group $\mathfrak{S}_n$, then

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

where $LI_n(\pi)$ is the length of the longest increasing subsequence in $\pi = (\pi_1, \dots, \pi_n)$, while $LC_n((1, 2, \dots, n), (\pi_1, \pi_2, \dots, \pi_n))$, 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 with an arbitrary permutation ρ and taking for π a uniform random permutation in $\mathfrak{S}_n$ lead to:

Proposition 3.1. *(i) Let $\rho = (\rho_1, \rho_2, \dots, \rho_n)$ be a fixed permutation in $\mathfrak{S}_n$ and let π be a uniform random permutation in $\mathfrak{S}_n$. Then,*

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

where $=_d$ denotes equality in distribution.

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

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

Proof. To begin the proof of (i), let $\pi' \in \mathfrak{S}_n$ be such that $\pi'_i = \rho_i$. Then, $\pi'' := \pi\pi'$ is still a uniform random permutation of $[n]$, and so

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

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

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

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

Now, for (ii),

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

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 $\mathfrak{S}_n$ 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 on $LI_n(\pi)$ presented in Romik [34], the mean asymptotic result of Vershik and Kerov [38], and Logan and Shepp [30] thus implies that (is equivalent to):

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

Moreover, the distributional asymptotic result of Baik, Deift and Johansson [5] implies that (is equivalent to), 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 3.2. (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 is presented in a separate publication (see [17]), where multiple words are also tackled. Such a result requires, at first, to use variance estimates, generalizing [23], as stated in the concluding remarks of [22] and then to extend to higher dimensions the closeness to the diagonal results obtained in [24].

(ii) 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 also be presented elsewhere. Already for hidden Markov models (HMM), convergence results, with rates, are obtained for $\mathbb{E}LC_n/n$ in [19], while [20] shows how to transfer iid normal approximation results such as Theorem 2.1 to the HMM case.

(iii) It would, similarly, also be of interest to study the random permutations versions of (i) and (ii) above. As in the previous section, and as far as the multiple sequences framework is concerned, the study of the length of the longest common subsequences reduces to the study of the length of the longest common and increasing subsequences with one less sequence, e.g., see [25].

References

[1] D. Aldous, P. Diaconis. *Longest increasing subsequences: From patience sorting to the Baik-Deift-Johansson theorem*, Bull. Amer. Math. Soc. (N.S.) 36 (4):413–432, 1999. MR1694204

[2] K. S. Alexander. *The rate of convergence of the mean length of the longest common subsequence.* Ann. Appl. Probab., 4(4), 1074–1082, 1994. MR1304773

[3] B. Arras and C. Houdré. *On Stein's Method for Infinitely Divisible Laws With Finite First Moment.* SpringerBriefs in Probability and Mathematical Statistics. Springer, Cham, 2019. xi+104 pp. MR3931309

[4] S. Amsalu, C. Houdré and H. Matzinger. *Sparse long blocks and the variance of the longest common subsequences in random words.* arXiv:math/1204.1009v2, 2016. MR4488538

[5] J. Baik, P. Deift, and K. Johansson. *On the distribution of the length of the longest increasing subsequence of random permutations*, J. Amer. Math. Soc. 12 (4): 1119–1178, 1999. MR1682248

[6] F. Bonetto, H. Matzinger. *Fluctuations of the longest common subsequence in the asymmetric case of 2- and 3-letter alphabets*, ALEA Lat. Am. J. Probab. Math. Stat. 2, 195–216, 2006. MR2262762

[7] J. Boutet de Monvel. *Extensive simulations for longest common subsequences: Finite size scaling, a cavity solution and configuration space properties*, Eur. Phys. J. B7, 293–308, 1999.

[8] J.-C. Breton, C. Houdré. *On the limiting law of the length of the longest common and increasing subsequences in random words.* Stochastic Process. Appl. 127, no. 5, 1676–1720, 2017. MR3630241

[9] R.M. Capocelli, *Sequences: Combinatorics, Compression, Security, and Transmission*, Springer-Verlag New York, 1989. MR1040295

[10] S. Chatterjee. *A new method of normal approximation.* Ann. Probab. 36 no. 4, 1584–1610, 2008. MR2435859

[11] S. Chatterjee, S. Sen. *Minimal spanning trees and Stein's method.* Ann. Appl. Probab. 27, no. 3, 1588–1645, 2017. MR3678480

[12] L. H. Y. Chen, L. Goldstein, Q.-M. Shao, *Normal approximation by Stein's method.* Probability and its Applications. Springer, Heidelberg, 2011. MR2732624

[13] V. Chvátal, D. Sankoff. *Longest common subsequences of two random sequences.* J. Appl. Probab. 12, 306–315, 1975. MR0405531

[14] M. Damron, J. Hanson, C.Houdré, C. Xu. *Lower bounds for fluctuations in first-passage percolation for general distributions.* Ann. Inst. Henri Poincaré Probab. Stat. 56, no. 2, 1336–1357, 2020 MR4076786

[15] V. Dančík. *Common subsequences and supersequences and their expected length.* Combinatorics, Probability and Computing 7, 365–373, 1998. MR1680096

[16] C. Deslandes, C. Houdré. *On the limiting law of the length of the longest common and increasing subsequences in random words with arbitrary distribution.* Electron. J. Probab. 26, 1–27, 2021. MR4262342

[17] R. Gong, C. Houdré, Ü. Işlak. *A central limit theorem for the optimal alignments score in multiple random words.* arXiv:math/1512.05699, 2016.

[18] R. Gong, C. Houdré, J. Lember. *Lower bounds on the generalized central moments of the optimal alignments score of random sequences.* Journal of Theoretical Probability, 1–41, *DOI* 10.1007/s10959-016-0730-4, 2017. MR3803910

[19] C. Houdré, G. Kerchev. *On the rate of convergence for the length of the longest common subsequences in hidden Markov models.* J. Appl. Probab., 56, no. 2, 558–573, 2019. MR3986952

[20] C. Houdré, G. Kerchev. *Normal approximation for functions of hidden Markov models.* Adv. in Appl. Probab., 54, no. 2. 536–569, 2022. MR4434666

[21] C. Houdré, J. Lember, H. Matzinger. *On the longest common increasing binary subsequence.* C.R. Acad. Sci. Paris Ser. I 343, 589–594, 2006. MR2269870

[22] C. Houdré, J. Ma. *On the order of the central moments of the length of the longest common subsequences in random words.* High Dimensional Probability VII: The Cargèse Volume, Progress in Probability 71, Birkhäuser, 105–137, 2016. MR3565261

[23] C. Houdré, H. Matzinger. *On the variance of the optimal alignments score for binary random words and an asymmetric scoring function.* J. Stat. Phys. 164(3), 693–734, 2016. MR3519215

[24] C. Houdré, H. Matzinger. *Closeness to the diagonal for longest common subsequences in random words.* Electron. Commun. Probab. 21(2): 1–19, 2016. MR3492931

[25] C. Houdré, C. Xu. *A note on the expected length of the longest common subsequences of two i.i.d. random permutations.* Electron. J. Combin. 25(2), P. 2.50, 2018. MR3830132

[26] M. Kiwi, M. Loebl, J. Matoušek. *Expected length of the longest common subsequence for large alphabets.* Adv. Math. 197, no. 2, 480–498, 2005. MR2173842

[27] R. Lachièze-Rey, G. Peccati. *New Berry-Esseen bounds for functionals of binomial point processes.* Ann. Appl. Probab., 27, No. 4, 1992–2031, 2017. MR3693518

[28] J. Lember, H. Matzinger. *Standard deviation of the longest common subsequence.* Ann. Probab. 37, no. 3, 1192–1235, 2009. MR2537552

[29] Q. Liu, C. Houdré. *Simulations, Computations, and Statistics for Longest Common Subsequences.* arXiv:math/1705.06826, 2017.

[30] B. Logan, L. A. Shepp. *A variational problem for random Young tableaux.* Advances in mathematics, 26.2, 206–222, 1977. MR1417317

[31] 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. MR2177365

[32] P. A. Pevzner. *Computational molecular biology: An algorithmic approach* A Bradford Book. MIT Press, Cambridge, MA, 2000. MR1790966

[33] W. Rhee. *On rates of convergence for common subsequences and first passage time.* Ann. Appl. Probab. 5, no. 1, 44–48, 1995. MR1325040

[34] D. Romik. *The surprising mathematics of longest increasing subsequences*. Cambridge University Press, 2014. MR3468738

[35] N. F. Ross. *Fundamentals of Stein's method.* Probability Surveys, 8, 210–293 (electronic), 2011. MR2861132

[36] D. Sankoff, and J. Kruskal. *Time warps, string edits and macromolecules: The theory and practice of sequence comparison.* Center for the Study of Language and Information, 1999. MR0700350

[37] J. M. Steele. *An Efron-Stein inequality for nonsymmetric statistics.* Ann. Statist. 14, 753–758, 1986. MR0840528

[38] A.M. Vershik, S.V. Kerov. *Asymptotic behavior of the Plancherel measure of the symmetric group and the limiting form of Young tableaux (Russian).* Dokl. Akad. Nauk SSSR 233, no. 6, 1024–1027, 1977. MR0480398

[39] M.S. Waterman. *Estimating statistical significance of sequence alignments.* Phil. Trans. R. Soc. Lond. B, 344:383–390, 1994.

[40] M.S. Waterman. *Introduction to Computational Biology: Maps, Sequences and Genomes* (Interdisciplinary Statistics), CRC Press, 2000.

Acknowledgments. Many thanks to the LPMA of the Université Pierre et Marie Curie, to CIMAT and to Bogazici University for their hospitality while part of this research was carried out. ÜI is grateful to L. Goldstein for introducing him to Stein's method and, in particular, to Chatterjee's normal approximation results. Also, many thanks to the LPMA of the Université Pierre et Marie Curie for its hospitality while part of this research was carried out as well as to the School of Mathematics of the Georgia Institute of Technology while being a Hale postdoctoral Fellow.

Both authors would like to thank the French Scientific Attachés Fabien Agenès (Los Angeles) and Nicolas Florsch (Atlanta) for their consular help. Without them, this research might not have existed. Lastly many thanks to Ruoting Gong, George Kerchev, Chen Xu and referees for their detailed reading and many comments which have led to numerous improvements on this manuscript.