On the Order of the Central Moments of the Length of the Longest Common Subsequence

Christian Houdré ∗ Jinyong Ma †

August 6, 2014

Abstract

We investigate the order of the $r$-th, $1 \leq r < +\infty$, central moment of the length of the longest common subsequence of two independent random words of size $n$ whose letters are identically distributed and independently drawn from a finite alphabet. When all but one of the letters are drawn with small probabilities, which depend on the size of the alphabet, a lower bound is shown to be of order $n^{r/2}$. This result complements a generic upper bound also of order $n^{r/2}$.

AMS 2010 Subject Classification: 60K35, 60C05, 05A05

Keywords: Longest common subsequence, $r$-th central moment, Burkholder inequality, last passage percolation.

1 Introduction and statements of results

Let $X = (X_i)_{i \geq 1}$ and $Y = (Y_i)_{i \geq 1}$ be two independent sequences of iid random variables taking their values in a finite alphabet $A_m = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$, with $\mathbb{P}(X_1 = \alpha_k) = \mathbb{P}(Y_1 = \alpha_k) = p_k$, $k = 1, 2, \ldots, m$. Let $LC_n$ be the length of the longest common subsequence of the random words $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$, i.e.,
$LC_n := LC_n(X_1 \cdots X_n; Y_1 \cdots Y_n)$ is the largest $k$ such that there exist $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ and $1 \leq j_1 < j_2 < \cdots < j_k \leq n$, with $X_{i_s} = Y_{j_s}$, $s = 1, \ldots, k$.

The study of the asymptotic behavior of $LC_n$ has a long history starting with the well known result of Chvátal and Sankoff [4] asserting that

$$\lim_{n \to \infty} \frac{\mathbb{E}LC_n}{n} = \gamma^*_m.$$  \hspace{1cm} (1.1)

However, to this day, the exact value of $\gamma^*_m$ (which depends on the distribution of $X_1$ and on the size of the alphabet) is still unknown even in "simple cases" such as for uniform Bernoulli random variables. This first asymptotic result was sharpened by Alexander ([1]) who showed that

$$\sqrt{n} \log n \leq \mathbb{E}LC_n \leq \gamma^*_m n,$$  \hspace{1cm} (1.2)

where $C > 0$ is a constant depending neither on $n$ nor on the distribution of $X_1$. Next, Steele [10] was the first to investigate the order of the variance proving, in particular, that $\text{Var}LC_n \leq n$. However, finding the order of the lower bound is more illusive. For Bernoulli random variables and 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 $n$ ([5], [6], [7]). The uniform case is still unresolved and tight lower variance estimates seem to be lacking (however, see [2], where a situation "as close as we want" to uniformity is treated).

Below, starting with a generic upper bound, we investigate the order of the $r$-th, $r \geq 1$, central moment of $LC_n$ in case of finite alphabets (of course, as far as the order is concerned only the case $1 \leq r \leq 2$ is really of interest for this lower bound).

The upper bound obtained in [10] relies on an asymmetric version of the Efron-Stein inequality which can be viewed as a tensorization property of the variance. The symmetric Efron-Stein inequality has seen a generalization, due to Rhee and Talagrand [9], to the $r$-th moment, where it is, in turn, viewed as a consequence of Burkholder’s square function inequality. As shown next, in the asymmetric case, a similar extension also holds thus providing a generic upper bound on the $r$-th central moment of $LC_n$. First, let $S : \mathbb{R}^n \to \mathbb{R}$ be a Borel function and let $(Z_{i})_{1 \leq i \leq n}$ and $(\hat{Z}_{i})_{1 \leq i \leq n}$ be two independent families of iid random variables having the same law. Now, and with suboptimal notation, let $S = S(Z_1, Z_2, \ldots, Z_n)$, and let $S_i = S(Z_1, Z_2, \ldots, Z_{i-1}, \hat{Z}_i, Z_{i+1}, \ldots, Z_n)$, $1 \leq i \leq n$. Then, for any $r \geq 2$,

$$\|S - \mathbb{E}S\|_r := (\mathbb{E}|S - \mathbb{E}S|^r)^{1/r} \leq \left( \sum_{i=1}^n \|S - S_i\|_r^2 \right)^{1/2}.$$  \hspace{1cm} (1.3)
Indeed, for $i = 1, \ldots, n$, let $\mathcal{F}_i = \sigma(Z_1, \ldots, Z_i)$ be the $\sigma$-field generated by $Z_1, \ldots, Z_i$, let $\mathcal{F}_0 = \{\Omega, \emptyset\}$ be trivial, and let $d_i := \mathbb{E}(S|\mathcal{F}_i) - \mathbb{E}(S|\mathcal{F}_{i-1})$. Thus, $(d_i, \mathcal{F}_i)_{1 \leq i \leq n}$ is a martingale differences sequence, and from Burkholder’s square function inequality, with optimal constant, for $r \geq 2$,

$$
\|S - \mathbb{E}S\|_r = \left\| \sum_{i=1}^{n} d_i \right\|_r \leq (r - 1)^{1/2} \left( \sum_{i=1}^{n} \left\| d_i \right\|^{2r}_r \right)^{1/2} \leq (r - 1)^{1/2} \left( \sum_{i=1}^{n} \left\| d_i \right\|^{r}_r \right)^{1/2}.
$$

(1.4)

Moreover, and as in [9], letting $\mathcal{G}_i = \sigma(Z_1, Z_2, \ldots, Z_i, \hat{Z}_i)$, $1 \leq i \leq n$,

$$
\mathbb{E}\left| S - S_i \right|^r = \mathbb{E}(\mathbb{E}(\left| S - S_i \right|^r | \mathcal{G}_i)) \geq \mathbb{E}(\left( \mathbb{E}(S|\mathcal{G}_i) - \mathbb{E}(S|\mathcal{F}_{i-1}) + \mathbb{E}(S|\mathcal{F}_{i-1}) - \mathbb{E}(S|\mathcal{G}_i) \right)^r) := \mathbb{E}|U + V|^r,
$$

(1.5)

where $U = \mathbb{E}(S|\mathcal{G}_i) - \mathbb{E}(S|\mathcal{F}_{i-1})$ and $V = \mathbb{E}(S|\mathcal{F}_{i-1}) - \mathbb{E}(S|\mathcal{G}_i)$. But, given $\mathcal{F}_{i-1}$, $U$ and $V$ are independent, with moreover $\mathbb{E}(U|\mathcal{F}_{i-1}) = \mathbb{E}(V|\mathcal{F}_{i-1}) = 0$ and $\mathbb{E}|U|^r = \mathbb{E}|V|^r = \mathbb{E}|d_i|^r$. Thus,

$$
\mathbb{E}|U + V|^r = \mathbb{E}(\mathbb{E}(\left| U + V \right|^r | \mathcal{F}_{i-1})) \geq \mathbb{E}|U|^r + \mathbb{E}|V|^r = 2\mathbb{E}|d_i|^r.
$$

(1.6)

Combining (1.4), (1.5) and (1.6) gives (1.3).

Next, apply (1.3) to $LC_n$ viewed as a function of $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ and note at first that replacing $X_i$ (resp. $Y_i$) by $\hat{X}_i$ (resp. $\hat{Y}_i$) an independent copy of itself, will increases $|LC_n - LC_n(X_1 \cdots \hat{X}_i \cdots X_n; Y_1 \cdots Y_n)|$ (resp. $|LC_n - LC_n(X_1 \cdots X_n; Y_1 \cdots \hat{Y}_i \cdots Y_n)|$) by at most 1, thus, following Steele [10] and for each $i = 1, \ldots, n$,

$$
\left\| LC_n - LC_n(X_1 \cdots \hat{X}_i \cdots X_n; Y_1 \cdots Y_n) \right\|_r^2
= \left( \mathbb{P}(LC_n - LC_n(X_1 \cdots \hat{X}_i \cdots X_n; Y_1 \cdots Y_n) \neq 1_{X_i \neq \hat{X}_i}) \right)^{2/r}
\leq \left( 1 - \sum_{k=1}^{m} p_k^2 \right)^{2/r}.
$$

(1.7)

Combining (1.7) and its version for $(Y_i)_{1 \leq i \leq n}$, with (1.3) yields, for any $r \geq 2$,

$$
\mathbb{E}|LC_n - \mathbb{E}LC_n|^r \leq \frac{(r - 1)^r}{2} \left( 1 - \sum_{k=1}^{m} p_k^2 \right) (2n)^{r/2},
$$

(1.8)
which further yields,
\[
\mathbb{E}|LC_n - \mathbb{E}LC_n|^r \leq \left( \left( 1 - \sum_{k=1}^{m} p_k^2 \right) n \right)^{r/2},
\]
for any \( 0 < r \leq 2 \).

Now that an upper bound is obtained, let us state the main result of the paper which provides a lower bound on the \( r \)-th central moment of \( LC_n \), when all but one of the symbols are drawn with very small probabilities.

**Theorem 1.1** Let \( 1 \leq r < +\infty \), and let \((X_i)_{i \geq 1}\) and \((Y_i)_{i \geq 1}\) be two independent sequences of iid random variables with values in \( \mathcal{A}_m = \{\alpha_1, \alpha_2, \ldots, \alpha_m\} \), and with \( \mathbb{P}(X_1 = \alpha_k) = p_k, k = 1, 2, \ldots, m \). Let \( p_{j_0} > 1/2 \), for some \( j_0 \in \{1, \ldots, m\} \), 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, 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 \( n \geq 1 \),
\[
M_r(LC_n) := \mathbb{E}|LC_n - \mathbb{E}LC_n|^r \geq Cn^{r/2}.
\] (1.9)

An estimate on the constant \( C \) above is provided in Remark 2.1.

In contrast to [6], [5] or [7] which deal only with binary words, our results are proved for alphabets of arbitrary, but fixed size \( m \), and are thus novel in that context as well, even for the variance, i.e., \( r = 2 \). Moreover, our results are no longer asymptotic, but rather valid for all \( n \geq 1 \), and precise constants sharply depending on the alphabet size are provided. As pointed out in [2], the LCS problem is a last passage percolation (LPP) problem with dependent weights and in our context, the order of the variance is linear. For the LPP problem with independent weights the variance is conjectured to be sublinear. In view of (1.8) and (1.9), it is tempting to conjecture, and we do so, that when properly centered (by \( \gamma_m n \)) and normalized (by \( \sqrt{n} \)), asymptotically, \( LC_n \) has a normal component. This conjecture might appear surprising since in LPP with independent weights different limiting laws are conjectured and, in particular, proved to be such in the closely related Bernoulli matching model [8]. It should finally also be noted that, as seen in [3] with another closely related model, the order \( n^{r/2} \) on the central moments does not guarantee normal convergence.

As for the content of the rest of paper, Section 2 presents a proof of Theorem 1.1 which relies on a key preliminary result, Theorem 2.1 whose proof is given in Section 3.
2 Proof of Theorem 1.1

The strategy of proof to obtain the lower bound is to first represent $LC_n$ as a random function of the number of most probable letters $\alpha_{j_0}$. In turn, this random function satisfies locally a reversed Lipschitz condition, ultimately giving the lower bound in Theorem 1.1. This methodology extends, modifies and simplifies (and at times corrects) the binary strategy of proof of [6], [7] providing also a non-asymptotic result.

To start, and as in [6], pick a letter equiprobably at random from all the non-$\alpha_{j_0}$ letters in either one of the two finite sequences of length $n$, $X$ or $Y$ (Throughout the paper, by finite sequences $X$ and $Y$ of length $n$, it is meant that $X = (X_i)_{1 \leq i \leq n}$ and $Y = (Y_i)_{1 \leq i \leq n}$). Next, change it to the most probable letter $\alpha_{j_0}$ and call the two new finite sequences $\tilde{X}$ and $\tilde{Y}$. Then the length of the longest common subsequence of $\tilde{X}$ and $\tilde{Y}$, denoted by $\tilde{LC}_n$, tends, on an event of high probability, to be larger than $LC_n$. This is the content of the following theorem which is proved in the next section.

**Theorem 2.1** Let the hypothesis of Theorem 1.1 hold. Then, for all $n \geq 1$, there exists a set $B_n \subset A^n_m$, such that,

$$\mathbb{P}((X, Y) \in B_n) \geq 1 - 121 \exp \left( -\frac{n(\max_{j \neq j_0} p_j)^6}{5} \right),$$  \hspace{1cm} (2.1)

and such that for all $(x, y) \in B_n$,

$$\mathbb{P}(\tilde{LC}_n - LC_n = 1|X = x, Y = y) \geq \frac{K}{m},$$  \hspace{1cm} (2.2)

$$\mathbb{P}(\tilde{LC}_n - LC_n = -1|X = x, Y = y) \leq \frac{K}{2m},$$  \hspace{1cm} (2.3)

where $K = 2^{-4}10^{-2}e^{-67}$.

As already mentioned, the proof of Theorem 2.1 is given in the next section, let us nevertheless indicate now how it leads to the lower bound on $M_r(LC_n)$ given in Theorem 1.1.

From now on, assume without loss of generality that $p_1 > 1/2$ and that $p_2 = \max_{2 \leq j \leq m} p_j$.

To begin with, let us present a few definitions. For the two finite random sequences $X = (X_i)_{1 \leq i \leq n}$ and $Y = (Y_i)_{1 \leq i \leq n}$, let $N_1$ be the total number of letters $\alpha_1$ present in both sequences, i.e., $N_1$ is a binomial random variable with parameters $2n$ and $p_1$. Next, by induction, define a finite collection of pairs of finite random sequences $(X^k, Y^k)_{0 \leq k \leq 2n}$ as follows: First, let $X^0 = (X_i^0)_{1 \leq i \leq n}$
and \( Y^0 = (Y^0_i)_{1 \leq i \leq n} \) be independent, with \( X^0_i \) and \( Y^0_i \), \( i = 1, \ldots, n \), iid random variables with values in \( \{\alpha_2, \ldots, \alpha_m\} \) and such that \( \mathbb{P}(X^0_1 = \alpha_k) = \mathbb{P}(Y^0_1 = \alpha_k) = p_k/(1 - p_1) \), \( 2 \leq k \leq m \). In other words, \( X^0 \) and \( Y^0 \) are two independent finite sequences of iid random variables whose joint law is the law of \( (X, Y|N_1 = 0) \).

Once \( (X^k, Y^k) \) is defined, let \( (X^{k+1}, Y^{k+1}) \) be the pair of finite random sequences obtained by taking (pathwise) with equal probability, one letter from all the letters \( \alpha_2, \alpha_3, \ldots, \alpha_m \) in the pair \( (X^k, Y^k) \) and replacing it with \( \alpha_1 \), and for this path iterating the process till \( k = 2n \). Clearly, for \( 1 \leq k \leq 2n - 1 \), \( X^k \) and \( Y^k \) are not independent, while \( (X_{i}^{2n}, Y_{i}^{2n})_{1 \leq i \leq n} \) is a deterministic sequence made up only of \( \alpha_1 \).

Rigorously, the random variables can be defined as follows: let \( \Omega \) be our underlying probability space, and let \( \Omega^{2n+1} \) be its \((2n+1)\)-fold Cartesian product. For each \( \omega = (\omega_0, \omega_1, \ldots, \omega_{2n}) \in \Omega^{2n+1} \) and \( 0 \leq k \leq 2n \), \( (X^k(\omega), Y^k(\omega)) \) only depends on \( \omega_0, \omega_1, \ldots, \omega_k \). Then, \( (X^{k+1}(\omega), Y^{k+1}(\omega)) \) is obtained from \( (X^k(\omega), Y^k(\omega)) \) by replacing with equal probability any non-\( \alpha_1 \) letter by \( \alpha_1 \), while the choice of the non-\( \alpha_1 \) letter to be replaced in \( (X^k(\omega), Y^k(\omega)) \) is determined by \( \omega_{k+1} \).

Next, let \( LC_n(k) \) denote the length of the longest common subsequence of \( X^k \) and \( Y^k \) (with a slight abuse of notation and terminology with the identification of finite sequences and words). The lemma below shows that \( (X^k, Y^k) \) has the same law as \( (X, Y) \) conditional on \( N_1 = k \), and therefore the law of \( LC_n(k) \) is the same as the conditional law of \( LC_n \) given \( N_1 = k \).

\begin{lemma}
Let \( X = (X_i)_{1 \leq i \leq n} \) and \( Y = (Y_i)_{1 \leq i \leq n} \). Then, for \( k = 0, 1, \ldots, 2n \),
\[
(X^k, Y^k) \overset{d}{=} (X, Y|N_1 = k), \tag{2.4}
\]
and
\[
(X^{N_1}, Y^{N_1}) \overset{d}{=} (X, Y), \tag{2.5}
\]
where \( \overset{d}{=} \) denotes equality in distribution.
\end{lemma}

**Proof.** The proof is by induction on \( k \). By definition, \( (X^0, Y^0) \) has the same law as \( (X, Y) \) conditional on \( N_1 = 0 \).

For any \( (\alpha_{j_1}, \ldots, \alpha_{j_{2n}}) \in A_m^n \times A_m^n \), let
\[
q_\ell = |\{1 \leq i \leq 2n : \alpha_{j_i} = \alpha_\ell\}|, 
\]
1 \( \leq \ell \leq m \). Now assume that (2.4) is true for \( k \), i.e., assume that for any \( (\alpha_{j_1}, \ldots, \alpha_{j_{2n}}) \in A_m^n \times A_m^n \), with \( q_1 = k \),
\[
\mathbb{P} \left( (X^k_1, \ldots, X^k_n, Y^k_1, \ldots, Y^k_n) = (\alpha_{j_1}, \ldots, \alpha_{j_{2n}}) \right) = \left( \frac{2n}{k} \right)^{-1} \prod_{\ell=2}^m \left( \frac{p_\ell}{1 - p_1} \right)^{q_\ell}. \tag{2.6}
\]

6
Then, for any \((\alpha_{j_1}, \ldots, \alpha_{j_{2n}}) \in \mathcal{A}_m^n \times \mathcal{A}_m^n\), with \(q_1 = k + 1\),

\[
\mathbb{P}
\left(
\left(\begin{array}{c}
X_{1}^{k+1}, \ldots, X_{n}^{k+1}, Y_{1}^{k+1}, \ldots, Y_{n}^{k+1}
\end{array}\right) = \left(\begin{array}{c}
\alpha_{j_1}, \ldots, \alpha_{j_{2n}}
\end{array}\right)
\right) = \\
\sum_{i=1}^{k+1} \mathbb{P}
\left(
\left(\begin{array}{c}
X_{1}^{k+1}, \ldots, X_{n}^{k+1}, Y_{1}^{k+1}, \ldots, Y_{n}^{k+1}
\end{array}\right) = \left(\begin{array}{c}
\alpha_{j_1}, \ldots, \alpha_{j_{2n}}|B_i^{k+1}
\end{array}\right) \mathbb{P}(B_i^{k+1})
\right),
\tag{2.7}
\]

where \(B_i^{k+1}, 1 \leq i \leq k+1\), is the event that the \(i\)-th \(\alpha_1\) in \((\alpha_{j_1}, \ldots, \alpha_{j_{2n}})\) is changed from a non-\(\alpha_1\) letter when passing from \((X^k, Y^k)\) to \((X^{k+1}, Y^{k+1})\). Conditional on \(B_i^{k+1}\), the \(i\)-th \(\alpha_1\) in \((\alpha_{j_1}, \ldots, \alpha_{j_{2n}})\) could have been changed from any letter in \(\{\alpha_2, \alpha_3, \ldots, \alpha_m\}\). Assuming this \(\alpha_1\) has been changed from \(\alpha_s\), \(2 \leq s \leq m\), the corresponding probability is given by:

\[
\mathbb{P}
\left(
\left(\begin{array}{c}
X_{1}^{k+1}, \ldots, X_{n}^{k+1}, Y_{1}^{k+1}, \ldots, Y_{n}^{k+1}
\end{array}\right) = \left(\begin{array}{c}
\alpha_{j_1}, \ldots, \alpha_{j_{2n}}|B_i^{k+1}
\end{array}\right) \mathbb{P}(B_i^{k+1})
\right) = \\
\left(\begin{array}{c}
2n^k
\end{array}\right) - 1 \prod_{\ell=2}^{m} \left(\frac{p_\ell}{1-p_1}\right)^{q_\ell} \left(\frac{p_s}{1-p_1}\right),
\tag{2.8}
\]

which when incorporated into (2.7), gives

\[
\mathbb{P}
\left(
\left(\begin{array}{c}
X_{1}^{k+1}, \ldots, X_{n}^{k+1}, Y_{1}^{k+1}, \ldots, Y_{n}^{k+1}
\end{array}\right) = \left(\begin{array}{c}
\alpha_{j_1}, \ldots, \alpha_{j_{2n}}
\end{array}\right)
\right) = \\
\left(\begin{array}{c}
2n^k + 1
\end{array}\right) - 1 \prod_{\ell=2}^{m} \left(\frac{p_\ell}{1-p_1}\right)^{q_\ell},
\tag{2.8}
\]

finishing the proof of the first part of the lemma.

Next, from (2.4) and the independence of \(N_1\) and \(\{(X^k, Y^k)\}_{0 \leq k \leq 2n}\), for any
\((u, v) \in \mathbb{R}^n \times \mathbb{R}^n\),

\[
\mathbb{E}\left(e^{i < u, X > + i < v, Y >}\right) = \sum_{k=0}^{2n} \mathbb{E}\left(e^{i < u, X^k > + i < v, Y^k >}\big| N_1 = k\right) \mathbb{P}(N_1 = k)
= \sum_{k=0}^{2n} \mathbb{E}\left(e^{i < u, X^k > + i < v, Y^k >}\big| N_1 = k\right) \mathbb{P}(N_1 = k)
= \sum_{k=0}^{2n} \mathbb{E}\left(e^{i < u, X^{N_1} > + i < v, Y^{N_1} >}\big| N_1 = k\right) \mathbb{P}(N_1 = k)
= \mathbb{E}\left(e^{i < u, X^{N_1} > + i < v, Y^{N_1} >}\right),
\]
finishing the proof of the lemma.

Let now \(LC_n(N_1)\) be the length of the longest common subsequence of \(X^{N_1}\) and \(Y^{N_1}\). The above lemma implies that \(LC_n\) and \(LC_n(N_1)\) have the same law and, therefore,

\[
\mathbb{M}_r(LC_n(N_1)) = \mathbb{M}_r(LC_n). \tag{2.9}
\]

To lower bound the right hand side of (2.9) (and to prove Theorem 1.1) the following simple inequality will prove useful.

**Lemma 2.2** Let \(f : D \to \mathbb{Z}\) satisfy locally a reversed Lipschitz condition, i.e., let \(\ell \geq 0\) and let \(f\) be such that for any \(i, j \in D\) with \(j \geq i + \ell\),

\[
f(j) - f(i) \geq c(j - i),
\]
for some \(c > 0\). Let \(T\) be a \(D\)-valued random variable with \(\mathbb{E}|f(T)|^r < +\infty\), \(r \geq 1\), then

\[
\mathbb{M}_r(f(T)) \geq \left(\frac{c}{2}\right)^r \mathbb{M}_r(T) - \ell^r. \tag{2.10}
\]

**Proof.** Let \(r \geq 1\), and let \(\hat{T}\) be an independent copy of \(T\). First, and clearly,

\[
\mathbb{M}_r(T) \leq \mathbb{E}(|T - \hat{T}|^r) \leq 2^r \mathbb{M}_r(T).
\]
Hence,

\[
\mathbb{M}_r(f(T)) \geq \frac{1}{2^r} \mathbb{E}(|f(T) - f(\hat{T})|^r)
\geq \left(\frac{c}{2}\right)^r \left(\mathbb{E}(T - \hat{T})^r \mathbf{1}_{T - \hat{T} \geq \ell} + \mathbb{E}(\hat{T} - T)^r \mathbf{1}_{\hat{T} - T \geq \ell}\right)
\geq \left(\frac{c}{2}\right)^r \left(\mathbb{E}|T - \hat{T}|^r - \ell^r\right)
\geq \left(\frac{c}{2}\right)^r \left(\mathbb{M}_r(T) - \ell^r\right).
\]

\(\blacksquare\)
The above lemma will be useful to provide a lower bound on \( M_r(LC_n(N_1)) \) by showing that, after removing the randomness of \( LC_n(\cdot) \), \( LC_n(\cdot) \) satisfies locally a reversed Lipschitz condition. To do so, for a random variable with finite \( r \)-th moment \( U \) and for a random vector \( V \), let \( M_r(U|V) := \mathbb{E}(|U - \mathbb{E}(U|V)|^r|V) \). Clearly,

\[
2^r \mathbb{E} (|U - \mathbb{E}U|^r|V) \geq M_r(U|V),
\]

and so, for any \( n \geq 1 \),

\[
M_r(LC_n(N_1)) \geq \frac{1}{2^r} \mathbb{E}(M_r(LC_n(N_1)|(LC_n(k))_{0 \leq k \leq 2n}))
\]

\[
= \frac{1}{2^r} \int_{\Omega} M_r(LC_n(N_1)|(LC_n(k))_{0 \leq k \leq 2n}(\omega)) \mathbb{P}(d\omega)
\]

\[
\geq \frac{1}{2^r} \int_{O_n} M_r(LC_n(N_1)|(LC_n(k))_{0 \leq k \leq 2n}(\omega)) \mathbb{P}(d\omega),
\]

where for each \( n \geq 1 \),

\[
O_n := \bigcap_{i,j \in I, j \geq i + \ell(n)} \left\{ LC_n(j) - LC_n(i) \geq \frac{K}{4m} (j - i) \right\}, \quad (2.13)
\]

where \( K \) is given in Theorem 2.1 and where \( \ell(n) \geq 0 \) is to be chosen later. (Of course, above and everywhere, intersections, unions and sums are taken over countable sets of integers.) In words, on the event \( O_n \), the random function \( LC_n \) has a slope of at least \( K/4m \), when restricted to the interval \( I \) and when \( i \) and \( j \) are at least \( \ell(n) \) away from each other.

Since \( N_1 \) is independent of \( (LC_n(k))_{0 \leq k \leq 2n} \), and from (2.11), for each \( \omega \in \Omega \),

\[
M_r(LC_n(N_1)|(LC_n(k))_{0 \leq k \leq 2n}(\omega))
\]

\[
\geq \frac{1}{2^r} M_r(LC_n(N_1)|(LC_n(k))_{0 \leq k \leq 2n}(\omega), 1_{N_1 \in I} = 1) \mathbb{P}(N_1 \in I|(LC_n(k))_{0 \leq k \leq 2n}(\omega))
\]

\[
= \frac{1}{2^r} M_r(LC_n(N_1)|(LC_n(k))_{0 \leq k \leq 2n}(\omega), 1_{N_1 \in I} = 1) \mathbb{P}(N_1 \in I),
\]

where

\[
I = \left[ 2np_1 - \sqrt{2n(1 - p_1)p_1}, 2np_1 + \sqrt{2n(1 - p_1)p_1} \right]. \quad (2.15)
\]

For each \( \omega \in O_n \), from Lemma 2.2 and since \( N_1 \) is independent of \( (LC_n(k))_{0 \leq k \leq 2n} \),

\[
M_r(LC_n(N_1)|(LC_n(k))_{0 \leq k \leq 2n}(\omega), 1_{N_1 \in I} = 1)
\]

\[
\geq \left( \frac{K}{8m} \right)^r (M_r(N_1|1_{N_1 \in I} = 1) - \ell(n)^r). \quad (2.16)
\]
Now, (2.12), (2.14) and (2.16) lead to
\[
\mathbb{M}_r(LC_n(N_1)) \geq \frac{1}{4^r} \left( \frac{K}{8m} \right)^r (\mathbb{M}_r(N_1|1_{N_1 \in I} = 1) - \ell(n)^r) \mathbb{P}(N_1 \in I) \mathbb{P}(O_n),
\] (2.17)
and it remains to estimate each of the three terms on the right hand side of (2.17).
By the Berry-Esseen inequality, and all \( n \geq 1 \),
\[
\left| \mathbb{P}(N_1 \in I) - \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-x^2/2} dx \right| \leq \frac{1}{\sqrt{2np_1(1-p_1)}}.
\] (2.18)
Moreover,
\[
\mathbb{M}_r(N_1|1_{N_1 \in I} = 1) = \mathbb{E}(\{N_1 - 2np_1 + 2np_1 - \mathbb{E}(N_1|1_{N_1 \in I} = 1)\}^r|1_{N_1 \in I} = 1)
\geq \left| \mathbb{E}(\{N_1 - 2np_1\}^r|1_{N_1 \in I} = 1) - 2np_1 - \mathbb{E}(N_1|1_{N_1 \in I} = 1) \right|^r, \quad (2.19)
\]
and
\[
\mathbb{E}(\{N_1 - 2np_1\}^r|1_{N_1 \in I} = 1) - 2np_1
= \sqrt{2np_1(1-p_1)} \left| \mathbb{E} \left( \left. \frac{N_1 - 2np_1}{\sqrt{2np_1(1-p_1)}} \right| 1_{N_1 \in I} = 1 \right) \right|
= \sqrt{2np_1(1-p_1)} \left| \mathbb{E}(F_n(1) - \Phi(1) + F_n(-1) - \Phi(-1) - \int_{-1}^{1} (F_n(x) - \Phi(x)) dx \right| \mathbb{P}(N_1 \in I)
\leq \sqrt{2np_1(1-p_1)} 4 \max_{x \in [-1,1]} |F_n(x) - \Phi(x)| \mathbb{P}(N_1 \in I)
\leq \frac{2}{\int_{-1}^{1} e^{-x^2/2} dx / \sqrt{2\pi} - 1/\sqrt{2np_1(1-p_1)}}, \quad (2.20)
\]
where \( F_n \) is the distribution functions of \((N_1 - 2np_1)/\sqrt{2np_1(1-p_1)}\), while \( \Phi \) is the standard normal one. Likewise,
\[
\mathbb{E}(\{N_1 - 2np_1\}^r|1_{N_1 \in I} = 1)
\geq (2np_1(1-p_1))^{r/2} \int_{-1}^{1} |x|^r \Phi(x) - 4 \max_{x \in [-1,1]} |F_n(x) - \Phi(x)| \mathbb{P}(N_1 \in I)
\geq (2np_1(1-p_1))^{r/2} \int_{-1}^{1} e^{-x^2/2} dx - 2\sqrt{\pi}/\sqrt{np_1(1-p_1)} \mathbb{P}(N_1 \in I)
\geq (2np_1(1-p_1))^{r/2} \int_{-1}^{1} e^{-x^2/2} dx + \sqrt{\pi}/\sqrt{np_1(1-p_1)}. \quad (2.21)
\]
Next, (2.19)-(2.21) lead to:
\[ M_r(N_1|1_{N_1 \in I} = 1) \geq \left(2np_1(1 - p_1)\right)^{\frac{1}{2}} \left(\frac{\int_{-1}^{1} |x| e^{-\frac{x^2}{2}} dx - 2\sqrt{\pi} / \sqrt{np_1(1 - p_1)}}{\int_{-1}^{1} e^{-\frac{x^2}{2}} dx + \sqrt{\pi} / \sqrt{np_1(1 - p_1)}}\right)^r - \frac{2}{\int_{-1}^{1} e^{-\frac{x^2}{2}} dx / \sqrt{2\pi} - 1 / \sqrt{2np_1(1 - p_1)}}. \tag{2.22} \]

Finally, assuming Theorem 2.1 the estimates (2.17)-(2.22) combined with the estimate on \( P(O_n) \) obtained in the next lemma give the lower bound (1.9) whenever \( 33m^2 \log n/K^2 \leq \ell(n) \leq K_1 \sqrt{n} \) (see Remark 2.1 for an estimate on \( K_1 \)).

**Lemma 2.3** Let \( K = \min(2^{-4}10^{-2}e^{-67}, 1/800m) \), and assume, moreover, that \( p_2 \leq \min(2^{-2}e^{-5}K/m, K/2m^2) \). Then, for all \( n \geq 1 \),
\[ P(O_n) \geq 1 - \left(484\sqrt{\pi}e^2n \exp\left(-\frac{np_2^6}{5}\right) + 2n \exp\left(-\frac{K^2\ell(n)}{32m^2}\right)\right). \tag{2.23} \]

**Proof.** Let \( A_n := \{(X, Y) \in B_n\} \) and let \( A_n^k := \{(X^k, Y^k) \in B_n\}. \) Then,
\[ P\left(\bigcap_{k \in I} A_n^k \right)^c \leq \sum_{k \in I} P\left(\left( A_n^k \right)^c \right) = \sum_{k \in I} P\left( A_n^k | N_1 = k \right) \leq \sum_{k \in I} \frac{P(A_n^c)}{P(N_1 = k)}, \tag{2.24} \]
by Lemma 2.1. Next, by Stirling’s formula, for all \( k \in I \) and \( n \geq 1 \),
\[ P(N_1 = k) = \binom{2n}{k} p_1^k (1 - p_1)^{2n-k} \geq \frac{1}{\sqrt{2\pi e^2}} \frac{(2n)^{2n+1/2}}{k^{k+1/2}(2n-k)^{2n-k+1/2}} p_1^k (1 - p_1)^{2n-k} = \gamma(k, n, p_1). \]
Hence, for all \( k \in I \) and \( p_1 \geq 3/4 \) (which holds true since \( p_2 \leq K/m \)),
\[ \gamma(k, n, p_1) \geq \min \left( \gamma \left(2np_1 - \sqrt{2n(1 - p_1)p_1}, n, p_1 \right), \gamma \left(2np_1 + \sqrt{2n(1 - p_1)p_1}, n, p_1 \right) \right) \geq \frac{1}{2\sqrt{2\pi e^2 \sqrt{n}}}. \tag{2.25} \]
This last inequality in conjunction with (2.24) and Theorem 2.1 gives
\[ P\left(\bigcap_{k \in I} A_n^k \right)^c \leq 4\sqrt{\pi}e^2n P(A_n^c) \leq 484\sqrt{\pi}e^2n \exp\left(-\frac{np_2^6}{5}\right). \tag{2.26} \]
Next, for each $n \geq 1$, letting
\[
\Delta_{k+1} = \begin{cases} 
LC_n(k+1) - LC_n(k), & \text{when } A_n^k \text{ holds,} \\
1, & \text{otherwise},
\end{cases} \quad (2.27)
\]
it follows from Theorem 2.1 that,
\[
\mathbb{E}(\Delta_{k+1} | X^k, Y^k) \geq \frac{K}{2m}. \quad (2.28)
\]
Now, for each $k = 0, 1, \ldots, 2n$, let $\mathcal{F}_k := \sigma(X^0, Y^0, \ldots, X^k, Y^k)$, be the $\sigma$-field generated by $X^0, Y^0, \ldots, X^k, Y^k$. Clearly, $(\Delta_k - \mathbb{E}(\Delta_k | F_{k-1}), \mathcal{F}_k)_{1 \leq k \leq 2n}$ forms a martingale differences sequence and since $-1 \leq \Delta_k \leq 1$, it follows from Hoeffding’s martingale inequality that, for any $i < j$,
\[
\mathbb{P}\left( \sum_{k=i+1}^{j} (\Delta_k - \mathbb{E}(\Delta_k | F_{k-1})) < -\frac{K}{4m} (j - i) \right) \leq \exp\left( -\frac{K^2(j - i)}{32m^2} \right). \quad (2.29)
\]
Moreover, from (2.28),
\[
\sum_{k=i+1}^{j} \mathbb{E}(\Delta_k | X^{k-1}, Y^{k-1}) \geq K(j - i)/2m,
\]
and therefore
\[
\mathbb{P}\left( \sum_{k=i+1}^{j} \Delta_k \leq \frac{K}{4m} (j - i) \right) \leq \mathbb{P}\left( \sum_{k=i+1}^{j} (\Delta_k - \mathbb{E}(\Delta_k | F_{k-1})) < -\frac{K}{4m} (j - i) \right) \leq \exp\left( -\frac{K^2(j - i)}{32m^2} \right). \quad (2.30)
\]
For each $n \geq 1$, let now
\[
O_n^\Delta = \bigcap_{i,j \in I} \left\{ \sum_{i+1}^{j} \Delta_k \geq \frac{K}{4m} (j - i) \right\},
\]
then, from (2.30)
\[
\mathbb{P}\left( (O_n^\Delta)^c \right) \leq \sum_{i,j \in I} \mathbb{P}\left( \sum_{i+1}^{j} \Delta_k < \frac{K}{4m} (j - i) \right) \leq 2n \exp\left( -\frac{K^2\ell(n)}{32m^2} \right). \quad (2.31)
\]
From the very definition of $\Delta_k$ in (2.27), $\bigcap_{k \in I} A_n^k \cap O_n^\Delta \subset O_n$, and therefore
\[
\mathbb{P}\left( (O_n)^c \right) \leq \mathbb{P}\left( \left( \bigcap_{k \in I} A_n^k \right)^c \right) + \mathbb{P}\left( (O_n^\Delta)^c \right) \leq 484 \sqrt{\pi c^2 n} \exp\left( -\frac{np_0^6}{5} \right) + 2n \exp\left( -\frac{K^2\ell(n)}{32m^2} \right). \quad (2.32)
\]
Remark 2.1 The reader might wonder how to estimate the constant $C$ in Theorem 1.1. From (2.17), letting $n \geq p_2^{-12} + m^8$ and choosing

$$\ell(n) = e^{-\frac{1}{2} \left( np_1(1 - p_1) \right)^{\frac{1}{2}}} \left( \frac{1}{1 + r} \right)^{\frac{1}{2}} \leq \frac{1}{2\sqrt{e}} \sqrt{n},$$

it follows from (2.18) and (2.23) that:

$$\frac{M_r(LC_n)}{n^{r/2}} \geq 2^{-2 - 5r} (2^{r/2} - 1)(1 + r)^{-1} e^{-1/2} K^r m^{-r}(1 - p_1)^{r/2}.$$ 

Letting $C_1 = 2^{-2 - 5r} (2^{r/2} - 1)(1 + r)^{-1} e^{-1/2} K^r m^{-r}(1 - p_1)^{r/2}$, and

$$C_2 = \min_{n \leq p_2^{-12} + m^8} \frac{M_r(LC_n)}{n^{r/2}} \leq \frac{(r - 1)^r}{2} 2^{r/2} \left( 1 - \sum_{k=1}^{m} p_k^2 \right),$$

by (1.8), then one can choose $C = \min(C_1, C_2)$ in Theorem 1.1.

3 Proof of Theorem 2.1

3.1 Description of alignments

Let us begin with an example. Let $A_3 = \{1, 2, 3\}$ and, say, let

$$X = 1213131112, \quad Y = 1113121112.$$ 

An optimal alignment corresponding to the LCS 11311112 is

$$\begin{array}{cccccccccc}
1 & 2 & 1 & 3 & 1 & 3 & 1 & 1 & 1 & 2 \\
1 & 1 & 1 & 3 & 1 & 2 & 1 & 1 & 1 & 2 \\
\end{array} \quad (3.2)$$

and another possible optimal alignment is

$$\begin{array}{cccccccccc}
1 & 2 & 1 & 3 & 1 & 3 & 1 & 1 & 1 & 2 \\
1 & 1 & 1 & 3 & 1 & 2 & 1 & 1 & 1 & 2 \\
\end{array} \quad (3.3)$$

Comparing these two alignments, it is clearly seen that the way the letters $\alpha_1$, between aligned non-$\alpha_1$ letters, are aligned is not important as long as a maximal number of such letters $\alpha_1$ are aligned. Therefore, in general, we need only describe which non-$\alpha_1$ letters are aligned and assume that between pairs of aligned non-$\alpha_1$ letters a maximal number of letters $\alpha_1$ are aligned. In other words, we can identify the two alignments (3.2) and (3.3) as the same.
Next, call *cells* the parts of the alignment between pairs of aligned non-$\alpha_1$ letters. For example, the alignment \((3.2)\) can be decomposed into two cells \(C(1)\) and \(C(2)\) as

\[
\begin{array}{c|c|c|c|c|c}
1 & 2 & 1 & 3 & 1 & 3 \\
1 & 1 & 1 & 3 & 1 & 1 & 1 & 2 \\
\hline
C(1), \ v_1=-1 & C(2), \ v_2=0
\end{array}
\]

(3.4)

where, moreover, \(v_i\) denotes the difference between the number of letters \(\alpha_1\) in the \(X\)-strand and the \(Y\)-strand of the cell \(C(i)\). Note that any alignment can be represented as a finite vector of such differences. For the alignment \((3.2)\), this gives the representation \((v_1, v_2) = (-1, 0)\). Another optimal representation is via \((v_1, v_2) = (0, -1)\) corresponding to:

\[
\begin{array}{c|c|c|c|c|c}
1 & 2 & 1 & 3 & 1 & 3 \\
1 & 1 & 1 & 3 & 1 & 1 & 1 & 2 \\
\hline
C(1), \ v_1=0 & C(2), \ v_2=-1
\end{array}
\]

(3.5)

Let \(X = X_1X_2 \cdots X_n\) and \(Y = Y_1Y_2 \cdots Y_n\) be given. As just explained, to every optimal alignment corresponds a vector representation \(v := (v_1, \ldots, v_k)\) indicating the number of cells \((k, \text{here})\) in the alignment and the differences between the numbers of letters \(\alpha_1\) in the \(X\)-strand and the \(Y\)-strand of each cell. In every cell, the maximum amount of letters \(\alpha_1\) is aligned. On the other hand, to every \(v = (v_1, \ldots, v_k) \in \mathbb{Z}^k\) corresponds a (possible empty) family of alignments. All of these alignments have the same pairs of aligned non-$\alpha_1$ letters and between consecutive pairs of aligned non-$\alpha_1$ letters, a maximal number of letters \(\alpha_1\) are aligned. Since the alignments corresponding to the same \(v\) can only differ in the way the letters \(\alpha_1\) are aligned inside the cells, we again identify all the alignments in the family associated with \(v\) as a single alignment. In other words, we identify each vector \(v\) with an alignment and vice-versa.

Writing \(|v|\) for the number of coordinates of \(v\), i.e., \(|v| = k\), if \(v \in \mathbb{Z}^k\), the alignment associated with \(v = (v_1, \ldots, v_k) \in \mathbb{Z}^k\) can now precisely be defined:

**Definition 3.1** Let \(k \in \mathbb{N}\) and let \(v = (v_1, \ldots, v_k) \in \mathbb{Z}^k\). Let \(\pi_v(0) = \nu_v(0) = 0\), and for \(0 \leq i \leq k-1\), let \((\pi_v(i), \nu_v(i + 1))\) be the smallest pair of integers \((s, t)\) (where \((s_1, t_1) \leq (s_2, t_2)\) indicates that \(s_1 \leq s_2\) and \(t_1 \leq t_2\)) such that the following three conditions are satisfied.

1. \(\pi_v(i) < s\) and \(\nu_v(i) < t\);
2. \(X_s = Y_t \in \{\alpha_2, \ldots, \alpha_m\}\);
3. the difference between the number of letters \(\alpha_1\) in the (integer) intervals \([\pi_v(i), s]\) and \([\nu_v(i), t]\) is equal to \(v_{i+1}\).
If no such \((s, t)\) exists, then set \(\pi_v(i + 1) = \cdots = \pi_v(k) = \infty\) and \(\nu_v(i + 1) = \cdots = \nu_v(k) = \infty\).

In other words, above, \(\pi_v(i), \nu_v(i)\) are the indices corresponding to the \(i\)-th aligned non-\(\alpha_1\) pair in \(v\). The \(i\)-th cell \(C_v(i)\) is the pair

\[
C_v(i) := (X_{\pi_v(i-1)+1} \cdots X_{\pi_v(i)}; Y_{\nu_v(i-1)+1} \cdots Y_{\nu_v(i)}),
\]

and the cell \(C_v(i)\) is called a \(v_i\)-cell.

With the above definition, we can then let the alignment \(v\) be any alignment (provided one exists) satisfying the following three conditions:

1. \(X_{\pi_v(i)}\) is aligned with \(Y_{\nu_v(i)}\), for every \(i = 1, 2, \ldots, k\);

2. the number of aligned \(\alpha_1\) in the cell \(C_v(i)\), denoted by \(S_v(i)\), is the minimum number of letters \(\alpha_1\) present in either \(X_{\pi_v(i-1)+1} \cdots X_{\pi_v(i)}\) or \(Y_{\nu_v(i-1)+1} \cdots Y_{\nu_v(i)}\);

3. after aligning \(X_{\pi_v(k)}\) with \(Y_{\nu_v(k)}\), align as many letters \(\alpha_1\) as possible, and let that number be \(r_v\).

From these definitions, for any \(v \in \mathbb{Z}^k\), and if an alignment corresponding to \(v\) exists, then \(\pi_v(k) \leq n\) and \(\nu_v(k) \leq n\). Such a \(v\) is said to be admissible, and let \(V\) denote the set of all admissible alignments, that is,

\[
V := \left\{ v \in \bigcup_{k \geq 1} \mathbb{Z}^k : \pi_v(|v|) \leq n, \nu_v(|v|) \leq n \right\}. \tag{3.6}
\]

Then, for every \(v \in V\), the length of the common subsequence corresponding to this alignment is:

\[
\Lambda C_v = |v| + \sum_{i=1}^{|v|} S_v(i) + r_v. \tag{3.7}
\]

Therefore the length of the longest common subsequence of \(X\) and \(Y\) can be expressed as:

\[
LC_n = \max_{v \in V} \Lambda C_v, \tag{3.8}
\]

and, moreover, an admissible alignment is optimal if and only if \(\Lambda C_v = LC_n\).

### 3.2 The effect of changing a non-\(\alpha_1\) letter to \(\alpha_1\)

Again, the main idea behind Theorem 2.1 is that, by changing a randomly picked non-\(\alpha_1\) letter into \(\alpha_1\), the length of the longest common subsequence is more likely to increase by one than to decrease by one. More precisely, conditional on the
event $A_n = \{(X,Y) \in B_n\}$, the probability of an increase of $LC_n$ is at least $K/m$ while the probability of a decrease is at most $K/2m$. Let us illustrate this fact with an example. Let $X$ and $Y$ be given by,

$$X = 112113112131, \quad Y = 13111111131,$$

with optimal alignment:

$$
\begin{array}{cccccccc}
1 & 1 & 2 & 1 & 1 & 3 & 1 & 1 \\
1 & 3 & 1 & 1 & 1 & 1 & 1 & 1 \\
\end{array}
$$

(3.10)

Above, there are 6 non-$\alpha_1$ letters, $X_3, X_9, X_{11}, Y_2, Y_{11}$, and each one has probability $1/6$ to be picked and replaced by $\alpha_1$. Next, $X_3, X_6, X_9$ are on the top strand which contains a lesser number of letters $\alpha_1$, picking one of them and replacing it leads to an increase of one in the length of the LCS. On the other hand, since $X_{11}$ and $Y_{11}$ are aligned in this optimal alignment, picking one of them could potentially (but not necessarily) decrease the length of the LCS by one. Finally, picking $Y_2$ may only potentially increase the length of the LCS by modifying the alignment. In conclusion, in this example, by switching a randomly chosen non-$\alpha_1$ letter into $\alpha_1$, the probability of an increase of the length of the LCS is at least $1/2$, while the probability of a decrease is at most $1/3$.

To prove Theorem 2.1, we just need to prove that typically there exists an optimal alignment $v$ such that:

1. Among all the non-$\alpha_1$ letters in $X$ and $Y$, the proportion which are on the cell-strand with a smaller number of letters $\alpha_1$ is at least $K/m$.

2. Among all the non-$\alpha_1$ letters in $X$ and $Y$, the proportion which is aligned is at most $K/2m$.

Formally, let $v = (v_1, \ldots, v_k) \in \mathbb{Z}^k$ be admissible. For each $1 \leq i \leq k$, if $v_i \neq 0$, let $N_v^-(i)$ be the number of non-$\alpha_1$ letters on the cell-strand of $C_v(i)$ with a lesser number of letters $\alpha_1$, i.e.,

$$N_v^-(i) = \left\{ \begin{array}{ll}
\sum_{j=\pi_v(i-1)+1}^{\pi_v(i)-1} \mathbf{1}_{x_j \in \{\alpha_2, \ldots, \alpha_m\}}, & \text{if } v_i < 0, \\
\sum_{j=\pi_v(i-1)+1}^{\pi_v(i)-1} \mathbf{1}_{y_j \in \{\alpha_2, \ldots, \alpha_m\}}, & \text{if } v_i > 0,
\end{array} \right. \quad (3.11)$$

while if $v_i = 0$, let $N_v^-(i) = 0$. Then, the total number of non-$\alpha_1$ letters on the cell-strand with the smaller number of letters $\alpha_1$ is

$$N_v^- := \sum_{i=1}^{\|v\|} N_v^-(i). \quad (3.12)$$
Let $N_i$ be the number of letters $\alpha_i$ in the two finite sequences $X$ and $Y$, and let

$$N_{>1} = \sum_{i=2}^{m} N_i. \quad (3.13)$$

Next, let

$$B_n := \{ (x, y) \in A^n \times A^n : \text{there exists an optimal alignment } v \text{ of } (x, y)$$

$$\text{with } n_v^- \geq Kn_{>1}/m \text{ and } 2|v| \leq Kn_{>1}/2m \},$$

where, above, $n_v^-$ is the value of $N_v^-$ corresponding to $v$ and similarly for $n_{>1}$. Clearly, $B_n$ depends on $K$ and $m$. Letting $A_n = \{ (X, Y) \in B_n \}$, our goal is now to prove that for some $K_3 > 0$, independent of $n$, $\mathbb{P}(A_n) \geq 1 - e^{-K_3 n}$.

To continue, we need an optimal alignment having enough non-$\alpha_1$ letters in the cell-strands with a smaller number of letters $\alpha_1$. However, for many optimal alignments, most cells are 0-cells, i.e., cells with the same number of letters $\alpha_1$ on both strands. To resolve this hurdle, on an optimal alignment where most cells are 0-cells, some of the 0-cells are broken up in order to create enough nonzero-cells while at the same time, maintaining the optimality of the alignment after this breaking operation. Let us present this breaking operation on an example. Take the two sequences

$$X = 1121131123, \quad Y = 112131113,$$

one of their optimal alignment is

$$\begin{array}{c|c|c}
C(1), v_1=0 & C(2), v_2=0 \\
\hline
1 & 1 & 2 \quad & 1 & 1 & 3 & 1 & 1 & 2 & 3 \\
1 & 1 & 2 \quad & 1 & 3 & 1 & 1 & 1 & 3 & 1 & 1 & 3
\end{array} \quad (3.14)$$

where both cells $C(1)$ and $C(2)$ are 0-cells. Now in cell $C(2)$, $X_6$ and $Y_5$ are only one position away from being aligned. Thus aligning them, instead of the pair $X_5$ and $Y_6$, breaks the cell $C(2)$ into two new cells $\hat{C}(2)$ and $\hat{C}(3)$, with $\hat{v}_2 = 1$ and $\hat{v}_3 = -1$. The new optimal alignment is then:

$$\begin{array}{c|c|c|c}
\hat{C}(1), \hat{v}_1=0 & \hat{C}(2), \hat{v}_2=1 & \hat{C}(3), \hat{v}_3=-1 \\
\hline
1 & 1 & 2 & 1 & 1 & 3 & 1 & 1 & 2 & 3 \\
1 & 1 & 2 & 1 & 3 & 1 & 1 & 1 & 3 & 1 & 1 & 3
\end{array} \quad (3.15)$$

The advantage of breaking up a 0-cell is that the newly formed cells have different numbers of letters $\alpha_1$ on each strand, thus $N_v^-$ tends to increase in this
process while the length of the common subsequence remains the same. After applying this operation and getting enough cells with different numbers of letters $\alpha_1$ on the two strands, there is a high probability to find enough non-$\alpha_1$ letters on the strand with a smaller number of letters $\alpha_1$.

The previous example leads to our next definition.

**Definition 3.2** Let $k \in \mathbb{N}$, $v \in \mathbb{Z}^k \cap V$. Let $C_v(i)$ be any cell with $v_i = 0$, $1 \leq i \leq k$. Then, $C_v(i)$ is said to be breakable if there exists $j$ and $j'$ such that:

1. $X_j = Y_{j'} \in \{\alpha_2, \ldots, \alpha_m\}$;
2. $\pi_v(i - 1) < j < \pi_v(i)$ and $\nu_v(i - 1) < j' < \nu_v(i)$;
3. the difference between the number of letters $\alpha_1$ in $X_{\pi_v(i-1)+1}\cdots X_{j-1}$ and $Y_{\nu_v(i-1)+1}\cdots Y_{j'-1}$ is plus or minus one.

### 3.3 Probabilistic developments

After the combinatorial developments of the previous sections, let us now bring some probabilistic tools. We start by introducing a useful way of constructing alignments corresponding a given vector $v = (v_1, \ldots, v_k) \in \mathbb{R}^k$.

For $1 \leq i \leq n$ and $2 \leq j \leq m$, let $R^j_i$ (resp. $S^j_i$) be the number of letters $\alpha_j$ between the $(i - 1)$-th and $i$-th $\alpha_1$ in the infinite sequence ($X_i$)$_{i \geq 1}$ (resp. ($Y_i$)$_{i \geq 1}$), with $R^j_1$ (resp. $S^j_1$) be the number of letters $\alpha_j$ before the first $\alpha_1$.

Recall also from Definition 3.1 that to construct a 0-cell, we use the random time $T_0$, where

$$T_0 = \min_{2 \leq j \leq m} T^j_0,$$

where $T^j_0 := \min\{i = 1, 2, \ldots : R^j_i \neq 0, S^j_i \neq 0\}$. For a $-u$-cell ($u > 0$), the random time

$$T_{-u} = \min_{2 \leq j \leq m} T^j_{-u},$$

where $T^j_{-u} := \min\{i = 1, 2, \ldots : R^j_i \neq 0, S^j_{i+u} \neq 0\}$, and for a $u$-cell ($u > 0$),

$$T_u = \min_{2 \leq j \leq m} T^j_u,$$

where $T^j_u := \min\{i = 1, 2, \ldots : R^j_{i+u} \neq 0, S^j_i \neq 0\}$. In other words, a cell with $v_i = u$ can be constructed in the following way: First keep the first $u$ letters $\alpha_1$ in the $X$ strand, then align consecutive pairs of $\alpha_1$ until meeting the first pair of non-$\alpha_1$ letters.
Let us find the law of $R_i^j$ and, to do so, let $R_i^{j-1} = \sum_{j=2}^{m} R_i^j$, be the total number of non-$\alpha_1$ letters between the $(i-1)$-th and the $i$-th $\alpha_1$. Then, $R_i^{j-1} + 1$ is a geometric random variable with parameter $p_1$, i.e., $\Pr(R_i^{j} = k) = (1 - p_1)^{k} p_1$, $k = 0, 1, 2, \ldots$. Moreover, conditionally on $R_i^{j} > 0$, $(R_i^j)_{j=2}^m$ has a multinomial distribution and therefore

$$\Pr(R_i^j = k) = \sum_{l=k}^{\infty} \Pr(R_i^j = k|R_i^{j-1} = l)\Pr(R_i^{j-1} = l)$$

$$= \sum_{l=k}^{\infty} \binom{l}{k} \left(\frac{p_j}{1 - p_1}\right)^k \left(\frac{1 - p_1 - p_j}{1 - p_1}\right)^{l-k} (1 - p_1)^{j-l} p_1$$

$$= \left(\frac{p_1}{p_1 + p_j}\right) \left(\frac{p_j}{p_1 + p_j}\right)^k,$$

(3.19)

for $k = 0, 1, 2, \ldots$. Thus, $R_i^j + 1$ has a geometric distribution with parameter $p_1/(p_1 + p_j)$, $2 \leq j \leq m$.

To continue our probabilistic analysis, let us provide a rough lower bound on the length of the LCS. First, aligning as many letters $\alpha_1$ as possible in $X$ and $Y$, would get approximately a common subsequence of length $np_1$, then aligning as many letters $\alpha_2$ as possible without disturbing the already aligned $\alpha_1$, would give an additional $\sum_{i=1}^{np_1} \min\{R_i^2, S_i^2\}$ aligned $\alpha_2$. Moreover, since $R_i^2$ and $S_i^2$ are independent geometric random variables, $\min\{R_i^2, S_i^2\} + 1$ is a geometric random variable with parameter $1 - (p_2/(p_1 + p_2))^2$. So, on average, the aligned letters $\alpha_2$ contribute to the length of the LCS by an amount of:

$$np_1 \frac{p_2^2}{p_1 (p_1 + 2p_2)} = \frac{1}{p_1 + 2p_2} np_2^2 \geq (1 - p_2) np_2^2.$$

This heuristic argument leads to the following lemma:

**Lemma 3.1** Let $p_1 > 1/2$ and let $D_1 := \{LC_n \geq np_1 + (1 - p_2)^2 - p_2) np_2^2\}$. Then, $\Pr(D_1) \geq 1 - 4 \exp(-2np_2^2) - \exp(n(p_2^2 + \log(1 - p_2^2))(p_1 - p_2^2))$.

**Proof.** For $\delta > 0$, let $D_2^x(\delta) := \{|\sum_{i=1}^{n} 1_{\{x_i = \alpha_1\}} - np_1| \leq \delta n\}$, let $D_2^y(\delta) := \{|\sum_{i=1}^{n} 1_{\{y_i = \alpha_1\}} - np_1| \leq \delta n\}$, and let $D_2(\delta) := D_2^x(\delta) \cap D_2^y(\delta)$, so that on $D_2(\delta)$, at least $n_1(\delta) := n(p_1 - \delta)$ letters $\alpha_1$ can be aligned. Clearly, $1 + \min(R_i^2, S_i^2)$ has a geometric distribution with parameter $1 - (p_2/(p_1 + p_2))^2$. Also, if $G_1, \ldots, G_N$ are iid geometric random variables with parameter $p$, then for any $\beta < 1$,

$$\Pr\left(\sum_{i=1}^{N} G_i \leq \frac{\beta}{p} N\right) \leq \exp\left(- (\beta - 1 - \log \beta) N\right).$$

(3.20)
By taking \( p = 1 - (p_2/(p_1 + p_2))^2 \) and \( N = n_1(\delta) \), and since the sequences have length \( n \), the following equivalence holds true:

\[
\sum_{i=1}^{n_1(\delta)} \min(R_i^2, S_i^2) \overset{d}{=} \sum_{i=1}^{n_1(\delta)} (G_i \wedge n).
\]

For any \( \beta < 1 \), let us estimate

\[
P\left( \sum_{i=1}^{n_1(\delta)} \min(R_i^2, S_i^2) < \frac{\beta n_1(\delta)}{1 - \left( \frac{p_2}{p_1 + p_2} \right)^2} - n_1(\delta) \right).
\]

(3.21)

Since \( p_2 < p_1 \), then

\[
\frac{\beta n_1(\delta)}{1 - \left( \frac{p_2}{p_1 + p_2} \right)^2} - n_1(\delta) \leq n,
\]

and therefore,

\[
P\left( \sum_{i=1}^{n_1(\delta)} \min(R_i^2, S_i^2) < \frac{\beta n_1(\delta)}{1 - \left( \frac{p_2}{p_1 + p_2} \right)^2} - n_1(\delta) \right) \leq e^{-(\beta - 1 - \log \beta)n_1(\delta)}.
\]

Let

\[
D_3(\beta, \delta) := \left\{ \sum_{i=1}^{n_1(\delta)} \min(R_i^2, S_i^2) \geq \frac{\beta n_1(\delta)}{1 - \left( \frac{p_2}{p_1 + p_2} \right)^2} - n_1(\delta) \right\}.
\]

Choosing \( \delta = p_2^3 \) and \( \beta = 1 - p_2^3 \), and when \( D_2(\delta) \) and \( D_3(\beta, \delta) \) hold,

\[
LC_n \geq \frac{\beta n_1(\delta)}{1 - \left( \frac{p_2}{p_1 + p_2} \right)^2} - n_1(\delta) + n_1(\delta)
\]

\[
= np_2^2 \frac{p_1 - p_2^3}{(p_1 + p_2)^2 - p_2^2} + n(p_1 - p_2^3) - np_2^2 \frac{p_2(p_1 - p_2^3)}{1 - \left( \frac{p_2}{p_1 + p_2} \right)^2}
\]

\[
\geq np_1 + ((1 - p_2^3) - p_2) np_2.
\]

By Hoeffding’s inequality, for any \( \delta > 0 \),

\[
P(D_2^\delta(\delta)) \geq 1 - 2e^{-2n\delta^2}, \quad P(D_2^\delta(\delta)) \geq 1 - 2e^{-2n\delta^2},
\]

but since \( D_2(p_2^3) \cap D_3(1 - p_2^3, p_2^3) \subset D_1 \), it follows that

\[
P(D_1) \geq 1 - 4 \exp(-2np_2^6) - \exp \left( n(p_2^3 + \log(1 - p_2^3))(p_1 - p_2^3) \right).
\]
To state our next lemma, let us introduce some more notation. First, let
\[ V(k) := \{ (v_1, v_2, \ldots, v_k) \in \mathbb{Z}^k : |v_1| + \cdots + |v_k| \leq 2k \} \quad (3.22) \]
then, recalling that \( V \) as defined in (3.6) is the set of admissible alignments, let
\[ P := \bigcup_{2k \geq np_2^3} \left( V \cap V(k) \right) . \quad (3.23) \]
With these definitions, the previous lemma further yields:

**Lemma 3.2** Let \( D = \{ v \in P : v \text{ encodes an optimal alignment} \} \), and let \( p_2 < 1/10 \), then \( \mathbb{P}(D) \geq 1 - 5 \exp \left( -np_2^3/5 \right) \).

**Proof.** Let \( N_1^x \) be the number of letters \( \alpha_1 \) in \( X = (X_i)_{1 \leq i \leq n} \), and \( N_1^y \) be the corresponding number in \( Y = (Y_i)_{1 \leq i \leq n} \), so that \( N_1 = N_1^x + N_1^y \) is the number of letters \( \alpha_1 \) in \( X \) and \( Y \). From the proof of the previous lemma, we see that whenever \( D_2(p_3^2) \) holds then \( N_1 \) is upper bounded by \( 2n(p_1 + p_2^3) \). Therefore on \( D_1 \), and provided that \( p_2 < 1/10 \),
\[ LC_n \geq \frac{N_1}{2} - np_2^3 + ((1 - p_2)^3 - p_2) np_2^2 \geq \frac{N_1}{2} + \frac{1}{2} np_2^2 , \quad (3.24) \]
Now assume that \( v = (v_1, \ldots, v_k) \in \mathbb{R}^k \) is an optimal alignment, then
\[ LC_n \leq \frac{N_1}{2} - \frac{1}{2} \sum_{i=1}^{k} |v_i| + k , \quad (3.25) \]
which when combined with (3.24) yields \( \sum_{i=1}^{k} |v_i| \leq 2k \) and \( np_2^2 \leq 2k \), finishing the proof. \( \blacksquare \)

The previous lemma asserts that, with high probability, any optimal alignment belongs to the set \( P \). Hence, proving a property of the optimal alignments essentially only requires to prove it for alignments in \( P \).

### 3.4 High probability events

Recall from Definition 3.1 that to \( v \in \mathbb{Z}^k \) is associated an alignment which has \( |v| \) cells \( C_0(1), \ldots, C_0(|v|) \), and that a cell is called a nonzero-cell if it contains a different number of letters \( \alpha_1 \) on the \( X \) strand and on the \( Y \) strand. Let \( W \) be the subset of \( P \), consisting of the alignments for which the proportion of the nonzero-cells is at least \( \theta \), i.e.,
\[ W := \{ v \in P : |\{ i \in [1, k] : v_i \neq 0 \}| \geq \theta k \} , \]
and let \( W^c := P \setminus W \).

To finish the proof of the theorem, let us define some further relevant events.
Let $E_v$ be the event that, among the zero-cells in $C_v(1), \ldots, C_v(|v|)$, the proportion which are breakable is at least $\theta$. Then, let

$$E := \bigcap_{v \in W^c} E_v := \bigcap_{v \in W^c} \{I_b \geq \theta J_0\},$$

where $J_0$ is the number of zero-cells while $I_b$ is the number of breakable zero-cells for $v$. i.e., $E$ is the event that for all $v \in W^c$, the proportion of breakable zero-cells is at least $\theta$.

Recall also from (3.12) and (3.13), that $N_v^-$ is the number of non-$\alpha_1$ letters on the cell strands with a lesser number of $\alpha_1$, and that $N_{>1}$ is the total number of non-$\alpha_1$ letters in $X$ and $Y$. Let

$$F := \bigcap_{v \in W} F_v := \bigcap_{v \in W} \left\{N_v^- \geq \frac{K}{m}N_{>1}\right\},$$

i.e., $F$ is the event that for every $v \in W$, the proportion of non-$\alpha_1$ letters which are on the cell-strand with the smaller number of letters $\alpha_1$, is at least $K/m$.

Let

$$G := \bigcap_{v \in W} G_v := \bigcap_{v \in W} \left\{2|v| \leq \frac{K}{2m}N_{>1}\right\},$$

i.e., $G$ is the event that for every $v \in W$, the proportion of non-$\alpha_1$ letters which are aligned is at most $K/2m$.

Recall finally from Section 3.2, that $A_n = \{(X, Y) \in B_n\}$ is the event that there exists an optimal alignment $v$ such that $N_v^- \geq KN_{>1}/m$ and $2|v| \leq KN_{>1}/2m$, and therefore

$$D \cap E \cap F \cap G \subset A_n. \quad (3.26)$$

Our next task is to prove that there exists $K > 0$ such that the events $E, F, G$ hold with high probability. Let us start with $E$.

**Lemma 3.3** Let $0 < \theta < 1$, then

$$\mathbb{P}(E) \geq 1 - \sum_{2k \geq np_2^2} \exp \left( - \left( 2(1 - \theta) \left( \frac{p_1^2}{1 + p_1^2} - \theta \right)^2 - \log f(\theta) \right) k \right), \quad (3.27)$$

where $f(\theta) = ((4 + 2\theta)/\theta^2)^\theta \left((2 + \theta)/2\right)^2 \left((1/(1 - \theta))\right)^{1-\theta}$. 

22
Proof. For any \( v \in W^c \), let us compute the probability that a 0-cell in the alignment associated with \( v \) is breakable. Recalling the definition of \( T_0 \) in (3.16), for \( 2 \leq j \leq m \), let \( M_j \) be the event that this cell ends with a pair of letters \( \alpha_j \), and so when \( M_j \) holds \( T_0 = T_0^{j} \). For \( 2 \leq j \leq m \), let

\[
U_{1}^{j} := \min \{ i = 2, 3, \ldots : R_{i-1}^{j} \neq 0, S_{i-1}^{j} = 0, R_{i}^{j} = 0, S_{i}^{j} \neq 0 \},
\]

\[
U_{2}^{j} := \min \{ i = 2, 3, \ldots : R_{i-1}^{j} = 0, S_{i-1}^{j} \neq 0, R_{i}^{j} \neq 0, S_{i}^{j} = 0 \},
\]

\[
U^{j} := \min \{ U_{1}^{j}, U_{2}^{j} \}.
\]

With the above constructions, conditional on the event \( M_j \), if \( U^{j} < T_0^{j} \) then this 0-cell is breakable, thus to lower bound the probability that this 0-cell is breakable, it is enough to lower bound \( \mathbb{P}(U^{j} < T_0^{j}) \). To do so, let first \( (Z_{i}^{j})_{i \geq 1} \) be the independent random vectors given via:

\[
Z_{i}^{j} = (R_{2i-1}^{j}, S_{2i-1}^{j}, R_{2i}^{j}, S_{2i}^{j}).
\]

Then, let

\[
\tilde{U}^{j} = \min \{ i = 1, 2, \ldots : Z_{i}^{j} \in B_{1} \cup B_{2} \},
\]

\[
\tilde{T}_{0}^{j} = \min \{ i = 1, 2, \ldots : Z_{i}^{j} \in B_{3} \cup B_{4} \},
\]

where

\[
B_{1} := N^{*} \times \{ 0 \} \times \{ 0 \} \times N^{*}, B_{2} := \{ 0 \} \times N^{*} \times N^{*} \times \{ 0 \},
\]

\[
B_{3} := N^{*} \times N^{*} \times N \times N^{*}, B_{4} := N \times N \times N^{*} \times N^{*},
\]

and where as usual \( N \) is the set of non-negative integers, while \( N^{*} = N \setminus \{ 0 \} \). Clearly, \( 2\tilde{U}^{j} \geq U^{j}, 2\tilde{T}_{0}^{j} - 1 \leq T_0^{j} \), thus \( \mathbb{P}(U^{j} < T_0^{j}) \) \( \geq \mathbb{P}(2\tilde{U}^{j} < 2\tilde{T}_{0}^{j} - 1) = \mathbb{P}(\tilde{U}^{j} < \tilde{T}_{0}^{j}) \). Now, since the random vectors \( (Z_{i}^{j})_{i \geq 1} \) are iid, and since \( B_{1} \cup B_{2} \) and \( B_{3} \cup B_{4} \) are pairwise disjoint,

\[
\mathbb{P}(\tilde{U}^{j} < \tilde{T}_{0}^{j}) = \frac{\mathbb{P}(Z_{i}^{j} \in B_{1} \cup B_{2})}{\mathbb{P}(Z_{i}^{j} \in B_{1} \cup B_{2}) + \mathbb{P}(Z_{i}^{j} \in B_{3} \cup B_{4})}
\]

\[
= \frac{1 + p_{1}^{2}}{2p_{1}^{2} + 2(p_{1}^{2} + p_{j}^{2})} \geq \frac{p_{1}^{2}}{1 + p_{1}^{2}}.
\]

Therefore,

\[
\mathbb{P}(\text{a 0-cell is breakable}) = \sum_{j=2}^{m} \mathbb{P}(\text{a 0-cell is breakable}|M_j)\mathbb{P}(M_j)
\]

\[
= \sum_{j=2}^{m} \mathbb{P}(U^{j} < T_0^{j})\mathbb{P}(M_j) \geq \frac{p_{1}^{2}}{1 + p_{1}^{2}}.
\]
Let $J$ be the index set of 0-cells in the alignment associated with $v \in W^c$, and so $|J| \geq (1 - \theta)|v|$. For each $i \in J$, let $I_i$ be the Bernoulli random variable which is one if the cell $C_v(i)$ is breakable and 0 otherwise. Recall that $E_v$ is the event that the proportion of breakable cells in $v$ is at least $\theta$. Then, from Hoeffding’s inequality, and after subtracting the mean,

$$\mathbb{P}(E_v^c) = \mathbb{P}\left(\sum_{i \in J} I_i < \theta |J|\right) \leq \exp\left(-2(1 - \theta)|v| \left(\frac{p_1^2}{1 + p_1^2} - \theta\right)^2\right).$$

Recall the definition of $V(k)$ in (3.22), and let $W^c(k) := W^c \cap V(k)$. For any two integers, $\ell$ and $q\ell$, with $0 < q < 1$, Stirling’s formula in the form $1 \leq \ell! e^{\ell}/(\sqrt{2\pi\ell\ell^\ell}) \leq e/\sqrt{2\pi}$, gives

$$\left(\begin{array}{c}
\ell \\
q\ell
\end{array}\right) \leq q^{q\ell}(1 - q)^{-(\ell - q\ell)}, \quad (3.28)$$

which when combined with simple estimates yield,

$$|W^c(k)| \leq 2^{\theta k}\left(\frac{2k + \theta k}{\theta k}\right)^k \leq (f(\theta))^k$$

$$:= \left(\frac{4 + 2\theta}{\theta^2}\right)^\theta \left(\frac{2 + \theta}{2}\right)^2 \left(\frac{1}{1 - \theta}\right)^{1 - \theta}^k. \quad (3.29)$$

Next, let $E(k) = \bigcap_{v \in W^c(k)} E_v$, then

$$\mathbb{P}(E(k)^c) \leq \sum_{v \in W^c(k)} \mathbb{P}(E_v^c) \leq \exp\left(-2(1 - \theta)\left(\frac{p_1^2}{1 + p_1^2} - \theta\right)^2 - \log f(\theta)\right) k,$$

and, therefore,

$$\mathbb{P}(E^c) \leq \sum_{2k \geq np_2^2} \mathbb{P}(E(k)^c)$$

$$\leq \sum_{2k \geq np_2^2} \exp\left(-2(1 - \theta)\left(\frac{p_1^2}{1 + p_1^2} - \theta\right)^2 - \log f(\theta)\right) k. \quad (3.30)$$

Of course, in (3.30), one wants

$$2(1 - \theta)\left(\frac{p_1^2}{1 + p_1^2} - \theta\right)^2 - \log f(\theta) > 0, \quad (3.31)$$

24
and choices of $\theta$ for which this is indeed the case are given later.

Let $u$ be a non-negative integer. For any $-u$-cell ending with an aligned pair of letters $\alpha_j$ (the event $M_j$ holds for this cell), let $\tau^j_x(l)$ be the index of the $l$-th $R^j_i$ such that $R^j_i \neq 0$, i.e.,

$$\tau^j_x(1) = \min\{i \geq 1 : R^j_i \neq 0\},$$

and for any $l \geq 1$, $\tau^j_x(l + 1) = \min\{i > \tau^j_x(l) : R^j_i \neq 0\}$. Let

$$\rho^{j, -} := \min\{l = 1, 2, \ldots : S^j_{u+\tau^j_x(l)} \neq 0\}.$$

In words, $\rho^{j, -}$ is the number of nonzero values taken by $R^j = (R^j_i)_{1 \leq i \leq s}$ (where $s$ is the number of letters $\alpha_1$ in the $X$-strand) in the cell (including the last one corresponding to the aligned pair of letters $\alpha_j$). Since $X$ and $Y$ are independent,

$$P(\rho^{j, -} = k) = \left(\frac{p_1}{p_1 + p_j}\right)^{k-1} \left(\frac{p_j}{p_1 + p_j}\right),$$

for $k = 1, 2, \ldots$. Thus, $\rho^{j, -}$ has a geometric distribution with parameter $\tilde{p}_j = p_j/(p_1 + p_j)$, $2 \leq j \leq m$. When $-u < 0$, the number of letters $\alpha_j$ in the $X$-strand (which is the strand with the smaller number of letters $\alpha_1$) is at least $\rho^{j, -} - 1$ and, as shown in the next lemma, this provides a lower bound for $N_v^-$ (the number of non-$\alpha_1$ letters on the cell-strand with the lesser number of letters $\alpha_1$) in this $-u$-cell.

**Lemma 3.4** Let $K = \min(2^{-4}10^{-2}e^{-67}, 1/800m)$, and let $p_1 \geq 1 - e^{-67}/4$, then $P(F) \geq 1 - 38 \exp(-3np_2^2/200)$.

**Proof.** For any $v \in W$, let now $J$ be the index set of the nonzero cells of the alignment corresponding to $v$. Hence, $|J| \geq \theta|v|$. Then,

$$N_v^- = \sum_{i=1}^{|v|} N_v^-(i) = \sum_{i \in J} N_v^-(i) \geq \sum_{i \in J} \left(\rho^{j(i), -}_i - 1\right),$$

where $j(i)$ is the index of the last aligned pair of letters $\alpha_j$ in the cell $C_v(i)$, and $\rho^{j(i), -}_i$ is the number of nonzero $R^{j(i)} = (R^{j(i)}_j)_{1 \leq j \leq s}$ (s is the number of letters $\alpha_1$ in the $X$-strand of $C_v(i)$ or the number of non-zero values taken by $S^{j(i)}$, if the $Y$-strand has a lesser number of letters $\alpha_1$). From (3.32), $\rho^{j(i), -}_i$ is a geometric
random variable with parameter $\tilde{p}_{j(i)}$. Now, let $\varepsilon > 0$, let again $\tilde{p}_2 = p_2/(p_1 + p_2)$, and let $F_{1,v} := \left\{ N_v^- \geq \frac{\varepsilon}{p_2} |v| \right\}$. Then,

$$P(F_{1,v}^c) \leq P \left( \sum_{i \in J} \left( \frac{\rho_{i}^{j(i),-}}{\tilde{p}_2} - 1 \right) \leq \frac{\varepsilon}{\tilde{p}_2} |v| \right) \leq P \left( \sum_{i \in J} \rho_{i}^{j(i),-} \leq \frac{\varepsilon/\theta + \tilde{p}_2}{\tilde{p}_2} |J| \right) \leq P \left( \sum_{i \in J} \rho_{i}^{j(i),-} \leq \frac{\varepsilon/\theta + 2p_2}{\tilde{p}_2} |J| \right). \quad (3.33)$$

The geometric random variables $\rho_{i}^{j(i),-}, i \in J$, are independent each with parameter $\tilde{p}_{j(i)} \leq \tilde{p}_2$, and moreover the sequences have finite length $n$, therefore,

$$P \left( \sum_{i \in J} \rho_{i}^{j(i),-} \leq \frac{\varepsilon/\theta + 2p_2}{\tilde{p}_2} |J| \right) \leq P \left( \sum_{i \in J} G_i \leq \frac{\varepsilon/\theta + 2p_2}{\tilde{p}_2} |J| \right),$$

where the $G_i$ are iid geometric random variables with parameter $\tilde{p}_2$. As proved later, when

$$\frac{\varepsilon/\theta + 2p_2}{\tilde{p}_2} |J| < n,$$  

it follows, using (3.20), that

$$P \left( \sum_{i \in J} \rho_{i}^{j(i),-} \leq \frac{\varepsilon/\theta + 2p_2}{\tilde{p}_2} |J| \right) \leq \exp \left((1 + \log(\varepsilon/\theta + 2p_2)) \theta |v|) \right). \quad (3.35)$$

Let $F_1(k) := \cap_{v \in \mathcal{W} \cap V(k)} F_{1,v} = \left\{ N_v^- \geq \frac{\varepsilon}{p_2} |v| \right\}$, and let $F_1 := \cap_{n \geq n_2} F_1(k)$. From the very definition of $V(k)$ in (3.22), and using (3.28),

$$|V(k)| \leq 2^k \frac{3^k}{k} \leq 2^{k} 3^k \left( \frac{3}{2} \right)^{2k} = \left( \frac{27}{2} \right)^k,$$

which when combined with (3.35) leads to

$$P(F_1(k)) \geq 1 - \exp \left(k \log(27/2) + k (1 + \log(\varepsilon/\theta + 2p_2)) \theta \right). \quad (3.36)$$

Of course, one wants $\log(27/2) + (1 + \log(\varepsilon/\theta + 2p_2)) \theta < 0$. Choosing $\theta = 1/25$ and $\varepsilon = 10^{-2} e^{-67}$, then $P((F_1(k))^c) \leq e^{-3k/100}$, for any $p_1 \geq 1 - 2^{-2} e^{-67}$, and
so $\mathbb{P}(F^c_1) \leq \sum_{2k \geq np_2^2} \mathbb{P}((F_1(k))^{c}) \leq 34 \exp(-3np_2^2/200)$. Note also that for these choices of $\theta$ and $p_1$, (3.31) is satisfied and so $E$ also holds with high probability.

From the proof of Lemma 3.1 when $D_2((1 - p_1))$ holds, the total number of non-$\alpha_1$ letters in $X$ and $Y$ is at most $4n(1 - p_1)$. Thus $N_{>1} \leq 4n(1 - p_1)$, and so when $F_1 \cap D_2((1 - p_1))$ holds, for every $v \in W$,

$$\frac{N_v}{N_{>1}} \geq \frac{\varepsilon |v|}{\bar{p}_2 4n(1 - p_1)} \geq \frac{\varepsilon}{\bar{p}_2} \frac{np_2}{4n(1 - p_1)} \geq \frac{\varepsilon}{16m} \geq \frac{K}{m}.$$

We also note that, by properly choosing these constants and under the condition $400mK < 1$, (3.34) is true. Therefore,

$$\mathbb{P}(F^c) \leq \mathbb{P}(F^c_1) + \mathbb{P}((D_2(1 - p_1))^c) \leq 34 \exp(-3np_2^2/200) + 4 \exp(-2n(1 - p_1)^2) \leq 38 \exp(-3np_2^2/200).$$

Lemma 3.5 Let $K = \min(2^{-4}10^{-2}e^{-67}, 1/800m)$, and assume, moreover, that $p_2 \leq \min\{2^{-2}e^{-5K/m}, K/2m^2\}$. Then, $\mathbb{P}(G) \leq 1 - 4 \exp(-np_2^2/2)$.

Proof. For any $v \in W$, let $C_v(1), \ldots, C_v(|v|)$ be the corresponding cells. If the cell $C_v(i)$ ends with a pair of aligned $\alpha_j$, $2 \leq j \leq m$, then let $\rho_i^{(i)}$ be the number of non-zero values taken by $R_i^{(i)}$ in $C_v(i)$. If $v_i \leq 0$, by the same argument as in getting (3.32), $\rho_i^{(i)}$ has a geometric distribution with parameter $\bar{p}_j(i)$. If $v_i > 0$, there exists a geometric random variable $\rho_i^{(i),-}$ with parameter $\bar{p}_j(i)$ such that $\rho_i^{(i),-} \leq \rho_i^{(i)} \leq \rho_i^{(i),-} + v_i$. Let $N^x_{>1}$ (resp. $N^y_{>1}$) be the number of non-$\alpha_1$ letters in $X$ (resp. $Y$), so that $N_{>1} = N^x_{>1} + N^y_{>1}$. Let

$$G^x_v := \left\{|v| \leq \frac{K}{2m} N^x_{>1}\right\} \text{ and } G^y_v := \left\{|v| \leq \frac{K}{2m} N^y_{>1}\right\},$$

and so $G^x_v \cap G^y_v \subseteq G_v$. Since $N^x_{>1} \geq \sum_{i=1}^{\varepsilon} \rho_i^{(i)}$, then,

$$\mathbb{P}((G^x_v)^c) \leq \mathbb{P}\left(|v| > \frac{K}{2m} \sum_{i=1}^{\varepsilon} \rho_i^{(i)}\right) \leq \mathbb{P}\left(|v| > \frac{K}{2m} \left(\sum_{1 \leq i \leq |v|, v_i \leq 0} \rho_i^{(i)} + \sum_{1 \leq i \leq |v|, v_i > 0} \rho_i^{(i),-}\right)\right) \leq \mathbb{P}\left(\sum_{i=1}^{\varepsilon} (G_i \cap n) < \frac{2m|v|}{K}\right),$$

27
where the $G_i$ are iid geometric random variables with parameter $\tilde{p}_2$ and the truncation is at $n$, the sequences having such a length. Since $2m|v| \leq 2mn(1 - p_1) < 2m^2p_2n$, and if $2m^2p_2 < K$, then for any $p_2 \leq 2^{-2}e^{-5}K/m$,

$$
\mathbb{P}((G^v)^c) \leq \mathbb{P} \left( \sum_{i=1}^{|v|} G_i < \frac{2m|v|}{K} \right) \leq \mathbb{P} \left( \sum_{i=1}^{|v|} G_i < \frac{e^{-5}|v|}{\tilde{p}_2} \right) \leq \exp(-4|v|).
$$

Likewise, $\mathbb{P}((G^v_w)^c) \leq \exp(-4|v|)$, and thus $\mathbb{P}((G^v)^c) \leq 2\exp(-4|v|)$. As previously, let $G(k) := \bigcap_{v \in W \cap V(k)} G_v$ and $G = \bigcap_{2k \geq np_2^2} G(k)$, then $\mathbb{P}((G(k))^c) \leq |V(k)|2\exp(-4k) \leq 2\exp(-k)$, and

$$
\mathbb{P}(G^c) \leq \sum_{2k \geq np_2^2} \mathbb{P}((G(k))^c) \leq 4\exp(-np_2^2/2).
$$

Combining Lemma 3.2, 3.3, 3.4 and 3.5, using (3.26), letting $\theta = 1/25$, and $K = \min(2^{-4}10^{-2}e^{-67}, 1/800m)$, it follows that for $p_2 \leq \min\{2^{-2}e^{-5}K/m, K/2m^2\}$,

$$
\mathbb{P}(A_n^c) \leq \mathbb{P}(D^c) + \mathbb{P}(E^c) + \mathbb{P}(F^c) + \mathbb{P}(G^c)
\leq 5 \exp \left( -\frac{np_2^6}{5} \right) + 74 \exp \left( -\frac{np_2^6}{10^2} \right) + 38 \exp \left( -\frac{3np_2^2}{200} \right) + 4 \exp \left( -\frac{np_2^2}{2} \right)
\leq 121 \exp \left( -\frac{np_2^6}{5} \right), \quad (3.37)
$$

and this finishes the proof of Theorem 2.1.

**Remark 3.1** (i) Our results on the central absolute moments of the LCS continue to be valid for three sequences or more. First, the upper bound methods are very easily adapted to provide the same order. Next, for the lower bound, the alignments can still be represented with a series of cells, each of the cells ending with the same non-$\alpha_1$ letter from every strand. Then, with exponential bounds techniques, a similar high probability event can be exhibited, leading to the result.

(ii) With the methodology developed here, the results of [2] and [6] can also be generalized, beyond the variance or the Bernoulli case, to centered absolute moments and $m$-letters alphabets.

**Acknowledgements:** Many thanks to Ruoting Gong for his detailed comments on this paper.
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] S. Amsalu, C. Houdré and H. Matzinger. Sparse long blocks and the variance of the LCS. Preprint arXiv:math/1204.1009, 2012.

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

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

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

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

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

[8] 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.

[9] W. T. Rhee, M. Talagrand. Martingale inequalities and the jackknife estimate of variance. Statist. Probab. Lett. 4, no. 1, 5-6, 1986.

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