On the Variance of the Length of the Longest Common Subsequences in Random Words With an Omitted Letter

Christian Houdré∗ Qingqing Liu†

December 27, 2018

Abstract

We investigate the variance of the length of the longest common subsequences of two independent random words of size $n$, where the letters of one word are i.i.d. uniformly drawn from \{\alpha_1, \alpha_2, \ldots, \alpha_m\}, while the letters of the other word are i.i.d. drawn from \{\alpha_1, \alpha_2, \ldots, \alpha_m, \alpha_{m+1}\}, with probability $p > 0$ to be $\alpha_{m+1}$, and $(1 - p)/m > 0$ for all the other letters. The order of the variance of this length is shown to be linear in $n$.

1 Introduction and Statement of Results

Let $X = (X_i)_{i \geq 1}$ and $Y = (Y_i)_{i \geq 1}$ be two independent sequences of i.i.d. random variables taking their values in a finite common alphabet $\mathcal{A}$, with $P(X_1 = \alpha) = p_{x,\alpha} \geq 0$ and $P(Y_1 = \alpha) = p_{y,\alpha} \geq 0$, $\alpha \in \mathcal{A}$. Let $LC_n$ be the largest $k$ such that there exist $1 \leq i_1 < \cdots < i_k \leq n$ and $1 \leq j_1 < \cdots < j_k \leq n$ with $X_{i_s} = Y_{j_s}$ for $s = 1, \ldots, k$, i.e., $LC_n$ denotes the length of the longest common subsequences of the random words $X^{(n)} := X_1 \cdots X_n$ and $Y^{(n)} := Y_1 \cdots Y_n$. The limiting behavior of the expectation of $LC_n$ has been extensively studied. In particular, if for all $\alpha \in \mathcal{A}$, $p_{x,\alpha} = p_{y,\alpha} = 1/|\mathcal{A}|$, where $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$, the earliest result is due to Chvátal and Sankoff [3], who proved the existence of

$$\gamma^* = \lim_{n \to \infty} \frac{E(LC_n)}{n},$$

where $m$ denotes the alphabet size, showing also that $0.727273 \leq \gamma^* \leq 0.905118$. Much work has since been done to improve these bounds ([6], [3], [7], [5], . . .), and to date the best known bounds seem to be $0.788071 \leq \gamma^* \leq 0.826280$, see [15]. These results have also been extended to multiple sequences and alphabet of size larger than two, e.g., see [11], [13] and the references therein.

The study of the variance of $LC_n$ is less complete. In case $p_{x,k} = p_{y,k} = p_k$ for $k = 1, \ldots, m$, the Efron-Stein inequality implies, as shown in [10], that

$$\text{Var } LC_n \leq n \left(1 - \sum_{k=1}^m p_k^2\right).$$

For lower bounds, linear order results are also proved in various biased instances ([12], [9], [10], [13], [8], [1], [2], . . .). For example, [12] and [9] assume that one of the letters has a significantly higher probability of appearing than any of the other letters in the alphabet, while [2] assumes that one of the two sequences is binary while the other is a trinary one. Our paper extends the result of [2] by

∗School of Mathematics, Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA 30332-0160 USA, houdre@math.gatech.edu. Research supported in part by the grant #246283 and #524678 from the Simons Foundation.
†School of Mathematics, Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA 30332-0160 USA, qqliu@gatech.edu

Keywords: Longest common subsequences, variance, lower bound

MSC 2010: 60C05, 60F10, 05A05
removing the binary/trinary assumptions and provides precise estimates allowing us to go beyond the uniform case and to also deal with central moments.

To formally state our problem, let $\mathcal{A} := \mathcal{A}_{m+1} = \{\alpha_1, \alpha_2, \ldots, \alpha_m, \alpha_{m+1}\}$, and let the letters distribution of $X$ to be such that

$$P(X_1 = \alpha_1) = \cdots = P(X_1 = \alpha_m) = \frac{1-p}{m} > 0,\quad P(X_1 = \alpha_{m+1}) = p > 0,$$

while the letters distribution of $Y$ is such that

$$P(Y_1 = \alpha_1) = \cdots = P(Y_1 = \alpha_m) = \frac{1}{m}.$$

To start with, an upper bound on the variance of $LC_n$ is shown to be

$$\text{Var} \, LC_n \leq \frac{n}{2} \left(2 - p^2 - \frac{1 + (1-p)^2}{m}\right),$$

for all $n \in \mathbb{N}$. Indeed, the Efron–Stein inequality states that:

$$\text{Var} \, S \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E} (S - S_i)^2, \quad (1.1)$$

where, $S = S(Z_1, Z_2, \cdots, Z_n)$ and $S_i = S(Z_1, Z_2, \cdots, Z_{i-1}, \hat{Z}_i, Z_{i+1}, \cdots, Z_n)$, and where $(Z_i)_{1 \leq i \leq n}$ and $(\hat{Z}_i)_{1 \leq i \leq n}$ are independent copies of each other.

Now following [16],

$$\mathbb{E} |LC_n - LC_n(X_1 \cdots X_{i-1} \hat{X}_i X_{i+1} \cdots X_n; Y_1 \cdots Y_n)|^2$$

$$\quad = \mathbb{E} \left( |LC_n - LC_n(X_1 \cdots X_{i-1} \hat{X}_i X_{i+1} \cdots X_n; Y_1 \cdots Y_n)|^2 1_{\hat{X}_i \neq X_i} \right)$$

$$\quad \leq P(X_i \neq \hat{X}_i) = 1 - \sum_{i=1}^{m+1} \left(P(X_1 = \alpha_i)\right)^2$$

$$\quad = 1 - m \left(\frac{1-p}{m}\right)^2 - p^2$$

$$\quad = (1-p) \left(1 - \frac{1}{m} + p \left(1 + \frac{1}{m}\right)\right),$$

since when replacing $X_i$ by $\hat{X}_i$, $LC_n$ changes by at most 1 and at least $-1$. Similarly,

$$\mathbb{E} |LC_n - LC_n(X_1 \cdots X_n; Y_1 \cdots Y_{i-1} \hat{Y}_i Y_{i+1} \cdots Y_n)|^2 \leq 1 - \sum_{i=1}^{m} \left(P(Y_1 = \alpha_i)\right)^2$$

$$\quad = 1 - \frac{1}{m}.$$

Applying (1.1) and combining the two bounds above give,

$$\text{Var} \, LC_n \leq \frac{1}{2} \left\{ \left(1-p\right) \left(1 - \frac{1}{m} + p \left(1 + \frac{1}{m}\right)\right)\right\} n + \left(1 - \left(\frac{1}{m}\right)\right) n$$

$$\quad = \frac{n}{2} \left(2 - p^2 - \frac{1 + (1-p)^2}{m}\right). \quad (1.2)$$

To match the easy bound (1.2), we can now state the main result of this paper.

**Theorem 1.** There exists a constant $C = C(p,m) > 0$ independent of $n$, such that for all $n \geq 1$,

$$\text{Var} \, LC_n \geq Cn. \quad (1.3)$$

This theorem, combined with the upper bound (1.2), gives a linear order, in $n$, for the variance of $LC_n$, and we refer the reader to Section 3 for an estimate on $C$. 

2
2 Proof of Theorem 1

The scheme of the proof elaborates and extends elements of of \cite{2} and \cite{9}. So, let $N$ denote the number of letters $\alpha_{m+1}$ in the random word $X^{(n)}$. Clearly, $N$ is a binomial random variable with parameter $n$ and $p$. Moreover, let $X^{(n)} := X_{i_1} \cdots X_{i_k}$, where $1 \leq i_1 < \cdots < i_k \leq n$, $X_j = \alpha_{m+1}$ for all $j \in \{i_1, \ldots, i_k\}$ and $X_j = \alpha_{m+1}$ for all $j \in \{1, 2, \ldots, n\} \setminus \{i_1, \ldots, i_k\}$. In words, $X^{(n)}$ is the subword of $X^{(n)}$ made only of non-$\alpha_{m+1}$ letters. To prove our main theorem, we will recursively define a finite random sequence $Z^{(1)}, Z^{(2)}, \ldots, Z^{(n)}$, where each $Z^{(k)}$ has length $k$, by inserting uniformly at random and at a uniform random location a letter from $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ to the previous $Z^{(k-1)}$.

To formally describe the defining mechanism, let $\{U_k\}_{1 \leq k \leq n}$ and $\{T_k\}_{3 \leq k \leq n}$ be two independent sequences of random variables, where $\{U_k\}_{1 \leq k \leq n}$ is a sequence of i.i.d. uniform random variables on $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$, and $\{T_k\}_{3 \leq k \leq n}$ is a sequence of independent random variables uniform on $\{2, 3, \ldots, k - 1\}$, $k \geq 3$.

Then as in \cite{2}, recursively define the sequence $Z^{(k)}$ via:

1. $Z^{(1)} = U_1$.
2. $Z^{(2)} = U_1 U_2$.
3. For $k \geq 2$, given $Z^{(k)} = Z_1^k Z_2^k \cdots Z_k^k$, let $Z^{(k+1)}$ be as follows:
   - For all $j < T_{k+1}$, let $Z_{j+1}^{k+1} = Z_j^k$.
   - For $j = T_{k+1}$, let $Z_{j+1}^{k+1} = U_{k+1}$.
   - For all $j$ such that $T_{k+1} < j \leq k + 1$, let $Z_{j+1}^{k+1} = Z_{j-1}^k$.

Hence, $\{Z_k^1\}_{1 \leq k \leq n}$ is a triangular array of uniform random variables with values in $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$, and finding the relation between $Z^{(n-N)}$ and $X^{(n)}$ is the purpose of our next lemma whose proof is akin to a corresponding proof in \cite{9}.

Lemma 1. For any $n \geq 1$ and $1 \leq k \leq n$,

\[
Z^{(k)} \overset{d}{=} (\tilde{X}^{(n)} | N = n - k),
\]

and moreover,

\[
Z^{(n-N)} \overset{d}{=} \tilde{X}^{(n)},
\]

where $\overset{d}{=}$ denotes equality in distribution.

Proof. The proof is by induction on $k$. Let $k = 1$, by definition, $Z^{(1)} = U_1$, which has the same distribution as $(\tilde{X}^{(n)} | N = n - 1)$. Next, assume that

\[
Z^{(k)} \overset{d}{=} (\tilde{X}^{(n)} | N = n - k), \quad 2 \leq k \leq n - 1,
\]

and so for any $(\alpha_j_1, \alpha_j_2, \ldots, \alpha_j_k) \in A^k$,

\[
P\left((Z_1^k, Z_2^k, \ldots, Z_k^k) = (\alpha_j_1, \alpha_j_2, \ldots, \alpha_j_k)\right) = \left(\frac{1}{m}\right)^k.
\]
Then, 

\[ P \left( (Z_1^{k+1}, Z_2^{k+1}, \ldots, Z_{k+1}^{k+1}) = (\alpha_{j_1}, \alpha_{j_2}, \ldots, \alpha_{j_{k+1}}) \right) \]

\[ = \sum_{t=2}^{k} P \left( (Z_1^{k+1}, Z_2^{k+1}, \ldots, Z_{k+1}^{k+1}) = (\alpha_{j_1}, \alpha_{j_2}, \ldots, \alpha_{j_{k+1}})|T_{k+1} = t \right) P(T_{k+1} = t) \]

\[ = \sum_{t=2}^{k} P \left( (Z_t^{k}, Z_{t-1}^{k}, Z_t^{k}) = (\alpha_{j_1}, \ldots, \alpha_{j_i}, \ldots, \alpha_{j_{k+1}}) \right) P(U_{k+1} = \alpha_{j_t}) P(T_{k+1} = t) \]

\[ = \sum_{t=2}^{k} \left( \frac{1}{m} \right)^k \frac{1}{m} \frac{1}{k-1} \]

\[ = \left( \frac{1}{m} \right)^{k+1}. \]

Thus,

\[ Z^{(k+1)} \overset{d}{=} (\check{X}^{(n)}|N = n - k - 1). \]

To prove the second part of the lemma, from the independence of \( N \) and \( Z^{(n-k)} \), for any \( u \in \mathbb{R}^{n-k} \),

\[ E e^{i<u, \check{X}^{(n)}} > = \sum_{k=0}^{n} E \left( e^{i<u, \check{X}^{(n)}} > |N = k \right) P(N = k) \]

\[ = \sum_{k=0}^{n} E \left( e^{i<u, Z^{(n-k)}} > \right) P(N = k) \]

\[ = \sum_{k=0}^{n} E \left( e^{i<u, Z^{(n-k)}} > |N = k \right) P(N = k) \]

\[ = \sum_{k=0}^{n} E \left( e^{i<u, Z^{(n-N)}} > |N = k \right) P(N = k) \]

\[ = E e^{i<u, Z^{(n-N)}} > . \]

Thus,

\[ Z^{(n-N)} \overset{d}{=} \check{X}^{(n)}. \]

Now let \( LC_n \) be the length of the longest common subsequences of \( X^{(n)} \) and \( Y^{(n)} \), and let \( L_n(k) \) be the length of the longest common subsequences/subwords of \( Z^{(k)} \) and \( Y^{(n)} \). It follows from Lemma 1 that,

\[ LC_n \overset{d}{=} L_n(n - N), \quad (2.1) \]

and therefore,

\[ \text{Var} \, LC_n = \text{Var}(L_n(n - N)). \quad (2.2) \]

In order to prove the main result, we will also need the following result taken from [9].

**Lemma 2.** Let \( f : D \subset \mathbb{R} \rightarrow \mathbb{Z} \) satisfy a local reversed Lipschitz condition, i.e., let \( h \geq 0 \) and let \( f \) be such that for any \( i, j \in D \) with \( j \geq i + h \),

\[ f(j) - f(i) \geq c(j - i), \]

for some \( c > 0 \). Let \( T \) be a \( D \)-valued random variable with \( \mathbb{E}|f(T)|^2 < \infty \), then

\[ \text{Var} \, f(T) \geq \frac{c^2}{2} \left( \text{Var}(T) - h^2 \right). \]
Next, let
\[ O_n := \bigcap_{i,j \in I, j \geq i+h(n)} \{ L_n(j) - L_n(i) \geq K(j-i) \}, \tag{2.3} \]
where \( I = [np - \sqrt{np(1-p)}, np + \sqrt{np(1-p)}] \), \( K > 0 \) is a constant which does not depend on \( n \) \((K \leq 1/2m \) will do, see Lemma 10\), and where \( h(n) \) will also be made precise later. The event \( O_n \) can be viewed as the event where the map \( k \to L_n(k) \) locally satisfies a reversed Lipschitz condition.

In Section 3 we will prove

**Theorem 2.** For all \( n \geq 1 \),
\[ \mathbb{P}(O_n) \geq 1 - Ae^{-Bn} - ne^{-2K^2h(n)}, \tag{2.4} \]
where, \( K \) is given in Lemma 10, \( A = \max\{C_4, C_5, C_7\} \), and \( B = \min\{C_8\mu, C_6, C_8\} \), and these constants are given in (3.5), Lemma 6 and Lemma 8 respectively.

Now with the help of Theorem 2 we can provide the proof of our main result stated in Theorem 1.

**Proof of Theorem 1.** By (2.2), it is sufficient to prove the lower bound for \( \text{Var}(L_n(n-N)) \). First as in [9], with its notation,
\[ \text{Var}(U|V) \leq 2^2 \left( \mathbb{E}(U - \mathbb{E}U)^2 | V \right)/2 + \mathbb{E}(\mathbb{E}(U|V) - \mathbb{E}U)^2 | V \right)/2 \]
\[ \leq 2^2 \mathbb{E}(U - \mathbb{E}U)^2 | V, \tag{2.5} \]
and so, for any \( n \geq 1 \),
\[ \text{Var}(L_n(n-N)) \geq \frac{1}{2^2} \mathbb{E} \left( \text{Var}(L_n(n-N) | (L_n(n-k))_{0 \leq k \leq n}) \right) \]
\[ = \frac{1}{2^2} \int \Omega \text{Var}(L_n(n-N) | (L_n(n-k))_{0 \leq k \leq n}(\omega)) \mathbb{P}(d\omega) \]
\[ \geq \frac{1}{2^2} \int O_n \text{Var}(L_n(n-N) | (L_n(n-k))_{0 \leq k \leq n}(\omega)) \mathbb{P}(d\omega). \tag{2.6} \]

Since \( N \) is independent of \((L_n(n-k))_{0 \leq k \leq n} \), and from (2.7), for each \( \omega \in \Omega \),
\[ \text{Var}(L_n(n-N)|\{(L_n(n-k))_{0 \leq k \leq n}(\omega)\}) \]
\[ \geq \text{Var}(L_n(n-N)|\{(L_n(n-k))_{0 \leq k \leq n}(\omega), 1_{N \in I} = 1\}) \mathbb{P}(N \in I |{(L_n(n-k))_{0 \leq k \leq n}(\omega)}) \]
\[ = \text{Var}(L_n(n-N)|\{(L_n(n-k))_{0 \leq k \leq n}(\omega), 1_{N \in I} = 1\}) \mathbb{P}(N \in I), \tag{2.7} \]
where again,
\[ I = [np - \sqrt{np(1-p)}p, np + \sqrt{np(1-p)}]] \cdot \]

Again, for each \( \omega \in O_n \), from Lemma 2 and since \( N \) is independent of \((L_n(n-k))_{0 \leq k \leq n} \),
\[ \text{Var}(L_n(n-N)|\{(L_n(n-k))_{0 \leq k \leq n}(\omega), 1_{N \in I} = 1\}) \geq \frac{K^2}{8} (\text{Var}(N|1_{N \in I} = 1) - h(n)^2). \tag{2.8} \]

Now, (2.6), (2.7) and (2.8) give
\[ \text{Var}(L_n(n-N)) \geq \frac{K^2}{8} (\text{Var}(N|1_{N \in I} = 1) - h(n)^2) \mathbb{P}(N \in I) \mathbb{P}(O_n), \tag{2.9} \]
and it remains to estimate each one of the three terms on the right hand side of (2.9). By the Berry-Esséen inequality, for all \( n \geq 1 \),
\[ \left| \mathbb{P}(N \in I) - \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} x e^{-x^2} dx \right| \leq \frac{1}{\sqrt{np(1-p)}}. \tag{2.10} \]
Moreover, 
\[ \text{Var}(N|1_{N \in I} = 1) = \mathbb{E}((N - np + np - \mathbb{E}(N|1_{N \in I} = 1))^2 |1_{N \in I} = 1) \]
\[ \geq \left( \mathbb{E}((N - np)^2 |1_{N \in I} = 1)^{1/2} - |np - \mathbb{E}(N|1_{N \in I} = 1)| \right)^2, \quad (2.11) \]
and
\[ |\mathbb{E}(N|1_{N \in I} = 1) - np| \]
\[ = \sqrt{np(1 - p)} \left| \mathbb{E} \left( \frac{N - np}{\sqrt{np(1 - p)}} |1_{N \in I} = 1 \right) \right| \]
\[ = \sqrt{np(1 - p)} \left| F_n(1) - \Phi(1) + F_n(-1) - \Phi(-1) - \int_{-1}^1 (F_n(x) - \Phi(x))dx \right| \]
\[ \leq \sqrt{np(1 - p)} \frac{4 \max_{x \in [-1, 1]} |F_n(x) - \Phi(x)|}{\mathbb{P}(N \in I)} \]
\[ \leq \frac{2}{\int_{-1}^1 e^{-\frac{x^2}{2}} dx / \sqrt{2\pi} - 1 / \sqrt{np(1 - p)}}, \quad (2.12) \]
where \( F_n \) is the distribution functions of \( (N - np) / \sqrt{np(1 - p)} \), while \( \Phi \) is the standard normal one. Likewise,
\[ \mathbb{E}(|N - np|^2 |1_{N \in I} = 1) \]
\[ \geq (np(1 - p)) \frac{\int_{-1}^1 |x|^2 d\Phi(x) - 4 \max_{x \in [-1, 1]} |F_n(x) - \Phi(x)|}{\mathbb{P}(N \in I)} \]
\[ \geq (np(1 - p)) \left( \frac{\int_{-1}^1 e^{-\frac{x^2}{2}} dx - 2\sqrt{2\pi}/\sqrt{np(1 - p)}}{\int_{-1}^1 e^{-\frac{x^2}{2}} dx + 2\sqrt{2\pi}/\sqrt{np(1 - p)}} \right) \quad (2.13) \]
Next, using (2.11) – (2.13).
\[ \text{Var}(N|1_{N \in I} = 1) \]
\[ \geq \left( (np(1 - p))^{\frac{1}{2}} \left( \frac{\int_{-1}^1 e^{-\frac{x^2}{2}} dx + 2\sqrt{2\pi}/\sqrt{np(1 - p)}}{\int_{-1}^1 e^{-\frac{x^2}{2}} dx / \sqrt{2\pi} - 1 / \sqrt{np(1 - p)}} \right) \right)^2. \quad (2.14) \]
Finally, the estimates (2.11)–(2.14) combined with the estimate on \( \mathbb{P}(O_n) \) obtained in Theorem 2 give the lower bound in Theorem 1 whenever \( 2 \ln n/K^2 \leq h(n) \leq K_1 \sqrt{n} \), where the upper bound on \( h(n) \) stems from the requirement that the right hand side of (2.14) needs to be lower bounded and where \( K_1 \) is estimated in Section 4

\[ \square \]

3 Proof of Theorem 2

In this section, we prove the aforementioned theorem, therefore completing our proof of Theorem 1. Before doing so, we will need to state a few definitions and set some notations used throughout the rest of the paper:

The sequences \( Z^{(k)} \) and \( Y^{(n)} \) are said to have a common subsequence of length \( \ell \) if there exist increasing functions \( \pi: [1, \ell] \to [1, k] \) and \( \eta: [1, \ell] \to [1, n] \) such that
\[ Z_{\pi(i)}^k = Y_{\eta(i)}, \quad i = 1, 2, \ldots, \ell, \]
and \((\pi, \eta)\) is then called a pair of matching subsequences of \(Z^{(k)}\) and \(Y^{(n)}\). Also, throughout, \(M^k\) denotes the set of pairs of matching subsequences of \(Z^{(k)}\) and \(Y^{(n)}\) of maximal length.

Following the approach in [2], the proof of Theorem 2 is then divided into two cases, \(k < \beta n\) and \(k \geq \beta n\), where in each case \(\beta < 1/m\).

### 3.1 \(k < \beta n\) (\(\beta < 1/m\))

We begin with the simpler case \(k < \beta n\). In this situation, we show that with high probability all the letters of \(Z^{(k)}\) are matched with letters of \(Y^{(n)}\). Let

\[
E_k^{(n)} := \{L_n(k) = k\}.
\]

Then clearly, \(E_k^{(n)} \subset E_{k-1}^{(n)} \subset \cdots \subset E_1^{(n)}\), and so

\[
E^{(n)} := \bigcap_{k=1}^{\beta n} E_k^{(n)} = \{L_n(k+1) - L_n(k) = 1, \forall k < \beta n\}.
\]

**Lemma 3.** For \(\beta < 1/m\), there exists a constant \(C_1 = C_1(\beta, m) > 0\) such that,

\[
P(L_n(\beta n) = \beta n) \geq 1 - \exp(-C_1 n).
\]

**Proof.** We construct a pair of matching sequence \((\pi, \eta)\) for \(Z^{(k)} = Z_1^k Z_2^k \cdots Z_k^k\) and \(Y\) as follows,

\[
\begin{align*}
\pi(i) &= i, \\
\eta(i) &= \min\{\ell : \ell > \eta(i-1), Y_\ell = Z_i^k\},
\end{align*}
\]

for \(i \geq 1\), where we also set \(\eta(0) = 0\).

Thus, \(\eta(i)\) is the smallest index \(\ell\) such that \(Z_i^k \cdots Z_i^k\) is a subsequence of \(Y_1 Y_2 \cdots Y_\ell\). In this way, \(\eta(1), \eta(2), \eta(3), \cdots\) is a renewal process with geometrically distributed holding time, i.e., denoting the inter arrival times as

\[T_i = \eta(i) - \eta(i-1),\]

then \(\{T_i\}_{i \geq 1}\) is a sequence of independent geometric random variables with parameter \(1/m\), i.e.,

\[
P(T_i = t) = \frac{1}{m} \left(\frac{m-1}{m}\right)^{t-1}, \quad t = 1, 2, 3, \cdots.
\]

Thus, \(\mathbb{E}T_i = m\). Next,

\[
P(L_n(\beta n) = \beta n) \geq P\left(\sum_{i=1}^{\beta n} T_i < n\right) = 1 - P\left(\sum_{i=1}^{\beta n} \left(T_i - \frac{1}{\nu}\right) \geq 0\right),
\]

and from the independence of the \(\{T_i\}_{i \geq 1}\),

\[
P\left(\sum_{i=1}^{\beta n} \left(T_i - \frac{1}{\nu}\right) \geq 0\right) \leq \inf_{s > 0} E\left(e^s \sum_{i=1}^{\beta n} (T_i - \frac{1}{\nu})\right) = \inf_{s > 0} \left(E e^{s(T_0 - 1/\nu)}\right)^{\beta n} = \inf_{s > 0} e^{-ns} \left(\frac{e^s}{m - (m-1)e^s}\right)^{\beta n}.
\]

This last term is minimized at

\[
s = \ln \frac{m(1 - \nu)}{m - (m-1)e^s},
\]
thus,
\[
\inf_{s > 0} e^{-s} \left( \frac{e^{s}}{m - (m - 1)e^{s}} \right)^{\nu} = \frac{(1 - \nu)^{\nu-1}}{m(m - 1)^{\nu-1} \nu^{\nu}},
\]
which is increasing in \( \nu \) for \( \nu \in (0, 1 - 1/m) \). Thus,
\[
\frac{(1 - \nu)^{\nu-1}}{m(m - 1)^{\nu-1} \nu^{\nu}} \begin{cases} > 1 & \text{when } \nu \in (1/m, 1 - 1/m) \\ = 1 & \text{when } \nu = 1/m \\ < 1 & \text{when } \nu < 1/m. \end{cases}
\]

Since \( \nu < 1/m \), by taking \( C_1 = \ln \left( m(m - 1)^{\nu-1} \nu^{\nu} / (1 - \nu)^{\nu-1} \right) \), we have
\[
\mathbb{P}(E_{\nu n}^{(n)}) = \mathbb{P}(L_n(\nu m) = \nu n) \geq 1 - \exp(-C_1 n).
\]

Therefore, Lemma 3 asserts that
\[
\mathbb{P}(E^{(n)}) = \mathbb{P}(E_{\nu n}^{(n)}) \geq 1 - \exp(-C_1 n).
\]

### 3.2 \( k \geq \nu n \) (\( \nu < 1/m \))

To continue, we introduce some more definitions and notations of use throughout the section.

(i) Let \( \preceq \) denote the partial order between two increasing functions \( \pi_1, \pi_2 : [1, \ell] \to \mathbb{N} \), i.e., \( \pi_1 \preceq \pi_2 \) if for every \( i \in [1, \ell] \), \( \pi_1(i) \leq \pi_2(i) \). Further \( (\pi_1, \eta_1) \preceq (\pi_2, \eta_2) \) is short for \( \pi_1 \preceq \pi_2 \) and \( \eta_1 \leq \eta_2 \).

(ii) Let \( M^k_{min} \subset M^k \) be the set of \( (\pi, \eta) \in M^k \) which are minimal for the relation \( \preceq \), i.e., such that for \( (\pi_1, \eta_1) \in M^k_{min} \) and \( (\pi_2, \eta_2) \in M^k \), if \( (\pi_1, \eta_1) \preceq (\pi_2, \eta_2) \) then \( (\pi_1, \eta_1) = (\pi_2, \eta_2) \).

(iii) If \( (\pi, \eta) \) is a pair of matching subsequences of \( Z^{(k)} \) and \( Y^{(n)} \) of length \( \ell \), a match of \( (\pi, \eta) \) is then defined to be the quadruple
\[
(\pi(i), \pi(i + 1), \eta(i), \eta(i + 1)).
\]
Moreover, if \( \eta(i) + 2 \leq \eta(i + 1) \), the match is said to be non-empty. Therefore, for a non-empty match, there exists \( j \), such that \( \eta(i) < j < \eta(i + 1) \) and \( Y_j = \alpha \) for some \( \alpha \in \mathcal{A} \setminus \{\alpha_{m+1}\} \). In that case, the match is said to contain an \( \alpha \), and \( Y_j \) is called an unmatched letter of the match \( (\pi(i), \pi(i + 1), \eta(i), \eta(i + 1)) \).

(iv) The sequence \( Y^{(n)} \) can be uniquely divided into \( d \) compartments \( [j_1, j_2 - 1], [j_2, j_3 - 1], \ldots, [j_d, n] \), where \( 1 = j_1 < j_2 < \cdots < j_d \leq n \) are determined by the following recursive relations:
\[
\begin{align*}
j_1 &= 1 \\
j_i &= \min(n + 1, \{s \in [j_{i-1} + 1, n] : Y_{j_{i-1}+1}Y_{j_{i-1}+2}\cdots Y_{s} \text{ contains } m \text{ distinct letters}\}),
\end{align*}
\]
and \( d = \max\{i : j_i \leq n\} \).

To get a lower bound on the probability that the length of the longest common subsequence increases by one, we recall the construction of \( Z^{(k)} \) and note that there are \( (k - 1) \) possible positions for the letter \( U_{k+1} \) to be inserted. Therefore, \( U_{k+1} \) falls into a non-empty match with probability at least \( (\text{number of nonempty matches of } (\pi, \eta))/k \geq (\text{number of nonempty matches of } (\pi, \eta))/k \).

For each non-empty match, there is at least one unmatched letter, and the probability that \( U_{k+1} \) takes the same value as the unmatched letter is \( 1/m \), resulting in the following lower bound for \( (\pi, \eta) \in M^k \):
\[
\mathbb{P} \left( L_n(k + 1) - L_n(k) = 1 | Z^{(k)}, Y^{(n)} \right) \geq \frac{1}{m} \frac{\text{number of nonempty matches of } (\pi, \eta)}{k}. \tag{3.1}
\]

Therefore, a good estimate on the number of nonempty matches of \( (\pi, \eta) \) will provide a lower bound on the probability that \( LC_n \) increases by one.
Next we give the main ideas behind the proof that, with high probability, the map \( k \to L(k) \) is linearly increasing on \([mn, n]\). We use the letter-insertion scheme, described above, to prove that the random map \( k \to L(k) \) typically has positive drift \( \lambda \) (which will be determined later in Lemma \( \text{L9} \)). To do so, let

\[ F_k^{(n)} := \{ (\pi, \eta) \in M_{\min}^k \text{ such that the number of nonempty matches of } (\pi, \eta) \text{ is at least } \lambda n \}, \quad (3.2) \]

and let

\[ F^{(n)} := \bigcap_{k=m}^n F_k^{(n)}. \]

When \( F^{(n)} \) holds, every pair of \( (\pi, \eta) \in M_{\min}^k \) has at least \( \lambda n \) nonempty matches. Hence the number of non-empty matches divided by \( k \) is larger than or equal to \( \lambda n/k \). It follows from (3.1) that when \( F_k^{(n)} \) holds,

\[ P(L_n(k + 1) - L_n(k) = 1 | Z^{(k)}, Y^{(n)}) \geq \frac{1}{m} \frac{\lambda n}{k} \geq \frac{\lambda}{m} > 0. \quad (3.3) \]

The inequality (3.3) implies that when \( F^{(n)} \) holds, the map \( k \to L_n(k) \) has drift at least \( \lambda/m \) for \( k \in [mn, n] \). In other words, whenever \( F^{(n)} \) holds, with high probability \( k \to L_n(k) \) has positive slope on \([mn, n]\).

It remains to show that, by concentration, \( F^{(n)} \) holds with high probability, and this is proved by contradiction. Indeed if all the matches of \((\pi, \eta) \in M^k \) were empty, then the following two conditions would hold:

1. \( (\eta(1), \eta(2), \eta(3), \ldots, \eta(\ell)) = (\eta(1), \eta(1) + 1, \eta(1) + 2, \ldots, \eta(1) + \ell - 1) \) where \( \ell \) is the length of the LCS of \( Z^{(k)} \) and \( Y^{(n)} \), i.e., \( \ell = L_n(k) \).

2. The sequence \( Y_{\eta(1)} Y_{\eta(2)} \cdots Y_{\eta(\ell)} = Y_{\eta(1)} Y_{\eta(1) + 1} \cdots Y_{\eta(1) + \ell - 1} \)

would be a subsequence of \( Z^{k}_{\pi(1)} Z^{k}_{\pi(1) + 1} \cdots Z^{k}_{\pi(\ell)}. \)

Above, we have two independent sequences of i.i.d. uniform random variables with parameter \( 1/m \), where one is contained in the other as a subsequence. Thus, the longer one must approximately be at least \( m \) times as long as the shorter one, hence \( k \) is approximately at least \( m \) times as long as \( \ell = L_n(k) \). As a result, the ratio \( L_n(k)/k \) is to be at most \( 1/m \), which is very unlikely (Lemma \( \text{L6} \)), leading to contradiction.

From the previous arguments, it follows that with high probability any \( (\pi, \eta) \in M_{\min}^k \) contains a non-vanishing proportion \( \epsilon > 0 \) of unmatched letters, hence \( \eta(L_n(k)) - L_n(k) / \eta(L_n(k)) \geq \epsilon \), where \( \eta(L_n(k)) \) is the index of the last matching letter in \( Y^{(n)} \) of the match \((\pi, \eta) \). We then show that this proportion \( \epsilon \) of unmatched letters generates sufficiently many non-empty matches, i.e., that the unmatched letters should not be concentrated on a too small number of matches.

To prove that there are more than \( \lambda n \) nonempty matches, the following two arguments are used:

1. Any \((\pi, \eta) \in M_{\min}^k \) is such that every match of \((\pi, \eta) \) contains unmatched letters from at most one compartment of \( Y^{(n)} \).

2. There exists a \( D > 0 \), not depending on \( n \), such that, with high probability, the total number of integer points contained in the compartments of \( Y^{(n)} \) of length larger than \( D \), is small.

Henceforth, for \((\pi, \eta) \in M_{\min}^k \), the majority of unmatched letters are at most \( D \) per match, ensuring that a proportion \( \epsilon \) of unmatched letters implies a proportion of at least \( \epsilon/D \) non-empty matches.

Let us return to the proof, and let \( L_\ell(k) \) denote the length of the LCS of \( Z^{(k)} \) and \( Y^{(\ell)} = Y_1 \cdots Y_\ell \). In order for \( Y^{(\ell)} \) to be contained in \( Z^{(k)} \), \( k \) needs to be approximately \( m \) times as long as \( \ell \), and then, \( L_\ell(k) = \ell \). Therefore, if \( k = m\ell(1 - \delta) \), for some \( \delta = \delta(\epsilon) > 0 \) not depending on \( \ell \), then it is extremely unlikely that \( Y^{(\ell)} \) is a subsequence of \( Z^{(k)} \), as shown in the forthcoming lemma.
Lemma 4. For any $0 < \delta < (m-1)/m$ and $\ell \geq 1$, we have
\[
P(L_\ell(m\ell(1-\delta)) = \ell) \leq e^{-C_2 \delta^2 \ell},
\] (3.4)
where $C_2 = m/2(m-1)$.

Proof. The proof is similar to the proof of Lemma 3 and some of its notation is used.

First let $\tilde{X} := X^{(\infty)}$, be the (infinite) subword of $X$ with $\alpha_{m+1}$ removed, and therefore each $\tilde{X}^{(n)}$ is a subword of $\tilde{X}$. Next, construct a pair of matching sequence $(\pi, \eta)$ for $X$ and $Y^{(\ell)}$ as follows:
\[
\pi(0) = 0, \quad \text{and for } i \geq 1, \quad \left\{ \begin{array}{l}
\pi(i) = \min\{ j : j > \pi(i-1), \tilde{X}_j = Y_i \} \\
\eta(i) = i.
\end{array} \right.
\]
Thus, $\pi(i)$ is the smallest index $j$ such that $Y_1Y_2 \cdots Y_i$ is a subsequence of $\tilde{X}_1 \cdots \tilde{X}_j$. In this way, $\pi(1), \pi(2), \pi(3), \cdots$ is a renewal process with geometrically distributed holding time, i.e., denoting the interarrival times as $T_i = \pi(i) - \pi(i-1)$, then $\{T_i\}_{i \geq 1}$ is a sequence of independent geometric random variables with parameter $1/m$, i.e.,
\[
P(T_i = t) = \frac{1}{m} \left( \frac{m-1}{m} \right)^{t-1}, \quad t = 1, 2, 3, \ldots.
\]
Thus, $ET_i = m$. Then by Lemma 1 and for $0 < \delta < 1$, we have
\[
P(L_\ell(m\ell(1-\delta)) = \ell) = \P \left( \sum_{i=1}^\ell T_i \leq m\ell(1-\delta) \right) = \P \left( \sum_{i=1}^\ell (m(1-\delta) - T_i) \geq 0 \right) \leq \left( \inf_{s>0} e^{s(m(1-\delta)-T_1)} \right) ^\ell = \left( \inf_{s>0} \frac{e^{sm(1-\delta)}}{me^s - (m-1)} \right) ^\ell.
\]
This last term is minimized at
\[
s = \ln \left( 1 + \frac{\delta}{m(1-\delta) - 1} \right),
\]
thus setting,
\[
w := \inf_{s>0} \frac{e^{sm(1-\delta)}}{me^s - (m-1)} = \frac{(m(1-\delta) - 1) \left( 1 + \frac{\delta}{m(1-\delta) - 1} \right)^{m(1-\delta)}}{m - 1},
\]
it follows that,
\[
P \left( \sum_{i=1}^\ell (m(1-\delta) - T_i) \geq 0 \right) \leq e^{(\ln w)\ell}.
\]
Now, the Taylor expansion of $\ln w$ with Lagrange remainder gives
\[
\ln w = -\frac{m}{2(m-1)} \delta^2 + \frac{1}{6} \left( \frac{m}{(1-\xi)^2} - \frac{m^3}{(m(1-\xi) - 1)^2} \right) \delta^3 < -\frac{m}{2(m-1)} \delta^2,
\]
where $0 < \xi < \delta$. Letting $C_2 = m/2(m-1)$ finishes the proof.

Lemma 4 further entails, as shown next, that for any $0 < \epsilon < 1$ there exists $\delta(\epsilon) > 0$, small, such that $L_\ell(m\ell(1-\delta(\epsilon))) \geq \ell(1-\epsilon)$ is also very unlikely.

Lemma 5. For any $0 < \epsilon < 1$ and all $\ell \geq 1$, there exists $\delta(\epsilon) > 0$, with $\lim_{\epsilon \to 0} \delta(\epsilon) \to 0$, such that
\[
P((C^{(n)}_\ell(\epsilon))^\epsilon) \leq e^{-C_3 \epsilon},
\]
where $C_3 = C_2$. 

10
where \( G_\ell^{(n)}(\epsilon) = \{ L_\ell(m\ell(1 - \delta(\epsilon))) < \ell(1 - \epsilon) \} \), and where \( C_3 := (\delta(\epsilon) - \epsilon)^2C_2/2 \). Therefore, letting

\[
G^{(n)}(\epsilon) := \bigcap_{\ell=\nu n} G_\ell^{(n)}(\epsilon),
\]

it follows that,

\[
\mathbb{P}(G^{(n)}(\epsilon)) \geq 1 - \sum_{k=\nu n}^{n} e^{-C_3 k} \geq 1 - \frac{1}{1 - e^{-C_3}} e^{-C_3 \nu n} = 1 - C_4 e^{-C_3 \nu n}, \tag{3.5}
\]

where \( C_4 = 1/(1 - e^{-C_3}) \).

**Proof.** Let \( S \subset \{ 1, 2, \ldots, \ell \} \) have cardinality \((1 - \epsilon)\ell\). Clearly, there are \( (\ell(1 - \epsilon))/\ell = \ell(1 - \epsilon) \) such subsets \( S \). Now fixing the values of \( Y^{(n)} \) at the indices belonging to \( S \), there are \( m^{\ell(1 - \epsilon)} \) such \( Y^{(n)} \) agreeing on \( S \). Therefore,

\[
\mathbb{P}(\{(G^{(n)}_\ell(\epsilon))^c\}) \leq m^{\ell(1 - \epsilon)} \mathbb{P}(L_{\ell(1 - \epsilon)}(m\ell(1 - \delta(\epsilon))) = \ell(1 - \epsilon)).
\]

From (3.4),

\[
\mathbb{P}(L_{\ell(1 - \epsilon)}(m\ell(1 - \delta(\epsilon))) = \ell(1 - \epsilon)) = \mathbb{P}
\left(L_{\ell(1 - \epsilon)} \left(m \left(1 - \frac{\delta(\epsilon) - \epsilon}{1 - \epsilon}\right) \right) = \ell(1 - \epsilon) \right)
\leq e^{-C_2 \left(\frac{\delta(\epsilon) - \epsilon}{1 - \epsilon}\right)^2 \ell} \leq e^{-C_2 \delta(\epsilon - \epsilon)^2 \ell}.
\]

Collecting the above estimates,

\[
\mathbb{P}(\{(G^{(n)}_\ell(\epsilon))^c\}) \leq m^{\ell(1 - \epsilon) \ell} e^{-C_2 \delta(\epsilon - \epsilon)^2 \ell}. \tag{3.6}
\]

Since

\[
\ell^2 = (\ell\epsilon + (1 - \epsilon)\ell)^2 \geq \ell^2 \left(\frac{\ell}{(1 - \epsilon)\ell}\right) = \ell^2 \left(\frac{1}{\epsilon(1 - \epsilon)}\right),
\]

then

\[
\left(\frac{\ell}{(1 - \epsilon)\ell}\right) \leq \left(\frac{1}{\epsilon(1 - \epsilon)}\right)^{\ell}.
\]

Therefore, (3.6) becomes

\[
\mathbb{P}(\{(G^{(n)}_\ell(\epsilon))^c\}) \leq e^{\ell \ln m - \ln \epsilon - (1 - \epsilon) \ln(1 - \epsilon) - C_2 \delta(\epsilon - \epsilon)^2} \ell,
\]

and it is enough to choose

\[
\delta(\epsilon) = \epsilon + \sqrt{\frac{2}{C_2} \left(\epsilon(\ln m - \ln \epsilon) - (1 - \epsilon) \ln(1 - \epsilon)\right)}, \tag{3.7}
\]

and obtain the stated result.

Lemma 6 and Lemma 7 presented next, formalize our contradictory argument asserted above. To show that it is very unlikely that “the ratio \( L_n(k)/k \) is at most \( 1/m^2 \),” note, at first, that for \( n \geq 2 \),

\[
\mathbb{E}L_n(n) > \mathbb{E} \sum_{i=1}^{n} \mathbb{1}_{Y_i = Z_i} \geq n \mathbb{P}(Y_1 = Z_1^n) = \frac{n}{m}. \tag{3.8}
\]

Specifically, when \( n = 2 \), see [3],

\[
\mathbb{E}L_2(2)/2 = \frac{4m^2 - 5m + 3}{2m^3}. \tag{3.9}
\]
Now, choose $\xi_m$ such that

$$1/m < \xi_m < \mathbb{E} L_2(2)/2,$$

and let us show that very likely $L_n(k)/k$ is larger than $\xi_m$. To do so, let

$$H_k^{(n)} := \{L_n(k) \geq \xi_m k\},$$

and

$$H^{(n)} := \bigcap_{k=\nu n}^n H_k^{(n)}.$$

**Lemma 6.** There exist constants $C_5, C_6 > 0$, such that

$$\mathbb{P}(H^{(n)}) \geq 1 - C_5 e^{-C_6 n}.$$  

**Proof.** Divide the sequences $Z^{(k)}$ and $Y^{(n)}$ into subsequences of length 2, as given in the previous lemma. Then, by superadditivity, $L_k(2) \geq \sum_{i=1}^{k/2} \hat{\mathcal{L}}_i$, where $\hat{\mathcal{L}}_i$ is the length of the longest common subsequence between $Y_{2(i-1)+1}^2$ and $Z_{2(i-1)+1}^2$. Clearly, by the i.i.d. assumptions, $\mathbb{E}(\hat{\mathcal{L}}_i) = \mathbb{E}(L_2(2))$ is constant. Hence for $\tau > 0$,

$$\mathbb{P} \left( \sum_{i=1}^{k/2} \hat{\mathcal{L}}_i < k \left( \frac{\mathbb{E}(\hat{\mathcal{L}}_i) - \tau}{2} \right) \right) \leq \left( \inf_{\tau \leq 0} \mathbb{E} \left( e^{s(\mathbb{E}(L_2(2)) - \tau)} \right) \right)^{\frac{k}{2}}. \quad (3.12)$$

Now let $p(s, \tau) := \mathbb{E} \left( e^{s(\mathbb{E}(L_2(2)) - \tau)} \right)$, it is easy to see that $p(s, \tau)$ is smooth in $s$, and that

$$\begin{cases} p(0, \tau) = 1, \\ \partial_s p(s, \tau) \big|_{s=0} = \tau > 0, \end{cases}$$

for every $\tau > 0$. Hence,

$$\inf_{s < 0} p(s, \tau) < e^{-c(\tau)}, \quad (3.13)$$

for a suitable $c(\tau) > 0$. Thus,

$$\mathbb{P}( (H_k^{(n)})^c ) \leq \mathbb{P}(L_k(2) < \xi_m k) \leq \mathbb{P} \left( \sum_{i=1}^{k/2} \hat{\mathcal{L}}_i < k \left( \frac{\mathbb{E}(\hat{\mathcal{L}}_i) - \tau}{2} \right) \right) \leq e^{-c(\tau)k/2}.$$

Now, let $\tau = \tau_m := \mathbb{E}(L_2(2)) - 2\xi_m$, let $\xi_m = 11/10m$, and so

$$p(s, \tau_m) = \frac{e^{-11s/5m}(me^{2s} + (4m^2 - 7m + 3)e^s + m^3 - 4m^2 + 6m - 3)}{m^3}.$$  

Since $\inf_{s < 0} p(s, \tau_m) < e^{-1/1000m}$, one can choose $c(\tau_m) = 1/1000m$. Hence,

$$\mathbb{P}( (H^{(n)})^c ) \leq \sum_{k=\nu n}^n e^{-c(\tau_m)k/2} = \frac{e^{c(\tau_m)(1/\nu)/2} - e^{-c(\tau_m)\nu/2}}{e^{c(\tau_m)/2} - 1} \leq \frac{e^{c(\tau_m)/2} - 1}{e^{c(\tau_m)/2} - 1} e^{c(\tau_m)(-\nu)/2}.$$

Choosing $C_5 = \frac{e^{c(\tau_m)/2}}{e^{c(\tau_m)/2} - 1}$, and $C_6 = c(\tau_m)(\nu)/2$, we have,

$$\mathbb{P}(H^{(n)}) \geq 1 - C_5 e^{-C_6 n}.$$  

$\blacksquare$
We now finish our argument showing that, with high probability, any \((\pi, \eta) \in M_{\min}^k\) contains a non-vanishing proportion \(\epsilon > 0\) of unmatched letters. To do so, let

\[ I_k^{(n)} := \{ L_n(k) \leq (1 - \epsilon) \eta(L_n(k)), \text{ for } (\pi, \eta) \in M_{\min}^k \}, \]

be the event that any pair of matching subsequences \((\pi, \eta) \in M_{\min}^k\) has a proportion at least \(\epsilon\) of unmatched letters, and let

\[ I^{(n)} := \bigcap_{k=\nu n}^n I_k^{(n)}. \]

Above, \(\eta(L_n(k)) - L_n(k)\) is the number of unmatched letters, since \(\eta(L_n(k))\) is the position of the last matched letter, while \(L_n(k)\) is the number of matched letters.

**Lemma 7.** Let \(\epsilon > 0\) be small enough such that \(\delta(\epsilon)\), as given in (3.7), satisfies

\[ \frac{1}{1 - \delta(\epsilon)} < \xi m, \tag{3.14} \]

where \(\xi_m\) is as in (3.10). Then, for all \(k \geq \nu n\),

\[ G^{(n)}(\epsilon) \cap H_k^{(n)} \subset I_k^{(n)}, \tag{3.15} \]

and thus

\[ G^{(n)}(\epsilon) \cap H^{(n)} \subset I^{(n)}. \tag{3.16} \]

**Proof.** Let \(k \in [\nu n, n]\). In order to prove (3.15), we show that if \(I_k^{(n)}\) does not hold while \(G^{(n)}(\epsilon)\) does hold, then \(H_k^{(n)}\) does not hold either. Let \((\pi, \eta) \in M_{\min}^k\). If \(I_k^{(n)}\) does not hold, then the proportion of unmatched letters of \((\pi, \eta)\) is smaller than \(\epsilon\), i.e.,

\[ \frac{L_\ell(k)}{\ell} \geq 1 - \epsilon, \]

where \(\ell := \eta(L_n(k))\). (Note that \(L_\ell(k) = L_n(k)\), since \((\pi, \eta)\) is of maximal length.) Therefore,

\[ L_\ell(k) \geq \ell(1 - \epsilon). \tag{3.17} \]

Now, when \(G^{(n)}(\epsilon)\) holds, then

\[ L_\ell(ml(1 - \delta(\epsilon))) < \ell(1 - \epsilon). \tag{3.18} \]

Comparing (3.17) with (3.18) and noting that the (random) map \(x \mapsto L_\ell(x)\) is increasing, yield

\[ k \geq ml(1 - \delta(\epsilon)), \]

and thus

\[ k \geq m\eta(L_n(k))(1 - \delta(\epsilon)) \geq mL_n(k)(1 - \delta(\epsilon)). \]

Hence, from (3.14),

\[ \frac{L_n(k)}{k} \leq \frac{1}{m(1 - \delta(\epsilon))} < \xi m, \]

which implies that \(H_k^{(n)}\) cannot hold.

As an example, when \(\epsilon \leq e^{-9}/(1 + \ln m)\),

\[ \delta(\epsilon) = \epsilon + \sqrt{\frac{2}{C_2} (\epsilon(\ln m - \ln \epsilon) - (1 - \epsilon) \ln(1 - \epsilon))} \leq \epsilon + 2\sqrt{(1 + \ln m - \ln \epsilon) \epsilon} \leq e^{-9} + 2\sqrt{10}e^{-9} < \frac{1}{11}, \]
and therefore,
\[
\frac{1}{1 - \delta(\epsilon)} < \frac{10}{11} = \xi m.
\]

In order to estimate the event \( F^{(n)} \), we need to show that the unmatched letters of \( Y^{(n)} \) do not concentrate in a small number of matches of \((\pi, \eta) \in M_k^{\text{min}} \). From the minimality of \( M_k^{\text{min}} \), the unmatched letters of a match of \((\pi, \eta) \in M_k^{\text{min}} \) contain at most one compartment.

Let \( N^D \) be the total number of letters in the sequence \( Y^{(n)} \) contained in a compartment of length at least \( D \), and let,
\[
J^{(n)} := \{ N^D \leq \xi_m \epsilon \nu n / 2 \},
\]
where again \( \xi_m \) is given via (3.10).

**Lemma 8.** For any \( 0 < \epsilon < 1 \), there exist a positive integer \( D \), and positive constants \( C_7 \) and \( C_8 \) depending on \( D \), such that
\[
\mathbb{P}(J^{(n)}) \geq 1 - C_7 e^{-C_8 n},
\]
(3.19)

**Proof.** Let \( \tilde{N}^D \) be the number of integers \( s \in [0, n - D] \) such that
\[
(Y_s, Y_{s+1}, \ldots, Y_{s+D-1}) \text{ belongs to a compartment.}
\]
(3.20)

It is easy to check that
\[
N^D \leq D \tilde{N}^D.
\]
(3.21)

Let now \( \tilde{Y}_s, s \in [0, n - D] \), be equal to 1 if and only if (3.20) holds, and 0 otherwise. Clearly,
\[
\sum_{s=1}^{n} \tilde{Y}_s = \tilde{N}^D.
\]
(3.22)

To estimate the sum (3.22), decompose it into \( D \) subsums of i.i.d. random variables \( \Sigma_1, \Sigma_2, \ldots, \Sigma_D \) where
\[
\Sigma_i = \sum_{s=1}^{n} \tilde{Y}_s,
\]
so that
\[
\tilde{N}^D = \sum_{i=1}^{D} \Sigma_i.
\]
(3.23)

Then, from (3.21)
\[
\mathbb{P}
\left( N^D > \frac{\xi_m \epsilon \nu n}{2} \right) \leq \mathbb{P}
\left( \tilde{N}^D > \frac{\xi_m \epsilon \nu n}{2D} \right) \leq D \mathbb{P}
\left( \Sigma_1 > \frac{\xi_m \epsilon \nu n}{2D^2} \right),
\]
(3.24)

since in (3.23) at least one of the summands has to be larger than \( n \xi_m \epsilon \nu / 2D^2 \). Now, the \( \tilde{Y}_s \) appearing in the subsum \( \Sigma_1 \) are i.i.d. Bernoulli random variables with
\[
\mathbb{P}(\tilde{Y}_s = 1) \leq m \left( \frac{m - 1}{m} \right)^D.
\]

Therefore,
\[
\mathbb{P}
\left( \Sigma_1 > (E \tilde{Y}_s + \delta)^n \right) \leq e^{-c(\delta) \delta},
\]
(3.25)

with \( c(\delta) > 0 \) for \( \delta > 0 \). Take \( \delta = \mathbb{P}(\tilde{Y}_s = 0) = 1 - \mathbb{P}(\tilde{Y}_s = 1) \), then \( c(\delta) = -\ln \mathbb{P}(\tilde{Y}_s = 1) \). Thus it is enough to choose \( D \) such that
\[
2D m \left( (m - 1) / m \right)^D < \xi_m \epsilon \nu.
\]
(3.26)

Let \( x = (m - 1) / m, y = \xi_m \epsilon \nu / 2m \), we next show that,
\[
D = \frac{1}{y(\ln x)^2} = \frac{40c^y m^3 (1 + \ln m)}{11 \ln^2 \left( \frac{m-1}{m} \right)},
\]
(3.27)
does satisfy (3.20), or equivalently that \( Dx^D < y \). With the choice in (3.27), \( Dx^D < y \) is equivalent to
\[
2y \ln x \ln y + 2y(\ln x)^2 < 1,
\]
which is true since
\[
2y \ln x \ln y + 2y(\ln x)^2 = 2(-\ln x)(-y \ln y) + 2y(\ln x)^2 \leq 2 \ln 2 \cdot 9e^{-9} + 2(\ln 2)^2e^{-9} < 1.
\]
Choosing \( C_7 = D \) and \( C_8 = c(\delta)/D \), we have
\[
\mathbb{P}(J^{(n)}) \geq 1 - C_7e^{-C_8n}.
\]

We can now find a suitable \( \lambda \) such that when \( H^{(n)}, I^{(n)} \) and \( J^{(n)} \) all hold, then \( F^{(n)} \) (which depends on \( \lambda \), see (3.24)) also holds.

**Lemma 9.** Let \( \epsilon > 0 \) be as in Lemma 7 let \( D \) be such that \( 2Dm ((m - 1)/m)^D < \xi_m \nu \epsilon \), and let
\[
\lambda = \frac{\xi_m \nu \epsilon}{2 - D - 1}.
\]
Then, for \( k \geq \nu n \),
\[
H^{(n)} \cap J^{(n)} \cap I_k^{(n)} \subset F_k^{(n)},
\]
and thus
\[
H^{(n)} \cap J^{(n)} \cap I^{(n)} \subset F^{(n)}.
\]

**Proof.** We prove (3.28), from which (3.29) immediately follows. On \( I_k^{(n)} \), each \((\pi, \eta) \in M_{\min}^k\) has at least \( \epsilon \mu (L_n(k)) \) unmatched letters. But,
\[
\eta(L_n(k)) \geq L_n(k).
\]
When \( H^{(n)} \) holds,
\[
L_n(k) \geq \xi_m k.
\]
Since \( k \geq \nu n \), (3.30) and (3.31), together imply that the number of unmatched letters of \((\pi, \eta) \in M_{\min}^k\) is at least \( \epsilon \xi_m \nu n \). By \( J^{(n)} \), there are at most \( \xi_m \nu n / 2 \) letters contained in compartments of length at least \( D \). Thus, there are at least \( \xi_m \nu n / 2 \) letters contained in compartments of length less than \( D \). But, every match of \((\pi, \eta) \in M_{\min}^k\) contains unmatched letters from only one compartment, and as such every match can contain at most \( D - 1 \) unmatched letters from compartments of length less than \( D \). Therefore, these \( \epsilon \xi_m \nu n / 2 \) unmatched letters which are not in \( N^D \), must fill at least \( \epsilon \xi_m \nu n /(2D - 2) \) matches of \((\pi, \eta) \in M_{\min}^k\). Hence, \((\pi, \eta) \in M_{\min}^k\) has at least \( \epsilon \xi_m \nu n / (2D - 2) \) non-empty matches.

Combining Lemma 7 and Lemma 9 gives,
\[
\mathbb{P}(F_n^c) \leq \mathbb{P}(G_n^c(\epsilon)^c) + \mathbb{P}(H_n^c)^c + \mathbb{P}(J_n^c)^c,
\]
which via (3.5), (3.11), and (3.19) entails
\[
\mathbb{P}(F_n) \geq 1 - C_4 e^{-C_5 \nu n} - C_5 e^{-C_6 n} - C_7 e^{-C_8 n}.
\]
Next, recalling the definition of \( O_n \) in (2.23), observe that
\[
\mathbb{P}(O_n^c) \leq \mathbb{P}(O_n^c \cap E_n \cap F_n) + \mathbb{P}((F_n^c)^c) + \mathbb{P}((E_n^c)^c).
\]

The next result estimates the first probability, on the above right hand side, and, therefore, completes the proof of Theorem 2.

**Lemma 10.** Let \( K \leq 1/2m \), then
\[
\mathbb{P}(O_n \cap E_n \cap F_n) \leq ne^{-2K^3 h(n)}.
\]
Proof. Let $\lambda$ given as in Lemma 9 be at most 1, and let $K := \lambda/2m$, so that $K \leq 1/2m$. Let

$$\Delta(k) := \begin{cases} L_n(k + 1) - L_n(k) & \text{when } F_k^{(n)} \text{ holds,} \\ 1 & \text{otherwise.} \end{cases}$$

From (3.30), it follows that:

$$\mathbb{P}(\Delta(k) = 1 | \sigma_k) \geq \lambda/m, \quad (3.32)$$

where $\sigma_k$ denote the $\sigma$-field generated by the $Z_i^k$ and $Y_j$, namely,

$$\sigma(Z_i^k, Y_j | i \leq k, j \leq n).$$

Moreover, $\Delta(k)$ is equal to zero or one (since $L_n(\cdot)$ is non-decreasing on $\mathbb{N}$) and is also $\sigma_k$-measurable. Let

$$\tilde{L}_n(k) = \begin{cases} L_n(\nu n) + \sum_{i=\nu n}^{k-1} \Delta(i) & \text{for } k \in [\nu n, n], \\ L_n(k) & \text{for } k \in [0, \nu n]. \end{cases}$$

Note that when $F^{(n)}$ holds, then

$$L(k) = \tilde{L}(k), \quad (3.33)$$

for all $k \in [0, n - 1]$. Define

$$\tilde{O}_{i,j}^{(n)} = \{ \tilde{L}_n(j) - \tilde{L}_n(i) \geq K(j - i) \},$$

and

$$\tilde{O}_n = \bigcap_{i,j \in \{\nu n, n\}} \tilde{O}_{i,j}^{(n)}.$$

When $E^{(n)}$ holds, then $L_n(k)$ has a slope of one on the domain $[0, \nu n]$. Therefore, since $K \leq 1/2m$, the slope condition of $O_n$ holds on the domain $[0, \nu n] \cap I$. When $F^{(n)}$ holds, then $L_n(k)$ and $\tilde{L}_n(k)$ are equal. Therefore, when $F^{(n)}$ and $\tilde{O}_n$ both hold, then the slope condition of $O_n$ is verified on the domain $[\nu n, n] \cap I$. Hence,

$$E^{(n)} \cap F^{(n)} \cap \tilde{O}_n = E^{(n)} \cap F^{(n)} \cap O_n, \quad (3.34)$$

and thus

$$\mathbb{P}(O_n^c \cap E^{(n)} \cap F^{(n)}) = \mathbb{P}(\tilde{O}_n^c \cap E^{(n)} \cap F^{(n)}) \leq \mathbb{P}(\tilde{O}_n^c).$$

It only remains to estimate $\mathbb{P}(\tilde{O}_n^c)$. First,

$$\mathbb{P}(\tilde{O}_n^c) \leq \sum_{i,j \in \{\nu n, n\}} \mathbb{P}((\tilde{O}_{i,j}^{(n)})^c). \quad (3.35)$$

Then, from Hoeffding’s exponential inequality, for any $t > 0$,

$$\mathbb{P}\left(\frac{\sum_{j=\nu n}^n \Delta(s)}{j - i} < \mathbb{E}\Delta(i) - t\right) \leq e^{-2(j-i)^2}. \quad (3.36)$$

With the help of (3.32), and since $K = \lambda/2m$, by choosing $t = \mathbb{E}\Delta(i) - K$, (3.36) becomes

$$\mathbb{P}((\tilde{O}_{i,j}^{(n)})^c) \leq e^{-2i(j-i)(\mathbb{E}\Delta(i) - K)^2} \leq e^{-2K^2h(n)}, \quad (3.37)$$

for all $i, j \in [\nu n, n]$. Then, note that there are at most $n$ terms in the sum in (3.35). Thus (3.35) and (3.37) together imply that

$$\mathbb{P}(\tilde{O}_n^c) \leq n e^{-2K^2h(n)}. \quad (3.38)$$
4 Estimation of the Constants

To estimate $C$ in (1.3), we need to first estimate various constants.

First let $\nu = 1/2m$. Next, to estimate $K_1$, the right hand side of (2.9) needs to be lower bounded. When $n \geq 900/(p(1-p))$, (2.14) gives that

$$\text{Var}(N|1_{N \in I} = 1) \geq \frac{1}{1000} p(1-p)n.$$  

Therefore, any $K_1$ satisfying $0 < K_1 < \sqrt{p(1-p)/(10\sqrt{5})}$ is fine. Choosing $K_1 = \sqrt{p(1-p)/(20\sqrt{5})}$, then

$$\text{Var}(N|1_{N \in I} = 1) - h(n)^2 \geq \frac{1}{2000} p(1-p)n.$$  

To estimate $A$ and $B$ in (2.4) requires upper bounds on $C_4$, $C_5$, $C_7$ and lower bounds for $C_3$, $C_6$, $C_8$. As shown after Lemma 7, we can choose $\epsilon = e^{-9}/(1 + \ln m)$, then

$$C_3 = (\delta(\epsilon) - \epsilon)^2 C_2/2 = \epsilon \ln m - (1 - \epsilon) \ln (1 - \epsilon) - \epsilon \ln \epsilon \geq \epsilon \ln m \geq e^{-10},$$

and

$$C_4 = 1/(1 - e^{-C_3}) \leq e^{11}.$$  

Lemma 8 gives

$$C_5 = e^{1/2000m}/e^{1/2000m} - 1 \leq 1 + 2000m,$$

and

$$C_6 = \frac{1}{4000m^2}.$$  

Lemma 8 gives

$$C_7 = D \leq 20e^9,$$

and

$$C_8 = \frac{c(\delta)}{D} = -\ln \mathbb{P}(\bar{Y}_s = 1) \geq \ln \frac{m}{m-1} - \ln m \geq \ln \frac{m}{m-1} \left(1 - \frac{11 \ln m \ln \frac{m}{m-1}}{40e^9 m^3 (1 + \ln m)}\right) \geq \ln \frac{m}{m-1} \left(1 - \frac{1}{20e^9}\right) \geq \frac{1}{2m}.$$  

Therefore, one can take $A = \max\{1+2000m, 20e^9\}$ and $B = e^{-10}/m^2$. Then, for $n \geq e^{10}m^2 \ln (80e^9 + 8000m)$, $\mathbb{P}(O_n) \geq 1/2$.

Note that when $n \geq 400/(p(1-p))$, we also have $\mathbb{P}(N \in I) \geq 1/2$. Let

$$C_9 = \frac{K^2}{64000} p(1-p),$$

and let

$$C_{10} = \min_{n \leq \max\{900/(p(1-p)), e^{10}m^2 \ln (80e^9 + 8000m)\}} \frac{\text{Var} LC_n}{n},$$

then one can choose $C = \min\{C_9, C_{10}\}$ in (1.3).
5 Concluding Remarks

• The results of the paper show that we can approach as closely as we want the uniform case and have a linear order on the variance of $LC_n$. However, the lower order of the variance in the uniform case is still unknown although numerical results, see [14], leave little doubt that the variance is linear in the length of the words. (Unfortunately, the estimates of the previous section, on $C = C(p, m)$ in (1.29), converge to zero as $p \to 0$.)

• Combining the above results with techniques and results presented in [9], the upper and lower bound obtained above can be generalized to provide estimates of order $n^{r/2}$, $r \geq 1$, on the centered $r$-th moment of $LC_n$.

• Finally, the above results might also be extended to the general case where the letters of one sequence are taken with probability $p_i$, $i = 1, 2, \ldots, m$, where $p_i > 0$ and $\sum_{i=1}^{m} p_i = 1$, while for the other sequence the first $m$ letters are taken with probability $p_i - r_i > 0$ and the extra letter is taken with probability $p = \sum_{i=1}^{m} r_i$. Then many of the lemmas remain true replacing $1/m$ by $\inf_{i=1, \ldots, m} p_i$ or $\inf_{i=1, \ldots, m} (p_i - r_i)/(1 - \sum_{k=1}^{m} r_k)$. For example, in the heading of Section 3.1 and Section 3.2 in (3.1), (3.3), (3.8), and Lemma 10 the $1/m$ can be replaced by $\inf_{i=1, \ldots, m} p_i$. In (3.4) of Lemma 4, and in the definition of $G^{(m)}(\epsilon)$ in Lemma 5 the term $L_{\ell}(m(1-\delta))$ would have to be replaced with

$$L_{\ell} \left( \frac{\ell(1-\delta)(1-\sum_{k=1}^{m} r_k)}{\inf_{i=1, \ldots, m}(p_i - r_i)} \right).$$

However, some constants that needs delicate estimations, such as $\xi_m$, could be a further research topic.

References

[1] Saba Amsalu, Christian Houdré, and Heinrich Matzinger. Sparse Long Blocks and the Variance of the Longest Common Subsequences in Random Words. arXiv:1204.1009v2 [math-ph], September 2016.

[2] Federico Bonetto and Heinrich Matzinger. Fluctuations of the Longest Common Subsequence in the Asymmetric Case of 2- and 3-Letter Alphabets. Latin American Journal of Probability and Mathematical Statistics, 2:195–216, 2006.

[3] Vaclav Chvatal and David Sankoff. Longest Common Subsequences of Two Random Sequences. Journal of Applied Probability, 12(2):306–315, 1975.

[4] Vaclav Chvatal and David Sankoff. An Upper-bound Technique for Lengths of Common Subsequences. In David Sankoff and Joseph Kruskal, editors, Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Addison-Wesley, Reading, Massachusetts, 1983.

[5] Vladimir Dancik. Expected Length of Longest Common Subsequences. PhD thesis, 1994.

[6] Joseph G. Deken. Some Limit Results for Longest Common Subsequences. Discrete Mathematics, 26(1):17–31, January 1979.

[7] Joseph G. Deken. Probabilistic Behavior of Longest-Common-Subsequence Length. In David Sankoff and Joseph Kruskal, editors, Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Addison-Wesley, Reading, Massachusetts, 1983.

[8] Ruoting Gong, Christian Houdré, and Jüri Lember. Lower Bounds on the Generalized Central Moments of the Optimal Alignments Score of Random Sequences. Journal of Theoretical Probability, pages 1–41, December 2016.
[9] Christian Houdré and Jinyong Ma. On the Order of the Central Moments of the Length of the Longest Common Subsequences in Random Words. In *High Dimensional Probability VII*, pages 105–136. Birkhäuser, Cham, 2016. DOI: 10.1007/978-3-319-40519-3_5.

[10] Christian Houdré and Heinrich Matzinger. On the Variance of the Optimal Alignments Score for Binary Random Words and an Asymmetric Scoring Function. *Journal of Statistical Physics*, 164(3):693–734, August 2016.

[11] Marcos Kiwi and José Soto. On a Speculated Relation Between Chvátal–Sankoff Constants of Several Sequences. *Combinatorics, Probability and Computing*, 18(04):517–532, July 2009.

[12] Jüri Lember and Heinrich Matzinger. Standard Deviation of the Longest Common Subsequence. *The Annals of Probability*, 37(3):1192–1235, May 2009.

[13] Jüri Lember, Heinrich Matzinger, Joonas Sova, and Fabio Zucca. Lower Bounds for Moments of Global Scores of Pairwise Markov Chains. *arXiv:1602.05560 [math]*, February 2016. arXiv: 1602.05560.

[14] Qingqing Liu and Christian Houdré. Simulations, Computations, and Statistics for Longest Common Subsequences. *arXiv:1705.06826*, May 2017.

[15] George S. Lueker. Improved Bounds on the Average Length of Longest Common Subsequences. *J. ACM*, 56(3):17:1–17:38, May 2009.

[16] J. Michael Steele. An Efron-Stein Inequality for Nonsymmetric Statistics. *The Annals of Statistics*, 14(2):753–758, June 1986.