FIRST MOMENT PROBLEM IN LONGEST COMMON SUBSEQUENCES

by

Abdurrahman Demirelli

B.S., Mathematics, Boğaziçi University, 2016

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Mathematics
Boğaziçi University

2018

ACKNOWLEDGEMENTS

I would like to thank Assoc. Prof. Ayhan Günaydın and Assoc. Prof. Özgür Martin, for their participation in my thesis committee.

I am extremely thankful to Assoc. Prof. Atilla Yılmaz for his constant support and valuable advice. I would also like to thank Assoc. Prof. Alp Bassa ans Assoc. Prof. Ali Emre Pusane for believing in me.

I thank my professors and my co-workers from Bilgi University for providing me a peaceful work environment.

I deeply thank my parents, Hayriye and Fuat Demirelli, and Ümmühan, Ada and Çağlar Akkaş for their unconditional trust, timely encouragement and endless patience.

Finally, I thank with love to Zeynep Çetin, my fiancée. She has been my best friend and great companion, loved, supported, encouraged, entertained, and helped me get through this agonizing period in the most positive way.

ABSTRACT

FIRST MOMENT PROBLEM IN LONGEST COMMON SUBSEQUENCES

In this thesis, we investigate the properties of the longest common subsequences in random words, examine upper and lower bounds for the expected value of the longest common subsequences in this setting, and discuss the behavior of the asymptotic order of the longest common subsequences's variance. Besides this, we also study the relationship between longest common subsequences and longest increasing subsequences in random permutations and discuss some properties of the matrix $L^{(n)}$ that is generated by the length of the longest common subsequences of permutations. Our aim is to understand the details of the theory of the longest common subsequences whose study begun in 1970's, draw attention to the progress about the longest common subsequences in the recent studies, and state some open problems about the subject.

ÖZET

EN UZUN ORTAK ALTDİZİLERDE İLK KUVVET PROBLEMİ

Bu tezde, rastgele kelimelerin en uzun ortak altdizilerinin özellikleri, beklenen değerlerinin alt ve üst sınırları araştırılmış ve bu ortamda en uzun ortak altdizilerin varyanslarının asimptotik davranışları tartışılmıştır. Aynı zamanda, rastgele permütasyonlarda en uzun ortak altdizilerin ve en uzun artan altdizilerin ilişkileri çalışılmış ve rastgele permütasyonların en uzun ortak altdizilerin uzunlukları kullanılarak oluşturulmuş $L^{(n)}$ matrisinin bazı özellikleri tartışılmıştır. Bu tezin amacı, 1970'lerde çalışılmaya başlanan en uzun ortak diziler hakkında detaylı bilgi sahibi olup günümüzde yapılan araştırmalarla gelinen noktalara ve hala ucu açık sorulara dikkat çekmektir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS

$\mathbf{1}$	The indicator function
$Ber(p)$	Bernoulli distribution with parameter p
$C(F)$	The set of points where the function F is continuous
$\|\|d_i\|\|_\infty$	The infinity norm of d_i
S_n	Symmetric group on n-elements
$[n]$	The set of natural numbers from 1 to n
$\lceil n \rceil$	The smallest integer that is greater than n
$\lfloor n \rfloor$	The greatest integer that is smaller than n
$:=$	Equality that includes a definition
$=_d$	Equality in distribution
$\sim$	Asymptotic order
$\approx$	Approximately
$\succeq$	Component-wise greater than or equal to
$\preceq$	Component-wise smaller than or equal to

LIST OF ACRONYMS/ABBREVIATIONS

a.s.	Almost Surely
i.i.d.	Independent and Identically Distributed
LCS	Longest Common Subsequence
LIS	Longest Increasing Subsequence
WLOG	Without Loss of Generality

1. INTRODUCTION

Consider two sequences that take their elements from the same finite alphabet. A subsequence that is shared for both of these sequences is called a common subsequence and a common sequence with longest length is called a *longest common subsequence.* The theory of longest common subsequences is one of the most well-studied problems of probability theory. It has lots of applications from computer science to computational biology. In recent years, this problem has become more popular than ever with the improvements on the gene matchings and the similarity problems [1].

For instance, in order to determine the difference between two different versions of a file, a computer scientist can consider these two files as strings or sequences and check their longest common subsequence, or letting our alphabet to be $\mathcal{A} := \{A, C, T, G\}$ where A,C,T and G represent the four nucleotides of DNA, namely; adenine, cytosine, thymine and guanine, respectively, a computational biologist can easily determine the similarity between two genes.

The first algorithm that gives the longest common subsequence of two given sequences is given by Sankoff in [2] in 1972. Three years later, Chvatal and Sankoff studied this problem thoroughly in [3]. Their astonishing result was stating that

$$\lim_{n\to\infty} \frac{\mathbb{E}[L_n]}{n} = \gamma$$

where L_n (see Definition 3.1.) is the cardinality of the longest common subsequence of two sequences with length n and $\gamma = \sup_{n\geq 1} \mathbb{E}[L_n]/n$.

The computation of γ is fairly interesting. First of all, the value of γ is not known regardless of the size of the alphabet and the distribution of the sequences. Even if we take an alphabet with size 2 and even if the distribution of the sequences are $Ber(1/2)$, we cannot find the exact value of γ. In [4], Steele conjectured that $\gamma = \frac{2}{1+\sqrt{2}}$, but later, it has been shown that γ is closer to 0.81, than $2/(1+\sqrt{2})$ by Rinsma-Melchert in [5].

In this thesis, our main concern is to understand bounds of the expected value of the longest common subsequences in random words and relationship between longest common subsequences and longest increasing subsequences in random permutations. The organization of this thesis is as follows:

In the preliminaries chapter of this thesis, some basic definitions and results about probability theory are given.

In the third chapter, we introduce the formal definition of the longest common subsequences. After some basic facts and examples, we give an upper and a lower bound for the expected value of a longest common subsequence using Rhee's argument [6]. At the end of Chapter 3, we discuss the statistical behavior of the variance of the longest common subsequences.

In the fourth chapter, we start with the definition of longest increasing subsequences. Then, we examine the relationship between longest common subsequences and longest increasing subsequences.

In the last chapter, we conclude our work and state some open problems about the longest common subsequences.

I hope that this thesis provides a good guide in order to understand the first moment problem in longest common subsequences.

2. PRELIMINARIES

2.1. Some Basic Definitions

Definition 2.1. Consider a sequence of real numbers $\{a_n\}$, where $n \geq 1$. We say that $\{a_n\}$ is *super-additive* if for any $n, m \geq 1$, we have $a_{n+m} \geq a_n + a_m$. Similarly, the sequence is called *sub-additive* if for any $n, m \geq 1$, we have $a_{n+m} \leq a_n + a_m$.

Definition 2.2. If $\{d_i\}_{i=1}^n$ is a sequence of random variables such that $\mathbb{E}[d_{i_1} d_{i_2} \dots d_{i_k}] = 0$ for all $1 \leq i_1 < i_2 < \dots < i_k \leq n$, then $\{d_i\}_{i=1}^n$ is called a *multiplicative system.*

Definition 2.3. Let φ be a measure-preserving map on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A set $A \in \mathcal{F}$ is said to be *invariant* if $\varphi^{-1}A = A$. Let $\mathcal{I}$ be the set of invariant events. If $\mathcal{I}$ is trivial, i.e. for every $A \in \mathcal{I}$, we have $\mathbb{P}(A) \in \{0, 1\}$, then φ is called *ergodic.*

Definition 2.4. A sequence of random variables $X_0, X_1, \dots$ is said to be a *stationary sequence* if the shifted sequence $\{X_{k+n}, n \geq 0\}$ has the same distribution as the original sequence $\{X_n, n \geq 0\}$, i.e. for each $m \geq 0$, $\{X_0, X_1, \dots, X_m\}$ and $\{X_k, X_{k+1}, \dots, X_{k+m}\}$ have the same distribution, for every $k \geq 0$.

Definition 2.5. Let $X_1, X_2, \dots$ be random variables. We say that X_n converges almost surely (a.s.) to the random variable X as $n \to \infty$ if

$$\mathbb{P}(w : X_n(w) \to X(w) \quad as \quad n \to \infty) = 1.$$

In this case, we write $X_n \to_{a.s.} X$ as $n \to \infty$.

Definition 2.6. Let $X_1, X_2, \dots$ be random variables. We say that X_n converges in probability to the random variable X as $n \to \infty$ if, for every $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

In this case, we write $X_n \to_{\mathbb{P}} X$ as $n \to \infty$.

Definition 2.7. Let $X_1, X_2, \ldots$ be random variables and $r \geq 1$. We say that X_n converges in r-mean to the random variable X as $n \to \infty$ if

$$\mathbb{E}[|X_n - X|^r] \to 0 \quad \text{as} \quad n \to \infty.$$

In this case, we write $X_n \to_r X$ as $n \to \infty$.

If $r = 2$, then this convergence is also called as mean-square convergence.

Definition 2.8. Let $X_1, X_2, \ldots$ be random variables. We say that X_n converges in distribution to the random variable X as $n \to \infty$ if

$$F_{X_n}(x) \to F_X(x) \quad \text{as} \quad n \to \infty, \quad \text{for all} \quad x \in C(F_X)$$

where F_{X_n} and F_x are cumulative distribution functions of X_n and X, respectively. In this case, we write $X_n \to_d X$ as $n \to \infty$.

Definition 2.9. A sequence $X_1, X_2, \ldots$ is called uniformly integrable if

$$\mathbb{E}[|X_n| \mathbf{1}\{|X_n| > a\}] \to 0, \quad as \quad a \to \infty \quad \text{uniformly in} \quad n.$$

Definition 2.10. Let, for $n \geq 1$, S_n be the symmetric group on n elements. A random permutation is a random reordering of the elements of S_n.

Definition 2.11. The radius of the spectrum of a square matrix is the largest absolute value of its eigenvalues, i.e. $\rho(A) = \max\{|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|\}$, where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are eigenvalues of the square matrix A.

2.2. Some Basic Results

Linearity of Expectation: [7] For any random variables $X_1, X_2, \ldots, X_n$ and constants $c_1, c_2, \ldots, c_n$, we have

$$\mathbb{E}\Big[\sum_{i=1}^{n} c_i X_i\Big] = \sum_{i=1}^{n} c_i \mathbb{E}[X_i].$$

Fekete's Lemma: [8] If $\{a_n\}$ is a super-additive sequence, then

$$\lim_{n\to\infty} \frac{a_n}{n} = \sup_n \frac{a_n}{n}.$$

Azuma's Inequality: [8] For a multiplicative system $\{d_i\}_{i=1}^n$, we have

$$\mathbb{P}\Big(|\sum_{i=1}^{n} d_i| \geq \lambda\Big) \leq 2\exp\Big(\frac{-\lambda^2}{2\sum_{i=1}^n ||d_i||_\infty^2}\Big).$$

McDiarmid's Inequality: [8] Let $X_1, X_2, \ldots, X_n$ be independent random variables that take values in Ω and $f : \Omega^n \to \mathbb{R}$ be a function of $X_1, X_2, \ldots, X_n$ that satisfies for any $i \in [n]$ and for any $x_1, x_2, \ldots, x_n, x_i' \in \Omega$

$$|f(x_1, x_2, \ldots, x_i, \ldots, x_n) - f(x_1, x_2, \ldots, x_i', \ldots, x_n)| \leq c_i$$

for some $c_i \in \mathbb{R}$. Then, for all $\epsilon > 0$, we have

$$\mathbb{P}\Big((f - \mathbb{E}[f]) \geq \epsilon\Big) \leq \exp\Big(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\Big).$$

Kingman's Sub-Additive Ergodic Theorem: [9] Let $\{X_{m,n}\}$ be a sequence of random variables indexed by non-negative integers $0 \leq m < n < \infty$. Assume that, we have

- $X_{0,n} \leq X_{0,m} + X_{m,n}$ for all $0 \leq m < n$;
- $\{X_{m+1,n+1} : 0 \leq m < n\} =_d \{X_{m,n} : 0 \leq m < n\}$ for each n and this shift operation is ergodic;
- $\mathbb{E}[X_{0,n}] > -cn$ for some $c > 0$ and all n.

Then

$$\lim_{n\to\infty} \frac{X_{0,n}}{n} = \lim_{n\to\infty} \frac{\mathbb{E}[X_{0,n}]}{n} = \inf_n \frac{\mathbb{E}[X_{0,n}]}{n}.$$

Erdös-Szekeres Theorem: [8] Let $\{a_i\}_{i=1}^n$ be a sequence of real numbers where $n = ab + 1$ for some $a, b \in \mathbb{N}$. Then, this sequence either contains a monotonic non-decreasing (non-increasing) subsequence of $a+1$ terms, or a monotonic non-increasing (non-decreasing) subsequence of $b+1$ terms.

Relations Between Convergence Concepts: [7] Let X and $X_1, X_2, \ldots$ be random variables and $r \geq 1$. The following implications hold as $n \to \infty$:

$$X_n \to_{a.s.} X \quad \Rightarrow \quad X_n \to_{\mathbb{P}} X \quad \Rightarrow \quad X_n \to_d X$$

$$\Uparrow$$

$$X_n \to_r X.$$

Lemma 2.11. [7] Suppose that $X_1, X_2, \ldots$ are random variables such that

$$|X_n| \leq Y \quad a.s. \quad \text{for all} \quad n$$

where Y is a positive integrable random variable. Then $\{X_n, n \geq 1\}$ is uniformly integrable.

Lemma 2.12. [7] Let X and $X_1, X_2, \ldots$ be random variables, and suppose that $X_n \to_{a.s.} X$ as $n \to \infty$. Let $r > 0$. The following are equivalent:

(i) $\{|X_n|^r, n \geq 1\}$ is uniformly integrable;

(ii) $X_n \to_r X$ as $n \to \infty$;

(iii) $\mathbb{E}[|X_n|^r] \to \mathbb{E}[|X|^r]$ as $n \to \infty$.

Moreover, if $r \geq 1$ and one of the above holds, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$ as $n \to \infty$.

Stirling's Formula: [7] As $n \to \infty$, $n! \sim \frac{n^n}{e^n}\sqrt{2\pi n}$, i.e.

$$\lim_{n\to\infty} \frac{n!}{(n^n/e^n)\sqrt{2\pi n}} = 1.$$

Efron-Stein Inequality: [10] Let S be any function of n variables, X_i and $\tilde{X}_i$ independent random variables with the same distribution for any $i \in [n]$, $S := S(X_1, X_2, \ldots, X_n)$, and $S_i := S(X_1, X_2, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n)$, for any $i \in [n]$. Then,

$$Var(S) \leq \frac{1}{2}\mathbb{E}\Big[\sum_{i=1}^{n}(S - S_i)^2\Big].$$

3. LONGEST COMMON SUBSEQUENCES

3.1. Basic Facts and Examples

Definition 3.1. Let $\mathcal{A}$ be a finite alphabet and let $\mathbb{X} = \{X_i\}_{i=1}^n$ and $\mathbb{Y} = \{Y_i\}_{i=1}^n$ be two independent and identically distributed sequences where $n \geq 1$ and for any $i \in [n]$, we have $X_i, Y_i \in \mathcal{A}$. Let also $\mathbb{X}$ and $\mathbb{Y}$ be independent as sequences. The length of the longest common subsequence of $\mathbb{X}$ and $\mathbb{Y}$ is given by a random variable $L_n = L_n(\mathbb{X}, \mathbb{Y})$ that is defined by

$$L_n(\mathbb{X}, \mathbb{Y}) := \max\{k : X_{i_1} = Y_{j_1}; X_{i_2} = Y_{j_2}; \ldots; X_{i_k} = Y_{j_k}\}$$

where the maximum is taken over all pairs of subsequences $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ and $1 \leq j_1 < j_2 < \cdots < j_k \leq n$.

Example 3.2. Let $\mathcal{A} = \{A, B, \ldots, Z\}$, $\mathbb{X}$ be the word *ARMADILLO* and $\mathbb{Y}$ be the word *ALLIGATOR*. Then observe that the longest common subsequence of words *ARMADILLO* and *ALLIGATOR* is *ALLO*. So, $L_9(\mathbb{X}, \mathbb{Y}) = 4$.

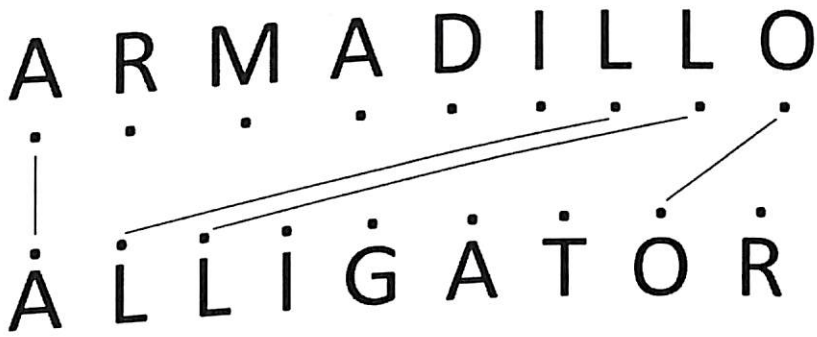

Figure 3.1. Illustration of the longest common subsequence for ARMADILLO and ALLIGATOR.

To determine the longest common subsequence, the following algorithm can be used

		A	R	M	A	D	I	L	L	O
	0	0	0	0	0	0	0	0	0	0
A	0									
L	0									
L	0									
I	0									
G	0									
A	0									
T	0									
O	0									
R	0									

Figure 3.2. Construction of the table for longest common subsequence of ARMADILLO and ALLIGATOR.

First construct a table with $n + 2$ columns and $m + 2$ rows, where n and m are lengths of the sequences. Write down the first sequence to the first row, starting from the third column and the second sequence to the first column, starting from the third row. Fill the second column and second row with zeros, leaving the boxes $1 \times 1, 1 \times 2$ and 2×1 empty.

		A	R	M	A	D	I	L	L	O
	0	0	0	0	0	0	0	0	0	0
A	0	1	1	1	1	1	1	1	1	1
L	0									
L	0									
I	0									
G	0									
A	0									
T	0									
O	0									
R	0									

Figure 3.3. Computation of the third row of the table.

Now, consider the first three rows and the first three columns. We have two subsequences and both of them are A. So, the longest common subsequence is 1. Next, consider the first three rows and the first four columns of the table. We have A and AR. If the new letter, in our case R, is the same with the letter compared, add 1 to the box above, if not, write down the same number. Since the second sequence is A, we have only 1 for the third row.

		A	R	M	A	D	I	L	L	O
	0	0	0	0	0	0	0	0	0	0
A	0	1	1	1	1	1	1	1	1	1
L	0	1	1	1	1	1	1	2	2	2
L	0	1	1	1	1	1	1	2	3	3
I	0	1	1	1	1	1	2	2	3	3
G	0	1	1	1	1	1	2	2	3	3
A	0	1	1	1	2	2	2	2	3	3
T	0	1	1	1	2	2	2	2	3	3
O	0	1	1	1	2	2	2	2	3	4
R	0	1	1	1	2	2	2	2	3	4

Figure 3.4. Computation of the whole table.

Then, move on to the fourth row. First, consider subsequences AL and A and fill the fourth row by moving one box to the right at each step. Fill the table by repeating the last step for each row. The last entry of the table, in our case 4, is the length of the longest common subsequence.

		A	R	M	A	D	I	L	L	O
	0	0	0	0	0	0	0	0	0	0
A	0	1	1	1	1	1	1	1	1	1
L	0	1	1	1	1	1	1	2	2	2
L	0	1	1	1	1	1	1	2	3	3
I	0	1	1	1	1	1	2	2	3	3
G	0	1	1	1	1	1	2	2	3	3
A	0	1	1	1	2	2	2	2	3	3
T	0	1	1	1	2	2	2	2	3	3
O	0	1	1	1	2	2	2	2	3	4
R	0	1	1	1	2	2	2	2	3	4

Figure 3.5. Finding the longest common subsequence and its length.

In order to find the longest common subsequence, check the last entry and the one above that. If they are equal, then move on to the above number. If not, then move to north-west neighbour. In our case, the last entry is 4 and the one above that, we also have a 4. So, we move upwards. Then, we have 4 and 3, hence we move to the north-west neighbour of 4. If we continue this procedure until we hit a 0, at the end, the letters on the first column with north-west jumps will be the longest common subsequence. In our case, it is *ALLO*.

Now, let us define $\mathbb{Z} := \{Z_i\}_{i=1}^{n+m}$ where $Z_i = (X_i, Y_i)$ for every $i \in [n+m]$. Observe that for any $n, m \geq 1$, we have

$$L_{n+m}(Z_1, Z_2, \ldots, Z_{n+m}) \geq L_n(Z_1, Z_2, \ldots, Z_n) + L_m(Z_{n+1}, Z_{n+2}, \ldots, Z_{n+m}). \quad (3.1)$$

The equality case is obvious. For the other case, consider the following example:

Example 3.3. Let again $\mathcal{A}$ be the alphabet that consists of letters and $\mathbb{Y}$ be words *ARMADILLO* and *ALLIGATOR*, respectively. We know that, from Example 3.2 that $L_9(\mathbb{X}, \mathbb{Y}) = L_9(Z_1, Z_2, \ldots, Z_9) = 4$. Let $\mathbb{X}' = (R)$ and $\mathbb{Y}' = (P)$. Then, $L_1(\mathbb{X}', \mathbb{Y}') = L_1(Z_{10}) = 0$. However, $L_{10}(Z_1, Z_2, \ldots, Z_{10}) = 5$ since now *ALLOR* is shared for both *ARMADILLOR* and *ALLIGATORP*.

If we let $a_n = \mathbb{E}[L_n]$ and if we take the expectation of both sides of the inequality in (3.1), from the linearity of expectation, we obtain that a_n is a superadditive sequence. So, we can use Fekete's lemma and get our first limit result [3]:

$$\lim_{n\to\infty} \frac{\mathbb{E}[L_n]}{n} = \sup_n \frac{\mathbb{E}[L_n]}{n} = \gamma \quad (3.2)$$

for some $\gamma \in [0, 1]$.

We have showed that $\frac{\mathbb{E}[L_n]}{n} \to \gamma$ as $n \to \infty$. Using Kingman's Subadditive Ergodic Theorem, we can get rid of the expectation in the above convergence and say that $\frac{L_n}{n} \to \gamma$ a.s., but first we need some preparation.

Let $L_{m,n}$ be the length of the longest common subsequence of two sequences of two sequences on indices $m \leq n$. Then, we know that $L_{0,m} + L_{m,n} \geq L_{0,n}$. Moreover, we know, from the construction of the longest common subsequence that $\{L_{m+1,n+1} : 0 \leq m < n\} \stackrel{d}{=} \{L_{m,n} : 0 \leq m < n\}$, i.e. sequences are stationary and the distribution does not depend on n. Then, lastly, $\mathbb{E}[L_{0,n}] > -cn$ for some $c > 0$ and all $n > 0$, which is obvious since the longest common subsequence is non-negative. This means that $L_{0,n}$ satisfies requirements of the Kingman's Subadditive Ergodic Theorem, which gives [9]

$$\lim_{n\to\infty} \frac{L_n}{n} = \lim_{n\to\infty} \frac{\mathbb{E}[L_n]}{n} = \sup_n \frac{\mathbb{E}[L_n]}{n} = \gamma.$$

We can also obtain the same result using McDiarmid's Inequality. First, observe that L_n is a function from Ω^n to $\mathbb{R}$. Moreover, changing a single letter from one of the sequences can change L_n by at most 1. So,

$$\mathbb{P}(L_n - \mathbb{E}[L_n] \geq \epsilon_n) \leq \exp\left(\frac{-2{\epsilon_n}^2}{n}\right).$$

Let $\epsilon_n = \sqrt{n\log(n)}$. Then we have,

$$\mathbb{P}(L_n - \mathbb{E}[L_n] \geq \sqrt{n\log(n)}) \leq \exp\left(\frac{-2n\log(n)}{n}\right) = \frac{1}{n^2}.$$

Now, we can use Borel-Cantelli Lemma since

$$\sum_{n=1}^{\infty} \mathbb{P}(L_n - \mathbb{E}[L_n] \geq \sqrt{n\log(n)}) \leq \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$$

and get that $L_n - \mathbb{E}[L_n] \geq \sqrt{n\log(n)}$ holds for only finitely many n's. So, with probability 1, there exists $N \in \mathbb{N}$ such that for every $n \geq N$, we have $L_n - \mathbb{E}[L_n] < \sqrt{n\log(n)}$. Therefore, $\lim_{n\to\infty} \frac{\mathbb{E}[L_n]}{n} = \gamma$ implies that

$$\lim_{n\to\infty} \frac{L_n}{n} = \gamma \quad \text{a.s.}$$

Lemma 3.4. *As $n \to \infty$, $\frac{L_n}{n}$ converges in mean square to γ, i.e. $\frac{L_n}{n} \to_2 \gamma$.*

Proof. Our aim is to show that $\mathbb{E}[\left|\frac{L_n}{n} - \gamma\right|^2] \to \infty$ as $n \to 0$. This can be done in two ways.

As a first proof, we know that $\frac{L_n}{n} \to \gamma$ as $n \to \infty$ a.s. So, Lemma 2.12. tells us instead of showing $\frac{L_n}{n} \to_2 \gamma$, we can show that $\{|\frac{L_n}{n}|^2, n \geq 1\}$ is uniformly integrable. So, using Lemma 2.11., it is enough to show that there exists a positive integrable random variable Y such that $|\frac{L_n}{n}|^2 \leq Y$ a.s. for all n. Since $L_n \leq n$ for any n, if we take $Y = 1$, the result will follow.

As a second proof, observe that

$$\begin{aligned}\mathbb{E}\Big[\Big|\frac{L_n}{n} - \gamma\Big|^2\Big] &= \mathbb{E}\Big[\big(\frac{L_n}{n} - \mathbb{E}[\frac{L_n}{n}] + \mathbb{E}[\frac{L_n}{n}] - \gamma\big)^2\Big] \\ &= \mathbb{E}\Big[\big(\frac{L_n}{n} - \mathbb{E}[\frac{L_n}{n}]\big)^2\Big] + 2\mathbb{E}\Big[\big(\frac{L_n}{n} - \mathbb{E}[\frac{L_n}{n}]\big)\big(\mathbb{E}[\frac{L_n}{n}] - \gamma\big)\Big] \\ &\quad + \mathbb{E}\Big[\big(\mathbb{E}[\frac{L_n}{n}] - \gamma\big)^2\Big].\end{aligned} \tag{3.3}$$

Since we know that $\mathbb{E}\Big[\big(\frac{L_n}{n} - \mathbb{E}[\frac{L_n}{n}]\big)^2\Big] = Var\Big(\frac{L_n}{n}\Big)$, we have

$$\begin{aligned}\mathbb{E}\Big[\Big|\frac{L_n}{n} - \gamma\Big|^2\Big] &= Var\Big(\frac{L_n}{n}\Big) + 2\Big(\mathbb{E}\Big[\frac{L_n}{n}\Big] - \gamma\Big)\mathbb{E}\Big[\big(\frac{L_n}{n} - \mathbb{E}[\frac{L_n}{n}]\big)\Big] + \Big(\mathbb{E}\Big[\frac{L_n}{n}\Big] - \gamma\Big)^2 \\ &= Var\Big(\frac{L_n}{n}\Big) + \Big(\mathbb{E}\Big[\frac{L_n}{n}\Big] - \gamma\Big)\Big(2\mathbb{E}\Big[\frac{L_n}{n} - \mathbb{E}[\mathbb{E}[\frac{L_n}{n}]]\Big] + \mathbb{E}[\frac{L_n}{n} - \gamma]\Big).\end{aligned} \tag{3.4}$$

In [4], Steele proved that if $\mathbb{E}[(S_{i,n} - S_{j,n})^2]$ is bounded for all $1 \leq i < j \leq n < \infty$, then $Var(S(V_1, V_2, \ldots, V_n)) = O(n)$, where S is a function from $(\mathbb{R}^d)^n$ to $\mathbb{R}$; V_i is any sequence of independent random vectors in $\mathbb{R}^d$; and $S_{i,n} := S(V_1, V_2, \ldots, V_{i-1}, V_{i+1}, \ldots, V_n)$. He used Tukey's jackknife estimate and an Efron-Stein type inequality to prove his claim.

When we take $V_i := Z_i = (X_i, Y_i)$, where $X_i, Y_i \in \mathbb{A}$, for some alphabet $\mathbb{A}$ and for any $i \in [n]$, and when we let S_n to be the random variable that gives the length of two longest common subsequences, i.e. L_n, we see that this setting satisfies the above requirements. Hence, we have $Var(L_n) = \mathrm{O}(n)$.

Since we have $Var(\frac{L_n}{n})$ in (3.4), Steele's bound will be $Var(\frac{L_n}{n}) \leq \frac{1}{n^2}n = \frac{1}{n}$. Observe that, as $n \to \infty$, we have $Var(\frac{L_n}{n}) \to 0$ and $\mathbb{E}[\frac{L_n}{n}] - \gamma \to 0$. Thus, we conclude that $\mathbb{E}[|\frac{L_n}{n} - \gamma|^2] \to 0$ and $\frac{L_n}{n} \to_2 \gamma$ as $n \to \infty$. □

In addition, Steele also shows that for any $\epsilon > 0$, $L_n - \mathbb{E}[L_n] = o(n^{3/4+\epsilon})$ with probability 1 as a corollary for his main theorem in [4].

3.2. Rate of Convergence

3.2.1. Introduction

In this section, we will focus on bounding $\mathbb{E}[L_n]$ in random words setting. Since we know that $\gamma = \sup_n \frac{\mathbb{E}[L_n]}{n}$, it is obvious that $\mathbb{E}[L_n] \leq n\gamma$. Finding a lower bound for $\mathbb{E}[L_n]$ is a bit more complicated.

Theorem 3.5. *[11] There is a constant K which does not depend on n, such that*

$$n\gamma - K\sqrt{n\log(n)} \leq \mathbb{E}[L_n] \leq n\gamma.$$

In [11], Alexander proved Theorem 3.5 with a rather complex proof. Then, Rhee introduced a significantly simpler proof of this result that relies on a reflection argument [6].

3.2.2. Rhee's Argument

Lemma 3.6. *[6] For all $x > 0$ and for all $n \in \mathbb{N}$,*

$$\mathbb{P}(L_{4n} \geq 4x) \leq (4n)^4 \big(\mathbb{P}(L_{2n} \geq 2x)\big)^{1/2}.$$

Proof. Let $\mathbb{X} = \{X_i\}_{i=1}^{4n}$ and $\mathbb{Y} = \{Y_i\}_{i=1}^{4n}$ be two sequences and $a_1, \ldots, a_\alpha; b_1, \ldots, b_\beta$ be integers such that $1 \leq a_1 < a_2 < \ldots < a_\alpha \leq 4n$ and $1 \leq b_1 < b_2 < \ldots < b_\beta \leq 4n$ for some $n \geq 1$. Define

$$L(\mathbb{X}_{a_1,a_\alpha}, \mathbb{Y}_{b_1,b_\beta}) := L(X_{a_1}, X_{a_2}, \ldots, X_{a_\alpha}; Y_{b_1}, Y_{b_2}, \ldots, Y_{b_\beta}).$$

Let $L(\mathbb{X}, \mathbb{Y})$ be the random variable that gives the length of the longest common subsequence of sequences $\mathbb{X}$ and $\mathbb{Y}$. We will prove that if $L(\mathbb{X}, \mathbb{Y}) \geq 4x$, then we can find $a_1, a_2, \ldots, a_\alpha, b_1, b_2, \ldots, b_\beta$ such that

$$\alpha + \beta = 2n; \qquad L(\mathbb{X}_{a_1,a_\alpha}, \mathbb{Y}_{b_1,b_\beta}) \geq x. \tag{3.5}$$

Indeed, for $i \in [5]$, we can find integers n_i, m_i with $1 = n_1 < n_2 < \ldots < n_5 = 4n$ and $1 = m_1 < m_2 < \ldots < m_5 = 4n$ such that for any i, we have

$$L(\mathbb{X}_{n_i,n_{i+1}}, \mathbb{Y}_{m_i,m_{i+1}}) = x$$

because otherwise, we would not have $L(\mathbb{X}, \mathbb{Y}) \geq 4x$. Since we partitioned $4n$ into 4 parts twice, we also know that

$$\sum_{i=1}^{4} \big((n_{i+1} - n_i) + (m_{i+1} - m_i)\big) = 8n.$$

So, for some $i \in [4]$, we have $|(n_{i+1} - n_i) + (m_{i+1} - m_i)| \leq 2n$. This means that we can choose $n_i, n_{i+1}, m_i, m_{i+1}$ such that $a_1 \leq n_i$, $n_{i+1} \leq a_\alpha$, $b_1 \leq m_i$, $m_{i+1} \leq b_\beta$ and (3.5) holds.

Then,

$$\mathbb{P}(L(\mathbb{X},\mathbb{Y}) \geq 4x) \leq \mathbb{P}(\bigcup_{[a_1,a_\alpha]\times[b_1,b_\beta]\in S} L_n(\mathbb{X}_{a_1,a_\alpha},\mathbb{Y}_{b_1,b_\beta} \geq x))$$

where $S := \{[a_1,a_\alpha]\times[b_1,b_\beta] : [a_1,a_\alpha]\times[b_1,b_\beta] \subset [1,4n]\times[1,4n]\}$. There are $\binom{4n}{2}$ $= \frac{4n(4n-1)}{2}$ possible choices for a_1 and a_α that satisfies the first requirement of (3.3). So, since $\frac{4n(4n-1)}{2} \leq (4n)^2$, we have

$$\mathbb{P}(L(\mathbb{X},\mathbb{Y}) \geq 4x) \leq (4n)^4\mathbb{P}(L(\mathbb{X}_{a_i,a_\alpha},\mathbb{Y}_{b_1,b_\beta}) \geq x).$$

Let $a := a_\alpha - a_1$ and $b := b_\beta - b_1$. Then we have

$$\mathbb{P}(L(\mathbb{X}_{0,a},\mathbb{Y}_{0,b}) \geq x) = \mathbb{P}(L(\mathbb{X}_{0,b},\mathbb{Y}_{0,a}) \geq x) = \mathbb{P}(L(\mathbb{X}_{a_1,a_\alpha},\mathbb{Y}_{b_1,b_\beta}) \geq x)$$

since we know that $\mathbb{X}_{1,4n}$ and $\mathbb{Y}_{1,4n}$ are independent and identically distributed and we know that shifting the sequence does not change the distribution and the length of the longest common subsequence. Now, the most important point here is the reflection argument and from super-additivity,

$$\mathbb{P}(L_{2n}(\mathbb{X},\mathbb{Y}) \geq 2x) \geq \mathbb{P}(L(\mathbb{X}_{0,a},\mathbb{Y}_{0,b}) \geq x)\mathbb{P}(L(\mathbb{X}_{0,b},\mathbb{Y}_{0,a}) \geq x).$$

Thus, we have

$$\mathbb{P}(L(\mathbb{X},\mathbb{Y}) \geq 4x) \leq (4n)^4\mathbb{P}(L_{2n}(\mathbb{X},\mathbb{Y}) \geq 2x)^{1/2}.$$

□

Lemma 3.7. *For all $n \in \mathbb{N}$ and for all $t > 0$, we have*

$$\mathbb{P}(L_n - \mathbb{E}[L_n] \geq t) \leq \exp(-t^2/8n).$$

Proof. We know, from Arzuma's inequality that

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n} d_i\Big| \geq \lambda\Big) \leq 2\exp\Big(\frac{-\lambda^2}{2\sum_{i=1}^{n}||d_i||_\infty^2}\Big)$$

where $\{d_i\}$ is multiplicative system and $\lambda > 0$. If we let $(L_n - \mathbb{E}[L_n]) = \sum_{i=1}^{n} d_i$, then we have a multiplicative system. From [8], we know that $||d_i||_\infty \leq 2$. Thus, the result follows. □

Proof of Theorem 3.4: In Lemma 3.7, consider L_{2n} and let $t := 2x - \mathbb{E}[L_{2n}]$. Then we have for $x \geq \mathbb{E}[L_{2n}]$

$$\mathbb{P}(L_{2n} \geq 2x) \leq 2\exp\Big(-(2x - \mathbb{E}[L_{2n}])^2/8n\Big).$$

Combining this inequality with the left hand side of the Lemma 3.6, we get

$$\mathbb{P}(L_{4n} \geq 4x) \leq 2(4n)^4 \exp\Big(\frac{-(2x - \mathbb{E}[L_{2n}])^2}{8n}\Big)$$

or, equivalently, for $x \geq \mathbb{E}[L_{2n}]$, we have

$$\mathbb{P}(L_{4n} \geq 2x) \leq 2(4n)^4 \exp\Big(\frac{-(x - \mathbb{E}[L_{2n}])^2)}{16n}\Big). \tag{3.6}$$

We know, from Theorem 12.1.(i) in [7], that for any non-negative random variable X, we have

$$\mathbb{E}[X] = \int_0^\infty (1 - F(x))dx = \int_0^\infty \mathbb{P}(X > x)dx$$

where F is the cumulative distribution function of X.

So, for any $u \geq 0$, we have

$$\begin{aligned}\mathbb{E}[X] &= \int_0^u \mathbb{P}(X > x)dx + \int_u^\infty \mathbb{P}(X > x)dx \\ &= \int_0^u (1 - F(x))dx + \int_u^\infty \mathbb{P}(X > x)dx \\ &= \int_0^u 1dx - \int_0^u F(x)dx + \int_u^\infty \mathbb{P}(X > x)dx \\ &\leq u + \int_u^\infty \mathbb{P}(X > x)dx.\end{aligned} \tag{3.7}$$

Since L_{4n} is a non-negative random variable and $2\mathbb{E}[L_{2n}] + K(n\log(n))^{1/2} \geq 0$, by letting $X = L_{4n}$ and $u = 2\mathbb{E}[L_{2n}] + K(n\log(n))^{1/2}$, we get

$$\mathbb{E}[L_{4n}] \leq 2\mathbb{E}[L_{2n}] + K(n\log(n))^{1/2} + \int_{2\mathbb{E}[L_{2n}]+K(n\log(n))^{1/2}}^\infty \mathbb{P}(L_{4n} \geq x)dx.$$

Our aim is to show that the integral contributes to the upper bound negligibly compared to the others.

Using (3.6), we get

$$\int_{2\mathbb{E}[L_{2n}]+K(n\log(n))^{1/2}}^\infty \mathbb{P}(L_{4n} \geq x)dx \leq 2(4n)^4 \int_{2\mathbb{E}[L_{2n}]+K(n\log(n))^{1/2}}^\infty \exp\left(\frac{-(x/2 - \mathbb{E}[L_{2n}])^2}{16n}\right)dx. \tag{3.8}$$

Let $v = \frac{x/2-\mathbb{E}[L_{2n}]}{2\sqrt{2n}}$. Then $dv = \frac{dx}{4\sqrt{2n}}$, and we have,

$$\int_{2\mathbb{E}[L_{2n}]+K(n\log(n))^{1/2}}^\infty \mathbb{P}(L_{4n} \geq x)dx \leq 2(4n)^4 \int_{\frac{K(\log(n))^{1/2}}{4\sqrt{2}}}^\infty e^{\frac{-v^2}{2}}dv.$$

We also know, from [7], that

$$\int_x^\infty e^{\frac{-t^2}{2}}dt \leq \int_x^\infty \frac{t}{x}e^{\frac{-t^2}{2}}dt.$$

So, our inequality becomes,

$$\int_{2\mathbb{E}[L_{2n}]+K(n\log(n))^{1/2}} \mathbb{P}(L_{4n} \geq x)dx \leq 2(4n)^4 \frac{4\sqrt{2}}{K(n\log(n))^{1/2}} \int_{\frac{K(n\log(n))^{1/2}}{4\sqrt{2}}}^{\infty} e^{\frac{-v^2}{2}} v dv.$$

Let $a = \frac{-v^2}{2}$. Then $da = -vdv$ and we have

$$\int_{2\mathbb{E}[L_{2n}]+K(n\log(n))^{1/2}}^{\infty} \mathbb{P}(L_{4n} \geq x)dx \leq -2(4n)^4 \frac{4\sqrt{2}}{K(n\log(n))^{1/2}} \int_{-\infty}^{\frac{-K^2\log(n)}{64}} e^a da.$$

Thus,

$$\int_{2\mathbb{E}[L_{2n}]+K(n\log(n))^{1/2}}^{\infty} \mathbb{P}(L_{4n} \geq x)dx \leq 2(4n)^4 \frac{4\sqrt{2}}{K(n\log(n))^{1/2}} \left(n^{\frac{-K^2}{64}}\right).$$

Now, we can choose K large enough so that the integral becomes negligible compared to the others in the upper bound. Therefore, we have

$$\mathbb{E}[L_{4n}] \leq 2\mathbb{E}[L_{2n}] + K(n\log(n))^{1/2}.$$

Dividing by $4n$, we get

$$\frac{\mathbb{E}[L_{4n}]}{4n} \leq \frac{\mathbb{E}[L_{2n}]}{2n} + K'(n\log(n))^{1/2}.$$

When we replace n with $2^k n$, we find

$$\frac{\mathbb{E}[L_{2^{k+2}n}]}{2^{k+2}n} \leq \frac{\mathbb{E}[L_{2^{k+1}n}]}{2^{k+1}n} + K''\sqrt{\frac{\log(2^k n)}{2^k n}}.$$

We can sum these inequalities over all $0 \leq k \leq s$ to find

$$\frac{\mathbb{E}[L_{2^{k+2}n}]}{2^{k+2}n} \leq \frac{\mathbb{E}[L_{2n}]}{2n} + K''\left(\sum_{i=0}^{s} \sqrt{\frac{\log(2^s n)}{2^s n}}\right).$$

Claim: For any $n \in \mathbb{N}$, we have $\sum_{i=0}^{\infty} \sqrt{\frac{\log(2^i n)}{2^i n}} = \mathrm{O}\Big(\sqrt{\frac{\log(n)}{n}}\Big)$.

Proof of the Claim: Let $\xi(s) := \sum_{i=0}^{s} \sqrt{\frac{\log(2)^i n}{2^i n}}$. Then,

$$\begin{aligned}
\xi(s) &= \sum_{i=0}^{s} \sqrt{\frac{\log(2^i)n}{2^i n}} = \sqrt{\frac{\log(n)}{n}} + \sqrt{\frac{\log(2n)}{2n}} + \ldots + \sqrt{\frac{\log(2^s)n}{2^s n}} \\
&= \sqrt{\frac{\log(n)}{n}} + \sqrt{\frac{\log(n)}{2n} + \frac{\log(2)}{2n}} + \ldots + \sqrt{\frac{\log(n)}{2^s n} + \frac{\log(2^s)}{2^s n}} \\
&= \sqrt{\frac{\log(n)}{n}}\Big(1 + \sqrt{\frac{1}{2} + \frac{\log(2)}{\log(n^2)}} + \ldots + \sqrt{\frac{1}{2^s} + \frac{\log(2^s)}{\log(n^{2^s})}}\Big) \\
&= \sqrt{\frac{\log(n)}{n}}\Big(1 + \sqrt{\frac{1}{2}(1 + \log_n(2))} + \ldots + \sqrt{\frac{1}{2^s}(1 + s\log_n(2))}\Big) \\
&= \sqrt{\frac{\log(n)}{n}} \sum_{i=0}^{s} \sqrt{\frac{1 + i\log_n(2)}{2^i}}.
\end{aligned} \tag{3.9}$$

Our aim is to show that $\sum_{i=0}^{\infty} \sqrt{\frac{1+i\log_n(2)}{2i}} < \infty$. Observe that

$$\begin{aligned}
\sum_{i=0}^{\infty} \sqrt{\frac{1 + i\log_n(2)}{2^i}} &\leq \sum_{i=0}^{\infty} \sqrt{\frac{1 + i\log_n(n)}{2^i}} \\
&= \sum_{i=0}^{\infty} \sqrt{\frac{1+i}{2^i}} \\
&\leq \sum_{i=1}^{\infty} \sqrt{\frac{2i}{2^i}} \\
&= K < \infty
\end{aligned} \tag{3.10}$$

for some $K > 0$. Thus we have

$$\sum_{i=0}^{\infty} \sqrt{\frac{1 + i\log_n(2)}{2^i}} \leq K\sqrt{\frac{\log(n)}{n}}.$$

Now, if we send s to the infinity in the inequality

$$\frac{\mathbb{E}[L_{2^{k+2}n}]}{2^{k+2}n} \leq \frac{\mathbb{E}[L_{2n}]}{2n} + K''\Big(\sum_{i=0}^{s}\sqrt{\frac{\log(2^s)n}{2^s n}}\Big)$$

we will have

$$\gamma \leq \frac{\mathbb{E}[L_{2n}]}{2n} + K\sqrt{\frac{\log(n)}{n}}.$$

Or, equivalently, for some $K' > 0$,

$$n\gamma - K'(n\log(n))^{1/2} \leq \mathbb{E}[L_{2n}].$$

The upper bound of the theorem is a consequence of the super-additivity of the sequence $a_n = \mathbb{E}[L_n]$.

Therefore, for any $n \in \mathbb{N}$, we have

$$2n\gamma - K(2n\log(2n))^{1/2} \leq \mathbb{E}[L_{2n}] \leq 2n\gamma$$

and we want to show that the theorem holds for odd n's as well.

Let $K' = K + \frac{\gamma}{(2n\log(2n))^{1/2}}$. Since $K' > K$, we have

$$2n\gamma - K'(2n\log(2n))^{1/2} \leq \mathbb{E}[L_{2n}]$$

for any $n \geq 0$.

Claim: For any $n \geq 0$, we have $(2n+1)\gamma - K'(2n\log(2n))^{1/2} \leq \mathbb{E}[L_{2n+1}]$.

Proof of the Claim: We know, from super-additivity of the sequence $\mathbb{E}[L_n]$, that for any $n \geq 0$, we have $\mathbb{E}[L_{2n+1}] \geq \mathbb{E}[L_{2n}]$. Then,

$$\begin{aligned}
\mathbb{E}[L_{2n+1}] \geq \mathbb{E}[L_{2n}] &\geq 2n\gamma - K(2n\log(2n))^{1/2} \\
&\geq 2n\gamma - K(2n\log(2n))^{1/2} + \gamma - \gamma \\
&\geq (2n+1)\gamma - K(2n\log(2n))^{1/2} - \gamma \\
&\geq (2n+1)\gamma - (2n\log(2n))^{1/2}\Big(K + \frac{\gamma}{(2n\log(2n))^{1/2}}\Big) \\
&\geq (2n+1)\gamma - K'(2n\log(2n))^{1/2}.
\end{aligned} \tag{3.11}$$

Since $((2n+1)\log(2n+1))^{1/2} \geq (2n\log(2n))^{1/2}$, we conclude that, for any $n \geq 0$, from from the claim and from the even case, we have

$$n\gamma - K'\sqrt{n\log(n)} \leq \mathbb{E}[L_n] \leq n\gamma.$$

The argument given above is adapted to different cases from score functions to hidden Markov models.

The function $S : \mathcal{A} \times \mathcal{A} \to \mathbb{R}^+$ that assigns a score to each couple from the finite alphabet $\mathcal{A}$ is called a pairwise scoring function. Suppose S is symmetric and let

$$F := \max_{a,b\in\mathcal{A}} S(a,b), \qquad A := \max_{a,b,c\in\mathcal{A}} |S(a,b) - S(a,c)|.$$

An alignment is a pair of increasing natural number sequences, say (π, μ) where $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ where $1 \leq \pi_1 < \pi_2 < \cdots < \pi_k \leq n$ and $1 \leq \mu_1 < \mu_2 < \cdots < \mu_k \leq n$. In this case, we say that k is the number of aligned letters and $n - k$ is the number of gaps in the alignment. We can also set a gap price, say γ. Given the scoring function S and a gap price Γ, the score of the alignment (π, μ) when aligning X and Y is given by

$$U_{(\pi,\mu)}(X,Y) := \sum_{i=1}^{k} S(X_{\pi_i}, Y_{\mu_i}) + \Gamma(n-k).$$

The optimal score for X and Y is defined as the best score over all possible alignments.

Observe that for $S(a,b) = 1$ when $a = b$ and $S(a,b) = 0$ when $a \neq b$ and $\Gamma = 0$, the optimal score is equivalent to the length of the longest common subsequence of X and Y.

In [12], Lember, Matzinger and Torres showed that Alexander's theorem can be improved by using the score function setting above. More specifically, they showed that for even $n \in \mathbb{N}$, and for any $c > \sqrt{A}$,

$$\gamma - \mathbb{E}\Big[\frac{L_n}{n}\Big] \leq c\sqrt{\frac{2}{n-1}\Big(\frac{n+1}{n-1} + \log(n-1)\Big)} + \frac{F}{n-1}.$$

This result generalizes Alexander's theorem in [11] since it is valid for not only longest common subsequences but also for any score function. Moreover, when this theorem is applied to the longest common subsequences, it gives a sharper estimate. In [11], the theorem was valid for $c > 3.42$, but in this we have $c > \sqrt{2}$.

More recent works support the conjecture. In [13], Gong, Houdré and Işlak extend the result to arbitrary alphabets and multisequences, using a similar score function setting. They also showed that Alexander's theorem is also valid when L_n is the length of the longest common subsequence of more than two sequences.

Finally, in [14], Houdré and Kerchev extend this result to hidden Markov models.

3.3. Variance of the Longest Common Subsequence

Despite the popularity of longest common subsequence problem, we do not know much about $Var(L_n)$ and its asymptotic order, even when the size of the alphabet is 2 and the distribution is Bernoulli with a parameter 1/2. In [3], Chvatal and Sankoff conjectured that $Var(L_n) = o(n^{2/3})$. Later, in [15], Steele altered Efron-Stein Inequality

An Efron Stein Inequality for Non-Symmetric Functions: [15] Let S be any function of n variables, X_i and $\tilde{X}_i$ be independent random variables with the same distribution for any $i \in [n]$ and let

$$S := S(X_1, X_2, \ldots, X_n)$$

and

$$S_i := S(X_1, X_2, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n).$$

Then,

$$Var(S) \leq \frac{1}{2}\mathbb{E}\Big[\sum_{i=1}^{n}(S - S_i)^2\Big].$$

Applying this inequality to the variance of the longest common subsequences, Steele showed that $Var(L_n) \leq 2\epsilon_\alpha(1 - \epsilon_\alpha)n$, where $\epsilon_\alpha = \mathbb{P}(X_i = \alpha) = \mathbb{P}(\tilde{X}_i = \alpha)$. Lastly in [16], Lember and Matzinger showed that the order of the standard deviation of L_n is $\sqrt{n}$ if the parameter of the Bernoulli variables is small enough.

In [17], Liu and Houdré studied the statistical behavior of the longest common subsequences in variance using Monte-Carlo approach. When the size of the alphabet is 2, they found the following result for n ranging from $50,000 : 50,000 : 1,000,000$:

- if $p_1 = 0.5$ and $p_2 = 0.5$, then $Var(L_n) \approx 0.0297n^{0.9080}$,
- if $p_1 = 0.1$ and $p_2 = 0.9$, then $Var(L_n) \approx 0.0208n^{0.9855}$,
- if $p_1 = 0.01$ and $p_2 = 0.99$, then $Var(L_n) \approx 0.0042n^{1.0021}$,

where $p_1 = \mathbb{P}(X = 0)$ and $p_2 = \mathbb{P}(X = 1)$. They also conjectured that $Var(L_n) \sim cn$, where c is a small constant.

4. LONGEST INCREASING SUBSEQUENCES IN RANDOM PERMUTATIONS

4.1. Longest Increasing Subsequences

Let $\mathcal{A}$ be some linearly ordered alphabet and $\mathbb{X} = \{X_i\}_{i=1}^n$ be an independent and identically distributed sequence where $n \geq 1$. For any $i \in [n]$, assume that $X_i \in \mathcal{A}$. The length of longest increasing subsequence of $\mathbb{X}$ is given by a random variable defined by

$$I_n := \max\{k : X_{i_1} < X_{i_2} < \cdots < X_{i_k}\}$$

where the maximum is taking taken all subsequences of $1 \leq i_1 < i_2 < \cdots < i_k \leq n$.

Example 4.1. Let $\mathbb{X} = (ARMADILLO)$ and let $\mathbb{Y}$ be the ASCII code representation of the letters of $\mathbb{X}$, i.e. $\mathbb{Y} = (065 \quad 082 \quad 077 \quad 065 \quad 068 \quad 073 \quad 076 \quad 076 \quad 079)$. Then $I_9(\mathbb{Y}) = 5$ since the longest increasing subsequence of $\mathbb{Y}$ is $(065 \quad 068 \quad 073 \quad 076 \quad 079)$.

Using symmetricity of longest increasing subsequences and longest decreasing subsequences, one can easily find a lower bound for the expected value of the length of longest increasing subsequences. Define I'_n to be the length of longest decreasing subsequence. Then, using Erdös-Szekeres theorem, we get

$$\max(I_n, I'_n) \geq \sqrt{n}.$$

Thus, we have $I_n + I'_n \geq \sqrt{n}$. From symmetry, we have $\mathbb{E}[I_n] = \mathbb{E}[I'_n]$. Therefore, $\mathbb{E}[I_n] \geq \frac{\sqrt{n}}{2}$.

In order to find an upper bound for $\mathbb{E}[I_n]$, we need the following lemma:

Lemma 4.2. *[8]*

Proof. First of all, observe that there are $\binom{n}{k}$ subsequences with length k. Moreover, each of these subsequences is monotone increasing with probability $\frac{1}{k!}$ since the probability that the first entry is strictly smaller than the second entry is $\frac{1}{k}$, the probability that the second entry is strictly smaller than the third entry given that the first one is strictly smaller than the second one is $\frac{1}{k-1}$, and so on. Thus, from Boole's inequality, we have $\mathbb{P}(I_n \geq k) \leq \binom{n}{k} / k!$. Let $k = \lceil 2e\sqrt{n} \rceil$. Then,

$$\mathbb{P}(I_n \geq k) \leq \frac{\binom{n}{k}}{k!} = \frac{n(n-1)\dots(n-k+1)}{k!k!} \leq \frac{n^k}{k!k!}.$$

Now, using Stirling's formula, we get

$$\mathbb{P}(I_n \geq k) \leq \frac{n^k}{k!k!} \leq \frac{e^k e^k n^k}{k^k k^k} \leq \left(\frac{e^2 n}{k^2}\right)^k \leq \left(\frac{e^2 n}{(2e\sqrt{n})^2}\right)^{2e\sqrt{n}}.$$

Observe that

$$\left(\frac{e^2 n}{(2e\sqrt{n})^2}\right)^{2e\sqrt{n}} \leq \left(1/4\right)^{2e\sqrt{n}} < \exp(-2e\sqrt{n}).$$

So, the proof is done. □

We are now ready to prove the upper bound. We know, from [7], that for any non-negative random variable I_n and $k > 0$, we have

$$\mathbb{E}[I_n] \leq k + n\mathbb{P}(I_n \geq k).$$

Combining this with the above lemma, we get the following inequality:

$$\begin{aligned} \mathbb{E}[I_n] &\leq 2e\sqrt{n} + n\mathbb{P}(I_n \geq 2e\sqrt{n}) \\ &\leq 2e\sqrt{n} + n\exp(-2e\sqrt{n}) \\ &\leq 2e\sqrt{n} + \frac{n}{e^{2e\sqrt{n}}}. \end{aligned} \tag{4.1}$$

This means that for all $c > 2e$ and $n \geq n(c)$ sufficiently large, we have

$$\mathbb{E}[I_n] \leq c\sqrt{n}.$$

4.2. Random Permutations and the Relationship Between LCS and LIS

At first glance, longest common subsequence and longest increasing subsequence might seem irrelevant. However, when we consider sequences as uniform random permutations, not as strings, they are closely related.

Let $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ be a permutation from S_n that moves an object from plae i to place $\pi(i)$. Last year, in [18], Houdré and Işlak showed that

$$L_n((1, 2, \ldots, n), (\pi_1, \pi_2, \ldots, \pi_n)) = I_n(\pi_1, \pi_2, \ldots, \pi_n) \tag{4.2}$$

where $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ is a random permutation from S_n. In fact, instead of identity permutation, we can take any fixed permutation from S_n.

Proposition 4.3. *[18] Let $\rho = (\rho_1, \rho_2, \ldots, \rho_n) \in S_n$ be fixed. Let also $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ be a uniformly random permutation in S_n. Then,*

$$L_n((\rho_1, \rho_2, \ldots, \rho_n), (\pi_1, \pi_2, \ldots, \pi_n)) =_d I_n(\pi_1, \pi_2, \ldots, \pi_n).$$

Proof. Since $\rho, \pi \in S_n$, $\pi' := \pi\rho$ is also a uniform random permutation of S_n. So, we have

$$L_n((\rho_1, \rho_2, \ldots, \rho_n), (\pi_1, \pi_2, \ldots, \pi_n)) =_d L_n((\rho_1, \rho_2, \ldots, \rho_n), (\pi_1', \pi_2', \ldots, \pi_n')).$$

Since $\pi_i' := \pi\rho_i = \pi_{\rho_i}$, we also have

$$L_n((\rho_1, \rho_2, \ldots, \rho_n), (\pi_1', \pi_2', \ldots, \pi_n')) = L_n((\rho_1, \rho_2, \ldots, \rho_n), (\pi_{\rho_1}, \pi_{\rho_2}, \ldots, \pi_{\rho_n})).$$

From (4.2), the proof is done since;

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

and

$$L_n((1, 2, \ldots, n), (\pi_1, \pi_2, \ldots, \pi_n)) =_d I_n(\pi_1, \pi_2, \ldots, \pi_n).$$

□

In [18], Houdré and Işlak also mentions about

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

Moreover, the distributional asymptotic result is given in their paper by interpreting the results of [19]

$$\frac{L_n(\pi, \rho) - 2\sqrt{n}}{n^{1/6}} \to_d F_2$$

where F_2 is the Tracy-Widom distribution with cumulative distribution function $F_2(x) = \exp(-\int_t^\infty (x-t)u^2(x)dx)$ and u is the solution to the Painlevé II equation.

Proposition 4.4. *If π_1, π_2 are independent uniform random permutations from S_n, and if $x \in \mathbb{R}$, then we have*

$$\mathbb{P}(L_n(\pi_1, \pi_2) \leq x) = \mathbb{P}(I_n(\pi_1) \leq x).$$

Proof. Observe that, we have

$$\begin{aligned}\mathbb{P}(L_n(\pi_1,\pi_2)\le x) &= \sum_{\gamma\in S_n}\mathbb{P}(L_n(\pi_1,\gamma)\le x|\pi_2=\gamma)\mathbb{P}(\pi_2=\gamma)\\ &= \frac{1}{n!}\sum_{\gamma\in S_n}\mathbb{P}(L_n((\pi_1,\pi_2,\dots,\pi_n),(\gamma_1,\gamma_2,\dots,\gamma_n))\le x)\\ &= \frac{1}{n!}\sum_{\gamma\in S_n}\mathbb{P}(I_n(\pi_1)\le x)\\ &= \mathbb{P}(I_n(\pi_1)\le x)\end{aligned} \tag{4.3}$$

where the third equality comes from (4.2). □

Now, since we know that for a fixed permutation $\rho\in S_n$ and a uniform random $\pi\in S_n$, we have $L_n(\rho,\pi)=I_n(\pi)$, it is natural to work on the case when π and ρ are not necessarily uniform but still i.i.d..

Let, for some $n\in\mathbb{N}$, $\pi_1,\pi_2\in S_n$ be random permutations, $\rho_1,\rho_2\in S_n$ be deterministic permutations sampled from a distribution $P=(p_i)_{i\in[n!]}$, where $\mathbb{P}_P(\pi=\rho_i)=p_i$. Then, we can represent expected value of the longest common subsequence of π_1 and π_2 as

$$\mathbb{E}_p[L_n(\pi_1,\pi_2)] = \sum_{i,j\in[n!]} p_i L_n(\rho_i,\rho_j)p_j = \sum_{i,j\in[n!]} p_i l_{ij} p_j = P^T L^{(n)} P$$

where $L^{(n)} := l_{ij} = L_n(\rho_i,\rho_j)$ and $\{l_{ij}\}_{(i,j)\in[n!]\times[n!]}$. It is obvious that for any $i,j\in[n!]$, $l_{ij}=l_{ji}$ and $l_{ii}=n$. So, $L^{(n)}$ is symmetric.

Lemma 4.5. *[20] For $n=2,3$, the matrix $L^{(n)}$ is positive semi-definite.*

Proof. It is easy to see that $L^{(2)}=\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}$. So, the characteristic function of $L^{(2)}$ is $\lambda^2-4\lambda+3$, which means $\lambda_1^{(2)}=1$ and $\lambda_2^{(2)}=3$. Thus, $L^{(2)}$ is positive semi-definite.

For $n = 3$, let us enumerate S_3 with the following order:

$$\{[123], [132], [312], [213], [231], [321]\}.$$

Then, it is easy to see that

$$L^{(3)} = \begin{bmatrix} 3 & 2 & 2 & 2 & 2 & 1 \\ 2 & 3 & 2 & 2 & 1 & 2 \\ 2 & 2 & 3 & 1 & 2 & 2 \\ 2 & 2 & 1 & 3 & 2 & 2 \\ 2 & 1 & 2 & 2 & 3 & 2 \\ 1 & 2 & 2 & 2 & 2 & 3 \end{bmatrix}.$$

A quick computation will give us the characteristic polynomial of $L^{(3)}$ as

$$\lambda^6 - 18\lambda^5 + 84\lambda^4 - 152\lambda^3 + 96\lambda^2 = \lambda^2(\lambda - 2)^3(\lambda - 12).$$

This means that, $\lambda_1^{(3)} = 0$, $\lambda_2^{(3)} = 2$ and $\lambda_3^{(3)} = 12$, and $L^{(3)}$ is positive semi-definite. □

Lemma 4.6. *[20] For $n \geq 4$, the smallest eigenvalue $\lambda_1^{(n)}$ of $L^{(n)}$ is negative.*

Proof. We know that $1 = \lambda_1^{(2)} > \lambda_1^{(3)} = 0$. So, if we show that $\lambda_1^{(k)} > \lambda_1^{(k+1)}$ for some $k \geq 3$, then we are done. In order to find the relationship between $L^{(k)}$ and $L^{(k+1)}$, we will use the following enumeration: Insert the new element $(k+1)$ into the permutation of S_k in $(k+1)$ different places, respectively. For instance, if S_2 is enumerated as $\{[12], [21]\}$, then the first enumeration of S_3 is $\{[123], [213]\}$, the second is $\{[132], [231]\}$, and the last one is $\{[312], [321]\}$. Then, overall, we have $\{[123], [213], [132], [231], [312], [321]\}$.

Using this, observe that the longest common subsequence of any two permutations, say $\pi_1, \pi_2 \in S_k$, will increase by 1 when we add the new element $(k+1)$ to π_1, π_2 to make $\pi_1', \pi_2' \in S_{k+1}$, respectively, since $\pi_1' = \pi_1(k+1)$ and $\pi_2' = \pi_2(k+1)$.

Since the $k! \times k!$ principal minor of $L^{(k+1)}$ is $L^{(k)} + E^{(k)}(E^{(k)})^T$, where $E^{(k)} \in \mathbb{R}^{k!}$ is a vector that consists of only ones.

As an example consider $L^{(2)}$ and $L^{(3)}$: the 2×2 principal minor of $L^{(3)}$ is

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}^T = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}.$$

Claim: The vector $E^{(k)}$ is an eigenvector of $L^{(k)}$.

Proof of the Claim: Since $E^{(k)} \in \mathbb{R}^{k!}$ is a vector that consits of only ones, it is enough to show that the sum of rows of $L^{(k)}$ are equal to each other. Observe that the sum of the π_i-indexed row of $L^{(k)}$ is

$$\sum_{j \in [k!]} L(\pi_i, \pi_j) = \sum_{j \in [k!]} L(id, \pi_i^{-1}\pi_j)$$

since changing π_i and π_j at the same time does not change the length of the longest common subsequence. Moreover, from [20]

$$\sum_{j \in [k!]} L(id, \pi_i^{-1}\pi_j) = \sum_{j \in [k!]} I(\pi_i^{-1}\pi_j).$$

So, the sum of the any row is equal to $\sum_{j \in [k!]} I(\pi_i^{-1}\pi_j) = \sum_{\pi \in S_k} I(\pi)$. Thus, the sum of rows of $L^{(k)}$ are equal to each other and $E^{(k)}$ is an eigenvector of $L^{(k)}$. However, since any longest increasing subsequence is non-negative, this eigenvalue is distinct from our smallest eigenvalue $\lambda_1^{(k)} \leq 0$.

Let $E^{(k)}$ and $R_1^{(k)}$ be eigenvectors associated with eigenvalues $\sum_{\pi \in S_k} I(\pi)$ and $\lambda_1^{(k)}$, respectively. Since we know that $L^{(k)}$ is symmetric, eigenvectors should be orthogonal, i.e.

$$(E^{(k)})^T R_1^{(k)} = 0.$$

WLOG, let $R_1^{(k)}$ be a unit vector. Then, $L^{(k)}R_1^{(k)} = \lambda_1^{(k)}R_1^{(k)}$ implies that $\lambda_1^{(k)} = \left(R_1^{(k)}\right)^T L^{(k)}\left(R_1^{(k)}\right)$. Let $g = \left(R_1^{(k)}\right)^T\left(L^{(k)} + E^{(k)}(E^{(k)})^T\right)\left(R_1^{(k)}\right)$. Observe that $\left(L^{(k)} + E^{(k)}(E^{(k)})^T\right)$ is the $k! \times k!$ principal minor of $L^{(k+1)}$. Let $\begin{bmatrix} R_1^{(k)} \\ 0 \end{bmatrix} \in \mathbb{R}^{(k+1)!}$. Then, from the orthogonality of eigenvectors, we have

$$\left(R_1^{(k)}\right)^T E^{(k)} = \begin{bmatrix} R_1^{(k)} \\ 0 \end{bmatrix}^T E^{(k+1)} = 0.$$

Moreover, since $R_1^{(k)}$ was a unit vector, $\begin{bmatrix} R_1^{(k)} \\ 0 \end{bmatrix}^T$ is also a unit vector. So,

$$\begin{bmatrix} R_1^{(k)} \\ 0 \end{bmatrix}^T L^{(k+1)} \begin{bmatrix} R_1^{(k)} \\ 0 \end{bmatrix}^T \geq \min_{R^T E = 0; \|R\| = 1} R^T L^{(k+1)} R = \lambda_1^{(k+1)}.$$

Moreover, this inequality holds if and only if $\begin{bmatrix} R_1^{(k)} \\ 0 \end{bmatrix}^T$ is an eigenvector of $L^{(k+1)}$ associated with $\lambda_1^{(k+1)}$. Assume, for contradiction, this is the case. Then, we have

$$L^{(k+1)} \begin{bmatrix} R_1^{(k)} \\ 0 \end{bmatrix} = \lambda_1^{(k+1)} \begin{bmatrix} R_1^{(k)} \\ 0 \end{bmatrix}^T.$$

Now, the $k! \times k!$ principal minor of $L^{(k+1)}$ is $L^{(k)}$ and we can represent this submatrix as $\{[(k+1)\pi_i]\}_{i \in [k!]}$ or as $\{[\pi_i(k+1)]\}_{i \in [k!]}$. Then, $(i,j)^{th}$ entry of this submatrix is

$$L\left([(k+1)\pi_i], [\pi_j(k+1)]\right) = L(\pi_i, \pi_j)$$

since $(k+1)$ can only be in the longest common subsequence if the left hand side is 1. Now, observe that the vector made of the bottom $k!$ elements of $L^{(k+1)} \begin{bmatrix} R_1^{(k)} \\ 0 \end{bmatrix}$ is $L^{(k)}R_1^{(k)} = \lambda_1^{(k)}R_1^{(k)}$, and this is a non-zero vector.

However, the vector made of the bottom $k!$ elements of $\lambda_1^{(k+1)} \begin{bmatrix} R_1^{(k)} \\ 0 \end{bmatrix}$ is zero vector.

Thus, this is a contradiction and we have

$$1 = \lambda_1^{(2)} > 0 = \lambda_1^{(3)} > \lambda_1^{(4)} > \cdots .$$

□

Theorem 4.7. *[20] Let $n \geq 1$ and $\pi_1, \pi_2 \in S_n$ be two i.i.d. random permutations sampled from a distribution P. Then, for $n \leq 3$, the uniform distribution U minimizes $\mathbb{E}[L(\pi_1, \pi_2)]$, while, for $n \geq 4$, U is sub-optimal.*

Proof. From previous lemma, we know that $\mathbb{E}_P[L(\pi_1, \pi_2)] = P^T L P$. Since we also know that U is an eigenvector of L and that $P^T U = 1$, we have

$$\mathbb{E}_P[L(\pi_1, \pi_2)] = (P - U)^T L (P - U) + 2P^T LU - U^T U.$$

Observe that $P^T LU = U^T LU$, so we have

$$\mathbb{E}_P[L(\pi_1, \pi_2)] = (P - U)^T L (P - U) + U^T LU.$$

Let $n \leq 3$. Then, $L^{(n)}$ is positive semi-definite and $(P - U)^T L (P - U) \geq 0$. Thus, $P^T LP \geq U^T LU$.

Now let $n > 3$. We know that the smallest eigenvalue of L, $\lambda_1^{(n)}$, is negative. Let $R_1^{(n)}$ be the eigenvector for L that is associated with $\lambda_1^{(n)}$. Then $U^T R_1^{(n)} = 0 = E^T R_1^{(n)}$, where E is a vector that consists of only ones. Thus, there exists a positive constant c such that $cR_1^{(n)} \succeq \frac{-1}{n!}$. Let P_0 be such that $P_0 - U = cR_1^{(n)}$. Then

$$E^T P_0 = E^T (U + cR_1^{(n)}) = 1 + 0 = 1$$

$$U + cR_1^{(n)} \succeq 0$$

Therefore, P_0 is a well-defined distribution on S_n. However, the expected value of the length of the longest common subsequence of π_1, π_2 under P_0 is

$$\begin{aligned}\mathbb{E}_{P_0}[L(\pi_1,\pi_2)] &= (P_0 - U)^T L(P_0 - U) + U^T LU \\ &= c^2 (R_1^{(n)})^T L(R_1^{(n)}) + U^T LU \\ &= c^2 \lambda_1^{(n)} + U^T LU \\ &< U^T LU\end{aligned} \tag{4.4}$$

Since $U^T LU$ is the expected value of the length of the longest common subsequence of π_1, π_2 under the uniform distribution, i.e. $\mathbb{E}_U[L(\pi_1,\pi_2)]$, we conclude that for $n > 3$, U is sub-optimal. □

Now, we will focus on some properties of the matrix $L^{(n)}$. From the previous part, we know that $\sum_{\pi \in S_n} I(\pi)$ is an eigenvalue of $L^{(n)}$. Moreover, the eigenvector associated with this eigenvector is $E^{(n)}$, which consists of only ones. This eigenvalue is, indeed, the radius of the spectrum of the matrix $L^{(n)}$.

Proposition 4.8. *[20] The radius of the spectrum of the matrix $L^{(n)}$ is the eigenvector $\sum_{\pi \in S_n} I(\pi)$.*

Proof. Let λ be an eigenvalue of $L^{(n)}$ and R be the eigenvector of $L^{(n)}$ associated with λ. WLOG, suppose $\max_{i \in [n!]} |r_i| = 1$, where $R = (r_1, r_2, \ldots, r_{n!})^T$ and let $|r_{i_0}| = 1$, for some $i_0 \in [n!]$. Then, since $L^{(n)} R = \lambda R$, we have,

$$|\lambda| = |\lambda r_{i_0}| = \Big| \sum_{j \in [n!]} L(\pi_{i_0}, \pi_j) r_j \Big|.$$

Since $\max_{i \in [n!]} |r_i| = 1$, we have

$$\Big| \sum_{j \in [n!]} L(\pi_{i_0}, \pi_j) r_j \Big| \leq \sum_{j \in [n!]} L(\pi_{i_0}, \pi_j) = \sum_{j \in [n!]} I(\pi_{i_0}^{-1} \pi_j) = \sum_{\pi \in S_n} I(\pi).$$

□

So, the smallest negative value of $\lambda_1^{(n)}$ is bounded from below by $-\sum_{\pi \in S_n} I(\pi)$. Moreover, since $\mathbb{E}[I(\pi)] \sim -2\sqrt{n}$, where π is a uniform random permutation, this result gives an asymptotic order of $-n!\sqrt{n}$. However, we are interested in an upper bound of $\lambda_1^{(n)}$. So, the next proposition will show the decrease rate of $\lambda_1^{(n)}$.

Proposition 4.9. *[20] For any $n \geq 4$, we have $\lambda_1^{(n)} \leq 2^{n-4}\lambda_1^{(4)} = -2^{n-3} < 0$.*

Proof. Instead of the above inequality, we will prove $\lambda_1^{(n+1)} \leq 2\lambda_1^{(n)}$. We know that

$$\lambda_1^{(n)} = \min_{E^T R = 0} \frac{R^T L^{(n+1)} R}{R^T R}. \tag{4.5}$$

Let $R^{(n)}$ be the eigenvector of $L^{(n)}$ that corresponds to the smallest eigenvalue $\lambda_1^{(n)}$ of $L^{(n)}$. Then, from Lemma 4.5, we know that $n! \times n!$ principal minor of $L^{(n+1)}$ is $L^{(n)} + EE^T$, while its bottom-left $n! \times n!$ submatrix is $L^{(n)}$. Since $L^{(n+1)}$ is symmetric, this means that the top right $n! \times n!$ submatrix of $L^{(n+1)}$ is $L^{(n)}$ and the bottom left $n! \times n!$ submatrix of $L^{(n+1)}$ is $L^{(n)} + EE^T$, i.e.

$$L^{(n+1)} = \begin{bmatrix} L^{(n)} + EE^T & \cdots & L^{(n)} \\ \vdots & \ddots & \vdots \\ L^{(n)} & \cdots & L^{(n)} + EE^T \end{bmatrix}.$$

Let also $R = \left[R_1^{(n)} 0 \ldots 0 R_1^{(n)}\right]^T$. Then, $E^T R = E^T R_1^{(n)} + E^T R_1^{(n)}$ and $||R||^2 = R^T R = 2||R_1^{(n)}|| = 2$. So, since $R^T R = 2$, from (4.4), we get that $2\lambda_1^{(n+1)} = R^T L^{(n+1)} R$ and we have the following equality.

$$R^T L^{(n+1)} R = \begin{bmatrix} R_1^{(n)} \\ 0 \\ \vdots \\ 0 \\ R_1^{(n)} \end{bmatrix}^T \begin{bmatrix} L^{(n)} + EE^T & \cdots & L^{(n)} \\ \vdots & \ddots & \vdots \\ L^{(n)} & \cdots & L^{(n)} + EE^T \end{bmatrix} \begin{bmatrix} R_1^{(n)} \\ 0 \\ \vdots \\ 0 \\ R_1^{(n)} \end{bmatrix}$$

$$= 2(R_1^{(n)})^T (L^{(n)} + EE^T)(R_1^{(n)}) + 2(R_1^{(n)})^T L^{(n)} (R_1^{(n)})$$

$$= 4(R_1^{(n)})^T L^{(n)} (R_1^{(n)})$$

$$= 4\lambda_1^{(n)}.$$

Therefore, $\lambda_1^{(n+1)} \leq 2\lambda_1^{(n)}$.

□

Lemma 4.10. *[20] Let π_1, π_2, π_3 be any permutations in S_n. Then,*

$$L_n(\pi_1, \pi_2) L_n(\pi_1, \pi_3) L_n(\pi_2, \pi_3) \geq n.$$

Proof. First of all, Houdré and Işlak's proof can be checked in [20].

As a second proof, let $n = 1$. There is only one permutation in S_1, which means for any $\pi_1, \pi_2, \pi_3 \in S_1$, we have $L_1(\pi_1, \pi_2)L_1(\pi_1, \pi_3)L_1(\pi_2, \pi_3) \geq 1$. Similarly, for $n = 2$, there are two permutations. So, at least two of the $\pi_1, \pi_2, \pi_3 \in S_2$, should be equal. Thus we have, $L_2(\pi_1, \pi_2)L_2(\pi_1, \pi_3)L_2(\pi_2, \pi_3) \geq 2$. Now, suppose the statement holds for some $k \in \mathbb{N}$, and assume, for contradiction, that it does not hold for $k+1$.

Let $\pi_1, \pi_2, \pi_3 \in S_k$, and define $\pi_i' \in S_{k+1}$, by adding $(k+1)$ at the end of each permutation for $i = 1, 2, 3$. Then, clearly, we have $L_{k+1}(\pi_i', \pi_j') = L_k(\pi_i, \pi_j) + 1$, for any $i, j \in \{1, 2, 3\}$. Thus,

$$\begin{aligned} L_{k+1}(\pi_1', \pi_2')L_{k+1}(\pi_1', \pi_3')L_{k+1}(\pi_2', \pi_3') &= (L_k(\pi_1, \pi_2) + 1)(L_k(\pi_1, \pi_3) + 1)(L_k(\pi_2, \pi_3) + 1) \\ &\geq k + 1 \end{aligned} \tag{4.6}$$

which gives a contradiction since we assumed $L_{k+1}(\pi_1', \pi_2')L_{k+1}(\pi_1', \pi_3')L_{k+1}(\pi_2', \pi_3') < (k+1)$. □

Theorem 4.11. *[20] Let π_1, π_2 be two arbitrary i.i.d. random permutations sampled from an arbitrary distribution P on S_n. Then, $\mathbb{E}[L(\pi_1, \pi_2)] \geq \sqrt[3]{n}$.*

Proof. Take $\sigma_1, \sigma_2, \sigma_3 \in S_n$ and define $l(\sigma_i) := \sum_{\sigma_1 \in S_n} P(\sigma_1)L(\sigma_1, \sigma_i)$, where $i = 1, 2, 3$. Then,

$$\begin{aligned} l(\sigma_2) + L(\sigma_2, \sigma_3) + l(\sigma_3) &= \sum_{\sigma_1 \in S_n} \big(P(\sigma_1)L(\sigma_1, \sigma_2)\big) + L(\sigma_2, \sigma_3) + \sum_{\sigma_1 \in S_n} \big(P(\sigma_1)L(\sigma_1, \sigma_3)\big) \\ &= \sum_{\sigma_1 \in S_n} P(\sigma_1)\Big(L(\sigma_1, \sigma_2) + L(\sigma_1, \sigma_3) + L(\sigma_2, \sigma_3)\Big). \end{aligned} \tag{4.7}$$

Then, from [21], we know that the arithmetic mean of three real numbers is always greater then their geometric mean. So, we have

$$\frac{1}{3}\big(L(\sigma_1, \sigma_2) + L(\sigma_1, \sigma_3) + L(\sigma_2, \sigma_3)\big) \geq \sqrt[3]{(L(\sigma_1, \sigma_2)L(\sigma_1, \sigma_3)L(\sigma_2, \sigma_3)}.$$

Therefore, using Lemma 4.10. we get

$$\begin{aligned} l(\sigma_2) + L(\sigma_2, \sigma_3) + l(\sigma_3) &\geq 3 \sum_{\sigma_1 \in S_n} P(\sigma_1) \sqrt[3]{(L(\sigma_1, \sigma_2) L(\sigma_1, \sigma_3) L(\sigma_2, \sigma_3)} \\ &\geq 3 \sum_{\sigma_1 \in S_n} P(\sigma_1) \sqrt[3]{n} \\ &= 3\sqrt[3]{n}. \end{aligned} \tag{4.8}$$

Moreover, summing the above equality over $P(\sigma_2)$ gives

$$\sum_{\sigma_2 \in S_n} P(\sigma_2) \Big(l(\sigma_2) + L(\sigma_2, \sigma_3) + l(\sigma_3) \Big) = \sum_{\sigma_2 \in S_n} \big(P(\sigma_2) l(\sigma_2) \big) + l(\sigma_3) + l(\sigma_3) \geq 3\sqrt[3]{n}.$$

Repeating the last step with weights over $P(\sigma_3)$ gives

$$\sum_{\sigma_2 \in S_n} P(\sigma_2) + 2 \sum_{\sigma_3 \in S_n} P(\sigma_3) l(\sigma_3) = 3 \sum_{\sigma \in S_n} P(\sigma) l(\sigma) \geq 3\sqrt[3]{n}.$$

However,

$$\begin{aligned} \mathbb{E}_P[L(\sigma_1, \sigma_2)] &= \sum_{\sigma_1 \in S_n} \sum_{\sigma_2 \in S_n} P(\sigma_1) L(\sigma_1, \sigma_2) P(\sigma_2) \\ &= \sum_{\sigma_1 \in S_n} P(\sigma_1) \sum_{\sigma_2 \in S_n} L(\sigma_1, \sigma_2) P(\sigma_2) \\ &= \sum_{\sigma \in S_n} P(\sigma) L(\sigma) \\ &\geq \sqrt[3]{n}. \end{aligned} \tag{4.9}$$

□

5. CONCLUSION

In this thesis, we investigate the properties of the longest common subsequences. Our aim is to understand the details of the theory of the longest common subsequences that become popular in 1970's, draw attention to the progress about the longest common subsequences in the recent studies, and state some open problems about the subject.

In the second and third chapters, some basic definitions and results about probability theory is given and the theory of the longest common subsequence is introduced. Two of the most important points about longest common subsequences, namely Rhee's argument and Alexander's theorem, are discussed. At the end of Chapter 3, some recent developments and variations of these two subjects are mentioned and some simulations about the variance of longest common subsequences are given. At the last chapter, modern improvements about the longest common subsequences are discussed. More specifically, recently discovered relationship about longest common subsequences and longest increasing subsequences in random permutations is stated and some properties of the matrix $L^{(n)}$, that is generated by the lengths of the longest common subsequences of random permutations are studied.

To sum up, even though longest common subsequences is one of the most studied problems of the discrete probability theory and have lots of applications from computer science to computational biology, there are lots of open problems. First of all, we know that given two random sequences, the limit of the length of the longest common subsequences, divided by the length of the sequence, converges to a constant. However, the value of this constant is unknown. Even if we take an alphabet with size 2 and even if the distribution of the sequences are $Ber(1/2)$, we do not know the value of this constant. Moreover, asymptotic order of the variance of the length of the longest common subsequence is also unknown.

In random permutation setting, one other open problem was introduced by Bukh and Zhou in [22]:

Conjecture: Let P be an arbitrary probability distribution on S_n. Let σ_1, σ_2 be two independent and identically distributed permutations sampled from P. Then, $\mathbb{E}_P[L(\sigma_1, \sigma_2)] \geq \sqrt{n}$. It might even be true that the uniform distribution U on S_n gives a minimizer.

In [20], with Lemma 4.11, Houdré and Xu showed that $\mathbb{E}_P[L(\sigma_1, \sigma_2)] \geq \sqrt[3]{n}$. However, the conjecture of Bukh and Zhou is still open.

Similarly, in [17], Liu and Houdré studied the statistical behavior of the longest common subsequences in variance using Monte-Carlo approach and they conjectured that if the alphabet size is 2, then $Var(L_n) \sim cn$ where c is a small constant. This remains open as well.

REFERENCES

1. Ning, K., H. K. Ng and H. W. Leong, "Analysis of the Relationships among Longest Common Subsequences, Shortest Common Supersequences and Patterns and its application on Pattern Discovery in Biological Sequences", *CoRR*, Vol. abs/0903.2310, 2009, http://arxiv.org/abs/0903.2310.

2. Sankoff, D., "Matching sequences under deletion/insertion constraints", *Proc. Nat. Acad. Sci. U.S.A.*, Vol. 69, pp. 4–6, 1972.

3. Chvatal, V. and D. Sankoff, "Longest common subsequences of two random sequences", *J. Appl. Probability*, Vol. 12, pp. 306–315, 1975.

4. Steele, J. M., "Long common subsequences and the proximity of two random strings", *SIAM J. Appl. Math.*, Vol. 42, No. 4, pp. 731–737, 1982, https://doi.org/10.1137/0142051.

5. Rinsma-Melchert, I., "The expected number of matches in optimal global sequence alignments", *New Zealand Journal of Botany*, Vol. 31, No. 3, pp. 219–230, 1993, https://doi.org/10.1080/0028825X.1993.10419499.

6. Rhee, W. T., "On rates of convergence for common subsequences and first passage time", *Ann. Appl. Probab.*, Vol. 5, No. 1, pp. 44–48, 1995.

7. Gut, A., *Probability: A Graduate Course*, Springer Texts in Statistics, Springer, New York, 2005.

8. Steele, J. M., *Probability theory and combinatorial optimization*, Vol. 69 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997, https://doi.org/10.1137/1.9781611970029.

9. Auffinger, A., M. Damron and J. Hanson, "Rate of convergence of the mean for sub-additive ergodic sequences", *Adv. Math.*, Vol. 285, pp. 138–181, 2015, https://doi.org/10.1016/j.aim.2015.07.028.

10. Efron, B. and C. Stein, "The jackknife estimate of variance", *Ann. Statist.*, Vol. 9, No. 3, pp. 586–596, 1981.

11. Alexander, K. S., "The rate of convergence of the mean length of the longest common subsequence", *Ann. Appl. Probab.*, Vol. 4, No. 4, pp. 1074–1082, 1994.

12. Lember, J., H. Matzinger and F. Torres, "The rate of the convergence of the mean score in random sequence comparison", *Ann. Appl. Probab.*, Vol. 22, No. 3, pp. 1046–1058, 2012, https://doi.org/10.1214/11-AAP778.

13. Ruoting Gong, C. H. and Ümit Işlak, "A Central Limit Theorem for the Optimal Alignments Score in Multiple Random Words", *(preprint)*, 2016, https://arxiv.org/pdf/1512.05699v2.pdf.

14. Houdré, C. and G. Kerchev, "Rate of convergence for the length of the longest common subsequences in hidden Markov models", *(preprint)*, 2017, https://arxiv.org/pdf/1712.09881v1.pdf.

15. Steele, J. M., "An Efron-Stein inequality for nonsymmetric statistics", *Ann. Statist.*, Vol. 14, No. 2, pp. 753–758, 1986, https://doi.org/10.1214/aos/1176349952.

16. Lember, J. and H. Matzinger, "Standard deviation of the longest common subsequence", *Ann. Probab.*, Vol. 37, No. 3, pp. 1192–1235, 2009, https://doi.org/10.1214/08-AOP436.

17. Liu, Q. and C. Houdré, "Simulations, Computations, and Statistics for Longest Common Subsequences", *(preprint)*, 2017, https://arxiv.org/pdf/1705.06826.pdf.

18. Houdré, C. and Ümit Işlak, "A Central Limit Theorem for the Length of the Longest Common Subsequences in Random Words", *(preprint)*, 2017, https://arxiv.org/pdf/1408.1559.pdf.

19. Baik, J., P. Deift and K. Johansson, "On the distribution of the length of the longest increasing subsequence of random permutations", *J. Amer. Math. Soc.*, Vol. 12, No. 4, pp. 1119–1178, 1999, https://doi.org/10.1090/S0894-0347-99-00307-0.

20. Houdré, C. and C. Xu, "A Note on the Expected Length of the Longest Common Subsequences of two i.i.d. Random Permutations", *(preprint)*, 2017, https://arxiv.org/pdf/1703.07691.pdf.

21. Borwein, J. M. and P. B. Borwein, *Pi and the AGM*, Vol. 4 of *Canadian Mathematical Society Series of Monographs and Advanced Texts*, John Wiley & Sons, Inc., New York, 1998, a study in analytic number theory and computational complexity, Reprint of the 1987 original, A Wiley-Interscience Publication.

22. Bukh, B. and L. Zhou, "Twins in words and long common subsequences in permutations", *Israel J. Math.*, Vol. 213, No. 1, pp. 183–209, 2016, https://doi.org/10.1007/s11856-016-1323-8.