FLUCTUATIONS OF THE LONGEST COMMON SUBSEQUENCE IN THE ASYMMETRIC CASE OF 2- AND 3-LETTER ALPHABETS

Federico Bonetto[*] and Heinrich Matzinger[†]

December 13, 2005

Abstract

We investigate the asymptotic standard deviation of the Longest Common Subsequence (LCS) of two independent i.i.d. sequences of length n. The first sequence is drawn from a three letter alphabet $\{0, 1, a\}$, whilst the second sequence is binary. The main result of this article is that in this asymmetric case, the standard deviation of the length of the LCS is of order $\Theta(\sqrt{n})$. This confirms Waterman's conjecture [21] for this special cases. This is very interesting considering that it is believed that for equal probability of 0 and 1 the order is $o(n^{1/3})$; (see the Sankoff-Chvatal conjecture in [10]).

The order of the fluctuation of the LCS of two i.i.d. binary sequences is a long open standing question. In a subsequent paper, we use the techniques developped in this article to solve this problem when the two sequences are binary, but 0 and 1 have sufficiently different probabilities.

The LCS problems can also be viewed as First Passage Problems (FPP) on a graph with correlated weights. For standard FPP the order of the fluctuations has been open for decades.

1 Introduction

For a sequence a_n, we say that a_n has order $\Theta(n)$, if there exists $k, K > 0$ not depending on n, such that $kn \leq a_n \leq Kn$ for all $n \in \mathbb{N}$.

In computational genetics and computational linguistics one of the basic problem is to find an optimal alignment between two given sequences $X := X_1 \dots X_n$ and $Y := Y_1 \dots Y_n$. This requires a scoring system which can rank the alignments. Typically a substitution matrix gives the score for each possible pair of letters. The

[*] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332. Email: bonetto@math.gatech.edu

[†] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332. Email: matzi@math.gatech.edu

total score of an alignment is the sum of terms for each aligned pair of residues, plus a usually negative term for each gap (gap penalty).

Let us look at an example. Take the sequences X and Y to be binary sequences. Let the substitution matrix be equal to:

$$\begin{array}{c|cc} & 0 & 1 \\ \hline 0 & 2 & 1 \\ 1 & 1 & 3 \end{array}$$

With the above matrix we get the following scores for pairs of letter:

$$s(0,0) = 2, s(0,1) = s(1,0) = 1, s(1,1) = 3.$$

(Here, $s(a,b)$ designates the score when we align letter a with letter b.) Take $X = 0101$ and $Y = 1100$ with the above substitution matrix and a zero gap penalty. The optimal alignment is:

$$\begin{array}{cccccc} 0 & 1 & 0 & 1 & \_ & \_ \\ \_ & 1 & \_ & 1 & 0 & 0 \end{array}$$

The above alignment gives the score $s(1,1) + s(1,1) = 3 + 3 = 6$. This is the alignment with maximal score.

Throughout this paper the substitution matrix is equal to the identity and there is no gap penalty. In this case, the optimal score is equal to the length of the Longest Common Subsequence (LCS) of X and Y. (A common subsequence of X and Y is a sequence which is a subsequence of X as well as of Y.)

LCS and optimal alignments are one of the main tools in computational linguistics. An example of an important application is the creation of large dictionaries for rare languages. Building the dictionarie manualy would necessitat years of work with a large stuff. Hence, one whishes the computer to build the dictionaries. For this one gives translated texts to the computer. An algorithm is then asked to identify corresponding words.
Let us next show how LCS's are used two identify pairs of corresponding words. Take as example two versions of the the first name "henry". Consider the Swiss version "heini" and the spanish version "enrique". When we align the two version and compare letter by letter

h	e	i	n	i		
e	n	r	i	q	u	e

the similarity is not obvious: there are zero coinciding letters in the same position. It follows that the computer is not able to recognize the great similarity of the two strings "heini" and "enrique", when comparing position by position.
Another method is needed to detect the similarity. One useful approach is based on the LCS. The LCS in this case is *eni*. The string *eni* can be obtained from both strings by only deleting letters. The relatively long common subsequence "eni" indicates that the two strings are related.

Let L_n designate the lenght of the LCS of two independent i.i.d. sequences of lenght n. Using a subadditivity argument, Chvatal and Sankoff [10] prove that the limit

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

exists. They consider two binary sequences. (This is the standart setting for this problem). The constant γ is called the Chvatal-Sankoff constant and its value is

unknown. Neither is the exact order of the fluctuation of the LCS lenght known. Steele [20] proved that $VAR[L_n] \leq n$.
The determination of the Chvatal-Sankoff constant and the order of fluctuations for the LCS problem are long standing open problems. Montecarlo simulations lead Chvatal and Sankoff to conjectured that $VAR[L_n] = o(n^{\frac{2}{3}})$. This order of magnitude is similar to the order for the longest increasing subsequence (LIS) of random permutations. (See Baik, Deift and Johansson [9] and also Aldous and Diaconis [1]).
This similarity of the order of magnitudes is not a complete surprise. As a matter of fact, the LCS can be formulated as an oriented First Passage Percolation (FPP) problem with correlated weights. On the other hand, the LIS problem is asymptotically equal to a Poisson-based oriented FPP model. For standart FPP the order of magnitude of the fluctuation has been open for decades despite FPP beeing one of the central research areas in discrete probability.
In [21], Waterman conjectured that in many cases the variance of L_n grows linearly. We believe that there are different possible order of magnitudes depending on the distribution of the strings X and Y.
In the present article, we consider the asymmetric case where X contains one symbol less than Y. For this case, we prove the variance $VAR[L_n]$ to be of order $\Theta(n)$. The same order is proven by Durringer, Lember and Matzinger [16] in the case that one sequence is not random but periodic. In a subsequent paper, we use the methods of this article to prove the same order in yet another case. This case is when we have two i.i.d. binary sequences where 0 and 1 have strongly different probabilities.

As mentioned, the exact value of γ remains unknown. Chvàtal-Sankoff [10] derive upper and lower bounds for γ, and similar upper bounds were found by Baeza-Yates, Gavalda, Navarro and Scheihing [8] using an entropy argument. These bounds have been improved by Deken [13], and subsequently by Dancik-Paterson [12, 18]. In [14], Hauser, Martinez and Matzinger developed a Monte Carlo and large deviation-based method which allows to further improve the upper bounds on γ. Their approach can be seen as a generalization of the method of Dancik-Paterson.
For sequence with many letters, Kiwi, Loebl and Matousek, [15] have the following interesting result:
when both sequences X and Y are drawn from the alphabet $\{1, 2, \ldots, k\}$ and the letters are equiprobable, then $\gamma \to 2/\sqrt{k}$ as $k \to \infty$.
Waterman-Arratia [7] derive a law of large deviation for L_n for fluctuations on scales larger than $\sqrt{n}$. In their ground breaking article [7], they show the existence of a critical phenomena.
Using first passage percolation methods, Alexander [2] proves that $E[L_n]/n$ converges at a rate of order at least $\sqrt{\log n/n}$. In [21], Waterman studies the statistical significance of the results produced by sequence alignment methods.

Another problem related to the LCS-problem is that of comparing sequences X and Y by looking for longest common words that appear both in X and Y, and generalizations of this problem where the words do not need to appear in exactly the same form in the two sequences. (This means that the words are more than common substrings. They need to appear in a continous string without additionnal letters inbetween.) The distributions that appear in this context have been studied by Arratia-Gordon-Goldstein-Waterman [3] and Neuhauser [17]. A crucial role is played by the Chen-Stein Method for the Poisson-Approximation. Arratia-Gordon-Waterman [4, 5] shed some light on the relation between the Erdös-Rényi law for random coin tossing and the above mentioned problem. In [6] the same authors also developed an extreme value theory for this problem.

For a general discussion of the relevance of string comparison for biology and of other similar problem in computational biology the reader can refer to the standard texts [19] and [11].

The reader might wonder why the case considered in the present article is relevant. Three letters in one sequence and two in the other might seem an unrealistic example. Our motivation is the following: in any i.i.d. sequence there are finite patterns (i.e. finite words) which tend to have below-average expected matching scores. The number of times any given finite pattern occurs in $X = X_1 \dots X_n$ is roughly a binomial variable with variance proportional to n. Hence, the number of times we observe a given pattern in Y behaves roughly like the number of a's in Y. The number of a's in Y, decrease the optimal score linearly. For a given finite pattern with low average matching score we hope that the same holds be true. (And we prove it in a subsequent article, when 0 and 1 have very different probabilities.)

2 Main result

Throughout this paper $\{X_i\}_{i\in\mathbb{N}}$ and $\{Y_i\}_{i\in\mathbb{N}}$ are two i.i.d. sequences which are independent of each other and which satisfy all of the following three conditions:

1. The variables $X_i, i \in \mathbb{N}$, have state space $\{0, 1, a\}$.
2. There exists p, $0 < p < 1$ such that
$$P(X_1 = a) = p, \qquad P(X_1 = 0) = P(X_1 = 1) = \frac{1-p}{2}. \tag{2.1}$$
3. The variables $Y_i, i \in \mathbb{N}$, are Bernoulli variables with parameter 1/2.

When all the three conditions above are satisfied, we say we are in *case I* . The main result of this paper is:

Theorem 2.1 *When we are in case I, there exists $k > 0$ not depending on n, such that for all $n \in \mathbb{N}$, we have*
$$VAR[L_n] \geq k \cdot n. \tag{2.2}$$

There is also an upper bound for the variance

$$VAR[L_n] \leq K \cdot n$$

where $K > 0$ is a constant not depending on n. This upper bound follows directly from the large deviation result for LCS of Waterman-Arratia [7]. Let us give this result:

Lemma 2.1 *Assume that we are in case I, then: there exists a constant $c > 0$ (not depending on n and Δ) such that for all n large enough and all $\Delta > 0$, we have that:*

$$P\left(|L_n - E\left[L_n\right]| \geq n\Delta\right) \leq e^{-cn\Delta^2} \tag{2.3}$$

Theorem 2.1 and lemma 2.1 together imply that the typical size of $L_n - E[L_n]$ is $o(\sqrt{n})$. More precisely, let $D_n := (L_n - E[L_n])/\sqrt{n}$ denote the rescaled fluctuation of L_n. Then:

Theorem 2.2 *The sequence $\{D_n\}$ is tight. Moreover, the limit of any weakly convergent subsequence of $\{D_n\}$ is not a Dirac measure.*

Theorem 2.2 is a rather direct consequence of theorem 2.1 and lemma 2.1. We refer the reader to [16] for the proof .

3 Proof of main theorem

Let N^a designate the numbers of a's in the sequence $X = X_1X_2\dots X_n$. Let X^{01} designate the subsequence of X consisting of all the 0's and 1's contained in X. In other words, X^{01} is obtained by removing the a's from the finite sequence X. Thus, X^{01} is a finite sequence of i.i.d. Bernoulli variables with parameter 1/2 with random length. The length of the random binary string X^{01} is equal to $(n - N^a)$.

Let us illustrate this with a practical example. For $n = 6$, assume that $X = 011a0a$ and $Y = 101011$. In this case $N^a = 2$ and $X^{01} = 0110$. Obviously the a's from sequence X can not be matched since Y does not contain any a's. Hence, The length L_6 of the LCS of X and Y is equal to the length of the LCS between X^{01} and Y. The length of the LCS is $L_6 = 3$. There are actual three longest common subsequences: 011, 010 and 110.

The main idea why L_n fluctuates on the scale $\sqrt{n}$ is the following: The binomial variable N^a has variance of order $o(n)$. The variable L_n tends to decrease linearly with an increase of N^a (since the a's are not matched and thus constitute losses). Hence L_n should also fluctuate on the scale $\sqrt{n}$.

To prove this rigorously, we simulate the variable L_n in a special way. We first simulate a variable with same distribution as N^a. (We can call it N^a.) Then we generate X^{01} by using a drop-scheme of random bits. Instead of flipping a coin independently $n - N^a$ times in a row we generate a sequence $Z^1, Z^2, \dots$ of binary strings where Z^k has length k. Z^{k+1} is obtained by adding to Z^k a random bit at a random location.

For example, assume that we have the binary string $Z^6 = 00010$. There are four possible positions where the next bit could come:

position 1	position 2	position 3	position 4
$0x0010$	$00x010$	$000x10$	$0001x0$

where x designates the possible position of the next bit. We assign the same probability to each of the four above possibilities and draw one of them at random. We flip a fair coin, and fill the previously chosen position with the number obtained from the fair coin.

If the position chosen is the second one and the fair coin gives us a 1, then we obtain $Z^7 = 001010$.

We apply this scheme recursively on k and obtain a sequence of random binary strings $Z^1, Z^2, \ldots, Z^n$. Let Z_i^k designate that i-th bit of the k-th string. With that notation:

$$Z^k = Z_1^k Z_2^k \ldots Z_k^k.$$

Hence, $\{Z_i^k\}_{i \leq k \leq n}$ is a triangular array of Bernoulli variables. Let us next define the Z^k's in a formal way: let $V_k, k \in \mathbb{N}$ be a sequence of i.i.d. Bernoulli variables with parameter $1/2$. Let $T_k, k \in \mathbb{N}$ be a sequence of independent integer variables, so that $\{V_k\}_{k \in \mathbb{N}}$ is independent of $\{T_k\}_{k \in \mathbb{N}}$. Furthermore, for $k \in \mathbb{N}$, let the distribution of T_{k+1} be the uniform distribution on the set $\{2, \ldots, k\}$, (i.e. for all $s \in \{1, \ldots, k\}$, we have that $P(T_k = s) = 1/(k-1)$.) We define Z^k recursively in k:

- Let $Z^2 := V_1 V_2$.
- Given the binary string $Z^k = Z_1^k Z_2^k \ldots Z_k^k$, we define Z^{k+1}:
  - For all $j < T_{k+1}$, let
    $$Z_j^{k+1} := Z_j^k.$$
  - For $j = T_{k+1}$, let
    $$Z_j^{k+1} = V_{k+1}.$$
  - For j, such that $T_{k+1} < j \leq k+1$, let
    $$Z_j^{k+1} := Z_{j-1}^k.$$

(Thus V_k designates the k-th bit added and T_k designates the position where it gets added.)

To prove the main result of this paper, we generate a variable having same distribution as L_n using the bit-drop-scheme. Instead of generating the sequence X, we generate the triangular array $\{Z_i^k\}_{i \leq k \leq n}$ and, independently, a random number N^a with binomial distribution with parameters p and n. Then, we look for the longest common subsequence of Y and Z^k with $k = n - N^a$.
More precisely, let $L_n^a(k)$ designate the length of the Longest Common Subsequence of Z^k and $Y = Y_1 Y_2 \ldots Y_n$. Then:

Lemma 3.1 *Assume that case I holds and Z^k is generated independently of Y and N^a, according to the mechanism described above. Then, L_n has same distribution as $L_n^a(n - N^a)$.*

Proof. For every $l, k \geq 0$ we have that $P(L_n = l | N^a = k) = P(L_n^a(n-k) = l)$. This gives the thesis. ■

We can now explain the main idea behind the proof of Theorem 2.1: assume f is a map with bounded slope so that $f'(x) \geq c > 0$ for all $x \in \mathbb{R}$. Let B be any random variable. Lemma 3.2 tells us, that in this case, the variance of $f(B)$ is bounded below by $c^2 \cdot VAR[B]$. On the other hand, the map $k \mapsto L_n^a(\cdot)$ is very likely to increase above a linear rate larger than a constant $k_1 > 0$. Hence $VAR[L_n] = VAR[L_n^a(n = N^a)]$ should be larger then $k_1^2 VAR[N^a]$. The most difficult part in the proof is showing that with high probability the slope of $k \mapsto L_n^a(k)$ is "everywhere" bounded below by a positive constant. This problem is solved in the next section. Let us look at the details of the proof of Theorem 2.1:

Lemma 3.2 *Let $c > 0$. Assume that $f : \mathbb{R} \to \mathbb{R}$ is a map which is everywhere differentiable and such that for all $x \in \mathbb{R}$:*

$$\frac{df}{dx} \geq c. \tag{3.1}$$

Let B be a random variable such that $E[|f(B)|] < +\infty$ Then:

$$VAR[f(B)] \geq c^2 \cdot VAR[B]. \tag{3.2}$$

Proof. We have that $E[B]$ and $E[f(B)]$ are finite. Observe that $\lim_{x \to \pm\infty} f(x) = \pm\infty$ and $f(x)$ is strictly increasing so that there exists $x_0 \in \mathbb{R}$ such that

$$f(x_0) = E[f(B)]. \tag{3.3}$$

By the mean value theorem, we know that there exists a map $\delta : \mathbb{R} \to \mathbb{R}$ such that for all $x \in \mathbb{R}$ we have

$$f(x) = f(x_0) + f'(\delta(x))\,(x - x_0)\,. \tag{3.4}$$

By definition of variance and eqs.(3.3)(3.4) we have:

$$VAR[f(B)] = E[(f(B) - f(x_0))^2] = E[f'(\delta(B))^2\,(B - x_0)^2] \tag{3.5}$$

Using eq.(3.1) we get:

$$VAR[f(B)] \geq c^2 E[(B - x_0)^2]. \tag{3.6}$$

Observe that

$$E[(B - x_0)^2] \geq \min_y E[(B - y)^2] = VAR[B] \tag{3.7}$$

where we used a well known minimizing property of the variance. This immediately gives

$$VAR[f(B)] \geq c^2 VAR[B] \tag{3.8}$$

which finishes this proof. ■

Typically, the (random) map $k \mapsto L^a(k)$ does not strictly increase for every $k \in [0, n]$. But it is likely that every order $o(\ln n)$ points, it increases by a linear quantity. Next we define an event which guarantees that the map $k \mapsto L^a(k)$ increases linearly on the scale $o(\ln n)$:

Definition 3.1 *Let E^n_{slope} designate the event that $\forall i, j$, such that $0 < i < j \leq n$ and $i + k_2 \ln n \leq j$, we have:*

$$L^a(j) - L^a(i) \geq k_1 |i - j|. \tag{3.9}$$

Here $k_1, k_2 > 0$ designate constants which do not depend on n and which will be fixed in the proofs in sects. 4,5.

The above definition gives the discrete equivalent of condition (3.1) in the case of a discrete function. Before proceeding we need a discrete version of Lemma 3.2.

Lemma 3.3 *Let $c, m > 0$ be two constants. Let $f : \mathbb{Z} \to \mathbb{Z}$ be a non decreasing map such that:*

- *for all $i < j$:*
$$f(j) - f(i) \leq (j - i) \tag{3.10}$$
- *for all i, j such that $i + m \leq j$:*
$$f(j) - f(i) \geq c \cdot (j - i). \tag{3.11}$$

Let B be an integer random variable such that $E[|f(B)|] \leq +\infty$. Then:

$$VAR[f(B)] \geq c^2 \left(1 - \frac{2m}{c\sqrt{VAR[B]}}\right) VAR[B]. \tag{3.12}$$

Proof. Because of conditions (3.10) and (3.11), we can find a continuously differentiable map $g : \mathbb{R} \to \mathbb{R}$ satisfying the following conditions:

- g agrees with f on every integer which is a multiple of m.
- $\forall x \in \mathbb{R}$, we have that
$$c \leq g'(x) \leq 1. \tag{3.13}$$

Thus, we can apply lemma 3.2 to $g(B)$ and find:

$$VAR[g(B)] \geq c^2 \cdot VAR[B]. \tag{3.14}$$

The random variable $g(B)$ approximates $f(B)$:

$$|f(B) - g(B)| \leq (1 - c) \cdot m \tag{3.15}$$

Hence,
$$VAR[f(B) - g(B)] \leq m^2 \tag{3.16}$$

Since, $f(B) = g(B) + (f(B) - g(B))$, we can apply the triangular inequality and find:
$$\sqrt{VAR[f(B)]} \geq \sqrt{VAR[g(B)]} - \sqrt{VAR[f(B) - g(B)]} \tag{3.17}$$

Hence:

$$\begin{aligned} VAR[f(B)] \geq VAR[g(B)] - 2\sqrt{VAR[g(B)]} \cdot \sqrt{VAR[f(B) - g(B)]} = \\ = VAR[g(B)] \left(1 - \frac{2\sqrt{VAR[f(B) - g(B)]}}{\sqrt{VAR[g(B)]}}\right). \end{aligned}$$

Applying the inequalities (3.14) and (3.16) to the last inequality above, yields

$$VAR[f(B)] \geq c^2 VAR[B] \left(1 - \frac{2\,m}{c\sqrt{VAR[B]}}\right) \tag{3.18}$$

which finishes this proof. ■

Let σ_Z designate the σ-algebra of the triangular array Z_i^k and σ_{YZ} the σ-algebra of the triangular array Z_i^k and of the Y_i. Thus:

$$\sigma_Z := \sigma(Z_i^k | i \leq k \leq n) \qquad \sigma_{YZ} := \sigma(Z_i^k, Y_j | i \leq k \leq n, j \leq n).$$

We are now ready for the proof of the main theorem 2.1 of this article.

Proof of theorem 2.1 By Lemma 3.1 it is enough to prove that there exits $k > 0$ not depending on n, such that:

$$VAR[L^a(n - N^a)] \geq kn. \tag{3.19}$$

Note that for any random variable D and any σ-field σ, we have

$$VAR[D] = VAR[\, E[D|\sigma]\,] + E[\, VAR[D|\sigma]\,]. \tag{3.20}$$

Thus, since the variance is never negative, we find that

$$VAR[D] \geq E[\, VAR[D|\sigma]\,]. \tag{3.21}$$

Taking $L^a(n - N^a)$ for D and σ_{YZ} for σ, we find:

$$VAR[L^a(n - N^a)] \geq E[\, VAR[L^a(n - N^a)|\sigma_{YZ}]\,] \tag{3.22}$$

Note that the map $L^a(\cdot)$ is σ_{YZ}-measurable. Thus, conditional on σ_{YZ}, $L^a(\cdot)$ becomes a non-random increasing map. The event E^n_{slope} is σ_{YZ}-measurable. When E^n_{slope} holds, then the hypotheses of Lemma 3.3 holds for $f = L^a(\cdot)$ with $c = k_1$ and $m = k_2 \ln n$. This implies that

$$VAR[L^a(n - N^a)|\sigma_{YZ}] \geq (k_1)^2 \left(1 - \frac{2k_2 \ln n}{k_1\sqrt{VAR[N^a|\sigma_{YZ}]}}\right) VAR[N^a|\sigma_{YZ}] \tag{3.23}$$

Since N^a is a binomial variable with parameter p and n and is independent from σ_{YZ}, we have that

$$VAR[N^a] = VAR[N^a|\sigma_{YZ}] = np(1-p). \tag{3.24}$$

Using the last equality with inequality (3.23), we obtain:

$$VAR[L^a(n - N^a)|\sigma_{YZ}] \geq np(1-p)\,(k_1)^2 \left(1 - \frac{2k_2 \ln n}{k_1\sqrt{p(1-p)n}}\right) \tag{3.25}$$

Since, $VAR[L^a(n - N^a)|\sigma_{YZ}]$ is never negative and since inequality 3.25 holds, whenever E^n_{slope} holds, we find

$$\begin{aligned} VAR[L_n] &\geq E[\, VAR[L^a(n - N^a)|\sigma_{YZ}]\,] \geq \\ &\geq n \cdot P(E^n_{\text{slope}}) \cdot \left[p(1-p)\,(k_1)^2 \left(1 - \frac{2k_2 \ln n}{k_1\sqrt{p(1-p)n}}\right)\right]. \end{aligned} \tag{3.26}$$

The expression on the right side of inequality (3.26) divided by n converges to

$$P(E^n_{\text{slope}})p(1-p)\,(k_1)^2\,.$$

We will show in Lemma 4.1 below that $P(E^n_{\text{slope}}) \to 1$ as $n \to \infty$. Hence, for all n big enough, $VAR[L_n]$ is larger than $np(1-p)\,(k_1)^2/2 > 0$. This finishes the proof of theorem 2.1.

4 Slope of $L^a(\cdot)$

This section is dedicated to the proof of the following lemma:

Lemma 4.1 *We have that:*

$$P(E^n_{\text{slope}}) \to 1 \tag{4.1}$$

as $n \to \infty$.

We first need a few definitions. A common subsequence of length m of the two sequences Z^k and Y, can be viewed as a pair of strictly increasing functions:

$$(\pi, \eta)$$

such that $\pi : [1, m] \to [1, k]$, $\eta : [1, m] \to [1, n]$ and

$$\forall i \in [1, m],\ Z^k_{\pi(i)} = Y_{\eta(i)}. \tag{4.2}$$

Definitions:

1. Let $\pi : [1, m] \to [1, k]$ and $\eta : [1, m] \to [1, n]$ be two increasing functions. The pair of (π, η) is called *a pair of matching subsequences of Z^k and Y* iff it satisfies condition (4.2).
2. Let M^k_1 designate the set of all pairs of matching subsequences of Z^k and Y.
3. Let M^k_2 designate the set of all pairs of matching subsequences of Z^k and Y of maximal length, (i.e. of maximal length in the set M^k_1.)
4. Let $\leq$ indicate the natural partial order relation between increasing functions $\pi : [1, m] \to \mathbb{N}$, i.e. $\pi_1 \leq \pi_2$ iff, for every $i \in [1, m]$, $\pi_1(i) \leq \pi_2(i)$. With a slight abuse of notation we will indicate with $\leq$ also the partial order induced on the pairs of increasing function (π, η), i.e. $(\pi_1, \eta_1) \leq (\pi_2, \eta_2)$ iff $\pi_1 \leq \pi_2$ and $\eta_1 \leq \eta_2$.
5. Let $M^k \subset M^k_2$ designate the set of all $(\pi, \eta) \in M^k_2$ which are minimal according to the relation $\leq$, (i.e. minimal in the set M^k_2).
6. Let (π, η) be a pair of matching subsequences of length m and let $i \in [0, m-1]$. We call the quadruple

$$(\pi(i), \pi(i+1), \eta(i), \eta(i+1))\,, \tag{4.3}$$

a *match of* (π, η). If $\eta(i) + 2 \leq \eta(i+1)$, we call the match a *non-empty* match. If there exists j, such that $\eta(i) < j < \eta(i+1)$ and $Y_j = 1$, resp. $Y_j = 0$, we say that the match *contains* a 1, resp. a 0. We also say that the match *contains* the point j and call the bit Y_j a *free bit* of the match $(\pi(i), \pi(i+1), \eta(i), \eta(i+1))$. Sometimes we identify the match $(\pi(i), \pi(i+1), \eta(i), \eta(i+1))$ with the couple of binary words:

$$\Big(\ Z^k_{\pi(i)} Z^k_{\pi(i)+1} \ldots Z^k_{\pi(i+1)}\ ,\ Y_{\eta(i)} Y_{\eta(i)+1} \ldots Y_{\eta(i+1)}\ \Big)\,.$$

7. Let $0 < s < t \leq n$. We call the integer interval $[s,t] = \{s, s+1, \dots, t\}$ a *block of* Y, if for all $r \in [s,t]$ we have $Y_r = Y_s$ but $Y_{s-1} \neq Y_s$ and $Y_t \neq Y_{t+1}$. The cardinality $|\,[s,t]\,| = s - t + 1$ is called *length* of the block $[s,t]$.

Let us give an illustrative example. Take $Z^6 = 101011$, $n = 9$ and $Y = 111000111$. Let (π, η) be defined as follows:

$$\pi(1) = 1, \pi(2) = 3, \pi(3) = 4, \pi(4) = 5, \pi(5) = 6$$

and

$$\eta(1) = 1, \eta(2) = 2, \eta(3) = 4, \eta(4) = 7, \eta(5) = 8.$$

Then, (π, η) is a pair of matching subsequences of Z^6 and Y. The common subsequence associated with it is:

$$Z_1^6 Z_3^6 Z_4^6 Z_5^6 Z_6^6 = Y_1 Y_2 Y_4 Y_7 Y_8 = 11011.$$

We represent the pair of matching subsequences (π, η) using an alignment of Z^6 and Y:

$$\begin{array}{cccccccccc} 1 & 0 & 1 & \_ & 0 & \_ & \_ & 1 & 1 & \_ \\ 1 & \_ & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}$$

In this example (π, η) contains the four following matches:

1. $$\begin{array}{ccc} 1 & 0 & 1 \\ 1 & \_ & 1 \end{array}$$

2. $$\begin{array}{ccc} 1 & \_ & 0 \\ 1 & 1 & 0 \end{array}$$

3. $$\begin{array}{cccc} 0 & \_ & \_ & 1 \\ 0 & 0 & 0 & 1 \end{array}$$

4. $$\begin{array}{cc} 1 & 1 \\ 1 & 1 \end{array}$$

The first match above is empty. The second match contains a one. Here, Y_3 is a free bit of the second match. The third match contains two zero's: Y_5 and Y_6 are free bits of the third match. The forth match is empty. The common subsequence 11011 is of maximal length (among all the common subsequences of Z^6 and Y). So, we have that $L^a(6) = 5$. Hence, $L^a(7)$ can only be equal to 5 or 6.

What is the probability that $L^a(7)$ is larger by one than $L^a(6)$? When we generate Z^7 by dropping the bit V_7 on Z^6, then there are five positions where it can fall:

position 1	position 2	position 3	position 4	position 5
$1x01011$	$10x1011$	$101x011$	$1010x11$	$10101x1$

where x designates the possible positions of the bit V^7. Each of these positions has same probability. Positions 1 and 2 correspond to the first match. Position 3 corresponds to the second match. Position 4 correspond to the third match and position 5 corresponds to match number four.

If $V_7 = 1$ and the bit drops on the match which contains a one (that is match number two corresponding to position three, i.e. $T_7 = 3$), then $L^a(7) = L^a(6) + 1$. The reason is that the bit V^7 can then get matched with the free 1-bit in match two and increase the score $L^a(6)$ by one. Similarly, if $V_7 = 0$ and the bit V^7 drops on match number three, the score gets increased by one, since then V^7 gets matched with the "free" zero contained in match number three. Hence, when V^7 drops on match number three, the result is: $L^a(7) = L^a(6) + 1$. In general $L^a(k+1) = L^a(k) + 1$, if the bit V_{k+1} drops on a match which contains a bit of the same color as to V_{k+1}. (By color, we mean 0 or 1.)

From the idea of the previous example, we can get a lower bound for the probability that the score $L^a(k)$ increases by one. The bit V_{k+1} is equally likely to be equal to one or equal to zero. So, when it drops on a nonempty match, the score has at least 50% probability to increase. Each nonempty match corresponds to at least one position. The bit V^{k+1} has $k-1$ equally likely positions. It follows: for any pair (π, η) of matching subsequences of Z^k and Y:

$$P\Big(\ L^a(k+1) = L^a(k) + 1 \ \mid\ Z^k, Y\ \Big) \geq \frac{1}{2} \cdot \frac{\#\text{ of nonempty matches of } (\pi,\eta)}{k} \tag{4.4}$$

if (π, η) is of maximal length.

Let us explain at this stage the main ideas for the proof of lemma 4.1. We distinguish two cases depending on the value of k.
We first deal with the case $k < 0.45n$. In this case it easy to show that with large probability all the bits in Z^k are matched. Let E^n_{1k} be the event:

$$E^n_{1k} := \{L^a_n(k) = k\} \tag{4.5}$$

and

$$E^n_1 := \bigcap_{k=1}^{0.45n} E^n_{1,k}. \tag{4.6}$$

Observe that we have

$$E^n_1 = \{L^a_n(k+1) - L^a_n(k) = 1,\ \forall k < 0.45n\} \tag{4.7}$$

i.e. the slope of $L^a_n(k)$ is equal to 1 for all $k < 0.45n$ if E^n_1 holds. In the next section we prove the following lemma:

Lemma 4.2 *We have*

$$\lim_{n\to\infty} P(E^n_1) = 1. \tag{4.8}$$

Assume that instead of looking for a LCS, we want to know if one sequence is contained in another. For example for given $l \in \mathbb{N}$, we may be interested in finding out if the sequence Z^k is a subsequence of $Y_1Y_2\dots Y_l$. For this let $\nu(i)$ be the smallest l such that Z^k_i is a subsequence of $Y_1Y_2\dots Y_l$. Then, $\nu(1), \nu(2), \nu(3), \dots$ defines a renewal process. The interarrival times $I_i = \nu(i+1) - \nu(i)$ have geometric distribution and expectation $E[I_i] = 2$. Thus, $E[\nu(i)] = 2i$ and $VAR[n] = o(n)$. From this it follows that if we want Z^k to be with high probability a subsequence of $Y_1Y_2\dots Y_l$, we need to take l somewhat above $2k$. Let us give a numerical example. Take $Z^3 = 001$ and $Y = 10101000111$. Then, $\nu(1)$ denotes the indices of the first Y_i equal to zero. In this case, $\nu(1) = 2$. Similarly, $\nu(2)$ is the smallest $i \geq \nu(1)$ such that $Y_i = Z_2^3 = 0$. Here: $\nu(2) = 4$. Finally, $\nu(3)$ is the smallest $i \geq \nu(2)$, such that $Y_3 = 1$, hence $\nu(3) = 5$.

Let us next give the main ideas, why with high probability, the slope of $k \mapsto L^a(k)$ is increasing linearly on the domain $[0.45n, n]$. We use the bit-drop scheme to prove this: we show that typically the random map $k \mapsto L^a(k)$ has a positive drift $\gamma > 0$. We define:

$$E_{2k}^n := \Big\{ \forall (\pi, \eta) \in M^k, \# \text{ of nonempty matches of } (\pi, \eta) \text{ is larger than } \gamma n \Big\}. \quad (4.9)$$

When E_{2k}^n holds, every pair $(\pi, \eta) \in M^k$ has at least γn non-empty matches. The proportion of non-empty matches to k hence is larger or equal to γ. Using inequality 4.4, it follows that

$$P\Big(\ L^a(k+1) = L^a(k) + 1 \ \mid \ Z^k, Y \ \Big) \geq 0.5 \cdot \gamma \quad (4.10)$$

when E_{2k}^n holds. Let E_2^n be the event:

$$E_2^n := \bigcap_{k=0.45n}^{n} E_{2k}^n. \quad (4.11)$$

Inequality 4.10 implies, that when E_2^n holds, the map $k \mapsto L^a(k)$ has positive drift $0.5\gamma > 0$ for $k \in [0.45n, n]$. By large deviation it follows, that with high probability $k \mapsto L^a(k)$ has positive slope on $[0.45n, n]$ as soon as E_2^n holds . (See lemma 4.9.)
It remains to explain why E_2^n holds with high probability.
Let us first summarize the general idea:
We proceed by contradiction. Assume all the matches of $(\pi, \eta) \in M_2^k$ were empty. Then all of the following would hold:

- $$(\eta(1), \eta(2), \eta(3), \ldots, \eta(m)) = (\eta(1), \eta(1)+1, \eta(1)+2, \ldots, \eta(1)+m)$$

  where m is the length of the LCS of Z^k and Y: $m = L^a(k)$.
- The sequence

  $$Y_{\eta(1)} Y_{\eta(2)} \ldots Y_{\eta(m)} = Y_{\eta(1)} Y_{\eta(1)+1} \ldots Y_{\eta(1)+m}$$

  is a subsequence of

  $$Z_{\pi(1)}^k Z_{\pi(1)+1}^k \ldots Z_{\pi(m)}^k.$$

Hence we would have two independent i.i.d. sequences of Bernoulli variables with parameter 1/2, where one is contained in the other as subsequence. This implies that the sequence containing the other must be approximately twice as long. Hence k is approximately at least twice as large as $m = L^a(k)$. Thus, the ratio $L^a(k)/k$ is close to 50% or below. This is very unlikely, since it is known that the $L^a(k)/k$ is typically above 80%. This is our contradiction.

From the previous argument it follows that with high probability any $(\pi, \eta) \in M^k$ contains a non-vanishing proportion $\epsilon > 0$ of free bits. (Hence, $L_n^a(k)/\eta(L_n^a(k)) \geq \epsilon$.) We need to show that this proportion ϵ of free bits generates sufficiently many non-empty matches: the free bits should not be concentrated in a too small number of matches.

Let us go back to the numerical example on page 9 to illustrate how we count the proportion of bits that are free. In that example, the first match of (π, η) contains no free bit. The second match contains one free bit which is a one. The third match contains two free bits which are zero's. The forth match contains no free bit. The sequence Y contains a total of 8 bits which are involved in a match of (π, η). (Note that the last bit Y_9 of Y is not counted since it is not involved in a match of (π, η).) We have a proportion of free bits to bits involved in matches equal to:

$$3/8 = (8-5)/5 = \frac{\eta(L_n^a(k)) - L_n^a(k)}{\eta(L_n^a(k))} = \frac{\eta(5) - 5}{\eta(5)}.$$

The 3 free bits generate two non-empty matches.

To prove that there are more than γn nonempty matches two arguments are used:

- Any pair of matching subsequence (π, η) which is minimal according to our partial order for pairs of matches satisfies:
  every match of (π, η) can contain zero's or one's but not both at the same time. Hence, each match of $(\pi, \eta) \in M^k$ contains free bits from at most one block of Y.
- With high probability, the total number of integer points in $[0, n]$ contained in blocks of Y of length $\geq D$ is very small. (By choosing D large, we make the total number of points contained in blocks longer than D, much smaller than the number of free bits.)

From the two points above, it follows that for $(\pi, \eta) \in M^k$, the majority of free bits are at most D per match. This ensures that the proportion ϵ of free bits, generates a proportion of at least order ϵ/D non-empty matches.

Let us look at an example of a pair (π, η) which is of maximal length but not minimal according to our order relation on M_2^k. Take $Z^7 = 0101101$ and $Y = 00110010111$. Define the pair of matching subsequences (π, η) as follows:

$$\pi(1) = 1, \pi(2) = 2, \pi(3) = 3, \pi(4) = 4, \pi(5) = 5, \pi(6) = 7$$

and

$$\eta(1) = 1, \eta(2) = 7, \eta(3) = 8, \eta(4) = 9, \eta(5) = 10, \eta(6) = 11.$$

Let us represent this pair of matching subsequences by an alignment:

$$\begin{array}{cccccccccccc} 0 & \_ & \_ & \_ & \_ & \_ & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & \_ & 1 \end{array}$$

This gives the common subsequence 010111. The pair (π, η) is of maximal length, but it is not minimal for our order relation on M_2^k: instead of $\eta(2) = 7$, take $\eta^*(2) = 3$. Let otherwise η^* be equal to η. Then (π, η^*) is strictly below (π, η). To construct η^* we used

the fact that a match of (π,η) contained both zero's and one's. It is always possible to find a strictly smaller pair $(\pi,\eta^*)\in m_2^K$ when a match of (π,η) contains hero's and one's at the same time.
Note that (π,η) contains 5 free bits, but only one non-empty match. All the free bits of (π,η) are concentrated in one match. The match containing all the free bits contains several blocks. By taking a minimal pair of matching subsequences, this kind of situation is avoided.

Let us look at the details of the proof of lemma 4.1. Let $L_l^a(k)$ denote the length of the LCS of Z^k and the sequence $Y^l := Y_1Y_2\dots Y_l$. For Y^l to be entirely contained as a subsequence in Z^k, one needs k to be approximately twice as long as l. (We have that Y^l is a subsequence of Z^k iff $L_l^a(k)=l$.) Hence, it is unlikely that that Y^l is a subsequence of Z^k, when $k=2l(1-\delta)$. (Here $\delta>0$ is a constant not depending on l.) In other words, it is unlikely that:

$$L_l^a(2l(1-\delta))\geq l.$$

Similarly, it is unlikely, that Y^l is "close to being a subsequence of Z^k", when $k=2l(1-\delta)$:

Lemma 4.3 *There exists a function $\delta:\mathbb{R}\rightarrow\mathbb{R}$ such that $\lim_{\epsilon\rightarrow 0}\delta(\epsilon)=0$ and*

$$P\left(L_l^a\left(2l\big(1-\delta(\epsilon)\big)\right)>l(1-\epsilon)\right)\leq Ce^{-cl} \tag{4.12}$$

for all $l>0$ and suitable constants $c>0$ and $C>0$ not depending on l. (Note that the constants $c>0$ and $C>0$ may depend on ϵ.)

We can now define:

$$E_{3l}^n=\{L_l^a\left(2l(1-\delta(\epsilon))\right)\leq(1-\epsilon)l\} \tag{4.13}$$

and

$$E_3^n:=\bigcap_{k=0.2n}^{n}E_{3k}^n \tag{4.14}$$

where ϵ is a suitable number, to be fixed in the following, and $\delta(\epsilon)$ is given by Lemma 4.3. It follows that:

Corollary 4.1 *If $\delta(\epsilon)$ in the definition of E_3^n is given by lemma 4.3, we have*

$$\lim_{n\rightarrow\infty}P(E_3^n)=1. \tag{4.15}$$

Typically, $L_n^a(k)$ is above $80\%\cdot k$. However, to make things easier, we prove only that it is above $65\%\cdot k$. We define:

$$E_{4k}^n:=\{L_n^a(k)\geq 0.65k\} \tag{4.16}$$

and

$$E_4^n:=\bigcap_{k=0.45n}^{n}E_{4,k}^n. \tag{4.17}$$

The next lemma is proven in the next section:

Lemma 4.4 *We have*

$$\lim_{n\to\infty} P(E_4^n) = 1. \tag{4.18}$$

Let us define the event E_{6k}^n:

$$E_{6k}^n := \left\{ L_n^a(k) \leq (1-\epsilon)\eta(L_n^a(k)),\ \forall(\pi,\eta) \in M^k \right\} \tag{4.19}$$

and

$$E_6^n := \bigcap_{k=0.45n}^{n} E_{6k}^n. \tag{4.20}$$

The event E_{6k}^n says that any pair of matching subsequences $(\pi,\eta) \in M^k$ has a proportion of at least ϵ free bits. (Note that $\eta(L_n^a(k))$ is the number of the last bit of Y involved in a match of (π,η). Furthermore, $L_n^a(k)$ represents the number of bits that are "matched" by (π,η). Hence, $\eta(L_n^a(k)) - L_n^a(k)$ is the number of "free" bits.)

Lemma 4.5 *Take $\epsilon > 0$ small enough, so that*

$$\frac{50\%}{1-\delta(\epsilon)} < 65\%. \tag{4.21}$$

Then, we have that, for all $k > 0.45n$,

$$E_3^n \cap E_{4k}^n \subset E_{6k}^n. \tag{4.22}$$

Thus

$$E_3^n \cap E_4^n \subset E_6^n. \tag{4.23}$$

Proof. Let $k \in [0.45n, n]$. We show that if E_{6k}^n does not hold and E_3^n holds, then E_{4k}^n can not hold. This in terms implies 4.22.
Let $(\pi,\eta) \in M^k$. If E_{6k}^n does not hold, than the proportion of "free" bits of (π,η) is below ϵ. In other words:

$$\frac{L_l^a(k)}{l} \geq 1-\epsilon$$

where $l := \eta(L_n^a(k))$. (Note that $L_l^a(k) = L_n^a(k)$, since (π,η) is of maximal length.) It follows that

$$L_l^a(k) \geq l(1-\epsilon). \tag{4.24}$$

Now, when E_{3k}^n holds, then

$$L_l^a(2l(1-\delta(\epsilon))) \leq l(1-\epsilon). \tag{4.25}$$

Comparing inequality 4.24, with 4.25 and noting that the (random) map $x \mapsto L_l^a(x)$ is increasing, yields:

$$k \geq 2l(1-\delta(\epsilon))$$

and hence

$$k \geq 2\eta(L_n^a(k))(1-\delta(\epsilon)) \geq 2L_n^a(k)(1-\delta(\epsilon)).$$

From this it follows, that:

$$\frac{L_n^a(k)}{k} \leq \frac{50\%}{1-\delta(\epsilon)} < 65\% \tag{4.26}$$

where the 65%-bound is obtained from inequality 4.21. Inequality 4.26 contradicts E_{4k}^n. ■

To obtain E_2^n we must be sure that the free bits of Y do not concentrate in a small amount of of matches of $(\pi, \eta) \in M^k$. As explained in the example on page 12, any match of $(\pi, \eta) \in M^k$ can contain 0's or 1's, (or nothing) but not 0's and 1's at the same time. This is due to the minimality respect to the ordering $<$. In fact if $(\pi(i), \pi(i+1), \eta(i), \eta(i+1))$ is a non empty match we must have that $Y_l \neq Y_{\eta(i+1)}$ for all $\eta(i) < l < \eta(i+1)$. Otherwise, we could match the bit $Z_{\pi(i+1)}$ with Y_l instead of $Y_{\eta(i+1)}$. This modification would yield a pair of matching subsequences of same length but strictly smaller according to our order relation on M_2^k. Thus, all the free bits of a match of $(\pi, \eta) \in M^k$ are contained in only one block of Y.

It is useful to see how many bits are contained in long blocks. Let $BLOCK^D$ designate the set of all blocks $[i, j] \subset [0, n]$ of Y of length at least D. (For the definition of blocks see the definitions at the beginning of this section.) Let N^D denote the total number of points in the sequence Y which are contained in a block of length at least D:

$$N^D := \left| \left\{ s \in [1, n] \mid \exists [i, j] \in BLOCK^D, s \in [i, j] \right\} \right|. \tag{4.27}$$

Let E_5^n designate the event:

$$E_5^n := \left\{ N^D \leq \epsilon n/4 \right\} \tag{4.28}$$

We will show in sec. 6 that:

Lemma 4.6 *For every ϵ there exists D such that*

$$\lim_{n\to\infty} P(E_5^n) = 1. \tag{4.29}$$

We then have the following combinatorial fact:

Lemma 4.7 *We have that, for all $k > 0.45n$:*

$$E_4^n \cap E_5^n \cap E_{6k}^n \subset E_{2k}^n \tag{4.30}$$

with $\gamma = \frac{0.0425\epsilon}{D-1}$. Thus also:

$$E_4^n \cap E_5^n \cap E_6^n \subset E_2^n. \tag{4.31}$$

Proof. We prove 4.30. The event E_{6k}^n implies that for each $(\pi, \eta) \in M^k$ there are at least $\epsilon\ \eta(L_n^a(k))$ free bits. We have:

$$\eta(L_n^a(k)) \geq L_n^a(k). \tag{4.32}$$

When E_4^n holds, we have that:

$$L_n^a(k) \geq 0.65k. \tag{4.33}$$

Since we take $k \geq 0.45n$, inequalities 4.32 and 4.33, together imply that the number of free bits of $(\pi, \eta) \in M^k$ is at least

$$\epsilon\, 0.65 \cdot 0.45n = \epsilon\, 0.2925n.$$

By E_5^n, there are at most $0.25\epsilon n$ bits contained in blocks of length $\geq D$. Thus, there are at least $0.0425\epsilon \cdot n$ free bits contained in blocks of length $< D$. Recall that every match of $(\pi, \eta) \in M^k$ contains free bits from only one block. Hence, every match of $(\pi, \eta) \in M^k$ can contain at most $D-1$ free bits from blocks of length $< D$. Hence, these $\epsilon\, 0.0425n$ free bits which are not in N^D, must fill at least $\epsilon\, 0.0425n/(D-1)$ matches of $(\pi, \eta) \in M^k$. It follows that $(\pi, \eta) \in M^k$ has at least $0.0425\epsilon \cdot n/(D-1)$ non-empty matches. ■

Lemmas 4.5 and 4.7 jointly imply that that $E_3^n \cap E_4^n \cap E_5^n \subset E_2^n$. Hence:

$$P(E_2^{nc}) \leq P(E_3^{nc}) + P(E_4^{nc}) + P(E_5^{nc}) \tag{4.34}$$

where E_x^{nc} denotes the complement of E_x^n. We have that $P(E_3^{nc})$, $P(E_4^{nc})$ and $P(E_5^{nc})$ all converge to zero when $n \to \infty$. (This follows from Lemmas 4.1, 4.4 and 4.6.) Hence, we have that:

$$\lim_{n\to\infty} P(E_2^n) = 1. \tag{4.35}$$

Let σ_k denote the σ-algebra:

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

It is easy to check that E_{2k}^n is σ_k-measurable. Note that $L^a(k+1) - L^a(k)$ is always equal to one or zero.

Lemma 4.8 *When E_{2k}^n holds, then*

$$P\left(L^a(k+1) - L^a(k) = 1 |\, \sigma_k\right) \geq 0.5\gamma. \tag{4.36}$$

Proof. This has already been explained. (See inequality 4.10). ■

We finally observe that

$$P(E_{\text{slope}}^{nc}) \leq P(E_{\text{slope}}^{nc} \cap (E_2^n \cap E_1^n)) + P(E_2^{nc}) + P(E_1^{nc}). \tag{4.37}$$

Since $P(E_1^{nc})$ and $P(E_2^{nc})$ both go to zero as n goes to infinity, we only need to prove that

$$P(E_{\text{slope}}^{nc} \cap (E_2^n \cap E_1^n)) \to 0 \text{ for } n \to \infty, \tag{4.38}$$

to establish lemma 4.1.

Lemma 4.9 *We have that*

$$P(E_{\text{slope}}^{nc} \cap (E_2^n \cap E_1^n)) \to 0$$

as $n \to \infty$.

Proof. We can assume that $\gamma < 1$. Define $k_1 := 0.4\gamma$, so that $k_1 \leq 0.4$. Let

$$\Delta(k) := L_n^a(k+1) - L_n^a(k)$$

when E_{2k}^n holds, and $\Delta(k) := 1$ otherwise. From eq.(4.36), it follows that:

$$P\left(\Delta(k) = 1 | \, \sigma_k\right) \geq 0.5\gamma. \tag{4.39}$$

Furthermore, $\Delta(k)$ is equal to zero or one and σ_k-measurable. For $k \in]0.45n, n]$, let

$$\tilde{L}_n^a(k) = L_n^a(0.45n) + \sum_{i=0.45n}^{k-1} \Delta(i).$$

For $k \in [0, 0.45n]$, let $\tilde{L}_n^a(k) := L_n^a(k)$. Note that when E_2^n holds, then

$$L^a(k) = \tilde{L}^a(k) \tag{4.40}$$

for all $k \in [0, n-1]$. Introduce the event $\tilde{E}_{slope}^n$ to be the event such that $\forall i, j$, with $0.45n < i < j \leq n$ and $i + k_2 \ln n \leq j$, we have:

$$\tilde{L}_n^a(j) - \tilde{L}_n^a(i) \geq k_1 |i - j|. \tag{4.41}$$

When E_1^n holds, then $L_n^a(k)$ has a slope of one on the domain $[0, 0.45]$. Hence, the slope condition of E_{slope}^n holds on the domain $[0, 0.45n]$, since we have $k_1 \leq 0.4$. When E_2^n holds, then $L_n^a(k)$ and $\tilde{L}_n^a(k)$ are equal. It follows that when E_2^n and $\tilde{E}_{slope}^n$ both hold, then the slope condition of E_{slope}^n is verified on the domain $[0.45n, n]$. Hence

$$E_1^n \cap E_2^n \cap \tilde{E}_{slope}^n = E_1^n \cap E_2^n \cap E_{slope}^n. \tag{4.42}$$

Thus

$$P(E_{slope}^{nc} \cap E_1^n \cap E_2^n) = P(\tilde{E}_{slope}^{nc} \cap E_1^n \cap E_2^n) \leq P(\tilde{E}_{slope}^{nc}).$$

It only remains to prove that $P(\tilde{E}_{slope}^{nc})$ goes to zero as $n \to \infty$. For this we can use large deviation. Let $\tilde{E}_{i,j}^n$ be the event that

$$\tilde{L}_n^a(j) - \tilde{L}_n^a(i) \geq k_1 |i - j|$$

Then

$$\tilde{E}_{slope}^n = \bigcap_{i,j} \tilde{E}_{i,j}^n$$

where the intersection in the last equation above is taken over all $i, j \in [0.45n, n]$ such that $i + k_2 \ln n \leq j$. It follows that

$$P(\tilde{E}_{slope}^{nc}) \leq \sum_{i,j} P(\tilde{E}_{i,j}^{nc}) \tag{4.43}$$

where the last sum is taken over all $i, j \in [0.45n, n]$ such that $i + k_2 \ln n \leq j$. Since we took $k_1 = 0.4\gamma$ and because of 4.39, large deviation tells us that there exists constants $c, C > 0$ such that

$$P(\tilde{E}_{i,j}^{nc}) \leq C e^{-c|i-j|} \tag{4.44}$$

for all $i, j \in \mathbb{N}$. (The constants C, c do not depend on i, j.) Take $k_2 := 3/c$. With this choice, 4.44 becomes:

$$P(\tilde{E}^{nc}_{i,j}) \leq Cn^{-3} \tag{4.45}$$

when $k_2 \ln n \leq |i-j|$. Note that there are less than n^2 terms in the sum in inequality 4.43. By 4.45, each term in the sum in inequality 4.43, is less or equal to Cn^{-3}. Thus inequality 4.43 and 4.45 together imply that

$$P(\tilde{E}^{nc}_{slope}) \leq \frac{C}{n}.$$

■

5 Bounds for the probabilities.

We report in this section several proofs of the lemmas used in sec. 4.

Lemma 5.1 *for every n and $\nu < 0.5$ we have*

$$P(L^a_n(\nu n) = \nu n) \geq 1 - e^{c(0.5-\nu)^2 n} \tag{5.1}$$

Proof. We can build a pair of matching subsequences has follows: start from Z^k_1 and match it with the first $Y_{i_1} = Z^k_1$, then match Z^k_2 with the first $Y_{i_2} = Z^k_2$ such that $i_2 > i_1$. We can proceed as before until we reach the end of the Z^k or of the Y. More precisely we can define a matching (π, η) such that $\pi(i) = i$ and $\nu(i) = \inf_{l > \nu(i-1)}\{Y_l = Z^k_i\}$ (see remark after Lemma 4.2 for an explicit example). Given Z^k and Y we call T_j the sequence of random variables defined by $T_j = \nu(j) - \nu(j-1)$. Observe that the T_j is a sequence of independent random variable all with geometric distribution of parameter $\frac{1}{2}$. It follows that

$$P(L^a_n(\nu n) = \nu n) \geq P\left(\sum_{i=0}^{\nu n} T_i < n\right) = P\left(\sum_{i=0}^{\nu n} T_i - \frac{1}{\nu} < 0\right) \tag{5.2}$$

but

$$P\left(\sum_{i=0}^{\nu l} T_i - \frac{1}{\nu} > 0\right) \leq \inf_{s>0} E\left(e^{s\left(\sum_{i=0}^{\nu n} T_i - \frac{1}{\nu}\right)}\right) \tag{5.3}$$

Due to the independence of the T_i we have

$$E\left(e^{s\left(\sum_{i=0}^{\nu n} T_i - \frac{1}{\nu}\right)}\right) = E\left(e^{s\left(T_0 - \frac{1}{\nu}\right)}\right)^{\nu n} = \left(\frac{e^s}{2 - e^s}\right)^{\nu n} e^{-ns} \tag{5.4}$$

It is easy to check that

$$\inf_{s>0} \left(\frac{e^s}{2 - e^s}\right)^{\nu} e^{-s} \leq e^{c(0.5-\nu)^2} \tag{5.5}$$

for a suitable constant c, so that we get

$$P(L^a_n(\nu n) = \nu n) \geq 1 - e^{c(\nu-0.5)^2 n}. \tag{5.6}$$

■

Proof of lemma 4.2. It follows immediately from the above lemma. ■

In a very similar way we can prove that

Lemma 5.2 *For every k*

$$P(L^a_k(2(1-\delta)k) = k) \leq Ce^{c\delta^2 k} \tag{5.7}$$

Proof. Observe that the only possibility for $L^a_n(k) = k$ is that the pair of matching subsequences constructed at the beginning of the proof of lemma 5.1 has length k. Using the notation of that proof we have that

$$P\left(L^a_{(2-\delta)k}(k) = k\right) = P\left(\sum_{i=0}^{k} T_i \leq (2-\delta)k\right) \tag{5.8}$$

This quantity can be evaluated as in the previous proof to obtain the lemma. ■

We can now estimate the probability of E^n_{3k}. **Proof of Lemma 4.3.** Consider a subset of $S \subset [0,l]$ containing $(1-\epsilon)l$ points. There are $\binom{l}{l(1-\epsilon)}$ such subset. We can fix the sequence Y on the subset S. We have $2^{\epsilon l}$ Y's that agree on S. Calling $\delta(\epsilon) = \epsilon + \delta'(\epsilon)$ we have, due to Lemma 5.2, that the probability of matching all Y in S is bounded by $e^{-\delta'(\epsilon)^2 l}$. Collecting the above estimates we get that

$$\begin{aligned} P\big(L^a_l\left(2l\left(1-\delta(\epsilon)\right)\right) > l(1-\epsilon)\big) &\leq 2^{\epsilon l}\binom{l}{l(1-\epsilon)}e^{-c\delta'(\epsilon)^2 l} \leq \\ &\leq Ce^{[\epsilon(\ln 2+\ln\epsilon)+(1-\epsilon)\ln(1-\epsilon)-c\delta'(\epsilon)^2]l} \end{aligned} \tag{5.9}$$

where we have used Stirling's formula. Thus it is enough to chose

$$\delta'(\epsilon) = \sqrt{\frac{2}{c}[\epsilon(\ln 2+\ln\epsilon)+(1-\epsilon)\ln(1-\epsilon)]} \tag{5.10}$$

to obtain the lemma. ■

Proof of lemma 4.4. We can divide the sequences Z^k and Y is subsequences of length 10 and write $L^a_k(k) < \sum_{i=1}^{k/10} L_i$ where L_i is the longest common subsequence between $Y_{10(i-1)+1}\dots Y_{10i}$ and $Z^k_{10(i-1)+1}\dots Z^k_{10i}$. From Chvatal we know that $E(L_i) = 6.97844$. From a standard large deviation argument we get

$$P\left(\sum_{i=1}^{k/10} L_i < k\left(\frac{E(L_i)}{10} - \delta\right)\right) < \left(\inf_{s<0} E\left(e^{s(L_0-(0.69-\delta))}\right)\right)^{\frac{k}{10}} \tag{5.11}$$

Calling $p(s,\delta) = E\left(e^{s(L_0-(0.69-\delta))}\right)$ it easy to see that $p(s,\delta)$ is smooth in s, $p(0,\delta) = 1$ and $\partial_s p(0,\delta) < 0$ for every $\delta > 0$. This implies that

$$\inf_{s<0} p(s,\delta) < e^{-c(\delta)} \tag{5.12}$$

for suitable $c(\delta) > 0$. This immediately give the thesis of the Lemma.

■

Finally we prove the lemma 4.6: **Proof of lemma 4.6.** Let $\tilde{N}^D$ be the number of integer points in $[0, n-D]$ which are followed by at least D times the same color in the sequence Y. Thus, $\tilde{N}^D$ is the number of integer points $s \in [0, n-D]$ so that

$$Y_s = Y_{s+1} = \ldots = Y_{s+D}. \tag{5.13}$$

It is easy to check that

$$N^D \leq D\tilde{N}^D. \tag{5.14}$$

Let now $\tilde{Y}_s$, $s \in [0, n-D]$, be equal to 1 iff 5.13 holds, and 0 otherwise. We find:

$$\sum_{s=1}^{n} \tilde{Y}_s = \tilde{N}^D. \tag{5.15}$$

To estimate the sum 5.15 we can decompose it into D sub sums $\Sigma_1, \Sigma_2, \ldots, \Sigma_D$ where

$$\Sigma_i = \sum_{\substack{s=1,\ldots,n \\ s \bmod D = i}} \tilde{Y}_s \tag{5.16}$$

so that

$$\tilde{N}^D = \sum_{i=1}^{D} \Sigma_i \tag{5.17}$$

It is easy to see that

$$P\left(N^D > \frac{\epsilon}{4}n\right) \leq P\left(\tilde{N}^D > \frac{\epsilon}{4D}n\right) \leq D \cdot P\left(\Sigma_0 > \frac{\epsilon}{4D^2}n\right) \tag{5.18}$$

where the last inequality follows from the fact that at least one of the addends in 5.17 has to be larger than $\frac{\epsilon}{4D^2}n$. Now, the Y_s appearing in the sub sum Σ_0 are i.i.d. Bernoulli random variable with $P(Y_s = 1) = 2^{-D}$. We can apply a large deviation argument analogous to the one used in the previous proof and obtain

$$P\left(\Sigma_0 > (2^{-D} + \delta)\frac{n}{D}\right) \leq e^{-c(\delta)\frac{n}{D}}. \tag{5.19}$$

with $c(\delta) > 0$ for $\delta > 0$. Thus it is enough to choose D such that $D2^{-D} < \frac{\epsilon}{4}$ ■

References

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

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

[3] R. Arratia, L. Goldstein, and L. Gordon. Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.*, 17(1):9–25, 1989.

[4] R. Arratia, L. Gordon, and M.S. Waterman. The Erdős-Rényi law in distribution, for coin tossing and sequence matching. *Ann. Statist.*, 18(2):539–570, 1990.

[5] R. Arratia and M.S. Waterman. The Erdős-Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab.*, 17(3):1152–1169, 1989.

[6] Richard Arratia, Louis Gordon, and Michael Waterman. An extreme value theory for sequence matching. *Ann. Statist.*, 14(3):971–993, 1986.

[7] Richard Arratia and Michael S. Waterman. A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.*, 4(1):200–225, 1994.

[8] R.A. Baeza-Yates, R. Gavaldà, G. Navarro, and R. Scheihing. Bounding the expected length of longest common subsequences and forests. *Theory Comput. Syst.*, 32(4):435–452, 1999.

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

[10] Václáv Chvatal and David Sankoff. Longest common subsequences of two random sequences. *J. Appl. Probability*, 12:306–315, 1975.

[11] Peter Clote and Rolf Backofen. *Computational molecular biology*. Wiley Series in Mathematical and Computational Biology. John Wiley & Sons Ltd., Chichester, 2000. An introduction.

[12] Vlado Dančík and Mike Paterson. Upper bounds for the expected length of a longest common subsequence of two binary sequences. *Random Structures Algorithms*, 6(4):449–458, 1995.

[13] Joseph G. Deken. Some limit results for longest common subsequences. *Discrete Math.*, 26(1):17–31, 1979.

[14] Hauser, Matzinger, and Martinez. Large deviation montecarlo method for lcs. submitted.

[15] Marcos Kiwi, Martin Loebl, and Jiri Matousek. Expected length of the longset common subsequence for large alphabets. *preprint*, 2003.

[16] Lember and Matzinger. Deviation from mean in sequence comparison with a periodic sequence. submitted.

[17] Claudia Neuhauser. A Poisson approximation for sequence comparisons with insertions and deletions. *Ann. Statist.*, 22(3):1603–1629, 1994.

[18] Mike Paterson and Vlado Dančík. Longest common subsequences. In *Mathematical foundations of computer science 1994 (Košice, 1994)*, volume 841 of *Lecture Notes in Comput. Sci.*, pages 127–142. Springer, Berlin, 1994.

[19] Pavel A. Pevzner. *Computational molecular biology*. Computational Molecular Biology. MIT Press, Cambridge, MA, 2000. An algorithmic approach, A Bradford Book.

[20] Michael J. Steele. An Efron- Stein inequality for non-symmetric statistics. *Annals of Statistics*, 14:753–758, 1986.

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