STANDARD DEVIATION OF THE LONGEST COMMON SUBSEQUENCE

J. Lember[*], H. Matzinger

August 23, 2006

TARTU UNIVERSITY
Institute of Mathematical Statistics
Liivi 2-513 50409, Tartu, Estonia
E-mail: jyril@ut.ee

BIELEFELD UNIVERSITY
Postfach 10031
D-33501 Bielefeld, Germany
E-mail: matzing@math.uni-bielefeld.de

Abstract. Let L_n designate the length of the Longest Common Subsequence of two independent i.i.d. sequences of Bernoulli variables of length n. We prove that the order of the standard deviation of L_n is $\sqrt{n}$, provided the parameter of the Bernoulli variables is small enough. This validates Waterman's conjecture in this situation [13]. The order conjectured by Chvatal-Sankoff [7], however, is different.

Keywords. *Longest common subsequence, variance bound, Chvatal-Sankoff conjecture.*

AMS. 60K35, 41A25, 60C05

1 Introduction

Throughout this paper $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are two independent sequence of i.i.d. Bernoulli variables with parameter $0.5 \geq \epsilon > 0$:

$$\epsilon = P(X_i = 1) = P(Y_i = 1) = 1 - P(X_i = 0) = 1 - P(Y_i = 0).$$

Let $X := X_1 X_2 \ldots X_n$ and let $Y := Y_1 Y_2 \ldots Y_n$. The longest common subsequence (LCS) of X and Y is any common subsequence that has the longest possible length. The length of LCS is denoted L_n. Formally, L_n is the biggest k such that there exists two subsets of indices $\{i_1, \ldots, i_k\}, \{j_1, \ldots, j_k\} \subset \{1, \ldots, n\}$ satisfying $i_1 < i_2 < \ldots < i_k$, $j_1 < j_2 < \ldots < j_k$ and $X_{i_1} = Y_{i_1}, X_{i_2} = Y_{i_2}, \ldots, X_{i_k} = Y_{i_k}$. The main result of this paper is, that for $\epsilon > 0$ small enough, the order of the standard deviation of L_n is $\sqrt{n}$.

[*]Supported by the Estonian Science Foundation Grant nr. 5694 and SFB 701 of Bielefeld University

LCS's are a very important tool in computational biology, where they are used for comparing DNA- and protein-alignments (see, e.g. [12, 14, 2]). They are also used in computational linguistics, speech recognition and so on. In all these applications, two strings with a relatively long LCS, are deemed related.

Example. Let us give an example of the practical use of LCS's. Take the two words: $X = fanthastic$ and $Y = fntastique$. These two words are very similar. They were obtained from the English word "fantastic" and the French word "fantastique" by adding spelling mistakes. We would like the computer to recognize the similarity. If the computer compares letter by letter,

f	a	n	t	h	a	s	t	h	a	s	t	i	c
f	n	t	a	s	t	i	q	u	e				

it finds that only one letter coincides. Comparing the i-th letter of the first word with the i-th letter of the second word for all the letters is not a good way to recognize any similarity. The reason are the missing letters. The original position of the letters in the words gets changed.

To take into account the missing letters or added letters, we align the two words allowing for gaps. We allow only same letters to be matched with each other. In such a way, we obtain a sequence of letters that is contained in X as well as in Y. Such a sequence is a common subsequence of X and Y. Hence, the longest common subsequence is the maximum number of same letters we can align allowing gaps. In our example the maximum is given by the alignment

f	a	n	t	h	a	s	t	i	c			
f		n	t		a	s	t	i		q	u	e

(1.1)

Hence f, n, t, a, s, t, i is the longest common subsequence of the two words and the length of the longest common subsequence, L_n, is 7. This indicates that the two words are very similar.

To distinguish related pairs of strings from unrelated via the LCS-method, we need to assess the order of the fluctuation of the LCS. For this reason the random variable L_n has received a lot of attention. Non the less, many questions remain open. In their pioneering paper [7], Chvatal and Sankoff prove that the limit

$$\gamma := \lim_{n\to\infty} \frac{EL_n}{n} \tag{1.2}$$

exists. In [1], Alexander investigated the rate of the convergence in (1.2) and showed that for a constant C, $EL_n - n\gamma \geq C\sqrt{n \ln n}$. Moreover, by a sub-additivity argument

$$\frac{L_n}{n} \to \gamma \text{ a.s and in } L_1. \tag{1.3}$$

(see, e.g. [1, 14]). The constant γ is called the Chvatal-Sankoff constant and its value is unknown for even as simple cases as i.i.d. Bernoulli sequences. In this case, the value of γ obviously depends on the Bernoulli parameter ϵ. When $\epsilon = 0.5$, the various bounds indicate that $\gamma \approx 0.81$ [11, 9, 3]. For a smaller ϵ, γ is even bigger. Hence, a common subsequence of two independent Bernoulli sequences typically makes up large part of the total length. This implies that to make some inference, the size of the variance $\mathrm{Var}[L_n]$ is essential. Unfortunately, not much is known about $\mathrm{Var}[L_n]$ and its asymptotic order is one of the central open problems in string matching theory.

Monte-Carlo simulations lead Chvatal and Sankoff in [7] to conjecture for $\epsilon = 0.5$ that

$\text{Var}[L_n] = o(n^{\frac{2}{3}})$. Using an Efron-Stein type of inequality, Steele [11] proved $\text{Var}[L_n] \leq 2\epsilon(1-\epsilon)n$. In [13], Waterman asks whether this linear bound can be improved. His simulations show that this is not the case and $\text{Var}(L_n)$ grows linearly. Boutet de Monvel [6] interprets his simulation the same way.

In a series of papers, we investigate the asymptotic behavior of $\text{Var}[L_n]$ in various setup. Our goal is to find out, whether there exists a constant $c > 0$ (not depending on n) such that $\text{Var}[L_n] \geq cn$. Together with Steele's bound, this means that $cn \leq \text{Var}[L_n] \leq n$, i.e. $\text{Var}[L_n] = \Theta(n)$ (a sequence a_n is of order $\Theta(n)$, if, for some constants $0 < c < C < \infty$, $cn \leq a_n \leq Cn$ for all n large enough). Simulations [5] indicate that $\text{Var}[L_n] = \Theta(n)$ when ϵ is not close to 0.5. In [4], Bonetto and Matzinger consider the asymmetric situation where the random variables in X are Bernoulli with 1/2, but Y is a random i.i.d. string with 3 symbols. They obtain that in this setting $\text{Var}[L_n] = \Theta(n)$. In [8], Houdre, Lember and Matzinger investigate the asymptotic behavior of the longest common increasing subsequence of two independent Bernoulli sequences. They find that under this additional restriction $n^{-1/2}(L_n - EL_n)$ converges in law to a functional of two Brownian motions implying that $\text{Var}[L_n] = \Theta(n)$ holds again (here L_n designates the length of the longest common increasing subsequence). Durringer, Lember and Matzinger [10] show that $\text{Var}[L_n] = \Theta(n)$ when Y is a non-random periodic binary sequence and X is i.i.d. Bernoulli 1/2 sequence.

The relatively long history shows that determining the exact order of the fluctuation of L_n is a difficult problem. In fact, as noted in [1, 2], the LCS-problem can be reformulated as a Last Passage Percolation (LPP) problem with correlated weights. But for standard LPP and First Passage Percolation, the question of the exact order of the fluctuation has been open for decades.

2 Main result

The main result of this paper, Theorem 2.1, asserts that when $\epsilon > 0$ is small, the fluctuation of L_n is of order $\sqrt{n}$. In fact, the theorem gives only a lower linear bound for the variance of L_n. The upper linear bound comes from the result of Steele [11]. Hence, Theorem 2.1 implies that $\text{Var}[L_n] = \Theta(n)$.

Theorem 2.1 *There exists $\epsilon_0 > 0$ such that for every $\epsilon < \epsilon_0$, there exists a constant $c > 0$ depending on ϵ but not depending on n, that satisfies*

$$VAR[L_n] \geq c \cdot n, \quad \forall n.$$

One of the main tools in this paper is a map that picks a one in the text X or Y at random and changes it into a zero. Let $\tilde{X}$ and $\tilde{Y}$ designate the texts obtained in this way.

Example. Let $n = 6$, $X = 001000$ and $Y = 101000$. The total number of ones in the two texts is 3. Hence, we pick one of these three ones at random with equal probability and switch it into a zero. Assume we pick the second one in text Y. Then $\tilde{X} = 001000$ and $\tilde{Y} = 100000$.

Let us define $\tilde{X}$ and $\tilde{Y}$ rigorously. For a binary string $x = x_1 x_2 \dots x_n$, we denote by N_1^x the total number of ones in x. So $N_1^x := \sum_{i=1}^n x_i$. Similarly, N_1^y is the total number of ones in $y = y_1 y_2 \dots y_n$. The binary random strings $\tilde{X}$ and $\tilde{Y}$ are defined by the following equations:

$$\sum_{i=1}^{n}(|\tilde{X}_i - X_i| + |\tilde{Y}_i - Y_i|) = \begin{cases} 1, & \text{if } \sum_{i=1}^n (X_i + Y_i) > 0\ ; \\ 0, & \text{else.} \end{cases}$$

$$\sum_{i=1}^{n}(\tilde{X}_i - X_i + \tilde{Y}_i - Y_i) = \begin{cases} -1, & \text{if } \sum_{i=1}^n (X_i + Y_i) > 0\ ; \\ 0, & \text{else.} \end{cases}$$

$$\mathbf{P}(\tilde{X}_i \neq X_i | X = x, Y = y) = \begin{cases} 0 & \text{if } x_i = 0\ ; \\ \frac{1}{N_1^x + N_1^y}, & \text{else.} \end{cases},$$

$$\mathbf{P}(\tilde{Y}_i \neq Y_i | X = x, Y = y) = \begin{cases} 0 & \text{if } y_i = 0\ ; \\ \frac{1}{N_1^x + N_1^y}, & \text{else.} \end{cases}$$

Let $\tilde{L}_n$ denote the length of the longest common subsequence of $\tilde{X}$ and $\tilde{Y}$. When we change one bit in X or Y and flip it to the opposite value, then the length of the LCS changes by at most one. The next theorem shows that in this case the length of the LCS L_n is more likely to increase by one unit than to decrease by one unit.

Theorem 2.2 *There exist constants α_1 and α_2, $\alpha_1 > \alpha_2$ and a set $B_n \subset \{0,1\}^n \times \{0,1\}^n$ such that for all $(x, y) \in B_n$*

$$\mathbf{P}(\tilde{L} - L = 1 | X = x, Y = y) \geq \alpha_1, \tag{2.1}$$

$$\mathbf{P}(\tilde{L} - L = -1 | X = x, Y = y) \leq \alpha_2. \tag{2.2}$$

Moreover, there exists an $\epsilon_0 > 0$ such that for every $0 < \epsilon \leq \epsilon_o$

$$\mathbf{P}((X, Y) \in B_n) \geq 1 - e^{-c_1 n}, \tag{2.3}$$

where $c_1 > 0$ does not depend on n, but may depend on ϵ.

In Section, 7 we prove that Theorem 2.2 implies Theorem 2.1. Let us briefly explain the main ideas behind the proof. We define two sequences of random binary strings $X^1, X^2, \dots, X^{2n}$ and $Y^1, Y^2, \dots, Y^{2n}$, all of them having the length n. The strings X^k and Y^k are define by induction on k: X^{2n} and Y^{2n} consist only of ones; X^{k-1} and Y^{k-1} are obtained by choosing a one at random in $X^k Y^k$ and replacing it by a zero. Hence we use the random map $\tilde{}$. We designate by $L(k)$ the length of the LCS of X^k and Y^k. Note that the total number of ones in the string X^k and Y^k is k. Let (X, Y) be independent of $\{(X^k, Y^k)\}_{k \in \{0,\dots,2n\}}$ and let N_1 designate the total number of ones in the two strings X and Y. It is not hard to see that (X^k, Y^k) has the same distribution as (X, Y) conditional on $N_1 = k$. This implies that $L(N_1)$ has same distribution, as L_n. The standard deviation of N_1 is of order $\sqrt{n}$. Moreover, from Theorem 2.2 directly follows that the (random) map $k \mapsto L(k)$ tends to increase linearly on a certain scale. These two facts together imply immediately that the standard deviation of $L(N_1)$ and hence also of L_n is of order $\sqrt{n}$.

Let us now give a heuristic argument why Theorem 2.2 holds. Recall that in this paper, we consider the situation where one has a small, but fixed probability. Hence, in the texts X and Y, there is a small proportions of ones. This implies that only a small percentage of ones can figure in a LCS. It will turn out that the number of ones in a LCS is typically of order $\epsilon^2 n$. This is much less than the total number of ones in the texts X and Y, which is of order $2\epsilon n$. It follows that the majority of ones in the texts X and Y constitute a "net loss" for the score L_n. Hence the number of ones tends to influence the score L_n negatively. Changing a randomly picked one into zero is not very likely to decrease the score. It can decrease the score only if the chosen one is used in a LCS. But the additional zero obtained in this way will in many cases increase the score.

Example. Let $X = 0001000001000000000000001$, $Y = 000100000000100000100000$. The longest common subsequence Z is $Z = 000100000000000000000$. An alignment corresponding to Z is

X	0	0	0	1	0	0	0	0	1	0	0	0	0		0	0	0	0		0	0	0	0	0	1
Y	0	0	0	1	0	0	0	0		0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	
Z	0	0	0	1	0	0	0	0		0	0	0	0		0	0	0	0		0	0	0	0	0	

The optimal solution is obtained by matching all the zeros, and the first one in both texts, but discarding all other ones. We see the general phenomena: since there are few ones, sometimes by chance some ones appear in respective positions in the two texts where they can be matched. The other ones in text X and Y appear in places in the text where we can not match them with a one. If we would match them we would loose too many zeros. That is why, most ones can not be used in the LCS.

The argument in the previous numerical example gives a first idea of what is happening. However, proving anything rigorously is difficult. The reason is as follows. We take ϵ small but fixed and let then n tend to infinity. The optimal alignment (optimal alignment is the alignment which defines the LCS) is then going to be a global alignment. Which means that typically some parts of the text X will be connected with parts of the text Y that are "far away". This introduces complicated correlations between the different parts of the optimal alignment. Microscopically it is easy to understand the approximate behavior of the optimal alignment. Macroscopically however, little is understood about the optimal alignment. It seems that there are complicated long range interactions between all the different parts.

3 Aligning the ones

Introducing the right notation to describe the alignments is a key ingredient to the solution of our problem. Throughout this paper only consider alignments which align a symbol with a gap or with the same symbol in the other text. We exclude alignments which align different symbols with each other. We start with a simple example.

Example. Take the two texts $X = 1000001$ and $Y = 1001$. The LCS of X and Y is $Z = 1001$. It is obtained by aligning the first one in both text and the last one and for the rest aligning as many zeros as possible. Text X contains 5 zeros and text Y contains 2. The maximum number of aligned zeros is thus $\min\{2,5\} = 2$. There are many alignments corresponding to the LCS $Z = 1001$. Let us present two alignments corresponding to this LCS:

X	1	0	0	0	0	0	1
Y	1	0	0				1

or another possibility:

X	1	0	0	0	0	0	1
Y	1				0	0	1

How the zeros are aligned between the ones is not important as long as we align the maximum number of zeros between the ones. Hence in general we will only describe which ones are aligned and assume the between ones we align the maximum number of zeros. Let us give a further example to illustrate this. Take the sequences:

$$X = 101010101$$
$$Y = 11010001$$

A LCS of X and Y is 1101001. This LCS can be obtained with the following alignment:

X	1	0	1	0	1	0		1	0	1
Y	1		1	0	1	0	0		0	1

(3.1)

We call the portions between pairs of aligned ones *cell*.
The first cell of alignment (3.1) is:

1
1

The first cell is an exception. It is the only cell which is not comprised between two pairs of aligned ones. Instead it consists of the first pair of aligned ones and everything to the left there of. We only introduce this special cell in order to simplify notations later on.
The second cell of alignment (3.1) is

0	1
	1

The third cell of alignment (3.1) is

0	1
0	1

The fourth cell of alignment (3.1) is

0		1	0	1
0	0		0	1

.

Note that the second cell has one more zero in the X-part than in the Y-part. The third cell has the same amount of zeros in both parts. The fourth cell has two zeros in the X-part and three zeros in the Y-part. Hence the X-part has one zero less. The difference of zeros between the X-part and the Y-part for cell 2,3 and 4 in this order is 1, 0 and -1. Cell number 1 has no zeros. Hence the difference of zeros for cell number number 1 is equal to zero. Let v_i denote the difference of zeros of cell i. We will represent alignments as the sequence of differences of zeros of their cells. For the alignment (3.1), this gives the representation $(v_1, v_2, v_3, v_4) = (0, 1, 0, -1)$. This sequence uniquely defines the alignment of the ones.

Let $X = X_1 \dots X_n$ and $Y = Y_1 \dots Y_n$ be given. As explained above, to every optimal alignment corresponds a vector $v := (v_1, \dots, v_k)$ that shows the number of cells in the alignment (k) and the difference of zeros in the cells. In every cell, the maximum amount of zeros is aligned. On the other hand, to every vector $v = (v_1, \ldots, v_k) \in \mathbb{Z}^k$ corresponds a (possible empty) family of alignments. All of them have the same pairs of aligned ones and between consecutive pairs of aligned ones, the maximum number of zeros is aligned. The alignments corresponding to v can differ only in the way the zeros between aligned ones (inside cells) are aligned. Since all the alignments associated with v have the same score (the same number of aligned zeros and ones), we do not care how the zeros inside a cell are aligned (as long as the maximal number of them is aligned). Therefore, in a slight imprecision we will speak of one alignment for the whole family associated with v. In other words, we identify each vector v with an alignment. In this alignment, the number of aligned ones (cells) is k, the difference in the number of zero's in cell number i is v_i and inside a cell, the maximal number of zeros is aligned. So, in a sense, it is the "smallest" alignment which aligns exactly k pairs of ones with each other and has the difference of zeros in cell i equal to v_i, for all $i \in \{1, 2, \ldots, k\}$.

We write $|v|$ for the length of v. If $v \in \mathbb{R}^k$, then $|v| = k$. Let us next define rigorously the alignment associated with $v = (v_1, \ldots, v_k) \in \mathbb{Z}^k$.

Definition 3.1 *Let $k \in \mathbb{N}$ and let $v = (v_1, \ldots, v_k) \in \mathbb{Z}^k$. Define $\pi(i), \nu(i)$ by induction on i:*

- *start with $\pi(0) = \nu(0) = 0$;*
- *for $i < k$, once $\pi(i), \nu(i)$ is defined, let $(\pi(i+1), \nu(i+1))$ be the smallest (s, t) such that all of the following three conditions are satisfied.*
  1. *$\pi(i) < s$ and $\nu(i) < t$;*
  2. *$X_s = Y_t = 1$;*
  3. *the difference between the number of zeros of X in the interval $[\pi(i), s]$ and the number of zeros of Y in the interval $[\nu(i), t]$ is equal to v_{i+1}. Hence,*

$$v_{i+1} := \Big((s - \pi(i)) - \sum_{j=\pi(i)}^{s} X_j \Big) - \Big((t - \nu(i)) - \sum_{j=\nu(i)}^{s} Y_j \Big).$$

If no such (s, t) exists, then $\pi(i+1) = \cdots = \pi(k) := \infty$ and $\nu(i+1) = \cdots = \nu(k) := \infty$.

The cell number i is equal to the pair of strings:

$$C(i) := \big((X_{\pi(i-1)+1}, \ldots, X_{\pi(i)}),\ \ (Y_{\nu(i-1)+1}, \ldots, Y_{\nu(i)}) \big) .$$

We define the alignment v as any alignment, if exists, such that:

- *$X_{\pi(i)}$ is aligned with $Y_{\nu(i)}$ for every $i = 1, \ldots, k$;*

- *the number of aligned zeros in the cell $C(i)$, denoted by $S_v(i)$, is the minimum between the number of zeros in the string $X_{\pi(i-1)+1}X_{\pi(i)+1}\dots X_{\pi(i)}$ and the number of zeros in the string $Y_{\nu(i-1)+1}Y_{\nu(i)+1}\dots Y_{\nu(i)}$;*

- *after aligning $X_{\pi(k)}$ with $Y_{\nu(k)}$, we align as many zeros as possible. Let that number be r.*

Hence, the number of aligned zeros up to the last pair of aligned ones equal to

$$S(i) := \min\left\{(\pi(i)-\pi(i-1)) - \sum_{j=\pi(i-1)+1}^{\pi(i)} X_j\ ,\ (\nu(i)-\nu(i-1)) - \sum_{j=\nu(i-1)+1}^{\nu(i)} Y_s\right\}.$$

To show that all $\pi(i), \nu(i), C(i), S(i)$ depend on v, we write also

$$\pi_v(i) := \pi(i),\ \nu_v(i) := \nu(i),\ C_v(i) := C(i),\ S_v(i) := S(i),\ r_v := r.$$

To summarize: every $v \in \mathbb{Z}^k$ defines an alignment. This alignment corresponds to aligning $X_{\pi_v(i)}$ with $Y_{\nu_v(i)}$, for each $i = 1, 2, \dots, k$. These are the aligned pairs of ones: $X_{\pi_v(i)} = Y_{\nu_v(i)} = 1$. Between the aligned pairs of ones we assume that we align as many zeros as possible. Hence in cell number i, we align $S_v(i)$ zeros (maximum possible amount). After last pair of aligned ones, we align as many zeros as possible. The length of the common subsequence defined by alignment v can now be computed as follows:
Each cell gives one aligned pair of ones. Hence, this part contributes $|v|$. Then we add for each cell the number of zeros aligned. This sums up to $\sum_{i=1}^{|v|} S_v(i)$. Finally we need to add the remaining amount of zeros r_v which can be aligned but which come after the last cell. When $v \in \mathbb{Z}^k$ is such that $\pi_v(k), \nu_v(k) \le n$, then r_v is the minimum between the number of zeros in the string $X_{\pi_v(k)}\dots X_n$ and the number of zeros in the string $Y_{\nu_v(k)}\dots Y_n$. The length of the common subsequence defined by the alignment v is now equal to:

$$S_v := |v| + \sum_{i=1}^{|v|} S_v(i) + r_v.$$

The number S_v is also called the *score* of the alignment v. This is the length of the common subsequence corresponding to v.

Of course, it can be that given $X = X_1 \dots X_n$ and $Y = Y_1 \dots Y_n$ there might not be any alignment corresponding to v. In this case $\pi(k) = \nu(k) = \infty$. On the other hand, if an alignment corresponding to v exists, then $\pi_v(k) \le n$ and $\nu_v(k) \le n$. A vector $v \in \mathbb{Z}^k$ satisfying the previous condition is called *admissible*. Let V designate the set of all admissible alignments, i.e.

$$V := \{v \in \cup_{k>0}\mathbb{Z}^k : \pi(|v|), \nu(|v|) \le n\}. \tag{3.2}$$

The set V, obviously, depends on X and Y. The next statement trivially holds.

Proposition 3.1

$$L_n = \max_{v \in V} \left(|v| + \sum_{i=1}^{|v|} S_v(i) + R_v \right). \tag{3.3}$$

We say an admissible alignment v is *optimal* if $S_v = L_n$.

Let $v \in \bigcup_{k>0} \mathbb{Z}^k$ be non-random and define $|v|$ random cells $C_v(1), \dots, C_v(|v|)$ as in Definition 3.1. One of the main advantages of defining alignments the way described above is that the cells $C_v(1), C_v(2), \dots, C_v(|v|)$ are independent so that we can use large deviation techniques.

3.1 An useful approach

In the sequel, we are often going to use the following way of modeling random sequences $X_1, X_2, \dots$ and $Y_1, Y_2, \dots$. Let $\xi_1, \xi_2, \dots$ be the sequence of iid random variables with the distribution of ξ being following:

$$P(\xi = \emptyset) = 1 - \epsilon, \quad P(\xi = 1) = \epsilon(1-\epsilon), \dots P(\xi = n) = \epsilon^n(1-\epsilon), \dots.$$

The distribution of ξ_i is geometric. The random variables ξ_i model the number of 1's between the 0's: ξ_1 is the number of ones before the first 0, ξ_2 is the number of ones between the first and second 0 and so on. For example, if $(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6) = (\emptyset, 2, \emptyset, \emptyset, 1, \emptyset)$, then the corresponding sequence $X_1, X_2, \dots$ begins with

$$\emptyset, 0, 2, 0, \emptyset, 0, \emptyset, 0, 1, 0, \emptyset, 0 = 0, 1, 1, 0, 0, 0, 1, 0, 0.$$

Similarly, let $\eta_1, \eta_2, \dots$ model the sequence $Y_1, Y_2, \dots$. With such construction, it is relatively easy to model cells. Indeed, to get a 0 cell, we look for the smallest time i such that $\xi_i \neq \emptyset$, $\eta_i \neq \emptyset$. So, the length of a 0-cell is modeled by the random variable T, where

$$T := \min\{i = 1, 2, \dots : \xi_i \neq \emptyset, \eta_i \neq \emptyset\}. \tag{3.4}$$

To model a $-u$ cell ($u > 0$), we look for the smallest time T such that $\xi_i \neq \emptyset$ and $\eta_{u+i} \neq \emptyset$. So, the length of a $-u$-cell is modeled by the random variable T, where

$$T := \min\{i = 1, 2, \dots : \xi_i \neq \emptyset, \eta_{u+i} \neq \emptyset\}. \tag{3.5}$$

In other words a cell with $v_i = u$ can be viewed in the following way: we first set u zeros aside on side X if $u \geq 0$ and on side Y otherwise. Then we align consecutive pairs of zeros, until we meet for the first time a pair of aligned zeros both directly followed by a one. Let us look at a numerical example:

Example. Take $v_1 = u = 2$. Let $X = 000101\dots$ and $Y = 001\dots$. We put a side the first two zeros in X. From there, we align all the zeros until we meet two zeros both followed directly by a one. And, this gives the cell:

X		0	0	0	1	0	1
Y				0		0	1

.

3.2 Optimal alignment contained in V_n

In Section 5, we will show that with high probability L_n is larger by $0.1\epsilon^2 n$ than half of the total amount of zeros in X or in Y. Let us briefly explain the use of this fact.
When

$$L_n \geq \frac{N_0}{2} + a,$$

where N_0 is the total number of zeros in X and Y and $a > 0$, there are two immediate consequences: 1) In any optimal alignment v there need to be at least a pairs of aligned ones. Hence, any optimal alignment v needs to be contained in the set $\cup_{k \geq a} \mathbb{Z}^k$.
2) Any optimal alignment v in $\mathbb{Z}^k$, satisfies

$$\sum_{i=1}^{k} |v_i| \leq 2k. \tag{3.6}$$

Otherwise the un-matched zeros (at least $\sum_{i=1}^{k} |v_i|$) would out-number the aligned ones (the number of aligned ones is $2k$) bringing the score below an alignment with only zeros aligned. Indeed, the number of non-aligned zeros in the alignment v is at least $\sum_{i=1}^{k} |v_i|$, so the number of aligned zeros is at most

$$\frac{N_0 - \sum_{i=1}^{k} |v_i|}{2}$$

and (3.6) follows from the inequalities

$$\frac{N_0}{2} < L_n \leq \frac{N_0 - \sum_{i=1}^{k} |v_i|}{2} + k.$$

When we take $0.1\epsilon^2 n$ for a, conditions 1) and 2) can be expressed by saying that any optimal alignment v is necessarily contained in the set V_n, where

$$V_n := \bigcup_{k \geq 0.1\epsilon^2 n} V(k), \tag{3.7}$$

and $V(k) \subset \mathbb{Z}^k$ is defined as follows

$$V(k) := \{(v_1, v_2, \ldots, v_k) \in \mathbb{Z} \mid |v_1| + \ldots + |v_k| \leq 2k\}. \tag{3.8}$$

The fact that optimal alignment is typically contained in V_n is very useful. The set V_n is relatively small (see the bound (5.7)). So, whenever we want to prove the likeliness of a property for the optimal alignment, we prove the property to hold typically for every alignment in V_n. The tremendous advantage of this approach is that for every (non-random) $v \in V_n$, the alignment associated with v has a simple distribution: the cells are independent. This allows us to use large deviation. For the optimal alignment on the other hand, the cells are correlated among each other in an complicated and poorly understood manner.

4 The effect of changing a one into a zero

4.1 The events B_n and A_n

This subsection is dedicated to proving theorem 2.2. We want to show that typically, when changing a randomly picked one into a zero, the score L_n is likelier to increase than to decrease. More precisely, we want the conditional probability of an increase in score to be above α_1, whilst the conditional probability of a decrease should be below α_2. The constants α_1 and α_2, do not depend on n and satisfy $\alpha_1 > \alpha_2$. By "conditional", we mean conditional on X and Y.

Example. Take the two texts $X = 01000001$ and $Y = 10010101$. An optimal alignment is given by

X			0	1	0		0	0	0	0	1
Y	1	0	0	1	0	1	0				1

The first cell in this alignment is

		0	1
1	0	0	1

whilst the second cell is:

0		0	0	0	0	1
0	1	0				1

Assume that the one which we switch into a zero is Y_6. This is a "non-aligned" one contained in the Y-part of cell number two. By switching Y_6 into a zero the LCS increases by one unit. The reason is that in cell number two, we can now align three zeros instead of only two. The new cell number two (after switching Y_6) looks as follows:

0	0	0	0	0	1
0	0	0			1

The score gets increased because Y_6 is on the side of the cell with strictly less zeros. We say that Y_6 is *on the side of a cell with less zeros.* Hence switching a one into a zero on that side increases the score by one unit. Let us imagine next that instead of Y_6 the one chosen would be X_8. This one is "used" in the alignment and hence switching it could result (and does in this case) in decreasing the optimal score L_n by one unit. (This is not always necessary though, as can be seen with X_2. When we flip X_2 into a zero, the score remains the same.) Finally note that changing Y_1 into a zero leaves the score unchanged. This, is because Y_1, unlike Y_6, is on the side of a cell with more zeros. We call the ones which are "used" in the alignment, ones that are *matched by the alignment.* In our example, X_2 is matched with Y_4 and X_8 is matched with Y_8, Y_5 is not matched, nor is Y_1.

In the present situation, we have six ones. Each one has a probability to get picked of 1/6. Only Y_6 increases the score when picked. Hence the probability of an increase in score is equal to 1/6. Four ones, X_2, X_8, Y_4 and Y_8 could potentially decrease the score. (In our example only X_8, Y_4 and Y_8 actually do). The conditional probability of a decrease is 3/6. Since, in general, with longer sequences we can not look in detail at every realization, we will use as upper-bound for the probability of a decrease: the proportions of matched ones to total number of ones. In our case, this gives 4/6 as upper bound for the probability of a decrease in score.

From our example, it becomes clear what we need to do. We need to prove that typically there exists an optimal alignment v for which:

1) The proportion of ones that are matched among all ones in X and Y, is below α_2.
2) The proportion of ones that are on a side of a cell with less zeros among all ones in X and Y is above α_1.
In other words, we need to show that there exists an optimal alignment, with much less aligned ones than ones that are on a side of a cell with less zeros.
Let $N_v^-(i)$ denote the number of ones on the side with less zeros in cell number i. Formally, let $k \in \mathbb{N}$ and let $v = (v_1, \ldots, v_k) \in \mathbb{Z}^k$ be admissible. For $i \in [0, k]$, we define

$$N_v^-(i) := \begin{cases} 0, & \text{if } v_i = 0 \text{ (there is no side with less zeros);} \\ \sum_{j=\nu(i)+1}^{\nu(i+1)-1} Y_j, & \text{if } v_i > 0 \text{ (Y part has less zeros);} \\ \sum_{j=\pi(i)+1}^{\pi(i+1)-1} X_j, & \text{if } v_i < 0 \text{ (X part has less zeros).} \end{cases}$$

The *total number of ones on sides with less zeros* is

$$N_v^- := \sum_{i=1}^{|v|} N_v^-(i).$$

Fix some constants α_1, α_2. Let A_n be the event that there exists an optimal alignment v such that

1. The proportion of aligned ones is below α_2: $2|v| \leq \alpha_2 N_1$, where N_1 is the total number of ones in X and Y.

2. The proportions of ones on sides with less zeros is above α_1. Hence, $N_v^- \geq \alpha_1 N_1$.

Obviously, A depends on chosen α_1, α_2. From what we explained it follows directly that on A_n, the desired inequalities hold:

$$\mathbf{P}(\tilde{L} - L = 1|X,Y) \geq \alpha_1 \text{ and } \mathbf{P}(\tilde{L} - L = -1|X,Y) \leq \alpha_2.$$

What is left to prove is that there exists $\alpha_1 > \alpha_2 > 0$ such that the event A_n has probability close to one:

$$\mathbf{P}(A_n) \geq 1 - \exp[-c_1 n], \quad \text{where } c_1 > 0. \tag{4.1}$$

To be consistent with the notation in Theorem 2.2, let B_n designate the set of pairs of strings (x, y) for which A_n holds. Hence, $(x, y) \in B_n$ if and only if

$$\{x = X, y = Y\} \subset A_n.$$

We have $A_n := \{(X, Y) \in B_n\}$ and for $(x, y) \in B_n$ inequalities (2.1) and (2.2) hold.

4.2 Breaking cells

In the previous section we argued, that we need an optimal alignment with enough ones in cell-sides with less zeros. The problem is that many optimal alignments can have most cells with the same number of zeros on both sides. For such alignments there will also be few ones on cell-sides with less zeros. This problem is circumvent by taking an optimal alignment with most cells having same number of zeros on both sides and applying some surgery, so as to create enough cells with different numbers of zeros on the sides. This is done in such a manner that the patient after operation is still an optimal alignment. Let us first look at an example.

Example. Take the texts $X = 00101001001001$ and $Y = 00100100010101$. Take the following optimal alignment

X	0	0	1	0	1	0		0	1	0	0	1	0		0	1
Y	0	0	1	0		0	1	0		0	0	1	0	1	0	1

The first cell is

0	0	1
0	0	1

The second cell is

0	1	0		0	1	0	0	1
0		0	1	0		0	0	1

The third cell is

0		0	1
0	1	0	1

All the cell in the above alignment have the same number of zeros. Hence $N_v^- = 0$. Now there is a way to remedy to this problem. Take cell number two. There are two ones which are "quasi" aligned: X_8 and Y_6. These two ones are only one position away from being aligned. So, if we align them, instead of the pair of zeros X_7 and Y_7, the score remains the same. When we align the pair of ones X_8 and Y_6 instead of the pair of zeros X_7 and Y_7, we split cell number two into two cells. This is how cell number two looks after this transformation:

0	1	0	0	1	0	0		1
0		0		1	0	0	0	1

Instead of the old cell number two, we observe the new cell number 2 followed by the new cell number 3. The old cell number three does not change but is renamed and becomes cell number 4. The new cell number two is equal to:

0	1	0	0	1
0		0		1

The new cell number three is

0	0		1
0	0	0	1

The advantage of breaking up a cell into two in this way, is that the new cells have different number of zeros on each side. Hence, N_v^- increases in the process whilst L_v remains the same. The breaking up process helps up get ride of the problem of having to many cells with the same number of zeros on both sides.

Let us define what we saw in the previous numerical example in a precise fashion.

Definition 4.1 *Let $k \in \mathbb{N}$, $v \in \mathbb{Z}^k \cap V$, $i \leq k$ and $v_i = 0$. We say that cell i of v can be broken up if there exists j and j' satisfying all of the following*

1. $X_j = Y_{j'} = 1$

2. $\pi(i) < j < \pi(i+1)$ *and* $\nu(i) < j' < \nu(i+1)$

3. The difference between the number of zeros in the strings

$$X_{\pi(i)+1}X_{\pi(i)+2}\dots X_{j-1} \text{ and } Y_{\nu(i)+1}Y_{\nu(i)+2}\dots Y_{j'-1}$$

is one or minus one. Hence

$$1 = \Big|\big(j - \pi(i) - \sum_{l=\pi(i)+1}^{j} X_l\big) - \big(j' - \nu(i) - \sum_{l=\nu(i)+1}^{j'} Y_l\big)\Big|$$

A cell which has different number of zeros in its X-part and in its Y-part is called a *non-zero cell.* We say that an alignment $v \in \mathbb{Z}^k$ has more than 1% non-zero cells if

$$|\{\ i \in [1,k] \mid v_i \neq 0\ \}| \geq 0.01k.$$

Recall the definition of V_n in (3.7). Let $V_{1\%}$ be the subset of V_n consisting of the alignments which have at least 1% of non-zero cells, i.e.

$$V_{1\%} := \{v \in V_n \mid v \text{ has more than } 1\% \text{ non} - \text{zero cells}\}.$$

Let

$$V^c_{1\%} := V_n - V_{1\%}.$$

4.3 The events

Recall that for a vector v we associate $|v|$ random cells $C_v(1), \dots, C_v(|v|)$ defined as a function of random i.i.d, Bernoulli random sequences $X_1, X_2, \dots$ and $Y_1, Y_2, \dots$. In the following we define some events that capture the typical behavior of these random cells. Recall that N_1 denotes the total number of ones in X and Y, $N_1 = \sum_{i=1}^{n}(X_i + Y_i)$. Let v be an admissible alignment, i.e. $v \in V$ or, equivalently, $\pi_v(|v|), \nu_v(|v|) \leq n$.
Let N_{1v} designate the number of ones up to the last cell of v:

$$N_{1v} := \Big(\sum_{j=1}^{\pi_v(|v|)} X_j + \sum_{j=1}^{\nu_v(|v|)} Y_j\Big).$$

Finally, we define the number of ones after the last cell

$$R_v = \sum_{j=\pi(|v|)+1}^{n} X_j + \sum_{j=\nu(|v|)+1}^{n} Y_j.$$

Definition 4.2

- *Let E_4 designate the event that every optimal alignment belongs to the set V_n.*
- *Let D be the event that for all $v \in V^c_{1\%}$, at least 1% of the cells can be broken up. So,*

  $$D := \bigcap_{v \in V^c_{1\%}} D_v,$$

  where D_v is the event that at least 1% of the cells $C_v(1), \ldots, C_v(|v|)$ can be broken up.
- *Let F be the event that every $v \in V_{1\%}$ has at least $2\alpha_1\%$ of ones in $C_v(1), \ldots C_v(|v|)$ on a side of less zeros. Hence,*

  $$F := \bigcap_{v \in V_{1\%}} F_v,$$

  where F_v is the event that

  $$N^-_v \geq 2\alpha_1 N_{1v}.$$
- *Let G be the event that every $v \in V_{1\%}$ has not more than $\alpha_2\%$ of matched ones. Hence*

  $$G := \bigcap_{v \in V_{1\%}} G_v,$$

  where G_v is the event that

  $$2|v| \leq \alpha_2 N_{1v}.$$
- *Let K be the event that for every optimal alignment v, the number of ones after the last cell is less than $0.1\alpha_1\%$ of the total number of ones. Hence,*

  $$K = \bigcap_{v \in V^*} K_v,$$

  where V^* is the set of optimal alignments and K_v is the event that

  $$R_v \leq 0.1\alpha_1 N_1.$$

In the next section, we shall prove that all the defined events hold with high probability. Note the importance of the breaking up notion. The events F and G together with the event K basically prove (2.1) and (2.2) for the case when the optimal alignment has at least 1% non-zero cell, i.e it belongs to $V_{1\%}$. But every optimal alignment need not belong to $V_{1\%}$. However, the event D ensures that for every alignment from $V^c_{1\%}$, there exists another alignment $v' \in V_{1\%}$ *with the same score.* So, when the events E_4 and D both hold, then there exists an optimal alignment in $V_{1\%}$. To this optimal alignment we can apply F, G and K and get the inequalities (2.1) and (2.2). These considerations lead to the next lemma, which is our main combinatorial lemma:

Lemma 4.1

$$E_4 \cap D \cap F \cap G \cap K \subset A_n. \tag{4.2}$$

Proof. Let v designate an optimal alignment, that is an alignment such that $L_v = L_n$. When E_4 holds, then v is contained in the set V_n. Assume that v contains less than 1% of cells with different number of zeros on their sides, i.e. $v \in V^c_{1\%}$. Then, the event D assures that we can break up v so that it gets more than 1% of non-zero cells and still remains optimal. Hence, there exists an optimal alignment w in V_n with at least 1% of non-zero cells. The event G guarantees that for such an alignment w there is a proportion of less than α_2% matched ones. In fact, the proportion in G is taken for the ones contained in cells, only, and not all the ones in X and Y. The proportion among all ones (hence $2|w|/N_1$ instead of $2|w|/N_{1w}$) is even smaller. Hence we get that the proportion of matched ones to N_1 is less or equal to α_2%.
The events F applies to w. Hence

$$\frac{N^-_w}{N_{1w}} \geq 2\alpha_1$$

and thus

$$\frac{N^-_w}{N_1} \cdot \frac{N_1}{N_{1w}} \geq 2\alpha_1. \tag{4.3}$$

The event K implies that

$$R_w \leq 0.1\alpha_1 N_1. \tag{4.4}$$

Since w is admissible, $N_1 = N_{1w} + R_w$. Using the last equality with (4.4) and (4.3) yields

$$\frac{N^-_w}{N_1} \cdot \frac{1}{1 - 0.1\alpha_1} \geq 2\alpha_1. \tag{4.5}$$

Since $\alpha_1 \leq 2/3$ we have that $1/(1 - 0.1\alpha_1)$ is less than 2. In this manner, equation (4.5) becomes

$$\frac{N^-_w}{N_1} \geq \alpha_1.$$

Summarizing: the alignment w is optimal. Its number of matched ones to N_1 is less equal than α_2. The proportion of ones on sides with less zeros to N_1 is above α_1. Hence the event A_n holds. We have just finished proving the inclusion (4.2). ■

4.4 Proof of Theorem 2.2

¿From (4.2) it follows that

$$\mathbf{P}(A^c_n) \leq \mathbf{P}(E^c_4) + \mathbf{P}(D^c) + \mathbf{P}(F^c) + \mathbf{P}(G^c) + \mathbf{P}(K^c). \tag{4.6}$$

So, the proof of Theorem 2.2 is accomplished, if we show that there exists $\alpha_1 > \alpha_2 > 0$ and ϵ_0 such that the events $\mathbf{P}(E^c_4)$, $\mathbf{P}(D^c)$, $\mathbf{P}(F^c)$, $\mathbf{P}(G^c)$ and $\mathbf{P}(K^c)$ are exponentially small in n, provided $\epsilon \leq \epsilon_0$. In Lemma 6.7, we prove the existence of constants $\alpha_1 > 0$ and C_F, not depending on ϵ, as well as a constant $c_F(\epsilon)$ such that $\mathbf{P}(F^c) \leq C_F \exp[-c_F n]$.

In Lemma 6.9, we prove that for every $0 < \alpha_2 < \alpha_1$, there exists ϵ_0, depending on α_2, such that for every $\epsilon \leq \epsilon_0$, $\mathbf{P}(G^c) \leq C_G \exp[-c_G n]$, where C_G and c_G are some constants (possibly depending on ϵ). In Lemmas 5.2, 6.3 and 6.10, we prove the existence of finite constants c_E, c_D, c_K as well as C_E, C_D, C_K, possibly depending on ϵ, such that

$$\mathbf{P}(D^c) \leq C_D \exp[-c_D n], \quad \mathbf{P}(E_4^c) \leq C_E \exp[-c_E n], \quad \mathbf{P}(K^c) \leq C_K \exp[-c_K n].$$

This finishes the proof of theorem 2.2.

The proofs that D^c, F^c, G^c and K^c all have exponentially small probability in n uses the representation of alignments as elements of V_n. All these events state that a certain property holds for every alignment in V_n. The proof that they have high probability goes as follows: For one non-random alignment $v \in V_n$, the cells are independent. Hence, one can use large deviation. It then only remains to prove that the large deviation rate beats the number of elements in the set V_n.

5 Preliminary bounds

A rough lower bound for the typical length of the LCS, is obtained as follows.

1. First only align all the zeros you can. You get approximately a common subsequence of length $(1-\epsilon)n$ consisting only of zero's.

2. Having aligned as many zeros as you could in 1, take the ones which can be aligned without disturbing the already aligned zeros. The sequence X has approximatively ϵn one's. The probability that a one in X can be matched with a one in Y without disturbing the already existing alignment of zero's is ϵ. Hence, the number of ones we get to align in this way is about $\epsilon^2 n$.

In the way described above we get a common subsequence of length about

$$[(1-\epsilon)+\epsilon^2]n. \tag{5.1}$$

To stay on the safe side, we bound L_n by a quantity that is little smaller than (5.1); we take $[(1-\epsilon)+0.9\epsilon^2]n$.
Let E denote the event that the LCS is longer than $\big((1-\epsilon)+0.9\epsilon^2\big)n$, i.e.

$$E := \{L_n \geq \big((1-\epsilon)+0.9\epsilon^2\big)n\}.$$

Lemma 5.1 *For every* $0.5 \geq \epsilon > 0$ *there exist a constant* $a(\epsilon) > 0$ *such that*

$$\mathbf{P}(E) \geq 1 - 8e^{-an}.$$

Proof. Let $\alpha \in (0, 0.5)$. Define the events (they depend on α)

$$E_2^x := \Big\{ |\sum_{i=1}^n X_i - n\epsilon| \leq \alpha\epsilon n \Big\} \quad E_2^y := \Big\{ |\sum_{i=1}^n Y_i - n\epsilon| \leq \alpha\epsilon n \Big\}.$$

When E_2^x holds, then $X_1,\ldots,X_n$ has at least $(1-(1+\alpha)\epsilon)n$ zeros and at least $\epsilon(1-\alpha)n$ ones. On E_2^y, the same holds for $Y_1,\ldots,Y_n$. Let

$$E_2 := E_2^x \cap E_2^y.$$

When E_2 holds, then the longest common subsequence is at least $(1-(1+\alpha)\epsilon)n$, because at least so many zeros can be aligned.
Let τ_x be the position of the last 0 in $X_1,\ldots,X_n$, let τ_y be the position of the last 0 in $Y_1,\ldots,Y_n$. Define

$$E_1^x := \{n-\tau_x \leq \alpha\epsilon n\}, \quad E_1^y := \{n-\tau_y \leq \alpha\epsilon n\}, \quad E_1 := E_1^x \cap E_1^y.$$

When $E_1 \cap E_2$ holds, then X and Y both have at least $\epsilon(1-\alpha)n$ ones and at least $m := \epsilon(1-2\alpha)n$ of them are located before the last 0. In terms of ξ_i's and η_i' as defined in Subsection 3.1, it means that

$$E_1^x \cap E_2^x \subset \{\sum_{i=1}^{N_0^x} \xi_i \geq m\}, \quad E_1^y \cap E_2^y \subset \{\sum_{i=1}^{N_0^y} \eta_i \geq m\}, \tag{5.2}$$

where N_0^x and N_0^y are the number of zero's in X and Y respectively. Here $\emptyset$ is identified with 0. We are interested in calculating the probability that among these m ones at least $\epsilon(1-\alpha)m$ can be aligned without destroying the already existing alignment of zero's. This event is $E_3 := E_3^x \cap E_3^y$, where

$$E_3^x := \Big\{\sum_{i=1}^{N_0^x} \eta_i I_{\{\eta_i \leq \xi_i\}} \geq m\epsilon(1-\alpha)\Big\}, \quad E_3^y := \Big\{\sum_{i=1}^{N_0^y} \eta_i I_{\{\xi_i \leq \eta_i\}} \geq m\epsilon(1-\alpha)\Big\}.$$

(here, again $\emptyset$ is identified with 0). The event E_3^x states that before the last zero in X, at least $\epsilon(1-\alpha)m$ ones can be aligned and the event E_3^y states that before the last zero in Y, at least $\epsilon(1-\alpha)m$ ones can be aligned. If they both hold, then at least $\epsilon(1-\alpha)m$ ones before the last aligned zero can be aligned, so

$$E_1 \cap E_2 \cap E_3 \subset \{L_n \geq (1-(1+\alpha)\epsilon)n + \epsilon(1-\alpha)m\} =: E(\alpha).$$

Let us bound the probabilities. Clearly, for an integer $a \geq 0$, $\mathbf{P}(n-\tau_x \geq a) = \epsilon^a$, implying that

$$\mathbf{P}(E_1^y) = \mathbf{P}(E_1^x) = 1-\epsilon^{(\alpha\epsilon n+1)} = 1-\epsilon\exp[(\ln\epsilon)\alpha\epsilon n] \geq 1-\exp[-2(\alpha\epsilon)^2 n],$$

where the last inequality follows from the fact that $\epsilon \leq \frac{1}{2}$. By Höffding's inequality,

$$\mathbf{P}\big((E_2^x)^c\big) \leq 2\exp[-2(\alpha\epsilon)^2 n], \quad \mathbf{P}\big((E_2^y)^c\big) \leq 2\exp[-2(\alpha\epsilon)^2 n].$$

Let X be such that at least m one's are located before the last zero. Then, it is not hard to see that

$$\mathbf{P}\big((E_3^x)^c|X\big) = \mathbf{P}(\sum_{i=1}^{m} \zeta_i < m\epsilon(1-\alpha)) \leq \exp[-2(\epsilon\alpha)^2 m],$$

where ζ_i are i.i.d. Bernoulli random variable with parameter ϵ. The last inequality follows from Höffding's inequality. Hence, from (5.2), it follows that $\mathbf{P}\big((E_3^x)^c \cap E_1^x \cap E_2^x\big) \leq \exp[-2(\epsilon\alpha)^2 m]$ and

$$\mathbf{P}\big((E_3^x)^c\big) \leq \exp[-2(\epsilon\alpha)^2 m]+\mathbf{P}(E_2^{xc})+\mathbf{P}(E_1^{xc}) \leq \exp[-2(\epsilon\alpha)^2\epsilon(1-2\alpha)n]+3\exp[-2(\alpha\epsilon)^2 n].$$

By symmetry, the same bound holds for $\mathbf{P}\big((E_3^y)^c\big)$ and so

$$\begin{aligned}\mathbf{P}(E^c(\alpha)) &\leq 2\cdot 3\exp[-2(\alpha\epsilon)^2 n] + 2\exp[-2(\epsilon\alpha)^2\epsilon(1-2\alpha)n]\\ &\leq 8\exp[-2(\epsilon\alpha)^2\epsilon(1-2\alpha)n].\end{aligned}$$

Let $\alpha_o(\epsilon)$ be so small that $(1-(1+\alpha)\epsilon)n + \epsilon(1-\alpha)\epsilon(1-2\alpha) > 1-\epsilon+0.9\epsilon^2$, if $\alpha < \alpha_o$. So, if $\alpha < \alpha_o$, then $E(\alpha) \subset E$ and

$$\mathbf{P}(E^c) \leq 8\exp[-2(\epsilon\alpha)^2\epsilon(1-2\alpha)n] = 8\exp[-an],$$

where $a(\epsilon) = 2(\epsilon\alpha)^2\epsilon(1-2\alpha)$. ■

Note that Lemma 5.1 gives a lower bound for the Chvatal-Sankoff constant: $(1-\epsilon)+\epsilon^2$. For $\epsilon = 0.5$, the lover bound is 0.75.

If $0 < \alpha \leq 0.8\epsilon$, then on E_2

$$N_0^x \leq n[(1-\epsilon)+0.8\epsilon^2], \quad N_0^y \leq n[(1-\epsilon)+0.8\epsilon^2], \tag{5.3}$$

where N_0^x and N_0^y are the number of zeros in X and Y, respectively. In this case, hence,

$$\frac{N_0}{2} \leq n[(1-\epsilon)+0.8\epsilon^2], \tag{5.4}$$

where N_0 is the number of 0's in X and Y. On the other hand, if E holds, then

$$L_n \geq n[(1-\epsilon)+0.9\epsilon^2]. \tag{5.5}$$

So, if $0 < \alpha \leq 0.8$ and $E(\alpha)\cap E$ hold, then

$$\frac{N_0}{2} + (0.1)\epsilon^2 n \leq L_n. \tag{5.6}$$

As explained in Subsection 3.2, (5.6) implies (3.6), i.e. $\sum_{i=1}^k |v_i| \leq 2k$. We also showed that (5.6) implies that in any optimal alignment there are at least $(0.1)\epsilon^2 n$ pears of aligned ones. Thus, if $0 < \alpha \leq 0.8$ and $E(\alpha)\cap E$ hold, then any optimal alignment must belong to V_n, where the set of alignments V_n has been defined in (3.7). Recall that E_4 designates the event that every optimal alignment belongs to V_n.

Lemma 5.2

$$\mathbf{P}(E_4) \geq 1 - 8\exp[-an] - 4\exp[-2(0.8\epsilon)^2\epsilon n].$$

Proof. We saw that $E_2(0.8\epsilon) \cap E \subset E_4$. Proposition 3.1 now finishes the proof. ■

Lemma 5.3

$$|V(k)| < 2^k C_k^{3k} < 16^k. \tag{5.7}$$

Proof. Let

$$V^+(k) = \{(v_1, \dots, v_k) \in \mathbb{Z}^+ : v_1 + \cdots + v_k \leq 2k\},$$

where $\mathbb{Z}^+ = \{0, 1, \dots\}$. Thus, $|V^+(k)|$ is the number of k-dimensional vectors with non-negative integer entries and summing up to at most $2k$. By adding one more component, we get that $|V^+(k)|$ is equal to the number of $k+1$-dimensional vectors with non-negative integer entries and summing up to exactly $2k$. The number of such vectors is $C_{k+1-1}^{2k+k+1-1} = C_k^{3k}$. It follows that

$$|V^+(k)| = C_k^{3k} < 2^{3k}.$$

For every k-dimensional vector, there are at most 2^k ways to assign the signs of the entries. This then yields

$$|V(k)| \leq 2^k C_k^{3k} < 2^{4k} = 16^k.$$

■

6 Bounding the probabilities

6.1 Combinatorics

Let

$$I(v_1, \dots, v_k) = |\{i \in \{0, \dots, k\} : v_i \neq 0\}|.$$

Lemma 6.1

$$|V_{1\%}^c(k)| \leq \exp[(2.01 \cdot 0.0315 + 0.7 \cdot 0.01)k] = \exp[(0.063315 + 0.007)k] = \exp[0.070315k], \tag{6.1}$$

where

$$V_{1\%}^c(k) := V(k) \cap \{(v_1, \dots, v_k) \in \mathbb{Z} : I(v_1, \dots, v_k) \leq 0.01k\}.$$

Proof. Without loss of generality assume that $0.01k$ is an integer. Consider the set of $0.01k$-dimensional vectors with non-negative integer entries and summing up to at most $2k$. Let this set be

$$W^+(k) := \{(w_1, \dots, w_{0.01k}) \in \mathbb{Z}^{+0.01k} : \sum_{i=1}^{0.01k} w_i \leq 2k\}.$$

We know that

$$|W^+(k)| = C_{0.01k+1-1}^{2k+0.01k+1-1} = C_{0.01k}^{2.01k} = C_{\frac{0.01}{2.01}(2.01)k}^{2.01k} < 2^{2.01kH(\frac{0.01}{2.01})} < 2^{2.01H(0.005)k},$$

where H is the binary entropy function. There is $2^{0.01k}$ ways to assign the signs. Thus, we find:

$$|W(k)| = 2^{((2.01)H(0.005)+0.01)k},$$

where

$$W(k) := \{(w_1, \ldots, w_{0.01k}) \in \mathbb{Z}^{0.01k} : \sum_{i=1}^{0.01k} |w_i| \le 2k\}.$$

Obviously,

$$|V^c_{10\%}(k)| = |W(k)|.$$

So

$$\begin{aligned} |V^c_{1\%}(k)| &= 2^{((2.01)H(0.005)+0.01)k} = \exp[\ln 2(2.01H(0.005) + 0.01)k] \\ &= \exp[(2.01H_e(0.005) + \ln 2(0.01))k]. \end{aligned}$$

Since $(\ln 2)H(0.005) = H_e(0.005) \le 0.0315$ and $\ln 2 < 0.7$, we get (6.1). ■

6.2 The event D

Recall that D_v denotes the event that 1% of the cells of the alignment v can be broken up.

Lemma 6.2 *Let $v \in V^c_{1\%}(k)$. Then*

$$\mathbf{P}(D^c_v) \le \exp[-0.089k]. \tag{6.2}$$

Proof. Let us calculate the probability that a 0-cell is breakable. For this, we use the approach introduced in Subsection 3.1. Recall the definition of T in (3.4). With this construction, being breakable means the existence of $(\xi_i, \eta_i), (\xi_{i+1}, \eta_{i+1})$ such that

$$\xi_i \neq \emptyset, \eta_i = \emptyset, \xi_{i+1} = \emptyset, \eta_i \neq \emptyset$$

or

$$\xi_i = \emptyset, \eta_i \neq \emptyset, \xi_{i+1} \neq \emptyset, \eta_i = \emptyset.$$

Let

$$\begin{aligned} U_1 &: = \min\{i = 2, \ldots : \xi_{i-1} \neq \emptyset, \eta_{i-1} = \emptyset, \xi_i = \emptyset, \eta_i \neq \emptyset\} \\ U_2 &: = \min\{i = 2, \ldots : \xi_{i-1} = \emptyset, \eta_{i-1} \neq \emptyset, \xi_i \neq \emptyset, \eta_i = \emptyset\}, \\ U &: = U_1 \wedge U_2. \end{aligned}$$

Let

$$\mathcal{X} := \{\emptyset, 1, 2, \ldots\}, \quad \mathcal{X}^+ := \{1, 2, \ldots\}.$$

With those stopping times, the probability that a 0 cell is breakable is $\mathbf{P}(U < T)$. Let us estimate it (from below).

An easy way is to consider the disjoint pairs of indexes $(1,2),(3,4),\ldots,(2j-1,2j),\ldots$ and restrict the stopping time U take the even integers only. So, we define the independent random vectors

$$Z_j = (\xi_{2j-1}, \eta_{2j-1}, \xi_{2j}, \eta_{2j}), j = 1,2,\ldots$$

$$\begin{aligned}
U_1' &: = \min\{j = 1,2,\ldots : \xi_{2j-1} \neq \emptyset, \eta_{2j-1} = \emptyset, \xi_{2j} = \emptyset, \eta_{2j} \neq \emptyset\} = \min\{j = 1,2,\ldots : Z_j \in A_1\}\\
U_2' &: = \min\{i = 1,2,\ldots : \xi_{2j-1} = \emptyset, \eta_{2j-1} \neq \emptyset, \xi_{2j} \neq \emptyset, \eta_{2j} = \emptyset\} = \min\{j = 1,2,\ldots : Z_j \in A_2\},\\
U' &: = U_1' \wedge U_2' = \min\{j = 1,2,\ldots : Z_j \in A_2 \cup A_1\},\\
T' &: = \{j = 1,2,\ldots : Z_j \in B_1 \cup B_2\},
\end{aligned}$$

where

$$A_1 := \mathcal{X}^+ \times \emptyset \times \emptyset \times \mathcal{X}^+, \quad A_2 := \emptyset \times \mathcal{X}^+ \times \mathcal{X}^+ \times \emptyset, \quad B_1 = \mathcal{X}^+ \times \mathcal{X}^+ \times \mathcal{X} \times \mathcal{X}, \quad B_2 = \mathcal{X} \times \mathcal{X} \times \mathcal{X}^+ \times \mathcal{X}^+.$$

Clearly,

$$U' \geq U, \quad \mathbf{P}(U < T) \geq \mathbf{P}(U' < T) = \mathbf{P}(U' < T').$$

Since the random variables Z_j are independent, the latter probability is easy to calculate:

$$\mathbf{P}(U' < T') = \frac{\mathbf{P}(Z_1 \in A_2 \cup A_1)}{\mathbf{P}(Z_1 \in A_2 \cup A_1) + \mathbf{P}(Z_1 \in B_2 \cup B_1)} = \frac{2\epsilon^2(1-\epsilon)^2}{2\epsilon^2(1-\epsilon)^2 + 2\epsilon^2 - \epsilon^4} = \frac{2(1-\epsilon)^2}{2(1-\epsilon)^2 + 2 - \epsilon^2}.$$

It is easy to check that the function

$$\epsilon \mapsto q(\epsilon) := \frac{2(1-\epsilon)^2}{2(1-\epsilon)^2 + 2 - \epsilon^2}$$

is decreasing in $[0, \frac{1}{2}]$, which implies

$$q(\epsilon) \geq \frac{2(\frac{1}{2})^2}{2(\frac{1}{2})^2 + 2 - (\frac{1}{2})^2} = \frac{2}{9}.$$

Let $v = (v_1, \ldots, v_k) \in V^c_{1\%}$. This means that the number of zero cells m is at least $0.99k$. Let J be the index set of zero-cells and let for every $j \in J$, I_j be the Bernoulli variable that is one if and only if the cell v_j is breakable. Clearly, the random variables I_j are iid and $p(\epsilon) := P(I_j = 1) \geq q(\epsilon)$. Let

$$c(\epsilon) := q(\epsilon) - 0.01 \geq \frac{2}{9} - 0.01 =: c.$$

With this notation, using Höffding's inequality

$$\begin{aligned}
\mathbf{P}(D_v^c) &= \mathbf{P}\Big(\frac{\sum_{j\in J} I_j}{m} < 0.01\Big) = \mathbf{P}\Big(\frac{\sum_{j\in J} I_j}{m} - p(\epsilon) < 0.01 - p(\epsilon)\Big)\\
&\leq \mathbf{P}\Big(\frac{\sum_{j\in J} I_j}{m} - p(\epsilon) < 0.01 - q(\epsilon)\Big) = \mathbf{P}\Big(\frac{\sum_{j\in J} I_j}{m} - p(\epsilon) < -c(\epsilon)\Big) \leq \exp[-2c^2(\epsilon)m]\\
&\leq \exp[-2c^2(\epsilon)0.99k] = \exp[-1.98c^2(\epsilon)k] \leq \exp[-1.98c^2k] \leq \exp[-0.089k].
\end{aligned}$$

■

Lemma 6.3 *There exists $C_D < \infty$ such that*

$$\mathbf{P}(D^c) \leq C_D \exp[-0.018685(0.1\epsilon^2)n]. \tag{6.3}$$

Proof.

$$D(k) := \bigcap_{v \in V^c_{1\%}(k)} D_v.$$

With (6.1) and (6.2), we get

$$\mathbf{P}(D^c(k)) \leq \sum_{v \in V^c_{1\%}(k)} \mathbf{P}(D^c_v) \leq \exp[(0.070315 - 0.089)k] = \exp[-0.018685k].$$

Since $k \geq (0.1\epsilon^2)n$, we find:

$$\mathbf{P}(D^c) \leq \sum_{k \geq (0.1\epsilon^2)n} \mathbf{P}(D^c(k)) \leq \sum_{k \geq (0.1\epsilon^2)n} \exp[-0.018685k] = C_D \exp[-0.018685(0.1\epsilon^2)n],$$

where

$$C_D := \big(1 - \exp[-0.018685]\big)^{-1}.$$

■

6.3 The event F

The following large deviation result is proven in the Appendix.

Lemma 6.4 (Large deviation for geometric random variables)
Let $G_1, \ldots, G_m$ be iid random variables with geometric distribution $G(p)$. There exists $0 < \alpha_0 < 1$, not depending on p, such that for every $\alpha \leq \alpha_0$, the inequality

$$\mathbf{P}\Big(\sum_{i=1}^m G_i \leq \frac{\alpha}{p} m\Big) \leq \exp[-300m], \quad \forall m \tag{6.4}$$

holds. Moreover, for every $C > 0$ there exists $1 < A_0(C) < \infty$, such that for every $A > A_0$

$$\mathbf{P}\Big(\sum_{i=1}^m G_i > \frac{A}{p} m\Big) \leq \exp[-Cm], \quad \forall m. \tag{6.5}$$

Let u be a non-negative integer. Let us model an $-u$-cell. Recall the random variables ξ_i and η_i as in Subsection 3.1 and recall the random variable T as in (3.5), which is the smallest time T such that $\xi_i \neq \emptyset$ and $\eta_{u+i} \neq \emptyset$. Let $T_x(j)$ be the index of j-th ξ_i such that $\xi_i \neq \emptyset$. So

$$T_x(1) = \min\{i \geq 1 : \xi_i \neq \emptyset\}, \quad \ldots, \quad T_x(j+1) = \min\{i > T_y(j) : \xi_i \neq \emptyset\}.$$

Let

$$\rho^- := \min\{j = 1, 2, \dots : \eta_{u+T_x(j)} \neq \emptyset\}. \tag{6.6}$$

Hence ρ^- is the number of ξ_i's (in the cell) that are not $\emptyset$. With this notation,

$$T = T_x(\rho^-).$$

For an $-u$ cell, the number of 0-s in X is smaller then the number of 0's in Y. Let us estimate (from below) the number of 1's in the X-side, N_1^-. This number is clearly at least ρ^-, so $N_1^- \geq \rho^-$, where the equality holds if and only if

$$\xi_{T_x(j)} = 1, \quad j = 1, \dots, \rho^-.$$

The random variable ρ^- has geometric distribution with parameter ϵ. Indeed, since X and Y are independent, from the right side of (6.6) follows

$$\mathbf{P}(\rho^- = n) = \mathbf{P}(\eta_{u+T_x(1)} = \emptyset, \dots, \eta_{u+T_x(n-1)} = \emptyset, \eta_{u+T_x(n)} \neq \emptyset) = (1-\epsilon)^{n-1}\epsilon.$$

Let $v = (v_1, \dots, v_k)$. Let N_v^- be the number of ones on the sides with fewer 0's of non-0 cells. At first, we give a lower bound on N_v^-.

Lemma 6.5 *There exists a $\gamma > 0$, not depending on ϵ, such that for every $v = (v_1, \dots, v_k) \in V_{1\%}$ it holds*

$$\mathbf{P}(F_{1v}^c) \leq \exp[-3k], \quad \textit{where} \quad F_{1v} = \{N_v^- \geq \frac{\gamma}{\epsilon} k\}. \tag{6.7}$$

Proof. Let $v = (v_1, \dots, v_k) \in V_{1\%}$. Let I be the index set of non 0-cells, $|I| \geq 0.01k$. Let us estimate (below) the number of 1's in the side of fewer 0's:

$$N_v^- = \sum_{i=1}^{|v|} N_v^-(i).$$

For a cell $v_i \neq 0$, we have that $N_v^-(i) \geq \rho_i^-$, where ρ_i^-, $i \in I$ are i.i.d. Geometrically distributed random variables with parameter ϵ as in (6.6). So,

$$N_v^- \geq \sum_{i \in I} \rho_i^-. \tag{6.8}$$

Let α_o be as in Lemma 6.4. It does not depend on ϵ. Take

$$m := 0.01k, \quad \gamma := 100\alpha_o.$$

and apply Lemma 6.4:

$$\begin{aligned}
\mathbf{P}(F_{1v}^c) &\leq \mathbf{P}\Big(\sum_{i\in I} \rho_i^- \leq \frac{\gamma}{\epsilon} k\Big) \\
&\leq \mathbf{P}\Big(\sum_{i=1}^{0.01k} \rho_i^- \leq \frac{\gamma}{\epsilon} k\Big) \\
&\leq \mathbf{P}\Big(\sum_{i=1}^{m} \rho_i^- \leq \frac{\gamma}{100\epsilon} m\Big) \\
&\leq \mathbf{P}\Big(\sum_{i=1}^{m} \rho_i^- \leq \frac{\alpha_o}{\epsilon} m\Big) \\
&\leq \exp[-300(0.01)k] \\
&= \exp[-3k].
\end{aligned}$$

■

Let

$$F_1(k) := \bigcap_{v\in V_{1\%}\cap V(k)} F_{1v} \quad \text{and} \quad F_1 := \bigcap_{k\geq (0.1\epsilon^2)n} F_1(k).$$

By (5.7) and (6.7),

$$\mathbf{P}(F_1(k)^c) \leq \sum_{v\in V(k)} \mathbf{P}(F_v^c) \leq 16^k \exp[-3k] = \exp[(\ln 16 - 3)k] \leq \exp[-0.2k].$$

Hence

$$\mathbf{P}(F_1^c) \leq \sum_{k\geq (0.1\epsilon^2)n} \mathbf{P}(F_1(k)^c) \leq \sum_{k\geq (0.1\epsilon^2)n} \exp[-0.2k] = C_{1,F} \exp[-0.2(0.1\epsilon^2)n], \tag{6.9}$$

where

$$C_{1,F} := (1-\exp[-0.2])^{-1}.$$

Let $v = (v_1, \ldots, v_k) \in V(k)$ be given. Let $C_v(1), \ldots, C_v(k)$ be the corresponding cells. Let $\rho_1, \ldots, \rho_k$ be the number of non-empty ξ_i' s in the cell $C_v(i)$. Clearly $\rho_1, \ldots, \rho_k$ are independent. The distribution of ρ_j is geometric with parameter ϵ, if $v_k \leq 0$. Otherwise, there exists a Geometric random variable with parameter ϵ, say ρ_j^- such that $\rho_j^- \leq \rho_j \leq \rho_j^- + v_k$. Since $v \in V(k)$, $\sum_j |v_j| \leq 2k$. Let us estimate from above the quantity $\rho_v := \sum^k \rho_j$.

Lemma 6.6 *There exist a constant B such that for every $v = (v_1, \ldots, v_k) \in V_{1\%}$ we have*

$$\mathbf{P}(F_{2v}^c) \leq \exp[-(\ln 16 + 1)k], \quad \textit{where } F_{2v} := \Big\{\rho_v < \frac{B}{\epsilon} k\Big\}.$$

Proof. Let B be such that $B-1 > A_0(\ln 16+1)$. By (6.5),

$$\begin{aligned}
\mathbf{P}(F_{2v}^c) &= \mathbf{P}\Big(\sum_{j=1}^{k}\rho_j \ge \frac{B}{\epsilon}k\Big)\\
&\le \mathbf{P}\Big(\sum_{j:v_j\le 0}\rho_j + \sum_{j:v_j>0}(\rho_j^- + v_j) \ge \frac{B}{\epsilon}k\Big)\\
&\le \mathbf{P}\Big(\sum_{j:v_j\le 0}\rho_j + \sum_{j:v_j>0}\rho_j^- + 2k \ge \frac{B}{\epsilon}k\Big)\\
&\le \mathbf{P}\Big(\sum_{j:v_j\le 0}\rho_j + \sum_{j:v_j>0}\rho_j^- \ge \frac{B-2\epsilon}{\epsilon}k\Big)\\
&\le \mathbf{P}\Big(\sum_{j:v_j\le 0}\rho_j + \sum_{j:v_j>0}\rho_j^- \ge \frac{B-1}{\epsilon}k\Big)\\
&\le \exp[-(\ln 16+1)k].
\end{aligned}$$

■

Let

$$F_2(k) := \bigcap_{v\in V(k)}\Big\{\rho_v < \frac{B}{\epsilon}k\Big\},\quad F_2 := \bigcap_{k\ge(0.1\epsilon^2)n} F_2(k).$$

Then, similarly to (6.9),

$$\begin{aligned}
\mathbf{P}(F_{2v}^c) &\le \sum_{v\in V(k)}\mathbf{P}(F_{v3}^c) \le \exp\big[-k\big((\ln 16+1)-\ln 16\big)\big] = \exp[-k]\\
\mathbf{P}(F_2^c) &\le C_{2F}\exp[-0.1\epsilon^2 n],
\end{aligned}$$

where

$$C_{2F} := (1-\exp[-1])^{-1}.$$

Next, using Lemma 6.6, we estimate from above the random number of ones in the X-side of the cells $C_v(1),\dots,C_v(|v|)$.

Lemma 6.7 *There exists a constant $A<\infty$, independent of ϵ, such that for every $v=(v_1,\dots,v_k)\in V_{1\%}$, we have*

$$\mathbf{P}\Big(\sum_{j=1}^{\pi(k)}X_j > \frac{Ak}{\epsilon(1-\epsilon)}\Big) \le 2\exp[-(\ln 16+1)k].$$

Proof. Let $v=(v_1,\dots,v_k)\in V_{1\%}$. Note that

$$\mathbf{P}(\xi_i = k|\xi_i\ne\emptyset) = \epsilon^{k-1}(1-\epsilon),\quad k=1,2,\dots.$$

The number of 1' s on the X-side of the cell $C_v(j)$ is

$$\sum_{i=1}^{\rho(j)} G_i, \tag{6.10}$$

where G_i are i.i.d. Geometrically distributed r.v-s with parameter $1-\epsilon$ independent of $\rho(j)$. Hence,

$$\sum_{j=1}^{\pi(k)} X_j = \sum_{i=1}^{\rho_v} G_i. \tag{6.11}$$

Let B be as in the previous lemma and let A be large enough so that

$$\frac{A}{B} > A_o\Big(\frac{(\ln 16+1)}{B}\Big)$$

and define

$$F_{3v} := \Big\{\sum_{i=1}^{\frac{B}{\epsilon}k} G_i < \frac{A}{\epsilon(1-\epsilon)}k\Big\}.$$

From Lemma 6.4 with $m = \frac{B}{\epsilon}k$

$$\begin{aligned}\mathbf{P}(F_{3v}^c) &= \mathbf{P}\Big(\sum_{i=1}^{\frac{B}{\epsilon}k} G_i \geq \frac{Ak}{\epsilon(1-\epsilon)}\Big) = \mathbf{P}\Big(\sum_{i=1}^{\frac{B}{\epsilon}k} G_i \geq \frac{k}{(1-\epsilon)}\frac{B}{\epsilon}\frac{A}{B}\Big)\\ &= \mathbf{P}\Big(\sum_{i=1}^{m} G_i \geq \frac{mA}{B(1-\epsilon)}\Big) \leq \exp[-\frac{(\ln 16+1)}{B}m]\\ &< \exp[-\frac{(\ln 16+1)\epsilon}{B}m] = \exp[-(\ln 16+1)k].\end{aligned}$$

Due to (6.11), for every v,

$$F_{2,v}\cap F_{3,v} \subset \Big\{\sum_{j=1}^{\pi(k)} X_j \leq \frac{Ak}{\epsilon(1-\epsilon)}\Big\} =: F_{4,v}.$$

Lemma 6.6 finishes the proof. ■

Let

$$F_4(k) := \bigcap_{v\in V(k)} F_{4v}, \quad F_4 := \bigcap_{k\geq(0.1\epsilon^2)n} F_3(k).$$

Then, by analogue of (6.9),

$$\mathbf{P}(F_4^c(k)) \leq 2\exp\big[-k\big((\ln 16+1)-\ln 16\big)\big] = 2\exp[-k] \tag{6.12}$$

$$\mathbf{P}(F_4^c) \leq 2C_{2F}\exp[-0.1(\epsilon^2)n]. \tag{6.13}$$

Lemma 6.8 *There exists $\alpha_1 > 0$, independent of ϵ, such that for a constant $C_F < \infty$*

$$\mathbf{P}(F^c) \leq C_F \exp[-0.02\epsilon^2 n].$$

Proof. It holds

$$F_{1,v} \cap F_{4,v} \subset \Big\{ N_v^- \geq \frac{(1-\epsilon)\gamma}{A} \sum_{j=1}^{\pi(k)} X_j \Big\}.$$

So,

$$\begin{aligned} F_1 \cap F_4 &\subset \Big(\bigcap_{v \in V_{1\%}} F_{1,v} \Big) \cap \Big(\bigcap_{v \in V_{1\%}} F_{4,v} \Big) = \bigcap_{v \in V_{1\%}} \Big(F_{1,v} \cap F_{4,v} \Big) \\ &\subset \bigcap_{v \in V_{1\%}} \Big\{ N_v^- \geq \frac{(1-\epsilon)\gamma}{A} \sum_{j=1}^{\pi(k)} X_j \Big\} =: F_x \end{aligned}$$

and by (6.9) and (6.13)

$$\mathbf{P}(F_x^c) \leq \mathbf{P}(F_1^c) + \mathbf{P}(F_4^c) \leq C_{F1} \exp[-0.02\epsilon^2 n] + 2C_{F2} \exp[-0.1\epsilon^2 n].$$

By symmetry, $\mathbf{P}(F_y^c) \leq C_{F1} \exp[-0.02\epsilon^2 n] + 2C_{F2} \exp[-0.1\epsilon^2 n]$, where

$$F_y := \Big\{ N_v^- \geq \frac{(1-\epsilon)\gamma}{A} \sum_{j=1}^{\pi(k)} Y_j \Big\}.$$

Thus

$$F_x \cap F_y \subset \Big\{ 2N_v^- \geq \frac{(1-\epsilon)\gamma}{A} \sum_{j=1}^{\pi(k)} (X_j + Y_j) \Big\} \subset \Big\{ N_v^- \geq 2\alpha_1 N_{1v} \Big\} = F,$$

where

$$\alpha_1 := \frac{\gamma}{8A} \leq \frac{(1-\epsilon)\gamma}{4A}, \tag{6.14}$$

provided $\epsilon \leq 0.5$ and

$$\mathbf{P}(F^c) \leq 2C_{F1} \exp[-0.02\epsilon^2 n] + 4C_{F2} \exp[-0.1\epsilon^2 n] < (2C_{F1} + 4C_{F2}) \exp[-0.02\epsilon^2 n].$$

■

6.4 The event G

We use the notations introduced in the previous subsection. Let α_1 be as in (6.14). Fix $0 < \alpha_2 < \alpha_1$.

Lemma 6.9 *There exists an constant $C_G < \infty$ and $\epsilon_o(\alpha_2) > 0$ such that for every $\epsilon \leq \epsilon_o$*

$$\mathbf{P}(G^c) \leq C_G \exp[-(300 - \ln 16)(0.1)\epsilon^2 n].$$

Proof. Let $v \in V(k)$. From (6.11)

$$\sum_{j=1}^{\pi(k)} X_j = \sum_{j=1}^{\rho_v} G_i \geq \rho_v = \sum_{i=1}^{k} \rho_j \geq \sum_{i=1}^{k} \rho_j^-.$$

Let

$$G_v := \Big\{k \leq \alpha_2 \sum_{j=1}^{\pi(k)} X_j\Big\} = \Big\{\frac{k}{\alpha_2} \leq \sum_{i=1}^{\rho_v} G_i\Big\}.$$

Then

$$\mathbf{P}(G_v^c) \leq \mathbf{P}\Big(\sum_{i=1}^{k} \rho_j^- < \frac{k}{\alpha_2}\Big) = \mathbf{P}\Big(\sum_{i=1}^{k} \rho_j^- < \frac{\epsilon}{\alpha_2}\frac{1}{\epsilon}k\Big).$$

Let α_o be as in Lemma 6.4. Let $\epsilon_o < \alpha_2$ be such that

$$\frac{\epsilon_o}{\alpha_2} = \alpha_o.$$

Recall that ρ_i^- are iid random variables with $G(\epsilon)$ distribution. Then, by Lemma 6.4, for every $\epsilon \leq \epsilon_o$,

$$\mathbf{P}(G_v^c) \leq \exp[-300k].$$

Let

$$G(k) := \bigcap_{v \in V(k)} G_v, \quad \bigcap_{k \geq 0.1\epsilon^2} G(k) = \bigcap_{v \in V_n} G(k) \subset \bigcap_{v \in V_{1\%}} \Big\{|v| \leq \alpha_2 \sum_{j=1}^{\pi(k)} X_j\Big\} =: G_x.$$

There exists a constant $0.5C_G$ such that, for $\epsilon \leq \epsilon_o$,

$$\mathbf{P}(G_v^c(k)) \leq \exp[-(300 - \ln 16)k], \quad \mathbf{P}(G_x^c) \leq 0.5C_G \exp[-(300 - \ln 16)(0.1\epsilon^2)n].$$

Similarly $\mathbf{P}(G_y^c) \leq 0.5C_G \exp[-(300 - \ln 16)(0.1\epsilon^2)n]$, where

$$G_y := \bigcap_{v \in V_{1\%}} \Big\{|v| \leq \alpha_2 \sum_{j=1}^{\pi(k)} Y_j\Big\}.$$

Since $G := G_x \cap G_y$, we have that

$$\mathbf{P}(F^c) \leq C_G \exp[-(300 - \ln 16)(0.1\epsilon^2)n],$$

provided $\epsilon \leq \epsilon_o$. ■

6.5 The event K

Lemma 6.10 *There exists a constant C_K such that*

$$\mathbf{P}(K^c) \leq C_K \exp[-c_K n],$$

where $c_K > 0$ is an constant, depending on ϵ.

Proof. Let v be an optimal alignment of X and Y. Consider the sequences after the last cell:

$$X_{\pi(|v|)+1}, X_{\pi(|v|)+2}, \ldots, X_n \quad \text{and} \quad Y_{\nu(|v|)+1}, Y_{\nu(|v|)+2}, \ldots, Y_n. \tag{6.15}$$

Writing these sequences in terms of ξ_i and η_i, we note that there are no i such that $\eta_i \neq \emptyset$ and $\xi_i \neq \emptyset$. Otherwise there would be one more cell, which contradicts the optimality of v. Hence, $X_{\pi(|v|)+1}, X_{\pi(|v|)+2}, \ldots, X_n$ and $Y_{\nu(|v|)+1}, Y_{\nu(|v|)+2}, \ldots, Y_n$ can be written as

$$\xi_1, 0, \xi_2, 0, \ldots, \xi_{U_x} \quad \text{and} \quad \eta_1, 0, \eta_2, 0, \ldots, \eta_{U_y}$$

respectively, where U_x and U_y are the random times that satisfy $U_x < T$ and $U_y < T$ with T being as in (3.4). Hence, conditioning on v, the random number of ones in (6.15),

$$R := \sum_{i=\pi(|v|)+1}^{n} X_i + \sum_{i=\nu(|v|)+1}^{n} Y_i,$$

is bounded by the number of ones in 0-cells, i.e. $\mathbf{P}(R > r|v) \leq \mathbf{P}(\zeta > r)$, $r = 0, 1, 2, \ldots$, where

$$\zeta := \sum_{i=1}^{T} (\xi_i + \eta_i).$$

(Here, by summing $\emptyset$ is identified with 0). Since the random variable R does not depend on v, it follows that

$$\mathbf{P}(R > r) \leq \mathbf{P}(\zeta > r) \quad \forall r \in \mathbb{N}. \tag{6.16}$$

Recall that

$$N_{1v} = \sum_{i=1}^{\pi(|v|)} X_i + \sum_{i=1}^{\nu(|v|)} Y_i.$$

Hence, the total number of ones in X and Y, $N_1 = N_{1v} + R$. Clearly, $R \leq \alpha_1 N_1 = \alpha_1(N_{1v} + R)$ holds if and only if

$$R \leq \frac{\alpha_1}{1-\alpha_1} N_{1v}.$$

Obviously $N_{1v} \geq 2|v|$, implying that

$$\mathbf{P}(K^c) \leq \mathbf{P}\Big(R > \frac{2\alpha_1}{1-\alpha_1}|v|\Big).$$

If E_4 holds, then every optimal v satisfies $|v| \geq (0.1)\epsilon^2 n$. Hence, by (6.16),

$$\mathbf{P}(E_4 \cap K^c) \leq \mathbf{P}\Big(R > \frac{2\alpha_1}{1-\alpha_1}(0.1)\epsilon^2 n\Big) \leq \mathbf{P}\Big(\zeta > \frac{2\alpha_1}{1-\alpha_1}(0.1)\epsilon^2 n\Big) \to 0.$$

So

$$\mathbf{P}(K^c) \leq \mathbf{P}\Big(\zeta > \frac{2\alpha_1}{1-\alpha_1}(0.1)\epsilon^2 n\Big) + \mathbf{P}\big(E_4^c\big).$$

By Lemma 5.2, there is a constant $a > 0$, depending on ϵ, such that

$$\mathbf{P}(E_4^c) \leq 8\exp[-an] - 4\exp[-2(0.8\epsilon)^2\epsilon n].$$

It remains to show that

$$\mathbf{P}\big(\zeta > \frac{2\alpha_1}{1-\alpha_1}(0.1)\epsilon^2 n\big)$$

decays exponentially fast. Since T corresponds to 0-cell, the number of non-empty ξ's before T, ρ, has $G(\epsilon)$ distribution. By (6.10),

$$\sum_{i=1}^{T} \xi_i = \sum_{i=1}^{\rho} G_i,$$

where $G_1, G_2, \ldots$ are i.i.d. random variables with $G(1-\epsilon)$ distribution independent of ρ. Let $A_0(1)$ be as in Lemma 6.4 and define

$$\delta := \frac{\beta(1-\epsilon)}{2A_0(1)}, \quad \beta := \frac{2\alpha_1}{1-\alpha_1}(0.1)\epsilon^2.$$

So

$$\mathbf{P}\Big(\sum_{i=1}^{T} \xi_i > \frac{\beta}{2}n\Big) = \mathbf{P}\Big(\sum_{i=1}^{\rho} G_i > \frac{\beta}{2}n\Big) \leq \mathbf{P}(\rho > \delta n) + \mathbf{P}\Big(\sum_{i=1}^{\delta n} G_i > \frac{\beta}{2}n\Big).$$

Clearly

$$\mathbf{P}(\rho > \delta n) = \exp[\ln(1-\epsilon)\delta n]$$

and by Lemma 6.4,

$$\mathbf{P}\Big(\sum_{i=1}^{\delta n} G_i > \frac{\beta}{2}n\Big) = \mathbf{P}\Big(\sum_{i=1}^{\delta n} G_i > \frac{A_0(1)}{1-\epsilon}\delta n\Big) \leq \exp[-\delta n].$$

Similarly,

$$\mathbf{P}\Big(\sum_{i=1}^{T} \eta_i > \frac{\beta}{2}n\Big) \leq \exp[\ln(1-\epsilon)\delta n] + \exp[-\delta n],$$

implying that

$$\mathbf{P}(\zeta > \beta n) \leq 2\exp[\ln(1-\epsilon)\delta n] + 2\exp[-\delta n] \leq 4\exp[\ln(1-\epsilon)\delta n].$$

■

7 Theorem 2.2 implies theorem 2.1. The proof

In this section, we prove that theorem 2.2 implies theorem 2.1. We use some of the techniques developed in [4].

Recall that N_1 is the total number of ones in the two strings X and Y. We already mentioned briefly the definition of the random pair of strings (X^k, Y^k) for $k \in [0, 2n]$. Let us give more details. Both strings X^k and Y^k are binary strings of length n. We proceed recursively on k. The strings X^{2n} and Y^{2n} consist only of 1's. We pick a 1 in the strings $X^{2n}Y^{2n}$ at random and change it into a 0. This way we obtain (X^{2n-1}, Y^{2n-1}). For general k, we obtain (X^{k-1}, Y^{k-1}) from (X^k, Y^k) by choosing a 1 at random in X^kY^k and changing it to the opposite value. Each one has the same probability to get chosen. We request that conditional on (X^k, Y^k), which one in (X^k, Y^k) gets chosen, is independent of $\{(X^i, Y^i)\}_{i \in [k,2n]}$. In other words, we apply the transformation $\tilde{}$, so that

$$X^{k-1} := \tilde{X^k} \quad \text{and} \quad Y^{k-1} := \tilde{Y^k}.$$

The distribution of (X^k, Y^k) is equal to the distribution of (X, Y) conditional on $N_1 = k$:

$$\mathcal{L}(X^k, Y^k) = \mathcal{L}(X, Y | N_1 = k), \tag{7.1}$$

where $\mathcal{L}(W)$ designates the distribution of the random variable W.

Let $L(k)$ designate the length of the LCS of X^k and Y^k.

Example. Take $n = 3$. Then

$$X^6 = 111, \;\; Y^6 = 111, \;\; L(6) = 3.$$

We chose in X^6Y^6 a one at random and flip it. There are 6 ones, and hence each one has a probability of $1/6$ to get chosen. Assume that the first bit of X^6 gets chosen. Then, we get that

$$X^5 = 011, \;\; Y^5 = 111, \;\; L(5) = 2.$$

There are now five ones and so each one has probability $1/5$ to get chosen. Assume it is this time the last bit of Y^5 which gets chosen. This leads to

$$X^4 = 011, \;\; Y^4 = 110, \;\; L(4) = 2.$$

We assume that $\{X^k, Y^k\}_{k \in [0,2n]}$ are independent of the random variable N_1. Picking N_1 according to its distribution gives us random strings (X^{N_1}, Y^{N_1}) that have the same distribution as (X, Y). Therefore, the length $L(N_1)$ of the LCS of (X^{N_1}, Y^{N_1}), has the same distribution as L_n. Hence

$$\mathrm{Var}[L_n] = \mathrm{Var}[L(N_1)].$$

Recall that our aim is to prove that $\mathrm{Var}[L(N_1)]$ it at least of order n. This follows from two facts: 1) the order of $\mathrm{Var}[N_1]$ is n; 2) the (random) map $k \mapsto L(k)$ typically decreases linearly on a certain scale.

The second point follows rather directly from Theorem 2.2 and is proven in Lemma 7.2.

This section is dedicated to showing that the 1) and 2) above imply the linear lower bound for $\mathrm{Var}[L(N_1)]$. There are two little technical difficulties: a) the map $k \mapsto L(k)$ does not increase in every point, but only on a certain scale; b) the increasing slope on a certain scale only holds in a domain where typically N_1 takes values, but not everywhere.

Recall that for any variables V and W,

$$\mathrm{Var}[V] = \mathrm{Var}[E[V|W]] + E[\mathrm{Var}[V|W]] \geq E[\mathrm{Var}[V|W]], \tag{7.2}$$

where $\mathrm{Var}[V|W]$ designate the variance of the conditional distribution $\mathcal{L}(V|W)$. Applying (7.2) to our case, we find:

$$\mathrm{Var}[L(N_1)] \geq E\big[\mathrm{Var}[L(N_1)\ |L(.)]\big], \tag{7.3}$$

where $L(\cdot)$ is the (random) map $k \mapsto L(k)$. Note that N_1 is independent of $L(\cdot)$.
Let I be the interval

$$I := [2\epsilon n - \sqrt{\epsilon(1-\epsilon)2n}, 2\epsilon n + \sqrt{\epsilon(1-\epsilon)2n}]. \tag{7.4}$$

Let $\tilde{N}_1$ be a random variable, independent of $L(\cdot)$ and having the distribution of N_1 conditioned on the interval I. From (7.2) follows for every fixed L that

$$\mathrm{Var}[L(N_1)] \geq \mathrm{Var}[L(N_1)|N_1 \in I]\mathbf{P}(N_1 \in I) = \mathrm{Var}[L(\tilde{N}_1)]\mathbf{P}(N_1 \in I).$$

Hence, since L and $\tilde{N}_1$ are independent,

$$E\big[\mathrm{Var}[L(N_1)|L(\cdot)]\big] \geq E\big[\mathrm{Var}[L(\tilde{N}_1)|L(\cdot)]\big]\mathbf{P}(N_1 \in I). \tag{7.5}$$

Assume that $f : \mathbb{R} \to \mathbb{R}$ is map such that, for a constant $c > 0$, $f'(x) > c$ for all $x \in \mathbb{R}$. Then, for any random variable Y, we have

$$\mathrm{Var}[f(Y)] \geq c^2 \mathrm{Var}[Y]. \tag{7.6}$$

(See [4] for the proof). Hence, if the map $L(\cdot)$ would have positive slope everywhere larger than $c > 0$, it would follow that $\mathrm{Var}[L(N_1)] \geq c \cdot \mathrm{Var}[N_1]$. Typically, the (random) map $k \mapsto L(k)$ does not strictly increase for every $k \in [0, 2n]$. But it is likely that in I it increases by a linear quantity. We are next going to formulate a lemma, proven on [4], which is a modification of inequality (7.6), for when the map $k \mapsto f(k)$ does not increase every k, but has a tendency to increase on some scale.

Lemma 7.1 *Let $c, m > 0$ be two constants. Let $f : \mathbb{Z} \to \mathbb{Z}$ be a non decreasing map that satisfies the following conditions*

$$f(j) - f(i) \leq (j - i), \quad \forall i < j \tag{7.7}$$

$$f(j) - f(i) \geq c \cdot (j - i), \quad \forall i, j \text{ such that } i + m \leq j. \tag{7.8}$$

Let B be an integer-valued random variable such that $E|f(B)| < \infty$. Then

$$\mathrm{Var}[f(B)] \geq c^2\Big(1 - \frac{2m}{c\sqrt{\mathrm{Var}[B]}}\Big)\mathrm{Var}[B]. \tag{7.9}$$

Recall the definition of I in (7.4). Let α_1 and α_2 be the constants from Theorem 2.2 and let E^n_{slope} designate the event that $\forall i, j \in I$, such that $i + n^{0.1} \leq j$, it holds

$$L(j) - L(i) \geq \alpha_3 |i - j|, \tag{7.10}$$

where

$$\alpha_3 := \frac{\alpha_1 - \alpha_2}{2}.$$

In other words, the event E^n_{slope} says that $L(\cdot)$ has a slope of at least α_3 on I, when we look only at points which are at least $n^{0.1}$ away from each other. The next lemma shows that the event E^n_{slope} has high probability, provided Theorem 2.2 holds.

Lemma 7.2 *For a constant $c_4 > 0$,*

$$\mathbf{P}(E^n_{\text{slope}}) \geq 1 - e^{c_4 \cdot n^{0.1}}, \tag{7.11}$$

provided n is sufficiently big.

Proof. Let A^k_n denote the event that the random vector (X^k, Y^k) takes the values in the set B_n defined in Theorem 2.2. So

$$A^k_n := \{(X^k, Y^k) \subset B_n\}.$$

Let A^{all}_n be the event

$$A^{\text{all}}_n := \bigcap_{k \in I} A^k_n.$$

Let

$$\Delta^k := \begin{cases} L(k-1) - L(k), & \text{when } A^k_n \text{ holds;} \\ 1, & \text{else.} \end{cases}.$$

Let $i < j$ and consider the random variable

$$\sum_{k=i+1}^{j} \Delta^k, \quad i < j.$$

When $(X^k, Y^k) = (x, y) \in B^k_n$, i.e. A^k_n holds, then Theorem 2.2 says that

$$\mathbf{P}(\Delta^k = 1 | X^k = x, Y^k = y) \geq \alpha_1,$$
$$\mathbf{P}(\Delta^k = -1 | X^k = x, Y^k = y) \leq \alpha_2,$$

implying that $E[\Delta^k | A^k_n, X^k, Y^k] \geq \alpha_1 - \alpha_2$. Since $E[\Delta^k | (A^k_n)^c] = 1 > \alpha_1 - \alpha_2$, we get

$$E(\Delta_k | X^k, Y^k) \geq \alpha_1 - \alpha_2. \tag{7.12}$$

Let, for every $k = 2n+1, \ldots, 2$,

$$\mathcal{F}_k := \sigma(X^{2n}, Y^{2n}, \ldots X^{k-1}, Y^{k-1}).$$

These σ-algebras perform a (reversed) filtration, because

$$\mathcal{F}_{2n+1} \subset \mathcal{F}_{2n} \subset \cdots \subset \mathcal{F}_2.$$

The random variable Δ_k is $\mathcal{F}_k$-measurable. Hence, $V_k := \Delta_k - E[\Delta_k|\mathcal{F}_{k+1}]$ are reversed martingale-differences. Since $-1 \leq \Delta_k \leq 1$, we can use Höffding-Azuma's inequality to obtain

$$\mathbf{P}\Big(\sum_{k=i+1}^{j} \Delta^k - \sum_{k=i+1}^{j} E[\Delta^k|\mathcal{F}_{k+1}] < -c\Big) \leq \exp[-\frac{2c^2}{4(j-i)}]. \tag{7.13}$$

The inequality (7.12) means

$$E[\Delta_k|\mathcal{F}_{k+1}] \geq \alpha_1 - \alpha_2$$

implying that

$$\sum_{k=i+1}^{j} E[\Delta^k|\mathcal{F}_{k+1}] \geq (\alpha_1 - \alpha_2)(j-i). \tag{7.14}$$

With $c = (\frac{\alpha_1 - \alpha_2}{2})(j-i)$, (7.13) and (7.14) yield

$$\mathbf{P}\Big(\sum_{k=i+1}^{j} \Delta^k < (\frac{\alpha_1 - \alpha_2}{2})(j-i)\Big) \leq$$
$$\mathbf{P}\Big(\sum_{k=i+1}^{j} \Delta^k - \sum_{k=i+1}^{j} E[\Delta^k|\mathcal{F}_{k+1}] < -(\frac{\alpha_1 - \alpha_2}{2})(j-i)\Big) \leq \exp[-\alpha(j-i)],$$

where $\alpha = \frac{1}{2}(\frac{\alpha_1 - \alpha_2}{2})^2$. So

$$\mathbf{P}\Big(\sum_{k=i+1}^{j} \Delta^k < \alpha_3(j-i)\Big) \leq \exp[-\alpha(j-i)]. \tag{7.15}$$

Let $E^n_{\Delta\text{ slope}}$ be the event that $\forall i, j \in I$, such that $2\epsilon n < i < j \leq 2\epsilon n + \sqrt{n}$ and $i + n^{0.1} \leq j$, we have:

$$\sum_{k=i}^{j} \Delta^k \geq \alpha_3|i-j|. \tag{7.16}$$

By (7.15), for n big enough, there exists a constant $c_2 > 0$ such that

$$\mathbf{P}\Big((E^n_{\Delta\text{ slope}})^c\Big) \leq n\exp[-(\alpha)n^{0.1}] \leq \exp[-c_2 \cdot n^{0.1}],$$

and hence

$$\mathbf{P}(E^n_{\Delta\text{ slope}}) \geq 1 - e^{-c_2 \cdot n^{0.1}}, \tag{7.17}$$

When the event A^{all}_n holds, then E^n_{slope} and $E^n_{\Delta\text{ slope}}$ are equivalent. Hence

$$A^{\text{all}}_n \cap E^n_{\Delta\text{ slope}} \subset E^n_{\text{slope}},$$

which implies

$$\mathbf{P}(E^{nc}_{\text{slope}}) \leq \mathbf{P}\big((A^{\text{all}}_n)^c\big) + \mathbf{P}(E^{nc}_{\Delta\,\text{slope}}). \tag{7.18}$$

Note that

$$\mathbf{P}\big((A^{\text{all}}_n)^c\big) \leq \sum_{k\in I} P(A^{kc}_n) = \sum_{k\in I} \mathbf{P}(A^c_n|N_1 = k) \leq \sum_{k\in I} \frac{\mathbf{P}(A^c_n)}{\mathbf{P}(N_1 = k)}, \tag{7.19}$$

where

$$A_n := \{(X,Y) \in B_n\}. \tag{7.20}$$

By the local central limit theorem, there exists $c_3 > 0$ such that for all $k \in I$

$$\mathbf{P}(N_1 = k) \geq \frac{1/c_3}{\sqrt{n}}.$$

Applying the last inequality to (7.19), yields

$$\mathbf{P}\big((A^{\text{all}}_n)^c\big) \leq \sqrt{2}nc_3\mathbf{P}(A^c_n). \tag{7.21}$$

Now the inequalities (7.17), (7.21) and (7.18) yield

$$\mathbf{P}(E^{nc}_{\text{slope}}) \leq \sqrt{2}nc_3\mathbf{P}(A^c_n) + e^{-c_2\cdot n^{0.1}}. \tag{7.22}$$

By Theorem 2.2, we have that $\mathbf{P}(A^c_n) \leq Ce^{-c_1 n}$. Applying this to (7.22) gives

$$\mathbf{P}(E^{nc}_{\text{slope}}) \leq c_3\sqrt{2}ne^{-c_1 n} + e^{-c_2\cdot n^{0.1}},$$

which finishes the proof. ■

When E^n_{slope} holds, then the map

$$L:\ I \to \mathbb{N}$$

satisfies the conditions of lemma 7.1 with $m = n^{0.1}$. Hence, when $L \in E^n_{\text{slope}}$, then

$$\text{Var}[L(\tilde{N}_1)] \geq {\alpha_3}^2\left(1 - \frac{2n^{0.1}}{\alpha_3\sqrt{\text{Var}[\tilde{N}_1]}}\right)\text{Var}[\tilde{N}_1].$$

Conditioning on E^n_{slope}, using the fact that the variance is non negative and $\tilde{N}_1$ and L are independent,

$$E\big[\text{Var}[L(\tilde{N}_1)]\big|L(\cdot)\big] \geq E\big[\ \text{Var}[L(\tilde{N}_1)|E^n_{\text{slope}}]\big]\mathbf{P}(E^n_{\text{slope}}) \geq {\alpha_3}^2\left(1 - \frac{2n^{0.1}}{\alpha_3\sqrt{\text{Var}[\tilde{N}_1]}}\right)\text{Var}[\tilde{N}_1]\mathbf{P}(E^n_{\text{slope}}).$$

Plugging the last inequality into (7.5) yields

$$E\big[\text{Var}[L(N_1)|L(\cdot)]\big] \geq {\alpha_3}^2\left(1 - \frac{2n^{0.1}}{\alpha_3\sqrt{\text{Var}[\tilde{N}_1]}}\right)\text{Var}[\tilde{N}_1]\mathbf{P}(E^n_{\text{slope}})\mathbf{P}(N_1 \in I). \tag{7.23}$$

By the central limit theorem, $\mathbf{P}(N_1 \in I)$ converges to

$$\mathbf{P}(\mathcal{N}(0,1) \in [-1,1]) > 0.$$

as $n \to \infty$. (Here $\mathcal{N}(0,1)$ designate the standard normal variable.)
Note that N_1 is a binomial variable with parameters $2n$ and ϵ. Hence, by the central limit theorem,

$$\frac{\mathrm{Var}[\tilde{N}_1]}{n} = \frac{\mathrm{Var}[N_1 | N_1 \in I]}{n} \to 2\epsilon(1-\epsilon)\mathbf{P}(\mathcal{N}(0,1) \in [-1,1])^{-1} \int_{-1}^{1} \phi(x) x^2 dx,$$

where ϕ is the standard normal density. Together with Lemma 7.2, this implies that the right side of inequality (7.23) divided by n converges to

$$\alpha_3^2 2\epsilon(1-\epsilon) \int_{-1}^{1} \phi(x) x^2 dx > 0.$$

The inequality (7.3) now finishes the proof.

8 Appendix

Proof of lemma 6.4 Let us recall a large deviation result for Bernoulli random variables. Let $X_i \sim B(1,p)$ be i.i.d. Then

$$\mathbf{P}\Big(\sum_{i=1}^{n} X_i - np > n\epsilon\Big) \leq \exp\Big[-\Big((p+\epsilon)\ln\frac{p+\epsilon}{p} + (1-(p+\epsilon))\ln\frac{1-(p+\epsilon)}{1-p}\Big)n\Big]. \quad (8.1)$$

If $p > \alpha$, then (6.4) trivially holds. If $p = \alpha$, then the probability in (6.4) equals $p^m = \exp[(\ln\alpha)m] = \exp[-\ln\frac{1}{\alpha}m]$. Hence, we consider only the case $p < \alpha < 1$. We have that,

$$\mathbf{P}\Big(\sum_{i=1}^{m} G_i \leq \frac{\alpha}{p}m\Big) = \mathbf{P}\Big(\sum_{j=1}^{\frac{\alpha}{p}m} Y_j \geq m\Big) = \mathbf{P}\Big(\sum_{j=1}^{\frac{\alpha}{p}m} Y_j \geq \frac{\alpha}{p}m\frac{p}{\alpha}\Big), \quad (8.2)$$

where Y_i are iid Bernoulli random variables with parameter p. With $n := \frac{\alpha}{p}m$ and $\frac{p}{\alpha} = p + \epsilon < 1$, we have

$$(p+\epsilon)\ln\frac{p+\epsilon}{p} + (1-(p+\epsilon))\ln\frac{1-(p+\epsilon)}{1-p} = \frac{p}{\alpha}\ln\frac{1}{\alpha} + (1-\frac{p}{\alpha})\ln\frac{1-\frac{p}{\alpha}}{1-p}, \quad p < \alpha.$$

The right side of (8.1) is

$$\exp\Big[-\Big((p+\epsilon)\ln\frac{p+\epsilon}{p} + (1-(p+\epsilon))\ln\frac{1-(p+\epsilon)}{1-p}\Big)n\Big] = \exp\Big[-\Big(\ln\frac{1}{\alpha} + (\frac{\alpha}{p}-1)\ln\frac{1-\frac{p}{\alpha}}{1-p}\Big)m\Big].$$

Let

$$L(\alpha,p) := \ln\frac{1}{\alpha} + (\frac{\alpha}{p}-1)\ln\frac{1-\frac{p}{\alpha}}{1-p}.$$

Since

$$\frac{d}{dp}\ln\frac{1-\frac{p}{\alpha}}{1-p} = \frac{1-p}{1-\frac{p}{\alpha}}\frac{d}{dp}\Big(\frac{1-\frac{p}{\alpha}}{1-p}\Big)= \frac{-\frac{1}{\alpha}(1-p)+(1-\frac{p}{\alpha})}{(1-p)(1-\frac{p}{\alpha})} = \frac{1-\frac{1}{\alpha}}{(1-p)(1-\frac{p}{\alpha})},$$

we have

$$\frac{d}{dp}L(\alpha,p) = \Big((\frac{\alpha}{p}-1)\ln\frac{1-\frac{p}{\alpha}}{1-p}\Big)' = (\frac{\alpha}{p}-1)'\ln\frac{1-\frac{p}{\alpha}}{1-p}+\frac{(\frac{\alpha}{p}-1)(1-\frac{1}{\alpha})}{(1-p)(1-\frac{p}{\alpha})} = \frac{\alpha}{p^2}\ln\frac{1-p}{1-\frac{p}{\alpha}}+\frac{1-\frac{1}{\alpha}}{\frac{p}{\alpha}(1-p)} > 0,$$

because for $y > 1$, $\ln\frac{1}{y} > 1-y$ (follows from $\ln x \leq x-1$) and so

$$\ln\frac{1-p}{1-\frac{p}{\alpha}} \geq 1-\frac{1-\frac{p}{\alpha}}{1-p} = \frac{p(\frac{1}{\alpha}-1)}{1-p}.$$

For every $\alpha < 1$, the function $p \mapsto L(\alpha,p)$, $p \in [0,\alpha]$ is increasing with a maximum at $L(\alpha,\alpha) = \ln\frac{1}{\alpha}$ and a minimum at $L(\alpha,0) = \ln\frac{1}{\alpha}+\alpha-1$. Take $\alpha_0 := \exp[-301]$. Then, if $\alpha \leq \alpha_0$, we have that $L(\alpha,p) \geq L(\alpha,0) = \ln\frac{1}{\alpha}+\alpha-1 > \ln\frac{1}{\alpha}-1 \geq \ln\frac{1}{\alpha_0}-1 = 300$.

By large deviation, for $A > 1$

$$\mathbf{P}\left(\sum_{i=1}^{m} G_i > \frac{A}{p}m\right) = \mathbf{P}\left(\sum_{i=1}^{m} G_i - \frac{A}{p}m > 0\right) \leq \exp[-\rho(A,p,s)m], \qquad (8.3)$$

where $\rho(A,p,s) = -\ln M_{A,p}(s)$, $s < \ln\frac{1}{1-p}$ and $M_{A,p}(s)$ is the moment generating function of $G_1 - \frac{A}{p}$,

$$M_{A,p}(s) = \frac{pe^{s(1-\frac{A}{p})}}{1-(1-p)e^s} = \frac{pe^{-\frac{As}{p}}}{e^{-s}-(1-p)}.$$

Hence

$$M_{A,p}(\frac{p}{2}) = \frac{pe^{-\frac{A}{2}}}{e^{-\frac{p}{2}}-(1-p)}.$$

Since

$$\Big(\frac{pe^{-\frac{A}{2}}}{e^{-\frac{p}{2}}-(1-p)}\Big)' = e^{-\frac{A}{2}}\frac{e^{-\frac{p}{2}}-(1-p)+\frac{p}{2}e^{-\frac{p}{2}}-p}{(e^{-\frac{p}{2}}-(1-p))^2} = e^{-\frac{A}{2}}\frac{e^{-\frac{p}{2}}(1+\frac{p}{2})-1}{(e^{-\frac{p}{2}}-(1-p))^2} \leq 0,$$

(because $e^x \geq 1+x$, for $x \geq 0$) the function $p \mapsto M_{A,p}(\frac{p}{2})$ is non-increasing. Since

$$\lim_{p\to 1} M_{A,p}(\frac{p}{2}) = \lim_{p\to 0}\frac{pe^{-\frac{A}{2}}}{e^{-\frac{p}{2}}-(1-p)} = \exp[\frac{1}{2}(1-A)]$$

$$\lim_{p\to 0} M_{A,p}(\frac{p}{2}) = \lim_{p\to 0}\frac{pe^{-\frac{A}{2}}}{e^{-\frac{p}{2}}-(1-p)} = 2e^{-\frac{A}{2}},$$

we have that the right side of (8.3) is smaller than $\exp[-Cm]$ as soon as A is so big that $2\exp[-\frac{A}{2}] < \exp[-C]$.

References

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

[2] 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.

[3] 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.

[4] Federico Bonetto and Heinrich Matzinger. Fluctuations of the longest common subsequence in the case of 2- and 3-letter alphabets. *submitted*, 2004.

[5] Federico Bonetto and Heinrich Matzinger. Simulations for the longest common subsequence problem. *in preparation*, 2004.

[6] J. Boutet de Monvel. Extensive simulations for longest common subsequences. *Eur. Phys. J. B*, 7:293–308, 1999.

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

[8] Christian Houdre, Jüri Lember, and Heinrich Matzinger. On the longest common increasing binary subsequence. *submitted*, 2005.

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

[10] Jüri Lember, Heinrich Matzinger, and Clemont Durringer. Deviation from mean in sequence comparison with a periodic sequence. *submitted*, 2004.

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

[12] M. Waterman. *Introduction to Computational Biology.* Chapman & Hall, 1995.

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

[14] M.S. Waterman and M. Vingron. Sequence comparison significance and Poisson approximation. *Statistical Science*, 9(3):367–381, 1994.