FLUCTUATION OF THE LENGTH OF THE LONGEST COMMON SUBSEQUENCE

J. Lember*, H. Matzinger

December 7, 2005

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

GEORGIA TECH
School of Mathematics
Atlanta, Georgia 30332-0160, U.S.A.
E-mail: matzing@math.gatech.edu

Abstract. Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be two independent sequences of i.i.d Bernoulli variables with parameter $\epsilon > 0$. Let X designate the string $X := X_1 X_2 \dots X_n$ and let $Y := Y_1 Y_2 \dots Y_n$. Let L_n designate the length of the longest common subsequence (LCS) of X and Y. We prove that for a constant $c > 0$, $\mathrm{VAR}[L_n] > cn$ if $\epsilon > 0$ is taken small enough. Hence for small ϵ, the order of magnitude of $\mathrm{VAR}[L_n]$ is $\Theta(n)$. For small ϵ, this rejects the Chvatal-Sankoff conjecture that $\mathrm{VAR}[L_n] = o(n^{\frac{2}{3}})$ in [7] and answers to Waterman's question, whether the linear bound on $\mathrm{VAR}[L_n]$ can be improved [14].

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

AMS. 60K35, 41A25, 60C05

1 Introduction

Throughout this paper $X_1, X_2, \dots$ and $Y_1, Y_2, \dots$ 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 $X := X_1 X_2 \dots X_n$, $Y := Y_1 Y_2 \dots Y_n$. A common subsequence of X and Y is a subsequence that is contained in X and in Y. Formally, a common subsequence of X and Y consists of two subsets of indices $\{i_1, \dots, i_k\}, \{j_1, \dots, j_k\} \subset \{1, \dots, n\}$ such that

$$X_{i_1} = Y_{i_1}, X_{i_2} = Y_{i_2}, \dots, X_{i_k} = Y_{i_k}.$$

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

The length of such a common subsequence is k. The longest common subsequence (LCS) of X and Y is any common subsequence that has the longest possible length, denoted by L_n. The random variable L_n is the main object of the paper.

Example. 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. So 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 the great similarity. The reason is that in the words there are missing letters. So the position of the letters in the words got shifted.
To take into account the missing letters or added letters, we align the two words allowing for gaps. We allow only the same letter 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 subsequence is a common subsequence of X and Y. Hence, the longest common subsequence is the maximum number of same letters we can align allowing caps. 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.

The longest common subsequence is a very important tool in computational biology, where it is used for comparing the DNA- and protein-alignments (see, e.g. [13, 15, 2]). It is also used in many other areas like computational linguistics, speech recognition and so on. In all those applications, when two strings have a relatively long common subsequence, then they are considered to be somehow related. On the other hand, it is clear that also two independent random strings have a longest common subsequence with length L_n. To be able to distinguish the related pairs from a random match, the asymptotic behavior of L_n (the length of the LCS of two independent random string) should be studied. For that reason the random variable L_n has been attracted the interests already for many decades. However, despite the relatively long history, its behavior is to large extent still unknown. 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 subadditivity argument

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

(see, e.g. [1, 15]). 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 $\gamma \approx 0.81$ [12, 9, 3]. For a smaller ϵ, γ is even bigger. Thus the proportion

of a common subsequence for two independent Bernoulli sequences is relatively big and, hence, to do some inferences, the information about the variance $\mathrm{VAR}[L_n]$ is essential. Unfortunately, not much is known about $\mathrm{VAR}[L_n]$ and its asymptotic order of the fluctuation is one of the main long standing open problems concerning LCS. Monte-Carlo simulations lead Chvatal and Sankoff in [7] to their famous conjecture that for $\epsilon = 0.5$, $\mathrm{VAR}[L_n] = o(n^{\frac{2}{3}})$. Using an Efron-Stein type of inequality, Steele [12] proved that in this case, $\mathrm{VAR}[L_n] \leq \mathbf{P}(X_1 \neq Y_1)n$. In [14], Waterman asks whether the linear bound can be improved. He performs several simulations which indicate that this is not the case and $\mathrm{VAR}[L_n]$ grows linearly in n, indeed. Boutet de Monvel [6] interprets his simulation in that way too. On the other hand, for a closely related Bernoulli matching model, Majumdar and Nechaev [11] obtained faster rate $O(n^{\frac{2}{3}})$.

In a series of papers, we investigate the asymptotic behavior of $\mathrm{VAR}[L_n]$ in various setup. The goal is to answer to the Waterman's question and show that the linear bound cannot be improved. More precisely, we conjecture the existence of a constant $c > 0$ such that $n\mathbf{P}(X_1 \neq Y_1) \geq \mathrm{VAR}[L_n] \geq cn$. This is written $\mathrm{VAR}[L_n] = \Theta(n)$. The simulations [5] indicate that except maybe for ϵ very close to 0.5, the conjecture holds true. In [4], Bonetto and Matzinger consider the asymmetric case where the random variables in X are Bernoulli with $1/2$, but the ones in Y can take 3 symbols. They prove that in this case $\mathrm{VAR}[L_n] = \Theta(n)$. In [8], the asymptotic behavior of the longest common increasing subsequence of two independent Bernoulli sequences was considered. This means that the common subsequences of interest must be increasing. Under this additional restriction, it is shown that $n^{-1/2}(L_n - EL_n)$ converges in law, so that $\mathrm{VAR}[L_n] = \Theta(n)$ holds again. In [10], it was showed that $\mathrm{VAR}[L_n] = \Theta(n)$ also when Y is a non-random periodic binary sequence and X consists of iid Bernoulli $1/2$ random variables. This results gives an insight that the linear growth might also hold for the case X and Y are both random (with arbitrary ϵ). Indeed, regarding L_n as a function of X and Y, by conditioning on Y, one obtains that $\mathrm{VAR}[L_n(X,Y)] \geq E\big(\mathrm{VAR}[L_n(X,Y)|Y]\big)$. So, to show that there exists a constant $c > 0$ such that $\mathrm{VAR}[L_n] \geq cn$, it suffices to show that $\mathrm{VAR}[L_n(X,Y)|Y] \geq cn$ holds for every possible outcome of Y. If Y consists of ones, only, then L_n is the number of ones in X and $\mathrm{VAR}[L_n] = \epsilon(1-\epsilon)n$. If Y is such that $Y_1 = \cdots = Y_{\frac{n}{2}} = 1$ and $Y_{\frac{n}{2}+1} = \cdots = Y_n = 0$, then it is intuitively clear that a longest common subsequence basically matches the ones in the first half of X and zeros in the second half and therefore the growth of $\mathrm{VAR}[L_n(X,Y)|Y]$ is linear as well. Here the reason of the linear growth of the variation is that, though Y has fifty percent ones, they are all gathered together so that Y has long unicolor blocks. A periodic Y has totally opposite nature – the ones and zeros are mixed as much as possible. As mentioned above, the desired constant c still exists. For a random Y, both considered realizations are highly untypical. However, since they represent, in some sense, extreme cases, we have a reason to believe that the linear growth of variance also holds for a typical realization, so that $\mathrm{VAR}[L_n] \geq cn$. In the present paper, we prove it when ϵ is sufficiently small.

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 non-standard First Passage Percolation (FPP) problem on an oriented graph with correlated weights. But for standard FPP, the question of the exact order of the fluctuation has been open for decades.

2 The main result

The main result of this paper, theorem 2.1, asserts that when $\epsilon > 0$ is small, then there exists a constant $c > 0$ such that $\mathrm{VAR}[L_n] > cn$. Then $\mathrm{VAR}[L_n] = \Theta(n)$.

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*

$$VAR[L_n] \geq c(\epsilon)n, \quad \forall n.$$

One of the main tools used in this paper is a map that we picks an one in the text X or Y at random and changes it into a zero. Let $\tilde{X}$ and $\tilde{Y}$ be 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 $x = x_1x_2 \dots x_n$, 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 $y = y_1y_2 \dots y_n$. Let, for given strings x and y, U be a random variable with uniform distribution in $\{1, 2, \dots, N_1^x + N_1^y\}$. Let $\tilde{x} = \tilde{x}_1\tilde{x}_2 \dots \tilde{x}_n$ and $\tilde{y} = \tilde{y}_1\tilde{y}_2 \dots \tilde{y}_n$ be 2 random vectors defined as follows

$$\tilde{x}_j := x_j I_{\{\sum_{i=1}^j x_i \neq U\}}, \quad \tilde{y}_j := y_j I_{\{\sum_{i=1}^j y_i \neq U - N_1^x\}}, \quad j = 1, \dots, n, .$$

Apply now the transformation $\tilde{}$ to the random vectors X and Y so that the additional randomness U depends on X and Y through N_1^x and N_1^y, only. So, the binary random vectors $\tilde{X}$ and $\tilde{Y}$ are such that

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

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

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

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 $\epsilon \leq \epsilon_o$ there exist a constant $c_1 > 0$, depending on ϵ but not depending on n such that

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

provided n is sufficiently big.

In section 3 we prove that theorem 2.2 implies theorem 2.1.

Let us give a heuristic argument for why theorem 2.2 holds. Let N_1 denote the total number of ones in both texts, X and Y, $N_1 := \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i$. Recall that we take $\epsilon > 0$ small. 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 one randomly picked one into zero reduces the number of ones and, most likely, increases the score. It can decrease the score only if the chosen one is used in the longest common subsequence. The proportion of such ones is small. The heuristic argument for theorem 2.1 is now also clear: L_n is approximately $n - \epsilon N_1$. Since, N_1 has variance of order n it follows that L_n must have the same order for its variance.

Example. Let $X = 00010000100000000000001$, $Y = 00010000000010000100000$. The longest common subsequence Z is $Z = 00010000000000000000$. 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. And hence the total number of ones appearing in the texts tends to be negatively correlated with L_n.

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 defined 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 extremely complicated correlations between the different part of the optimal alignment.

3 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 going define a random pair of strings X^k, Y^k, $k \in [0, 2n]$ recursively. 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. In other words, we apply the transformation $\tilde{}$, so that

$$X^{k-1} := \tilde{X^k}, Y^{k-1} := \tilde{Y^k}.$$

In this way, we find that the distribution of (X^k, Y^k) is equal to the distribution of (X, Y) conditional on $N_1 = k$,

$$\mathcal{L}(X^k, Y^k) = \mathcal{L}(X, Y | N_1 = k), \tag{3.1}$$

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. Picking now N_1 according to its distribution and proceeding in the above described manner gives us random vector (X^{N_1}, Y^{N_1}) that have same distribution as (X, Y). Therefore, the length of LCS of (X^{N_1}, Y^{N_1}), $L(N_1)$, has the same distribution as L_n. Hence

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

Recall that our aim is to bound $\mathrm{VAR}[L(N_1)]$ below. The variance of $L(N_1)$ has two sources: the random number of ones N_1 and the random mapping L. In the following, we show that N_1 has such a big influence on $L(N_1)$, so that the variance $\mathrm{VAR}[L(N_1)]$ is essentially bounded by the variance of N_1. The latter is, obviously, linear on n.
Recall that for any variables V and W we have

$$\mathrm{VAR}[V] = \mathrm{VAR}[E[V|W]] + E[\mathrm{VAR}[V|W]] \geq E[\mathrm{VAR}[V|W]], \tag{3.2}$$

where $\mathrm{VAR}[V|W]$ designate the variance of the conditional distribution $\mathcal{L}(V|W)$. Applying (3.2) to our case, we find:

$$\mathrm{VAR}[L(N_1)] \geq E[\ \mathrm{VAR}[L(N_1)\ |L(.)], \tag{3.3}$$

where $L(\cdot)$ is the (random) map $k \mapsto L(k)$. The law of total probability implies that

$$E[\mathrm{VAR}[L(N_1)|L(\cdot)] =$$
$$E[\mathrm{VAR}[L(N_1)|L(\cdot), N_1 \in I] \cdot \mathbf{P}(N_1 \in I) + E[\mathrm{VAR}[L(N_1)|L(\cdot), N_1 \notin I] \cdot \mathbf{P}(N_1 \notin I),$$

where I is the interval

$$I := [2\epsilon n - \sqrt{\epsilon(1-\epsilon)2n}, 2\epsilon n + \sqrt{\epsilon(1-\epsilon)2n}]. \tag{3.4}$$

The last equality above together with (3.3) implies

$$\mathrm{VAR}[L(N_1)] \geq E[\ \mathrm{VAR}[L(N_1)|L(\cdot), N_1 \in I]\ ]\mathbf{P}(N_1 \in I). \tag{3.5}$$

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

$$\mathrm{VAR}[f(Y)] \geq c^2 \mathrm{VAR}[Y]. \tag{3.6}$$

(See [4] for the proof). Hence, if the map $L(\cdot)$ would have positive slope everywhere larger than $c > 0$, is 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 $k \in [0, n]$. 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 (3.6), for when the map $f(\cdot)$ does not increase every k, but has a tendency to increase on some scale.

Lemma 3.1 *Let $c, m > 0$ be two constants. Let $f : \mathbb{Z} \to \mathbb{Z}$ be a non decreasing map such that:*

- *for all $i < j$:*
$$f(j) - f(i) \leq (j - i) \tag{3.7}$$
- *for all i, j such that $i + m \leq j$:*
$$f(j) - f(i) \geq c \cdot (j - i). \tag{3.8}$$

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

$$\mathrm{VAR}[f(B)] \geq c^2 \left(1 - \frac{2m}{c\sqrt{\mathrm{VAR}[B]}}\right) \mathrm{VAR}[B]. \tag{3.9}$$

Recall the definition of I in (3.4). 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

$$L(j) - L(i) \geq \alpha_3 |i - j|, \tag{3.10}$$

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 if theorem 2.2 holds, then the event E^n_{slope} has high probability.

Lemma 3.2 *Assume that theorem 2.2 holds. Then for a constant $c_4 > 0$,*

$$\mathbf{P}(E^n_{\text{slope}}) \geq 1 - e^{c_4 \cdot n^{0.1}}, \tag{3.11}$$

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 from 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

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

We consider the random variable

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

Let X^j, Y^j be given. Then $\Delta^j = 1$, when $(X^j, Y^j) \notin A^j_n$, otherwise it depends on a random variable U_j taking values on $\{1, \ldots, j\}$ with equal probabilities. The random variable U_j determines X^{j-1}, Y^{j-1}, which together with a random variable U_{j-1} that is independent of U_j and takes values on $\{1, \ldots, j-1\}$ with equal probabilities, determine $\Delta^{j-1} = 1$ and so on. Hence,

$$\sum_{k=i+1}^{j} \Delta^k = g(U_{i+1}, \ldots, U_j),$$

where $U_{i+1}, \ldots, U_j$ are independent random variables, the distribution of U_k is uniform on $\{1, \ldots, k\}$. The function g depends on X^j, Y^j. As noted before, by changing an U_k, the value of $g(U_{i+1}, \ldots, U_j)$ can change at most 1. Hence, by McDiarmid's inequality,

$$\mathbf{P}\big(g(U_{i+1}, \ldots, U_j) - Eg(U_{i+1}, \ldots, U_j) < -c\big) \leq \exp[-\frac{2c^2}{(j-i)}]. \tag{3.12}$$

When $(X^k, Y^k) = (x, y) \in B^k_n$, then theorem 2.2 says that, with $k \leq j$,

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

From the last inequalities, we get

$$\mathbf{P}(\Delta^k = 1 | A^k_n, X^j, Y^j) \geq \alpha_1, \quad \mathbf{P}(\Delta^k = -1 | A^k_n, X^j, Y^j) \leq \alpha_2,$$

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

$$E(\Delta_k|X^j, Y^j) \geq \alpha_1 - \alpha_2.$$

Hence,

$$Eg(U_{i+1}, \ldots, U_j) = E\Big(\sum_{k=i+1}^{j} \Delta^k|X^j, Y^j\Big) \geq (\alpha_1 - \alpha_2)(j-i). \tag{3.13}$$

With $c = (\frac{\alpha_1-\alpha_2}{2})(j-i)$, (3.12) and (3.13) yield

$$\begin{aligned}&\mathbf{P}\Big(g(U_{i+1}, \ldots, U_j) < (\frac{\alpha_1 - \alpha_2}{2})(j-i)\Big) \leq \\ &\mathbf{P}\Big(g(U_{i+1}, \ldots, U_j) - Eg(U_{i+1}, \ldots, U_j) < -(\frac{\alpha_1 - \alpha_2}{2})(j-i)\Big) \leq \exp[-\alpha(j-i)],\end{aligned}$$

where $\alpha = 2(\frac{\alpha_1-\alpha_2}{2})^2$. The right side of the last inequality does not depend on X^j, Y^j, so

$$\mathbf{P}\Big(\sum_{k=i+1}^{j} \Delta^k < \alpha_3(j-i)\Big) \leq \exp[-\alpha(j-i)]. \tag{3.14}$$

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:

$$\sum_{k=i}^{j} \Delta^k \geq \alpha_3|i-j|. \tag{3.15}$$

By (3.14), for n big enough, there exists a constant $c_2 > 0$ such that

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

so

$$\mathbf{P}(E^n_{\Delta \text{ slope}}) \geq 1 - e^{-c_2 \cdot n^{0.1}}, \tag{3.16}$$

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

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

which implies

$$\mathbf{P}(E^{nc}_{\text{slope}}) \leq \mathbf{P}\big((A_n^{\text{all}})^c\big) + \mathbf{P}(E^{nc}_{\Delta \text{ slope}}). \tag{3.17}$$

Note

$$\mathbf{P}\big((A_n^{\text{all}})^c\big) \leq \sum_{k \in I} P(A_n^{kc}) = \sum_{k \in I} \mathbf{P}(A_n^c|N_1 = k) \leq \sum_{k \in I} \frac{\mathbf{P}(A_n^c)}{\mathbf{P}(N_1 = k)}, \tag{3.18}$$

where

$$A_n := \{(X, Y) \in B_n\}. \tag{3.19}$$

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

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

Applying the last inequality to (3.18), yields

$$\mathbf{P}\big((A_n^{\text{all}})^c\big) \leq \sqrt{2}nc_3\mathbf{P}(A_n^c). \tag{3.20}$$

Now the inequalities (3.16), (3.20) and (3.17) yields

$$\mathbf{P}(E_{\text{slope}}^{nc}) \leq \sqrt{2}nc_3\mathbf{P}(A_n^c) + e^{-c_2 \cdot n^{0.1}}. \tag{3.21}$$

By theorem 2.2, we have that $\mathbf{P}(A_n^c) \leq Ce^{-c_1 n}$. Applying this to (3.21) gives

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

which finishes the proof. ■

Conditioning on E_{slope}^n and using the law of total probability with the fact that variance is non negative, inequality (3.5) becomes

$$\text{VAR}[L(N_1)] \geq E[\ \text{VAR}[L(N_1)|L(\cdot), N_1 \in I]\ \mid\ E_{\text{slope}}^n\ ]\mathbf{P}(E_{\text{slope}}^n)\mathbf{P}(N_1 \in I). \tag{3.22}$$

However, when E_{slope}^n holds, then the map

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

satisfies the conditions of lemma 3.1 with $m = n^{0.1}$. Hence, when E_{slope}^n holds, then

$$\text{VAR}[L(N_1)|L(\cdot), N_1 \in I] \geq {\alpha_3}^2\left(1 - \frac{2n^{0.1}}{\alpha_3\sqrt{\text{VAR}[N_1|N_1 \in I]}}\right)\text{VAR}[N_1|N_1 \in I].$$

Plugging the last inequality into (3.22) yields

$$\text{VAR}[L(N_1)] \geq \alpha_3^2\left(1 - \frac{2n^{0.1}}{\alpha_3\sqrt{\text{VAR}[N_1|N_1 \in I]}}\right)\text{VAR}[N_1|N_1 \in I]\mathbf{P}(E_{\text{slope}}^n)\mathbf{P}(N_1 \in I). \tag{3.23}$$

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,

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

where ϕ is the standard normal density. Together with lemma 3.2, this implies that the right side of inequality (3.23) divided by n converges to

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

This finishes the proof.

4 Aligning the ones

Introducing the right notation to define the "alignments of ones" is a key ingredient to the solution of our problem. The best way is to start with simple numerical examples.

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 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

(4.1)

We call the portions between aligned ones *cells*.
The first cell of alignment (4.1) 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 pari 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 (4.1) is

0	1
1	1

The third cell of alignment (4.1) is

0	1
0	1

The fourth cell of alignment (4.1) is

0		1	0	1
0	0		0	1

.

Note that in the last example above, 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. The notation we are going to use is to write the sequence $(v_1, v_2, v_3, v_4) = (0, 1, 0, -1)$ for the alignment 4.1. More precisely, which one's get aligned with each other will be defined by indicating the differences of number of ones for each cell.

Let X and Y be given. As explained above, to every alignment (between X and Y) corresponds a vector $v := (v_1, \ldots, v_k)$ that shows the number of cells in the alignment - k - and the difference of zeros in the cells. And, vice versa, to each vector $v = (v_1, \ldots, v_k) \in \mathbb{Z}^k$ corresponds a (possible empty) family of alignments. The alignments associated with given v have the same pairs of aligned ones. Between consecutive pairs of aligned ones, they align a maximum number of zeros. Hence, all the alignments associated with one v have the same score. In a slight imprecision we will often speak of one alignment for the whole family associated with v. The number v_i indicates the difference in the number of zero's in cell number i. Of course, the alignment associated with v might in some cases not be feasible because it attempt to align ones from the sequences $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ which are outside the strings X and Y. (Hence they might align an X_i with an Y_j for which $i > n$ or $j > n$.)
Let us next define rigourously how $v = (v_1, \ldots, v_k) \in \mathbb{Z}^k$ defines an alignment. Recall that X and Y are fixed.

Definition 4.1 *Let $k \in \mathbb{N}$ and let $v = (v_1, \ldots, v_k) \in \mathbb{Z}^k$. We write $|v|$ for the length of v. Hence, if $v \in \mathbb{R}^k$, then $|v| = k$.*
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. *We have $\pi(i) < s$ and $\nu(i) < t$.*
  2. *$X_{\pi(i+1)} = Y_{\nu(i+1)} = 1$*
  3. *The difference between the 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,*

$$v_{i+1} := \left((s - \pi(i)) - \sum_{j=\pi(i)}^{s} X_{\pi(j)} \right) - \left((t - \nu(i)) - \sum_{j=\nu(i)}^{s} Y_{\nu(j)} \right).$$

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

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

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

Hence, the number of aligned zeros is 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

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

To summarize: every $v \in \mathbb{Z}^k$ defines an alignment of ones. This alignment corresponds to aligning the one $X_{\pi_v(i)}$ with $Y_{\nu_v(i)}$ for each $i = 1, 2, \ldots, k$. Between the aligned ones we assume that we align as many zeros as possible. Hence in cell number i, we align $S_v(i)$ zeros. Searching through all the alignment in $\cup_k \mathbb{Z}^k$ defined in this way, yields the optimal alignment. We have to take care of the zeros after the last cell. We denote by r_v the maximal number of zeros we can align, which come after the last cell. Hence when $v \in \mathbb{Z}^k$ is such that $\pi_v(k), \nu_v(k) \leq n$, we define r_v to be the minimum between the number of zeros in the string $X_{\pi_v(k)} \ldots X_n$ and the number of zeros in the string $Y_{\nu_v(k)} \ldots Y_v(n)$. The score obtained by the alignment π_v, ν_v can be calculated as follows. Each cell gives one aligned pair of ones. Hence, this part contributes $|v|$, the length of the alignment. Then we add for each cell the number of zeros aligned. This gives $\sum_{i=1}^{|v|} S_v(i)$. Finally we need to add the remaining amount of zeros r_v which can be aligned but which come after the last cell. Of course the alignment can only align letters from the text $X = X_1 \ldots X_n$ and $Y = Y_1 \ldots Y_n$. Hence, if $v \in \mathbb{Z}^k$ is an alignment, then $\pi_v(k)$ and $\nu_v(k)$ should not fall outside the interval $[0, n]$, i.e. $\pi_v(k) \leq n$ and $\nu_v(k) \leq n$. Such an $v \in \mathbb{Z}^k$ is called *admissible*. Let V be the set of all admissible alignments, i.e.

$$V := \{v \in \cup_{k>0} \mathbb{Z}^k : \pi(|v|), \nu(|v|) \leq n\}. \tag{4.2}$$

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

Proposition 4.1

$$L_n = \max_{v \in V} \left(|v| + \sum_{i=1}^{|v|} S_v(i) + r_v \right). \tag{4.3}$$

The *score* of an alignment $v \in V$ is defined the following way:

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

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

Let now $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ be independent iid sequences of Bernoulli random variables with parameter ϵ. Let $v \in \bigcup_{k>0} \mathbb{Z}^k$ be fixed and define $|v|$ random cells $C_v(1), \ldots, C_v(|v|)$ as in Definition 4.1. One of the main advantages of defining alignments the way described is that the cells $C_v(1), C_v(2), \ldots, C_v(|v|)$ are independent so that we can use large deviation techniques.

4.1 An useful approach

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

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

So, 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 $X_1, X_2, \ldots$ 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 $\eta_1, \eta_2, \ldots$ model the sequence $Y_1, Y_2, \ldots$. 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

$$T := \min\{i = 1, 2, \ldots : \xi_i \neq \emptyset, \eta_i \neq \emptyset\}. \tag{4.4}$$

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

$$T := \min\{i = 1, 2, \ldots : \xi_i \neq \emptyset, \eta_{u+i} \neq \emptyset\}. \tag{4.5}$$

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 exists a constant $a(\epsilon) > 0$ such that*

$$\mathbf{P}(E) \geq 1 - 16e^{-an}.$$

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 4.1, 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

$$\mathbf{P}(E_1^y) = \mathbf{P}(E_2^y) = 1 - \exp[\alpha\epsilon n - 1].$$

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 $\mathbf{P}\big((E_3^x)^c|E_1^x\cap E_2^x\big)\leq \exp[-2(\epsilon\alpha)^2 m]$ and

$$\mathbf{P}\big((E_3^x)^c\big)\leq\exp[-2(\epsilon\alpha)^2 m]+\mathbf{P}(E_2^{xc})+\mathbf{P}(E_3^{xc})\leq\exp[-2(\epsilon\alpha)^2\epsilon(1-2\alpha)n]+4\exp[-2(\alpha\epsilon)^2 n].$$

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

$$\begin{aligned}\mathbf{P}(E^c(\alpha))&\leq 2\exp[-(\alpha\epsilon)n]+12\exp[-2(\alpha\epsilon)^2 n]+2\exp[-2(\epsilon\alpha)^2\epsilon(1-2\alpha)n]\\&\leq 16\exp[-2(\epsilon\alpha)^2\epsilon(1-2\alpha)n].\end{aligned}$$

Let $\alpha(\epsilon)$ 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

$$\mathbf{P}(E^c)\leq 16\exp[-2(\epsilon\alpha)^2\epsilon(1-2\alpha)n]=16\exp[-an],$$

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

Note that lemma 5.1 gives an lower bound to 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's in X and Y, respectively. So, if $0<\alpha\leq 0.8$ and $E(\alpha)\cap E$ holds, then (5.3) and $L_n\geq(1-\epsilon)+0.9\epsilon^2$ simultaneously hold. Then

$$\frac{N_0}{2}=\frac{N_0^x+N_0^y}{2}\leq\big((1-\epsilon)+(0.8\epsilon^2)\big)n<\big((1-\epsilon)+0.9\epsilon^2\big)n\leq L_n,\tag{5.4}$$

where N_0 is the number of 0's in X and Y. Consider now an optimal alignment $(v_1,\ldots,v_k)$. Then k is the number of aligned 1's, and there is at least $\sum_{i=1}^k|v_i|$ not aligned 0's. Hence, the number of aligned 0's is at most $N_0-\sum_{i=1}^k|v_i|$, and so

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

The last equation together with (5.4) implies

$$\sum_{i=1}^{k} |v_i| < 2k. \tag{5.5}$$

Thus, in this case any optimal alignment must satisfy (5.5).

Let us formalize the foregoing argument. We define $V(k) \subset \mathbb{Z}^k$ as follows

$$V(k) = \{(v_1, v_2, \ldots, v_k) \in \mathbb{Z} \mid |v_1| + \ldots + |v_k| \leq 2k\}. \tag{5.6}$$

We define the set V_n

$$V_n := \bigcup_{k \geq 0.1\epsilon^2 n} V(k). \tag{5.7}$$

Lemma 5.2 *There exists an event E_4 such that*

$$\mathbf{P}(E_4) \geq 1 - 16\exp[-an] - \exp[-(0.8\epsilon)^2 \epsilon n]$$

and on E_4

$$L_n = \max_{v \in V_n \cap V} \Big(|v| + \sum_{i=1}^{|v|} L_v(i) + r_v \Big).$$

Proof. Take $E_4 := E_2(0.8\epsilon) \cap E$. By (5.5), every optimal v such that $|v| = k$ belongs to $V(k)$. So, all optimal alignments belong to $\cup_k V(k)$. From (5.4)

$$L_n - \frac{N_0}{2} \geq 0.1\epsilon^2 n,$$

implying that the optimal alignment must have at least $0.1\epsilon^2 n$ $1's$. So, k must be greater that $0.1\epsilon^2 n$. So, all optimal alignments belong to $\cup_{k \geq 0.1\epsilon^2} V(k) = V_n$. Proposition 4.1 now finishes the proof. ■

Lemma 5.3

$$|V(k)| < 2^k C_k^{3k} < 16^k, \tag{5.8}$$

Proof. Let

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

where $\mathbb{Z}^+ = \{0, 1, \ldots\}$. Thus, $|V^+(k)|$ is number of k-dimensional vectors with components being non-negative integers and summing up at most $2k$. By adding one more component, we get that $|V^+(k)|$ is number of $k+1$-dimensional vectors with components being non-negative integers and summing up exactly $2k$. The number of such vectors is $C_{k+1-1}^{2k+k+1-1} = C_k^{3k}$. So,

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

For every k-dimensional vector, there are at most 2^k ways to assign the signs. So

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

■

6 The effect of changing a one into a zero

6.1 The events B_n and A_n

In this subsection we explain why theorem 2.2 holds. We want to show, that typically, when changing a randomly picked one into a zero, the score L_n tends to increase.

Example. Take the two texts $X = 01000001$ and $Y = 0010101$. An optimal alignment is given by

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

The first cell in this alignment is

	0	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_5. This is a "non-aligned" one contained in the Y-part of cell number two. By switching Y_5 into a zero the LCS increases by one. The reason is that in cell number two, we can now align three zeros instead of only two:

0	0	0	0	0	1
0	0	0			1

In this example, the score gets increased because Y_5 is on the side of the cell with strictly less zeros. We say that Y_5 is *on the side of a cell with less zeros*. Hence adding on zero on that side increases the score by one. Let us imagine next that instead of Y_5 the one chosen would be X_2. This one is "used" in the alignment and hence switching it could result in decreasing the optimal score L_n by one. (This is not necessary though, since there could exist another alignment where the effect of switching X_2 would not be detrimental to the score.) We call the ones which are "used" in the alignment, ones that are *matched by the alignment*". In our example, the matched ones are: X_2 is matched with Y_3 and X_8 is matched with Y_7, Y_5 is not matched.

From our example, it should be clear what we need to do: To show that the score has a tendency to increase when we switch a randomly picked one, we need to prove that typically, there are many more ones which will increase the score than ones which will decrease the score if switched. In other words, we need to show that in an optimal alignment, there typically are much less aligned ones as the 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 $v = (v_1, \ldots, v_k) \in \mathbb{Z}^k$. For $i \in [0, k]$, we define

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

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

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

Recall now the set B_n from theorem 2.2. The set B_n contains the outcomes of X and Y such that there exists an optimal alignment v having the proportion of matched ones below α_2, but more than $\alpha_1\%$ of ones on cell-sides with less zeros. More formally: $(x, y) \in B_n$ if there exists an $v \in V(x, y)$ (v is admissible) such that

1. The alignment v is optimal: $L_v = L_n$;
2. The proportion of aligned ones is below α_2: $|v| \leq \alpha_2 N_1$, where N_1 is the total number of ones in x and y;
3. The proportions of ones on sides with less ones is above α_1: $N_v^- \geq \alpha_1 N_1$.

From what we explained it follows directly that when $(x, y) \in B_n$ then

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

i.e. (2.1) and (2.2) hold. Hence, what is left to prove is that the event $\{(X, Y) \in B_n\}$ has big probability. In other words, theorem 2.2 is proven if we show that

$$\mathbf{P}(A_n) \geq 1 - \exp[-c_1 n], \quad \text{where } c_1 > 0, \quad A_n := \{(X, Y) \in B_n\}. \tag{6.1}$$

6.2 Breaking cells

The rest of the paper is devoted to proving (6.1). The main problem is that we could easily not have enough ones on the sides with less zeros. Let us 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 what happens to cell number two:

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

The old cell number two is split into two cells: The new cell number two and the new cell number 3. The old cell number three does not change but i s 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. Hence the breaking up process can help up get ride of the problem of having to may cells with the same number of zeros.

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

Definition 6.1 *Let $k \in \mathbb{N}$, $v \in \mathbb{Z}^k$, $i \leq k$ 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*

$$X_{\pi(i)+1} X_{\pi(i)+2} \dots X_{j-1}$$

and

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

is one or minus one. Hence

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

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 (5.7). Let $V_{1\%}$ be the subset of V_n consisting of the alignments of V 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

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

6.3 The events

Recall that for a vector v we associate $|v|$ random cells $C_v(1), \ldots, C_v(|v|)$ defined as a function of random i.i.d, Bernoulli random sequences $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$. In the following we define some events that capture the typical behavior of these random cells.

Definition 6.2

- *Let D be the event that for all $v \in V^c_{1\%}$, we have that 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
$$N_v^- \geq 2\alpha_1 \Big(\sum_{j=1}^{\pi_v(|v|)} X_j + \sum_{j=1}^{\nu_v(|v|)} Y_j \Big).$$

- *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
$$|v| \leq \alpha_2 \Big(\sum_{j=1}^{\pi_v(|v|)} X_j + \sum_{j=1}^{\nu_v(|v|)} Y_j \Big).$$

- *Let K be the event that for every optimal alignment v we have that 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
$$R_v := \sum_{j=\pi(|v|)+1}^{n} X_j + \sum_{j=\nu(|v|)+1}^{n} Y_j \leq \alpha_1 \left(\sum_{j=1}^{n} X_j + \sum_{j=1}^{n} Y_j \right).$$

Recall that v is optimal if $L_n = L_v$ and v is admissible, i.e. $v \in V$ or, equivalently, $\pi_v(|v|), \nu_v(|v|) \leq n$.

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 $V_{1\%}^c$, there exists another alignment $v' \in V_{1\%}$ *with the same score.* So, when the event D holds, we can restrict the search space to $V_{1\%}$ instead of using the whole set V_n.

The main combinatorial lemma of this paper is

Lemma 6.1

$$E_4 \cap D \cap F \cap G \cap K \subset A_n. \tag{6.2}$$

Proof. When E_4 holds, then by lemma 5.2 any optimal alignment belongs to the set V_n, i.e. $V^* \subset V \cap V_n$. Let $v \in V^*$ be optimal. Suppose that $v \in V \cap V_{1\%}^c$. If D_v holds, then at least 1% of the cells can be broken up. That means, there exists another $v' \in V_{1\%}$ with the same score, i.e. $L_v = L_{v'}$. Hence, we can assume $v \in V \cap V_{1\%}$. If $F_v \cap K_v$ holds, then N_1, the number of ones in X and Y satisfies

$$N_1 = R_v + \sum_{j=1}^{\pi_v(|v|)} X_j + \sum_{j=1}^{\nu_v(|v|)} Y_j \leq \alpha_1 N_1 + \sum_{j=1}^{\pi_v(|v|)} X_j + \sum_{j=1}^{\nu_v(|v|)} Y_j \leq \alpha_1 N_1 + \frac{N_v^-}{2\alpha_1}$$

implying that

$$N_v^- \geq (1-\alpha_1) 2\alpha_1 N_1 \geq \alpha_1 N_1,$$

because $\alpha_1 \leq \frac{2}{3}$. Then $\mathbf{P}(\tilde{L} - L = 1 | X = x, Y = y) \geq \alpha_1$ i.e. (2.2) holds. If G_v holds, then $\mathbf{P}(\tilde{L} - L = -1 | X = x, Y = y) \leq \alpha_2$ i.e. (2.1) holds. ■

7 Bounding the probabilities

From (6.2) it follows that

$$\mathbf{P}(A_n^c) \leq \mathbf{P}(E_4^c) + \mathbf{P}(D^c) + \mathbf{P}(F^c) + \mathbf{P}(G^c) + \mathbf{P}(K^c) \tag{7.1}$$

By lemma 5.2, $\mathbf{P}(E_4^c)$ is exponentially small in n. So, it only remains to prove that $\mathbf{P}(D^c)$, $\mathbf{P}(F^c)$, $\mathbf{P}(G^c)$ and $\mathbf{P}(K^c)$ are exponentially small in n, provided ϵ is sufficiently small. In the following we show the existence of finite constants C_D, C_F, C_G, C_K as well as positive constants c_D, c_F, c_G, c_K that all depend on ϵ such that

$$\mathbf{P}(D^c) \leq C_D \exp[-c_D n], \quad \mathbf{P}(F^c) \leq C_F \exp[-c_F n]$$
$$\mathbf{P}(G^c) \leq C_G \exp[-c_G n], \quad \mathbf{P}(K^c) \leq C_K \exp[-c_K n].$$

(Lemmas 7.3, 7.8, 7.9,7.10, respectively.) Moreover, the bound $\mathbf{P}(G^c) \leq C_G \exp[-c_G n]$ holds only for ϵ being sufficiently small. With lemma 5.2, this finishes the proof of theorem 2.2.

7.1 Combinatorics

Let

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

Lemma 7.1

$$|V^c_{1\%}(k)| \leq \exp[(2.01 \cdot 0.0315 + 0.7 \cdot 0.01)k] = \exp[(0.063315 + 0.007)k] = \exp[0.070315k], \tag{7.2}$$

where

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

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

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

We know that

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

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

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

where

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

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 $(\ln 2)H(0.005) = H_e(0.005) \leq 0.0315$ and $\ln 2 < 0.7$, we get (7.2). ■

7.2 The event D

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

Lemma 7.2 *Let $v \in V^c_{1\%}(k)$. Then*

$$\mathbf{P}(D^c_v) \le \exp[-0.089k]. \tag{7.3}$$

Proof. Let us calculate the probability that a 0-cell is breakable. For this, we use the approach introduced in subsection 4.1. Recall the definition of T in (4.4). 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,

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

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 $[0, \frac{1}{2}]$, so

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

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

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

■

Lemma 7.3 *There exists $C_D < \infty$ such that*

$$\mathbf{P}(D^c) \leq C_D \exp[-0.018685(0.1\epsilon^2)n]. \tag{7.4}$$

Proof.

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

With (7.2) and (7.3), 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 we consider $k \geq (0.1\epsilon^2)n$,

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

where

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

■

7.3 The event F

In the following, we use the following large deviation result proven in Appendix.

Lemma 7.4 (A large deviation for geometric random variables) *Let $G_1, \dots, G_m$ be iid random variables with geometric distribution $G(p)$. There exists $0 < \alpha_0 < 1$ such that for every $\alpha \le \alpha_0$, it holds*

$$\mathbf{P}\Big(\sum_{i=1}^{m} G_i \le \frac{\alpha}{p} m\Big) \le \exp[-300m] \quad \forall m \tag{7.5}$$

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

$$\mathbf{P}\Big(\sum_{i=1}^{m} G_i > \frac{A}{p} m\Big) \le \exp[-Cm] \quad \forall m. \tag{7.6}$$

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 4.1 and recall the random variable T as in (4.5), 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

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

Let

$$\rho^- := \min\{j = 1, 2, \dots : \eta_{u+T_x(j)} \neq \emptyset\}. \tag{7.7}$$

So, 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 (below) the number of 1's in X side, N_1^-. This number is clearly at least ρ^-, so $N_1^- \ge \rho^-$, 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 (7.7) 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 in the sides of fewer 0's of non-0 cells.

Lemma 7.5 *There exists a $\gamma > 0$ such that for every $v = (v_1, \dots, v_k) \in V_{1\%}$ it holds*

$$\mathbf{P}(F_{1v}^c) \le \exp[-3k], \quad \textit{where} \quad F_{1v} = \{N_v^- \ge \frac{\gamma}{\epsilon} k\}. \tag{7.8}$$

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 (7.7). So,

$$N_v^- \geq \sum_{i \in I} \rho_i^-. \tag{7.9}$$

Let α_o be as in Lemma 7.4. Take

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

and apply Lemma 7.4:

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

■

Let

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

By (5.8) and (7.8),

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

Hence

$$\mathbf{P}(F_1^c) \leq \sum_{k \geq (0.1\epsilon^2)n} \mathbf{P}(F_1(k)^c) \leq \sum_{k \geq (0.1\epsilon^2)n} \exp[-0.2k] = C_{1,F} \exp[-0.2(0.1\epsilon^2)n], \tag{7.10}$$

where

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

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

Lemma 7.6 *There exist constant B such that for every $v = (v_1, \dots, v_k) \in V_{1\%}$ it holds*

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

Proof. Let B be such that $B - 1 > A_0(\ln 16 + 1)$. By (7.6),

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

■

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, by analogue of (7.10),

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

where

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

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

Lemma 7.7 *There exists constant $A < \infty$ such that that for every $v = (v_1, \dots, v_k) \in V_{1\%}$ it holds*

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

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

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

So, the number of 1' s in X side of the cell $C_v(j)$ is

$$\sum_{i=1}^{\rho(j)} G_i, \tag{7.11}$$

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

$$\sum_{j=1}^{\pi(k)} X_j = \sum_{i=1}^{\rho_v} G_i. \tag{7.12}$$

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

$$\frac{A}{B} > A_o\left(\frac{(\ln 16 + 1)\epsilon}{B}\right)$$

and define

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

From lemma 7.4 with $m = \frac{B}{\epsilon}k$

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

Due to (7.12), for every v,

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

Lemma 7.6 finishes the proof. ■

Let

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

Then, by analogue of (7.10),

$$\mathbf{P}(F_4^c(k)) \le \sum_{v\in V(k)} \mathbf{P}\Big(\sum_{j=1}^{k} \rho_j \ge \frac{B}{\epsilon}k\Big) \le 2\exp\big[-k\big((\ln 16+1)-\ln 16\big)\big] = 2\exp[-k] \tag{7.13}$$

$$\mathbf{P}(F_4^c) \le 2C_{2F}\exp[-0.1(\epsilon^2)n]. \tag{7.14}$$

Lemma 7.8 *There exists $\alpha_1 > 0$ such that for a constant $C_F < \infty$*

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

Proof. It holds

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

So,

$$F_1 \cap F_4 \subset \Big(\bigcap_{v\in V_{1\%}} F_{1,v}\Big) \cap \Big(\bigcap_{v\in V_{1\%}} F_{4,v}\Big) = \bigcap_{v\in V_{1\%}} \Big(F_{1,v}\cap F_{4,v}\Big)$$

$$\subset \bigcap_{v\in V_{1\%}} \Big\{N_v^- \ge \frac{(1-\epsilon)\gamma}{A}\sum_{j=1}^{\pi(k)} X_j\Big\} =: F_x$$

and by (7.10) and (7.14)

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

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

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

Thus

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

where

$$\alpha_1 := \frac{\gamma}{8A} \le \frac{(1-\epsilon)\gamma}{4A}, \tag{7.15}$$

provided $\epsilon \le 0.5$ and

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

■

7.4 The event G

We use the notations introduced in the previous subsection. Let α_1 be as in (7.15). Fix $\alpha_2 < \alpha_1$.

Lemma 7.9 *There exists an constant $C_G < \infty$ and $\epsilon_o(\alpha_2) > 0$ such that for every $\epsilon \le \epsilon_o$*

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

Proof. Let $v \in V(k)$. From (7.12)

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

Let

$$G_v := \Big\{ k \le 2\alpha_2 \sum_{j=1}^{\pi(k)} X_j \Big\} = \Big\{ \frac{k}{2\alpha_2} \le \sum_{i=1}^{\rho_v} G_i \Big\}.$$

Then

$$\mathbf{P}(G_v^c) \le \mathbf{P}\Big(\sum_{i=1}^{k} \rho_j^- < \frac{k}{2\alpha_2} \Big) = \mathbf{P}\Big(\sum_{i=1}^{k} \rho_j^- < \frac{\epsilon}{2\alpha_2} \frac{1}{\epsilon} k \Big).$$

Let α_o be as in Lemma 7.4. Let $\epsilon_o < 2\alpha_2$ be such that

$$\frac{\epsilon_o}{2\alpha_2} = \alpha_o.$$

Then, by Lemma 7.4, for every $\epsilon \le \epsilon_o$,

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

Let

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

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

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

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

$$G_y := \bigcap_{v \in V_{1\%}} \Big\{ |v| \le 2\alpha_2 \sum_{j=1}^{\pi(k)} Y_j \Big\}.$$

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

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

provided $\epsilon \le \epsilon_o$. ■

7.5 The event K

Lemma 7.10 *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:

$$X_{\pi(|v|)+1}, X_{\pi(|v|)+2}, \ldots, X_n \quad \text{and} \quad Y_{\nu(|v|)+1}, Y_{\nu(|v|)+2}, \ldots, Y_n. \tag{7.16}$$

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 (4.4). Hence, conditioning on v, we note that the random number of ones in the (7.16),

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

is bounded by the number on ones in a 0-cell, 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 holds

$$\mathbf{P}(R > r) \leq \mathbf{P}(\zeta > r) \quad \forall r \in \mathbb{N}. \tag{7.17}$$

Let

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

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 (7.17),

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

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

$$\mathbf{P}(E_4^c) \leq 16\exp[-an] - \exp[-(0.8\epsilon)^2\epsilon n].$$

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 (7.11),

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

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 7.4 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 7.4,

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

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

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

■

8 Appendix

Proof of lemma 7.4 Let us recall a large deviation result for Bernoulli random variables. Let $X_i \sim B(1,p)$ be iid. Then

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

If $p > \alpha$, then (7.4) trivially holds. If $p = \alpha$, then the probability in (7.4) equals $p^m = \exp[(\ln\alpha)m] = \exp[-\ln\frac{1}{\alpha}m]$. Hence, we consider the case $p < \alpha < 1$, only. It holds,

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

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

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

So 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

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

Now, 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 $\ln x \le x-1$) and so

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

So, for every $\alpha < 1$, the function $p \mapsto L(\alpha,p)$, $p \in [0,\alpha]$ is increasing with maximum being $L(\alpha,\alpha) = \ln\frac{1}{\alpha}$ and mimimum is $L(\alpha,0) = \ln\frac{1}{\alpha} + \alpha - 1$. Take $\alpha_0 := \exp[-301]$.

Then, if $\alpha \le \alpha_0$, it holds that $L(\alpha, p) \ge L(\alpha, 0) = \ln \frac{1}{\alpha} + \alpha - 1 > \ln \frac{1}{\alpha} - 1 \ge \ln \frac{1}{\alpha_0} - 1 = 300$.

By large deviation, for $A > 1$

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

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 $G_1 - \frac{A}{p}$,

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

Hence

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

Since

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

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

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

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

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] S.N. Majumdar and S. Nechaev. Exact asymptotic results for a model of sequence alignment. *preprint*, 2004.

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

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

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

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