On the Probabilistic Longest Common Subsequence Problem for Sequences of Independent Blocks

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Mathematik
der Universität Bielefeld

vorgelegt von

Felipe Torres

März 2009

Gedruckt auf alterungsbeständigem Papier nach DIN ISO-9706

Contents

ACKNOWLEDGEMENTS

This thesis not only respresents the end of hard graduate studies but also the testimony of a challenging adventure in a foregin country full of unexpected situations, sad and happy moments, multiple chances for knowing new people, new fascinating places and new ways of doing Mathematics. I want to dedicate this work and the whole experience to my family, fundamental and unconditional support, I am proud to be motivated by them.

The list of people who somehow had contributed to this thesis is large to show it up here in detail, though I would like to begin sincerely thanking my advisors Prof. Friederich Goetze and Ph.D Heinrich Matzinger who always trusted me and belived in me and my work.

The International Graduate College (IGK) and all the people involved on it from Germany and from China took me in, made me feel like at home and gave me all the advantages of a cooperation between Germany and China, turning my Ph.D. time in Bielefeld into a rich experience which inherited me a legacy I will try to keep alive, not only in academic terms. My complete gratitude to this program, to the whole IGK-team and to the whole Chinese-side team (CAS and GUCAS).

I would also like to thank the economic support from the Collaborative Research Centre 701: Spectral Structures and Topological Methods in Mathematics, especially for financing me a trip to Atlanta to finish this thesis and in general for all the additional support in other academic activities.

Finally, but not less important, thanks to all secretaries, my dear friends, my friends and the friends of my friends who helped me since the very beginning (for example, to look for a place to live in) until the very end. The day after day was easier with you around.

Chapter 1

Introduction

Let X and Y be two binary random strings of length n independent of each other. Let L_n denote the length of the Longest Common Subsequence (LCS) of X and Y. In general the order of magnitude in n of $\mathrm{VAR}[L_n]$ is not known. So far, Matzinger and his collaborators had been able to prove that $\mathrm{VAR}[L_n]$ has order $\Theta(n)$ in few cases [15, 16, 18, 19, 20]. In this thesis, we prove the same result for a model which is not low entropy. Previous cases were all low entropy models. The model for the distribution is an i.i.d. sequences of blocks, where blocks are words consisting only of one symbol. In the present case all the blocks have length $l-1$, l or $l+1$ with probability $1/3$ for a given integer parameter $l > 5$. We reduce the problem of proving that $\mathrm{VAR}[L_n] = \Theta(n)$ to showing that a function under an entropy constraint does not go below zero. The method which we develop could be used for many other more complex cases whenever one pattern tends to influence the LCS score in a biased way. Also, a natural question is what happens if one has a more realistic situation than i.i.d, like Markov chains of words for instance (DNA models). This thesis partially answers this question since the model considered is one particular example of Markov chain of words. More general cases with more possible words are still open for future research though the techniques shown here give us new tools for approaching them.

1.1 The LCS history

Let X and Y be two finite strings over a finite alphabet Σ. A common subsequence of X and Y is a subsequence which is a subsequence of X as well as of Y. A Longest Common Subsequence (LCS) of X and Y is a common subsequence of X and Y of maximal length.

Example 1.1.1 *Let us consider the DNA-alphabet $\Sigma = \{A, G, C, T\}$. In Bioinformatics, an usual problem is to decide if two sequences are related, which means that they evolved from a common ancestor. If they are related, they should look similar. Biologist try to determine which parts are related by finding an alignment* *which aligns the related parts. Let us consider two sequences* $x = ACGTAGCA$ *and* $y = ACCGTATA$. *If we compare them letter by letter the great similarity does not become obvious:*

x		A	C	G	T	A	G	T	A
y		A	C	C	G	T	A	C	A

(1.1.1)

The reason is that some letters "got lost" so that they are present only in one of the two sequences. When we align without leaving any gaps for those lost letters, we mostly align non-corresponding letter pairs. It is better to use gaps and to allow aligning a letter with a gap. Then the letters which are present only in one of the two sequences get aligned with gaps. A good alignment is provided by:

x		A	C		G	T	A	G		T	A
y		A	C	C	G	T	A		C		A

(1.1.2)

We now see a much better coincidence between the two sequences. We displayed in 1.1.1 and 1.1.2 two possible alignments between x *and* y, *1.1.1 without gaps and 1.1.2 with gaps. From here on, and throughout this thesis, we only allow alignments which align same letter pairs and letters with gaps. This kind of alignments are useful when the evolution process only looses letters (there is no mutations). So, when we speak about an alignment, we automatically assume that it only aligns same-letter-pairs or letters with gaps. Each such an alignment defines a common subsequence. The number of aligned letters will be sometimes refered to as* alignment-score. *An alignment aligning a maximum number of letter pairs is called an* optimal alignment. *The Common Subsequence defined by an optimal alignment is hence a* LCS. *For example, the alignment 1.1.2 defines the common subsequence* $z = ACGTAA$, *which consists of the pair of matched letters:*

x		A	C		G	T	A	G		T	A
y		A	C	C	G	T	A		C		A
z		A	C		G	T	A				A

(1.1.3)

In the alignment 1.1.3, the sequence $z = ACGTAA$ *is a common subsequence of* X *and* Y *with maximal length, therefore a* LCS *of* x *and* y.

Longest Common Subsequences are used in computational biology and linguistics (among other areas) to recognize when strings are similar. A relatively long LCS indicates that the strings are related. How long does the LCS need to be to imply relatedness? In the direction of an answer, one could try to understand the typical behavior of the length of the LCS. But even the asymptotic behavior of the LCS of two independent random strings of length n is one of the major open problems in probability theory.

Assume now that $X = X_1X_2 \dots X_n$ and $Y = Y_1Y_2 \dots Y_n$ are two i.i.d. strings independent of each other over the same finite alphabet Σ. Let L_n denote the length of the LCS of X and Y. Using a sub-additivity argument, Chvàtal-Sankoff [6] prove that the limit

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

exists. The constant γ depends on the distribution of X_1 and Y_1. However until the date the exact value of γ is not known in even such simple cases as when one has two equally likely symbols. Neither it is known in general if $\mathrm{VAR}[L_n]$ is of linear order in n. There exists conflicting conjectures on that topic: Waterman [10] thinks that $\mathrm{VAR}[L_n] = \Theta(n)$ and Chvàtal-Sankoff thought they were observing in their simulation that $\mathrm{VAR}[L_n] = \Theta(n^{2/3})$. The order conjectured by Chvàtal-Sankoff would be similar to the situation in the Longest Increasing Subsequence (LIS) of a random permutation problem where the fluctuation is of order third root of the expectation.

Chvàtal-Sankoff [6] derived upper and lower bounds for γ, and similar upper bounds were found by Baeza-Yates, Gavalda, Navarro and Scheihing [7] using an entropy argument. These bounds have been improved by Deken [23], and subsequently by Dancik-Paterson [24, 25]. For sequence with many equiprobable letters (i.e. when Σ is large) Kiwi, Loebl and Matousek [8] determined the asymptotic value of γ. Arratia-Waterman [11] derived a law of large deviation for L_n for fluctuations on scales larger than $\sqrt{n}$. In [9], Steele was able to prove that there exists a constant $c > 0$ not depending on n such that $\mathrm{VAR}[L_n] \leq c\,n$. The LCS-problem can be formulated as a last passage percolation problem with correlated weights, moreover Alexander [12] proved that $\mathrm{E}[L_n]/n$ converges at a rate of order $\sqrt{\log n/n}$ by using first passage percolation methods.

One of the only cases where for first/last passage percolation models the asymptotic order of the fluctuation is known is for Longest Increasing Subsequence (LIS) of a random permutation of natural numbers (see Baik, Deift and Johansson [26] and also Aldous and Diaconis [27]). The LIS problem is asymptotically equivalent to a special last passage percolation process on a Poisson graph. Furthermore, the LIS problem can be formulated as a special LCS problem: the LIS is the LCS of two sequences where one is a sequence of randomly permuted numbers and the other is the sequence of increasing integers.

In [26] Baik, Deift and Johansson denoted l_n as the length of the LIS of a random permutation drawn from the symmetric group $\mathcal{S}_n$ with the uniform distribution. They proved that the centered and scaled expression

$$\frac{l_n - 2\sqrt{n}}{n^{1/6}}$$

converges in distribution as $n \to \infty$ to the so called Tracy-Widom distribution. The Tracy-Widom distribution was first obtained by Tracy and Widom [26] in the framework of Random Matrix Theory where it gives the limit distribution for the (centered and scaled) largest eigenvalues in the Gaussian Unitary Ensemble of Hermitian matrices. The problem of the asymptotics of l_n was first raised by Ulam [28]. Substantial contributions to the solution of the problem have been made by Hammersley [30], Logan and Shepp [29], Vershik and Kerov (Vershik/Kerov 1977 Soviet math dokl).

After the break through of Baik, Deift and Johansson [26] on the LIS problem it was natural to try to use these techniques to tackle the LCS-problem, but for most situations it has not yet worked to adapt those methods to solving the LCS-problem.

However, Matzinger and his collaborators have been able to develop and to publish a set of new techniques [15, 16, 18, 19, 20] which allow to determine the order of the fluctuation for the LCS in several special cases of X and Y. In all this cases the order of the fluctuation of $\mathrm{VAR}[L_n]$ is $\Theta(n)$. For giving a short description of those cases above, let us call a random binary sequence $X = X_1X_2\cdots X_n$ to be a *Bernoulli sequence* if $\mathrm{P}(X_i = 0) = \mathrm{P}(X_i = 1)$ for every $i = 1, \ldots, n$. We have that:

- in [15] the sequence X is a Bernoulli sequence and Y is a non-random periodic sequence.
- in [16] the sequence X is a Bernoulli sequence and Y is an i.i.d. random sequence over a $3-$symbols alphabet.
- in [18] both sequences X and Y are Bernoulli sequences but they are aligned by using a score function which gives more weight when aligning ones than aligning zeros.
- in [19] the sequences are i.i.d. and one symbol has much smaller probability than the other. That is a case where the considered sequences have low entropy.

Nevertheless, the most basic situation of i.i.d. sequences with equiprobable symbols remains open. The techniques go together with a deep understanding of the path structure of the optimal alignments.

The LCS-problem has received a lot of attention also because LCS and the related optimal alignments are some of the main tools in computational biology and string treatment (see for example [2], [1, 5, 10]).

Chapter 2

Main Ideas

2.1 Aim and definitions

Let $l \in \mathbb{N}$ be a parameter. Let $B_{X1}, B_{X2}, \ldots$ and $B_{Y1}, B_{Y2}, \ldots$ be two i.i.d. sequences independent of each other such that:

$$\mathrm{P}(B_{Xi} = l-1) = \mathrm{P}(B_{Xi} = l) = \mathrm{P}(B_{Xi} = l+1) = 1/3$$

and

$$\mathrm{P}(B_{Yi} = l-1) = \mathrm{P}(B_{Yi} = l) = \mathrm{P}(B_{Yi} = l+1) = 1/3.$$

We call the runs of 0's and 1's blocks. Let

$$X^\infty = X_1 X_2 X_3 \ldots$$

be the binary sequence so that the i-th block has length B_{Xi}. Similarly let

$$Y^\infty = Y_1 Y_2 Y_3 \ldots$$

be the binary sequence so that the i-th block has length B_{Yi}.

Example 2.1.1 *Assume that $X_1 = 1$ and $B_{X1} = 2$, $B_{X2} = 3$ and $B_{X3} = 1$. Then we have that the sequence X^∞ starts as follows $X^\infty = 1100010\cdots$ meaning that in X^∞ the first block consists of two 1's, the second block consists of three 0's, the third block consists of one 1's, etc.*

Let X denote the sequence obtained by only taking the first n bits of X^∞:

$$X = X_1 X_2 X_3 \ldots X_n$$

and similarly

$$Y = Y_1 Y_2 Y_3 \ldots Y_n.$$

Let L_n denote the length of the LCS of X and Y, $L_n := |\mathrm{LCS}(X, Y)|$. **The main result of this paper states that for l large enough, the order of the fluctuation of L_n is n:**

Theorem 2.1.1 *There exists l_0 so that for all $l \geq l_0$ we have that*

$$\mathrm{VAR}[L_n] = \Theta(n).$$

We show that the above theorem is equivalent to proving that "a certain random modification has a biased effect on L_n". This is a technique with similar approches in other papers (for instance see [16], [19]) . So the main difficulty is actually proving that the random modification has typically a biased effect on the LCS.

We choose at random in X a block of length $l-1$ and at random one block of length $l+1$. This means that all the blocks in X of length $l-1$ have the same probability to be chosen and then we pick one of those blocks of length $l-1$ up and also that all the blocks in X of length $l+1$ have the same probability to be chosen and we pick one of those blocks of length $l+1$ up. Then we change the length of both these blocks to l. The resulting new sequence is denoted by $\tilde{X}$. Let $\tilde{L}_n$ denote the length of the LCS after our modification of X. Hence:

$$\tilde{L}_n := |\mathrm{LCS}(\tilde{X}, Y)|.$$

If we can prove that our block length changing operation has typically a biased effect on the LCS than the order of the fluctuation of L_n is $\sqrt{n}$. This is the content of the next theorem:

Theorem 2.1.2 *Assume that there exists $\epsilon > 0$ and $\alpha > 0$ not depending on n such that for all n large enough we have:*

$$\mathrm{P}\Big(\;\mathrm{E}[\tilde{L}_n - L_n | X, Y] \geq \epsilon\; \Big) \geq 1 - \exp(-n^{\alpha}). \tag{2.1.1}$$

Then,

$$\mathrm{VAR}[L_n] = \Theta(n).$$

The above theorem reduces the order of fluctuation problem to proving that our random modification has typically a higher probability to lead to an increase than to a decrease in score. Note that our random modification can change the score by at most one unit. Hence, we always have:

$$|\tilde{L}_n - L_n| \leq 1. \tag{2.1.2}$$

In theorem 2.1.3 we reduce proving 2.1.1 to showing that a minimizing problem has a positive solution. The minimizing problem is on a 9 dimensional space but in chapter 6, by using Lagrange multiplyers, we are able to further reduce it to a parametrized 3 dimensional problem. We numerically and graphically verify that the positive minimum condition in chapter 6 is already verified for $l > 5$. This implies that $\mathrm{VAR}[L_n] = \Theta(n)$ already for $l = 6$.

Let us next look when the random modification introduces an increase or a decrease.

Example 2.1.2 *Let us look at a situation where $l = 3$. Let us take two sequences $x = 00110011110000111$ and $y = 0011100001100001111$. An optimal alignment (in the sen of the Example 1.1.1) would be:*

x		0	0	1	1		0	0			1	1	1	1	0	0	0	0	1	1	1	
y		0	0	1	1	1	0	0	0	0	1	1			0	0	0	0	1	1	1	1
LCS		0	0	1	1		0	0			1	1			0	0	0	0	1	1	1	

(2.1.3)

In this example no block gets left out completely. By this we mean that no block is only aligned with gaps. The first block of x is aligned with the first block of y. The second block of x is aligned with the second block of y. By this we mean that all the bits from the second block of x are either aligned with bits of the second block of y or with gaps and vice versa. We have that the second block of the LCS *is hence obtained from the second blocks of x and y by taking the minimum of their respective lengths. In our current special example, we have that for all $i = 1, 2, \ldots, 6$, the i-th block of X gets aligned with the i-th block of y. We could represent this idea visually by viewing the alignment as an alignment of blocks in the following manner:*

x		00	11	00	1111	0000	111
y		00	111	0000	11	0000	1111
LCS		00	11	00	11	0000	111

(2.1.4)

Let us next analyze what is the expected change when we perform our random modification. In x there are exactly 3 blocks of length $l - 1 = 2$. These are the first three blocks of x. The first block of x of length 2 is aligned with a block of y of length 2, the second one with a block of length 3 and the fourth with a block of length 4. Hence, when we increase the length of the first block of length 2 of x by one the score does not increase. When we increase the second or third, however, the score increases by one unit. Each of these blocks has the same probability $1/3$ to get drawn. Hence, the conditional expected increase due to the enlargement of a randomly chosen block of length 2 in this case, is equal to $2/3$. In our random modification we also choose a block of length $l + 1$ and decrease it to length l. In our example, there are two blocks in x of length $l + 1 = 4$. These blocks are the fourth and fifth block of x. The fourth block is aligned with a block of length 2 whilst the fifth is aligned with a block of length 4. Hence, when we decrease the length of the fourth block we get no change in score whilst when we decrease the fifth we get a decrease by one unit. Each of the two blocks have same probability to get drawn. This implies that the expected change due to decreasing a randomly chosen block of length 4 is equal to $-1/2$. Adding the two changes, we find that *for x and y defined as in the current example, the conditional expected change is equal to:*

$$E[\,\tilde{L}_n - L_n\,|X = x, Y = y] = \frac{2}{3} - \frac{1}{2} = \frac{1}{6} \tag{2.1.5}$$

In our example we have six aligned block pairs leading to the following set of pairs of lengths:

$$\{(2,2);(2,3);(2,4);(4,2);(4,4);(3,4)\}.$$

Let p_{ij} designate the proportion of aligned block pairs which have the x-block having length i and the y-block having length j.

Example 2.1.3 *For our example above we have:*

$$\begin{pmatrix} p_{22} & p_{23} & p_{24} \\ p_{32} & p_{33} & p_{34} \\ p_{42} & p_{43} & p_{44} \end{pmatrix} = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & \frac{1}{6} \\ \frac{1}{6} & 0 & \frac{1}{6} \end{pmatrix} \tag{2.1.6}$$

With this notation, equality 2.1.5 can be written as:

$$E[\,\tilde{L}_n - L_n\,|X = x, Y = y] \geq \frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}} - \frac{p_{l+1,l+1}}{p_{l+1,l-1} + p_{l+1,l} + p_{l+1,l+1}} \tag{2.1.7}$$

The inequality 2.1.7 holds if there exists an optimal alignment a of x and y leaving out no blocks, and having a proportion p_{ij} of aligned block pairs such that the x-block has length i and the y-block has length j (for every $i, j \in \{l-1, l, l+1\}$).

Typically, for large n, the optimal alignment will not be like in the example above, but there will be blocks which are left out, which implies also that some blocks are aligned with several blocks at the same time. Let us check an example.

Example 2.1.4 *Let* $x = 00110011100011000$ *and* $y = 00001111000011000$. *In this situation the* LCS *is equal to* $\text{LCS}(x,y) = 000011100011000$ *and corresponds to the following optimal alignment:*

x	0	0	1	1	0	0	1	1	1		0	0	0		1	1	0	0	0
y	0	0			0	0	1	1	1	1	0	0	0	0	1	1	0	0	0
LCS	0	0			0	0	1	1	1		0	0	0		1	1	0	0	0

(2.1.8)

which in block representation would be:

x	001100	111	000	11	000
y	0000	1111	0000	11	000
LCS	0000	111	000	11	000

(2.1.9)

In the last alignment above we see that the first block of y is aligned with the first and third block of x. This implies that the second block of x is "completely left *out", which means all its bits are aligned with gaps. The other blocks are aligned one block with one block: the fourth block of x is aligned with the second block of y, whilst the fifth block of x is aligned with the third block of y. Finally the last blocks of x and y are aligned with each other.*

In everything that follows, the proportions p_{ij} will only refer to the block pairs aligned one block with one block. Hence, in the alignment 2.1.8, the first three blocks of x and the first block of y do not contribute to $\{p_{ij}\}_{i,j}$.

Example 2.1.5 *In the last example above there are 4 block-pairs aligned one block with one block. The corresponding pairs of block-lengths are:*

$$(3,4);(3,4);(2,2);((3,3)$$

Hence for the alignment 2.1.8, we find $p_{3,4} = 2/4$, $p_{2,2} = 1/4$, $p_{3,3} = 1/4$ and $p_{ij} = 0$ for all $(i,j) \notin \{(3,4),(2,2),(3,3)\}$. We will denote by q_1, resp. q_2, the proportion of left out blocks in x, resp. in y. In the alignment 2.1.8, in the sequence x there is one left out block from a total of 7 blocks. This implies that $q_1 = 1/7$. There is no left out block in y so that $q_2 = 0$. In section 3.1, we will see that typically, for n large enough, q_1 and q_2 can be taken as close to each other as we want to. When $q_1 = q_2$ we denote the proportion of left out blocks by q. When we choose a block of length $l-1$ in x to increase its length we will have to consider the probability that the block is not aligned one block with one block. In the alignment 2.1.8, there are 4 blocks in x of length $l-1=2$. The first three are not aligned one block with one block: the second is left out, whilst the first and the third block are aligned with the same block of y. Hence in 2.1.13 the proportion of blocks not aligned one to one among the blocks of length 2 is 3/4. On the other hand, the blocks of length $l+1=4$ in x are all aligned one to one. So, for the alignment 2.1.8, we have that the proportion of blocks not aligned one to one among the blocks of length 4 is 0.

Using some combinatorial arguments in section 3.1 we will see that typically the proportion among the blocks of x of length $l-1$ which are not aligned one block with one block is not more than $9q$. Similarly for the blocks of length $l+1$ in x one gets a bound $3q$ for the proportion of blocks aligned with several blocks of y or left out. We can rewrite the lower bound on the right side of inequality 2.1.5, taking also into account the left out blocks. Assuming that there is an equal proportion of blocks q which are not aligned one to one in x and in y we get the following lower bound for the conditional expected increase in the LCS:

$$\frac{p_{l-1,l}+p_{l-1,l+1}}{p_{l-1,l-1}+p_{l-1,l}+p_{l-1,l+1}}(1-9q)-\frac{p_{l+1,l+1}}{p_{l+1,l-1}+p_{l+1,l}+p_{l+1,l+1}}(1-3q)-3q \quad (2.1.10)$$

The above lower bound for the conditional expected increase in LCS holds assuming that the following conditions holds:

- There exists an optimal alignment leaving out exactly the same proportion q of blocks in X and in Y. For that optimal alignment a, let $\{p_{ij}\}_{i,j}$ denote the empirical distribution of the aligned block pairs, so that $p_{ij} = P_{ij}(a)$.
- There is exactly the same number of blocks in X and in Y.
- In X, each block lenghts $l-1, l, l+1$ constitutes exactly 1/3 of the blocks. Same thing in Y.

The above conditions do not typically hold exactly but only approximately. We first look at this somehow simplified case before looking at the general case (for the general case, see the proof of theorem 2.1.3). Let us next explain how we get the bound 2.1.10 for this somehow simplified case (also, the reader should compare it to the version 2.1.7 with no gaps). Assume next that we have an optimal alignment a with given empirical distribution $\{p_{ij}\}_{i,j\in\{l-1,l,l+1\}}$ of the aligned block pairs and leaving out in both sequences x and y a proportion q of blocks. What is now the effect of our random change on the score of the alignment a? First let us look at the randomly chosen block of length $l-1$ which gets its length changed to l. If that block is aligned with a block of length l or $l+1$ the alignment gets increased by one unit. So, conditional that the randomly chosen block of length $l-1$ is a block aligned one to one, we get that the probability of an increase is equal at least to:

$$\frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}}.$$

Now, if the randomly chosen block of length $l-1$ is aligned with two or more blocks, then we also get an increase by one unit. If the chosen block however is aligned with a block of Y which is aligned with several blocks of Y (let us call it a *polygamist* block), then we have no increase. The same happens if the block is not aligned with a block of Y. There are at most a proportion of $3q$ blocks which are not aligned with any block or aligned together with polygamist block of Y. There are about a proportion of 1/3 blocks of length $l-1$. Hence among the blocks of length $l-1$, there is a proportion of at least $1-9q$ which are aligned one block with one block or aligned one with several. Hence we get that the conditional expected change due to changing the randomly chosen block of length $l-1$ to l is equal at least to:

$$\frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}}(1-9q). \tag{2.1.11}$$

Similarly we can analyze the effect of the randomly chosen block of length $l+1$ which gets reduced to length l. If the block is aligned one block to one block and the length of the aligned block of Y is $l+1$ then the score can get reduced by one. If the block is aligned with a block of Y of length l or $l-1$ the score does not get reduced. Hence, given that the block of length $l+1$ chosen is aligned one block to one block, the conditional expected change is not less than:

$$-\frac{p_{l+1,l+1}}{p_{l+1,l-1}+p_{l+1,l}+p_{l+1,l+1}}.$$

On the other hand, when the chosen block of length $l+1$ is aligned with several blocks of Y then the score goes down by one unit. There are at most a proportion q of blocks of X aligned with several blocks of Y. So, among the blocks of length $l+1$ this represents a proportion of at most $3q$. Hence we get that at worst the expected change due to changing a random block from $l+1$ to l is equal to:

$$-\frac{p_{l+1,l+1}}{p_{l+1,l-1}+p_{l+1,l}+p_{l+1,l+1}}(1-3q)-3q \tag{2.1.12}$$

Putting 2.1.11 and 2.1.12 together we get that the expected conditional change of the alignment score is bounded below as follows:

$$\mathrm{E}[\Delta L_a|X,Y] \geq \frac{p_{l-1,l}+p_{l-1,l+1}}{p_{l-1,l-1}+p_{l-1,l}+p_{l-1,l+1}}(1-9q)-\frac{p_{l+1,l+1}}{p_{l+1,l-1}+p_{l+1,l}+p_{l+1,l+1}}(1-3q)-3q$$

where ΔL_a denotes the change in score of the alignment a due to the random modification of X.

Then, to prove inequality 2.1.1 in theorem 2.1.2, it is thus sufficient to show that for all optimal alignments a of X and Y, expression 2.1.10 is positive and bounded away from zero with high probability. Hence the next question is how can we prove that typically, for large n, expression 2.1.10 is larger than a positive constant not depending on n?

Example 2.1.6 *Let us return back to the example of alignment 2.1.8. That alignment left out only one block, and that was the second block of X. We could now proceed in a different order. We could first decide which blocks get left out before generating the random sequences X and Y. The resulting alignment is in general not optimal. On the other hand, such an alignment has the property that the block pairs aligned one to one are i.i.d. This is a very nice property for large deviation estimations, for instance. Let us give an example. Assume we request that the only left out block is the second block of X (as in alignment 2.1.8). Assume we redraw X and Y and obtain $X = 00111001100011000$ and $Y = 00011110000111000$. Then we get as alignment and Common Subsequence (CS) the following:*

x	0	0	1	1	1	0	0	1	1			0	0	0		1	1		0	0	0
y	0	0				0		1	1	1	1	0	0	0	0	1	1	1	0	0	0
CS	0	0				0		1	1			0	0	0		1	1		0	0	0

(2.1.13)

which can be represented as an alignment of blocks by:

x	0011100	11	000	11	000
y	000	1111	0000	111	000
CS	000	11	000	11	000

In this case we use the term of common subsequence instead of the longest common subsequence because we are leaving a block out of the alignment, if we do not leave it out we might get a longer common subsequence (which does not happen in this case neither but might happens in the general case). So, in this last example, before drawing X and Y, we know that the fourth block of X gets aligned with the second block of Y and this aligned pair builts the second block in the CS. The length of the second block of the CS has thus length equal to $\min\{B_{X4}, B_{Y2}\}$. Similarly, before even drawing X and Y, we know that the fifth block of X gets aligned with the third block of Y. Hence, we have that the pair of lengths in the second block pair is (B_{X5}, B_{Y3}) whilst the third block of the CS has length $\min\{B_{X5}, B_{Y3}\}$. Note that (B_{X4}, B_{Y2}) is independent of (B_{X5}, B_{Y3}) and B_{X4} is independent of B_{Y2} whilst B_{X5} is independent of B_{Y3}. The distribution of each of the blocks B_{X4}, B_{Y2}, B_{X5} and B_{Y3} is unchanged, they take value $l-1$, l or $l+1$ with equal probability 1/3. Hence, (B_{X4}, B_{Y2}) can take any of the nine values in the set $\{(i,j)|i,j = l-1, l, l+1\}$ with probability 1/9.

When we specify an alignment by deciding which blocks we leave out before drawing X and Y, the aligned block pairs are "almost" i.i.d. Why do we say "almost"? In the above example (B_{X4}, B_{Y2}) and (B_{X5}, B_{Y3}) are i.i.d. and not just close to be i.i.d. On the other hand, block B_{X7} in the case 2.1.8 is no longer in X if the first, third and fourth blocks get each increase by one unit. In this sense the blocks are not completely independent. But since we take n large this is only a minor effect. We will take care of this detail in section 4 and until then pretend that the aligned block pairs are i.i.d.

Note that for each alignment a defined by specifying which blocks we left out before drawing X and Y, the empirical distribution of the aligned blocks is random. We write $\{P_{ij}(a)\}_{i,j\in\{l-1,l,l+1\}}$ for this empirical distribution. Thus, $P_{ij}(a)$ denotes the proportion of aligned block pairs where the block of X has length i and the block of Y has length j. Given a non-random distribution $\{p_{ij}\}_{i,j\in\{l-1,l,l+1\}}$ we can ask what is the probability for the empirical distribution to be equal to the $\{p_{ij}\}_{i,j}$. The answer is, since the block pairs are close to i.i.d, the distribution is close to a multinomial distribution:

$$\mathrm{P}\,(\ P_{ij}(a) = p_{ij}, \forall i,j \in I_l\ ) \approx \binom{n^*}{p_{l-1,l-1}n^*\ p_{l-1,l}n^*\ \ldots\ p_{l+1,l}n^*\ p_{l+1,l+1}n^*} \left(\frac{1}{9}\right)^{n^*} \tag{2.1.14}$$

where n^* designates the total number of aligned block pairs (here we act as if that number would be non-random). By using Stirling, the expression 2.1.14 is approximately equal to:

$$e^{(\ln(1/9)+H(p))n^*} \tag{2.1.15}$$

where $H(p)$ designates the entropy of the empirical distribution:

$$H(p) = \sum_{i,j\in\{l-1,l,l+1\}} p_{ij}\ln(1/p_{ij}).$$

A question arises: for a given aligned block pairs distribution $\{p_{ij}\}$, is it likely that there exist an alignment with that distribution and having a proportion q of left out blocks? Let $\mathcal{A}(q)$ denote the set of alignments leaving out a proportion q of blocks. Let A denote the event that there exists an alignment in $\mathcal{A}(q)$ having its empirical distribution equal to $\{p_{ij}\}$. An upper bound for the probability $\mathrm{P}(A)$ is given by the number of elements in $\mathcal{A}(q)$ times the probability 2.1.14. By using 2.1.15, this product is close to:

$$|\mathcal{A}(q)|\cdot e^{(\ln(1/9)+H(p))n^*}. \tag{2.1.16}$$

But the size of the set $\mathcal{A}(q)$ is approximately equal to $e^{2H(q)n/l}$, since there are about n/l blocks. Hence, expression 2.1.16 is approximately equal to:

$$e^{(2H(q)n/l)+(\ln(1/9)+H(p))n^*}. \tag{2.1.17}$$

If we want the event A to not have exponentially small probability in n, we need the logarithm of 2.1.17 to be non-negative, which leads to the condition:

$$2H(q)+(1-4q)(\ln(1/9)+H(p))\geq 0, \tag{2.1.18}$$

where we used as lower bound on n^* the number $(n/l)(1-4q)$.

We can now explain how we prove that typically, for all optimal alignment, expression 2.1.10 is larger than a positive constant not depending on n. For this we simply need to find a q_0 so that we can prove that the optimal alignment leaves out at most a proportion of $q\leq q_0$ blocks and then show that expression 2.1.10 is bounded away from zero under condition 2.1.18 for $q\in[0,q_0]$.

Let $F^n(q)$ be the event that any optimal alignment of X and Y leaves out at most a proportion q of bocks in X and leaves out the same proportion q of blocks in Y. In more details, given $q>0$ and an optimal alignment a of X and Y in $F^n(q)$, we can count the number of blocks that are left out (not used in a) and divide this number by the total number of blocks in X to obtain q_1, and also divide this number by the total number of blocks in Y to obtain q_2, then we know that $q_1\leq q$ and $q_2\leq q$.

Example 2.1.7 *Let us take again the case where* $X = 00111001100011000$ *and* $Y = 00011110000111000$*, then we have as before the following common subsequence (CS) represented in an alignment:*

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

and represented as an alignment of blocks by:

X	0011100	11	000	11	000
Y	000	1111	0000	111	000
CS	000	11	000	11	000

Let us compute q_1 *and* q_2 *in this case. For* X *we have a total of 7 blocks and only 1 block is left out in the alignment, so* $q_1 = 1/7$*. For* Y *we do not have left out blocks so* $q_2 = 0$*. Then given* $q > 0$*, this alignment belongs to* $F^n(q)$ *if and only if* $q_1 = 1/7 \leq q$ *and* $q_2 = 0 \leq q$*.*

The next theorem says that if we can bound expression 2.1.10 away from zero under condition 2.1.18, then we have typically the desired bias for $\mathrm{E}[\tilde{L}_n - L_n|X, Y]$ the conditional expected increase in score:

Theorem 2.1.3 *Assume that there exists* $q_0 \in [0, (1/3)[$ *such that the following minimizing problem:*

$$\min\left(\frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}}(1 - 9q) - \frac{p_{l+1,l+1}}{p_{l+1,l-1} + p_{l+1,l} + p_{l+1,l+1}}(1 - 3q) - 3q\right) \tag{2.1.19}$$

under the conditions:

$$q \in [0, q_0], \sum_j p_{l-1,j} \geq ((1/3) - q_0)/2 \ , \ \sum_j p_{l+1,j} \geq ((1/3) - q_0)/2 \tag{2.1.20}$$

$$\sum_{i,j \in I} p_{ij} = 1, p_{ij} \geq 0, \forall i, j \in I \tag{2.1.21}$$

$$2H(q) + (1 - 4q)\left(\ln(1/9) + H(p)\right) \geq 0 \tag{2.1.22}$$

has a strictly positive solution. Let this minimum be equal to $2\epsilon > 0$*. Then we have that:*

$$\mathrm{P}\Big(\ \mathrm{E}[\tilde{L}_n - L_n|X, Y] \geq \epsilon \ \Big) \geq 1 - e^{-n^\beta} - \mathrm{P}(F^{nc}(q_0)) \tag{2.1.23}$$

where $\beta > 0$ *is a constant not depending on* n*.*

The last theorem above reduces proving condition 2.1.1 to a minimizing problem. Note that this minimizing problem depends on the parameter q_0. Also, if the probability to have less than q_0 gaps is not a likely event, that is if $P(F^{nc}(q_0))$ is not small, then 2.1.23 is useless. To be able to use the above theorem, we first need to find a q_0 such that $P(F^{nc}(q_0))$ is small. Having such a q_0, we try to show that the minimizing problem above has a strictly positive solution. If we succeed in proving that, then we get the likely biased effect of the random modification (inequality 2.1.23) which, according to theorem 2.1.2, implies the fluctuation order $\text{VAR}[L_n] = \Theta(n)$.

The probability $\text{P}(F^{nc}(q))$ depends on the parameter l. In chapter 3.2, we show how to find upper bounds on the proportion of left out blocks. In general, for l larger, the bounds gets better. Actually the bounds even converge to zero as l goes to infinity. As q goes to zero, expression 2.1.19 gets close to 1/3 on the domain. That is why the minimizing problem in theorem 2.1.3 has a strictly positive solution when l is large enough. Next we are going to prove formally that from the last theorem 2.1.3 it follows, for l large enough, that we have $\text{VAR}[L_n] = \Theta(n)$. In other words, we prove that theorem 2.1.3 and theorem 2.1.2 implies theorem 2.1.1. Here comes the proof:

Proof. We suppose that $F^{nc}(q_0)$ has exponentially small probability for any fixed $q_0 > 0$ provided l is large enough (see section 3.2 and 4.2). In section 4 we will show how large l should be depending on q_0 but not on n. The conditions in theorem 2.1.3 are satisfied when $q_0 > 0$ (hence $q \leq q_0$ small enough) is taken small enough. Let us explain why. First note that inequality 2.1.22 can be written:

$$H(p) \geq \frac{-2H(q)}{1-4q} + \ln(1/9) \tag{2.1.24}$$

When q goes to zero, then $H(q)$ also goes to zero and so does $2H(q)/1-4q$. But we have that $H(p)$ is always less or equal to $\ln(1/9)$, with equality iff all the p_{ij}'s are equal to 1/9.

It follows that by taking $q > 0$ small enough, we get condition 2.1.24 to imply that the distribution p_{ij} gets as close as we want to the equiprobable distribution. On the other hand, when q goes to zero and all the p_{ij}'s converge to 1/3, then the quantity

$$\frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}}(1-9q) - \frac{p_{l+1,l+1}}{p_{l+1,l-1} + p_{l+1,l} + p_{l+1,l+1}}(1-3q) - 3q,$$

converges to $2/3 - 1/3 = 1/3 > 0$. This shows that by taking $q_0 > 0$ small enough we get that the minimizing problem in theorem 2.1.3, has a strictly positive solution. So, assume that $q_0 > 0$ is such that the following two things hold:

- $F^{nc}(q_0)$ has exponentially small probability in n.

- The minimizing problem in theorem 2.1.3, has a strictly positive solution. Call this solution 2ϵ, where $\epsilon > 0$.

By theorem 2.1.3, we then have that inequality 2.1.23 holds. But since $F^{nc}(q_0)$ is exponentially small in n, we get that the expression on the right hand side of inequality 2.1.23 is smaller than $\exp(-n^{\alpha})$ for all n and $\alpha > 0$ not depending on n. This implies that condition 2.1.1 in theorem 2.1.2 is satisfied. Then theorem 2.1.2 implies that:

$$\mathrm{VAR}[L_n] = \Theta(n). \qquad \blacksquare$$

Chapter 3

Left out blocks in an optimal alignment

3.1 Combinatorics of the left out blocks

Example 3.1.1 *Let* $X = 0011100$ *and* $Y = 0001100$. *The* LCS *is* 001100. *This corresponds to the following alignment:*

$$\begin{array}{c|c|c|c|c|c|c|c|c|c}
X & & 0 & 0 & & 1 & 1 & 1 & 0 & 0 \\
\hline
Y & & 0 & 0 & 0 & 1 & 1 & & 0 & 0 \\
\hline
\text{LCS} & & 0 & 0 & & 1 & 1 & & 0 & 0
\end{array} \tag{3.1.1}$$

In this example, the first block of the LCS *has length* 2. *It is obtained from the first block of* X *and the first block of* Y. *The first block of* X *has length* 2 *whilst the first block of* Y *has length* 3. *The length of the first block of the* LCS *is equal to the minimum of these two numbers. In this kind of situation we say that the first block of* X *is aligned to the first block of* Y. *Similarly the length of the second block of the* LCS *is the minimum of the lengths of the second block of* X *and of* Y. *We say that in this alignment the second block of* X *gets aligned with the second block of* Y. *Finally the third block of* X *gets aligned with the third block of* Y *to yield the third block of the* LCS. *In this present example no block of* X *or* Y *got left out completely: every block "contributed" some bits to the* LCS. *All the blocks are aligned one block of* X *with one block of* Y. *Each such pair of aligned blocks is responsible for one block in the* LCS.

In some other cases, some blocks of X and Y are completely left out. Let us look at such a situation.

Example 3.1.2 *Consider* $X = 00100000111$ *and* $Y = 00000100011$. *The* LCS *would be* 000000011. *The* LCS *corresponds to the alignment:*

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|c|c|c|c|c}
X & & 0 & 0 & 1 & 0 & 0 & 0 & & 0 & 0 & & 1 & 1 & 1 \\
\hline
Y & & 0 & 0 & & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & \\
\hline
LCS & & 0 & 0 & & 0 & 0 & 0 & & 0 & 0 & & 1 & 1 &
\end{array} \tag{3.1.2}$$

In the last example above we have that the second block of X and of Y are totally left out and do not contribute to the LCS. We say that these blocks are *left out* blocks. The last block of X and the last block of Y "get aligned" together to yield the last block of the LCS. We say that this is an *aligned block pair* or also that these two blocks are *aligned one block to one block*. One way of thinking about the LCS defined by the alignment 3.1.2 above is as follows: we first decide which blocks we leave out in X and Y. Then from the two obtained sequences, we align block by block without leaving out any blocks. So the alignment 3.1.2 can be seen as the alignment in which we leave out the second block of X and the second block of Y. This gives then the modified sequences $X^* = 0000000111$ and $Y^* = 0000000000011$. Then we align X^* and Y^* block by block. The common subsequence we obtain has its i-th block having length equal to the minimum of the length of block i of X^* and of Y^*. In this example we have that the first and the third block of X get aligned with the first and third block of Y. By this we mean that in both sequences the first and third block are made into one block and these blocks are then matched. We will be able to prove that in the case we study here this is untypical: for optimal alignment we will only have one block aligned with several at the same time, but not several with several.

Let us look at one more example.

Example 3.1.3 *Let* $X = 001001111$ *and* $Y = 000011011$. *The* LCS *is* 00001111. *This corresponds to the alignment:*

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

(3.1.3)

Here the second block of X is left out. Hence the first and the third block of X get aligned with the first block of Y. Similarly the fourth block of X gets aligned with the second and fourth block of Y. The third block of Y is left out.

This situation will happen in optimal alignment: one block aligned with several blocks of the other sequence.

Assume that we know for an alignment a which blocks are left out. Assume that X^*, resp. Y^* denotes the modified sequence X, resp. Y where we left out the specified blocks. Let Z denote a common subsequence defined by the alignment a. The alignment must then align all the blocks of X^* with the blocks of Y^* one to one, otherwise there would be more left-out blocks. Hence, the first block of X^* gets aligned, then the second block of X^* and so on. If the alignment wants to stand a chance to be an optimal one (and hence Z to be a LCS) for each pair of aligned blocks from X^* and Y^* aligned to one another, it needs to extract a maximum of bits of each such pair. Hence, for every $i = 1, 2, \ldots, j$ we have that the length of the block number i of Z must be equal to the minimum between the length of the i-th block of X^* and the length of the i-th block of Y^* (here j denotes the number of blocks in Z.) Hence, since we are interested in LCS's (and hence in optimal alignments) we will only consider alignments defined in the following manner: first we define exactly which blocks get left out. Second we align the resulting sequences X^* and Y^* one block with one block. The next lemma says that in our setting an optimal alignment cannot align several blocks with several blocks.

Another useful fact is that for optimal alignments we do not need to consider adjacent left-out blocks except maybe at the end of the sequences. But in section 4 we prove that only a small percentage of bits could be left out at the end of X and Y in an optimal alignment. Hence, the practical implication is that we only need to consider left out blocks at least separated by one non-left out block. Let us first explain what we mean by adjacent left out blocks between aligned blocks:

Example 3.1.4 *Take* $x = 11001100$ *and* $y = 00001100$. *Let us align as follows:*

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|c|c|c|c} x & & & & & & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ \hline y & & 0 & 0 & 0 & 0 & 1 & 1 & & & & & 0 & 0 \end{array} \tag{3.1.4}$$

We see a typical situation where the second and third block of x in the alignment above get left out (i.e. entirely aligned with gaps). These two blocks are adjacent and they are comprised between aligned blocks (i.e. in our example they are comprised between the first and fourth block of x which are "aligned", by aligned we mean aligned with another block hence not entirely aligned with gaps). The next lemma below states that for our LCS problem (i.e. optimal alignments) the kind of situation we face in the current numerical example 3.1.4 can be discarded. The reason is as follows. In the current example y_7 gets aligned with x_7. Now instead align y_7 with x_3 and keep all the rest of alignment 3.1.4 identical otherwise. Then by doing this you have not decreased the score but have destroyed the situation of two adjacent completely left out blocks. The next lemma shows what we explained in our example in a rigorous way:

Lemma 3.1.1 *There exists an optimal alignment of* X *and* Y *having no adjacent left-out blocks between aligned blocks.*

Proof. View an alignment as a finite sequence of points in $\mathbb{N} \times \mathbb{N}$, so that if x_i gets aligned with y_j, then (i, j) is a point in the set representing the alignment. Introduce for two alignments $a, b \in \mathbb{N}\times\mathbb{N}$ the order relation $a \leq b$ iff all a contains the same number of points as b and if we numerate in both sets the points from down left to up right then the i-th point $a_i = (a_{ix}, a_{iy})$ of a and the i-th point $b_i = (b_{ix}, b_{iy})$ of b satisfy $a_{ix} \leq b_{ix}$ and $a_{iy} \leq b_{iy}$ for all $i \leq |a|$. Here $|a|$ designates the number of points in a. Take now an optimal alignment which is minimal according to the relation $\leq$. That optimal alignment satisfies the property of not having several adjacent left out blocks between aligned blocks. ■

Next we show the relation between left out blocks at the end of each sequence and the total left out blocks in each sequence:

Lemma 3.1.2 *Let $x, y \in \{0,1\}^n$ be two sequences of length n. Let the number of blocks of x, resp. y be denoted by $n_1^* = (n/l) + \Delta_1$, resp. $n_2^* = (n/l) + \Delta_2$. Assume that $|\Delta_1|, |\Delta_2| \leq \Delta$. Assume also that a is an alignment of x and y which does never leave out adjacent blocks except maybe a contiguous group at the very end of x and of y. Let $\delta_1 \geq 0$ denote the proportion of blocks which are entirely left out at the end of x, resp. y, among all the blocks of x, resp. y. Let q_1, resp. q_2 denote the proportion of blocks left out in x, resp. in y. Then we find that:*

$$|q_1 - q_2| \leq 1.5|\delta_1 - \delta_2| + \frac{4l\Delta}{n} \tag{3.1.5}$$

Proof. Let x^*, resp. y^* denote the sequence we obtain after we removed the blocks which are completely left out by a. Since there are no other completely left out blocks, we have that the number of blocks in x^* must be equal to the number of blocks in y^*. Note that for every left out blocks which has no adjacent left out block the number of blocks is reduced by 2. for the adjacent left out blocks at the end, for each left out block there is one block less. Since there are no adjacent left out blocks except the adjacent blocks at the end, we get that the number of blocks of x^*, resp. of y^* is equal to

$$n_1^*(1 - 2(q_1 - \delta_1) + \delta_1)), \tag{3.1.6}$$

resp.

$$n_2^*(1 - 2(q_2 - \delta_2) + \delta_2)). \tag{3.1.7}$$

Taking the difference of 3.1.6 and 3.1.7 and dividing by $(l/2n)$, we find

$$q_1 - q_2 = 1.5(\delta 2 - \delta_1) + \frac{bl\Delta}{n} \tag{3.1.8}$$

where

$$2b = 1 - 2(q_1 - \delta_1) + \delta_1) - (1 - 2(q_2 - \delta_2) + \delta_2))$$

we see that b is always smaller than 4 which ends the proof. ■

Lemma 3.1.3 *For $l > 4$ any optimal alignment of X and Y does not align several blocks in X with several blocks in Y.*

Proof. Let us explain the idea behind through an example. Let us take $x =$ 0001111000111100000 and $y =$ 0001111000001110000 two realizations of X and

Y, respectively, with $l = 4$. An alignment using all blocks of x and y in block representation becomes:

x		000	1111	000	1111	00000
y		000	1111	00000	111	0000
LCS		000	1111	000	111	0000

(3.1.9)

Let us now suppose that we leave out the second block of x and the second block of y, then the alignment in block representation looks like:

x		0001111000	1111	00000
y		000111100000	111	0000
LCS		000000	111	0000

(3.1.10)

One clearly sees that in alignment 3.1.10 we lost the entire block of 1's of length 4 and we did not gain any new aligned symbol, so the LCS decreased on 4 units compared to alignment 3.1.9. In this particular example, the neighbour blocks of the left out block in y had all together at least as many symbols (8 zeros all together) as the neighbour blocks of the left out block in x had all together (6 zeros). In general we could gain at most 2 new symbols from the neighbour blocks of the left out block but we always loose at least $l - 1$ symbols leaving a block out and aligning its neighbours together instead. The other blocks do not get involved in the change on the score. Hence, when one leaves out a block and tries to align the neighbour blocks together the LCS changes in $2-(l-1) = 3-l$. Then for blocks of length $l > 4$, to align several blocks with several blocks decreases the LCS rather than to increase it. ■

3.2 Maximum number of left out blocks

The first key question is the percentage of blocks which are at most left out in an optimal alignment. Since the blocks have length $l - 1$, l or $l + 1$ with equal probability 1/3 the expected block length is l. Hence, the expected number of blocks in a sequence of length n is about n/l. Now let us define the limit:

$$\gamma_l = \lim_{n\to\infty} \frac{\mathrm{E}[L_n]}{n}. \tag{3.2.1}$$

Hence, the number of bits in the sequence X (and also in the sequence Y) which are not used for the LCS is about $(1 - \gamma_l)n$. Every block we leave out means at least $l - 1$ non-used bits. Hence, the number of left out blocks for long sequences can typically not be much above:

$$\frac{(1 - \gamma_l)n}{l - 1}.$$

This represents typically a proportion of:

$$\frac{(1-\gamma_l)n/(l-1)}{n/l} = \frac{1-\gamma_l}{1-(1/l)}$$

from the total number of blocks. Hence we find that the proportion of left out blocks in the optimal alignment is typically close or below the following bound:

$$\frac{1-\gamma_l}{1-(1/l)}. \tag{3.2.2}$$

Let us next find a simple lower bound for γ_l which we can use in expression 3.2.2. Assume we choose an alignment which leaves out no blocks. The typical score of such an alignment gives a lower bound for γ_l. In this case the common subsequence defined by such an alignment has its i-th block having length:

$$B_i := \min\{B_{Xi}, B_{Yi}\}.$$

where B_{Xi} (resp. B_{Yi}) is the length of the i-th block of X (resp. Y). Recall that B_{Xi} (resp. B_{Yi}) has uniform distribution on the set $\{l-1, l, l+1\}$. The distribution of the minimum above is as follows:

$$\mathrm{P}(B_i = l-1) = 5/9, \mathrm{P}(B_i = l) = 3/9, \mathrm{P}(B_i = l+1) = 1/9.$$

The expected length is thus:

$$\mathrm{E}[B_i] = \frac{5}{9}(l-1) + \frac{3}{9}l + \frac{1}{9}(l+1) = l - \frac{4}{9}. \tag{3.2.3}$$

Since there are about n/l blocks, the score aligning all the blocks gives thus about a score of:

$$\frac{n}{l} \cdot \mathrm{E}[B_i] = n\left(1 - \frac{4}{9l}\right),$$

so that we obtain:

$$\gamma_l \geq \left(1 - \frac{4}{9l}\right).$$

The last inequality together with the bound 3.2.2 implies that the proportion of left out blocks should typically not be much above the following bound:

$$\frac{1-(1-(4/9l))}{1-(1/l)} = \frac{4/9}{l-1} \tag{3.2.4}$$

Another similar approach is to get a lower bound for γ_l by simulations. As a matter of fact we have for any n that $\mathrm{E}[L_n]/n$ is a lower bound for γ_l. By Montecarlo we can find an estimate of $\mathrm{E}[L_n]/n$ and a very likely lower bound γ_{lb}. We then replace in inequality 3.2.2 γ_l by γ_{lb}.

Chapter 4

High probability events

Let $\delta > 0$ be a parameter not depending on n. We will define a number of events related with the combinatorial properties of the optimal alignments, called C^n, $D^n(\delta)$, $G^n(\delta)$ and $J^n(\delta)$. In the following we will prove that these events have high probability for n large. By high probability, we mean a quantity which is negatively exponential close to one in n. It will turn out that this is true for the above events for any parameters $\delta > 0$ not depending on n. Also we will prove that $F^n(q)$ has high probability for n large in the same sense as above but restricted to some values of q.

A very useful tool we use often is the Azuma-Hoeffding theorem. The following is a version of it for martingales (for a proof see [14]):

Theorem 4.0.1 *(Hoeffding's inequality) Let $(V, \mathfrak{F})$ be a martingale, and suppose that there exists a sequence $\mathfrak{a}_1, \mathfrak{a}_2, \cdots$ of real numbers such that*

$$\mathrm{P}(|V_n - V_{n-1}| \leq \mathfrak{a}_n) = 1$$

for all n. Then:

$$\mathrm{P}(|V_n - V_0| \geq v) \leq 2 \exp\Big\{ -\frac{1}{2} v^2 \Big/ \sum_{i=1}^{n} \mathfrak{a}_i^2 \Big\} \tag{4.0.1}$$

for every $v > 0$.

We also will use a corollary of the above theorem, for some intermediate bounds:

Corollary 4.0.1 *Let $a > 0$ be constant and $V_1, V_2, \ldots$ be an i.i.d sequence of random bounded variables such that:*

$$\mathrm{P}(|V_i - \mathrm{E}[V_i]| \leq a) = 1$$

for every $i = 1, 2, \ldots$ Then for every $\Delta > 0$, we have that:

$$\mathrm{P}\left(\left| \frac{V_1 + \cdots + V_n}{n} - \mathrm{E}[V_1] \right| \geq \Delta \right) \leq 2 \exp\left(-\frac{\Delta^2}{2a^2} \cdot n \right) \tag{4.0.2}$$

4.1 Number of blocks as renewal process

For $k > 0$ let us define the sum of the length of the first k blocks in X as:

$$S_k^X = B_{X1} + \cdots + B_{Xk}$$

Let us define the number of blocks used in a sequence of length t in X as:

$$N_t^X = \max\{k > 0 : S_k^X \leq t\} \tag{4.1.1}$$

Note that there might be at the end of X a block which has length smaller than $l-1$. Since this is at most one block it plays little role and we will not mention it every time, only when it is relevant (the same will apply to Y in what follows).

Due to the standard theory of renewal processes, for every $k, t > 0$ the following relation holds between the two random variables defined above:

$$N_t^X \geq k \Leftrightarrow S_k^X \leq t. \tag{4.1.2}$$

In the same way we define for Y the same variables as before:

$$\begin{array}{rcl} S_k^Y & = & B_{Y1} + \cdots + B_{Yk} \\ N_t^Y & = & \max\{k > 0 : S_k^Y \leq t\} \end{array}$$

where still the relation $N_t^X \geq k \Leftrightarrow S_k^X \leq t$, for every $k, t > 0$ holds true.

Let C^n be the event that the number of blocks in X and in Y lies in the interval

$$I_n := \left[\frac{n}{l} - n^{0.6}, \frac{n}{l} + n^{0.6}\right].$$

Lemma 4.1.1 *There exists a constant $b_1 > 0$ depending on l such that:*

$$\mathrm{P}(C^{nc}) \leq 8\, e^{-b_1 \cdot n^{0.2}}$$

for every $n > 0$ large enough.

Proof. It is easy to see that:

$$C_n = \{N_n^X \in I_n\} \cap \{N_n^Y \in I_n\} \tag{4.1.3}$$

It is sufficient to compute directly $\mathrm{P}(\{N_n^X \in I_n\}^c)$:

$$\mathrm{P}(\{N_n^X \in I_n\}^c) \leq \mathrm{P}\left(N_n^X \leq \frac{n}{l} - n^{0.6}\right) + \mathrm{P}\left(N_n^X \geq \frac{n}{l} + n^{0.6}\right) \tag{4.1.4}$$

Now let us compute each expression separately. Let $m_1 := \lceil \frac{n}{l} - n^{0.6} \rceil$ be an auxiliar variable. We have at the beginning:

$$\begin{aligned} \mathrm{P}\left(N_n^X \leq \frac{n}{l} - n^{0.6}\right) &\leq \mathrm{P}(N_n^X \leq m_1) \\ \text{(by using } N_t^X \geq k \Leftrightarrow S_k^X \leq t) &= \mathrm{P}\left(S_{m_1}^X \geq n\right) \\ &= \mathrm{P}\left(\frac{S_{m_1}^X}{m_1} - l \geq \frac{n}{m_1} - l\right) \\ \text{(by 4.0.2 with } \mathrm{P}(|B_{X1} - l| \leq 1) = 1) &\leq 2\exp\left(-\frac{m_1}{2}\left(\frac{n}{m_1} - l\right)^2\right) \qquad (4.1.5) \end{aligned}$$

Now we need to bound m_1 in order to get the right order for moderate deviations. Let us start looking at the following:

$$\begin{aligned} \left(\frac{n}{m_1} - l\right)^2 &\geq l^2\left(\frac{n}{n - ln^{0.6} + l} - 1\right)^2, \text{ by using } m_1 \leq \tfrac{n}{l} - n^{0.6} + 1 \\ &\geq l^2\left(\frac{1}{1 - \frac{l}{n^{0.4}} + \frac{l}{n}} - 1\right)^2 \\ &\geq l^4\left(\frac{1}{n^{0.4}} - \frac{1}{n}\right)^2\left(\frac{1}{1 - \frac{l}{n^{0.4}} + \frac{l}{n}}\right)^2 \\ &\geq \frac{l^4}{n^{0.8}}\left(1 - \frac{1}{n^{0.6}}\right)^2\left(\frac{1}{1 - \frac{l}{n^{0.4}} + \frac{l}{n}}\right)^2 \qquad (4.1.6) \end{aligned}$$

We have:

$$\lim_{n\to\infty}\left(1 - \frac{1}{n^{0.6}}\right)^2\left(\frac{1}{1 - \frac{l}{n^{0.4}} + \frac{l}{n}}\right)^2 = 1 > \frac{1}{4}$$

Hence for n large enough, the expression on the right hand side of 4.1.6 is larger than $l^4/(4n^{0.8})$ so that:

$$\left(\frac{n}{m_1} - l\right)^2 \geq \frac{l^4}{4n^{0.8}} \qquad (4.1.7)$$

Also, for $n > 0$ large enough we can take:

$$m_1 = \left\lceil \frac{n}{l} - n^{0.6} \right\rceil \geq \frac{n}{l} - n^{0.6} + 1 \geq \frac{n}{2l} + 1 = \frac{n}{2l}\left(1 + \frac{2l}{n}\right) \geq \frac{n}{4l} \qquad (4.1.8)$$

Finally we can use 4.1.7, 4.1.8 in 4.1.5 to get:

$$\begin{aligned} \mathrm{P}\left(N_n^X \leq \frac{n}{l} - n^{0.6}\right) &\leq 2\exp\left(-\frac{m_1}{2}\left(\frac{n}{m_1} - l\right)^2\right) \\ &\leq 2\exp\left(-\frac{l^3}{32} \cdot n^{0.2}\right) \qquad (4.1.9) \end{aligned}$$

for $n > 0$ large enough. For the other term, let $m_2 := \left\lfloor \frac{n}{l} + n^{0.6} \right\rfloor$ be an auxiliar variable and do the same as before. We have at the begining:

$$\begin{aligned}
\mathrm{P}\left(N_n^X \geq \frac{n}{l} + n^{0.6}\right) &\leq \mathrm{P}(N_n^X \geq m_2) \\
\text{(by using } N_t^X \geq k \Leftrightarrow S_k^X \leq t) &= \mathrm{P}\left(S_{m_2}^X \leq n\right) \\
&= \mathrm{P}\left(\frac{S_{m_2}^X}{m_2} - l \leq \frac{n}{m_2} - l\right) \\
\text{(by 4.0.2 with } \mathrm{P}(|B_{X1} - l| \leq 1) = 1) &\leq 2\exp\left(-\frac{m_2}{2}\left(\frac{n}{m_2} - l\right)^2\right) \quad (4.1.10)
\end{aligned}$$

Now we need to bound m_2 in order to get the right order for moderate deviations. Let us start looking at the following:

$$\begin{aligned}
\left(\frac{n}{m_2} - l\right)^2 &\geq l^2\left(\frac{n}{n + ln^{0.6}} - 1\right)^2, \text{ by using } m_2 \leq \tfrac{n}{l} + n^{0.6} \\
&\geq l^2\left(\frac{1}{1 + \frac{l}{n^{0.4}}} - 1\right)^2 \\
&\geq \frac{l^4}{n^{0.8}}\left(\frac{1}{1 + \frac{l}{n^{0.4}}}\right)^2 \quad (4.1.11)
\end{aligned}$$

where the very last inequality was obtained by assuming n large enough and noticing that:

$$\lim_{n\to\infty}\left(\frac{1}{1 + \frac{l}{n^{0.4}}}\right)^2 \geq \frac{1}{4}$$

Also, for $n > 0$ large enough we can take:

$$m_2 = \left\lfloor \frac{n}{l} + n^{0.6} \right\rfloor \geq \frac{n}{2l} \quad (4.1.12)$$

Finally we can use 4.1.11, 4.1.12 in 4.1.10 to get:

$$\begin{aligned}
\mathrm{P}\left(N_n^X \geq \frac{n}{l} + n^{0.6}\right) &\leq 2\exp\left(-\frac{m_2}{2}\left(\frac{n}{m_2} - l\right)^2\right) \\
&\leq 2\exp\left(-\frac{l^3}{16} \cdot n^{0.2}\right) \quad (4.1.13)
\end{aligned}$$

Then combining 4.1.4, 4.1.9 and 4.1.13 we obtain:

$$\mathrm{P}(\{N_n^X \in I_n\}^c) \leq 4\,\exp\left(-n^{0.2} \cdot \frac{l^3}{32}\right)$$

and by symmetry we finally get:

$$\mathrm{P}(C_n^c) \leq 8 \exp\left(-n^{0.2} \cdot \frac{l^3}{32}\right)$$

for every $n > 0$ large enough. Taking $b_1 = \frac{l^3}{32} > 0$ the proof is finished. ■

4.2 Left out blocks in an optimal alignment

Let $F^n(q)$ denote the already defined event that any optimal alignment of X and Y leaves out at most a proportion q of blocks in X as well as in Y.

Lemma 4.2.1 *For any q satisfying $q > \frac{4}{9(l-1)}$, we have that there exists $\beta > 0$ such that:*

$$\mathrm{P}(F^{nc}(q)) \leq e^{-\beta n}$$

for all n. Note that here q does not depend on n and also $\beta > 0$ does not depend on n but on l and q.

Proof. For the proof we need two events. Let $\delta > 0$ and let $A^n(\delta)$ be the event that the two strings X and Y both have more than $n/l - \delta n$ blocks. Let K^n denote the event that when we align the first $\lceil (n/l) - \delta n \rceil$ blocks of X and Y without leaving out a single block, the score for these aligned blocks is above expectation minus $\delta(n/l)$. Recall that in the formula 3.2.3 we computed the expected length of B_i as:

$$\mathrm{E}[B_i] = l - \frac{4}{9}$$

where $B_i = \min\{B_{Xi}, B_{Yi}\}$ each $i = 1, 2, \ldots$. Hence K^n is the event that the following inequality holds:

$$\sum_{i=1}^{\lceil (n/l) - \delta n \rceil} B_i \geq \left(l - \frac{4}{9}\right) \left\lceil \frac{n}{l} - \delta n \right\rceil - \delta \frac{n}{l}.$$

Now we find that when both events A^n and K^n hold, then:

$$L_n \geq \left(l - \frac{4}{9}\right) \left\lceil \frac{n}{l} - \delta n \right\rceil - \delta \frac{n}{l},$$

which is the same as saying that the number of unused (not used for the LCS) bits are less or equal to:

$$n - \left(l - \frac{4}{9}\right) \left\lceil \frac{n}{l} - \delta n \right\rceil + \delta \frac{n}{l}.$$

Since each block has length at least $n-1$, this then implies that the number of left out blocks (left out by an optimal alignment) is at most the last bound above divided by $(l-1)$. As a proportion of the total number of blocks in X and Y, this gives the following upper bound:

$$\begin{aligned}\frac{n-\left(l-\frac{4}{9}\right)\left\lceil\frac{n}{l}-\delta n\right\rceil+\delta\frac{n}{l}}{(l-1)((n/l)-\delta n)} &\leq \frac{n-\left(l-\frac{4}{9}\right)\left(\frac{n}{l}-\delta n\right)+\delta\frac{n}{l}}{(l-1)((n/l)-\delta n)}\\ &= \frac{n-l\left(\frac{n}{l}-\delta n\right)+\frac{4}{9}\left(\frac{n}{l}-\delta n\right)}{(l-1)((n/l)-\delta n)}\\ &= \frac{4}{9(l-1)}+\frac{\delta nl+\delta n/l}{(l-1)((n/l)-\delta n)}\\ &= \frac{4}{9(l-1)}+\delta\left(\frac{1+l^2}{(l-1)(1-\delta l)}\right) \quad (4.2.1)\end{aligned}$$

where we used that by $A^n(\delta)$ the number of blocks in X and Y is above $(n/l)-\delta n$ (here $\delta<1/l$). In other words, we have just proved that when we take q to be equal to:

$$\frac{4}{9(l-1)}+\delta\left(\frac{1+l^2}{(l-1)(1-\delta l)}\right)$$

then,

$$A^n(\delta)\cap K^n(\delta)\subset F^n(q)$$

and hence:

$$\mathrm{P}(F^{nc}(q))\leq \mathrm{P}(A^{nc}(\delta))+\mathrm{P}(K^{nc}(\delta)).$$

Note that holding l fixed and letting $\delta>0$ go to zero, we find that $\delta\left(\frac{1+l^2}{(l-1)(1-\delta l)}\right)$ goes to zero. This implies that for any q satisfying

$$q>\frac{4}{9(l-1)}, \quad (4.2.2)$$

there exists $\delta>0$ such that:

$$\delta\left(\frac{1+l^2}{(l-1)(1-\delta l)}\right)<q-\frac{4}{9(l-1)}. \quad (4.2.3)$$

For given q satisfying 4.2.2, let thus $\delta>0$ be such that inequality 4.2.3 is satisfied. In that case and with that choice of q and $\delta>0$ we get that:

$$\mathrm{P}(F^{nc}(q))\leq \mathrm{P}(A^{nc}(\delta))+\mathrm{P}(K^{nc}(\delta)).$$

So to prove that $\mathrm{P}(F^{nc}(q))$ has exponentially small probability in n, it is enough to show that $\mathrm{P}(A^{nc}(\delta))$ and $\mathrm{P}(K^{nc}(\delta))$ have both exponentially small probability in n for any $\delta>0$ not depending on n. This is going to be proved (for any $\delta>0$) in the next two lemmas.

Lemma 4.2.2 *For every $0 < \delta < \frac{1}{l}$ there exists a constant $b_2 > 0$ depending on δ and l but not on n, such that:*

$$\mathrm{P}(A^{nc}(\delta)) \leq 4e^{-b_2 \cdot n}$$

for n large enough.

Proof. With our definitions on renewal processes, we can re-write $A^n(\delta)$ as follows:

$$A^n(\delta) = \left\{ N_n^X \geq \frac{n}{l} - \delta n \right\} \cap \left\{ N_n^Y \geq \frac{n}{l} - \delta n \right\} \tag{4.2.4}$$

Then let us compute directly, setting $m = \lceil \frac{n}{l} - \delta n \rceil$:

$$\begin{aligned} \mathrm{P}\left(N_n^X \leq \frac{n}{l} - \delta n \right) &\leq \mathrm{P}\left(N_n^X \leq m \right) \\ \text{(by using } N_t^X \geq k \Leftrightarrow S_k^X \leq t\text{)} &= \mathrm{P}\left(S_m^X \geq n \right) \\ &= \mathrm{P}\left(\frac{S_m^X}{m} - l \geq \frac{n}{m} - l \right) \\ \text{(by 4.0.2 with } \mathrm{P}(|B_{X1} - l| \leq 1) = 1\text{)} &\leq 2\exp\left(-\frac{m}{2}\left(\frac{n}{m} - l \right)^2 \right) \end{aligned} \tag{4.2.5}$$

From one side we have that:

$$\begin{aligned} \left(\frac{n}{m} - l \right)^2 &\geq \left(\frac{nl}{n - l\delta n + l} - l \right)^2, \text{ by using } m \leq \tfrac{n}{l} - \delta n + 1 \\ &= l^2 \left(\frac{1}{1 - l\delta + \frac{l}{n}} - 1 \right)^2 \\ &= l^2 \left(l\delta - \frac{l}{n} \right)^2 \left(\frac{1}{1 - l\delta + \frac{l}{n}} \right)^2 \\ &\geq l^2 \cdot \frac{l^2\delta^2}{4} \cdot \frac{1}{4} \left(\frac{1}{1 - l\delta} \right)^2 \end{aligned} \tag{4.2.6}$$

for $n > 0$ large enough, where the last inequality is due to:

$$\lim_{n\to\infty} \left(l\delta - \frac{l}{n} \right)^2 \left(\frac{1}{1 - l\delta + \frac{l}{n}} \right)^2 \geq \frac{l^2\delta^2}{4} \cdot \frac{1}{4} \left(\frac{1}{1 - l\delta} \right)^2$$

From the other side we also know that:

$$m \geq \frac{n}{l} - \delta n = n\left(\frac{1}{l} - \delta \right) \tag{4.2.7}$$

Then by using inequalities 4.2.6 and 4.2.7 in 4.2.5 we get:

$$\begin{aligned} \mathrm{P}\left(N_n^X \leq \frac{n}{l} - \delta n\right) &\leq 2\exp\left(-\frac{m}{2}\left(\frac{n}{m} - l\right)^2\right) \\ &\leq 2\exp\left(-n \cdot \frac{l^3\delta^2}{32(1-l\delta)}\right) \qquad (4.2.8) \end{aligned}$$

Finally by symmetry and using 4.2.8 we get:

$$\mathrm{P}(A^{nc}(\delta)) \leq 4\exp\left(-n \cdot \frac{l^3\delta^2}{32(1-l\delta)}\right)$$

which ends the proof by taking

$$b_2 = \frac{l^3\delta^2}{32(1-l\delta)} > 0. \quad \blacksquare$$

Lemma 4.2.3 *For every $0 < \delta < \frac{1}{l}$ there exists a constant $b_3 > 0$ depending on δ and l but not on n such that:*

$$\mathrm{P}(K^{nc}(\delta)) \leq e^{-b_3 \cdot n}$$

for n large enough.

Proof. First let us remember again that

$$\mathrm{E}[B_i] = l - \frac{4}{9}$$

where $B_i = \min\{B_{Xi}, B_{Yi}\}$ for each $i = 1, \ldots, n$. Now, after setting an auxiliar variable $m := \lceil \frac{n}{l} - \delta n \rceil$, we can write:

$$\begin{aligned} \mathrm{P}(K^{nc}(\delta)) &\leq \mathrm{P}\left(\frac{1}{m}\sum_{i=1}^{m} B_i \leq \left(l - \frac{4}{9}\right) - \delta\frac{n}{ml}\right) \\ &\leq \mathrm{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} B_i - \left(l - \frac{4}{9}\right)\right| \geq \frac{\delta}{1-\delta l}\right) \\ &\leq 2\exp\left(-\frac{1}{2}\cdot\frac{1}{4}\left(\frac{\delta}{1-\delta l}\right)^2 m\right), \text{ by 4.0.2 and } \mathrm{P}(|B_i - \mathrm{E}[B_i]| \leq 2) = 1 \\ &\leq 2\exp\left(-\frac{1}{8}\left(\frac{\delta}{1-\delta l}\right)^2 \left(\frac{1}{l} - \delta\right) n\right), \text{ by using } m \geq \tfrac{n}{l} - \delta n \\ &= 2\exp\left(-\frac{1}{8}\cdot\frac{\delta^2}{l(1-\delta l)}\, n\right) \end{aligned}$$

To finish the proof choose $b_3 = \frac{1}{8} \cdot \frac{\delta^2}{l(1-\delta l)}$. $\blacksquare$

4.3 Proportion of blocks in X and Y

Let X^m be the sequence X^∞ taken up to the m-th block. Similarly, let Y^m be the sequence Y^∞ taken up to the m-th block. Let $D_m(\delta)$ be the event that the proportions of blocks in X^m and Y^m of length $l-1$, l and $l+1$ are not further from 1/3 than δ. Let $D^n(\delta)$ be the event:

$$D^n(\delta) = \bigcap_{m \in I_n} D_m(\delta)$$

where we defined the interval $I_n = [\, n/l - n^{0.6} \,,\, n/l + n^{0.6} \,]$.

Lemma 4.3.1 *For every $\delta > 0$ we have that:*

$$\mathrm{P}(D^{nc}(\delta)) \leq 2n^{0.6} \left(\frac{1}{1+3\delta} \right)^{n\frac{(1+3\delta)}{2l}}$$

for n large enough.

Proof. Let $\mathcal{B}$ be the set of all possible blocks in X^m (or in Y^m), namely:

$$\mathcal{B} = \{b_{l-1}, b_l, b_{l+1}\}$$

where b_i denotes a block of length i. Let $W_1, W_2, \ldots$ be i.i.d. random variables on $\mathcal{B}$ with distribution:

$$p_i := \mathrm{P}(W_k = b_i) = 1/3 \qquad \text{every } i = l-1, l, l+1 \text{ and } k > 0.$$

Let $Z_1, Z_2, \ldots$ be new random variables in $\{0,1\}^3$ such that for $j > 0$ and $i = l-1, l, l+1$ the components are defined as follows:

$$\left(Z_j\right)_i := \begin{cases} 1 & \text{if } W_j = b_i \\ 0 & \text{otherwise} \end{cases}$$

Then we can see that $W_1, W_2, \ldots$ has the same distribution as $B_{X1}, B_{X2}, \ldots$ (or as $B_{Y1}, B_{Y2}, \ldots$) and then for $m > 0$ the expression:

$$\frac{Z_1 + \cdots + Z_m}{m}$$

is the empirical distribution of X^m (or of Y^m). Then for $\vec{q} = (q_1, q_2, q_3) \in \mathbb{R}^3_+$ we know that (from [13]; chapter 2):

$$\mathrm{P}\left(\frac{Z_1 + \cdots + Z_m}{m} = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix} \right) \leq r(\vec{q})^m \tag{4.3.1}$$

for every given $m > 0$, where

$$r(\vec{q}) := \frac{p_1{}^{q_1} p_2{}^{q_2} p_3{}^{q_3}}{q_1{}^{q_1} q_2{}^{q_2} q_3{}^{q_3}}. \tag{4.3.2}$$

Given $\delta > 0$, for $\vec{q_\delta} = (1/3+\delta, 1/3+\delta, 1/3+\delta)$ we get by replacing in 4.3.2:

$$r(\vec{q_\delta}) = \left(\frac{1}{1+3\delta}\right)^{1+3\delta}$$

and later by using 4.3.1:

$$\begin{aligned} \mathrm{P}(D^c_m(\delta)) &\leq P\left(\frac{Z_1 + \cdots + Z_m}{m} = \begin{pmatrix} 1/3+\delta \\ 1/3+\delta \\ 1/3+\delta \end{pmatrix}\right) \\ &\leq \left(\frac{1}{1+3\delta}\right)^{(1+3\delta)m} \end{aligned} \tag{4.3.3}$$

Finally by using 4.3.3 we can write down:

$$\begin{aligned} \mathrm{P}(D^{nc}(\delta)) \leq \sum_{m \in I_n} \left(\frac{1}{1+3\delta}\right)^{(1+3\delta)m} &\leq 2n^{0.6}\left(\frac{1}{1+3\delta}\right)^{(1+3\delta)(n/l - n^{0.6})} \\ &= 2n^{0.6}\left(\frac{1}{1+3\delta}\right)^{n\left(\frac{1}{l} - \frac{1}{n^{0.4}}\right)(1+3\delta)} \\ &\leq 2n^{0.6}\left(\frac{1}{1+3\delta}\right)^{\frac{n}{2l}(1+3\delta)} \end{aligned}$$

where the very last inequality is due to:

$$\frac{1}{l} - \frac{1}{n^{0.4}} \geq \frac{1}{2l}$$

for n large enough. ■

4.4 Number of blocks for an optimal alignment

Recall that N_n^X (resp. N_n^Y) is the number of blocks in X (resp. in Y) having lenghts in $\{l-1, l, l+1\}$ as in expression 4.1.1. Let $G^n(\delta)$ be the event that the following inequality holds:

$$\frac{N_n^Y}{N_n^X} \leq 1 + \delta.$$

Lemma 4.4.1 *For every $\delta > 0$ there exist two constantsl $b_4, b_5 > 0$ depending on l and on δ such that:*

$$\mathrm{P}(G^{nc}(\delta)) \leq 2e^{-b_4 \cdot n^{0.2}} + 2e^{-b_5 \cdot n^{0.2}}$$

for every n large enough.

Proof. Let us denote $\alpha := \frac{n}{l} + n^{0.6}$. Now computing directly we obtain:

$$\begin{aligned}
\mathrm{P}(G^{nc}(\delta)) &= \mathrm{P}(N_n^X(1+\delta) \leq N_n^Y) \\
&= \mathrm{P}(N_n^X(1+\delta) \leq N_n^Y \mid N_n^Y \leq \alpha)\mathrm{P}(N_n^Y \leq \alpha) \\
&+ \mathrm{P}(N_n^X(1+\delta) \leq N_n^Y \mid N_n^Y > \alpha)\mathrm{P}(N_n^Y > \alpha) \\
&\leq \mathrm{P}\left(N_n^X \leq \frac{\alpha}{1+\delta}\right)\mathrm{P}(N_n^Y \leq \alpha) + \mathrm{P}(N_n^Y > \alpha) \\
&\leq \mathrm{P}\left(N_n^X \leq \frac{\alpha}{1+\delta}\right) + \mathrm{P}(N_n^Y > \alpha) \qquad (4.4.1)
\end{aligned}$$

since $\mathrm{P}(N_n^Y \leq \alpha) \leq 1$. Let us work on each term as before. First, let us define $m_2 = \left\lfloor \frac{n}{l} + n^{0.6} \right\rfloor$ and write down:

$$\begin{aligned}
\mathrm{P}(N_n^Y > \alpha) &\leq \mathrm{P}(N_n^Y \geq m_2) \\
(\text{by using } N_t^Y \geq k \Leftrightarrow S_k^Y \leq t) &= \mathrm{P}\left(S_{m_2}^Y \leq n\right) \\
&= \mathrm{P}\left(\frac{S_{m_2}^Y}{m_2} - l \leq \frac{n}{m_2} - l\right) \\
(\text{by 4.0.2 with } \mathrm{P}(|B_{Y1} - l| \leq 1) = 1) &\leq 2\exp\left(-\frac{m_2}{2}\left(\frac{n}{m_2} - l\right)^2\right) \qquad (4.4.2)
\end{aligned}$$

Now we need to bound m_2 in order to get the right order for moderate deviations. Let us start looking at the following:

$$\begin{aligned}
\left(\frac{n}{m_2} - l\right)^2 &\geq l^2\left(\frac{n}{n + ln^{0.6}} - 1\right)^2, \text{ by using } m_2 \leq \tfrac{n}{l} + n^{0.6} \\
&\geq l^2\left(\frac{1}{1 + \frac{l}{n^{0.4}}} - 1\right)^2 \\
&\geq \frac{l^4}{n^{0.8}}\left(\frac{1}{1 + \frac{l}{n^{0.4}}}\right)^2 \\
&\geq \frac{l^4}{4n^{0.8}} \qquad (4.4.3)
\end{aligned}$$

where the very last inequality above holds for n large enough since:

$$\lim_{n\to\infty}\left(\frac{1}{1 + \frac{l}{n^{0.4}}}\right)^2 = 1 > \frac{1}{4}$$

Also, for $n > 0$ large enough we can take:

$$m_2 = \left\lfloor \frac{n}{l} + n^{0.6} \right\rfloor \geq \frac{n}{2l} \qquad (4.4.4)$$

Finally we can use 4.4.3, 4.4.4 in 4.4.2 to get:

$$\begin{aligned} \mathrm{P}\left(N_n^Y > \alpha\right) &\leq 2\exp\left(-\frac{m_2}{2}\left(\frac{n}{m_2} - l\right)^2\right) \\ &\leq 2\exp\left(-\frac{l^3}{16}\cdot n^{0.2}\right) \qquad (4.4.5) \end{aligned}$$

for $n > 0$ large enough. In the same way, calling $m = \left\lceil \frac{\alpha}{1+\delta} \right\rceil$ we have:

$$\begin{aligned} \mathrm{P}\left(N_n^X \leq \frac{\alpha}{1+\delta}\right) &\leq \mathrm{P}\left(N_n^X \leq m\right) \\ \text{(by using } N_t^X \geq k \Leftrightarrow S_k^X \leq t\text{)} &= \mathrm{P}\left(S_m^X \geq n\right) \\ &= \mathrm{P}\left(\frac{S_m^X}{m} - l \geq \frac{n}{m} - l\right) \\ \text{(by 4.0.2 with } \mathrm{P}(|B_{X1} - l| \leq 1) = 1\text{)} &\leq 2\exp\left(-\frac{m}{2}\left(\frac{n}{m} - l\right)^2\right) \qquad (4.4.6) \end{aligned}$$

Now we need again to bound m in order to get the right order for moderate deviations. Let us start looking at the following:

$$\begin{aligned} \left(\frac{n}{m} - l\right)^2 &\geq l^2\left(\frac{(1+\delta)n}{n + ln^{0.6}} - 1\right)^2, \text{ by using } m \leq \tfrac{1}{1+\delta}\left(\tfrac{n}{l} + n^{0.6}\right) \\ &\geq l^2\left(\frac{1+\delta}{1 + \frac{l}{n^{0.4}}} - 1\right)^2 \\ &\geq l^2\left(\frac{1+\delta}{1 + \frac{l}{n^{0.4}}} - (1+\delta)\right)^2 \\ &\geq \frac{l^4(1+\delta)^2}{n^{0.8}}\left(\frac{1}{1 + \frac{l}{n^{0.4}}}\right)^2 \\ &\geq \frac{l^4(1+\delta)^2}{4n^{0.8}} \qquad (4.4.7) \end{aligned}$$

where the very last inequality holds for n large enough since:

$$\lim_{n\to\infty}\left(\frac{1}{1 + \frac{l}{n^{0.4}}}\right)^2 = 1 > \frac{1}{4}$$

Also, for $n > 0$ large enough we can take:

$$m \geq \frac{1}{1+\delta}\left(1 + \frac{l}{n^{0.4}}\right)\frac{n}{l} \geq \frac{1}{2(1+\delta)}\cdot\frac{n}{l} \qquad (4.4.8)$$

Finally we can use 4.4.7, 4.4.8 in 4.4.6 to get:

$$\begin{aligned} \mathrm{P}\left(N_n^X \leq \frac{\alpha}{1+\delta}\right) &\leq 2\exp\left(-\frac{m}{2}\left(\frac{n}{m}-l\right)^2\right) \\ &\leq 2\exp\left(-\frac{l^3(1+\delta)}{8}\cdot n^{0.2}\right) \end{aligned} \tag{4.4.9}$$

for $n > 0$ large enough. Combining inequalities 4.4.5 and 4.4.9 in 4.4.1 we have:

$$\begin{aligned} \mathrm{P}(G^{nc}(\delta)) &\leq \mathrm{P}\left(N_n^X \leq \frac{\alpha}{1+\delta}\right) + \mathrm{P}(N_n^Y > \alpha) \\ &\leq 2\exp\left(-\frac{l^3}{16}\cdot n^{0.2}\right) + 2\exp\left(-\frac{l^3(1+\delta)}{8}\cdot n^{0.2}\right) \end{aligned} \tag{4.4.10}$$

from where we can take the constants

$$\begin{aligned} b_4 &= \frac{l^3}{16} > 0 \\ b_5 &= \frac{l^3(1+\delta)}{8} > 0 \end{aligned}$$

and finish the proof. ■

4.5 Cut blocks at the end

Let $J^n(\delta)$ denote the event that the proportion of left out blocks at the end of X or Y in any optimal alignment is at most a proportion δ of the total number of blocks in each of these sequences. As all events before, we want to prove that $J^n(\delta)$ has high probability to happen for every $\delta > 0$ provided n is large enough. We need an extra definition and a previous lemma in order to show the high probability of $J^n(\delta)$.

For an integer number $s \in [1, n]$ we denote:

$$L_1^s := |\mathrm{LCS}(X_1X_2\cdots X_s, Y_1Y_2\cdots Y_n)| \tag{4.5.1}$$

Lemma 4.5.1 *Given $\delta > 0$, there exists a constant $c_* > 0$ not depending on n but on δ such that:*

$$\mathrm{E}[L_n - L_1^{n-\delta n}] \geq c_* \cdot n \tag{4.5.2}$$

for every $n > 0$ large enough.

Proof. Given $n > 0$ and $t \in [-1, 1]$ let us define the number $\gamma(t, n) > 0$ as follows:

$$\gamma(t,n) := \frac{\mathrm{E}[\,|\mathrm{LCS}(X_1\cdots X_{n+nt}, Y_1\cdots Y_{n-nt})|\,]}{n}$$

This number $\gamma(t,n)$ is a kind of extension for the Chvatal-Sankoff constant γ (see [6]), or more precisely in the case of our paper an extension of γ_l defined as in expression 3.2.1. An extended motivation for this definition can be found in [21]. For any fixed $t \in [-1,1]$ it is known that $\gamma(t,n)$ converges as $n \to \infty$ (see [12] or [21]), let us denote that limit by

$$\gamma(t) := \lim_{n\to\infty} \gamma(t,n).$$

The speed of convergence to that limit is also known due to theorem 2.1 in [12]. This theorem says that there exists $\theta_1 > 0$ a constant not depending on n such that:

$$|\gamma(t,n) - \gamma(t)| \leq \frac{\theta_1 \ln(n)}{\sqrt{n}} \tag{4.5.3}$$

for any fixed $t \in [-1,1]$ provided $n > 0$ is large enough. On the other hand, it is known that the map $t \in [-1,1] \mapsto \gamma(t) \in [0,1]$ is concave and symmetric in the origin (see [21]). Hence, for every $t \in [-1,1]$ we have

$$\gamma(0) \geq \gamma(t) \tag{4.5.4}$$

Let us set an auxiliar variable n^* as follows:

$$n^* := n\left(1 - \frac{\delta}{2}\right)$$

Note that with the last definition, the inequality

$$\frac{n\ln(n)}{\sqrt{n}} = \sqrt{n}\,\ln(n) \geq \sqrt{n^*}\,\ln(n^*) = \frac{n^*\ln(n^*)}{\sqrt{n^*}} \tag{4.5.5}$$

holds due to $\sqrt{\cdot}$ and $\ln(\cdot)$ being increasing functions. By using the previous definitions, the inequality for the speed of convergence 4.5.3, the concave inequality 4.5.4 and inequality 4.5.5 (following this order), we can write:

$$\begin{aligned}
\mathrm{E}[L_n - L_1^{n-\delta n}] &= n\,\gamma(0,n) - n^*\gamma\left(t^*, n^*\right) \\
&\geq n\left(\gamma(0) - \frac{\theta_1\ln(n)}{\sqrt{n}}\right) - n^*\left(\gamma(t^*) + \frac{\theta_1\ln(n^*)}{\sqrt{n^*}}\right) \\
&\geq n\left(\gamma(0) - \frac{\theta_1\ln(n)}{\sqrt{n}}\right) - n^*\left(\gamma(0) + \frac{\theta_1\ln(n^*)}{\sqrt{n^*}}\right) \\
&= (n-n^*)\gamma(0) - \theta_1\left(\frac{n\ln(n)}{\sqrt{n}} + \frac{n^*\ln(n^*)}{\sqrt{n^*}}\right) \\
&\geq (n-n^*)\gamma(0) - 2\theta_1\frac{n\ln(n)}{\sqrt{n}} \\
&= \frac{n\,\delta\,\gamma(0)}{2} - 2\theta_1\frac{n\ln(n)}{\sqrt{n}} \\
&= \left(\frac{\delta\,\gamma(0)}{2} - 2\theta_1\frac{\ln(n)}{\sqrt{n}}\right)n \\
&\geq \frac{\delta\,\gamma(0)}{4}n
\end{aligned} \tag{4.5.6}$$

where the very last inequality above holds for n large enough, since

$$\lim_{n\to\infty}\left(2\theta_1\frac{\ln(n)}{\sqrt{n}}\right)=0<\frac{\delta\,\gamma(0)}{4}$$

To finish the proof we take $c_*=\frac{\delta\,\gamma(0)}{4}$. ■

Now comes the main result of this section which establishes the high probability of the event $J^n(\delta)$:

Proposition 4.5.1 *For every $\delta>0$, there exists a constant $\theta>0$ not depending on n but on δ such that:*

$$\mathrm{P}(J^{nc}(\delta))\le 2e^{-\theta\cdot n}$$

for every $n>0$ large enough.

Proof. With the notation as in 4.5.1 we write:

$$\begin{aligned}\mathrm{P}(J^{nc}(\delta)) &\le 2\,\mathrm{P}(\,|\mathrm{LCS}(X_1\cdots X_{n-\delta n},Y_1\cdots Y_n)|-L_n\ge 0\,)\\ &= 2\,\mathrm{P}(\,L_1^{n-\delta n}-L_n\ge 0\,)\\ &= 2\,\mathrm{P}(\,L_1^{n-\delta n}-L_n-\mathrm{E}[\,L_1^{n-\delta n}-L_n\,]\ge \mathrm{E}[\,L_n-L_1^{n-\delta n}\,]\,)\end{aligned} \tag{4.5.7}$$

Let us define

$$M_n(\delta):=L_1^{n-\delta n}-L_n-\mathrm{E}[\,L_1^{n-\delta n}-L_n\,]$$

It is not difficult to see that $M_n(\delta)$ is a martingale with respect to the filtration $\mathfrak{F}_n=\sigma\{(X_k,Y_k):k\le n\}$ and that $M_0=0$. The following inequality also holds:

$$|M_n(\delta)-M_{n-1}(\delta)|\le 4$$

for $\delta>0$ with probability 1. So, we can use the theorem 4.0.1 (Azuma-Hoeffding inequality for martingales) with $\mathfrak{a}_i=4$ and $v=\mathrm{E}[\,L_n-L_1^{n-\delta n}\,]$ to estimate:

$$\begin{aligned}\mathrm{P}(\,L_1^{n-\delta n}-L_n-\mathrm{E}[\,L_1^{n-\delta n}-L_n\,]\ge v\,) &\le 2\exp\left(-\frac{v^2}{2\cdot 4n}\right)\\ &= 2\exp\left(-\frac{n}{8}\left(\frac{\mathrm{E}[\,L_n-L_1^{n-\delta n}\,]}{n}\right)^2\right)\\ \text{(by 4.5.2 and } c_* \text{ from lemma 4.5.1)} &\le 2\exp\left(-\frac{c_*^2}{8}\cdot n\right)\end{aligned}$$

Taking $\theta=\frac{c_*^2}{8}>0$ finishes the proof. ■

4.6 Optimal events

In theorem 2.1.3, q represents the proportion of left out blocks in X and in Y. In reality, typically, the proportion of left out blocks in X will not be exactly equal to the proportion of left out blocks in Y. Because of this, q_1 will designate the proportion of left out blocks in X and q_2 will designate the proportion of left out blocks in Y. We will have that q_1 can be made as close to q_2 as we want to by taking a large n. Now we need to rewrite all our conditions as in theorem 2.1.3 with q_1 and q_2 instead of q.

Let us define the following events:

- Given any m_1, m_2, q_1, q_2, let $E_{m_1,m_2,q_1,q_2}(\epsilon)$ denote the event that there is no optimal alignment of X^{m_1} with Y^{m_2} leaving out a proportion of q_1 blocks in X^{m_1} and a proportion of q_2 blocks in Y^{m_2} and such that:

$$H(q_1)+H(q_2)+(1-\max\{q_1+3q_2, 3q_1+q_2\})\left(\ln(1/9)+H(p)\right) \leq -\epsilon. \tag{4.6.1}$$

- Let $E^n(\epsilon)$ be the event :

$$E^n(\epsilon) = \bigcap_{m_1,m_2\in I^n, q_1,q_2} E_{m_1,m_2,q_1,q_2}(\epsilon). \tag{4.6.2}$$

If δ designates the difference between q_1 and q_2, then note that the system

$$\begin{array}{c}
\min\left[\dfrac{p_{l-1,l}+p_{l-1,l+1}}{p_{l-1,l-1}+p_{l-1,l}+p_{l-1,l+1}}\left(1-\dfrac{3q_1}{(1/3)-\delta}\right)+\right. \\
\left. -\left(1-\dfrac{\delta+2(q_1-\delta)}{(1/3+\delta)}\right)\dfrac{p_{l+1,l+1}}{p_{l+1,l-1}+p_{l+1,l}+p_{l+1,l+1}} - q_2\cdot\dfrac{1+\delta}{(1/3)-\delta}\right]
\end{array}$$

$$\begin{aligned}
|q_1-q_2| &\leq 2\delta \\
H(q_1)+H(q_2)+(1-\max\{q_1+3q_2,3q_1+q_2\})\left(\ln(1/9)+H(p)\right) &\geq 0
\end{aligned} \tag{4.6.3}$$

converges to the conditions in theorem 2.1.3 when δ goes to zero (when q_1 is as close to q_2 as we want to by taking a large q_1). Note also that replacing q_1 and q_2 by q and taking $\delta = 0$ in the minimized function and in the last inequality of 4.6.3, they become equal to 2.1.19 respectively 2.1.22. If the minimizing problem in theorem 2.1.3 has a strictly positive solution 2ϵ and if expression 2.1.19 is less than ϵ_1, this implies that 2.1.22 must be smaller than a $-\epsilon_2$ for $\epsilon_2 > 0$ (we are assuming that 2.1.20, 2.1.21 and 2.1.22 hold). The next lemma shows that the same holds true for the system 4.6.3 if we take δ small enough.

Lemma 4.6.1 *Assume there exists* $0 < q_0 < (1/3)$ *and* $\epsilon_1 > 0$ *such that for all* $\{p_{ij}\}_{i,j}$ *and* $q \in [0, q_0]$ *satisfying all the conditions 2.1.20, 2.1.21 and 2.1.22 in theorem 2.1.3, we have that expression 2.1.19 is larger or equal to* $2\epsilon_1$ *(in other words, the condition that the minimizing problem in theorem 2.1.3 has a strictly positive solution* $2\epsilon_1$ *is satisfied). Then, we have that there exists* $\epsilon_2 > 0$ *and* $\delta_0 > 0$ *such that for all* $\{p_{ij}\}_{i,j\in\{l-1,l,l+1\}}$ *and* $q_1, q_2 \in [0, q_0]$ *and* $\delta \in [0, \delta_0]$ *satisfying 2.1.20 and 2.1.21, we have that if* $|q_1 - q_2| \leq 2\delta_0$ *and if*

$$\frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}}\left(1 - \frac{3q_1}{(1/3) - \delta_0}\right) \quad + \\ -\left(1 - \frac{\delta_0 + 2(q_1 - \delta_0)}{(1/3 + \delta_0)}\right)\frac{p_{l+1,l+1}}{p_{l+1,l-1} + p_{l+1,l} + p_{l+1,l+1}} - q_2 \cdot \frac{1 + \delta_0}{(1/3) - \delta_0} \leq \epsilon_1 \tag{4.6.4}$$

then

$$H(q_1) + H(q_2) + (1 - \max\{q_1 + 3q_2, 3q_1 + q_2\})\,(\ln(1/9) + H(p)) \leq -\epsilon_2 \tag{4.6.5}$$

Proof. We are going to do the proof by *reductio ad absurdum* (reduction to the absurd). Assume for this that for all $(p_{ij})_{i,j\in I}$ and $q \in [0, q_0]$ satisfying all the conditions 2.1.20, 2.1.21 and 2.1.22 in theorem 2.1.3 we have that expression 2.1.19 is larger equal to $2\epsilon_1$. Assume that the rest of the lemma would not hold. Then for every $\delta > 0$ (as small as we want) we could find a vector $\vec{p}$:

$$\vec{p} := (p_{l-1,l-1}, p_{l-1,l}, \ldots, p_{l+1,l+1}, q_1, q_2, \delta)$$

such that the components satisfy $|q_1 - q_2| \leq \delta$, and the components of $\vec{p}$ satisfy 2.1.20, 2.1.21 whilst inequality 4.6.4 is satisfied and we can take the expression

$$H(q_1) + H(q_2) + (1 - \max\{q_1 + 3q_2, 3q_1 + q_2\})\,(\ln(1/9) + H(p)) \tag{4.6.6}$$

as close to zero as we want. Hence there exists a sequence $\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_t, \ldots$ of such vectors with notation:

$$\vec{p}(t) := (p_{l-1,l-1}(t), p_{l-1,l}(t), \ldots, p_{l+1,l+1}(t), q_1(t), q_2(t), \delta(t))$$

so that for each $t \in \mathbb{N}$ the vector $\vec{p}(t)$ satisfies all the conditions 2.1.20, 2.1.21 and 4.6.4, whilst

$$\lim_{t\to\infty} |q_1(t) - q_2(t)| = 0$$

and expression 4.6.6 converges to zero as t goes to infinity.

The vectors $\vec{p}(t)$ are contained in a bounded domain and hence in a compact domain. This implies that there exists a converging subsequence. Hence there exists an increasing map $\pi : \mathbb{N} \to \mathbb{N}$ so that $\vec{p}(\pi(t))$ converges as t goes to infinity. Let the limit be denoted by

$$\vec{p} := (p_{l-1,l-1}, p_{l-1,l}, \ldots, p_{l+1,l+1}, q_1, q_2, 0).$$

We have that $q_1 = q_2$, so let us denote $q_1 = q_2 = q$. We find that our limit satisfies all the conditions 2.1.20, 2.1.21. Furthermore, at the limit expression 4.6.6 becomes equal to zero. Replacing then q_1 and q_2 in 4.6.6 by q, we find that condition 2.1.22 is satisfied. finally since for our sequence $\vec{p}(\pi(t))$, we have that 4.6.4 is satisfied, by continuity it must also be satisfied for the limit. Hence, noting that at the limit $q_1 = q_2 = q$ and $\delta = 0$, we get that expression 2.1.19 is less or equal to ϵ_1. This contradicts our assumption, since our limit vector satisfies all the conditions2.1.20, 2.1.21 and 2.1.22 and should thus have expression 2.1.19 larger equal to $2\epsilon_1$. Hence, we have that when all the conditions 2.1.20 and 2.1.21, are satisfied and when δ goes to zero then expression 4.6.4 should be bounded away from zero. This means that for $\delta > 0$ small enough and when all the conditions 2.1.20 and 2.1.21 are satisfied, we have that there exists $\epsilon_2 > 0$ so that 4.6.4 is less or equal to $-\epsilon_2$. ■

Let us now show that event $E_{m_1,m_2,q_1,q_2}(\epsilon)$ holds with high probability.

Lemma 4.6.2 *Assume that there exists* $0 < q_0 < (1/3)$ *and* $\epsilon_1 > 0$ *such that for all* $\{p_{ij}\}_{i,j}$ *and* $q \in [0, q_0]$ *satisfying all the conditions 2.1.20, 2.1.21 and 2.1.22 in theorem 2.1.3, we have that expression 2.1.19 is larger or equal to* $2\epsilon_1$. *Then, for every* $\epsilon > 0$ *there exist a polynomial* $w(n) > 0$ *and a constant* $\vartheta > 0$, *both only depending on* l *such that:*

$$\mathrm{P}(E^{nc}(\epsilon)) \leq w(n)e^{-\vartheta \cdot n}$$

for every n *large enough.*

Proof. Let $\vec{a}$ denote an alignment of the X^{m_1} and Y^{m_2}. Hence a consists of two binary vectors $\vec{a} = (\vec{a}_X, \vec{a}_Y)$ the first one having length m_1 and the second one having length m_2.

Hence $\vec{a}_x \in \{0,1\}^{m_1}, \vec{a}_Y \in \{0,1\}^{m_2}$ when the i-th entry a_{Xi} of $\vec{a}_X$ is a 1 that means that the i-th block of X_{m_1} is discarded (entirely aligned with gaps)by the alignment $\vec{a}$, otherwise the i-th block of X^{m_1} is not discarded. Similarly when $a_{Yi} = 1$ then the i-th block of Y^{m_1} is discarded. Here we use the same way of defining alignment as explained before in the first section: we specify which blocks get entirely discarded and then align the rest block by block. Doing so and assuming that the alignment $\vec{a}$ is not random, we get that the aligned block pairs are i.i.d.. For the lengths of aligned block pairs we have nine possibilities each having the same probability. Hence, given the alignment a, the empirical frequencies of the aligned block pair lengths is simply a multinomial distribution. Let $p = \{p_{ij}\}_{i,j\in\{l-1,l,l+1\}}$ be a (non-random) probability distribution. Let $E_a(p)$ denote the event that the empirical distribution of the aligned block pairs by the alignment a is not p.

From what we said we have that the probability $\mathrm{P}(E_a^c(p))$ is equal to the probability that a 9-nomial variable with parameter m^* and all probability parameters equal to 1/9 gives the frequencies given by p. Here m^* designates the number of aligned block pairs by a. Hence, we get

$$\mathrm{P}(E_a^c(p)) = \binom{m^*}{m^*p_{l-1,l-1} \;\; m^*p_{l-1,l} \;\; \ldots \;\; m^*p_{l+1,l+1}} \left(\frac{1}{9}\right)^{m^*} \tag{4.6.7}$$

where

$$\binom{a}{a_1 \ldots a_k} = \frac{a!}{a_1! \cdots a_k!}$$

is the multinomial factorial coefficient. Let us define

$$\begin{aligned}
B(p) &:= \binom{m^*}{p_{l-1,l-1}m^* \;\; p_{l-1,l}m^* \;\; \ldots \;\; p_{l+1,l+1}m^*} \\
M(p) &:= \prod_{p_i \in \{p_{l-1,l-1},\ldots,p_{l+1,l+1}\}} p_i^{p_i} \\
H(p) &:= \sum_{p_i \in \{p_{l-1,l-1},\ldots,p_{l+1,l+1}\}} p_i \ln(1/p_i) \;=\; \ln\left(\frac{1}{M(p)}\right)
\end{aligned} \tag{4.6.8}$$

note that $B(p) \cdot (M(p))^{m^*}$ is the probability distribution of a multinomial random variable with parameters m^* and vector $(m^*p_{l-1,l-1}, \ldots, m^*p_{l+1,l+1})$. Hence

$$B(p) \cdot (M(p))^{m^*} \leq 1. \tag{4.6.9}$$

Then, by using 4.6.9 we can bound expression 4.6.7 as follows:

$$\begin{aligned}
B(p)\left(\frac{1}{9}\right)^{m^*} &= B(p) \cdot (M(p))^{m^*} \left(\frac{1/9}{M(p)}\right)^{m^*} \\
&\leq \left(\frac{1/9}{M(p)}\right)^{m^*} \\
&= \exp\left(\left[\ln\left(\frac{1}{9}\right) + H(p)\right] m^*\right)
\end{aligned} \tag{4.6.10}$$

On the other hand, we have at least $(1 - \max\{q_1 + 3q_2, 3q_1 + q_2\}) \min\{m_1, m_2\}$ aligned block pairs. Let us give an intuition for this. There are three situations for aligning a fixed block in X with blocks in Y. First, when we align one block in X with one block in Y one to one, the resulting length contributing to the LCS is the minimun between their lenghts, so at most if all the blocks of X and Y are aligned one to one then we will have at most a contribution of $\min\{m_1, m_2\}$ aligned blocks pairs. Second, when we align one block in X with several blocks in Y then we at least leave $q_1 \cdot m_1$ blocks in X. Third, when know that we cannot align two adjacent blocks in X with the same block in Y, then we leave at least $2q_1 \cdot m_1$ blocks in X also. In total, in the worse case, looking first at blocks in X, we are leaving $(3q_1 + q_2) \min\{m_1, m_2\}$ blocks in both sequences X and Y. Similarly, but

looking first at Y, we can leave $(3q_2 + q_1)\min\{m_1, m_2\}$ blocks in both sequences X and Y. Finally, at least we have $(1 - \max\{q_1 + 3q_2, 3q_1 + q_2\})\min\{m_1, m_2\}$ aligned block pairs due to the considerations above.
Since $m_1, m_2 \in I_n$, this gives the lower bound for m^*

$$m^* \geq (1 - \max\{q_1 + 3q_2, 3q_1 + q_2\}) \cdot ((n/l) - n^{0.6}). \tag{4.6.11}$$

and hence together with the bound 4.6.10, we obtain

$$\mathrm{P}((E_a^c(p)) \leq \exp\big((\ln(1/9) + H(p))(1 - \max\{q_1 + 3q_2, 3q_1 + q_2\})((n/l) - n^{0.6}) \big) \tag{4.6.12}$$

Let $\mathcal{A}_{m_1,m_2,q_1,q_2}$ denote the set of all alignments aligning X^{m_1} with Y^{m_2} and leaving out a proportion of q_1 blocks in X^{m_1} and a proportion of q_2 blocks in Y^{m_2}. In other words, the set $\mathcal{A}_{m_1,m_2,q_1,q_2}$ is the set of all elements $\vec{a} = (\vec{a}_X, \vec{a}_Y)$ of $\{0,1\}^{m_1} \times \{0,1\}^{m_2}$ for which $|\vec{a}_X| = q_1 m_1$ and $|\vec{a}_Y| = q_2 m_2$.
Let $\mathcal{P}_{\epsilon,q_1,q_2}$ denote the set of those distributions p (for aligned block pairs, hence on the space $\Omega = \{(l-1, l-1), (l-1, l), \ldots, (l+1, l+1)\}$) for which inequality 4.6.1 is satisfied and which are possible in our case. Before we continue with the proof, let us look at an example:

Example 4.6.1 *Assume we look at binary strings of length* 5*. Then there can be 0,1,2,3,4 or 5 ones. Hence, the empirical distribution for side one when we flip a coin exactly five times can only be* 0*,* 20%*,* 40%*,* 60%*,* 80% *or* 100%*. In general for a string of length n and k symbols, there are no more than* $(n+1)^{k-1}$ *possible empirical distributions (see [13], Lemma 2.1.2 (a)). In the case above we have an empirical distribution for* m^* *aligned block pairs. For each block pairs there are* 9 *possibilities. Hence, there are no more than* $(m^* + 1)^8$ *possible empirical distributions. However* m^* *is not known. It could potentially take on any value between* 1 *and* $(n/l) + n^{0.6}$*. Hence, we find that for the number of empirical distributions we need to consider the following upper bound:*

$$((n/l) + n^{0.6}) \cdot ((n/l) + n^{0.6} + 1)^8 \leq ((n/l) + n^{0.6} + 1)^9.$$

Let us continue with the proof. We have that:

$$\bigcap_{a \in \mathcal{A}_{m_1,m_2,q_1,q_2}, \mathcal{P}_{\epsilon,q_1,q_2}} E_a(p) = E_{m_1,m_2,q_1,q_2}(\epsilon)$$

and hence:

$$\mathrm{P}(E^c_{m_1,m_2,q_1,q_2}(\epsilon)) \leq \sum_{a \in \mathcal{A}_{m_1,m_2,q_1,q_2}, \, p \in \mathcal{P}_{\epsilon,q_1,q_2}} \mathrm{P}(E_a^c(p)). \tag{4.6.13}$$

By using 4.6.12, the inequality 4.6.13 above becomes:

$$\mathrm{P}(E^c_{m_1,m_2,q_1,q_2}(\epsilon)) \leq \sum_{a \in \mathcal{A}_{m_1,m_2,q_1,q_2}, \, p \in \mathcal{P}_{\epsilon,q_1,q_2}} \exp\big((\ln(1/9) + H(p))(1 - \max\{q_1 + 3q_2, 3q_1 + q_2\})((n/l) - n^{0.6}) \big) .$$

Note that the number of alignment considered in the sum on the right hand side of the last inequality above can be bound as follows:

$$\begin{aligned}
|\mathcal{A}_{m_1,m_2,q_1,q_2}| &= \binom{m_1}{q_1m_1(1-q_1)m_1}\binom{m_2}{q_2m_2(1-q_2)m_2} \\
&\leq \left(\frac{1}{q_1^{q_1}(1-q_1)^{1-q_1}}\right)^{m_1}\left(\frac{1}{q_2^{q_2}(1-q_2)^{1-q_2}}\right)^{m_2} \\
&= \exp(H(q_1)m_1 + H(q_2)m_2) \\
&\leq \exp((H(q_1)+H(q_2))m^*) \\
&\leq \exp\big(\,(H(q_1)+H(q_2))((n/l)+n^{0.6})\,\big) \qquad (4.6.14)
\end{aligned}$$

where for $i = 1, 2$ we denote

$$H(q_i) := q_i \ln(1/q_i) + (1-q_i)\ln(1/(1-q_i)).$$

The number of distributions in $\mathcal{P}_{\epsilon,q_1,q_2}$ we need to consider is (as explained above) less or equal to $((n/l)+n^{0.6}+1)^9$. Combining all of this we find that $\mathrm{P}(E^c_{m_1,m_2,q_1,q_2}(\epsilon))$ is less or equal to:

$$\exp((H(q_1)+H(q_2))((n/l)+n^{0.6}))\cdot b\cdot\exp\big(\,(\ln(1/9)+H(p))(1-\max\{q_1+3q_2,3q_1+q_2\})((n/l)-n^{0.6})\,\big)\,.$$

where $b := ((n/l)+n^{0.6}+1)^9$. In other words, we found that:

$$\mathrm{P}(E^c_{m_1,m_2,q_1,q_2}(\epsilon)) \leq b\exp\Big(\frac{n}{l}(H(q_1)+H(q_2)+(\ln(1/9)+H(p))(1-\max\{q_1+3q_2,3q_1+q_2\})+r)\,\Big) \qquad (4.6.15)$$

where the rest term r is equal to:

$$r = ln^{-0.4}\,(H(q_1)+H(q_2)-(\ln(1/9)+H(p))(1-\max\{q_1+3q_2,3q_1+q_2\}))$$

being bounded as follows:

$$|r| \leq ln^{-0.4}(|H(q_1)|+|H(q_2)|+(|\ln(1/9)|+|H(p)|)) = ln^{-0.4}(3+|\ln(1/9)|).$$

Note that r is bounded from above by a constant times $n^{-0.4}$ where the constant does not depend on l, q_1, q_2, p. Hence for n large enough:

$$r \leq \epsilon/2 \qquad (4.6.16)$$

Note also that in the sum 4.6.13, we only took distributions $p \in \mathcal{P}_{\epsilon,q_1,q_2}$ hence satisfying inequality 4.6.1. This implies that in the bound 4.6.15, we can assume that inequality 4.6.1 holds. This then implies

$$\mathrm{P}(E^c_{m_1,m_2,q_1,q_2}(\epsilon)) \leq b\exp\Big(\frac{n}{l}(-\epsilon+r)\,\Big) \qquad (4.6.17)$$

Assuming now that 4.6.16 holds, we obtain:

$$\mathrm{P}(E^c_{m_1,m_2,q_1,q_2}(\epsilon)) \leq b\exp\left(\frac{n}{l}(-\epsilon/2)\right) \tag{4.6.18}$$

Note that the bound on the right side of the last inequality above is negatively exponentially small in n, since b is an expression which is only polynomial in n. Using the equation 4.6.2, we obtain:

$$\mathrm{P}(E^{nc}(\epsilon)) \leq \sum_{m_1,m_2\in I^n,q_1,q_2} \mathrm{P}(E^c_{m_1,m_2,q_1,q_2}(\epsilon)).$$

Applying inequality 4.6.18 to the last inequality above, we obtain:

$$\mathrm{P}(E^{nc}(\epsilon)) \leq \sum_{m_1,m_2\in I^n,q_1,q_2} b\exp\left(\frac{n}{l}(-\epsilon/2)\right). \tag{4.6.19}$$

Note that when m_1 is given, the number of possibilities for the number of left out blocks in X^{m_1} is at most m_1. Hence, for given m_1 we have that q_1 can take on at most m_1 values. Similarly for given m_2 we have that q_2 can take on at most m_2 values. But m_1 and m_2 are less then $(n/l)+n^{0.6}$. Also, both m_1 and m_2 are in I_n hence they can take on at most $2n^{0.6}$ values. This implies that in the sum 4.6.19, the number of terms is bound above by the expression:

$$\left((n/l)+n^{0.6}\right)^2 4n^{1.2}$$

This upper bound applied to inequality 4.6.19 yields:

$$\mathrm{P}(E^{nc}(\epsilon)) \leq b\left((n/l)+n^{0.6}\right)^2 4n^{1.2}\exp\left(\frac{n}{l}(-\epsilon/2)\right). \tag{4.6.20}$$

which is the negative exponential upper bound we where looking for. ■

4.7 Positive expected change in the score

Let us recall the events that we have proven to have high probability:

- C^n is the event that the number of blocks in X and in Y lies in the interval
$$I_n = \left[\frac{n}{l} - n^{0.6}, \frac{n}{l} + n^{0.6}\right].$$
- $D^n(\delta)$ is the intersection
$$D^n(\delta) = \bigcap_{m\in I_n} D_m(\delta),$$
where $D_m(\delta)$ is the event that the proportion of blocks in X^m and in Y^m of length $l-1, l$ and $l+1$ are not further from $1/3$ than δ, where X^m (resp. Y^m) denotes the sequence X^∞ taken up to the m-th block (resp. the sequence Y^∞ taken up to the m-th block).

- $F^n(q)$ is the event that any optimal alignment of X and Y leaves out at most a proportion q of blocks in X as well as in Y.
- $G^n(\delta)$ is the event that the following inequality holds:

  $$\frac{N_n^Y}{N_n^X} \leq 1+\delta$$

  where N_n^X (resp. N_n^Y) is the number of blocks in X (resp. in Y) having length in $\{l-1, l, l+1\}$.
- $E^n(\epsilon)$ is the intersection

  $$E^n(\epsilon) = \bigcap_{m_1,m_2 \in I_n\,;\, q_1,q_2 \in [0,1]} E_{m_1,m_2,q_1,q_2}(\epsilon)$$

  where $E_{m_1,m_2,q_1,q_2}(\epsilon)$ is the event that there is no optimal alignment of X^{m_1} with Y^{m_2} leaving out a proportion of q_1 blocks in X^{m_1} and a proportion of q_2 blocks in Y^{m_2} and such that:

  $$H(q_1) + H(q_2) + (1 - \max\{q_1 + 3q_2, q_2, 3q_1\})(\ln(1/9) + H(p)) \leq -\epsilon_2$$

  where $\epsilon_2 > 0$ depends on ϵ, δ_0 and q_0 and comes from lemma 4.6.1, X^{m_1} (resp. Y^{m_2}) denotes the sequence X^∞ taken up to the m_1-th block (resp. the sequence Y^∞ taken up to the m_2-th block) and $H(p)$ denotes the entropy as in 4.6.8 for an alignment.

We can now formulate our combinatorial lemma based on those events:

Lemma 4.7.1 *Let us consider the constants $q_0, \epsilon_1, \delta_0$ and ϵ_2 from lemma 4.6.1. Assume that C^n, $D^n(\delta_0)$, $F^n(q_0)$, $G^n(\delta_0)$ and $E^n(\epsilon_2)$ all hold. Then, we have that*

$$\mathrm{E}[\tilde{L}_n - L_n | X, Y] \geq \epsilon_1$$

Proof. For any $x, y \in \{0,1\}^n$ let $L(x,y)$ denote the length of the LCS of x and y. Let now $x, y \in \{0,1\}$ be any two realizations so that if $X = x$ and $Y = y$, then the events C^n, $D^n(\delta_0)$, $F^n(q_0)$ and $E^n(\epsilon_2)$ all hold. Let a be a left most optimal alignment of x and y. Let $\tilde{x}$ denote the sequence x on which we performed our random changes. That is $\tilde{x}$ is obtained by selecting a block of length $l-1$ at random and changing it to length l and also selecting a block of length $l+1$ at random and reducing it to length l. Let x^* be the sequence we obtain by applying to x only the first one of the two random changes. That is x^* is obtained be increasing the length of a randomly chosen block of x of length $l-1$ to length l. So, we start with x. Then we apply the first change and obtain x^*. Then in x^* we choose a block of length $l+1$ at random, decrease it by one unit to obtain $\tilde{x}$.

For all $i, j \in \{l-1, l, l+1\}$, let p_{ij} denote the proportion of aligned block pairs with lengths (i, j) in the alignment a of x and y. Let q_1, resp. q_2 denote the proportion of blocks not aligned by a in x, resp. in y. Let p^I_{l-1} denote the proportion of blocks which get aligned by a one block to one block, among all blocks of x of length $1-1$. Let p^{II}_{l-1} denote the proportion among all blocks of x of length $l-1$ which are aligned with several blocks of y. Finally, let p^{III}_{l-1} denote the proportion among the blocks of length $l-1$ in x which are left out or are together with other blocks of x aligned with the same block of y. Note that when we increase by one unit a block in this third category, then in general the score does not get any increase. On the other hand, assume that the block of x length $l-1$ chosen randomly and increased by one unit, is aligned one block with one block. Then if this chosen block is aligned with a block of length l or $l+1$ the score is going to increase. Let $G_{l-1,I}$ be the event that the block of length $l-1$ chosen is aligned one block with one block. From what we said it follows that:

$$\mathrm{P}(L(x^*, y) - L(x, y) = 1 \mid G_{l-1,I}) \geq \frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}}.$$

Note that by only adding a bit the score cannot decrease, so that the last inequality above means:

$$\mathrm{E}[L(x^*, y) - L(x, y) \mid G_{l-1,I}] \geq \frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}}. \tag{4.7.1}$$

When the block of length $l-1$ chosen and increased is aligned with several blocks of y at the same time, then we will always observe and increase of one unit. This yields:

$$\mathrm{E}[L(x^*, y) - L(x, y) \mid G_{l-1,II}] = 1, \tag{4.7.2}$$

where $G_{l-1,II}$ denotes the event that the chosen block of length $l-1$ is aligned with several blocks of y. By law of total probability we find thus:

$$\begin{aligned} \mathrm{E}[L(x^*, y) - L(x, y)] &\geq \mathrm{P}(G_{l-1,I}) \frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}} + \mathrm{P}(G_{l-1,II}) \\ &\geq (1 - \mathrm{P}(G_{l-1,III})) \frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}} \end{aligned}$$

where $G_{l-1,III}$ denotes the event that the block of length $l-1$ chosen is left out or aligned to the same block of y at the same time as other blocks of x. The last inequality above yields:

$$\mathrm{E}[L(x^*, y) - L(x, y)] \geq (1 - p^{III}_{l-1}) \frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}}. \tag{4.7.3}$$

Note that the proportion of left out blocks in x is q_1. There can not be two adjacent blocks of x aligned with the same block of y (this is so because a is an optimal left most alignment, see lemma 3.1.1). So between blocks of x aligned with the same block of y, there is at least one left out block of x. Hence the maximum proportion of blocks of x, which are aligned at the same time as other blocks of x to the same block of y, can not exceed twice the number of left out blocks of x. This yields a lower bound equal to $2q_1$. This is as a proportion among all blocks in x, but we are interested in the number as a proportion of the total number of blocks of length $l-1$ of x. So, we get as lower bound $2q_1/p_{l-1}$, where p_{l-1} is the proportion of blocks of x which have length $l-1$. Adding the blocks in x which are left out and the blocks which are aligned with several blocks of y, we get:

$$p_{l-1}^{III} \leq \frac{3q_1}{p_{l-1}}. \tag{4.7.4}$$

By $D^n(\delta_0)$, we have that $p_{l-1} \geq (1/3) - \delta_0$, so that together with 4.7.4, we obtain:

$$p_{l-1}^{III} \leq \frac{3q_1}{(1/3) - \delta_0}.$$

By using the above inequality in 4.7.3 we obtain:

$$\mathrm{E}[L(x^*, y) - L(x, y)] \geq \frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}} \left(1 - \frac{3q_1}{(1/3) - \delta_0}\right). \tag{4.7.5}$$

Next we are going to investigate the effect of decreasing a randomly chosen block of length $l+1$ by one unit. The score can decrease when the selected block of x of length $l+1$ is aligned with a block of length $l+1$ of y. If it is aligned with one block and that block has length l or $l-1$, then there is no decrease. This leads to:

$$\mathrm{E}[L(\tilde{x}, y) - L(x^*, y) | G_{l+1,I}] \geq -\frac{p_{l+1,l+1}}{p_{l+1,l-1} + p_{l+1,l} + p_{l+1,l+1}},$$

where $G_{l+1,I}$ denotes the event that the block of length $l+1$ chosen is aligned one block with one block. When the selected block of x of length $l+1$ is aligned with several blocks of y then the score decreases by one unit. When the selected block of length $l+1$ in x is left out or is aligned at the same time as other blocks of x to the same block of y then there is no decrease. This leads to:

$$\mathrm{E}[L(\tilde{x}, y) - L(x^*, y)] \geq -\mathrm{P}(G_{l+1,I}) \frac{p_{l+1,l+1}}{p_{l+1,l-1} + p_{l+1,l} + p_{l+1,l+1}} - \mathrm{P}(G_{l+1,II}), \tag{4.7.6}$$

where $G_{l+1,II}$ denotes the event that the selected block of length $l+1$ is aligned with several blocks of y at the same time. Let p_{l+1} denote the total proportion of blocks of length $l+1$ among all blocks of x. Let $p_{l+1,I}$ denote the proportion among all blocks of x of length $l+1$, of blocks which are aligned one to one. There is a proportion of q_1 totally left out blocks in x. At most a proportion δ_0 are at the end of the alignment a contiguous group of left out blocks. That means, (assuming $q_1 \geq \delta_0$), the proportion of left out blocks in x which are not adjacent to another left out block of x is at least $q_1 - \delta_0$. Going with each left out block which is not adjacent to another left out block, there is at least one adjacent block which is aligned together with several other blocks of x to the same block of y. This gives a lower bound for the blocks of x which are not aligned one block to one block of $\delta_0 + 2(q_1 - \delta_0)$. This is taken as a proportion among all blocks of x. This gives among all blocks of length $l + 1$ a proportion of at least:

$$\frac{\delta_0 + 2(q_1 - \delta_0)}{(1/3 + \delta_0)},$$

since by the event $D^n(\delta_0)$ we know that among all blocks of x the proportion of the blocks of length $l + 1$ is less than $(1/3) + \delta_0$. Hence,

$$P(G_{l+1,I}) \leq 1 - \frac{\delta_0 + 2(q_1 - \delta_0)}{(1/3 + \delta_0)}. \tag{4.7.7}$$

Next let us note that we can give an upper bound for the number of blocks of x aligned with several blocks of y. Since we never have several blocks aligned with several blocks, we have that the number of blocks of x aligned with several blocks of y is not more than the total number of left out blocks of y. This is so because between two blocks aligned with the same block there is always at least one left out block. The proportion of left out blocks in y is q_2. but this is taken as proportion among all the blocks of y. Since the total amount of blocks in x and y could not be exactly the same, that number can get slightly changed when we report it as proportion of the total number of blocks in x. Let p_{l+1} denote the proportion among the blocks of x which are of length $l + 1$. We have thus that the probability to select a block of length $l + 1$ of x which is aligned with several blocks of y is less or equal to

$$P(G_{l+1,II}) \leq \frac{q_2 N_n^Y}{p_{l+1} N_n^X}. \tag{4.7.8}$$

By the event $D^n(\delta_0)$ we have

$$p_{l+1} \geq \frac{1}{3} - \delta_0 \tag{4.7.9}$$

and by the event $G^n(\delta_0)$ we have

$$\frac{N_n^Y}{N_n^X} \leq 1 + \delta_0. \tag{4.7.10}$$

Applying now 4.7.9 and 4.7.10 to 4.7.8, we find

$$P(G_{l+1,II}) \leq q_2 \cdot \frac{1 + \delta_0}{(1/3) - \delta_0} \tag{4.7.11}$$

Finally, using inequalities 4.7.11, 4.7.7 in 4.7.6 we get:

$$\begin{aligned} \mathrm{E}[L(\tilde{x},y)-L(x^*,y)] \;\; &\geq \;\; -\left(1-\frac{\delta_0+2(q_1-\delta_0)}{(1/3+\delta_0)}\right)\frac{p_{l+1,l+1}}{p_{l+1,l-1}+p_{l+1,l}+p_{l+1,l+1}} \\ & \quad -q_2\cdot\frac{1+\delta_0}{(1/3)-\delta_0}, \qquad (4.7.12) \end{aligned}$$

Using inequalities 4.7.5 and 4.7.12 together we find:

$$\begin{aligned} \mathrm{E}[L(\tilde{x},y)-L(x,y)] \;\; &\geq \;\; \mathrm{E}[L(\tilde{x},y)-L(x^*,y)]+\mathrm{E}[L(x^*,y)-L(x,y)] \\ &\geq \;\; \frac{p_{l-1,l}+p_{l-1,l+1}}{p_{l-1,l-1}+p_{l-1,l}+p_{l-1,l+1}}\left(1-\frac{3q_1}{(1/3)-\delta_0}\right) \\ &- \;\; \left(1-\frac{\delta_0+2(q_1-\delta_0)}{(1/3+\delta_0)}\right)\frac{p_{l+1,l+1}}{p_{l+1,l-1}+p_{l+1,l}+p_{l+1,l+1}} \\ & \quad -q_2\cdot\frac{1+\delta_0}{(1/3)-\delta_0}, \qquad (4.7.13) \end{aligned}$$

Note next that we can apply lemma 3.1.2 with $\Delta = n^{0.6}$ because of C^n and $\delta_1, \delta_2 \leq \delta_0$ thanks to $E^n(\delta_0)$. Hence we find that:

$$|q_1-q_2| \leq 1.5|\delta_0| + \frac{4ln^{0.6}}{n}$$

We assume that n is large enough so that:

$$|q_1-q_2| \leq 2|\delta_0|.$$

With the last inequality holding, we get from lemma 4.6.1 that if inequality 4.6.4 holds, then 4.6.5 should be satisfied. By the event $E^n(\epsilon_2)$, we have that 4.6.5 can not be satisfied. Hence, the inequality 4.6.4 cannot hold, which implies that the expression on the left side of 4.6.4 is larger or equal to ϵ_1. Together with inequality 4.7.13, this implies that:

$$\mathrm{E}[L(\tilde{x},y)-L(x,y)] \geq \epsilon_1. \quad \blacksquare$$

Chapter 5

Random modification & fluctuation

In this chapter, we show that the biased effect of our random modification implies the fluctuation order $\Theta(n)$. In other words, we are going to prove theorem 2.1.2 which states that $\mathrm{VAR}[L_n] = \Theta(n)$ holds if there exist $\epsilon, \alpha > 0$ not depending on n such that:

$$\mathrm{P}\left(\ \mathrm{E}[\tilde{L}_n - L_n | X, Y] \geq \epsilon\ \right) \geq 1 - \exp(-n^{\alpha}). \tag{5.0.1}$$

for all n large enough. Note that if $\mathcal{Z}$ is a random variable with $\mathrm{VAR}[\mathcal{Z}] = \Theta(n)$ and f is a map which tends to increase linearly, then for $\mathcal{W} = f(\mathcal{Z})$, we also have the order $\mathrm{VAR}[\mathcal{W}] = \Theta(n)$. The map f can be even a random map but must be independent of $\mathcal{Z}$. The exact basic result ([16], lemma 3.2) goes as follows:

Lemma 5.0.2 *Let $c > 0$ be a constant. Assume that $g : \mathbb{R} \to \mathbb{R}$ is a map which is everywhere differentiable and such that for all $x \in \mathbb{R}$ we have:*

$$\frac{dg(x)}{dx} \geq c.$$

Let B be a random variable such that $\mathrm{E}[|g(B)|] < +\infty$. Then:

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

In the present context, we need a slightly different version:

Lemma 5.0.3 *Let $\epsilon, m > 0$ be constants and $f : \mathbb{Z} \to \mathbb{Z}$ be a map such that for all $z_1 \leq z_2$ the following two conditions hold:*

$$z_2 - z_1 \geq m \Rightarrow f(z_2) - f(z_1) \geq \frac{\epsilon}{8}(z_2 - z_1) \tag{5.0.2}$$

$$\exists\ \beta > 0 :\ z_2 - z_1 < m \Rightarrow f(z_2) - f(z_1) \leq \beta(z_2 - z_1) \tag{5.0.3}$$

Let B be a random variable such that $\mathrm{E}[|f(B)|] \leq +\infty$. Then:

$$\mathrm{VAR}[(f(B)] \geq \frac{\epsilon^2}{64}\left(1 - 16\frac{(\epsilon/8 + \beta)m}{\epsilon\sqrt{\mathrm{VAR}[B]}}\right)\mathrm{VAR}[B] \tag{5.0.4}$$

Proof. Let $h : \mathbb{Z} \to \mathbb{Z}$ be a map defined from f as follows: for a given $z \in \mathbb{Z}$ choose $k \geq 2$ such that $z \in [km, (k+1)m]$ and compute

$$h(z) = \left(\frac{f((k+1)m) - f(km)}{m} \right) (z - km) + f(km)$$

then $h(z)$ is just the linear interpolation of $f(z)$ in $[km, (k+1)m]$. It is easy to see that h satisfies the conditions of lemma 5.0.2 for $c = \epsilon/8$. Then:

$$\mathrm{VAR}[h(B)] \geq \frac{\epsilon^2}{64} \mathrm{VAR}[B] \tag{5.0.5}$$

We want to estimate the distance between the random variables $h(B)$ and $f(B)$. First, we note that from 5.0.2 and by the definition of h, the following inequalities hold for $km \leq B \leq (k+1)m$:

$$\frac{\epsilon}{8}(B - km) + f(km) \leq f(B), h(B) \leq \frac{\epsilon}{8}(B - (k+1)m) + f((k+1)m)$$

looking at conditions 5.0.2, 5.0.3 and the inequalities above we get

$$\begin{aligned} |h(B) - f(B)| &\leq |\frac{\epsilon}{8}(B - km) + f(km) - \frac{\epsilon}{8}(B - (k+1)m) + f((k+1)m)| \\ &\leq \frac{\epsilon}{8} m + |f((k+1)m) - f(km)| \\ &\leq \left(\frac{\epsilon}{8} + \beta \right) m \end{aligned}$$

and by using the last inequality above:

$$\mathrm{VAR}[f(B) - h(B)] \leq \left(\frac{\epsilon}{8} + \beta \right)^2 m^2. \tag{5.0.6}$$

Since $f(B) = h(B) + (f(B) - h(B))$ we can apply triangular inequality and find:

$$\sqrt{\mathrm{VAR}[f(B)]} \geq \sqrt{\mathrm{VAR}[h(B)]} - \sqrt{\mathrm{VAR}[f(B) - h(B)]},$$

hence we have:

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

Finally, applying the inequalities 5.0.5 and 5.0.6 to the last inequality above, we get:

$$\mathrm{VAR}[(f(B)] \geq \frac{\epsilon^2}{64} \left(1 - 16 \frac{(\epsilon/8 + \beta)m}{\epsilon \sqrt{\mathrm{VAR}[B]}} \right) \mathrm{VAR}[B]. \quad \blacksquare$$

Hence to prove that $\mathrm{VAR}[L_n] = \Theta(n)$, we try to represent L_n as $f(\mathcal{Z})$ where f is a random map which tends to increase linearly on a certain scale and $\mathcal{Z}$ is a random variable having fluctuation of order $\sqrt{n}$.

5.1 Random modifications and variables (T, Z)

Let N_l denote the number of blocks in X of length l, whilst N_{l-1}, resp. N_{l+1} denote the number of blocks of length $l-1$, resp $l+1$ in X. Let us define the following three random variables:

$$\begin{aligned} T &:= N_l + N_{l-1} + N_{l+1} && (5.1.1) \\ Z &:= N_l - N_{l-1} - N_{l+1} && (5.1.2) \\ R &:= n - (\, l\, N_l + (l+1)\, N_{l+1} + (l-1)\, N_{l-1}\,) && (5.1.3) \end{aligned}$$

Note that when we know the values of (T, Z, R) we can determine the values of N_{l-1}, N_l and N_{l+1} as a linear function by using the definitions of T, Z and R as follows:

$$\begin{pmatrix} N_{l-1}(T,Z,R) \\ N_l(T,Z,R) \\ N_{l+1}(T,Z,R) \end{pmatrix} = \begin{pmatrix} (2l+1)/4 & -1/4 \\ 1/2 & 1/2 \\ -(2l-1)/4 & -1/4 \end{pmatrix} \begin{pmatrix} T \\ Z \end{pmatrix} + \begin{pmatrix} -(n-R)/2 \\ 0 \\ (n-R)/2 \end{pmatrix} \tag{5.1.4}$$

The variable R represents what is left in X after the last block of length $l-1, l$ or $l+1$.

Example 5.1.1 *Let us consider the sequence $X = 000111100011001$ for $l = 3$ and $n = 15$. We see that $N_{l-1} = 2$, $N_l = 2$ and $N_{l+1} = 1$, hence $T = 5, Z = -1$ and $R = 1$. Also, the block 1 at the end of X has length strictly smaller than $l-1$ which also means that $R = 1$. In this case is easy to interpret what R is since the last block in X has length strictly less than $l-1$. Let us see a different situation. Let us take again $l = 3$ and now consider $B_{X1} = 2, B_{X2} = 3, B_{X3} = 4, B_{X4} = 3, B_{X5} = 2, B_{X6} = 4, \ldots$ such that $X^\infty = 001110000111001111\cdots$ Take $n = 16$ so that $X = 0011100001110011$. Here the last block of X has length $l-1 = 2$ which should imply (using the point of view of the last situation) that $R = 0$. But, notice that the block in X^∞ corresponding to B_{X6} was cut when we took X. In this case, we say that the last block in X corresponds to the rest so $R = 2$ and therefore $N_{l-1} = 2, N_l = 2$ and $N_{l+1} = 1$, then $T = 5$ and $Z = -1$. We take this convention on R, even if the definition 5.1.3 is not the exact one, because of the simplifications later during the computation of the joint distribution of N_{l-1}, N_l, N_{l+1}.*

Let us roughly explain the main idea behind this subsection. Assume that we have a random couple (V, W) which can take on a finite number of values only. We also assume the joint distribution $\mathcal{L}(V, W)$ to be given. To simulate (V, W), we could first simulate V using the marginal law $\mathcal{L}(V)$. We would obtain a numeric value v_0. Then, we could simulate W using the conditional law $\mathcal{L}(W|V = v_0)$ and obtain the numeric value w_0. The couple (v_0, w_0) has joint distribution $\mathcal{L}(V, W)$. Another less efficient possibility is to simulate for each (non-random) value v that V can take, a value for W with distribution $\mathcal{L}(W|V = v)$. Call the numeric value

$w(v)$. Then, we would simulate V with distribution $\mathcal{L}(V)$ and obtain a numeric value v_0. Then, for W we would take among all the values which we have simulated, the one corresponding to $V = v_0$. In this manner, we get $(v_0, w(v_0))$. This couple has the distribution $\mathcal{L}(V, W)$ and this does not even require that we simulate the different $w(v)$'s independently of each other. Only, V needs to be simulated independently of the assignment $v \mapsto w(v)$.

We are going to do the above simulation scheme with V being (T, Z, R) and W being the rest of the information in (X, Y). More precisely, for all possible (t, z, r) non-random values, we simulate X conditional on $(T, Z, R) = (t, z, r)$. The resulting string is denoted by $X_{(t,z,r)}$ and has thus distribution

$$\mathcal{L}(X_{(t,z,r)}) = \mathcal{L}(X \mid (T, Z, R) = (t, z, r)\ ).$$

Let $L_n(t, z, r)$ denote the length of the LCS

$$L_n(t, z, r) := |\text{LCS}(X_{(t,z,r)}, Y)|.$$

We assume that the simulation of the string $X_{(t,z,r)}$ is done independently of (T, R, Z) and of Y. In this manner, we get that $L_n(T, Z, R)$ has same distribution as $L_n = |\text{LCS}(X, Y)|$. So to prove that $\text{VAR}[L_n] = \Theta(n)$, it is enough to prove that

$$\text{VAR}[L_n(T, Z, R)] = \Theta(n). \tag{5.1.5}$$

We saw at the beginning of this section (see lemma 5.0.2 and 5.0.3), that when we transform a variable having variance of order $\Theta(n)$ with a map which tends to increase linearly, then the resulting variable has variance of order $\Theta(n)$. It is easy to see that $\text{VAR}[Z] = \Theta(n)$ (see also lemma 5.1.6). Hence to prove 5.1.5, it is enough to show that with high probability the (random) map

$$z \mapsto L_n(T, z, R)$$

tends to increase linearly (on the appropriate scale and on a domain on which Z typically takes its value). That means, we need to show that we can simulate the values $L_n(t, z, r)$ in such a manner to get the desired distribution $\mathcal{L}(X|(T, Z, R) = (t, z, r))$ as well as the desired linear increase of the map $z \mapsto L_n(T, z, R)$. This is achieved by simulating $X_{(t,z,r)}$ in the following way: for a given value (t, r), so that $P((T, R) = (t, r)) \neq 0$, we take a left most (left most to be defined later) value z_0 and simulate a string with distribution equal to the conditional distribution of X given $(T, Z, R) = (t, z_0, r)$. That resulting string is denoted by $X_{(t,z_0,r)}$. Then, we apply the random modification to $X_{(t,z_0,r)}$. This means, we choose one block of length $l - 1$ and one block of length $l + 1$ at random in $X_{(t,z_0,r)}$ and turn them both into length l. The resulting string is denoted by $X_{(t,z_0+4,r)}$. Then, we choose at random in $X_{(t,z_0+4,r)}$ a block of lenght $l - 1$ and a block of lenght $l + 1$ and turn them both into length l. The new string which we obtain in this manner is denoted by $X_{(t,z_0+8,r)}$. We keep repeating this same operation to obtain the sequence of strings

$$X_{(t,z_0,r)}, X_{(t,z_0+4,r)}, X_{(t,z_0+8,r)}, \ldots. \tag{5.1.6}$$

For each value of (t, r) with $\mathrm{P}((T, R) = (t, r)) \neq 0$ we obtain two finite sequences of strings: first 5.1.6 and then

$$X_{(t,z_0+2,r)}, X_{(t,z_0+6,r)}, X_{(t,z_0+10,r)}, \ldots.$$

by a similar procedure. Namely, after $X_{(t,z_0+2,r)}$ is generated with distribution X conditional on $(T, Z, R) = (t, z + 2, R)$, the subsequent strings are obtained by applying sucessively the random modification tilde, which chooses at random in the string a block of length $l - 1$ and a block of length $l + 1$ and turn them both into length l.

Recall that in this section we assume that our random modification has a biased effect of $\epsilon > 0$ on the LCS, so that with high probability

$$\mathrm{E}[\tilde{L}_n - L_n \mid X, Y\,] \geq \epsilon.$$

Hence, it follows that the map $z \mapsto L_n(T, z, R)$ tends with high probability to increase with slope close to ϵ on a constant time scale $\ln n$ (the constant must be taken large enough though, see lemma 5.1.4 and proposition 5.1.2). In other words, since the random modification has a biased positive effect, the map $z \mapsto L_n(T, z, R)$ behaves like a random walk with drift ϵ. The only thing which remains to be proved is that with our scheme of using the random modification, the strings $X_{(t,z,r)}$ have the right distribution, i.e. the distribution of X conditional on $(T, Z, R) = (t, z, r)$. This is proved in lemma 5.1.3.

We have so far summarized the idea which explains why the biased effect of the random modification implies $\mathrm{VAR}[L_n] = \Theta(n)$. There is one more detail which we should mention and which makes notations a little more difficult. To prove that $z \mapsto L_n(T, z, R)$ tends to increase linearly we use the biased effect on the LCS for the random modification. However, this bias holds with high probability for X and not for $X_{(t,z,r)}$. When we look at the conditional distribution of X given $(T, Z, R) = (t, z, r)$, we divide by the probability

$$\mathrm{P}((T, Z, R) = (t, z, r)). \tag{5.1.7}$$

The string $X_{(t,z,r)}$ has distribution of X conditional on $(T, Z, R) = (t, z, r)$. So for the biased effect to have large probability also for $X_{(t,z,r)}$ (and not just for X), we need the probability 5.1.7 to not be too small. To assure this, we will restrict ourselves to "typical" values for (T, Z, R). We will consider only values for (T, Z) which lie in an interval $D = D_T \times D_Z$ (see definition below 5.1.19) and prove that any possible value $(t, z) \in D_z \times D_t$ has polynomially bounded probability (see lemma 5.1.2).

Let us now give all the details:

Proposition 5.1.1 *Given* $\epsilon > 0$ *there exist constants* $1 \leq k_1, k_2, k_3 \leq k^*$ *all not depending on* n *but on* ϵ *such that:*

$$\mathrm{P}\left(\left|\frac{N_{l-1}-\frac{n}{3l}}{\sqrt{n}}\right| \leq k_1\right), \quad \mathrm{P}\left(\left|\frac{N_l-\frac{n}{3l}}{\sqrt{n}}\right| \leq k_2\right), \quad \mathrm{P}\left(\left|\frac{N_{l+1}-\frac{n}{3l}}{\sqrt{n}}\right| \leq k_3\right) \geq 1-\epsilon \tag{5.1.8}$$

for every n *large enough.*

Proof. We will prove the result only for N_l in the positive case, namely

$$\mathrm{P}\left(\frac{N_l-\frac{n}{3l}}{\sqrt{n}} \leq k_2\right) \geq 1-\epsilon \tag{5.1.9}$$

since the technique is the same for all the other cases and for N_{l-1}, N_{l+1} as well. Given $\alpha, \beta, \pi, m, n > 0$, let us define the following events:

$$\begin{aligned} A(\alpha,m) &= \left\{\xi_1+\cdots+\xi_m \leq \frac{m}{3}+\alpha\sqrt{m}\right\} \\ B(\beta,n) &= \left\{N_{l-1}+N_l+N_{l+1} \leq \frac{n}{l}+\beta\sqrt{n}\right\} \\ C(\pi,n) &= \left\{N_l \leq \frac{n}{3l}+\pi\sqrt{n}\right\} \end{aligned}$$

where the random variables ξ_i are defined as

$$\xi_i = \begin{cases} 1 & \text{if } B_i = l \\ 0 & \text{otherwise} \end{cases}$$

Now to prove 5.1.9 is the same as to find π depending on ϵ but not on n such that:

$$\mathrm{P}(C(\pi,n)) \geq 1-\epsilon \tag{5.1.10}$$

for every n large enough. For given α, β we define

$$\begin{aligned} m^* &= \frac{n}{l}+\beta\sqrt{n} \\ \pi^* &= \frac{\beta}{3}+\alpha\sqrt{\frac{1}{l}+\beta} \end{aligned}$$

Taking m^*, π^* as before we have that:

$$\left.\begin{array}{l} \xi_1+\cdots+\xi_{m^*} \leq \frac{m^*}{3}+\alpha\sqrt{m^*} \\ N_{l-1}+N_l+N_{l-1} \leq \frac{n}{l}+\beta\sqrt{n} \end{array}\right\} \Rightarrow N_l \leq \frac{n}{3l}+\pi^*\sqrt{n}$$

which is equivalent to the inclusion

$$A(\alpha, m^*) \cap B(\beta, n) \subseteq C(\pi^*, n).$$

Hence, proving 5.1.10 is equivalent to finding α and β depending on ϵ such that:

$$\mathrm{P}(C^c(\pi^*, n)) \leq \mathrm{P}(A^c(\alpha, m^*)) + \mathrm{P}(B^c(\beta, n)) \leq \epsilon \tag{5.1.11}$$

For this we will use a special version of Chebichev inequality: let U be a random variable, then for every positive constant u we have

$$\mathrm{P}\left(\ |U - \mathrm{E}[U]| \geq u\sqrt{\mathrm{VAR}[U]} \ \right) \leq \frac{1}{u^2} \tag{5.1.12}$$

For the event $A(\alpha, m^*)$ we have, that taking $\alpha = \sqrt{2/(3\epsilon)}$ and using 5.1.12 the following inequality holds:

$$\mathrm{P}(A^c(\alpha, m^*)) = \mathrm{P}\left(\frac{\xi_1 + \cdots + \xi_{m^*}}{m^*} - \frac{1}{3} \geq \frac{\alpha}{\sqrt{m^*}}\right) \leq \frac{\epsilon}{2}. \tag{5.1.13}$$

To choose β in the event $B(\beta, n)$ such that $\mathrm{P}(B^c(\beta, n))$, we will use the same techniques as in the previous chapter. Let $m := \frac{n}{l} + \beta\sqrt{n}$ be an auxiliar variable. Recall that N_n^X is the total number of blocks in X defined in 4.1.1. We have

$$\begin{aligned}
\mathrm{P}(B^c(\beta, n)) &\leq \mathrm{P}\left(N_n^X \geq \frac{n}{l} + \beta\sqrt{n}\right) \\
\text{(by using } N_n^X \geq m \Leftrightarrow S_m^X \leq n \text{)} &\leq \mathrm{P}\left(S_m^X \leq n\right) \\
&= \mathrm{P}\left(\frac{S_m^X}{m} - l \leq \frac{n}{m} - l\right) \\
\text{(by 4.0.2 with } \mathrm{P}(|B_{X1} - l| \leq 1) = 1) &\leq 2\exp\left(-\frac{m}{2}\left(\frac{n}{m} - l\right)^2\right) \\
&= 2\exp\left(-\frac{l^3\beta^2}{2}n\right)
\end{aligned}$$

which finally says

$$\mathrm{P}(B^c(\beta, n)) \leq \exp\left(-\frac{l^3\beta^2}{2}n\right). \tag{5.1.14}$$

Then, taking β such that

$$\exp\left(-\frac{l^3\beta^2}{2}n\right) \leq \frac{\epsilon}{4}$$

for n large enough, in 5.1.14 we have

$$\mathrm{P}(B^c(\beta, n)) \leq \frac{\epsilon}{2} \tag{5.1.15}$$

Combining 5.1.13 and 5.1.15 in 5.1.11, we finish the proof with $k_2 = \pi^*$. ■

We will need later the following lemma:

Lemma 5.1.1 *There exists $c > 0$ not depending on n such that:*

$$\mathrm{P}\left(T \in \left[\frac{n}{l} - c\sqrt{n}, \frac{n}{l} + c\sqrt{n}\right], Z \in \left[-\frac{n}{3l} - c\sqrt{n}, -\frac{n}{3l} + c\sqrt{n}\right]\right) \geq 0.9 \quad (5.1.16)$$

Proof. Given any $\varepsilon > 0$ we have that:

$$\begin{aligned} \mathrm{P}\left(\left|\frac{N_{l-1} + N_l + N_{l+1} - \frac{n}{l}}{\sqrt{n}}\right| > 3k^*\right) &\leq \mathrm{P}\left(\left|\frac{N_{l-1} - \frac{n}{3l}}{\sqrt{n}}\right| + \left|\frac{N_l - \frac{n}{3l}}{\sqrt{n}}\right| + \left|\frac{N_{l+1} - \frac{n}{3l}}{\sqrt{n}}\right| > 3k^*\right) \\ &\leq 3 \cdot \mathrm{P}\left(\left|\frac{N_l - \frac{n}{3l}}{\sqrt{n}}\right| > k^*\right) < 3\varepsilon \end{aligned} \quad (5.1.17)$$

$$\begin{aligned} \mathrm{P}\left(\left|\frac{N_l - N_{l-1} - N_{l+1} + \frac{n}{3l}}{\sqrt{n}}\right| > 3k^*\right) &\leq \mathrm{P}\left(\left|\frac{N_l - \frac{n}{3l}}{\sqrt{n}}\right| + \left|\frac{\frac{n}{3l} - N_{l-1}}{\sqrt{n}}\right| + \left|\frac{\frac{n}{3l} - N_{l+1}}{\sqrt{n}}\right| > 3k^*\right) \\ &\leq 3 \cdot \mathrm{P}\left(\left|\frac{N_l - \frac{n}{3l}}{\sqrt{n}}\right| > k^*\right) < 3\varepsilon \end{aligned} \quad (5.1.18)$$

both inequalities hold due to proposition 5.1.1. Then, it is not difficult to see that

$$\mathrm{P}\left(\left\{T \notin \left[\frac{n}{l} - 3k^*\sqrt{n}, \frac{n}{l} + 3k^*\sqrt{n}\right]\right\} \cup \left\{Z \notin \left[-\frac{n}{3l} - 3k^*\sqrt{n}, -\frac{n}{3l} + 3k^*\sqrt{n}\right]\right\}\right) \leq 6\varepsilon$$

from where the proof is complete with $c = 3k^*$ and $\varepsilon = 1/60$. ■

Let D denote the domain

$$D := \left[\frac{n}{l} - c\sqrt{n}, \frac{n}{l} + c\sqrt{n}\right] \times \left[-\frac{n}{3l} - c\sqrt{n}, -\frac{n}{3l} + c\sqrt{n}\right] \quad (5.1.19)$$

and let

$$\begin{aligned} D_T &:= \left[\frac{n}{l} - c\sqrt{n}, \frac{n}{l} + c\sqrt{n}\right] \\ D_Z &:= \left[-\frac{n}{3l} - c\sqrt{n}, -\frac{n}{3l} + c\sqrt{n}\right] \end{aligned}$$

hence,

$$D = D_T \times D_Z.$$

Given $(t, z) \in D$ such that $(T, Z) = (t, z)$ we have

$$N_{l-1}(t, z) + N_l(t, z) + N_{l+1}(t, z) = t.$$

The probability for a realization of N_{l-1}, N_l and N_{l+1} is given by:

$$\begin{aligned} \mathrm{P}(T = t, Z = z, R = r) &= \binom{N_{l-1}(t,z) + N_l(t,z) + N_{l+1}(t,z)}{N_{l-1}(t,z) \;\; N_l(t,z) \;\; N_{l+1}(t,z)} \left(\frac{1}{3}\right)^t \cdot \mathrm{P}(B_{X1} > r) \\ &= \frac{t!}{(N_{l-1}(t,z))!\,(N_l(t,z))!\,(N_{l+1}(t,z))!} \left(\frac{1}{3}\right)^t \cdot \mathrm{P}(B_{X1} > r) \end{aligned} \quad (5.1.20)$$

where the probability $\mathrm{P}(B_{X1} > r) = \mathrm{P}(R = r)$ is due to the convention of R described in the example 5.1.1. Finally, due to 5.1.4, for any $n_1, n_2, n_3 \in \mathbb{N}$ the conditional joint distribution

$$\mathrm{P}(N_{l-1}(T, Z, R) = n_1, N_l(T, Z, R) = n_2, N_{l+1}(T, Z, R) = n_3 \,|\, R = r)$$

is multinomial.

Lemma 5.1.2 *There exists $k_0 > 0$ not depending on n (but depending on c) such that for every $(t, z) \in D$ and $r < l + 1$ for which the probability $\mathrm{P}((T, Z, R) = (t, z, r)) \neq 0$, we have that:*

$$\mathrm{P}((T, Z, R) = (t, z, r)) \geq \frac{k_0}{n}$$

for every n large enough.

Proof. In all what follows, we always suppose that every expression in factorial numbers is an integer number. As we saw in 5.1.20, given $(t, z) \in D$ we have the following expression for the probability $\mathrm{P}((T, Z, R) = (t, z, r))$:

$$\mathrm{P}((T, Z, R) = (t, z, r)) = \frac{t!}{n_1!\; n_2! n_3!} \left(\frac{1}{3}\right)^t \cdot \mathrm{P}(B_{X1} > r)$$

where for simplicity $n_1 := N_{l-1}(t, z, r), n_2 := N_l(t, z, r), n_3 := N_{l+1}(t, z, r)$. Now we will develop the expression in order to find a lower bound. Let us keep in mind that in all what follows $n_1 + n_2 + n_3 = t$. Consider the Stirling's approximations:

$$\begin{aligned} t! &= \sqrt{2\pi} \cdot t^{t+\frac{1}{2}} \cdot e^{-t} \left(1 + O\left(\frac{1}{t}\right)\right) \\ n_i! &= \sqrt{2\pi} \cdot n_i^{n_i+\frac{1}{2}} \cdot e^{-n_i} \left(1 + O\left(\frac{1}{n_i}\right)\right) \end{aligned}$$

for $i = 1, 2, 3$. For $r < l + 1$ natural number, note that $\mathrm{P}(B_{X1} > r) \geq 1/3$ since when $r < l - 1$ then $\mathrm{P}(B_{X1} > r) = 1$, otherwise $\mathrm{P}(B_{X1} > r) = 2/3$. With all above, we may write:

$$\begin{aligned} \mathrm{P}((T, Z, R) = (t, z, r)) &= \frac{t!}{n_1!\; n_2! n_3!} \left(\frac{1}{3}\right)^t \cdot \mathrm{P}(B_{X1} > r) \\ &\geq \frac{3^{-t}\, t^{t+\frac{1}{2}} e^{-t}}{2\pi \cdot n_1^{n_1+\frac{1}{2}} e^{-n_1} \cdot n_2^{n_2+\frac{1}{2}} e^{-n_2} \cdot n_3^{n_3+\frac{1}{2}} e^{-n_3}} \cdot \frac{1}{3} \cdot \left(1 + O\left(\frac{1}{t}\right)\right) \\ &= \frac{3^{-t}\, t^{t+\frac{1}{2}}}{6\pi \cdot n_1^{n_1+\frac{1}{2}} \cdot_2^{n_2+\frac{1}{2}} \cdot n_3^{n_3+\frac{1}{2}}} \left(1 + O\left(\frac{1}{t}\right)\right) \\ &= \frac{3^{-t}}{6\pi} \left(\frac{t}{n_1}\right)^{n_1} \left(\frac{t}{n_2}\right)^{n_2} \left(\frac{t}{n_3}\right)^{n_3} \left(\frac{t}{n_1\; n_2\; n_3}\right)^{\frac{1}{2}} \left(1 + O\left(\frac{1}{t}\right)\right), \end{aligned}$$

which says that:

$$\mathrm{P}((T,Z,R) = \frac{3^{-t}}{6\pi}\left(\frac{t}{n_1}\right)^{n_1}\left(\frac{t}{n_2}\right)^{n_2}\left(\frac{t}{n_3}\right)^{n_3}\left(\frac{t}{n_1\;n_2\;n_3}\right)^{\frac{1}{2}}\left(1+O\left(\frac{1}{t}\right)\right) \tag{5.1.21}$$

Recall that $t \in D_T$ implies

$$t \geq \frac{n}{l} - c\sqrt{n} \tag{5.1.22}$$

and that, from lemma 5.1.1 the following inequalities hold almost sure

$$n_1 \leq \frac{n}{3l} + k^*\sqrt{n}, \quad n_2 \leq \frac{n}{3l} + k^*\sqrt{n}, \quad n_3 \leq \frac{n}{3l} + k^*\sqrt{n} \tag{5.1.23}$$

for every n large enough. So, by using 5.1.22 and 5.1.23, equality 5.1.21 becomes

$$\begin{aligned}
\mathrm{P}((T,Z,R) \quad &\geq \quad \frac{3^{-\frac{n}{l}-c\sqrt{n}}}{6\pi}\left(\frac{\frac{n}{l}-c\sqrt{n}}{\frac{n}{3l}+k^*\sqrt{n}}\right)^{\frac{n}{l}+3k^*\sqrt{n}}\left(\frac{\frac{n}{l}-c\sqrt{n}}{\left(\frac{n}{3l}+k^*\sqrt{n}\right)^3}\right)^{\frac{1}{2}}\left(1+O\left(\frac{1}{n}\right)\right), \\
&= \quad \frac{3^{2c\sqrt{n}+\frac{3}{2}}\;\;l}{6\pi n}\cdot\frac{\left(1-\frac{cl}{\sqrt{n}}\right)^{3k^*\sqrt{n}}}{\left(1+\frac{3k^*l}{\sqrt{n}}\right)^{3k^*\sqrt{n}}}\cdot\left(\frac{1-\frac{cl}{\sqrt{n}}}{1+\frac{3k^*l}{\sqrt{n}}}\right)^{\frac{n}{l}}\cdot\frac{\left(1-\frac{cl}{\sqrt{n}}\right)^{\frac{1}{2}}}{\left(1+\frac{3k^*l}{\sqrt{n}}\right)^{\frac{3}{2}}}\left(1+O\left(\frac{1}{n}\right)\right) \\
&\geq \quad \frac{3^{2c\sqrt{n}+\frac{3}{2}}\;\;l}{6\pi n}\cdot\frac{\left(1-\frac{cl}{\sqrt{n}}\right)^{3k^*\sqrt{n}}}{\left(1+\frac{3k^*l}{\sqrt{n}}\right)^{3k^*\sqrt{n}}}\cdot\left(\frac{1-\frac{cl}{\sqrt{n}}}{1+\frac{3k^*l}{\sqrt{n}}}\right)^{\frac{\sqrt{n}}{l}}\cdot\frac{\left(1-\frac{cl}{\sqrt{n}}\right)^{\frac{1}{2}}}{\left(1+\frac{3k^*l}{\sqrt{n}}\right)^{\frac{3}{2}}}\left(1+O\left(\frac{1}{n}\right)\right)
\end{aligned}$$

So we get:

$$\mathrm{P}((T,Z,R) = (t,z,r)) \geq \frac{3^{2c\sqrt{n}+\frac{3}{2}}\;\;l}{6\pi n}\cdot\frac{\left(1-\frac{cl}{\sqrt{n}}\right)^{3k^*\sqrt{n}}}{\left(1+\frac{3k^*l}{\sqrt{n}}\right)^{3k^*\sqrt{n}}}\cdot\left(\frac{1-\frac{cl}{\sqrt{n}}}{1+\frac{3k^*l}{\sqrt{n}}}\right)^{\frac{\sqrt{n}}{l}}\cdot\frac{\left(1-\frac{cl}{\sqrt{n}}\right)^{\frac{1}{2}}}{\left(1+\frac{3k^*l}{\sqrt{n}}\right)^{\frac{3}{2}}}\left(1+O\left(\frac{1}{n}\right)\right) \tag{5.1.24}$$

But we have the following inequalities for the limits:

$$\begin{aligned}
\lim_{n\to 0}\left(1-\frac{cl}{\sqrt{n}}\right)^{3k^*\sqrt{n}} &= e^{-3ck^*l} \quad \geq \quad \frac{e^{-3ck^*l}}{2} \\
\lim_{n\to 0}\left(1+\frac{3k^*l}{\sqrt{n}}\right)^{3k^*\sqrt{n}} &= e^{9{k^*}^2 l} \quad \leq \quad 2e^{9{k^*}^2 l} \\
\lim_{n\to 0}\left(\frac{1-\frac{cl}{\sqrt{n}}}{1+\frac{3k^*l}{\sqrt{n}}}\right)^{\frac{\sqrt{n}}{l}} &= e^{-(c+3k^*)} \quad \geq \quad \frac{e^{-(c+3k^*)}}{2} \\
\lim_{n\to 0}\frac{\left(1-\frac{cl}{\sqrt{n}}\right)^{\frac{1}{2}}}{\left(1+\frac{3k^*l}{\sqrt{n}}\right)^{\frac{3}{2}}} &= 1 \quad \geq \quad \frac{1}{2}
\end{aligned}$$

Hence, by using these last inequalities and $\sqrt{n} \geq 1$ in 5.1.24 we have:

$$P((T, Z, R) = (t, z, r)) \geq \frac{3^{2c+\frac{3}{2}}\ l}{96\pi\ e^{3ck^*l+c+3k^*-9k^{*2}l}} \cdot \frac{1}{n}$$

for every n large enough. Therefore, taking

$$k_0 = \frac{3^{2c+\frac{3}{2}}\ l}{96\pi\ e^{3ck^*l+c+3k^*-9k^{*2}l}}$$

proves the result. ■

Note that for any variables X and Y we have (see for example [22])

$$\mathrm{VAR}[Y] = \mathrm{E}[\mathrm{VAR}[Y|X]] + \mathrm{VAR}[\mathrm{E}[Y|X]] \geq \mathrm{E}[\mathrm{VAR}[Y|X]]. \tag{5.1.25}$$

Let O be the random variable which is equal to one when (T, Z) is in D and 0 otherwise.
We can now use inequalities 5.1.16 and 5.1.25 to find

$$\mathrm{VAR}[L_n] \geq \mathrm{E}[\mathrm{VAR}[L_n|O]] \geq \mathrm{VAR}[L_n|O = 1] \cdot \mathrm{P}(O = 1) \geq 0.9\mathrm{VAR}[L_n|O = 1] \tag{5.1.26}$$

Next for every (t, z) in D and $r < l + 1$ we are going to simulate the random variable L_n conditional on $(T, Z, R) = (t, z, r)$. We denote the result by $L_n(t, z, r)$. In other words, the distribution of $L_n(t, z)$ is equal to

$$\mathcal{L}(L_n(t, z, r)) = \mathcal{L}(L_n|(T, Z, R) = (t, z, r)).$$

Let (T_D, Z_D) denote a variable having the distribution of (T, Z) conditional on the event $(T, Z) \in D$. We assume that all the $L_n(t, z, r)$ are independent of (T_D, Z_D). Then, we get that

$$L_n(T_D, Z_D, R)$$

has same distribution as L_n conditional on $(T, Z) \in D$. Hence, we get

$$\mathrm{VAR}[L_n|O = 1] = \mathrm{VAR}[L_n(T_D, Z_D, R)] \tag{5.1.27}$$

By using 5.1.25, we find

$$\mathrm{VAR}[L_n(T_D, Z_D, R)]] \geq \mathrm{E}[\ \mathrm{VAR}[L_n(T_D, Z_D, R)|T_D, R]\ ]. \tag{5.1.28}$$

Note that for $L_n(T_D, Z_D, R)$ to have the same distribution as L_n conditional on $(T, Z) \in D$ and on $R = r$, the variables $L_n(t, z, r)$ do not need to be independent of each other. We are next going to explain how we simulate the variables $L_n(t, z, r)$ a bit more in detail as before. We simulate a string $X_{(t,z,r)}$ having the distribution of the string X conditional on the event $(T, Z, R) = (t, z, r)$. Then we put

$$L_n(t, z, r) = |\mathrm{LCS}(X_{(t,z,r)}, Y)|.$$

Next, let us describe how we simulate $X_{(t,z,r)}$ based on what was roughly explained at the beginning of subsection 5.1. Given $t_0 \in D_T$ the most left element in D_T and $r_0 < l-1$, we are going to simulate $X_{(t_0,z,r_0)}$ for $z \in D_Z$ only if $\mathrm{P}((T,Z,R) = (t_0,z,r_0)) \neq 0$. We simulate $X_{(t_0,z_0,r_0)}$ so that it has distribution $\mathcal{L}(\, X|(T,Z,R) = (t_0,z_0,r_0)\,)$. Next, we simulate $X_{(t_0,z_0+2,r_0)}$ by choosing in X, with the same probability, a block of length $l-1$ either a block of length $l+1$ and change its length to l. The next realization we simulate is $X_{(t_0,z_0+4,r_0)}$ by choosing in X, with the same probability, a block of length $l-1$ and a block of length $l+1$ and change their lengths to l (this is our usual random modification). Then by induction we simulate

$$\{X_{(t_0,z_0+4i,r_0)} : i = 1, 2, \dots\}$$

with our usual random modification and later

$$\{X_{(t_0,z_0+2+4i,r_0)} : i = 1, 2, \dots\}$$

just starting with $X_{(t_0,z_0+2,r_0)}$ and performing our usual random modification to get $X_{(t_0,z_0+6,r_0)}, X_{(t_0,z_0+10,r_0)}, X_{(t_0,z_0+14,r_0)}$, etc. Both inductions run untill indexes i_0 and i_0^*, resp., satisfying:

$$\begin{aligned} z_0 + 4i_0 &\leq -\frac{n}{3l} + c\sqrt{n} \quad \Rightarrow \quad i_0 \leq \sqrt{n} \\ z_0 + 2 + 4i_0^* &\leq -\frac{n}{3l} + c\sqrt{n} \quad \Rightarrow \quad i_0^* \leq \frac{\sqrt{n}-1}{2} \end{aligned}$$

For simplicity, let us call $z_0, z_1 = z_0 + 2, z_2 = z_0 + 4, \dots, z_d$ all the values which Z takes. After we have simulated $X_{(t_0,z_0,r_0)}, X_{(t_0,z_1,r_0)}, \dots, X_{(t_0,z_d,r_0)}$ we fix $t_1 = t_0 + 1$ and repeat all the procedure again starting with the simulation of $X_{(t_1,z_0,r_0)}$. We keep taking $t_2 < t_3 < t_4 \dots$ all natural numbers in D_T to finish all the simulation of $\{X_{(t,z,r_0)} : t \in D_T,\ z = z_0, z_1, \dots, z_d\}$. Once we have finished with that, we take $r_1 < l-1$ natural number and do all the simulation above starting with $X_{(t_0,z_0,r_1)}$ only if $\mathrm{P}((t_0,z_0,r_1)) \neq 0$. Finally, we obtain the complete sequence $\{X_{(t,z,r)} : t \in D_T,\ z = z_0, z_1, \dots, z_d\,,\ r = 0, \dots, l-2\}$, where each (t,z,r) has probability $\mathrm{P}((T,Z,R) = (t,z,r)) \neq 0$.

We need to verify that this operation give us the equiprobable distribution. This is the content of the next lemma:

Lemma 5.1.3 *Assume that $X_{(t,z,r)}$ is distributed according to*

$$\mathcal{L}(X|(T,Z,R) = (t,z,r)).$$

Choose at random (with equal probability) in the string $X_{(t,z,r)}$ a block of length $l+1$ and $l-1$ and modify them to have both length l. Then the resulting string has distribution

$$\mathcal{L}(X|(T,Z,R) = (t,z+4,r)).$$

Proof. Because of our linear equation system 5.1.4, we have that conditioning on T, Z, R is equivalent to conditioning on (N_{l-1}, N_l, N_{l+1}). As mentioned, $X_{(t,z,r)}$ denotes a string of length n, having the distribution of X conditional on $(T, Z, R) = (t, z, r)$. We denote by $\tilde{X}_{(t,z,r)}$ the string we obtain by performing our random modification on $X_{(t,z,r)}$. In other words, $\tilde{X}_{(t,z,r)}$ is obtained by choosing a block of length $l+1$ and a block of length $l-1$ at random in $X_{(t,z,r)}$ and changing them both to length l. Let (n_1, n_2, n_3) be the number of blocks of length $l-1$, l and $l+1$ corresponding to (t, z, r). In other words, n_1, n_2 and n_3 are given by the linear system of equation 5.1.4 when $N_{l-1} = n_1, N_l = n_2, N_{l+1} = n_3$ and $T = t, Z = z, R = r$. We have

$$\mathrm{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3 | T = t, Z = z, R = r) = 1.$$

The distribution of the random string $X_{(t,z,r)}$ is the uniform distribution on $\xi^n(t, z, r)$. Here, $\xi^n(t, z, r)$ denotes the set of strings of length n, which consists only of blocks of length $l-1$, l and $l+1$, such that the total number of blocks is t, whilst the number of blocks of length l minus the number of blocks of length $l-1$ and $l+1$ is z. We also request that the rest block at the end has length r. We can describe $\xi^n(t, z, r)$ equivalently as the set of all strings consisting exactly of n_1 blocks of length $l-1$, n_2 blocks of length l and n_3 blocks of length $l+1$, no other blocks allowed except a rest block at the end which has length strictly less than $l-1$. In other words, the random string $X_{(t,z,r)}$ is such that the number of blocks of length $l-1$, l and $l+1$ is determined, only the order in which these blocks appear varies. Among others, each possible realization for $X_{(t,z,r)}$ which has non-zero probability has the same probability:

$$\binom{n_1 + n_2 + n_3}{n_1 \; n_2 \; n_3}^{-1} \tag{5.1.29}$$

When we apply the random modification, the variable T stays the same, the variable Z increases by 4 and the variable R stays the same.

Since the distribution of X conditional on (T, Z, R) is the uniform distribution on the appropriate set of strings, we have the following: for proving that $\tilde{X}_{(t,z,r)}$ has distribution of X conditional on $(T, Z, R) = (t, z+4, r)$ it is enough to show that its distribution is the uniform distribution on $\xi^n(t, z+4, r)$. For this, let $\tilde{x}$ denote a (non-random) element of $\xi^n(t, z+4, r)$. Hence, the number of blocks in $\tilde{x}$ of length $l-1$, l, resp $l+1$ is $n_1 - 1$, $n_2 + 2$, resp. $n_3 - 1$. The probability

$$P(\tilde{X}_{(t,z,r)} = \tilde{x})$$

can be calculated as follows: if we only know $\tilde{x}$, any block of length l of $\tilde{x}$ could be the block which had lenght $l-1$ and has been turned into length l by the *tilde operation* (choosing blocks at random and changing their lenghts). Same thing for the block which had length $l+1$. But when we know these two blocks, then the string before the random modification is uniquely determined. Let x be such a string which could lead to $\tilde{x}$ after the random modification. There are hence $\tilde{n}_2 \cdot (\tilde{n}_2 - 1)$ such strings (here, $\tilde{n}_2 = n_2 + 2$, so that $\tilde{n}_2$ denotes the number of blocks of length l in $\tilde{x}$). The probability, given $X_{(t,z,r)} = x$, that the random string turns out to be $\tilde{x}$ is equal to $1/(n_1 \cdot n_3)$. As a matter of fact, among the n_1 blocks of lenght $l-1$, there is exactly one which needs to be randomly modified. Similarly, among the n_3 blocks of lenght $l+1$, there is exactly one which needs to be changed into length l in order to obtain the string $\tilde{x}$. Hence,

$$\mathrm{P}(\tilde{X}_{(t,r,z)} = \tilde{x} | X_{(t,z,r)} = x) = \frac{1}{n_1 \cdot n_3}. \tag{5.1.30}$$

Let ξ^{n*} denote the set of all strings which could lead to $\tilde{x}$ if we apply the random modification to them. We saw that there are $(n_2+2)(n_2+1)$ elements in the set ξ^{n*}. By law of total probability, we have

$$\mathrm{P}(\tilde{X}_{(t,z,r)} = \tilde{x}) = \sum_{x \in \xi^{n*}} \mathrm{P}(\tilde{X} = \tilde{x} | X = x) \mathrm{P}(X_{(t,z,r)} = x) = \sum_{x \in \xi^{n*}} \frac{1}{n_1 \cdot n_3} \binom{n_1 + n_2 + n_3}{n_1 \; n_2 \; n_3}^{-1} \tag{5.1.31}$$

The last equation above was obtained using 5.1.30 and 5.1.29. Note that the sum on the most right of equation in 5.1.31, is a sum of $(n_2+2)(n_2+1)$ equal terms. This leads to

$$\mathrm{P}(\tilde{X}_{(t,z,r)} = \tilde{x}) = \frac{(n_2+2)(n_2+1)}{n_1 \cdot n_3} \binom{n_1 + n_2 + n_3}{n_1 \; n_2 \; n_3}^{-1}.$$

The formula on the right side above does not depend on $\tilde{x}$. Hence, this proves that $\tilde{X}_{(t,z,r)}$ has the uniform distribution on the set of strings $\xi^n(t, z+4, r)$. But the uniform distribution is the distribution of X conditional on $(T, Z, R) = (t, z+4, r)$. That is, we have proven that

$$\mathcal{L}(\tilde{X}_{(t,r,z)}) = \mathcal{L}(X | (T, Z, R) = (t, z+4, r)),$$

which finishes this proof. ■

Note that we have seen what happens with the variables T, Z, R after our random modification, let us see what happens with the length of the LCS after our random modification. In what follows, we always consider a triplet of values (t, z, r) such that $\mathrm{P}((T, Z, R) = (t, z, r)) \neq 0$. For any $\epsilon > 0$ let $U^n_{t,r}(\epsilon)$ denote the event that the map

$$D_Z \to \mathbb{N} \; : \; z \mapsto L_n(t, z, r)$$

is increasing with a slope of at least $\epsilon/8$ on a scale $c_2 \ln(n)$ where $c_2 > 0$ is a large constant not depending on n. More precisely, $U^n_{t,r}(\epsilon)$ is the event that for any z_1, z_2 in D_Z, with $z_2 - z_1 \geq c_2 \ln(n)$ we have

$$L_n(t, z_2, r) - L_n(t, z_1, r) \geq (z_2 - z_1)\epsilon/8.$$

The event $U^n_{t,r}(\epsilon)$ has large probability because we assumed that inequality 5.0.1 holds. Hence $z \mapsto L_n(t,z,r)$ can be viewed somehow as behaving like a random walk with drift ϵ. In the next lemma we will show this looking at the event $U^n(\epsilon)$:

$$U^n(\epsilon) := \bigcap_{t\in D_T,\ r<l+1} U^n_{t,r}(\epsilon).$$

Lemma 5.1.4 *Given $\epsilon > 0$, take α from inequality 5.0.1 (theorem 2.1.2) and c_2 to be big enough but not depending on n, for example $c_2 \geq \frac{80}{\epsilon^2}$ depending on ϵ. Then, there exists a constant $k_* > 0$ not depending on n but on α and on c_2 such that:*

$$\mathrm{P}(U^{nc}(\epsilon)) \leq \frac{k_*}{n^2} \tag{5.1.32}$$

for n large enough, provided 5.0.1 holds.

Proof. We are going to define an event $\mathcal{U}(\epsilon)$ for any $\epsilon > 0$. Let $\mathcal{U}_{(t,z,r)}(\epsilon)$ be the event that the expected conditional increase is larger than ϵ when we introduce the random change into $X_{(t,z,r)}$. More precisely, let $\mathcal{U}^n_{(t,z,r)}(\epsilon)$ be the event that

$$\mathrm{E}[\ L_n(t,z+4,r) - L_n(t,z,r)|X_{(t,z,r)}, Y\ ] \geq \epsilon \tag{5.1.33}$$

Let

$$\mathcal{U}^n(\epsilon) := \bigcap_{(t,z)\in D,\ r<l+1} \mathcal{U}^n_{(t,z,r)}(\epsilon).$$

hence

$$\mathrm{P}(\mathcal{U}^{nc}(\epsilon)) \leq \sum_{(t,z)\in D,\ r<l+1} \mathrm{P}(\mathcal{U}^{nc}_{(t,z,r)}(\epsilon)). \tag{5.1.34}$$

Note that inequality 5.0.1 provides a bound for the probability that the conditional expected increase of LCS due to our random modification not being larger or equal to ϵ. That probability bound is $\exp(-n^\alpha)$. The only problem is that the bound is for X and Y whilst the event $\mathcal{U}^n_{(t,z,r)}(\epsilon)$ is for $X_{(t,z,r)}$ and Y. By going on to conditional probability we must multiply the probability by $\mathrm{P}((T,Z,R) = (t,z,r))$. Hence we find

$$\mathrm{P}(\mathcal{U}^{nc}_{(t,z,r)}(\epsilon)) \leq \frac{\exp(-n^\alpha)}{\mathrm{P}((T,Z,R) = (t,z,r))}. \tag{5.1.35}$$

We can next use the lower bound on $\mathrm{P}((T,Z,R) = (t,z,r))$ provided by lemma 5.1.2 for all values $(t,z) \in D$ and $r < l+1$ to inequality 5.1.35 and obtain

$$\mathrm{P}(\mathcal{U}^{nc}_{(t,z,r)}(\epsilon)) \leq \frac{1}{k_0} \cdot n \cdot \exp(-n^\alpha). \tag{5.1.36}$$

which still gives an exponentially small bound in n. Applying now 5.1.36 to inequality 5.1.34, we obtain

$$P(\mathcal{U}^{nc}(\epsilon)) \leq \frac{4lc^2}{k_0} \cdot n^2 \cdot \exp(-n^\alpha). \tag{5.1.37}$$

Which is an exponentially small bound in n. Note that when the event $\mathcal{U}^n(\epsilon)$ holds, we have that $z \mapsto L_n(t,z,r)$ behaves like a random walk with drift ϵ. Let us formalize this. As before, let $\{z_0, z_1, z_2, \dots, z_d\}$ be the set for the admissible values of Z. For fixed $t \in D_T$ and $r < l+1$, we are going to define $L_n^*(t,z)$ inductively for $z \in \{z_0, z_1, z_2, \dots, z_d\}$. Let us define $L_n^*(t,z,r) := L_n(t,z,r)$ for every $z \in \{z_0, z_1, z_2, \dots, z_d\}$. Given $\tilde{z} \in \{z_0, z_1, z_2, \dots, z_d - 4\}$ let us define $L_n^*(t, \tilde{z}+4, r)$ as follows:

$$L_n^*(t,\tilde{z}+4,r) = \begin{cases} L_n(t,\tilde{z}+4,r) & \text{if } \mathcal{U}^n_{(t,s,r)}(\epsilon) \text{ hold for all } s \in \{z_0, z_1, \dots, \tilde{z}\} \\ L_n^*(t,\tilde{z},r) + \epsilon & \text{otherwise} \end{cases}$$

Note that when the event $\mathcal{U}^n(\epsilon)$ holds, then $L_n(t,z,r)$ and $L_n^*(t,z,r)$ are identical for all $t \in D_T$, $r < l+1$ and $z \in \{z_0, z_1, \dots, z_d\}$. Let $\mathcal{V}^n_{t,r}(\epsilon)$ be the event that the map

$$D_Z \to \mathbb{N} \; : \; z \mapsto L_n^*(t,z,r)$$

is increasing with a slope of at least $\epsilon/8$ on a scale $c_2 \ln n$.

Let $\mathcal{V}^n(\epsilon)$ be the event

$$\mathcal{V}^n(\epsilon) := \bigcap_{t \in D_T,\, r<l+1} \mathcal{V}^n_{t,r}(\epsilon).$$

Hence by using proposition 5.1.2 we have that:

$$P(\mathcal{V}^{nc}(\epsilon)) \leq \sum_{t \in D_T,\, r<l+1} P(\mathcal{V}^{nc}_t(\epsilon)) \leq \sum_{t \in D_T,\, r<l+1} 2n^{-\tau} \leq 4lc\, n^{0.5-\tau} \tag{5.1.38}$$

where $\tau = \frac{\epsilon^2 c_2}{32}$. When $\mathcal{U}^n(\epsilon)$ holds then $\mathcal{V}^n(\epsilon)$ and $U^n(\epsilon)$ are equivalent. Hence

$$\mathcal{U}^n(\epsilon) \cap \mathcal{V}^n(\epsilon) \subset U^n(\epsilon)$$

Hence by using 5.1.37 and 5.1.38 we get:

$$P(U^{nc}(\epsilon)) \leq P(\mathcal{U}^{nc}(\epsilon)) + P(\mathcal{V}^{nc}(\epsilon)) \leq \frac{4lc^2}{k_0} \cdot n^2 \cdot \exp(-n^\alpha) + 4lc\, n^{0.5-\tau} \tag{5.1.39}$$

To show that the last inequality gives us a rate of convergence to zero as a constant divided by a polynomial in n, we try now to get a closed form for the inequality supposing extra information for the involved constants.

Taking $c_2 \geq \frac{80}{\epsilon^2}$ we have the following bound for the exponent:

$$0.5 - \tau \leq -2$$

therefore we can bound

$$4lc\, n^{0.5-\tau} \leq \frac{4lc}{n^2}\,. \tag{5.1.40}$$

Also, we have that:

$$n^2 \exp\left(n^{-\alpha}\right) \leq \frac{1}{n^2} \tag{5.1.41}$$

holds for n large enough. So, by using 5.1.40 and 5.1.41 in 5.1.39 we can finally bound:

$$\begin{aligned} \mathrm{P}(U^{nc}(\epsilon)) \leq \mathrm{P}(\mathcal{U}^{nc}(\epsilon)) + \mathrm{P}(\mathcal{V}^{nc}(\epsilon)) &\leq 4lc^2\tilde{c}_2 n^2 \cdot \exp(-n^{\alpha}) + 4lc\, n^{0.5-\tau} \\ &\leq (4lc^2\tilde{c}_2 + 4lc) \cdot \frac{1}{n^2} \end{aligned}$$

for n large enough, which ends the proof with $k_* = 4lc^2 k_0 + 4lc$. ■

Proposition 5.1.2 *Given $\epsilon > 0$, let $\mathcal{V}^n_{t,r}(\epsilon)$ denote the event that the map $z \mapsto L^*(t,z,r)$ is increasing with a slope at least $\epsilon/8$ on a scale $c_2 \ln(n)$. Given $t \in D_T$, $r < l+1$ and $z_1, z_2 \in D_Z$ such that $z_2 - z_1 \geq c_2 \ln(n)$ we have the following inequality:*

$$\mathrm{P}\left(\mathcal{V}^{nc}_{t,r}(\epsilon)\right) \leq 2\, n^{-\tau}$$

where $\tau = \frac{\epsilon^2\, c_2}{32}$.

Proof. Let $z_1, z_2 \in D_Z$ such that $z_1 < z_2$. In order to simplify the notation, let us assume that $z_2 - z_1$ can be dived by 4 and denote $\frac{z_2 - z_1}{4} = m \in \mathbb{N}$. Let z_0 be the most left point of D_Z. Given $\epsilon > 0$, let us remember that $\mathcal{V}^n_{t,r}(\epsilon)$ is the event such that the following inequality holds:

$$L^*(t,z_2,r) - L^*(t,z_1,r) \geq \frac{\epsilon}{8}.$$

Now let us define the filtration $\mathfrak{F}_0 \subset \mathfrak{F}_1 \subset \cdots \subset \mathfrak{F}_m$ as follows:

$$\mathfrak{F}_i := \sigma\left(X_{(t,z_0,r)}, X_{(t,z_1,r)}, \ldots, X_{(t,z_1+4i,r)}\,;\, Y\right)$$

for $i = 1, \ldots, m$. Let us denote

$$e_i = E[\, L^*_n(t, z_1 + 4(i+1), r) - L^*_n(t, z_1 + 4i, r) \,|\, \mathfrak{F}_i\,]$$

and define a martingale $M_0, M_1, \ldots, M_m$ with respect to the filtration $\mathfrak{F}_0 \subset \mathfrak{F}_1 \subset \cdots \subset \mathfrak{F}_m$ as follows:

$$\begin{aligned} M_0 &:= L^*_n(t,z_1,r) \\ M_{i+1} - M_i &:= L^*_n(t, z_1 + 4(i+1), r) - L^*_n(t, z_1 + 4i, r) - e_i \end{aligned}$$

for $i = 1, \dots, m$. By definition of the map $z \mapsto L_n^*(t, z, r)$ we have an expected increase of at least ϵ every time z gets increased by 4, so that the expected increase of

$$E[\, L_n^*(t, z_1 + 4(i+1), r) - L_n^*(t, z_1 + 4i, r)\,]$$

is at least ϵ which implies that the following inequality

$$e_i \geq \epsilon \tag{5.1.42}$$

is satisfied almost surely for every $0 = 1, \dots, m$. We can write the increase of the map $z \mapsto L_n^*(t, z)$ in terms of the martingale $M_0, \dots, M_m$ in the following way:

$$L^*(t, z_2, r) - L^*(t, z_1, r) = M_m - M_0 + \sum_{i=0}^{m-1} e_i \tag{5.1.43}$$

Now, we are ready to estimate the probability of $\mathcal{V}_{t,r}^{nc}(\epsilon)$:

$$\begin{aligned}
\mathrm{P}\left(\mathcal{V}_{t,r}^{nc}(\epsilon)\right) &= \mathrm{P}\left(L^*(t, z_2, r) - L^*(t, z_1, r) \leq \frac{\epsilon}{8}(z_2 - z_1)\right) \\
\text{(by equality 5.1.43)} &\leq \mathrm{P}\left(M_m - M_0 + \sum_{i=0}^{m-1} e_i \leq \frac{\epsilon}{8}(z_2 - z_1)\right) \\
&= \mathrm{P}\left(M_m - M_0 \leq \frac{\epsilon}{8}(z_2 - z_1) - \sum_{i=0}^{m-1} e_i\right) \\
\text{(by 5.1.42 and } z_2 - z_1 = 4m) &\leq \mathrm{P}\left(M_m - M_0 \leq \frac{\epsilon}{8}(z_2 - z_1) - \frac{\epsilon}{4}(z_2 - z_1)\right) \\
&= \mathrm{P}\left(M_m - M_0 \leq -\frac{\epsilon}{8}(z_2 - z_1)\right) \qquad (5.1.44)
\end{aligned}$$

At this point we want to use Azuma-Hoeffding inequality 4.0.1. For this, we note that for every $i = 1, \dots, m$ we have

$$\mathrm{P}(|M_{i+1} - M_i| \leq 1) = 1$$

since $\epsilon < 1$ and we take $v = \frac{\epsilon}{8}(z_2 - z_1)$ for writing down:

$$\begin{aligned}
\mathrm{P}\left(M_m - M_0 \leq -\frac{\epsilon}{8}(z_2 - z_1)\right) &\leq 2\exp\left(-\frac{v^2}{2m}\right) \\
\text{(by using } z_2 - z_1 = 4m) &= 2\exp\left(-\frac{\epsilon^2}{32}(z_2 - z_1)\right) \qquad (5.1.45)
\end{aligned}$$

Combining together 5.1.44 and 5.1.45 we finally have:

$$\mathrm{P}\left(\mathcal{V}_{t,r}^{nc}(\epsilon)\right) \leq 2\exp\left(-\frac{\epsilon^2}{32}(z_2 - z_1)\right)$$

from where, after taking $z_2 - z_1 \leq c_2 \ln(n)$, we have:

$$\mathrm{P}\left(\mathcal{V}^{nc}_{t,r}(\epsilon)\right) \leq 2\exp\left(-\frac{\epsilon^2\,c_2}{32}\ln(n)\right) = 2\,n^{-\frac{\epsilon^2\,c_2}{32}}$$

which finishes the proof ■

Note that by law of total probability E[VAR[$L_n(T_D, Z_D, R)|T_D, R$]] is equal to :

$\mathrm{P}(U^n(\epsilon))\mathrm{E}[\ \mathrm{VAR}[L_n(T_D,Z_D,R)|T_D,R]\ |U^n(\epsilon)]+\mathrm{P}(U^{nc}(\epsilon))\mathrm{E}[\ \mathrm{VAR}[L_n(T_D,Z_D,R)|T_D,R]\ |U^{nc}(\epsilon)]$,

for every $\epsilon > 0$ and hence:

$$\mathrm{E}[\ \mathrm{VAR}[L_n(T_D,Z_D,R)|T_D,R]\ ] \geq \mathrm{P}(U^n(\epsilon))\mathrm{E}[\ \mathrm{VAR}[L_n(T_D,Z_D,R)|T_D,R]\ |U^n(\epsilon)] \tag{5.1.46}$$

Now, conditional on the event $U^n(\epsilon)$ holding, we have that the random map:

$$D_Z \to \mathbb{N}\ :\ z \mapsto L_n(t,z,r)$$

has a slope of at least $\epsilon/8$ on a scale of $c_2\ln(n)$ (as in proposition 5.1.2) for any $t \in D_T$ and $r < l+1$, then:

$$\begin{aligned} z_2 - z_1 \geq c_2\ln(n) \quad &\Rightarrow \quad L_n(t,z_2,r) - L_n(t,z_1,r) \leq \frac{\epsilon}{8}(z_2 - z_1) \\ z_2 - z_1 < c_2\ln(n) \quad &\Rightarrow \quad L_n(t,z_2,r) - L_n(t,z_1,r) \leq 2\,(z_2 - z_1) \end{aligned}$$

hold. Hence, conditional on $U^n(\epsilon)$, we can apply lemma 5.0.3 and obtain:

$$\mathrm{VAR}[L_n(t,Z_D,R)|T_D=t,R=r,U^n(\epsilon)] \geq \frac{\epsilon^2}{64}\left(1 - 16\frac{(\epsilon/8+2)c_2\ln(n)}{\epsilon\sqrt{\mathrm{VAR}[Z_D|T_D=t,R=r]}}\right)\mathrm{VAR}[Z_D|T_D=t,R=r] \tag{5.1.47}$$

The next results give us an uniform bound for VAR[$Z_D|T_D = t, R = r$] for all $t \in D_T$.

Lemma 5.1.5 *There exists a constant $K > 0$ not depending on n such that:*

$$1 - \frac{K}{\sqrt{n}} \leq \frac{P(Z_D = z+4|T_D=t,R=r)}{P(Z_D=z|T_D=t,R=r)} \leq 1 + \frac{K}{\sqrt{n}} \tag{5.1.48}$$

for every $(t,z) \in D$, $r < l+1$ and n large enough.

Proof. Note that from 5.1.4 we can get the following relations:

$$\begin{aligned} N_{l-1}(t,z+4,r) &= N_{l-1}(t,z,r) - 1 \\ N_l(t,z+4,r) &= N_l(t,z,r) + 2 \\ N_{l+1}(t,z+4,r) &= N_{l+1}(t,z,r) - 1 \end{aligned}$$

therefore we have an explicit formula for the joint probability by using the above last relations and 5.1.20:

$$\frac{P(Z = z + 4|T = t, R = r)}{P(Z = z|T = t, R = r)} = \frac{N_{l-1}(t, z, r)N_{l+1}(t, z, r)}{(N_l(t, z, r) + 1)(N_l(t, z, r) + 2)} \geq 0 \tag{5.1.49}$$

By using 5.1.8 we can bound the expression in 5.1.49 as follows:

$$\begin{aligned}\frac{\left(\frac{n}{3l} - k_1\sqrt{n}\right)\left(\frac{n}{3l} - k_3\sqrt{n}\right)}{\left(\frac{n}{3l} + 1 + k_2\sqrt{n}\right)\left(\frac{n}{3l} + 2 + k_2\sqrt{n}\right)} &\leq \frac{N_{l-1}(t, z, r)N_{l+1}(t, z, r)}{(N_l(t, z, r) + 1)(N_l(t, z, r) + 2)} \\ &\leq \frac{\left(\frac{n}{3l} + k_1\sqrt{n}\right)\left(\frac{n}{3l} + k_3\sqrt{n}\right)}{\left(\frac{n}{3l} + 1 - k_2\sqrt{n}\right)\left(\frac{n}{3l} + 2 - k_2\sqrt{n}\right)}\end{aligned} \tag{5.1.50}$$

By using the following inequalities for logarithm :

$$\begin{aligned}-\frac{3x}{2} &\leq \ln(1 - x), \quad \text{for } 0 < x \leq 0.5 \\ \ln(1 + x) &\leq x, \quad \text{for } x > -1\end{aligned} \tag{5.1.51}$$

we have on the right hand side:

$$\begin{aligned}\frac{\left(\frac{n}{3l} + k_1\sqrt{n}\right)\left(\frac{n}{3l} + k_3\sqrt{n}\right)}{\left(\frac{n}{3l} + 1 - k_2\sqrt{n}\right)\left(\frac{n}{3l} + 2 - k_2\sqrt{n}\right)} &\leq \frac{\left(1 + \frac{3lk_1}{\sqrt{n}}\right)\left(1 + \frac{3lk_3}{\sqrt{n}}\right)}{\left(1 - \frac{3lk_2}{\sqrt{n}}\right)^2} \\ &= \exp\left[\ln\left(1 + \frac{3lk_1}{\sqrt{n}}\right) + \ln\left(1 + \frac{3lk_3}{\sqrt{n}}\right)\right. \\ &\quad \left.-2\ln\left(1 - \frac{3lk_2}{\sqrt{n}}\right)\right] \\ &\leq \exp\left[\frac{15lk^*}{\sqrt{n}}\right] \leq 1 + \frac{15lk^*}{\sqrt{n}} + |R(\xi)| \\ &\leq 1 + \frac{15lk^*}{\sqrt{n}} + \varepsilon\end{aligned} \tag{5.1.52}$$

after considering the rest form for a Taylor expansion of the function $f(x) = e^x$ and $\xi = \frac{15lk^*}{\sqrt{n}}$ as follows:

$$R(\xi) = \left|\frac{f''(\xi)}{2}\right|\xi^2 \leq \frac{(15lk^*)^2}{2n}\exp\left(\frac{15lk^*}{\sqrt{n}}\right) \leq \varepsilon$$

for n large enough and a given precision ε, and on the left hand side:

$$\begin{aligned}
\frac{\left(\frac{n}{3l}-k_1\sqrt{n}\right)\left(\frac{n}{3l}-k_3\sqrt{n}\right)}{\left(\frac{n}{3l}+1+k_2\sqrt{n}\right)\left(\frac{n}{3l}+2+k_2\sqrt{n}\right)} &\geq \frac{\left(1-\frac{3lk_1}{2\sqrt{n}}\right)\left(1-\frac{3lk_3}{2\sqrt{n}}\right)}{\left(1+\frac{3lk_2}{2\sqrt{n}}\right)^2} \\
&= \exp\left[\ln\left(1-\frac{3lk_1}{2\sqrt{n}}\right)+\ln\left(1-\frac{3lk_3}{2\sqrt{n}}\right)\right. \\
&\quad \left.-2\ln\left(1+\frac{3lk_2}{2\sqrt{n}}\right)\right] \\
&\geq \exp\left[-\frac{9lk^*}{4\sqrt{n}}\right] \geq 1-\frac{9lk^*}{4\sqrt{n}} \qquad (5.1.53)
\end{aligned}$$

Finally, taking $K = 15lk^*$ we have our desired result combining 5.1.52, 5.1.53 and 5.1.50 for any given $\varepsilon > 0$. ■

Lemma 5.1.6 *There exists a constant $C > 0$ not depending on n such that:*

$$\mathrm{VAR}[Z_D|T_D = t, R = r] \geq C \cdot n$$

for every $t \in D_T$, $r < l+1$ and for every n large enough.

Proof. Since 5.1.48 is satisfied we have that Z_D takes almost surely the same value on D_Z conditional on $T_D = t, R = r$. More in details, given $(t,z) \in D$, $r < l+1$ and $k \in \mathbb{Z}$ such that $z + 4k \in D_Z$ we can write down

$$\mathrm{P}(Z_D = z+4k|T_D = t, R = r) = \mathrm{P}(Z_D = z|T_D = t, R = r)\cdot a_1 \cdot a_2 \cdots a_k \quad (5.1.54)$$

where the notation is

$$a_i = \frac{\mathrm{P}(Z_D = z+4(k-i)+4|T_D = t, R = r)}{\mathrm{P}(Z_D = z+4(k-i)|T_D = t, R = r)}$$

for every $i = 1, \ldots, k$. By using 5.1.48 there exists $K > 0$ such that:

$$1 - \frac{K}{\sqrt{n}} \leq a_i \leq 1 + \frac{K}{\sqrt{n}}$$

for every n large enough and every $i = 1, \ldots, k$ which means:

$$\left(1-\frac{K}{\sqrt{n}}\right)^k \leq a_1 \cdot a_2 \cdots a_k \leq \left(1+\frac{K}{\sqrt{n}}\right)^k$$

and finally:

$$\left(1-\frac{K}{\sqrt{n}}\right)^k \leq \frac{\mathrm{P}(Z_D = z+4k|T_D = t, R = r)}{\mathrm{P}(Z_D = z|T_D = t, R = r)} \leq \left(1+\frac{K}{\sqrt{n}}\right)^k \quad (5.1.55)$$

Moreover, when n is large enough, we have better bounds from 5.1.55:

$$\exp\left(-\frac{3K}{2\sqrt{n}}k\right) \leq \frac{\mathrm{P}(Z_D = z + 4k|T_D = t, R = r)}{\mathrm{P}(Z_D = z|T_D = t, R = r)} \leq \exp\left(\frac{K}{\sqrt{n}}k\right) \tag{5.1.56}$$

due to inequalities 5.1.51, which means that for n large enough Z_D has the same probability of taking two values which are $4k$ far away from each other. This only can happen if $\mathrm{VAR}[Z_D|T_D = t, R = r]$ is at least of order n. More precisely, let us take $\delta > 0$ such that:

$$\mathrm{P}\left(|Z_D - \mathrm{E}[Z_D|T_D = t, R = r]| \geq \frac{1}{\sqrt{\delta}}\sigma_Z \,\middle|\, T_D = t, R = r\right) \leq \delta \tag{5.1.57}$$

by using Chebyshev inequality 5.1.12 where $\sigma_Z = \sqrt{\mathrm{VAR}[Z_D|T_D = t, R = r]}$. If we suppose $\sigma_Z < 4\delta\sqrt{\delta}\sqrt{n}$ then 5.1.57 tells us that:

$$\mathrm{P}\left(|Z_D - \mathrm{E}[Z_D|T_D = t, R = r]| \geq 4\delta\sqrt{n} \,\middle|\, T_D = t, R = r\right) \leq \delta \tag{5.1.58}$$

But at the same time 5.1.56 also tells us that:

$$\exp\left(-\frac{3K}{2}\delta - \frac{3K}{2\sqrt{n}}k\right) \leq \frac{\mathrm{P}(Z_D = z + 4(\delta\sqrt{n} + k)|T_D = t, R = r)}{\mathrm{P}(Z_D = z + 4k|T_D = t, R = r)} \leq \exp\left(K\delta + \frac{K}{\sqrt{n}}k\right) \tag{5.1.59}$$

for every $z \in D_Z$ and $k \in \mathbb{Z}$ such that $z + 4(\delta\sqrt{n} + k) \in D_Z$. In particular for $z = \mathrm{E}[Z_D|T_D = t, R = r]$ we have:

$$\begin{aligned}\exp\left(-\frac{3K}{2}\delta - \frac{3K}{2\sqrt{n}}k\right) &\leq \frac{\mathrm{P}(Z_D = \mathrm{E}[Z_D|T_D = t, R = r] + 4(\delta\sqrt{n} + k)|T_D = t, R = r)}{\mathrm{P}(Z_D = \mathrm{E}[Z_D|T_D = t, R = r] + 4k|T_D = t, R = r)} \\ &\leq \exp\left(K\delta + \frac{K}{\sqrt{n}}k\right)\end{aligned}$$

which says that when we choose $\delta > 0$ small enough the probability of taking values which are $4\delta\sqrt{n}$ far away is the same for n large enough, showing a contradiction with 5.1.58. Finally we can take $C = 16\delta^3$ for a choosen δ depending on the precision we want to have in 5.1.58. ■

Using the bound in lemma 5.1.6 we get the following inequality:

$$\left(1 - 16\frac{(\epsilon/8 + 2)c_2 \ln(n)}{\epsilon\sqrt{\mathrm{VAR}[Z_D|T_D = t, R = r]}}\right) \geq \left(1 - 16\frac{(\epsilon/8 + 2)c_2}{\epsilon\sqrt{C}} \cdot \frac{\ln(n)}{\sqrt{n}}\right) \geq 0.5 \tag{5.1.60}$$

for n large enough. Using inequality 5.1.60 above with inequality 5.1.47 we find:

$$\mathrm{VAR}[L_n(t, Z_D, R)|T_D = t, R = r, U^n(\epsilon)] \geq \frac{\epsilon^2}{64}\, 0.5 \cdot \mathrm{VAR}[Z_D|T_D = t, R = r].$$

Using again lemma 5.1.6 we find that the left side of the above inequality is larger than

$$\frac{C\epsilon^2}{128}n$$

and hence:

$$\mathrm{E}[\ \mathrm{VAR}[L_n(T_D, Z_D, R)|T_D, R]\ |U^n(\epsilon)\ ] \geq \frac{C\epsilon^2}{128}n \tag{5.1.61}$$

We can now combine inequalities 5.1.26, 5.1.27, 5.1.28, 5.1.46 and 5.1.61 to obtain:

$$\mathrm{VAR}[L_n] \geq \mathrm{P}(U^n(\epsilon))\frac{C\epsilon^2}{1000}n$$

and plugging in the lower bound for $\mathrm{P}(U^n(\epsilon))$ obtained in 5.1.32 (lemma 5.1.4) we get:

$$\mathrm{VAR}[L_n] \geq \frac{C\epsilon^2}{1000}\,n\left(1 - \frac{k_*}{n^2}\right)$$

with $k_* > 0$ is the constant from lemma 5.1.4. This expression is a lower bound of order $\Theta(n)$ for $\mathrm{VAR}[L_n]$. Hence, we have finished proving the statement of our main result in theorem 2.1.1.

Chapter 6

Solution of the Optimization Problem

We have so far shown in section 5 that a likely biased effect of the random modification leads to the desired order $\mathrm{VAR}[L_n] = \Theta(n)$. This means that if

$$\mathrm{E}[\tilde{L}_n - L_n \mid X, Y] \geq \epsilon \tag{6.0.1}$$

holds with high enough probability, then the desired fluctuation order $\mathrm{VAR}[L_n] = \Theta(n)$ follows. Theorem 2.1.3 allows to prove the high probability bias for the random modification as in 6.0.1. More, precisely according to theorem 2.1.3, the bias 6.0.1 follows as soon as the minimizing problem in theorem 2.1.3 has a solution greater or equal to 2ϵ. For the bias 6.0.1 to be have high probability (see 2.1.23), one needs to take the input variable q_0 in the optimization problem to be such that with high probability there is not more than a proportion q_0 of left out blocks in the optimal alignment. This means that we need $F^{nc}(q_0)$ to have exponentially small probability in a fractional power in n. Note that the whole system depends on l only through $q_0 = q_0(l)$. One simple upper bound for the number of left out blocks is $4/(9(l-1))$ (see 3.2.4). We can take q_0 equal to $4/(9(l-1))$ and then check if the optimization problem in theorem 2.1.3 has a strictly positive solution. If yes, then it follows that $\mathrm{VAR}[L_n] = \Theta(n)$. We have already explained that if we take l large enough (and hence $q_0 = q_0(l)$ small enough) then the optimization problem in theorem 2.1.3 has a strictly positive solution and hence $\mathrm{VAR}[L_n] = \Theta(n)$ for l large enough (see the proof of 2.1.3). The goal of this section is to verify numerically, that this already holds for l not too large. We actually find that linear order for every $l \geq 5$, i.e.

$$\mathrm{VAR}[L_n] = \Theta(n) \ , \ \forall\, l \geq 5.$$

To prove this we simply have to check that for $l \geq 5$ the optimization problem in theorem 2.1.3, namely

$$\min\left(\frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}}(1-9q) - \frac{p_{l+1,l+1}}{p_{l+1,l-1} + p_{l+1,l} + p_{l+1,l+1}}(1-3q) - 3q\right)$$

under the conditions

$$q \in [0, q_0], \sum_j p_{l-1,j} \geq ((1/3) - q_0)/2 \ , \ \sum_j p_{l+1,j} \geq ((1/3) - q_0)/2$$

$$\sum_{i,j \in I} p_{ij} = 1, p_{ij} \geq 0, \forall i, j \in I$$

$$2H(q) + (1 - 4q)\,(\ln(1/9) + H(p)) \geq 0$$

has a strictly positive solution. Again, we need a bound for q_0. We also use a better bound than $4/9(l-1))$ which is obtained by Montercarlo simulation (see subsection 6.2.2).

One important thing about the minimizing problem is that when we take q smaller, the constraint becomes stricter and the objective function increases. Hence, we can replace in the objective function and the constrain q by q_0.

6.1 Parametrical solution

We are going to transform the minimizing problem into a problem with a linear objective function by introducing two help variables κ_1 and κ_2. We would like to use the so called *Lagrange method.* As a first approach to do so, for fixed l and $q = q_0$, let us re-write the minimization problem above in an equivalent linear version parametrized by $\kappa_1, \kappa_2 > 0$ as follows:

$$\begin{aligned}
\min \quad F(p) &:= \alpha p_{12} + \alpha p_{13} - \beta p_{33} - 3q_0 && \\
G_1(p) := p_{11} + p_{12} + p_{13} - \kappa_1 &= 0 && (6.1.1)\\
G_2(p) := p_{31} + p_{32} + p_{33} - \kappa_2 &= 0 && (6.1.2)\\
\sum_{i,j=1}^{3} p_{ij} - 1 &= 0 && \\
-p_{ij} &\leq 0 && \\
\mathcal{H}(p) := -2H(q_0) - (1 - 4q_0)(\ln(1/9) + H(p)) &\leq 0 && (6.1.3)
\end{aligned}$$

where the simplified notation is the following:

$$\begin{pmatrix} p_{l-1,l-1} & p_{l-1,l} & p_{l-1,l+1} \\ p_{l,l-1} & p_{l,l} & p_{l,l+1} \\ p_{l+1,l-1} & p_{l+1,l} & p_{l+1,l+1} \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$

$$\alpha = \frac{1 - 9q_0}{\kappa_1}$$

$$\beta = \frac{1 - 3q_0}{\kappa_2}$$

Note that the straight forward relation between the variables in our simplified notation and the variables in the orginal problem. When we read conditions 6.1.1 and 6.1.2 we get:

$$\begin{aligned}\kappa_1 &= p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}\\ \kappa_2 &= p_{l+1,l-1} + p_{l+1,l} + p_{l+1,l+1}\end{aligned}$$

We consider the conditions

$$\sum_j p_{l-1,j} \geq ((1/3) - q_0)/2 \ , \ \sum_j p_{l+1,j} \geq ((1/3) - q_0)/2$$

as inside an optimal region for the parameters κ_1 and κ_2 described by the inequalities:

$$\frac{1/3 - q_0}{2} \leq \kappa_1, \kappa_2 \leq 2/3 + q_0$$

The idea behind this scheme is to reduce the complexity of computing the optimal conditions for the matrix p since the original expression

$$\min\left(\frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}}(1-9q) - \frac{p_{l+1,l+1}}{p_{l+1,l-1} + p_{l+1,l} + p_{l+1,l+1}}(1-3q) - 3q\right)$$

is non-linear in $(p_{ij})_{i,j}$. We will only consider 6.1.1,6.1.2 and 6.1.3 as active constrains for the computation of the gradients in the *Lagrange* formulation, the rest of the constraints will be considered especifically to reduce the degree of freedom through the computations.

Now we start the *Lagrange method*. Let us compute the gradients:

$$\begin{aligned}\left(\frac{\partial G_1}{\partial p_{ij}}\right)_{ij} &= \begin{pmatrix} 1 & 1 & 1\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}\\ \left(\frac{\partial G_2}{\partial p_{ij}}\right)_{ij} &= \begin{pmatrix} 0 & 0 & 0\\ 0 & 0 & 0\\ 1 & 1 & 1\end{pmatrix}\end{aligned}$$

$$\begin{aligned}\left(\frac{\partial \mathcal{H}}{\partial p_{ij}}\right)_{ij} &= (1-4q_0)\left\{\begin{pmatrix} 1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{pmatrix} + \begin{pmatrix} \ln p_{11} & \ln p_{12} & \ln p_{12}\\ \ln p_{21} & \ln p_{22} & \ln p_{23}\\ \ln p_{31} & \ln p_{32} & \ln p_{33}\end{pmatrix}\right\}\\ \left(\frac{\partial F}{\partial p_{ij}}\right)_{ij} &= \begin{pmatrix} 0 & \alpha & \alpha\\ 0 & 0 & 0\\ 0 & 0 & -\beta\end{pmatrix}\end{aligned}$$

Then the *Lagrange* optimal condition for the lagrange multipliers λ for $\mathcal{H}$, η_1 and η_2 for G_1 and G_2 respectively is

$$\left(\frac{\partial F}{\partial p_{ij}}\right)_{ij} = \lambda \left(\frac{\partial \mathcal{H}}{\partial p_{ij}}\right)_{ij} + \eta_1 \left(\frac{\partial G_1}{\partial p_{ij}}\right)_{ij} + \eta_2 \left(\frac{\partial G_2}{\partial p_{ij}}\right)_{ij}$$

which leads to the following expressions for $(p_{ij})_{ij}$ as functions of λ, η_1, η_2 :

$$\begin{aligned}
p_{11} &= \exp\left(\frac{\eta_1}{\lambda(1-4q_0)} - 1\right) = c^{-\eta_1}e^{-1} \\
p_{12} = p_{13} &= \exp\left(\frac{\alpha-\eta_1}{\lambda(1-4q_0)} - 1\right) = c^{\alpha-\eta_1}e^{-1} \\
p_{21} = p_{22} = p_{23} &= \frac{1-\kappa_1-\kappa_2}{3} =: d \\
p_{31} = p_{32} &= \exp\left(\frac{-\eta_2}{\lambda(1-4q_0)} - 1\right) = c^{-\eta_2}e^{-1} \\
p_{33} &= \exp\left(\frac{-\beta-\eta_2}{\lambda(1-4q_0)} - 1\right) = c^{-\beta-\eta_2}e^{-1}
\end{aligned}$$

where $c := \exp(\frac{1}{\lambda(1-4q_0)})$. Here, to determine the constant d we used that all the $\{p_{ij}\}_{ij}$ sum up to one,

$$\kappa_1 + 3d + \kappa_2 = 1.$$

In matrix notation the optimal form of p looks like:

$$p = \begin{pmatrix} c^{-\eta_1}e^{-1} & c^{\alpha-\eta_1}e^{-1} & c^{\alpha-\eta_1}e^{-1} \\ d & d & d \\ c^{-\eta_2}e^{-1} & c^{-\eta_2}e^{-1} & c^{-\beta-\eta_2}e^{-1} \end{pmatrix} \tag{6.1.4}$$

Note that η_1, η_2 can be expressed both in terms of κ_1, κ_2 and c in the following way by using expressions 6.1.1, 6.1.2 and the positivity and normalized conditions on $(p_{ij})_{ij}$:

$$\eta_1 = -\frac{\ln \frac{\kappa_1 e}{1+2c^{\alpha}}}{\ln c} \tag{6.1.5}$$

$$\eta_2 = -\frac{\ln \frac{\kappa_2 e}{2+c^{-\beta}}}{\ln c} \tag{6.1.6}$$

6.2 Numerical solution

For a numerical solution, we should decide between having a difficult expression for p as a function of (κ_1, κ_2, c) after replacing η_1, η_2 from 6.1.5, 6.1.6 in the optimal form 6.1.4 or having a more simplified expression for p as a function of more parameters $(\kappa_1, \kappa_2, c, \eta_1, \eta_2)$ but performing an extra searching for the values of

η_1, η_2.

We prefer to use the second strategy since the expression in the first case is higly implicit for the method of numerical solution implemented in MATLAB which could lead to less accuracy in the results.

We will proceed in detail as follows: let us fix the constants $l > 0$ and κ_1, κ_2 for now on. In general we will consider that q depends on l, so we compute q. The parameter c depends on the multiplier λ which is related to the inequality of the entropy 6.1.3 then we determine c looking at the entropy condition:

$$\begin{aligned} c^{-\eta_1}(1+\eta_1 \ln c) + 2c^{\alpha-\eta_1}(1-(\alpha-\eta_1)\ln c) + 2c^{-\eta_2}(1+\eta_2 \ln c) &\quad + \\ c^{-\beta-\eta_2}(1+(\beta+\eta_2)\ln c) + 3de\ln(1/d) + e\cdot\left(\frac{2H(q_0)}{1-4q_0} + \ln(1/9)\right) &\quad = \quad 0 \end{aligned}$$

For each value of η_1, η_2 we get a $c = c(\eta_1, \eta_2)$ solution of the equation above. This solution is obtained by using in MATLAB the routine *fsolve* which basically uses an optimization method to find the zeros of an non-linear function.

With that c we compute the updated values of η_1^*, η_2^* as a function of c by using expressions 6.1.5, 6.1.6. After that we use expression 6.1.4 for writing down the last form of $p^* = p(\eta_1^*, \eta_2^*)$. Then we perform our searching of $\eta_1^* \in (\alpha, 1]$ and $\eta_2^* \in (-\beta, 1]$ such that: $\min\{F(p^*)\} > 0$, since we would like to fullfill the condition $\mathrm{VAR}[L_n] = \Theta(n)$. Finally the next iteration is to pick up another value of κ_1, κ_2 and repeat the same as before. At the end we will have an optimal value for κ_1, κ_2, let us denote them κ_1^*, κ_2^*.

In order to have more realistic results we will study three types of dependency of q on l, namely the *basic form*, the *simulation form* and the *entropy form*.

6.2.1 The basic form

As in equation 3.2.4 we consider q_0 to be:

$$q_0 = \frac{4}{9(l-1)}$$

The numerical results for the minimum, in this case, are shown in table 6.1.

6.2.2 The simulation form

We consider q_0 to be:

$$q_0 = \frac{1-\gamma_l}{l-1}$$

where $\gamma_l = \lim_{n\to\infty} \mathrm{E}[L_n]/n$ is the value of the Chvatal-Sankoff constant [6] for X, Y being two sequences of uniformly i.i.d blocks. It is not difficult to prove

l	p^*	$F(p^*)$	$H(p^*)$	q^*
10	$\begin{pmatrix} 0.191 & 0.2255 & 0.2255 \\ 0.072 & 0.072 & 0.072 \\ 0.0613 & 0.0613 & 0.0194 \end{pmatrix}$	0.1257	0.948	0.0494

Table 6.1: solution values for $l = 10, \dots, 20$ taking the *basic* form for q.

that γ_l exists due to a sub-additivity argument. We also simulated γ_l by using a *home made* script in MATLAB. The numerical results for the minimum, in this case, are shown in table 6.2.

l	γ_l	p^*	$F(p^*)$	$H(p^*)$	q^*
5	0.9131	$\begin{pmatrix} 0.1778 & 0.239 & 0.239 \\ 0.0295 & 0.0295 & 0.0295 \\ 0.106 & 0.106 & 0.0439 \end{pmatrix}$	0.3609	0.7803	0.0217

Table 6.2: solution values for $l = 5, \dots, 20$ taking the *simulation* form for q.

6.2.3 The entropy form

Let $W_1, W_2, \dots$ be a sequence of i.i.d variables with distribution:

$$P(W_i = l-1) = 5/9, P(W_i = l) = 3/9, P(W_i = l+1) = 1/9$$

for $i = 1, 2, \dots$ Hence W_i is distributed like B_i the minimum of two independent blocks of X and Y. We argued in section 3.1 that we only need to consider alignments where there are no left out blocks next to each other. In the part combinatorics we already noted that we never have several blocks of X aligned together with several blocks of Y. Note also that when in an alignment we specify the left out blocks in a non-random manner, then the aligned block pairs become independent of each other. They also are independent of the polygamist blocks and their families. We consider here alignments with all of those above properties. Hence if we have n_1^* blocks aligned one block with one block and n_2^* polygamist blocks aligned with several blocks, then we have that the score of an optimal alignment, under this properties, behaves at least like:

$$W_1 + W_2 + \dots + W_{n_l^*} + (l-1)n_2^*.$$

With q_0 blocks missing we have in the alignment at least a proportion $(1 - 3q_0)$ of blocks aligned one to one and at most a proportion of q_0 polygamist blocks. Hence, we have at least $(n/l)(1-3q_0)$ blocks which are aligned on block with one block and at most $(n/l)q_0$ polygamist blocks, then the score of such an optimal alignment behaves at least like:

$$W_1 + \ldots + W_{\frac{n}{l}(1-3q_0)} + (l-1)\frac{n}{l}q_0 \qquad (6.2.1)$$

On the other hand, note that typically for large n the LCS is close to $\gamma_l n$. Hence, if an alignment with q_0 left-out blocks should not be optimal we would need the sum 6.2.1 to be close to $\gamma_l n$ or above:

$$W_1 + \ldots + W_{\frac{n}{l}(1-3q_0)} + (l-1)\frac{n}{l}q_0 \geq \gamma_l\, n \qquad (6.2.2)$$

Assuming that we leave out a proportion q_0 of blocks, there are about $e^{2H(q_0)n/l}$ ways of specifying which blocks are left out in the alignment. Hence for q_0 to not get excluded as candidate for the proportion of left out blocks in an optimal alignment, we need that the probability for the sum 6.2.2 times $e^{2H(q_0)n/l}$ must have exponentially small probability. If we denote

$$\begin{aligned} m &:= (1-3q_0)\frac{n}{l} \\ c &:= -\left(\frac{(l-1)q_0 - l\gamma_l}{1-3q_0}\right) \\ \kappa(l,c) &:= -\ln\left[\min_{t\geq 0}\left\{e^{-tc}\left(\frac{5}{9}e^{(l-1)t} + \frac{3}{9}e^{lt} + \frac{1}{9}e^{(l+1)t}\right)\right\}\right], \end{aligned}$$

we should verify that:

$$\begin{aligned} \mathrm{P}\left(\sum_{i=1}^{\frac{n}{l}(1-3q_0)} W_i + (l-1)\frac{n}{l}q_0 \geq \gamma_l\, n\right)\cdot e^{2H(q_0)n/l} &= \mathrm{P}\left(\frac{1}{m}\sum_{i=1}^{m} W_i - c \geq 0\right)\cdot e^{2H(q_0)n/l} \\ &\leq e^{(-\kappa(l,c)+\frac{2H(q_0)}{l})n} \end{aligned}$$

is an exponential small bound, which holds if:

$$-\kappa(l,c) + \frac{2H(q)}{l} \leq 0 \qquad (6.2.3)$$

Note that $\kappa(l,c)$ above comes from a large deviation approach. Based in this way of thinking and given l, γ_l we consider q_0 to be a solution of the inequality 6.2.3. Here we need to solve an extra minimization problem associated to computing $\kappa(l,c)$. The numerical results for the minimum of $F(p)$, in this case, are shown in table 6.3.

Finally, we compare in figure 6.1 the minimun values of $F(p)$ taking the *basic*, *simulation* and *entropy* forms for q_0 with values $l = 6, \ldots, 30$. We clearly see that the so called *entropy form* is the most accurate one in reaching $F(p) = 0$.

l	γ_l	p^*	$F(p^*)$	$H(p^*)$	q^*
5	0.9131	$\begin{pmatrix} 0.1785 & 0.2398 & 0.2398 \\ 0.0279 & 0.0279 & 0.0279 \\ 0.1062 & 0.1062 & 0.0457 \end{pmatrix}$	0.3983	0.7486	0.017

Table 6.3: solution values for $l = 5, \ldots, 20$ taking the *entropy* form for q.

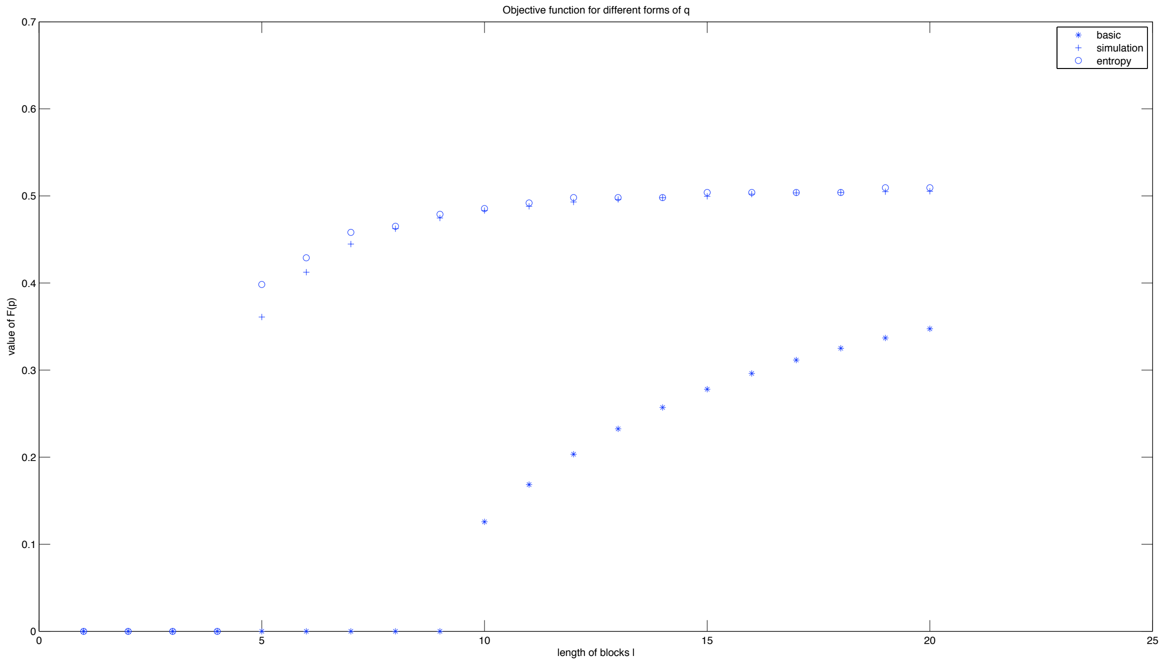

Figure 6.1: values of the objective function using the *basic, simulation* and *entropy* form for q_0.

Appendix A

Proof of Theorem 2.1.3

Chapter 4 contains all the lemmas and definitions needed for the proof of theorem 2.1.3. We already have a rough idea why theorem 2.1.3 holds: we found for the expected score increase

$$\mathrm{E}[\tilde{L}_n - L_n|X, Y] \tag{A.0.1}$$

a lower bound, given by the expression 2.1.10:

$$\frac{p_{l-1,l} + p_{l-1,l+1}}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}}(1 - 9q) - \frac{p_{l+1,l+1}}{p_{l+1,l-1} + p_{l+1,l} + p_{l+1,l+1}}(1 - 3q) - 3q,$$

where $\{p_{i,j}\}_{i,j}$ denotes the empirical distribution of the aligned block pairs in an optimal alignment. We also found that only the distributions satisfying 2.1.18 are likely to happen. So, if the expression 2.1.10 is bounded from below by 2ϵ under the condition 2.1.18, then it is likely that for all optimal alignments, the expression 2.1.10 stays above 2ϵ. This gives a lower bound of 2ϵ for the expected increase. In reality, we will only get a likely expected increase of ϵ instead of 2ϵ. The reason is that to stay on the safe side, we consider that the expression on the left hand side of inequality 2.1.18 needs to be smaller than a negative quantity $-\epsilon_2 < 0$ (in order to have an unlikely corrresponding alignment) instead of just being smaller than 0.

The main problem with the above simplified description of the proof is that to obtain the lower bound 2.1.10 we had to make assumptions which do not exactly hold. These assumptions are the following:

1. We assumed that the propotion of left out blocks in X and in Y is the same. In reality, it is not exactly the same. We denote instead by q_1, resp. q_2 the proportion of left out blocks in X and in Y.

2. We assumed that the proportions of blocks in X for the different lengths $l - 1, l, l + 1$ are exactly equal to 1/3. We assumed the same for Y. In reality, these proportions are typically of order $\Theta(1/\sqrt{n})$ away from 1/3. In

section 4.3, we define $D^n(\delta)$ to be the event that all these proportions are not further from 1/3 than δ.

3. We assumed that the number of blocks in X and Y are exactly equal. In reality, there is typically a difference of order $\Theta(\sqrt{n})$. In section 4.4, we define $G^n(\delta)$ to be the event that the number of blocks in X is not way bigger than the number of blocks in Y. More precisely, $G^n(\delta)$ is defined to be the event that
$$\frac{N_n^X}{N_n^Y} \leq 1+\delta,$$
where N_n^X, resp. N_n^Y, is the number of blocks in X, resp. Y.

In the proof of lemma 4.7.1, it is shown that when $D^n(\delta)$ and $G^n(\delta)$ hold, then the expression

$$\begin{aligned} &\frac{p_{l-1,l}+p_{l-1,l+1}}{p_{l-1,l-1}+p_{l-1,l}+p_{l-1,l+1}}\left(1-\frac{3q_1}{(1/3)-\delta}\right) + \\ -&\frac{p_{l+1,l+1}}{p_{l+1,l-1}+p_{l+1,l}+p_{l+1,l+1}}\left(1-\frac{\delta+2(q_1-\delta)}{(1/3)-\delta}\right)-\frac{q_2(1+\delta)}{(1/3)-\delta} \end{aligned} \tag{A.0.2}$$

bounds from below the expected increase $\mathrm{E}[\tilde{L}_n - L_n|X,Y]$ instead of the expression 2.1.10 (here $\{p_{ij}\}_{i,j}$ is the empirical distribution of the aligned block pairs of an optimal alignment). When we consider different proportions of left out blocks in X and Y, then the expression on the left hand side of 2.1.22 is replaced by the following expression:

$$H(q_1)+H(q_2)+(1-\max\{q_1+3q_2, 3q_1+q_2\})\left(\sum_{i,j\in\{l-1,l,l+1\}} p_{ij}\ln(1/3)+H(p)\right) \tag{A.0.3}$$

To prove theorem 2.1.3, we assume that there exists $q_0>0$ such that the minimum in theorem 2.1.3 under the constrains of theorem 2.1.3 is bounded from below by $2\epsilon>0$. We need to prove that with high probability, if $F^n(q_0)$ holds then the expected increase in score is larger than ϵ, i.e.

$$\mathrm{E}[\tilde{L}_n - L_n|X,Y] \ \geq \epsilon.$$

Take the additional condition

$$|q_1-q_2| \leq 2\delta \tag{A.0.4}$$

Note that with this additional condition, when $\delta \to 0$ then the expression A.0.2 converges to the expression 2.1.19 in theorem 2.1.3. Also, the expression A.0.3 converges to the expression on the right hand side of inequality 2.1.22 in theorem 2.1.3. So, by continuity, we get that for $\delta>0$ small enough "the conditions A.0.3

and A.0.2" behave similarly to the conditions in theorem 2.1.3. More precisely, there exists $\delta_0 > 0$ such that: if $|q_1 - q_2| \leq 2\delta_0$, $0 < q_1, q_2 \leq q_0$ and $|\delta| \leq \delta_0$, then

$$H(q_1)+H(q_2)+(1-\max\{q_1+3q_2, 3q_1+q_2\})\left(\sum_{i,j\in\{l-1,l,l+1\}} p_{ij}\ln(1/3) + H(p)\right) \geq -\epsilon_2 \tag{A.0.5}$$

implies that

$$\begin{aligned} &\frac{p_{l-1,l}+p_{l-1,l+1}}{p_{l-1,l-1}+p_{l-1,l}+p_{l-1,l+1}}\left(1-\frac{3q_1}{(1/3)-\delta}\right) + \\ -&\frac{p_{l+1,l+1}}{p_{l+1,l-1}+p_{l+1,l}+p_{l+1,l+1}}\left(1-\frac{\delta+2(q_1+\delta)}{(1/3)-\delta}\right)-\frac{q_2(1+\delta)}{(1/3)-\delta} \end{aligned} \tag{A.0.6}$$

is larger than ϵ (we also assumed that conditions 2.1.20 and 2.1.21 of theorem 2.1.3 hold). In other words, the fact that the minimizing problem in theorem 2.1.3 has a lower bound $2\epsilon > 0$, implies that when we take $\delta > 0$ small enough then, under constrain A.0.5 and $|q_1 - q_2| \leq 2\delta$, the expression A.0.6 must be larger than ϵ. This is proved in lemma 4.6.1.

Alignments for which condition A.0.5 is not satisfied are unlikely to happen. Actually, the probability that an alignment does not satisfy A.0.5 is about $e^{-m^*\epsilon_2}$, where m^* designates the number of aligned block pairs. In section 4.6, we define the event $E^n(\epsilon_2)$, which states that there is no optimal alignment not satisfying A.0.5. In the same section, we prove the high probability of this event. The next question is how do we get the condition $|q_1 - q_2| \leq 2\delta_0$ to be satisfied? The answer is the following: in section 4.5, we define the event $J^n(\delta)$ which states that the proportion of left out blocks at the end of X and Y is less than δ, for any optimal alignment. Also, in section 4.1, we defined the event C^n which states that the numbers of blocks in X and in Y are not further from n/l than $n^{0.6}$. In section 3.1, lemma 3.1.1 shows that we need to consider only optimal alignments which do not leave out adjacent blocks between aligned blocks. For such alignments, we get by lemma 3.1.2 that when the events I^n and $J^n(\delta)$ both hold, for n large enough we have then the following: for any optimal alignment not leaving out adjacent blocks between aligned blocks, we have

$$|q_1 - q_2| \leq 2\delta. \tag{A.0.7}$$

Next, we were showing that when all the events $F^n(q_0)$, $D^n(\delta_0)$, $G^n(\delta_0)$, $J^n(\delta_0)$, C^n and $E^n(\epsilon_2)$ hold, then

$$\mathrm{E}[\tilde{L}_n - L_n|X, Y] \geq \epsilon.$$

For this, let a be an optimal alignment of X and Y leaving out no adjacent blocks between aligned block pairs. Let q_1, resp. q_2 designate the proportion of block left out by a in X, resp. in Y. Let $P_{i,j}(a)$ denote the empirical distribution of the lenght of the aligned block pairs by a. Then, as mentioned, when $D^n(\delta_0)$ and $G^n(\delta_0)$ hold, then the expression A.0.2 is a lower bound for the expected increase $\mathrm{E}[\tilde{L}_n - L_n|X,Y]$. For this, take p_{ij} in A.0.2 equal to $P_{ij}(a)$ and δ equal to δ_0. When the event $F^n(q_0)$ holds, then every optimal alignment has less than a proportion q_0 of left out blocks in X as well as in Y. Since a is an optimal alignment, this implies that $q_1, q_2 \leq q_0$. Now, assume that I^n and $J^n(\delta_0)$ both hold. As explained, by lemma 3.1.2, this implies that $|q_1 - q_2| \leq 2\delta_0$. The last inequality together with $q_1, q_2 \leq q_0$ imply that the expression A.0.6 is larger than ϵ if A.0.3 is larger than $-\epsilon_2$ (this is how we had defined δ_0 in the first place). But by the event $E^n(\epsilon_2)$, we have that all optimal alignments satisfy that A.0.3 is more than $-\epsilon_2$. Hence, the expression A.0.6 is larger than ϵ if the event $E^n(\epsilon_2)$ holds. Note that we have argued the following: when $D^n(\delta_0)$ and $G^n(\delta_0)$ hold, then the expected increase $\mathrm{E}[\tilde{L}_n - L_n|X,Y]$ is bounded from below by the expression A.0.6. But, with $E^n(\epsilon_2)$, $J^n(\delta_0)$, $F^n(q_0)$ and I^n all holding, the expression A.0.2 is larger than ϵ. Summarizing, we have just shown that when all the events $F^n(q_0)$, $D^n(\delta_0)$, $G^n(\delta_0)$, $J^n(\delta_0)$, C^n and $E^n(\epsilon_2)$ hold, then

$$\mathrm{E}[\tilde{L}_n - L_n|X,Y] \geq \epsilon.$$

Hence,

$$\mathrm{P}(\ \mathrm{E}[\tilde{L}_n - L_n|X,Y])$$

is not less than

$$1 - \mathrm{P}(F^{nc}(q_0)) - \mathrm{P}(D^{nc}(\delta_0)) - \mathrm{P}(G^{nc}(\delta_0)) - \mathrm{P}(J^{nc}(\delta_0)) - \mathrm{P}(C^{nc}) - P(E^{nc}(\epsilon_2)). \tag{A.0.8}$$

We can now apply the bounds for the probabilites $\mathrm{P}(D^{nc}(\delta_0))$, $\mathrm{P}(G^{nc}(\delta_0))$, $\mathrm{P}(C^{nc})$, $\mathrm{P}(J^{nc}(\delta_0)$ and $\mathrm{P}(E^{nc}(\epsilon_2))$ which we obtained in section 4, to A.0.8. These bounds are:

- In section 4.1, we obtain $\mathrm{P}(C^{nc}) \leq 8e^{-b_1 n^{0.2}}$ where $b_1 > 0$ is a constant only depending on l.
- In subsection 4.3, we show that
$$\mathrm{P}(D^{nc}(\delta)) \leq 2n^{0.6}\left(\frac{1}{1+3\delta}\right)^{n(1+3\delta)/2l}.$$
- In section 4.4, we show that
$$\mathrm{P}(G^{nc}(\delta_0)) \leq 4e^{-b_6 n^{0.2}},$$
where $b_6 > 0$ is a constant only depending on l and δ_0.

- In subsection 4.5, we show that

  $$\mathrm{P}(J^{nc}(\delta_0) \leq e^{-\theta n}$$

  where $\theta > 0$ only depends on l and δ_0.

- In the proof of lemma 4.6.2, we prove that

  $$P(E^{nc}(\epsilon_2)) \leq w(n)e^{-\vartheta n}$$

  where $\vartheta > 0$ is a constant only depending on l and δ_0 and $w(n)$ is a polynomial in n.

With the above bounds, we find that if $0 < \beta < 0.2$ does not depend on n, then

$$\mathrm{P}(D^{nc}(\delta_0)) + \mathrm{P}(G^{nc}(\delta_0)) + \mathrm{P}(J^{nc}(\delta_0)) + \mathrm{P}(C^{nc}) + \mathrm{P}(E^{nc}(\epsilon_2)) \tag{A.0.9}$$

is smaller than e^{-n^β} for n large enough. Or alternatively, taking $0 < \beta < 0.2$ small enough but not depending on n, we have that the expression A.0.9 is smaller than e^{-n^β} for all n. Applying this to the bound A.0.8, yields:

$$\mathrm{P}(\ \mathrm{E}[\tilde{L}_n - L_n|X,Y] \ \geq\ \epsilon) \ \ \geq\ \ 1 - \mathrm{P}(F^{nc}(q_0)) - e^{-n^\beta},$$

where $\beta > 0$ does not depend on n. This finishes the proof of theorem 2.1.3. ■

Bibliography

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

[2] P. Pevzner. *Computational Molecular Biology*, MIT Press, Cambridge, MA, 2000. An algorithmic approach, A Bradford Book.

[3] Dayhoff, M.O., Schwartz, R.M. y Orcutt, B.C. *A model of evolutionary change in proteins.* In M.O. Dayhoff editor. Atlas of Protein Sequence an Structure, volume 5 suppl. 3. Natl. Biomed. Res. Found. pp. 354-352, 1978.

[4] Henikoff, S. y Henikoff, J.G. *Amino acid substitution matrices from protein blocks.* Porc. Natl. Aca. Sci. USA 89: 10925-10919, 1992.

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

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

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

[8] M. Kiwi, M. Loebl, and J. Matousek. *Expected length of the longest common subsequence for large alphabets.* preprint, 2003.

[9] M. J. Steele. *An Efron- Stein inequality for non-symmetric statistics.* Annals of Statistics, 14:75–758, 1986.

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

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

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

[13] A. Dembo and O. Zeitouni. *Large Deviations: Techniques and Applications.* Stochastic Modelling and Applied Probability Series. Springer, 1998. Second edition.

[14] G. Grimmett. and D. Strizaker. *Probability and Random Processes*, Oxford University Press, 2001. Third edition.

[15] J. Lember, H. Matzinger, and C. Durringer. Deviation from mean in sequence comparison with a periodic sequence. Alea, Volume 3:1–29, 2007.

[16] F. Bonetto and H. Matzinger. Fluctuations of the longest common subsequence in the case of 2- and 3-letter alphabets. Latin American Journal of Probability and Mathematics, Volume 2:195–216, 2006.

[17] C.Houdre, J. Lember, H. Matzinger. *On the longest common increasing binary subsequence.* C.R. Acad. Sci. Paris, Ser. I 343:589–594, 2006.

[18] C. Houdre and H. Matzinger. *Fluctuations of the Optimal Alignment Score with and Asymmetric Scoring Function.* Submitted 2006.

[19] J. Lember, H. Matzinger. *Stardard Deviation of the Longest Common Subsequence when zero and one have different probabilities.* Accepted in Annals of Probability, 2008.

[20] S. Amsalu, C. Houdre and H. Matzinger. *Fluctuation of the LCS for close to i.i.d. distribution.* In preparation, 2007.

[21] S. Amsalu, H. Matzinger and M. Vachkovskaia. *Thermodynamical Approach to the Longest Common Subsequence Problem.* Journal of Statistical Physics, Volume 131, Number 6, June 2008.

[22] S.M. Ross, *Introduction Probability Models.* Academic Press, 8 edition, 2002.

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

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

[25] V. Dancik and M. Paterson, *Longest common subsequences.* Lecture Notes in Comput. Sci, Volume 841:127–142. Springer, 1994.

[26] J. Baik, P. Deift and K. Johansson *On the distribution of the length of the longest increasing subsequence of random permutations.* J. Amer. Math. Soc., 12(4):1119-1178, 1999.

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

[28] S.M. Ulam *Monte Carlo calculations in problems of mathematical physics.* Modern mathematics for the engineer. Second series: 261–281. McGraw-Hill, 1961.

[29] B.F. Logan, L.A. Shepp *A variational problem for random Young tableaux.* Advances in Math. 26(2): 206–222, 1977.

[30] J.M. Hammersley *A few seedlings of research.* Proceedings of the Sixth Berkeley Symposium on Statistics and Probability (Univ. California, Calif., 1970/1971), Vol. I: Theory of statistics. 345-394, 1972.