DEVIATION FROM MEAN IN SEQUENCE COMPARISON WITH A PERIODIC SEQUENCE

Heinrich Matzinger, Jüri Lember[1], Clement Durringer

Abstract. Let L_n denote the length of the longest common subsequence of two sequences of length n. We draw one of the sequences i.i.d., but the other is non-random and periodic. We prove that $\mathrm{VAR}[L_n] = \Theta(n)$. This confirms the conjecture of Waterman [9] in the special case when one sequence is periodic.

Keywords. *Longest common subsequence, Waterman conjecture.*

AMS. 60K35, 41A25, 60C05

1 Introduction

Let $\{X_i\}_{i\in\mathbb{N}}$ and $\{Y_i\}_{i\in\mathbb{N}}$ be two ergodic processes independent of each other. We assume that the variables X_i and Y_i have a common state space. Let L_n denote the length of the longest common subsequence of the two finite strings $X := X_1X_2\dots X_n$ and $Y := Y_1Y_2\dots Y_n$. (A common subsequence of X and Y is a subsequence of X which is also a subsequence of Y.) The investigation of the longest common subsequences (LCS) of two finite words is one of the main problems in the theory of pattern matching. The LCS-problem plays a role for DNA- and Protein-alignments, file-comparison, speech-recognition and so forth. The random variable L_n and several of its variants have been studied intensively by probabilists, computer-scientists and mathematical biologists; for applications of LCS-algorithms in biology see Waterman [8].
Using a subadditivity argument, Chvatal and Sankoff [5] prove that the limit

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

exists. The constant γ is called the Chvatal-Sankoff constant and its value is unknown for even as simple cases as i.i.d. binary sequences. Neither is the exact order of the fluctuation of the LCS lenght known. Steele [7] proved that $\mathrm{VAR}[L_n] \leq n$.
The determination of the Chvatal-Sankoff constant and the order of fluctuations for the LCS problem are long standing open problems. Montecarlo simulations lead Chvatal and Sankoff to conjectured that $\mathrm{VAR}[L_n] = o(n^{\frac{2}{3}})$. For a closely related Bernoulli matching model, Majumdar and Nechaev [6] obtained the rate $O(n^{\frac{2}{3}})$.
In [9], Waterman conjectured that in many cases the variance of L_n grows linearly. Boutet de Monvel [4] interprets his simulation in that way too.

[1]Author is supported by Estonian Science Foundation Grant nr. 5694

We believe that there are different possible order of magnitudes depending on the distribution of the strings X and Y.
In [3], Bonetto and Matzinger consider the asymmetric case where X contains two symbols while Y contains three. They prove the variance $\mathrm{VAR}[L_n]$ to be of order $\Theta(n)$.
In reality, the models like a language or a genetic code are often more complicated than an i.i.d. sequence. Therefore, in order to understand what determines the size of the variance of L_n, it becomes essential to investigate different kind of models. Every model might capture one aspect of a complicated real life system. This is why, through a series of papers, we analyze the order of magnitude of the fluctuations of L_n for different cases. In a forthcoming paper, we prove that the order of variance is $\Theta(n)$ for two i.i.d. sequences provided that one and zero have sufficiently different probabilities. Having this result in mind, it is interesting to know if this order of magnitude comes from the fact that both X and Y are random. In other words, we wanted to know what happens if one text is taken non-random. For this we take one text periodic. In the present paper, we show that when one of the sequence is non-random and periodic with a short period, then there exists constants $0 < k < K < \infty$ so that for big n,

$$kn < \mathrm{VAR}[L_n] < Kn.$$

So, in this case $\mathrm{VAR}[L_n] = \Theta(n)$.

Let us mention a little bit more about the history of this field. The most widely used method for the comparison of genetic data is a generalization of the LCS-method. (For an excellent overview of this subject see Waterman-Vingron [10].) In this generalization a maximal score is sought over the set of all possible alignments of the two sequences, where gaps are penalized with a fixed parameter $\delta > 0$ and mismatches are penalized by a fixed amount $\mu > 0$: consider for example the two words "brot" and "bat". One possible alignment $\mathbb{A}$ of these words is

b	r	o	t
b	a	$-$	t

The score of this alignment is $1 - \mu - \delta + 1 = S(\mathbb{A})$. The matching pairs of letters "b" and "t" are each valued with a weight of 1. The gap $-$ in "bat" after the "a" costs $-\delta$. Furthermore, the mismatch between "r" and "a" is penalized by adding $-\mu$ to the total score. If $M_{\mu,\delta}(X, Y)$ denotes the maximal score amongst all possible alignments of the two words X and Y, and if $M_n(\mu, \delta)$ is the random variable defined by $M_n(\mu, \delta) = M_{\mu,\delta}(X, Y)$, where X and Y are two i.i.d. random sequences of length n, then the LCS-problem is a special case of the investigation of $M_n(\mu, \delta)$, because $L_n = M_n(\infty, 0)$. Generalizing the arguments from the LCS-problem, one can prove that the limit

$$a(\mu, \delta) = \lim_{n\to\infty} \frac{\mathbb{E}[M_n]}{n}$$

exists. Arratia-Waterman [2] showed that there is a phase transition phenomenon defined by critical values of μ and δ. In one phase M_n is of linear order in n, whereas in the other

it is logarithmically small in n. Waterman [9] conjectures that the standard deviation of M_n from its mean behaves like $\sqrt{n}$. Waterman-Arratia [2] derive a law of large deviation for L_n for fluctuations on scales larger than $\sqrt{n}$. Using first passage percolation methods, Alexander [1] proves that $\mathbb{E}[L_n]/n$ converges at a rate of order $\sqrt{\log n/n}$.

2 Main result

Let $X_1, X_2, \ldots$ be an i.i.d. sequence of Bernoulli variable with parameter $1/2$. Let $Y_1, Y_2, \ldots$ be a non-random periodic sequence with period p, that is fixed throughout the paper. This means that $p > 1$ is the smallest natural number such that: $Y_{p+n} = Y_n$ for all $n \in \mathbb{N}$. Let L_n be the length of the longest common subsequence of the two finite sequences, $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$. A similar argument as in [5] implies that

$$\frac{L_n}{n} \to \gamma_Y, \quad \text{a.s.},$$

where γ_Y is an unknown constant. Of course, γ_Y depends on the periodic scenery Y. In this paper, we study the asymptotic deviation from the mean of the random variable L_n. Let D_n be defined as follows:

$$D_n := \frac{L_n - E[L_n]}{\sqrt{n}} \tag{2.1}$$

The main result of this paper is Theorem 2.1, which states that $L_n - E[L_n]$ is typically of order $\sqrt{n}$. To prove theorem 2.1, we show in Lemma 2.2 that the standard deviation of L_n is of order $\sqrt{n}$.
We need the following large deviation result, (which is similar to a result of Arratia and Waterman [2]):

Lemma 2.1 *There exists a constant $b > 0$ not depending on n and $\Delta > 0$ such that for all n large enough, we have:*

$$P\left(|L_n - EL_n| \geq n\Delta\right) \leq e^{-bn\Delta^2} \tag{2.2}$$

Proof. The inequality (2.2) is a straightforward application of the McDiarmid inequality: Let $X_1, \ldots, X_n$ independent A-valued random variables. Let $f : A^n \mapsto \mathbb{R}$ be a function that satisfies

$$\sup_{x_1,\ldots,x_n,x_i' \in A} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i, \quad i = 1, \ldots, n.$$

Then for any $\Delta > 0$

$$P\Big(|f(X_1, \ldots, X_n) - Ef(X_1, \ldots, X_n)| \geq \Delta\Big) \leq 2\exp\big[-\frac{2\Delta^2}{\sum_{i=1}^n c_i^2}\big]. \tag{2.3}$$

Take $f : \{0,1\}^n \to \mathbb{R}$ to be the length of the longest common subsequence between i.i.d. random variables $X_1, \ldots, X_n$ and non-random $Y_1, \ldots, Y_n$. So $L_n = f(X_1, \ldots, X_n)$. Clearly

the following holds: by changing an element in a binary sequence $(x_1, \ldots, x_n) \in \{0,1\}^n$, the length of a longest common subsequence of $x_1, \ldots, x_n$ and $Y_1, \ldots, Y_n$ changes at most by one. Thus, the assumptions of McDiarmid inequality are satisfied with $c_i = 1$, $i = 1, \ldots, n$. Hence, the inequality (2.3) holds, and (2.2) trivially follows. ■

Our main result about the variance is the following.

Lemma 2.2 *There exist $0 < k < K < \infty$ not depending on n, such that for all n large enough:*

$$Kn \geq \mathrm{VAR}\,[L_n] \geq kn.$$

The proof of Lemma 2.2 is presented at the end of Section 3.

Our main theorem studies the sequence $\{D_n\}$ as defined in (2.1).

Theorem 2.1 *The sequence $\{D_n\}$ is tight. Moreover, the limit of any weakly convergent subsequence of $\{D_n\}$ is not a Dirac measure.*

Proof. For $s > 0$, the inequality (2.2) with $\Delta = \frac{s}{\sqrt{n}}$ implies

$$P(|D_n| \geq s) = P(|D_n| \geq \sqrt{n}\frac{s}{\sqrt{n}}) \leq \exp[-cn\frac{s^2}{n}] = \exp[-cs^2].$$

The last inequality implies that for any $r \geq 1$, the sequence $\{D_n\}$ is uniformly bounded in L_r, i.e.

$$\sup_n E|D_n|^r = \sup_n \int_0^\infty P(|D_n|^r \geq s)ds \leq \int_0^\infty \exp[-cs^{\frac{2}{r}}]ds < \infty. \tag{2.4}$$

Hence, the sequence $\{D_n\}$ is uniformly integrable and, therefore, tight.
Let $D_{n_i} \Rightarrow Q$ be a weakly converging subsequence of $\{D_n\}$. Suppose $Q = \delta_c$, for a $c \in (-\infty, \infty)$. By the continuous mapping theorem, $D_{n_i}^2 \Rightarrow \delta_{c^2}$ or, equivalently, the sequence $D_{n_i}^2$ converges to the constant c^2 in probability. Since $\sup_n E|D_n|^3 < \infty$, the sequence $\{D_n^2\}$ is uniformly integrable, as well. Hence, the weak convergence implies that: $ED_{n_i}^2 = \mathrm{VAR}D_{n_i} \to 0$, which contradicts Lemma 2.2. ■

3 Proof of Lemma 2.2

This section is dedicated to the proof of Lemma 2.2.

3.1 Main idea and numerical example

Lemma 2.2 states that the variance of L_n is of order n. To prove this, we show that L_n can be written as the sum of two independent parts: $Z_{\vec{T}}$ and $L_n^{\vec{T}}$ (see 3.7). The variance of $Z_{\vec{T}}$ is of order n, and so is the variance of L_n.

Let us present a simple numerical example: Let the periodic sequence Y have period 2, such that:

$$Y_1Y_2Y_3Y_4Y_5Y_6\ldots = 010101\ldots$$

Let $l \in 16\mathbb{N}$. (Here the number 16 corresponds to $4p^2$). Assume that in the neighborhood of l, the sequence X is equal to the periodic sequence Y (except possibly in l). More precisely, assume that we observe:

$$Y_{l-16}Y_{l-15}Y_{l-14}\ldots Y_{l+13}Y_{l+14}Y_{l+15} = 0101010101010101a1010101010101010,$$

where a can be equal to either zero or one. A point l satisfying the last equality above is called a *replica point*. If a coincides with the periodic pattern, we say that the replica point l *matches*. In our example, this would happen if $a = 0$. We call $[l - 4p^2, l + 4p^2 - 1]$ the *interval of the replica point* l. The main combinatorial idea in this article is contained in Lemma 3.1. It states that for a replica point l, the score L_n is increased by one when l matches. Furthermore, this not influenced by the sequence X outside the interval of the replica point l. This fact is intuitively clear and it is simple to find a heuristic proof. However, the formal proof of Lemma 3.1 is difficult. The whole Section 4 is dedicated to it.
The variable $Z_{\vec{T}}$ is defined to be the number of replica points that mach (among the first cn replica points, where $c > 0$ is a constant not depending on n). From Lemma 3.1, it follows directly that L_n can be written as a sum of $Z_{\vec{T}}$ and a term which depends only on the sequence X "outside the replica points intervals". This leads directly to the independence $Z_{\vec{T}}$ and $L_n^{\vec{T}}$.

3.2 Replica points

We can assume without restriction that $Y_0 = 1$. For $l \in \mathbb{N}$ we define the integer interval:

$$J_l := [l - 4p^2, l + 4p^2 - 1].$$

Let I_l designate J_l minus its center:

$$I_l := J_l - \{l\}.$$

Definition 3.1 *Let $l \in \mathbb{N}$, with $l > 4p^2$. We say that l is a* **replica point** *if the following condition holds:*

$$Y_z = X_z, \forall z \in I_l.$$

If l is a replica point and $X_l = Y_l$, then we say that the **replica point l matches**.

We need some more notation. We denote by A_l the event that l is a replica point and denote by Z_l the Bernoulli variable which is equal to one if and only if l is a replica point which matches. Thus, $Z_l = 1$ if A_l and $X_l = Y_l$ both hold, otherwise $Z_l = 0$.
We denote by $X_{|l}$ the finite sequence obtained from $X_1, \ldots, X_n$ by removing X_l, i.e.

$$X_{|l} := (X_1, X_2, \ldots, X_{l-2}, X_{l-1}, X_{l+1}, X_{l+2}, \ldots, X_n).$$

We denote by Σ_l the σ-algebra generated by $X_{|l}$, i.e.

$$\Sigma_l := \sigma(X_i | 1 \leq i \leq n, i \neq l).$$

Let L_n^l designate the length of the longest common subsequence of $X_{|l}$ and $Y_1, \ldots, Y_n$.

The next Lemma is the fundamental combinatorial idea for replica points. It says that when l is a replica point, then the length of the longest common subsequence can be decomposed as $L_n = Z_l + L_n^l$, where Z_l comes from the replica point and L_n^l depends on $X_{|l}$, only. Such a decomposition is useful, because $A_l \in \Sigma_l$, i.e. whether l is a replica point or not does not depend on X_l. The proof of Lemma 3.1 is given in Section 4.

Lemma 3.1 *Let $l \in \mathbb{N}$ so that $4p^2 < l \leq n - 4p^2 - 1$. If A_l holds, then*

$$L_n = Z_l + L_n^l. \tag{3.1}$$

3.3 Several replica points

In the following, $c > 0$ is a constant not depending on n such that $cn \in \mathbb{N}$. (We choose $c > 0$ to be small enough, so that with high probability there are at least cn replica points in $[0, n]$. By Lemma 3.4, it is enough to take c such that: $0 < c < (0.5)^{8p^2-1}$.) Let $K^n \subset \mathbb{N}^{cn}$ designate the set of all integer vectors

$$\vec{k} = (k_1, k_2, \ldots, k_{cn})$$

such that $k_i + 8p^2 \leq k_{i+1}, \forall i = 1, \ldots, cn - 1$ and $4p^2 < k_1$ and $k_{cn} < n - 4p^2$.

Let $\vec{k} = (k_1, k_2, \ldots, k_{cn}) \in K^n$. We define the σ-algebra:

$$\Sigma_{\vec{k}} := \sigma\left(X_i |\ i \in [0, n] \text{ and } i \neq k_j, \forall j \in [1, cn]\right).$$

We denote by $A_{\vec{k}}$ the event that k_i is a replica point for all $i = 1, \ldots, cn$. Clearly $A_{\vec{k}} \in \Sigma_{\vec{k}}$.

Suppose $A_{\vec{k}}$ holds. Let $Z_{\vec{k}}$ designate the number of replica points among $k_1, k_2, \ldots, k_{cn}$ which are matches. So, if $A_{\vec{k}}$ holds, and $\vec{k} = (k_1, \ldots, k_{cn}) \in K^n$, then

$$Z_{\vec{k}} := \sum_{i=1}^{cn} Z_{k_i}.$$

Let $X_{|\vec{k}}$ designate the finite sequence one obtains by removing from X the bits $X_{k_i}, i = 1, \ldots, cn$. Hence, for $\vec{k} = (k_1, \ldots, k_{cn}) \in K^n$,

$$X_{|\vec{k}} := \left\{X_i |\ i \in [0, n] \text{ and } i \neq k_j, \forall j \in [1, cn]\right\}.$$

Finally, let $L_n^{\vec{k}}$ designate the length of the longest common subsequence of $X_{|\vec{k}}$ and Y.

Lemma 3.2 *Let $\vec{k} \in K^n$. When $A_{\vec{k}}$ holds, then*

$$L_n = Z_{\vec{k}} + L_n^{\vec{k}}. \tag{3.2}$$

Proof. The proof follows from Lemma 3.1 by induction.
Let $cn = 2$, i.e. $\vec{k} = (l_1, l_2)$. Let $Z_i = Z_{l_i}$, $i = 1, 2$. Let us show that

$$L_n = L_n^{\vec{k}} + Z_1 + Z_2. \tag{3.3}$$

Let L_n^{1+} be length of the longest common subsequence of $X|_{l_1}$ and $Y_1, \ldots, Y_n$ provided that $Z_2 = 1$. Let L_n^{1-} be length of the longest common subsequence of $X|_{l_1}$ and $Y_1, \ldots, Y_n$ provided that $Z_2 = 0$. Finally, let $L_n^1 := L_n^{l_1}$, so L_n^1 is either L_n^{1+} or L_n^{1-}.
At first note,

$$L_n^{1-} + 1 = L_n^{1+}. \tag{3.4}$$

Let L_n^+ and L_n^- denote the length of the longest common subsequence of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ provided that $Z_2 = 1$ and $Z_2 = 0$, respectively. ¿From Lemma 3.1 follows that $L_n^+ = L_n^- + 1$ as well as $L_n^{1+} + Z_1 = L_n^+$ and $L_n^{1-} + Z_1 = L_n^-$. Hence, (3.4) holds.
Clearly, $L_n^1 \geq L_n^{\vec{k}} \geq L_n^1 - 1$. Hence, $L_n^{\vec{k}}$ is equal to L_n^{1+} or $L_n^{1+} - 1 = L_n^{1-}$. If $L_n^{\vec{k}} = L_n^{1+}$, we would have that $L_n^{\vec{k}} > L_n^{1-}$, a contradiction. Hence $L_n^{\vec{k}} = L_n^{1+} - 1 = L_n^{1-}$. Suppose $Z_2 = 1$. Then $L_n = L_n^+ = L_n^{1+} + Z_1$, so

$$L_n^{\vec{k}} + Z_1 + Z_2 = L_n^{\vec{k}} + Z_1 + 1 = L_n^{1+} + Z_1 = L_n^+ = L_n.$$

Suppose $Z_2 = 0$. Then $L_n = L_n^- = L_n^{1-} + Z_1$, so

$$L_n^{\vec{k}} + Z_1 + Z_2 = L_n^{\vec{k}} + Z_1 = L_n^{1-} + Z_1 = L_n^- = L_n.$$

Let $cn = m + 1$, i.e. $\vec{k} = (l_1, l_2, \ldots, l_{m+1})$. Let $\vec{m} := (l_1, l_2, \ldots, l_m)$, $Z_m = \sum_{i=1}^m Z_{l_i}$, $Z_{m+1} := Z_{l_{m+1}}$. Suppose (3.2) holds for $cn = m$, i.e.

$$L_n = L_n^{\vec{m}} + Z_m. \tag{3.5}$$

Let us show that

$$L_n = L_n^{\vec{k}} + Z_m + Z_{m+1}.$$

The argument is similar to the case $m = 2$. Let L_n^{m+} be equal to $L_n^{\vec{m}}$ provided that $Z_{m+1} = 1$. Let L_n^{m-} be equal to $L_n^{\vec{m}}$ provided that $Z_{m+1} = 0$. At the first, we prove that

$$L_n^{m-} + 1 = L_n^{m+}. \tag{3.6}$$

Let L_n^+ and L_n^- denote the length of the longest common subsequence of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ provided that $Z_{m+1} = 1$ and $Z_{m+1} = 0$, respectively. From Lemma 3.1 follows that $L_n^+ = L_n^- + 1$. ¿From (3.5) follows $L_n^{m+} + Z_m = L_n^+$ and $L_n^{m-} + Z_m = L_n^- = L_n^+ - 1$. Hence, (3.6) holds.
Clearly, $L_n^{\vec{m}} \geq L_n^{\vec{k}} \geq L_n^{\vec{m}} - 1$. Hence, $L_n^{\vec{k}}$ is equal to L_n^{m+} or $L_n^{m+} - 1 = L_n^{m-}$. If $L_n^{\vec{k}} = L_n^{m+}$,

we would have that $L_n^{\vec{k}} > L_n^{m-}$, a contradiction. Hence $L_n^{\vec{k}} = L_n^{m+} - 1 = L_n^{m-}$.
Suppose $Z_{m+1} = 1$. Then by (3.5), $L_n = L_n^+ = L_n^{m+} + Z_m$, so

$$L_n^{\vec{k}} + Z_m + Z_{m+1} = L_n^{\vec{k}} + Z_m + 1 = L_n^{m+} + Z_m = L_n^+ = L_n.$$

Suppose $Z_{m+1} = 0$. Then by (3.5), $L_n = L_n^- = L_n^{m-} + Z_m$, so

$$L_n^{\vec{k}} + Z_m + Z_{m+1} = L_n^{\vec{k}} + Z_m = L_n^{m-} + Z_m = L_n^- = L_n.$$

■

3.4 Intervals

Let U_i, $i = 1, 2, \ldots$ be the disjoint consecutive intervals with length $8p^2$, i.e. (recall the definition of J_l)

$$U_i := J_{i4p^2+1} = [(i-1)8p^2 + 1, i8p^2], \quad i = 1, 2, \ldots.$$

Let $u_i := i4p^2 + 1$. Whether u_i is a replica point or not, depends on $\{X_z : z \in U_i, z \neq u_i\}$.

Let T_i designate the i-th replica point. Formally, we define T_i by induction on i. For $i = 1$, we put:

$$T_1 := \min\{u_j | u_j \text{ is a replica point}, j > 0\}.$$

Once, T_i is defined, we define T_{i+1} in the following way:

$$T_{i+1} := \min\{u_j > T_i | u_j \text{ is a replica point }, j > 0\}.$$

Let $c > 0$ be a constant not depending on n. We define the event

$$E_n := \{T_{cn} \leq n\}$$

which guarantees that there are at least cn replica points in $[0, n]$.

Let

$$\vec{T} := \begin{cases} (T_1, T_2, \ldots, T_{cn}), & \text{if } E_n \text{ holds}, \\ 0, & \text{otherwise} \end{cases},$$

$$X_{|\vec{T}} := \begin{cases} X_{|\vec{k}}, & \text{if } \vec{T} = \vec{k}, \\ X, & \text{if } \vec{T} = 0. \end{cases} \qquad Z_{\vec{T}} := \begin{cases} Z_{\vec{k}}, & \text{if } \vec{T} = \vec{k}. \\ 0, & \text{if } \vec{T} = 0. \end{cases}$$

In other words, when E_n holds, $X_{|\vec{T}}$ is the sequence obtained by removing the bits $X_{T_1}, X_{T_2}, \ldots, X_{T_{cn}}$ from the sequence X and $Z_{\vec{T}}$ is the number of matching replica points

in $\vec{T}$.
With $L_n^0 := L_n$, we obviously have

$$L_n = Z_{\vec{T}} + L_n^{\vec{T}}. \tag{3.7}$$

Finally, let

$$\Sigma := \sigma(\vec{T}, X_{|\vec{T}}).$$

Clearly, $L_n^{\vec{T}}$ is Σ-measurable and $E_n \in \Sigma$.

Lemma 3.3 *Conditional on Σ and E_n, $Z_{\vec{T}}$ has binomial distribution with parameters $1/2$ and cn:*

$$\mathcal{L}\big(Z_{\vec{T}}\big|\vec{T} = \vec{k}, X_{|\vec{k}}\big) = B(1/2, cn),$$

for all $\vec{k} \in K^n$.

Proof. By interval construction, it holds that $\{\vec{T} = \vec{k}\} \in \sigma(X_{|\vec{k}})$. The vector $\vec{Z} := (Z_{k_1}, \dots Z_{k_{cn}})$ is $\sigma(X_{k_1}, \dots, X_{k_{cn}})$-measurable. Those σ-algebras are independent, hence $\vec{Z}$ is independent of $\sigma(X_{|\vec{k}})$. By interval-construction, $\vec{Z}$ consists of independent components. Since X_i is a Bernoulli $1/2$-random variable, the statement holds. ■

The next Lemma shows that we can choose $c > 0$ so that for big n, there are typically at least cn replica points in $[0, n]$.

Lemma 3.4 *If $c < (0.5)^{8p^2-1}$, then $\lim_{n\to+\infty} P(E_n) = 1$.*

Proof. Let ξ_i be a Bernoulli random variable that is 1 if and only if u_i is a replica point. Clearly, $P(\xi_i = 1) = (0.5)^{8p^2-1} =: q$ and

$$E_n = \Big\{\sum_{i=1}^{n} \xi_i \geq cn\Big\}.$$

Then, by Hoeffding inequality,

$$P(E_n^c) = P\Big(\sum_{i=1}^{n} \xi_i < cn\Big) = P\Big(\sum_{i=1}^{n} \xi_i - qn < (c-q)n\Big) \leq \exp[-2(c-q)^2 n] \to 0.$$

■

3.5 Proof of Lemma 2.2

From (2.4) it follows: $\exists K < \infty$ such that

$$\sup_n ED_n^2 = \sup_n \frac{\mathrm{VAR}[L_n]}{n} < K.$$

We now prove the existence of $k > 0$.
Clearly

$$\mathrm{VAR}[L_n] = E\,(\,\mathrm{VAR}[L_n|\Sigma]\,) + \mathrm{VAR}\,(\,E[L_n|\Sigma]\,) \geq E\,(\,\mathrm{VAR}[L_n|\Sigma]\,)\,.$$

By (3.7), $L_n = Z_{\vec{T}} + L_n^{\vec{T}}$. Since $L_n^{\vec{T}}$ is Σ-measurable, it holds that:

$$\mathrm{VAR}[L_n|\Sigma] = \mathrm{VAR}[Z_{\vec{T}}|\Sigma]. \tag{3.8}$$

By Lemma 3.3, on $E_n = \{T \neq 0\}$, the conditional distribution of $Z_{\vec{T}}$ is binomial. On E_n^c, $Z_{\vec{T}} = 0$ and hence $E(I_{E_n^c}\mathrm{VAR}[Z_{\vec{T}}|\Sigma]) = 0$. Therefore:

$$E(\mathrm{VAR}[L_n|\Sigma]) = E(\mathrm{VAR}[Z_{\vec{T}}|\Sigma]) = E(I_{E_n}\mathrm{VAR}[Z_{\vec{T}}|\Sigma]) + E(I_{E_n^c}\mathrm{VAR}[Z_{\vec{T}}|\Sigma]) = 0.25cn \cdot P(E_n).$$

By Lemma 3.4, for all n large enough we have:

$$0.25cn \cdot P(E_n) \geq kn,$$

for any $k > 0$ not depending on n, such that $k < 0.25c$.

4 Combinatorics

The rest of this paper is devoted to the proof of Lemma 3.1.

4.1 Preliminaries

4.1.1 Blocks

We need to introduce some necessary formalism. In the present Section, we consider the non-random sequences, only. At first, we formalize the common subsequence.

Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ be two fixed finite sequences. A *common subsequence* of $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ is a strictly increasing mapping

$$v : \{1, \ldots, n\} \hookrightarrow \{1, \ldots, m\}. \tag{4.1}$$

Notation (4.1) means: There exists $I \subset \{1, \ldots, n\}$ and a mapping

$$v : I \to \{1, \ldots, m\}$$

such that

$$y_{v(i)} = x_i, \quad \forall i \in I$$

and v is strictly increasing: $v(i_2) > v(i_1)$, if $i_2 > i_1$.

Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ be two sequences and let v be a common subsequence.

Since v is defined as a mapping (4.1), in what follows, we would like to distinguish the sequence on which v is defined from the image sequence of v. Therefore, we say: v is a common subsequence between $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$, implying that v is defined as (4.1), i.e. from the sequence $x_1, \ldots, x_n$ into $y_1, \ldots, y_m$.

The set I in (4.1) shall be denoted by

$$\mathrm{Dom}(v).$$

The length of v, denoted as $|v|$, is $|\mathrm{Dom}(v)|$.

With $J \subset \{1, \ldots, n\}$, we denote by $v|_J$ the restriction of v to J. The restriction as a subsequence of the common sequence v is defined even when J is not a subset of $\mathrm{Dom}(v)$.

For $a \in \{1, \ldots, n\}$, we define

$$\underline{v}(a) = v(\max\{i \in \mathrm{Dom}(v) : i < a\}) + 1, \quad \bar{v}(a) = v(\min\{i \in \mathrm{Dom}(v) : i > a\}) - 1.$$

Our analysis is based on the *optimality principle*: If v is a longest common subsequence, then for any $[a, b] \subset \{1, \ldots, n\}$, the subsequences:

$$\begin{aligned} v|_{[1,a-1]} &: \{1, \ldots, a-1\} \hookrightarrow \{1, \ldots, \bar{v}(a-1)\} \\ v|_{[a,b]} &: \{a, \ldots, b\} \hookrightarrow \{\underline{v}(a), \ldots, \bar{v}(b)\} \\ v|_{[b+1,n]} &: \{b+1, \ldots, n\} \hookrightarrow \{\underline{v}(b+1), \ldots, m\} \end{aligned}$$

are all with the longest possible length.

Note: $[\underline{v}(a), \bar{v}(b)]$ can also be empty. Moreover, the intervals $[1, \bar{v}(a-1)]$ and $[\underline{v}(a), \bar{v}(b)]$ as well as $[\underline{v}(a), \bar{v}(b)]$ and $[\underline{v}(b+1), m]$ can be overlapping, but the overlapping region does not contain any elements of common subsequence v.

Let v be a common subsequence, i.e. a mapping satisfying (4.1). Let $\{A_1, \ldots, A_l\}$ be a partition of $\mathrm{Dom}(v)$ that satisfies:

i) A_i is an integer interval for every i, i.e. $A_i = \{j, j+1, \ldots, j+s\}$ for some $s \geq 0$.

ii) v is linear on A_i, i.e.

$$v(j+1) = v(j) + 1, \quad \text{for every } j \in A_i \text{ such that } j+1 \in A_i.$$

Clearly there exists at least one partition that satisfies **i)** and **ii)**: the partition, where $A_i = \{i\}$ for every $i \in \mathrm{Dom}(v)$. This is the maximal partition. Let $B^*(v) = B^* = \{B_1, \cdots, B_r\}$ be the minimal partition that satisfies **i)** and **ii)**, i.e. every other partition is a subpartition of B^*. Clearly B^* exists and is unique. We call the elements of B^* the *blocks* of v. By **i)**, every block $B \in B^*$ is an interval, the *length* of a block B is the number of the elements in B.

Proposition 4.1 *Let* $\{B_1, \ldots, B_r\}$ *be the blocks of*

$$v : \{1, \ldots, n\} \hookrightarrow \{1, \ldots, m\}.$$

Then

$$\max\{n, m\} \geq \lfloor \frac{r-1}{2} \rfloor + \sum_{i}^{r} |B_i| = \lfloor \frac{r-1}{2} \rfloor + |\mathrm{Dom}(v)|. \tag{4.2}$$

Proof. Let $n_j := \max B_j$, $j = 1, 2, \ldots, r$. From the definition of blocks, it follows: $n_2 \geq |B_1| + |B_2| + 1$ or $v(n_2) \geq |B_1| + |B_2| + 1$, i.e by changing the block, v "loses" an element either in the set on which v is defined or in the image set of v. Similarly, $n_4 \geq |B_1| + |B_2| + 2$ or $v(n_4) \geq |B_1| + |B_2| + 2$. Hence, for an even r,

$$\max\{n_r, v(n_r)\} \geq \sum_{i}^{r} |B_i| + \frac{r}{2}.$$

Since $\max\{n, m\} \geq \max\{n_r, v(n_r)\}$, (4.2) follows. ■

4.1.2 The blocks between two subsequences of a periodic sequence

In the following, we investigate common subsequences between finite periodic sequences. We start with a simple but yet useful observation, proved in the Appendix.

Proposition 4.2 *Let* $x_1, x_2, \ldots$ *be a periodic sequence with period* p*. If* $k \leq p$ *is a non-negative integer such that*

$$x_j = x_{k+j}, \quad \forall j = 1, \ldots, p, \tag{4.3}$$

then $k = p$*.*

Assume now that $x_1, \ldots, x_n$ and $x_{m+1}, \ldots, x_{m+n}$ are two subsequences of a periodic sequence $\{x_n\}$ with period p. Let v be a common subsequence of $x_1, \ldots, x_n$ and $y_1, \ldots, y_n = x_{m+1}, \ldots, x_{m+n}$, i.e.

$$v : \{1, \ldots, n\} \hookrightarrow \{1, \ldots, n\}.$$

Let B be a block of v. The difference $v(i) + m - i$, where $i \in B$ is called the *bias* of B.

What is the meaning of the bias? Suppose v is a common subsequence, $B = \{j, \ldots, j+s\}$ is a block of v with the bias 2. This means that the common subsequence v includes the elements $x_j, \ldots, x_{j+s}$ of $x_1, \ldots, x_n$. We also know, how these elements are matched with the elements of $y_1, \ldots, y_n$: $x_j = y_{j+2-m}$, $x_{j+1} = y_{j+3-m}, \ldots, x_{j+s} = y_{j+s+2-m}$. Since $y_j = x_{j+m}$, we get $x_j = x_{j+2}$, $x_{j+1} = x_{j+3}, \ldots, x_{j+s} = x_{j+s+2}$. Moreover, for x_{j-1} (x_{j+m+1}), it holds: x_{j-1} (x_{j+m+1}) either does not belong to the common subsequence or it is matched with an element not equal to x_{j+1} (x_{j+m+3}).
Hence, the bias 0 means that every element of B is matched with itself – the *identity matching*. By periodicity, the bias np means essentially the same. We say that B is *unbiased*, if the bias of B is np for a $n \in \mathbb{N}$. Otherwise B is *biased*. Proposition 4.2 can be restated:

Proposition 4.3 *Let B be a biased block. Then the length of B is at most $p-1$.*

Example 4.1 *Let us give a numerical example. Let*

$$\begin{aligned}(x_1,\dots,x_{20}) &= (00111001110011100111),\\ (y_1,\dots,y_{20}) := (x_2,\dots,x_{21}) &= (01110011100111001110).\end{aligned}$$

So, we consider the subsequences of a periodic sequence with the period $p=5$. Let

$$v : \{1,\dots,20\} \hookrightarrow \{1,\dots,20\},$$

with

$$\begin{aligned}&v(1)=1, v(3)=3, v(4)=4, v(5)=7, v(6)=10, v(7)=11, v(8)=12\\ &v(14)=13, v(15)=14, v(16)=15, v(17)=16, v(18)=17, v(19)=18\end{aligned}$$

be a common subsequence. Obviously,

$$\mathrm{Dom}(v) = \{1,3,4,5,6,7,8,14,15,16,17,18,19\}$$

and v has 5 blocks:

$$B_1=\{1\},\ B_2=\{3,4\},\ B_3=\{5\},\ B_4=\{6,7,8\},\ B_5=\{14,15,16,17,18,19\}.$$

Since $m=1$, the corresponding biases are

$$b(B_1)=1-1+1=1,\ b(B_2)=1,\ b(B_3)=7-5+1=3,\ b(B_4)=5,\ b(B_5)=0.$$

Hence, the blocks B_4 and B_5 are unbiased. The lengths of the blocks are, respectively, 1,2,1,3,6. The length of v, is $|v| = |B_1|+|B_2|+|B_3|+|B_4|+|B_5| = 1+2+1+3+6 = 13$.

Sometimes we regard v as a subsequence between

$$\begin{aligned}(x_1,\dots,x_{20}) &= (00111001110011100111),\\ (x_2,\dots,x_{21}) &= \ \ (01110011100111001110),\end{aligned}$$

i.e. v is a mapping

$$v : \{1,\dots,20\} \hookrightarrow \{2,\dots,21\}.$$

with

$$\begin{aligned}&v(1)=2, v(3)=4, v(4)=5, v(5)=8, v(6)=11, v(7)=12, v(8)=13\\ &v(14)=14, v(15)=15, v(16)=16, v(17)=17, v(18)=18, v(19)=19.\end{aligned}$$

With this notation, the blocks and their biases remain unchanged, the bias of a block $B=\{i,\dots,j\}$ is just defined as $v(i)-i$.

4.2 The structure of a common subsequence between periodic subsequences

4.2.1 The structure of a common subsequence between periodic subsequences with length $8p^2$

In the present Subsection, we consider the subsequences of a periodic sequence with length $8p^2$, i.e. we consider the sequences $x_1, \ldots, x_{8p^2}$ and $x_{m+1}, \ldots, x_{m+8p^2}$. We are interested in the length and the structure of (any) longest common subsequence of these two subsequences. Of course, when m is a multiple of p, then the longest common subsequence is just the identity matching. Hence, we assume that m is not a multiple of p. Without loss of generality, we assume that $0 < m < p$. Moreover, it is easy to see that without loss of generality we can (and we do) assume that

$$0 < m \leq \frac{p}{2}.$$

Obviously, there exists a common subsequence v with length $8p^2 - m$: the identity matching. Such a v has only one block with bias 0.

Proposition 4.4 *Let $x_1, \ldots, x_{8p^2}$ and $x_{m+1}, \ldots, x_{m+8p^2}$ be the subsequences of a periodic sequence, $0 \leq m \leq \frac{p}{2}$. Then the length of the longest common subsequence is $8p^2 - m$.*

Proof. Let v be a longest common subsequence, let $\{B_1, \ldots, B_r\}$ be the blocks of v. Note: if v has an unbiased block, then the length of v is at most $8p^2 - m$. Indeed: suppose that the bias of $B_j = \{i_j, i_j + 1, \ldots, i_j + s\}$ $s \geq 0$ is 0. Let $n_{j-1} = \max B_{j-1}$. Since $v(n_{j-1}) \leq v(i_j) - 1 = i_j - 1 - m$, we have that the length of $v|_{B_1 \cup \cdots \cup B_{j-1}}$ is at most $v(n_{j-1}) = i_j - m - 1$. Similarly, the length of $v|_{B_{j+1} \cup \cdots \cup B_r}$ is at most $8p^2 - (i_j + s)$. So the length of v is at most $(i_j - m - 1) + (s + 1) + (8p^2 - (i_j + s)) = 8p^2 - m$.
If the bias of B_j is kp for a $k \in \mathbb{N}, k \neq 0$ the same argument holds.
Hence, if the length of v is bigger than $8p^2 - m$, then all blocks $\{B_1, \ldots, B_r\}$ must be biased. By Proposition 4.3, the length of a biased block is at most $p - 1$. Thus, the number of blocks is bounded below $r \geq \frac{8p^2 - m + 1}{p}$ and

$$\lfloor \frac{r-1}{2} \rfloor \geq \lfloor \frac{8p^2 - m + 1 - p}{2p} \rfloor \geq \lfloor 4p - \frac{m-1}{2p} - \frac{1}{2} \rfloor \geq 4p - 1 > m + 1. \tag{4.4}$$

From Proposition 4.1, it follows $|\mathrm{Dom}(v)| < 8p^2 - m - 1$ that contradicts the assumption that the length of v is at least $8p^2 - m + 1$. ■

Corollary 4.1 *Let v be a longest common subsequence, and let $\{B_1, \ldots, B_r\}$ be its blocks. Then there exists one and only one block B_j that is unbiased. Moreover, the bias of B_j is 0 or p, and it can be p only, when $m = \frac{p}{2}$.*

Proof. From (4.4) follows that v has at least one unbiased block. Since v is the longest, Proposition 4.1 implies that v has only one unbiased block, say B_j. If $m < \frac{p}{2}$, the argument used in the beginning of the proof of Proposition 4.4 yields that the bias of B_j is 0. If $m = \frac{p}{2}$, then the bias of B_j can be p as well. ■

Corollary 4.2 *Let v be a longest common subsequence, let $\{B_1, \ldots, B_r\}$ be its blocks. Let $B_j = \{i_j, \ldots, i_j + s\}$ be its unbiased block. Let $b \in \{0, p\}$ be the bias of B_j. Then the length of $v|_{B_1 \cup \cdots \cup B_{j-1}}$ is $i_j - m - 1 + \frac{b}{2}$ and the length of $v|_{B_{j+1} \cup \cdots \cup B_r}$ is $8p^2 - (i_j + s) - \frac{b}{2}$.*

Proposition 4.5 *Let v be a longest common subsequence, let $B_j = \{i_j, \ldots, i_j + s\}$ be the unbiased block of v. Let $b \in \{0, p\}$ be the bias of B_j. Then the integer interval $[mp + 1 - \frac{b}{2}, 8p^2 - m(p-1) - \frac{b}{2}] \subset B_j$. In particular, $[mp + 1, 8p^2 - mp] \subset B_j$.*

Proof. Let us first consider the case $b = 0$. By Corollary 4.2, the length of $v|_{B_1 \cup \cdots \cup B_{j-1}}$ is $i_j - m - 1$. Since

$$v|_{B_1 \cup \cdots \cup B_{j-1}} : \{1, \ldots, i_j - 1\} \hookrightarrow \{1, \ldots, i_j - m - 1\},$$

it holds that:

$$v|_{B_1 \cup \cdots \cup B_{j-1}}(\{1, \ldots, i_j - 1\}) = \{1, \ldots, i_j - m - 1\}.$$

This means that

$$v(n_{j-1}) = i_j - m - 1 = |B_1| + \cdots + |B_{j-1}|, \tag{4.5}$$

where $n_{j-1} = \max B_{j-1}$. Hence, by changing the blocks, v loses only the elements on the set where it is defined. Up to the block B_j there are $j - 1$ changes. Hence, v loses at least $j - 1$ elements, so that:

$$i_j > |B_1| + \cdots + |B_{j-1}| + j - 1.$$

On the other hand, by (4.5):

$$i_j = |B_1| + \cdots + |B_{j-1}| + (m + 1),$$

and thus $j - 1 < m + 1$ or $j - 1 \leq m$. Since the blocks $B_1, \ldots, B_{j-1}$ are biased, their length is at most $p - 1$. Therefore, $i_j \leq m(p-1) + (m+1) = mp + 1$.
By Corollary 4.2, the length of $v|_{B_{j+1} \cup \cdots \cup B_r}$ is at most $8p^2 - (i_j + s)$. Since

$$v|_{B_{j+1} \cup \cdots \cup B_r} : \{i_j + s + 1, \ldots, 8p^2\} \hookrightarrow \{i_j + s - m + 1, \ldots, 8p^2\},$$

it holds:

$$\mathrm{Dom}(v|_{B_{j+1} \cup \cdots \cup B_r}) = \{i_j + s + 1, \ldots, 8p^2\}.$$

The last equality implies that:

$$8p^2 - (i_j + s) = |B_{j+1}| + \cdots + |B_r|. \tag{4.6}$$

Hence, after B_j, by changing the blocks, v loses the elements on the image set, only. From B_j to B_r there are $r - j$ changes, so that:

$$v(i_j + s) + (r - j) + |B_{j+1}| + \cdots + |B_r| \leq 8p^2.$$

Hence, with $v(i_j + s) = i_j + s - m$, we have that:

$$(r - j) \leq 8p^2 - (|B_{j+1}| + \cdots + |B_r|) - v(i_j + s) = i_j + s - v(i_j + s) = m.$$

Therefore, (4.6) implies $8p^2 - (i_j + s) \leq m(p-1)$, so $i_j + s \geq 8p^2 - m(p-1)$.

Finally, let us consider the case $b = p$. This can happen only, when $m = \frac{p}{2}$. Then

$$v|_{B_1 \cup \cdots \cup B_{j-1}} : \{1, \ldots, i_j - 1\} \hookrightarrow \{1, \ldots, i_j + m - 1\},$$
$$v|_{B_{j+1} \cup \cdots \cup B_r} : \{i_j + s + 1, \ldots, 8p^2\} \hookrightarrow \{i_j + s + m + 1, \ldots, 8p^2\}$$

and the arguments used before yield $i_j \leq (p-1)m + 1$ and $8p^2 - (i_j + s) \leq mp$. ■

Proposition 4.5 states that a certain neighborhood of $(4p^2 + 1)$ belongs to the unbiased block. This means that, for every longest common subsequence, the elements

$$x_{(4p^2+1)-p^2}, x_{(4p^2+1)-p^2+1}, \ldots, x_{4p^2+1}, \ldots, x_{(4p^2+1)+p^2}$$

are included and directly matched. In particular, the element x_{4p^2+1} belongs to the same block and are directly matched. Similarly, x_{2p^2+1+m} is directly matched. This implies that we can define $x_1, \ldots, x_n = x_{m+1}, \ldots, x_{m+n}$ and $y_1, \ldots, y_n = x_1, \ldots, x_n$. Then, for every longest common subsequence, the element x_{2p^2+1} is directly matched.

4.2.2 The structure of a common subsequence between periodic subsequences with unequal length

In the previous Subsection, we analyzed the longest common subsequences of two periodic subsequences with length $8p^2$ in detail. We now consider the longest common subsequences between two finite periodic subsequence with unequal length. We study the case, when one sequence is still with length $8p^2$ and length of the other sequence differs from $8p^2$ by at most $2(p-1)$. Our aim is still to show that any longest common subsequence contains a unbiased block that is located in the center.
The proofs used in the present Subsection are essentially the same as the ones in the previous Subsection, but for a few additional technicalities. Therefore, we leave the proofs for the Appendix.

Proposition 4.6 *Let $x_1, \ldots, x_{8p^2}$ and $x_{l-m_1+1}, \ldots, x_{l+8p^2+m_2}$ be the subsequences of a periodic sequence, with $0 \leq m_1 \leq p-1$, $-m_1 \leq m_2 \leq p-1$ and $l = jp$, for a $j \in \mathbb{Z}$. Let $t_1 = (p - m_1) \mod p$, $t_2 = \max\{-m_2, 0\}$. Then the length of the longest common subsequence is $8p^2 - \min\{t_1, t_2\}$ and any longest common subsequence between $x_1, \ldots, x_{8p^2}$ and $x_{l-m_1+1}, \ldots, x_{l+8p^2-1}$ includes an unbiased block which contains x_{4p^2+1}.*

Proposition 4.7 *Let $x_1, \ldots, x_{8p^2}$ and $x_{l+m_1+1}, \ldots, x_{l-m_2+8p^2}$ be the subsequences of a periodic sequence, $0 \leq m_1 \leq p-1$, $-m_1 \leq m_2 \leq p-1$ and $l = jp$, for a $j \in \mathbb{Z}$. Let $t_1 = (p - m_1) \mod p$, $t_2 = \max\{-m_2, 0\}$. Then the length of the longest common subsequence is $8p^2 - m_1 - m_2$, if $m_2 \geq 0$ and $8p^2 - \min\{m_1, p + m_2\}$, else. Moreover, any longest common subsequence between $x_1, \ldots, x_{8p^2}$ and $x_{l+m_1+1}, \ldots, x_{l-m_2+8p^2}$ includes an unbiased block which contains x_{4p^2+1}.*

4.2.3 The structure of a common subsequence between periodic subsequences with mismatch

In the present Subsection, we consider the subsequences of a periodic sequence with the length $8p^2$. The only difference is that sequence $x_1, \ldots, x_{8p^2}$ has a *mismatch :* the element x_{4p^2+1} has been changed. So, formally, we consider the sequences $z_1, \ldots, z_{8p^2}$ and $x_{m+1}, \ldots, x_{m+8p^2}$, where $z_i = x_i$, $i = 1, \ldots, 4p^2, 4p^2+2, \ldots, 8p^2$ and $z_{4p^2+1} \neq x_{4p^2+1}$.

Proposition 4.8 *Let $z_1, \ldots, z_{8p^2}$ and $x_{t+1}, \ldots, x_{t+8p^2+h}$ be the subsequences of a periodic sequence with mismatch, $0 \leq t \leq \frac{p}{2}, 0 \leq h \leq p - 2t$. Then the length of the longest common subsequence is $8p^2 - t - 1$.*

Proof. Let v be a longest common subsequence of $z_1, \ldots, z_{8p^2}$ and $x_{t+1}, \ldots, x_{t+8p^2+h}$. The length of v is clearly at least $8p^2 - m - 1$.
Let us show that both subsequences $v|_{[1,4p^2]}$ and $v|_{[4p^2+2,8p^2]}$ have an unbiased block. By (4.4), v has at least one unbiased block $B_j = \{i_j, \ldots, n_j\}$. Assume $i_j > 4p^2 + 1$. It holds that:

$$v|_{[1,i_j-1]} : \{1, \ldots, i_j - 1\} \hookrightarrow \{1, \ldots, i_j - 1 - m + b\},$$

where $b \in \{0, 2t\}$ is the bias of B_j. Clearly the length of $v|_{[1,i_j-1]}$ is at least $i_j - 1 - t + \frac{b}{2}$. Let $B_1, \ldots, B_{r_1}$ be the blocks of $v|_{[1,i_j-1]}$. Suppose they all are biased. Then, with $u = i_j - (4p^2 + 2)$, we find:

$$\frac{r_1 - 1}{2} \geq \frac{i_j - 1 - (p-1) - t + \frac{b}{2}}{2(p-1)} = \frac{4p(p-1) + 2 + 3p + u - t + \frac{b}{2}}{2(p-1)} > 2p.$$

By Proposition 4.1, $i_j - 1 + \frac{b}{2} \geq 2p + |v|_{[1,i_j-1]}|$ or $|v|_{[1,i_j-1]}| \leq i_j - 1 + \frac{b}{2} - 2p$, which is a contradiction. Since the argument holds for any u, the unbiased block is contained in $\{1, \ldots, 4p^2\}$.
Hence, $B_1, \ldots, B_{r_1}$ contain at least one unbiased block.

Suppose the unbiased block B_j is contained in $\{1, \ldots, 4p^2\}$. It holds that:

$$v|_{[n_j+1,8p^2]} : \{n_j + 1, \ldots, 8p^2\} \hookrightarrow \{n_j + 1 + b, \ldots, t + 8p^2 + h\},$$

where $h = 0$, if $b = 2t$. Then $|v|_{[n_j+1,8p^2]}| \geq 8p^2 - n_j - 1 - \frac{b}{2}$. Let $C_1, \ldots, C_{r_2}$ be the blocks of $v|_{[n_j+1,8p^2]}$. Suppose they all are biased, hence, with $u = 4p^2 - n_j$,

$$\frac{r_2 - 1}{2} \geq \frac{4p(p-1) + 3p + u - \frac{b}{2}}{2(p-1)} > 2p.$$

By Proposition 4.1,

$$h + t + 8p^2 - n_j - 1 \geq |v|_{[n_j+1,8p^2]}| + 2p \geq 8p^2 - n_j - 1 + 2p - \frac{b}{2},$$

which is a contradiction. Since the argument holds for any u, the unbiased block is contained in $\{4p^2 + 1, \ldots, 8p^2\}$.

Let $l > j$ and B_j,B_l be unbiased blocks: $B_i \subset \{1, \ldots, 4p^2\}$, $B_l \subset \{4p^2+2, \ldots, 8p^2\}$.
If $t < \frac{p}{2}$, then the bias of both blocks is 0. Since v is the longest common subsequence, it follows that $|v| = 8p^2 - t - 1$ and the blocks are consecutive: $l = j + 1$ and

$$B_j = \{i_j, \ldots, 4p^2\}, \quad B_l = B_{j+1} = \{4p^2+2, \ldots, 4p^2+s\}. \tag{4.7}$$

If $\frac{t}{2}, p > 2$, then the bias of both blocks can be p as well. However, the length of v is still $8p^2 - t - 1$ and (4.7) holds. In both cases, the element z_{4p^2+1} is not included in v.
Finally, if $t = 1$ and $p = 2$, it might be that the bias of B_j is 0, the bias of B_{j+1} is 2 and the element z_{4p^2+1} is included in v. The length of v is still however equal to $8p^2 - t - 1$ ■

Proposition 4.9 *Let $z_1, \ldots, z_{8p^2}$ and $x_{m+1}, \ldots, x_{m+8p^2-h}$ be the subsequences of a periodic sequence with mismatch, $0 \leq 2m \leq p + h, 0 \leq h \leq m$. Then the length of the longest common subsequence is $8p^2 - m - 1$.*

Proof. The proof of Proposition 4.8 holds without changes. ■

Proposition 4.10 *Let $z_1, \ldots, z_{8p^2}$ and $x_{l-m_1+1}, \ldots, x_{l+8p^2+m_2}$ be the subsequences of a periodic sequence with mismatch, where $m_1 \leq p-1$, $-m_1 \leq m_2 \leq p-1$ and $l = jp$, for a $j \in \mathbb{Z}$. Let $t_1 = (p - m_1) \mod p$, $t_2 = \max\{-m_2, 0\}$. Then the length of the longest common subsequence is $8p^2 - \min\{t_1, t_2\} - 1$.*

Proposition 4.11 *Let $z_1, \ldots, z_{8p^2}$ and $x_{l+m_1+1}, \ldots, x_{l-m_2+8p^2}$ be the subsequences of a periodic sequence with mismatch, $m_1 \leq p-1$, $-m_1 \leq m_2 \leq p-1$ and $l = jp$, for a $j \in \mathbb{Z}$. Let $t_1 = (p - m_1) \mod p$, $t_2 = \max\{-m_2, 0\}$. Then the length of the longest common subsequence is $8p^2 - m_1 - m_2 - 1$, if $m_2 \geq 0$ and $8p^2 - \min\{m_1, p + m_2\} - 1$, else.*

4.3 Sequences with periodic pieces

4.3.1 Sequence with a periodic piece

Let $y_1, \ldots, y_n$ be a periodic sequence. Let $x_1, \ldots, x_n$ be a sequence with property:

$$\exists k \leq n - 8p^2 \text{ such that } x_{k+1} = y_{k+1}, x_{k+2} = y_{k+2}, \ldots, x_{k+8p^2} = y_{k+8p^2}. \tag{4.8}$$

So, the sequence $x_1, \ldots, x_n$ contains a periodic piece of length $8p^2$.
Let v be a longest common subsequence between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. We consider the integer interval $[\underline{v}(k+1), \bar{v}(k+8p^2)]$, and we show that the length of $[\underline{v}(k+1), \bar{v}(k+8p^2)]$ is about $8p^2$. The proofs of the following two propositions can be found in the Appendix.

Proposition 4.12 *Suppose the length of $[\underline{v}(k+1), \bar{v}(k+8p^2)]$ is not smaller than $8p^2$. Then there exist integers l, m_1, m_2 such that*

$$[\underline{v}(k+1), \bar{v}(k+8p^2)] = [l + 1 - m_1, l + 8p^2 + m_2], \tag{4.9}$$

where $|k - l| = jp$, for a non-negative $j \in \mathbb{N}$, $0 \leq m_1 \leq p-1$ and $-m_1 \leq m_2 \leq p-1$. In particular, the length of $[\underline{v}(k+1), \bar{v}(k+8p^2)]$ is at most $8p^2 + 2(p-1)$.

Proposition 4.13 *Suppose the length of* $[\underline{v}(k+1), \bar{v}(k+8p^2)]$ *is not bigger than* $8p^2$. *Then there exist integers* l, m_1, m_2 *such that*

$$[\underline{v}(k+1), \bar{v}(k+8p^2)] = [l+1+m_1, l+8p^2-m_2], \tag{4.10}$$

where $|k-l| = jp$, *for a non-negative* $j \in \mathbb{N}$ *and* $0 \leq m_1 \leq p-1$, $-m_1 \leq m_1 \leq p-1$. *In particular, the length of* $[\underline{v}(k+1), \bar{v}(k+8p^2)]$ *is at least* $8p^2 - 2(p-1)$.

4.3.2 Subsequence with a periodic piece and mismatch

Let $y_1, \ldots, y_n$ be a periodic sequence. Let $z_1, \ldots, z_n$ be a sequence with property: $\exists k \leq n - 8p^2$ such that

$$z_{k+1} = y_{k+1}, \ldots, z_{k+4p^2} = y_{k+4p^2}, z_{k+4p^2+1} \neq y_{k+4p^2+1}, z_{k+4p^2+2} = y_{k+4p^2+2}, \ldots, z_{k+8p^2} = y_{k+8p^2}. \tag{4.11}$$

Hence, the sequence $z_1, \ldots, z_n$ contains a periodic piece of length $8p^2$ with mismatch. ¿From the proofs of Propositions 4.12 and 4.13, the following corollaries can be deduced.

Corollary 4.3 *There exists a longest common subsequence* v *between* $z_1, \ldots, z_n$ *and* $y_1, \ldots, y_n$ *such that either (4.9) or (4.10) holds.*

4.4 Proof of Lemma 3.1

Corollary 4.4 *Let* $y_1, \ldots, y_n$ *be a periodic sequence. Let* $x_1, \ldots, x_n$ *be a sequence with property (4.8). Then any longest common subsequence between* $x_1, \ldots, x_n$ *and* $y_1, \ldots, y_n$ *has an unbiased block that contains the element* x_{k+4p^2+1}.

Proof. Let v be a longest common subsequence between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. We consider $[\underline{v}(k+1), \bar{v}(k+8p^2)]$. By optimality principle,

$$v|_{[k+1,k+8p^2]} : \{k+1, \ldots, k+8p^2\} \hookrightarrow \{\underline{v}(k+1), \ldots, \bar{v}(k+8p^2)\}$$

must be the longest common subsequence.
Suppose that the length of $[\underline{v}(k+1), \bar{v}(k+8p^2)]$ is bigger than $8p^2$. Then Proposition 4.12 and Proposition 4.6 apply.
Suppose that the length of $[\underline{v}(k+1), \bar{v}(k+8p^2)]$ is smaller than $8p^2$. Then Proposition 4.13 and Proposition 4.7 apply. ■

Corollary 4.5 *Let* L_n *be the length of the longest common subsequence of a periodic sequence* $y_1, \ldots, y_n$ *and a sequence* $x_1, \ldots, x_n$ *with the property (4.8). Let* $z_1, \ldots, z_n$ *be a sequence with the property (4.11). Then the length of the longest common subsequence of* $y_1, \ldots, y_n$ *and* $z_1, \ldots, z_n$ *is* $L_n - 1$.

Proof. Let v be a longest common subsequence between $z_1, \dots, z_n$ and $y_1, \dots, y_n$ that satisfies (4.9) ((4.10), resp.). By Corollary 4.3, such a v exists. Recall that $|L_n - |v|| \geq 1$. The length of v is the sum of the length of restrictions:

$$\begin{aligned} v|_{[1,k]} :&\{1, \dots, k\} \hookrightarrow \{1, \dots, \underline{v}(k+1) - 1\} \\ v|_{[k+1,k+8p^2]} :&\{k+1, \dots, k+8p^2\} \hookrightarrow \{\underline{v}(k+1), \dots, \bar{v}(k+8p^2)\} \\ v|_{[k+8p^2+1,n]} :&\{k+8p^2+1, \dots, n\} \hookrightarrow \{\underline{v}(k+8p^2+1), \dots, \bar{v}(n)\}. \end{aligned}$$

In this case, Proposition 4.10 (Proposition 4.11 resp.) specifies the length of $v|_{[k+1,k+8p^2]}$. Proposition 4.6 (Proposition 4.7 resp.) states: if $z_{k+1}, \dots, z_{k+8p^2}$ is replaced with $x_{k+1}, \dots, x_{k+8p^2}$, i.e. the mismatch has been removed, then there exists a common subsequence

$$v' : \{k+1, \dots, k+8p^2\} \hookrightarrow \{\underline{v}(k+1), \dots, \bar{v}(k+8p^2)\}$$

with length $|v|_{[k+1,k+8p^2]}| + 1$. Hence, the sequence v^* between $x_1, \dots, x_n$ and $y_1, \dots, y_n$, defined as

$$v^*|_{[1,k]} = v|_{[1,k]}, \;\; v^*|_{[k+1,k+8p^2]} = v', \;\; v^*|_{[k+8p^2+1,n]} = v|_{[k+8p^2+1,n]}$$

has length $|v| + 1$ and is, therefore, the longest common subsequence of $x_1, \dots, x_n$ and $y_1, \dots, y_n$. This proves the statement. ■

Proof of Lemma 3.1. Let $x_1, \dots, x_n$ be a realization of $X_1, \dots, X_n$ such that l is a replica point. Denote $y_1, \dots, y_n := Y_1, \dots, Y_n$. Recall that L_n is the length of the longest common subsequence of $x_1, \dots, x_n$ and $y_1, \dots, y_n$, and L_n^l is the length of the longest common subsequence of $x_1, \dots, x_{l-1}, x_{l+1}, \dots x_n$ and $y_1, \dots, y_n$. Recall

$$L_n - 1 \leq L_n^l \leq L_n. \tag{4.12}$$

Assume that A_l holds, i.e. l is a replica point. If the replica point matches, then $x_1, \dots, x_n$ is a sequence satisfying (4.8) with $x_{k+4p^2+1} = x_l$ being the replica point. Let L_n^+ be the length of the longest common subsequence of $x_1, \dots, x_n$ and $y_1, \dots, y_n$ with matching replica point. Suppose $L_n^+ = L_n^l$. Then any longest common subsequence of s $x_1, \dots, x_{l-1}, x_{l+1}, \dots x_n$ and $y_1, \dots, y_n$ would also be a longest common subsequence of $x_1, \dots, x_n$ and $y_1, \dots, y_n$. This contradicts Corollary 4.4 which states that any longest common subsequence of $x_1, \dots, x_n$ and $y_1, \dots, y_n$ contains x_l. Hence, $L_n^+ = L_n^l + 1 = L_n^l + Z_n$.
Suppose that the replica point does not match. Then $x_1, \dots, x_n$ is a sequence as in (4.11) with $x_{k+4p^2+1} = x_l$ being the mismatching replica point. Let L_n^- be the length of the longest common subsequence of $x_1, \dots, x_n$ and $y_1, \dots, y_n$ with mismatching replica point. By Corollary 4.5, $L_n^- = L_n^+ - 1$. By (4.12), $L_n^l \leq L_n^- = L_n^+ - 1 \leq L_n^l$, i.e. $L_n^- = L_n^l$.

References

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

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

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

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

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

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

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

[8] Michael S. Waterman. General methods of sequence comparison. *Bull. Math. Biol.*, 46(4):473–500, 1984.

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

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

5 Appendix

Proof of Proposition 4.2. Assume that there exists $k < p$ such that (4.3) hold. Then

$$x_{mk+j} = x_j \quad \forall m \geq 1, \quad j = 1, \ldots, p. \tag{5.1}$$

The latter implies

$$x_{k+n} = x_n \quad \forall n \geq 1.$$

that contradicts the definition of p.
Let us proof (5.1). Use induction: For $m = 1$, (5.1) is equivalent to (4.3).
Suppose that (5.1) holds for m. Let $k + j \leq p$. Then $x_{(m+1)k+j} = x_{mk+(k+j)} = x_{k+j} = x_j$. If $k + j > p$, then $x_{(m+1)k+j} = x_{mk+(k+j)} = x_{mk+k+j-p} = x_{k+j-p} = x_{k+j} = x_j$, To get the third inequality note that from $j \leq p$ follows $k + j - p < p$, and use (5.1).

5.1 Proofs of Propositions 4.6 and 4.7

Proposition 5.1 *Let $x_1, \ldots, x_{8p^2}$ and $x_{t+1}, \ldots, x_{t+8p^2+h}$ be the subsequences of a periodic sequence, $0 \le t \le \frac{p}{2}$, $0 \le h \le p - 2t$. Then the length of the longest common subsequence is $8p^2 - t$. Moreover, any longest common subsequence has an unbiased block B_j that contains the integer-interval $[tp + 1, 7p^2] \subset B_j$.*

Proof. Since $h \le p - 2t$, we have $p - (t + h) \ge t$, so t is the minimal bias between the two subsequences. In the proof of Proposition 4.4, replace the inequalities (4.4) with

$$\lfloor \frac{r-1}{2} \rfloor \ge \lfloor \frac{8p^2 - t + 1 - p}{2p} \rfloor \ge \lfloor 4p - \frac{t-1}{2p} - \frac{1}{2} \rfloor \ge 4p - 1 \ge t + h, \tag{5.2}$$

where the last inequality holds, because $t \le \frac{p}{2}$ and $h \le p$.
Let assume $b = 0$. Then the first half of the proof of Proposition 4.5 holds with any changes. For the second half, replace $8p^2$ by $8p^2 + h$. Then $8p^2 - (i_j + s) \le (t + h)(p - 1) \le p(p - 1)$ implying $(i_j + s) \le 7p^2 + p$. For $t = \frac{p}{2}$, $h = 0$. ■

Proposition 5.2 *Let $x_1, \ldots, x_{8p^2}$ and $x_{m+1}, \ldots, x_{m+8p^2-h}$ be the subsequences of a periodic sequence, $0 \le 2m \le p + h$, $0 \le h \le m$. Then the length of the longest common subsequence is $8p^2 - m$. Moreover, any longest common subsequence has an unbiased block B_j that contains the integer-interval $[mp + 1, 8p^2 - mp] \subset B_j$.*

Proof. By assumption, $2m \le p + h \le m + p$, i.e., $m \le p$. It holds, $p - m + h \ge m$, i.e. m is the minimal bias between the two subsequences. But it might be that $m > \frac{p}{2}$. The proof of Proposition 4.4 holds without any changes. Since $0 \le h \le m$, Proposition 4.5 holds, the only formal change is

$$v|_{B_{j+1} \cup \cdots \cup B_r} : \{i_j + s + 1, \ldots, 8p^2\} \hookrightarrow \{i_j + s - m + 1, \ldots, 8p^2 - h\}. \tag{5.3}$$

■

Proposition 5.3 *Let $x_{m_1+1}, \ldots, x_{m_1+8p^2}$ and $x_1, \ldots, x_{m_1+8p^2+m_2}$ be the subsequences of a periodic sequence, $0 \le m_1, m_2 \le p - 1$. The length of the longest common subsequence is $8p^2$ and each such subsequence of $x_{m_1+1}, \ldots, x_{m_1+8p^2}$ and $x_1, \ldots, x_{m_1+8p^2+m_2}$ includes an unbiased block which contains the interval $[p^2, 7p^2]$.*

Proof. Let $v : [1, 8p^2] \hookrightarrow [1, 8p^2 + m_1 + m_2]$ be a longest common subsequence, the length of v is clearly $8p^2$. Let $\{B_1, \ldots, B_r\}$ be the blocks of v. Suppose that all blocks are unbiased. Then $r \ge \frac{8p^2}{p-1}$. Since all the elements of the smallest subsequence are included in the longest common subsequence, by changing the blocks, v loses the elements on the bigger subsequence, only. Thus,

$$8p^2 + (r - 1) = \sum_{i}^{r} |B_i| + (r - 1) \le 8p^2 + m_1 + m_2,$$

implying that $r - 1 \le m_1 + m_2 \le 2(p - 1)$. This contradicts the lower bound for r.
Hence, there exists one and only one unbiased block $B_j = \{i_j, \ldots, i_j + s\}$. The bias of B_j can only be 0. Before the unbiased block, there are at most m_1 biased blocks, implying: $i_j \le m_1(p - 1) < p^2$. Similarly, $i_j + s \ge 7p^2$. ■

Proposition 5.4 *Let $x_1, \ldots, x_{8p^2}$ and $x_{m_1+1}, \ldots, x_{8p^2-m_2}$ be the subsequences of a periodic sequence, $0 \leq m_1, m_2 \leq p-1$. The length of a longest common subsequence is $8p^2 - (m_1+m_2)$ and every such a subsequence of $x_1, \ldots, x_{8p^2}$ and $x_{m_1+1}, \ldots, x_{8p^2-m_2}$ includes an unbiased block which contains the interval $[p^2, 7p^2]$.*

Proof. Let v be a longest common subsequence, the length of v is clearly $8p^2 - m_1 - m_2$. Let $\{B_1, \ldots, B_r\}$ be the blocks of v. Suppose that all blocks are unbiased. Then $r \geq \frac{8p^2-2p}{p-1}$. Since all the elements of the smallest subsequence are included in the longest common subsequence, by changing the blocks, v loses the elements on the bigger subsequence, hence. Thus,

$$8p^2 - (m_1+m_2) + (r-1) = \sum_{i}^{r} |B_i| + (r-1) \leq 8p^2,$$

implying that $r - 1 \leq m_1 + m_2 \leq 2p$. This contradicts with the lower bound of r.
So, there exists one and only one unbiased block $B_j = \{i_j, \ldots, i_j + s\}$. The bias of B_j is 0. Since before the unbiased block, there are at most t_1 biased blocks, we have: $i_j \leq m_1(p-1) + m_1 \leq p^2$. Similarly, $i_j + s + m_2(p-1) + m_2 \geq 8p^2$, so $i_j + s \geq 7p^2$. ■

Proof of Proposition 4.6. Suppose $t_2 = 0$. Then Proposition 5.3 applies.
If $t_1 = 0$, then $m_1 = 0$ and $t_2 = 0$, Proposition 5.3 applies again.
Suppose $t_1 > 0, t_2 > 0$. Assume $t_1 \leq t_2$. Note that $t_1 \leq \frac{p}{2}$. If not, then $m_1 = p - t_1 \leq \frac{p}{2}$, a contradiction with the assumption $m_1 \geq t_2$.
Since $l - m_1 = (l-1)p + t_1 = l^* + t_1$, we have

$$x_{l-m_1+1}, \ldots, x_{l-m_1+m_1+8p^2+m_2} = x_{l^*+t_1+1}, \ldots, x_{l^*+t_1+8p^2+m_2+m_1}.$$

Let $h = m_2 + m_1$. Clearly, $h = m_2 + m_1 \geq 0$ and $h = m_2 + m_1 = p - t_1 - t_2 \leq p - 2t_1$ since $-t_2 \leq -t_1$. Hence Proposition 5.1 applies.
Assume $t_1 \geq t_2$. Then $t_2 > \frac{1}{2}$ would imply that $m_1 > \frac{1}{2}$ and $t_1 \geq \frac{1}{2}$, a contradiction. We reverse the sequences, i.e we define

$$x'_1 = x_{8p^2}, x'_2 = x_{8p^2-1}, \ldots, x'_{8p^2} = x_1.$$

Then $x'_{t_2+1} = x'_{8p^2+m_2}, x'_{t_2+2} = x'_{8p^2+m_2-1}, \ldots, x'_{t_2+8p^2} = x_{m_2+1}, \ldots, x'_{t_2+8p^2+m_1-t_2} = x_{-m_1+1}$.
Take $h = m_1 - t_2 = m_1 + m_2 \geq 0$. It holds: $p - 2t_2 \geq p - t_1 - t_2 = m_1 - t_2 = h$. Now apply Proposition 5.1 to the reversed sequences. The reversing does not change the longest common subsequences (except reversing them). The element x_{4p^2+1} in the original sequence is the element x'_{4p^2}. By Proposition 5.1, it belongs to the unbiased block of any longest common subsequence.

Proof of Proposition 4.7. If $t_1 = 0$ then $m_2 \geq 0$. If $m_2 \geq 0$, then apply Proposition 5.4.
Let $0 < m_1 \leq p + m_2$. Define $h = m_1 + m_2 \geq 0$. Since $2m_1 \leq p + m_2 + m_1 = p + h$, Proposition 5.2 applies.
Let $0 < m_2 + p \leq m_1$. Then reverse the sequences as in the proof of Proposition 4.6 and apply Proposition 5.2.

5.2 Proofs of Proposition 4.10 and 4.11

Proposition 5.5 *Let $z_{m_1+1}, \ldots, z_{m_1+8p^2}$ and $m_1, \ldots, x_{m_1+8p^2+t_2}$ be the subsequences of a periodic sequence with mismatch, $0 \leq m_1, m_2 \leq p-1$. The length of the longest common subsequence is $8p^2-1$.*

Proof. Let v be a longest common subsequence of $z_{m_1+1}, \ldots, z_{m_1+8p^2}$ and $x_1, \ldots, x_{m_1+8p^2+m_2}$. By the argument used in the proof of Proposition 5.3, v has at least one unbiased block. The same argument, applied again, yields that the subsequences $v|_{[1,4p^2]}$ and $v|_{[4p^2+1,8p^2]}$ both have an unbiased block. If $p > 2$, then the bias of the unbiased blocks is 0, implying that the length of the longest common subsequence is $8p^2-1$.
When $p = 2$, the statement is easy to see. ■

Proposition 5.6 *Let $z_{m_1+1}, \ldots, z_{m_1+8p^2}$ and $x_1, \ldots, x_{8p^2-m_2}$ be the subsequences of a periodic sequence with mismatch, $0 \leq m_1, m_2 \leq p-1$. The length of the longest common subsequence is $8p^2-1-(m_1+m_2)$.*

Proof. Let v be a longest common subsequence of $z_{m_1+1}, \ldots, z_{m_1+8p^2}$ and $x_1, \ldots, x_{m_1+8p^2-m_2}$. By the argument used in the proof of Proposition 5.4, v has at least one unbiased block, by the same argument, $v|_{[1,4p^2]}$ and $v|_{[4p^2+1,8p^2]}$ both have an unbiased block. If $p > 2$, then the bias of the unbiased blocks is 0, implying that the length of the longest common subsequence is $8p^2-(m_1+m_2)-1$.
When $p = 2$, the statement is easy to see. ■

Proof of Proposition 4.10. Suppose $t_2 = 0$, i.e. $m_2 \geq 0$. If $t_1 = 0$, then $m_1 = 0$ and $t_2 = 0$. For $m_2 \geq 0$, Proposition 5.5 applies.
Suppose $t_1 > 0, t_2 > 0$. Assume $t_1 \leq t_2$. Then $t_1 \leq \frac{p}{2}$.
Since $l - m_1 = (l-1)p + t_1 = l^* + t_1$, we have

$$x_{l-m_1+1}, \ldots, x_{l-m_1+m_1+8p^2+m_2} = x_{l^*+t_1+1}, \ldots, x_{l^*+t_1+8p^2+m_2+m_1}.$$

Let $h = m_2 + m_1$. Clearly, $h = m_2 + m_1 \geq 0$ and $h = m_2 + m_1 = p - t_1 - t_2 \leq p - 2t_1$ since $-t_2 \leq -t_1$. Hence Proposition 4.8 applies.
Assume $t_1 \geq t_2$. Then $t_2 \leq \frac{1}{2}$. Reverse the sequences as in the proof of Proposition 4.6, i.e. we define

$$z'_1 = z_{8p^2}, z'_2 = z_{8p^2-1}, \ldots, z'_{8p^2} = z_1.$$

Note that in the reversed sequence, the mismatching element is z'_{4p^2} instead of z'_{4p^2+1}. However, it is easy to see that the proof of Propositions 4.8 holds also in this case.

Proof of Proposition 4.11. If $t_1 = 0$ then $m_2 \geq 0$. If $m_2 \geq 0$, then apply Proposition 5.6.
Let $0 < m_1 \leq p + m_2$. Define $h = m_1 + m_2 \geq 0$. Since $2m_1 \leq p + m_2 + m_1 = p + h$, Proposition 4.8 applies.
Let $0 < m_2 + p \leq m_1$. Then reverse the sequences as in the proof of Proposition 4.10 and apply Proposition 4.8.

5.3 Proofs of Propositions 4.12 and 4.13

Proof of Proposition 4.12. If $|[\underline{v}(k+1), \bar{v}(k+8p^2)]| = 8p^2$, the statement clearly holds. Suppose $|[\underline{v}(k+1), \bar{v}(k+8p^2)]| > 8p^2$. Then it holds: either $k+1 > \underline{v}(k+1)$ or $\bar{v}(k+8p^2) > (k+8p^2)$. Without loss of generality assume

$$\underline{v}(k+1) < k+1. \tag{5.4}$$

There $\exists l \geq 0$ such that $|k-l| = jp$, for a non-negative $j \in \mathbb{N}$ and

$$\underline{v}(k+1) = l - ip - m_1 + 1, \quad \bar{v}(k+8p^2) = l + 8p^2 + m_2,$$

where $0 \leq m_1 \leq p-1$ and $-m_1 \leq m_2 \leq p-1$, when $i = 0$ and $0 \leq m \leq p-1$, when $i \geq 1$.
The proposition is proven, if we show that $i = 0$. Suppose not. Then $0 \leq m \leq p-1$.
By the optimality principle, the subsequence

$$v|_{[k+1,k+8p^2]} : \{k+1, \ldots, k+8p^2\} \hookrightarrow \{l - ip - m_1 + 1, \ldots, l + 8p^2 + m_2\}$$

is the longest possible and its length is therefore equal to $8p^2$. Let

$$v' : \{k+1, \ldots, k+8p^2\} \hookrightarrow \{l+1, \ldots, l+8p^2\}$$

be a common subsequence that consists of a direct match:

$$v'(k+1) = l+1, \ldots, v'(k+8p^2) = l+8p^2.$$

The length of v' is also $8p^2$.
Let

$$w : \{1, \ldots, n\} \hookrightarrow \{1, \ldots, n\}$$

be a common subsequence of $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ that is defined as follows:

$$\begin{aligned} w|_{[1,k]} &= v|_{[1,k]} \\ w|_{[k+1,k+8p^2]} &= v' \\ w|_{[k+8p^2+1,n]} &= v|_{[k+8p^2+1,n]} \end{aligned}$$

Hence, w is a modification of v obtained by $v|_{[k+1,k+8p^2]}$ replaced by a direct matching v'. Of course, the length of w is the same as the length of v, hence, w is the longest common subsequence.
The subsequence w has the following property: $[1, \bar{w}(k)] = [1, l]$, but

$$\begin{aligned} \underline{w}(k+1) &= w(\max\{i \leq k : i \in \mathrm{Dom}(w)\}) + 1 \\ &= v(\max\{i \leq k : i \in \mathrm{Dom}(v)\}) + 1 = \underline{v}(k+1) = l - ip - m_1 + 1. \end{aligned}$$

Hence, the interval $[l - ip - m_1 + 1, l]$ does not contain any element of w. This means that the subsequence

$$w|_{[1,k]} : \{1, \ldots, k\} \hookrightarrow \{1, \ldots, l\} \tag{5.5}$$

is actually a subsequence

$$w|_{[1,k]} : \{1, \dots, k\} \hookrightarrow \{1, \dots, l - ip - m_1\}.$$

We shall show that this property contradicts the optimality principle.
By (5.4), $k > l - m_1 - ip$. Let

$$t = \max\{i \le k : i \notin \mathrm{Dom}(v)\}.$$

We have: $w(t+1), \dots, w(k) \le l - ip - m_1$. Define $w' : \{1, \dots, k\} \hookrightarrow \{1, \dots, l\}$,

$$\begin{aligned} w'|_{[1,t]} &= w|_{[1,t]} \\ w'(t+1) = w(t+1) + p, \dots, w'(k) &= w(k) + p. \end{aligned}$$

Since $w(k) \le l$, the sequence w' is well defined and has the length as (5.5). Let s be the last element of w before t, i.e. $s = \max\{i < t : i \in \mathrm{Dom}(w)\}$. By definition of w', $w'(t+1) = w(t+1) + p \ge w'(s) + 1 + p$, so the interval $[w'(s)+1, w'(s)+p]$ does not contain any elements of w'. By periodicity, the interval $[y_{w'(s)+1}, y_{w'(s)+p}]$ contains at least one 0 and at least one 1. On the other hand, the unconnected element x_t is either 0 or 1. Therefore, we can connect the element x_t with an element of $[y_{w'(s)+1}, y_{w'(s)+p}]$. The possibility of such a connection shows that w' is not the longest common subsequence. This, in turn, implies that (5.5) can not be the longest common subsequence. By the optimality principle, the latter implies that w and, hence, v cannot be the longest common subsequences as well. This is a contradiction. The reason for the contradiction is the assumption $i \ge 1$.

Proof of Proposition 4.13. If $|[\underline{v}(k+1), \bar{v}(k+8p^2)]| = 8p^2$, the statement clearly holds. Suppose $|[\underline{v}(k+1), \bar{v}(k+8p^2)]| < 8p^2$. Then it holds: either $k+1 < \underline{v}(k+1)$ or $\bar{v}(k+8p^2) < (k+8p^2)$. Without loss of generality assume

$$\bar{v}(k+8p^2) < (k+8p^2). \tag{5.6}$$

There $\exists l \ge 0$ such that $|k-l| = jp$, for a non-negative $j \in \mathbb{N}$ and

$$\underline{v}(k+1) = l + m_1 + 1, \quad \bar{v}(k+8p^2) = l - ip + 8p^2 - m_2 =: u_l,$$

where $0 \le m_1 \le p-1$ and $-m_1 \le m_2 \le p-1$, when $i = 0$, and $0 \le m \le p-1$, when $i \ge 1$. Proposition is proved, if we show that $i = 0$. Suppose $i > 0$. Then $0 \le m_1 \le p-1$. By the optimality principle, the subsequence

$$v|_{[k+1,k+8p^2]} : \{k+1, \dots, k+8p^2\} \hookrightarrow \{l + m_1 + 1, \dots, u_l\}$$

is the longest possible, the length of it is, therefore, $L := 8p^2 - (m_1 + m_2 + ip)$. Let

$$v' : \{k+1, \dots, k+8p^2\} \hookrightarrow \{l + m_1 + 1, \dots, u_l\}$$

be a common subsequence that consists of a direct match:

$$v'(k+1+m_1) = l+m_1+1, \ldots, v'(k+8p^2-ip-m_2) = l-ip+8p^2-m_2 = u_l.$$

The length of v' is also L.
Let

$$w : \{1, \ldots, n\} \hookrightarrow \{1, \ldots, n\}$$

be a common subsequence of $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ that is defined as follows:

$$\begin{aligned} w|_{[1,k]} &= v|_{[1,k]} \\ w|_{[k+1,k+8p^2]} &= v' \\ w|_{[k+8p^2+1,n]} &= v|_{[k+8p^2+1,n]} \end{aligned}$$

Of course, the length of w is the same as the length of v, hence, w is the longest common subsequence of $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$.
The subsequence w has the following property: $u_k := k+8p^2-ip-m_2 \in \mathrm{Dom}(w)$, and the next element in $\mathrm{Dom}(w)$ is not earlier as $k+8p^2+1$: $\min\{i \geq u_k : i \in \mathrm{Dom}(w)\} \geq k+8p^2+1$. In particular, this implies: $\underline{w}(k+8p^2+1) = u_l+1$ or

$$|w|_{[u_k+1,n]}| = |w|_{[k+8p^2+1,n]}|. \tag{5.7}$$

Note:

$$w|_{[k+8p^2+1,n]} : \{k+8p^2+1, \ldots, n\} \hookrightarrow \{u_l+1, \ldots, n\}.$$

By (5.6), $k+8p^2+1 < u_l+1$, so there exists at least one element $j \in [u_l+1, n]$ such that y_j does not belong to the subsequence $w|_{[k+8p^2+1,n]}$. Let

$$t = \min\{j \geq u_l+1 : j \notin w([k+8p^2+1, n])\}. \tag{5.8}$$

Suppose $t \in [u_l+1, l+8p^2]$. Let r be such that $w(r) = t-1$, i.e. $r = w^{-1}(t-1)$. Obviously, $r \in [k+8p^2+1, n]$. Define

$$v'' : \{k+1, \ldots, u_k+(t-u_l)\} \hookrightarrow \{l+m_1+1, \ldots, t\}$$

be a common subsequence that consists of a direct match:

$$v''(k+1+m_1) = l+m_1+1, \ldots, v''(u_k) = u_l, v''(u_k+1) = u_l+1, \ldots, v''(u_k+(t-u_l)) = t.$$

The definition of v'' is possible, since $t-u_l \leq ip+m_2$ and $(t-u_l) \leq k+8p^2$.
The length of v'' is $L+(t-u_l)$. Define $w' : \{k+1, \ldots, n\} \hookrightarrow \{l+m_1+1, \ldots, n\}$,

$$\begin{aligned} w'|_{[k+1,u_k+(t-u_l)]} &= v'' \\ w'|_{[u_k+(t-u_l)+1,n]} &= v|_{[r+1,n]}. \end{aligned} \tag{5.9}$$

By the definition of t and r, $|v|_{[k+8p^2+1,n]}| - |v|_{[r+1,n]}| = |v|_{[k+8p^2+1,r]}| = (t-u_r)-1$. Hence, the length of w' is $L+(t-u_l)+|v|_{[k+8p^2+1,n]}| - (t-u_r)+1 = L+|v|_{[k+8p^2+1,n]}|+1 =$

$|w|_{[k+1,n]}|+1$ which contradicts the assumption that w is a longest common subsequence.

Suppose $t \in [l+8p^2+1, n]$. Then $t-p \geq u_l+1$ and by the definition of t, the elements $y_{u_l+1}, \ldots, y_{t-1}$ all belong to the common subsequence w. Let

$$v'' : \{u_k+1, \ldots, w^{-1}(t-p)\} \hookrightarrow \{u_l+1, \ldots, t\}$$

be defined as follows:

$$v''(u_k+1) = u_l+1, \ldots, v''(u_k+p) = u_l+p, v''(w^{-1}(u_l+1)) = u_l+1+p, \ldots, v''(w^{-1}(t-p)) = t.$$

The definition of v'' is possible, because $w^{-1}(u_l+1) \geq k+8p^2+1$. The length of v'' is $|w|_{[k+8p^2+1, w^{-1}(t-p)]}|+p$. Note

$$|w|_{[w^{-1}(t-p)+1, w^{-1}(t-1)]} = |w|_{[k+8p^2+1, w^{-1}(t-1)]}| - |w|_{[k+8p^2+1, w^{-1}(t-p)]}| = p-1.$$

We define $w' : \{u_k+1, \ldots, n\} \hookrightarrow \{u_l+1, \ldots, n\}$, where

$$w'|_{[u_k+1, w^{-1}(t-p)]} = v'',$$
$$w'|_{[w^{-1}(t-p)+1, n]} = w|_{[w^{-1}(t-1)+1, n]}.$$

The definition of w' is correct, because $w(w^{-1}(t-1)+1) > t$. The length of w' is

$$|w|_{[k+8p^2+1, w^{-1}(t-p)]}| + p + |w|_{[w^{-1}(t-1)+1, n]}| = |w|_{[k+8p^2+1, w^{-1}(t-1)]}| + |w|_{[w^{-1}(t-1)+1, n]}| + 1 =$$
$$= |w|_{[k+8p^2+1, n]}| + 1.$$

By (5.7), the length of w' is strictly bigger than that of

$$w|_{[u_k+1, n]} : \{u_k+1, \ldots, n\} \hookrightarrow \{u_l+1, \ldots, n\}.$$

This contradicts the assumption that w is a longest common subsequence.

Proof of Corollary 4.3 Let v be a longest common subsequence of $z_1, \ldots, z_n$ and $y_1, \ldots, y_n$. Suppose $[\underline{v}(k+1), \bar{v}(k+1)]$ is bigger than $8p^2$ but does not satisfy (4.9). Then there exists $0 \leq m_1, m_2 \leq p-1$, $i > 0$, such that

$$v|_{[k+1, k+8p^2]} : \{k+1, \ldots, k+8p^2\} \hookrightarrow \{l-ip-m_1+1, \ldots, l+8p^2+m_2\}. \qquad (5.10)$$

Suppose $i \geq 2$. Then, assuming (5.4), it holds $\underline{v}(k+1)+2 < k+1$. The length of $|v|_{[k+1, k+8p^2]}|$ is $8p^2$. Define the common subsequence w as in the proof of Proposition 4.12. The length of w is $|v|-1$, but the length of the empty interval $[l-ip-m_1+1, l]$ is at least $2p$. Since there are at least two elements in $[1,k]$, say t_1 and t_2, not included into $\mathrm{Dom}(v)$, by rearranging the elements of $w|_{[1,k]}$ as in the proof of Proposition 4.12, both $z_{t_1} = x_{t_1}$ and $z_{t_2} = x_{t_2}$ can be matched with an empty period. So, the length of $w|_{[1,k]}$ can be increased by 2. This contradicts the assumption that v is the longest common subsequence.

This means that in (5.10), $i = 1$. Now, again, use the argument of Proposition 4.12: Define the common subsequence w and note that the length of w is $|v| - 1$. Then rearrange the elements of $w|_{[t,k]}$ by defining $w'(t+1) = w(t+1)+p, \ldots, w'(k) = w(k)+p = l-p-m_1+p = l - m_1$ and connect the element x_t with some element on $[y_{w'(s)+1}, y_{w'(s)+p}]$. Let

$$w^* : \{1, \ldots, k\} \hookrightarrow \{1, \ldots, l - m_1\},$$

be modification of w' with connected x_t so the length of w^* is $|w|_{[1,k]}| + 1$. Hence, the sequence v^* with

$$\begin{aligned} v^*|_{[1,k]} &= w^* \\ v^*|_{[k+1,k+8p^2]} &= w \\ v^*|_{[k+8p^2+1,n]} &= w|_{[k+8p^2+1,n]} \end{aligned}$$

has length $|w| + 1$, which is the same as the length of v. Since $v^*(k) = w^*(k) = w'(k) = l - m_1$, the sequence v^* satisfies (4.9).

Suppose $[\underline{v}(k + 1), \bar{v}(k + 1)]$ is bigger than $8p^2$ but does not satisfy (4.10). The proof is similar: as in the proof of Proposition 4.13, define the subsequence w and note that the length of w is $|v| - 1$. Define t as in (5.8), and $w^{'}$ as in (5.9). With the help of w', construct the common subsequence v^* with $v^*|_{[1,k]} = v|_{[1,k]}$ and $v^*|_{[k+1,n]} = w^{'}$. The length of v^* is the same as the length of v. If v^* does not satisfy (4.10), then use the proof of Proposition 4.13 to see that w^* satisfies (4.10).