LOCAL UNIQUENESS OF ALIGNMENTS WITH A FIXED PROPORTION OF GAPS

RAPHAEL HAUSER[†] AND HEINRICH MATZINGER [‡]

Abstract. We consider two independent random strings with i.i.d. characters and examine their optimal alignments containing a fixed proportion of gaps. We prove that when the proportion of gaps is small then with high probability optimal alignments differ only in a small number of places and are locally unique everywhere else. The result is somewhat surprising, as one might expect unrelated sequences to admit many near-optimal alignments, whereas for related sequences one might expect an optimal alignment that is unique in many places.

AMS subject classifications. Primary 60F10, Secondary 92D20.

Key words. Random strings, sequence alignment, large deviations.

1. Introduction. To motivate the main issues under investigation in this paper, let us compare the German name 'heinrich' with its Spanish equivalent 'enrique'. A naive comparison of the two strings would be to align them character by character and count the number of matched entries,

$$\begin{array}{cccccccc} h & e & i & n & r & i & c & h \\ e & n & r & i & q & u & e & \end{array}$$

We see that in this case the similarity score would be zero, as no aligned characters match each other, despite the close resemblance of the two names. Such a string comparison technique is too simplistic to yield any interesting results.

A better measure for the similarity of the two strings is the length of a *longest common subsequence* (LCS). A common subsequence is any string that can be obtained by deleting some characters of either string and keeping the remaining ones in the original order. In the case of our example, 'enri' turns out to be the longest common subsequence, a result whose significance is further underlined by its close similarity to the French version of the name, 'henri'.

Of course, the similarity measure could be further improved: A linguist would be quick to point out that the groups of characters 'ch', 'k' and 'qu' are all closely related, and so are 'he' and 'e' if they occur at the very beginning of a word. Matching up these and other similar groups of characters to count towards a total similarity score, one could design a scoring function that is particularly well suited as a tool in the etymological comparison of Indo-European languages for example.

A fairly general and useful technique to identify high quality alignments of two strings $x = x_1 \dots x_m$ and $y = y_1 \dots y_n$ with characters from a finite alphabet $\mathbb{A}$ is to consider alignments with gaps $\sqcup$,

$$\begin{array}{ccccc} \sqcup & x_1 & x_2 & x_3 & \sqcup \\ y_1 & y_2 & \sqcup & y_3 & y_4 \end{array}$$

[†]Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford, OX1 3QD, United Kingdom, (hauser@comlab.ox.ac.uk). This author was supported through grant GR/S34472 from the Engineering and Physical Sciences Research Council of the UK and through grant NAL/00720/G from the Nuffield Foundation.

[‡]Fakultät für Mathematik, Universität Bielefeld, D-33501 Bielefeld, Germany, (matzing@mathematik.uni-bielefeld.de). Also: School of Mathematics, Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA 30332-0160, USA, (matzi@math.gatech.edu).

and to quantify the quality of such an alignment with a score of the form

$$S(x,y) = s(\sqcup, y_1) + s(x_1, y_2) + s(x_2, \sqcup) + s(x_3, y_3) + s(\sqcup, y_4). \tag{1.1}$$

The choice of individual scores $s(a,b)$ of matched symbols a and b depends on the specific area of application. The matching of any symbol $a \in \mathbb{A}$ with a gap $\sqcup$ is typically penalized by a negative score term $s(a,\sqcup)$, $s(\sqcup,a) < 0$.

Note that the LCS score is a special case of this type of scoring function: setting $s(a,b) = 1$ if $a = b \neq \sqcup$, $s(a,b) = \infty$ if $a \neq b$ and $a, b \neq \sqcup$, and $s(a,\sqcup) = s(\sqcup,a) = 0$ for all $a \neq \sqcup$, the LCS string 'enri' is found via the score maximizing alignment

$$\begin{array}{cccccccccccc} h & e & i & n & r & i & c & h & \sqcup & \sqcup & \sqcup \\ \sqcup & e & \sqcup & n & r & i & \sqcup & \sqcup & q & u & e \end{array}$$

Note that this optimal alignment is not unique, as

$$\begin{array}{cccccccccccc} h & e & i & n & r & i & c & \sqcup & h & \sqcup & \sqcup \\ \sqcup & e & \sqcup & n & r & i & \sqcup & q & \sqcup & u & e \end{array}$$

leads to the same score (and here to the same LCS, although this is not necessarily the case in general).

Although such string comparison techniques play an overwhelmingly important role in various domains of applications (particularly in biology, see e.g. [18], but also in speech recognition, pattern recognition and other areas where hidden Markov models are used as an analytic tool), their mathematical underpinning is relatively poorly understood. A central question concerns whether a high score provides significant evidence of a close relationship between two strings or whether the same score can be explained in terms of the statistical fluctuation of the scores of random strings. The study of optimal alignments of random strings is also highly interesting to statistical physicists, because it can be shown that optimal alignments occur naturally in some first passage percolation problems with correlated weigths.

The LCS setting is mathematically the best understood case of maximum score alignments of random strings. We briefly summarize some of the existing literature: Let L_n denote the length of the LCS of two independent binary i.i.d. sequences of length n. Using a subadditivity argument, Chvàtal-Sankoff [8] showed that the limit

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

exists. Determining the exact value of γ – the so-called *Chvàtal-Sankoff constant* – is a long standing open problem. However, Chvàtal-Sankoff [8] derived both upper and lower bounds on γ. Using an entropy argument, similar upper bounds were found by Baeza-Yates-Gavalda-Navarro-Scheihing [5]. These bounds were later improved by Deken [10], and subsequently by Dancik-Paterson [9, 15]. Hauser-Martinez-Matzinger [11] extended the Dancik-Paterson method further to generate large-deviation based probabilistic upper bounds on γ which approximate the exact value to two correct digits on a 95% confidence level. Other results concerning γ are as follows: Kiwi-Loebl-Matousek [13] considered the case of random strings with i.i.d. entries uniformly distributed on a finite alphabet $\mathbb{A}$ and determined the exact asymptotic dependence of γ on the cardinality $k = \#\mathbb{A}$ of the alphabet when k goes to infinity. And finally, using first passage percolation methods, Alexander [2] proved that $E[L_n]/n$ converges to γ at a rate of order $\sqrt{\log n/n}$.

Another long standing open problem in the LCS context is to determine the exact order of the fluctuation of the LCS length. Steele [17] proved that $VAR[L_n] \leq n$. Montecarlo simulations led Chvàtal-Sankoff [8] to conjecture that $VAR[L_n] = o(n^{2/3})$. This order of magnitude is similar to the order for the so-called *longest increasing subsequence* (LIS) of random permutations, see Baik-Deift-Johansson [6] and Aldous-Diaconis [1]). Waterman [19] conjectured that in many cases the variance of L_n grows linearly. We believe that there may exist different possible orders of magnitude for these fluctuations, depending on the distribution of the strings X and Y. Bonetto-Matzinger [7] considered the asymmetric case where X contains one symbol less than Y and proved that in this case $VAR[L_n] = \Theta(n)$. The same order was shown by Lember-Matzinger [14] for the setup where one sequence is not random but periodic. Further contributions to the order of fluctuations were made by Arratia-Waterman [3] by deriving a law of large deviation for L_n for fluctuations on scales larger than $\sqrt{n}$. In the same ground-breaking article the existence of interesting phase transition phenomena was shown.

The present paper forms part of a forthcoming series of research reports that investigate different aspects of the probabilistic behavior of scoring functions $S(x, y)$ applied to random strings. To motivate the particular aspect investigated here, let us go back to the optimal LCS alignments of the strings 'heinrich' and 'enrique': If we were to give an exhaustive list of optimal alignments, one would observe that all of these start as follows,

$$\begin{array}{cccccc} h & e & i & n & r & i \\ \sqcup & e & \sqcup & n & r & i \end{array}$$

We say that in this region the optimal alignment is *locally unique*. Thus, in this example the optimal LCS alignment is locally unique on a large proportion of the two strings.

This observation is typical not only in the LCS setting, but for general scoring functions $S(x, y)$ as introduced in (1.1): When two strings are closely related to one another, the optimal alignments are locally unique in many places. Conversely, the optimal alignments of two random strings with i.i.d. entries often are locally unique only in very few places. One might be tempted to exploit this observation in an algorithm to extract the most significantly related parts of two (nonrandom) sequences. However, if one were to do this, one would have to select the gap penalty $s(a, \sqcup)$ quite carefully: When gaps are strongly penalized, then no more than a constant proportion of gaps are observed in optimal alignments, and if the proportion of gaps is small then the optimal alignment is locally unique in most places even for i.i.d. random (and hence totally unrelated) strings.

Our paper concerns a theoretical analysis of this phenomenon. To make the analysis transparent, we chose a simplified setup in which $X = X_1 \dots X_m$ and $Y = Y_1 \dots Y_n$ are random strings consisting of i.i.d. standard Bernoulli variables, where $m = \lfloor (1 - \delta)n \rfloor$ depends on n via a fixed gap proportion δ. We then investigate alignments of X and Y that contain gaps only in X, and we use a scoring function in which matching symbols contribute to the total score with unit weight and non-matching ones with zero weight. Our interest is in the random number $U \leq m$ of indices i for which X_i is aligned with more than one Y_j under the different optimal alignments. The main theorem of this paper shows that when δ is small and n is large, optimal alignments are locally unique in an arbitrarily large proportion of places with arbitrarily high probability:

THEOREM 1.1. *For all $\varepsilon > 0$ there exists $\delta_0 > 0$ and $n_0 \in \mathbb{N}$ such that for all $\delta \in (0, \delta_0)$ and $n > n_0$,* $\mathrm{P}\left[U \geq m\varepsilon\right] < \varepsilon$.

While it is clear that $\mathrm{P}[U > m\varepsilon] = 0$ when $\delta = 0$ for any n, the theorem shows the nontrivial fact that the limit $\lim_{n\to\infty} \mathrm{P}[U > m\varepsilon]$ is continuous in δ at $\delta = 0$. Furthermore, the proof provides the quantitative estimate

$$\mathrm{P}\left[U \geq m\varepsilon\right] \leq \frac{\mathcal{O}\left(\delta^{\frac{1}{2}}\right)}{\mathcal{O}(\varepsilon) + \mathcal{O}\left(\delta^{\frac{1}{2}}\right)} + \mathcal{O}\left(\mathrm{e}^{-n\delta}\right).$$

Theorem 1.1 is also very interesting in the context of the Chvàtal-Sankoff conjecture which concerns the order of magnitude of the fluctuation of the LCS of two random texts. In a paper currently under preparation we show that there exists a deep connection between this order of magnitude of fluctuation and the points where the optimal alignment is unique.

2. The Intuition Behind the Main Theorem. We recall the assumptions under which we prove our main result and which will remain valid throughout the rest of this paper: Let $n \in \mathbb{N}$ and let $0 < \delta < 1$ be a fixed constant not depending on n. We set $m = \lfloor n - \delta n \rfloor$ and define two independent random strings $X = X_1 \ldots X_m$ and $Y = Y_1 \ldots Y_n$ by choosing $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ as i.i.d. Bernoulli variables with parameter $1/2$ (i.e., coin tossing experiments). We then consider alignments

$$\begin{array}{cccccccccccc} \sqcup & \ldots & \sqcup & X_1 & \sqcup & \ldots & \sqcup & X_m & \sqcup & \ldots & \sqcup \\ Y_1 & \ldots & Y_{\xi(1)-1} & Y_{\xi(1)} & Y_{\xi(1)+1} & \ldots & Y_{\xi(m)-1} & Y_{\xi(m)} & Y_{\xi(m)+1} & \ldots & Y_n \end{array}$$

with gaps in X only and attribute to it the score

$$S(X, Y; \xi) = \#\{i \in \mathbb{N}_m : X_i = Y_{\xi(i)}\},$$

where $\#$ denotes the cardinality of a set and $\mathbb{N}_n := \{1, \ldots, n\}$. The set of alignments ξ that maximize $S(X, Y; \xi)$ is denoted by $\mathcal{O}A_{X,Y}$. Of course, this is a random set of alignments, since X and Y are random. Finally, we write

$$U := \#\{i \in \mathbb{N}_m : \exists\, \xi, \lambda \in \mathcal{O}A_{X,Y} \text{ s.t. } \xi(i) \neq \lambda(i)\}$$

for the number of positions where X is not uniquely aligned with Y among the alignments with maximum score. U is a random variable.

Theorem 1.1 states that $\mathrm{P}\left[U \geq m\varepsilon\right] < \varepsilon$ for all n large enough and δ small enough. In other words, for large n and small gap proportion δ the optimal alignment is locally unique in a $(1-\epsilon)$-proportion of the string X with probability greater than $1 - \epsilon$. We believe that this result is qualitatively representative for what occurs with regards to the local uniqueness of optimal alignments of random strings under *arbitrary* scoring functions $S(X, Y)$ as defined in (1.1) whenever gaps are strongly penalized, i.e., $s(a, \sqcup)$ is a negative number of not too small a modulus. Indeed, strong gap penalization prevents optimal alignments from having more than a small proportion of gaps. Furthermore, allowing gaps only in one of the strings is merely a technical assumption that vastly simplifies the analysis; extra work allows in principle to extend our analysis to the case where a small number of gaps occurs in both strings.

Let us now briefly explain the main idea behind the proof of Theorem 1.1. We define a measure preserving map by picking an entry of X at random and flipping it to the opposite value. By this we mean that if the entry was 1 we change it to 0, and if it was 0 we change it to 1 (the term "flipping" stems from imagining X as a line-up of randomly tossed fair coins). We denote the string obtained in this fashion by $\tilde{X}$. Since this operation is measure preserving, we have

$$\mathrm{E}[\Delta] = 0, \tag{2.1}$$

where $\Delta := S^*(\tilde{X}, Y) - S^*(X, Y)$. The crucial observation is now that when the optimal alignment is nonunique in a large proportion of places, then the optimal score tends to increase under this measure-preserving map. We illustrate this phenomenon in Example 2.1 below. Together with 2.1 this observation implies that the probability that the optimal alignment is nonunique in many points is small.

EXAMPLE 2.1. *Consider the case where $n = 8$, $m = 6$, $\delta = 1/4$ and X and Y take the values $x = 001110$ and $y = 11110011$. There are two optimal alignments, ξ given by $\xi(i) = i$ for $(i = 1, \ldots, 6)$ or*

$$\begin{array}{cccccccc} 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & \sqcup & \sqcup, \end{array}$$

and λ given by $\lambda(i) = i$ for $(i = 1, \ldots, 4)$ and $\lambda(5) = 7$, $\lambda(6) = 8$ or

$$\begin{array}{cccccccc} 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & \sqcup & \sqcup & 1 & 0. \end{array}$$

The optimal score is $S^*(x, y) = S(x, y; \xi) = S(x, y; \lambda) = 3$.

The following combinatorial property holds for arbitrary alignments ξ, λ of x and y, not only optimal ones: If $i \in \{1, \ldots, m\}$ is such that

$$y_{\xi(i)} \neq y_{\lambda(i)} \tag{2.2}$$

then flipping x_i to the opposite value increases at least one of the scores $S(x, y; \xi)$, $S(x, y; \lambda)$ by one unit. In particular, if ξ and λ are both optimal alignments and condition (2.2) holds, then flipping x_i to the opposite value increases the optimal score by one unit. For the chosen values of x and y we find that $i = 5, 6$ satisfy condition 2.2. Flipping the value of x_5 from 1 to 0, we find that the score of ξ increases to 4 and the score of λ decreases to 2. The maximum score is now 4. Similarly, flipping x_6 from 0 to 1, the score of ξ decreases to 2 whereas the score of λ increases to 4. The new maximum score is again 4. If a random entry x_T of x is flipped (with $T \in \mathbb{N}_m$), then this implies

$$\mathrm{E}[\Delta \| X = x, Y = y, \xi(T) \neq \lambda(T), Y_{\xi(T)} \neq Y_{\lambda(T)}] = 1. \tag{2.3}$$

On the other hand, if one of the entries $x_1, \ldots, x_4$ that do not satisfy (2.2) is flipped, then the maximum score can either increase or decrease: For $i = 1 \ldots, 4$, we have $\xi(i) = \lambda(i)$. The entries x_1 and x_2 are aligned with non-matching symbols, so that flipping one of these entries increases the optimal score by one. The entries x_3 and x_4 are aligned with matching symbols. In the present example, switching one of these entries results in a decrease of the optimal score by one unit, though in other cases the maximum score can remain unchanged (but it will then be attained by a

different alignment). Thus, we find that if a random entry x_T of x is flipped (where $T \in \mathbb{N}_m$) then

$$\mathrm{E}[\Delta \| X = x, Y = y, \xi(T) = \lambda(T)] \geq 0. \tag{2.4}$$

For the same reason, if there were any indices i such that $\xi(i) \neq \lambda(i)$ where (2.2) does not hold, then we would find

$$E[\Delta \| X = x, Y = y, \xi(T) \neq \lambda(T), Y_{\xi(T)} \neq Y_{\lambda(T)}] \geq 0.$$

Together with (2.3) this implies

$$\begin{aligned} &\mathrm{E}[\Delta \| X = x, Y = y, \xi(T) \neq \lambda(T)] \\ &\qquad \geq \mathrm{P}[Y_{\xi(T)} \neq Y_{\lambda(T)} \| X = x, Y = y, \xi(T) \neq \lambda(T)]. \end{aligned} \tag{2.5}$$

In the proof of Theorem 1.1 we exploit a generalization of the same mechanism. Lemma 3.3 of Section 3 shows that there exist two optimal alignments ξ and λ that differ from each other exactly in those positions i where the optimal alignment of X and Y is not locally unique. In Section 4 we show that up to negatively exponentially small probability in n the following are true,

i) approximately half the points $i \in \mathbb{N}_m$ with $\xi(i) = \lambda(i)$ satisfy $X_i \neq Y_{\lambda(i)}$,
ii) approximately half the points $i \in \mathbb{N}_m$ with $\xi(i) \neq \lambda(i)$ satisfy $Y_{\xi(i)} \neq Y_{\lambda(i)}$,
iii) approximately a quarter of points $i \in \mathbb{N}_m$ with $\xi(i) \neq \lambda(i)$ satisfy $X_i \neq Y_{\xi(i)} = Y_{\lambda(i)}$,
iv) approximately a quarter of points $i \in \mathbb{N}_m$ with $\xi(i) \neq \lambda(i)$ satisfy $X_i = Y_{\xi(i)} = Y_{\lambda(i)}$,
v) for all $e_1, e_2 \in \{0, 1\}$ approximately a quarter of points $i \in \mathbb{N}_m$ satisfy $X_i = e_1$ and $Y_{\xi(i)} = e_2$.

Let T be the random index in $\mathbb{N}_m$ that corresponds to the entry of X that is flipped. By the observations of Example 2.1, i)–iv) lead to the following generalization of (2.5),

$$\mathrm{E}[\Delta \| X = x, Y = y, \xi(T) \neq \lambda(T)] \geq \frac{1}{2}. \tag{2.6}$$

Likewise, v) leads to the following generalization of (2.4),

$$\mathrm{E}[\Delta \| X = x, Y = y, \xi(T) = \lambda(T)] \geq 0. \tag{2.7}$$

Of course, (2.6) and (2.7) hold only approximately. Much of the work of Section 4 is devoted to overcoming these imprecisions. For now, let us work with the simplified assumption that (2.6) and (2.7) hold true except on a set F^c of pairs (X, Y) with negatively exponentially small probability $\mathrm{P}[F^c] = \exp(-\mathcal{O}(n))$. Equations (2.1), (2.6) and (2.7) then imply

$$0 = \mathrm{E}[\Delta] \geq \frac{1}{2} \times \mathrm{P}[\xi(T) \neq \lambda(T)] + 0 \times \mathrm{P}[\xi(T) = \lambda(T)] - 1 \times \mathrm{P}[F^c],$$

so that

$$\mathrm{P}[\xi(T) \neq \lambda(T)] \leq \exp(-\mathcal{O}(n)). \tag{2.8}$$

When the approximate statements (2.6) and (2.7) are replaced with correct inequalities, (2.8) turns into the weaker claim of Theorem 1.1.

The structure of the remaining sections of this paper is as follows. Section 3 serves to introduce the main notation relevant to scoring functions, alignments and local uniqueness of alignments. We also discuss illustrative examples and prove two technical results of preliminary nature. In Section 4 we introduce events defined in terms of certain empirical distributions and their large deviations. These events allow putting the above-made approximate statements i)–v) onto a rigorous footing. In Section 5 we define formally the measure-preserving map which flips a random entry of X to its opposite value. In Lemma 5.1 of that section we prove that the locations where the optimal alignment is nonunique tend to introduce a positive bias into $\mathrm{E}[\Delta]$. Section 6 finally brings all the elements together in the proof of Theorem 1.1.

3. Alignments and Scores. Let $(x, y) \in \{0,1\}^m \times \{0,1\}^n$ be a pair of strings of lengths $m < n$ over the binary alphabet. Let us consider alignments

$$\begin{array}{ccccccccccc} y_1 & \cdots & y_{\xi(1)-1} & y_{\xi(1)} & y_{\xi(1)+1} & \cdots & y_{\xi(m)-1} & y_{\xi(m)} & y_{\xi(m)+1} & \cdots & y_n \\ \sqcup & \ldots & \sqcup & x_1 & \sqcup & \ldots & \sqcup & x_m & \sqcup & \ldots & \sqcup \end{array}$$

of x and y that have δn gaps in x, where $m = \lfloor (1-\delta)n \rfloor$. In the above display we marked gaps with the symbol $\sqcup$. We identify the set of such alignments with the set $\mathcal{A}_{m,n}$ of order-preserving injections of $\mathbb{N}_m$ into $\mathbb{N}_n := \{1, \ldots, n\}$, that is, $\xi \in \mathcal{A}_{m,n}$ if and only if $\xi : \mathbb{N}_m \hookrightarrow \mathbb{N}_n$ and $i < j$ implies $\xi(i) < \xi(j)$.

LEMMA 3.1. *If $\delta < 5/6$, then*

$$\#\mathcal{A}_{m,n} \le \mathrm{e}^{nH(\delta)},$$

where $H(\delta) = -\big(\delta \ln \delta + (1-\delta)\ln(1-\delta)\big)$ is the entropy function.

Proof. Robbins' inequality [16] says that

$$\sqrt{2\pi} n^{-n+\frac{1}{12+1}} \le n! \le \sqrt{2\pi} n^{n+\frac{1}{2}} \mathrm{e}^{-n+\frac{1}{12n}}.$$

Therefore,

$$\begin{aligned} \#\mathcal{A}_{m,n} &= \binom{n}{n(1-\delta)} \\ &\le \frac{\sqrt{2\pi} n^{n+\frac{1}{2}} \mathrm{e}^{-n+\frac{1}{12n}}}{\sqrt{2\pi}\big(n(1-\delta)\big)^{n(1-\delta)+\frac{1}{2}} \mathrm{e}^{-n(1-\delta)+\frac{1}{12n(1-\delta)+1}} \times \sqrt{2\pi}(n\delta)^{n\delta+\frac{1}{2}} \mathrm{e}^{-n\delta+\frac{1}{12n\delta+1}}} \\ &= \mathrm{e}^{nH(\delta)} \times \frac{\exp\left(\frac{1}{12n} - \frac{1}{12n(1-\delta)+1} - \frac{1}{12n\delta+1}\right)}{\sqrt{2\pi(1-\delta)(n-m)}}. \end{aligned}$$

Note that the second factor converges to zero for fixed δ and $n \to \infty$. Moreover, the numerator is < 1 and for $\delta < 5/6$ the denominator is > 1 since $2\pi(1-\delta)(n-m) > 6(1-\delta) > 1$. □

We define a scoring function $\{0,1\}^m \times \{0,1\}^n \times \mathcal{A}_{m,n} \to \mathbb{N}_0$ as follows,

$$S(x,y;\xi) := \sum_{i=1}^{m} s(x_i, y_{\xi(i)}),$$

where $s(0,0) = s(1,1) = 1$ and $s(0,1) = s(1,0) = 0$. The set of optimal alignments of (x,y) is the set of alignments with maximum score,

$$\mathcal{O}A_{x,y} := \{\xi \in \mathcal{A}_{m,n} : \ S(x,y;\xi) \geq S(x,y;\lambda)\ \forall\, \lambda \in \mathcal{A}_{m,n}\}\,.$$

We write $S^*(x,y) := \max\{S(x,y;\xi) : \ \xi \in \mathcal{A}_{m,n}\}$ for the maximum score.

For each $i \in \mathbb{N}_m$ we define the variable

$$u_i(x,y) := \begin{cases} 1 & \text{if } \exists\, \xi, \lambda \in \mathcal{O}A_{x,y} \text{ s.t. } \xi(i) \neq \lambda(i), \\ 0 & \text{otherwise} \end{cases}$$

that indicates when the image of i under the set of optimal alignments is nonunique. We say that the optimal alignment is *locally nonunique* at i if $u_i(x,y) = 1$. We write

$$u(x,y) := \sum_{i=1}^{m} u_i(x,y)$$

for the number of indices where the optimal alignment is locally nonunique.

The sets $\mathcal{O}A_{x,y}$ and $\{i \in \mathbb{N}_m : \ u_i(x,y) = 1\}$ can be found via dynamic programming: A $m \times n$ matrix $\big(score(i,j)\big)$ is recursively computed, using the rules

r.i) $score(i,j) = -1$ for $i > j$ or $j > i + \delta n$,

r.ii) $score(1,j) = s(x_1, y_j)$ for $j = 1, \ldots, 1 + \delta n$,

r.iii) $score(i,j) = s(x_i, y_j) + \max\{score(i-1,k) : \ k < j\}$ for all other (i,j).

Arguing recursively, one immediately verifies that

$$S^*(x,y) = \max\{score(m,j) : \ j \in \mathbb{N}_n\},$$

and furthermore that $\xi \in \mathcal{O}A_{x,y}$ if and only if the following conditions are satisfied:

c.i) $\xi(m) \in \big\{j \in \mathbb{N}_n : \ score(m,j) = S^*(x,y)\big\}$,

c.ii) $\xi(i-1) \in \big\{j < \xi(i) : \ score(i-1,j) = \max_{k<\xi(i)} score(i-1,k)\big\}$ for all $i = 2, \ldots, m$.

EXAMPLE 3.2. *Let* $x = 01010101$ *and* $y = 110101101100$. *Then the above described dynamic programming algorithm generates the matrix of Figure 3.1, where it is displayed in tableau format. Optimal paths follow the arrows and pass through the shaded entries. The tableau is annotated with the generating sequences* x *and* y*, so that the optimal alignments can easily be read off. The following table lists* y *in the top row, followed by a complete list of optimal alignments of* x *with* y*,*

1	**1**	**0**	**1**	**0**	**1**	**1**	**0**	**1**	**1**	**0**	**0**
0	1	0	1	0	1	⊔	0	1	⊔	⊔	⊔
0	1	0	1	0	1	⊔	0	⊔	1	⊔	⊔
0	1	0	1	0	⊔	1	0	1	⊔	⊔	⊔
0	1	0	1	0	⊔	1	0	⊔	1	⊔	⊔
⊔	⊔	0	1	0	1	⊔	0	1	⊔	0	1
⊔	⊔	0	1	0	1	⊔	0	⊔	1	0	1
⊔	⊔	0	1	0	⊔	1	0	1	⊔	0	1
⊔	⊔	0	1	0	⊔	1	0	⊔	1	0	1

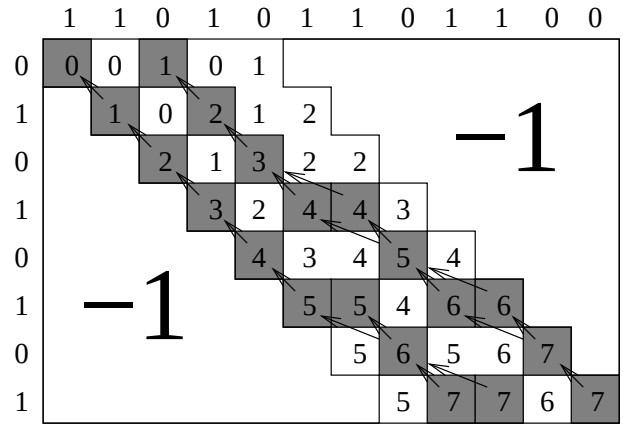

FIG. 3.1. *The scoring matrix of Example 3.2.*

Every line of the the tableau in Figure 3.1 contains multiple shaded entries. Therefore, $u_i(x,y) = 1$ *for all* $i \in \mathbb{N}_m$ *and* $u(x,y) = m$.

In the above example we ordered the optimal alignments from leftmost to rightmost as located within the tableau, that is, alignments are listed in inverse alphabetical order with respect to the lateness of gaps. This also provides the idea of proof for the following result, which shows that there exist two optimal alignments that differ from one another at every point where the optimal alignment is locally nonunique.

LEMMA 3.3. *For all* $(x,y) \in \{0,1\}^m \times \{0,1\}^n$ *there exist* $\xi, \lambda \in \mathcal{OA}_{x,y}$ *such that*

$$\xi(i) \neq \lambda(i) \Leftrightarrow u_i(x,y) = 1.$$

Proof. The claim is clearly true if we can prove that $\xi, \lambda \in \mathcal{OA}_{x,y}$, where

$$\begin{aligned} \xi(i) &:= \min\{\psi(i) : \psi \in \mathcal{OA}_{x,y}\} \quad \forall\, i \in \mathbb{N}_m, \\ \lambda(i) &:= \max\{\psi(i) : \psi \in \mathcal{OA}_{x,y}\} \quad \forall\, i \in \mathbb{N}_m. \end{aligned}$$

If $\lambda \notin \mathcal{OA}_{x,y}$ then there exists an index $i \in \mathbb{N}_m \setminus \{1\}$ such that $\widehat{\psi}(i-1) < \lambda(i-1)$ for all $\widehat{\psi} \in \mathcal{OA}_{x,y}$ such that $\widehat{\psi}(i) = \lambda(i)$. On the other hand, there exists $\widehat{\lambda} \in \mathcal{OA}_{x,y}$ such that $\widehat{\lambda}(i-1) = \lambda(i-1)$, and therefore it is necessarily the case that $\widehat{\lambda}(i) < \widehat{\psi}(i) = \lambda(i)$. But $\widehat{\lambda}$ and $\widehat{\psi}$ satisfy condition c.ii), that is,

$$\begin{aligned} \widehat{\lambda}(i-1) &\in \Big\{ j < \widehat{\lambda}(i) : \mathit{score}(i-1,j) = \max_{k<\widehat{\lambda}(i)} \mathit{score}(i-1,k) \Big\}, \\ \widehat{\psi}(i-1) &\in \Big\{ j < \widehat{\psi}(i) : \mathit{score}(i-1,j) = \max_{k<\widehat{\psi}(i)} \mathit{score}(i-1,k) \Big\}. \end{aligned}$$

Therefore, either $\max_{k<\widehat{\lambda}(i)} \mathit{score}(i-1,k) = \max_{k<\widehat{\psi}(i)} \mathit{score}(i-1,k)$ and then there exists $\psi \in \mathcal{OA}_{x,y}$ such that $\psi(i) = \widehat{\psi}(i) = \lambda(i)$ and $\psi(i-1) = \widehat{\lambda}(i-1) = \lambda(i-1)$, or else $\max_{k<\widehat{\lambda}(i)} \mathit{score}(i-1,k) < \max_{k<\widehat{\psi}(i)} \mathit{score}(i-1,k)$ and then $\widehat{\psi}(i-1) > \widehat{\lambda}(i-1) = \lambda(i-1)$. In either case we have a contradiction, and this shows that $\lambda \in \mathcal{OA}_{x,y}$ indeed.

The proof that $\xi \in \mathcal{O}A_{x,y}$ is analogous. □

Note that the alignments ξ and λ constructed in the proof of Lemma 3.3 are uniquely determined by (x, y). Furthermore, they satisfy the relation $\xi \leq \lambda$ which is defined by $\xi(i) \leq \lambda(i)$ for all $i \in \mathbb{N}_m$.

4. Large Deviations of Some Empirical Distributions. In this section we establish a rigorous version of the approximate inequalities (2.6),(2.7) and statements i)–v) of Section 2. Recall that $X = X_1 \dots X_m$ and $Y = Y_1 \dots Y_n$ are two independent random strings that consist of i.i.d. standard Bernoulli variables $X_i, Y_j \sim \mathscr{B}(1/2)$ defined on some probability space $(\Omega, \mathscr{F}, \mathrm{P})$. We write U_i and U for the random variables $u_i(X, Y)$ and $u(X, Y)$ respectively. Furthermore, we think of $\delta \in (0, 1)$ as a fixed gap proportion that relates m to n via $m = \lfloor (1-\delta)n \rfloor$.

For $\xi \in \mathcal{A}_{m,n}$ fixed and $\omega \in \Omega$ let $\widehat{\mathscr{D}}_\xi(\omega)$ be the empirical distribution of $(X_i(\omega), Y_{\xi(i)}(\omega))$ over $i \in \mathbb{N}_m$, i.e., the distribution of $(X_T(\omega), Y_{\xi(T)}(\omega))$ when $T \sim \mathscr{U}(\mathbb{N}_m)$ is a random index with uniform distribution on $\mathbb{N}_m$. Yet another way to define this distribution is to require that for all $(e_1, e_2) \in \{0,1\}^2$,

$$\mathrm{P}_{\widehat{\mathscr{D}}_\xi(\omega)}\big[(e_1, e_2)\big] = \frac{1}{m} \times \#\Big\{ i \in \mathbb{N}_m \,:\, X_i(\omega) = e_1,\, Y_{\xi(i)}(\omega) = e_2 \Big\}.$$

EXAMPLE 4.1. *Let x, y and ξ be chosen as in Example 2.1. If $\omega \in \Omega$ is chosen such that $X(\omega) = x$ and $Y(\omega) = y$ then*

$$\mathrm{P}_{\widehat{\mathscr{D}}_\xi(\omega)}\big[(0,0)\big] = \frac{1}{6},\ \mathrm{P}_{\widehat{\mathscr{D}}_\xi(\omega)}\big[(0,1)\big] = \frac{2}{6},\ \mathrm{P}_{\widehat{\mathscr{D}}_\xi(\omega)}\big[(1,0)\big] = \frac{1}{6},\ \mathrm{P}_{\widehat{\mathscr{D}}_\xi(\omega)}\big[(1,1)\big] = \frac{2}{6}.$$

Note that $\widehat{\mathscr{D}}_\xi$ only depends on x and y.

Let E_ξ be the event that

$$\max_{(e_1,e_2)\in\{0,1\}^2} \Big| \mathrm{P}_{\widehat{\mathscr{D}}_\xi(\omega)}\big[(e_1, e_2)\big] - 1/4 \Big| < \sqrt{\frac{9H(\delta)}{4(1-\delta)}} =: \epsilon_1(\delta), \tag{4.1}$$

in other words,

$$E_\xi := \Big\{ \omega \in \Omega : \Big\| \widehat{\mathscr{D}}_\xi(\omega) - \mathscr{B}(1/2) \otimes \mathscr{B}(1/2) \Big\| < \epsilon_1(\delta) \Big\}.$$

Let us furthermore define the event

$$E_{m,n} := \bigcap_{\xi \in \mathcal{A}_{m,n}} E_\xi.$$

In a similar vein we define empirical distributions and events relating to $(X_i, Y_{\xi(i)}, Y_{\lambda(i)})$ as follows: Let $\varepsilon > 0$ and $\xi, \lambda \in \mathcal{A}_{m,n}$ be fixed such that $\xi \leq \lambda$ and $d(\xi, \lambda) := \#\{i : \xi(i) \neq \lambda(i)\} \geq m\varepsilon$. For all i such that $\xi(i) = \lambda(i)$ we define the random variables

$$R_i^{\xi,\lambda} := \begin{cases} 1 & \text{if } X_i \neq Y_{\xi(i)} \\ -1 & \text{if } X_i = Y_{\xi(i)}. \end{cases}$$

Likewise, for all i such that $\xi(i) \neq \lambda(i)$ we define the random variables

$$R_i^{\xi,\lambda} := \begin{cases} 0 & \text{if } Y_{\xi(i)} \neq Y_{\lambda(i)}, \\ 1 & \text{if } X_i \neq Y_{\xi(i)} = Y_{\lambda(i)}, \\ -1 & \text{if } X_i = Y_{\xi(i)} = Y_{\lambda(i)}. \end{cases}$$

Let us now consider the empirical distributions

$$\begin{aligned} \widehat{\mathscr{L}}_{\xi,\lambda}^{agree}(\omega) &:= \mathscr{D}(R_T^{\xi,\lambda}(\omega)), \quad \text{where } T \sim \mathscr{U}(\{i \in \mathbb{N}_m : \xi(i) = \lambda(i)\}), \\ \widehat{\mathscr{L}}_{\xi,\lambda}^{disag}(\omega) &:= \mathscr{D}(R_T^{\xi,\lambda}(\omega)), \quad \text{where } T \sim \mathscr{U}(\{i \in \mathbb{N}_m : \xi(i) \neq \lambda(i)\}), \\ \widehat{\mathscr{L}}_{\xi,\lambda}^{unif}(\omega) &:= \mathscr{D}(R_T^{\xi,\lambda}(\omega)), \quad \text{where } T \sim \mathscr{U}(\mathbb{N}_m). \end{aligned}$$

Let $\mathscr{J}^{agree}$ be the distribution on $\{-1,1\}$ defined by $\mathrm{P}_{\mathscr{J}^{agree}}[-1] = 1/2 = \mathrm{P}_{\mathscr{J}^{agree}}[1]$, and let $\mathscr{J}^{disag} = \mathscr{J}^{unif}$ be the distribution on $\{-1,0,1\}$ defined by $\mathrm{P}_{\mathscr{J}^{disag}}[-1] = 1/4 = \mathrm{P}_{\mathscr{J}^{disag}}[1]$ and $\mathrm{P}_{\mathscr{J}^{disag}}[0] = 1/2$. Let finally

$$\epsilon_2(\delta,\varepsilon) := \sqrt{\frac{3H(\delta)}{2(1-\delta)\varepsilon}}, \quad \epsilon_3(\delta,\varepsilon) := \sqrt{\frac{27H(\delta)}{8(1-\delta)\varepsilon}}, \quad \epsilon_4(\delta,\varepsilon) := \sqrt{\frac{3H(\delta)}{2(1-\delta)(1-\varepsilon)}}.$$

These notions define the following events,

$$\begin{aligned} F_{\xi,\lambda}^{\varepsilon} &:= \left\{\omega \in \Omega : \left\| \widehat{\mathscr{L}}_{\xi,\lambda}^{agree} - \mathscr{J}^{agree} \right\| < \epsilon_2(\delta,\varepsilon)\right\}, \\ G_{\xi,\lambda}^{\varepsilon} &:= \left\{\omega \in \Omega : \left\| \widehat{\mathscr{L}}_{\xi,\lambda}^{disag} - \mathscr{J}^{disag} \right\| < \epsilon_3(\delta,\varepsilon)\right\}, \\ H_{\xi,\lambda}^{\varepsilon} &:= \left\{\omega \in \Omega : \left\| \widehat{\mathscr{L}}_{\xi,\lambda}^{unif} - \mathscr{J}^{unif} \right\| < \epsilon_4(\delta,\varepsilon) + 2\varepsilon\right\}, \\ F_{m,n,\varepsilon} &:= \bigcap_{\{(\xi,\lambda):\xi\leq\lambda, m\varepsilon\leq d(\xi,\lambda)\leq m(1-\varepsilon)\}} \left(F_{\xi,\lambda}^{\varepsilon} \cap G_{\xi,\lambda}^{\varepsilon}\right) \quad \cap \bigcap_{\{(\xi,\lambda):\xi\leq\lambda, m(1-\varepsilon)<d(\xi,\lambda)\}} H_{\xi,\lambda}^{\varepsilon}. \end{aligned}$$

The following estimates play an important role in our later analysis:

LEMMA 4.2. *For all $\delta < 5/6$ and $\varepsilon > 0$,*
i) $\mathrm{P}\big[E_{m,n}^c\big] \leq 8\,\mathrm{e}^{-nH(\delta)}$,
ii) $\mathrm{P}\big[F_{m,n,\varepsilon}^c\big] \leq 10\,\mathrm{e}^{-nH(\delta)}$.

Proof. The Azuma-Hoeffding Theorem [4, 12] says that if $(V_0,\ldots,V_m)$ is a martingale with $V_0 \equiv 0$ and $\mathrm{P}[|V_k - V_{k-1}| \leq a] = 1$ for all $k \in \mathbb{N}_m$ then

$$\mathrm{P}[V_m \geq mb] \leq \exp\left(-\frac{mb^2}{2a^2}\right)$$

for all $b > 0$. For $\xi \in \mathcal{A}_{m,n}$ and $(e_1,e_2) \in \{0,1\}^2$ fixed let

$$Z_i(\omega) := \begin{cases} 1 & \text{if } \big(X_i(\omega), Y_{\xi(i)}(\omega)\big) = (e_1,e_2), \\ 0 & \text{otherwise.} \end{cases}$$

Then Z_i ($i \in \mathbb{N}_m$ are i.i.d. random variables with expectation $\mathrm{E}[Z_i] = 1/4$. If we set $V_0 := 0$ and

$$V_k := \sum_{i=1}^{k} \big(Z_i - \mathrm{E}[Z_i]\big) \qquad (k \in \mathbb{N}_m),$$

then $(V_0, \ldots, V_m)$ is a martingale with $|V_k - V_{k-1}| \leq 3/4$ for all k. By the Azuma-Hoeffding Theorem, we have

$$\mathrm{P}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i - \frac{1}{4} \geq \epsilon_1(\delta)\right] = \mathrm{P}\left[\frac{1}{m}V_m \geq \epsilon_1(\delta)\right] \leq \exp\left(-\frac{8m\epsilon_1^2(\delta)}{9}\right) \leq \mathrm{e}^{-2nH(\delta)}\,.$$

Applying the same reasoning to the martingale $(-V_0, \ldots, -V_m)$, we find

$$\mathrm{P}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i - \frac{1}{4} \leq -\epsilon_1(\delta)\right] \leq \mathrm{e}^{-2nH(\delta)},$$

so that

$$\mathrm{P}\left[\left|\mathrm{P}_{\widehat{\mathscr{D}}_\xi}\left[(e_1, e_2)\right] - \frac{1}{4}\right| \geq \epsilon_1(\delta)\right] \leq 2\,\mathrm{e}^{-2nH(\delta)}\,.$$

Since this is true for all $(e_1, e_2) \in \{0,1\}^2$, simple union bounds show that

$$\mathrm{P}\left[E_\xi^c\right] \leq 8\,\mathrm{e}^{-2nH(\delta)}$$

and

$$\mathrm{P}\left[E_{m,n}^c\right] = \mathrm{P}\left[\bigcup_{\xi \in \mathcal{A}_{m,n}} E_\xi^c\right] \leq \#\mathcal{A}_{m,n} \times 8\,\mathrm{e}^{-2nH(\delta)} \overset{Lem3.1}{\leq} 8\,\mathrm{e}^{-nH(\delta)}\,.$$

ii) The proof of the second part is similar: Let (ξ, λ) be such that $\xi \leq \lambda$ and $d(\xi, \lambda) \geq m\varepsilon$. For all i such that $\xi(i) \neq \lambda(i)$ we have $\xi(i) < \lambda(i)$. For $e \in \{-1, 0, 1\}$ fixed let

$$Z_i := \begin{cases} 1 & \text{if } R_i^{\xi,\lambda} = e, \\ 0 & \text{otherwise} \end{cases} \qquad (i \in \{k : \xi(k) \neq \lambda(k)\}.$$

Then we have

$$\mathrm{E}[Z_i] = \begin{cases} \frac{1}{4} & \text{if } e \in \{-1, 1\}, \\ \frac{1}{2} & \text{if } e = 0., \end{cases}$$

so that $|Z_i - \mathrm{E}[Z_i]| \leq 3/4$ in all three cases. Furthermore, the random variables

$$(Z_i - E[Z_i]), \qquad (i \in \{k : \xi(k) \neq \lambda(k)\})$$

are i.i.d. with distribution $\mathscr{J}^{disag}$. This is seen by induction, using the observation that for all index sets

$$I \subset \{k : \xi(k) \neq \lambda(k)\},$$

if $i_{\max} = \max I$ then $X_{i_{\max}}$ and $Y_{\lambda(i_{\max})}$ do not appear in any of the expressions $(X_i, Y_{\xi(i)}, Y_{\lambda(i)})$ $(i \in I \setminus \{i_{\max}\})$, so that independently of the value of $Y_{\xi(i_{\max})}$ (which could have appeared in the above expressions at most once as $Y_{\lambda(i)}$), we have

$$\mathrm{P}\left[Y_{\lambda(i_{\max})} \neq Y_{\xi(i_{\max})}\right] = \frac{1}{2}, \quad \mathrm{P}\left[X_{i_{\max}} \neq Y_{\xi(i_{\max})} = Y_{\lambda(i_{\max})}\right] = \frac{1}{4},$$
$$\text{and} \quad \mathrm{P}\left[X_{i_{\max}} = Y_{\xi(i_{\max})} = Y_{\lambda(i_{\max})}\right] = \frac{1}{4}.$$

We define $V_0 \equiv 0$ and for $k \in \mathbb{N}_{d(\xi,\lambda)}$, $V_k := V_{k-1} + Z_i - \mathrm{E}[Z_i]$, where

$$i = \min\{l \in \mathbb{N}_m : \#\{j \le l : \xi(j) \neq \lambda(j)\} = k\}.$$

Then $(V_0, \ldots, V_{d(\xi,\lambda)})$ is a martingale, and arguing as above by ways of the Azuma-Hoeffding Theorem, we find

$$\begin{aligned}
&\mathrm{P}\left[\left|\mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\xi,\lambda}}[e] - \mathrm{P}_{\mathscr{J}^{disag}}[e]\right| \ge \epsilon_3(\delta,\varepsilon)\right] \\
&\qquad = \mathrm{P}\left[\left|\frac{1}{d(\xi,\eta)} V_{d(\xi,\eta)}\right| \ge \epsilon_3(\delta,\varepsilon)\right] \le 2\exp\left(-\frac{8d(\xi,\lambda)\epsilon_3^2(\delta,\varepsilon)}{9}\right) \le 2\,\mathrm{e}^{-3nH(\delta)}.
\end{aligned}$$

Since this holds for all $e \in \{-1, 0, 1\}$, we have

$$\mathrm{P}\left[G^c_{\xi,\lambda}\right] \le 6\,\mathrm{e}^{-3nH(\delta)} \tag{4.2}$$

whenever $d(\xi,\lambda) \ge m\varepsilon$ and $\xi \le \lambda$.

If it is even the case that $d(\xi,\lambda) > m(1-\varepsilon)$, then we find

$$\mathrm{P}\left[\left|\mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\xi,\lambda}}[e] - \mathrm{P}_{\mathscr{J}^{disag}}[e]\right| \ge \epsilon_4(\delta,\varepsilon)\right] \le 2\exp\left(-\frac{8d(\xi,\lambda)\epsilon_4^2(\delta,\varepsilon)}{9}\right) \le 2\,\mathrm{e}^{-3nH(\delta)},$$

and also

$$\left|\mathrm{P}_{\widehat{\mathscr{L}}^{unif}_{\xi,\lambda}}[e] - \mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\xi,\lambda}}[e]\right| \le 2\varepsilon.$$

Therefore,

$$\begin{aligned}
&\mathrm{P}\left[\left|\mathrm{P}_{\widehat{\mathscr{L}}^{unif}_{\xi,\lambda}}[e] - \mathrm{P}_{\mathscr{J}^{unif}}[e]\right| \ge \epsilon_4(\delta,\varepsilon) + 2\varepsilon\right] \\
&\qquad\qquad \le \mathrm{P}\left[\left|\mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\xi,\lambda}}[e] - \mathrm{P}_{\mathscr{J}^{disag}}[e]\right| \ge \epsilon_4(\delta,\varepsilon)\right] \le 2\,\mathrm{e}^{-3nH(\delta)}.
\end{aligned}$$

Since this holds for all $e \in \{-1, 0, 1\}$, we have

$$\mathrm{P}\left[H^c_{\xi,\lambda}\right] \le 6\,\mathrm{e}^{-3nH(\delta)} \tag{4.3}$$

whenever $d(\xi,\lambda) > m(1-\varepsilon)$ and $\xi \le \lambda$.

Next, let $\xi \le \lambda$ be such that $d(\xi,\lambda) \le m(1-\varepsilon)$, and for $e \in \{-1, 1\}$ fixed let

$$Z_i := \begin{cases} 1 & \text{if } R_i^{\xi,\lambda} = e, \\ 0 & \text{otherwise} \end{cases} \qquad (i \in \{k : \xi(k) = \lambda(k)\}).$$

Then $\mathrm{E}[Z_i] = 1/2$ so that $|Z_i - \mathrm{E}[Z_i]| = 1/2$, and

$$(Z_i - E[Z_i]), \qquad (i \in \{k : \xi(k) = \lambda(k)\})$$

are i.i.d. random variables with distribution $\mathscr{J}^{agree}$. Let $V_0 := 0$, and for $k \in \mathbb{N}_{m-d(\xi,\lambda)}$, $V_k := V_{k-1} + Z_i - \mathrm{E}[Z_i]$, where

$$i = \min\{l \in \mathbb{N} : \#\{j \le l : \xi(j) = \lambda(j)\} = k\}.$$

Then $(V_0, \ldots, V_{m-d(\xi,\lambda)})$ is a martingale and the large deviations argument from above shows that

$$\begin{aligned}
&\mathrm{P}\left[\left|\mathrm{P}_{\widehat{\mathscr{L}}^{agree}_{\xi,\lambda}}[e] - \mathrm{P}_{\mathscr{J}^{agree}}[e]\right| \geq \epsilon_2(\delta,\varepsilon)\right] \\
&\qquad = \mathrm{P}\left[\left|\frac{1}{m-d(\xi,\lambda)} V_{m-d(\xi,\lambda)}\right| \geq \epsilon_2(\delta,\varepsilon)\right] \leq 2\,\mathrm{e}^{-2(m-d(\xi,\lambda))\epsilon_2^2} \leq 2\,\mathrm{e}^{-3nH(\delta)}\,.
\end{aligned}$$

Since this holds for both $e \in \{1,-1\}$, we find

$$\mathrm{P}\left[F^c_{\xi,\lambda}\right] \leq 4\,\mathrm{e}^{-3nH(\delta)} \tag{4.4}$$

whenever $d(\xi,\lambda) \leq m(1-\varepsilon)$ and $\xi \leq \lambda$.

Finally, the combination of equations (4.2), (4.3), (4.4) and Lemma 3.1 shows that

$$\begin{aligned}
\mathrm{P}\left[F^c_{m,n}\right] &\leq \sum_{\{(\xi,\lambda):\xi\leq\lambda, m\varepsilon\leq d(\xi,\lambda)\leq m(1-\varepsilon)\}} \left(\mathrm{P}\left[F^c_{\xi,\lambda}\right] + \mathrm{P}\left[G^c_{\xi,\lambda}\right]\right) + \sum_{\{(\xi,\lambda):\xi\leq\lambda, m(1-\varepsilon)<d(\xi,\lambda)\}} \mathrm{P}\left[H^c_{\xi,\lambda}\right] \\
&\leq (\mathcal{A}_{m,n})^2 \times \left(10\,\mathrm{e}^{-3nH(\delta)}\right) \\
&\leq 10\,\mathrm{e}^{-nH(\delta)}\,.
\end{aligned}$$

□

5. An Ergodic Map. Let us now introduce an ergodic map as follows: Let $T \sim \mathscr{U}(\mathbb{N}_m)$ be a uniform random variable on $\mathbb{N}_m$. By Kolmogorov's theorem we may assume without loss of generality that $(\Omega, \mathscr{F}, \mathrm{P})$ is extended so that T is defined on Ω and independent of the X_i and Y_j. Let us define new random strings $\tilde{X} = \tilde{X}_1 \ldots \tilde{X}_m$ and $\tilde{Y} = \tilde{Y}_1 \ldots \tilde{Y}_n$ by setting $\tilde{Y} := Y$, $\tilde{X}_i := X_i$ for all $i \neq T$ and

$$\tilde{X}_T := X_T + 1 \mod 2.$$

In other words, $(\tilde{X}, \tilde{Y})$ is obtained from (X, Y) by flipping one random bit of X and keeping all other entries of X and Y unchanged. The map $(X,Y) \mapsto (\tilde{X}, \tilde{Y})$ is measure-preserving, since $\tilde{X}$ and $\tilde{Y}$ again consist of i.i.d. standard Bernoulli variables. Therefore,

$$\mathrm{E}[\Delta] = 0 \tag{5.1}$$

where $\Delta := S^*(\tilde{X}, \tilde{Y}) - S^*(X, Y)$. The construction in the proof of Lemma 3.3 shows that there exists a $\sigma(X,Y)$-measurable map

$$\begin{aligned}
(\Xi, \Lambda) : \Omega &\to \mathcal{A}_{m,n} \times \mathcal{A}_{m,n}, \\
\omega &\mapsto \left(\Xi_\omega, \Lambda_\omega\right)
\end{aligned}$$

such that for all $\omega \in \Omega$, $\Xi_\omega \leq \Lambda_\omega$ and

$$\{i : \Xi_\omega(i) \neq \Lambda_\omega(i)\} = \{i : U_i(\omega) = 1\}\,.$$

Furthermore, X and Y define the $\sigma(X,Y)$-measurable events $E_{m,n}$, $F_{m,n}$ and the $\sigma(X,Y)$-measurable random variable U introduced in Section 4. The following two

lemmas show how these objects affect Δ and will be the key tools in the proof of the main theorem of this paper.

LEMMA 5.1. *For all $\delta < 5/6$ and $\varepsilon > 0$,*

$$\mathrm{E}\left[\Delta \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \geq \left[\frac{1}{2} - 3\max\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon, \epsilon_3(\delta,\varepsilon)\right)\right] \times \varepsilon + \min\left[\frac{1}{2} - 3\left(\epsilon_4(\delta,\varepsilon) + \varepsilon\right), -2\epsilon_2(\delta,\varepsilon)\right] \times (1-\varepsilon).$$

Proof. A key observation is that $Y_{\Xi(T)} \neq Y_{\Lambda(T)}$ implies $\Delta = 1$: Without loss of generality we may assume that $\tilde{X}_T = Y_{\Xi(T)}$, so that $X_T \neq Y_{\Xi(T)}$ and $S^*(X,Y) = \sum_{i \neq T} s(X_i, Y_{\Xi(i)})$. But then we have

$$S^*(X,Y) + 1 \geq S^*(\tilde{X},\tilde{Y}) \geq S(\tilde{X},\tilde{Y};\Xi) = \sum_{i \neq T} s(X_i, Y_{\Xi(i)}) + s(\tilde{X}_T, Y_{\Xi(T)}) = S^*(X,Y) + 1,$$

so that $\Delta = 1$ indeed. Likewise, $X_T \neq Y_{\Xi(T)} = Y_{\Lambda(T)}$ implies $\Delta = 1$. Using these facts and the trivial lower bound $\Delta \geq -1$, we have

$$\begin{aligned}
&\mathrm{E}\left[\Delta \,\middle\|\, U \geq m(1-\varepsilon), F_{m,n}\right] \\
&\quad \geq \mathrm{E}\left[1 \times \mathrm{P}_{\widehat{\mathscr{L}}^{unif}_{\Xi,\Lambda}}[0] + 1 \times \mathrm{P}_{\widehat{\mathscr{L}}^{unif}_{\Xi,\Lambda}}[1] - 1 \times \mathrm{P}_{\widehat{\mathscr{L}}^{unif}_{\Xi,\Lambda}}[-1] \,\middle\|\, U \geq m(1-\varepsilon), F_{m,n}\right] \\
&\quad \geq 1 \times \left(\mathrm{P}_{\mathscr{J}^{unif}}[0] - \epsilon_4(\delta,\varepsilon) - 2\varepsilon\right) + 1 \times \left(\mathrm{P}_{\mathscr{J}^{unif}}[1] - \epsilon_4(\delta,\varepsilon) - 2\varepsilon\right) \\
&\qquad\qquad - 1 \times \left(\mathrm{P}_{\mathscr{J}^{unif}}[-1] + \epsilon_4(\delta,\varepsilon) + 2\varepsilon\right) \\
&\quad = \frac{1}{2} - 3\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon\right), \qquad (5.2)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{E}\left[\Delta \,\middle\|\, m(1-\varepsilon) \geq U \geq m\varepsilon, F_{m,n}, \Xi(T) \neq \Lambda(T)\right] \\
&\quad \geq \mathrm{E}\Big[1 \times \mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\Xi,\Lambda}}[0] + 1 \times \mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\Xi,\Lambda}}[1] - 1 \times \mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\Xi,\Lambda}}[-1] \,\Big\|\, m(1-\varepsilon) \geq U \geq m\varepsilon, \\
&\qquad\qquad F_{m,n}, \Xi(T) \neq \Lambda(T)\Big] \\
&\quad \geq 1 \times \left(\mathrm{P}_{\mathscr{J}^{disag}}[0] - \epsilon_3(\delta,\varepsilon)\right) + 1 \times \left(\mathrm{P}_{\mathscr{J}^{disag}}[1] - \epsilon_3(\delta,\varepsilon)\right) \\
&\qquad\qquad - 1 \times \left(\mathrm{P}_{\mathscr{J}^{disag}}[-1] + \epsilon_3(\delta,\varepsilon)\right) \\
&\quad = \frac{1}{2} - 3\epsilon_3(\delta,\varepsilon), \qquad (5.3)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{E}\left[\Delta \,\middle\|\, m(1-\varepsilon) \geq U \geq m\varepsilon, F_{m,n}, \Xi(T) = \Lambda(T)\right] \\
&\quad \geq \mathrm{E}\Big[1 \times \mathrm{P}_{\widehat{\mathscr{L}}^{agree}_{\Xi,\Lambda}}[1] - 1 \times \mathrm{P}_{\widehat{\mathscr{L}}^{agree}_{\Xi,\Lambda}}[-1] \,\Big\|\, m(1-\varepsilon) \geq U \geq m\varepsilon, F_{m,n}, \Xi(T) = \Lambda(T)\Big] \\
&\quad \geq 1 \times \left(\mathrm{P}_{\mathscr{J}^{agree}}[1] - \epsilon_2(\delta,\varepsilon)\right) - 1 \times \left(\mathrm{P}_{\mathscr{J}^{agree}}[-1] + \epsilon_2(\delta,\varepsilon)\right) \\
&\quad = -2\epsilon_2(\delta,\varepsilon), \qquad (5.4)
\end{aligned}$$

Putting the pieces together, we find

$$
\begin{aligned}
&\mathrm{E}\left[\Delta \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \\
&\quad \geq \mathrm{E}\left[\Delta \,\middle\|\, U \geq m(1-\varepsilon), F_{m,n}\right] \times \Big(\mathrm{P}\left[U \geq m(1-\varepsilon), \Xi(T) \neq \Lambda(T) \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \\
&\qquad\qquad + \mathrm{P}\left[U \geq m(1-\varepsilon), \Xi(T) = \Lambda(T) \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right]\Big) \\
&\quad + \mathrm{E}\left[\Delta \,\middle\|\, m(1-\varepsilon) \geq U \geq m\varepsilon, F_{m,n}, \Xi(T) \neq \Lambda(T)\right] \\
&\qquad\qquad \times \mathrm{P}\left[m(1-\varepsilon) \geq U \geq m\varepsilon, \Xi(T) \neq \Lambda(T) \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \\
&\quad + \mathrm{E}\left[\Delta \,\middle\|\, m(1-\varepsilon) \geq U \geq m\varepsilon, F_{m,n}, \Xi(T) = \Lambda(T)\right] \\
&\qquad\qquad \times \mathrm{P}\left[m(1-\varepsilon) \geq U \geq m\varepsilon, \Xi(T) = \Lambda(T) \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \\
&\overset{(5.2),(5.3),(5.4)}{\geq} \left[\frac{1}{2} - 3\max\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon, \epsilon_3(\delta,\varepsilon)\right)\right] \times \mathrm{P}\left[\Xi(T) \neq \Lambda(T) \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \\
&\qquad + \min\left[\frac{1}{2} - 3\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon\right), -2\epsilon_2(\delta,\varepsilon)\right] \times \mathrm{P}\left[\Xi(T) = \Lambda(T) \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \\
&\quad \geq \left[\frac{1}{2} - 3\max\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon, \epsilon_3(\delta,\varepsilon)\right)\right] \times \varepsilon + \min\left[\frac{1}{2} - 3\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon\right), -2\epsilon_2(\delta,\varepsilon)\right] \times (1-\varepsilon).
\end{aligned}
$$

□

LEMMA 5.2. *For any $\sigma(X,Y)$-measurable event B and all $\delta < 5/6$, we have*

$$\int_B \Delta \, d\,\mathrm{P} \geq -4\epsilon_1(\delta) - 8\,\mathrm{e}^{-nH(\delta)}.$$

Proof.

$$\int_B \Delta \, d\,\mathrm{P} \geq \int_{B\cap E_{m,n}} \Delta \, d\,\mathrm{P} - \mathrm{P}[E_{m,n}^c] \overset{Lem4.2}{\geq} \int_{B\cap E_{m,n}} \Delta \, d\,\mathrm{P} - 8\,\mathrm{e}^{-nH(\delta)}. \tag{5.5}$$

Clearly, for all $\omega \in \Omega$,

$$\Delta(\omega) \geq S\big(\tilde{X}(\omega), \tilde{Y}(\omega); \Xi_\omega\big) - S\big(X(\omega), Y(\omega); \Xi_\omega\big).$$

Therefore,

$$
\begin{aligned}
\int_{B\cap E_{m,n}} \Delta \, d\,\mathrm{P} &\geq \int_{B\cap E_{m,n}} \left[S\left(\tilde{X}, \tilde{Y}; \Xi\right) - S\left(X, Y; \Xi\right)\right] d\,\mathrm{P} \\
&= \int_{B\cap E_{m,n}} \left[s\left(\tilde{X}_T, Y_{\Xi(T)}\right) - s\left(X_T, Y_{\Xi(T)}\right)\right] d\,\mathrm{P}\left[(X(\omega), Y(\omega), T(\omega))\right] \\
&= \int_{B\cap E_{m,n}} \mathrm{E}\left[s\left(\tilde{X}_T, Y_{\Xi(T)}\right) - s\left(X_T, Y_{\Xi(T)}\right) \,\middle\|\, (X,Y)\right] d\,\mathrm{P}\left[(X(\omega), Y(\omega))\right] \\
&\overset{(4.1)}{\geq} \int_{B\cap E_{m,n}} -1 \times 4\epsilon_1(\delta) \, d\,\mathrm{P}\left[(X(\omega), Y(\omega))\right] \geq -4\epsilon_1(\delta).
\end{aligned}
$$

Together with (5.5) this implies the claim. □

6. Proof of The Main Theorem. After introducing the tools of Sections 4 and 5, the stage is set for a proof of Theorem 1.1.

Proof. Since $\{\omega : U < m\varepsilon\} \cup F_{m,n}^c$ is $\sigma(X,Y)$-measurable, we have

$$\begin{aligned}
0 = \mathrm{E}[\Delta] &= \mathrm{E}\left[\Delta \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \times \mathrm{P}\left[U \geq m\varepsilon, F_{m,n}\right] + \int_{\{U<m\varepsilon\}\cup F_{m,n}^c} \Delta d\,\mathrm{P} \overset{Lem5.1,Lem5.2}{\geq} \\
&\geq \left(\left[\frac{1}{2} - 3\max\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon, \epsilon_3(\delta,\varepsilon)\right)\right] \times \varepsilon + \min\left[\frac{1}{2} - 3\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon\right), -2\epsilon_2(\delta,\varepsilon)\right] \times (1-\varepsilon)\right) \\
&\qquad \times \left(\mathrm{P}\left[U \geq m\varepsilon\right] - \mathrm{P}\left[F_{m,n}^c\right]\right) \\
&\quad - \left(4\epsilon_1(\delta) + 8\,\mathrm{e}^{-nH(\delta)}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathrm{P}\left[U \geq m\varepsilon\right] \\
&\leq \frac{4\epsilon_1(\delta) + 8\,\mathrm{e}^{-nH(\delta)}}{\left[\frac{1}{2} - 3\max\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon, \epsilon_3(\delta,\varepsilon)\right)\right] \times \varepsilon + \min\left[\frac{1}{2} - 3\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon\right), -2\epsilon_2(\delta,\varepsilon)\right] \times (1-\varepsilon)} \\
&\qquad + 10\,\mathrm{e}^{-nH(\delta)} = \frac{\mathcal{O}\left(\delta^{\frac{1}{2}}\right)}{\mathcal{O}(\varepsilon) + \mathcal{O}\left(\delta^{\frac{1}{2}}\right)} + \mathcal{O}\left(\mathrm{e}^{-n\delta}\right).
\end{aligned}$$

□

REFERENCES

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

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

[3] 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.

[4] K. Azuma. Weighted sums of certain dependent random variables. *Tohuku Math. J.*, 19:357–367, 1967.

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

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

[7] Federico Bonetto and Heinrich Matzinger. Fluctuations of the longest common subsequence in the case of 2- and 3-letter alphabets. Preprint at http://arxiv.org/abs/math/0410404, 2004.

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

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

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

[11] Raphael Hauser, Servet Martinez, and Heinrich Matzinger. Large deviation based upper bounds for the lcs problem. Technical Report NA03/13 Oxford University Computing Laboratory, 2003.

[12] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statis. Assoc.*, 58:13–30, 1963.

[13] Marcos Kiwi, Martin Loebl, and Ji˘ r Matou˘ sek. Expected length of the longset common subsequence for large alphabets. In *LATIN 2004: Theoretical informatics*, volume 2976 of *Lecture Notes in Comput. Sc.*, pages 302–311. Springer-Verlag, 2004.

[14] Jyri Lember Lember, Heinrich Matzinger, and Clement Dürringer. Deviation from mean in sequence comparison with a periodic sequence. In preparation, 2004.

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

[16] H. Robbins. A remark on stirling's formula. *Amer. Math. Monthly*, 62:26–29, 1955.

[17] Michael J. Steele. An Efron- Stein inequality for nonsymmetric statistics. *Ann. Statist.*, 14:753–758, 1986.

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

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