

# AN UPPER BOUND ON THE CONVERGENCE RATE OF A SECOND FUNCTIONAL IN OPTIMAL SEQUENCE ALIGNMENT

RAPHAEL HAUSER, HEINRICH MATZINGER, AND IONEL POPESCU

ABSTRACT. Consider finite sequences  $X_{[1,n]} = X_1 \dots X_n$  and  $Y_{[1,n]} = Y_1 \dots Y_n$  of length  $n$ , consisting of i.i.d. samples of random letters from a finite alphabet, and let  $S$  and  $T$  be chosen i.i.d. randomly from the unit ball in the space of symmetric scoring functions over this alphabet augmented by a gap symbol. We prove a probabilistic upper bound of linear order in  $(\ln(n))^{1/4} n^{3/4}$  for the deviation of the score relative to  $T$  of optimal alignments with gaps of  $X_{[1,n]}$  and  $Y_{[1,n]}$  relative to  $S$ . It remains an open problem to prove a lower bound. Our result contributes to the understanding of the microstructure of optimal alignments relative to one given scoring function, extending a theory begun in [4].

## 1. INTRODUCTION AND MAIN RESULTS

The subject of this paper is concerned with the asymptotics of optimal sequence alignments for random sequences whose lengths tend to infinity. An important problem that occurs both in bioinformatics and in natural language processing is to decide on the homology of two (or more) finite sequences consisting of symbols from a fixed finite alphabet. A highly successful approach is to fix a *scoring function* and maximise the total score over the set of all alignments with gaps of the two sequences (for a precise definition, see the text below). Despite the combinatorially many alignments to be considered, the total score can be maximised in polynomial time by use of a dynamic

---

1991 *Mathematics Subject Classification.* Primary 60F10; Secondary 92D20, 60K35.

*Key words and phrases.* Sequence alignment, convex geometry, large deviations, percolation theory.

Raphael Hauser was supported by the Engineering and Physical Sciences Research Council [grant number EP/H02686X/1].

Ionel Popescu was partially supported by a grant of the Romanian National Authority for Scientific Research, CNCS - UEFISCDI, project number PN-II-RU-TE-2011-3-0259 and Marie Curie Action Grant PIRG.GA.2009.249200.

programming recursion [5]. Using this approach, two sequences can be considered as homologous if the total score of their optimal alignment relative to a salient scoring function significantly exceeds the typical total score of an optimal alignment of two random sequences of the same length. Rigorous statistical tests on this basis require an understanding of relevant null models, thus giving the initial motivation for the theoretical study of optimal sequence alignments of random sequences and their total scores [6].

The purpose of this paper is to contribute to this theory by studying the following question: given two *symmetric* scoring functions  $S$  and  $T$ , and given two i.i.d. random sequences of length  $n$ , does the rescaled total score (the score divided by  $n$ ) relative to  $T$  of an optimal alignment of the two sequences relative to  $S$  converge as  $n$  tends to infinity, and if the answer to this question is ‘yes’, can we bound the convergence rate? We will answer both questions in the affirmative. Before we go into the technical details of our analysis, we introduce the necessary notation and background and give further details on the main contributions of this paper in relation to the existing literature.

**1.1. Alignments with Gaps.** Let  $n \in \mathbb{N}$  and write  $[1, n] := \{1, \dots, n\}$ . Consider two sequences of length  $n$ ,  $x_{[1, n]} := (x_i)_{i \in [1, n]}$  and  $y_{[1, n]} := (y_j)_{j \in [1, n]}$  consisting of letters from a finite alphabet  $\mathcal{A}$ . Let us augment this alphabet by a symbol  $G$  for a *gap* and write  $\mathcal{A}^* = \mathcal{A} \cup \{G\}$ . We define an *alignment* (with gaps) of  $x_{[1, n]}$  and  $y_{[1, n]}$  as a pair of increasing subsequences  $(i_\ell)_{\ell \in [1, k]}$  and  $(j_\ell)_{\ell \in [1, k]}$  of  $[1, n]$ . For  $\ell \in [1, k]$ , each letter  $x_{i_\ell}$  of the first sequence is then interpreted as aligned with the letter  $y_{j_\ell}$  from the second sequence, while all remaining letters of either sequence are thought of as aligned with gaps.

For example the pair of increasing subsequences  $(\{1, 5, 6, 8\}, \{2, 4, 5, 6\})$  of  $[1, 8]$  correspond to the alignment

$$\begin{array}{cccccccccccc} G & x_1 & x_2 & x_3 & x_4 & G & x_5 & x_6 & x_7 & x_8 & G & G \\ y_1 & y_2 & G & G & G & y_3 & y_4 & y_5 & G & y_6 & y_7 & y_8 \end{array}$$

Note that the same subsequences also correspond to the alignment

$$\begin{array}{cccccccccccc} G & x_1 & x_2 & G & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & G & G \\ y_1 & y_2 & G & y_3 & G & G & y_4 & y_5 & G & y_6 & y_7 & y_8 \end{array}$$

and other arrangements obtained by permuting the order of consecutive letters aligned with gaps, so that the pair  $(\{1, 5, 6, 8\}, \{2, 4, 5, 6\})$  represent in fact an equivalence class of alignments. By slight abuse of language, we will speak about an *alignment* when in fact referring to an entire equivalence class. In order to refer to the set of alignments of two sequences of length  $n$ , we introduce the following notation,

$$\Lambda_{n,k} := \left\{ ((i_\ell)_{\ell \in [1,k]}, (j_\ell)_{\ell \in [1,k]}) : 1 \leq i_1 < \dots < i_k \leq n, 1 \leq j_1 < \dots < j_k \leq n \right\}, \quad (k \in [0, n]),$$

$$\Lambda_n := \bigcup_{k=0}^n \Lambda_{n,k}.$$

**1.2. Scoring Functions and Optimal Alignments.** A function  $R : \mathcal{A}^* \times \mathcal{A}^* \rightarrow \mathbb{R}$  will be called a *symmetric scoring function* if  $R(\alpha, \beta) = R(\beta, \alpha)$  for all  $\alpha, \beta \in \mathcal{A}^*$ , and  $R(G, G) = 0$ . Given a symmetric scoring function  $R$  and two finite sequences  $x_{[1,n]}$  and  $y_{[1,n]}$  consisting of letters from the alphabet  $\mathcal{A}$ , we define the *total score* of  $x_{[1,n]}$  and  $y_{[1,n]}$  under an alignment  $\nu = ((i_\ell), (j_\ell)) \in \Lambda_{n,k}$  as the sum of the scores of individually aligned letter pairs,

$$R_\nu(x_{[1,n]}, y_{[1,n]}) := \sum_{\ell=1}^k R(x_{i_\ell}, y_{j_\ell}) + \sum_{i \in [1,n] \setminus \{i_\ell : \ell \in [1,k]\}} R(x_i, G) + \sum_{j \in [1,n] \setminus \{j_\ell : \ell \in [1,k]\}} R(G, y_j).$$

Note that since our definition of alignments with gaps disallows the situation where a gap is aligned with a gap, the value of  $R(G, G)$  should be inconsequential. Our rationale for requiring  $R(G, G) = 0$  is to simplify some of our formulas, notably the ones defined in Section 1.6.

The *optimal alignment score* of  $x_{[1,n]}$  and  $y_{[1,n]}$  relative to  $R$  is defined by

$$R^*(x_{[1,n]}, y_{[1,n]}) := \max_{\nu \in \Lambda_n} R_\nu(x_{[1,n]}, y_{[1,n]}),$$

while the set of *optimal alignments* of  $x_{[1,n]}$  and  $y_{[1,n]}$  relative to  $R$  is the set of alignments

$$\nu_R^*(x_{[1,n]}, y_{[1,n]}) := \left\{ \nu \in \Lambda_n : R_\nu(x_{[1,n]}, y_{[1,n]}) = R^*(x_{[1,n]}, y_{[1,n]}) \right\}$$

on which the maximum is achieved. Note that in general,  $\nu_R^*$  is not a singleton.

For a scoring function  $R$ , we define

$$|R|_\infty = \max_{(\alpha, \beta)} |R(\alpha, \beta)|.$$

With this notation observe the following general bound on the change of the optimal alignment for two given scoring functions  $R_1$  and  $R_2$ ,

$$(1.1) \quad |R_1^*(x_{[1,n]}, y_{[1,n]}) - R_2^*(x_{[1,n]}, y_{[1,n]})| \leq 2n|R_1 - R_2|_\infty.$$

**1.3. Random Sequences.** Let us now consider two sequences  $(X_i)_{i \in \mathbb{N}} : \Omega \rightarrow \mathcal{A}^\mathbb{N}$  and  $(Y_j)_{j \in \mathbb{N}} : \Omega \rightarrow \mathcal{A}^\mathbb{N}$ , defined on some appropriate probability space  $(\Omega, \mathcal{F}, P)$  so as to consist of i.i.d. random letters  $X_i$  (respectively  $Y_i$ ) drawn from a fixed probability distribution over a finite alphabet  $\mathcal{A}$ . Let us again augment this alphabet by a symbol  $G$  for a *gap* and write  $\mathcal{A}^* = \mathcal{A} \cup \{G\}$ . We write  $X_{[1,n]} = (X_i)_{i=1}^n$  for the finite sequence consisting of the first  $n$  terms of  $(X_i)_{i \in \mathbb{N}}$  and use a similar notation for the second sequence.

Let a symmetric scoring function  $R$  be given on  $\mathcal{A}^* \times \mathcal{A}^*$ . The following is then a well defined random variable for any  $n \in \mathbb{N}$

$$\begin{aligned} L_{n,R} : \Omega &\rightarrow \mathbb{R}, \\ \omega &\mapsto R^*(X_{[1,n]}(\omega), Y_{[1,n]}(\omega)), \end{aligned}$$

and we write

$$\begin{aligned} \nu_{n,R}^* : \Omega &\rightarrow \mathcal{P}(\Lambda_n), \\ \omega &\mapsto \nu_R^*(X_{[1,n]}(\omega), Y_{[1,n]}(\omega)) \end{aligned}$$

for the random set of optimal alignments of  $X_{[1,n]}$  and  $Y_{[1,n]}$  relative to  $R$ .

It was shown in [2] that

$$(1.2) \quad \frac{E[L_{n,R}]}{n} \xrightarrow{n \rightarrow \infty} \lambda_R,$$

where  $\lambda_R$  is some deterministic constant that depends only on  $R$ . Lemma 2.1 and 2.2 contain a proof and some quantitative convergence bound for the convergence of  $L_{n,R}/n$  and  $E[L_{n,R}]/n$  to  $\lambda_R$ .

We close this section observing that for two given scoring functions,  $R_1$  and  $R_2$ ,

$$(1.3) \quad |L_{n,R_1} - L_{n,R_2}| \leq 2n|R_1 - R_2|_\infty \text{ and } |\lambda_{R_1} - \lambda_{R_2}| \leq 2|R_1 - R_2|_\infty$$

which follow directly from (1.1) and (1.2).

**1.4. The Problem Setting of this Paper.** Let us now consider two different symmetric scoring functions  $S$  and  $T$  and investigate the total score relative to  $T$  of an optimal alignment relative to  $S$ . Using the random sequences introduced above, we define the following random subsets of  $\mathbb{R}^2$ ,

$$\begin{aligned} \text{SCORES}_{S,T}^n &:= \left\{ \left( \frac{S_\nu(X_{[1,n]}, Y_{[1,n]})}{n}, \frac{T_\nu(X_{[1,n]}, Y_{[1,n]})}{n} \right) : \nu \in \Lambda_n \right\} \\ \text{SET}_{S,T}^n &:= \text{cl}(\text{conv}(\text{SCORES}_{S,T}^n)), \end{aligned}$$

where  $\text{cl}(\cdot)$  denotes the topological closure in the canonical topology of  $\mathbb{R}^2$  and  $\text{conv}(\cdot)$  denotes the convex hull.

Next, consider a symmetric scoring function  $R = aS + bT$  given as a linear combination of  $S$  and  $T$ . It follows from our definition of  $\text{SET}_{S,T}^n$  that

$$(1.4) \quad \frac{L_{n,R}}{n} = \max_{(x,y) \in \text{SET}_{S,T}^n} f_{(a,b)}(x, y),$$

where  $f_{(a,b)} : (x, y) \mapsto ax + by$  is the linear form on  $\mathbb{R}^2$  defined by the weights  $a, b$ . Combining Equations (1.2) and (1.4), it follows that

$$(1.5) \quad \max_{(x,y) \in \text{SET}_{S,T}^n} f_{(a,b)}(x, y) \xrightarrow{n \rightarrow \infty} \lambda_{aS+bT}, \quad \text{a.s..}$$

Since any linear functional  $f \in (\mathbb{R}^2)^*$ , can be written in the form  $f(x, y) = ax + by$ , it follows that  $\lambda_f = \lambda_{a,b}$  is well defined for any linear functional  $f$ .

Using (1.5) for the case of  $a, b$  rational numbers combined with a density argument and the estimate from (1.3), we can conclude that almost surely,

$$(1.6) \quad \max_{(x,y) \in \text{SET}_{S,T}^n} f(x, y) \xrightarrow{n \rightarrow \infty} \xi_f$$

for any linear functional  $f \in (\mathbb{R}^2)^*$ .

We observe that, if a sequence of random compact convex sets  $A_1, A_2, \dots \subset \mathbb{R}^2$  has the property that for any linear functional  $f \in (\mathbb{R}^2)^*$ ,

$$(1.7) \quad \max_{(x,y) \in A_n} f(x,y) \xrightarrow{n \rightarrow \infty} \xi_f, \quad \text{a.s.},$$

where  $\xi_f \in \mathbb{R}$  is a deterministic constant that depends only on  $f$ , then the sequence  $(A_n)_{n \in \mathbb{N}}$  converges in Hausdorff distance to a convex compact set  $A$ . We will prove this claim in Lemma 2.4. For compact sets  $A, B \subset \mathbb{R}^2$ , the Hausdorff distance is defined as

$$(1.8) \quad d_H(A, B) = \max\left\{\sup_{x \in A} \inf_{y \in B} d(x, y), \sup_{y \in B} \inf_{x \in A} d(x, y)\right\},$$

where  $d(x, y) = \|x - y\|_2$  denotes the Euclidean distance.

Using (1.6) and the sequence of sets  $A_n = (\text{SET}_{S,T}^n)_{n \in \mathbb{N}}$  in (1.7), we conclude from (1.8) that

$$(1.9) \quad d_H(\text{SET}_{S,T}^n, \text{SET}_{S,T}) \xrightarrow{n \rightarrow \infty} 0, \quad \text{a.s..}$$

One of our goals is to refine this analysis and quantify an upper-bound on the rate of convergence. An upper bound on the convergence was given in [4] for scoring functions that are not necessarily symmetric. In this paper we give a much simpler proof that is made possible by exploiting the symmetry of scoring functions. Since most scoring functions used in applications are symmetric, the simplification is of interest.

Another goal is to study how much the total score relative to  $T$  varies when two random strings are aligned optimally relative to  $S$ . Note that we have

$$L_{n,S} = \max_{(x,y) \in \text{SET}_{S,T}^n} x.$$

In general, we should not expect that  $\nu_{n,S}^*$  to be a singleton. In other words, there may exist multiple optimal alignments of  $X_{[1,n]}$  and  $Y_{[1,n]}$  relative to  $S$ . Therefore, we need to consider the following quantities,

$$(1.10) \quad \max_{\pi \in \nu_{n,S}^*} \frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} = \max \left\{ y : (x, y) \in \text{SET}_{S,T}^n, x = \frac{L_{n,S}}{n} \right\},$$

$$(1.11) \quad \min_{\pi \in \nu_{n,S}^*} \frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} = \min \left\{ y : (x, y) \in \text{SET}_{S,T}^n, x = \frac{L_{n,S}}{n} \right\},$$

Lemma 2.5 will establish that if  $\max_{(x,y) \in \text{SET}_{S,T}} x$  has a unique maximizer  $(x_0, y_0)$ , then the upper and lower bounds (1.10), (1.11) both converge to  $y_0$  almost surely.

**1.5. Relation to percolation and empirical distribution of aligned letter pairs.** The current paper is a continuation of the paper [4] on the empirical distribution of letter pairs in optimal alignment of random sequences. To explain what this is about let us start with an example.

Consider the following alignment with gaps:

$A$	$T$		$T$	$A$	$T$
$A$	$T$	$A$	$T$	$A$	$A$

Take here  $x = AT T A T$  and  $y = A T A T A A$  to be the two strings of letters and denote the above alignment by  $\pi$ . The frequency of the aligned letters in the alignment  $\pi$  is given by

$$\vec{p}_\pi(x, y) = (p_{AA}, p_{AT}, p_{AG}, p_{TA}, p_{TT}, p_{TG}, p_{GA}, p_{GT}) = \left( \frac{1}{3}, 0, 0, \frac{1}{6}, \frac{1}{3}, 0, \frac{1}{6}, 0 \right).$$

For instance, two columns are  $A$  aligned with  $A$  and thus the frequency of the pair  $(A, A)$  is given by  $p_{AA} = 2/6 = 1/3$ . In the same fashion, two columns are  $T$  aligned with  $T$  and this corresponds to  $p_{TT} = 2/6 = 1/3$ . Furthermore, there is one gap aligned with  $A$  and a  $T$  aligned with an  $A$  and this completely describe the vector  $\vec{p}_\pi(x, y)$ .

The vector  $\vec{p}_\pi(x, y)$  is called the *empirical distribution of the aligned letter pairs of the alignment  $\pi$* . One of the basic questions is how does it behave when we have long random strings?

In [4], it is proven that the empirical distribution of the aligned letter pairs of an optimal alignment converges to a limit when the length of the strings goes to infinity and the letters are taken to be independent and identically distributed. However, one serious restriction is that the scoring function needs to be chosen at random and thus the result in [4] holds only for almost all scoring functions rather than for all. A further technical detail is that in this paper and in [4] the empirical distribution of the aligned letter pairs is taken by rescaling by  $n$  and not by the number of columns. This is of minor importance and we are going to ignore it.

Assume now that  $X_1, X_2, \dots$  and  $Y_1, Y_2, \dots$  be i.i.d. sequences of letters from a finite alphabet which is fixed. Let  $\pi_n$  denote any optimal alignment of  $X_1 \dots X_n$  with  $Y_1 \dots Y_n$  according to the scoring function  $S$ . The result in [4], states that for almost all scoring function  $S$ , the empirical distribution of the aligned letter pairs

$$\vec{p}_{\pi_n}(X_1 \dots X_n, Y_1 \dots Y_n)$$

converges almost surely as  $n$  goes to infinity. Notice that we can choose  $\pi_n$  in any way we want among the alignments of  $X_1 \dots X_n$  with  $Y_1 \dots Y_n$  which are optimal according to  $S$ .

The practical importance of this comes from the fact that in many cases an optimal alignment score can be used to determine if two DNA strings are related or not. Nevertheless, there are also many real life situations where such a one dimensional score may not conclude about the similarity of the strings. Typically, in those hard-to-discriminate-cases, the optimal alignments “look completely different” depending on whether the strings are related or not. This suggests that a different approach should be used. For instance, at first, compute the (an) optimal alignment with a given scoring function. Then, use some statistic related to the empirical distribution of the aligned letter pairs of an optimal alignment. (For example how many same letters are aligned with each other. We can also consider Meta-letters, which represents words of a finite given length from our alphabet  $\mathbb{A}$ . The same analysis can be carried out for the empirical distribution of those Meta-letters. Hence, we can extend our analysis to the joint distribution of subsequently aligned letter pairs and even arbitrary strings).

If there is a unique limiting distribution for the aligned letter pairs by optimal alignments, things become much easier for testing. When the limiting distribution of optimal alignments is unique, then when the strings are long enough, all empirical distributions along different optimal paths should be similar, thus, we can just consider for a statistical test, one optimal alignment with its empirical distribution of aligned letter pairs. If there are several coexisting possible limits of optimal alignments, we never know all the different empirical distributions and thus it not possible to construct testings for similarity of the sequences as for the unique limiting case.

In preliminary testing [7], there is very encouraging evidence for cases where relatedness is difficult to recognize with a single scoring function. Our approach from [7] without fine tuning and optimization, already works at least as well as BLAST (a dedicated software for DNA analysis).

In [4] we establish only a result on the convergence to the empirical letter pair distribution but not on the speed of that convergence. In the present work, we also have an estimate on the convergence speed.

In the current paper, we do not need the whole empirical distribution, only a one dimensional functional of it. (Since with a finite number of linear functionals we can reconstitute the whole empirical distribution of the aligned letter pairs.)

As an example of such a linear functional, take  $T$  to be defined by:

$$T(p_{AA}, p_{AT}, \dots, p_{G,T}) := p_{AA}$$

So, then  $T$  is simply for an alignment the proportion of columns aligning an  $A$  with an  $A$ .

Let again  $\pi_n$  be an alignment of the strings  $X_1 \dots X_n$  and  $Y_1 \dots Y_n$  which is optimal according to  $S$  (chosen arbitrarily from the many possible ones). The main result of the current paper, namely Theorems 1.1 and 1.2, imply that

$$T(\vec{p}_{\pi_n}(X_1 \dots X_n, Y_1 \dots Y_n))$$

converges at a rate

$$(1.12) \quad \text{constant} \times \left( \frac{\ln(n)}{n} \right)^{1/4}$$

again, provided that we chose the scoring functions  $S$  and  $T$  at random. The rate (1.12), holds then for almost all  $S$  and  $T$ .

The difficulty in analyzing the empirical distribution function of optimal alignments, resides in the fact that concentration inequalities are not available. Whilst Azuma-Hoeffding applies to the optimal alignment score, it does not apriori apply to the empirical aligned letter pair distribution of an optimal alignment. The reason is the following. When you change one letter only in the strings  $X$  or  $Y$ , then the score according to  $S$  changes by at most  $|S|$ . Hence, when the strings  $X$  and  $Y$  are i.i.d. we can apply Azuma-Hoeffding to the optimal alignment score. This argument does not work for the empirical letter pair distribution of an optimal alignment. Indeed, a one-letter-change only could potentially lead to an entirely new optimal alignment. This in turn,

would result in a massive change for the empirical letter pair distribution. And since we assume long term correlation in optimal alignments, we believe that a one letter change will often result in an entirely different optimal alignment.

Optimal alignments of random strings can be reformulated as a Last Passage Percolation problem with correlated weights. In First Passage Percolation (FPP) and Last Passage Percolation (LPP), the question of the empirical weight-distribution along an optimal path, is known to be difficult. For example, one long standing open problem [8], is the question of how finding how many vertical and horizontal vertices are in an optimal path from say  $(0, 0)$  to  $(n, n)$ . Formulated differently, we consider FPP on the  $\mathbb{Z}^2$  grid. The weights are i.i.d. and we ask if the proportion of vertical to horizontal edges in any shortest path from  $(0, 0)$  to  $(n, n)$  converges as  $n \rightarrow \infty$ .

We believe that in certain cases, this percolation problem can be solved with the techniques of the current paper. Think for example of a situation where the distribution of the weights is different for horizontal and for vertical edges (for the FPP on  $\mathbb{Z}^2$ ). Assume also that these distributions have finite support which are disjoint. Then we could count along a shortest path the proportion of weights of each type encountered. This would yield the empirical distribution of the weights along an optimal path. The same result as in this article applies to this scenario provided we chose the weights at random.

In FPP an open problem is to determine in which directions the asymptotically rescaled wet zone is strictly convex. This is one of the hardest problems. In this paper we bypass this difficulty by the fact that we choose the direction at random with uniform probability. Consequently, here, as well as in [4], we get strict convexity at the boundary for almost every point. What we do then is to choose the scoring functions at random and then automatically get a.s. the desired strict convexity at the places where these scoring functions reach their optimal values on the convex sets involved. This convex set is not exactly the same as the rescaled wet zone, but the idea is very similar.

Finally we should mention that as in the case with the wet-zone proving, for a specific scoring function that there is strict convexity in the place where the scoring function reaches its maximum on the convex set is very difficult. But, then again, if the

scoring function is chosen at random, the result will hold. And we can consider, the scoring functions used by biologists to be chosen at random so that our result holds. Let us explain why. The scoring function is the logarithm of an evolution probability. Nature does not tend to give to two different letter pairs exactly the same alignment score. Hence, you can think of the scoring function used by biologist as a point where a little random disturbance has been introduced. That random disturbance would then guarantee the a.s. strict convexity which we need for our result.

Finally let us mention a counter example. Recall that the Longest Common Subsequence (LCS) of two strings  $x$  and  $y$  is a subsequence of both strings of maximal lengths. If the LCS of two strings is large, then, we can assume that the strings are related. Now, the length of the LCS, can be viewed as an optimal alignment score. Indeed, it suffices to take a scoring function with 0 gap penalty and which gives one point for identical aligned letters and 0 otherwise. Our result does not apply to LCS. The LCS-scoring function is not chosen at random. For instance, every letter aligned with itself yields precisely 1 point. This is never the case, in the scoring functions which arise naturally as the logarithms of pairwise evolution probabilities and we can view these as small random perturbations.

**1.6. Statement of the Main Results.** To state the main results of this paper, we introduce the following norms on the set of symmetric scoring functions  $R : \mathcal{A}^* \times \mathcal{A}^* \rightarrow \mathbb{R}$ ,

$$(1.13) \quad |R| := \max_{a,b,c \in \mathcal{A}^*} |R(a,b) - R(a,c)|, \quad (\text{the change norm}),$$

$$(1.14) \quad |R|_2 := \sqrt{\sum_{a,b \in \mathcal{A}^*} R^2(a,b)}, \quad (\text{the Frobenius norm}).$$

Notice that the change norm plays the following important role: given two finite sequences and a fixed alignment with gaps, changing a single letter of one of the two sequences into an arbitrary other letter from the alphabet  $\mathcal{A}$  changes the total score of the alignment by at most  $|R|$ . It is an abuse of notation as  $|\cdot|$  is not really a norm. Indeed, for instance in the case of a scoring function  $R$  which is diagonal, that is  $R(a,b) \neq 0$  if and only if  $a = b$ , we have  $|R| = 0$ . Again, by the abuse of notation we will continue to call  $|\cdot|$  a norm.

**Theorem 1.1.** *Let  $S$  and  $T$  be two symmetric scoring functions on  $\mathcal{A}^* \times \mathcal{A}^*$  such that the optimisation problem  $\max_{(x,y) \in \text{SET}_{S,T}^n} x$  has a unique maximizer  $(x_0, y_0)$  and the boundary of  $\text{SET}_{S,T}$  has curvature at least  $k > 0$  at this point, then the following bound applies for  $n \geq n_0$  with  $n_0$  large enough,*

$$\mathbb{P} \left[ \left| \frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} - y_0 \right| \leq \frac{5|T| + 2\sqrt{30|S|}}{k} \left( \frac{\ln(n e)}{n} \right)^{1/4}, \quad \forall \pi \in \nu_{n,S}^* \right] \geq 1 - 3n^{-\ln n}.$$

Here,  $e$  is the Euler constant and  $n_0$  depends on the geometry of the boundary  $\partial \text{SET}_{S,T}$  near  $(x_0, y_0)$ .

It is clear that  $n_0$  depends (in our argument) on the local behavior of the boundary of  $\partial \text{SET}_{S,T}$  near the point  $(x_0, y_0)$ . Given more information about the local structure we can eventually extract a concrete bound on  $n_0$  from the proof of the Theorem.

In particular if both  $S$  and  $T$  have change norm less than 1, the statement of Theorem 1.1 simplifies to

$$\mathbb{P} \left[ \left| \frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} - y_0 \right| \leq \frac{17}{k} \left( \frac{\ln(n e)}{n} \right)^{1/4}, \quad \forall \pi \in \nu_{n,S}^* \right] \geq 1 - 3n^{-\ln n}, \quad \forall n \geq n_0$$

where  $n_0$  depends on the local geometry of the boundary of  $\text{SET}_{S,T}$  near  $(x_0, y_0)$ .

The curvature condition at the point  $(x_0, y_0)$  means that one can parametrize the boundary  $\partial \text{SET}_{S,T}$  of the set  $\text{SET}_{S,T}$  by a curve  $c(t)$  for  $t$  in a neighbourhood of 0, with  $c(0) = (x_0, y_0)$  and  $\|\dot{c}(t)\|_2 = 1$  for all allowable  $t$ , where  $\dot{c}$  denotes the derivative with respect to  $t$ , the curvature

$$\kappa(\partial \text{SET}_{S,T}, (x_0, y_0)) := \|\ddot{c}(0)\|_2$$

then being defined as the standard curvature of this curve at  $t = 0$ . Here we denoted by  $\|\cdot\|_2$  the standard Euclidian norm of a vector in  $\mathbb{R}^2$ . By convention, we define the curvature at vertices of  $\partial \text{SET}_{S,T}$  (points on the boundary where  $\text{SET}_{S,T}$  has a normal cone with nonempty interior) to be  $+\infty$ . We postpone the proof of Theorem 1.1 until Section 3.

While Theorem 1.1 establishes that if the boundary of  $\text{SET}_{S,T}$  has positive curvature at  $(x_0, y_0)$ , then the  $T$ -score on an  $S$ -optimal alignment has a fluctuation of order at most

$O([\ln(n)/n]^{0.25})$ , the conditions of this result are difficult to verify in practice. However, as the following result shows, they apply generically:

**Theorem 1.2.** *Let  $S$  and  $T$  be chosen i.i.d. uniformly at random from the Frobenius-unit sphere in the space of symmetric scoring functions. Then the following hold true,*

- (1)  $\max_{(x,y) \in \text{SET}_{S,T}^n} x$  has a unique maximizer  $(x_0, y_0)$  almost surely,
- (2) for any real number  $k > 0$ ,

$$\mathbb{P} [\kappa(\partial \text{SET}_{S,T}, (x_0, y_0)) < k] \leq \frac{4k}{\pi},$$

where  $\kappa(\partial \text{SET}_{S,T}, (x_0, y_0))$  is the curvature at  $(x_0, y_0)$  of the boundary of  $\text{SET}_{S,T}$ .

The Frobenius norm is defined as in (1.14) and is a natural norm because it has invariant properties under the orthogonal transformations.

Combining Theorems 1.1 and 1.2, we arrive at the following conclusion:

**Corollary 1.1.** *If the symmetric scoring functions  $S$  and  $T$  are chosen as in Theorem 1.2, then almost surely there exists  $k > 0$  such that*

$$\mathbb{P} \left[ \left| \frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} - y_0 \right| \leq \frac{17}{\max(k, 1)} \left( \frac{\ln(n e)}{n} \right)^{1/4}, \quad \forall \pi \in \nu_{n,S}^* \right] \geq 1 - 3n^{-\ln n}, \quad \forall n \text{ large enough.}$$

## 2. PRELIMINARY RESULTS AND THEIR PROOFS

In this section we derive the main estimates on which the proofs of our main theorems rely. We begin by giving the classical Azuma-Hoeffding – McDiarmid Inequality.

**Theorem 2.1.** *Let  $W_1, \dots, W_n$  i.i.d. random variables that take values in some set  $D$ , let  $a \geq 0$  be a constant and  $f : D^n \rightarrow \mathbb{R}$  a  $n$ -variate real function with the property that for any  $i \in [1, n]$ ,  $w \in D^n$  and  $z \in D$ ,*

$$|f(w_1, w_2, \dots, w_n) - f(w_1, w_2, \dots, w_{i-1}, z, w_{i+1}, \dots, w_n)| \leq a.$$

*Then, for any  $\epsilon > 0$ , the following inequalities hold true,*

$$\mathbb{P} [|f(W_1, W_2, \dots, W_n) - \mathbb{E}[f(W_1, \dots, W_n)]| \geq \epsilon n] \leq 2 \exp(-\epsilon^2 n / (2a^2)),$$

$$\mathbb{P} [f(W_1, W_2, \dots, W_n) - \mathbb{E}[f(W_1, \dots, W_n)] \geq \epsilon n] \leq \exp(-\epsilon^2 n / (2a^2)).$$

For a proof, see e.g. [1].

**Lemma 2.1.** *For any symmetric scoring function  $R : \mathcal{A}^* \times \mathcal{A}^* \rightarrow \mathbb{R}$  there exists a deterministic constant  $\lambda_R$  such that*

$$\frac{L_{n,R}}{n} \xrightarrow{n \rightarrow \infty} \lambda_R, \quad a.s.$$

*Proof.* It is trivial to see that the function  $n \mapsto E[L_{n,R}]$  is superadditive. Therefore and since the scoring function is bounded, we have

$$(2.1) \quad E[L_{n,R}]/n \xrightarrow{n \rightarrow \infty} \lambda_R := \sup_{n \geq 1} E[L_{n,R}]/n,$$

where  $\sup_{n \geq 1} E[L_{n,R}]/n$  is finite. For any  $\epsilon > 0$ , let  $D_{n,R}(\epsilon)$  denote the event

$$D_{n,R}(\epsilon) = \{|L_{n,R} - E[L_{n,R}]| \geq \epsilon \ln(n) \sqrt{n}\}.$$

Applying Theorem 2.1 with  $a = |R|$ , we obtain

$$(2.2) \quad \mathbb{P} [D_{n,R}(\epsilon) \leq 2 \exp(-\epsilon^2 (\ln n)^2 / 2 |R|^2)] = 2n^{-\frac{\epsilon^2 \ln n}{2 |R|^2}}.$$

By virtue of Borel-Cantelli, the finite summability of (2.2) implies that almost surely at most a finite number of the events  $D_{n,R}(\epsilon)$  will hold. Combined with (2.1), and using the fact that  $\epsilon > 0$  was arbitrary, this implies the claim.  $\square$

The next result gives the rate of convergence for  $E[L_n(R)]/n$  toward  $\lambda_R$ . A bound for non-symmetric scoring functions was given in [4]. Here we exploit the symmetry of  $R$  to give a tighter bound that we will use to prove our main theorems.

**Lemma 2.2.** *For any symmetric scoring function  $R : \mathcal{A}^* \times \mathcal{A}^* \rightarrow \mathbb{R}$ , the following convergence bound applies,*

$$\left| \lambda_R - E \left[ \frac{L_{n,R}}{n} \right] \right| \leq 3|R| \sqrt{\frac{\ln(n e)}{n}}.$$

*Proof.* To simplify the notation, let us write  $\lambda_{n,R} = E[L_{n,R}]/n$ . Let  $m = kn$  for some  $k \in \mathbb{N}$ , and let  $\mathcal{P}^{m,n}$  be the set of pairs of partitions of the integer interval  $[1, m]$  into  $2k$  pieces for which the sum of the lengths of the  $i$ -th pieces is always  $n$ . In other words,

$$\mathbf{p} = (i_0, i_1, \dots, i_{2k}, j_0, j_1, \dots, j_{2k})$$

is in  $\mathcal{P}^{m,n}$  if

$$\begin{aligned} 0 &= i_0 < i_1 < \dots < i_{2k} = m, \\ 0 &= j_0 < j_1 < \dots < j_{2k} = m, \quad \text{and} \\ i_\ell - i_{\ell-1} + j_\ell - j_{\ell-1} &= n, \quad \forall \ell \in [1, 2k]. \end{aligned}$$

For a partition  $p \in \mathcal{P}^{n,m}$ , let  $L_{m,R}^p$  denote the optimal alignment score of  $X_{[1,n]}$  and  $Y_{[1,n]}$  relative to  $R$  under the extra constraint that the  $l$ -th pieces of the two partitions are aligned with each other, hence imposing that  $X_{i_{l-1}+1} \dots X_{i_l}$  be aligned with  $Y_{j_{l-1}+1} \dots Y_{j_l}$  for  $l = 1, \dots, 2k$ . In other words, we have

$$(2.3) \quad L_{m,R}^p = \sum_{l=2}^{2k} R(X_{i_{l-1}+1} \dots X_{i_l}, Y_{j_{l-1}+1} \dots Y_{j_l}).$$

We can apply Azuma-Hoeffding to our constrained optimal alignment score  $L_{m,R}^p$  to justify that for any constant  $\epsilon > 0$ ,

$$(2.4) \quad P(L_{m,R}^p - E[L_{m,R}^p] \geq \epsilon m) \leq \exp\left(-\frac{\epsilon^2 \cdot m}{2|R|^2}\right).$$

The optimal alignment score  $L_{m,R}$  is not always equal to one of the constrained alignment scores  $L_{m,R}^p$ , however we can argue that it is not far from this. In fact, it is not hard to see that for some partition  $p$

$$(2.5) \quad |L_{m,R} - L_{m,R}^p| \leq 4k|R|.$$

To argue about this, we reason as follows. Ideally we take one column after the other in the optimal alignment. We do this starting from the left and going to the right until we have sufficiently many columns so the total number of letters from the strings of  $X$  and  $Y$  equals  $n$ . That is we stop when the columns we have chosen contain  $n$  letters. This gives then  $i_1$  and  $j_1$  for the first part of the partition  $p$ . That is the last letter of  $X$  in the columns we chose has index  $i_1$  and the last letter of  $Y$  in the columns we chose has index  $j_1$ . After we have done this we start again with the remaining of the sequence and take the next columns until we have again  $n$  letters. This then yields  $i_2$  and  $j_2$ . We keep repeating this process, until  $i_1, \dots, i_{2k}$  and  $j_1, \dots, j_{2k}$  are all determined. The only problem with this procedure is that sometimes when we add a column, we could

overshoot by 1 letter. For instance if we have already  $n - 1$  letters and the next column contains two letters, then we get  $n + 1$  letters. Thus, the alignment defined by  $i_1, \dots, i_{2k}$  and  $j_1, \dots, j_{2k}$ , may be off by one letter per pair of string-pieces

$$(2.6) \quad X_{i_{l-1}+1} \dots X_{i_l}, Y_{j_{l-1}+1} \dots Y_{j_l}.$$

That gives a total of  $2k$  letters differently aligned. Each of these letters can lead to a difference of  $2|R|$  in the alignment score. So, the total difference between the optimal alignment and our alignment which aligns all the pieces (2.6), is at most  $2k \times 2|R|$ .

Continuing now, from (2.5), if the alignment score  $L_{m,R}$  is to exceed a given benchmark, at least one of the constrained scores  $L_{m,R}^p$  must exceed this benchmark shifted by the correction term (2.5). This implies

$$(2.7) \quad \mathbb{P}[L_{m,R} \geq n\lambda_{n,R}k + \epsilon m] \leq \sum_{p \in \mathcal{P}^{m,n}} \mathbb{P}[L_{m,R}^p \geq n\lambda_{n,R}m + \epsilon m - 4k|R|].$$

We claim that by symmetry of  $R$ , we have

$$(2.8) \quad \mathbb{E}[L_{m,R}^p] \leq n\lambda_{n,R}k, \quad \forall p \in \mathcal{P}^{n,m}.$$

Our claim holds for two reasons: Firstly,  $i_l - i_{l-1} + j_l - j_{l-1} = n$  implies  $i_l < i_{l-1} + n$  and  $j_l < j_{l-1} + n$  and

$$\begin{aligned} R(X_{i_{l-1}+1} \dots X_{i_l}, Y_{j_{l-1}+1} \dots Y_{j_l}) + R(X_{i_l+1} \dots X_{i_{l-1}+n}, Y_{j_l+1} \dots Y_{j_{l-1}+n}) \\ \leq R(X_{i_{l-1}+1} \dots X_{i_{l-1}+n}, Y_{j_{l-1}+1} \dots Y_{j_{l-1}+n}). \end{aligned}$$

Taking expectations on both sides, we find

$$(2.9) \quad \begin{aligned} \mathbb{E}[R(X_{i_{l-1}+1} \dots X_{i_l}, Y_{j_{l-1}+1} \dots Y_{j_l})] + \mathbb{E}[R(X_{i_l+1} \dots X_{i_{l-1}+n}, Y_{j_l+1} \dots Y_{j_{l-1}+n})] \\ \leq \mathbb{E}[R(X_{i_{l-1}+1} \dots X_{i_{l-1}+n}, Y_{j_{l-1}+1} \dots Y_{j_{l-1}+n})]. \end{aligned}$$

Secondly, the crucial assumption that  $R$  be symmetric implies that the two terms on the left-hand side of (2.9) are equal, thus yielding

$$(2.10) \quad \begin{aligned} 2 \mathbb{E}[R(X_{i_{l-1}+1} \dots X_{i_l}, Y_{j_{l-1}+1} \dots Y_{j_l})] &\leq \mathbb{E}[R(X_{i_{l-1}+1} \dots X_{i_{l-1}+n}, Y_{j_{l-1}+1} \dots Y_{j_{l-1}+n})] \\ &= \mathbb{E}[R(X_1 \dots X_n, Y_1 \dots Y_n)] \\ &= n\lambda_{n,R}. \end{aligned}$$

Taking the expectation on both sides of (2.3) and applying (2.10) to each term on the right-hand side yields the claimed inequality, (2.8).

Substitution of (2.8) into (2.7) now yields

$$(2.11) \quad \mathbb{P}[L_{m,R} \geq n\lambda_{n,R}k + \epsilon m] \leq \sum_{\mathbf{p} \in \mathcal{P}^{m,n}} \mathbb{P}[L_{m,R}^{\mathbf{p}} \geq \mathbb{E}[L_{m,R}^{\mathbf{p}}] + \epsilon m - 4k|R|].$$

Using (2.4) and the fact that  $\mathcal{P}^{n,m}$  has fewer than  $\binom{m}{k}^2$  elements yields that for large  $n$  and  $k$ ,

$$(2.12) \quad \mathbb{P}[L_{m,R} \geq n\lambda_{n,R}k + \epsilon m] \leq \binom{m}{k}^2 \exp\left(-\frac{(\epsilon - 4|R|/n)^2 \cdot m}{2|R|^2}\right).$$

Let  $Z$  be a binomial variable with parameters  $m$  and  $p = 1/n$ , so that we have

$$\mathbb{P}[Z = k] = \binom{m}{k} \left(\frac{1}{n}\right)^k \cdot \left(\frac{n-1}{n}\right)^{m-k} \leq 1,$$

and hence,

$$(2.13) \quad \binom{m}{k} \leq n^k \cdot \left(\frac{1}{1 - \frac{1}{n}}\right)^{k(n-1)} \leq (e \cdot n)^k, \quad (n \gg 1).$$

Substituting (2.13) into (2.12), we find that for large  $n$ ,

$$(2.14) \quad \mathbb{P}\left[\frac{L_{m,R}}{m} \geq \lambda_{n,R} + \epsilon\right] \leq \exp\left(k \left[2 \ln(e \cdot n) - \frac{(\epsilon - 4|R|/n)^2 \cdot n}{4|R|^2}\right]\right).$$

The key now is to let  $k$  tend to infinity. In doing so, we know on the one hand that  $L_{m,R}/m \rightarrow \lambda_R$ , and on the other that the right-hand side of (2.14) converges either to 0 or  $+\infty$ . It does converge to 0 only if

$$2 \ln(e \cdot n) - \frac{(\epsilon - 4|R|/n)^2 \cdot n}{4|R|^2} < 0$$

which is certainly satisfied if  $n$  is chosen large enough ( $n > 10$  suffices) and

$$\epsilon = 3|R|\sqrt{\frac{\ln(ne)}{n}}.$$

Therefore, we find

$$\mathbb{P}\left[\lambda_R \geq \lambda_{n,R} + 3|R|\sqrt{\frac{\ln(ne)}{n}}\right] = 0,$$

and since  $\lambda_R$  is a constant, and similarly  $\lambda_{n,R}$ , we actually deduce that

$$\lambda_R \leq \lambda_{n,R} + 3|R| \sqrt{\frac{\ln(ne)}{n}}.$$

On the other hand, we also know from (2.1) that  $\lambda_n/n \leq \lambda_R$ , thus concluding the proof.  $\square$

**Lemma 2.3.** *Let  $R : \mathcal{A}^* \times \mathcal{A}^* \rightarrow \mathbb{R}$  be a symmetric scoring function and let  $A^n(R)$  denote the event*

$$A^n(R) = \left\{ \left| \lambda_R - \frac{L_{n,R}}{n} \right| \leq 5|R| \sqrt{\frac{\ln(ne)}{n}} \right\}.$$

Then for large  $n$ ,

$$\mathbb{P}[A^n(R)] \geq 1 - n^{-\ln n}.$$

*Proof.* This follows by combining (2.2) with  $\epsilon = 2|R|$ , Lemma 2.2, Theorem 2.1 and Lemma 2.1.  $\square$

The next result is about the convergence of convex compact sets.

**Lemma 2.4.** *Let  $(A_n)_{n \in \mathbb{N}}$  be a sequence of random compact convex sets in  $\mathbb{R}^2$  such that for any linear form  $f \in (\mathbb{R}^2)^*$  there exists a deterministic constant  $\xi_f \in \mathbb{R}$  for which*

$$\max_{(x,y) \in A_n} f(x,y) \xrightarrow{n \rightarrow \infty} \xi_f, \quad a.s.$$

Then there exists a deterministic compact convex set  $A \subset \mathbb{R}^2$  for which

$$d_H(A_n, A) \xrightarrow{n \rightarrow \infty} 0, \quad a.s.,$$

where  $d_H$  is the Hausdorff distance.

*Proof.* Let  $F$  be a dense countable subset of the unit sphere in  $(\mathbb{R}^2)^*$ . Then

$$A := \{(x, y) : f(x, y) \leq \xi_f, \forall f \in (\mathbb{R}^2)^*\} = \{(x, y) : f(x, y) \leq \xi_f, \forall f \in F\}.$$

Furthermore,  $A$  is compact and convex, the condition of the lemma implies that

$$(2.15) \quad \mathbb{P} \left[ \max_{(x,y) \in A_n} f(x,y) \xrightarrow{n \rightarrow \infty} \xi_f, \forall f \in F \right] = 1,$$

and we have

$$(2.16) \quad \max_{(x,y) \in A} f(x,y) = \xi_f, \quad \forall f \in F.$$

Suppose it is not the case that  $d_H(A_n, A) \rightarrow 0$  almost surely. Then there exists  $\delta > 0$  and a set  $\mathcal{E} \subset \Omega$  such that  $P[\mathcal{E}] > 0$  and  $\forall \omega \in \mathcal{E}$  there exists a sequence of points  $(\alpha_n(\omega))_{n \in \mathbb{N}}$  such that  $\alpha_n(\omega) \in A_n(\omega)$  and

$$d(\alpha_n(\omega), A) := \min_{\beta \in A} d(\alpha_n(\omega), \beta) \geq \delta.$$

We claim that all the sets  $A_n(\omega)$  are contained in a big fixed box  $\Delta(\omega)$ . This can be seen by considering the functionals  $f(x,y) = x$ ,  $f(x,y) = -x$ ,  $f(x,y) = y$  and  $f(x,y) = -y$  combined with the conditions of the lemma to show that each of the coordinates stays bounded.

Since all sets  $A_n(\omega)$  are contained in some large closed box, there exists a convergent subsequence  $(\alpha_{n_k}(\omega))_{k \in \mathbb{N}} \rightarrow \alpha(\omega)$ . The continuity of the function  $\alpha \mapsto d(\alpha, A)$  implies that we have  $d(\alpha(\omega), A) \geq \delta > 0$ , and in particular that  $\alpha(\omega) \notin A$ . By virtue of the Hahn-Banach separation theorem, there exists  $g_\omega \in (\mathbb{R}^2)^*$  such that  $A \subset \{(x,y) : g_\omega(x,y) \leq \max_{(s,t) \in A} g_\omega(s,t)\}$  and  $g_\omega(\alpha) > \max_{(s,t) \in A} g_\omega(s,t) + \epsilon$  for some  $\epsilon > 0$ . Let  $(f_\ell)_{\ell \in \mathbb{N}} \subset F$  be a sequence such that  $f_\ell \rightarrow g_\omega$  in the weak topology. By (2.16), we have  $A \subset \{(x,y) : f_\ell(x,y) \leq \xi_{f_\ell}\}$ , and for  $\ell$  large enough it is the case that  $f_\ell(\alpha) > \xi_{f_\ell} + 2\epsilon/3$ . Pick and fix now such an  $\ell$ . If it were the case that

$$(2.17) \quad \max_{(x,y) \in A_n(\omega)} f_\ell(x,y) \rightarrow \xi_{f_\ell},$$

then for large enough  $k$ ,

$$f_\ell(\alpha(\omega)) > \xi_{f_\ell} + 2\epsilon/3 > \max_{(x,y) \in A_{n_k}(\omega)} f_\ell(x,y) + \epsilon/3 \geq f_\ell(\alpha_{n_k}(\omega)) + \epsilon/3.$$

But this is a contradiction, since by continuity of  $f_\ell$ , we have  $f_\ell(\alpha_{n_k}(\omega)) \xrightarrow{k \rightarrow \infty} f_\ell(\alpha(\omega))$ . We conclude that for each  $\omega \in \mathcal{E}$  there exists  $f_\ell \in F$  for which (2.17) does not apply, and since  $P[\mathcal{E}] > 0$ , this contradicts (2.15).  $\square$

**Lemma 2.5.** *Let  $S, T$  be two symmetric scoring functions on  $\mathcal{A}^* \times \mathcal{A}^*$ . If the optimization problem  $\max_{(x,y) \in \text{SET}_{S,T}} x$  has a unique maximizer  $(x_0, y_0)$ , then*

$$(2.18) \quad \max_{\pi \in \nu_{n,S}^*} \frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} \xrightarrow{n \rightarrow \infty} y_0, \quad a.s.,$$

$$(2.19) \quad \min_{\pi \in \nu_{n,S}^*} \frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} \xrightarrow{n \rightarrow \infty} y_0, \quad a.s.$$

*Proof.* By virtue of (1.5) and Lemma 2.4,  $d_H(\text{SET}_{S,T}^n, \text{SET}_{S,T}) \rightarrow 0$  almost surely. Keeping in mind (1.10) and (1.11), taking any convergent subsequence  $((x_{n_\ell}, y_{n_\ell}))_{\ell \in \mathbb{N}}$  of a sequence  $((x_n, y_n))_{n \in \mathbb{N}}$  of maximizers

$$(2.20) \quad (x_n, y_n) \in \arg \max \left\{ y : (x, y) \in \text{SET}_{S,T}^n, x = \frac{L_{n,S}}{n} \right\},$$

and writing  $(x^*, y^*) = \lim_{\ell \rightarrow \infty} (x_{n_\ell}, y_{n_\ell})$ , we have  $x^* = x_0$  almost surely (by virtue of (1.5)), and  $(x^*, y^*) \in \text{SET}_{S,T}$  almost surely. By the assumptions of the lemma, we thus have  $(x^*, y^*) = (x_0, y_0)$ . Furthermore, a convergent subsequence of  $((x_n, y_n))_{n \in \mathbb{N}}$  always exists, since all sets  $\text{SET}_{S,T}^n$  are contained in a compact box, and the argument above shows that  $(x_0, y_0)$  is the only accumulation point. Therefore,  $(x_n, y_n) \rightarrow (x_0, y_0)$  almost surely, and since the choice of  $(x_n, y_n)$  among the maximizers of (2.20) was arbitrary, (2.18) and (2.19) both follow.  $\square$

**Lemma 2.6.** *Let  $K \subset \mathbb{R}^2$  be a deterministic convex compact set. Then the maximizer*

$$(x_0, y_0) = \arg \max_{(x,y) \in K} ax + by$$

*is unique for all but countable many points  $(a, b)$  on the unit sphere in  $\mathbb{R}^2$ . Furthermore, if  $(a, b)$  is chosen uniformly at random from the unit sphere in  $\mathbb{R}^2$ , then*

$$(2.21) \quad \mathbb{P} [\kappa(\partial K, (x_0, y_0)) \leq k] \leq \frac{k \cdot l}{2\pi},$$

*where  $\kappa(\partial K, (x_0, y_0))$  is the curvature of the boundary of  $K$  at the point  $(x_0, y_0)$ , and where  $l$  denotes the length of the boundary of  $K$ . Here  $P(\cdot)$  is the uniform probability on the unit sphere of  $\mathbb{R}^2$ .*

*Proof.* The first part of the lemma is well known. The mapping

$$H : (a, b) \mapsto (x_0, y_0) := \arg \max_{(x,y) \in K} ax + by$$

is thus well defined for all but a countable number of points  $(a, b)$  on the unit circle. If the interior of  $K$  is empty, then  $K$  lives on a line segment. The maximizer  $(x_0, y_0)$  is then one of the two endpoints of this segment for almost all  $(a, b)$ , and since the curvature is infinite at these points, the claim of the lemma is trivially true.

If  $K$  has nonempty interior, then its boundary  $\partial K$  is locally the graph of a convex function, and hence it is continuous. Given an interior point  $z \in K$  and  $r > 0$  such that  $B(z, r) \subset K$ , we define the spherical projection to be the map which assigns to any point  $u \in \partial K$ , the point on the circle  $\partial B(z, r)$  which is on the line joining  $z$  and  $u$ . The spherical projection with respect to an interior point defines a parametrization  $u(\theta)$  of  $\partial K$ , with  $\theta$  running on the unit circle. This can be used to show that the boundary of  $K$  is a locally the graph of a convex function.

Since the boundary  $\partial K$  is locally the graph of a convex function,  $u(\theta)$  is differentiable everywhere except at a countable number of points, and it is twice differentiable everywhere except on a set of Lebesgue measure 0, see e.g. [3, Theorem 1, page 242]. The length  $l = \int_0^{2\pi} \|du(\theta)/d\theta\|_2 d\theta$  is thus well defined and finite, and so is the reparametrization  $c(t)$  of  $u(\theta)$  with respect to the length  $t = \int_0^\theta \|du(\tau)/d\tau\|_2 d\tau$ . Furthermore, we have  $\|\dot{c}(t)\|_2 = 1$  for all  $t \in [0, l]$ , and  $\ddot{c}(t)$  is defined except on a Lebesgue-null set. Let  $A$  be the subset of  $t \in [0, l]$  where  $\dot{c}(t)$  is defined, and  $B$  the subset where  $\ddot{c}(t)$  is defined. Without loss of generality, we may assume that the orientation of the curve  $c(t)$  is positive, so that  $G(t) = i\dot{c}(t)$  is the unit normal vector to  $K$  at  $c(t)$  (orthogonal to  $\dot{c}(t)$  and pointing away from  $K$ ). This defines a mapping  $t \mapsto G(t)$  from  $A$  to the unit circle. We make the following two observations:

- (a)  $\kappa(t) := \kappa(\partial K, c(t)) = \|\ddot{c}(t)\|_2 = \|\dot{G}(t)\|_2$  equals the curvature of  $\partial K$  at  $c(t)$ .
- (b) Given  $(a, b)$  on the unit circle, if  $c(t) = \arg \max_{(x,y) \in K} ax + by$  for some  $t \in A$ , and if this is the unique maximizer, then  $(a, b) = G(t)$ .

Let  $T := \{t \in [0, l] : \kappa(t) \leq k\}$  and  $\tilde{P}(\cdot)$  denote the uniform probability on  $T$ . The fact that  $G(t)$  is defined at all points where  $\kappa(t)$  is defined combined with Observations (a)

and (b) imply that

$$\begin{aligned}
 (2.22) \quad \mathbb{P}[\kappa(\partial K, (x_0, y_0)) \leq k] &= \tilde{\mathbb{P}}[\kappa(t) \leq k] = \mathbb{P}[G(T)] \\
 &= \int_T |\dot{G}(t)| \frac{dt}{2\pi} \\
 &= \int_T k(t) \frac{dt}{2\pi}, \\
 (2.23) \quad &\leq \frac{k \cdot l}{2\pi}.
 \end{aligned}$$

Here  $G(T)$  is a subset of the unit sphere in  $\mathbb{R}^2$  and  $P(G(T))$  is its probability.

Equation (2.23) establishes the claim (2.21) of the lemma. The only nontrivial step that needs further explanation is (2.22). Let  $g : [0, l] \rightarrow [0, 2\pi]$  be such that  $G(t) = \exp(i g(t))$ . Then  $g(t)$  is well defined except at a countable number of points. By convexity of  $K$ ,  $g(t)$  is a non-decreasing function, and without loss of generality we may assume that it is right continuous. Equation (2.22) can thus be reformulated as follows,

$$(2.24) \quad \mu[g(T)] = \int_T \frac{\dot{g}(t)}{2\pi} dt,$$

where  $\mu$  is the uniform probability measure on the interval  $[0, 2\pi]$ . If  $g$  is smooth and increasing, then (2.24) is simply a change of variable formula. In the general case we can approximate using smooth functions. Thus take a standard mollifier  $\phi_\epsilon$  and  $g_{\epsilon, \delta} = (g + \delta h) \star \phi_\epsilon$  where  $h(x) = x$ . The rationale for taking  $g + \delta h$  is to render the derivative positive and  $g_\epsilon$  increasing. Equation (2.24) is true for  $g_{\epsilon, \delta}$ , and its general validity is obtained by first passing  $\epsilon$  to zero, followed by  $\delta$ .  $\square$

### 3. PROOFS OF THE MAIN THEOREMS

#### 3.1. Proof of Theorem 1.1.

*Proof.* With the notations from Lemma 2.3 combined with the results of Theorem 2.1, we know that for all large  $n$ ,

$$\mathbb{P}[A^n(R)] \geq 1 - n^{-\ln n}.$$

By the assumptions of the theorem,  $x_0 = \lambda_S$  and

$$(3.1) \quad \kappa(\partial \text{SET}_{S,T}, (x_0, y_0)) \geq k > 0.$$

For small  $\epsilon > 0$  the point  $P_\epsilon := (x_\epsilon, y_\epsilon)$  on the boundary  $\partial \text{SET}_{S,T}$  with  $y$ -coordinate  $y_\epsilon := y_0 + \epsilon/k$  nearest to  $(x_0, y_0)$  is well defined. For the main objects which appear in this proof below we refer to Figure 1. Choose  $a_\epsilon$  such that the linear form  $f_{(1,a_\epsilon)} : (x, y) \mapsto x + a_\epsilon y$  has its maximizer over the set  $\text{SET}_{S,T}$  at  $P_\epsilon$ . The existence of  $a_\epsilon$  follows from the convexity of  $\text{SET}_{S,T}$  and the continuity of the boundary of  $\text{SET}_{S,T}$ .

This implies that for any  $(x, y) \in \text{SET}_{S,T}$ ,

$$(3.2) \quad x + a_\epsilon y \leq x_\epsilon + a_\epsilon y_\epsilon.$$

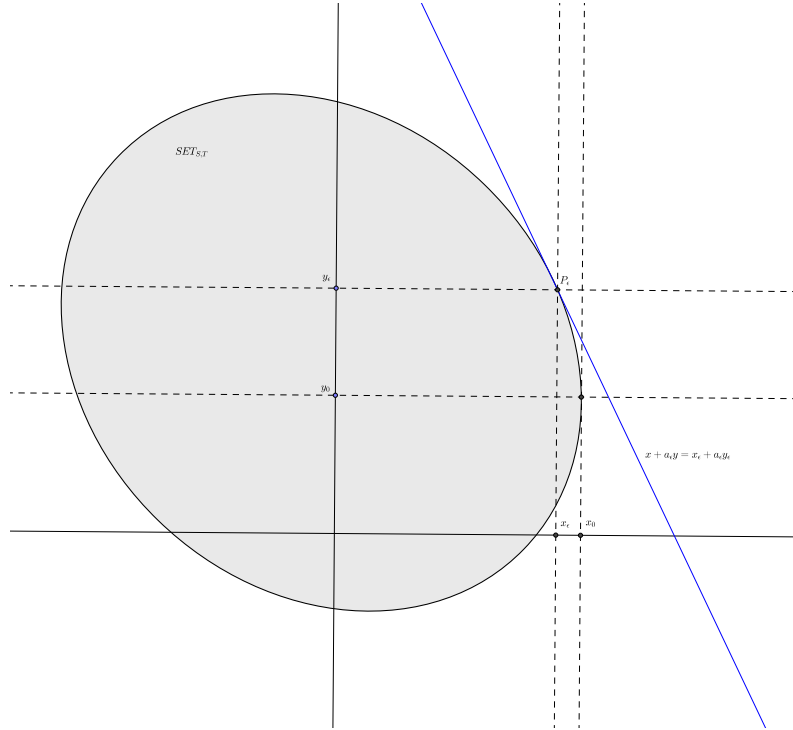


FIGURE 1. The shaded area is  $\text{SET}_{S,T}$  such that  $(x_0, y_0)$  is the maximizer of the functional  $f(x, y) = x$ . The line in blue is a line which passes through  $P_\epsilon$  and leaves  $\text{SET}_{S,T}$  on one side of it.

The curvature condition (3.1) implies that for all  $\epsilon$  small enough,

$$x_\epsilon \leq x_0 - \frac{\epsilon^2}{3}.$$

Combined with (3.2) for  $(x, y) = (x_0, y_0)$ , this yields  $(x - x_\epsilon) + a_\epsilon(y - y_\epsilon) \leq 0$ , and since furthermore  $x_0 > x_\epsilon$ , it follows that

$$(3.3) \quad a_\epsilon \geq \frac{x_0 - x_\epsilon}{y_\epsilon - y_0} \geq \frac{\epsilon k}{3}.$$

If  $A^n(S)$  holds, then for any optimal alignment  $\pi$  relative to  $S$  we have

$$(3.4) \quad \left| \frac{S_\pi(X_{[1,n]}, Y_{[1,n]})}{n} - x_0 \right| \leq \frac{5|S|\sqrt{\ln(en)}}{\sqrt{n}},$$

and similarly, if the event  $A^n(S + a_\epsilon T)$  holds, then

$$(3.5) \quad \left| \frac{L_{n,S+a_\epsilon T}}{n} - \lambda_{S+a_\epsilon T} \right| \leq \frac{5|S|\sqrt{\ln(en)}}{\sqrt{n}}.$$

On the other hand,

$$\lambda_{S+a_\epsilon T} = \max_{(x,y) \in \text{SET}_{S,T}} f_{(1,a_\epsilon)}(x, y) = x_\epsilon + a_\epsilon y_\epsilon,$$

and substituted into (3.5) this yields

$$(3.6) \quad \left| \frac{L_{n,S+a_\epsilon T}}{n} - (x_\epsilon + a_\epsilon y_\epsilon) \right| \leq \frac{5|S + a_\epsilon T|\sqrt{\ln(en)}}{\sqrt{n}}.$$

Next, for any optimal alignment  $\pi$  relative to  $S$ , we have

$$\frac{(S + a_\epsilon T)_\pi(X_{[1,n]}, Y_{[1,n]})}{n} \leq \frac{L_{n,S+a_\epsilon T}}{n} \stackrel{(3.6)}{\leq} x_\epsilon + a_\epsilon y_\epsilon + \frac{5|S + a_\epsilon T|\sqrt{\ln(en)}}{\sqrt{n}}.$$

It now follows from (3.4) that

$$a_\epsilon \left( \frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} - y_0 \right) \leq x_\epsilon - x_0 + a_\epsilon(y_\epsilon - y_0) + \frac{5(|S + a_\epsilon T| + |S|)\sqrt{\ln(en)}}{\sqrt{n}},$$

and since  $x_\epsilon - x_0 \leq 0 < a_\epsilon$ , this finally yields that for large  $n$  and small  $\epsilon > 0$ ,

$$\frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} - y_0 \leq \frac{\epsilon}{k} + \frac{5(|S + a_\epsilon T| + |S|)\sqrt{\ln(en)}}{a_\epsilon \sqrt{n}} \leq \frac{5(2|S| + a_\epsilon |T|)\sqrt{\ln(en)}}{a_\epsilon \sqrt{n}}.$$

In combination with (3.3) this yields

$$\frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} - y_0 \leq \frac{\epsilon}{k} + \frac{5(6|S| + \epsilon k|T|)}{\epsilon k} \sqrt{\frac{\ln(en)}{n}}.$$

For large  $n$  (depending on the geometry of the boundary of  $SET_{S,T}$  near  $(x_0, y_0)$ ), we can minimize the right-hand side over  $\epsilon$  (the minimizing value being  $\epsilon = \sqrt{30|S|\sqrt{\ln(n)/n}}$ ), yielding for large  $n$  that

$$\frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} - y_0 \leq \frac{5|T| + 2\sqrt{30|S|}}{k} \left( \frac{\ln(en)}{n} \right)^{1/4}$$

By changing the scoring function  $T$  to  $-T$ , an analogous argument also shows that

$$-\left( \frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} - y_0 \right) \leq \frac{5|T| + 2\sqrt{30|S|}}{k} \left( \frac{\ln(en)}{n} \right)^{1/4},$$

and hence,

$$(3.7) \quad \left| \frac{T_\pi(X_{[1,n]}, Y_{[1,n]})}{n} - y_0 \right| \leq \frac{5|T| + 2\sqrt{30|S|}}{k} \left( \frac{\ln(en)}{n} \right)^{1/4}.$$

We conclude that if all of the events  $A^n(S)$  and  $A^n(S + a_\epsilon T)$  and  $A^n(S - a_\epsilon T)$  hold, then (3.7) applies, and since the probability that any individual event fails to hold is bounded by  $n^{-\ln n}$ , the claim of the theorem follows.  $\square$

### 3.2. Proof of Theorem 1.2.

*Proof.* Let  $V = \arccos\langle S, T \rangle_F$ , where  $\langle \cdot, \cdot \rangle_F$  is the inner product on the space of symmetric scoring functions that corresponds to the Frobenius norm. Then  $V$  is uniformly distributed on  $[-\pi/2, \pi/2]$ . Let  $T_1$  be the Gram-Schmidt orthogonalization of  $T$  with respect to  $S_1 := S$ , and let  $U$  be a uniform random variable on  $[0, 2\pi]$ , independent of  $S$  and  $T$ , and hence also of  $V$ , and let us define  $(S_2, T_2) = \Phi(S_1, T_1)$ , where  $\Phi$  is the rotation

$$\begin{aligned} \Phi : \mathbb{R}^2 &\rightarrow \mathbb{R}^2, \\ (x, y) &\mapsto (\cos(U)x + \sin(U)y, -\sin(U)x + \cos(U)y) \end{aligned}$$

by the angle  $U$ . It is easy to see that  $SET_{S_2, T_2} = \Phi(SET_{S_1, T_1})$ , and that under  $\Phi^{-1}$ , the point where  $SET_{S_2, T_2}$  has a point of maximal first coordinate corresponds to the point where the random linear form  $f : (x, y) \mapsto \cos(U)x + \sin(U)y$  takes a maximum value

on  $\text{SET}_{S_1, T_1}$ . Furthermore, since  $\Phi$  is angle-preserving, the curvature  $\kappa_1$  of  $\partial\text{SET}_{S_2, T_2}$  and  $\partial\text{SET}_{S_1, T_1}$  at these points is also the same. Lemma 2.6 applies, and we have  $P[\kappa_1 \leq k] \leq k \cdot l / (2\pi)$ , where  $l$  is the length of the boundary of  $\text{SET}_{S_1, T_1}$ . Since the scoring functions under considerations have unit norm, the rescaled alignment score cannot exceed 2, implying that  $l \leq 8$  and

$$(3.8) \quad P[\kappa_1 \leq k] \leq \frac{4k}{\pi}.$$

It remains to relate  $\kappa_1$  to the curvature  $\kappa$  of  $\partial\text{SET}_{S, T}$  at the point where its first coordinate is maximized. Since  $\text{SET}_{S, T} = \Psi(\text{SET}_{S_1, T_1})$ , where  $\Psi$  is the linear transformation

$$\begin{aligned} \Psi : \mathbb{R}^2 &\rightarrow \mathbb{R}^2, \\ (x, y) &\mapsto (x, \cos(V)x + \sin(V)y), \end{aligned}$$

we have  $\kappa = \kappa_1 / |\sin V| \geq \kappa_1$ , so that

$$P[\kappa < k] \leq P[\kappa_1 < k] \leq \frac{4k}{\pi},$$

as claimed in the statement of the theorem.  $\square$

#### 4. ACKNOWLEDGEMENTS

We would like to thank the reviewers for their careful reading of the paper. Their suggestions and corrections improved this paper significantly.

#### REFERENCES

- [1] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Math. J.*, 19:357–367, 1967.
- [2] Václav Chvatal and David Sankoff. Longest common subsequences of two random sequences. *J. Appl. Probability*, 12:306–315, 1975.
- [3] Lawrence C. Evans and Ronald F. Gariepy. *Measure theory and fine properties of functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.
- [4] Raphael Hauser and Heinrich Matzinger. Distribution of aligned letter pairs in optimal alignments of random sequences. *arXiv:1211.5491*, 2013.
- [5] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

- [6] M.S. Waterman and M. Vingron. Sequence comparison significance and poisson approximation. *Statistical Science*, 9(3):367–381, 1994.
- [7] J. Lember, and H. Matzinger. Detecting the homology of DNA-sequences based on the variety of optimal alignments: a case study. *arXiv:1210.3771 [stat.AP]*, 2012.
- [8] H. Kesten, editor. Probability on discrete structures, *volume 110 of Encyclopaedia of Mathematical Sciences*, Berlin, 2004. Springer-Verlag.

RAPHAEL HAUSER, MATHEMATICAL INSTITUTE, UNIVERSITY OF OXFORD, RADCLIFFE OBSERVATORY QUARTER, WOODSTOCK ROAD, OXFORD OX2 6GG, UNITED KINGDOM, AND PEMBROKE COLLEGE, ST ALDATES, OXFORD, OX1 1DW, UNITED KINGDOM.

*E-mail address:* hauser@maths.ox.ac.uk

HEINRICH MATZINGER, SCHOOL OF MATHEMATICS, GEORGIA INSTITUTE OF TECHNOLOGY, 686 CHERRY STREET, ATLANTA, GA 30332-0160 USA. CORRESPONDING AUTHOR.

*E-mail address:* matzi@math.gatech.edu

IONEL POPESCU, SCHOOL OF MATHEMATICS, GEORGIA INSTITUTE OF TECHNOLOGY, 686 CHERRY STREET, ATLANTA, GA 30332, USA, AND “SIMION STOILOW” INSTITUTE OF MATHEMATICS OF ROMANIAN ACADEMY, 21 CALEA GRIVIȚEI, BUCHAREST, ROMANIA.

*E-mail address:* ipopescu@math.gatech.edu, ionel.popescu@imar.ro