The rate of the convergence of the mean score in random sequence comparison

J. Lember, H. Matzinger, F. Torres

November 12, 2010

Jüri Lember, Tartu University, Institute of Mathematical Statistics
Liivi 2-513 50409, Tartu, Estonia. E-mail address: jyril@ut.ee
Heinrich Matzinger, Georgia Tech, School of Mathematics
Atlanta, Georgia 30332-0160, U.S.A. E-mail address: matzing@math.gatech.edu
Felipe Torres, University of Bielefeld, Faculty of Mathematics
Postfach 100131 - 33501 Bielefeld, Germany. E-mail address: ftorres@math.uni-bielefeld.de

Abstract

We consider a general class of superadditive scores measuring the similarity of two independent sequences of \( n \) i.i.d. letters from a finite alphabet. Our object of interest is the mean score by letter \( l_n \). By the subadditivity \( l_n \) is nondecreasing and converges to a limit \( l \). We give a simple method of bounding the difference \( l - l_n \) and obtaining the rate of convergence. Our result generalizes the previous result of Alexander [1], where only the special case of the longest common subsequence is considered.

Keywords. Random sequence comparison, longest common sequence, rate of convergence.

AMS. 60K35, 41A25, 60C05

1 Introduction

Throughout this paper \( X_1, X_2, \ldots \) and \( Y_1, Y_2, \ldots \) are two independent sequences of i.i.d. random variables drawn from a finite alphabet \( A \). We do not assume that \( X_1 \) and \( Y_1 \) have the same distribution. Since we mostly study the finite strings of length \( n \), let \( X = (X_1, X_2, \ldots X_n) \) and let \( Y = (Y_1, Y_2, \ldots Y_n) \) be the corresponding \( n \)-dimensional random vectors. We shall usually refer to \( X \) and \( Y \) as random
The problem of measuring the similarity of $X$ and $Y$ is central in many areas of applications including computational molecular biology [7, 14, 23, 24, 25] and computational linguistics [8, 18, 20, 21]. In this paper, we consider a general scoring scheme, where $S : A \times A \rightarrow \mathbb{R}^+$ is a pairwise scoring function that assigns a score to each couple of letters from $A$. We assume $S$ to be symmetric and we denote by $F$ and $A$ the largest possible score and the largest possible change of score by one variable, respectively. Formally

$$F := \max_{a, b \in A} S(a, b), \quad A := \max_{a, b, c \in A} |S(a, b) - S(a, c)|.$$  

In the definition of $A$ recall that $S$ is symmetric.

An alignment is a pair $(\pi, \mu)$ where $\pi = (\pi_1, \pi_2, \ldots, \pi_k)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ are two increasing sequences of natural numbers, i.e. $1 \leq \pi_1 < \pi_2 < \ldots < \pi_k \leq n$ and $1 \leq \mu_1 < \mu_2 < \ldots < \mu_k \leq n$. The integer $k$ is the number of aligned letters and $n - k$ is the number of gaps in the alignment. We also consider a gap price $\delta$. With pairwise scoring function $S$ and the gap price $\delta$, the score of the alignment $(\pi, \mu)$ when aligning $X$ and $Y$ is

$$U_{(\pi, \mu)}(X, Y) := \sum_{i=1}^{k} S(X_{\pi_i}, Y_{\mu_i}) + \delta(k - n).$$

The (optimal) score of $X$ and $Y$ is defined to be best score over all possible alignments, i.e.

$$L_n := L(X, Y) := \max_{(\pi, \mu)} U_{(\pi, \mu)}(X, Y),$$

The alignments achieving the maximum are called optimal. Such a similarity criterion is most commonly used in sequence comparison [3, 14, 24, 25, 26]. Usually the gap price is negative and in this case it will be called gap penalty. We naturally assume $\delta \leq F$. When $S(a, b) = 1$ for $a = b$ and $S(a, b) = 0$ for $a \neq b$, then for $\delta = 0$ the optimal score is equal to the length of the longest common subsequence (LCS) of $X$ and $Y$.

It is well-known that the sequence $EL_n$, $n = 1, 2 \ldots$ is superadditive, i.e. $EL_{n+m} \geq EL_n + EL_m$ for all $n, m \geq 1$. Hence, by Fekete’s lemma the ratios $l_n := \frac{EL_n}{n}$ are nondecreasing and converge to the limit

$$l := \lim_{n} l_n = \sup l_n.$$  

In fact, from Kingman’s subadditivity ergodic theorem, it follows that $l$ is also the a.s. limit of $\frac{L_n}{n}$. The limit $l$ (which for the LCS-case is called Chvatal-Sankov constant) is not known exactly even for the simplest scoring scheme and the simplest model for $X$ and $Y$, so it is usually estimated by simulations. Using McDiarmid’s inequality (see (3.6)) one can estimates the constant $L_n$ with prescribed preciseness; to obtain confidence intervals for $l$, the difference $l - L_n$ should be estimated. This is the aim of the present paper.

To our best knowledge, the difference $l - L_n$ has been theoretically studied only by
Alexander in [1], though there exist many numeric results on the value of $l_n$ or its
distribution in various contexts [4, 6, 9, 11, 12, 15, 16, 17, 22]. Alexander proved
that in the case of the LCS, for any $C > (2 + \sqrt{2})$ there exists integer $n_o(C)$ such that

$$l - l_n \leq C \sqrt{\frac{\log n}{n}}, \quad \text{provided } n > n_o(C).$$

(1.1)

The bound (1.1) is independent of the common law of $X$ and $Y$, and the integer
$n_o(C)$ can be exactly determined. Hence the bound (1.1) can be used for the cal-
culation of the explicit confidence intervals.

Our main result is the following:

**Theorem 1.1** Let $n \in \mathbb{N}$ be even. Then, with any $c > \sqrt{A}$,

$$l - l_n \leq c \sqrt{\frac{2}{n-1} \left( \frac{n+1}{n-1} + \ln(n-1) \right) + \frac{F}{n-1}}.$$

(1.2)

Note that by the monotonicity of $l_n$, the assumption on even $n$ is not actually
restrictive. In fact, Alexander’s main result (Prop. 2.4 in [1]) is also proven for even
$n$. Theorem 1.1 and its proof generalize Alexander’s result in many ways:

1. Theorem 1.1 applies for a general scoring scheme, not just for the LCS. This is
due to the fact that our proof is based solely on McDiarmid’s large deviation
equality, whilst Alexander’s proof, although using also McDiarmid’s inequality,
is mainly based on first passage percolation techniques. Despite the fact that
the percolation approach applies in many other situations rather than sequence
comparison (see [2]), it is not clear whether it can be efficiently applied to
our general scoring scheme. For McDiarmid’s inequality, however, makes no
difference what kind of scoring is used. This gives us reasons to believe that
our proof is somehow ”easier” than the one in [1].

2. The proof of Theorem 1.1 relates the rate of the convergence of $l_n$ to the
cardinality of the set of partitions $\mathcal{B}_{k,n}$ (see Lemma 3.1) so that finding the good
rate boils down to the good estimation of $|\mathcal{B}_{k,n}|$. The bound (1.2) corresponds
to a particular estimate of $|\mathcal{B}_{k,n}|$, any better estimate would give a sharper
bound and, probably, also a faster rate. In a sense, the cardinality $|\mathcal{B}_{k,n}|$ could
be interpreted as the complexity of the model and the relation between the
rate of convergence and the complexity of the model is well-known in statistics
(see e.g. [5]).

3. When applied to the LCS, our bound (1.2) is sharper than (1.1). Indeed, for
the case of LCS the constants $A$ and $F$ in (1.2) can be taken equal to one and
the smaller constants make the difference. In other words, for the case of LCS
both results yield the rate $C \sqrt{\frac{\ln n}{n}}$, but the constant $C$ is different ($C > 3.42$
in Alexander’s result and $c > \sqrt{2}$ in ours).
We can easily compare (1.2) and (1.1) by comparing the decay of the two following functions:

\( R : \{1, \ldots, 10000\} \to \mathbb{R}^+ \)

\[ R(n) = (3.42 + 0.1) \sqrt{\frac{\ln n}{n}} \quad (1.3) \]

\( Q_F : \{1, \ldots, 10000\} \times \{0.1, \ldots, 2\} \to \mathbb{R}^+ \)

\[ Q_F(n, A) = \sqrt{A} \sqrt{\frac{2}{n-1} \left( \frac{n+1}{n-1} + \ln(n-1) \right)} + \frac{F}{n-1} \quad (1.4) \]

In figure 1, we can see the improved bound (1.2) given by function (1.4) (changes of \( A \) are represented in colours, \( F = 1 \)) over the bound by Alexander (1.1) given by function (1.3) (in black). Note that the dark blue curve corresponds to \( A = 0.1 \) whilst the light violet curve to \( A = 2 \) (namely, the colour gets lighter as \( A \) increases). The curve in green corresponds to the case \( A = 1 \) (ie, our bound for the LCS case).

![Figure 1: comparison of the bounds (1.2) and (1.1) through the functions (1.4) and (1.3), respectively.](image)

2 Confidence bounds for \( l \)

Suppose that \( N \) samples of \( X^i = X^i_1, \ldots, X^i_n \) and \( Y^i = Y^i_1, \ldots, Y^i_n, i = 1, \ldots, N \) are generated. Let \( L^i_n \) be the score of \( i \)-th sample. Thus \( E L^i_n = n l_n \). By McDiarmid’s
inequality (see (3.5) below), for every $\rho > 0$

$$P\left(\frac{1}{kn} \sum_{i=1}^{N} L_{n}^{i} - l < -\rho\right) = P\left(\sum_{i=1}^{N} L_{n}^{i} - kn l < -kn\rho\right) \leq \exp\left[-\frac{\rho^2 k n}{A^2}\right]$$  \hspace{1cm} (2.1)

Let

$$\bar{L}_{n} := \frac{1}{kn} \sum_{i=1}^{N} L_{n}^{i}.$$  

If $n$ is even, by (1.2) and (1.4) we have that $l \leq l_{n} + Q_{F}(n, A)$ and then

$$P(\bar{L}_{n} + \rho + Q_{F}(n, A) \geq l) \geq P(\bar{L}_{n} + \rho \geq l_{n}) = P(\bar{L}_{n} - l \geq -\rho) \geq 1 - \exp\left[-\frac{\rho^2 k n}{A^2}\right]$$  \hspace{1cm} (2.2)

Now, given $\varepsilon > 0$, choose $\rho = \rho(n, \varepsilon)$ so that the right hand side in the last inequality is equal to $1 - \varepsilon$:

$$\rho(n, \varepsilon) = A\sqrt{\frac{\ln(1/\varepsilon)}{kn}}.$$  

So, with probability $1 - \varepsilon$, we obtain one side confidence interval as follows:

$$l \leq \bar{L}_{n} + Q_{F}(n, A) + A\sqrt{\frac{\ln(1/\varepsilon)}{kn}}.$$  \hspace{1cm} (2.3)

In statistical learning, the inequalities of type (2.3) are known as PAC inequality (probably almost correct inequalities). The two-sided confidence bounds are, with probability $1 - \varepsilon$, as follows:

$$\bar{L}_{n} - A\sqrt{\frac{\ln(1/\varepsilon)}{kn}} \leq l \leq \bar{L}_{n} + Q_{F}(n, A) + A\sqrt{\frac{\ln(1/\varepsilon)}{kn}}.$$  \hspace{1cm} (2.4)

The bounds in (2.4) suggest to use the estimate

$$\hat{i}_{n} := \bar{L}_{n} + \frac{Q_{F}(n, A)}{2}$$  

so that the confidence bounds for this estimate are

$$P\left(|\hat{i}_{n} - l| \leq A\sqrt{\frac{\ln(1/\varepsilon)}{kn}} + \frac{Q_{F}(n, A)}{2}\right) \geq 1 - \varepsilon$$  \hspace{1cm} (2.5)

Alexander [1] obtained, for $n = 100000$, $k = 2$ and $A = F = 1$ (for the LCS case), the following bounds:

$$P(|\hat{i}_{n} - l| \leq 0.0264) \geq 0.95.$$  \hspace{1cm} (2.6)

By using (2.5) and (1.4) we obtain, for $n = 100000$, $k = 2$ and $A = F = 1$ (for the LCS case), the following bounds:

$$P(|\hat{i}_{n} - l| \leq 0.0119) \geq 0.95.$$  \hspace{1cm} (2.7)
It is clear that (2.7) is sharper than (2.6). To our best knowledge, the best previous lower and upper bounds for \( l \), in the LCS context for \( A = \{0, 1\} \), were due to Dancik [10], Dancik and Paterson [11, 22] (0.773911 and 0.837623, respectively) and Lueker [19] (0.788071 and 0.826280, respectively).

**Remark:** The inequality (2.3) confirms the well-known fact that it is better to generate one sample of length \( kn \) rather than \( k \) samples of length \( n \). Indeed, with one sample of length \( kn \), the inequality (2.3) becomes

\[
l \leq L_n + Q_F(kn, A) + A \sqrt{\frac{\ln(1/\varepsilon)}{kn}}
\]

and since \( Q_F(kn, A) < Q_F(n, A) \), the bounds get narrower.

### 3 Proof of the main result

#### 3.1 The set of partitions \( \mathcal{B}_{k,n} \)

In this section, we shall consider the sequences \( X \) and \( Y \) with length \( kn \) where \( k, n \) are nonnegative integers. Let \((\pi, \mu)\) be an arbitrary alignment of \( X \) and \( Y \). Let \( \nu = (\nu_1, \ldots, \nu_{r+1}) \) and \( \tau = (\tau_1, \ldots, \tau_{r+1}) \) be vectors satisfying \( 1 = \nu_1 \leq \nu_2 \leq \cdots \leq \nu_r \leq \nu_{r+1} = kn + 1 \) and \( 1 = \tau_1 \leq \tau_2 \leq \cdots \leq \tau_r \leq \tau_{r+1} = kn + 1 \). We say that the pair \((\nu, \tau)\) forms a \( r \)-partition of the alignment \((\pi, \mu)\) if for any \( j = 1, \ldots, r \), the following conditions simultaneously hold:

1) if, for some \( i = 1, \ldots, k \), it holds that \( \nu_j \leq \pi_i < \nu_{j+1} \), then \( \tau_j \leq \mu_i < \tau_{j+1} \);
2) if, for some \( i = 1, \ldots, k \), it holds that \( \tau_j \leq \mu_i < \tau_{j+1} \), then \( \nu_j \leq \pi_i < \nu_{j+1} \).

Thus \((\nu, \tau)\) is a \( r \)-partition, if the sequences \( X \) and \( Y \) can be partitioned into \( r \) pieces

\[
(X_1, \ldots, X_{\nu_2-1}), (X_{\nu_2}, \ldots, X_{\nu_3-1}), \ldots, (X_{\nu_r}, \ldots, X_{kn})
\]
\[
(Y_1, \ldots, Y_{\tau_2-1}), (Y_{\tau_2}, \ldots, Y_{\tau_3-1}), \ldots, (Y_{\tau_r}, \ldots, Y_{kn})
\]

such that the alignment \((\pi, \mu)\) aligns a piece \((X_{\nu_j}, \ldots, X_{\nu_{j+1}-1})\) with the piece \((Y_{\tau_j}, \ldots, Y_{\tau_{j+1}-1})\), where \( j = 1, \ldots, r \). It is important to note that the pieces might be empty, i.e. it might be that \( \nu_j = \nu_{j+1} \) (or \( \tau_j = \tau_{j+1} \)). It then means that \((\tau_j, \ldots, \tau_{j+1} - 1)\) cannot contain any elements of \( \mu \), otherwise the requirement 2) would be violated (or \((\mu_j, \ldots, \mu_{j+1} - 1)\) cannot contain any elements of \( \tau \), otherwise the requirement 1) would be violated). Hence, if for a partition a piece of \( X \) is empty, then the corresponding piece of \( Y \) must be totally aligned with gaps.

The following observation shows that any alignment of \( X \) and \( Y \) can be partitioned into \( r \) pieces such that \( k \leq r \leq \lceil \frac{2kn}{2n-1} \rceil \) and such that the sum of the lengths of aligned pairs in each partition is always at most \( 2n \). We believe that the proof as well as the meaning of the partition becomes transparent by an example.
Example. Let $n = 3, k = 4$. Let $\pi = (1, 5, 6, 9, 10, 12)$ and $\mu = (2, 3, 4, 6, 9, 10)$. The alignment $(\pi, \mu)$ can be represented as follows

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Y</td>
<td>1</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

The table above indicates that $X_1$ is aligned with $Y_2$, $X_5$ is aligned with $Y_3$ and so on; the rest of the letters are unaligned, so we say that they are aligned with gaps. In the table, there are two types of columns: the columns with two figures (aligned pairs) and the columns with one figure (unaligned pairs). Let $u_i \in \{1, 2\}$ be the number of figures in the $i$-th column, and let $s_j = u_1 + \cdots + u_j$ be the corresponding cumulative sum. To get an $r$-partition proceed as follows: start from the beginning of the table (most left position) and find $j$ such that $s_j = 2n$. Since the cumulative sum increases by one or two, such a $j$ might not exist. In this case find $j$ such that $s_j = 2n - 1$. In the present example $n = 3$, thus we are looking for $j$ such that $s_j = 6$. Such a $j$ is 5. The first five columns thus form the first part of the partition and there are exactly $2n = 6$ elements in the first part. Now disregard the first five columns from the table and start the same procedure afresh. Then the second part is obtained and so on. In the following table the vertical lines indicate the different parts obtained by the aforementioned procedure: the first two parts have six elements, the third and fourth has five elements and the last part consists of one element:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Y</td>
<td>1</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

From the table, we read the corresponding pieces from the $X$-side: $(1, 4), (5, 8), (9, 9), (10, 12), \emptyset$ as well as the ones from the $Y$-side: $(1, 2), (3, 4), (5, 8), (9, 11), (12, 12)$. The corresponding vectors $\nu$ and $\tau$ are thus $\nu = (1, 5, 9, 10, 13, 13), \tau = (1, 3, 5, 9, 13)$. The number of parts in such a partition is clearly at least $k$ (corresponding to the case that all pairs sum up to $2n$) and at most $\left\lceil \frac{2kn}{2n-1} \right\rceil$ (corresponding to the case that all pairs except the last one sum up to $2n - 1$). In our example is $r = 5 = \left\lceil \frac{24}{7} \right\rceil$. Now, it is clear that the following claim holds.

Claim 3.1 Let $X, Y$ be sequences of length $kn$ and let $(\pi, \mu)$ be an arbitrary alignment of $X$ and $Y$. Then there exist an integer $r$ such that $k \leq r \leq \left\lfloor \frac{2kn}{2n-1} \right\rfloor$ and an $r$-partition $(\nu, \tau)$ of $(\pi, \mu)$ such that for every $j = 1, \ldots, r - 1$, it holds

$$
(\nu_{j+1} - \nu_j) + (\tau_{j+1} - \tau_j) \leq (2n, 2n - 1) \quad \text{and} \quad (\nu_{r+1} - \nu_r) + (\tau_{r+1} - \tau_r) \leq 2n. \quad (3.1)
$$

Let, for every $r$, $B^r_{k,n}$ be the set of all $r$-partitions satisfying (3.1). Let

$$
B_{k,n} = \bigcup_{r=k}^{\left\lfloor \frac{2kn}{2n-1} \right\rfloor} B^r_{k,n}.
$$

If $(\pi, \mu)$ is optimal for $X,Y$ and $(\nu, \tau)$ is a $r$-partition of $(\pi, \mu)$, then obviously

$$
L_{kn}(\nu, \tau) := \sum_{i=1}^{r} L(X_{\nu_j}, \ldots, X_{\nu_{j+1}-1}; Y_{\tau_j}, \ldots, Y_{\tau_{j+1}-1}), \quad (3.2)
$$

7
where \( L(X_{\nu_1}, \ldots, X_{\nu_{j+1}-1}; Y_{\tau_j}, \ldots, X_{\tau_{j+1}-1}) \) is the optimal score between \( X_{\nu_1}, \ldots, X_{\nu_{j+1}} \) and \( Y_{\tau_j}, \ldots, X_{\tau_{j+1}} \). Note that the number of the members in the sum (3.2) might be actually smaller than \( r \). By Claim 3.1, every alignment has at least one partition from the set \( \mathcal{B}_{k,n} \), hence from (3.2) it follows that
\[
L_{kn} = \max_{(\nu, \tau) \in \mathcal{B}_{k,n}} L_{kn}(\nu, \tau). \tag{3.3}
\]

**Claim 3.2** For every \( r \)-partition \((\nu, \tau) \in \mathcal{B}_{k,n}\),

\[
E(L_{kn}(\nu, \tau)) \leq \frac{r}{2} EL_2n \leq \frac{1}{2} \left[ \frac{2kn}{2n - 1} \right] EL_2n. \tag{3.4}
\]

**Proof.** Let \((\nu, \tau) \in \mathcal{B}_{k,n}^r\) with \( r \leq \left\lfloor \frac{2nk}{2n-1} \right\rfloor \). Let \( j \) be such that \((\nu_{j+1} - \nu_j) + (\tau_{j+1} - \tau_j) = 2n\). Thus, there exists an integer \( u \in \{-n, \ldots, n\} \) such that \( \nu_{j+1} - \nu_j = n - u \) and \( \tau_{j+1} - \tau_j = n + u \). Since \( X_1, X_2, \ldots, Y_1, Y_2, \ldots \) are i.i.d., we have
\[
E(L(X_{\nu_1}, \ldots, X_{\nu_{j+1}-1}; Y_{\tau_j}, \ldots, X_{\tau_{j+1}-1})) = E(L(X_1, \ldots, X_{n-u}; Y_1, \ldots, Y_{n+u})) = E(L(X_{n-u+1}, \ldots, X_{2n}; Y_{n+u+1}, \ldots, Y_{2n})) \leq \frac{1}{2} E(L(X_1, \ldots, X_{2n}; Y_1, \ldots, Y_{2n})) = \frac{1}{2} EL_2n.
\]
The last inequality follows from the superadditivity:
\[
L(X_1, \ldots, X_{n-u}; Y_1, \ldots, Y_{n+u}) + L(X_{n-u+1}, \ldots, X_{2n}; Y_{n+u+1}, \ldots, Y_{2n}) \leq L(X_1, \ldots, X_{2n}; Y_1, \ldots, Y_{2n}).
\]
If \((\nu_{j+1} - \nu_j) + (\tau_{j+1} - \tau_j) < 2n\), then by the same argument
\[
E(L(X_{\nu_1}, \ldots, X_{\nu_{j+1}-1}; Y_{\tau_j}, \ldots, X_{\tau_{j+1}-1})) \leq E(L(X_1, \ldots, X_{n-u}; Y_1, \ldots, Y_{n+u})) \leq \frac{1}{2} EL_2n.
\]
Hence the first inequality in (3.4) follows. The second inequality follows from the condition \( r \leq \left\lfloor \frac{2nk}{2n-1} \right\rfloor \). \( \blacksquare \)

### 3.2 The size of \( \mathcal{B}_{k,n} \) and the rate of convergence

In the following we prove the main theoretical result that links the rate of the convergence to the rate at which the number of elements in \(|\mathcal{B}_{k,n}|\) grows as \( k \) increases. Our proof is entirely based on McDiarmid’s inequality, so let us recall it for the sake of completeness: Let \( Z_1, \ldots, Z_{2kn} \) be independent random variables and \( f(Z_1, \ldots, Z_{2kn}) \) be a function so that changing one variable changes the value at most \( A \). Then for any \( \Delta > 0 \),
\[
P\left(f(Z_1, \ldots, Z_{2kn}) - Ef(Z_1, \ldots, Z_{2kn}) > \Delta \right) \leq \exp\left[-\frac{\Delta^2}{knA^2}\right]. \tag{3.5}
\]
For the proof, we refer [13]. We apply (3.5) with \( L \) in the role of \( f \) to the independent (but not necessarily identically distributed) random variables \( X_1, \ldots, X_{kn}, Y_1, \ldots, Y_{kn} \). It is easy to see that independently of the value of \( \delta \), changing one random variable changes the score at most by \( A \) so that in our case (3.5) is
\[
P\left(L_n - EL_n > \Delta \right) \leq \exp\left[-\frac{\Delta^2}{knA^2}\right]. \tag{3.6}
\]
Lemma 3.1 Suppose that for any \( n \) and for \( k \) big enough
\[
|B_{k,n}| \leq \exp[(\psi(n) + o(k))kn],
\] (3.7)
where \( \psi(n) \) does not depend on \( k \). Let \( u(n) > A\sqrt{\psi(n)} \). Then
\[
l - l_{2n} \leq u(n) + \frac{l_{2n}}{2n-1} \leq u(n) + \frac{l}{2n-1} \leq u(n) + \frac{F}{2n-1}.
\] (3.8)

Proof. Let \( (\nu, \tau) \in B_{k,n} \). Recall (3.4). Thus, from (3.6), we get that for any \( \rho > 0 \),
\[
P\left( L_{kn}(\nu, \tau) - \frac{1}{2} \left[ \frac{2kn}{2n-1} \right] EL_{2n} > \rho kn \right) \leq P\left( L_{kn}(\nu, \tau) - \frac{1}{2} \left[ \frac{2kn}{2n-1} \right] EL_{2n} > \rho kn \right) \leq \exp\left[ -\frac{\rho^2 kn}{A^2} \right].
\] (3.9)

From (3.3) and (3.7) it now follows that, for big \( k \)
\[
P\left( \frac{L_{kn}}{kn} - \frac{1}{k} \left[ \frac{2kn}{2n-1} \right] l_{2n} > \rho \right) \leq \sum_{(\nu, \tau) \in B_{k,n}} P\left( L_{kn}(\nu, \tau) - \frac{1}{2} \left[ \frac{2kn}{2n-1} \right] EL_{2n} > \rho kn \right) \leq |B_{k,n}| \exp\left[ -\frac{\rho^2 kn}{A^2} \right] \leq \exp\left[ (\psi(n) + o(k) - \left( \frac{\rho}{A} \right)^2 ) kn \right].
\]

We consider \( n \) fixed and let \( k \) go to infinity. If \( u(n) > A\sqrt{\psi(n)} \), then there exists \( K(n) < \infty \) so that for every \( k > K(n) \),
\[
\psi(n) + o(k) - \left( \frac{u(n)}{A} \right)^2 < \frac{1}{2} \left( \psi(n) - \left( \frac{u(n)}{A} \right)^2 \right).
\]
Hence, replacing in the inequalities above \( \rho \) with \( u(n) \), we obtain for every \( k > K(n) \),
\[
P\left( \frac{L_{kn}}{kn} - \frac{1}{k} \left[ \frac{2kn}{2n-1} \right] l_{2n} > u(n) \right) \leq \exp\left[ \frac{1}{2} \left( \psi(n) - \left( \frac{u(n)}{A} \right)^2 \right) nk \right] = \exp[-d_n k],
\] (3.10)
where
\[
d_n := \left( \left( \frac{u(n)}{A} \right)^2 - \psi(n) \right) n > 0.
\]

Now recall the assumption that \( \delta \leq F \). Hence for any \( n \) and \( k \), the random variable \( L_{kn} l_{kn} \) is bounded by \( F \). From (3.10), it thus follows that for any \( k \)
\[
E\left( \frac{L_{kn}}{kn} \right) = l_{kn} \leq \frac{1}{k} \left[ \frac{2kn}{2n-1} \right] l_{2n} + u(n) + F \exp[-d_n k].
\]
Since \( l_{kn} \to l \) as \( k \to \infty \) and
\[
\frac{1}{k} \left[ \frac{2kn}{2n-1} \right] \leq \frac{2n}{2n-1} + \frac{1}{k},
\]
we obtain that for any \( n \),
\[
l \leq \left( \frac{2n}{2n-1} \right) l_{2n} + u(n) = l_{2n} \left( 1 + \frac{1}{2n-1} \right) + u(n).
\]
The proof of Theorem 1.1. From Lemma 3.1, it follows that to obtain a bound to \( l - l_n \), a suitable estimator of \( |\mathcal{B}_{k,n}| \) satisfying (3.7) should be found. Let us estimate \( |\mathcal{B}_{k,n}^r| \). The number of parts in the \( X \) side is bounded above by the number of combination with repetition from \( nk + 1 \) by \( r - 1 \). The repetitions allow empty parts. When the size of a part in \( X \)-side is \( m \), then, except from the last part, the size of the corresponding part on \( Y \) side has two possibilities: \( 2n - 1 - m \) or \( 2n - m \). Hence to any \( r \)-partition of \( X \)-size corresponds at most \( 2^{r-1}2n \) options in \( Y \) side. In the following we use the fact that the number of combination with repetition from \( nk + 1 \) by \( r - 1 \) is \( \binom{nk+r-1}{r-1} \) and for any non-negative integers \( a > b \) it holds
\[
\binom{a}{b} \leq \exp \left[ h_e \left( \frac{b}{a} \right) a \right],
\]
where \( h_e(q) := -q \ln q - (1 - q) \ln(1 - q) \) is the binary entropy function. Since \( r \leq \lceil \frac{2n}{2n-1} \rceil \) implies that \( r - 1 \leq \frac{2nk}{2n-1} \), we thus have for \( n \geq 2 \)
\[
|\mathcal{B}_{k,n}^r| \leq (2^{r-1}2n) \binom{nk+r-1}{r-1}
\leq \exp \left[ (r - 1)(\ln 2) + \ln(2n) + h_e \left( \frac{r - 1}{nk + r - 1} \right) (nk + r - 1) \right]
\leq \exp \left[ \left( \frac{\ln 4}{2n - 1} + \frac{\ln(2n)}{nk} \right) + h_e \left( \frac{r - 1}{nk + r - 1} \right) \left( 1 + \frac{2}{2n - 1} \right) \right] n^k
\leq \exp \left[ \left( \frac{\ln 4}{2n - 1} + \frac{\ln(2n)}{nk} \right) + h_e \left( \frac{2}{2n + 1} \right) \left( 2n + 1 \right) \left( 2n - 1 \right) \right] n^k.
\]
The last inequality follows from the inequalities
\[
\frac{r - 1}{nk + r - 1} \leq \frac{\frac{2nk}{2n-1}}{nk + \frac{2n}{2n-1}} = \frac{2}{2n + 1}
\]
so that if \( n \geq 2 \), then \( \frac{2n}{2n+1} \leq 0.5 \) and
\[
h_e \left( \frac{r - 1}{nk + r - 1} \right) \leq h_e \left( \frac{2}{2n + 1} \right).
\]
Hence
\[
|\mathcal{B}_{k,n}| \leq \left( \frac{2nk}{2n - 1} - k + 2 \right) \exp \left[ \left( \frac{\ln 4}{2n - 1} + \frac{\ln(2n)}{nk} \right) + h_e \left( \frac{2}{2n + 1} \right) \left( 2n + 1 \right) \left( 2n - 1 \right) \right] n^k
\leq \left( \frac{k}{2n - 1} + 2 \right) \exp \left[ \left( \frac{\ln 4}{2n - 1} + \frac{\ln(2n)}{nk} \right) + h_e \left( \frac{2}{2n + 1} \right) \left( 2n + 1 \right) \left( 2n - 1 \right) \right] n^k
= \exp \left[ \ln \left( \frac{k}{2n - 1} + 2 \right) + \frac{\ln 4}{2n - 1} + \frac{\ln(2n)}{nk} + h_e \left( \frac{2}{2n + 1} \right) \left( 2n + 1 \right) \left( 2n - 1 \right) \right] n^k
\leq \exp \left[ \ln \left( \frac{k}{2n - 1} + 2 \right) + \frac{\ln(2n)}{nk} + \frac{\ln 4}{2n - 1} + h_e \left( \frac{2}{2n + 1} \right) \left( 2n + 1 \right) \left( 2n - 1 \right) \right] n^k
\]
\[ = \exp \left( o(k) + \frac{\ln 4}{2n-1} + h_e \left( \frac{2}{2n+1} \right) \left( \frac{2n+1}{2n-1} \right) nk \right) \]
\[ \leq \exp \left( o(k) + \frac{2}{2n-1} \left( \frac{2n+1}{2n-1} + \ln(2n-1) \right) \right) nk , \]

where the last inequality follows from the inequality

\[ h_e \left( \frac{2}{2n+1} \right) \leq \frac{2}{2n+1} \left( \frac{2n+1}{2n-1} + \ln \left( \frac{2n-1}{2} \right) \right) . \]  

Hence (3.7) holds with

\[ \psi(n) = \frac{2}{2n-1} \left( \frac{2n+1}{2n-1} + \ln(2n-1) \right) . \]

The inequality (1.2) now follows from Lemma 3.1.  

**Acknowledgments**

The authors would like to thank the support of the German Science Foundation (DFG) through the Collaborative Research Center 701 "Spectral Structures and Topological Methods in Mathematics" (CRC 701) at Bielefeld University and the support of Barbara Gentz with the research stay of J. Lember at the CRC 701. Additionally, F. Torres would like to thank the partial support of the International Graduate College "Stochastics and Real World Models" (IRTG 1132) at Bielefeld University.
References


