LIMIT THEOREM FOR MAXIMAL SEGMENTAL SCORE FOR RANDOM SEQUENCES OF RANDOM LENGTH

UDC 519.21

B. L. S. PRAKASA RAO AND M. SREEHARI

Abstract. We obtain the limiting distribution of the maximal segmental score for the partial sums for a random number of independent and identically distributed random variables.

1. Introduction

Sequence alignment is useful for discovering functional, structural and evolutionary information in biological sequences. If two sequences from different organisms are similar, it could be because the organisms had a common ancestor sequence which transformed into these two organisms due to evolutionary changes (mutation). Alignment refers to the procedure of comparing two or more biological sequences by looking for a series of individual characters or character patterns that are in the same order in the sequences. A major problem is to identify interesting patterns in sequences. Suppose \( A_1, \ldots, A_n \) is an observed sequence (DNA, protein, etc.) from a finite alphabet (nucleotides or amino acids). Let \( \sigma \) be a scoring function. The local score of the sequence \( A_1, \ldots, A_n \) according to the scoring scheme \( \sigma \) is defined by

\[
H_n = \max_{1 \leq i \leq j \leq n} \left( \sum_{k=i}^{j} \sigma(A_k) \right).
\]

The local score \( H_n \) corresponds to a segment of the sequence with maximal aggregate score. The properties of the local score \( H_n \) have been investigated using the probability model that the successive letters of a sequence are generated by independent and identically distributed (i.i.d.) random variables or with a Markov chain model. Arratia and Waterman [2] proved that, for a sequence of length \( n \), as \( n \to \infty \), there exists a transition phase; when the average score is positive, there is a linear growth of the local score in \( n \), \( H_n = O(n) \); and when the average score is negative, it is called the logarithmic case, \( H_n = O(\log n) \). Arratia et al. [1] investigated the approximation of counts of matches in the best matching segment of specified length when comparing two long sequences of i.i.d. letters. Mercier and Daudin [7] obtain the exact distribution for the local score of one i.i.d. sequence. Mercier et al. [8] established the exact distribution of the maximum partial sum of a sequence of i.i.d. random variables using random walk theory, which led to a new approximation of the distribution of the local score of a sequence.

Let \( \{X_n\} \) be a sequence of independent and identically distributed (i.i.d.) random
variables defined on a probability space \((\Omega, \mathcal{F}, P)\) satisfying the following assumptions:

(i) \(P(X_1 > 0) > 0\),
(ii) \(E(X_1) < 0\), and
(iii) the random variable \(X_1\) is bounded; i.e., \(P(|X_1| < c) = 1\) for some constant \(c > 0\).

Let \(S_0 = 0, S_n = \sum_{k=1}^{n} X_k, n \geq 1,\) and \(Z_n = \max_{1 \leq i \leq j \leq n} (S_j - S_i)\). Iglehart [5] proved the following theorem.

**Theorem 1.1.** If \(\{X_n\}\) is a sequence of non-lattice i.i.d. random variables satisfying the conditions given above, then

\[
P[Z_n - \theta \log n \leq x] \to G(x) = \exp \left[ -ke^{-x/\theta} \right] \quad \text{as } n \to \infty
\]

for every \(x \in \mathbb{R}\), where \(\theta\) and \(k\) are positive constants depending on the distribution of \(X_1\).

Karlin and Dembo [6] extended this result to the lattice case. We obtain a random version of Theorem 1.1 in Section 2.

**2. Main result**

Let \(\{N_n\}\) be a sequence of positive integer-valued random variables defined on the probability space \((\Omega, \mathcal{F}, P)\) satisfying the condition

\[
\frac{N_n}{k_n} \xrightarrow{p} N
\]

for some sequence of integers \(\{k_n\}\), \(0 < k_n \uparrow \infty\), where \(N\) is a positive random variable. We prove the following random indexed version of the result in (1.1).

**Theorem 2.1.** Suppose \(\{X_n\}\) is a sequence of non-lattice i.i.d. random variables satisfying the conditions stated above. Then

\[
P(Z_{N_n} - \theta \log N_n \leq x) \to G(x)
\]

as \(n \to \infty\) for every \(x \in \mathbb{R}\), where \(G(x)\) is as defined by (1.1), and, for every \(x \in \mathbb{R}\),

\[
P(Z_{N_n} - \theta \log k_n \leq x) \to \int_{0}^{\infty} G(x - \theta \log u) dP(N \leq u) \quad \text{as } n \to \infty.
\]

The proof of the theorem depends on the following lemmas.

**Lemma 2.2.** Let \(\{r_n\}\) and \(\{k_n\}\) with \(r_n < k_n\) be two increasing sequences of positive integers with \(r_n \to \infty\) and let \(\{A_n\}\) be a sequence of events such that \(A_n\) depends only on the random variables \(X_{r_1}, \ldots, X_{k_n}\). Then, for any event \(A\) not depending on \(n\) such that \(P(A) > 0\),

\[
P(A_n \mid A) - P(A_n) \to 0 \quad \text{as } n \to \infty.
\]

The proof of the result is given in Lemma 1 in Barndorff-Nielsen [3].

Let \(r_n = o(\log k_n)\), and \(Z_n^* = \max_{r_n+1 \leq i \leq j \leq k_n} (S_j - S_i)\). Then

\[
Z_n^* \overset{\Delta}{=} Z_{k_n - r_n},
\]

in the sense that they have identical distributions, since \(X_1, X_2, \ldots, X_n\) are i.i.d. random variables. Hence, as \(n \to \infty\), we have

\[
P(Z_n^* \leq x + \theta \log k_n) \simeq P(Z_n^* \leq x + \theta \log(k_n - r_n))
\]

because \(r_n = o(\log k_n)\) in the sense that their difference tends to zero. This can be proved by the following arguments. Let

\[
U_n = Z_n^* - \theta \log(k_n - r_n)
\]
and
\[ F_n(x) = P(U_n \leq x). \]
In view of (1.1) and (2.3), it follows that
\[ F_n(x) \xrightarrow{w} G(x) \quad \text{as } n \to \infty. \]

Note that
\[ P(Z^*_n \leq x + \theta \log k_n) = P\left(Z^*_n \leq x + \theta \log(k_n - r_n) - \theta \log \frac{k_n - r_n}{k_n}\right) \]
\[ = P\left(U_n \leq x - \theta \log \frac{k_n - r_n}{k_n}\right) = F_n\left(x - \theta \log \frac{k_n - r_n}{k_n}\right). \]

Note that the distribution function \( G(x) \) is continuous for all \( x \). An application of Polya’s theorem and the fact that \( r_n = o(\log k_n) \) implies that
\[ F_n\left(x - \theta \log \frac{k_n - r_n}{k_n}\right) \xrightarrow{w} G(x) \quad \text{as } n \to \infty. \]

Hence
\[ (2.4) \quad P(Z^*_n \leq x + \theta \log k_n) \to G(x) \quad \text{as } n \to \infty. \]

We now establish the Rényi mixing property for the sequence of events \( \{A_n\} \), where
\[ A_n = \{Z_{k_n} \leq x + \theta \log k_n\}. \]

**Lemma 2.3.** For \( x \in \mathbb{R} \) and \( \theta \) as in (1.1) and any event \( A \) independent of \( n \) such that \( P(A) > 0 \),
\[ P(A_n \mid A) - P(A_n) \to 0 \quad \text{as } n \to \infty. \]

**Proof.** In view of Lemma 2.2,
\[ P(Z^*_n \leq x + \theta \log k_n \mid A) - P(Z^*_n \leq x + \theta \log k_n) \to 0 \quad \text{as } n \to \infty. \]

The proof of Lemma 2.3 will be complete if we prove that
\[ (2.5) \quad P(A_n \mid A) - P(Z^*_n \leq x + \theta \log k_n \mid A) \to 0 \quad \text{as } n \to \infty \]
for any event \( A \) with \( P(A) > 0 \). Let
\[ W_n = \max_{1 \leq i \leq n} \max_{i \leq j \leq k_n} (S_j - S_i). \]

We observe that
\[ A_n = \{Z_{k_n} \leq x + \theta \log k_n\} = \{\max(Z^*_n, W_n) \leq x + \theta \log k_n\} \]
and hence
\[ (2.6) \quad P(A)\left|P(A_n \mid A) - P(Z^*_n \leq x + \theta \log k_n \mid A)\right| \leq P(Z^*_n \leq x + \theta \log k_n < W_n). \]

Writing \( x + \theta \log k_n = a_n(x) \), we observe that \( W_n > a_n(x) \) if and only if there exist integers \( r \) and \( s \) such that \( r < s \leq k_n, r \leq r_n \) for which
\[ (2.7) \quad \sum_{j=r}^{s} X_j > a_n(x). \]

Clearly \( s > r_n \) because otherwise
\[ \sum_{j=r}^{s} X_j < c(s - r + 1) \]
and the inequality (2.7) cannot hold for large \( n \). Thus, the event \( [W_n > a_n(x)] \) implies that for some \( r \leq r_n < s \leq k_n \), the inequality (2.7) holds and for such values of \( r \) and \( s \),

\[
Z_n^* \geq X_{r_n+1} + \cdots + X_s = \sum_{j=r}^{s} X_j - \sum_{j=r}^{r_n} X_j
\]

\[
> a_n(x) - c(r_n - r + 1) \quad \text{(because } |X_i| < c \text{)}
\]

\[
\geq a_n(x) - cr_n.
\]

Thus

\[
P(Z_n^* \leq x + \theta \log k_n < W_n) \leq P[a_n(x) - cr_n < Z_n^* \leq a_n(x)] \to 0 \quad \text{as } n \to \infty,
\]

because of the fact at (2.4) and the choice of the sequence \( r_n \). The proof now follows from (2.6).

The results stated in Theorem 2.1 follow directly from Theorem 1 and Corollary 1 of Csörgő [4] in view of the Rényi mixing property proved in Lemma 2.3.

**Bibliography**


University of Hyderabad, Hyderabad 500046, India

E-mail address: blsprsm@uohyd.ernet.in

M. S. University, Vadodara, India

E-mail address: msreehari03@yahoo.co.uk

Received 17/JUL/2006

Originally published in English