isid/ms/2004/09
March 26, 2004
http://www.isid.ac.in/~statmath/eprints

Limit distribution of maximal segmental score for partial sums for random number of I.I.D. random variables

B. L. S. Prakasa Rao
M. Sreehari

Indian Statistical Institute, Delhi Centre
7, SJSS Marg, New Delhi–110 016, India

Limit Distribution of Maximal Segmental Score for Partial Sums for Random Number of I.I.D. Random Variables

B. L. S. PRAKASA RAO and M. SREEHARI
Indian Statistical Institute, New Delhi and M. S. University, Vadodara

Abstract

Molecular sequence analysis via the study of local score is an important technique in the study of molecular biology. Local score is the maximum segmental score for the partial sums obtained via the scoring scheme. Suppose the scoring scheme is modeled by a sequence of independent and identically distributed random variables. 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.

Key words : Maximum segmental score for partial sums; Random Limit Theorem; Local Score; Molecular sequence analysis.

AMS 2000 Subject Classification : Primary 60 G 50.

1 Introduction

Molecular sequence analysis is an important technique useful in the study of molecular biology. 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 σ be a scoring function. Scoring assignments for nucleotides arise from considerations such as biochemical categorization, physical properties or association with secondary structures. The local score of the sequence $A_1, \ldots, A_n$ according to the scoring scheme σ is defined by

$$H_n = \max_{1 \le i \le j \le n} (\Sigma_{k=i}^{j} \sigma(A_k)).$$

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 sucessive letters of a sequence are generated by independent and identically distributed random variables. For a discussion on biological sequence analysis, see Durbin et al. (1998) and Waterman (1995).

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 $S_n = \sum_{k=1}^{n} X_k$ and $Z_n = \max_{1 \leq i \leq j \leq n} (S_j - S_i)$. Iglehart (1972) proved the following theorem.

Theorem 1.1 : *If $\{X_n\}$ is a sequence of non-lattice random variables satisfying the conditions given above, then, as $n \to \infty$,*

$$P\left[Z_n - \theta \log n \leq x\right] \to G(x) = \exp\left[-ke^{-x/\theta}\right] \tag{1.1}$$

where θ and k are positive constants depending on the distribution of X_1.

Karlin and Dembo (1992) extended this result to the lattice case. Let $S_0 = 0$. Using the random walk theory, Mercier et. al. (2003) establish the exact distribution of the maximum partial sum $M = \sup_{k \geq 0} S_k$ of independent and identically distributed random variables $\{X_i, i \geq 1\}$ taking values from the alphabet $\mathcal{A} = \{+u, \ldots, -v\}$ where u and v are positive integers with $E(X_1) < 0$. They also obtain an approximation to the distribution of the local score. Note that the random variable Z_n defined above is the 'local score' in the study of molecular sequence analysis in molecular biology when X_i models $\sigma(A_i)$ and it also arises in the study of the maximal waiting-time distribution for the single server queue GI / G / 1, etc. Further details of these results and other examples can be found in the refrences cited above. Our main aim in this paper is to obtain a random version of Theorem 1.1.

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

$$\frac{N_n}{k_n} \xrightarrow{p} N \tag{2.1}$$

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

Theorem 2.1: *Under the conditions stated above,*

$$P\left(Z_{N_n} - \theta \log N_n \leq x\right) \to G(x) \tag{2.2}$$

as $n \to \infty$ for every $x \in R$, where $G(x)$ is as defined by (1.1).

The proof of the theorem depends on the following Lemmas.

Lemma 2.2 : Let $\{r_n\}$ and $\{m_n\}$ with $r_n \leq m_n$ be two increasing sequences of positive integers and let $\{A_n\}$ be a sequence of events such that A_n depends only on the random variables $X_{r_n}, \ldots, X_{m_n}$. Then, for any event A independent of n such that $P(A) > 0$,

$$P(A_n|A) - P(A_n) \to 0$$

as $n \to \infty$.

Proof of the result is given in Lemma 1 in Barndorff-Nielsen (1964).

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

$$Z_n^\star \stackrel{d}{=} Z_{k_n - r_n} \tag{2.3}$$

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

$$P\left(Z_n^\star \leq x + \theta \log k_n\right) \simeq P\left(Z_n^\star \leq x + \theta \log(k_n - r_n)\right)$$

because $r_n = o(\log\ k_n)$. This can be proved by the following arguments. Let

$$U_n = Z_n^\star - \theta \log(k_n - r_n)$$

and

$$F_n(x) = P(U_n \leq x).$$

In view of (1.1) and (2.3), it follows that

$$F_n(x) \stackrel{w}{\to} G(x)$$

as $n \to \infty$. Note that

$$\begin{aligned}
P\left(Z_n^\star \leq x + \theta \log k_n\right) &= P\left(Z_n^\star \leq x + \theta \log(k_n - r_n) - \theta \log \frac{k_n - r_n}{k_n}\right) \\
&= P\left(U_n \leq x - \theta \log \frac{k_n - r_n}{k_n}\right) \\
&= F_n\left(x - \theta \log \frac{k_n - r_n}{k_n}\right).
\end{aligned}$$

Note that the distribution function $G(x)$ is continuous for all x. An application of the Polya's theorem and the fact that $r_n = o(\log\ k_n)$ implies that

$$F_n\left(x - \theta \log \frac{k_n - r_n}{k_n}\right) \stackrel{w}{\to} G(x)$$

as $n \to \infty$. Hence

$$P\left(Z_n^\star \le x+\theta \log k_n\right) \to G(x) \tag{2.4}$$

as $n \to \infty$. Let

$$W_n = \max_{1\le i\le r_n-1}\max_{i\le j\le k_n}\left(S_j - S_i\right).$$

We now establish the mixing property for the sequence of events $\{A_n\}$ where

$$A_n = \left\{Z_{k_n} \le x+\theta \log k_n\right\}.$$

Lemma 2.3 : For $x \in R$ and θ as in (1.1) and any event A independent of n such that $P(A) > 0$,

$$P\left(A_n|A\right) - P\left(A_n\right) \to 0$$

as $n \to \infty$.

Proof. In view of Lemma 2.2

$$P\left(Z_n^\star \le x+\theta\log k_n|A\right) - P\left(Z_n^\star \le x+\theta\log k_n\right) \to 0$$

as $n \to \infty$. Furthermore, in view of the observation at (2.3) and the assumption $r_n = o(\log\, k_n)$, we have from (1.1) and (2.4) that,

$$\begin{aligned} P\left(Z_n^\star \le x+\theta\log k_n\right) &= P\left(Z_{k_n-r_n} \le x+\theta\log k_n\right) \\ &\to G(x) \end{aligned} \tag{2.5}$$

as $n \to \infty$. The proof of the Lemma 2.3 will be complete if we prove that

$$P\left(A_n|A\right) - P\left(Z_n^\star \le x+\theta\log k_n|A\right) \to 0$$

as $n \to \infty$ for any event A with $P(A) > 0$. We observe that

$$\begin{aligned} A_n &= \left\{Z_{k_n} \le x+\theta\log k_n\right\} \\ &= \left\{\max\left(Z_n^\star, W_n\right) \le x+\theta\log k_n\right\} \end{aligned}$$

and hence

$$\begin{aligned} &P(A)\left|P(A_n|A) - P(Z_n^\star \le x+\theta\log k_n|A)\right| \\ &\le P\left(Z_n^\star \le x+\theta\log k_n < W_n\right). \end{aligned} \tag{2.6}$$

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 $r < s \le k_n$, $r \le r_n - 1$ for which

$$\sum_{j=r}^{s} X_j > a_n(x). \tag{2.7}$$

Clearly $s > r_n$ because otherwise

$$\sum_{j=r}^{s} X_j < c(s-r)$$

and the inequality (2.7) can not hold for large n. Thus, the event $[W_n > a_n(x)]$ implies that for some $r < r_n < s \le k_n$, the inequality (2.7) holds and for such values of r and s

$$\begin{aligned} Z_n^\star &\ge X_{r_n} + X_{r_n+1} + \cdots + X_s = \sum_{j=r}^{s} X_j - \sum_{j=r}^{r_n-1} X_j \\ &> a_n(x) - c(r_n - r) \quad (\text{because } |X_i| < c) \\ &> a_n(x) - cr_n. \end{aligned}$$

Thus

$$\begin{aligned} &P\left(Z_n^\star \le x + \theta \log k_n < W_n\right) \\ &\le P\left[a_n(x) - cr_n < Z_n^\star \le a_n(x)\right] \\ &\to 0, \end{aligned}$$

as $n \to \infty$, because of the fact at (2.5) and the choice of the sequence r_n. The lemma now follows from (2.6).

We now prove the main Theorem 2.1.

Proof : Let i_0 and m be two positive integers. Write

$$\begin{aligned} \alpha_n &= P\left(Z_{N_n} \le x + \theta \log N_n, \left|\frac{N_n}{k_n} - N\right| > \frac{1}{m}\right), \\ \beta_n &= P\left(Z_{N_n} \le x + \theta \log N_n, \left|\frac{N_n}{k_n} - N\right| \le \frac{1}{m}, N \le \frac{i_0}{m}\right), \end{aligned}$$

and

$$\gamma_{ni} = P\left[Z_{N_n} \le x + \theta \log N_n, \left|\frac{N_n}{k_n} - N\right| \le \frac{1}{m}, \frac{i}{m} < N \le \frac{i+1}{m}\right].$$

Then

$$P\left(Z_{N_n} \le x + \theta \log N_n\right) = \alpha_n + \beta_n + \sum_{i=i_0}^{\infty} \gamma_{ni}. \tag{2.8}$$

Writing

$$n_{1i} = \left[\frac{k_n(i-1)}{m}\right] \quad \text{and} \quad n_{2i} = \left[\frac{k_n(i+2)}{m}\right],$$

we observe that

$$\gamma_{ni} \le P\left[Z_{n_{1i}} \le x + \theta \log n_{2i} | M_i\right] \pi_i \tag{2.9}$$

where

$$M_i = \left\{\frac{i}{m} < N \le \frac{i+1}{m}\right\} \quad \text{and} \quad \pi_i = P(M_i).$$

Also

$$\gamma_{ni} \geq P\left(Z_{n_{2i}} \leq x + \theta \log n_{1i} | M_i\right) \pi_i - P\left(\left|\frac{N_n}{k_n} - N\right| > \frac{1}{m}, M_i\right). \tag{2.10}$$

Let $\varepsilon > 0$. Choose i_0 so large such that for $i \geq i_0$

$$G(x - \theta \log \frac{i+2}{i-1}) \geq G(x - \varepsilon)$$

and

$$G(x + \theta \log \frac{i+2}{i-1}) \leq G(x + \varepsilon).$$

Then choose m large enough so that $P(N \leq \frac{i_0}{m}) < \varepsilon$. Finally choose n_0 large so that for $n \geq n_0$

$$P\left(\left|\frac{N_n}{k_n} - N\right| > \frac{1}{m}\right) < \varepsilon.$$

Then, by Lemma 2.3, we get from (1.1),(2.8), (2.9) and (2.10), that as $n \to \infty$

$$\begin{aligned} G(x-\varepsilon) \sum_{i=i_0}^{\infty} \pi_i - \varepsilon &\leq \underline{\lim} P(Z_{N_n} \leq x + \theta \log N_n) \\ &\leq \overline{\lim} P(Z_{N_n} \leq x + \theta \log N_n) \\ &\leq G(x+\varepsilon) \sum_{i=i_0}^{\infty} \pi_i + 2\varepsilon. \end{aligned}$$

In view of the continuity of the function $G(x)$ at x, we have

$$P\left(Z_{N_n} \leq x + \theta \log N_n\right) \to G(x)$$

as $n \to \infty$. This completes the proof of Theorem 2.1.

In analogy to Theorem 2 in Sreehari (1968), we can derive the following theorem.

Theorem 2.4 : Under the conditions stated earlier,

$$P\left(Z_{N_n} \leq x + \theta \log k_n\right) \to \int_0^{\infty} G(x - \theta \log u)\, dP(N \leq u)$$

as $n \to \infty$ for every x.

Proof : Let m be a positive integer. Write

$$\begin{aligned} \alpha_n^{\star} &= P\left(Z_{N_n} \leq x + \theta \log k_n,\ \left|\frac{N_n}{k_n} - N\right| > \frac{1}{m}\right) \\ \beta_n^{\star} &= P\left(Z_{N_n} \leq x + \theta \log k_n, \left|\frac{N_n}{k_n} - N\right| \leq \frac{1}{m},\ N \leq \frac{2}{m}\right) \end{aligned}$$

and

$$\gamma_{ni}^{\star} = P\left(Z_{N_n} \le x + \theta \log k_n, \left|\frac{N_n}{k_n} - N\right| \le \frac{1}{m}, \frac{i}{m} < N \le \frac{i+1}{m}\right).$$

Then

$$P\left(Z_{N_n} \le x + \theta \log k_n\right) = \alpha_n^{\star} + \beta_n^{\star} + \sum_{i=2}^{\infty} \gamma_{ni}^{\star}, \tag{2.11}$$

and

$$\gamma_{ni}^{\star} \le P\left[Z_{n_{1i}} \le x + \theta \log k_n | M_i\right] \pi_i \tag{2.12}$$

where n_{1i}, M_i and π_i are as defined in Theorem 2.1. Also

$$\gamma_{ni}^{\star} \ge P\left(Z_{n_{2i}} \le x + \theta \log k_n | M_i\right) \pi_i - P\left(\left|\frac{N_n}{k_n} - N\right| > \frac{1}{m}, M_i\right). \tag{2.13}$$

Let $\varepsilon > 0$. Choose m large enough so that

$$P\left(N < \frac{2}{m}\right) < \varepsilon.$$

Choose n_0 as in Theorem 2.1. Then, by Lemma 2.3, we get from (1.1),(2.11), (2.12) and (2.13), as $n \to \infty$,

$$\begin{aligned} \sum_{i=2}^{\infty} G\left(x + \theta \log \frac{m}{i+2}\right) \pi_i - \varepsilon &\le \underline{\lim}_{n} P\left(Z_{N_n} \le x + \theta \log k_n\right) \\ \le \quad \overline{\lim}_{n} P\left(Z_{N_n} \le x + \theta \log k_n\right) &\le \sum_{i=2}^{\infty} G\left(x + \theta \log \frac{m}{i-1}\right) \pi_i + 2\varepsilon. \end{aligned} \tag{2.14}$$

Since both

$$\sum_{i=2}^{\infty} G\left(x + \theta \log \frac{m}{i+2}\right) \pi_i = \sum_{i=2}^{\infty} G\left(x - \theta \log \frac{i+2}{m}\right) \pi_i$$

and

$$\begin{aligned} \sum_{i=2}^{\infty} G\left(x + \theta \log \frac{m}{i-1}\right) \pi_i &= \sum_{i=2}^{\infty} G\left(x - \theta \log \frac{i-1}{m}\right) \pi_i \\ \to \quad \int_0^{\infty} G\left(x - \theta \log u\right) dP\left(N \le u\right) \end{aligned}$$

as $m \to \infty$, by the usual arguments concerning the Riemann-Stieltjes integral, it follows from (2.14) that

$$P\left(Z_{N_n} \le x + \theta \log k_n\right) \to \int_0^{\infty} P\left(Z + \theta \log u \le x\right) \, dP(N \le u)$$

as $n \to \infty$ where Z is a random variable with distribution function $G(.)$.

Note : If the random variable N is a constant, C, say, then

$$P\left(Z_{N_n} \le x + \theta \log k_n\right) \to P\left(Z + \theta \log C \le x\right)$$

as $n \to \infty$.

This result also follows from Theorem 2.1 and the Slutsky's theorem.

Remarks :(1) It may be possible to deduce the Theorem 2.1 from general results in Csorgo (1974) for sequences of random variables $\{Z_n\}$ by using our Lemma 2.3 and then checking the Anscombe's condition. But our proof is direct. For general results on weak convergence of probability measures on complete separable metric spaces indexed by random sequences, see Prakasa Rao (1973).

(2) If the random variable N is independent of the sequence of random variables $\{X_n\}$, we do not need Lemma 2.3.

(3) If the random sequence $\{N_n\}$ is independent of the sequence of random variables $\{X_n\}$, then perhaps the assumption that

$$\frac{N_n}{k_n} \xrightarrow{p} N \quad \text{as} \quad n \to \infty$$

may be weakened to convergence in distribution of the sequence N_n/k_n to N as $n \to \infty$ as in Korolev (1994).

References

1. Barndorff-Nielsen, O. (1964) On the limit distribution of the maximum of a random number of independent random variables, *Acta Math. Acad. Sci. Hung.* **15**, 399-403.

2. Csorgo, S. (1974) On limit distributions of sequences of random variables with random indices, *Acta Math. Acad. Sci. Hung.* **25**, 227-232.

3. Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) Biological sequence analysis, In *Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press.

4. Iglehart, D. L. (1972) Extreme values in the GI/G/1 queues, *Ann. Math. Statist.,* **43**, 627-635.

5. Karlin, S. and Dembo, A. (1992) Limit-distribution of maximal segmental score among Markov-dependent partial sums, *Adv. Appl. Probab.* **24**, 113-140.

6. Korolev, V. Yu. (1994) Convergence of random sequences with independent random indices I, *Theor. Probab. Appl.* **39**, 282-297.

7. Mercier, S., Cellier, D. and Charlot, D. (2003) An improved approximation for assessing the statistical significance of molecular sequence features, *J. Appl. Probab.,* **40**, 427-441.

8. Prakasa Rao, B.L.S. (1973) Limit theorems for random number of elements on complete separable metric spaces, *Acta Math. Acad. Sci. Hung.* **24**, 1-4.

9. Sreehari, M. (1968) A limit theorem for the maximum of cumulative sums, *Acta. Math. Acad. Sci. Hung.* **19**, 117-120.

10. Waterman, M.S. (1995) *Introduction to Computational Biology*, Chapman and Hall. London.