The Annals of Probability
1991, Vol. 19, No. 4, 1737–1755

STRONG LIMIT THEOREMS OF EMPIRICAL FUNCTIONALS FOR LARGE EXCEEDANCES OF PARTIAL SUMS OF I.I.D. VARIABLES[1]

BY AMIR DEMBO AND SAMUEL KARLIN

Stanford University

Let (X_i, U_i) be pairs of i.i.d. bounded real-valued random variables (X_i and U_i are generally mutually dependent). Assume $E[X_i] < 0$ and $\Pr\{X_i > 0\} > 0$. For the (rare) partial sum segments where $\sum_{i=k}^{l} X_i \to \infty$, strong limit laws are derived for the sums $\sum_{i=k}^{l} U_i$. In particular a strong law for the length $(l - k + 1)$ and the empirical distribution of U_i in the event of large segmental sums of $\sum X_i$ are obtained. Applications are given in characterizing the composition of high scoring segments in letter sequences and for evaluating statistical hypotheses of sudden change points in engineering systems.

1. Introduction. The following problems are of interest in connection with molecular (DNA and protein) sequence comparisons [see Section 4, Karlin and Altschul (1990) and Karlin, Dembo and Kawabata (1990)] and motivated the developments of Theorems 1 and 2. The results can also be used in ascertaining the asymptotic maximal waiting time distribution for the general one server queue system $GI/G/1$ and in characterizing the biases of interarrival and service times over the busy period of the maximal waiting time. Other applications relate to insurance risk models and traffic flow.

In the simplest model the sequence

$$X_1, X_2, \dots, X_n, \dots \tag{1}$$

are i.i.d. random variables based on observations from a finite alphabet $\{a_i\}_1^r$, where

$$\Pr\{X = s_i\} = p_i, \qquad i = 1, 2, \dots, r, \qquad p_i > 0, \ \sum p_i = 1,$$

is interpreted in the manner that sampling the letter a_i yields a score s_i (see Section 2 for our general formulation). Let $\{S_m\}_1^n$, $S_0 = 0$, be the partial sum process subtended from (1). We assume

$$E[X] < 0 \tag{2}$$

and also $\Pr\{X > 0\} > 0$ so that $\{S_k\}$ entails a negative drift but positive probability of attaining early a positive score. The quantity

$$M(n) = \sup_{0 \le k < l \le n} (S_l - S_k) \tag{3}$$

Received December 1989; revised November 1990.

[1] Supported in part by NSF Grant DMS-86-06244 and NIH Grants GM-39907-02 and GM-10452-27.

AMS 1980 *subject classifications*. Primary 60F15, 60F10, 60G50.

Key words and phrases. Strong laws, large segmental sums, empirical functionals.

corresponds to a segment of the sequence $\{S_m\}_0^n$ with maximal score. Recent studies of the random variables $M(n)$ in various contexts occur in Deheuvels and Devroye (1987), Rootzén (1988), Iglehart (1972), Siegmund (1988) and Karlin, Dembo and Kawabata (1990). It is convenient to characterize $M(n)$ through successive excursions of positive values as follows: For the sample path $X_1, X_2, \dots,$ we define

$$K_1 = \min_k \{k \geq 1, S_k \leq 0\} \tag{4}$$

and sequentially the stopping times

$$K_0 = 0, \qquad K_\nu = \min\{k \geq K_{\nu-1} + 1, S_k - S_{K_{\nu-1}} \leq 0\}, \qquad \nu = 1, 2 \dots . \tag{5}$$

By virtue of the negative drift of $\{S_m\}$, these random variables are finite-valued. Fluctuation theory for sums of i.i.d. variables affirms that $K_j - K_{j-1}$ are i.i.d. positive-valued integer random variables having a distribution function with tails of at least exponential decay.

The time frame $K_{\nu-1} + 1$ to K_ν designates the νth excursion epoch encompassing the νth segment of the process $\{S_m\}$ starting from zero until hitting a nonpositive level.

For each $y > 0$, it is useful to define the stopping time

$$\begin{aligned} T_1(y) = \min\{m\colon 0 < S_k < y, k = 1, \dots, m-1 \\ \text{and either } S_m \geq y \text{ or } S_m \leq 0\}, \end{aligned} \tag{6}$$

indicating the elapsed time until the first departure of $\{S_k\}$ from the open interval $(0, y)$.

The realizations in (6) are of two kinds:

$$\begin{aligned} &I_1(y) = 1 \text{ or } 0 \text{ if } 0 < S_k < y,\ 0 < k < T_1(y) \quad \text{and} \\ &S_{T_1(y)} \geq y \text{ or } S_{T_1(y)} \leq 0, \text{ respectively.} \end{aligned} \tag{7}$$

Starting fresh in each excursion, we define [cf. (5)] successively

$$T_{\nu+1}(y) = \min\{m\colon m > K_\nu \text{ and either } S_m - S_{K_\nu} \leq 0 \text{ or } S_m - S_{K_\nu} \geq y\}.$$

Let $I_\nu(y) = 1$ when $S_{T_\nu(y)} - S_{K_{\nu-1}} \geq y$ and $I_\nu(y) = 0$ otherwise and $L_\nu(y)$ be the length of the νth segment, namely $L_\nu(y) = T_\nu(y) - K_{\nu-1}$, so $M(n) \geq y$ iff $I_\nu(y) = 1$ and $T_\nu(y) \leq n$ for some $\nu = 1, 2, \dots$. We prove the following limit theorem.

THEOREM 1. *Under the assumptions and notation above, let $T_\nu(y)$ be the first time in the above succession where $I_\nu(y) = 1$. Then*

$$\frac{L_\nu(y)}{y} \to \frac{1}{w^*} \quad a.s.\ as\ y \to \infty, \tag{8}$$

$w^* = E[Xe^{\theta^* X}]$, where θ^* is the unique positive root of the equation $E[e^{\theta X}] = 1$ (see Lemma 1 where $w^* > 0$ is proved).

Let $\{U_i\}$ be a sequence of bounded i.i.d. random variables; U_i may depend on X_i, but is independent of X_j, $j \neq i$. We form $W_m = \Sigma_{i=1}^{m} U_i$ and let

$$(9) \qquad W_\nu(y) = W_{T_\nu(y)} - W_{K_{\nu-1}}, \qquad \mu_\nu(y) = \frac{W_\nu(y)}{L_\nu(y)}.$$

THEOREM 2. *Excluding a set of measure zero with the index ν determined as in Theorem* 1, *then*

$$(10) \qquad \frac{W_\nu(y)}{L_\nu(y)} \to u^* \quad a.s.\ as\ y \to \infty,$$

where $u^* = E[Ue^{\theta^* X}]$.

We generally omit the index ν when referring to ν determined as in Theorem 1. The following application of Theorem 2 is important. Define $U = U(X) = 1$ if $X = s_1$ and 0 otherwise. In this case the variable $W(y)/L(y)$ counts the proportion of occurrences of the letter a_1 during an excursion confined strictly to positive values over a time segment achieving a level beyond y, $y \to \infty$, that is, conditioned on the sample realization $I(y) = 1$.

COROLLARY 1. *The empirical frequency distribution $\mu(y)$ of the letters $\{a_1, \dots, a_r\}$ observed during the νth excursion epoch, ν defined in Theorems* 1 *and* 2, *converges* (*as $y \to \infty$*) *with probability* 1 *to the frequency measure μ^* which takes the value a_i with probability $q_i = p_i e^{\theta^* s_i}$, $i = 1, 2, \dots, r$.*

For X nondiscrete-valued but of bounded range, let $U(X) = 1$ if $X \in A$ (A a Borel set of the real line) and zero otherwise, then $\mu(y) \to E[e^{\theta^* X} I_A(X)]$, $I_A(\cdot)$ the indicator function of A.

As a simple consequence, we can calculate the following random size ballot count probability. Consider two contestants, I and II, cumulating votes assigned to I versus II with probability p and q, respectively, $p + q = 1$, $q > p$. We inquire, assuming I strictly leads II throughout the count, as to the probability of the proportion of votes I received until the vote gap between I and II first attains a level y. By identifying

$$X_i = \begin{cases} 1, & \text{prob. } p, \\ -1, & \text{prob. } q, \end{cases} \qquad (q > p),$$

we ascertain that the limiting conditioned ballot count probability (as $y \to \infty$) of this event converges to q.

For the variable $M(n)$ defined in (3), let $L(M(n))$ be the duration of the excursion until first reaching the level $M(n)$ before returning to a nonpositive value. The strong law $\theta^* M(n)/\log n \to 1$ a.s. is proved, for example, in Karlin and Dembo (1992). It follows from Theorems 1 and 2 that

$$(11) \qquad \frac{\theta^* L(M(n))}{\log n} \to \frac{1}{w^*} \quad \text{a.s. as } n \to \infty, \quad \text{where } w^* = E[Xe^{\theta^* X}],$$

and the frequency distribution of scores during the maximal excursion interval converges to

$$\mu(M(n)) \to \mu^*, \tag{12}$$

where μ^* takes value a_i with probability $p_i e^{\theta^* s_i}$ (or more generally, $\mu^*(A) = E[e^{\theta^* X} I_A(X)]$).

Note that $L(M(n))$ differs from the longest positive excursion of the partial sum process. For example, the latter when divided by $\log n$ approaches a constant exceeding $1/\theta^* w^*$ [see Deheuvels and Steinebach (1989), Theorem 6].

The proofs of Theorems 1 and 2 rely on properties of the Wald martingale family

$$P_n(\theta) = \frac{e^{\theta S_n}}{[\phi(\theta)]^n}, \qquad \phi(\theta) = E[e^{\theta X}], \tag{13}$$

coupled with information from renewal processes and fluctuation theory for sums of independent random variables. Extensions of the results to the case where $\{X_k\}$ are generated as Markov-dependent are presented in the companion paper [Dembo and Karlin (1991)].

A conditional central limit theorem applicable to the variable

$$\frac{W(y) - u^* L(y)}{\sqrt{L(y)}} \quad \text{conditioned on } I(y) = 1 \quad [\text{see } (7)] \tag{14}$$

and other limit laws will be discussed elsewhere.

Various applications of Theorems 1 and 2 are presented in Section 4 describing the segment composition for periodic score patterns, matching among multiple letter sequences and connections with likelihood ratio tests for detecting a change of measure.

2. Formulation and preliminaries. Let $\{X_k\}$ be real i.i.d. random variables with bounded range $|X_k| \le K$ (we shall drop the subscript writing X for X_k whenever possible), negative expectation, $E[X] < 0$, and positive with nonzero probability, $\Pr\{X > 0\} > 0$.

Let $\{U_k\}$ be another sequence of random variables with bounded range such that $\{X_k, U_k\}$ are jointly i.i.d. pairs but typically U_k depends on X_k. Denote

$$S_m = \sum_{k=1}^{m} X_k \tag{15}$$

with $S_0 = 0$ and define the sequence of stopping times

$$\begin{aligned} &K_0 = 0, \\ &K_\nu = \min\{k \colon k > K_{\nu-1}, S_k - S_{K_{\nu-1}} \le 0\}, \qquad \nu = 1, 2, \dots . \end{aligned} \tag{16}$$

The stopping times K_ν delineate a sequence of epochs with largest attained height in the νth epoch given by

$$Q_\nu = \sup\{S_k - S_{K_{\nu-1}}, K_{\nu-1} \le k < K_\nu\}. \tag{17}$$

For $y > 0$, let for $\nu = 1, 2, \dots,$

$$T_\nu(y) = \inf\{k\colon k > K_{\nu-1} \text{ such that either } S_k - S_{K_{\nu-1}} \le 0 \text{ or } S_k - S_{K_{\nu-1}} \ge y \text{ holds}\}, \tag{18}$$

so that $T_\nu(y)$ is the first exit time of $S_k - S_{K_{\nu-1}}$, $k > K_{\nu-1}$, from the range $(0, y)$. $T_\nu(y) < K_\nu$ iff $Q_\nu \ge y$ in which case we indicate this event by

$$I_\nu(y) = 1 \tag{19}$$

and refer to the epoch $(K_{\nu-1}, T_\nu)$ as a *y-excursion* (y is positive throughout this paper). Otherwise the realization is designated by $I_\nu(y) = 0$. We shall investigate the distribution of

$$L_\nu(y) = T_\nu(y) - K_{\nu-1} \quad \text{and} \quad W_\nu(y) = \sum_{i=K_{\nu-1}+1}^{T_\nu(y)} U_i \tag{20}$$

for large y-excursions and during the epoch of the maximal $M(n)$-excursion; see (3). Because of the independence of $\{X_k, U_k\}$ with respect to k, the variables $\{W_\nu(y)\}_{\nu=1}^\infty$ and $\{L_\nu(y)\}_{\nu=1}^\infty$ are i.i.d. sharing the distribution of $W(y) = W_1(y)$ and $L(y) = L_1(y)$, respectively. Likewise $\{I_\nu(y)\}_{\nu=1}^\infty$ are i.i.d. sharing the distribution of $I(y)$.

The principal results are stated in Theorems 1 and 2 of Section 1. The proofs are organized centered on the following series of lemmas. Henceforth the assumptions ($|X|$ bounded range, $E[X] < 0$, $\Pr\{X > 0\} > 0$) prevail unless explicitly stated to the contrary.

LEMMA 1 (Wald [see, e.g., Karlin and Taylor (1975), pages 264–265]). *There exists a unique positive solution θ^* of $E[e^{\theta X}] = 1$ and*

$$w^* = E[Xe^{\theta^* X}] > 0. \tag{21}$$

LEMMA 2. *For any $\nu \ge 1$, the probability that $I_\nu(y) = 1$ [see, e.g.,* (19)*] has exponential decay*

$$0 < \delta \le \Pr\{I_\nu(y) = 1\}e^{\theta^* y} \le 1 \quad \textit{for all } y > 0 \textit{ and some } \delta > 0. \tag{22}$$

[Actually, $\Pr\{I(y) = 1\}e^{\theta^* y} \to C^*$ as $y \to \infty$, where C^* has an explicit evaluation; see Iglehart (1972), Karlin and Dembo (1992).]

Define

$$u^* = E[Ue^{\theta^* X}]. \tag{23}$$

LEMMA 3. *For $y \to \infty$ [for notation, see* (20)*],*

$$E\left[\left|\frac{L(y)}{y} - \frac{1}{w^*}\right|^4 \middle| I(y) = 1\right] = O\left(\frac{1}{y^2}\right). \tag{24}$$

LEMMA 4. *For $y \to \infty$ [for notation, see* (20)],

$$E\left[\left|\frac{W(y)}{L(y)} - u^*\right|^4 \middle| I(y) = 1\right] = O\left(\frac{1}{y^2}\right). \tag{25}$$

[Henceforth $O(y)$ (or $O(y^2)$) will signify a bound Cy (or Cy^2) for y large where C is a generic positive constant which can change over successive equations.]

These estimates applied via the Chebyshev inequality and the Borel–Cantelli lemma lead to the following conclusion.

LEMMA 5. *For any sequence $y_1, y_2, \ldots$ and $\nu_1^*, \nu_2^*, \ldots,$ integer-valued random variables such that $I_{\nu_n^*}(y_n) = 1$, $\sum_{n=1}^{\infty} 1/y_n^2 < \infty$ (where the choice of $\nu_1^*, \nu_2^*, \ldots,$ do not alter the distribution of any of L_ν, W_ν corresponding to $\{L_1, W_1\}$), the random variables $L_{\nu_n^*}(y_n)/y_n$ and $W_{\nu_n^*}(y_n)/L_{\nu_n^*}(y_n)$ each converge to a finite constant with probability* 1.

Let $L(y)$ and $W(y)$ correspond to the initial y-excursion to which Theorems 1 and 2 refer. These theorems exclude a set Γ of probability 0 such that for every sample path not in Γ,

$$\frac{L(y)}{y} \to \frac{1}{w^*} \quad \text{and} \quad \frac{W(y)}{L(y)} \to u^* \quad \text{as } y \to \infty.$$

We proceed by specifying levels in blocks to which Lemma 5 applies such that

$$\begin{aligned} &y_1 = y_2 = \cdots = y_{\alpha_1} = 1; \\ &y_{\alpha_1+1} = \cdots = y_{\alpha_2} = 2; \ldots ; y_{\alpha_{n-1}+1} = \cdots = y_{\alpha_n} = n; \ldots \end{aligned} \tag{26}$$

with the level $y_{\alpha_{n-1}+j}$ or greater attained in the jth epoch of the type $(K_{\nu-1}, T_\nu(n))$ in which $I_\nu(n) = 1$. These successive excursions are denoted by

$$\left(K_{\nu-1}, T_\nu(y_{\alpha_{n-1}+j})\right) \equiv \beta_j(n), \qquad j = 1, \ldots, \alpha_n - \alpha_{n-1}. \tag{27}$$

Let $\alpha_n - \alpha_{n-1} = [A \log n]$ with A a fixed large constant to be determined later and $\alpha_0 = 0$. Note that

$$\sum_{k=1}^{\infty} \frac{1}{y_k^2} \le A \sum_{n=1}^{\infty} \frac{\log n}{n^2} < \infty$$

so that the conditions of Lemma 5 are met. Therefore $L_j(n)/n \to 1/w^*$ and $W_j(n)/L_j(n) \to u^*$ with probability 1 [here $L_j(n)$ and $W_j(n)$ are $L_\nu(y)$ and $W_\nu(y)$ corresponding to the epoch of $\beta_j(n)$].

We will estimate the probability of the following events:

$$\begin{aligned} \mathscr{E}_n = \{&\text{the first excursion to level } n + 1 \text{ does not include any} \\ &\text{of the first } [A \log n] \text{ consecutive excursions to height } n.\} \end{aligned} \tag{28}$$

By suitable choice of A large enough but fixed, the convergence $\sum \Pr\{\mathscr{E}_n\} < \infty$ will be established. Application of Borel–Cantelli implies:

Lemma 6. *With probability* 1, *the first excursion to level* $n + 1$ *with* n *large enough* (*dependent on the sample path*) *is coincident with one among the first* $[A \log n]$ *excursions to level* n.

We then deduce that the first excursion to level y or greater for any y satisfying $n \le y \le n + 1$ is also coincident, with probability 1, with one of the initial $[A \log n]$ excursions that reach height n.

For the sample realizations $\beta_j(n)$, where the level n at least is first attained at the index time $\tau_j(n)$, let $\tilde{J}_j(n)$ be determined satisfying

$$\max_{k \ge \tau_j(n)} \left[S_k - S_{\tau_j(n)}\right] = S_{\tilde{J}_j(n)+\tau_j(n)} - S_{\tau_j(n)}.$$

The variable $\tilde{J}_j(n)$ is well-defined by virtue of the negative drift of the process $\{S_k\}$. Moreover, $\tilde{J}_j(n)$ is governed by the same distribution as the time span required in achieving the maximum of $\{S_k\}$ starting from time zero. Thus, for $n \le y \le n + 1$ and where the event $\mathscr{E}_n$ *does not occur* then for some j ($1 \le j \le [A \log n]$),

$$|L(y) - L_j(n)| \le \max\left(\tilde{J}_1(n), \tilde{J}_2(n), \dots, \tilde{J}_{[A \log n]}(n)\right) = J_n^*. \tag{29}$$

We will prove:

Lemma 7.

$$\frac{J_n^*}{n} \to 0 \quad a.s.$$

In view of (29) and Lemma 7, since $L_j(n)/n \to 1/w^*$ with probability 1 and with $n \le y \le n + 1$, we infer $L(y)/y \to 1/w^*$ for the same sample realization as asserted in Theorem 1.

3. Proofs of lemmas and theorems.

Proof of Lemma 1. Because X has a nondegenerate bounded range, $\phi(\theta) = E[e^{\theta X}]$ is analytic and strictly convex for all real θ. Also $\phi(0) = 1$, $\phi'(0) < 0$ by hypothesis and $\lim_{\theta \to +\infty} \phi(\theta) = \infty$, because $\Pr\{X > 0\} > 0$. The existence of a unique $\theta^* > 0$ satisfying $\phi(\theta^*) = 1$ and $\phi'(\theta^*) > 0$ is accordingly assured.

Let

$$\psi(\theta, t) = E[e^{\theta X + tU}] \tag{30}$$

and set $S_m = \sum_{i=1}^m X_i$, $W_m = \sum_{i=1}^m U_i$. The family of random variables

$$P_m = \frac{e^{\theta S_m + t W_m}}{[\psi(\theta, t)]^m}, \qquad m = 0, 1, \dots, \tag{31}$$

constitutes the familiar Wald martingale (a two-dimensional version). Since the distribution function of the stopping time random variable $L = L_1(y)$ of the first exit time from the interval $(0, y)$ tails down exponentially fast [see (61) below], we can apply the optional sampling theorem yielding

$$E[e^{\theta S_L + tW_L - L\zeta(\theta, t)}] \equiv 1, \tag{32}$$

where $W_L = W_1(y)$, $\zeta(\theta, t) = \log \psi(\theta, t)$, and this equation is valid for $|t|$ sufficiently small and all real $\theta \geq \theta^*$ as $|S_L| \leq y + K$ and the U_i terms are bounded ensuring $|W_L| \leq L(y)K'$.

We need to assess the asymptotic probability that an excursion attains height y or more, $\Pr\{I(y) = 1\}$. □

LEMMA 2.

$$0 < \delta \leq e^{\theta^* y} \Pr\{I(y) = 1\} \leq 1. \tag{33}$$

PROOF. Take $\theta = \theta^*$, $t = 0$ in (32) [so that $\psi(\theta^*, 0) = \phi(\theta^*) = 1$], then

$$\begin{aligned} 1 = E[e^{\theta^* S_{L(y)}}] &= \Pr\{I(y) = 1\} e^{\theta^* y} E\big[e^{\theta^*(S_{L(y)} - y)} | I(y) = 1\big] \\ &\quad + \Pr\{I(y) = 0\} E\big[e^{\theta^* S_{L(y)}} | I(y) = 0\big]. \end{aligned} \tag{34}$$

Since $S_{L(y)} \geq y$ conditioned on $I(y) = 1$ and $\theta^* > 0$, the upper bound in (33) results simply by discarding the second term of (34). For the lower bound, observe that $S_{L(y)} \leq y + K$ and if $I(y) = 0$, then $S_{L(y)} \leq 0$. Moreover, the event $\{X_1 \leq 0\}$ is included in $I(y) = 0$. These facts combined into (34) yield

$$1 \leq \Pr\{I(y) = 1\} e^{\theta^* y} e^{\theta^* K} + E\big[e^{\theta^* S_{L(y)}}; I(y) = 0\big].$$

Rearranging the terms produces

$$1 - E\big[e^{\theta^* X_1}; X_1 \leq 0\big] - E\big[e^{\theta^* S_L}; X_1 > 0, I(y) = 0\big] \leq \Pr\{I(y) = 1\} e^{\theta^* y} e^{\theta^* K}.$$

The left side exceeds

$$1 - \Pr\{X_1 > 0\} - E\big[e^{\theta^* X_1}; X_1 \leq 0\big] = \Pr\{X_1 \leq 0\} E\big[(1 - e^{\theta^* X_1}) | X_1 \leq 0\big]$$

and therefore

$$\delta = \Pr\{X \leq 0\} E\big[(1 - e^{\theta^* X}) | X \leq 0\big] e^{-\theta^* K} \leq \Pr\{I(y) = 1\} e^{\theta^* y}.$$

Since $E[X] < 0$ entails $\Pr\{X < 0\} > 0$ and $E[e^{\theta^* X} | X \leq 0] < 1$ (as $\theta^* > 0$), we deduce that $\delta > 0$ as was to be shown. □

PROOF OF LEMMA 3. The expression (32) is analytic for all θ and $|t|$ small. Differentiating in θ successively four times with $t = 0$ and afterwards setting $\theta = \theta^*$ [since $\psi(\theta^*, 0) = \phi(\theta^*) = 1$] produces the formulas

$$E\big[(S_L - w^* L) e^{\theta^* S_L}\big] = 0, \qquad w^* = \phi'(\theta^*) > 0, \tag{35a}$$

$$E\big[(S_L - w^* L)^2 e^{\theta^* S_L}\big] = k(\theta^*) E[L e^{\theta^* S_L}], \qquad k(\theta) = \frac{d}{d\theta}\left[\frac{\phi'(\theta)}{\phi(\theta)}\right], \tag{35b}$$

$$\begin{aligned} E\big[(S_L - w^* L)^4 e^{\theta^* S_L}\big] &= 6k(\theta^*) E\big[(S_L - w^* L)^2 L e^{\theta^* S_L}\big] \\ &\quad + 4k'(\theta^*) E\big[(S_L - w^* L) L e^{\theta^* S_L}\big] \\ &\quad - 3[k(\theta^*)]^2 E[L^2 e^{\theta^* S_L}] + k''(\theta^*) E[L e^{\theta^* S_L}]. \end{aligned} \tag{36}$$

We rewrite (35a), decomposing by the event $I(y) = 0$ or 1, yielding

$$w^*E[Le^{\theta^*S_L}] = e^{\theta^*y}\Pr\{I(y) = 1\}E[S_Le^{\theta^*(S_L-y)}|I(y) = 1] \\ + \Pr\{I(y) = 0\}E[S_Le^{\theta^*S_L}|I(y) = 0]. \tag{37}$$

Since $y \leq S_L \leq y + K$ conditioned on $I(y) = 1$ while S_L is bounded conditioned on $I(y) = 0$, by virtue of Lemma 2 and since K is fixed, we deduce for y large

$$C_2y \leq E[Le^{\theta^*S_L}] \leq C_1y, \tag{38}$$

where C_1 and C_2 are appropriate positive constants.

We ascertain from (35b) that

$$E\left[(S_L - w^*L)^2e^{\theta^*S_L}\right] = O(y). \tag{39}$$

Paraphrasing the analysis of (37) leads to the estimate $D^2 = E[S_L^2e^{\theta^*S_L}] = O(y^2)$.

Next, expanding the square on the left of (39) gives

$$(w^*)^2E[L^2e^{\theta^*S_L}] - 2w^*E\left[(S_LL)e^{\theta^*S_L}\right] + E\left[S_L^2e^{\theta^*S_L}\right] = O(y). \tag{40}$$

Setting $\xi^2 = E[L^2e^{\theta^*S_L}]$ and applying the Schwarz inequality (using $E[e^{\theta^*S_L}] = 1$) to the second term of (40) produces a quadratic inequality for the (positive square root) quantity ξ, namely

$$(w^*)^2\xi^2 \leq 2w^*\xi D + E, \tag{41}$$

where $D = O(y)$ and $E = O(y^2)$. It follows that the largest root of this quadratic equation is certainly $O(y)$ and ξ, being positive and obeying the inequality (41), has the order growth

$$\xi^2 = E[L^2e^{\theta^*S_L}] = O(y^2). \tag{42}$$

Applying the Schwarz inequality in (36) to the terms $E[(S_L - w^*L)^2Le^{\theta^*S_L}]$ and $E[(S_L - w^*L)Le^{\theta^*S_L}]$ yields the inequality

$$\eta^2 \leq \tilde{A}(y)\eta + \tilde{B}(y),$$

where $\tilde{A}(y) = O(y)$ and $\tilde{B}(y) = O(y^2)$, for the variable

$$\eta^2 = E\left[(S_L - w^*L)^4e^{\theta^*S_L}\right],$$

taking account of the result of (42). The reasoning attendant to (41) applies, mutatis mutandis, implying

$$E\left[(S_L - w^*L)^4e^{\theta^*S_L}\right] = O(y^2). \tag{43}$$

In particular, we obtain (since $\Pr\{I(y) = 1\}e^{\theta^*y} \geq \delta > 0$)

$$E\left[(S_L - w^*L(y))^4e^{\theta^*(S_L-y)}|I(y) = 1\right] = O(y^2). \tag{44}$$

Since $S_L - y$ is bounded conditioned that $I(y) = 1$, we can convert (44) (use the Minkowski inequality) into

$$E\left[(y - w^*L(y))^4 | I(y) = 1\right] = O(y^2).$$

To sum up, we have

$$E\left[\left(\frac{L(y)}{y} - \frac{1}{w^*}\right)^4 | I(y) = 1\right] = O\left(\frac{1}{y^2}\right). \tag{45}$$ □

Proof of Lemma 4. Paraphrasing the analyses leading to (45) by differentiating the identity (32) with respect to t and evaluating at $\theta = \theta^*$ and $t = 0$ will yield

$$E\left[(W_L - u^*L)^4 | I(y) = 1\right] = O(y^2), \tag{46}$$

where $u(\theta, t) = (\partial/\partial t)\zeta(\theta, t)$ and $u^* = u(\theta^*, 0)$. Here is how it is done.

Successive differentiation of (32) in t [cf. (35)–(36)] produces

$$E\left[(W_L - u^*L)e^{\theta^* S_L}\right] = 0, \tag{47a}$$

$$E\left[(W_L - u^*L)^2 e^{\theta^* S_L}\right] = \kappa(\theta^*, 0)E\left[Le^{\theta^* S_L}\right] \quad \text{where } \kappa(\theta, t) = \frac{\partial}{\partial t}u(\theta, t), \tag{47b}$$

$$\begin{aligned} E\left[(W_L - u^*L)^4 e^{\theta^* S_L}\right] &= 6\kappa(\theta^*, 0)E\left[(W_L - u^*L)^2 Le^{\theta^* S_L}\right] \\ &\quad + 4\frac{\partial}{\partial t}\kappa(\theta^*, 0)E\left[(W_L - u^*L)Le^{\theta^* S_L}\right] \\ &\quad - 3\left[\kappa(\theta^*, 0)\right]^2 E\left[L^2 e^{\theta^* S_L}\right] + \frac{\partial^2}{\partial t^2}\kappa(\theta^*, 0)E\left[Le^{\theta^* S_L}\right]. \end{aligned} \tag{48}$$

From previous estimates we deduce sequentially

$$\begin{aligned} E\left[W_L e^{\theta^* S_L}\right] &= O(y), \\ E\left[(W_L - u^*L)^2 e^{\theta^* S_L}\right] &= O(y), \\ E\left[W_L^2 e^{\theta^* S_L}\right] &= O(y^2). \end{aligned} \tag{49}$$

Observe next that

$$\begin{aligned} E\left[L(W_L - u^*L)^2[e^{\theta^* S_L}] \le \sqrt{E[L^2 e^{\theta^* S_L}]\, e^{\theta^* S_L}}\right] \\ \times\sqrt{E\left[(W_L - u^*L)^4 e^{\theta^* S_L}\right]} \\ = O(y)\zeta_0, \qquad \text{[we use (42)]} \end{aligned} \tag{50}$$

where $\zeta_0 = (E[(W_L - u^*L)^4 e^{\theta^* S_L}])^{1/2}$.

With (50) in hand, returning to (48) we see that ζ_0 satisfies the quadratic inequality

$$\zeta_0^2 \le A(y)\zeta_0 + B(y), \tag{51}$$

where $A(y) = O(y)$, $B(y) = O(y^2)$. On the basis of (51) we deduce as before [cf. (41) and after (42)] that

$$E\left[(W_L - u^*L)^4 e^{\theta^* S_L}\right] = O(y^2). \tag{52}$$

In particular, following the method of (44), we further infer (46), that is,

$$E\left[(W_L - u^*L)^4 | I(y) = 1\right] = O(y^2).$$

The inequalities (45) and (46) imply

$$E\left[\left(\frac{W_{L(y)}}{L(y)} - u^*\right)^4 | I(y) = 1\right] = O\left(\frac{1}{y^2}\right). \tag{53}$$

We are now prepared to prove Lemma 6. Let $\tau_1(y)$ be the time duration of the excursion of $\{S_k\}_{k=1}^{\infty}$ on the positive axis where $\{S_k\}_{k=1}^{\infty}$ first attains a level at least y, that is, $\tau_1(y)$ is the first time index satisfying

$$y \le \max_{0 \le k \le l \le \tau_1(y)} (S_l - S_k) = S_{\tau_1(y)} - S_{\kappa_1(y)},$$

where $\kappa_1(y)$ indicates the starting time of the relevant part of the excursion. After $\tau_1(y)$, the $\{S_k\}$ process returns to the nonpositive axis, say at time $\sigma_1(y)$. Commencing the process $\{S_k\}$ fresh after time $\sigma_1(y)$, we determine the first passage time defined by the relations

$$\max_{\sigma_1(y) < k \le l < \tau_2(y)} [S_l - S_k] < y \quad \text{but } S_{\tau_2(y)} - S_{\kappa_2(y)} \ge y.$$

Continuing in this way, the time indices

$$(\kappa_1(y), \tau_1(y), \sigma_1(y)), (\kappa_2(y), \tau_2(y), \sigma_2(y)), \dots, (\kappa_\nu(y), \tau_\nu(y), \sigma_\nu(y)), \dots$$

represent successive epochs with $S_0 = 0$ specifying the νth epoch such that the process $\{S_k\}$ starting at $\kappa_\nu(y)$ first departs $(0, y)$ at time $\tau_\nu(y)$ at a level at least y and returns afterwards to the nonpositive axis at $\sigma_\nu(y)$. Therefore, $\tau_\nu(y) - \kappa_\nu(y)$ has the distribution of $L(y)$ conditioned on $I(y) = 1$.

For the level $y = n$, we construct $[A \log n]$ such n-excursion epochs characterized by stopping times

$$\begin{aligned} &\tau_1(n), \tau_2(n), \dots, \tau_{[A \log n]}(n) \\ &\qquad \text{with corresponding } \kappa_1(n), \kappa_2(n), \dots, \kappa_{[A \log n]}(n). \end{aligned} \tag{54}$$ □

Lemma 8.

$$\lim_{n \to \infty} \max_{j = 1, \dots, [A \log n]} \left| \frac{\tau_j(n) - \kappa_j(n)}{n} - \frac{1}{w^*} \right| = 0 \quad a.s. \tag{55}$$

Arrange these random variables in lexicographic order such that (n, j) precedes (m, k) if $n < m$ or if $n = m$ and $j < k$.

Proof. We have the estimate [see (45)]

$$E\left[\left(\frac{\tau(y)-\kappa(y)}{y}-\frac{1}{w^*}\right)^4\right]=O\left(\frac{1}{y^2}\right)$$

and in particular for $n = 1, 2, \dots$

$$E\left[\left(\frac{\tau_j(n)-\kappa_j(n)}{n}-\frac{1}{w^*}\right)^4\right]\le\frac{C}{n^2}\quad\text{for some } C<\infty. \tag{56}$$

Applying the simple Markov inequality based on (56), we have for every prescribed $\varepsilon > 0$, for some $C_\varepsilon < \infty$,

$$\sum_{n=1}^{\infty}\sum_{j=1}^{[A\log n]}\Pr\left\{\left|\frac{\tau_j(n)-\kappa_j(n)}{n}-\frac{1}{w^*}\right|>\varepsilon\right\}\le C_\varepsilon A\sum_{n=1}^{\infty}\frac{\log n}{n^2}<\infty. \tag{57}$$

Invoking the Borel–Cantelli theorem completes the proof of (55). □

By using the bounds of (53) paralleling the analysis leading to (55), we deduce the following result.

Lemma 9.

$$\lim_{n\to\infty}\max_{j=1,\dots,[A\log n]}\left|\frac{\sum_{i=\kappa_j(n)+1}^{\tau_j(n)}U_i}{\tau_j(n)-\kappa_j(n)}-u^*\right|=0\quad a.s. \tag{58}$$

To continue the development of Theorem 1, we first prove Lemma 6.

Proof of Lemma 6. Since $\Pr\{X > 0\} > 0$, we have for some finite m_0 (possibly $m_0 = 1$),

$$\Pr\{0<S_1<S_2<\cdots<S_{m_0-1}<1<S_{m_0}\}=a>0. \tag{59}$$

Determine A large enough such that $A((-1)\log(1-a)) = \gamma > 1$. For $y = n$, consider the first $j_n = [A\log n]$ distinct successive excursions to level n occuring at times $\tau_1(n), \tau_2(n), \dots, \tau_{j_n}(n)$ as determined in (54).

Consider the event [defined earlier in (28)]

(60) $\mathscr{E}_n$ = none of the j_n excursions characterized by $\{\kappa_j(n), \tau_j(n), \sigma_j(n)\}_1^{j_n}$ attain level $n+1$ or higher.

Since these j_n excursions are clearly independent, by virtue of (59),

$$\Pr\{\mathscr{E}_n\}\le(1-a)^{[A\log n]}\le\frac{C}{n^\gamma}$$

and consequently

$$\sum_{n=1}^{\infty}\Pr\{\mathscr{E}_n\}<\infty.$$

Applying again Borel–Cantelli, we conclude that

$$\Pr\{\mathscr{E}_n \text{ infinitely often}\} = 0.$$

Therefore with probability 1, for n large enough (depending on the sample path) at least one of the level n-excursions corresponding to $\{\tau_j(n)\}_1^{j_n}$ also reaches level $n + 1$. This confirms Lemma 6. □

Proof of Lemma 7. Consider the partial sum process $\{S_m\}_{m=0}^{\infty}$, $S_0 = 0$. The maximum $M = \max S_k = S_J$ is well-defined, since $E(X) < 0$ and J represents the index in which M is first achieved.

Note for $k \ge 1$ that

$$\Pr\{J = k\} \le \Pr\{S_k > 0\} \le \phi^k(\theta) = e^{-bk}$$

by Markov's inequality, where $\phi(\theta) < 1$ for $0 < \theta < \theta^*$. Accordingly,

$$\Pr\{J \ge k\} \le \tilde{C}e^{-bk} \tag{61}$$

and the random variable J has at least an exponential decay tail probability.

For each of the level n-excursions of time duration $\tau_\nu(n) - \kappa_\nu(n)$, $\nu = 1, 2, \dots, j_n$ [see (54) and before for notation],

(62) let $\tilde{J}_\nu(n)$ be the time span beyond $\tau_\nu(n)$ where $\max_{\tau_\nu(n) \le k}\{S_k - S_{\tau_\nu(n)}\}$ is first achieved.

Obviously the tail behavior estimate in (61) applies to each $\tilde{J}_\nu(n)$.

Consider an arbitrary level y and determined $n(y)$ satisfying $n(y) \le y < n(y) + 1$. Let $\tau(y)$ be the first time index such that $\{S_k\}$ reaches a height greater than or equal to y.

Consider

$$T_n^* = \sup_{n \le y < n+1} \min_{\substack{1 \le \nu \le j_n \\ \tau_\nu(n) \le \tau(y)}} \left[\tau(y) - \tau_\nu(n)\right]. \tag{63}$$

By Lemma 6, for almost every sample path of $\{S_k\}$ for n large enough, the excursion to level y agrees with one of the excursions among $\{\kappa_\nu(n), \tau_\nu(n), \sigma_\nu(n)\}_1^{j_n}$ reaching the level n (one of these actually reaches level $n + 1$). On this basis it follows that

$$T_n^* \le \max\left\{\tilde{J}_1(n), \tilde{J}_2(n), \dots, \tilde{J}_{j_n}(n)\right\} = J_n^*, \tag{64}$$

with $\tilde{J}_\nu(n)$ defined as in (62). We claim that $T_n^*/n \to 0$ with probability 1. Indeed, for each $\varepsilon > 0$, by virtue of (61), we have

$$\sum_{n=1}^{\infty} \Pr\left\{\frac{T_n^*}{n} > \varepsilon\right\} \le \sum_{n=1}^{\infty} A \log n \Pr\{J \ge n\varepsilon\} \le \tilde{C} \sum_{n=1}^{\infty} A \log n e^{-\varepsilon n b} < \infty.$$

Now apply Borel–Cantelli, yielding

$$\frac{T_n^*}{n} \to 0 \quad \text{a.s.} \tag{65}$$

□

Proof of Theorem 1. From Lemma 8, we know that

$$(66)\qquad \max_{\nu=1,\dots,jn}\left|\frac{\tau_\nu(n)-\kappa_\nu(n)}{n}-\frac{1}{w^*}\right|\to 0\quad \text{a.s. as } n\to\infty.$$

Combining the facts of Lemma 6 with (62)–(65) entails the conclusion

$$(67)\qquad \frac{\tau(y)-\kappa(y)}{y}\to\frac{1}{w^*}\quad\text{a.s.}$$

The sets of measure zero precluded relate only to the convergence statements involving the excursions of $\{\kappa_\nu(n),\tau_\nu(n),\sigma_\nu(n)\}_1^{j_n}$, $n=1,\dots,$ those occurring in Lemmas 6, 7 and 8. Therefore, the statement of (67) applies with probability 1 as $y\uparrow\infty$ in any manner. The proof is complete. □

Proof of Theorem 2. For a level y, let n be determined as before, namely $n(y)\le y\le n(y)+1$ and the $\tau(y)$, $\kappa(y)$, $\sigma(y)$, $\{\tau_\nu(n),\kappa_\nu(n),\sigma_\nu(n)\}_{\nu=1}^{j_n}$. Consider

$$(68)\qquad W_n^*=\sup_{n\le y<n+1}\ \min_{1\le\nu\le j_n}\left|\sum_{i=\kappa(y)+1}^{\tau(y)}U_i-\sum_{i=\kappa_\nu(n)+1}^{\tau_\nu(n)}U_i\right|.$$

For almost every sample path and all n large enough, on the basis of Lemma 6 we know that $\kappa(y)=\kappa_\nu(n)$ for an appropriate ν. Thus (with $K=\max|U_i|$) we have

$$(69)\qquad \begin{aligned}\frac{W_n^*}{n}&\le K\sup_{n\le y<n+1}\ \min_{\substack{1\le\nu\le j_n\\ \tau_\nu(n)\le\tau(y)}}\frac{|\tau(y)-\tau_\nu(n)|}{n}\\ &\le K\frac{\max\left(\tilde{J}_1(n),\tilde{J}_2(n),\dots,\tilde{J}_{j_n}(n)\right)}{n}=K\frac{J_n^*}{n},\end{aligned}$$

which by Lemma 7 tends to 0 with probability 1. We have already established that $(\tau(y)-\kappa(y))/y$ and $(\tau_\nu(n)-\kappa_\nu(n))/n$ both converge to $1/w^*$ a.s. and therefore Lemma 9, that is,

$$\max_{\nu=1,\dots,j_n}\left|\frac{1}{\tau_\nu(n)-\kappa_\nu(n)}\sum_{i=\kappa_\nu(n)+1}^{\tau_\nu(n)}U_i-u^*\right|\to 0\quad\text{a.s. as } n\to\infty,$$

implies Theorem 2:

$$\frac{1}{\tau(y)-\kappa(y)}\sum_{i=\kappa(y)+1}^{\tau(y)}U_i\to u^*\quad\text{a.s. as } y\uparrow\infty.\qquad\square$$

4. Applications.

1. Assume $\Pr\{X=s_i\}=p_i$, $i=1,\dots,r$ and $E[X]<0$. The segment of the maximal excursion from $\{S_m\}_1^n$ corresponds to the index $\nu^0(n)$ with the property that $0\le S_m-S_{K^0_{\nu(n)}}<M(n)$, $K_{\nu^0(n)}<m<T_{\nu^0(n)}$ and $S_{T_{\nu^0(n)}}-S_{K_{\nu^0(n)}}=M(n)$. Since $\theta^*M(n)/\ln n\to 1$ a.s. as $n\to\infty$, the result of

Theorem 1 entails

$$\frac{L_{\nu^0(n)}}{M(n)} \to \frac{1}{E[Xe^{\theta^* X}]} \quad \text{a.s.,} \tag{70}$$

where $L_{\nu^0(n)} = T_{\nu^0(n)} - K_{\nu^0(n)}$. Moreover, Theorem 2 posits that the letter frequencies in the maximal segment are approximated almost surely by the probabilities

$$\Pr\{X = s_i\} \approx p_i e^{\theta^* s_i}. \tag{71}$$

2. The conclusions of paragraph 1 apply to the second highest excursion, third highest excursion and to the several top segmental score values which all with probability 1 have asymptotic growth rate $\ln n/\theta^*$ and their content realize scores with the biased frequencies $\{p_i e^{\theta^* s_i}\}$.

3. Suppose in evaluating the segments $S_l - S_k$, $1 \le k \le l \le n$, we are allowed to delete up to d (d fixed) summands. In this context we seek to assess

$$M_d(n) = \max_{0 \le k \le l \le n} \left\{ S_l - S_k - \sum_{\substack{k < \lambda_i \le l \\ i = 1, \dots, e;\, e \le d}} X_{\lambda_i} \right\}. \tag{72}$$

Because X_i have bounded range, $\theta^* M_d(n)/\ln n \to 1$ and for the segment achieving $M_d(n)$, the results of Theorems 1 and 2 are again in force.

4. A periodic version of $M(n)$ relevant for molecular sequence studies pertains to an alternating sequence of random variables $\{X_k\}$, where $\{X_{2m}\}_1^n$ are identically distributed and separately $\{X_{2m+1}\}$ are identically distributed with $\Pr\{X_{2m} = s_i\} = p_i$ and $\Pr\{X_{2m+1} = s_i'\} = p_i'$ and all $\{X_k\}$ are independent. As usual we assume negative drift, namely $\sum_{i=1}^r (p_i s_i + p_i' s_i') < 0$. In this case the θ^* parameter is determined as the unique positive root of the equation

$$\left(\sum_{i=1}^r p_i e^{\theta s_i} \right) \left(\sum_{i=1}^r p_i' e^{\theta s_i'} \right) = 1. \tag{73}$$

The composition of the maximal segment score (or any high scoring segment) has letter a_i occuring with approximate frequency

$$\tfrac{1}{2} p_i e^{\theta^* s_i} \left(\sum_{j=1}^r p_j' e^{\theta^* s_j'} \right) + \tfrac{1}{2} p_i' e^{\theta^* s_i'} \left(\sum_{j=1}^r p_j e^{\theta^* s_j} \right).$$

Analogous results prevail for more general periodic patterns in sequences.

5. In molecular sequence analysis it is frequently of interest to compare two letter sequences

$$A_1, A_2, \dots, A_n \quad \text{and} \quad A_1', A_2', \dots, A_n', \tag{74}$$

both i.i.d. independently distributed, with

$$\Pr\{A = \text{letter } a_\alpha,\ A' = \text{letter } a_\beta\} = p_\alpha p_\beta'$$

and the score for such a match being $s_{\alpha\beta}$. We assume as before that

$\sum_{\alpha,\beta} s_{\alpha\beta} p_\alpha p'_\beta < 0$ and Pr{of obtaining a positive score} > 0 or, more generally, replace $p_\alpha p'_\beta$ by $p_{\alpha\beta}$.

For convenience of notation, let $X(A, A') = s_{\alpha\beta}$ when A samples a_α and A' samples a_β. In the aligned case, we define

$$S_m = X(A_1, A'_1) + \cdots + X(A_m, A'_m), \qquad m \geq 1, S_0 = 0,$$

and for $n = n'$,

$$M(n) = \max_{0 \leq k \leq l \leq n} (S_l - S_k).$$

The analogs of Theorems 1 and 2 apply to this situation without alteration. There are r^2 generic values for X in this model and θ^* is the unique positive root of the equation

$$\sum_{\alpha,\beta} p_\alpha p'_\beta e^{\theta s_{\alpha\beta}} = 1. \tag{75}$$

Another problem of interest is to ascertain the maximal score among all possible, not necessarily aligned, intervals, that is, of obtaining the asymptotic distributional properties of $\max_{k,m,n} \tilde{S}_k(m,n)$, where

$$\tilde{S}_k(m,n) = \sum_{\substack{i=m\\ j=n}}^{\substack{m+k\\ n+k}} X(A_i, A'_j). \tag{76}$$

The asymptotic behavior of the quantity (76) is as yet unresolved. A special case was dealt with in Arratia, Morris and Waterman (1988).

6. Consider aligned sequences and score assignments for every pair of three independent sequences: $\{A_i^{(1)}\}_1^n$, $\{A_i^{(2)}\}_1^n$ and $\{A_i^{(3)}\}_1^n$ generated with probabilities $\{p_i^{(1)}\}$, $\{p_i^{(2)}\}$ and $\{p_i^{(3)}\}$, respectively. The scoring arrays are $\{s_{\alpha,\beta}^{(1,2)}\}$, $\{s_{\alpha,\beta}^{(1,3)}\}$ and $\{s_{\alpha,\beta}^{(2,3)}\}$. We seek to characterize the maximal scoring segments among any two of the three sequences, that is,

$$M(n) = \max_{0 \leq k \leq l \leq n} \left[\left(S_l^{(1,2)} - S_k^{(1,2)}\right), \left(S_l^{(1,3)} - S_k^{(1,3)}\right), \left(S_l^{(2,3)} - S_k^{(2,3)}\right) \right]. \tag{77}$$

For each pair there is the critical parameter $\theta_{1,2}^*, \theta_{1,3}^*, \theta_{2,3}^*$ determined as in (75). If, say, $\theta_{1,2}^* < \min(\theta_{1,3}^*, \theta_{2,3}^*)$, then the maximal segment score in (77) derives exclusively from the comparisons of the first and second sequences.

We established in (33) the exponential decay $\Pr\{I(y) = 1\} = O(e^{-\theta^* y})$. Indeed, the precise asymptotic behavior is indicated in Karlin, Dembo and Kawabata (1990), giving

$$\lim_{y \to \infty} \Pr\{T(y) e^{\theta^* y} \leq t\} = 1 - e^{-K^* t}, \qquad t > 0, \tag{78}$$

where $T(y)$ is the dual of $M(n)$ and K^* is an explicit constant represented in general by a geometrically fast convergent series involving familiar quantities from the fluctuation theory of partial sums of i.i.d. random variables. Let $K_{1,2}^*$, $K_{1,3}^*$ and $K_{2,3}^*$ be the corresponding parameters for the respective sequence pairings. In the case $\theta_{1,2}^* = \theta_{1,3}^* < \theta_{2,3}^*$, then the maximal segment occurs with probability $K_{1,2}^*/(K_{1,2}^* + K_{1,3}^*)$ for the sequence pair 1 and 2 and with probability $K_{1,3}^*/(K_{1,2}^* + K_{1,3}^*)$ in the matching of sequences 1 and 3. When $\theta_{1,2}^* =$

$\theta^*_{1,3} = \theta^*_{2,3}$, then the maximal segment score can be achieved from any pair of sequence comparisons with appropriate probabilities $K^*_{i,j}/(K^*_{1,2} + K^*_{1,3} + K^*_{2,3})$ for sequences i and j. Consider a desirable target set of frequencies $\{q_{\alpha\beta}\}$, $\sum^r_{\alpha,\beta=1} q_{\alpha\beta} = 1$. The likelihood ratio scores

$$s^{(i,j)}_{\alpha,\beta} = \ln\left(\frac{q_{\alpha\beta}}{p^{(i)}_\alpha p^{(j)}_\beta}\right)$$

are meaningful in many circumstances [the rationale for these assignments is elaborated in Karlin and Altschul (1990)]. For these scores, necessarily $\theta^*_{1,2} = \theta^*_{1,3} = \theta^*_{2,3} = 1$ and the interpretations above on mixtures pertain.

Adapting the above arguments in a straightforward manner allows us to apply Theorems 1 and 2 for objectives of ascertaining the high scoring aligned segments, involving comparing all subsequences from r out of s sequences.

7. The following scenario underlies realizations of many engineering systems and in other social and managerial contexts. Observations $A_1, A_2, \dots, A_n$ accumulate sequentially. For simplicity of exposition we assume $\{A_i\}$ come from r possible values and are mutually independent but not necessarily of identical distribution. Specifically, to time t (unknown), the A_i are governed by the probability law P with $\{p_i\}^r_1$ and then abruptly change to follow a different probability law Q with $\{q_i\}^r_1$. The change in the probability law could reflect sudden failure of the measuring equipment or a more critical phase transition during the system process. Where the change occured after the epoch t, $t < n$, the joint probability distribution of $\{A_i\}^n_1$ is

$$\left(\prod_{i=1}^{t} p_{A_i}\right)\left(\prod_{j=t+1}^{n} q_{A_j}\right).$$

A common statistical test used to decide if a change in probability law has occured is founded on likelihood ratio statistics such as

$$\log\left(\frac{Q(A_{t+1}, \dots, A_n)}{P(A_{t+1}, \dots, A_n)}\right). \tag{79}$$

We have the problem of discriminating between a set of hypotheses

$$H_0 : t \geq n \quad \text{versus} \quad H_1 : t < n.$$

A generalized likelihood test for this problem decides in favor of H_1 once

$$\max_{0 \leq t \leq n} \log\left(\prod_{i=t+1}^{n} \frac{q_{A_i}}{p_{A_i}}\right) \geq y \tag{80}$$

for some critical level y. Interpreting the numbers $\log(q_i/p_i) = s_i$ as scores, the random variables (80) ascertain the maximal segment score among the collection of segments extending from t to n, $1 \leq t \leq n$. Note under the hypothesis H_0 that $-E_0[X_n] = \sum_i p_i \log(p_i/q_i)$ is the Kullback–Leibler distance $D(\mathbf{p}, \mathbf{q})$ of the measure $\mathbf{p}$ relative to the measure $\mathbf{q}$, so $E_0(X_n)$ is always negative here. Under the hypothesis H_1, let $L = n - t$ be the delay (penalty) in detecting a change. Consider the partial sum process $S_m = \sum^m_{k=1} X_k$, where

X_k takes values s_i with probability p_i. By the law of large numbers,

$$\lim_{n-t\to\infty} \frac{1}{n-t}(S_n - S_t) = D(\mathbf{q},\mathbf{p}) \quad \text{in probability (under } H_1).$$

It thus follows that

$$\lim_{y\to\infty} \frac{L}{y} \le \frac{1}{D(\mathbf{q},\mathbf{p})} \quad \text{in probability (under } H_1). \tag{81}$$

Thus the threshold y of the test could be set to the level $L^* D(\mathbf{q},\mathbf{p})$ with L^* the tolerated delay. Large deviation estimates establish the behavior of L under H_1.

Under H_0, eventually false positives (detection) result since

$$M(N) = \max_{1\le n\le N} \max_{0\le t<n} (S_n - S_t) \to \infty \quad \text{a.s. (at the rate } \log N);$$

see, for example, Karlin, Dembo and Kawabata (1990). Actually the probability estimate of a wrong decision is ascertained there from the limit law (78) with $\theta^* = 1$ and an explicit formula for K^*. As $E_{\mathbf{p}}[U(X)e^{\theta^* X}] = E_{\mathbf{q}}[U(X)]$, Theorem 1 requires equality in (81) under H_0. Moreover, by Theorem 2, any statistic averaged over the segment $\{K_\nu(y), T_\nu(y)\}$ under H_0 is distributed in the limit ($y \to \infty$) as if the change of probability law occured at K_ν. The foregoing calculations indicate that the generalized likelihood test is a sufficient statistic in testing hypothesis H_0 versus H_1. For other analyses of related statistical problems and for further references, see Siegmund (1985).

8. Portfolio management. Suppose that a certain portfolio management strategy in use leads over successive time periods to wealth accumulation $W_0, W_1, \dots, W_n \dots$. Then $X_n = \log(W_n/(W_{n-1}))$ is a random variable usually of positive mean. The randomness in X_n reflects the uncertainty of the market economy and in particular of the investments comprising the portfolio. For example, investing proportions $b_1, \dots, b_l$ on l stocks whose daily price gains (losses) are $(Y_n)_1, \dots, (Y_n)_l$ yields $X_n = \sum_{i=1}^{l} b_i (Y_n)_i$. An i.i.d. assumption on the changes $X_1, \dots, X_n$ provides a first approximation model. $S_n = \log(W_n/W_0) = X_n + \cdots + X_1$ assesses the (logarithmic) growth of wealth realized by this portfolio. The largest negative exceedance

$$M_n = \min_{0\le k\le l\le n} (S_l - S_k),$$

measures the *largest segmental* loss incurred during the time frame $1, \dots, n$. The variable $L(M_n)$ is the time duration during this loss and $\mu(M_n)$ the composition of the wealth increments during such bad periods.

Our theorems, for example, provide predictions on the depth of the worst stock market decline, expected length of such a recession and so on.

9. Population extinction. Now W_n is the size of the population of a certain organism at time period n and $X_n = \log(W_n/W_{n-1})$, the logarithmic reproduction rate in the population at this time. Under a resource unlimited model X_n can be considered i.i.d. (or Markov) random variables of positive mean. In this case the size of the population grows without bound.

Here M_n assesses the largest extinction percentage experienced by the population, while $L(M_n)$ indicates its time duration and $\mu(M_n)$ the empirical distribution of $\{X_i\}$ during this period.

REFERENCES

Arratia, R., Morris, P and Waterman, M. S. (1988). Stochastic scrabble: Large deviations for sequences with scores. *J. Appl. Probab.* **25** 106–119.

Deheuvels, P. and Devroye, L. (1987). Limit laws of Erdös–Rényi–Shepp type. *Ann. Probab.* **15** 1363–1386.

Deheuvels, P. and Steinebach, J. (1989). Sharp rates for increments of renewal processes. *Ann. Probab.* **17** 700–722.

Dembo, A. and Karlin S. (1991). Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of Markov variables. *Ann. Probab.* **19** 1756–1767.

Iglehart, D. (1972). Extreme values in the *GI/G/*1 gene. *Ann. Math. Statist.* **43** 627–635.

Karlin, S. and Altschul, S. F. (1990). New methods for assessing the statistical significance of molecular sequence features using general scoring schemes. *Proc. Nat. Acad. Sci. U.S.A.* **87** 2264–2268.

Karlin, S. and Dembo, A. (1992). Limit distributions of maximal segmental scòre among Markov dependent partial sums. *Adv. in Appl. Probab.* To appear.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes*, 2nd ed. Academic, New York.

Karlin, S., Dembo, A. and Kawabata, T. (1990). Statistical composition of high scoring segments from molecular sequences. *Ann. Statist.* **18** 571–581.

Rootzén, H. (1988). Maxima and exceedances of stationary Markov chains. *Adv. Appl. Probab.* **20** 371–390.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer, New York.

Siegmund, D. (1988). Approximate tail probabilities for the maxima of some random fields. *Ann. Probab.* **16** 487–501.

Departments of Mathematics and Statistics
Stanford University
Stanford, California 94305