The Annals of Probability
1988, Vol. 16, No. 2, 535–563

MAXIMAL LENGTH OF COMMON WORDS AMONG RANDOM LETTER SEQUENCES[1]

By Samuel Karlin and Friedemann Ost

Stanford University and Technische Universität München

Consider random letter sequences $\{\xi_t^{(\sigma)},\ t = 1, \dots, N;\ \sigma = 1, \dots, s\}$ based on a finite alphabet generated by uniformly mixing stationary processes. The asymptotic distributional properties of the length of the longest common word in r or more of the s sequences $K_{r,s}(N)$, are investigated. When the probability measures of the different sequences are not too dissimilar, a classical extremal type limit law holds for $K_{r,s}(N) - (r \log N/(-\log \lambda))$, λ being an appropriate local match parameter. The distributional properties of other long-word relationships and patterns among the sequences are also discussed.

1. Introduction. Consider s independent random stationary letter sequences $\{\mathscr{S}_\sigma\} = \{\xi_t^{(\sigma)},\ \sigma = 1, 2, \dots, s,\ t \in \mathbb{N}^+\}$, generated from a finite alphabet $\mathscr{A} = \{a_1, \dots, a_m\}$, $m \geq 2$, each process having the following properties.

(i) *Uniform mixing*. For each $\delta > 0$ there exists a $d(\delta) \in \mathbb{N}^+$ such that for all $t \in \mathbb{N}^+$, $d > d(\delta)$, and events $A \in \mathscr{F}(\xi_1, \dots, \xi_t)$, $B \in \mathscr{F}(\xi_{t+d}, \xi_{t+d+1}, \dots)$, we have

$$(1.1)\qquad (1 - \delta)\Pr\{A\}\Pr\{B\} \leq \Pr\{A \cap B\} \leq (1 + \delta)\Pr\{A\}\Pr\{B\},$$

where $\mathscr{F}(\cdots)$ comprises the σ-field of events induced by the indicated random variables.

(ii) *Positivity*. There exists a universal constant $\varepsilon > 0$ such that

$$(1.2)\qquad \begin{aligned} &\Pr\{\xi_t = a_t | \xi_1 = a_1, \dots, \xi_{t-1} = a_{t-1}, \xi_{t+1} = a_{t+1}, \xi_{t+2} = a_{t+2}, \dots\} \\ &\qquad > \varepsilon, \end{aligned}$$

for all $a_1, a_2, \dots \in \mathscr{A}$, and $t \in \mathbb{N}^+$.

Random stationary sequences fulfilling (i) and (ii) include irreducible Markov stationary sequences of any order, sequences based on functions of these Markov chains, semi-Markov processes and ARMA sequences.

A k-word is a set of k contiguous letters in a sequence identified by its starting position. We will compare words across the different sequences $\mathscr{S}_\sigma$ with positions traversing the range $1 \leq t \leq N_\sigma$, $\sigma = 1, \dots, s$, and N_σ of the same order. For ease of the typography we take all $N_\sigma = N$.

We investigate the asymptotic distributional properties ($N \to \infty$) of the random variables, $K_{r,s}(N)$, defined as the length of the longest common word in r or more of the s sequences. For Markov-dependent stationary sequences

Received March 1986; revised November 1986.

[1] Supported in part by NIH Grant GM-10452-22 and NSF Grant MCS-82-15131.

AMS 1980 *subject classifications*. 60F10, 60F99.

Key words and phrases. Random letter sequences, extremal distributions, uniformly mixing stationary processes, local match distribution, longest common word.

governed by probability transition matrices $P(\sigma)$, $\sigma = 1, 2, \dots, s$, each $P(\sigma)$ appropriately irreducible aperiodic, $K_{r,s}(N)$ has order growth $(\log N^r)/(-\log \lambda)$, where λ is a characteristic local match parameter of the aggregate processes (Theorem 2.2) provided $P(\sigma_i)$ are not too different from each other. An anomalous strong law emerges when some letters are strongly favored in some sequences while other sequences strongly favor a complementary set of letters [Arratia and Waterman (1985)]. Sufficient conditions of wide scope that assure a limit law for the variable $K_{r,s}(N) - (\log N^r)/(-\log \lambda)$ are set forth. The limit distribution is one of the classical extremal types depending on two parameters λ and γ. The λ is a generalized principal eigenvalue which can be defined for any uniformly mixing stationary process (Theorem 3.1). For independent i.i.d. sequences $\gamma = 1$, but generally γ is an explicit nontrivial function of the eigenvectors associated with λ.

Section 2 formalizes the concepts and problems of the foregoing discussion and the main results are stated in precise terms. The proofs are developed in Sections 3–6.

The distributional properties of other long word relationships can also be discerned. These include relationships such as relabelings, transpositions on word positions, grouping letters, masking letters and allowances for a limited number of errors between matches. For example, with two sequences let $R(N)$ be the length of the longest word in $\mathscr{S}_1$ whose reversed word occurs in $\mathscr{S}_2$. The growth of $R(N)$ is of order $(\log N^2)/(-\log \mu)$ for suitable μ, $0 < \mu < 1$, but $-\log \mu > -\log \lambda$. The evaluation of μ is given in Karlin and Ost (1987). The methods of this paper for assessing long matching words also work for general comparisons of word relationships within and between sequences.

Characterizing the longest word aligned across r sequences is related to the problem of the longest success run in a random sequence. For the case of iterated logarithmic type bounds in i.i.d. sequences see Erdös and Révész (1975), Guibas and Odlyzko (1980), Samarova (1981) and in the context of semi-Markov letter sequences, Foulser and Karlin (1987) and references therein. Arratia and Waterman (1985) and Arratia, Gordon and Waterman (1986) consider the functional $K_{2,2}(N)$ for two sequences of i.i.d. variables. The results of this paper for independent letter sequences with applications to DNA sequences were reported in Karlin, Ghandour, Ost, Tavare and Korn (1983) and Karlin, Ghandour and Foulser (1985). The results of Theorem 2.2 are useful for practical objectives in assessing the statistical significance of patterns in molecular sequences of similar genes within and between species.

2. Formulation and results.

Consider an alphabet $\mathscr{A}$ of m letters and s independent $\mathscr{A}$-valued *uniformly mixing positive* stationary letter sequences as characterized in (1.1) and (1.2). We denote a sample realization of the process $\mathscr{S}_\sigma$ by $\xi_t^{(\sigma)}$, $t = 1, 2, \dots$. An unrestricted shared word of length at least k relative to the sequences $\mathscr{S}_{\sigma_1}, \mathscr{S}_{\sigma_2}, \dots, \mathscr{S}_{\sigma_r}$ is such that there exist positions $t_1, t_2, \dots, t_r$, $1 \le t_\nu \le N$, with the property that

$$(2.1) \qquad \xi_{t_1+\kappa}^{(\sigma_1)} = \xi_{t_2+\kappa}^{(\sigma_2)} = \cdots = \xi_{t_r+\kappa}^{(\sigma_r)}, \qquad 0 \le \kappa \le k - 1.$$

We abbreviate $(\sigma_1, \dots, \sigma_r) = \boldsymbol{\sigma}$ and the corresponding positions by $(t_1, \dots, t_r) = \mathbf{t}$.

Let $A(k;\, \boldsymbol{\sigma}, \mathbf{t})$ indicate the event that (2.1) holds. The principal focus of this paper will concern the asymptotic ($N \to \infty$) distributional properties of the random variable

$$K(N) = K_{r,s}(N) = \max\{k | A(k;\, \boldsymbol{\sigma}, \mathbf{t}) \text{ holds for some } r \text{ sequences } \boldsymbol{\sigma} \text{ at some positions } \mathbf{t}\}. \tag{2.2}$$

The probability of observing a given word w in the sequence $\mathscr{S}_\sigma$ at position t is denoted by $\mathrm{Pr}_\sigma\{w;\, t\}$. The probability of realizing a common word of length k at positions t_ν in the sequences $\mathscr{S}_{\sigma_\nu}$ (as the sequences are independently generated) is $\sum_{w \in \mathscr{W}_k} \prod_{\nu=1}^{r} \mathrm{Pr}_{\sigma_\nu}\{w;\, t_\nu\}$, where the set $\mathscr{W}_k$ encompasses all words of length k.

We write

$$\mathrm{Pr}_{\boldsymbol{\sigma}}\{w\} = \mathrm{Pr}_{(\sigma_1, \dots, \sigma_r)}\{w\} = \prod_{\nu=1}^{r} \mathrm{Pr}_{\sigma_\nu}\{w;\, t_\nu\} = \prod_{\nu=1}^{r} \mathrm{Pr}_{\sigma_\nu}\{w\}, \tag{2.3}$$

suppressing the t_ν in the last equation due to the assumption of stationarity.

Definition 2.1. Given $r \geq 2$ independent stationary letter sequences $\mathscr{S}_{\sigma_\nu} = \{\xi_t^{(\sigma_\nu)}\}$, $\nu = 1, \dots, r$, governed by measures $\mathrm{Pr}_{\sigma_1}, \dots, \mathrm{Pr}_{\sigma_r}$, we refer to

$$F_{\boldsymbol{\sigma}}(k) = F_{(\sigma_1, \dots, \sigma_r)}(k) = \sum_{w \in \mathscr{W}_k} \prod_{\nu=1}^{r} \mathrm{Pr}_{\sigma_\nu}\{w\}, \tag{2.4}$$

as the *local match distribution*.

We have

Theorem 2.1. *If the sequences $\mathscr{S}_\sigma$ are stationary, uniformly mixing and positive* [*as defined in* (1.1) *and* (1.2)], *then the limiting geometric mean probability of a local match between sequences,*

$$\lim_{k \to \infty} [F_{\boldsymbol{\sigma}}(k)]^{1/k} = \lambda(\sigma_1, \dots, \sigma_r) = \lambda(\boldsymbol{\sigma}), \quad \textit{with } 0 < \lambda(\boldsymbol{\sigma}) < 1, \textit{ exists.} \tag{2.5}$$

When all the processes $\{\mathscr{S}_\sigma\}$ are generated by the same probability measure $P = \mathrm{Pr}$, we write for (2.5)

$$\lambda^{[r]} = \lambda(P^{[r]}) = \lim_{k \to \infty} \left(\sum_{w \in \mathscr{W}_k} [\mathrm{Pr}\{w\}]^r \right)^{1/k}. \tag{2.6}$$

In Theorem 3.2 it is proved that the limit

$$\mu(P) = \lim_{r \to \infty} [\lambda(P^{[r]})]^{1/r} \quad \text{exists and } 0 < \mu(P) < 1. \tag{2.7}$$

The value $\mu(P)$ can be interpreted as the *limiting geometric mean probability of the most likely word* on a sequence with measure P.

For the case of Markov sequences with underlying transition matrices $P(\sigma_\nu)$, $\lambda(\boldsymbol{\sigma}) = \lambda(P(\sigma_1) \circ P(\sigma_2) \circ \cdots \circ P(\sigma_r))$ is the spectral radius of the indicated Schur product matrix. (The Schur product $P \circ Q$ is the matrix obtained from elementwise multiplication of P and Q.)

We further assume for each r-tuple of sequences the existence of the limit

$$\lim_{k\to\infty}\frac{F_{\sigma}(k)}{[\lambda(\sigma)]^{k}}=\gamma(\sigma_1,\dots,\sigma_r)=\gamma(\sigma)>0. \tag{2.8}$$

The condition (2.8) is satisfied for Markov sequences provided the Schur product matrix $P(\sigma_1)\circ P(\sigma_2)\circ\cdots\circ P(\sigma_r)$ is primitive (irreducible aperiodic).

It is convenient to define $S_1(N,k)=\Sigma_{\sigma,\mathbf{t}}\Pr\{A(k;\,\sigma,\mathbf{t})\}$ which is the expected count of common k-words among the s sequences $\{\mathscr{S}_{\sigma}\}$. The correct order relation between k and N as will be established in Section 5 is

$$k=\left[\frac{\log N^{r}}{-\log\lambda}+x+1\right]=\frac{\log N^{r}}{-\log\lambda}+x+\rho(N,x) \tag{2.9}$$

($[h]$ denotes the integer part of the quantity h) such that $0<\rho(N,x)\le 1$.

In the case of identically distributed stationary sequences, $\Pr\{A(k;\,\sigma,\mathbf{t})\}$ is independent of σ and $\mathbf{t}$ and

$$\begin{aligned}S_1(N,k)&=N^{r}\binom{s}{r}\Pr\{A(k)\}=N^{r}\binom{s}{r}\gamma\lambda^{k}(1+o(1))\\&=\binom{s}{r}\gamma\lambda^{x+\rho(N,x)}(1+o(1)),\end{aligned}$$

the last equations due to (2.8) provided k is given by (2.9).

For each collection $\{\mathscr{S}_{\sigma_i}\}_1^r$ we ascertain $\lambda(\sigma)$ according to (2.5) and set

$$\lambda^{*}=\max_{\sigma}\lambda(\sigma). \tag{2.10}$$

For each individual process we determine [see (2.7)] $\mu(P_{\sigma})$. The processes $\mathscr{S}_{\sigma_\nu}$ with probability measure P_{σ_ν} are said to be *proximal* in the presence of the inequalities

$$\mu(P_{\sigma})<(\lambda^{*})^{1/r},\quad \text{for } \sigma=1,2,\dots,s. \tag{2.11}$$

When all the separate processes are identically distributed, then (2.11) *holds and these are proximal as they should be.*

For the special case where each P underlies i.i.d. letters, such that the probability of sampling letter a_i has probability $p_i^{(\sigma)}$, $i=1,2,\dots,m$, then $\mu(P_{\sigma})=\max_{1\le i\le m}(p_i^{(\sigma)})$ and (2.11) reduces to

$$\max_{i,\sigma}\left(p_i^{(\sigma)}\right)<\max_{\sigma}\left[\sum_{i=1}^{m}\prod_{\nu=1}^{r}p_i^{(\sigma_\nu)}\right]^{1/r}.$$

We are now prepared to state the main theorem of this paper.

THEOREM 2.2. *Consider s independent stationary letter sequences obeying* (1.1) *and* (1.2). *Denote for all $1\le\sigma_1<\sigma_2<\cdots<\sigma_r\le s$ by $\lambda(\sigma)$ the limit of* (2.5) *and assume the existence of* (2.8). *Let $\lambda^{*}=\max_{\sigma}\lambda(\sigma)$ [see* (2.10)*] and set $\gamma^{*}=\Sigma^{*}\gamma(\sigma)$, where the sum extends over those sets of indices σ satisfying $\lambda(\sigma)=\lambda^{*}$.*

Let r be given. Assume the processes $\{\mathscr{S}_\sigma\}_1^s$ are proximal in the sense of (2.11). *Then as $N \to \infty$, k determined as in* (2.9), *we have*

$$\lim_{N\to\infty}\left[\Pr\left\{K_{r,s}(N) < \left[\frac{\log N^r}{-\log\lambda^*} + x + 1\right]\right\} - \exp\{-(1-\lambda^*)S_1(N,k)\}\right] = 0, \tag{2.12}$$

or in an equivalent form,

$$\frac{-\log\Pr\{K_{r,s}(N) < k\}}{(\lambda^*)^{\rho(N,x)}} \xrightarrow[N\to\infty]{} (1-\lambda^*)\gamma^*(\lambda^*)^x, \tag{2.13}$$

with $\rho(N,x)$ uniquely determined by the identity

$$\left[\frac{\log N^r}{(-\log\lambda^*)} + x + 1\right] = \frac{\log N^r}{(-\log\lambda^*)} + x + \rho(N,x).$$

REMARK. We indicate the modification in the statement of (2.10) for unequal sequence lengths $N_1, N_2, \dots, N_s$. Assume all N_i are of the same order such that as $N \to \infty$, $(N_{\sigma_1}N_{\sigma_2}\cdots N_{\sigma_r})/N^r$ converges to $\alpha(\sigma)$, $0 < \alpha(\sigma) < \infty$. Adjust the definition of γ^* to be $\gamma^* = \Sigma_{\tilde\sigma}\gamma(\tilde\sigma)\alpha(\tilde\sigma)$, where the sum extends as previously over those $\tilde\sigma$ for which $\lambda(\tilde\sigma) = \lambda^* = \max_\sigma\lambda(\sigma)$. With these modifications the limit relations (2.12) and (2.13) hold.

3. The local match distribution and properties of mixing sequences.

Consider s independent stationary uniformly mixing sequences $\mathscr{S} = \{\xi_t^{(\sigma)}, t \in \mathbb{N}^+, \sigma = 1,\dots,s\}$. We start with a general inequality emanating from the simple mixing assumption (1.1).

The total set of letter variables may be designated in the form $\mathscr{I} = \{(\sigma,t) | \sigma \in \{1,\dots,s\}, t \in \mathbb{N}^+\}$.

LEMMA 3.1. *Let δ, $0 < \delta < 1$, $d_\sigma(\delta) \in \mathbb{N}$ be determined such that* (1.1) *holds for $d > d_\sigma(\delta)$. Define $d(\delta) = \max_\sigma d_\sigma(\delta)$. Let $I_1,\dots,I_r \subset \mathscr{I}$ be disjoint sets of indices satisfying the property that for all $\sigma = 1,\dots,s$, $\rho_1 \neq \rho_2$,*

$$(\sigma,t_1) \in I_{\rho_1}, (\sigma,t_2) \in I_{\rho_2} \text{ entails } |t_1 - t_2| > d(\delta). \tag{3.1}$$

A set $\{(\sigma,t+1),(\sigma,t+2),\dots,(\sigma,t+\Delta)\}$ of indices is called a gap segment relative to $\{I_\rho\}$ when (σ,t) and $(\sigma,t+\Delta+1)$ belong to index sets I_{ρ_1} and I_{ρ_2}, respectively, $\rho_1 \neq \rho_2$, while $(\sigma,t+\nu) \notin \bigcup_{\rho=1}^r I_\rho$ for all $\nu = 1,\dots,\Delta$. Certainly (3.1) *implies $\Delta \geq d(\delta)$. Define n^* to be the maximum number of different gap segments corresponding to the index sets $I_1,\dots,I_r$.*

Then

$$\left(\frac{1-\delta}{1+\delta}\right)^{n^*}\prod_{i=1}^r\Pr\{A_i\} \leq \Pr\left\{\bigcap_{i=1}^r A_i\right\} \leq \left(\frac{1+\delta}{1-\delta}\right)^{n^*}\prod_{i=1}^r\Pr\{A_i\}, \tag{3.2}$$

for all

$$A_\rho \in \mathscr{F}\{\xi_t^{(\sigma)} | (\sigma, t) \in I_\rho\}, \qquad \rho = 1, \dots, r.$$

The proof of this lemma is given in Karlin and Ost (1987).

In addition to the asymptotic independence probability inequalities (3.2), we have the general boundedness inequalities described next.

LEMMA 3.2. *Assume $\mathscr{S} = \{\xi_t\}_{t=1}^\infty$ is the realization of a uniformly mixing positive letter sequence. Let A and B be two events defined on ordered sets of positions $t_1 < t_2 < \cdots < t_\nu$ and $s_1 < s_2 < \cdots < s_\mu$, such that $t_\nu < s_1$. There exists a uniform constant $K > 0$ such that*

$$\Pr\{A \cap B\} \le K \Pr\{A\}\Pr\{B\}. \tag{3.3a}$$

For events $A_1, A_2, \dots, A_\rho$ defined on mutually disjoint ordered sets of position indices $I_1, \dots, I_\rho$ with $i_\alpha < i_\beta$ for $i_\alpha \in I_\alpha$, $i_\beta \in I_\beta$, $\alpha < \beta$, then

$$\Pr\{A_1 \cap A_2 \cap \cdots \cap A_\rho\} \le K \prod_{j=1}^{\rho} \Pr\{A_j\}, \tag{3.3b}$$

where K is an absolute constant.

The probability of realizing some common word of length k at the positions $t_1, \dots, t_r$ for the sequences $\mathscr{S}_\sigma$, $\sigma = 1, \dots, r$, respectively, is

$$F_{[r]}(k) = \sum_{w \in \mathscr{W}_k} \prod_{\sigma=1}^{r} \Pr_\sigma\{w; t_\sigma\} \quad [\text{cf. } (2.4)]. \tag{3.4}$$

In the stationary case this expression does not depend on $t_1, t_2, \dots, t_r$ and unless stated otherwise we take $t_1 = t_2 = \cdots = t_r = 1$ with this index suppressed. Let $X^{[r]}$ be the random variable indicating the length of a word match over the r sequences $\{\mathscr{S}_\sigma\}_{\sigma=1}^r$. Then $F_{[r]}(\cdot)$ is the distribution function of $X^{[r]}$ called the *local match distribution.*

For our objectives we can weaken the positivity condition (1.2) to the following form. We require for the collection of processes $\xi^{(\rho)}$,

$$\begin{aligned}\Pr\{\xi_{k+1}^{(\rho)} &\text{ match} | \xi_t^{(\rho)} \text{ match at indices} \\ &t = 1, \dots, k, k+d+1, \dots, k+l \text{ for } \rho = \sigma_1, \sigma_2, \dots, \sigma_r\} \ge \varepsilon,\end{aligned} \tag{3.5}$$

where ε is a positive constant independent of k, d, l and σ. This property is satisfied in comparing identically generated stationary letter processes and for independent Markov sequences which are jointly aperiodic and irreducible. Certainly (1.2) implies (3.5).

We next state a key theorem on the local match distribution.

THEOREM 3.1. *Consider the match random variable $X^{[r]}$ associated with r stationary sequences taking values in a finite state space obeying the uniform* *mixing condition* (1.1) *and the positivity condition* (3.5) *or* (1.2). *Then*

(i)

$$\lim_{k\to\infty}\left(\Pr\{X^{[r]}\ge k\}\right)^{1/k} = \lim_{k\to\infty}\left(\sum_{w\in\mathscr{W}_k}\prod_{i=1}^{r}\Pr_{\sigma_i}\{w\}\right)^{1/k} = \lambda^{[r]} = \lambda(\sigma) = \lambda, \tag{3.6}$$

with $0 < \lambda^{[r]} < 1$. [*In the Markov chain case* $\lambda^{[r]}$ *is explicitly given by* $\lambda^{[r]} = \lambda(P(1)\circ P(2)\circ\cdots\circ P(r))$, *equal to the spectral radius of the Schur product matrix* $P(1)\circ P(2)\circ\cdots\circ P(r)$, *where* $P(\rho)$ *is the Markov transition matrix of the* $\xi^{(\rho)}$ *process.*]

(ii) $\Pr\{X^{[r]}\ge k\}/(\lambda^{[r]})^k$ *is bounded above and away from* 0.

The λ value of (3.6) is referred to as the *characteristic match parameter*.

PROOF OF (i). Consider

$$\Pr\{X^{[r]}\ge k+l\} = \Pr\{\xi_t^{(1)} = \cdots = \xi_t^{(r)},\ t = 1,\dots,k+l\},$$

which is obviously estimated above by $\le \Pr\{\xi_t^{(1)} = \cdots = \xi_t^{(r)},\ t = 1,\dots,k,k+d+1,\dots,k+l\}$. By virtue of uniform mixing and stationarity for $\delta > 0$, there exists an integer $d(\delta)$ such that

$$\Pr\{X^{[r]}\ge k+l\} \le \left(\frac{1+\delta}{1-\delta}\right)^r \Pr\{X^{[r]}\ge k\}\Pr\{X^{[r]}\ge l-d\}. \tag{3.7}$$

By virtue of the positivity condition (3.5) we have (suppressing the superscript $[r]$)

$$\Pr\{X\ge l-d\} \le \frac{1}{\varepsilon^d}\Pr\{X\ge l\}. \tag{3.8}$$

Introducing the function $f(k) = -\log\Pr\{X\ge k\}$ and combining (3.7) and (3.8) leads to the set of inequalities

$$f(k+l) \ge f(k) + f(l) - \alpha, \quad \text{for all integers } k,l\ge 0, \tag{3.9}$$

with α a fixed positive constant.

Next, we find for events $A = \{\xi_t^{(\rho)},\ \rho = 1,\dots,r$ match at positions $t = 1,2,\dots,k,k+d+1,k+d+2,\dots,k+l\}$, $B = \{\xi_t^{(\rho)}$ match at $t = k+1,\dots,k+d\}$ such that $\Pr\{X\ge k+l\} = \Pr\{A\cap B\} = \Pr\{B|A\}\Pr\{A\}$. The first factor is bounded by ε^d [due to (3.5)] while for the second factor, owing to mixing and a trivial inequality,

$$\begin{aligned}\Pr\{A\} &\ge \left(\frac{1+\delta}{1-\delta}\right)^{-r}\Pr\{X\ge k\}\Pr\{X\ge l-d\}\\ &\ge \left(\frac{1+\delta}{1-\delta}\right)^{-r}\Pr\{X\ge k\}\Pr\{X\ge l\}.\end{aligned} \tag{3.10}$$

Therefore

$$f(k+l) \le f(k) + f(l) + \alpha', \tag{3.11}$$

with some positive constant α'. Without loss of generality we can take $\alpha' = \alpha$. The relations (3.9) and (3.11) entail that

$$\begin{aligned} &g(k) = f(k) + \alpha \text{ is subadditive} \quad \text{and} \\ &\tilde{g}(k) = f(k) - \alpha \text{ is superadditive.} \end{aligned} \tag{3.12}$$

From standard properties of such functions, we have

$$\lim_{k\to\infty} \frac{g(k)}{k} = \inf_{k\ge 1} \frac{g(k)}{k} = \inf_{k\ge 1} \frac{f(k)+\alpha}{k} = \lim_{k\to\infty} \frac{f(k)}{k} = c \ge 0 \tag{3.13}$$

and

$$\lim_{k\to\infty} \frac{f(k)}{k} = \lim_{k\to\infty} \frac{\tilde{g}(k)}{k} = \sup_{k\ge 1} \frac{\tilde{g}(k)}{k} = \sup_{k\ge 1} \frac{f(k)-\alpha}{k} = c. \tag{3.14}$$

We claim that $c > 0$. Suppose to the contrary that $c = 0$, then by virtue of the inf and sup characterization of the limits in (3.13) and (3.14), respectively, we necessarily have $-\alpha \le f(k) \le \alpha$ for all $k \ge 1$. We argue next that $f(k)$ is unbounded by virtue of the property

$$\Pr\{X^{[r]} \ge k\} \to 0, \quad \text{as } k \to \infty. \tag{3.15}$$

Indeed, on account of the uniformly mixing attribute we know for any $\delta > 0$ and the corresponding $d(\delta)$ that

$$\begin{aligned} &\Pr\{X^{[r]} \ge n(d+1)\} \\ &\quad \le \Pr\left\{\xi^{(1)}_{\nu(d+1)} = \xi^{(2)}_{\nu(d+1)} = \cdots = \xi^{(r)}_{\nu(d+1)},\ \nu = 1,2,\dots,n\right\} \\ &\quad \le (1+\delta)^{rn}\left[\Pr\{X^{[r]} \ge 1\}\right]^n. \end{aligned}$$

But $\Pr\{X^{[r]} \ge 1\} < 1$ implying for δ small enough $\Pr\{X^{[r]} \ge n(d+1)\} \le [b(1+\delta)^r]^n = v^n$ with $v < 1$ and thereby (3.15) obtains. Since $f(k) = -\log \Pr\{X^{[r]} \ge k\}$ is unbounded, precluding that $-\alpha \le f(k) \le \alpha$ holds for all k, it follows that

$$\lim_{k\to\infty} \frac{f(k)}{k} = c^{[r]} = c, \quad \text{with } 0 < c < \infty \tag{3.16}$$

and synonymously

$$\lim_{k\to\infty} \left(\Pr\{X^{[r]} \ge k\}\right)^{1/k} = \lambda^{[r]} = v^{-c}. \tag{3.17}$$

Proof of (ii). We assert for c determined as in (3.16) that

$$u(k) = f(k) - ck \text{ is bounded.} \tag{3.18}$$

Suppose to the contrary that $u(k)$ is unbounded above. Then the superadditive function $u(k) - \alpha = \tilde{g}(k) - ck$ is also unbounded and therefore

$$\lim_{k\to\infty} \frac{u(k) - \alpha}{k} = \sup_{k\geq 1} \frac{\tilde{g}(k) - ck}{k} > 0,$$

which contradicts the limit relation

$$\lim_{k\to\infty} \frac{\tilde{g}(k)}{k} = c.$$

In a similar manner we deduce that $f(k) - ck$ is bounded below.

The conclusion (3.18) is equivalent to the assertion that with $\lambda = \lambda^{[r]}$, $\Pr\{X \geq k\}/\lambda^k$ is bounded away from 0 and ∞. This completes the proof of Theorem 3.1. □

The convergence

$$\text{(3.19)}\qquad \frac{\Pr\{X^{[r]} \geq k\}}{(\lambda^{[r]})^k} = \frac{F(k)}{(\lambda^{[r]})^k} \xrightarrow[k\to\infty]{} \gamma^{[r]} = \gamma(P(1),\dots,P(r)) = \gamma > 0$$

holds in many models but *not* always. A collection of sequence processes satisfying (3.19) is *called a* γ*-process*. When each $\mathscr{S}_\sigma$ corresponds to a Markov chain with transition matrix $P(\sigma)$ and every Schur product matrix $P(\sigma) = P(\sigma_1)\circ P(\sigma_2)\circ \cdots \circ P(\sigma_r)$ is primitive, then the system $\{\mathscr{S}_\sigma\}$ qualifies as a γ-process [see Karlin and Ost (1987)].

THEOREM 3.2. *In the context of Theorem* 3.1, *where the processes* $\mathscr{S}_\sigma$ *are governed by the same measure* $P = \Pr$, *we denote the characteristic match parameter by* $\lambda^{[r]} = \lambda(P^{[r]}) = \lambda(P,\dots,P)$ *and then for* $r \geq 1$ *we have* [*compare to Karlin and Ost* (1985)].

$$\text{(3.20)(i)}\qquad \lambda^{[r_1+r_2]} \leq \lambda^{[r_1]}\lambda^{[r_2]} \qquad (\log \textit{subadditivity}),$$

$$\text{(3.20)(ii)}\qquad (\lambda^{[r]})^2 < \lambda^{[r-1]}\lambda^{[r+1]} \qquad (\textit{strict} \log \textit{convexity}),$$

$$\text{(3.20)(iii)}\qquad (\lambda^{[r]})^{1/(r-1)} \quad \textit{increases while } (\lambda^{[r]})^{1/r} \textit{ strictly decreases}.$$

For (3.20)(ii), we apply the Schwarz inequality to $(\Sigma_{w\in\mathscr{W}_k}(\Pr\{w\})^r)^2$ in conjunction with the limit relation (3.6).

With (3.20)(i) and (3.20)(ii) the monotonicity properties of (3.20)(iii) without strictness follow.

Because the processes $\{\mathscr{S}_\sigma\}$ are independent, stationary, and identically distributed, the positivity postulate (3.5) is automatic. From this postulate with suitable $\varepsilon = \varepsilon_{[r]} > 0$, we deduce that

$$\text{(3.21)}\qquad \Pr\{X^{[r]} \geq k\} \leq (1-\varepsilon)^k.$$

Comparing to (3.6) we infer that $\lambda^{[r]} \leq 1 - \varepsilon$. Certainly the exponential polynomial $\Sigma_{w\in\mathscr{W}_k}[\Pr\{w\}]^r$ is analytic in the variable r and consequently $\lambda^{[r]}$ is analytic and log convex. Since $\lambda^{[1]} = 1$ and $\lambda^{[r]} < 1$ for integral $r \geq 2$ by virtue of (3.21) it follows that $\lambda^{[r]}$ is strictly log convex and concomitantly $(\lambda^{[r]})^{1/r}$ is strictly decreasing.

Define

$$\mu = \lim_{r\to\infty} (\lambda^{[r]})^{1/r} = \inf_{r\geq 1} (\lambda^{[r]})^{1/r}. \tag{3.22}$$

We know from (3.20)(iii) that $0 < \mu < 1$.

THEOREM 3.3. *The μ parameter of (3.22) is the limiting geometric mean probability of the most likely word interpreted by the formula*

$$\mu = \lim_{k\to\infty} \Big(\max_{w\in\mathscr{W}_k} \Pr\{w\}\Big)^{1/k}. \tag{3.23}$$

Moreover,

$$\max_{w\in\mathscr{W}_k} \frac{\Pr\{w\}}{\mu^k} \text{ is bounded above.} \tag{3.24}$$

PROOF OF (3.23). Obviously,

$$\sum_{w\in\mathscr{W}_k} (\Pr\{w\})^{r+1} \leq \Big(\max_{w\in\mathscr{W}_k} \Pr\{w\}\Big)^{r},$$

and now on account of (3.4), (3.6) and (3.22) we obtain

$$\mu \leq \liminf_{k\to\infty} \Big(\max_{w\in\mathscr{W}_k} \Pr\{w\}\Big)^{1/k}. \tag{3.25}$$

Plainly for each r

$$(\lambda^{[r]})^{1/r} = \lim_{k\to\infty} \Big(\sum_{w\in\mathscr{W}_k} (\Pr\{w\})^{r}\Big)^{1/kr} \geq \limsup_{k\to\infty} \Big(\max_{w\in\mathscr{W}_k} \Pr\{w\}\Big)^{1/k}. \tag{3.26}$$

Letting $r \to \infty$ and comparing with (3.25) the conclusion of (3.23) ensues.

PROOF OF (3.24). Define $G(k) = \max_{w\in\mathscr{W}_k} \Pr\{w\}$. Consider the $(k + l)$-word $u = (a_1, \ldots, a_{k+l})$ in $\mathscr{W}_{k+l}$ and set $v = (a_1, \ldots, a_k)$ as the word of the first k letters of u, w the word of the last l letters of u, and $\bar{w} = (a_{k+d+1}, \ldots, a_{k+l})$ the contracted word of the final $l - d$ letters. For given δ and $d = d(\delta)$ determined such that (1.1) applies we have $\Pr\{u\} \leq ((1 + \delta)/(1 - \delta))\Pr\{v\}\Pr\{\bar{w}\}$ and therefore $G(k + l) \leq ((1 + \delta)/(1 - \delta))G(k)G(l - d)$. Invoking the positivity postulate (3.5) or (1.2) and specifying $\bar{w}$ as the word which gives $G(l - d)$ we get $G(l) \geq \Pr\{w\} \geq G(l - d)\varepsilon^d$. These combined give $G(k + l) \leq ((1 + \delta)/(1 - \delta))G(k)G(l)\varepsilon^{-d}$, and expressed in terms of $H(k) = \log G(k)$ we have the subadditive inequality

$$H(k + l) \leq H(k) + H(l) + \alpha, \qquad k, l \geq 1, \tag{3.27}$$

for α a suitable positive constant.

On the other hand, with $z = (a_{k+1}, \dots, a_{k+d})$ and again appealing to the positivity postulate (3.5) [or (1.2)] and (1.1), we obtain

$$\begin{aligned} G(k+l) &\geq \Pr\{uz\bar{w}\} = \Pr\{u\bar{w}\}\Pr\{z|u \text{ and } \bar{w}\} \\ &\geq \Pr\{u\bar{w}\}\varepsilon^d \geq \frac{1-\delta}{1+\delta}\Pr\{u\}\Pr\{\bar{w}\}\varepsilon^d \end{aligned} \tag{3.28}$$

and therefore

$$\geq \frac{1-\delta}{1+\delta}G(k)G(l-d)\varepsilon^d \geq \left(\frac{1-\delta}{1+\delta}\right)\varepsilon^d G(k)G(l).$$

This leads to the opposite inequality of (3.27), namely,

$$H(k+l) \geq H(k) + H(l) - \alpha. \tag{3.29}$$

It is proved in the course of the development of Theorem 3.1 that the pair of complementary subadditivity inequalities (3.27) and (3.29) imply that

$$\lim_{k \to \infty} [G(k)]^{1/k} \text{ exists,} \tag{3.30}$$

which we have already identified in (3.22), (3.25) and (3.26) as μ. The foregoing analysis embodies the result of (3.24). □

We conclude this section with a basic lemma on the decay of probabilities when an additional match between two collections of letter variables is imposed.

Lemma 3.3. *Consider $s \geq 1$ positive stationary sequences $\mathscr{S}_\sigma$ satisfying* (1.1) *and* (1.2) $\{\xi_t^{(\sigma)}|\sigma = 1, \dots, s,\ t \in \mathbb{N}\}$, *with values from $\mathscr{A}$. Let J_1 and J_2 contained in $\mathscr{I} = \{(\sigma, t)|\sigma \in \{1, \dots, s\},\ t \in \mathbb{N}\}$ be disjoint sets of indices with the number of indices of J_2, $|J_2| \leq r$, bounded by a given r.*

Consider any event A of the processes $\{\mathscr{S}_\sigma\}$ defined on the indices exterior to $J_1 \cup J_2$. For any set of indices I we designate the event $\mathscr{M}(I)$ to indicate the existence of some $a \in \mathscr{A}$ such that $\xi_t^{(\sigma)} = a$ for all $(\sigma, t) \in I$, i.e., the processes match on I when I contains two or more indices and take on a prescribed letter value when I consists of a single position. There exists a constant $\beta < 1$ satisfying

$$\Pr\{\mathscr{M}(J_1 \cup J_2)|\mathscr{M}(J_1), \mathscr{M}(J_2), A\} \leq \beta, \tag{3.31}$$

with β depending only on r and s and the underlying probability measures of $\{\mathscr{S}_\sigma\}$.

[In other words, the condition $\mathscr{M}(J_1)$ asserts a match on the index set J_1 and the condition $\mathscr{M}(J_2)$ asserts a match on the index set J_2 but not necessarily the same match. Then the event of the same match has probability less than $\beta < 1$.]

Proof. Let $\varepsilon > 0$ be a positive constant ensuring (1.2) with any conditioning simultaneously for all the sequences $\mathscr{S}_\sigma = \{\xi_t^{(\sigma)}\}$. Let $\mathscr{M}_a(J)$ for an index set J signify the event of a match coincident with letter a occurring at the indices of

J. Then

$$\begin{aligned}
&\Pr\{\mathscr{M}(J_1 \cup J_2)|\mathscr{M}(J_1), \mathscr{M}(J_2), A\}\\
&\quad = \frac{\Pr\{\mathscr{M}(J_1 \cup J_2)|\mathscr{M}(J_2), A\}}{\Pr\{\mathscr{M}(J_1)|\mathscr{M}(J_2), A\}}\\
&\quad = \frac{\Sigma_a \Pr\{\mathscr{M}_a(J_1 \cup J_2)|\mathscr{M}(J_2), A\}}{\Pr\{\mathscr{M}(J_1)|\mathscr{M}(J_2), A\}}\\
&\quad = \frac{\Sigma_a \Pr\{\mathscr{M}_a(J_1 \cup J_2)|\mathscr{M}_a(J_1), \mathscr{M}(J_2), A\}\Pr\{\mathscr{M}_a(J_1)|\mathscr{M}(J_2), A\}}{\Pr\{\mathscr{M}(J_1)|\mathscr{M}(J_2), A\}}\\
&\quad \le \max_a \Pr\{\mathscr{M}_a(J_1 \cup J_2)|\mathscr{M}_a(J_1), \mathscr{M}(J_2), A\}.
\end{aligned}$$

For every $a \in \mathscr{A}$ (since J_1 and J_2 are disjoint index sets),

$$\begin{aligned}
&\Pr\{\mathscr{M}_a(J_1 \cup J_2)|\mathscr{M}_a(J_1), \mathscr{M}(J_2), A\}\\
&\quad = \frac{\Pr\{\mathscr{M}_a(J_1), \mathscr{M}_a(J_2), A\}}{\Pr\{\mathscr{M}_a(J_1), \mathscr{M}(J_2), A\}}\\
&\quad = \frac{\Pr\{\mathscr{M}_a(J_2)|\mathscr{M}_a(J_1), A\}}{\Sigma_{b \in \mathscr{A}} \Pr\{\mathscr{M}_b(J_2)|\mathscr{M}_a(J_1), A\}}.
\end{aligned} \tag{3.32}$$

Since $|J_2| \le r$ the positivity postulate (1.2) implies for every pair of letters $a, c \in \mathscr{A}$,

$$\Pr\{\mathscr{M}_c(J_2)|\mathscr{M}_a(J_1), A\} \ge \varepsilon^r. \tag{3.33}$$

Thus the right-hand side of (3.32) can be represented in the form $\varepsilon_a/(\varepsilon_a + \Sigma_{b \ne a}\varepsilon_b)$ with each $\varepsilon_c \ge \varepsilon^r$. Set

$$\beta = \sup_{\varepsilon_b, \varepsilon_a}\left(\varepsilon_a \Big/ \left(\varepsilon_a + \sum_{b \ne a} \varepsilon_b\right)\right) \le \left(1 + \frac{\varepsilon^r}{1 - \varepsilon^r}(m-1)\right)^{-1} < 1.$$

This completes the proof of Lemma 3.3. □

4. Outline of the proof of Theorem 2.2. The proof of Theorem 2.2 is elaborate, involving a number of technical lemmas set forth in Sections 5 and 6. It is helpful to review the main ideas and steps. In the present discussion we focus on the variable $K(N) = K_{r,r}(N)$. With each index set $\boldsymbol{\alpha} = (t_1, t_2, \dots, t_r)$, $1 \le t_\rho \le N$, of positions we consider the local match event

$$A_{\boldsymbol{\alpha}}(k) = \left\{\xi^{(1)}_{t_1+\kappa} = \cdots = \xi^{(r)}_{t_r+\kappa},\ \kappa = 0, 1, \dots, k-1\right\}, \tag{4.1}$$

indicating a coincident word of length k appearing at positions $t_1, t_2, \dots, t_r$ in the sequences $\mathscr{S}_1, \dots, \mathscr{S}_r$, respectively. We proved in Theorem 3.1 for every fixed $\boldsymbol{\alpha}$,

$$\lim_{k \to \infty} (\Pr\{A_{\boldsymbol{\alpha}}(k)\})^{1/k} = \lambda, \qquad 0 < \lambda < 1. \tag{4.2}$$

The bar $\{\overline{\xi^{(1)}_{t_1} = \xi^{(2)}_{t_2} = \cdots = \xi^{(r)}_{t_r}}\}$ shall signify that the sequences do not perfectly match at the index positions α. It is convenient to introduce for $\alpha = (t_1, t_2, \dots, t_r)$, $1 \le t_\rho \le N$, the constrained local event

$$(4.3) \qquad \hat{A}_\alpha(k) = \left\{\overline{\xi^{(1)}_{t_1-1} = \xi^{(2)}_{t_2-1} = \cdots = \xi^{(r)}_{t_r-1}}\right\} \cap A_\alpha(k),$$

indicating the occurrence of an identical k-word at position α but *preceded by a mismatch* (the condition at a zero coordinate is disregarded). We deduce easily $\lim_{k \to \infty}(\Pr\{\hat{A}_\alpha(k)\})^{1/k} = \lambda$ with the same λ.

The global event (arrange the α in lexicographic order) $A(N, k) = \bigcup A_\alpha(k)$ expresses the existence of at least one common k-word across the r sequences. Observe that $K(N) = K_{r,r}(N) = \max_k\{k | A(N, k) \text{ holds}\}$. We also form $\hat{A}(N, k) = \bigcup \hat{A}_\alpha(k)$ and define $\hat{K}(N) = \max_k\{k | \hat{A}(N, k) \text{ holds}\}$.

In order to achieve an asymptotic distribution, we take the integer k of the precise order growth

$$(4.4) \qquad k = k(N) = \left[\frac{r \log N}{-\log \lambda} + x + 1\right] = \frac{r \log N}{-\log \lambda} + x + \rho(N, x),$$

where $0 < \rho(N, x) \le 1$ is uniquely determined and λ is the characteristic match parameter of (4.2). Under the assumption that the sequences $\{\mathscr{S}_\rho\}$ constitute a γ-process [see Theorem 3.1 and discussion of (2.8)] we have

$$(4.5) \qquad \lim_{k \to \infty} \frac{\Pr\{A_\alpha(k)\}}{\lambda^k} = \gamma$$

and also

$$(4.6) \qquad \lim_{k \to \infty} \frac{\Pr\{\hat{A}_\alpha(k)\}}{\lambda^k} = (1 - \lambda)\gamma.$$

Obviously, $\hat{K}(N) \le K(N) \le \hat{K}(N) + l_0$, where l_0 is the length of the longest common word having at least one initial position equal to 1. It is easy to demonstrate that with probability 1, $l_0/K(N) \to 0$ indicating that $K(N)$ and $\hat{K}(N)$ have the same distributional limit properties.

Implementing the inclusion–exclusion method, we introduce the sums [with $k = k(N)$ of (4.4)],

$$(4.7) \qquad \hat{S}_p(N) = \hat{S}_p(N, k) = \sum_{\alpha_1 < \alpha_2 < \cdots < \alpha_p} \Pr\left\{\hat{A}_{\alpha_1}(k)\hat{A}_{\alpha_2}(k) \cdots \hat{A}_{\alpha_p}(k)\right\}, \qquad p = 1, 2, \dots .$$

The key steps proved in Sections 5 and 6 are the facts

$$(4.8)(\text{i}) \qquad \hat{S}_1(N) \text{ is bounded away from 0 and } \infty,$$

$$(4.8)(\text{ii}) \qquad \hat{S}_p(N) - \frac{(\hat{S}_1(N))^p}{p!} \xrightarrow[N \to \infty]{} 0.$$

After (4.8) is established we exploit the probability bounds (Bonferoni inequalities for unions of events) to deduce

$$(4.9) \qquad \Pr\{\hat{A}(N, k(N))\} - \{1 - \exp\{-\hat{S}_1(N, k)\} \to 0.$$

For a γ-process [i.e., (4.5) and (4.6) hold] we have the explicit limit relation $\hat{S}_1(N, k)/\lambda^{\rho(N,x)} \to \lambda^x(1-\lambda)\gamma$ and then (4.9) passes into

$$\frac{-\log \Pr\left\{K(N) \le \left[\frac{r \log N}{(-\log \lambda)} + x\right]\right\}}{\lambda^{\rho(N,x)}} \xrightarrow[N\to\infty]{} \lambda^x(1-\lambda)\gamma. \tag{4.10}$$

5. Proof of Theorem 2.2 for $r = s = 2$. It helps clarify the ideas and techniques to detail the proof first for the case of $r = s = 2$. Let $P_1 = \Pr_1$ and $P_2 = \Pr_2$ denote the probability measures underlying the two sequences $\mathscr{S}_1$ and $\mathscr{S}_2$. For ready reference we display the condition (2.11) of Theorem 2.2

$$\mu_1 = \lim_{r\to\infty} \left[\lambda(P_1^{[r]})\right]^{1/r} < \sqrt{\lambda^*}, \tag{5.1}$$

where

$$\lambda(P_1^{[r]}) = \lim_{k\to\infty}\left(\sum_{w\in\mathscr{W}_k} [\Pr_1\{w\}]^r\right)^{1/k}, \tag{5.2}$$

and as proved in Theorem 3.1,

$$\lambda = \lambda^* = \lim_{k\to\infty}\left(\sum_{w\in\mathscr{W}_k} \Pr_1\{w\}\Pr_2\{w\}\right)^{1/k}. \tag{5.3}$$

Similarly,

$$\mu_2 = \lim_{r\to\infty}\left[\lambda(P_2^{[r]})\right]^{1/r} < \sqrt{\lambda}. \tag{5.4}$$

There exists a universal constant C_1 independent of k with the property (Theorem 3.3)

$$\max_{w\in\mathscr{W}_k} \Pr_1\{w\} \le C_1\mu_1^k, \quad \text{for all } k. \tag{5.5}$$

Similarly, we have

$$\max_{w\in\mathscr{W}_k} \Pr_2\{w\} \le C_2\mu_2^k, \quad \text{for all } k. \tag{5.6}$$

In the present context, k grows with N by the relationship

$$k = \left[\frac{\log N^2}{(-\log\lambda)} + x + 1\right] = \frac{\log N^2}{(-\log\lambda)} + x + \rho(N, x), \tag{5.7}$$

where $\rho(N,x)$, $0 < \rho(N,x) \le 1$, is uniquely determined maintaining k an integer. This prescription of k entails that

(5.8) $N^2\lambda^k$ is uniformly bounded away from 0 and ∞ as $N \to \infty$.

For simplicity of exposition, we divide the proof into a series of lemmas.

Henceforth, all the hypotheses of Theorem 2.2 are in force and we will not repeat their statements.

Lemma 5.1. *For k determined as in* (5.7) *and $\hat{S}_1(N,k)$ defined in* (4.7) *we have*

$$\hat{S}_1(N,k) - \lambda^{x+\rho(N,x)}\gamma(1-\lambda) \to 0, \quad \text{as } N \to \infty. \tag{5.9}$$

[γ *is defined in* (2.8).]

Proof. Observe that

$$\begin{aligned}\hat{S}_1(N,k) &= \sum_{\alpha} \Pr\{\hat{A}_\alpha(k)\} = (N-1)^2 \Pr\{\hat{A}_{\alpha_0}(k)\}\\ &= (N-1)^2\lambda^k \frac{\Pr\{\hat{A}_{\alpha_0}(k)\}}{\lambda^k} = \left(\frac{N-1}{N}\right)^2 \lambda^{x+\rho(N,x)} \frac{\Pr\{\hat{A}_{\alpha_0}(k)\}}{\lambda^k}.\end{aligned}$$

Thus

$$\begin{aligned}&\hat{S}_1(N,k) - \lambda^{x+\rho(N,x)}\gamma(1-\lambda)\left(\frac{N-1}{N}\right)^2\\ &\quad = \left(\frac{N-1}{N}\right)^2 \lambda^{x+\rho(N,x)}\left[\frac{\Pr\{\hat{A}_{\alpha_0}(k)\}}{\lambda^k} - \gamma(1-\lambda)\right].\end{aligned} \tag{5.10}$$

This quantity goes to zero as $N \to \infty$ by virtue of (4.6) which is the assumption that the sequences $\{\mathscr{S}_1, \mathscr{S}_2\}$ are generated as a γ-process. The conclusion of (5.9) is now clear. □

We turn next to assess the asymptotic behavior of

$$\hat{S}_2(N,k) = \sum_{\alpha_1 < \alpha_2} \Pr\{\hat{A}_{\alpha_1}\hat{A}_{\alpha_2}\}$$

involving pairs of match events.

The collection of index vectors $\alpha_1 = (t_1^{(1)}, t_1^{(2)})$ and $\alpha_2 = (t_2^{(1)}, t_2^{(2)})$ are classified into four types.

First fix $\delta > 0$ and determine $d = d(\delta)$ so that the uniform mixing inequality (1.1) applies.

(i) $\hat{A}_{\alpha_1}(k)$ and $\hat{A}_{\alpha_2}(k)$ are said to be *far* [abbreviated (F)] if

$$|t_1^{(1)} - t_2^{(1)}| > k + d(\delta) \quad \text{and} \quad |t_1^{(2)} - t_2^{(2)}| > k + d(\delta) \text{ hold.} \tag{5.11}$$

(ii) $\hat{A}_{\alpha_1}(k)$ and $\hat{A}_{\alpha_2}(k)$ are *partially far* (PF) if either $|t_1^{(1)} - t_2^{(1)}| > k + d$ or $|t_1^{(2)} - t_2^{(2)}| > k + d$ but *not* both.

(iii) $\hat{A}_{\alpha_1}(k)$ and $\hat{A}_{\alpha_2}(k)$ are said to be *close and synchronous* (CS) if

$$\left|t_1^{(1)} - t_2^{(1)}\right| \le k + d \quad \text{and} \quad t_1^{(1)} - t_2^{(1)} = t_1^{(2)} - t_2^{(2)}. \tag{5.12}$$

(iv) $\hat{A}_{\alpha_1}(k)$ and $\hat{A}_{\alpha_2}(k)$ are said to be *close but asynchronous* (CA) if $|t_1^{(1)} - t_2^{(1)}| \le k + d$ and $|t_1^{(2)} - t_2^{(2)}| \le k + d$ but (5.12) *does not hold.*

We decompose the summands of $\hat{S}_2$ into four parts as follows:

$$\begin{aligned}\hat{S}_2(N,k) &= \sum_{\substack{\alpha_1 < \alpha_2\\ (\mathrm{F})}} \Pr\{\hat{A}_{\alpha_1}(k)\hat{A}_{\alpha_2}(k)\} + \sum_{(\mathrm{PF})} + \sum_{(\mathrm{CS})} + \sum_{(\mathrm{CA})}\\ &= \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}.\end{aligned} \tag{5.13}$$

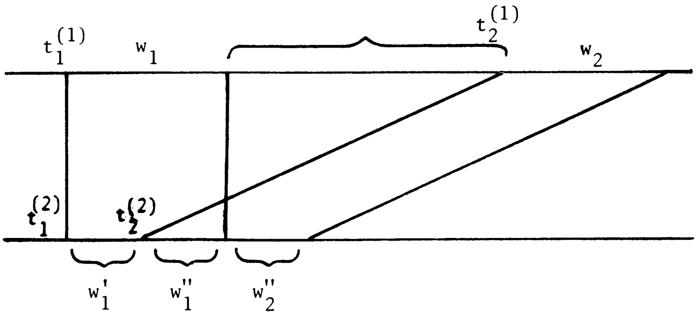
Lemma 5.2. *The sum* III *is identically zero.*

Proof. The event of $\hat{A}_\alpha(k)$ connotes a k-word match preceded by a mismatch. Accordingly close synchronous block matches entail incompatible conditions as they require a match and mismatch for the same positions. □

Lemma 5.3. *The sum* II *of* (5.13) *tends to* 0 *as* $N \to \infty$.

We refer to the diagram displaying $w_1 = w_1' w_1''$ as the word match of A_{α_1} and $w_2 = w_1'' w_2''$ the word match of A_{α_2},

(5.14)



where w_1' is of length a, w_1'' of length b, w_2'' of length c with $a + b = k$, $b + c = k$. Using the bounds of (3.3) (C stands for a generic constant which may change over successive equations)

$$\begin{aligned}\Pr\{A_{\alpha_1}A_{\alpha_2}\} &\le C \sum_{w_1', w_1'', w_2''} \Pr_1\{w_1'\}\Pr_1\{w_1''\}\Pr_1\{w_1''\}\Pr_1\{w_2''\} \\ &\qquad \times \Pr_2\{w_1'\}\Pr_2\{w_1''\}\Pr_2\{w_2''\} \\ &\le C \max_{w \in \mathscr{W}_b} \Pr_1(w)\left[\sum_{w_1'} \Pr_1\{w_1'\}\Pr_2\{w_1'\}\right] \\ &\qquad \times \left[\sum_{w_1''} \Pr_1\{w_1''\}\Pr_2\{w_1''\}\right]\left[\sum_{w_2''} \Pr_1\{w_2''\}\Pr_2\{w_2''\}\right] \\ &\le C\mu_1^b \lambda^a \lambda^b \lambda^c.\end{aligned}$$

Since $\mu_1 < \sqrt{\lambda}$ by hypothesis letting $\theta_1 = \mu_1/\sqrt{\lambda}$, we have the estimate

$$\Pr\{A_{\alpha_1}A_{\alpha_2}\} \le C\theta_1^b \lambda^{a+3b/2+c} = C\theta_1^b \lambda^{c/2}\lambda^{3k/2} \le C\theta^k \lambda^{3k/2}$$

having set $\max(\theta_1, \sqrt{\lambda}) = \theta$. Certainly $\theta < 1$.

Enumerating the number of such terms, we have

$$\text{sum of II} \le 4C(k + d(\delta))N^3 \theta^k \lambda^{3k/2}.$$

Since $N^3\lambda^{3k/2}$ is bounded and $(k + d(\delta))\theta^k \to 0$ the assertion of the lemma is proved.

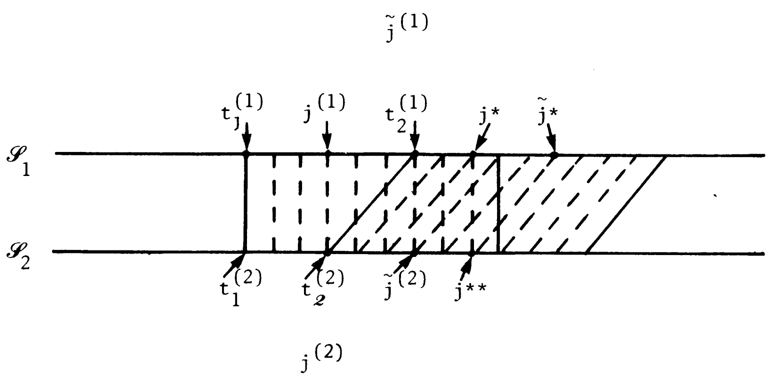
LEMMA 5.4. *The sum* IV *of* (5.13) *goes to zero as* $N \to \infty$.

PROOF. In this proof we use the result of Lemma 3.3. We will specifically prove that the terms of IV can be bounded above by

$$\Pr\{\hat{A}_{\alpha_1}(k)\hat{A}_{\alpha_2}(k)\} \le \Pr\{\hat{A}_{\alpha_1}(k)\}\beta^k, \tag{5.15}$$

with β the factor occurring in Lemma 3.3 In order to establish (5.15) it is convenient to graphically display a close asynchronous pair of word matches

(5.16)



Since α_1 and α_2 are close asynchronous we necessarily have $|t_1^{(1)} - t_2^{(1)}| \le k + d$ and $|t_1^{(2)} - t_2^{(2)}| \le k + d$, but $t_1^{(1)} - t_2^{(1)} \ne t_1^{(2)} - t_2^{(2)}$ and the drawing (5.16) is the only contingency requiring attention apart from relabeling indices.

Let J be the match positions $J = \{j^{(1)}, j^{(2)}\}$ between the two sequences in the k-word match of $\hat{A}_{\alpha_1}(k)$ with a further match condition, the match positions of $\hat{A}_{\alpha_1}(k)$ corresponding to the first match position of $\hat{A}_{\alpha_2}(k)$, namely at $\tilde{J} = \{\tilde{j}^{(1)}, \tilde{j}^{(2)}\}$. Let A indicate the event representing all the remaining identities of $\hat{A}_{\alpha_1}$ and $\hat{A}_{\alpha_2}$ excluding those of the indices of J and $\tilde{J}$ but including the condition implied by the match $\hat{A}_{\alpha_2}$ at position $\tilde{j}^{(2)}$. On the basis of (3.31) we have

$$\Pr\{\mathscr{M}(J, \tilde{J}) | \mathscr{M}(J), \mathscr{M}(\tilde{J}), A\} \le \beta < 1. \tag{5.17}$$

Therefore, in this case

$$\Pr\{\hat{A}_{\alpha_1}\hat{A}_{\alpha_2}\} < \beta \Pr\{\mathscr{M}(J), \mathscr{M}(\tilde{J}), A\}. \tag{5.18}$$

Another necessary application of Lemma 3.3 involves the index sets [see the figure in (5.16)] $J^* = \{j^*, j^{**}\}$ and $\tilde{J}^* = \{\tilde{j}^*\}$. In this case the event A includes all matches implied by $\hat{A}_{\alpha_1}(k)$ and $\hat{A}_{\alpha_2}(k)$ precluding the matches of the indices $\{j^*, j^{**}, \tilde{j}^*\}$. Appealing to Lemma 3.3, relation (3.31), gives

$\Pr\{\mathscr{M}(J^*, J^{**})|\mathscr{M}(J^*), \mathscr{M}(J^{**}), \mathrm{A}\} \le \beta$ which implies

$$\Pr\{\hat{A}_{\alpha_1}(k)\hat{A}_{\alpha_2}(k)\} \le \beta \Pr\{\mathscr{M}(J^*), \mathscr{M}(\tilde{J}^*), A\}. \tag{5.19}$$

Let $\tilde{A}_{\alpha_2}(k-1)$ be the matching event involving the same conditions $A_{\alpha_2}(k)$ excluding its first match at positions $\tilde{j}^{(1)}$ and $j^{(2)}$. Clearly,

$$\Pr\{\mathscr{M}(J), \mathscr{M}(\tilde{J}), A\} \le \Pr\{\hat{A}_{\alpha_1}(k)\tilde{A}_{\alpha_2}(k-1)\}. \tag{5.20}$$

We apply the same analysis next to the positions $j^{(1)}+1$, $j^{(2)}+1$, $\tilde{j}^{(1)}+1$, $\tilde{j}^{(2)}+1$ and continue to iterate the conclusions of (5.18) and (5.20) and the inequality of (5.19) leading directly to (5.15).

With (5.15) validated and counting the summands the series IV possesses the upper bound

$$\text{sum of IV} \le [2(k+d(\delta))]^2 N^2\beta^k \Pr\{\hat{A}_{\alpha_0}(k)\} \le Ck^2N^2\lambda^k\beta^k.$$

Since $N^2\lambda^k$ is bounded and $k^2\beta^k \to 0$ the conclusion of Lemma 5.4 is established. □

It remains to deal with the sum I of (5.13). Invoking the uniform mixing inequality (1.1) we have

$$\begin{aligned} &\frac{1}{2}\left(\frac{1-\delta}{1+\delta}\right)^2 \sum_{\substack{\alpha_1,\alpha_2\\(\mathrm{F})}} \Pr\{\hat{A}_{\alpha_1}\}\Pr\{\hat{A}_{\alpha_2}\} \\ &\le \mathrm{I} \le \frac{1}{2}\left(\frac{1+\delta}{1-\delta}\right)^2 \sum_{\substack{\alpha_1,\alpha_2\\(\mathrm{F})}} \Pr\{\hat{A}_{\alpha_1}\}\Pr\{\hat{A}_{\alpha_2}\}. \end{aligned} \tag{5.21}$$

Now the difference of

$$\sum_{\substack{\alpha_1,\alpha_2\\(\mathrm{F})}} \Pr\{\hat{A}_{\alpha_1}(k)\}\Pr\{\hat{A}_{\alpha_2}(k)\} \text{ and } (\hat{S}_1(N,k))^2$$

is bounded above by $(4(k+d)N^3)\lambda^{2k}$, adapting the argument of Lemma 5.3 [easier in the present case since each term $\Pr\{A_\alpha(k)\} \le C\lambda^k$].

Using this fact in (5.21), we easily deduce

$$\limsup_{N\to\infty}\left|\hat{S}_2(N,k) - \frac{(\hat{S}_1(N,k))^2}{2!}\right| \le C\delta. \tag{5.22}$$

But as δ is arbitrarily prescribed, it follows that

LEMMA 5.5.

$$\lim_{N\to\infty}\left[\hat{S}_2(N,k) - \frac{(\hat{S}_1(N,k))^2}{2!}\right] = 0.$$

Arranging the index vectors $\boldsymbol{\alpha}$ in order (say lexicographically) we examine next the sum

$$\hat{S}_p(N,k) = \sum_{\boldsymbol{\alpha}_1 < \boldsymbol{\alpha}_2 < \cdots < \boldsymbol{\alpha}_p} \Pr\left\{\hat{A}_{\boldsymbol{\alpha}_1}(k)\hat{A}_{\boldsymbol{\alpha}_2}(k) \cdots \hat{A}_{\boldsymbol{\alpha}_p}(k)\right\}. \tag{5.23}$$

These partition naturally into four groups of terms [cf. (5.13)]. Fix δ and determine $d(\delta)$ as previously done so that the uniformly mixing inequalities (3.2) apply.

Group I includes all index vectors $\boldsymbol{\alpha}_1 = (t_i^{(1)}, t_i^{(2)})$ *far* (F) apart, i.e.,

$$|t_i^{(\sigma)} - t_j^{(\sigma)}| \geq d + k, \qquad \sigma = 1, 2,\ i \neq j. \tag{5.24}$$

Group II consists of vector tuples embracing the combinations that are *close* and in *synchrony* such that

$$t_{\nu_i}^{(1)} - t_{\nu_j}^{(1)} = t_{\nu_i}^{(2)} - t_{\nu_j}^{(2)} \leq k, \tag{5.25}$$

for an appropriate succession $\nu_1, \nu_2, \dots, \nu_p$ which constitute a permutation of the subscripts $1, 2, \dots, p$.

Group III includes those index vectors satisfying

$$\left|t_{\nu_i}^{(1)} - t_{\nu_j}^{(1)}\right| \leq k \quad \text{and} \quad \left|t_{\mu_i}^{(2)} - t_{\mu_j}^{(2)}\right| \leq k, \tag{5.26}$$

for some successions $\nu_1, \nu_2, \dots, \nu_p$ and $\mu_1, \mu_2, \dots, \mu_p$ but

$$t_{i_0}^{(1)} - t_{j_0}^{(1)} \neq t_{i_0}^{(2)} - t_{j_0}^{(2)}, \tag{5.27}$$

for at least one pairing from $(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_p)$, say starting on $\mathscr{S}_1$ at positions $t_{i_0}^{(1)}$ and $t_{j_0}^{(1)}$. Thus, these contingencies incorporate all matching words associated with positions $(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_p)$ that are *close* but at least one pair which is *not in synchrony*.

Group IV consists of the remaining terms of (5.23).

By these criteria the sum (5.23) [analogous to (5.13)] can be divided into four sums,

$$\begin{aligned}\hat{S}_p(N,k) &= \text{I (sets of far indices)} + \text{II (close and synchronous)} \\ &\quad + \text{III (close and asynchronous)} + \text{IV (remaining terms)}.\end{aligned} \tag{5.28}$$

In computing the probabilities

$$\Pr\left\{\hat{A}_{\boldsymbol{\alpha}_1}\hat{A}_{\boldsymbol{\alpha}_2} \cdots \hat{A}_{\boldsymbol{\alpha}_p}\right\}, \tag{5.29}$$

we can assume that *no* pair among the index sets $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_p$ are close and in synchrony [see (5.25)] since otherwise the quantity (5.29) is zero for the same reason as that of Lemma 5.2. Moreover, on this basis we can delete the sum of II forthwith and also contract III and IV to III* and IV*, respectively, by removing all terms which involve at least one pair of close synchronous index vectors.

Lemma 5.6. *The sum* III* *goes to zero as* $N \to \infty$.

Proof. We are dealing here with the cumulation of all terms $\Pr\{\hat{A}_{\alpha_1}\hat{A}_{\alpha_2} \cdots \hat{A}_{\alpha_p}\}$ having index vectors with all α_i close and all pairs asynchronous. Trivially, this probability is bounded above by $\Pr\{\hat{A}_{\alpha_1}\hat{A}_{\alpha_2}\} \le C\beta^k \Pr\{\hat{A}_{\alpha_1}\}$, where the last inequality derives from (5.15). Now there are at most $(2k + 2d)^{2p}N^2$ such summands. Therefore $\text{III} = \text{III}^* \le C(2k + 2d)^{2p}N^2\beta^k\lambda^k$, which goes to zero geometrically fast since $N^2\lambda^k$ is bounded and $\beta < 1$.

The proof of Lemma 5.6 is complete. □

It is convenient henceforth to use the notation $B(\alpha_1, \ldots, \alpha_p) = \hat{A}_{\alpha_1}\hat{A}_{\alpha_2} \cdots \hat{A}_{\alpha_p}$. In analyzing the terms of IV* it is convenient to partition them further into the sums $\text{IV}^*(a, b)$ depending on the integer parameters a and b with $1 \le a, b \le p$, $a + b \le 2p - 1$. Explicitly, the sum $\text{IV}^*(a, b)$ selects all terms of IV* with the property that the position vectors of the matching words in $B(\alpha_1, \alpha_2, \ldots, \alpha_p)$ divide into a separated (far) groups $\{G_i\}_1^a$, each G_i comprised of close words on $\mathscr{S}_1$ [close shall mean here that the span of all the words of G_i covers at most $p(k + d(\delta))$ positions] but such that *all* the distances of words comparing different groups are far (at least $k + d$ apart). Similarly, there exist b separated groups $\{H_i\}_1^b$ of close words on $\mathscr{S}_2$.

We assume inductively for events involving at most $p - 1$ matching k-word pairs between the sequences $\mathscr{S}_1$ and $\mathscr{S}_2$,

$$\text{IV}^*(a, b) \le CN^{a+b}\lambda^{(k(a+b))/2}\theta^k \text{ applies provided} \qquad (5.30)$$

$$a + b \le 2p - 1,\ p \ge 2,$$

with an appropriate $\theta < 1$.

The case $p = 2$ has already been dealt with in Lemmas 5.3 and 5.4. We proceed to advance the induction from $p - 1$ to p.

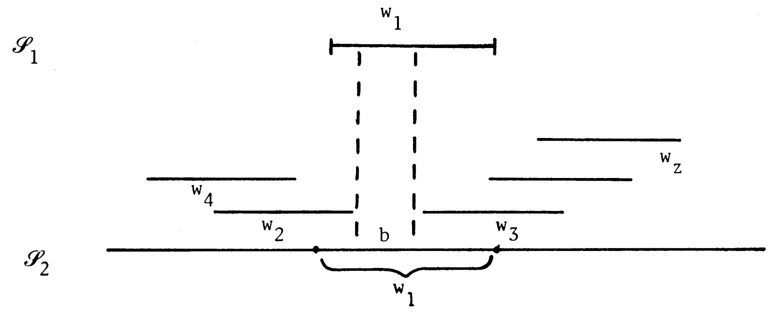
Case 1. Suppose there exists a group, say G_1 consisting of a single word and its matching word also belongs to a single separated group, say H_1. Let this word be that of the index α_1. Obviously, using the bounds (3.2), we have $\Pr\{B(\alpha_1, \ldots, \alpha_p)\} \le C\lambda^k \Pr\{B(\alpha_2, \ldots, \alpha_p)\}$. Appeal to the induction hypothesis of (5.30) applied to $\Pr\{B(\alpha_2, \ldots, \alpha_p)\}$ produces the upper estimate $C\lambda^k N^{a+b-2}\lambda^{k(a+b-2)/2}\theta^k$. The maximum number of such contingencies is at most N^2 (the choices in positions for G_1 and H_1) and therefore (5.30) prevails in this situation.

Case 2. Suppose there exists a group, say G_1 consisting of a single word (say word w_1 associated with index α_1) and its required matching word belongs to H_1 which involves $z \ge 2$ close words (see the figure below).

The analysis of Lemma 5.3 is readily adapted to establish the inequality

$$\Pr\{B(\alpha_1, \ldots, \alpha_p)\} \le C\mu_1^h\lambda^l \Pr\{B(\alpha_2, \ldots, \alpha_p)\},$$

where l is the length of the part in w_1 not covered among the other matching



words of H_1, $h = k - l$ and $\mu_1 = \lim_{k \to \infty}(\max_{w \in \mathscr{W}_k} \Pr_1\{w\})^{1/k}$. The hypothesis (5.1) and the method of Lemma 5.3 further lead to the inequality $(\mu_1)^h \lambda^l \leq \lambda^{k/2}\theta_1^k$ with $\theta_1 < 1$. Invoking the induction hypothesis

$$\sum_{\alpha_1, \dots, \alpha_p} \Pr\{B(\alpha_1, \dots, \alpha_p)\} \leq CN^{a+b-1}\lambda^{(k/2)(a+b-1)}\lambda^{k/2}\theta_2^k.$$

The number of indices of this type are at most N corresponding to the freedom of positions of G_1. Combining these estimates the induction is again advanced.

Case 3. All the groups G_i and H_j contain at least two close words. In this case we delete the matching condition represented at the positions of α_1 and obviously the number of groups of these matches remains a in $\mathscr{S}_1$ and b in $\mathscr{S}_2$ based on $B(\alpha_2, \dots, \alpha_p)$. Since $p \geq 3$, the induction hypothesis and bound (5.30) directly apply.

This completes the proof of Lemma 5.7 highlighted next.

LEMMA 5.7. *The sum* IV* $\to 0$ *as* $N \to \infty$.

We consider finally the sum I of (5.28). Employing the uniformly mixing inequalities (3.2) yields

$$\begin{aligned} \frac{1}{p!}\left(\frac{1-\delta}{1+\delta}\right)^{2(p-1)} &\sum_{\alpha_i, \text{ far}} \prod_{i=1}^{p} \Pr\{\hat{A}_{\alpha_i}\} \\ \leq \text{I} \leq \frac{1}{p!}\left(\frac{1+\delta}{1-\delta}\right)^{2(p-1)} &\sum_{\alpha_i, \text{ far}} \prod_{i=1}^{p} \Pr\{\hat{A}_{\alpha_i}\}. \end{aligned} \tag{5.31}$$

The bound on the difference

$$\left| \left(\hat{S}_1(N,k)\right)^p - \sum_{\alpha_i, \text{ far}} \prod_{i=1}^{p} \Pr\{\hat{A}_{\alpha_i}\} \right| \leq C(k+d)\lambda^{pk}N^{2p-1}, \tag{5.32}$$

and the fact of $\lambda^k N^2$ being bounded imply that (5.32) $\to 0$ as $N \to \infty$. Next, paraphrasing the proof of Lemma 5.5, we achieve the result

LEMMA 5.8. *$\hat{S}_p(N,k) - (1/p!)[\hat{S}_1(N,k)]^p \to 0$ as $N \to \infty$ with k as defined in* (5.7).

We are now prepared to complete the proof of Theorem 2.2, specifically the limit law of (4.9) and (4.10) in the case of $r = s = 2$. For this objective we use the two established facts

$$\hat{S}_p(N,k) - \frac{(\hat{S}_1(N,k))^p}{p!} \xrightarrow[N\to\infty]{} 0, \qquad p = 1,2,\dots, \tag{5.33}$$

and

$$\hat{S}_1(N,k) \text{ is bounded away from 0 and } \infty. \tag{5.34}$$

Recall that

$$\hat{A}(N,k) = \Pr\left\{K_{2,2}(N) \geq \left[\frac{\log N^2}{-\log \lambda} + x + 1\right]\right\}.$$

For any $\varepsilon > 0$ and odd $l \in \mathbb{N}$, the Bonferoni inequalities yield

$$\begin{aligned} &-\varepsilon + \sum_{p=1}^{l-1} (-1)^{p-1} \frac{(\hat{S}_1(N,k))^p}{p!} \\ &\leq \Pr\{\hat{A}(N,k)\} \leq \sum_{p=1}^{l} (-1)^{p-1} \frac{(\hat{S}_1(N,k))^p}{p!} + \varepsilon, \end{aligned} \tag{5.35}$$

for all $N \geq N(\varepsilon, l)$.

Since for x bounded the uniform inequality

$$\left| e^{-x} - 1 + x - \frac{x^2}{2!} + \cdots + (-1)^{l-1} \frac{x^l}{l!} \right| \leq \varepsilon \tag{5.36}$$

prevails, merely taking l large enough, we join the facts of (5.35) and (5.36) to deduce

$$\lim_{N\to\infty} \left[\Pr\left\{K_{2,2}(N) \leq \left[\frac{\log N^2}{-\log \lambda} + x\right]\right\} - e^{-\hat{S}_1(N,k)} \right] = 0. \tag{5.37}$$

The proof of Theorem 2.2 with $r = s = 2$, $N_1 = N_2 = N$ is complete.

With (5.37) in hand the affirmation of Theorem 2.2 or relation (4.10) follows on the basis of the asymptotics of $\hat{S}_1(N,k)$.

6. Proof of Theorem 2.2 of the maximal length word match in r out of s sequences. Consider s independent stationary letter sequences $\{\mathscr{S}_\sigma\}_1^s$ each of length N generated, respectively, by the probability measures $\Pr_\sigma$, $\sigma = 1,\dots,s$.

$K_{r,s}(N)$ is defined as the length of the maximal segmental match extant in at least $r \geq 2$ from the sequences $\{\mathscr{S}_\sigma\}_1^s$ the match starting but not necessarily completed within the positions $1, \dots, N$.

There are obviously $n(N) = \binom{s}{r} N^r$ different configurations of starting positions at which the longest word match can be realized. Each index $\boldsymbol{\alpha} = (\mathbf{t}, \boldsymbol{\sigma})$ embodies a vector position $\mathbf{t} = (t_1, \dots, t_r)$, $1 \leq t_i \leq N$, and a selection of $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_r)$ sequences from among the s sequences. We order these alternatives $\boldsymbol{\alpha} = 1, 2, \dots, n(N)$. Let $A_{\boldsymbol{\alpha}}(k)$ indicate a k-word match at index $\boldsymbol{\alpha}$ and let $\hat{A}_{\boldsymbol{\alpha}}(k) = \hat{A}_{\boldsymbol{\alpha}}$ (suppressing k when possible) signify the event of a common k-word occurring at index $\boldsymbol{\alpha}$ preceded by a mismatch position; see (4.3).

The identity of events

$$\{\hat{K}_{r,s}(N) \geq k\} = \bigcup_{\boldsymbol{\alpha}=1}^{n(N)} A_{\boldsymbol{\alpha}}(k) \tag{6.1}$$

clearly holds. The asymptotic calculation of $\Pr\{\hat{K}_{r,s}(N) \geq k\}$ is done by the inclusion–exclusion method as previously. For this end, we need to determine limits of the sums

$$\hat{S}_p(N, k) = \sum_{1 \leq \boldsymbol{\alpha}_1 < \boldsymbol{\alpha}_2 < \cdots < \boldsymbol{\alpha}_p \leq n(N)} \Pr\left\{\hat{A}_{\boldsymbol{\alpha}_1} \hat{A}_{\boldsymbol{\alpha}_2} \cdots \hat{A}_{\boldsymbol{\alpha}_p}\right\}. \tag{6.2}$$

By stationarity,

$$\hat{S}_1(N) = \sum_{\alpha=1}^{n(N)} \Pr\{\hat{A}_{\boldsymbol{\alpha}}\} = \sum_{\boldsymbol{\sigma}} N^r \Pr\{X[\boldsymbol{\sigma}] \geq k\}, \tag{6.3}$$

where $X(\sigma_1, \dots, \sigma_r) = X[\boldsymbol{\sigma}]$ is the "local word match" random variable assessing the length of a common word match across the sequences $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_r)$ constrained with a preceding mismatch. According to Theorem 3.1 for each collection of r sequences $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_r)$ we have the local match limit theorem

$$\lim_{k \to \infty} \Pr\{X[\boldsymbol{\sigma}] \geq k\}^{1/k} = \lambda(\sigma_1, \sigma_2, \dots, \sigma_r) = \lambda(\boldsymbol{\sigma}). \tag{6.4}$$

We will require the stronger property that the sequences are γ-processes [see (4.6)], meaning that

$$\frac{\Pr\{X[\boldsymbol{\sigma}] \geq k\}}{[\lambda(\boldsymbol{\sigma})]^k} \xrightarrow[k \to \infty]{} \gamma(\boldsymbol{\sigma})(1 - \lambda(\boldsymbol{\sigma})). \tag{6.5}$$

Let $\lambda^* = \lambda^{[r,s]} = \max_{\boldsymbol{\sigma}} \lambda(\boldsymbol{\sigma})$ be the maximal characteristic parameter among all collections of r out of the s sequences $\{\mathscr{S}_\sigma\}_1^s$.

Since $\Pr\{X[\boldsymbol{\sigma}] \geq k\} \approx (\lambda(\boldsymbol{\sigma}))^k$ we can expect (and this is correct) that the overall longest common words found in at least r out of s sequences $\{\mathscr{S}_\sigma\}_{\sigma=1}^s$ occur only in the collections $\boldsymbol{\sigma}$ for which $\lambda(\boldsymbol{\sigma}) = \lambda^*$.

We define $\gamma^* = \sum_{\boldsymbol{\sigma}}^* \gamma(\boldsymbol{\sigma})$, where the sum is restricted to those $\boldsymbol{\sigma}$ for which $\lambda(\boldsymbol{\sigma}) = \lambda^*$.

Set

$$k = \left[\frac{\log N^r}{-\log \lambda^*} + x + 1\right] = \frac{\log N^r}{-\log \lambda^*} + x + \rho(N, x), \tag{6.6}$$

with $0 < \rho(N, x) \leq 1$ uniquely determined so that k is an integer. For the specification (6.6), we readily deduce

$$\lim_{N\to\infty} \left[\hat{S}_1(N, k) - (1-\lambda^*)\gamma^*(\lambda^*)^{x+\rho(N,x)}\right] = 0. \tag{6.7}$$

Comparison of (6.3) and (6.7) with k determined as in (6.6) reveals that

$$N^r\lambda^k \text{ is uniformly bounded as } N \to \infty. \tag{6.8}$$

In the lemmas that follow, we develop the asymptotics of $\hat{S}_2(N, k)$. We will prove explicitly that

$$\hat{S}_2(N, k) - \frac{(\hat{S}_1(N, k))^2}{2!} \xrightarrow[N\to\infty]{} 0.$$

In evaluating

$$\hat{S}_2(N, k) = \sum_{\alpha_1 < \alpha_2} \Pr\left\{\hat{A}_{\alpha_1}\hat{A}_{\alpha_2}\right\}, \tag{6.9}$$

we partition the sum into four parts [cf. (5.13)] as follows. First, for specified $\delta > 0$ we determine $d(\delta)$ ensuring that the uniformly mixing inequalities (3.2) apply simultaneously relative to all sequences $\{\mathscr{S}_\sigma\}_1^s$.

(i) The partial sum I of (6.9) is defined to include all index pairings $\boldsymbol{\alpha}_1 = (\mathbf{t}_1, \boldsymbol{\sigma}_1)$ and $\boldsymbol{\alpha}_2 = (\mathbf{t}_2, \boldsymbol{\sigma}_2)$ for which $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are *far* apart which in the present context will mean that for those sequences represented in both $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ we have

$$\left|t_1^{(\sigma_\nu)} - t_2^{(\sigma_\nu)}\right| > k + d(\delta). \tag{6.10}$$

No restriction is imposed on $t_i^{(\sigma_0)}$ if $\mathscr{S}_{\sigma_0}$ is involved in one of the $\boldsymbol{\alpha}$ but not in the other $\boldsymbol{\alpha}$. We also exclude from the sum I all indices $\boldsymbol{\alpha} = (\mathbf{t}, \boldsymbol{\sigma})$ for which $\lambda(\boldsymbol{\sigma}) < \lambda^*$.

(ii) The sum II consists of all terms $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ of (6.9) which are *close* and *synchronous* signifying that $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2$ $(= \boldsymbol{\sigma})$ and

$$t_1^{(\sigma_i)} - t_2^{(\sigma_i)} = t_1^{(\sigma_j)} - t_2^{(\sigma_j)}, \quad \text{for all } 1 \leq i, j \leq r, \tag{6.11a}$$

$$\left|t_1^{(\sigma_i)} - t_2^{(\sigma_i)}\right| \leq k \tag{6.11b}$$

hold.

(iii) The terms of III are ascertained to be close but asynchronous. Thus, these summands embrace all pairs $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ with the properties $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2$, $|t_1^{(\sigma_i)} - t_2^{(\sigma_i)}| \leq k + d(\delta)$, $1 \leq i \leq r$, that do not belong to II.

(iv) The collection IV contains all the remaining summands.

LEMMA 6.1. *The sum* II *is identically zero.*

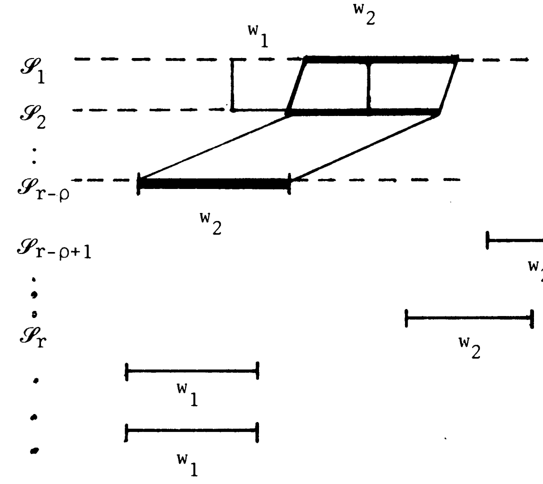
PROOF. Same as Lemma 5.2, mutatis mutandis. □

LEMMA 6.2. *The sum* III $\to 0$ *as* $N \to \infty$.

Proof. As in Lemma 5.4 we apply Lemma 3.3 to establish the estimate $\Pr\{\hat{A}_{\alpha_1}\hat{A}_{\alpha_2}\} \le \beta^k \Pr\{\hat{A}_{\alpha_1}\}$ with $\beta < 1$. An upper bound of III is therefore III $\le Ck^r\beta^k\hat{S}_1(N, k)$ and since $\beta < 1$ and $\hat{S}_1(N, k)$ is bounded, we see that the sum III $\to 0$. □

We examine next a typical term of the sum IV. Consider index vectors $\boldsymbol{\alpha}_1 = (\mathbf{t}_1, \boldsymbol{\sigma}_1)$ and $\boldsymbol{\alpha}_2 = (\mathbf{t}_2, \boldsymbol{\sigma}_2)$ satisfying $\lambda(\boldsymbol{\sigma}_1) = \lambda(\boldsymbol{\sigma}_2) = \lambda^*$. We label the sequences of $\boldsymbol{\sigma}_2 = (\sigma_1, \sigma_2, \dots, \sigma_r)$. Assume that $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ share the sequences $(\sigma_1, \dots, \sigma_{r-\rho})$, $1 \le \rho \le r - 1$, and $|t_1^{(\sigma_i)} - t_2^{(\sigma_i)}| \le k + d$, $i = 1, \dots, r - \rho$, holds while either $|t_1^{(\sigma_j)} - t_2^{(\sigma_j)}| > k + d$, $j = r - \rho + 1, \dots, r$, or only one of $\boldsymbol{\alpha}_1$ or $\boldsymbol{\alpha}_2$ has a sequence $\mathscr{S}_{\sigma_0}$ with the required word match. Note $r - \rho \ge 1$ because the summand under consideration is otherwise included in I, and $\rho \ge 1$ because otherwise the term belongs to III.

(6.12)



Say the matching word in $\hat{A}_{\alpha_1}$ is w_1 with w_2 that of $\hat{A}_{\alpha_2}$.

Overlap between the matching segments associated with $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ entails identities of some components of the words w_1 and w_2.

Let a be the cumulative length of the part of w_1 occurring in w_2 and $b = k - a$. Recall that $\mu_\sigma = \lim_{l\to\infty}(\max_{w\in\mathscr{W}_l}\Pr_\sigma\{w\})^{1/l}$ is the maximum geometric mean probability of words in the sequence $\mathscr{S}_\sigma$ (see Theorem 3.3).

Using the bounds of (3.3) and (3.24) we get

$$\Pr\{\hat{A}_{\alpha_1}\hat{A}_{\alpha_2}\} \le C\left(\prod_{j=r-\rho+1}^{r}\mu_{\sigma_j}\right)^a (\lambda^*)^b \Pr\{\hat{A}_{\alpha_1}\}. \tag{6.13}$$

Set $\tilde{\theta} = \max_i(\mu_i/(\lambda^*)^{1/r}) < 1$ by (2.11) and $\theta = \max(\tilde{\theta}, (\lambda^*)^{(r-\rho)/r}) < 1$ since $r - \rho \ge 1$.

Observe that because of $\rho \ge 1$ (with $\lambda = \lambda^*$),

$$\left(\prod_{j=r-\rho+1}^{r}\mu_{\sigma_j}\right)^a \lambda^b < \lambda^{(\rho/r)a}\tilde{\theta}^a\lambda^b = \lambda^{(\rho/r)a}\lambda^{(\rho/r)b}\lambda^{((r-\rho)/r)b}\tilde{\theta}^a \le \lambda^{(\rho/r)k}\theta^k. \tag{6.14}$$

By virtue of (6.14), the bound of (6.13) can be converted to

$$\Pr\left\{\hat{A}_{\alpha_1}\hat{A}_{\alpha_2}\right\} \le C\theta^k\lambda^{(\rho/r)k}\Pr\left\{\hat{A}_{\alpha_1}\right\} \le C\theta^k\lambda^{(\rho/r)k}\lambda^k, \tag{6.15}$$

the final λ^k arises from the asymptotics of the local match probability $\Pr\{\hat{A}_\alpha(k)\}$.

The number of positions of the type exemplified in Figure (6.12) is certainly less than $C(k+d)^{r-\rho}N^{r+\rho}$. The accumulation of these terms using the bound (6.15) on each term provides the total upper estimate

$$CN^{r+\rho}\lambda^{k+(\rho/r)k}\theta^k(k+d)^s.$$

This quantity goes to zero as $N \to \infty$ because $N^r\lambda^k$ is bounded while $\theta^k k^s$ tends to zero at an exponential rate. These kinds of estimates obviously can be adapted to cover all type of terms in IV.

The foregoing analysis proves

LEMMA 6.3. *The sum of* IV *goes to zero.*

LEMMA 6.4.

$$\textit{The sum } \mathrm{I} - \frac{\left(\hat{S}_1(N,k)\right)^2}{2!} \xrightarrow[N\to\infty]{} 0. \tag{6.16}$$

The proof completely parallels that of Lemma 5.5.

The conjunction of Lemmas 6.1–6.4 yields the asymptotic relation

$$\hat{S}_2(N,k) - \frac{\left[\hat{S}_1(N,k)\right]^2}{2!} \xrightarrow[N\to\infty]{} 0. \tag{6.17}$$

The next step proves the limit

$$\hat{S}_p(N,k) - \frac{\left[\hat{S}_1(N,k)\right]^p}{p!} \longrightarrow 0, \quad \text{as } N \to \infty. \tag{6.18}$$

The proof of (6.18) is more elaborate but in principle the same using an induction procedure paraphrasing Lemmas 5.5–5.7. We omit the details. With (6.18) validated the remainder of the proof of Theorem 2.2 is identical to that of the $r = s = 2$ case detailed in Section 5.

7. Comments and extensions.

7. Comments and extensions. There are various extensions and refinements on the methods and ideas of the preceding sections.

I. *γ-processes.* For r independent realizations from a uniformly mixing stationary process of a finite letter state space we proved in Theorem 3.1 that the probability of a match of length k or more, $F(k) = \Pr\{X^{[r]} \ge k\}$ obeys the asymptotic law $[F(k)]^{1/k} \to_{k\to\infty} \lambda^{[r]} = \lambda$ for some λ, $0 < \lambda < 1$, and that $F(k)/\lambda^k$ is bounded away from zero and infinity. When the actual limit exists

$$\lim_{k\to\infty} \frac{F(k)}{\lambda^k} = \gamma > 0, \tag{7.1}$$

we call this a γ-process [see (3.19)]. More generally, a collection of r uniformly mixing sequences with probability measures $\Pr_\rho\{w\}$, $\rho = 1, \ldots, r$, is said to form a γ-process if the limit (7.1) exists for $F(k) = \sum_{w\in\mathscr{W}_k}[\Pr_1\{w\}\Pr_2\{w\}\cdots\Pr_r\{w\}]$.

This holds for r Markov generated sequences with transition matrices $P(\rho)$, $\rho = 1, \dots, r$, if the matrix $P(1) \circ P(2) \circ \cdots \circ P(r)$ is primitive.

In the Markov case the γ limit can be explicitly identified. Let φ and ψ be the right and left eigenvectors of the matrix $P^{[r]} = P(1) \circ P(2) \circ \cdots \circ P(r)$ corresponding to the principal eigenvalue $\lambda(P^{[r]})$ normalized such that $\langle \varphi, \psi \rangle = \sum_{i=1}^{m} \varphi_i \psi_i = 1$. Let $\pi^{[r]}$ be the Schur product of the stationary frequency vectors $\pi^{(\rho)}$ associated with $P(\rho)$. Then $\gamma = \langle \varphi, \pi^{[r]} \rangle \langle \psi, \mathbf{u} \rangle / \lambda^{[r]}$, where $\mathbf{u}$ is the vector of all unit components.

The limit relations $\lim_{k \to \infty}(F(k))^{1/k} = \lambda$ and (7.1) serve decisively in establishing the limit laws set forth in Sections 2–6. The characterization of γ-processes is of independent interest. We describe next another class of γ-processes. Consider a collection of independent stationary Markov chains $\{\xi_t^{(\rho)}\}$ and let $\eta_t^{(\rho)} = f(\xi_t^{(\rho)})$, where f is some function on the state space (letter alphabet) which may coalesce groups of letters when f is not bijective. We can consider the random variable $Y^{[r]}$ indicating the length of the longest common word across the sequences $\{\eta_t^{(\rho)}, \ \rho = 1, \dots, r\}$. For simplicity, we take $r = 2$ with associated transition probability matrices P and Q. In order to calculate $F(k) = \Pr\{Y^{[r]} \geq k\} = \Pr\{\eta_{t+j}^{(1)} = \eta_{t+j}^{(2)}, \ j = 0, 1, \dots, k-1\}$, we form the Kronecker product matrix $P \otimes Q$ and let H be the principal submatrix restricted to the transitions $(i_1, i_2) \to (j_1, j_2)$ obeying $f(i_1) = f(i_2)$ and $f(j_1) = f(j_2)$, i.e., from an η-match to an η-match. For the processes $\{\eta^{(\rho)}\}$, the characteristic match parameter is the maximal Perron–Frobenius eigenvalue of the matrix H. The $\eta^{(\rho)}$ sequences form a γ-process if H is aperiodic. This is valid when P and Q are both irreducible, aperiodic with $p_{ii} q_{ii} > 0$ for all $i = 1, \dots, m$, or H is irreducible aperiodic so that $P \circ Q$ is primitive and there exists a state i_0 such that $p_{i_0 j} q_{i_0 j} > 0$ for all j.

II. *The eigenvalue condition* (2.11). Consider two stationary Markov dependent sequences $\mathcal{S}_1$ and $\mathcal{S}_2$ generated by the probability transition matrices P and Q, respectively. Assume $P \circ Q$ is a primitive matrix. Let P and Q be symmetric (or symmetrizable by a positive diagonal matrix) and commuting with eigenvalues $\{\mu_i\}_i^m$ and $\{\theta_i\}_1^m$, respectively. If P and Q are positive definite, then $\lambda(P \circ Q) \geq (1/m) \sum_{i=1}^{m} \mu_i \theta_i$. Condition (2.11) holds provided

$$\lim_{r \to \infty} \left(\lambda(P^{[r]}) \right)^{1/r} \text{ and } \lim_{r \to \infty} \left(\lambda(Q^{[r]}) \right)^{1/r}$$

$$\text{are bounded above by } \sqrt{\sum_{i=1}^{m} \mu_i \theta_i / m}\,. \tag{7.2}$$

In particular, (7.2) holds if $P = Q^l$ for l a positive integer provided Q is positive definite and $\lim_{r \to \infty}(\lambda(Q^{[r]}))^{1/r} < 1/\sqrt{m}$.

III. *Counts of long common words.* Let $W_{r,s}(N, k)$ be the number of common words extant in r out of s sequences exceeding length $k =$

$\log(N^r)/(-\log\lambda) + x$ (x is a fixed parameter), $\lambda = \lambda^*$ defined in Theorem 2.2.

$$\hat{W}_{r,s}(N,k) = \#\{\text{positions } \mathbf{t} \text{ and sequences } \boldsymbol{\sigma} | \text{event (2.1) holds for } \boldsymbol{\sigma} = (\sigma_1,\dots,\sigma_r) \text{ and } \mathbf{t} = (t_1,\dots,t_r) \text{ with a mismatch at } \mathbf{t} - \mathbf{1}\}. \tag{7.3}$$

The following theorem can be proved adapting the analysis of Sections 5 and 6.

THEOREM 7.1. *The random variable $\hat{W}_{r,s}(N,k)$ possesses the limit "Poisson law" as follows. For $k = \log N^r/(-\log\lambda) + x + \rho(N,x)$ [with $\rho(N,x)$, $0 < \rho(N,x) \le 1$, uniquely determined so that k is an integer], we have*

$$\Pr\{\hat{W}_{r,s}(N,k) = \nu\} - e^{-\phi(N,x)}\frac{[\phi(N,x)]^\nu}{\nu!} \xrightarrow[N\to\infty]{} 0, \tag{7.4}$$

with $\phi(N,x) = (1-\lambda)\gamma^*\lambda^{x+\rho(N,x)}$ ($\lambda = \lambda^*$; the parameter γ^* is defined in Theorem 2.2).

IV. *The limit distribution of repeats on one sequence.* We indicate the nature of the asymptotic distribution of $L_r(N)$ defined as the maximum length of a word that appears at least r times starting within the first N positions of a letter sequence generated as a uniformly mixing γ-process.

THEOREM 7.2. *Let $\mathscr{S} = \{\xi_t\}$ be a stationary uniformly mixing letter sequence. Assume the limits $\lambda = \lambda^{[r]} = \lim_{k\to\infty}(\sum_{w\in\mathscr{W}_k}[\Pr\{w\}]^r)^{1/k}$ and $\gamma = \lim_{k\to\infty}(\sum_{w\in\mathscr{W}_k}[\Pr\{w\}]^r)/(\lambda^{[r]})^k$ exist. Then*

$$|-\log\Pr\{L_r(N) < k\} - \hat{S}_1(N,k)| \xrightarrow[N\to\infty]{} 0, \tag{7.5}$$

where

$$\hat{S}_1(N,k) = \binom{N}{r}\Pr\{\xi_{t_1+\kappa} = \xi_{t_2+\kappa} = \cdots = \xi_{t_r+\kappa},\ \kappa = 1,2,\dots,k, \textit{ and } \xi_{t_1} = \xi_{t_2} = \cdots = \xi_{t_r} \textit{ fails}\}$$

$$\textit{for some } 1 \le t_1 < t_2 < \cdots < t_r \le N,\ k = \left[\frac{\log\binom{N}{r}}{-\log\lambda} + x + 1\right].$$

Compare to Zubkov and Mikhailov (1974) and Mikhailov (1974).

REFERENCES

ARRATIA, R., GORDON, L. and WATERMAN, M. (1986). An extreme value theory for sequence matching. *Ann. Statist.* **14** 971–994.

ARRATIA, R. and WATERMAN, M. (1985). Critical phenomena in sequence matching. *Ann. Probab.* **13** 1236–1249.

ERDÖS, P. and RÉVÉSZ, P. (1975). On the length of the longest head run. *Colloq. Math. Soc. János Bolyai. Topics in Information Theory* (I. Csiszár and P. Elias, eds.) **16** 219–228. North-Holland, Amsterdam.

FOULSER, D. and KARLIN, S. (1987). Maximal success runs for semi-Markov processes. *Stochastic Process. Appl.* **29** 203–224.

GUIBAS, L. J. and ODLYZKO, A. M. (1980). Long repetitive patterns in random sequences. *Z. Wahrsch. verw. Gebeite* **53** 241–262.

KARLIN, S., GHANDOUR, G. and FOULSER, D. (1985). DNA sequence comparisons of the human, mouse, and rabbit immunoglobulin kappa-gene. *Molecular Biol. Evol.* **2** 35–52.

KARLIN, S., GHANDOUR, G., OST, F., TAVARE, S. and KORN, L. J. (1983). New approaches for computer analysis of nucleic acid sequences. *Proc. Nat. Acad. Sci. U.S.A.* **80** 5660–5664.

KARLIN, S. and OST, F. (1985). Some montonicity properties of Schur powers of matrices and related inequalities. *Linear Algebra Appl.* **68** 47–65.

KARLIN, S. and OST, F. (1987). Counts of long aligned word matches among random letter sequences. *Adv. in Appl. Probab.* **19** 293–351.

MIKHAILOV, V. G. (1974). Limit distributions of random variables associated with multiple long duplications in a sequence of independent trials. *Theory Probab. Appl.* **19** 180–184.

SAMAROVA, S. S. (1981). On the length of the longest head-run for a Markov chain with two states. *Theory Probab. Appl.* **26** 498–509.

ZUBKOV, A. M. and MIKHAILOV, V. G. (1974). Limit distributions of random variables associated with long duplications in a sequence of independent trials. *Theory Probab. Appl.* **19** 172–179.

DEPARTMENT OF MATHEMATICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

INSTITUT FÜR ANGEWANDTE MATHEMATIK UND STATISTIK
TECHNISCHE UNIVERSITÄT MÜNCHEN
ARCISSTRASSE 21
D-8000 MÜNCHEN 2
WEST GERMANY