LIMIT DISTRIBUTION OF MAXIMAL NON-ALIGNED TWO-SEQUENCE SEGMENTAL SCORE

Amir Dembo*
Department of Mathematics and Department of Statistics
Stanford University
Stanford, CA 94305, U.S.A.

Samuel Karlin†
Department of Mathematics
Stanford University
Stanford, CA 94305 U.S.A.

Ofer Zeitouni††
Department of Electrical Engineering
Technion
Haifa 32000, Israel

1 January 1994; Revised: 9 February 1994

Abstract

Consider two independent sequences $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$. Suppose that $X_1, \cdots, X_n$ are i.i.d. μ_X and $Y_1, \cdots, Y_n$ are i.i.d. μ_Y, where μ_X and μ_Y are distributions on finite alphabets Σ_X and Σ_Y, respectively. A score $F : \Sigma_X \times \Sigma_Y \to I\!R$ is assigned to each pair (X_i, Y_j) and the maximal non-aligned segment score is $M_n = \max_{\substack{0\le i,j\le n-\Delta \\ \Delta\ge 0}} \{\sum_{k=1}^{\Delta} F(X_{i+k}, Y_{j+k})\}$. The limit distribution of M_n is derived here when μ_X and μ_Y are not too far apart and F is slightly constrained.

1. Introduction.

Our motivation derives from DNA and protein score based multiple sequence comparisons. Consider two sequences of length n, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, where the letters X_i take values in a finite alphabet Σ_X and the letters Y_i take values in a finite alphabet Σ_Y. A real-valued score $F(\cdot, \cdot)$ is assigned to each pair of letters (X_i, Y_j). The maximal segment score allowing shifts, is

$$M_n = \max_{\substack{0\le i,j\le n-\Delta \\ \Delta\ge 0}} \{\sum_{k=1}^{\Delta} F(X_{i+k}, Y_{j+k})\}.$$

Suppose the two sequences are independent: $X_1, \ldots, X_n$ i.i.d. following the distribution law μ_X and $Y_1, \ldots, Y_n$ i.i.d. following the distribution law μ_Y, where μ_X and μ_Y refer to probabilities on Σ_X and Σ_Y, respectively.

$*$ Partially supported by NSF DMS92-09712 grant and by a US-ISRAEL BSF grant.

† Partially supported by grants NIH 5R01GM10452-29, 8R01HG00335-04 and NSF DMS86-06244.

‡ Partially supported by a US-ISRAEL BSF grant and Fund for promotion of research at the Technion.

Key Words: Large deviations, Chen-Stein method, sequence matching, large segmental sums.

AMS (1991) subject classifications: Primary: 60F10; Secondary: 60G70

Of primary relevance is the case where the expected score per pair is negative and there is positive probability of attaining some positive pair score. Thus, we assume

$$\text{(H)} \qquad E_{\mu_X \times \mu_Y}(F) < 0 \ , \quad \mu_X \times \mu_Y(F > 0) > 0,$$

in which case $M_n \to \infty$ is the maximum of segmental scores of negative mean. The hypothesis (H) is in force throughout this paper, and it is also assumed that μ_X and μ_Y are strictly positive on Σ_X and Σ_Y, respectively.

It was shown in [DKZ, Theorem 1] that $M_n/\log n$ converges a.s. to a positive finite constant γ^* defined in terms of appropriate relative entropies. Here, we address the problem, mentioned in [DKZ], of evaluating limit laws for M_n, or equivalently, for the dual variables $T_y = \inf\{n : M_n > y\}$. These are closely related to Poisson limit laws for the count

$$\bar{W}_y = \sum_{i \le t_y} \sum_{j \le t_y} \sum_{\Delta=1}^{\min\{i,j\}} 1_{\{\sum_{k=1}^{\Delta} F(X_{i+k-\Delta}, Y_{j+k-\Delta}) > y\}},$$

with the proviso that when (i, j, Δ) is counted then the triplets (i, j, Δ') for $\Delta' > \Delta$ and $(i+k, j+k, \Delta')$ for $\Delta' \ge k \ge 1$ are not counted (the value of t_y is specified in Theorem 1 below). To state our main result we need some additional notation. Let $d(\cdot, \cdot)$ denote the variational norm between the indicated distributions, and $\mathrm{Po}(\lambda)$ denotes the Poisson random variable of parameter λ. Let θ^* and α^* denote the conjugate exponent and conjugate measure, respectively, defined in [DKZ]. That is, determine θ^* as the positive constant (unique, by (H)) satisfying

$$E_{\mu_X \times \mu_Y}(e^{\theta^* F}) = 1$$

and

$$\frac{d\alpha^*}{d(\mu_X \times \mu_Y)} = e^{\theta^* F}.$$

Let $\Sigma = \Sigma_X \times \Sigma_Y$ be the alphabet of letter pairs, and let $M_1(\Sigma)$ denote the set of all probability measures on Σ. The relative entropy of $\nu \in M_1(\Sigma)$ with respect to $\mu \in M_1(\Sigma)$, denoted by $H(\nu|\mu)$, is given for $\Sigma = \{b_1, \cdots, b_N\}$ by the formula:

$$H(\nu|\mu) = \sum_{i=1}^{N} \nu(b_i) \log \frac{\nu(b_i)}{\mu(b_i)},$$

with $0 \log 0$ interpreted as 0. In addition to (H), we impose throughout the assumption

$$\text{(E}'\text{)} \qquad H(\alpha^*|\mu_X \times \mu_Y) > 2\max(H(\alpha^*_X|\mu_X), H(\alpha^*_Y|\mu_Y))$$

where, for any $\nu \in M_1(\Sigma)$, ν_X and ν_Y denote the marginals of ν on Σ_X and Σ_Y, respectively. In particular we shall use μ to denote the product measure $\mu_X \times \mu_Y$. Note that the condition (E') requires strict inequality compared to (E) of [DKZ], which permits equality.

While in general, $\gamma^* \leq 2/\theta^*$, it is shown in [DKZ, Theorem 4] that, under (E'), $\gamma^* = 2/\theta^*$, and that for identical alphabets, (E') holds whenever $\mu_X = \mu_Y$ and $F(x,y) = F(y,x)$ is not of the form $F(x) + F(y)$. It is easy to check that (E') entails $\alpha^* \neq \alpha_X^* \times \alpha_Y^*$. Let

$$R_n = \max_{\substack{0 \leq i \leq n-\Delta \\ \Delta \geq 0}} \{\sum_{k=1}^{\Delta} F(X_{i+k}, Y_{i+k})\}, \tag{1.1}$$

be the maximal segment score between two aligned sequences. It is shown in [KD, Theorem A] (following [Igl]) that when $F(X,Y)$ is non-lattice, then

$$\lim_{n\to\infty} P(R_n - \frac{\log n}{\theta^*} \leq x) = \exp(-K^* \exp(-\theta^* x)) \ , \tag{1.2}$$

while if $F(X,Y)$ is a lattice variable then

$$\lim_{n\to\infty} \exp(K^* \exp(-\theta^* x_n)) P(R_n - \frac{\log n}{\theta^*} \leq x_n) = 1$$

for any bounded sequence x_n such that $x_n + \frac{\log n}{\theta^*}$ are lattice points. The constant K^* is determined from fluctuation sum series identities (see for example [KD, (1.8) and (1.11)]), and examples for which K^* is explicitly computed are given in [KD, Section 3].

The analysis of [DKZ] shows that under condition (E'), the constant limits of $M_n/\log n$ and $R_{n^2}/\log n$ are the same (i.e., then $\gamma^* = 2/\theta^*$). Our main result here establishes that the limit distribution of M_n is the same as that of R_{n^2}.

Theorem 1. *Assume (E') and (H). If $F(X,Y)$ is non-lattice, then*

$$\lim_{n\to\infty} P(M_n - \frac{2\log n}{\theta^*} \leq x) = \exp(-K^* \exp(-\theta^* x)) \ , \tag{1.3}$$

and if $F(X,Y)$ is a lattice variable then

$$\lim_{n\to\infty} \exp(K^* \exp(-\theta^* x_n)) P(M_n - \frac{2\log n}{\theta^*} \leq x_n) = 1 \tag{1.4}$$

for any bounded sequence x_n such that $x_n + \frac{2\log n}{\theta^*}$ are lattice points. Moreover, for $t_y = \sqrt{t} e^{\theta^* y/2}$,

$$\lim_{y\to\infty} d(\bar{W}_y, \mathrm{Po}(tK^*)) = 0 \tag{1.5}$$

implying that

$$\lim_{y\to\infty} P(T_y \leq t_y) = 1 - e^{-K^* t} \tag{1.6}$$

where if $F(X,Y)$ is a lattice variable then $y \to \infty$ in (1.5) and (1.6) via lattice points.

Remark 1. In deriving Theorem 1 we assume $F(\cdot,\cdot)$ to be finite valued, although the possibility of $F(x,y) = -\infty$ for some values of (x,y) is easily accommodated (see also the discussion of [DKZ, Theorem 3]). Thus, in the special case of $F(x,x) = 1$ and $F(x,y) = -\infty$ for all $x \neq y$ (with

$\Sigma_X = \Sigma_Y$), the limit (1.4) corresponds to the limit distribution of the longest segmental match between the two sequences. In this context, condition (H) holds as soon as $|\Sigma_X| > 1$, while condition (E') reduces to

$$\max\{\sum_{i\in\Sigma_X} \mu_X(i)\mu_Y(i)\log\mu_Y(i), \sum_{i\in\Sigma_X} \mu_X(i)\mu_Y(i)\log\mu_X(i)\} < \frac{1}{2}\lambda^*\log\lambda^* \tag{1.7}$$

where $\lambda^* = e^{-\theta^*} = \sum_i \mu_X(i)\mu_Y(i)$ (and in this case $K^* = 1-\lambda^*$, see [KD, Example 2]). For this special case, Theorem 1 was proved earlier in [KO, Theorem 2.2] encompassing a wide class of *proximal* ψ-mixing stationary sequences (see [KO, (2.11)] for the technical definition of proximal sequences). It is easy to check that for i.i.d. sequence letters (1.7) improves upon the proximality condition of [KO, (2.11)]. For related results in the context of longest quality match see [AGW86, AGW90].

Remark 2. Theorem 1 putatively extends to the maximal intersequence segment score involving any subset of r out of s independent sequences, of possibly different lengths $n_1, \cdots, n_s$ provided (H) applies for each r-subset and there is a unique dominant subset (having the maximal value of θ^*) for which condition (E_λ) of [DKZ, Section 5] holds with strict inequality.

Remark 3. In [DKZ, Theorem 4] it is shown that $\gamma^* = 2/\theta^*$ if and only if either (E') holds, or

$$H(\alpha^*|\mu_X\times\mu_Y) = 2\max(H(\alpha_X^*|\mu_X), H(\alpha_Y^*|\mu_Y)),$$

in which case $\alpha^* = \alpha_X^*\times\alpha_Y^*$. For example, this latter situation occurs for identical alphabets when $\mu_X = \mu_Y$ and $F(x,y) = F(x)+F(y)$. In this context, $M_n \le R_n^X + R_n^Y$, where for each fixed n,

$$R_n^X = \max_{\substack{0\le i\le n-\Delta\\ \Delta\ge 0}}\{\sum_{k=1}^{\Delta} F(X_{i+k})\},\quad R_n^Y = \max_{\substack{0\le j\le n-\Delta\\ \Delta\ge 0}}\{\sum_{k=1}^{\Delta} F(Y_{j+k})\},$$

are two i.i.d. random variables. Assuming for simplicity that $F(X)$ is non-lattice, it follows from (1.2) that

$$\lim_{n\to\infty} P(M_n - \frac{2\log n}{\theta^*} \le x) \ge \lim_{n\to\infty} P((R_n^X - \frac{\log n}{\theta^*}) + (R_n^Y - \frac{\log n}{\theta^*}) \le x) = h(K^* e^{-\theta^* x/2})\ ,$$

where

$$h(u) = \int_{-\infty}^{\infty} \exp(-(u^2/K^*)\exp(\theta^* z))d[\exp(-K^*\exp(-\theta^* z))] = u\int_0^\infty e^{-u(t+1/t)}dt \ge 1.5ue^{-2.5u}.$$

Since $K^* > 0$, considering $x\to-\infty$, it is clear that (1.3)-(1.6) do not hold in this case.

Remark 4. Even when (E') does not hold, M_n may still possess a limiting extremal distribution of type I (with a different constant $1/\theta^* < \gamma^* < 2/\theta^*$), and this might happen even when the set $\mathcal{M}$ of optimal measures as characterized in [DKZ, Theorem 2] is infinite. For example, let $G_Y(y) = \max_x\{F(x,y)\}$ and

$$\bar{R}_n^Y = \max_{\substack{0\le j\le n-\Delta \\ \Delta\ge 0}} \{\sum_{k=1}^{\Delta} G_Y(Y_{j+k})\} .$$

Suppose that $E_{\mu_Y}(G_Y) < 0$ and let $\bar{\theta}^*$ denote the unique positive solution of $E_{\mu_Y}(e^{\theta G_Y}) = 1$. Then $\bar{R}_n^Y - \log n/\bar{\theta}^*$ possesses a limit distribution of type I (cf. [KD, Theorem A]). Let $\bar{\Sigma} = \{(x,y) : F(x,y) = G_Y(y)\}$ and define $\beta^* \in M_1(\Sigma_Y)$ such that $\frac{d\beta^*}{d\mu_Y} = e^{\bar{\theta}^* G_Y}$. If

$$(E_Y) \qquad 2H(\beta^*|\mu_Y) > \min_{\nu:\nu(\bar{\Sigma})=1,\nu_Y=\beta^*} H(\nu|\mu_X \times \mu_Y) ,$$

then $\gamma^* = 1/\bar{\theta}^*$ (see [DKZ, (1) and (13)]). Clearly, $\bar{R}_n^Y \ge M_n$. In Section 3 we provide a specific example for which (E_Y) holds and show that (E_Y) results with

$$\lim_{n\to\infty} P(M_n = \bar{R}_n^Y) = 1 . \tag{1.8}$$

Consequently, M_n possesses the same limit distribution of type I as does $\bar{R}_n^Y$.

Remark 5. In comparison with the recent works of [AW] and [Neu], we allow for a general score $F(\cdot,\cdot)$ but accommodate neither insertions nor deletions. Note however that in [AW] only the growth order of M_n is found, while in [Neu] the Poisson approximation is established under an additional assumption of a limited number of insertions/deletions.

2. Proof of Theorem 1.

Since $\{\bar{W}_y \ne 0\} = \{T_y \le t_y\} = \{M_n > y\}$ for $n = [t_y]$, (1.3) and (1.6) are direct consequences of (1.5) while (1.4) holds provided (1.5) applies to any bounded $t = t(y)$. Hence, Theorem 1 amounts to proving that (1.5) holds for any bounded $t = t(y)$. We start with an outline of the main steps in proving this result.

The large deviations analysis of [DKZ] allows us to concentrate on segments of length not exceeding $c_1 y$ whose empirical measure is near α^*. Hence, partitioning both sequences into disjoint blocks of size ℓ_y such that $e^{\theta^* y} \gg \ell_y \gg y$, the probability $P(\bar{W}_y \ne W_y)$ approaches 0 as $y \to \infty$, where $W_y = \sum_{i,j,\xi} I_{i,j,\xi}$ and the indicator $I_{i,j,\xi}$ equals one if there exists a segmental score exceeding y involving the i-th block of the X-sequence, the j-th block of the Y-sequence and a relative shift (alignment) ξ between the indices of the X-letters and the corresponding Y-letters. Adapting the arguments of [KD] and [Igl], we see in Lemma 1 that $|E[W_y] - tK^*| \to 0$ as $y \to \infty$. Applying the Chen-Stein method we show that $d(W_y, \mathrm{Po}(tK^*)) \to 0$ from which (1.5) follows. The main task is in bounding the correlation terms $E(I_{i,j,\xi} I_{i',j',\xi'})$, where large deviations estimates are again decisive, and where the condition (E') and the restriction to an empirical measure near α^* are needed (see Lemma 2 below).

Turning now to the detailed proof, let $\|\cdot\|$ denote the variational norm between distributions on Σ and $G_\eta = \{\nu \in M_1(\Sigma) : \|\nu - \alpha^*\| < \eta\}$ denote the corresponding open ball of radius $\eta > 0$, centered at α^*. Let $T^i\mathbf{X} = (X_{i+1}, X_{i+2}, \cdots), T^j\mathbf{Y} = (Y_{j+1}, Y_{j+2}, \cdots)$, and define the empirical measure

$$L_\Delta^{(T^i\mathbf{X},T^j\mathbf{Y})} = \frac{1}{\Delta}\sum_{k=1}^{\Delta} \delta_{(X_{i+k},Y_{j+k})}.$$

For $U \in M_1(\Sigma)$ let

$$M_n^U = \max\{\sum_{k=1}^{\Delta} F(X_{i+k}, Y_{j+k}) : 0 \le \Delta \le n, \quad i,j \le n-\Delta, L_\Delta^{(T^i\mathbf{X},T^j\mathbf{Y})} \in U\} ,$$

i.e. M_n^U is the maximal score among segments with letter pairs having empirical measure in the set U. It is shown in [DKZ, Theorem 3] that if U is a closed set such that $\alpha^* \notin U$ then a.s.

$$\limsup_{n\to\infty} M_n^U / \log n < 2/\theta^* .$$

Let

$$\bar{M}_n^U = \max\{\sum_{k=1}^{\Delta} F(X_{i+k}, Y_{j+k}) : 0 \le \Delta \le c_0 \log n, \quad i,j \le n-\Delta, L_\Delta^{(T^i\mathbf{X},T^j\mathbf{Y})} \in U\} , \tag{2.1}$$

be the maximal score among segments of length not exceeding $c_0 \log n$ and letter pairs having empirical measure in the set U. It follows from [DKZ, Lemma 1] that for c_0 large enough

$$\sum_{n=1}^{\infty} P(\bar{M}_n^U \neq M_n^U) < \infty .$$

Consequently, for all $\eta > 0$

$$\lim_{n\to\infty} P(\bar{M}_n^{G_\eta} \neq M_n) = 0. \tag{2.2}$$

In particular, for c_1 large enough and all $\eta > 0$, suffices to prove (1.5) with the count $\bar{W}_y$ restricted to triplets (i,j,Δ) for which $\Delta \le c_1 y$ and $L_\Delta^{(T^{i-\Delta}\mathbf{X},T^{j-\Delta}\mathbf{Y})} \in G_\eta$. Now let $\ell_y \ge 3c_1 y$ be a sequence of integers such that $\log \ell_y / y \to 0$ and $y^2/\ell_y \to 0$ as $y \to \infty$. Set $m_y = t_y/\ell_y$. Obviously, $m_y \to \infty$. Since $d(\mathrm{Po}(\lambda), \mathrm{Po}(\lambda')) \le |\lambda - \lambda'|$, we may assume without loss of generality that m_y (and hence t_y) are integers. Partition the sequence $(X_1, \ldots, X_{t_y})$ into blocks of ℓ_y letters each, such that the i-th block is $X^i = (X_0^i, X_1^i, \ldots, X_{\ell_y-1}^i)$ where $X_k^i = X_{i\ell_y+k+1}$. Similarly, partition the sequence $(Y_1, \ldots, Y_{t_y})$ into blocks of ℓ_y letters each. For $j = 0, \ldots, m_y - 1$ and $\xi = 0, 1, \ldots, \ell_y - 1$, let $Y^{j,\xi} = (Y_0^{j,\xi}, Y_1^{j,\xi}, \ldots, Y_{\ell_y-1}^{j,\xi})$ denote the ξ-cyclically-shifted j-th block, such that $Y_k^{j,\xi} = Y_{j\ell_y+1+(\xi+k)_{\mathrm{mod}\,\ell_y}}$. Let

$$W_y = \sum_{i=0}^{m_y-1} \sum_{j=0}^{m_y-1} \sum_{\xi=0}^{\ell_y-1} I_{i,j,\xi} , \tag{2.3}$$

where

$$I_{i,j,\xi} = \begin{cases} 1 & \text{if } \max\{\sum_{k=r}^{r+\Delta-1} F(X_k^i, Y_k^{j,\xi}) : \quad \ell_y - \Delta \geq r \geq 0,\ c_1 y \geq \Delta \geq 0,\ L_\Delta^{i,j,\xi,r} \in G_\eta\} > y \\ 0 & \text{otherwise,} \end{cases}$$

and

$$L_\Delta^{i,j,\xi,r} = \frac{1}{\Delta} \sum_{k=r}^{r+\Delta-1} \delta_{(X_k^i, Y_k^{j,\xi})}.$$

For $k \leq c_1 y$ let $\mathcal{E}_1(k)$ be the event of a score exceeding y in at least one of the segments of length k which cross the block boundaries in either the X-sequence or the Y-sequence. Similarly, let $\mathcal{E}_2(k)$ be the event of a score exceeding y in at least one of the segments of length k in which the ξ-shift in $Y^{j,\xi}$ causes a gap in the Y-letters. It is easy to check that at most $2t_y m_y(k-1)$ segments are contributing to $\mathcal{E}_i(k)$ for $i = 1, 2$, and therefore by the union of events bound

$$P(\bigcup_{k \leq c_1 y} \mathcal{E}_1(k) \bigcup_{k \leq c_1 y} \mathcal{E}_2(k)) \leq 2t_y m_y (c_1 y)^2 \sup_{k \geq 1} P(\sum_{i=1}^{k} F(X_i, Y_i) > y)\ .$$

Because $E[e^{\theta^* F(X,Y)}] = 1$ and independence

$$P(\sum_{i=1}^{k} F(X_i, Y_i) > y) \leq E(\exp(\sum_{i=1}^{k} \theta^* F(X_i, Y_i)))e^{-\theta^* y} = e^{-\theta^* y}$$

and since by definition $t_y = \sqrt{t} e^{\theta^* y/2}$ we obtain that

$$P(\bigcup_{k \leq c_1 y} \mathcal{E}_1(k) \bigcup_{k \leq c_1 y} \mathcal{E}_2(k)) \leq \frac{2t(c_1 y)^2}{\ell_y} \to 0 \quad \text{as} \quad y \to \infty$$

Let $\mathcal{E}_3(i,j,\xi)$ be the event that there are $\Delta \leq r$ and $r + \Delta' \leq r' \leq \ell_y$ such that

$$\sum_{k=r-\Delta+1}^{r} F(X_k^i, Y_k^{j,\xi}) > y, \qquad \sum_{k=r'-\Delta'+1}^{r'} F(X_k^i, Y_k^{j,\xi}) > y.$$

Since R_n is monotone in n (see (1.1)), it follows by conditioning on $\{X_k^i, Y_k^{j,\xi}, k \leq r\}$ that $P(\mathcal{E}_3(i,j,\xi)) \leq P(R_{\ell_y} > y)^2$. Consequently, by the union of events bound

$$P(\bigcup_{i,j,\xi} \mathcal{E}_3(i,j,\xi)) \leq m_y^2 \ell_y P(R_{\ell_y} > y)^2 = t\frac{e^{\theta^* y}}{\ell_y} P(R_{\ell_y} > y)^2$$

Hence, the next lemma implies that $P(\cup_{i,j,\xi} \mathcal{E}_3(i,j,\xi)) \to 0$ as $y \to \infty$.

Lemma 1.

$$\lim_{y \to \infty} \frac{e^{\theta^* y}}{\ell_y} P(R_{\ell_y} > y) = K^*\ .$$

It is not hard to check that

$$\{\bar{W}_y \neq W_y\} \subset \bigcup_{k\le c_1 y} \mathcal{E}_1(k) \bigcup_{k\le c_1 y} \mathcal{E}_2(k) \bigcup_{i,j,\xi} \mathcal{E}_3(i,j,\xi)\,.$$

Consequently, in order to prove (1.5), it suffices to show that

$$d(W_y, \mathrm{Po}(tK^*)) \underset{y\to\infty}{\to} 0 \tag{2.4}$$

Proof of Lemma 1: Following [KD], divide the realization of $S_n = \sum_{i=1}^n F(X_i,Y_i)$ into successive nonnegative excursions:

$$K_0 = 0,\ \ K_\nu = \min\{k : k \ge K_{\nu-1}+1,\ S_k - S_{K_{\nu-1}} < 0\},\ \ \nu = 1,2,\ldots$$

with excursions extremes

$$Q_\nu = \max_{K_{\nu-1}\le k<K_\nu} (S_k - S_{K_{\nu-1}}).$$

Note that Q_ν are i.i.d. random variables, with common distribution function denoted $G(y)$. Thus, $P(R_{K_m} > y) = 1-[G(y)]^m$. Fix $\delta > 0$ arbitrarily small and define next $m_\pm = \gamma_\pm \ell_y / E(K_1)$ with $E(K_1) < \infty$ due to $E_\mu(F) < 0$, where $\gamma_+ \ge (1+\delta)$ and $\gamma_- \le (1-\delta)$ are chosen as the minimal (maximal) values such that m_+ (and m_-, respectively) are integers (as $y\to\infty$ we have $\gamma_+ \to 1+\delta$ and $\gamma_- \to (1-\delta)$). Using

$$\lim_{y\to\infty} (1-G(y))e^{\theta^* y} = E(K_1)K^*,$$

which is provided by [KD, Lemma A] and the identification of K^* in [KD, below (1.12)] (see also [Igl]), one sees that

$$\lim_{y\to\infty} \frac{e^{\theta^* y}}{\ell_y} P(R_{K_{m_+}} > y) = \lim_{y\to\infty} \frac{e^{\theta^* y}}{\ell_y}[1-G(y)^{\gamma_+\ell_y/E(K_1)}] = (1+\delta)K^*, \tag{2.5}$$

and

$$\lim_{y\to\infty} \frac{e^{\theta^* y}}{\ell_y} P(R_{K_{m_-}} > y) = (1-\delta)K^*. \tag{2.6}$$

Since R_n is monotone in n,

$$P(R_{K_{m_-}} > y) - P(K_{m_-} > \ell_y) \le P(R_{\ell_y} > y) \le P(R_{K_{m_+}} > y) + P(K_{m_+} < \ell_y). \tag{2.7}$$

Let $g(\theta) = -\theta + \frac{(1-\delta)}{E(K_1)}\log E(e^{\theta K_1})$. Note that, for each m, K_m is a sum of i.i.d. positive random variables. Hence, using Chebycheff's bound,

$$P(K_{m_-} > \ell_y) \le \inf_{\theta\ge 0}\{e^{-\theta\ell_y}E(e^{\theta K_1})^{m_-}\} \le \inf_{\theta\ge 0}\ e^{g(\theta)\ell_y}\,.$$

Note that for $\lambda_0 > 0$ such that $\Lambda(\lambda_0) = \log E(e^{\lambda_0 F(X_1,Y_1)}) < 0$ (λ_0 exists due to the boundedness of F and (H), see [DKZ, proof of Lemma 1]) we have

$$P(K_1 > n) \leq P(\sum_{i=1}^{n} F(X_i, Y_i) \geq 0) \leq e^{n\Lambda(\lambda_0)}$$

Therefore, $g(\theta) < \infty$ for all θ in a small enough neighborhood of 0. It follows that $g'(0) = -\delta < 0$, leading to

$$\frac{e^{\theta^* y}}{\ell_y} P(K_{m_-} > \ell_y) \leq e^{-c(\delta)\ell_y} e^{\theta^* y} \underset{y\to\infty}{\to} 0 \tag{2.8}$$

for some constant $c(\delta) > 0$. A similar computation yields

$$\frac{e^{\theta^* y}}{\ell_y} P(K_{m_+} < \ell_y) \underset{y\to\infty}{\to} 0 \tag{2.9}$$

Substituting (2.5)-(2.6) and (2.8)-(2.9) into (2.7) and taking $\delta \to 0$ yields the lemma. ∎

For the objective of proving (2.4) we employ a version of the Chen-Stein method given in [AGG]. Let $\alpha = (i, j, \xi)$ and let $\mathcal{B}_\alpha = \{(i', j', \xi') : i = i' \text{ or } j = j'\}$ denote the associated neighborhood of dependence. With this definition, note that I_α is independent of $\{I_\gamma : \gamma \notin \mathcal{B}_\alpha\}$. Thus, from [AGG] (see also [DK92, inequalities (2.4) and (2.7)]), one has

$$d(W_y, \mathrm{Po}(tK^*)) \leq (b_1 + b_2)\frac{(1 - e^{-\lambda_y})}{\lambda_y} + |\lambda_y - tK^*|$$

where $\lambda_y = E(W_y)$, and

$$b_1 = \sum_{\alpha} \sum_{\beta\in\mathcal{B}_\alpha} P(I_\alpha = 1)P(I_\beta = 1)$$
$$b_2 = \sum_{\alpha} \sum_{\substack{\beta\in\mathcal{B}_\alpha \\ \beta\neq\alpha}} P(I_\alpha = 1, I_\beta = 1)$$

(in the notations of [AGG], $b_3 = 0$). Let

$$R_{\ell_y}^{G_\eta} = \max\{\sum_{k=1}^{\Delta} F(X_{i+k}, Y_{i+k}) : 0 \leq i \leq \ell_y - \Delta, 0 \leq \Delta \leq c_1 y, \quad L_\Delta^{T^i\mathbf{X}, T^i\mathbf{Y}} \in G_\eta\}$$

and $p_y = P(R_{\ell_y}^{G_\eta} > y)$. Note that for any α, $P(I_\alpha = 1) = p_y$, and $|\mathcal{B}_\alpha| \leq 2m_y\ell_y$. Therefore,

$$\lambda_y = m_y^2 \ell_y p_y = t(\frac{p_y}{\ell_y})e^{\theta^* y}. \tag{2.10}$$

and

$$b_1 = p_y^2 \sum_{\alpha} |\mathcal{B}_\alpha| \leq 2m_y\ell_y p_y^2(m_y^2\ell_y) = \frac{2\lambda_y^2}{m_y}$$

Since $R_{\ell_y} \geq R_{\ell_y}^{G_\eta}$, it follows that

$$P(R_{\ell_y} > y) \geq p_y \geq P(R_{\ell_y}^{G_\eta} = R_{\ell_y} | R_{\ell_y} > y)P(R_{\ell_y} > y) .$$

The strong laws of [DK91, Theorems 1 and 2] imply that $P(R_{\ell_y}^{G_\eta} = R_{\ell_y} | R_{\ell_y} > y) \to 1$ for every $\eta > 0$, and hence by (2.10) and Lemma 1

$$\lim_{y\to\infty} |\lambda_y - tK^*| = \lim_{y\to\infty} t|(\frac{p_y}{\ell_y})e^{\theta^* y} - K^*| = 0 \tag{2.11}$$

(recall that $t = t(y)$ is bounded). In particular, (2.11) implies that $b_1 \to 0$, and (2.4) thus follows from the next lemma, completing the proof of Theorem 1.

Lemma 2. For all $\eta > 0$ small enough, $\quad b_2 \to 0$ as $y \to \infty$.

Proof of Lemma 2: Using I_0 to abbreviate $I_{(0,0,0)}$ let $Q_0(y) = e^{\theta^* y/2} P(I_{(1,0,0)} = 1 | I_0 = 1)$, $Q_1(y) = e^{\theta^* y/2} P(I_{(0,1,0)} = 1 | I_0 = 1)$, and $Q_2(y) = \sum_{\xi=1}^{\ell_y - 1} P(I_{(0,0,\xi)} = 1 | I_0 = 1)$. By the symmetry of the problem,

$$\begin{aligned} b_2 &= \sum_\alpha p_y \sum_{\substack{\beta \in \mathcal{B}_{(0,0,0)} \\ \beta \neq (0,0,0)}} P(I_\beta = 1 | I_0 = 1) \\ &\le p_y m_y^2 \ell_y m_y \ell_y [P(I_{(1,0,0)} = 1 | I_0 = 1) + P(I_{(0,1,0)} = 1 | I_0 = 1) + \frac{1}{m_y \ell_y} \sum_{\xi=1}^{\ell_y - 1} P(I_{(0,0,\xi)} = 1 | I_0 = 1)] \\ &= a_y (Q_0(y) + Q_1(y)) + \tilde{a}_y Q_2(y) \end{aligned} \tag{2.13}$$

where $a_y = \frac{p_y}{\ell_y} m_y^3 \ell_y^3 e^{-\theta^* y/2}$ is such that $|a_y - t^{3/2} K^*| \to 0$ as $y \to \infty$ (see (2.11)), and $\tilde{a}_y = a_y e^{\theta^* y/2} / \ell_y m_y$ is such that $|\tilde{a}_y - tK^*| \to 0$ as $y \to \infty$. Proving Lemma 2 thus requires showing that $Q_i(y) \to 0$, $i = 0, 1, 2$ as $y \to \infty$. Due to the symmetric roles played by μ_X and μ_Y, it is enough to consider only $i = 1$ and $i = 2$.

It is now useful to decompose the events I_0, $I_{(0,1,0)}$ and $I_{(0,0,\xi)}$. Thus, let

$$J_{x,k,\nu} = \{\omega : \frac{1}{k} \sum_{j=0}^{k-1} \delta_{(X_{x+j}, Y_{x+\ell_y+j})} = \nu \in G_\eta, \quad kE_\nu(F) > y\},$$

with $x = 1, \ldots, \ell_y - k + 1$, $k \le c_1 y$ and ν ranges over all possible k-types ($\nu \in M_1(\Sigma)$ with $k\nu(i)$ an integer for all $i \in \Sigma$); thus, the range of the pair (k, ν) is of cardinality at most $(c_1 y + 1)^{|\Sigma|}$. Similarly, define

$$J_{x',k',\nu',\xi} = \{\omega : \frac{1}{k'} \sum_{j=0}^{k'-1} \delta_{(X_{x'+j}, Y_{x'+(\xi+j)_{\mathrm{mod}\,\ell_y}})} = \nu' \in G_\eta, \quad k' E_{\nu'}(F) > y\}$$

with $x' = 1, \ldots, \ell_y - k' + 1$, $k' \le c_1 y$, ν' ranges over all possible k'-types, and $\xi = 0, 1, \ldots, \ell_y - 1$.

(1) Starting with $Q_1(y)$, note that

$$P(I_{(0,1,0)} = 1 | I_0 = 1) = P(\bigcup_{x,k,\nu} J_{x,k,\nu} | \bigcup_{x',k',\nu'} J_{x',k',\nu',0}) \le \sum_{x,k,\nu} \sum_{x',k',\nu'} P(J_{x,k,\nu} | J_{x',k',\nu',0}).$$

There are two distinct classes of four-tuples $e = (x, x', k, k')$ to consider, $e \in \mathcal{E}_a$ if $[x, x+k-1] \cap [x', x'+k'-1] = \emptyset$ and $e \in \mathcal{E}_b$ otherwise. For $e \in \mathcal{E}_a$,

$$P(J_{x,k,\nu}|J_{x',k',\nu',0}) = P(J_{x,k,\nu}) \leq P(I_{(0,1,0)} = 1) = p_y \tag{2.14}$$

Since the only connection between the conditioning event and $J_{x,k,\nu}$ is through the X-sequence

$$\begin{aligned}
\sup_{e \in \mathcal{E}_b, \nu, \nu'} P(J_{x,k,\nu}|J_{x',k',\nu',0}) = & \sup_{\substack{k,k',\nu,\nu' \\ 1 \leq x \leq k'}} P(J_{x,k,\nu}|J_{1,k',\nu',0}) \\
\leq & \sup_{\substack{(a_1,\dots,a_k) \\ k',k,\nu,1\leq x \leq k'}} P(J_{x,k,\nu}|X_x = a_1, \dots, X_{x+k-1} = a_k)
\end{aligned}$$

$$= \sup_{k,\nu} P\Big(\frac{1}{k}\sum_{j=1}^{k} \delta_{(X_j, Y_{\ell_y+j})} = \nu \Big| \frac{1}{k}\sum_{j=1}^{k} \delta_{X_j} = \nu_X\Big) = \sup_{k,\nu} \frac{P(\frac{1}{k}\sum_{j=1}^{k} \delta_{(X_j,Y_j)} = \nu)}{P(\frac{1}{k}\sum_{j=1}^{k} \delta_{X_j} = \nu_X)} \tag{2.15}$$

Using simple combinatorial bounds (see, e.g. [DKZ, (3) and (4)]), one sees that

$$\sup_{k,\nu} \frac{P(\frac{1}{k}\sum_{j=1}^{k} \delta_{(X_j,Y_j)} = \nu)}{P(\frac{1}{k}\sum_{j=1}^{k} \delta_{X_j} = \nu_X)} \leq \sup_{k,\nu} (c_1 y + 1)^{|\Sigma|} e^{-k[H(\nu|\mu) - H(\nu_X|\mu_X)]} \tag{2.16}$$

By (E') and the continuity in a of $H(a|b)$, for $\eta > 0$ small enough

$$\beta(\eta) = \inf_{\nu \in G_\eta} \{H(\nu|\mu) - 2\max[H(\nu_X|\mu_X), H(\nu_Y|\mu_Y)]\} > 0$$

Thus, for $\nu \in G_\eta$ such that $kE_\nu(F) > y$, one has that $H(\nu|\mu) \geq 2H(\nu_X|\mu_X) + \beta(\eta)$ while $kH(\nu|\mu) \geq \theta^* y$. Hence, using (2.14)-(2.16) and (2.11),

$$\begin{aligned}
Q_1(y) =& e^{\theta^* y/2} P(I_{(0,1,0)} = 1|I_0 = 1) \\
\leq & e^{\theta^* y/2} [\ell_y^2 (c_1 y + 1)^{3|\Sigma|} e^{-\theta^* y/2} e^{-\beta(\eta) y/2\|F\|_\infty} + p_y (c_1 y + 1)^{2|\Sigma|} \ell_y^2] \underset{y \to \infty}{\to} 0.
\end{aligned}$$

(2) It remains to deal with $Q_2(y)$. As in the above computation, note that,

$$p_y = P\Big(\bigcup_{x,k,\nu} J_{x,k,\nu,\xi}\Big) = P\Big(\bigcup_{x',k',\nu'} J_{x',k',\nu',0}\Big),$$

and one has

$$\begin{aligned}
Q_2(y) = \sum_{\xi=1}^{\ell_y - 1} P(I_{(0,0,\xi)} = 1|I_0 = 1) =& \sum_{\xi=1}^{\ell_y - 1} P\Big(\bigcup_{x,k,\nu} J_{x,k,\nu,\xi}, \bigcup_{x',k',\nu'} J_{x',k',\nu',0}\Big)/p_y \\
\leq & 2 \sum_{\substack{\xi,x,k,\nu \\ x',k' \leq k,\nu'}} \frac{P(J_{x,k,\nu,\xi}, J_{x',k',\nu',0})}{p_y}
\end{aligned} \tag{2.17}$$

For any five-tuple $e = (\xi, x, x', k, k')$, let Δ_X (Δ_Y) denote the set of X_i (Y_i) letters occurring in the definition of $J_{x,k,\nu,\xi}$ which do not occur in the definition of $J_{x',k',\nu',0}$. Three distinct cases are possible:

(a) $|\Delta_X| \vee |\Delta_Y| \geq (1-\eta)k$ (denoted $e \in \mathcal{E}_a$).

(b) $(1-\eta)k \geq |\Delta_X| \vee |\Delta_Y| \geq \delta y$ (denoted $e \in \mathcal{E}_b$).

(c) $|\Delta_X| \vee |\Delta_Y| \leq \delta y$ (denoted $e \in \mathcal{E}_c$).

Here, δ is a small fixed constant which depends on η and will be chosen below. We analyze the three cases separately. The argument for $|\Delta_X| > |\Delta_Y|$ being the same as for $|\Delta_X| \leq |\Delta_Y|$, we may assume the latter in subsequent computations.

Case (a) To simplify the notations we assume that ηk is an integer (otherwise replace ηk by its integer part), and let $L_\eta = \sum_{i=1}^{\eta k} \delta_{Y_i}/\eta k$, $L_{1-\eta} = \sum_{i=\eta k+1}^{k} \delta_{Y_i}/(1-\eta)k$. Note that, after re-labeling the random variables involved, since $\nu \in G_\eta$, for $\eta \leq 1/2$

$$\begin{aligned} P(J_{x,k,\nu,\xi}|J_{x',k',\nu',0}) &\leq \sup_{(b_1,b_2,\ldots,b_{\eta k})} P(\frac{1}{k}\sum_{i=1}^{k} \delta_{Y_i} = \nu_Y | Y_1 = b_1, Y_2 = b_2, \ldots, Y_{\eta k} = b_{\eta k}) \\ &= \sup_{(b_1,b_2,\ldots,b_{\eta k})} P((1-\eta)L_{1-\eta} + \eta L_\eta = \nu_Y | Y_1 = b_1, Y_2 = b_2, \ldots, Y_{\eta k} = b_{\eta k}) \\ &\leq \sup_{\phi \in G_{4\eta}} P(L_{1-\eta} = \phi_Y) \end{aligned}$$

With $\alpha_Y^* \neq \mu_Y$, one may find an η small enough such that $\rho(\eta) = \inf_{\phi\in G_{4\eta}} H(\phi_Y|\mu_Y) > 0$. Choosing η at least that small, by the combinatorial upper bound of [DZ, Lemma 2.1.9]

$$\sup_{\phi \in G_{4\eta}} P(L_{1-\eta} = \phi_Y) \leq e^{-(1-\eta)\rho(\eta)k} \leq e^{-(1-\eta)\rho(\eta)y/||F||_\infty}$$

(recall that $kE_\nu(F) > y$). Since $p_y \geq P(J_{x',k',\nu',0})$, we are led to the conclusion that, for all $e \in \mathcal{E}_a$,

$$\frac{P(J_{x,k,\nu,\xi}, J_{x',k',\nu',0})}{p_y} \leq P(J_{x,k,\nu,\xi}|J_{x',k',\nu',0}) \leq e^{-(1-\eta)\rho(\eta)y/||F||_\infty}. \tag{2.18}$$

Note that in both cases (b) and (c), since the overlap between the sequences involved in the definition of $J_{x,k,\nu,\xi}$ and $J_{x',k',\nu',0}$ is at least of one symbol whereas $\ell_y \geq 3c_1 y \geq 3k$, one may re-label the sequences such that $x' = 1$, x may assume both positive and negative values and the modulus operation is omitted from the definition of $J_{x,k,\nu,\xi}$. We will henceforth work with this re-labeling without further mentioning it.

Case (b) Let here

$$L_{x,k,\xi} = \frac{1}{k}\sum_{\ell=x}^{x+k-1} \delta_{X_\ell, Y_{\ell+\xi}},$$

and

$$L^{\Delta}_{x,k,\xi} = \frac{1}{k-|\Delta_Y|}\sum_{\ell+\xi\in[x+\xi,x+\xi+k-1]\setminus\Delta_Y} \delta_{X_\ell, Y_{\ell+\xi}}.$$

Note that now,

$$P(J_{x,k,\nu,\xi}, J_{1,k',\nu',0}) = P(L_{x,k,\xi} = \nu, L_{1,k',0} = \nu') \leq P(L_{x,k,\xi} = \nu, L^{\Delta}_{x,k,\xi} \notin G_{2\eta})$$
$$+ \sup_{\phi \in G_\eta} P(L_{x,k,\xi} = \phi, L_{1,k',0} = \nu', L^{\Delta}_{x,k,\xi} \in G_{2\eta}) = A_1 + A_2$$

Turning our attention to A_1, note that, by combining [DZ, (2.1.32) and (2.1.34)],

$$P(L^{\Delta}_{x,k,\xi} = \psi | L_{x,k,\xi} = \nu) \leq (k+1)^{2(|\Sigma|+1)} e^{-(k-|\Delta_Y|)H(\psi|\nu)} .$$

Hence, for $\nu \in G_\eta$ such that $kE_\nu(F) > y$

$$\begin{aligned} \frac{A_1}{p_y} &\leq P(L^{\Delta}_{x,k,\xi} \notin G_{2\eta} | L_{x,k,\xi} = \nu) \leq (k+1)^{3(|\Sigma|+1)} e^{-\eta k \inf_{\psi \notin G_{2\eta}} H(\psi|\nu)} \\ &\leq (c_1 y + 1)^{3(|\Sigma|+1)} e^{-y\eta^3/2||F||_\infty} \end{aligned} \tag{2.19}$$

where we have used in the last inequality the relation (see [DZ, Exercise 6.2.17])

$$H(\psi|\phi) \geq \|\psi - \phi\|^2/2 . \tag{2.20}$$

To evaluate A_2, let L^{Δ_Y} denote the empirical measure of the Y_i letters in the set Δ_Y, and note that, denoting $v_\Delta = |\Delta_Y|/k$,

$$\begin{aligned} P(L_{x,k,\xi} = \phi, L^{\Delta}_{x,k,\xi} \in G_{2\eta} | L_{1,k',0} = \nu') &\leq (c_1 y + 1)^{|\Sigma|} \sup_{\phi \in G_\eta, \psi \in G_{2\eta}} P(v_\Delta L^{\Delta_Y} + (1 - v_\Delta)\psi_Y = \phi_Y) \\ &\leq (c_1 y + 1)^{|\Sigma|} P(\|L^{\Delta_Y} - \alpha^*_Y\| \leq 3c_1\eta/\delta) . \end{aligned}$$

Therefore, using again (2.20) and the combinatorial upper bound from [DZ, Lemma 2.1.9], and choosing $\delta = \delta(\eta)$ not too small such that $3c_1\eta/\delta < \|\alpha^*_Y - \mu_Y\|/2$ (this is always possible for small η since $\alpha^*_Y \neq \mu_Y$), one obtains

$$\frac{A_2}{p_y} \leq (c_1 y + 1)^{2|\Sigma|} e^{-\delta y \|\alpha^*_Y - \mu_Y\|^2/8} . \tag{2.21}$$

Note that one may have both η small and $\delta = \delta(\eta)$ small (for example, by choosing $\delta = \delta(\eta) = \sqrt{\eta}$ and taking η small enough). Combining (2.19) and (2.21), one obtains that for any $e \in \mathcal{E}_b$, and every $\eta > 0$ small enough,

$$\frac{P(J_{x,k,\nu,\xi}, J_{x',k',\nu',0})}{p_y} \leq g_1(y) e^{-\kappa(\eta)y} \tag{2.22}$$

where $g_1(y)$ is independent of e and of η, $y^{-1} \log g_1(y) \to 0$ with y, and $\kappa(\eta) > 0$.

Case (c) Note that since $k \geq k'$ and $|\Delta_X| \leq |\Delta_Y| \leq \delta y$, necessarily $k - k' \leq \delta y$, and $\xi \leq 2\delta y$. Let now $Z_i = ((Z_i)_X, (Z_i)_Y)$ denote the following (relabeled) random variables:

$$(Z_i)_X = X_{x-1+i\xi}, (Z_i)_Y = Y_{x-1+(i+1)\xi}, \ , i = 0, 1, \ldots, ([\frac{k}{\xi}] - 1)$$

$$(Z_i)_X = X_{x+(i-[\frac{k}{\xi}])\xi}, (Z_i)_Y = Y_{x+(i-[\frac{k}{\xi}]+1)\xi}, \ , i = [\frac{k}{\xi}], \ldots, (2[\frac{k}{\xi}]-1),$$

etc, up to $i = [\frac{k}{\xi}]\xi - 1$. Complete this construction up to $i = k$ in such a way that the empirical measure of $(Z_1, \ldots, Z_k)$ is $L_{x,k,\xi}$. Define next the empirical measure

$$L_k = \frac{1}{k} \sum_{i=0}^{k-1} \delta_{Z_i Z_{i+1}} \in M_1(\Sigma^2).$$

For any $\theta \in M_1(\Sigma_X \times \Sigma_Y \times \Sigma_X \times \Sigma_Y)$, let

$$(\theta)_1 = \sum_{\substack{x_2 \in \Sigma_X \\ y_2 \in \Sigma_Y}} \theta(\cdot, \cdot, x_2, y_2) \in M_1(\Sigma)$$

$$(\theta)_2 = \sum_{\substack{x_1 \in \Sigma_X \\ y_1 \in \Sigma_Y}} \theta(x_1, y_1, \cdot, \cdot) \in M_1(\Sigma)$$

and

$$(\theta)_{12} = \sum_{\substack{x_1 \in \Sigma_X \\ y_2 \in \Sigma_Y}} \theta(x_1, \cdot, \cdot, y_2) \in M_1(\Sigma).$$

Note that $(L_k)_2 = L_{x,k,\xi}$, $\|(L_k)_1 - L_{x,k,\xi}\| \leq 2/k$, and $\|(L_k)_{12} - L_{1,k',0}\| \leq (4\xi + 4\delta y)/k \leq 12\delta y/k \leq 12\delta ||F||_\infty$. Hence, with $\epsilon = \eta + 12\delta ||F||_\infty$, for all large y

$$P(J_{x,k,\nu,\xi}, J_{1,k',\nu',0}) \leq (c_1 y + 1)^{2|\Sigma|} \sup_{\theta_1, \theta_2 \in G_\epsilon} P((L_k)_1 = \theta_1, (L_k)_{12} = \theta_2, (L_k)_2 = \nu). \tag{2.23}$$

For any ν, it follows from the Markov structure of the chain $\{(Z_i Z_{i+1})\}_i$, that

$$P(L_k = \nu) \leq e^{-kH(\nu|(\nu)_1 \times \mu_X \times \mu_Y)} \tag{2.24}$$

(see [CCC, Lemma 3], or [DZ, Exercise 3.1.21]). Using (2.23) and (2.24), one obtains that

$$\frac{P(J_{x,k,\nu,\xi}, J_{x',k',\nu',0})}{p_y} \leq g_2(y) \exp(-k \inf_{\theta \in \Theta_\epsilon} (H(\theta|(\theta)_1 \times \mu_X \times \mu_Y) - H((\theta)_2|\mu_X \times \mu_Y))),$$

where $\Theta_\epsilon = \{\theta \in M_1(\Sigma^2) : (\theta)_1, (\theta)_2, (\theta)_{12} \in G_\epsilon\}$ and $y^{-1} \log g_2(y) \to 0$ with y, independently of $e \in \mathcal{E}_c$ and of η.

It is easy to check that for all $\theta \in M_1(\Sigma^2)$

$$H(\theta|(\theta)_1 \times \mu_X \times \mu_Y) - H((\theta)_2|\mu_X \times \mu_Y) = H(\theta|(\theta)_1 \times (\theta)_2) \geq 0, \tag{2.25}$$

with equality iff $\theta = (\theta)_1 \times (\theta)_2$. Equality cannot be achieved in (2.25) when $(\theta)_1 = (\theta)_2 = (\theta)_{12} = \alpha^*$ since by (E'), $(\alpha^* \times \alpha^*)_{12} = \alpha^*_X \times \alpha^*_Y \neq \alpha^*$. In view of the continuity of $\theta \mapsto H(\theta|(\theta)_1 \times (\theta)_2)$ and the compactness of $M_1(\Sigma^2)$, it follows that for all $\epsilon = \eta + 12\delta||F||_\infty$ small enough

$$\beta'(\epsilon) = \inf_{\theta \in \Theta_\epsilon} \{H(\theta|(\theta)_1 \times \mu_X \times \mu_Y) - H((\theta)_2|\mu_X \times \mu_Y)\} > 0.$$

This in turn implies, for η, δ small enough (again, the choice $\delta = \sqrt{\eta}$ with η small enough will do) and $\beta = \beta'(\epsilon)/||F||_\infty > 0$, that for each $e \in \mathcal{E}_c$,

$$\frac{P(J_{x,k,\nu,\xi}, J_{x',k',\nu',0})}{p_y} \leq g_2(y) e^{-\beta y} . \tag{2.26}$$

Combining now (2.18), (2.22) and (2.26), one sees that $\lim_{y\to\infty} Q_2(y) = 0$ (see (2.17)), completing the proof of the lemma. ∎

3. Proof of (1.8) and an example satisfying (E_Y).

Proof of (1.8): By (E_Y) and the continuity of $H(\cdot|\mu_X \times \mu_Y)$ there exists a relatively open subset U of $\{\nu : \nu(\bar{\Sigma}) = 1\}$ such that $U_Y = \{\nu_Y : \nu \in U\}$ is an open neighborhood of β^* and

$$\sup_{\nu \in U} \{H(\nu|\mu_X \times \nu_Y) - H(\nu_Y|\mu_Y)\} \leq \frac{1-\delta}{1+\delta} H(\beta^*|\mu_Y) ,$$

for some $\delta > 0$. Let $I_n = \{\Delta : |H(\beta^*|\mu_Y)\Delta/\log n - 1| \leq \delta\}$ and set Δ_n, $j_n \leq n - \Delta_n$ to be such that $\bar{R}_n^Y = \sum_{k=1}^{\Delta_n} G_Y(Y_{j_n+k})$. Note that $M_n = \bar{R}_n^Y$ if for some $i = 0, \ldots, [n/\Delta_n] - 1$ the empirical measure $L_{\Delta_n}^{T^{i\Delta_n}\mathbf{X}, T^{j_n}\mathbf{Y}}$ of the pairings $(X_{i\Delta_n+k}, Y_{j_n+k})$ is supported on $\bar{\Sigma}$. By [DKZ, Theorem 2],

$$q_n = P(\Delta_n \in I_n, L_{\Delta_n}^{T^{j_n}\mathbf{Y}} \in U_Y) \to_{n\to\infty} 1 .$$

For n large enough, every $\Delta_n \in I_n$ and all i,

$$P(L_{\Delta_n}^{T^{i\Delta_n}\mathbf{X}, T^{j_n}\mathbf{Y}} \in U | \Delta_n, j_n, L_{\Delta_n}^{T^{j_n}\mathbf{Y}} \in U_Y) \geq (\Delta_n + 1)^{-(|\Sigma|-1)} e^{-\Delta_n(1-\delta)H(\beta^*|\mu_Y)/(1+\delta)} = p(\Delta_n)$$

(see [DKZ, (3) and (5)]). For some $c > 0$ and all n large enough, $\inf_{\Delta \in I_n} [n/\Delta] p(\Delta) \geq c n^{\delta/2}$. Hence, by the independence of $(X_{i\Delta_n+1}, \ldots, X_{i\Delta_n+\Delta_n})$,

$$P(M_n = \bar{R}_n^Y) \geq q_n \inf_{\Delta \in I_n} \{1 - (1 - p(\Delta))^{[n/\Delta]}\} \geq q_n (1 - e^{-cn^{\delta/2}}) \to_{n\to\infty} 1. \quad ∎$$

The following example satisfies (E_Y) for $\Sigma_X = \Sigma_Y = \{0, 1, 2\}$. Let $\mu_X(i) = 1/3$, $i = 0, 1, 2$, $\mu_Y(0) = \mu_Y(1) = 1/6$ and consider the symmetric score $F(x, y) = 1$ for $x + y < 2$ while $F(x, y) = -\infty$ otherwise (so $F(x, y) \neq F(x) + F(y)$). Here, $E_{\mu_Y}(G_Y) = -\infty$ and $\bar{\Sigma} = \{(0,0), (0,1), (1,0)\}$, with $\bar{\theta}^* = H(\beta^*|\mu_Y) = \log 3$, $\beta^*(0) = \beta^*(1) = 1/2$ and $E_\nu(F) = 1$ as soon as $\nu(\bar{\Sigma}) = 1$. Thus, (E_Y) holds since $H(\nu|\mu_X \times \mu_Y) < 2\log 3$ for $\nu((0,1)) = 1/2$, $\nu((0,0)) = \nu((1,0)) = 1/4$. In this particular example, $\theta^* = \log 6$ hence $1/\theta^* < \gamma^* < 2/\theta^*$, while $\mathcal{M} = \{\nu : \nu((0,1)) = 1/2, \nu((0,0)) + \nu((1,0)) = 1/2\}$ is the set of limit points of the empirical measure of pairings $(X_{i+\ell}, Y_{j+\ell})$ over the segment where M_n is achieved (cf. [DKZ, Theorem 2]). In particular, $|\mathcal{M}| = \infty$, $\alpha^* \notin \mathcal{M}$ and (E') fails while M_n possesses a limit distribution of type I (up to lattice effects as in (1.4)).

References

[AW] Arratia R. and Waterman M. S. (1994). A phase transition for the score in matching random sequences allowing deletions. *Annals of Applied Probability.* To appear.

[AGG] Arratia R., Goldstein L. and Gordon L. (1989). Two moments suffice for Poisson approximations: The Chen-Stein method. *Annals of Probability*, **17** 9-25.

[AGW86] Arratia, R., Gordon, L. and Waterman, M. S. (1986). An extreme value theory for sequence matching. *Ann. Statist.* **14** 971-993.

[AGW90] Arratia, R., Gordon, L. and Waterman, M. S. (1990). The Erdös-Rényi law in distribution for coin tossing and sequence matching. *Ann. Statist.* **18** 539-570.

[CCC] Csiszár, I., Cover, T. M. and Choi, B. S. (1987). Conditional limit theorems under Markov conditioning. *IEEE Trans. Inf. Theory* **33** 788-801.

[DK91] Dembo A. and Karlin S. (1991). Strong Limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables. *Annals of Probability*, **19** 1737-1755.

[DK92] Dembo A. and Karlin S. (1992). Poisson Approximations for r-scan processes. *Annals of Applied Probability*, **2** 329-357.

[DKZ] Dembo A. Karlin S. and Zeitouni O. (1994). Critical phenomena for sequence matching with scoring. *Annals of Probability.* To appear.

[DZ] Dembo, A. and Zeitouni, O. (1993). *Large Deviations Techniques and Applications*, Jones and Bartlett, Boston.

[Igl] Iglehart D. (1972). Extreme values in the GI/G/1 queue. *Annals Mathematical Statistics* **43** 627-635.

[KD] Karlin S. and Dembo A. (1992). Limit distributions of maximal segmental score among Markov dependent partial sums. *Advances in Applied Probability*, **24** 113-140.

[KO] Karlin, S. and Ost, F. (1988). Maximal length of common words among random letter sequences. *Annals of Probability* **16** 535-563.

[Neu] Neuhauser, C. (1994). A Poisson approximation for sequence comparisons with insertions and deletions. *Annals of Statistics.* To appear.