A Comparison of Scores of Two Protein Structures With Foldings

Tiefeng Jiang [1]

January 1, 2000

Abstract Let $\{X_i; i \geq 1\}, \{Y_i; i \geq 1\}, \{U, U_i; i \geq 1\}$ and $\{V, V_i; i \geq 1\}$ be four i.i.d. sequences of random variables. Suppose U and V are uniformly distributed on $[0,1]^3$. For each realization of $\{U_j; 1 \leq j \leq n\}$, $\{X_{i,p}; 1 \leq p \leq n\}$ is constructed as a certain permutation of $\{X_p; 1 \leq p \leq n\}$ for any $1 \leq i \leq n$. Also, $\{Y_{j,p}; 1 \leq p \leq n\}, 1 \leq j \leq n$, are constructed the same way based on $\{Y_j\}$ and $\{V_j\}$. For a score function F, we show that

$$W_n := \max_{1 \leq i,j,m \leq n} \sum_{p=1}^{m} F(X_{i,p}, Y_{j,p})$$

has an asymptotic extreme distribution with the same parameters as in the one-dimensional case. This model is constructed for a comparison of scores of protein structures with foldings.

1 Introduction

During the last fifteen years, a number of authors studied biomolecular problems, for example, Arratia and Waterman in [6], Arratia, Morris and Waterman in [5], Karlin and Ost in [16], Arratia, Gordon and Waterman in [2] and [3], Karlin and Altschul in [15], Karlin in [14], Dembo, Karlin and Zeitouni in [8] and [9]. Based on these works, the BLAST program (see Altschul, Gish, Miller, Myers and Lipman in [1] and States, Gish and Altschul in [19]) was established and is being used widely in the area of bioinformatics. The above papers are also used for algorithms founded on information (likelihood ratio) scoring matrices as in Stormo and Hartzell in [20] and Henikoff and Henikoff in [12]. We next review some bimolecular background relevant to this paper.

A protein is a polymer with a linear single chain called backbone composed of peptide bonds. Amino acids are the building blocks of protein. An amino acid has three functional ends: an amino end, a carboxyl end, and a side chain. There are 20 different amino acids

[1] School of Statistics, University of Minnesota.

Key Words: maxima, Chen-Stein method and large deviations.

AMS (1991) subject classifications: 60F10, 60B10

and they differ only in their side chain composition in either charge, hydrophobic or chemical properties. The amino backbone of one amino acid links to the carboxyl backbone end of another amino acid to form the peptide bond that is the backbone of a protein chain. This single chain protein folds into a stable complex three dimensional structure in solution. Although certain secondary structures (alpha helices and beta sheets) can be predicted from the primary linear sequence, the overall three dimensional structure is still beyond the ability of even the best structural prediction algorithms available today.

Suppose the letters of amino acids in a primary linear sequence are $X_1, X_2, \cdots, X_n$. The positions of these amino acids in the three dimensional space are $U_1, U_2, \cdots, U_n$. We usually call $\{U_i;\ i = 1, 2 \cdots, n\}$ the folding of the protein structure. The letter-position pairs of another protein chain are $\{(Y_i, V_i);\ i = 1, 2 \cdots, n\}$. Biologists are interested in local similarities between the two chains, namely, there exist two neighborhoods B_U and B_V such that the alphabets of those amino acids with positions U_i in B_U and V_i in B_V are similar. For example, they may be completely the same or partially matched.

Previous work mentioned so far essentially use the following paradigm. For a given real score function $F(x, y)$, which primarily has negative mean and an essential positive part, they constructed statistics based only on the values of X's and Y's without considering their foldings U's and V's. For example, in [8] and [9], the statistic is the maximum of all $\sum_{l=1}^{\Delta} F(X_{i+l}, Y_{j+l})$ over all possible i, j and Δ running in $\{1, 2, \cdots, n\}$. Since U and V are ignored, statistics used there are not accurate. People have not taken this into account because the folding is very complicated.

Building on [13], we construct a model to compare the scores of two protein structures with foldings $\{U\}$ and $\{V\}$, and then give its asymptotic distribution. Now let us state our main result.

Let $\{X, X_i,\ i = 1, 2, \cdots\}$ be a sequence of i.i.d. random variables with values in a metric space Σ (not necessarily $\mathbb{R}^d$) and let the same be true of $\{Y, Y_i,\ i = 1, 2, \cdots\}$. Let $\{U, U_i,\ i = 1, 2, \cdots\}$ and $\{V, V_i,\ i = 1, 2, \cdots\}$ be two sequences of i.i.d. random variables with both the law of U and that of V being the uniform distribution on $[0, 1]^3$. Throughout this paper, we assume that the above four sequences are independent. For any $i \in \{1, 2, \cdots, n\}$, let $\{u_{i,p},\ p = 1, 2, \cdots, n\}$ be a permutation of $\{1, 2, \cdots, n\}$ such that

$$0 = \|U_{u_{i,1}} - U_i\| < \|U_{u_{i,2}} - U_i\| < \|U_{u_{i,3}} - U_i\| < \cdots < \|U_{u_{i,n}} - U_i\|,$$

where $\|x\| = \max\{|x_1|, |x_2|, |x_3|\}$ for any $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. In other words, we list U_j's in a row such that their distances to U_i are in an increasing order. Then the corresponding indices are $\{u_{i,p}, 1 \le p \le n\}$. They are well defined with probability one. By the same way, we obtain $\{v_{i,p}\}$ from $\{V_k,\ k=1,2,\cdots,n\}$. For simplicity, write $U_{i,p}=U_{u_{i,p}},\ V_{i,p}=V_{v_{i,p}},\ X_{i,p}=X_{u_{i,p}}$ and $Y_{i,p}=Y_{v_{i,p}}$ for all $1\le i,p\le n$. See Figure 1 below.

$$
\begin{array}{lcl}
X_{1,1},\ X_{1,2},\cdots,\ X_{1,n} & \qquad\qquad & Y_{1,1},\ Y_{1,2},\cdots,\ Y_{1,n} \\
X_{2,1},\ X_{2,2},\cdots,\ X_{2,n} & & Y_{2,1},\ Y_{2,2},\cdots,\ Y_{2,n} \\
\cdots\cdots\cdots & & \cdots\cdots\cdots \\
X_{n,1},\ X_{n,2},\cdots,\ X_{n,n} & & Y_{n,1},\ Y_{n,2},\cdots,\ Y_{n,n}
\end{array}
$$

Figure 1

Define

$$W_n := \max_{1\le i,j,m\le n}\sum_{p=1}^{m} F(X_{i,p},Y_{j,p}),$$

where $F(\cdot,\cdot):\Sigma^2\to\mathbb{R}$ is a given real-valued function. The function $F(\cdot,\cdot)$ is usually called a score function in terms of protein matching problems. The following is the only theorem in this paper.

THEOREM 1 *There are positive constants θ and K given in (1.3) and (1.4) below, respectively, such that as $n\to\infty$, for any $x\in\mathbb{R}$,*

$$P\left(W_n > 2\log n/\theta + x\right)\to 1-e^{-Ke^{-\theta x}}$$

provided (1.2), (1.5) and (1.6) below hold.

In practical problems, letters X's and Y's and distances $\|U_i-U_j\|$'s and $\|V_i-V_j\|$'s between amino acids described as above can be obtained by X-ray. The score function $F(\cdot,\cdot)$ is given according to needs. Then W_n, θ and K can be calculated. Consequently, a statistical hypothesis test can be carried out. For a given score function F, a conclusion on certain local similarities can be made.

Now, we state conditions used in Theorem 1. Denote the logarithm of moment generating function of $F(X,Y)$ and its rate function, respectively, by

$$\Lambda_F(t)=\log E\exp(tF(X,Y)) \quad\text{and}\quad \Lambda_F^*(x)=\sup_{t\in\mathbb{R}}\{tx-\Lambda_F(t)\},\quad t,x\in\mathbb{R}. \tag{1.1}$$

If there is no confusion, we also write $\Lambda_F(t)=\Lambda(t)$. The following condition is standard in this context:

$$F(X,Y)\text{ is non-lattice},\quad \mu_F:=EF(X,Y)<0 \text{ and } \Lambda_F(t)<\infty \text{ for all } t\in\mathbb{R}. \tag{1.2}$$

It is obvious that the above condition implies that there exists an unique $\theta > 0$ so that

$$\Lambda_F(\theta) = 0. \tag{1.3}$$

Also, under (1.2), Spitzer (E4 on page 217 from [18], see also (5.13) in [11] or Lemma A.2 in [13]) has shown that there is a constant $K > 0$ depending on $F(X, Y)$ such that

$$\lim_{x \to +\infty} e^{\theta x} P\left(\max_{n \geq 1} \sum_{i=1}^{n} F(X_i, Y_i) > x\right) \to K. \tag{1.4}$$

Define a measure α^* on Σ^2 by

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

where μ_X and μ_Y are the distributions of X and Y, respectively. For any two probability measures μ, ν on Σ^2 recall the definition of relative entropy $H(\nu|\mu)$:

$$H(\nu|\mu) = \begin{cases} \int_{\Sigma^2} \left(\log \frac{d\nu}{d\mu}\right) d\nu, \text{ if } \nu \ll \mu; \\ +\infty, \text{ otherwise.} \end{cases}$$

The following Dembo-Karlin-Zeitouni condition will be used in our theorem:

$$H(\alpha^*|\mu_X \times \mu_Y) > 2\max(H(\alpha_X^*|\mu_X), H(\alpha_Y^*|\mu_Y)). \tag{1.5}$$

A detailed discussion about (1.5) is given in [8] and [9].

The following is the last condition we need in our theorem. Suppose $\{X, X_1, X_2\}$ are three i.i.d. random variables with values in Σ and $\{Y, Y_1, Y_2\}$ are another three i.i.d. random variables with values in Σ, which are independent of the previous three. Assume

$$E\{F(X_1, Y)e^{\theta F(X_2, Y)}\} < 0 \text{ and } E\{F(X, Y_1)e^{\theta F(X, Y_2)}\} < 0. \tag{1.6}$$

The example below is motivated by the analysis of protein sequences. See Remark 1 in [9].
Example. On some space Σ with $1 < |\Sigma| < \infty$, define $F(x, x) = 1$ and $F(x, y) = -m\sqrt{2}$ for $x \neq y$, $x, y \in \Sigma$ with $m \in \mathbb{N}$. Then, $F(\cdot, \cdot)$ is non-lattice. It is easy to check that conditions (1.2) and (1.6) hold for sufficiently large m. Furthermore, condition (1.5) holds for sufficiently large m provided the following condition ((1.7) in [9]) is true:

$$\max\left\{\sum_{i \in \Sigma} \mu_X(i)\mu_Y(i) \log \mu_Y(i), \sum_{i \in \Sigma} \mu_X(i)\mu_Y(i) \log \mu_X(i)\right\} < -\frac{1}{2}\theta e^{-\theta}.$$

Remark 1. In a real protein structure, the physical positions of amino acids, i.e, U's and V's, are not uniformly distributed in a cube. However, the proof of Theorem 1 indicates that it does not depend on the specific geometry of a cube but will apply in the case of any regular geometric shape on which $\{U_i; i = 1, 2, \cdots\}$ are uniformly or close to being uniformly distributed. Because of convenience of mathematical proofs, the condition that letters X's and their positions U's are independent is assumed in the above theorem. The proof indicates that this can be relaxed too. However, it is open to what extent independence can be relaxed.

Remark 2. In the theorem, we use the maximum norm $\|(x_1, x_2, x_3)\| = \max\{|x_1|, |x_2|, |x_3|\}$ to order a sequence of points $\{U_1, U_2, \cdots, U_n\}$ in $\mathbb{R}^3$. According to the proof, the theorem still holds if the maximum distance is replaced by other distances equivalent to the Euclidian distance.

Remark 3. As shown in Remark 3 from [9], condition (1.5) is almost necessary in the one-dimensional setting. By contrast, (1.6) is not required in the one dimensional counterpart of Theorem 1. We need this condition here because we have a more complicated structure than in the one-dimensional case as in [9]. Also, we do not consider the case of indels in this paper. So it is open if condition (1.6) can be dropped and how W_n can be adjusted in the indel-case.

Finally, we make some remarks about the proof of the theorem. The main tool of the proof is the Chen-Stein Poisson approximation method. Certain sharp large deviation results are used to estimate rare events. Unlike the one-dimensional case, the renewal phenomenon no longer occurs in our three-dimensional model. Thus the usual declumping appearing in sequence by sequence matching (see e.g., [9]) disappears from our proof. This is because the model is constructed in three-dimensional spaces in which some phenomena are much different from the one-dimensional counterpart. For example, a simple random walk on one or two-dimensional lattice points is recurrent; but it is transient in the three-dimensional case. One can see the fact of non-declumping clearly from the proof of Theorem 9 in [13], which is the motivating paper for the current one. Also, a detailed explantation of declumping can be found in section 2 on page 543 of [2].

This paper is organized as follows: In section 2, the proof of Theorem 1 is given; in section 3, some technical lemmas used in section 2 are proved.

2 The Proof of the Theorem

The following Poisson approximation theorem is a straightforward application of Theorem 1 in [4] (see also lemma 2.2 in [13]), which is a special case of the Chen-Stein method. The lemma is used quite often in analyzing maxima of random variables. It is the starting point of the proof of Theorem 1.

LEMMA 2.1 *Let Ω be a finite set and $\mathcal{A}$ is a collection of some subsets of Ω. Let $\{X_\alpha,\ \alpha \in \Omega\}$ be a collection of random variables. Write $S_A = \sum_{\alpha\in A} X_\alpha$ and $\lambda = \sum_{A\in\mathcal{A}} P(S_A > t)$ for some $t \in \mathbb{R}$. Then*

$$|P(\max_{A\in\mathcal{A}} S_A \le t) - e^{-\lambda}| \le (1 \wedge \lambda^{-1})(b_1 + b_2 + b_3),$$

where

$$b_1 = \sum_{A\in\mathcal{A}} \sum_{B: B\cap A\ne\emptyset} P(S_A > t)P(S_B > t), \quad b_2 = \sum_{A\in\mathcal{A}} \sum_{B: B\cap A\ne\emptyset} P(S_A > t,\ S_B > t),$$
$$b_3 = \sum_{A\in\mathcal{A}} E|P(S_A > t|\sigma\{S_B; B\cap A = \emptyset\}) - P(S_A > t)|,$$

where $\sigma\{S_B; B\cap A = \emptyset\}$ is the σ-algebra generated by the collection of random variables $\{S_B; B\cap A = \emptyset\}$. In particular, if $\{X_\alpha,\ \alpha\in\Omega\}$ is a set of independent random variables, then $b_3 = 0$.

Now, let us sketch the proof. The first step is to get rid of many subcubes appearing in the definition of W_n. Then W_n is reduced to a simple form $W_{n,1}$. This is shown in Lemma 2.2 next. The second step is analyzing $W_{n,1}$ by applying the Poisson-approximation method Lemma 2.1 through the following two sub-steps: (i) show that λ in lemma 2.1 has a limit, which is given in Lemma 2.4; (ii) prove that b_1 and b_2 in Lemma 2.1 go to zero as $n\to\infty$.

For any $x = (x_1, x_2, x_3) \in \mathbb{R}^3$, $\|x\| = \max\{|x_1|, |x_2|, |x_3|\}$ is the maximum norm. A ball centered at x and with radius r under this norm is denoted by $B(x, r)$. Thus, the volume of such a box is $8r^3$. Recall θ is given in (1.3). For two numbers a and b, by $a\wedge b$ we mean $\min\{a, b\}$. For any positive numbers ρ and λ, define

$$\begin{aligned}
l_n^{\pm} &= \frac{1}{2}\left(\frac{2\log n}{\theta\Lambda'(\theta)n}\right)^{1/3}\left(1 \pm \rho\sqrt{\frac{\log_2 n}{\log n}}\right),\\
T_{i,j}^{\pm} &= \sum_{p=1}^{n} 1_{B(U_i, l_n^{\pm})}(U_p) \wedge \sum_{p=1}^{n} 1_{B(V_j, l_n^{\pm})}(V_p),\\
\Omega_n &= \{m \ge 1;\ |m - 2\log n/\theta\Lambda'(\theta)| \le \lambda\sqrt{(\log n)\log_2 n}\,\}, \qquad (2.1)\\
\Psi_{i,j} &= \left\{m;\ T_{i,j}^{-} \le m \le T_{i,j}^{+}\right\} \cap \Omega_n,\\
W_{n,1} &= \max_{\substack{m\in\Psi_{i,j}\\ 1\le i,j\le n}} \sum_{p=1}^{m} F(X_{i,p}, Y_{j,p}),
\end{aligned}$$

where $\log_2 n = \log(\log n)$. By $l_n^{\pm}$ we mean l_n^{+} or l_n^{-} depending on the "+" sign or the "−" sign occurring in the first assertion of (2.1). This interpretation applies to $T_{i,j}^{\pm}$ too.

Suppose $\xi_1, \xi_2, \cdots, \xi_n$ are random variables. Let $f(x_1, x_2, \cdots, x_n)$ be a real-valued function on $\mathbb{R}^n$. Define

$$E^A f(\xi_1, \xi_2, \cdots, \xi_n) = E(f(\xi_1, \xi_2, \cdots, \xi_n)|\mathcal{B}), \tag{2.2}$$

the conditional expectation of $f(\xi_1, \xi_2, \cdots, \xi_n)$ given $\mathcal{B}$, where $\mathcal{B}$ is the σ-algebra generated by $\{\xi_k, k \notin A\}$ if $A \subset \{1, 2, \cdots, n\}$ or by $\{\xi_1, \cdots, \xi_n\} \backslash A$ if A is a subset of $\{\xi_k, 1 \le k \le n\}$. The same interpretation applies to P^A too.

For convenience, throughout all this section, we set $z_n = 2 \log n/\theta + x$.

Accept, for now, the following two lemmas. They together with other lemmas in this section will be proved in Section 3.

LEMMA 2.2 *Suppose condition (1.2) holds. Then, for any $x \in \mathbb{R}$,*

$$P(W_n > z_n) - P(W_{n,1} > z_n) \to 0 \tag{2.3}$$

for sufficiently large λ and ρ.

Set

$$\begin{aligned} &\phi_X(x) = \log\{E^Y e^{\theta F(x,Y)}\}, \ h_X = E\left\{e^{\theta F(X,Y)} \phi_X(X)\right\}; \\ &\phi_Y(y) = \log\{E^X e^{\theta F(X,y)}\}, \ h_Y = E\left\{e^{\theta F(X,Y)} \phi_Y(Y)\right\} \text{ and} \\ &G_{(i,j,m)}(\epsilon) = \left\{ |\frac{1}{m} \sum_{p=1}^{m} \phi_X(X_{i,p}) - h_X| < \epsilon \right\} \bigcap \left\{ |\frac{1}{m} \sum_{p=1}^{m} \phi_Y(Y_{j,p}) - h_Y| < \epsilon \right\}. \end{aligned} \tag{2.4}$$

Since $E \exp(\theta F(X,Y)) = 1$, h_X and h_Y are actually mean values of $\phi_X(x)$ and $\phi_Y(y)$, respectively, under the measure induced by $\exp(\theta F(X,Y))$. We need the following lemma.

LEMMA 2.3 *Suppose condition (1.2) holds. Then, for any $a \in (0,1), \epsilon > 0$ and sequence $\{\gamma_n;\ n \ge 1\}$ so that $\gamma_n \to \gamma \in \mathbb{R}$, there exists $\delta > 0$ such that*

$$\max_{an \le k \le n} P\left(\frac{1}{n} \sum_{i=1}^{n} F(X_i, Y_i) \ge \gamma_n,\ \left| \frac{1}{k} \sum_{i=1}^{k} \phi_Y(Y_i) - h_Y \right| \ge \epsilon \right) = o(e^{-(\theta\gamma + \delta)n}).$$

By Cramér's large deviation result, the probability of the first event in the above lemma is roughly $e^{-\theta\gamma n}$. So the interpretation of Lemma 2.3 is as follows: given that the rare event $(1/n) \sum_{i=1}^{n} F(X_i, Y_i) \ge \gamma_n$ occurs, the second event in Lemma 2.3 is also rare. As a consequence, as k is large, $(1/k) \sum_{i=1}^{k} \phi_Y(Y_i)$ is around its mean h_Y under the measure induced by $\exp(\theta F(X,Y))$ rather than the product measure induced by X and Y. This is a key observation in the proof.

Lemma 2.4 *Suppose condition (1.2) holds. Then for any $\epsilon > 0$ small enough, there exist $\lambda > 0$ and $\rho > 0$ such that*

$$b_{1,n} := \sum_{1 \le i,j \le n} P^{X,Y}\left(\bigcup_{m \in \Psi_{i,j}} \left\{ \sum_{p=1}^{m} F(X_{i,p}, Y_{j,p}) \ge z_n,\, G_{(i,j,m)}(\epsilon) \right\}\right) \to Ke^{-\theta x}$$

in probability (on U and V) for any $x \in \mathbb{R}$, where K is as in (1.4).

With the explantation given before Lemma 2.4, the intersection of the two events inside the $\{\cdot\}$ in Lemma 2.4 is roughly equal to the first one. Therefore, $b_{1,n}$ is close to

$$\sum_{1 \le i,j \le n} P^{X,Y}\left(\max_{m \in \Psi_{i,j}} \sum_{p=1}^{m} F(X_{i,p}, Y_{j,p}) \ge z_n\right) \sim n^2 P\left(\max_{m \ge 1} \sum_{p=1}^{m} F(X_p, Y_p) \ge z_n\right) \to Ke^{-\theta x}$$

by (1.4). This is what exactly happens in Lemma 2.4.

One fact we constantly use in proofs is that the law of $\sum_{p=1}^{m} F(X_{i,p}, Y_{j,p})$ conditioned, or not, on U and V is equal to the law of $\sum_{p=1}^{m} F(X_p, Y_p)$ for any i, j and m.

Proof of Theorem 1. By Lemma 2.2, it is enough to show that

$$P\left(W_{n,1} > z_n\right) \to 1 - e^{-Ke^{-\theta x}}. \tag{2.5}$$

Define

$$A_{i,j} = \bigcup_{m \in \Psi_{i,j}} \left\{ \sum_{p=1}^{m} F(X_{i,p}, Y_{j,p}) \ge z_n, G_{(i,j,m)}(\epsilon) \right\}, \quad 1 \le i,j \le n. \tag{2.6}$$

Note that the difference between $\{W_{n,1} > z_n\}$ and $\cup_{1 \le i,j \le n} A_{i,j}$ is a subset of $\cup_{1 \le i,j \le n} \cup_{m \in \Psi_{i,j}}$ $\left\{\sum_{p=1}^{m} F(X_{i,p}, Y_{j,p}) \ge z_n, G_{(i,j,m)}(\epsilon)^c\right\}$. By Lemma 2.3, we have

$$\begin{aligned} & P\left(W_{n,1} > z_n\right) - P\left(\cup_{1 \le i,j \le n} A_{i,j}\right) \\ \le \; & n^2 |\Omega_n| \max_{m \in \Omega_n} P\left(\sum_{p=1}^{m} F(X_{1,p}, Y_{1,p}) \ge z_n, G_{(1,1,m)}(\epsilon)^c\right) = O(n^{-\delta_0}) \end{aligned}$$

for some constant $\delta_0 > 0$. Thus, to prove (2.5), it suffices to prove that

$$P\left(\cup_{1 \le i,j \le n} A_{i,j}\right) \to 1 - e^{-Ke^{-\theta x}}. \tag{2.7}$$

Recalling $b_{1,n}$ defined in Lemma 2.4, we have

$$\begin{aligned} & |P\left(\cup_{1 \le i,j \le n} A_{i,j}\right) - e^{-Ke^{-\theta x}}| \\ \le \; & E^{U,V}\left|P^{X,Y}\left(\cup_{1 \le i,j \le n} A_{i,j}\right) - e^{-b_{1,n}}\right| + E^{U,V}\left|e^{-b_{1,n}} - e^{-Ke^{-\theta x}}\right|. \end{aligned} \tag{2.8}$$

By Lemma 2.4, $E^{U,V}|e^{-b_{1,n}} - e^{-Ke^{-\theta x}}| \to 0$. It is enough to show the first term in (2.8) goes to 0. Actually, by Lemma 2.1,

$$E^{U,V}\left| P^{X,Y}\left(W_{n,1} \le z_n\right) - e^{-b_{1,n}} \right| \le E^{U,V} b_{2,n} + E^{U,V} b_{3,n},$$

where

$$b_{2,n} = \sum_{i,j=1}^{n} \sum_{(k,l)\in\Gamma_{i,j}} P(A_{i,j})P(A_{k,l}),\ \ b_{3,n} = \sum_{i,j=1}^{n} \sum_{(k,l)\in\Gamma_{i,j}} P(A_{i,j}\cap A_{k,l}), \quad \text{and}$$

$$\Gamma_{i,j} = \{(k,l) \in \{1,2,\cdots,n\}^2 \backslash (i,j); B(U_k, l_n^+) \cap B(U_i, l_n^+) \ne \emptyset \text{ or}$$
$$B(V_l, l_n^+) \cap B(V_j, l_n^+) \ne \emptyset\}.$$

In words, (k,l) belongs to $\Gamma_{i,j}$ if one of the following is true: (i) two balls with the same radius l_n^+ and centered at U_k and U_i, respectively, intersect; (ii) two balls with the same radius and centered at V_l and V_j, respectively, intersect.

It is easy to see by Doob's submartingale inequality that $P^{X,Y}(A_{1,1}) \le e^{-\theta x} n^{-2}$ *a.s.* for each $n \ge 1$. It follows that $E^{U,V} b_{2,n} \le e^{-2\theta x} n^{-2} E^{U,V}(\#\Gamma_{1,1}) \le 2e^{-2\theta x} n^{-1} E^{U,V}(\#\Xi)$, where $\Xi = \{2 \le k \le n;\ B(U_k, l_n^+) \cap B(U_1, l_n^+) \ne \emptyset\}$. But note that $\Xi = \sum_{k=2}^{n} 1\{d(U_k, U_1) \le 2l_n^+\}$, so $E\Xi = O(\log n)$, and thus $E^{U,V} b_{2,n} = O(e^{-2\theta x} n^{-1} \log n)$. So the remaining task is to show that $E^{U,V} b_{3,n} \to 0$.

By using symmetry, we see that

$$E^{U,V} b_{3,n} = n^2 E^{U,V} \sum_{(k,l)\in\Gamma_{1,1}} P(A_{k,l} \cap A_{1,1}) \le n^3 E^{U,V} P(A_{2,1} \cap A_{1,1})$$
$$+ n^3 E^{U,V} P(A_{1,2} \cap A_{1,1}) + 2n^4 E^{U,V}[P(A_{2,2} \cap A_{1,1}) 1\{d(U_1, U_2) \le 2l_n^+\}].$$

Lemmas 2.5 and 2.6 next show that the first and last terms on the right hand side of the inequality above go to zero. By symmetry and Lemma 2.6 again, the middle term goes to zero too. The proof is completed. ■

Lemma 2.5 *Under conditions of Theorem 1,*

$$E^{U,V}[P(A_{2,2} \cap A_{1,1}) 1\{d(U_1, U_2) \le 2l_n^+\}] = o(n^{-4})$$

for sufficiently small $\epsilon > 0$, where ϵ is as in (2.6).

Lemma 2.6 *Under conditions of Theorem 1,*

$$E^{U,V} P(A_{1,1} \cap A_{1,2}) = o(n^{-3})$$

for sufficiently small $\epsilon > 0$, where ϵ is as in (2.6).

3 Technical Lemmas

In this section, we will prove the lemmas stated in Section 2. The following four results are needed for doing that.

Lemma 3.1 *Suppose condition (1.2) holds. Then, for any $\epsilon > 0$ and $n \geq 1$,*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\phi_X(X_i) - h_X\right| \leq \epsilon\right) \leq 2e^{-n(h_X-\epsilon)}.$$

Proof. By Lemma A.2, the above probability is at most $2e^{-n\inf_{x\in A}J(x)}$, where $J(x) = \sup_{t\in\mathbb{R}}\{tx - \log E\left(E^Y e^{\theta F(X,Y)}\right)^t\}$ and $A = \{x \in \mathbb{R};\ |x - h_X| \leq \epsilon\}$. By taking $t = 1$ in the definition of $J(x)$ we have that $J(x) \geq x$. The lemma is proved. ■

Lemma 3.2 *Suppose the second inequality of (1.6) holds. Let $M(t) = E\exp(tF(X,Y_1) + \theta F(X,Y_2))$ and $t_0 = \sup\{t > 0; M(t) < 1\}$. Then*
(i) $t_0 \in (0,\theta)$,
(ii) There exists $\delta \in (\mu_F, 0)$ such that $\gamma_1 := \sup_{0<t<t_0}\{\delta t - \Lambda_F(t)\} > 0$ and $\gamma_2 := \sup_{0<t<t_0}\{\delta t - \log M(t)\} > 0$, where $\mu_F = EF(X,Y)$ and $\Lambda_F(t)$ is as in (1.1).

Proof. (i) Note that $M(0) = 1$ and $M'(0) = EF(X,Y_1)e^{\theta F(X,Y_2)} < 0$, so there exists some $t > 0$ such that $M(t) < 1$. Since $Ee^{\theta F(X,Y)} = 1$, by Hölder's inequality, $M(t) \geq M(\theta)^{t/\theta} = (E^X(E^Y e^{\theta F(X,Y)})^2)^{t/\theta} > (Ee^{\theta F(X,Y)})^{2t/\theta} = 1$ for all $t > \theta$. Thus, $t_0 < \theta$.
(ii) By (i), $\Lambda_F(t_0/2) < 0$ and $\log M(t_0/2) < 0$. The conclusion follows by taking $t = t_0/2$ and

$$\delta = \frac{1}{2}\max\left\{\mu_F,\ 2t_0^{-1}\Lambda_F(t_0/2),\ 2t_0^{-1}\log M(t_0/2)\right\}. \quad ■$$

The following lemma is almost immediate. We state it without proof.

Lemma 3.3 *Denote $\xi_X = E^Y e^{\theta F(X,Y)}$ and $\xi_Y = E^X e^{\theta F(X,Y)}$. Then, $h_X = E^X(\xi_X\log\xi_X)$, and $h_Y = E^Y(\xi_Y\log\xi_Y)$, and (1.5) is equivalent to*

$$\frac{1}{2}\theta\Lambda'(\theta) > \max\{E^X(\xi_X\log\xi_X), E^Y(\xi_Y\log\xi_Y)\}.$$

In the following proofs, one should keep in mind that the law of $\sum_{p=1}^m F(X_{i,p}, Y_{j,p})$, regardless of conditioning on U and V, is equal to that of $\sum_{p=1}^m F(X_p, Y_p)$ for any i, j and m.

Lemma 3.4 *Under conditions of Theorem 1, for any sufficiently small $\epsilon > 0$, there exists a constant $\gamma > 0$ so that*

$$\max_{m_1,m_2\in\Omega_n} E^X\left\{\prod_{k=1}^{2} P^Y\left(\sum_{p=1}^{m_k} F(X_{k,p}, Y_{1,p}) \geq z_n, G_{(k,1,m_k)}(\epsilon)\right)\right\} = O(n^{-3-\gamma}),$$

where Ω_n is as in (2.1).

Proof. Recall from (2.4) that $G_{(k,1,m_k)}(\epsilon) \subset H_k := \{|(1/m)\sum_{p=1}^m \phi(X_{k,p}) - h_X| < \epsilon\}$. Then, by Chebyshev's inequality,

$$P^Y\left(\sum_{p=1}^{m_k} F(X_{k,p}, Y_{1,p}) \geq z_n, G_{(k,1,m_k)}(\epsilon)\right) \leq e^{-\theta z_n} E^Y\left[\exp(\theta \sum_{p=1}^{m_k} F(X_{k,p}, Y_{1,p})) \cdot I_{H_k}\right]$$

for $k = 1, 2$. Since H_k does not depend on the Y's, by independence, the expectation above is equal to $\Pi_{p=1}^{m_k}[E^Y \exp(\theta F(X_{k,p}, Y))] \cdot I_{H_k} = \exp(\sum_{p=1}^{m_k} \phi_X(X_{k,p})) \cdot I_{H_k} \leq e^{m_k(h_X+\epsilon)} \cdot I_{H_k}$. Therefore,

$$E^X\left\{\prod_{k=1}^{2} P^Y\left(\sum_{p=1}^{m_k} F(X_{k,p}, Y_{1,p}) \geq z_n, G_{(k,1,m_k)}(\epsilon)\right)\right\} \leq e^{-2\theta z_n} e^{2(m_1 \vee m_2)(h_X+\epsilon)} P^X(H_1 \cap H_2).$$

By Lemma 3.1, $P^X(H_1 \cap H_2) \leq 2\exp(-(m_1 \vee m_2)(h_X - \epsilon))$. Also, by Lemma 3.3, there exists $\epsilon_0 > 0$ such that $\theta\Lambda'(\theta)/2 > \max\{h_X, h_Y\} + \epsilon_0$. Then the desired result follows by noting that $m_1 \vee m_2 \sim (2\log n)/\theta\Lambda'(\theta)$ uniformly for $m_1, m_2 \in \Omega_n$. ∎

The Proof of Lemma 2.2. Let $S_m = \sum_{p=1}^m F(X_p, Y_p)$. Obviously, the left hand side of (2.3) is bounded by the expectation of

$$n^2 \max_{n \geq m \notin \Omega_n} P(S_m \geq z_n) + n^2 \max_{T_{1,1}^+ < m \in \Omega_n} P(S_m \geq z_n) + n^2 \max_{T_{1,1}^- > m \in \Omega_n} P(S_m \geq z_n). \tag{3.9}$$

By Lemmas A.4 and A.5, the first term above $\leq 2e^{-\theta x}(\log n)^{(1/2)-C\lambda^2}$ for large n. Here and throughout the rest of the paper, C always stands for a positive constant depending on X, Y and F and may vary from line to line. By symmetry, we only need to show that the second term of (3.9) goes to 0. Actually, it is no more than

$$n^2 \sum_{m \in \Omega_n} P\left(S_m \geq z_n\right) P^{U,V}(T_{1,1}^+ \leq m). \tag{3.10}$$

By Lemma A.3,

$$P\left(S_m \geq z_n\right) \sim \frac{C^{-1}e^{-\theta x}}{n^2\sqrt{\log n}} \exp(-\overbrace{C(m\Lambda'(\theta) - z_n)^2/\log n}^{I_n})$$

uniformly for all $m \in \Omega_n$. By Bernstein's inequality, there exist two constant C and C' such that

$$\begin{aligned} P(T_{1,1}^+ < m) &\leq 2P\left(\sum_{i=1}^n 1_{B(U_1, l_n^+)} < m\right) \\ &\leq 2\exp(-\underbrace{C'(m\Lambda'(\theta) - z_n - C\rho\sqrt{(\log n)\log_2 n}\ )^2/\log n}_{I_n'}). \end{aligned}$$

Since $a^2+(a-b)^2 \geq b^2/2$ for any $a, b \in \mathbb{R}$, we have that $I_n + I_n' \geq C\rho^2(\log_2 n)$. Therefore, by observing that $|\Omega_n| \leq 2\lambda\sqrt{(\log n)\log_2 n}$, the term in (3.10) is less than

$$\frac{2C^{-1}e^{-\theta x}}{\sqrt{\log n}} \sum_{m\in\Omega_n} e^{-I_n - I_n'} \leq (4C^{-1}\lambda e^{-\theta x})\sqrt{\log_2 n}/(\log n)^{C\rho^2}.$$

The proof is completed by choosing λ and ρ sufficiently large. ■

The Proof of Lemma 2.3. For any sequence $\{k_n, n \geq 1\}$, define

$$Z_n = \left(\frac{1}{n}\sum_{i=1}^{n} F(X_i, Y_i), \frac{1}{k_n}\sum_{i=1}^{k_n}\phi_Y(Y_i)\right) \in \mathbb{R}^2.$$

To prove this lemma, it is enough to show that there exists $\delta > 0$ such that

$$p_n := P\left(Z_n \in [\gamma_n, \infty) \times (h_Y - \epsilon, h_Y + \epsilon)^c\right) = o(e^{-(\theta\gamma+\delta)n}) \tag{3.11}$$

for all subsequences $k_n \leq n$ such that $k_n/n \to a' \in [a, 1]$. Let $F_\eta = [\gamma-\eta, \infty)\times\{y;\ |y-h_Y| \geq \epsilon\}$. Then F_η is a closed set in $\mathbb{R}^2$. It is easy to see that $p_n \leq P\left(Z_n \in F_\eta\right)$ for any $\eta > 0$ and n large enough (depending on η). For any $(u, v) \in \mathbb{R}^2$

$$\begin{aligned}\frac{1}{n}\log Ee^{n(u,v)\cdot Z_n} &= \frac{n-k_n}{n}\log\left\{Ee^{uF(X,Y)}\right\} + \frac{k_n}{n}\log\left\{Ee^{uF(X,Y)+(nv/k_n)\phi_Y(Y)}\right\} \\ &\to (1-a')\log E\left\{e^{uF(X,Y)}\right\} + a'\log\left\{Ee^{uF(X,Y)}(E^X e^{\theta F(X,Y)})^{v/a'}\right\} := g(u,v).\end{aligned}$$

Clearly, $g(u, v)$ is finite and differentiable for any $(u, v) \in \mathbb{R}^2$. For any $\eta > 0$, by the Gätner-Ellis Theorem (see, e.g., Theorem 2.3.6 of [10]),

$$\limsup_{n\to\infty}(\log p_n)/n \leq -\inf_{(x,y)\in F_\eta} I((x,y)),$$

where $I((x,y)) = \sup_{(u,v)\in\mathbb{R}^2}\{ux + vy - g(u,v)\}$. Note that $F_\eta \downarrow F_0$ as $\eta \downarrow 0$. Therefore, by letting $\eta \downarrow 0$, we obtain

$$\limsup_{n\to\infty}(\log p_n)/n \leq -\inf_{(x,y)\in F_0} I((x,y)). \tag{3.12}$$

It is easy to see that $g(\theta, v) \leq a\log\left\{Ee^{\theta F(X,Y)}(E^X e^{\theta F(X,Y)})^{v/a}\right\}$, since $E^{\theta F(X,Y)} = 1$ and $a' \in [a, 1]$. Therefore, for any $x \geq \gamma$ and $y \leq h_Y - \epsilon$, by taking $u = \theta$ and $v = aw$ in the definition of $I((x,y))$, we have that

$$I((x,y)) \geq \theta\gamma + a\cdot\sup_{w\leq 0}\{\underbrace{wh_Y - w\epsilon - \log(Ee^{\theta F(X,Y)}(\overbrace{E^X e^{\theta F(X,Y)})^w}^{\xi(w)})}_{\psi(w)}\}. \tag{3.13}$$

Obviously, $\psi(0)=0$, and observe that $E(e^{\theta F(X,Y)}\xi'(w)) \to h_Y$ as $w \uparrow 0$. It then follows that

$$\psi'(w) = h_Y - \epsilon - \left(Ee^{\theta F(X,Y)}\xi(w)\right)^{-1} E\left(e^{\theta F(X,Y)}\xi'(w)\right) < 0$$

for $w < 0$ and $|w|$ sufficiently small. Therefore, by (3.13), there exists $\delta > 0$ such that $I((x,y)) > \theta\gamma + 2\delta$ for all $x \geq \gamma$ and $y \leq h_Y - \epsilon$. By the same arguments, the inequality $I((x,y)) > \theta\gamma + 2\delta$ is also valid for $x \geq \gamma$ and $y \geq h_Y + \epsilon$. Thus, $\inf_{(x,y)\in F_0} I((x,y)) > \theta\gamma + 2\delta$ for some $\delta > 0$, which together with (3.12), yields (3.11). ■

The Proof of Lemma 2.4. By Lemma 2.3

$$\begin{aligned} & E^{U,V} \sum_{1\leq i,j\leq n} P^{X,Y}\left(\bigcup_{m\in\Psi_{i,j}} \left\{\sum_{p=1}^{m} F(X_{i,p}, Y_{j,p}) \geq z_n,\, G_{(i,j,m)}(\epsilon)^c\right\}\right) \\ \leq\ & n^2|\Omega_n| \max_{m\in\Omega_n} P\left(\sum_{p=1}^{m} F(X_{1,p}, Y_{1,p}) \geq z_n,\, G_{(1,1,m)}(\epsilon)^c\right) = O(n^{-\delta}) \end{aligned}$$

for some $\delta = \delta_{F,\epsilon} > 0$. Therefore, to prove the lemma, it is enough to prove that

$$b'_{1,n} := \sum_{1\leq i,j\leq n} P^{X,Y}\left(\max_{m\in\Psi_{i,j}} \sum_{p=1}^{m} F(X_{i,p}, Y_{j,p}) \geq z_n\right) \to Ke^{-\theta x} \tag{3.14}$$

in probability, where $P^{X,Y}$ is defined in (2.2). Set $J_n^{\pm} = 2\log n/\theta\Lambda'(\theta) \pm \lambda\sqrt{(\log n)\log_2 n}$. Then

$$-n^2 P^{X,Y}\left(\max_{m\notin\Omega_n} \sum_{p=1}^{m} F(X_{1,p}, Y_{1,p}) \geq z_n\right) \tag{3.15}$$

$$\leq\ n^2 P^{X,Y}\left(\max_{m\in\Omega_n} \sum_{p=1}^{m} F(X_{1,p}, Y_{1,p}) \geq z_n\right) - b'_{1,n}$$

$$\leq\ \sum_{1\leq i,j\leq n} P^{X,Y}\left(\max_{m\in\Omega_n} \sum_{p=1}^{m} F(X_{i,p}, Y_{j,p}) \geq z_n\right) 1(T_{i,j} \geq J_n^+ \text{ or } T_{i,j} \leq J_n^-). \tag{3.16}$$

By (1.4), we only need to show both terms in (3.15) and (3.16) go to zero in probability. First, by Lemmas A.4 and A.5, we have that

$$n^2 P^{X,Y}\left(\max_{m\notin\Omega_n} \sum_{p=1}^{m} F(X_{1,p}, Y_{1,p}) \geq z_n\right) \to 0 \quad a.s. \tag{3.17}$$

for sufficiently large λ, which is as in (2.1). Second, since the probability given in (3.16) is less than $Ke^{-\theta x}n^{-2}$, the expectation of the expression in (3.16) is less than

$$Ke^{-\theta x} P(T_{1,1}^+ \geq J_n^+ \text{ or } T_{1,1}^- \leq J_n^-).$$

For ρ given in the definition of $l_n^{\pm}$ in (2.1), by Bernstein's inequality, for sufficiently large λ,

$$P(T_{1,1}^+ \geq J_n^+ \text{ or } T_{1,1}^- \leq J_n^-) \leq 4\exp\left\{-C\left(\frac{6\rho}{\theta\Lambda'(\theta)} - \lambda\right)^2 \log_2 n\right\} \to 0$$

as $n \to \infty$. Thus, (3.14) follows. ■

The Proof of Lemma 2.5. Let $c_n = (\log n)^{-\delta} n^{-1/3}$, $\delta \in (1/3, 2/3)$. Obviously,

$$P(d(U_1, U_2) \leq 2l_n^+) \leq C(\log n)/n, \ P(d(U_1, U_2) \leq c_n) \leq 8/n(\log n)^{3\delta}.$$

Since $P(A_{2,2} \cap A_{1,1}) \leq P(A_{1,1}) \leq n^{-2}$, the above two inequalities yield

$$\begin{aligned} & E^{U,V} P(A_{2,2} \cap A_{1,1}) 1\{d(U_1, U_2) \leq 2l_n^+\} \\ \leq \ & E^{U,V} P(A_{2,2} \cap A_{1,1})(1_{E_{U,1} \cap E_{V,1}} + 1_{E_{U,2} \cap E_{V,2}}) + O(n^{-4}(\log n)^{1-3\delta}), \end{aligned} \tag{3.18}$$

where $E_{U,1} = \{d(U_1, U_2) \in (c_n, 2l_n^+)\}$, $E_{V,1} = \{d(V_1, V_2) \in (c_n, 2l_n^+)\}$, $E_{U,2} = \{d(U_1, U_2) \leq 2l_n^+\}$ and $E_{V,2} = \{d(V_1, V_2) > 2l_n^+\}$. On $E_{U,2} \cap E_{V,2}$, $B(V_1, l_n^+)$ and $B(V_2, l_n^+)$ are disjoint. By the definition of $Y_{1,p}$'s, $Y_{2,p}$'s, $\Psi_{1,1}$ and $\Psi_{2,2}$, we have that

$$\begin{aligned} & P(A_{2,2} \cap A_{1,1}) = E^X(P^Y(A_{2,2}) P^Y(A_{1,1})) \\ \leq \ & \left(4\lambda^2 (\log n) \log_2 n\right) \cdot \max_{m_1, m_2 \in \Omega_n} E^X \left\{\prod_{k=1}^{2} P^Y\left(\sum_{p=1}^{m_k} F(X_{k,p}, Y_{1,p}) \geq z_n, G_{(k,1,m_k)}\right)\right\}, \end{aligned}$$

where Ω_n is as in (2.1). Consequently, by Lemma 3.4,

$$n^4 \cdot E^{U,V}\left\{P(A_{2,2} \cap A_{1,1}) 1_{E_{U,2} \cap E_{V,2}}\right\} = o\left((\log n)^3 n^{-\gamma}\right) \tag{3.19}$$

for some $\gamma > 0$. Thus, it remains to show that

$$E^{U,V}\left(P(A_{2,2} \cap A_{1,1}) 1_{E_{U,1} \cap E_{V,1}}\right) \to 0. \tag{3.20}$$

Define

$$\tau_{i,j}(r) = \sum_{p=1}^{n} 1_{B(U_i, r)}(U_p) \wedge \sum_{p=1}^{n} 1_{B(V_j, r)}(V_p). \tag{3.21}$$

Then, $\Psi_{i,j} = \{\tau_{i,j}(r);\ l_n^- \leq r \leq l_n^+,\ \tau_{i,j}(r) \in \Omega_n\}$. Now

$$\begin{aligned} P^{X,Y}(A_{1,1} \cap A_{2,2}) \ \leq \ & \lambda^2 (\log n)(\log_2 n) \cdot \\ & \max P^{X,Y}(\underbrace{\sum_{p=1}^{\tau_{1,1}(r)} F(X_{1,p}, Y_{1,p}) \geq z_n}_{A_1(r)}, \underbrace{\sum_{p=1}^{\tau_{2,2}(s)} F(X_{2,p}, Y_{2,p}) \geq z_n}_{A_2(s)}) \end{aligned} \tag{3.22}$$

where the maximum is taken over all r, s such that $r, s \in (l_n^-, l_n^+)$ and $\tau_{1,1}(r), \tau_{2,2}(s) \in \Omega_n$. For any such pair r, s, without loss of generality, assume $r > s$. Then it is easy to check that on $E_{U,1} \cap E_{V,1}$

$$\text{the volume of } B(V_1, r) \backslash B(V_2, s) \geq 4s^2 d(V_1, V_2) \geq \kappa (\log n)^{2/3-\delta}/n \tag{3.23}$$

for some constant $\kappa > 0$. Recalling the definition of $\tau_{1,1}(r)$, by symmetry, we assume without loss of generality that

$$\sum_{i=1}^{n} 1_{B(V_1,r)}(V_i) \leq \sum_{i=1}^{n} 1_{B(U_1,r)}(U_i). \tag{3.24}$$

By Bernstein's inequality

$$P^{V_3,\cdots,V_n}(\underbrace{\sum_{i=1}^{n} 1_{B(V_1,r)\backslash B(V_2,s)}(V_i) \leq \frac{\kappa}{2}(\log n)^{2/3-\delta}}_{H_n}) \leq \exp(-C(\log n)^{2/3-\delta}).$$

Define

$$\Gamma_1 = \{1 \leq p \leq n; U_{1,p} \in B(U_1, r) \backslash B(U_2, s) \text{ and } V_{1,p} \in B(V_1, r) \backslash B(V_2, s)\},$$
$$\Gamma_2 = \{1 \leq p \leq n; U_{1,p} \in B(U_1, r) \cap B(U_2, s) \text{ and } V_{1,p} \in B(V_1, r) \backslash B(V_2, s)\}.$$

Recall (3.24). On H_n^c, there are only two possibilities: either $|\Gamma_1| \geq (\kappa/4)(\log n)^{2/3-\delta}$ or $|\Gamma_2| \geq (\kappa/4)(\log n)^{2/3-\delta}$. Now we deal with these two cases separately.

Case 1. $|\Gamma_1| \geq (\kappa/4)(\log n)^{2/3-\delta}$ on H_n^c.

In this case, the cardinality of the symmetric difference between $\{(X_{1,p}, Y_{2,p}); 1 \leq p \leq \tau_{1,1}(r)\}$ and $\{(X_{2,p}, Y_{2,p}); 1 \leq p \leq \tau_{2,2}(s)\}$ is at least $(\kappa/4)(\log n)^{2/3-\delta}$. It follows by Lemma A.6 that on $E_{U,1} \cap E_{V,1} \cap H_n^c$,

$$\max P^{X,Y}(A_1(r) \cap A_2(s)) = O(n^{-2} e^{-C(\log n)^{2/3-\delta}}),$$

where the maximum is taken over all r and s as in (3.22). On the other hand, it is trivial to see that $E^{U,V}\left[P(A_{2,2} \cap A_{1,1}) 1_{E_{U,1} \cap E_{V,1} \cap H_n}\right] = O\left(n^{-4}(\log n)^2 e^{-C(\log n)^{2/3-\delta}}\right)$. Therefore, by (3.22),

$$\begin{aligned} & E^{U,V}\left[P(A_{2,2} \cap A_{1,1}) 1(E_{U,1} \cap E_{V,1} \cap \{|\Gamma_1| \geq (\kappa/4)(\log n)^{2/3-\delta}\})\right] \\ = \; & O\left(n^{-4}(\log n)^4 e^{-C(\log n)^{2/3-\delta}}\right) \end{aligned} \tag{3.25}$$

for some $C > 0$.

Case 2. $|\Gamma_2| \geq (\kappa/4)(\log n)^{2/3-\delta}$ on H_n^c.

By (1.3) and Chebyshev's inequality, $P^{X,Y}(\sum_{p\in D} F(X_{1,p},Y_{1,p}) \geq z_n - w|\Gamma_2|) = O(n^{-2}e^{\theta w|\Gamma_2|})$ for any subset $D \subset \{1,2,\cdots,n\}$. It follows that on $E_{U,1}\cap E_{V,1}\cap H_n^c$,

$$
\begin{aligned}
& P^{X,Y}(A_1(r)\cap A_2(s)) \\
\leq\ & P^{X,Y}\left(\sum_{p\in\Gamma_2} F(X_{1,p},Y_{1,p}) \geq w|\Gamma_2|,\ \sum_{p=1}^{\tau_{2,2}(s)} F(X_{2,p},Y_{2,p}) \geq z_n\right) 1(\tau_{2,2}(s)\in\Omega_n) \\
& + O\left(n^{-2}e^{Cw(\log n)^{2/3-\delta}}\right)
\end{aligned}
$$

for any fixed $w \in (\mu_F, 0)$, where $\mu_F = EF(X,Y)$. Define

$$
\begin{aligned}
A_U &= \{1\leq i\leq n;\ U_i = U_{1,p} = U_{2,q} \text{ for some } p\in\Gamma_2 \text{ and some } 1\leq q\leq \tau_{2,2}(s)\}, \\
B_U &= \{1\leq i\leq n;\ U_i = U_{1,p} \neq U_{2,q} \text{ for some } p\in\Gamma_2 \text{ and for all } 1\leq q\leq \tau_{2,2}(s)\}.
\end{aligned}
$$

Consequently, $A_U\cap B_U = \emptyset$ and $|A_U\cup B_U| = |\Gamma_2|$. Then, by Chebyshev's inequality,

$$
\begin{aligned}
T_n &:= P^{X,Y}\left(\sum_{p\in\Gamma_2} F(X_{1,p},Y_{1,p}) \geq w|\Gamma_2|,\ \sum_{q=1}^{\tau_{2,2}(s)} F(X_{2,q},Y_{2,q}) \geq z_n\right) 1(\tau_{2,2}(s)\in\Omega_n) \\
&\leq n^{-2}e^{-wt|\Gamma_2|}E^{X,Y}\{\prod_{p\in A_U\cup B_U} e^{tF(X_{1,p},Y_{1,p})}\cdot \prod_{\substack{1\leq q\leq\tau_{2,2}(s)\\ \tau_{2,2}(s)\in\Omega_n}} e^{\theta F(X_{2,q},Y_{2,q})}\} \\
&= n^{-2}e^{-wt|\Gamma_2|}N(t)^{|B_U|}M(t)^{|A_U|}
\end{aligned}
$$

for all $t>0$, where $N(t) := Ee^{tF(X,Y)}$ and $M(t) = Ee^{tF(X,Y_1)+\theta F(X,Y_2)}$ as in Lemma 3.2. If $|B_U| \geq |\Gamma_2|/2$, then by Lemma 3.2, $T_n \leq n^{-2}(e^{-wt}\sqrt{N(t)}\ )^{|\Gamma_2|}$ for all $t\in(0,t_0)$. By Lemma 3.2 again and choosing $w=\delta/2$, $T_n \leq n^{-2}\exp(-|\Gamma_2|\gamma_1)$ for some $\gamma_1>0$. Note that $N(t)<1$ for all $t\in(0,t_0)$. If $|A_U|\geq|\Gamma_2|/2$, by repeating the same arguments as above, we then obtain the same bound for T_n with another constant γ_2. Consequently, $T_n = O(n^{-2}\exp(-C(\log n)^{2/3-\delta}))$. Therefore, as in (3.25), we obtain

$$
\begin{aligned}
& E^{U,V}\left[P^{X,Y}(A_{2,2}\cap A_{1,1})1(E_U\cap E_{V,1}\cap\{|\Gamma_2|\geq(\kappa/4)(\log n)^{2/3-\delta}\})\right] \\
=\ & O(n^{-4}(\log n)^4 e^{-C(\log n)^{2/3-\delta}}),
\end{aligned}
\tag{3.26}
$$

which together with (3.25) implies that

$$
E^{U,V}\left(P^{X,Y}(A_{2,2}\cap A_{1,1})1(E_{U,1}\cap E_{V,1})\right) = O(n^{-4}(\log n)^4 e^{-C(\log n)^{2/3-\delta}}).
$$

Thus, (3.20) is justified. The proof is completed. ■

The Proof of Lemma 2.6. Set $c_n = (\log n)^{-h} n^{-1/3}$, $h \in (0, 1/6)$. Since $P(A_{1,1} \cap A_{1,2}) \leq n^{-2}$ and $P(d(V_1, V_2) \leq c_n) \leq 8/n(\log n)^{3h}$,

$$\begin{aligned} & E^{U,V} P(A_{1,1} \cap A_{1,2}) \\ \leq \;\; & E^{U,V} \left[P(A_{1,1} \cap A_{1,2}) \{ I(d(V_1, V_2) \geq 2l_n^+) + I(c_n < d(V_1, V_2) < 2l_n^+) \} \right] + 8n^{-3} (\log n)^{-3h}. \end{aligned}$$

As in the case of the estimate of $E^{U,V} P(A_{2,2} \cap A_{1,1}) I_{E_{U,2} \cap E_{V,2}}$ in (3.19), we obtain

$$E^{U,V} P(A_{1,1} \cap A_{1,2}) I(d(V_1, V_2) \geq 2l_n^+) = o((\log n)^3 n^{-\gamma})$$

for some $\gamma > 0$. So we only need to estimate $E^{U,V}[P(A_{1,1} \cap A_{1,2}) I(c_n < d(V_1, V_2) < 2l_n^+)]$. Recall the definition of $\tau_{i,j}(r)$ in (3.21). If $d(V_1, V_2) \in (c_n, 2l_n^+)$, then we have that

$$\begin{aligned} & E^{U,V} P(A_{1,1} \cap A_{1,2}) \\ \leq \;\; & \lambda^2 (\log n)(\log_2 n) E^{U,V} \left[\max P^{X,Y} \Big(\sum_{p=1}^{\tau_{1,1}(r)} F(X_{1,p}, Y_{1,p}) \geq z_n, \; \sum_{p=1}^{\tau_{1,2}(s)} F(X_{1,p}, Y_{2,p}) \geq z_n \Big) \right], \end{aligned}$$

where the maximum is taken over all r, s such that $r, s \in (l_n^-, l_n^+)$ and $\tau_{1,1}(r), \tau_{1,2}(s) \in \Omega_n$. Assume, without loss of generality, $r > s$. As in (3.23), the volume of $B(V_1, r) \backslash B(V_2, s) \geq C(\log n)^{2/3-h}/n$. It is easy to check by Bernstein's inequality that with probability at least $1 - 2\exp(-\sqrt{\log n})$,

$$\begin{aligned} & \#\{3 \leq p \leq n; \; V_p \in B(V_1, r) \backslash B(V_2, s)\} \geq C(\log n)^{2/3-h} \qquad \text{and} \\ & \#\{3 \leq p \leq n; \; U_p \in B(U_1, r) \backslash B(U_1, s)\} \leq C\sqrt{(\log n) \log_2 n}, \end{aligned}$$

which implies that there exists a set $\Gamma_3 \subset \{1, 2, \cdots, n\}$ such that $|\Gamma_3| \geq C(\log n)^{2/3-h}$, $U_{1,p} \in B(U_1, s)$ and $V_{1,p} \in B(V_1, r) \backslash B(V_2, s)$ for all $p \in \Gamma_3$. Then by using the same argument as was used for obtaining (3.25) and (3.26) , we have that

$$P\left(\sum_{p=1}^{\tau_{1,1}(r)} F(X_{1,p}, Y_{1,p}) \geq z_n, \; \sum_{p=1}^{\tau_{1,2}(s)} F(X_{1,p}, Y_{2,p}) \geq z_n \right) = o(n^{-2} \exp\{-C(\log n)^{2/3-h}\}).$$

Combining all the above arguments, we have

$$E^{U,V} P(A_{1,1} \cap A_{1,2}) = O\left(n^{-3} e^{-C\sqrt{\log n}} \right).$$

The proof is complete. ■

4 Appendix

The following inequality provides us with bounds for tails of sums of independent and bounded random variables, see Exercise 14 in [7] or page 193 in [17].

Lemma A.1 *(Bernstein's Inequality). Let* $\{\epsilon_i;\ 1 \le i \le n\}$ *be a sequence of independent random variables with* $E\epsilon_i = 0,\ E\epsilon_i^2 = \sigma_i^2$ *and* $|\epsilon_i| \le 1$. *Let* $S_n = \sum_{i=1}^n \epsilon_i,\ s_n^2 = \sum_{i=1}^n \sigma_i^2$. *Then*

$$P(S_n > x) \le \exp\{-x^2/2(s_n^2 + x)\}, \qquad x > 0.$$

In the following lemmas, we assume that $\{\xi, \xi_i, i \ge 1\}$ is a sequence of i.i.d. random variables. Let $S_n = \sum_{i=1}^n \xi_i$. The next inequality is called Chernoff's bound (see e.g. p.p. 31 in [10]).

Lemma A.2 *(Chernoff's bound). Let* $\Lambda_\xi^*(x) = \sup_{t\in\mathbb{R}}\{tx - \log E\exp(t\xi)\}$. *Then,*

$$P(S_n/n \ge x) \le \exp(-n\Lambda_\xi^*(x)), \quad n \ge 1,\ x \ge E\xi.$$

The lemma below is a refinement of a large deviation result from corollary 2.1 in [13].

Lemma A.3 *Suppose* ξ *is non-lattice and* $Ee^{t\xi} < \infty$ *for all* $t \in \mathbb{R}$. *Then*

$$\sup_{a\le\eta\le b}\ \sup_{|x|\le\delta\sqrt{n\log n}} |C_n(x,\eta)P(S_n \ge n\Lambda'(\eta) + x) - 1| \to 0 \quad \text{as } n \to \infty,$$

for any positive constants a, b *and* δ, *where*

$$C_n(x,\eta) = \eta\sqrt{2\pi n\Lambda''(\eta)}\ \exp\{n\Lambda^*(\Lambda'(\eta)) + \eta x + (x^2/2\Lambda''(\eta)n)\}.$$

Next, we need a condition on ξ as follows:

$$\xi \text{ is non-lattice},\ E(\xi) < 0,\ P(\xi > 0) > 0 \text{ and } E\exp(t\xi) < \infty \ \ \forall t \in \mathbb{R}. \tag{4.1}$$

Under this condition, there is an unique constant $\theta > 0$ so that $E\exp(\theta\xi) = 1$. Set $\Lambda(t) = \log(E\exp(t\xi))$ for $t \in \mathbb{R}$. The following two facts are lemmas 2.5 and 2.6 from [13], respectively.

Lemma A.4 *Assume* ξ *satisfies condition (4.1), then there exist constants* $r > 1$ *and* $t_0 > \theta$ *such that*

$$\sum_{k\ge rz} P(S_k \ge z) + \sum_{k\le r^{-1}z} P(S_k \ge z) = o(e^{-t_0 z}) \quad \text{as } z \to \infty.$$

Lemma A.5 *Suppose condition (4.1) holds. For any two positive functions $a(z)$ and $b(z)$ such that $(a(z)+b(z))/z \to 0$, and any two positive numbers s and r satisfying $s < \Lambda'(\theta) < r$, we have that*

$$(e^{\theta z}/\sqrt{z}) \sum_{k \in \Gamma_z} P(S_k \geq z) = O\left(e^{-c(z)^2/z}\right) \text{ as } z \to \infty,$$

where $\Gamma_z = \{k \in \mathbb{N};\ sz \leq k \leq \Lambda'(\theta)z - b(z) \text{ or } \Lambda'(\theta)z + a(z) \leq k \leq rz\}$ and $c(z) = a(z) \wedge b(z),\ z > 0$.

Lemma A.6 *Suppose A, B, and C are disjoint sets of indices and $\{X, X_\alpha; \alpha \in A \cup B \cup C\}$ are i.i.d. random variables with X satisfying condition (4.1) and $\mu := EX$. For any subset $D \subset A \cup B \cup C$, we use the notation $S_D := \sum_{\alpha \in D} X_\alpha$. Then,*

$$P(S_{A \cup B} \geq z, S_{B \cup C} \geq z) \leq 2e^{-\theta z - m_1 \zeta} \leq 2e^{-\theta z - m_2 \zeta},$$

where $\zeta = \sup_{\mu < x < 0}\{\Lambda^*(x) \wedge \theta|x|\} > 0,\ m_1 = |A| \vee |C|$ and $m_2 = |A \cup C|/2$.

Acknowledgment. This is a part of the author's dissertation in the Department of Statistics, Stanford University. The author would like to thank his advisor Amir Dembo for his stimulating supervision. The author also thanks Samuel Karlin for his inspirational discussion on constructing protein models, the proof and the application of Theorem 1. Finally, the author thanks an anonymous referee for his/her constructive suggestions.

References

[1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local aligment search tool. J. Mol. Biol. 215 403-410.

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

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

[4] Arratia, R. and Goldstein, L. and Gordon, L. (1989). Two Moments Suffice for Poisson Approximation: The Chen-Stein Method. Ann. Probab. 17 9-25.

[5] Arratia, R. and Moris, P. and Waterman, M. S. (1988). Stochastic scrabble:large deviations for sequences with scores. J. Appl. Probab. 25 106-119.

[6] Arratia, R. and Waterman, M. S. (1985). Critical phenomena in sequence matching. Ann. Probab. 13 1236-1249.

[7] Chow, Y.S. and Teicher, H. (1988). Probability Theory, Independence, Interchangeability, Martingales. Springer-Verlag, New York, Second edition.

[8] Dembo, A. and Karlin, S. and Zeitouni, O. (1994). Critical Phenomena for sequence matching with scoring. Ann. Probab. 22 1993-2021.

[9] Dembo, A. and Karlin, S. and Zeitouni, O. (1994). Limit distribution of maximal non-aligned two-sequences segmental score. Ann. Probab. 22 2022-2039.

[10] Dembo, A. and Zeitouni, O. (1998). Large Deviations Techniques and Applications. Springer, Second edition.

[11] Feller, W. (1971). An Introduction to Probability Theory and Its Applications. Wiley, New York, Vol. 2, Second edition.

[12] Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc. Nat. Acad. Sci. U.S.A. 89 10915-10919.

[13] Jiang, T. Maxima of partial sums indexed by geometrical structures. Preprint.

[14] Karlin, S. (1994). Statistical studies of bimolecular sequences: Score-based methods. Phi. Trans. R. Soc. Lond. B. 344 391-402.

[15] Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. natn. Acad. Sci. U.S.A. 87 2264-2268.

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

[17] Pollard, D. (1984). Convergence of Stochastic Processes. Springer-Verlag, New York.

[18] Spitzer, F. (1964). Principles of Random Walk. Van Nostrand, Princeton.

[19] States, D. J., Gish, W. and Altschul, S. F. (1991). Improved sensitivity of nucleic acid database searches using applications-specific scoring matrices. Methods 3 66-70.

[20] Stormo, G. D. and Hartzell, G. W. (1989) Identifying protein-binding sites from unaligned DNA fragments. Proc. Nat. Acad. Sci. U.S.A. 86 1183-1187.

email: tjiang@stat.umn.edu
Mailing Address: School of Statistics, 313 Ford Hall, 224 Church Street S.E., Minneapolis, MN55455