The Annals of Applied Probability
1995, Vol. 5, No. 4, 1236–1240

A LIMIT THEOREM FOR MATCHING RANDOM SEQUENCES ALLOWING DELETIONS[1]

BY YU ZHANG

University of Colorado

We consider a sequence matching problem involving the optimal alignment score for contiguous sequences, rewarding matches by one unit and penalizing for deletions and mismatches by parameters δ and μ, respectively. Let M_n be the optimal score over all possible choices of two contiguous regions. Arratia and Waterman conjectured that, when the score constant $a(\mu, \delta) < 0$,

$$P\left(\frac{M_n}{\log n} \to 2b\right) = 1$$

for some constant b. Here we prove the conjecture affirmatively.

1. Introduction. Let $A_1, A_2, \dots$ and $B_1, B_2, \dots$ be two independent sequences of i.i.d. random variables such that A_i and B_i have the same distribution on a finite number set $\{0, 1, \dots, \tau\}$. Let $I = (A_{g+1}, \dots, A_{g+i})$ and $J = (B_{h+1}, \dots, B_{h+j})$ with $1 \le g+1 \le g+i \le n$ and $1 \le h+1 \le h+j \le n$. The *alignment score* $S(I, J)$ is defined to be

$$S(I, J) = \max\left\{-\delta(i - l + j - l) + \sum_{k=1}^{l} s\left(A_{a(k)}, B_{b(k)}\right)\right\}, \tag{1.1}$$

where the maximum is taken over all alignments, given by increasing sequences

$$g = a(0) \le a(1) < a(2) < \cdots < a(l) \le a(l+1) = g + i + 1$$

and

$$h = b(0) \le b(1) < b(2) < \cdots < b(l) \le b(l+1) = h + j + 1.$$

In particular, if we restrict $a(0) = g$, $a(1) = g + 1, \dots,$ $a(l) = g + l$ and $b(0) = h$, $b(1) = h + 1, \dots,$ $b(l) = h + l$, the corresponding score is called the *nonalignment score* or the *score without deletions*. The score function $s(x, y)$ for aligned pairs is 1 if $x = y$ and $-\mu$ if $x \ne y$. In words, each match is rewarded by 1, each mismatch is penalized by μ and each deletion by δ. Let $S_n = S(A_1, \dots, A_n, B_1, \dots, B_n)$. That is, $I = A_1, \dots, A_n$ and $J = B_1, \dots, B_n$. By a standard subadditive argument (see [2]), it is easy to see that, for

Received December 1994; revised August 1995.
[1]Research supported in part by NSF Grant DMS-94-00467 and Grigsby.
AMS 1991 *subject classifications*. 62E20, 62P10.
Key words and phrases. Sequence matching.

$\mu, \delta \geq 0$, there exists a nonrandom constant $a(\mu, \delta)$ such that

$$\lim_{n \to \infty} \frac{S_n}{n} = a(\mu, \delta) \quad \text{a.s. and in } L_1. \tag{1.2}$$

Here $a(\mu, \delta)$ is called the *score constant*. On the other hand, denote the *large deviation rate* by

$$r(q) = \lim_{n \to \infty} \frac{-\log P(S_n \geq qn)}{n} = \inf\left\{ \frac{-\log P(S_n \geq qn)}{n} \right\}, \tag{1.3}$$

where log means the natural logarithm. The limit in (1.3) exists and equals the infimum also using the subadditive property. Let M_n be the optimal aligned score over all possible choices of two contiguous regions I and J for $I \subset \{A_1, \dots, A_n\}$ and $J \subset \{B_1, \dots, B_n\}$. Formally,

$$M_n = \max_{I, J} S(I, J). \tag{1.4}$$

Let $b = \max_{q \geq 0} q/r(q)$. It can be proved by applying the Borel–Cantelli lemma along a suitable skeleton in Lemma 2 in [2] that, if $a(\mu, \delta) < 0$, then

$$b \leq \liminf \frac{M_n}{\log n} \leq \limsup \frac{M_n}{\log n} \leq 2b \quad \text{a.s.} \tag{1.5}$$

On the other hand, it was also proved in [2] that, if $a(\mu, \delta) > 0$,

$$\lim_{n \to \infty} \frac{M_n}{n} = a(\mu, \delta) \quad \text{a.s. and in } L_1.$$

The phenomenon of the two different behaviors of M_n is called a *phase transition*. When $a(\mu, \delta) < 0$, one of the most important problems is to decide whether $M_n/(\log n)$ converges. In fact, Arratia and Waterman conjectured that $M_n/(\log n)$ converges to $2b$ in probability. Note that it was verified in [1] that the conjecture is true for the nonaligned case. Furthermore, Dembo, Karlin and Zeitouni [3] gave a more general discussion for the nonaligned case. In the following theorem we prove that the conjecture is true for any δ and μ.

THEOREM 1. *For each μ and δ, if $a(\mu, \delta) < 0$, then*

$$\lim_{n \to \infty} \frac{M_n}{\log n} = 2b \quad a.s. \tag{1.6}$$

REMARK 1. Here we prove that the theorem holds on a finite number set. We can also show that the theorem holds on Polish alphabets by the same proof of the theorem and Theorem 4′ in [3]. On the other hand, the theorem also holds for a more general score function $s(x, y)$.

REMARK 2. Amir Dembo pointed out that the same proof of the theorem carries over to generalized scoring and gapping, repeats in a sequence, and matching Markov chains (see the detailed definitions in [2]).

2. Proof of the theorem. The proof is based on Theorem 3 in [3]. For a positive integer m, consider two independent sequences $\{X_i\}$ and $\{Y_i\}$ with

$$X_i = (A_{im+1}, \dots, A_{(i+1)m}) \quad \text{and} \quad Y_j = (B_{jm+1}, \dots, B_{(j+1)m}).$$

Clearly, $\{X_i\}$ is i.i.d. and so is $\{Y_j\}$. Let X_i and Y_j have the probability laws π_X and π_Y on finite sets Γ_X and Γ_Y, respectively, where

$$\Gamma_X = \Gamma_Y = \{0, 1, \dots, \tau\}^m = \{(x_1, \dots, x_m): x_i \in \{0, 1, \dots, \tau\} \text{ for } i = 1, \dots, m\}.$$

Clearly,

$$\pi_X(X_i = \mathscr{X}) = P((A_{im+1}, \dots, A_{(i+1)m}) = \mathscr{X}),$$
$$\pi_Y(Y_j = \mathscr{Y}) = P((B_{jm+1}, \dots, B_{(j+1)m}) = \mathscr{Y})$$

for $\mathscr{X} \in \Gamma_X$ and $\mathscr{Y} \in \Gamma_Y$. A general score F: $\Gamma_X \times \Gamma_Y \to \mathscr{R}$ is assigned to each pair (X_i, Y_j) and the maximal nonaligned segment score is

$$\mathscr{M}_n = \max_{0 \le i, j \le n-k;\, k \ge 0} \left\{ \sum_{l=1}^{k} F(X_{i+l}, Y_{j+l}) \right\}. \tag{2.1}$$

It was proved in [3] that, if

$$E_{\pi_X \times \pi_Y} F < 0 \quad \text{and} \quad \pi_X \times \pi_Y(F > 0) > 0, \tag{2.2}$$

then

$$\frac{\mathscr{M}_n}{\log n} \to \gamma(\pi_X, \pi_Y). \tag{2.3}$$

Furthermore, if (2.2) holds, there exists a unique positive value θ such that

$$E_{\pi_X \times \pi_y}[e^{\theta F}] = 1. \tag{2.4}$$

Let α denote the conjugate measure associated with θ, that is,

$$\frac{d\alpha}{d(\pi_X \times \pi_Y)} = e^{\theta F},$$

and let α_X and α_Y denote the marginals of α on Γ_X and Γ_Y, respectively. Dembo, Karlin and Zeitouni [3] also showed that, if

$$H(\alpha | \pi_X \times \pi_Y) \ge 2 \max\{H(\alpha_X | \pi_X), H(\alpha_Y | \pi_Y)\}, \tag{2.5}$$

then

$$\frac{\mathscr{M}_n}{\log n} \to \frac{2}{\theta}, \tag{2.6}$$

where the relative entropy $H(\nu|\pi)$ is defined to be

$$H(\nu | \pi) = \sum_{i=1}^{K} \nu(b_i) \log \frac{\nu(b_i)}{\pi(b_i)}$$

for $\{b_1, \dots, b_K\} = \Gamma_X \times \Gamma_Y$. Now we apply (2.6) to our purpose. Note that the score defined in (2.1) is nonaligned so that we have to choose some special F

and use F to approximate the aligned score. Set

$$F(X_i, Y_j) = S(X_i, Y_j) \quad [\text{see (1.1) for the definition of } S].$$

If $a(\mu, \delta) < 0$, by (1.2) and our definition, with a large m,

$$\text{(2.7)} \qquad \begin{gathered} E_{\pi_X \times \pi_Y} F = ES_m < \frac{a(\mu, \delta) m}{2} < 0, \\ \pi_X \times \pi_Y (F > 0) \geq P(A_1 = B_1, \dots, A_m = B_m) > 0. \end{gathered}$$

It follows from (2.7) and (2.3) that

$$\text{(2.8)} \qquad \frac{\mathscr{M}_n}{\log n} \to \gamma(\pi_X, \pi_Y)$$

for our special definition of $\{X\}$ and $\{Y\}$. It also follows from (2.1) that

$$\text{(2.9)} \qquad \mathscr{M}_n \leq M_{nm},$$

where mn is the product of m and n. On the other hand, by a standard information inequality (see (13) in [3])

$$\text{(2.10)} \qquad H(\alpha | \pi_X \times \pi_Y) \geq H(\alpha_X | \pi_X) + H(\alpha_Y | \pi_Y).$$

Note that $\Gamma_X = \Gamma_Y = \{0, 1, \dots, \tau\}^m$, $\pi_X = \pi_Y$ and $F(\mathscr{X}, \mathscr{Y}) = F(\mathscr{Y}, \mathscr{X}) = S(\mathscr{X}, \mathscr{Y}) = S(\mathscr{Y}, \mathscr{X})$ so that $\alpha_X = \alpha_Y$. By (2.10), (2.5) holds. It follows from (2.10) and (2.6) that, for the m satisfying (2.7),

$$\text{(2.11)} \qquad \frac{\mathscr{M}_n}{\log n} \to \frac{2}{\theta} \quad \text{as } n \to \infty,$$

where θ, which may depend on m, is a positive constant such that $E_{\pi_X \times \pi_Y}[e^{\theta F}] = 1$. For a given $\varepsilon > 0$, it follows from Theorem 2 in [2] that we can pick $q' > 0$ such that

$$r(q') > 0 \quad \text{and} \quad b < \frac{q'}{r(q')} + \varepsilon.$$

Furthermore, by (1.3), we can also pick m large such that

$$\text{(2.12)} \qquad P(F \geq q'm) = P(S_m \geq q'm) \geq \exp[(-r(q') - \varepsilon)m].$$

Note that $\theta > 0$ so that, by (2.12),

$$\text{(2.13)} \qquad \begin{aligned} 1 = E_{\pi_X \times \pi_Y} \exp(\theta F) &\geq \exp(\theta q' m) P(F \geq q'm) \\ &\geq \exp[m(\theta q' - r(q') - \varepsilon)]. \end{aligned}$$

Note also that $m \geq 1$ and $q' > 0$ so that

$$\text{(2.14)} \qquad 0 < \theta \leq \frac{r(q') + \varepsilon}{q'}.$$

By (2.9), (2.11) and (2.14), we choose a large n such that, for the m satisfying (2.7) and (2.13),

$$\text{(2.15)} \qquad \frac{M_{nm}}{\log n} \geq \frac{\mathscr{M}_n}{\log n} \geq \frac{2}{\theta} - \varepsilon \geq 2(b - \varepsilon) \frac{r(q')}{r(q') + \varepsilon} - \varepsilon.$$

Note that $r(q') \geq r(0)$ and $r(0)$ is a positive constant which does not depend on n, m and ε so that, by (2.15),

$$\frac{M_{nm}}{\log n} \geq 2(b - \varepsilon)\left(1 - \frac{\varepsilon}{r(0)}\right) - \varepsilon. \tag{2.16}$$

For any $t = nm + k$ with $k < m$, note that

$$M_{nm} \leq M_t \leq M_{n(m+1)}$$

so that, for the m satisfying (2.7) and (2.13) by (2.16),

$$\begin{aligned} \liminf_t \frac{M_t}{\log t} &\geq \liminf_n \frac{M_{nm}}{\log[n(m+1)]} \\ &= \liminf_n \frac{M_{nm}}{\log n} \geq 2(b - \varepsilon)\left(1 - \frac{\varepsilon}{r(0)}\right) - \varepsilon \quad \text{a.s.} \end{aligned} \tag{2.17}$$

The theorem holds by (1.5) and (2.17). □

Acknowledgment. The author would like to thank Amir Dembo for pointing out Remark 2.

REFERENCES

[1] ARRATIA, R. and WATERMAN, M. S. (1985). An Erdös–Rényi law with shifts. *Adv. in Math.* **55** 13–23.

[2] ARRATIA, R. and WATERMAN, M. S. (1994). A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.* **4** 200–225.

[3] DEMBO, A., KARLIN, S. and ZEITOUNI, O. (1994). Critical phenomena for sequence matching with scoring. *Ann. Probab.* **22** 1993–2021.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF COLORADO
COLORADO SPRINGS, COLORADO 80933