Matching problem

November 26, 2000

Abstract

1 Introduction

Let \( \{X,X_1,X_2,\cdots\} \) be a sequence of i.i.d. random variables. \( X \) may or may not take infinite values. Let \( r \) be a fixed positive integer. We are interested in the following pattern: \( a_1a_2\cdots a_r \), where each \( a_i \) is a certain realization of \( X \) with probability \( p_i, 1 \leq i \leq r \). For a number \( x \), as in convention, \( \lfloor x \rfloor \) denotes its integer part. For any positive integer \( i \), the quantity \( t(i) := i - r[i/r] \) is the remainder of \( i \) after divided by \( r \). Denote

\[
b_i = a_{t(i)} \text{ and } p_i = p_{t(i)}, i = 1,2,\cdots.
\]

Then the sequence of periodic numbers \( \{a_1a_2\cdots a_1a_2\cdots a_r a_1a_2\cdots a_r \cdots\} \) and \( \{p_1p_2\cdots p rp_1p_2\cdots p_r \cdots\} \) can be written as \( \{b_1b_2b_3\cdots\} \) and \( \{p_1p_2p_3\cdots\} \), respectively. We say \( X_{i+1}\cdots X_{i+l} \) matches \( b_{j+1}\cdots b_{j+l} \) if \( X_{i+k} = b_{j+k} \) for \( k = 0,1,2,\cdots, l \). The purpose of this note is to measure the length of longest such matches between the sequence \( \{X_1,X_2,\cdots,X_n\} \) and \( \{b_1,b_2,\cdots,b_n\} \). Precisely, define a function \( F(x,y) \) such that \( F(x,y) = 1 \) if \( x = y \), and \( -\infty \) otherwise, then the length of longest matches is the following statistic

\[
W_n = \max_{(i,j,\Delta) \in A_n} \sum_{k=0}^{\Delta} F(X_{i+k},b_{j+k}),
\]

where \( A_n = \{(i,j,\Delta); 1 \leq \Delta \leq n, 1 \leq i,j \leq n - \Delta\} \). Here is our first result.

**Theorem 1** Let \( p = p_1p_2\cdots p_r \). Then

\[
\lim_{n \to \infty} \frac{W_n}{\log n} \to \frac{1}{\theta} \text{ a.s.,}
\]

where \( \theta = (\log(1/p))/r \).

\(^1\)School of Statistics, University of Minnesota.

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

AMS (1991) subject classifications: 60F10, 60B10
For any \( x \in \mathbb{R} \), let \( d_n(x) = 1 + [x + (\log n)/\theta] \) and \( g_n(x) = \sum_{i=1}^{r} \left( \prod_{l=0}^{(d_n(x))} p_{l+i} \right) \), and
\( h_n(x) = [d_n(x)/r] \log(1/p) \log(ng_n(x)) \). The next is the limiting distribution of \( W_n \).

**Theorem 2** Suppose \( (b_1,b_2,\cdots,b_r) \neq (b_{i+1},b_{i+2},\cdots,b_{i+r}) \) for any \( i \in \{1,2,\cdots,r-1\} \). Then,
\[
 e^{K e^{h_n(x)}} P \left( W_n \leq \frac{\log n}{\theta} + x \right) \to 1, \ x \in \mathbb{R},
\]
as \( n \to \infty \), where \( \theta \) is as in Theorem 1 and
\[
 K^{-1} = \sum_{j=2}^{\infty} \left( \sum_{i=1}^{j} p_{i} p_{i+1} \cdots p_{i+j-2} (1-p_{i+j-1}) \right) < \infty. \tag{1.1}
\]

**2 Proofs**

**The Proof of Theorem 1.** We first show that
\[
 \limsup_{n \to \infty} \frac{W_n}{\log n} \leq \frac{r}{\log(1/p)} \ a.s. \tag{2.2}
\]
For any \( \epsilon > 0 \), denote by \( m_n \) the least number which is greater than or equal to \( (1+\epsilon)(r/\log(1/p)) \log n \). If \( W_n \geq (1+\epsilon)(r/\log(1/p)) \log n \), then there exists \( 1 \leq i \leq m_n[r/m_n] \) such that \( X_i X_{i+1} \cdots X_{i+m_n-1} \) completely matches a part of \( \{b_1 b_2 \cdots b_n\} \). Therefore,
\[
 P(W_n \geq m_n) \leq \sum_{i=1}^{m_n[r/m_n]} \sum_{s=1}^{r} P(X_i = b_s, X_{i+1} = b_{s+1}, X_{i+m_n-1} = b_{s+m_n-1} - b_{s+m_n-1})
\]
\[
 \leq r m_n \left( \frac{n}{m_n} \right) \cdot \max_{1 \leq s \leq r} P(X_1 = b_s, X_2 = b_{s+1}, X_{m_n-1} = b_{s+m_n-1}).
\]
Since there exists at least \([m_n/r]\)’s many \( b_i \) in \( b_s b_{s+1} \cdots b_{s+m_n-1} \) for each \( i = 1,2,\cdots,r \),
\[
 P(W_n \geq m_n) \leq r n p^{[m_n/r]} \leq (r/p)n^{-\epsilon}.
\]
Let \( l_n = [n^{-2/\epsilon}] \), then \( P(W_n \geq m_n) = O(n^{-2}) \). Consequently, \( \sum_{n \geq 1} P(W_n \geq m_n) < \infty \).
By the Borel-Cantelli lemma, we have that \( \limsup_{n \to \infty} W_n / \log l_n \leq (1+\epsilon)r/\log(1/p) \) a.s.
Observe that \( W_n \) is increasing with \( n \) and \( l_{n+1}/l_n \to 1 \) as \( n \to \infty \). We have that
\[
 \limsup_{n \to \infty} W_n / \log n \leq (1+\epsilon)r/\log(1/p) \ a.s.
\]
for any \( \epsilon > 0 \). Let \( \epsilon \downarrow 0 \), we then obtain (2.2).

Now we turn to prove the lower bound:
\[
 \liminf_{n \to \infty} \frac{W_n}{\log n} \geq \frac{r}{\log(1/p)} \ a.s. \tag{2.3}
\]
For any $\epsilon \in (0, 1)$, denote $q_n = 1 + [(1 - \epsilon)(r \log n) / \log(1/p)]$. Break the sequence $X_1X_2 \cdots X_n$ into $[n/q_n]$'s many disjoint subsequences as $X_1X_2 \cdots X_{q_n}, X_{q_n+1}X_{q_n+2} \cdots X_{2q_n}, \cdots$. Let

$$A_m = \{ X_{mq_n+1} = b_1, X_{mq_n+2} = b_2, \ldots, X_{(m+1)q_n} = b_q \}$$

for $m = 0, 1, 2, \cdots, [n/q_n] - 1$. Then

$$\{W_n \leq (1 - \epsilon)r(\log n) / \log(1/p)\} \subset \bigcap_{m=0}^{[n/q_n]-1} A_m^c.$$

Note that $A_i$'s are i.i.d. events and $P(A_1) \geq p^{[q_n/r]+1}$. Then, the probability of the first event above is bounded by

$$(1 - p^{[q_n/r]+1})^{[n/q_n]-1} \leq \exp \left( -([n/q_n] - 1)p^{[q_n/r]+1} \right) \leq \exp(-Cn^\epsilon / \log n)$$

for sufficiently large $n$, where $C$ is a constant depending on $\epsilon, p$ and $r$. Consequently

$$\sum_{n \geq 1} P(W_n/\log n \leq (1 - \epsilon)r/\log(1/p)) < \infty$$

for any $\epsilon \in (0, 1)$. By the Borel-Cantelli lemma again, we obtain (2.3). The proof is completed. ■

**Lemma 2.1** Suppose $(b_1, b_2, \cdots, b_r) = (b_{i+1}, b_{i+2}, \cdots, b_{i+r})$ for some $i \in \{1, 2, \cdots, r-1\}$. Let $d$ be a divisor of $i$ and $r$. Then, $(b_1, b_2, \cdots, b_r)$ can be decomposed into the following form:

$$(b_1, b_2, \cdots, b_r) = (b_1, b_2, \cdots, b_d, b_1, b_2, \cdots, b_d, \cdots, b_1, b_2, \cdots, b_d).$$

**Proof.** Recall the definition of $b_i$'s, we have

$$(b_1, b_2, \cdots, b_r, b_{r+1}, \cdots) = (b_{i+1}, b_{i+2}, \cdots, b_{i+r}, b_{i+r+1}, \cdots).$$

It follows that $b_j = b_{ki+j}$ for $k = 1, 2, \cdots$ and $j = 1, 2, \cdots, r$. Define $l = (r/d) - 1$ and

$$A_j = \{ b_j, b_{i+j}, b_{2i+j}, \cdots, b_{li+j} \}, \ j = 1, 2, \cdots, d.$$  

First, we claim that

$$\{b_j, b_{i+j}, b_{2i+j}, \cdots \} = A_j, \ j = 1, 2, \cdots, d. \quad (2.4)$$

Actually, for any $s > l$, write $s = m(r/d) + s_1$, where $0 \leq s_1 \leq l$. Thus, $si + j = mr(i/d) + s_1i + j$. Note that $d$ is a divisor of $i$, by the periodicity of $b_i$'s we have that $b_{si+j} = b_{s_1i+j} \in A_j$.  

3
Write $s_i + j \equiv m_s \pmod{r}$ for $s = 0, 1, 2, \ldots, l$, where $0 \leq m_s \leq r - 1$. We next show that

$$m_{s_1} \neq m_{s_2}$$

(2.5)

for all $0 \leq s_1 < s_2 \leq l$. Suppose the conclusion is not true, then there exist $s_1$ and $s_2$ such that $0 \leq s_1 < s_2 \leq l$ and that $r$ is a divisor of $(s_2 - s_1)i$. Since $(i, r) = d$, then $r/d$ has to be a divisor of $s_2 - s_1$. But this is impossible because $1 \leq s_2 - s_1 \leq (r/d) - 1$.

At last, we claim that $A_j = \{b_j, b_{j+d}, b_{j+2d}, \ldots, b_{j+td}\}$. If this is true, the proof is completed. From (2.4) and (2.5), we only need to show that for any $1 \leq t \leq l$ there exists an integer $x \geq 1$ such that $xi + j \equiv td + j \pmod{r}$, or equivalently, $xi \equiv td \pmod{r}$. Since $(i, r) = d$, there exist integers $g$ and $h$ such that $ig + rh = d$. Thus, $r(th - |tg|i) = td - (r|tg| + tg)i$. Set $x = r|tg| + tg$, then $x$ is a positive integer and $r$ is a divisor of $td - xi$. The proof is completed. ■

With the above lemma, we may assume that no periodic patterns appear in $b_1b_2\cdots b_r$. This is because we can count the shortest pattern if a periodic pattern occurs. Define

$$Z_i = \max_{j \geq 1, k \geq 1} \sum_{k=0}^{\Delta} F(X_{i+k}, b_{j+k}), i = 1, 2, \ldots.$$ 

Obviously, $Z \geq 1$. For any integer $x \geq 1$, let $N(x) = x - r[x/r]$.

**Lemma 2.2** Suppose $(b_1, b_2, \cdots, b_r) \neq (b_{i+1}, b_{i+2}, \cdots, b_{i+t})$ for any $i \in \{1, 2, \cdots, r-1\}$. Then,

$$p^{-[m/r]} P(Z_1 \geq m) = \sum_{i=1}^{r} \left( \prod_{l=0}^{N(m)-1} p^{l+i} \right)$$

for any integer $m > 0$, where the above product is understood to be 1 if $N(m) = 0$.

**Proof.** It is easy to see that

$$\{Z_1 \geq m\} = \bigcup_{s=0}^{r-1} \{X_i = b_{i+s}, i = 1, 2, \ldots, m\}.$$ 

By the given assumption, the $r$ events above after the union notation are disjoint. Then $P(Z_1 \geq m)$ is the sum of probabilities of these $r$ events. Let $A_s = \{X_i = b_{i+s}, i = 1, 2, \ldots, m\}$ for $s = 0, 1, 2, \ldots, r - 1$. In $A_s$, every $b_i$ occurs $[m/r]$ times in the part

$$b_1b_2\cdots b_r\; b_{r+1}b_{r+2}\cdots b_{2r}\cdots b_{[m/r]-1}b_{[m/r]-1}b_{[m/r]-1}r+2\cdots b_{[m/r]}r$$
The $b_i$'s appeared in the left $b_i$ sequence in $A_1$ are: $b_1, b_2, \cdots, b_{N(m)}$ if $N(m) \geq 1$. Therefore,

$$P(A_1) = \begin{cases} 
  p_{[m/r]} \prod_{i=1}^{N(m)} p_i, & \text{if } N(m) > 0; \\
  p_{[m/r]}, & \text{if } N(m) = 0.
\end{cases}$$

By the same arguments, we generally have that

$$P(A_i) = \begin{cases} 
  p_{[m/r]} \prod_{i=0}^{N(m)-1} p_{i+i}, & \text{if } N(m) > 0; \\
  p_{[m/r]}, & \text{if } N(m) = 0.
\end{cases}$$

Then the desired result follows. ■

Let $Y_{i,j} = \max_{1 \leq i \leq j} \sum_{k=i}^{j} F(X_k, b_{k+k})$ for any $j \geq i \geq 1$. Define $\tau_0 = 0,$

$$\tau_1 = \inf \{ j \geq 1; Y_{1,j} = -\infty \} \quad \text{and} \quad \tau_k = \inf \{ j \geq \tau_{k-1}; Y_{\tau_{k-1}, j} = -\infty \} \quad k = 1, 2, \cdots .$$

Also, $T_n^- = \inf \{ k \geq 1; \tau_k \leq n \}$ and $T_n^+ = \inf \{ k \geq 1; \tau_k \geq n \}$. Then, $Y_{\tau_{k-1}, \tau_k-1} = \tau_k - \tau_{k-1}$ for any $k \geq 1$ and

$$\max \{ \tau_k - \tau_{k-1}, 1 \leq k \leq T_n^- \} \leq W_n \leq \max \{ \tau_k - \tau_{k-1}, 1 \leq k \leq T_n^+ \}. \quad (2.6)$$

**Lemma 2.3** Suppose $(b_1, b_2, \cdots, b_r) \neq (b_{i+1}, b_{i+2}, \cdots, b_{i+r})$ for any $i \in \{ 1, 2, \cdots, r-1 \}$. Let

$$K^{-1} = \sum_{j=2}^{\infty} j \left( \sum_{i=1}^{j} p_i p_{i+1} \cdots p_{i+j-2} (1 - p_{i+j-1}) \right) \quad (2.7)$$

Then $E(\tau_1) = 1/K \in (0, \infty)$ and

$$\lim_{n \to \infty} \frac{T_n^-}{n} = K \ a.s. \quad \text{and} \quad \lim_{n \to \infty} \frac{T_n^+}{n} = K \ a.s. \quad (2.8)$$

**Proof.** It is easy to see that $1/K < \infty$. First, $E\tau_1 = \sum_{j=2}^{\infty} j P(\tau_1 = j)$. Second, Set $B_i = \{ X_1 = b_i, X_2 = b_{i+1}, \cdots, X_{j-1} = b_{i+j-2}, X_j \neq b_{i+j-1} \}, \ i = 1, 2, \cdots$. Then, $B_1, B_2, \cdots B_j$ are disjoint. Since $\{ \tau_1 = j \} = \cup_{i=1}^{j} B_i$, $P(\tau_1 = j) = \sum_{i=1}^{j} P(B_i) = \sum_{i=1}^{j} p_i p_{i+1} \cdots p_{i+j-2} (1 - p_{i+j-1})$. So $E\tau_1 = K^{-1}$ is justified.

It is not difficult to see that $\{ \tau_1, \tau_2 - \tau_1, \tau_3 - \tau_2, \cdots \}$ is a sequence of i.i.d. random variables. Then, $E\tau_1 < \infty$ implies that $\tau_k < \infty \ a.s.$ for any $k \geq 1$. By definition, $\tau_k - \tau_{k-1} \geq 1$ for any $k \geq 1$, so $\tau_k \geq k$. Thus,

$$+\infty > T_n^+ \geq T_n^- \to +\infty \ a.s. \quad (2.8)$$
as \( n \to \infty \). We claim that
\[
\lim_{n \to \infty} \frac{\tau_n}{n} \to 1/K \; \text{a.s.} \tag{2.9}
\]

Once this is true, since \( \tau_{T_n^-} \leq n \) and \( \tau_{T_n^+} \geq n \), we obtain from (2.8) that \( T_n^-/n \to K \; \text{a.s.} \) as \( n \to \infty \). By the same arguments, we obtain that \( T_n^+/n \to K \; \text{a.s.} \) as \( n \to \infty \) too.

Now we turn to prove (2.9). Since \( \{X_1, X_2, \cdots \} \) is a sequence of i.i.d. random variables, it is easily seen that \( \{\tau_1, \tau_2 - \tau_1, \tau_3 - \tau_2, \cdots \} \) is also an i.i.d. sequence. By the Strong Law of Large Numbers, we have that \( \tau_n/n \to E\tau_1 \; \text{a.s.} \) as \( n \to \infty \). The (2.9) follows. \( \blacksquare \)

**The Proof of Theorem 2.** First observe that \( \tau_1 - 1 = Z_1 \). Let \( \{Z_1, \xi_i, i \geq 1\} \) be a sequence of i.i.d. random variables. Recall (2.6), for any \( \epsilon \in (0, 1) \), since \( T_n^-/n \to K \), we have that
\[
P\left(W_n \leq \frac{\log n}{\theta} + x\right) - \epsilon \leq P\left(\max_{k \leq (1-\epsilon)Kn} \xi_k \leq d_n(x) - 1\right) = (1 - P(Z_1 \geq d_n(x)))^m_n \tag{2.10}
\]
where \( m_n = [(1 - \epsilon)Kn] \). By Lemma 2.2, we have that \( P(Z_1 \geq d_n(x)) = p^{[d_n(x)/r]}gn(x) \), it follows that
\[
m_n P(Z_1 \geq d_n(x)) = K \exp\left(- \log(1/p)[d_n(x)/r] + \log(m_n/K) + \log g_n(x)\right).
\]

Note that
\[
|\log (m_n/K) - \log n| \leq |\log(1 - \epsilon - (1/nK))|.
\]

Recall \( h_n(x) = [d_n(x)/r] \log(1/p) - \log(ng_n(x)) \), we finally have that
\[
m_n P(Z_1 \geq d_n(x)) = K \exp(-h_n(x) + C_{n,\epsilon}), \tag{2.11}
\]
where \( C_{n,\epsilon} \) is a constant and \( \limsup_{\epsilon \downarrow 0} \limsup_{n \to \infty} C_{n,\epsilon} = 0 \). Since \( P(Z_1 \geq d_n(x)) \to 0 \) as \( n \to \infty \), by (2.10) and (2.11), we have
\[
\limsup_{n \to \infty} e^{h_n(x)} P\left(W_n \leq \frac{\log n}{\theta} + x\right) \leq 1
\]

By applying the above arguments to \( T_n^+ \) and the right hand side of (2.6), we obtain that
\[
\liminf_{n \to \infty} e^{h_n(x)} P\left(W_n \leq \frac{\log n}{\theta} + x\right) \geq 1
\]

The proof is completed. \( \blacksquare \)