An Erdős-Révész Type Law for the Length of the Longest Match of Two Coin-Tossing Sequences

Karl Grill *

Posted Date: 18 November 2024

doi: 10.20944/preprints202411.1254.v1

Keywords: coin-tossing; runs; matching subsequences; strong asymptotics

 doi:10.20944/preprints202411.1254.v1

Article

An Erdős-Révész Type Law for the Length of the Longest Match of Two Coin-Tossing Sequences

Karl Grill †

Institute of Statistics and Mathematical Methods in Economy, TU Wien; karl.grill@tuwien.ac.at

† Current address: Institute of Statistics and Mathematical Methods in Economy, TU Wien, Wiedner Hauptstraße 8-10, 1040 Wien, Austria.

Abstract: Consider a coin-tossing sequence, i.e., a sequence of independent variables taking values 0 and 1 with probability 1/2. The famous Erdős-Rényi (1970) law of large numbers implies that the longest run of ones in the first n observations has a length R_n that behaves like $\log_2(n)$ as n tends to infinity. Erdős and Révész (1976) refined this result by giving a description of the Lévy upper and lower classes of the process R_n. In another direction, Arratia and Waterman (1985) extended the Erdős-Rényi result to the longest matching subsequence (with shifts) of two coin-tossing sequences, finding that it behaves asymptotically like $2\log_2(n)$. The present paper gives some Erdős-Révész-type results in this situation, obtaining a complete description of the upper classes and a partial result on the lower ones.

Keywords: coin-tossing; runs; matching subsequences; strong asymptotics;

MSC: 60F15

1. Introduction

Consider a coin-tossing sequence (X_n), i.e. a sequence of independent random variables satisfying $\mathbb{P}(X_n = 0) = \mathbb{P}(X_n = 1) = 1/2$. Let R_n be the length of the longest head-run, i.e., the largest integer r for which there is an i, $0 \le i \le n - r$, for which $X_{i+j} = 1$ for $j = 1, \dots r$. A result of Erdős and Rényi [2] implies that

$$\lim_{n\to\infty} \frac{R_n}{\log(n)} = 1 \tag{1}$$

(throughout this paper, log will denote base 2 logarithms. The notation $\log_k$ will be used for its iterates: $\log_2(x) = \log(\log(x))$, $\log_{k+1}(x) = \log(\log_k(x))$. Also C and c, with or without index, are used to denote generic constants that may have different values at each occurrence). The simple result (1) has seen a number of improvements. Erdős and Révész [3] gave a detailed description of the asymptotic behavior of R_n. In order to formulate their result, let us recall

Definition 1 (Lévy classes). *Let (Y_n) be a sequence of random variables. We say that a sequence (a_n) of real numbers belongs to*

- *The upper-upper class of (Y_n) ($UUC(Y_n)$), if, with probability 1 as $n \to \infty$, $Y_n \le a_n$ eventually.*
- *The upper-lower class of (Y_n) ($ULC(Y_n)$), if, with probability 1 as $n \to \infty$, $Y_n > a_n$ for infinitely many n.*
- *The lower-upper class of (Y_n) ($LUC(Y_n)$), if, with probability 1 as $n \to \infty$, $Y_n < a_n$ for infinitely many n.*
- *The lower-lower class of (Y_n) ($ULC(Y_n)$), if, with probability 1 as $n \to \infty$, $Y_n \ge a_n$ eventually.*

Of course, these definitions work best, if the sequence (Y_n) obeys some zero-one law.

Their result is as follows:
Let (a_n) be a nondecreasing integer sequence. Then

- $(a_n) \in UUC(R_n)$ if $\sum_n 2^{-a_n} < \infty$,

 doi:10.20944/preprints202411.1254.v1

- $(a_n) \in ULC(R_n)$ if $\sum_n 2^{-a_n} = \infty$,
- for any $\epsilon > 0$, $a_n = \lfloor \log(n) - \log_3(n) + \log_2(e) - 1 + \epsilon \rfloor \in LUC(R_n)$,
- for any $\epsilon > 0$, $a_n = \lfloor \log(n) - \log_3(n) + \log_2(e) - 2 - \epsilon \rfloor \in LLC(R_n)$.

Arratia and Waterman [1] extend Erdős and Rényi's result in another direction: they consider two independent coin-tossing sequences (X_n) and (Y_n) and look for the longest matching subsequences when shifting is allowed. Formally, let $M(n)$ be the the largest integer m for which there are i, j with $0 \le i, j \le n - m$ and $X_{i+k} = Y_{j+k}$ for all $k = 1, \ldots, m$. They prove that, with probability 1

$$\lim_{n\to\infty} \frac{M_n}{\log(n)} = 2. \tag{2}$$

In the present paper, we will make this more precise by giving a description for the upper classes of (M_n) and also some results on its lower classes:

Theorem 1. *Let (a_n) be a nondecreasing integer sequence. We have*

- $(a_n) \in UUC(M_n)$ *if* $\sum_n n2^{-a_n} < \infty$.
- $(a_n) \in ULC(M_n)$ *if* $\sum_n n2^{-a_n} = \infty$.
- *for some c,* $a_n = \lfloor 2\log(n) - \log_3(n) + c \rfloor \in LUC(M_n)$.
- *for some c,* $a_n = \lfloor 2\log(n) - \log_2(n) - log_3(n) + c \rfloor \in LLC(M_n)$.

2. Discussion

We leave the proof of Theorem 1 for later and rather discuss some of the concepts that are connected to this problem. One of them is the so-called independence principle: in many, though not all, situations, one may pretend that the waiting times until a given pattern of length l is observed have an exponential distribution with parameter 2^{-l}, and that the waiting times for different patterns are independent. Móri [4] and Móri and Székely [6] give an account of this principle and its limitations. In our case, all results but the lower-lower class one are more or less in tune with this principle.

Another question that is closely related is that of the number $N(n, l)$ of different length l subsequences of $(X_1, \ldots, X_n)$. This question doesn't seem to have been touched by literature very much; one remarkable result by Móri [5] states that for $l = \lceil \log(n) - \log_2(n) \rceil$ and n large enough, all 2^l possible patterns occur as subsequences of $(X_1, \ldots, X_n)$. The independence principle would suggest that $N(n, \log(n))/n$ is bounded away from 0 with probability one, and this or even the easier $N(n, \log(n)) \ge n(\log_2(n))^{-c}$ eventually would serve to remove the double log term from the LLC result. Unfortunately, we are only able to get $N(n, \log(n)) \ge cn/\log(n)$, which is also implied by Móri's result.

3. Proofs

Proof of the upper-upper class result. Both upper class statements are fairly easy to prove. First observe that under our assumptions, the convergence of

$$\sum_{n=1}^{\infty} n2^{-a_n} \tag{3}$$

is equivalent to that of

$$\sum_{k=1}^{\infty} n_k^2 2^{-a_{n_k}} \tag{4}$$

with $n_k = 2^k$.
Now, define events

$$A_k = [M_{n_k} \ge a_{n_{k-1}}]. \tag{5}$$

 doi:10.20944/preprints202411.1254.v1

A_k occurs if in one of the $(n_k + 1 - a_{n_{k-1}})^2$ pairs of sequences

$$((X_{i+1}, \dots, X_{i+a_{k-1}}), ((Y_{j+1}, \dots, Y_{j+a_{k-1}})) \tag{6}$$

both sequences agree. That gives the trivial upper bound

$$\mathbb{P}(A_k) \leq n_k^2 2^{-a_{n_{k-1}}}, \tag{7}$$

so, by our assumptions, $\sum_n \mathbb{P}(A_k) < \infty$, and the Borel-Cantelli lemma implies that, with probability 1, only finitely many events A_k occur. Thus, for sufficiently large k, $M_{n_k} \leq a_{n_{k-1}}$, and for $n_{k-1} \leq n \leq n_k$, we have

$$M_n \leq M_{n_k} \leq a_{n_{k-1}} \leq a_n. \tag{8}$$

This shows that $(a_n) \in UUC(M_n)$, as claimed. □

Proof of the upper-lower class result. We may assume without loss of generality that $n^2 2^{-a_n} \leq 1/4$.

Again, let $n_k = 2^k$. We want to use the second Borel-Cantelli lemma, so we are defining independent events

$$A_k = [\exists i, j : n_{k-1} < i, j \leq n_k : X_{i+s} = Y_{j+s}, s = 0, \dots, a_{n_k} - 1] \tag{9}$$

This is the union of the events

$$B_{ij} == [X_{i+s} = Y_{j+s}, s = 0, \dots, a_{n_k} - 1] \tag{10}$$

with $n_{k-1} < i, j \leq n_k$. We endow the set of pairs (i, j) with the lexicographic order. For a subset I of the real numbers, Bonferroni's inequality gives

$$\mathbb{P}(\bigcup_{(i,j)\in I\times I} B_{ij}) \geq \sum_{(i,j)\in I\times I} \mathbb{P}(B_{ij}) - \sum_{(i,j),(i',j')\in I\times I,(i,j)<(i',j')} \mathbb{P}(B_{ij} \cap B_{i'j'}). \tag{11}$$

Let $d((i, j), (i', j')) = \max(|i - i'|, |j - j'|)$. If $d((i, j), (i', j')) \geq a_{n_k}$, then $\mathbb{P}(B_{ij} \cap B_{i'j'}) = 2^{-2a_{n_k}}$, otherwise $\mathbb{P}(B_{ij} \cap B_{i'j'}) = 2^{-(a_{n_k} + d((i,j),(i',j')))}$.

Setting $I = \{i : n_{k-1} < i \leq n_k : 2|c\}$ in (11) yields, after some calculation

$$\mathbb{P}(A_k) \geq \frac{1}{48} n_k^2 2^{-a_{n_k}}, \tag{12}$$

and $\sum_k \mathbb{P}(A_k) = \infty$. Borel-Cantelli implies that, with probability 1, infinitely many events A_k occur. Thus, for infinitely many k, $M_{n_k} \geq a_{n_k}$, so $(a_n) \in ULC(M_n)$. □

For the lower class results, we first prove some lemmas:

Lemma 1. *For any $c \in (0, 1)$, with probability 1 eventually*

$$c\frac{n}{\log(n)} \leq N(n, \log(n)) \leq n \tag{13}$$

Proof of Lemma 1. The lower part is a direct consequence of Móri's result: this states that, for sufficiently large n $N(n, \lfloor \log(n) - \log_2(n) \rfloor) = 2^{\lfloor \log(n) - \log_2(n) \rfloor} = \frac{n}{\log(n)}(1 + o(1))$ and, obviously $N(n, \log(n)) \geq N(n, \log(n) - \log_2(n)$, as extending two different sequences from length $\log(n) - \log_2(n)$ keeps them different; it can only happen that some of them are extended beyond index n. □

 doi:10.20944/preprints202411.1254.v1

Lemma 2. *Let S be a set of $m < 2^l$ sequences of length $l < n$, and let A be the event that none of the sequences in S occurs as a subsequence of $(X_1, \ldots, X_n)$. Assume $nml^2 2^{-2l} < \gamma < 1$. Then Then there are positive constants C_1, C_2, c_1 and c_2 (depending on γ such that*

$$C_1 \exp(-c_1 mn 2^{-l}) \leq \mathbb{P}(A) \leq C_2 \exp(-c_2 mn 2^{-l}). \tag{14}$$

Proof of Lemma 2. The assumptions imply that we can find n' such that both $n'm2^{-l} < 1$ and $\frac{n}{n'} lm2^{-l} < 1$. The probability that there is a sequence from S in a coin-tossing sequence of length n' can be trivially bounded above by $n'm2^{-l}$, and a Bonferroni-type argument like the one we have seen before shows that this is bounded below by $cn'm2^{-l}$ (with c depending on γ, of course). So, we get an upper bound for the probability of A from the probability that there is no sequence from S in any of the n/n' Blocks of length n', amounting to

$$(1 - cn'm2^{-l})^{n/n'} \leq \exp(-cnm2^{-l}). \tag{15}$$

For the lower bound, the probability that there is no sequence from S in any Block of length n' cannot directly be used as a lower bound for the probability of A, as there may be sequences crossing the border between two blocks. We can fix this by subtracting the sum of the probabilities of a sequence crossing a border multiplied by the product of the probabilities of all blocks except the two adjacent two the border in question. Thus, the lower bound looks like

$$(1 - n'm2^{-l})^{n/n'-2} \Big((1 - n'm2^{-l})^2 - \frac{n}{n'} ml2^{-l} \Big). \tag{16}$$

This can be bounded below by

$$C_1 \exp(-c_1 nm2^{-l}) \tag{17}$$

as claimed. □

Proof of the lower-lower class result. Combining Lemmas 1 and 2, we get an "almost" upper bound for the probability (more to the point, an upper bound for the conditional probability with respect to the event that the number of different sequences of length a_n among $Y_1, \ldots, Y_n$ is at least $cn/\log(n)$) that the longest match is shorter than a_n amounting to

$$C_1 \exp(-c \frac{n^2}{\log(n)} 2^{-a_n}). \tag{18}$$

Letting $n = 2^k$ again, and substituting a_n as defined in the theorem, choosing the constants appropriately we get an upper bound $O(k^{-\alpha})$ with $\alpha > 1$, giving a convergent series again, so another appeal to Borel Cantelli finishes this part. □

Proof of the lower-upper class result. The main obstacle in this proof is the need to find independent or almost independent events that can be fed into an appropriate Borel-Cantelli lemma. We go for conditional independence. With an appropriate constant C, we start with a sufficiently large τ_0 and define τ_k as the minimum of $\tau_{k-1} + C\sqrt{2^k \log(k)}$ and the smallest n for which there is a match of length k between $X_1, \ldots, X_n$ and $Y_{\tau_{k-1}+1}, \ldots, Y_n$ or between $Y_1, \ldots, Y_n$ and $X_{\tau_{k-1}+1}, \ldots, X_n$. We let A_k denote the event that $\tau_k - \tau_{k-1} > \delta\sqrt{2^k \log(k)}$. By our construction, $\tau_k = O(\sqrt{2^k \log(k)})$, and we can use this upper bound as m in Lemma 2. This gives a lower bound $\exp(-c\delta \log(k))$ for $\mathbb{P}(A_k)$. By Borel-Cantelli, we conclude that, with probability one, infinitely many of the events A_k occur. This puts us almost at our goal, the only possible problem is that there may be a match of length k crossing one of the boundaries τ_k. The probability for this is easily bounded above by $O(k\sqrt{2^{-k} \log(k)})$, and another Borel-Cantelli argument shows that eventually this does not happen. □

 doi:10.20944/preprints202411.1254.v1

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Arratia, R; Waterman, S. An Erdős-Rényi law with shifts. *Adv. Math.* **1985** *55(1)*, 13–23.
2. Erdős, P.; Rényi, A. On a new law of large numbers. *J. Analyse Math.* **1970**, *23*, 103–111.
3. Erdős, P.; Révész, P. On the length of the longest head-run. *Coll. Math. soc. J. Bolyai: Topics in Information Theory*; Csiszár, I., Elias, P., Eds.; **1976** *23*, 219–228.
4. Móri, T. Large deviation results for waiting times in repeated experiments. *Acta Math. Hung.* **1985**, *45(1–2)*, 213–221.
5. Móri, T. On the waiting time till each of some given patterns occurs as a run. *Probab. Th. Rel. Fields* **1991**, *87*, 313–323.
6. Móri, T.; Székely, G. Asymptotic independence of pure head stopping times. *Stat. Probabil. Lett.* **1984**, *2*, 5–8.