Minimum frequencies of occurrences of squares and letters in infinite words

Pascal Ochem
CNRS-LRI-Université Paris Sud 11, France
ochem@lri.fr

Michaël Rao
CNRS-LIRMM-Université Montpellier 2, France
rao@lirmm.fr

Abstract

We prove that the limit of the ratio the minimal number of squares occurrences in a binary word over its size is $\frac{103}{187} = 0.5508021\ldots$. The same proof technique is applied to compute lower bounds on the function $\rho(x)$ corresponding to the minimal letter frequency in an infinite x-free word. This leads to some exact values of $\rho(x)$ for $x < \frac{5+\sqrt{5}}{2}$. Finally, we give a conjecture for the value of $\rho(x)$ for $x \geq \frac{5+\sqrt{5}}{2}$.

1 Introduction

A *square* is a factor of the form uu where u is a non-empty word. Thue's famous result show that squares can be avoided in an infinite ternary word [7, 8]. We are interested in the minimum number of square occurrences in a binary word.

Let $\Sigma_2 = \{0, 1\}$. For $w \in \Sigma_2^*$, let $s(w)$ be the number of (possibly overlapping) square occurrences in w. For $n \in \mathbb{N}$, let $m(n) = \min_{w \in \Sigma_2^n} s(w)$. Let $\alpha = \lim_{n \to \infty} \frac{m(n)}{n}$.

We have shown [5] that $\frac{1815}{3297} \leq \alpha \leq \frac{103}{187}$. We prove here that:

Theorem 1. *The exact value of α is $\frac{103}{187}$ $(= 0.5508021390\ldots)$.*

Let $x \in \mathbb{R}$. A word w is an x-power if there exists a k such that $\frac{|w|}{k} = x$ and $w[i-k] = w[i]$ for all $i \in \{k+1, \ldots, |w|\}$. A square is a 2-power. A word is *x-free* (resp. *(x^+)-free*) if it does not contain as factor any x-power such that $y \geq x$ (resp $y > x$).

Let $\rho(x)$ (resp. $\rho(x^+)$) be the minimal density of a letter in an infinite binary word with no repetition of exponent $\geq x$ (resp. $> x$). The function ρ has been defined in [4] and also studied in [6]. This function is defined starting from 2^+ since

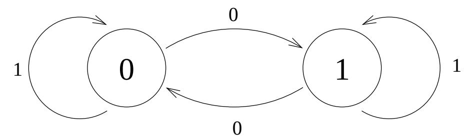

Figure 1: De Bruijn graph of words of size 1 ($\lambda^*(G) = 0$).

square are unavoidable in an infinite binary word, and there exists an infinite (2^+)-free binary word [8]. Moreover, ρ is decreasing and is equal to $1/2$ on the interval $[2^+, 7/3]$ [4].

The same proof technique can be applied to compute lower bounds on the function $\rho(x)$ corresponding to the minimal letter frequency in an infinite x-free word. This leads to new exact values of $\rho(x)$ for $x < \frac{5+\sqrt{5}}{2}$. We also propose a conjecture for the value of $\rho(x)$ for $x \geq \frac{5+\sqrt{5}}{2}$.

2 Suffix graphs

Let $v \in \Sigma_2^* \setminus \epsilon$. Let $v^\sharp$ be the last letter of v, and let $v^\bullet$ be the prefix of v of size $|v| - 1$. Note that $v = v^\bullet v^\sharp$.

Definition 2. A *good suffix cover* is a set of words V such that

(a) $\emptyset \subsetneq V \subseteq \Sigma_2^* \setminus \{\epsilon\}$.

(b) For every $u, v \in V$ with $u \neq v$, u is not a suffix of v.

(c) For every left-infinite word w, there is a $v \in V$ such that v is a suffix of w.

(d) For every $u \in V$, there is a $v \in V$ such that $u^\bullet$ is a suffix of v.

Definition 3. A *suffix graph* $G = (V, A, w)$ is a directed graph (V, A) with weight function $w : A \to \mathbb{N}$ such that:

- V is a good suffix cover.
- There is an arc (u, v) if $v^\bullet$ is a suffix of u.
- The weight of an arc (u, v) is $s(uv^\sharp) - s(u)$, (*i.e.* the number of squares involving the last letter in $uv^\sharp$).

For example, De Bruijn graphs with the appropriate weight function are suffix graphs. Note that a suffix graph is uniquely determined by the good suffix cover.

Lemma 4. *If $G = (V, A, w)$ is a suffix graph, then we have:*

1. *For every $w \in \Sigma_2^*$, there exists $v \in V$ such that v is a suffix of w or w is a suffix of v.*

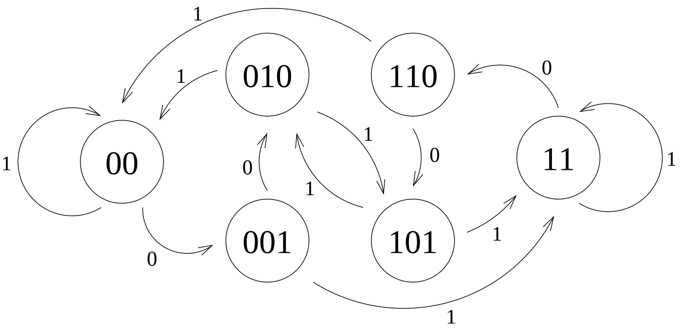

Figure 2: A suffix graph with $\lambda^*(G) = 1/3$.

2. Every vertex has out-degree two.

3. Every vertex has in-degree at least one.

Proof. (1) Let $w \in \Sigma_2^*$ and let w' be a left-infinite word with suffix w. By (c), there exists $v \in V$ which is a suffix of w'. Then either w is a suffix of v, or v is a suffix of w.

(2) Let $v \in V$ and $x \in \Sigma_2$. Let $u_x \in V$ be such that either u_x is a suffix of vx or vx is a suffix of u_x. If u_x is a suffix of vx then $(v, u_x) \in A$ by definition. Otherwise, by (d), $u_x^\bullet$ is a suffix of some $w \in V$. Then v is a suffix of w, and thus $v = w$ by (b).

Thus $v \in V$ has exactly two distinct out-neighbors since $u_0 \neq u_1$.

(3) Let $v \in V$. By (d), there exists $u \in V$ such that $v^\bullet$ is a suffix of u. Thus $(u, v) \in A$. □

Let $G = (V, A, w)$ be a suffix graph. A *walk* is a sequence $P = (v_1, \ldots, v_k)$ of vertices in V such that for all $i \in \{1, \ldots, k-1\}$, $(v_i, v_{i+1}) \in A$. A *circuit* is a circular sequence $C = (v_1, \ldots, v_k)$ of vertices in V such that for all $i \in \{1, \ldots, k\}$, $(v_i, v_{i+1}) \in A$ (indices are taken modulo k). The *size* $l(C)$ of a circuit (resp. walk) is k. The weight $w(C)$ of a circuit (resp walk) is $\sum_{i\in\{1,\ldots,k\}} w((v_i, v_{i+1}))$ (resp. $\sum_{i\in\{1,\ldots,k-1\}} w((v_i, v_{i+1}))$).

The *minimum mean circuit* of G is $\lambda^*(G) = \min_{C \text{ circuit of } G} \frac{w(C)}{l(C)}$. A circuit C with $\frac{w(C)}{l(C)} = \lambda^*(G)$ can be found in polynomial time with a dynamic approach [3].

Lemma 5. *Let G be a suffix graph. Then $\lambda^*(G) \leq \alpha$.*

Proof. Similar to the proof of Lemma 9 in [5]. □

We show in [5] that $\alpha \leq \frac{103}{187}$. We explain how to construct a suffix graph with $\lambda^*(G) \geq \frac{103}{187}$ in the next section. This proves that $\alpha = \frac{103}{187}$.

3 Construction of a suffix graph with $\lambda^* = \frac{103}{187}$

Proposition 6. *Let $(u, v) \in A$ such that $|u| < |v|$. Then $|u| = |v| - 1$, and u is the only in-neighbor of v.*

Proof. By definition, $|u| = |v| - 1$ and there exists $x \in \Sigma_2$ such that $ux = v$. Suppose that v has an other in-neighbor w. Then there exists $x' \in \Sigma_2$ such that v is a suffix of wx'. Thus $x = x'$ and u is a suffix of w. Contradiction. □

We say that a vertex $v \in V$ is *critical* if there exists $u \in V$ such that u is the suffix of v of size $|v| - 1$. The critical vertices of the graph in Figure 2 are 001 and 110.

Lemma 7. *Let $G = (V, A, w)$ be a suffix graph, and let $v \in V$ be a non-critical vertex. Then there exists a unique suffix graph $G * v$ with vertex set $V' = (V \setminus \{v\}) \cup \{0v, 1v\}$.*

Proof. We only need to show that V' is a good prefix cover. Clearly, V' respects (a), (b) and (c). Suppose that (d) is not fulfilled and let $u \in V'$ such that $u^\bullet$ is not a suffix of any word in V'. Then $u \in \{0v, 1v\}$. W.l.o.g. $u = 0v$. Let $w \in V$ be such that either w is a suffix of $0v^\bullet$ or $0v^\bullet$ is a suffix of w. We have $w \neq v$, otherwise $0v^\bullet$ will be a suffix of $0w \in V'$. Thus $w \in V'$. If w is a suffix of $0v^\bullet$, then $w' = 0v^\bullet$ otherwise w' would be a suffix of $v^\bullet$ and thus v would be critical. In all cases, $0v^\bullet$ is suffix of $w \in V'$. Contradiction. □

We describe now the algorithm used to obtain the graph. We start with $G = DB_1$ (Figure 1). While $\lambda^*(G) < \frac{103}{187}$, we take a circuit C of ratio $\frac{w(C)}{l(C)} = \lambda^*(G)$, we take a vertex v in C of minimum length, and we replace G by $G * v$. Note that a vertex of minimum length on the cycle cannot be critical.

This algorithm stops with a graph G of size 62739. For this graph, $\lambda^*(G) \geq \frac{103}{187}$ thus by Lemma 5, $\alpha \geq \frac{103}{187}$. With the result of [5], this proves Theorem 1.

4 Minimal letter frequency in infinite repetition-free words

A similar technique can be applied to obtain a lower bounds on the minimal letter frequency in an infinite x-free binary word.

Using the technique described in previous sections, and techniques described in [6], we get:

Theorem 8.

$$
\begin{array}{rclcl}
\rho(2+) = \rho(7/3) & = & 1/2 & = & 0.5 \\
\rho(7/3+) = \rho(407/172) & = & 327/703 & = & 0.4651493598\ldots \\
\rho(407/172+) = \rho(833/344) & = & 347/746 & = & 0.4651474530\ldots \\
\rho(833/344+) & \leq & 6012/12925 & = & 0.4651450676\ldots \\
\rho(17/7) & \geq & 754/1621 & = & 0.4651449722\ldots \\
\rho(17/7+) & \leq & 2129/4600 & = & 0.4628260869\ldots \\
\rho(298/121) & \geq & 3318/7169 & = & 0.4628260566\ldots
\end{array}
$$

$\rho(298/121+)$	$\le$	$6841/14781$	$= 0.4628238955\ldots$
$\rho(5/2)$	$\ge$	$54286/117293$	$= 0.4628238684\ldots$
$\rho(5/2+)$	$\le$	$2767/6258$	$= 0.4421540428\ldots$
$\rho(131/52)$	$\ge$	$3818/8635$	$= 0.4421540243\ldots$
$\rho(131/52+) = \rho(43/17)$	$=$	$191/432$	$= 0.4421296296\ldots$
$\rho(43/17+)$	$\le$	$4309/9753$	$= 0.4418127755\ldots$
$\rho(23/9)$	$\ge$	$6678/15115$	$= 0.4418127687\ldots$
$\rho(23/9+)$	$\le$	$8437/19101$	$= 0.4417046227\ldots$
$\rho(41/16)$	$\ge$	$197/446$	$= 0.4417040358\ldots$
$\rho(41/16+) = \rho(18/7)$	$=$	$79/179$	$= 0.4413407821\ldots$
$\rho(18/7+)$	$\le$	$3983/9035$	$= 0.4408411732\ldots$
$\rho(631/245)$	$\ge$	$1740/3947$	$= 0.4408411451\ldots$
$\rho(631/245+)$	$\le$	$2306/5231$	$= 0.4408334926\ldots$
$\rho(2900/1107)$	$\ge$	$5480/12431$	$= 0.4408334003\ldots$
$\rho(2900/1107+)$	$\le$	$1926/4369$	$= 0.4408331425\ldots$
$\rho(2917/1107)$	$\ge$	$4720/10707$	$= 0.4408330998\ldots$
$\rho(2917/1107+)$	$\le$	$5696/12921$	$= 0.4408327528\ldots$
$\rho(8/3)$	$\ge$	$10144/23011$	$= 0.4408326452\ldots$
$\rho(8/3+)$	$\le$	$241/593$	$= 0.4064080944\ldots$
$\rho(886/315)$	$\ge$	$12152/29901$	$= 0.4064078124\ldots$
$\rho(886/315+)$	$\le$	$6520/16043$	$= 0.4064077790\ldots$
$\rho(197/69)$	$\ge$	$5430/13361$	$= 0.4064067060\ldots$
$\rho(197/69+)$	$\le$	$1459/3590$	$= 0.4064066852\ldots$
$\rho(901/315)$	$\ge$	$7473/18388$	$= 0.4064063519\ldots$
$\rho(901/315+)$	$\le$	$38131/93825$	$= 0.4064055422\ldots$
$\rho(26/9)$	$\ge$	$1561/3841$	$= 0.4064045821\ldots$
$\rho(26/9+) = \rho(79/27)$	$=$	$89/219$	$= 0.4063926940\ldots$
$\rho(79/27+) = \rho(202/69)$	$=$	$662/1629$	$= 0.4063842848\ldots$
$\rho(202/69+)$	$\le$	$853/2099$	$= 0.4063839923\ldots$
$\rho(44/15)$	$\ge$	$675/1661$	$= 0.4063816977\ldots$
$\rho(44/15+)$	$\le$	$447/1100$	$= 0.4063636363\ldots$
$\rho(3)$	$\ge$	$5570/13707$	$= 0.4063617129\ldots$
$\rho(3+)$	$\le$	$332/1149$	$= 0.2889469103\ldots$
$\rho(31/10)$	$\ge$	$1981/6856$	$= 0.2889439906\ldots$
$\rho(31/10+)$	$\le$	$4442/15393$	$= 0.2885727278\ldots$
$\rho(1554/499)$	$\ge$	$6389/22140$	$= 0.2885727190\ldots$
$\rho(1554/499+)$	$\le$	$2149/7447$	$= 0.2885725795\ldots$
$\rho(22/7)$	$\ge$	$2899/10046$	$= 0.2885725661\ldots$
$\rho(22/7+) = \rho(67/21)$	$=$	$126/437$	$= 0.2883295194\ldots$
$\rho(67/21+)$	$\le$	$1781/6180$	$= 0.2881877022\ldots$
$\rho(11501/3581)$	$\ge$	$4594/15941$	$= 0.2881876921\ldots$
$\rho(11501/3581+)$	$\le$	$7407/25702$	$= 0.2881876896\ldots$
$\rho(68/21)$	$\ge$	$2813/9761$	$= 0.2881876856\ldots$
$\rho(68/21+)$	$\le$	$2777/9643$	$= 0.2879809188\ldots$
$\rho(13/4)$	$\ge$	$4828/16765$	$= 0.2879809126\ldots$

$\rho(13/4+)$	$\leq$	$10289/36400$	$=$	$0.2826648351\ldots$
$\rho(36/11)$	$\geq$	$1642/5809$	$=$	$0.2826648304\ldots$
$\rho(36/11+) = \rho(23/7)$	$=$	$13/46$	$=$	$0.2826086956\ldots$
$\rho(23/7+) = \rho(83/25)$	$=$	$37/132$	$=$	$0.2803030303\ldots$
$\rho(83/25+) = \rho(37/11)$	$=$	$442/1577$	$=$	$0.2802790107\ldots$
$\rho(37/11+) = \rho(38/11)$	$=$	$44/157$	$=$	$0.2802547770\ldots$
$\rho(38/11+) = \rho(7/2)$	$=$	$27/97$	$=$	$0.2783505154\ldots$
$\rho(7/2+) = \rho(103/29)$	$=$	$5/18$	$=$	$0.2777777777\ldots$
$\rho(103/29+) = \rho(168/47)$	$=$	$23/83$	$=$	$0.2771084337\ldots$
$\rho(168/47+) = \rho(273/76)$	$=$	$129/466$	$=$	$0.2768240343\ldots$
$\rho(273/76+) = \rho(443/123)$	$=$	$109/394$	$=$	$0.2766497461\ldots$
$\rho(443/123+) = \rho(718/199)$	$=$	$112/405$	$=$	$0.2765432098\ldots$
$\rho(718/199+) = \rho(1163/322)$	$=$	$569/2058$	$=$	$0.2764820213\ldots$
$\rho(1163/322+) = \rho(1883/521)$	$=$	$473/1711$	$=$	$0.2764465225\ldots$
$\rho(1883/521+) = \rho(1016/281)$	$=$	$1556/5629$	$=$	$0.2764256528\ldots$
$\rho(1016/281+) = \rho(4933/1364)$	$=$	$225/814$	$=$	$0.2764127764\ldots$
$\rho(4933/1364+) = \rho(7983/2207)$	$=$	$1018/3683$	$=$	$0.2764051045\ldots$
$\rho(7983/2207+)$	$\leq$	$6656/24081$	$=$	$0.2764004817\ldots$
$\rho(4)$	$\geq$	$2584/9349$	$=$	$0.2763931971\ldots$

Whereas our previous method for lower bounds [6] was not well suited for $x > 3$, the new method also handles this case. Theorem 8 gives in particular the exact value for ρ on the intervals $[2^+, 833/344]$, $[131/52+, 43/17]$, $[41/16+, 18/7]$, $[26/9+, 202/69]$, $[22/7+, 67/21]$, and $[36/11+, 1016/281]$. Moreover, ρ is piecewise constant on these intervals. We calculated that the decreasing between $\rho(2^+) = 1/2$ and $\rho(4) \geq 2584/9349$ is now almost completely due to the jumps except for an amount smaller than 2×10^{-5}.

5 A conjecture for $x \geq \frac{5+\sqrt{5}}{2}$

We propose the following conjecture for $x \geq \frac{5+\sqrt{5}}{2}$. Note that the conjectured values are irrational, thus the techniques presented in [6] and in this article cannot prove these values.

Conjecture. For every integer $n \geq 4$,

1. $\rho([n-1, \overline{1, n-3}]) = \rho(n) = [0, n-1, \overline{1, n-3}]$,

2. for $k \in \mathbb{N}, \rho(U_{n,k}^+) = \rho(U_{n,k+1}) = [0, n(, 1, n-2)^k, \overline{1, n-3}]$.

where $[a, b, c, \ldots]$ denotes the continued fraction $a + 1/(b + 1/(c + \ldots))$, and $U_{n,k} = n + 1 - \frac{D_{n,k-1}+2}{D_{n,k}},\ D_{n,-1} = -1,\ D_{n,0} = 1,\ D_{n,k+1} = nD_{n,k} - D_{n,k-1}$.
The values of $\rho(x)$ are given by the sturmian word of density (or slope) $\rho(x)$.

We need a result of Damanik and Lenz [1] in order to prove the upper bounds of the conjecture. Every irrational $\alpha \in (0,1)$ has a unique continued fraction expansion $\alpha = [0, a_1, a_2, a_3, \ldots]$. The rational approximants $\frac{p_t}{q_t}$ of α are defined by

$$\begin{aligned} p_0 = 0, \quad p_1 = 1, \quad p_t = a_t p_{t-1} + p_{t-2}, \\ q_{-1} = 0, \quad q_0 = 1, \quad q_t = a_t q_{t-1} + q_{t-2}. \end{aligned}$$

Theorem 9. *[1]*
The largest exponent of a repetition in the sturmian word of slope α is

$$2 + \sup_{t \in \mathbb{N}} \left\{ a_{t+1} + \frac{q_{t-1} - 2}{q_t} \right\}.$$

Theorem 10. *For every integer $n \geq 4$,*

1. $\rho([n-1, \overline{1, n-3}]) \leq [0, n-1, \overline{1, n-3}]$,
2. *for* $k \in \mathbb{N}, \rho(U^+_{n,k}) \leq [0, n(, 1, n-2)^k, \overline{1, n-3}]$.

Proof.
[2]. Let $n \geq 4$, $k \in \mathbb{N}$ and let w be the Sturmian word of slope $[0, n(, 1, n-2)^k, \overline{1, n-3}]$. We show that the largest exponent of a repetition in w is $U_{n,k}$. Let $\beta_i = 2 + a_{i+1} + \frac{q_{i-1}-2}{q_i}$. It is not hard to see that $0 \leq \frac{q_{i-1}-2}{q_i} \leq 1$ for all $i > 1$. Thus if $q = 0$, then the greatest exponent in w is $\beta_0 = n = U_{n,0}$. Otherwise, the greatest exponent is $\sup_{i \in \{1,\ldots k\}} \beta_{2i}$. One can easily show by induction than $D_i = q_{2i}$ for all i. Thus for all $i \in \{1, \ldots k\}$:

$$\beta_{2i} = n + \frac{q_{i-1} - 2}{q_i} = n + \frac{q_i - q_{i-2} - 2}{q_i} = n + 1 - \frac{q_{i-2} + 2}{q_i} = U_{n,k}.$$

To conclude, we show that $\{U_{n,i}\}_i$ is increasing (note that $D^2_{n,i} - D_{n,i+1} D_{n,i-1} = n+2$ for all i):

$$\begin{aligned} U_{n,i+1} - U_{n,i} &= \frac{1}{D_{n,i+1} D_{n,i}} \{D_{n,i+1}(D_{n,i-1} + 2) - D_{n,i}(D_{n,i} + 2)\} \\ &= \frac{1}{D_{n,i+1} D_{n,i}} \{2D_{n,i+1} - 2D_{n,i} - (n+2)\} \geq 0. \end{aligned}$$

[1, $n > 4$]. Let w be the Sturmian word of slope $[0, n-1, \overline{1, n-3}]$. With the same arguments, the greatest exponent in w is $\lim_{i \to \infty} U_{n-1,i}$.

$$\begin{aligned} \lim_{i \to \infty} U_{n-1,i} &= n - \lim_{i \to \infty} \frac{D_{n-1,i-1}}{D_{n-1,i}} \\ &= n - \frac{2}{n - 1 + \sqrt{(n-1)^2 - 4}} = \frac{n + 1 + \sqrt{(n-1)^2 - 4}}{2} \\ &= [n-1, \overline{1, n-3}]. \end{aligned}$$

[1, $n = 4$]. Let w be the Sturmian word of slope $[0,3,\overline{1}]$. For $i \in \mathbb{N}$, let $\beta_i = 3+\frac{q_{i-1}-2}{q_i}$. Note that $q_i = \mathcal{F}_{i+1}$ (the i+1-th Fibonacci number), and $\lim_{i\to\infty} \beta_i = 3 + \frac{2}{1+\sqrt{5}} = \frac{5+\sqrt{5}}{2}$. Now:

$$\begin{aligned}\beta_{i+1} - \beta_i &= \frac{1}{q_i q_{i+1}} \left\{ q_i^2 - q_{i+1}q_{i-1} + 2q_{i+1} - 2q_i \right\} \\ &= \frac{1}{q_i q_{i+1}} \left\{ (-1)^{i+1} + 2q_{i+1} - 2q_i \right\} \geq 0.\end{aligned}$$

Thus β_i is increasing, and the largest exponent in w is $\frac{5+\sqrt{5}}{2} = [3,\overline{1}]$. □

References

[1] D. Damanik and D. Lenz, The index of sturmian sequences. *Europ. J. Combinatorics*, 23 (2002), 23–29.

[2] F. Dejean, Sur un théorème de Thue. *J. Combinatorial Th. (A)*, 13 (1972), 90–99.

[3] R.M. Karp, A characterization of the minimum mean cycle in a digraph. *Discrete Math.* 23 (1978), 309–311.

[4] R. Kolpakov, G. Kucherov, and Y. Tarannikov, On repetition-free binary words of minimal density. *Theoret. Comput. Sci.*, 218(1) (1999), 161–175.

[5] G. Kucherov, P. Ochem, and M. Rao, How many square occurrences must a binary sequence contain? *The Electronic Journal of Combinatorics*, 10(1), R12 (2003).

[6] P. Ochem, Letter frequency in infinite repetition-free words. *Theoret. Comput. Sci.*, 380 (2007), 388–392.

[7] A. Thue, Über unendliche Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania*, 7 (1906), 1–22.

[8] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania*, 10 (1912), 1–67.