RAIRO-Theor. Inf. Appl. 56 (2022) 3
https://doi.org/10.1051/ita/2022003

RAIRO - Theoretical Informatics and Applications
www.rairo-ita.org

CRITICAL FACTORISATION IN SQUARE-FREE WORDS

Tero Harju*

Abstract. A position p in a word w is critical if the minimal local period at p is equal to the global period of w. According to the Critical Factorisation Theorem all words of length at least two have a critical point. We study the number $\eta(w)$ of critical points of square-free ternary words w, *i.e.*, words over a three letter alphabet. We show that the sufficiently long square-free words w satisfy $\eta(w) \leq |w| - 5$ where $|w|$ denotes the length of w. Moreover, the bound $|w| - 5$ is reached by infinitely many words. On the other hand, every square-free word w has at least $|w|/4$ critical points, and there is a sequence of these words closing to this bound.

Mathematics Subject Classification. 68R15.

Received August 10, 2021. Accepted January 28, 2022.

1. Introduction

The Critical Factorisation Theorem [2, 4] is one of the gems in combinatorics on words. It states that each word w with $|w| \geq 2$ has a critical point, *i.e.*, a position where the local period $\partial(w, p)$ is equal to the global period $\partial(w)$ of the word. For a word w with a factorisation $w = xy$, $\partial(w, |x|)$ denotes the length of the shortest word u such that of u and x one is a suffix of the other, and of u and y one is a prefix of the other.

In the binary case, say $w \in \{0, 1\}^*$, it was shown in [5] that there are words having only one critical point; *e.g.*, the Fibonacci words of length at least five are such. Also, it was shown there that each binary word w of length $|w| \geq 5$ and period $\partial(w) > |w|/2$ has less than $|w|/2$ critical points.

We shall now study the number of critical points in ternary square-free words. We show that, each sufficiently long square-free word w can have at most $|w| - 5$ critical points, and the bound $|w| - 5$ is obtained by infinitely many square-free w. Also, we prove that a square-free word w has at least $|w|/4$ critical points, and that there is a sequence of square-free words closing to this bound.

2. Preliminaries

For a more extensive introduction to combinatorics on words, including square-freeness and critical factorisation, we refer to Lothaire [6].

For a finite alphabet Σ, let Σ^* denote the monoid of all finite words over Σ under concatenation. The empty word is denoted by ε. Let $w \in \Sigma^*$. The length $|w|$ of w is the number of the occurrences of its letters. If $w = w_1 u w_2$ then u is a *factor* of w. It is a *prefix* if $w_1 = \varepsilon$, and a *suffix* if $w_2 = \varepsilon$. The word w is said to be *bordered* if there exists a nonempty word v, with $v \neq w$, that is both a prefix and a suffix of w.

Keywords and phrases: Critical point, critical factorisation theorem, ternary words, square-free words.

Department of Mathematics and Statistics, University of Turku, Turku, Finland.

* Corresponding author: harju@utu.fi

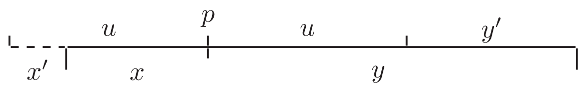

Figure 1. A repetition word u of $w = xy$ having left overflow at position $p = |x|$.

A word $w \in \Sigma^*$ is *square-free* if it has no factors of the form vv for nonempty words v. Axel Thue [8] showed in 1912 that there are square-free words over a ternary alphabet $\Sigma_3 = \{0, 1, 2\}$. One such word is obtained by iterating the following morphism $\tau \colon \Sigma_3^* \to \Sigma_3^*$ on the initial letter 0:

$$\tau(0) = 012, \quad \tau(1) = 02, \quad \tau(2) = 1\,.$$

The iteration ultimately gives an infinite square-free word

$$\mathbf{m} = 012021012102012\cdots$$

By the form of the morphism τ, the word $\mathbf{m}$ does not contain the short words 010, 212 and 01201 as its factors. The infinite word $\mathbf{m}$ is sometimes called a *variation of Thue-Morse word*; see [1].

Lemma 2.1. *Let x be a nonempty factor of a square-free word w. Then "x does not overlap with itself in w", meaning that if $w = ux_1x_2x_3v$ where $x = x_1x_2 = x_2x_3$ and $x_2 \neq \varepsilon$ then $x_1 = \varepsilon = x_3$.*

Proof. Overlapping means, see *e.g.* [6], that x_1 and x_3 are conjugates: $x_1 = rs$, $x_3 = sr$ and $x_2 = (rs)^k r$ for some r, s and $k \geq 0$. But $x_1x_2x_3 = (rs)^{k+2}r$ does contain a square even if $k = 0$. □

3. Critical factorisations

We follow the main notations of [5].

An integer p, with $1 \leq p \leq |w|$, is a *period* of w if for the prefix u of w of length p, w is a prefix of u^n for some n. The *minimal period* of w is denoted by $\partial(w)$. We have that w is unbordered if and only if $\partial(w) = |w|$.

An integer p with $1 \leq p < |w|$ is called a *position* or a *point* in w. It denotes the place after the prefix x of length p: $w = x{\cdot}y$, $|x| = p$. Thus there are $|w| - 1$ positions in w. A nonempty word u is a *repetition word* of w at p if there are words x' and y' (possibly empty) such that $u = x'x$ or $x = x'u$, and $u = yy'$ or $y = uy'$. If here $|u| > |x|$ (resp. $|u| > |y|$) then u is said to have *left overflow* (resp., *right overflow*) at p; see Figure 1.

The length of a repetition word of w at p is called a *local period* at p. The *minimal local period* of w at p is denoted by

$$\partial(w, p) = \min\{q \mid q \text{ a local period of } w \text{ at } p\}.$$

Clearly, the (global) period $\partial(w)$ is a local period at every point, and hence $\partial(w, p) \leq \partial(w)$ for all p. A position p of w is said to be *critical* if $\partial(w, p) = \partial(w)$.

The following result follows from the minimality assumption on $\partial(w, p)$.

Lemma 3.1. *A repetition word u of w at p of length $\partial(w, p)$ is unique and it is unbordered.*

For a word w, we let

$$\eta(w) = \text{the number of critical points of } w.$$

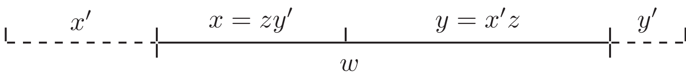

Figure 2. Left and right overflows imply criticality.

The number

$$\frac{\eta(w)}{|w|-1}$$

is called the *density* of the critical points in w.

Example 3.2. Let $w = 0120201202021021021$ be an unbordered word of length 19, *i.e.*, $\partial(w) = |w|$. It is not square-free. The minimal local periods of w are in order of the 18 positions

$$3, 5, 5, 2, 5, 5, 19, 19, 2, 2, 19, 19, 3, 3, 3, 3, 3, 3\,.$$

In this example, $\eta(w) = 4$, and the density of critical points is $4/18 = 0.222\ldots$

The Critical Factorisation Theorem is due to Césari and Vincent [2]. The present form of the theorem was developed by Duval [4]; for the proofs, see also [3], [5] and Chapter 8 in [7].

Theorem 3.3 (Critical Factorisation Theorem)**.** *Every word w of length $|w| \geq 2$ has a critical point. Moreover, there is a critical point p satisfying $p \leq \partial(w)$.*

Lemma 3.4. *Let u be a repetition word of w at p with $|w| \geq 2$ of length $\partial(w,p)$. If u has both left and right overflows at p then p is a critical point.*

Proof. Let $w = xy$ where $u = x'x = yy'$ for nonempty words x', y'; see Figure 2. By symmetry, we may assume that $|x'| \leq |y|$ (otherwise $|y'| \leq |x|$). Therefore $y = x'z$ and $x = zy'$ for some z. Now, $w = xy = zy'x'z$, and hence $|zy'x'|$ is a period of w, *i.e.*, $\partial(w) \leq |zy'x'|$. But $|zy'x'| = |x'zy'| = |u|$ which shows that $\partial(w,p) = |u| = \partial(w)$ implying that p is a critical point. ☐

4. Maximum number of critical points

We notice first that if w is a square-free word with $|w| \geq 2$, then $\partial(w,p) \geq 2$ for all positions p, since $\partial(w,p) = 1$ would imply a factor of the form aa in w.

The next lemma follows from the observation that if a point p of w has neither left nor right overflow, the minimal repetition word u at p supplies a square uu in w.

Lemma 4.1. *A word w with $|w| \geq 2$ is square-free if and only if each repetition word at each position p has left or right overflow, or both.*

Example 4.2. The square-free word $w = 01020120210201021$ of length 17 is unbordered, *i.e.*, $\partial(w) = 17$. It has 9 critical points at the consecutive positions $p = 5, 6, \ldots, 13$. This gives the density number $9/16 \approx 0.56$. For instance, the position $p = 4$ has the minimal repetition word $u = 012021020102$, since u is the shortest factor after the prefix 0102 that ends with 0102. Thus $\partial(w, 4) = 12$.

For a word w, let

$$M(w) = \left\lfloor \frac{|w|+1}{2} \right\rfloor$$

denote the *midpoint* of w. For odd length $|w|$, it is just a choice of the two points nearest to the centre of w.

Lemma 4.3. *For a square-free word $w \in \Sigma_3^*$, the position $M(w)$ is critical.*

Proof. For even $|w|$, the claim is clear from Lemma 4.1.

Suppose then that $|w| = 2k+1$, and let u be the minimal local repetition word of w at $M(w) = k+1$. Suppose u has right but not left overflow. Then $|u| = k+1$, and hence $w = vav$ where $u = va$ for a prefix v and an overflow letter a. But then $\partial(w, k+1) = |u| = \partial(w)$, and the claim follows. □

Theorem 4.4. *The minimal local periods form a unimodular sequence for square-free ternary words $w \in \Sigma_3^*$ with $|w| \geq 2$, i.e.,*

$$\partial(w, p-1) \leq \partial(w, p) \quad \text{for} \quad 2 \leq p \leq M(w),$$
$$\partial(w, p) \leq \partial(w, p-1) \quad \text{for} \quad p-1 \geq M(w)\,.$$

In particular, the critical points p of w form an interval $q_0 \leq p \leq q_1$ for some $q_1 \leq M(w)$ and $q_2 \geq M(w)$.

Proof. Let $2 \leq p \leq M(w)$. The cases for $p \geq M(w)$ follow by considering the reverse of the word w which is also square-free. Let the minimal repetition word of w at p be u, *i.e.*, $|u| = \partial(w, p)$. Since w is square-free and $|u| \geq 2$, u has left overflow. If it also has right overflow then p is critical by Lemma 3.4. Let a be the letter such that $u = va$. Then $|av|$ is a local period at $p-1$ since the position $p-1$ has a repetition word av. (It need not be minimal.) Hence $\partial(w, p-1) \leq \partial(w, p)$.

For the second claim, by Lemma 4.3, w has a critical point p with $p \leq M(n)$ and a critical point $q \geq M(w)$. This proves the claim. □

Example 4.5. Consider the prefix $w = \tau^5(0)$ of the square-free word $\mathbf{m}$, *i.e.*,

$$w = 012021012102012021020121.$$

It is unbordered with $|w[= 24$. The sequence of the 23 minimal local periods is

$$3, 6, 6, 12, 12, 12, 12, 24, \ldots, 24, 14, 14, 6, 2.$$

Thus $\eta(w) = 12$, *i.e.*, just over one half of the positions are critical.

Theorem 4.6. *For each square-free ternary word w of length $|w| \geq 26$, we have $\eta(w) \leq |w| - 5$.*

Proof. Let $w \in \Sigma_3^*$ be a square-free ternary word of length $n \geq 26$. We show that w has at least four non-critical points among the $n-1$ positions. The points 1 and $n-1$ are always non-critical, since every letter of Σ_3 occurs in every factor of length four. Let us then assume that w has exactly three non-critical points. Therefore at least one of the positions 2 or $n-2$ is critical. Without restriction, we can say that $p = 2$ is critical.

Without restriction we may assume that 01 is a prefix of w. It can be checked that there are no such square-free words of length 15 where 01 occurs only as a prefix. After inspection, we find that the only such word of length 14 is $v = 01210212021020$.

Thus since $|w| \geq 15$, we have $w = 01x01y$ for some words x and y such that 01 does not occur in x. Now, the word $x01$ is a repetition of w at position 2 and hence, by the criticality of $p = 2$, we have $|x01| = \partial(w)$. Because w is square-free, y must be a proper prefix of x. The word x does not have any occurrences of 01 and it cannot end in the letter 0. Thus $|01x| \leq 13$. But now $n \leq 25$; a contradiction. □

Example 4.7. In contrast to Theorem 4.6, the word $w = 01210212021020121021202$ of length 23 with $\partial(w) = 13$ has only three non-critical points, $p = 1, 2, 22$.

The upper bound on the critical points is optimal:

Theorem 4.8. *There are arbitrarily long square-free words $w \in \Sigma_3^*$ with $\eta(w) = |w| - 5$.*

TABLE 1. Local periods of non-critical points.

p	$\partial(w,p)$	Rep. word
1	2	10
2	4	0201
$\lvert w\rvert - 2$	4	1202
$\lvert w\rvert - 1$	2	21

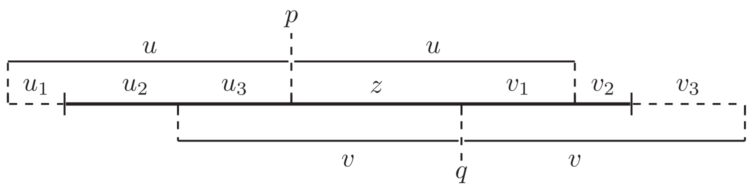

FIGURE 3. The repetition words u and v for non-critical points p and q.

Proof. We rely on the infinite square-free word $\mathbf{m}$ that is a fixed point of the morphism τ. Consider the factors of $\mathbf{m}$ of the form $\beta = 10201\alpha12021$. For our purpose, it suffices to choose the words β that start after the position 9 of $\mathbf{m}$, *i.e.*, just after the prefix 012021012. There are infinitely many words β since the suffix 12021 is a factor of $\tau^2(0)$.

For fixed middle word α, consider $w = 0\beta2 = 010201\alpha120212$ that begins and ends in the 'forbidden' words 010 and 212 that do not occur in $\mathbf{m}$. It is, clearly, square-free and unbordered. Each point p with $2 < p < |w| - 2$ is critical, since the minimal repetition word at p must have both left and right overflow in order to leap over a factor 010 or 212; see Lemma 3.4. Table 1 lists the local periods and the minimal repetition words for the remaining four (non-critical) points. □

5. Minimum number of critical points

We now turn to the minimality problem of critical points in square-free words.

Theorem 5.1. *For each square-free word $w \in \Sigma_3^*$ with $|w| \geq 2$, we have $\eta(w) \geq |w|/4$.*

Proof. Let $w \in \Sigma_3^*$ be a square-free word of length $|w| = n$. We remind that, by Lemma 4.3, the middle point $M(w)$ is always critical in w. We show that the distance between two non-critical points on the opposite sides of the middle point is at least $n/4$. The claim then follows from Theorem 4.4.

Assume, contrary to the claim, that p and q are non-critical points such that

$$p < n/2 < q \quad \text{and} \quad q - p < n/4\,. \tag{5.1}$$

Let u and v be the minimal repetition words at p and q, respectively. Consequently, the word u has left overflow, and v has right overflow. Observe that $p > n/4$ and $q < 3n/4$. From $p > n/4$ it follows that $|u| > n/4$. Similarly $|v| > n/4$ and $q - |v| < n/2$. Since $q - p < n/4$, we have $p + |u| \geq q$, *i.e.*, the second occurrence of u reaches over the position q. Similarly the first occurrence of v starts before the position p; see Figure 3, where $|z| = q - p$.

We now rely on the notations of the factors in Figure 3.

The words u_3 and v_1 are both prefixes of v and suffixes of u. If $|v_1| > |u_3|$ then, as prefixes of v, we have $v_1 = u_3x$ for some nonempty x. But then x is a border of u since u_3 cannot overlap with itself at the end of the first occurrence of u; a contradiction. Similarly, if $|u_3| > |v_1|$ then, as suffixes of u, we have $u_3 = xv_1$ for some nonempty x yielding that x is is a border of v; a contradiction. Therefore $v_1 = u_3$. In this case $z = u_1u_2 = v_2v_3$, and

$$w = u_2v_1zv_1v_2 = u_2v_1u_1u_2v_1v_2\,.$$

Since $u_1u_2 = v_2v_3$, one of u_1 or v_2 is a prefix of the other. To avoid $(u_2v_1u_1)^2$ in w, the word v_2 must be a proper prefix of u_1. But now $\partial(w) \le |u_2v_1u_1| = |u| = \partial(w,p)$ contradicting the assumption that p was not critical. This proves the claim. □

For the existence part of the next theorem, we take a quick technical analysis of the prefixes of the word $\mathbf{m}$. An induction argument gives $|\tau^n(0)| = 3\cdot 2^{n-1}$, $|\tau^n(1)| = 2^n$ and $|\tau^n(2)| = 2^{n-1}$. For instance,

$$|\tau^{n+1}(0)| = |\tau^n(012)| = 3{\cdot}2^{n-1} + 2^n + 2^{n-1} = 3{\cdot}2^n\,.$$

Define the words $\mathbf{m}_n$, for $n \ge 1$, as follows

$$\mathbf{m}_n = \tau^{2n-1}(0)\tau^{2n-3}(0)\cdots\tau^3(0)\tau(0)\,. \tag{5.2}$$

We show that $\mathbf{m}_n0$ is a prefix of $\mathbf{m}$ of length 4^n. First $\mathbf{m}_10 = 0120 = \tau(0)0$ is a prefix of $\mathbf{m}$. Inductively, we have

$$\tau^2(\mathbf{m}_n0) = \tau^{2n+1}(0)\tau^{2n-1}(0)\cdots\tau^3(0)\tau^2(0) = \mathbf{m}_{n+1}0{\cdot}21\,.$$

and hence also $\mathbf{m}_{n+1}0$ is a prefix of $\mathbf{m}$.

For the length of $\mathbf{m}_n$, we obtain

$$|\mathbf{m}_n| = \sum_{i=1}^{n} 3\cdot 2^{2(n-i)} = 3\sum_{i=1}^{n} 4^{n-i} = 4^n - 1.$$

As a prefix of $\mathbf{m}$, the word $\mathbf{m}_n$ is square-free.

Theorem 5.2. *For all real numbers $\delta > 0$, there exists a square-free ternary word $w = w(\delta)$ the density of which satisfies*

$$0.25 < \frac{\eta(w)}{|w|} < 0.25 + \delta\,.$$

Proof. For any square-free word $x \in \Sigma_3^*$, let

$$w_x = 0x02x10x02x0\,. \tag{5.3}$$

Suppose first that w_x is square-free, and thus that x does not overlap with itself in w_x. The suffix $2x0$ of w_x does not occur elsewhere in w_x, and hence the point $3|x|+6$ is critical, since it must have both overflows. It is the rightmost critical point. Indeed, $\partial(w_x, 3|x|+7) = |x|+2$. For the point $2|x|+3$, the minimal repetition word is $10x02x$ of length $2|x|+4 < \partial(w)$ since $\partial(w) > 3|x|+7$. By Lemma 4.3, the middle point is critical, and hence the position $2|x|+4$ is the leftmost critical point. It follows that w_x has $(3|x|+7)-(2|x|+4) = |x|+3$ critical points. Thus

$$\frac{\eta(w)}{|w_x|} = \frac{|x|+3}{4|x|+8} = 0.25 + \frac{1}{|w_x|}\,,$$

which has the limit 0.25 as $|x| \to \infty$.

It remains to show that there are arbitrarily long square-free words x for which w_x is square-free. Again, we lean on the word $\mathbf{m}$. We consider the words w_{x_n} where

$$x_n = 120102\,\mathbf{m}_n\,.$$

We have

$$w_{x_n} = 0\cdot 120102\mathbf{m}_n \cdot 02\cdot 120102\mathbf{m}_n\cdot 10\cdot 120102\mathbf{m}_n \cdot 02\cdot 120102\mathbf{m}_n\cdot 0$$

Since 010 and 212 do not occur in $\mathbf{m}$, both 010 and 212 would have to be aligned in any square uu of w_{x_n}, which is not possible by the 'markers' 02, 1 and 0 dividing the word. Also, since $\mathbf{m}_n$ has a border $\tau(0)$, one easily checks that there are no short squares uu in w_x for $|u| \le 4$. Hence a possible square must be inside one of the words (a) $102\mathbf{m}_n021$, (b) $102\mathbf{m}_n101201$, or (c) $102\mathbf{m}_n0$. We consider these cases separately. Recall that $\mathbf{m}_1 = 012 = \tau(0)$. Also, since $\mathbf{m}$ is a fixed point of the morphism τ, whenever v is a factor of $\mathbf{m}$, so is $\tau(v)$.

(a) Let $\alpha_n = 102\mathbf{m}_n021$. The word $\alpha_1 = 102012021$ occurs in $\mathbf{m}$ after position 9. We prove by induction that each α_n is a factor of $\mathbf{m}$, and thus they are square-free. Suppose, using (5.2), that

$$\alpha_i = 102\mathbf{m}_i021 = 102\tau^{2i-1}(0)\cdots\tau(0)021$$

is a factor of $\mathbf{m}$. Then

$$\tau(\alpha_i) = 0201\cdot 21\tau^{2i}(0)\cdots\tau^2(0)012\cdot 021$$

where the indicated factor will be denoted by $z = 21\tau^{2i}(0)\cdots 012$. By mapping with τ, we obtain

$$\tau(z) = 102\tau^{2i+1}(0)\cdots\tau^3(0)\tau(0)021 = 102\mathbf{m}_{i+1}021 = \alpha_{i+1}.$$

Hence α_n is a factor of $\mathbf{m}$ for all n.

(b) We employ in this case the same techniques as in (a) except that we need to eliminate the last letter 1 of the word. In order for $102\mathbf{m}_n101201$ to have a square uu, the former occurrence of u in the square must be a factor of $102\mathbf{m}$. However, $\mathbf{m}$ does not have a factor 01201 since it would have to be part of the square 012012. Therefore we can, and must, choose $\beta_n = 102\mathbf{m}_n101202$.
The first occurrence of $\beta_1 = 102\mathbf{m}_1101202 = 102\tau(0)101202$ in $\mathbf{m}$ starts after position 17. We proceed inductively as in case (a). Suppose that

$$\beta_i = 102\mathbf{m}_i101202 = 102\tau^{2i-1}(0)\cdots\tau(0)101202$$

is a factor of $\mathbf{m}$. Mapping by τ gives

$$\tau(\beta_i) = 0201\cdot 21\tau^{2i}(0)\cdots\tau^2(0)0201\cdot 20210121\,,$$

where the indicated portion $z = 21\tau^{2i}(0)\cdots\tau^2(0)0201$ gives

$$\tau(z) = 102\tau^{2i+1}(0)\cdots\tau^3(0)\tau(0)101202 = 102\mathbf{m}_{i+1}101202 = \beta_{i+1}\,.$$

Hence β_n is a factor of $\mathbf{m}$, and thus square-free, for all n.

(c) The word $102\mathbf{m}_n0$ is a factor of α_n and thus square-free. This proves the claim.

□

The chosen words $x_n = 120102\,\mathbf{m}_n$ are not the only ones that give a square-free word w_x.

Problem 5.3. Does there exist, for all sufficiently large n, a word x of length n such that w_x is square-free?

Problem 5.4. Does there exist a word w such that $\eta(w) = |w|/4$?

Acknowledgements. The author thanks for the kind referees for their comments that clairifed the proofs of the article.

References

[1] F. Blanchet-Sadri, J.D. Currie, N. Rampersad and N. Fox, Abelian complexity of fixed point of morphism $0 \mapsto 012, 1 \mapsto 02, 2 \mapsto 1$. *Integers* **14** (2014) a11.
[2] Y. Césari and M. Vincent, Une caractérisation des mots périodiques. *C. R. Acad. Sci. Paris* **286(A)** (1978) 1175–1177.
[3] M. Crochemore and D. Perrin, Two-way string-matching. *J. Assoc. Comput. Mach.* **38** (1991) 651–675.
[4] J.-P. Duval, Périodes et répétitions des mots du monoïde libre. *Theoret. Comput. Sci.* **9** (1979) 17–26.
[5] T. Harju and D. Nowotka, Density of critical factorizations. *Theor. Inform. Appl.* **36** (2002) 315–327.
[6] M. Lothaire, Combinatorics on words. Vol. 17 of *Encyclopedia of Mathematics*. Addison-Wesley, Reading, Massachusetts (1983).
[7] M. Lothaire, Algebraic Combinatorics on Words. Cambridge University Press, Cambridge, United Kingdom (2002).
[8] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske Vid. Selsk. Skr., I Mat.-nat. Kl. Christiania* **1** (1912) 1–67.