On canonical number systems

Shigeki Akiyama* and Attila Pethő†

Abstract. Let \( P(x) = p_dx^d + \ldots + p_0 \in \mathbb{Z}[x] \) be such that \( d \geq 1, p_d = 1, p_0 \geq 2 \) and \( \mathcal{N} = \{0, 1, \ldots, p_0 - 1\} \). We are proving in this note a new criterion for the pair \( \{P(x), \mathcal{N}\} \) to be a canonical number system. This enables us to prove that if \( p_2, \ldots, p_{d-1}, \sum_{i=1}^{d} p_i \geq 0 \) and \( p_0 > 2 \sum_{i=1}^{d} |p_i| \), then \( \{P(x), \mathcal{N}\} \) is a canonical number system.

Key words and phrases: canonical number system, radix representation, algebraic number field, height.

1 Introduction

Let \( P(x) = p_dx^d + \ldots + p_0 \in \mathbb{Z}[x] \) be such that \( d \geq 1 \) and \( p_d = 1 \). Let \( R \) denote the quotient ring \( \mathbb{Z}[x]/P(x)\mathbb{Z}[x] \). Then all \( \alpha \in R \) can be represented in the form

\[
\alpha = a_0 + a_1 x + \ldots + a_{d-1} x^{d-1}
\]

with \( a_i \in \mathbb{Z}, i = 0, \ldots, d-1 \).

The pair \( \{P(x), \mathcal{N}\} \) with \( \mathcal{N} = \{0, 1, \ldots, |p_0| - 1\} \) is called canonical number system, CNS, if every \( \alpha \in R, \alpha \neq 0 \) can be written uniquely in the form

\[
\alpha = \sum_{j=0}^{\ell(\alpha)} a_j x^j,
\]

where \( a_j \in \mathcal{N}, j = 0, \ldots, \ell(\alpha), a_{\ell(\alpha)} \neq 0 \).

If \( P(x) \) is irreducible, then let \( \gamma \) denote one of its zeros. In this case \( \mathbb{Z}[x]/P(x)\mathbb{Z}[x] \) is isomorphic to \( \mathbb{Z}[\gamma] \), the minimal ring generated by \( \gamma \) and \( \mathbb{Z} \), hence we may replace \( x \) by \( \gamma \) in the above expansions. Moreover \( \mathcal{N} \) forms a complete representative system mod \( \gamma \) in \( \mathbb{Z}[\gamma] \). We simplify in this case the notation \( \{P(x), \mathcal{N}\} \) to \( \{\gamma, \mathcal{N}\} \).

Extending the results of [7] and [3], I. Kátaí and B. Kovács and independently W.J. Gilbert [2] classified all quadratic CNS, provided the corresponding \( P(x) \) is irreducible. B. Kovács [8] proved that in any algebraic number field there exists an element \( \gamma \) such that \( \{\gamma, \mathcal{N}\} \) is a CNS. J. Thuswaldner [13] gave in the quadratic and K. Scheicher [12]

†Research supported in part by the Hungarian Foundation for Scientific Research, Grant No. 25157/98.
*We need a slight explanation of their results, since their definition of canonical number system is more restricted than ours. In fact, they assumed still more that \( \mathbb{Z}[\gamma] \) coincides with the integer ring of \( \mathbb{Q}(\gamma) \), the field generated by \( \gamma \) over the field of rational numbers.
in the general case a new proof of the above theorems based on automaton theory. B. Kovács [8] proved further that if \( p_d \leq p_{d-1} \leq p_{d-2} \leq \ldots \leq p_0, p_0 \geq 2 \), and if \( P(x) \) is irreducible and \( \gamma \) is a zero of \( P(x) \) then \( \{ \gamma, N \} \) is a CNS in \( \mathbb{Z}[\gamma] \). In [9] B. Kovács and A. Pethő gave also a characterization of those irreducible polynomials \( P(x) \), whose zeros are bases of CNS.

Interesting connections between CNS and fractal tilings of the Euclidean space were discussed by several mathematicians. D.E. Knuth [7] seems to be the first discoverer of this phenomenon in the case \( x = -1 + \sqrt{-1} \). For the recent results on this topic, the reader can consult [4] or [1] and their references.

The concept of CNS for irreducible polynomials was generalized to arbitrary polynomials with leading coefficient one by the second author [11]. He extended most of the results of [8] and [9] and proved among others that if \( \{ P(x), N \} \) is a CNS then all real zeroes of \( P(x) \) are less than \(-1\) and the absolute value of all the complex roots are larger than 1. This implies that if \( \{ P(x), N \} \) is a CNS then \( p_0 > 0 \), which we will assume throughout this paper.\(^2\)

The aim of the present paper is to give a new characterization of CNS provided \( p_0 \) is large enough. It enables us to prove for a large class of polynomials that their zeros together with the corresponding set \( N \) yield a CNS. Unfortunately our criterion in Theorem 1 cannot be adapted to polynomials with small \( p_0 \), but it suggests us that the characterization problem of CNS does not depend on the structure of the corresponding field, such as fundamental units, ramifications or discriminants, but only on the coefficients of its defining polynomials.

### 2 Notations and results

For a polynomial \( P(x) = p_dx^d + \ldots + p_0 \in \mathbb{Z}[x] \), let

\[
L(P) = \sum_{i=1}^{d} |p_i|,
\]

which we call the length of \( P \). Every \( \alpha \in R = \mathbb{Z}[x]/P(x)\mathbb{Z}[x] \) has a unique representation in the form

\[
\alpha = \sum_{j=0}^{d-1} a_j x^j.
\]

Put \( q = \lfloor \frac{a_0}{p_0} \rfloor \), where \( \lfloor \cdot \rfloor \) denotes the integer part function. Let us define the map \( T : R \to R \) by

\[
T(\alpha) = \sum_{j=0}^{d-1} (a_{j+1} - qp_{j+1})x^j,
\]

where \( a_d = 0 \). Putting

\[
T^{(0)}(\alpha) = \alpha \quad \text{and} \quad T^{(i+1)}(\alpha) = T(T^{(i)}(\alpha))
\]

\(^2\)In Theorem 6.1 of [11] it is assumed that \( g(t) \) is square-free, but this assumption is necessary only for the proof of (iii).
we define the iterates of $T$. As $T^{(i)}(\alpha) \in R$ for all non-negative integers $i$, and $\alpha \in R$, the element $T^{(i)}(\alpha)$ can be represented with integer coefficients in the basis $1, x, \ldots, x^{d-1}$. The coefficients of this representation will be denoted by $T^{(i)}_j(\alpha), i \geq 0, 0 \leq j \leq d - 1$. It is sometimes convenient to extend this definition by putting $T^{(i)}_j(\alpha) = 0$ for $j \geq d$. This map $T$ obviously describes the algorithm to express any $\alpha \in R$ in a form (1) since we have

$$\alpha = \sum_{j=0}^{\ell(\alpha)} \left\lfloor \frac{T^{(j)}_0(\alpha)}{p_0} \right\rfloor x^j,$$

when $\{P(x), N\}$ is a CNS. With this notation we have

$$\alpha = \sum_{j=0}^{d-1} T^{(0)}_j(\alpha)x^j,$$

and

$$T^{(i)}(\alpha) = \sum_{j=0}^{d-1} T^{(i)}_j(\alpha)x^j,$$

$$= \sum_{j=0}^{d-1} (T^{(i-1)}_j(\alpha) - q_{i-1}p_{j+1})x^j,$$

where $q_{i-1} = \left\lfloor \frac{T^{(i-1)}_0(\alpha)}{p_0} \right\rfloor$ for $i \geq 1$.

After this preparation we are in the position to formulate our results. The first assertion is a new characterization of CNS provided $p_0 > L(P)$. By Lemma 1 in §3, the roots of such a $P$ have moduli greater than 1, which is a necessary condition for a CNS. So we are interested in such a class of polynomials. The spirit of Theorem 1 below and Theorems 3 of [9] and 6.1 of [11] is the same: it is proved that $\{P(x), N\}$ is a CNS in $R$ if and only if every element of bounded size of $R$ is representable in $\{P(x), N\}$. The difference is in the choice of the size. Whereas Kovács and Pethő used the height, max $\{|T^{(0)}_j(\alpha)|, 0 \leq j \leq d - 1\}$, we use the weight, defined by (13) in §4.

**Theorem 1** Let $M$ be a positive integer. Assume that $p_0 \geq (1 + 1/M)L(P)$, if $p_i \neq 0$ for $i = 1, \ldots, d - 1$, and assume that $p_0 > (1 + 1/M)L(P)$ otherwise. The pair $\{P(x), N\}$ is a CNS in $R$ if and only if each of the following elements $\alpha \in R$ has a representation in $\{P(x), N\}$:

$$\alpha = \sum_{i=0}^{d-1} \left( \sum_{j=i}^{d-1} \varepsilon_{j}p_{d+i-j} \right) x^i,$$

where $\varepsilon_j \in [1 - M, M] \cap \mathbb{Z}$ for $0 \leq j \leq d - 1$.

Our algorithm is easier and more suitable for hand calculation than the ones in [9] and [11], since we do not need any information on the roots of $P$. We need only to check whether $(2M)^d$ elements have representations in $\{P(x), N\}$ or not. Running time estimates for the Kovács and Pethő algorithm of [9] is difficult, since it depends on the distribution of the roots of $P$. But in many cases, our method is very rapid when $p_0$ or $d$ is large.
Example 1 We compare for three CNS polynomials the number of elements needed to be checked for representability in \( \{ P(x), N \} \) by our algorithm and by the algorithm of Kovács and Pethő.

Case \( x^3 + x^2 + 5 \):
(Our algorithm) 8 elements (M=1),
(Kovács and Pethő algorithm) 89 elements.

Case \( x^3 + 2x^2 - x + 7 \):
(Our algorithm) 64 elements (M=2),
(Kovács and Pethő algorithm) 123 elements.

Case \( x^4 + x^3 - x^2 + x + 8 \):
(Our algorithm) 16 elements (M=1),
(Kovács and Pethő algorithm) 1427 elements.

Using Theorem 1 we are able to prove that a wide class of polynomials correspond to a CNS. Similar results were proven in [8] and in [11]. Using the idea of B. Kovács [8] it was proved in [11] that if \( 0 < p_{d-1} \leq \ldots \leq p_0, p_0 \geq 2 \) then \( \{ P(x), N \} \) is a CNS. We however do not assume the monotonicity of the sequence of the coefficients. Moreover \( p_1 \) is allowed to be negative.

Theorem 2 Assume that \( p_2, \ldots, p_{d-1}, \sum_{i=1}^{d} p_i \geq 0 \) and \( p_0 > 2 \sum_{i=1}^{d} |p_i| \) Then \( \{ P(x), N \} \) is a CNS in \( R \). The last inequality can be replaced by \( p_0 \geq 2 \sum_{i=1}^{d} |p_i| \) when all \( p_i \neq 0 \).

Note that the conditions \( p_2, \ldots, p_{d-1}, \sum_{i=1}^{d} p_i \geq 0 \) are necessary if \( d = 3 \) by Proposition 1 in §3. So Theorem 2 gives us a characterization of all cubic CNS provided \( p_0 > 2L(P) \). Generally, the inequality \( \sum_{i=1}^{d} p_i \geq 0 \) is by Lemma 4 below necessary for \( \{ P(x), N \} \) to be a CNS. On the other hand the following examples show that the inequalities \( p_2, \ldots, p_{d-1} \geq 0 \) are not necessary if \( d \geq 4 \).

Example 2 In fact, we can show that the roots of each polynomials
\[
x^4 + 2x^3 - x^2 - x + 5, \ x^4 - x^3 + 2x^2 - 2x + 3, \ x^5 + x^4 + x^3 - x^2 - x + 4
\]
form a CNS by the criterion of [9].

We are also able to prove that \( p_{d-1} \) cannot be too small. More precisely the following theorem is true.

Theorem 3 If \( p_0 \geq \sum_{i=1}^{d} |p_i| \) and \( \{ P(x), N \} \) is a CNS then \( p_l + \sum_{j=l+1}^{d} |p_j| \geq 0 \) holds for all \( l \geq 0 \). In particular \( p_{d-1} \geq -1 \).

The characterization of higher dimensional CNS where \( p_0 \) is large is an interesting problem left to the reader. Numerical evidence supports the following:

Conjecture 1 Assume that \( p_2, \ldots, p_{d-1}, \sum_{i=1}^{d} p_i \geq 0 \) and \( p_0 > \sum_{i=1}^{d} |p_i| \). Then \( \{ P(x), N \} \) is a CNS.

Conjecture 2 The pair \( \{ P(x), N \} \) is a CNS in \( R \) if and only if all \( \alpha \in R \) of the form (4) with \( \varepsilon_j \in \{-1, 0, 1\} \), \( 0 \leq j \leq d - 1 \), have a representation in \( \{ P(x), N \} \).

This conjecture is best possible in the sense that that we can not remove \(-1\) or \(1\) from the allowed set of \( \varepsilon_j \). Considering polynomial \( P(x) = x^3 + 4x^2 - 2x + 6 \), the element \(-x^2 - 5x - 1\) does not have a representation in \( \{ P(x), \{0, 1, 2, 3, 4, 5\} \} \).
3 Auxiliary results

Several general results of CNS are shown in this section. Some of them are used in the proof of our Theorems.

**Lemma 1** If \( p_0 > L(P) \) then each root of \( P \) has modulus greater than \( 1 \).

**Proof:** Assume that \( \gamma \) is a root of \( P \) with \(|\gamma| \leq 1\). Then we have
\[
\left| \sum_{i=1}^{d} p_i \gamma^i \right| \leq L(P) < p_0,
\]
which is absurd. \( \Box \)

In the sequel we will put \( T_j^{(i)}(\alpha) = 0 \) for \( j > d - 1 \) and \( p_j = 0 \) for \( j > d \).

**Lemma 2** Let \( \alpha \in R \) and \( i, j, k \) be non-negative integers such that \( k \geq i \). Let \( q_k = \left\lfloor \frac{T^{(k)}(\alpha)}{p_0} \right\rfloor \). Then
\[
T_j^{(k)}(\alpha) = T_{j+i}^{(k-i)}(\alpha) - \sum_{\ell=1}^{i} q_k - \ell p_{j+\ell}, \tag{5}
\]
\[
\alpha = \sum_{\ell=0}^{k-1} (T_0^{(\ell)}(\alpha) - q_k p_0) x^\ell + x^k T^{(k)}(\alpha). \tag{6}
\]

**Proof:** Identity (5) is obviously true if \( i = 0 \). Assume that it is true for an \( i \) such that \( 0 \leq i < k \). We have
\[
T_{j+i}^{(k-i)}(\alpha) = T_{j+i+1}^{(k-i-1)}(\alpha) - q_{k-i-1} p_{j+i+1}
\]
by (3). Inserting this into (5) we obtain at once the stated identity for \( i + 1 \).

Identity (6) is obviously true for \( k = 0 \). Assume that it is true for \( k - 1 \geq 0 \). Using that \( P(x) = 0 \) in \( R \) we have
\[
T^{(k-1)}(\alpha) = \sum_{j=0}^{d-1} T_j^{(k-1)}(\alpha) x^j
\]
\[
= \sum_{j=0}^{d-1} T_j^{(k-1)}(\alpha) x^j - q_{k-1} \sum_{j=0}^{d} p_j x^j
\]
\[
= \sum_{j=0}^{d} (T_j^{(k-1)}(\alpha) - q_{k-1} p_j) x^j
\]
\[
= (T_0^{(k-1)}(\alpha) - q_{k-1} p_0) + x T^{(k)}(\alpha).
\]
Considering (6) for \( k - 1 \) and using the last identity we obtain
\[
\alpha = \sum_{\ell=0}^{k-2} (T_0^{(\ell)}(\alpha) - q_\ell p_0) x^\ell + x^{k-1} T^{(k-1)}(\alpha)
\]
\[
= \sum_{\ell=0}^{k-2} (T_0^{(\ell)}(\alpha) - q_\ell p_0) x^\ell + x^{k-1} ((T_0^{(k-1)}(\alpha) - q_{k-1} p_0) + x T^{(k)}(\alpha))
\]
\[
= \sum_{\ell=0}^{k-1} (T_0^{(\ell)}(\alpha) - q_\ell p_0) x^\ell + x^k T^{(k)}(\alpha).
\]
Thus (6) is proved for all $k \geq 0$. □

Lemma 3 The element $\alpha \in R$ is representable in $\{P(x), \mathcal{N}\}$ if and only if there exists a $k \geq 0$ for which $T^{(k)}(\alpha) = 0$.

**Proof:** The condition is sufficient, because if $\alpha$ is representable in $\{P(x), \mathcal{N}\}$ then we can take $k = \ell(\alpha)$.

To prove the necessity, assume that there exists a $k \geq 0$ for which $T^{(k)}(\alpha) = 0$. Then

$$\alpha = \sum_{\ell=0}^{k-1} (T^{(\ell)}(\alpha) - q_0 p_0) x^\ell$$

by Lemma 2, and since $T^{(\ell)}(\alpha) - q_0 p_0 \in \mathcal{N}$ this is a representation of $\alpha$ in $\{P(x), \mathcal{N}\}$. □

Lemma 4 If $\{P(x), \mathcal{N}\}$ is a CNS, then $\sum_{i=1}^{d} p_i \geq 0$.

**Proof:** By the results of [11], stated in the introduction, we have $P(1) = \sum_{i=0}^{d} p_i > 0$, since otherwise $P(x)$ would have a real root greater or equal to 1.

Assume that $\sum_{i=1}^{d} p_i < 0$. Then $P(1) = p_0 + \sum_{i=1}^{d} p_i < p_0$, i.e., $P(1) \in \mathcal{N}$. Let

$$\alpha = \sum_{i=0}^{d-1} \sum_{j=i}^{d-1} p_{d+i-j} x^i.$$ 

Then $T^{(0)}(\alpha) = \sum_{i=1}^{d} p_i$, hence $-p_0 < T^{(0)}(\alpha) < 0$, which implies $q = \lfloor T^{(0)}(\alpha)/p_0 \rfloor = -1$. Thus $T(\alpha) = \alpha \neq 0$ and $\alpha$ does not have a representation in $\{P(x), \mathcal{N}\}$ by Lemma 3. □

We wish to summarize some inequalities satisfied by a cubic CNS. These were proved by W.J. Gilbert [2]. For the sake of completeness we are given here a slightly different proof.

**Proposition 1** Let $\{P(x), \mathcal{N}\}$ be a cubic CNS. Then we have the following inequalities:

\begin{align*}
1 + p_1 + p_2 & \geq 0, \quad (7) \\
p_0 + p_2 & > 1 + p_1, \quad (8) \\
p_0 p_2 + 1 & < p_0^2 + p_1, \quad (9) \\
p_2 & \leq p_0 + 1, \quad (10) \\
p_1 & < 2p_0, \quad (11) \\
p_2 & \geq 0. \quad (12)
\end{align*}

**Proof:** Lemma 4 implies (7). By a similar argument to Lemma 4, we see $P(-1) > 0$. This shows (8). If $P(-p_0) \geq 0$ then there exists a real root less than or equal to $-p_0$. Since $p_0$ is the product of the three roots of $P(x)$, this implies that there exists a root whose modulus is less than or equal to 1. This shows $P(-p_0) < 0$ which is (9).

Let $\gamma_i$ ($i = 1, 2, 3$) be the roots of $P(x)$. Noting $xy + 1 > x + y$ for $x, y > 1$, we see

$$|p_2| = |\gamma_1 + \gamma_2 + \gamma_3| < |\gamma_1 \gamma_2| + |\gamma_3| + 1 < |\gamma_1 \gamma_2 \gamma_3| + 2 = p_0 + 2.$$
Thus we have (10). Using (8) we have (11).

Finally we want to show (12). By (7), if \( p_2 < 0 \) then \( p_1 \geq 0 \). Let \( w = x + p_2 \). By (8), we have \( p_2 > -p_0 \). Thus

\[
T(w) = x^2 + p_2x + p_1 + 1.
\]

Since \( 1 \leq p_1 + 1 \leq p_0 + 2 < p_0 \), we see \( p_1 + 1 \in \mathcal{N} \). Thus we have

\[
T^{(2)}(w) = x + p_2 = w.
\]

Hence \( T^{(2k)}(w) = w \) and \( T^{(2k+1)}(w) = x^2 + p_2x + p_1 + 1 \) for all \( k \geq 0 \), i.e., \( T^{(j)}(w) \neq 0 \) holds for all \( j \geq 0 \). By Lemma 4 \( w \) is not representable in \( \{ P(x), \mathcal{N} \} \). This completes the proof of the proposition. \( \square \)

We can find a CNS with \( p_{d-1} = -1 \) when \( d = 2 \) or \( d \geq 4 \).

4 Proof of Theorem 1.

Proof:

Let \( \eta \) be a positive number and put \( p_i^* = p_i \) if \( p_i \neq 0 \) and \( p_i^* = \eta \) otherwise. Taking a small \( \eta \), we may assume

\[
p_0 \geq (1 + 1/M) \sum_{i=1}^{d} |p_i^*|.
\]

Define the weight of \( \alpha \in R \) by

\[
\mathcal{W}(\alpha) = \max \left\{ M, \max_{i=0,1,\ldots,d-1} \frac{|T_i^{(0)}(\alpha)|}{\sum_{k=i+1}^{d} |p_k^*|} \right\}.
\]

(13)

Obviously the weight of \( \alpha \) takes discrete values. We have

\[
|T_i^{(0)}(\alpha)| \leq \mathcal{W}(\alpha) \sum_{k=i+1}^{d} |p_k^*|,
\]

by definition. Remark that this inequality is also valid when \( i = d \).

First we show that \( \mathcal{W}(T(\alpha)) \leq \mathcal{W}(\alpha) \) for any \( \alpha \in R \). If \( |T_0^{(0)}(\alpha)/p_0| \geq M \) then we have

\[
\left| \frac{T_0^{(0)}(\alpha)}{p_0} \right| < \left| \frac{T_0^{(0)}(\alpha)}{p_0} \right| + 1 \leq \left( 1 + \frac{1}{M} \right) \left| \frac{T_0^{(0)}(\alpha)}{p_0} \right| \leq \frac{|T_0^{(0)}(\alpha)|}{\sum_{k=1}^{d} |p_k^*|} \leq \mathcal{W}(\alpha).
\]

If \( |T_0^{(0)}(\alpha)/p_0| < M \), we see \( |T_0^{(0)}(\alpha)/p_0| \in [-M, M - 1] \cap \mathbb{Z} \). (Here we used the fact that \( M \) is a positive integer.) This shows \( |T_0^{(0)}(\alpha)/p_0| \leq M \leq \mathcal{W}(\alpha) \). So we have shown

\[
\left| \frac{T_0^{(0)}(\alpha)}{p_0} \right| \leq \mathcal{W}(\alpha)
\]

7
for any $\alpha$. We note that the equality holds only when $q_0 = [T_0^{(0)}(\alpha)/p_0] = -M$. This fact will be used later. Recall the relation:

$$T(\alpha) = \sum_{i=0}^{d-1} (T_{i+1}(\alpha) - q_0 p_{i+1}) x^i$$

with $q_0 = [T_0^{(0)}(\alpha)/p_0]$. So we have

$$\left| \frac{T_{i+1}(\alpha) - q_0 p_{i+1}}{\sum_{k=i+1}^d |p_k^*|} \right| \leq \frac{\mathcal{W}(\alpha) \sum_{k=i+2}^d |p_k^*| + \mathcal{W}(\alpha)|p_{i+1}|}{\sum_{k=i+1}^d |p_k^*|} \leq \mathcal{W}(\alpha),$$

which shows $\mathcal{W}(T(\alpha)) \leq \mathcal{W}(\alpha)$.

If $\{P(x), \mathcal{N}\}$ is a CNS then every element of form (4) must have a representation in $\{P(x), \mathcal{N}\}$. Assume that $\{P(x), \mathcal{N}\}$ is not a CNS. Then there exist elements of $R$ which do not have any representation in $\{P(x), \mathcal{N}\}$. Let $\kappa \in R$ be such an element of minimum weight. Our purpose is to prove that there exists some $m$ such that $T^{(m)}(\kappa)$ must have the form (4). First we show $\mathcal{W}(\kappa) = M$. So assume that $\mathcal{W}(\kappa) > M$. Then we have

$$\mathcal{W}(\kappa) = \max_{i=0,1,\ldots,d-1} \left| \frac{T_i^{(0)}(\kappa)}{p_0} \right|.$$  

Since $p_i^* \neq 0$, reviewing the above proof, we easily see $\mathcal{W}(T(\kappa)) < \mathcal{W}(\kappa)$ when $q_0 \neq -M$. By the minimality of $\kappa$, we see $[T_0^{(0)}(\kappa)/p_0] = -M$ and $\mathcal{W}(T(\kappa)) = \mathcal{W}(\kappa)$. Repeating this argument we have

$$q_j = \left| \frac{T_0^{(j)}(\kappa)}{p_0} \right| = -M, \quad j = 0, 1, \ldots, d - 1.$$

By (5) with $k = i = d$ and $\alpha = \kappa$, we have

$$T_j^{(d)}(\kappa) = -\sum_{\ell=1}^{d-j} q_{d-\ell} p_{j+\ell} = -\sum_{\ell=j+1}^d q_{d-\ell} p_{\ell} = M \sum_{\ell=j+1}^d p_{\ell},$$

but this implies $\mathcal{W}(T^{(d)}(\kappa)) = M$, which contradicts the inequality $\mathcal{W}(\kappa) > M$. This shows $\mathcal{W}(\kappa) = M$ and moreover $\mathcal{W}(T^{(j)}(\kappa)) = M$ for any $j$. So we have

$$\left| \frac{T_0^{(j)}(\kappa)}{p_0} \right| \leq \frac{M^2}{1 + 1/M} \leq \frac{M^2}{1 + M} < M,$$
which shows \( q_j = [-M, M - 1] \cap \mathbb{Z} \) for \( j \geq 0 \). Again by (5) with \( k = i = d \) and \( \alpha = \kappa \), we have

\[
T^{(d)}(\kappa) = -\sum_{j=\ell}^{d-1} q_j p_{d+\ell-j}.
\]

Letting \( \varepsilon_j = -q_j \in [1 - M, M] \cap \mathbb{Z} \), we have

\[
T^{(d)}(\kappa) = \sum_{\ell=0}^{d-1} \left( \sum_{j=\ell}^{d-1} \varepsilon_j p_{d+\ell-j} \right) x^\ell,
\]

which has the form (4). This proves the assertion. \( \square \)

**Remark 1** The integer assumption on \( M \) is not necessary for the above proof but we cannot get a better bound by choosing non-integer \( M \geq 1 \).

**Remark 2** To derive a result of this type, we first used the length of \( \alpha (\sum_{i=0}^{d-1} |T_i^{(0)}|) \) instead of the weight and used a technique inspired by the analysis of the running time of the euclidean algorithm. (See e.g. [10].) Under this choice, we could only show a rather bad bound but it was an inspiring experience for us.

## 5 Proof of Theorem 2.

**Proof:** Define

\[
\alpha(\varepsilon_0, \ldots, \varepsilon_{d-1}) = \sum_{i=0}^{d-1} \left( \sum_{j=i}^{d-1} \varepsilon_j p_{d+i-j} \right) x^i.
\]

Since the assumption of Theorem 1 is satisfied with \( M = 1 \), it is enough to prove that every element of the form \( \alpha = \alpha(\varepsilon_0, \ldots, \varepsilon_{d-1}) \) with \( \varepsilon_j \in \{0, 1\}, 0 \leq j \leq d - 1 \) is representable in \( \{P(x), \mathcal{N}\} \). A simple computation shows that

\[
|T_i^{(0)}(\alpha)| \leq L(P) < p_0.
\]

This means that if \( T_i^{(0)}(\alpha) \geq 0 \) for some \( i \), then \( T_i^{(0)}(\alpha) \in \mathcal{N} \), otherwise \( p_0 - T_i^{(0)}(\alpha) \in \mathcal{N} \).

If \( p_1 \geq 0 \), then \( T_i^{(0)}(\alpha) \geq 0 \) for all \( i \), such that \( 0 \leq i \leq d - 1 \) and for all choices of \( \varepsilon_j \in \{0, 1\}, 0 \leq j \leq d - 1 \). Similarly, as \( p_2, \ldots, p_{d-1} \) are non-negative \( T_i^{(0)}(\alpha) \geq 0 \) for all \( i \), such that \( 1 \leq i \leq d - 1 \). If \( \varepsilon_{d-1} = 0 \) then \( T_0^{(0)}(\alpha) = \sum_{j=0}^{d-2} \varepsilon_j p_{d-j} > 0 \). In these cases every \( \alpha \) of form (4) is representable in \( \{P(x), \mathcal{N}\} \).

We assume \( p_1 < 0 \) and \( \varepsilon_{d-1} = 1 \) in the sequel. Let \( \varepsilon_j \in \{0, 1\}, 0 \leq j \leq d - 1 \) be fixed. Put \( \alpha = \alpha(\varepsilon_0, \ldots, \varepsilon_{d-1}) \). If \( T_0^{(0)}(\alpha) \geq 0 \), then \( \alpha \) is representable in \( \{P(x), \mathcal{N}\} \). Thus we may assume \( T_0^{(0)}(\alpha) < 0 \). Then there exists an \( i \) with \( 0 \leq i < d - 1 \) such that \( \varepsilon_i = 0 \) because \( \sum_{j=1}^{d} p_j \geq 0 \) by Lemma 2. Let \( j \) be the index such that \( \varepsilon_j = \ldots = \varepsilon_{d-1} = 1 \), but \( \varepsilon_{j-1} = 0 \). We apply to \( \alpha \) the transformation \( T \) several times and ultimately we obtain an element, which is represented in \( \{P(x), \mathcal{N}\} \).
Indeed, as $T_0(0)(\alpha) < 0$ we have $q_0 = \left\lfloor \frac{T_0(0)(\alpha)}{p_0} \right\rfloor = -1$. Putting $\varepsilon_d = 1$ we obtain

$$T(1)(\alpha) = \sum_{i=0}^{d-1} \left( \sum_{j=i}^{d-1} \varepsilon_{j+1} p_{d+i-j} \right) x^i.$$ 

Hence $T(1)(\alpha) = \alpha(\varepsilon_1, \ldots, \varepsilon_d)$. If $T_0(1)(\alpha) \geq 0$ then this is already the representation of $T(1)(\alpha)$ in $\{P(x), \mathcal{N}\}$. Otherwise, i.e., if $T_0(1)(\alpha) < 0$ we continue the process with $q_1 = \left\lfloor \frac{T_0(1)(\alpha)}{p_0} \right\rfloor = -1$ and $\varepsilon_{d+1} = 1$. Hence either $T_0(k)(\alpha) \geq 0$ for some $k < j - 1$ or $T_0(k)(\alpha) < 0$ for all $k$ with $0 \leq k < j - 1$. In the second case we have $T^{(j-1)}(\alpha) = \alpha(1, \ldots, 1)$.

Thus there exists always a $k \geq 0$ such that $T^{(k)}(\alpha)$ is representable in $\{P(x), \mathcal{N}\}$. Theorem 2 follows now immediately from Lemma 3. □

6 Proof of Theorem 3.

For

$$\alpha = \alpha(\varepsilon_0, \ldots, \varepsilon_{d-1}) = \sum_{i=0}^{d-1} \left( \sum_{j=i}^{d-1} \varepsilon_j p_{d+i-j} \right) x^i$$

with $\varepsilon_i \in \mathbb{Z}, i = 0, \ldots, d - 1$ let

$$E(\alpha) = \max\{|\varepsilon_i|, i = 0, \ldots, d - 1\}.$$ 

With this notation we prove the following useful lemma.

Lemma 5 Assume that $p_0 \geq L(P)$ and that $\alpha$ is given in the form (14). Then

$$E(T(\alpha)) \leq E(\alpha).$$

Proof: Taking

$$q = \left\lfloor \frac{1}{p_0} \sum_{j=0}^{d-1} \varepsilon_j p_{d-j} \right\rfloor$$

we have

$$\frac{1}{p_0} \sum_{j=0}^{d-1} \varepsilon_j p_{d-j} - 1 < q \leq \frac{1}{p_0} \sum_{j=0}^{d-1} \varepsilon_j p_{d-j}.$$ 

The inequality

$$\left| \frac{1}{p_0} \sum_{j=0}^{d-1} \varepsilon_j p_{d-j} \right| \leq \frac{E(\alpha) L(P)}{p_0} \leq E(\alpha)$$

implies

$$|q| \leq E(\alpha).$$
Putting $\epsilon_d = -q$ we obtain

$$T(\alpha) = \sum_{i=0}^{d-1} (\sum_{j=1}^{d-1} \epsilon_j p_{d+i-j}) x^i,$$

which implies

$$E(T(\alpha)) = \max\{|\epsilon_1|, \ldots, |\epsilon_{d-1}|, |\epsilon_d|\} \leq E(\alpha).$$

The lemma is proved. \(\square\)

Now we are in the position to prove Theorem 3.

Assume that there exists some $\ell$ with $0 < \ell < d$, such that $p_\ell + \sum_{j=\ell+1}^{d} |p_j| < 0$. We show that $-1$ is not representable in $\{P(x), N\}$. More precisely we prove for all $k \geq 0$ that at least one of the $T_j^{(k)}(-1), j = 0, \ldots, d-1$, is negative.

This assertion is obviously true for $k = 0$. Let $k \geq 0$ and assume that at least one of the $T_j^{(k)}(-1), j = 0, \ldots, d-1$, is negative. We have

$$-1 = \sum_{i=0}^{d-1} (\sum_{j=1}^{d-1} \epsilon_j p_{d+i-j}) x^i$$

with $\epsilon_0 = -1$ and $\epsilon_j = 0, j = 1, \ldots, d-1$. Hence

$$T^{(k)}(-1) = \sum_{i=0}^{d-1} (\sum_{j=1}^{d-1} \epsilon_{j+k} p_{d+i-j}) x^i$$

holds with $|\epsilon_{j+k}| \leq 1, j = 0, \ldots, d-1$, by Lemma 5 for all $k \geq 0$. Hence we have

$$T^{(k+1)}(-1) = \sum_{i=0}^{d-1} (\sum_{j=1}^{d-1} \epsilon_{j+k+1} p_{d+i-j}) x^i$$

with $\epsilon_{d+k} = -|T_0^{(k)}(-1)/p_0|$. We distinguish three cases according to the values of $\epsilon_{d+k}$.

Case 1: $\epsilon_{d+k} = -1$. Then $T_{d-1}^{(k+1)}(-1) = \epsilon_{d+k} p_d = -1$. Hence the assertion is true for $k + 1$.

Case 2: $\epsilon_{d+k} = 0$. Then $T_j^{(k+1)}(-1) = T_{j+1}^{(k)}(-1)$ for $j = 0, \ldots, d-2$, and $T_{d-1}^{(k+1)}(-1) = 0$. There exists by the hypothesis a $j$ with $0 \leq j \leq d-1$ such that $T_j^{(k)}(-1) < 0$. This index cannot be zero because $\epsilon_{d+k} = 0$. Hence $j > 0$ and $T_{j-1}^{(k+1)}(-1) = T_j^{(k)}(-1) < 0$. The assertion is true again.

Case 3: $\epsilon_{d+k} = 1$. In this case we have

$$T_{\ell-1}^{(k+1)}(-1) = \epsilon_{k+\ell} p_d + \ldots + \epsilon_{k+d-1} p_{\ell+1} + \epsilon_{k+d} p_\ell$$

$$= \epsilon_{k+\ell} p_d + \ldots + \epsilon_{k+d-1} p_{\ell+1} + p_\ell \leq p_\ell + \sum_{j=\ell+1}^{d} |p_j| < 0$$

because $|\epsilon_{k+j}| \leq 1, j = \ell, \ldots, d-1$, by Lemma 5. Theorem 3 is proved.
References


Shigeki Akiyama
Department of Mathematics, Faculty of Science, Niigata University,
Ikarashi 2-8050, Niigata 950-2181, Japan
e-mail: akiyama@math.sc.niigata-u.ac.jp
Attila Pethő
Institute of Mathematics and Computer Science, University of Debrecen,
H-4010 Debrecen P.O.Box 12, Hungary
e-mail: pethoe@math.klte.hu