On number expansions in lattices

A. Kovács
Department of Computer Algebra
Eötvös Loránd University, Budapest, Hungary
attila@compalg.inf.elte.hu

Abstract
In this paper, some problems regarding number systems will be analyzed in finite dimensional Euclidean spaces and an effective algorithm will be presented in order to classify number expansions for a given radix and digit set.
Keywords: Digital expansion, Number system

The author is grateful to Prof. I. Kátai for his remarks and suggestions regarding Assertion 4.

1 Introduction

Let \( \theta \) be any rational integer greater than one. It is well-known that every non-negative integer \( n \) has a unique representation of the form

\[
n = a_0 + a_1 \theta + \ldots + a_i \theta^i = (a_1 a_{i-1} \ldots a_1 a_0)_\theta, \tag{1}
\]

where the integers \( a_j \) are selected from the set \( D = \{0, 1, \ldots, \theta - 1\} \). The decimal \((\theta = 10)\) and binary \((\theta = 2)\) systems are the most familiar. Here, \( D \) is called the digit set and \( \theta \) is called the base or radix. In fact, for any \( \theta \leq -2 \) every rational integer, including the negatives, has a base \( \theta \) representation with digit set \( D = \{0, 1, \ldots, |\theta| - 1\} \). Conditions under which each rational integer has a unique radix representation have been investigated by D. W. ¹

¹The research was supported partly by OTKA-T031877 and Ericsson-ELTE CNL.
Matula [1], A. M. Odlyzko [2] and by B. Kovács, A. Pethő [3]. A straightforward way to extend radix systems is choosing the radix to an algebraic integer. The systematic research of positional number systems in algebraic extensions was initiated by I. Kátai and J. Szabó [4]. There is a more general framework for radix expansions including number representations in algebraic number fields. Recall that a lattice in \( \mathbb{R}^k \) is the set of all integer combinations of \( k \) linearly independent vectors. Let \( \Lambda \) be a lattice, which can be viewed either geometrically as a set of points in a Euclidean space, or algebraically, a \( \mathbb{Z} \)-module or as a finitely generated free Abelian group. Let \( M : \Lambda \to \Lambda \) be a group endomorphism, and let \( D \) be a finite subset of \( \Lambda \) containing \( 0 \). Clearly, \( M \) can be taken as a square non-singular matrix. Moreover, if the basis of \( M \) is chosen in \( \Lambda \), then \( M \) is an integer matrix.

**Definition:** The triple \((\Lambda, M, D)\) is called a *number system* (or having the unique representation property) if every element \( n \) of \( \Lambda \) has a unique finite representation of the form

\[
n = a_0 + Ma_1 + M^2a_2 + \ldots + M^n a_i = (a_0a_{i-1} \ldots a_1a_0)_{M},
\]

where \( a_i \in D \). Now, the endomorphism \( M \) is called the *base* or *radix*, \( D \) is the *digit set*. Observe that if \( \Lambda \) is a ring and \( M \) represents the endomorphism \( x \mapsto \beta x \) for every \( x \in \Lambda \) then (2) reduces to (1).

We say that \( M \) is *Jordan-diagonalizable* if it is similar to the companion matrix (or Frobenius matrix) of an irreducible monic polynomial over \( \mathbb{Z} \). It is known that the problems regarding number expansions in algebraic number fields are special cases of problems in \( \mathbb{Z}^k \). To be more precise, for every radix expansion in an algebraic number field there is a uniquely defined radix expansion in the system \((\mathbb{Z}^k, M, D)\) for some Jordan-diagonalizable \( M \) and for some digit set \( D \).

## 2 Some problems

### 2.1 Problem #1

One main problem concerning radix representation is to give conditions under which \((\Lambda, M, D)\) is a number system. For a lattice \( \Lambda \), both \( \Lambda \) and \( \Lambda A \) are Abelian groups under addition. The order of the factor group \( \Lambda / \Lambda A \) is \( | \det M | \). Let \( A_j \) denote the cosets of this group. If \( z_1, z_2 \in A_j \), so they are
in the same residue class then we will say that they are congruent modulo 
M and we will denote this by \( z_1 \equiv z_2 \pmod{M} \).

The following result was known and used by I. Kátaï and co-workers (see e.g. [5]) as well as by W. Gilbert [6] in algebraic number fields. Moreover, it can be found implicitly in A. Vince’s paper [7]. Recall that a linear map is called expansive if all eigenvalues have modulus greater than one.

**Assertion 1.** *(Necessary conditions for the number system property)*

\( (\Lambda, M, D) \) is a number system then

(a) \( D \) must be a complete set of residues modulo \( M \);

(b) \( M \) must be expansive and

(c) \( \det(I - M) \neq \pm 1 \).

**Proof:** Considering (a) if \( z \in \Lambda \) is represented by \( (a_m a_{m-1} \ldots a_1 a_0)_M \) then
\( z \equiv a_0 \pmod{M} \). Hence the digit set \( D \) must contain a complete residue system modulo \( M \). Now suppose that two digits \( c \) and \( d \) are congruent modulo \( M \). Then \( c - d = Me \) for some \( e \in \Lambda \). Represent \( e \) by \( (a_i a_{i-1} \ldots a_1 a_0)_M \) so that \( (c)_M = c = Me + d = (a_i a_{i-1} \ldots a_1 a_0 d)_M \). Hence \( c \in \Lambda \) has two different representations, which is a contradiction. Statement (b) was proved in [7]. Concerning (c) first observe that \( (I - M^n) \) is nonsingular for any positive integer \( n \). Otherwise 1 would be an eigenvalue of \( M^n \), hence \( M \) would have an eigenvalue of modulus one. Second, it is also clear that if \( (\Lambda, M, D) \) is a number system then there is not any \( \pi \in \Lambda \) and \( l \in \mathbb{N} \) for which \( \pi = a_0 + Ma_1 + \ldots + M^{l-1}a_{l-1} + M^l \pi \), where \( a_i \in D \). In other words \( (I - M^l)^{-1}(a_0 + Ma_1 + \ldots + M^{l-1}a_{l-1}) \in \Lambda \) can never be happen. But if \( \det(I - M) = \pm 1 \) then \( (I - M) \Lambda = (I - M)^{-1} \Lambda = \Lambda \), which is a contradiction. □

**Corollary.** Suppose that an arbitrary \( z \in \Lambda \) has a finite expansion of form (2). Then, the uniqueness of the representation follows from the assumption that any two elements of \( D \) are incongruent modulo \( M \).

If for a given triple \( (\Lambda, M, D) \) the conditions (a) and (b) in Assertion 1 hold then we say that it is a *radix system*. Assertion 1(b) answers the question of T. Safer [8]. Assertion 1(c) explains why it is impossible to find appropriate digit sets for the companion matrix of the polynomial \( x^2 + mx - m \), \( m \in \mathbb{Z} \), or for the matrix \( 2I + T \), where \( T \) is strictly upper (or lower) triangular.
2.2 Problem #2

The second question concerning number systems is the following: for a given $M$ satisfying criterion (b) and (c) in Assertion 1 is there any digit set $D$ for which $(\Lambda, M, D)$ is a number system? How many such digit sets exist and how to construct them? In imaginary quadratic fields due to G. Steidl [9] and I. Kátai [10] we know that to be able to construct number systems the conditions in Assertion 1 are also sufficient. Remarkable results are obtained by G. Farkas in real quadratic fields [11, 12, 13]. The above mentioned authors gave the constructions as well. The general problem seems to be hard. Nevertheless, if $M$ is Jordan-diagonalizable then some results are available [14]. Now, we prove the following.

Assertion 2. (Sufficient condition for the number system property)

Suppose that the conditions for $M, D$ in Assertion 1 hold. Let us denote in $\mathbb{R}^k$ a vector norm and the corresponding operator norm by $\| \cdot \|$ for which $r = \| M^{-1} \| < 1$. Let $K = \max \{ \| d \|, d \in D \}$ and $L = Kr/(1 - r)$. Let furthermore $R$ be a positive real number for which $z \in \Lambda, \| z \| \leq R$ implies $z \in D$. If $r \leq R/(R + K)$ then $(\Lambda, M, D)$ is a number system.

Proof: It is known [15] that if $\pi$ is a periodic element then $\| \pi \| \leq L$. Hence, if we could prove that $L \leq R$ then we would be ready, since in this case the only periodic element is the null vector. But if $r = \| M^{-1} \| \leq R/(R + K)$ then $Kr \leq R(1 - r)$, by which $L = Kr/(1 - r) \leq R$.

The construction of the digit set is as follows: enumerate all lattice points in a ‘big enough’ ball around the origin, order them using the appropriate norm and select a full residue system keeping the norm of the elements as small as possible.

Assertion 2 has an important corollary. First, observe that a basis transformation does not change the number system property. To be more precise, if $M_1$ and $M_2$ are similar via the matrix $Q$ then the number system property of $(\Lambda, M_1, D)$ and $(Q\Lambda, M_2, QD)$ holds at exactly the same time. Let $U = [-\frac{1}{2}, \frac{1}{2}]^k$ denote the $k$-dimensional half-open unit cube centered at the origin. Recall that the $k$-dimensional parallelepiped $V = MU$ has volume $|\det(M)|$ and the appropriate lattice points in $V$ constitute a full residue system modulo $M$. Suppose that the norm in $\mathbb{R}^k$ is the Euclidean norm. Then, performing a basis transformation, the full residue system $V$ can be transformed to the half-open unit cube $U$, in which case $K/R$ is equal to $\sqrt{k}$. Hence, we proved the following:

Assertion 3. For a given expansive $M$ suppose that $\| M^{-1} \|_2 \leq 1/(1 + \sqrt{k})$. 

4
Then, there exists a digit set $D$ for which $(\Lambda, M, D)$ is a number system.

Our result is sharper than that of A. Vince [7] except in dimension 2. Applying Assertions 1 and 3 in dimension 1 shows that if $2 < \theta \in \mathbb{Z}$ then every rational integer has a unique base $\theta$ radix representation with $D = \{-\lfloor\theta | -1/2\rfloor, \ldots, \lfloor\theta | /2\rfloor\}$, which is well-known. Consider the ring of Gaussian integers $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$ and let $\theta = A + Bi \in \mathbb{Z}[i]$. In this case $M_\theta = \left(\frac{A^2 - B^2}{A^2 + B^2}\right)$ and $\|M_\theta^{-1}\|_2 = 1/\sqrt{A^2 + B^2}$, which is, apart from a few cases, always smaller than $1/(1 + \sqrt{2})$. Keeping in mind Assertion 1, Assertion 2 and [4] these cases are easy to handle. We got the following: for any Gaussian integer $\theta$ of modulus larger than one, except 2 and $1 \pm i$, there exists a full residue system $D$ such that $(\mathbb{Z}^2, M_\theta, D)$ is a number system. Hence, as a special case of Assertion 3 we have the result of G. Stein [1].

If we consider the Eisenstein integers $\mathbb{Z}[\omega] = \{a + b\omega : a, b \in \mathbb{Z}\}$, where $\omega$ is the complex cube root of unity, and we perform the above mentioned computations, we obtain the same conclusion. Nevertheless, it is not any surprise: I. Kátai solved the problem in all imaginary quadratic fields. If we consider the real quadratic fields — without going into the details — it is possible to reprove the result of G. Farkas [11]. The interesting is that the above mentioned authors gave the digit sets explicitly which is different from our construction. This suggests that the unique representation property depends mainly on the radix, and if any, than several different digit sets can be constructed.

2.3 Problem #3

The third problem regarding number expansions is to give an estimate for the length of expansions in the radix system $(\Lambda, M, D)$.

Let us denote in $\mathbb{R}^k$ a vector norm and the corresponding operator norm by $\|\cdot\|$ for which $r = \|M^{-1}\| < 1$, let $K = \max\{\|d\|, d \in D\}$ and $L = Kr/(1-r)$ as before. Let $z \in \Lambda \setminus \{0\}$ be fixed. Let us define the path of $z = z_0$ in $\Lambda$ by $z_j = a_j + Mz_{j+1}$ ($j = 0, \ldots$). Let $T = l(z)$ be the smallest non-negative integer for which $\|z_T\| \leq L$. The existence of such a $T$ was proved in [15].

--Historical remark: for the first proof of this result there is a research report by M. Davio, J. P. Deschamps and C. Gossart [16] dated back to 1978.
Assertion 4. There is a constant $c$ for which

$$l(z) \leq \frac{\log \| z \|}{\log (1/\| M^{-1} \|)} + c. \quad (3)$$

**Proof:** It is enough to examine the case $\| z \| > L$, $z \in \Lambda$. Since $z_j = a_j + Mz_{j+1}$ therefore $z_{j+1} = M^{-1}z_j - M^{-1}a_j$, hence $\| z_{j+1} \| \leq r(\| z_j \| + K)$. Let $t = t(z_0)$ be the smallest non-negative integer for which $\| z_t \| \leq 2KL$. Since the ball $\| \omega \| \leq 2KL$ contains finitely many lattice points therefore the inequality

$$l(z) \leq t(z) + c_1 \quad (4)$$

holds for an appropriate constant $c_1$. On the other hand $2KL < \| z_{t-1} \| \leq r(\| z_{t-1} \| + K) \leq r^2(\| z_{t-2} \| + K) + rK \leq \ldots \leq r^{t-1}\| z_0 \| + KL$. It means that $KL \leq r^{t-1}\| z_0 \|$, hence

$$\log KL \leq (t - 1) \log r + \log \| z_0 \|,$$

from which we can deduce that

$$(t - 1) \log 1/r \leq \log \| z_0 \| - \log KL,$$

i.e.,

$$t \leq \frac{\log \| z_0 \|}{\log (1/r)} + c_2$$

for an appropriate $c_2$. Using the inequality (4) the assertion follows immediately. \qed


### 2.4 Problem #4

The fourth problem concerning radix systems is to give an efficient algorithm to decide whether $(\Lambda, M, D)$ is a number system. Such an algorithm was given in [15]. The main idea is the following. First observe that if we change the basis in $\Lambda$, a similar integer matrix $M_2 : \mathbb{Z}^k \to \mathbb{Z}^k$ can be obtained. Hence the number system property can be examined without loss of generality on
the cubic lattice $\mathbb{Z}^k$. This has a computational advantage, since $M$, and its characteristic polynomial have integer coefficients.

Suppose that for a given radix system $(\mathbb{Z}^k, M, D)$ the conditions in Assertion 1 hold. It is not hard to see that for an arbitrary $z \in \mathbb{Z}^k$ the path $z, \Phi(z), \Phi^2(z), \ldots$ is ultimately periodic, where

$$\Phi : \mathbb{Z}^k \rightarrow \mathbb{Z}^k, \Phi(z) = M^{-1}(z - d), d \in D, z \equiv d \mod M.$$  \hfill (5)

The path must form a cycle inside the set $\mathbb{Z}^k \cap (-H)$, where $H$ denotes the set of fractions, i.e.

$$H = \left\{ \sum_{n=1}^{\infty} M^{-n} d_n : d_n \in D \right\} \subseteq \mathbb{R}^k.$$

Since $H$ is compact, the number of cycles is finite, their length is bounded. These cycles define a classification of $\mathbb{Z}^k$, i.e. two integers $x, y \in \mathbb{Z}^k$ are in the same class iff $\Phi^{l_1}(x) = \Phi^{l_2}(y)$ for some non-negative integers $l_1, l_2$. But how to calculate the integers in the set $-H$ (or, which is computational equivalent, in the set $H$)? Clearly, it is enough to determine a set $G, H \subseteq G$, for which the set of integers in $G$ can be computed simply. Then, applying $\Phi$ for these vectors one can obtain all the cycles. If the only periodic element is $0 \in \mathbb{Z}^k$ then $(\mathbb{Z}^k, M, D)$ is a number system. In [15] the set $G$, as a $k$-dimensional rectangle was constructed. In the next section we show how to construct in an effective way the set $W, H \subseteq W \subseteq G$, by which we can (in higher dimensions even drastically) reduce the number of integer vectors, which still cover the integers in $H$. We remark that in general there is not known any fast method to resolve the fourth problem different from the above mentioned classification. There is not known any procedure of polynomial complexity to calculate the shortest cycle different from $0 \rightarrow 0$.

**2.5 Problem #5**

The fifth problem concerning radix representation is closely related to the fourth: characterize the number, location and structural properties of the periodic elements, when the system $(\mathbb{Z}^k, M, D)$ is not a number system. It seems to be a hard problem. In case of imaginary quadratic fields using canonical digit sets, i.e. $D = \{0, 1, \ldots, \lfloor \text{det}(M) \rfloor - 1\}$, it was completely answered [19, 20]. There are also some results in the real quadratic field $\mathbb{Q}(\sqrt{a})$ [21]. Recently, J. Thuswaldner [22] described the systems $(\mathbb{Z}^2, M, D)$, where the digit sets are the canonical digit sets and $M$ have the special form $\left( \begin{array}{cc} A & B \\ z_1 & z_2 \end{array} \right)$. 

\textit{7}
3 Classification of number expansions

A finite set of contractions \( \{ f_i \} \) mapping from \( \mathbb{R}^k \) to \( \mathbb{R}^k \) is called an iterated function system (IFS). On the space \( S \) of nonempty compact subsets of \( \mathbb{R}^k \), with respect to the Hausdorff metric \( \delta(A, B) = \inf \{ r : A \subseteq N_r(B) \text{ and } B \subseteq N_r(A) \} \), where \( N_r(A) \) is the open \( r \)-neighborhood of \( A \), define \( f : S \to S \) by \( f(X) = \bigcup_{i=1}^{i} f_i(X) \), for any compact set \( X \). Clearly, \( f \) is a contraction on \( S \) and hence, by Hutchinson’s theorem [23], \( f \) has a unique fixed point or attractor \( T \) satisfying

\[
T = \bigcup_{i=1}^{i} f_i(T)
\]

and given by

\[
T = \lim_{n \to \infty} f^{(n)}(X_0),
\]

where \( f^{(n)} \) denotes the \( n \)th iterate of \( f \), \( X_0 \) is an arbitrary compact subset of \( \mathbb{R}^k \), and the limit is with respect to the Hausdorff metric.

For each digit \( d \in D \) we define the function \( f_d : \mathbb{R}^k \to \mathbb{R}^k \) by \( f_d(z) = M^{-1}(z + d) \). These are linear contraction maps. If \( z \in H \) then \( f_d(z) \in H \). Clearly, \( f_d \) is a right-shift map and furthermore \( H = \bigcup_{d \in D} f_d(H) \) so \( H \) is the unique invariant set determined by Hutchinson’s theorem applied to the functions \( f_d \). The set \( H \) is self-affine with respect to these functions.

It was mentioned in the previous section that we are interested in the integers in the set \(-H\). Let \( \pi \in -H \). Then

\[
-\pi - (M^{-1}d_1 + \ldots + M^{-J}d_J) = M^{-(J+1)}d_{J+1} + M^{-(J+2)}d_{J+2} + \ldots,
\]

for the appropriate sequence \( d_i \in D \). Fortunately, for the right hand side of (6) a good estimate can be given. The following algorithm provides the set \( W \), for which the integers in \( W \) cover the integers in \( H \).

**Number Expansion Classification Algorithm** in \( \mathbb{Z}^k \) for a given expansive matrix \( M \) and digit set \( D \). Let \( M \in \mathbb{Z}^{k \times k} \) be similar to \( M \) via the matrix \( Q \) and let \( Q \) be an optional argument of the algorithm. If it is not given then let \( Q \) be the identity matrix. Let \( \tilde{D} = QD \). Further, \( B \) and \( C \) are constants depending on the given computer hardware (word size, memory capacity) and on the matrix \( M \). \( B \) is an integer and \( C < 1 \) a real number.
1. \( q := \min \{ j \in \mathbb{N}, \| \hat{M}^{-j} \|_\infty < 1 \} \);
2. \( s := \min \{ j \in \mathbb{N}, r := \| \hat{M}^{-j} \|_\infty < C \} \);
3. \( f := (f_1, \ldots, f_k)^T \in \mathbb{R}^k, \ f_m = \frac{1}{1-r} \sum_{i=1}^r \max_{b \in D} |c_m^{(i)}(b)|, \ 1 \leq m \leq k, \)
   where \( (c_1^{(i)}(b), \ldots, c_k^{(i)}(b))^T = M^{-1}b \);
4. \( \minvol := \infty \); Choose an appropriate \( B, q \leq B \leq s \);
5. for \( j \) from \( q \) to \( B \) do {
   if \( \| \hat{M}^{-j} \|_\infty < 1 \) {
      Compute the vector \( v^{(i)} = (v_1^{(i)}, \ldots, v_k^{(i)})^T \in \mathbb{R}^k, \ v_m^{(j)} = \sum_{i=1}^k |M^{-1}f_i|, \ 1 \leq m \leq k; \)
      if \( (\omega := \prod_{i=1}^k v_i^{(j)} < \minvol) \) \{ \minvol := \omega; J := j; \}\}
6. \( U := \{ -\sum_{i=1}^j M^{-1}b, b \in D \} \);
7. \( S := \bigcup_{u \in U} (u + P) \), where \( P \) denotes the \( k \)-dimensional rectangle
   \( P = \{(p_1, \ldots, p_k)^T \in \mathbb{R}^k, \ p_i \leq v_i^{(j)}, \ 1 \leq i \leq k \} \);
8. \( W := \{ w = (w_1, \ldots, w_k)^T \in \mathbb{Z}^k, \ qw \in S \} \);
9. Apply the function \( \Phi \) determined by the system \( (\mathbb{Z}^k, M, D) \) for the points of \( W \) and the arising cycles mean the required classification.

The lines 1-3 provide the \( k \)-dimensional rectangle \( \hat{G} = \{(g_1, \ldots, g_k)^T \in \mathbb{R}^k, \ |g_i| \leq f_i, \ 1 \leq i \leq k \} \) as was suggested in [15]. Let us analyze the second assignment in line 4. If we increase \( B \), the time complexity of the algorithm grows exponentially in \( t = | \det(M) | \). Unfortunately, in some cases \( q \) can be "rather big", which means that the convergence of \( M^{-i} \) \( (i \to \infty) \) is slow. In these cases this algorithm can be ineffective, even keeping the running time moderate one choose \( B \) close to \( q \). The reason is that the set \( \hat{G} \) can also be rather big. Let an example be the Frobenius matrix of the irreducible polynomial \( 2 + 3x + 4x^2 + 4x^3 + 4x^4 + 3x^5 + 2x^6 + x^7 \) with the canonical (binary) digit set, \( Q = I, C = 0.01 \). Then \( s = 188, q = 53 \) and the number of integers in \( \hat{G} \) is 15319297125. Using other kinds of matrices, during the computation of \( s \) problems can arise with the matrix elements (see [15], section 3.1, remark 2). Line 5 try to keep the index \( J \) small. The lines 6-8 are the application of Hutchinson’s theorem in (6). Considering line 1 one can observe that the number of elements of the set \( W \) depends also on \( | \det(Q) | \). Line 9 was through out in [15]. The termination of the algorithm is clear.

It must be emphasized that the running time of the algorithm depends strongly on the matrices \( M \) and \( Q \), i.e., on the basis of the lattice determined by the matrix \( M \). In other words one has to choose the matrix \( Q \) in a way that the convergence of \( M^{-i} = (QM^{-1})^{-i} (i \to \infty) \) is fast, \( | \det(Q) | \)
is big and the volume of \( \hat{G} \) is as small as possible. It seems to be rather hard. Sometimes the simple idea of choosing the matrix \( Q \) in a way that \( M = M^T \) can help. Fortunately, for a large class of matrices the algorithm is quite effective even if we choose \( Q \) for the identity matrix. The author implemented the CLASSIFICATION ALGORITHM in C language. In order to perform computations in the lattice effectively the elements of \( \mathbb{Z}^k \) were transformed to \( \mathbb{Z} \) using mixed radix representation. During the computation of elements of the set \( S \) a hashing table was used.

Without giving the exact values of \( B \) and \( C \) let us see some examples. Let \( Q = I \). Consider example 1 in [15]. With the old method the set \( G \) contains 25 integer elements, with the new algorithm the set \( W \) only 4, which is equal to \( \#P \), the number of periodic elements. In example 2 these values are: \( \#\{\text{integers in } G\} = 35, \#\{\text{integers in } W\} = 6, \#P = 6 \). Finally, consider the Frobenius matrix of the polynomial \( 2 - x^3 + x^6 \) with the binary digit set. Then \( \#\{\text{integers in } G\} = 42875, \#\{\text{integers in } W\} = 1134, \#P = 1 \).

References


A. Kovács
Department of Computer Algebra
Eötvös Loránd University
Pázmány P. sétány I/C.
H-1117 Budapest, Hungary