Canonical number systems for complex integers

By I. KÁTAI and J. SZABÓ in Budapest

1. It is a well-known fact that every non-negative integer \( N \) has a unique representation of the form
\[
N = a_0 + a_1 A + \ldots + a_k A^k,
\]
where the integers \( a_j \) are selected from the set \( \{0, 1, \ldots, A-1\} \), and \( A \) is an integer, \( A \geq 2 \). Furthermore, choosing a negative integer \( -A \) (\( A \geq 2 \)), we can represent every integer \( N \) as a sum:
\[
N = a_0 + a_1 (-A) + \ldots + a_k (-A)^k, \quad 0 \leq a_j \leq A-1 \quad (j = 0, 1, \ldots, k-1),
\]
where \( a_j \) are integers. The representation (1.2) is also unique.

The number systems of negative base have some applications in the theory of computations.

The following question seems to be interesting: Given a Gaussian integer \( \mathfrak{g} \), can we represent every Gaussian integer \( \alpha \) in the form
\[
\alpha = r_0 + r_1 \mathfrak{g} + \ldots + r_k \mathfrak{g}^k
\]
or not? Here \( r_j \in \mathcal{U} \), \( \mathcal{U} \) being a fixed complete residue system mod \( \mathfrak{g} \).

If the answer is affirmative, we say that \( (\mathfrak{g}, \mathcal{U}) \) is a number system.
We shall investigate only the case \( \mathcal{U} = \mathcal{U}_0 \) where
\[
\mathcal{U}_0 = \{0, 1, \ldots, N(\mathfrak{g})-1\},
\]
and \( N(\mathfrak{g}) \) denotes the "norm"
\[
N(\mathfrak{g}) = \mathfrak{g} \cdot \overline{\mathfrak{g}} = (\text{Re} \, \mathfrak{g})^2 + (\text{Im} \, \mathfrak{g})^2.
\]

It is known that for \( \mathfrak{g} = -1+i \), \( (\mathfrak{g}, \mathcal{U}_0) \) is a number system; see [1]
We prove:

Theorem 1. \( (\mathfrak{g}, \mathcal{U}_0) \) is a number system if and only if
a) \( \text{Re} \, \mathfrak{g} < 0 \) and
b) \( \text{Im} \, \mathfrak{g} = \pm 1 \).

For \( \mathfrak{g} = -A \pm i \) the representation of \( \alpha \) in the form (1.3) is unique.
Theorem 2. Let $\mathfrak{g} = -A \pm i$, $z$ an arbitrary complex number. Then

$$(1.5) \quad z = a_0 \mathfrak{g}^l + \ldots + a_0 + \frac{a_{-1}}{\mathfrak{g}} + \frac{a_{-2}}{\mathfrak{g}^2} + \ldots,$$

where $a_j \in \mathfrak{U}_0$ ($j = l, l-1, \ldots, 0, -1, -2, \ldots$).

We do not assert the uniqueness of the representation of $z$ in the form (1.5).

2. Proof of Theorem 1. Necessity. Let $\mathfrak{g} = A + Bi$. Then

$$\mathfrak{U}_0 = \{0, 1, \ldots, A^2 + B^2 - 1\}.$$ 

It is obvious that $\mathfrak{U}_0$ must be a complete residue system mod $\mathfrak{g}$ if $(\mathfrak{g}, \mathfrak{U}_0)$ is a number system. In the opposite case there is an $\alpha$ which is incongruent to $k$ for every $k$ in $\mathfrak{U}_0$, but from (1.3) $\alpha \equiv r_0 \pmod{\mathfrak{g}}$, $r_0 \in \mathfrak{U}_0$ follows, and this is a contradiction.

Suppose that $A > 0$. We prove that $\alpha = (1 - A) + iB = 1 - \mathfrak{g}$ has no representation of type (1.3). Suppose in the contrary that

$$(2.1) \quad \alpha = r_0 + r_1 \mathfrak{g} + \ldots + r_k \mathfrak{g}^k.$$ 

Let $q = \alpha(1 - \mathfrak{g}) = (1 - A)^2 + B^2 = A^2 + B^2 - 2A + 1.$

Since $A \equiv 1$, we have $q \in \mathfrak{U}_0$. From (2.1) we get

$$q = r_0 + (r_1 - r_0) \mathfrak{g} + \ldots + (r_k - r_{k-1}) \mathfrak{g}^k - r_k \mathfrak{g}^{k+1}.$$ 

Hence $q \equiv r_0 \pmod{\mathfrak{g}}$, and by $q \in \mathfrak{U}_0$, $r_0 \in \mathfrak{U}_0$ we get: $q = r_0$. So

$$(r_1 - r_0) \mathfrak{g} + \ldots + (r_k - r_{k-1}) \mathfrak{g}^k - r_k \mathfrak{g}^{k+1} = 0.$$ 

Hence it follows immediately that

$$r_1 - r_0 = 0, \ldots, r_k - r_{k-1} = 0, \quad r_k = 0,$$

whence $r_k = r_{k-1} = \ldots = r_1 = r_0 = 0$. Therefore $q = 0$, and so $A = 1$, $B = 0$. But it is obvious that $\mathfrak{g} = 1$ is not a base of a number system. Similarly, $\mathfrak{g} = \pm i$ ($A = 0$, $B = \pm 1$) is not a base of a number system, either.

Let now $\text{Im} \mathfrak{g} = B \neq \pm 1$. Let us take into account that $B$ is a divisor of $\text{Im} \mathfrak{g}^v$ ($v = 1, 2, \ldots$). Hence, for an $\alpha$ of (1.3) we get:

$$\text{Im} \alpha = r_1 \text{Im} \mathfrak{g} + \ldots + r_k \text{Im} \mathfrak{g}^k,$$

and so $B | \text{Im} \alpha$. Consequently, (1.3) will not hold for $\alpha = i$ ($B \neq \pm 1$).

Sufficiency. Let now $\mathfrak{g} = -A + i$ ($A \equiv 1$). Then $\mathfrak{U}_0$ is a complete residue system mod $\mathfrak{g}$ as it is well known. Let us take into account, that

$$(2.2) \quad \mathfrak{g}^2 + 2A\mathfrak{g} + A^2 + 1 = 0.$$
Let $\alpha = E + Fi$ be an arbitrary Gaussian integer. Taking $D = F$, $C = E + AF$, we get
\begin{equation}
\alpha = C + D\mathfrak{d}.
\end{equation}

First we prove that every $\alpha$ has the form
\begin{equation}
\alpha = U + V\mathfrak{d} + X\mathfrak{d}^2 + Y\mathfrak{d}^3,
\end{equation}
where $U, V, X, Y$ are non-negative integers. From (2.2) we have
\[-1 = \mathfrak{d}^3 + 2A\mathfrak{d} + A^2.
\]
Assuming that $C < 0$ we can substitute $C$ in (2.3) by
\[|C| \cdot \mathfrak{d}^3 + 2A|C| \cdot \mathfrak{d} + A^2|C|.
\]
In the case $D = 0$ we take a similar substitution, and get (2.4).

We shall use the following relation:
\begin{equation}
A^3 + 1 = \mathfrak{d}^3 + (2A - 1)\mathfrak{d}^2 + (A - 1)^2 \mathfrak{d}.
\end{equation}

Let
\begin{equation}
\alpha = d_0 + d_1 \mathfrak{d} + \ldots + d_k \mathfrak{d}^k \quad (k \geq 3), \quad d_j \geq 0 \quad (j = 0, \ldots, k).
\end{equation}

Let
\begin{equation}
t(\alpha, d) = d_0 + d_1 + \ldots + d_k;
\end{equation}
$t(\alpha, d)$ is a non-negative integer, $t(\alpha, d) = 0$ only if $\alpha = 0$.

We take
\[d_0 = r_0 + tN(\mathfrak{d}) = r_0 + t(A^2 + 1),
\]
$t \geq 0$, integer, $0 \leq r_0 \leq A^2$. From (2.5) we have
\begin{equation}
d_0 = r_0 + t(A^2 + 1) = r_0 + t(A - 1)^2 \mathfrak{d} + t(2A - 1)\mathfrak{d}^2 + t\mathfrak{d}^3.
\end{equation}
We take the right hand side of (2.8) into (2.6). Then
\[\alpha = r_0 + (d_1 + t(A - 1)^2)\mathfrak{d} + (d_2 + t(2A - 1))\mathfrak{d}^2 + (d_3 + t)\mathfrak{d}^3 + d_4 \mathfrak{d}^4 + \ldots + d_k \mathfrak{d}^k =
\begin{equation}
d_0^* + d_1^* \mathfrak{d} + \ldots + d_k^* \mathfrak{d}^k.
\end{equation}
Since
\[-t(A + 1)^2 + t(A - 1)^2 + t(2A - 1) + t = 0,
\]
therefore
\[t(\alpha, d^*) = d_0^* + \ldots + d_k^* = t(\alpha, d), \quad d_j^* \equiv 0 \quad (j = 0, \ldots, k).
\]
Let
\begin{equation}
\alpha_1 = d_1^* + d_2^* \mathfrak{d} + \ldots + d_k^* \mathfrak{d}^{k - 1}.
\end{equation}
We have
\[ \alpha = \alpha_1 \mathcal{G} + r_0 \quad (r_0 \in \mathfrak{A}_0), \]
\[ t(\alpha_1, d^*) = d_1^* + d_2^* + \ldots + d_k^*. \]

It is obvious that \( t(\alpha_1, d^*) < t(\alpha, d) \), when \( r_0 \neq 0 \). For \( r_0 = 0 \), \( t(\alpha_1, d^*) = t(\alpha, d) \).

Now we write \( t(\alpha, d) = t(\alpha_1, d^*) = t(\alpha_1), \ldots \). We repeat the algorithm (2.9), (2.11):
\[ \alpha = \alpha_1 \mathcal{G} + r_0, \quad \alpha_1 = \alpha_2 \mathcal{G} + r_1, \quad \ldots, \quad \alpha_{j-1} = \alpha_j \mathcal{G} + r_{j-1} \quad (r_i \in \mathfrak{A}_0). \]

Then \( t(\alpha) \equiv t(\alpha_i) \equiv \ldots \) and \( t(\alpha_i) \equiv t(\alpha_{i+1}) \) when \( r_i \neq 0 \). This process is terminated at the \( j \)th step if \( \alpha_j = 0 \). In this case we get
\[ \alpha = r_0 + r_1 \mathcal{G} + \ldots + r_{j-1} \mathcal{G}^{j-1} \quad (r_i \in \mathfrak{A}_0). \]

Suppose that the process is not terminated. Then for a suitably large \( i \)
\[ t(\alpha_i) = t(\alpha_{i+1}) = \ldots \quad (\neq 0). \]
Hence
\[ \alpha_i = \alpha_{i+1} \mathcal{G}, \quad \ldots, \quad \alpha_{i+k-1} = \alpha_{i+k} \mathcal{G} \]
and, therefore, \( \mathcal{G}^k \mid \alpha_i \) \( (k = 1, 2, \ldots) \). This holds only if \( \alpha_i = 0 \).

We proved the existence of the representation of \( \alpha \) in the form (1.3).

Let us suppose now that there is an \( \alpha \) which has two different representations:
\[ \alpha = r_0 + r_1 \mathcal{G} + \ldots + r_k \mathcal{G}^k = s_0 + s_1 \mathcal{G} + \ldots + s_k \mathcal{G}^k, \quad r_i, s_i \in \mathfrak{A}_0. \]

Then \( 0 = (r_0 - s_0) + (r_1 - s_1) \mathcal{G} + \ldots + (r_k - s_k) \mathcal{G}^k \) and therefore \( r_0 \equiv s_0 \mod \mathcal{G} \); as \( r_0, s_0 \in \mathfrak{A}_0 \) we get \( r_0 = s_0 \). Dividing by \( \mathcal{G} \), we get
\[ 0 = (r_1 - s_1) + \ldots + (r_k - s_k) \mathcal{G}^{k-1}. \]

We repeat the argument. Finally we get:
\[ r_0 = s_0, \quad r_1 = s_1, \ldots, \quad r_k = s_k. \]

We have proved the theorem for \( \mathcal{G} = -A + i \).

Let now \( \mathcal{G} = -A - i \). Using the theorem for \( \mathcal{G} = -A + i \), we get
\[ \bar{\alpha} = r_0 + r_1 \mathcal{G} + \ldots + r_k \mathcal{G}^k \quad (r_i \in \mathfrak{A}_0) \]
for every Gaussian integer \( \bar{\alpha} \). Hence
\[ \alpha = r_0 + r_1 \mathcal{G} + \ldots + r_k \mathcal{G}^k, \]
and so the theorem holds for \( \mathcal{G} = -A - i \), too.
3. Proof of Theorem 2. Let \( z \) be an arbitrary complex number, \( z = x + iy \). Let

\[
\mathcal{g}^k = U_k + iV_k.
\]

We have

\[
z = \frac{z^\mathcal{g}^k}{\mathcal{g}^k} = \frac{(x + iy)(U_k + iV_k)}{\mathcal{g}^k} = \frac{C_k + D_k i}{\mathcal{g}^k} + \frac{u_k + v_k i}{\mathcal{g}^k},
\]

where \( C_k, D_k \) are rational integers, \( |u_k| < 1, |v_k| < 1 \). Let

\[
z_k = \frac{C_k + iD_k}{\mathcal{g}^k}, \quad \delta_k = \frac{u_k + iv_k}{\mathcal{g}^k}.
\]

It is obvious that \( \delta_k \to 0 \) (\( k \to \infty \)), and so \( z_k \to z \). Since \( C_k + iD_k \) is a Gaussian integer, by Theorem 1 we have

\[
C_k + iD_k = a_t^* \mathcal{g}^t + \cdots + a_0^*, \quad t = t(k).
\]

First we prove that the sequence \( t(k) - k \) (\( k = 1, 2, \ldots \)) has an upper bound. Indeed, from (3.4)

\[
z_k = a_t^* \mathcal{g}^{t-k} + \cdots + a_0^* \mathcal{g}^{-k}.
\]

Hence

\[
a_t^* \mathcal{g}^{t-k} + \cdots + a_0^* = z_k - \frac{a_{k-1}^*}{\mathcal{g}} - \cdots - \frac{a_0^*}{\mathcal{g}^k},
\]

and so

\[
|a_t^* \mathcal{g}^{t-k} + \cdots + a_0^*| = |z_k| + \frac{a_{k-1}^*}{|\mathcal{g}|} + \cdots + \frac{a_0^*}{|\mathcal{g}|^k} \leq |z| + |\delta_k| + A^2 \left( \frac{1}{|\mathcal{g}|} + \frac{1}{|\mathcal{g}|^2} + \cdots \right) \leq |z| + |\delta_k| + \frac{A^2}{|\mathcal{g}| - 1}.
\]

Hence it follows that

\[
|a_t^* \mathcal{g}^{t-k} + \cdots + a_0^*| \leq c,
\]

\( c = c(z) \) being a suitable positive constant.

Since the representation of Gaussian integers in the form (1.3) is unique, and the circle \( |w| \leq c \) contains only a finite set of Gaussian integers, therefore \( t(k) - k \) has an upper bound. Let \( K \) be an integer, \( t - k \leq K \). Then we can write \( z_k \) as

\[
z_k = a_k^{(1)} \mathcal{g}^1 + \cdots + a_0^{(1)} + \frac{a_k^{(2)}}{\mathcal{g}} + \frac{a_0^{(2)}}{\mathcal{g}^2} + \cdots,
\]

where \( a_j^{(1)} \in \mathbb{Z}_0 \) (\( j = K, K - 1, \ldots, 0, -1, \ldots \)). Let \( b_K \in \mathbb{Z}_0 \) be an integer so that \( a_K^{(1)} = -b_K \) for infinitely many \( k \). Let \( S_K \) be the subset of those integers \( k \) satisfying \( a_K^{(1)} = -b_K \)
Suppose that \( S_K, \ldots, S_{t+1} \) is constructed \( (S_K \supseteq \ldots \supseteq S_{t+1}) \). Then there is a
\( b_t \in \mathbb{U}_0 \), such that for infinitely many \( k \) in \( S_{t+1} \), \( d_t^{(k)} = b_t \). Let \( S_t \) be the set of these
\( k' \)'s. \( S_t \) has infinitely many elements. We repeat this argument for \( K, K-1, \ldots, 0, -1, \ldots \). Let
\[
  w = b_K \alpha^k + \ldots + b_0 + \frac{b_{-1}}{\alpha} + \ldots
\]
Let \( k_1 < k_2 < \ldots \) be an infinite sequence, so that
\[
  k_v \in S_{K-v+1} \quad (v=1, 2, \ldots).
\]
Since
\[
  z_k = b_K \alpha^k + \ldots + b_{K-v+1} \alpha^{K-v+1} + a_{K-v}^{(k_v)} \alpha^{K-v} + \ldots,
\]
therefore
\[
  \lim_{v \to \infty} z_{k_v} = w.
\]
Taking into account that \( \lim_{k \to \infty} z_k = z \), we have \( w = z \). Hence it follows that (3.9) is a
suitable representation of \( z \).

We have proved Theorem 2.

Reference


(Received December 28, 1974)