Chapter 10

Arithmetic Codes

§10.1 AN Codes

In this chapter we shall give a brief introduction to codes which are used to check and correct arithmetic operations performed by a computer. Operations are now ordinary arithmetic and as a result the theory is quite different from the preceding chapters. However, there is in several places a similarity to the theory of cyclic codes. In some cases we shall leave the details of proofs to the reader. For further information on this area see the references mentioned in Section 10.4.

The arithmetic operations in this chapter are carried out with numbers represented in the number system with base r ($r \in \mathbb{N}$, $r \geq 2$). For practical purposes the binary case ($r = 2$) and the decimal case ($r = 10$) are the most important. The first thing we have to do is to find a suitable distance function. In the previous chapters we have used Hamming distance but that is not a suitable distance function for the present purposes. One error in an addition can cause many incorrect digits in the answer because of carry. We need a distance function which corresponds to arithmetical errors in the same way as Hamming distance corresponds to misprints in words.

(10.1.1) Definition. The *arithmetic weight* $w(x)$ of an integer x is the minimal $t \geq 0$ such that there is a representation

$$x = \sum_{i=1}^{t} a_i r^{n(i)},$$

with integers a_i, $n(i)$ for which $|a_i| < r$, $n(i) \geq 0$ $(i = 1, 2, \ldots, t)$. The *arithmetic distance* $d(x, y)$ of two integers is defined by

$$d(x, y) := w(x - y).$$

It is easily checked that this is indeed a metric on $\mathbb{Z}$. Arithmetic distance is translation invariant, i.e. $d(x, y) = d(x + z, y + z)$. This is not true for the Hamming distance of two integers (in r-ary representation). Arithmetic distance is at most equal to the Hamming distance.

We shall consider codes C of the form

$$C := \{AN \mid N \in \mathbb{Z}, 0 \leq N < B\},$$

where A and B are fixed positive integers. Such codes are called *AN codes*. These codes are used in the following way. Suppose we wish to add two integers N_1 and N_2 (both positive and small compared to B). These are encoded as AN_1 and AN_2 and then these two integers are added. Let S be the sum. If no errors have been made then we find $N_1 + N_2$ by dividing by A. If S is not divisible by A, i.e. errors have been made, we look for the codeword AN_3 such that $d(S, AN_3)$ is minimal. The most likely value of $N_1 + N_2$ is N_3. In order to be able to correct all possible patterns of at most e errors it is again necessary and sufficient that the code C has minimum distance $\geq 2e + 1$. As before, that is equivalent to requiring that C has minimum weight at least $2e + 1$. These properties of the code C are based on the resemblance of C to the subgroup $H := \{AN \mid N \in \mathbb{Z}\}$ of $\mathbb{Z}$. It would not be a good idea to take H as our code because H has minimum weight ≤ 2 (see Problem 10.5.1).

In order to avoid this difficulty we shall consider so-called *modular AN codes*. Define $m := AB$. Now we can consider C as a subgroup of $\mathbb{Z}/m\mathbb{Z}$. This makes it necessary to modify our distance function. Consider the elements of $\mathbb{Z}/m\mathbb{Z}$ as vertices of a graph Γ_m and let $x \pmod m$ and $x' \pmod m$ be joined by an edge iff

$$x - x' \equiv \pm c \cdot r^j \pmod m$$

for some integers c, j with $0 < c < r, j \geq 0$.

(10.1.2) Definition. The *modular distance* $d_m(x, y)$ of two integers x and y (considered as elements of $\mathbb{Z}/m\mathbb{Z}$) is the distance of x and y in the graph Γ_m. The *modular weight* $w_m(x)$ of x is $d_m(x, 0)$. Note that

$$w_m(x) = \min\{w(y) \mid y \in \mathbb{Z}, y \equiv x \pmod m\}.$$

Although we now have achieved a strong resemblance to linear codes there is another difficulty. Not every choice of m makes good sense. For example, if we take $r = 3$, $A = 5$, $B = 7$, i.e. $m = 35$, then by (10.1.2) we have $d_m(0, 4) = 1$ because $4 \equiv 3^{10} \pmod{35}$. But it is not very realistic to consider errors in the position corresponding to 3^{10} when adding integers less than 35. Restricting j in the definition of edges of Γ_m also has drawbacks. It turns out that we get an acceptable theory if we take $m = r^n - 1$ ($n \in \mathbb{Z}$, $n \geq 2$). In practice this is also a good choice because many computers do arithmetic operations mod $2^n - 1$.

Every integer x has a unique representation

$$x \equiv \sum_{i=0}^{n-1} c_i r^i \pmod{r^n - 1},$$

with $c_i \in \{0, 1, \ldots, r-1\}$ $(0 \leq i < n)$, not all $c_i = 0$. Hence $\mathbb{Z}/(r^n - 1)$ can be interpreted as the set of nonzero words of length n over the alphabet $\{0, 1, \ldots, r-1\}$. Of course it would not have been necessary to exclude $\mathbf{0}$ if we had taken $m = r^n$, which is again a practical choice because many computers work mod 2^n. However, we cannot expect good codes for $r = 2$, $m = 2^n$. We would have to take $A = 2^k$ for some k and then the code C would consist of the integers $\sum_{i=0}^{n-1} c_i 2^i$, $c_i \in \{0, 1\}$, for which $c_0 = c_1 = \cdots = c_{k-1} = 0$. An integer $x \pmod B$ would be encoded by adding k 0s to its representation. This would serve no purpose. For arbitrary r there are similar objections. The reader should convince himself that in the case $AB = m = r^n - 1$ the modular distance is the natural function for arithmetic in $\mathbb{Z}/m\mathbb{Z}$ and that C behaves as a linear code. In fact we have an even stronger analogy with earlier chapters.

(10.1.3) Definition. A *cyclic AN code* of length n and base r is a subgroup C of $\mathbb{Z}/(r^n - 1)$. Such a code is a principal ideal in this ring, i.e. there are integers A and B such that $AB = r^n - 1$ and

$$C = \{AN \mid N \in \mathbb{Z}, 0 \leq N < B\}.$$

As in Section 6.1 we call A the *generator* of C. By now it will not be surprising to the reader that we are primarily interested in codes C with a large *rate* $(= (1/n)\log_r B)$ and a large minimum distance. The terminology of (10.1.3) is in accordance with (6.1.1). If $x \in C$ then $rx \pmod{r^n - 1}$ is also a codeword because C is a group and $rx \pmod{r^n - 1}$ is indeed a cyclic shift of x (both represented in base r). The integer B can be compared with the check polynomial of a cyclic code.

The idea of negacyclic codes can be generalized in the same way by taking $m = r^n + 1$ and then considering subgroups of $\mathbb{Z}/m\mathbb{Z}$.

(10.1.4) Example. Let $r = 2$, $n = 11$. Then $m = r^n - 1 = 2047$. We take $A = 23$, $B = 89$. We obtain the cyclic AN code consisting of 89 multiples of 23 up to 2047. There are 22 ways to make one error, corresponding to the integers $\pm 2^j$ $(0 \leq j < 11)$. These are exactly the integers mod 23 except 0. Therefore every integer in [1,2047] has modular distance 0 or 1 to exactly one codeword. This cyclic AN code is therefore *perfect*. It is a generalization of the Hamming codes.

§10.2 The Arithmetic and Modular Weight

In order to be able to construct AN codes which correct more than one error we need an easy way to calculate the arithmetic or modular weight of an integer.

By Definition 10.1.1 every integer x can be written as

$$x = \sum_{i=1}^{w(x)} a_i r^{n(i)}$$

with integers a_i, $n(i)$, $|a_i| < r$, $n(i) \geq 0$ $(i = 1, \ldots, w(x))$. It is easy to find examples which show that this representation is not unique. We shall put some more restrictions on the coefficients which will make the representation unique.

(10.2.1) Definition. Let $b \in \mathbb{Z}$, $c \in \mathbb{Z}$, $|b| < r$, $|c| < r$. The pair (b, c) is called *admissible* if one of the following holds

(i) $bc = 0$,
(ii) $bc > 0$ and $|b + c| < r$,
(iii) $bc < 0$ and $|b| > |c|$.

Note that if $r = 2$ we must have possibility (i). Therefore a representation $x = \sum_{i=0}^{\infty} c_i 2^i$ in which all pairs (c_{i+1}, c_i) are admissible has no two adjacent nonzero digits. This led to the name *nonadjacent form* (*NAF*) which we now generalize.

(10.2.2) Definition. A representation

$$x = \sum_{i=0}^{\infty} c_i r^i,$$

with $c_i \in \mathbb{Z}$, $|c_i| < r$ for all i and $c_i = 0$ for all large i is called an *NAF* for x if for every $i \geq 0$ the pair (c_{i+1}, c_i) is admissible.

(10.2.3) Theorem. *Every integer x has exactly one* NAF. *If this is*

$$x = \sum_{i=0}^{\infty} c_i r^i,$$

then

$$w(x) = |\{i \mid i \geq 0, c_i \neq 0\}|.$$

Proof.

(a) Suppose x is represented as $\sum_{i=0}^{\infty} b_i r^i$, $|b_i| < r$. Let i be the minimal value such that the pair (b_{i+1}, b_i) is not admissible. W.l.o.g. $b_i > 0$ (otherwise consider $-x$). Replace b_i by $b'_i := b_i - r$ and replace b_{i+1} by $b'_{i+1} := b_{i+1} + 1$ (if $b_{i+1} + 1 = r$ we carry forward). If $b_{i+1} > 0$ we either have $b'_{i+1} = 0$ or $b'_i b'_{i+1} < 0$ and $b'_{i+1} = b_{i+1} + 1 > r - b_i = |b'_i|$ since (b_{i+1}, b_i) was not admissible. If $b_{i+1} < 0$ then $b'_{i+1} = 0$ or $b'_i b'_{i+1} > 0$ and $|b'_i + b'_{i+1}| = r - b_i - b_{i+1} < r$ because $-b_{i+1} \leq b_i$ as (b_{i+1}, b_i) is not admissible. So (b'_{i+1}, b'_i) is admissible and one checks in the same way that (b'_i, b_{i-1}) is admissible. In this way we can construct an NAF and in the process the weight of the representation does not increase.

(b) It remains to show that the NAF is unique. Suppose some x has two such representations $x = \sum_{i=0}^{\infty} c_i r^i = \sum_{i=0}^{\infty} c_i' r^i$. W.l.o.g. we may assume $c_0 \neq c_0', c_0 > 0$. Therefore $c_0' = c_0 - r$. It follows that $c_1' \in \{c_1 + 1 - r, c_1 + 1, c_1 + 1 + r\}$. If $c_1' = c_1 + 1 - r$ then $c_1 \geq 0$ and hence $c_0 + c_1 \leq r - 1$. Since $c_0' c_1' > 0$ we must have $-c_0' - c_1' < r$, i.e. $r - c_0 + r - c_1 - 1 < r$, so $c_0 + c_1 > r - 1$, a contradiction. In the same way the assumptions $c_1' = c_1 + 1$ resp. $c_1' = c_1 + 1 + r$ lead to a contradiction. Therefore the NAF is unique. □

A direct way to find the NAF of an integer x is provided in the next theorem.

(10.2.4) Theorem. *Let $x \in \mathbb{Z}$, $x \geq 0$. Let the r-ary representations of $(r+1)x$ and x be*

$$(r+1)x = \sum_{j=0}^{\infty} a_j r^j, \qquad x = \sum_{j=0}^{\infty} b_j r^j.$$

with $a_j, b_j \in \{0, 1, \ldots, r-1\}$ for all j and $a_j = b_j = 0$ for j sufficiently large. Then the NAF *for x is*

$$x = \sum_{j=0}^{\infty} (a_{j+1} - b_{j+1}) r^j.$$

PROOF. We calculate the numbers a_j by adding $\sum_{j=0}^{\infty} b_j r^j$ and $\sum_{j=0}^{\infty} b_j r^{j+1}$. Let the carry sequence be $\varepsilon_0, \varepsilon_1, \ldots$, so $\varepsilon_0 = 0$ and $\varepsilon_i := \lfloor (\varepsilon_{i-1} + b_{i-1} + b_i)/r \rfloor$. We find that $a_i = \varepsilon_{i-1} + b_{i-1} + b_i - \varepsilon_i r$. If we denote $a_i - b_i$ by c_i then $c_i = \varepsilon_{i-1} + b_{i-1} - \varepsilon_i r$. We must now check whether (c_{i+1}, c_i) is an admissible pair. That $|c_{i+1} + c_i| < r$ is a trivial consequence of the definition of ε_i. Suppose $c_i > 0$, $c_{i+1} < 0$. Then $\varepsilon_i = 0$. We then have $c_i = \varepsilon_{i-1} + b_{i-1}$, $c_{i+1} = b_i - r$ and the condition $|c_{i+1}| > |c_i|$ is equivalent to $\varepsilon_{i-1} + b_{i-1} + b_i < r$, i.e. $\varepsilon_i = 0$. The final case is similar. □

The NAF for x provides us with a simple estimate for x as shown by the next theorem.

(10.2.5) Theorem. *If we denote the maximal value of i for which $c_i \neq 0$ in an* NAF *for x by $i(x)$ and define $i(0) := -1$ then*

$$i(x) \leq k \Leftrightarrow |x| < \frac{r^{k+2}}{r+1}.$$

We leave the completely elementary proof to the reader.

From Section 10.1 it will be clear that we must now generalize these ideas in some way to modular representations. We take $m = r^n - 1$, $n \geq 2$.

(10.2.6) Definition. A representation

$$x \equiv \sum_{i=0}^{n-1} c_i r^i \pmod{m},$$

with $c_i \in \mathbb{Z}$, $|c_i| < r$ is called a *CNAF (= cyclic NAF)* for x if (c_{i+1}, c_i) is admissible for $i = 0, 1, \ldots, n-1$; here $c_n := c_0$.

The next two theorems on CNAF's are straightforward generalizations of Theorem 10.2.3 and can be obtained from this theorem or by using Theorem 10.2.4. A little care is necessary because of the exception but the reader should have no difficulty proving the theorems.

(10.2.7) Theorem. *Every integer x has a* CNAF mod m; *this* CNAF *is unique except if*

$$(r+1)x \equiv 0 \not\equiv x \pmod{m}$$

in which case there are two CNAFs *for x* (mod m). *If $x \equiv \sum_{i=0}^{n-1} c_i r^i$* (mod m) *is a* CNAF *for x then*

$$w_m(x) = |\{i \mid 0 \leq i < n, c_i \neq 0\}|.$$

(10.2.8) Theorem. *If $(r+1)x \equiv 0 \not\equiv x$* (mod m) *then $w_m(x) = n$ except if $n \equiv 0$* (mod 2) *and $x \equiv \pm[m/(r+1)]$*(mod m), *in which case $w_m(x) = \frac{1}{2}n$.*

If we have an NAF for x for which $c_{n-1} = 0$ then the additional requirement for this to be a CNAF is satisfied. Therefore Theorem 10.2.5 implies the following theorem.

(10.2.9) Theorem. *An integer x has a* CNAF *with $c_{n-1} = 0$ iff there is a $y \in \mathbb{Z}$ with $x \equiv y$* (mod m), $|y| \leq m/(r+1)$.

This theorem leads to another way of finding the modular weight of an integer.

(10.2.10) Theorem. *For $x \in \mathbb{Z}$ we have $w_m(x) = |\{j \mid 0 \leq j < n$, there is a $y \in \mathbb{Z}$ with*

$$m/(r+1) < y \leq mr/(r+1),\ y \equiv r^j x \ (mod\ m)\}|.$$

PROOF. Clearly a CNAF for rx is a cyclic shift of a CNAF for x, i.e. $w_m(rx) = w_m(x)$. Suppose $x \equiv \sum_{i=0}^{n-1} c_i r^i$ (mod m) is a CNAF and $c_{n-1-j} = 0$. Then $r^j x$ has a CNAF with 0 as coefficient of r^{n-1}. By Theorem 10.2.9 this is the case iff there is a y with $y \equiv r^j x$ (mod m) and $|y| \leq m/(r+1)$. Since the modular weight is the number of nonzero coefficients the assertion now follows unless we are in one of the exceptional cases of Theorem 10.2.7 but then the result follows from Theorem 10.2.8. □

§10.3 Mandelbaum–Barrows Codes

We now introduce a class of cyclic multiple-error-correcting AN codes which is a generalization of codes introduced by J. T. Barrows and D. Mandelbaum. We first need a theorem on modular weights in cyclic AN codes.

(10.3.1) Theorem. *Let $C \subset \mathbb{Z}/(r^n - 1)$ be a cyclic* AN *code with generator A and let*

$$B := (r^n - 1)/A = |C|.$$

Then

$$\sum_{x \in C} w_m(x) = n\left(\left\lfloor \frac{rB}{r+1} \right\rfloor - \left\lfloor \frac{B}{r+1} \right\rfloor\right).$$

PROOF. We assume that every $x \in C$ has a unique CNAF

$$x \equiv \sum_{i=0}^{n-1} c_{i,x} r^i \pmod{r^n - 1}.$$

The case where C has an element with two CNAFs is slightly more difficult. We leave it to the reader. We must determine the number of nonzero coefficients $c_{i,x}$, where $0 \le i \le n-1$ and $x \in C$, which we consider as elements of a matrix. Since C is cyclic every column of this matrix has the same number of zeros. So the number to be determined is equal to $n|\{x \in C | c_{n-1,x} \neq 0\}|$. By Theorem 10.2.9 we have $c_{n-1,x} \neq 0$ iff there is a $y \in \mathbb{Z}$ with $y \equiv x \pmod{r^n - 1}$, $m/(r+1) < y \le mr/(r+1)$. Since x has the form AN $(\mathrm{mod}\ r^n - 1)$ $(0 \le N < B)$, we must have $B/(r+1) < N \le Br/(r+1)$. □

The expression in Theorem 10.1.3 is nearly equal to $n|C|[(r-1)/(r+1)]$ and hence the theorem resembles our earlier result

$$\sum_{\mathbf{x} \in C} w(\mathbf{x}) = n|C| \cdot \frac{q-1}{q}$$

for a linear code C (cf. (3.7.5)).

The next theorem introduces the generalized *Mandelbaum–Barrows codes* and shows that these codes are *equidistant.*

(10.3.2) Theorem. *Let B be a prime number that does not divide r with the property that $(\mathbb{Z}/B\mathbb{Z})$ is generated by the elements r and -1. Let n be a positive integer with $r^n \equiv 1 \pmod{B}$ and let $A := (r^n - 1)/B$. Then the code $C \subset \mathbb{Z}/(r^n - 1)$ generated by A is an equidistant code with distance*

$$\frac{n}{(B-1)}\left(\left\lfloor \frac{rB}{r+1} \right\rfloor - \left\lfloor \frac{B}{r+1} \right\rfloor\right).$$

PROOF. Let $x \in C$, $x \neq 0$. Then $x = AN \pmod{r^n - 1}$, with $N \not\equiv 0 \pmod{B}$. Our assumptions imply that there is a j such that $N \equiv \pm r^j \pmod{B}$. Therefore $w_m(x) = w_m(\pm r^j A) = w_m(A)$. This shows that C is equidistant and then the constant weight follows from Theorem 10.3.1. □

The Mandelbaum–Barrows codes correspond to the minimal cyclic codes M_i^- of Section 6.2. Notice that these codes have word length at least $\frac{1}{2}(B-1)$ which is large with respect to the number of codewords which is B. So, for practical purposes these codes do not seem to be important.

§10.4 Comments

The reader interested in more information about arithmetic codes is referred to W. W. Peterson and E. J. Weldon [53], J. L. Massey and O. N. Garcia [48], T. R. N. Rao [58]. Perfect single error-correcting cyclic AN codes have been studied extensively. We refer to M. Goto [28], M. Goto and T. Fukumara [29], and V. M. Gritsenko [31]. A perfect single error-correcting cyclic AN code with $r = 10$ or $r = 2^k$ $(k > 1)$ does not exist.

For more details about the NAF and CNAF we refer to W. E. Clark and J. J. Liang [14], [15]. References for binary Mandelbaum–Barrows codes can be found in [48]. There is a class of cyclic AN codes which has some resemblance to BCH codes. These can be found in C. L. Chen, R. T. Chien and C. K. Liu [12].

For more information about perfect arithmetic codes we refer to a contribution with that title by H. W. Lenstra in the Séminaire Delange–Pisot–Poitou (Théorie des Nombres, 1977/78).

§10.5 Problems

10.5.1. Prove that $\min\{w(AN) | N \in \mathbb{Z}, N \neq 0\} \le 2$ for every $A \in \mathbb{Z}$ if w is as defined in (10.1.1).

10.5.2. Generalize (10.1.4). Find an example with $r = 3$.

10.5.3. Consider ternary representations mod $3^6 - 1$. Find a CNAF for 455 using the method of the proof of Theorem 10.2.3.

10.5.4. Determine the words of the Mandelbaum–Barrows code with B = 11, $r = 3$, $n = 5$.