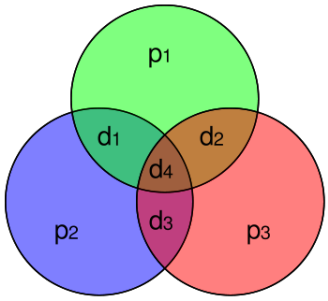
# Smart Codes II

UCLA Math Circle

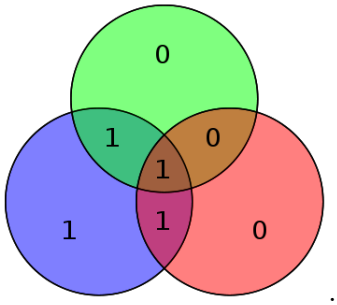
10/07/2012

## 1 Hamming Codes and Hamming Distance

Last time, we saw an efficient way to encode words with 4 bits using 7-bit long codeword. Suppose the 4-bit word is $\overline{d_1d_2d_3d_4}$. Define three bits $p_1, p_2, p_3 \in \{0, 1\}$ such that in the diagram below, four bits in the same circle add up to 0.




The codeword is then $\overline{p_1p_2d_1p_3d_2d_3d_4}$. For example, if the bits are 1011, then the diagram above becomes




.

So $p_1 = 0, p_2 = 1, p_3 = 0$ and the codeword is 0110011.

This encoding method can correct single-digit errors. In fact, given a 7-bit word $\overline{a_1a_2\ldots a_7}$, we can define the check bits $c_1, c_2, c_4$ by

$$c_1 = a_1 + a_3 + a_5 + a_7,$$
$$c_2 = a_2 + a_3 + a_6 + a_7,$$
$$c_4 = a_4 + a_5 + a_6 + a_7.$$

Then $\overline{a_1a_2\dots a_7}$ is a codeword if $c_1 = c_2 = c_4 = 0$, and the binary number $\overline{c_4c_2c_1}$ gives the position of the error if $\overline{a_1a_2\dots a_7}$ differs from a codeword by a single digit. For example, if we received 0111011 instead of the codeword 0110011, then we can calculate $c_1 = 0, c_2 = 0, c_4 = 1$ and the binary number 100, which is 4 in base 10, tells us that the $4^{\text{th}}$ digit is incorrect. Hamming's $[7, 4]$-code is a special case of Hamming's code, which we will describe in the next section.

To describe this ability to correct single-digit errors, it is useful to define the concept of *Hamming distance*. Let $\mathbb{F}_2^n$ be the set of all words with length $n$ composed of the digits 0 and 1, i.e.

$$\mathbb{F}_2^n := \{(a_1, a_2, \dots, a_n) : a_k \in \{0, 1\} \text{ for all } 1 \leq k \leq n\}.$$

The ***Hamming Distance*** between two words $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ in $\mathbb{F}_2^n$ is defined by

$$d_H(x, y) := \#\{j : x_j \neq y_j\}. \tag{1}$$

For example, the Hamming distance between $(1, 0, 1, 0), (0, 0, 0, 1) \in \mathbb{F}_2^4$ is 3 since they differ in 3 digits (the first, third and fourth digits). As a measure of closeness, the Hamming distance satisfies properties that are familiar to us

**Proposition 1.** *For words $x, y, z \in \mathbb{F}_2^n$, we have*

1. $d_H(x, y) \geq 0$ *with equality if and only if* $x = y$.
2. $d_H(x, y) = d_H(y, x)$.
3. $d_H(x, z) \leq d_H(x, y) + d_H(y, z)$.

Given the concept of distance between two points in $\mathbb{F}_2^n$, we can make sense of a lot of concepts familiar to us in the Euclidean world. For example, given a point $x \in \mathbb{F}_2^n$ and a non-negative integer $r$, the ***ball of radius*** $r$ ***and center*** $x$ is denoted by $B_r(x)$ and defined by

$$B_r(x) := \{y \in \mathbb{F}_2^n : d_H(x, y) \leq r\}.$$

Since the set $\mathbb{F}_2^n$ is finite, each ball $B_r(x)$ contains finitely many points. The following lemma tells us how many points there are in the ball $B_r(x)$.

**Lemma 1.** *For any integer $0 \leq r \leq n$, the number of points in the ball $B_r(x)$ is*

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{r}, \tag{2}$$

*which is independent of the center $x$.*

Since a code of length $n$ is a subset $C \subseteq \mathbb{F}_2^n$, we define the minimum distance of a code as $d_H(C)$ by

$$d_H(C) := \min_{x,y \in C} d_H(x, y). \tag{3}$$

If we want to fill a square box with balls of the same size in every day life, there will be empty spaces not covered by any ball. However, it is possible to cover the entire set $\mathbb{F}_2^n$ with balls without any overlapping. To cover $\mathbb{F}_2^7$ for example, we can take all the balls with center being a codeword in Hamming's $[7, 4]$-code and radius 1. Then every string of 0's and 1's in $\mathbb{F}_2^7$ falls within exactly one of these balls. This makes Hamming's $[7, 4]$-code a perfect code.

**Definition 1.** *A code $C \subseteq \mathbb{F}_2^n$ is called **perfect** if there is an integer $t$ such that for all $x \in \mathbb{F}_2^n$ there is a unique codeword $c \in C$ satisfying $d_H(x, c) \leq t$.*

***Exercise 1*** List all the codewords in Hamming's $[7, 4]$-code and find the minimum distance of this code.

***Exercise 2*** Prove Proposition 1 above.

***Exercise 3*** Prove Lemma 1 above.

***Exercise 4*** Check if the following codes are perfect. If so, what is the minimum distance of the code?

- Trivial code $C = \mathbb{F}_2^n$.
- $C = \{00, 11\} \subset \mathbb{F}_2^2$.
- $C = \{000, 111\} \subset \mathbb{F}_2^3$.
- Hamming's $[7, 4]$-code, which is a subset of $\mathbb{F}_2^7$.

# 2 Linear Codes, Generator and Parity Matrix

We know that a code is determined by its codewords, which form a subset $C$ of $\mathbb{F}_2^n$ for some positive integer $n$. It is cumbersome to store all the codewords when the size of $C$ is large. For certain codes though, it is enough to store $k$ of the codewords to generate all the other codewords, for some integer $0 \leq k \leq n$. These codes and their properties will be the topic of this section.

First, recall the addition law in the set with two element $\mathbb{F}_2 = \{0, 1\}$ is given by

$$0 + 0 = 0, 1 + 0 = 1, 1 + 1 = 0. \tag{4}$$

If we treat elements in $\mathbb{F}_2^n$ as an $n$-tuple of 0's and 1's, we can define addition on elements $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n) \in \mathbb{F}_2^n$ via

$$x + y := (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n) \in \mathbb{F}_2^n. \tag{5}$$

For example, suppose $x = (0, 1, 0, 1), y = (0, 1, 1, 0) \in \mathbb{F}_2^4$, then

$$x + y = (0 + 0, 1 + 1, 0 + 1, 1 + 0) = (0, 0, 1, 1) \in \mathbb{F}_2^4.$$

Using this concept of addition, we can define a class of code called linear code.

**Definition 2.** *A code $C \subset \mathbb{F}_2^n$ is a **linear code** if for all $x, y \in C$, we have $x + y \in C$.*

Here is simple example of linear code:

$$C = \{000000, 100011, 010101, 001110, 011011, 101101, 110110, 111000\} \subset \mathbb{F}_2^6 \tag{6}$$

To find the minimum distance of a general code $C \subseteq \mathbb{F}_2^n$, one has to vary $x, y$ over all of $\mathbb{F}_2^n$, calculate $d_H(x, y)$, and find the minimum. This requires calculating the Hamming distance $|C| \cdot (|C| - 1)/2$ times. If $C$ is a linear code, then it is much easier to find the minimum distance $d_H(C)$. Also, there is a restriction on the size of $C$.

**Proposition 2.** *The minimum distance $d_H(C)$ of a linear code $C \subseteq \mathbb{F}_2^n$ is the least number of 1's contained in a nonzero codeword in $C$. Furthermore, the size of $C \subseteq \mathbb{F}_2^n$ is of the form $2^k$ for some integer $0 \leq k \leq n$.*

For a linear code $C$ of size $2^k$, we can generate all of $C$ from $k$ codewords by repeated additions. In example (6), we can use $x_1 = 100011, x_2 = 010101, x_3 = 001110$ to generate the rest of the codewords since

$$x_1 + x_1 = 000000, x_2 + x_3 = 011011, x_1 + x_3 = 101101,$$
$$x_1 + x_2 = 110110, x_1 + x_2 + x_3 = 111000.$$

There are many choices of such a subset of $k$ codewords. Any such choice is called a ***basis*** of the code $C$. The $k \times n$ matrix whose $k$ rows form a basis of the code $C$ is called a ***generator matrix***. For example, since $\{100011, 010101, 001110\}$ is a basis of the linear code in example (6), the matrix below is a generator matrix

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

Now we can encode strings in $\mathbb{F}_2^k$ by treating it as a $1 \times k$ matrix and multiplying the $k \times n$ generator matrix on the right. For example, if $x = (1, 0, 1) \in \mathbb{F}_2^3$, then the corresponding codeword is

$$(1,0,1) \cdot \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix} = (1,0,1,1,0,1) \in C.$$

Suppose we are given $x \in \mathbb{F}_2^n$ and want to check if it is a codeword in $C \subseteq \mathbb{F}_2^n$. If $C$ is not linear, we generally need to compare $x$ with all the codewords in $C$. However if $C$ is linear, then there is a simple way to tell whether of not $x \in C$. First, we can find a generator matrix $G_C$ of $C$. Then define the complementary code $C^\perp \subseteq \mathbb{F}_2^n$ by

$$C^\perp := \{x \in \mathbb{F}_2^n : G_C \cdot x = (0, 0, \ldots, 0) \in \mathbb{F}_2^k\}. \tag{7}$$

Here $\cdot$ is the usual matrix multiplication. It turns out that $C^\perp$ is also a linear code, so we can find its generator matrix $G_{C^\perp}$. This is called the ***parity matrix*** of the code $C$, and can be used to test if $y \in \mathbb{F}_2^n$ is in $C$.

**Proposition 3.** *An element $y \in \mathbb{F}_2^n$ is a codeword in $C$ if and only if $G_{C^\perp} \cdot y = 0$.*

From the definition of the generator matrix and Proposition 3, it is clear that either the generator matrix or the parity matrix *uniquely* determines the linear code $C$. Now let $n = 2^r - 1$ for a positive integer $r$. Let $H$ be the $r \times n$ matrix whose columns are the non-zero elements in $\mathbb{F}_2^r$. A code whose parity matrix is $H$ is called a ***Hamming code***. For example, Hamming's $[7, 4]$-code is a special case of Hamming code with $r = 3, n = 2^3 - 1 = 7$. In this case, the non-zero elements in $\mathbb{F}_2^3$ are

$$\{(1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1)\}.$$

They form the matrix

$$H = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}. \tag{8}$$

Given any $x \in \mathbb{F}_2^7$, one can check that $H \cdot x = (0, 0, 0)$ if and only if $x$ is a codeword in Hamming's $[7, 4]$-code.

***Exercise 5*** Show that the addition law defined in (5) has an additive identity and additive inverse for each element in $\mathbb{F}_2^n$. Using these to define subtraction.

***Exercise 6*** Given a linear code $C \subseteq \mathbb{F}_2^n$, show that $(0, 0, \ldots, 0) \in C$. In general, using the subtraction defined in the previous exercise, show that $x, y \in C$ implies $x - y \in C$.

***Exercise 7*** Check whether the code $C \subset \mathbb{F}_2^6$ defined in (6) is linear and compute its minimum distance $d_H(C)$. If it is linear, find a basis and compute the generator and parity matrix using this basis. Do the same for Hamming's $[7, 4]$-code and the following codes

$$C_1 = \{0000, 1000, 0100, 1100\} \subset \mathbb{F}_2^4,$$
$$C_2 = \{0000, 1010, 0101, 1001\} \subset \mathbb{F}_2^4.$$

***Exercise 8*** Use the parity matrix in the previous exercise to check if 0101010 is a codeword in Hamming's $[7, 4]$-code. If not, what is the codeword closest to it? What is their Hamming distance?

***Exercise 9*** Let $H$ be as in equation (8). Check that $H \cdot x = (0, 0, 0)$ if and only if $x$ is a codeword in Hamming's $[7, 4]$-code.

***Exercise 10*** Find a parity matrix of the Hamming code with $n = 15, r = 4$. Check if 010101010101010 is a codeword in this code. If not, find the nearest codeword and their Hamming distance.

***Exercise 11*** Prove Propositions 2 and 3

***Exercise 12*** Try to generalize some of the concepts above (such as the addition law (5), linear code, generator and parity matrix) to the setting where strings have digits in the set $\mathbb{F}_p := \{0, 1, 2, \ldots, p-1\}$ for a prime number $p$. What are the differences?