CSE 533: Error-Correcting Codes **(Autumn 2006)**

Lecture 2: More on Linear Codes

6 October 2006

Lecturer: Venkatesan Guruswami *Scribe: Dan Halperin*

Definitions and notation from Lecture 1

We use Σ to denote the *alphabet*, the set of symbols we use to write *codewords*. The *block length*, or length of codewords is n. The set of codewords is a *code* $C \subseteq \Sigma^n$.

A code C distinguishes between $|C|$ messages and therefore encodes $\log |C|$ bits of information. In an error-free world, you could encode $n \log |\Sigma|$ bits using all of Σ^n. We define the *rate* of a code C, $R(C)$, to be the ratio of the information it encodes to the maximum possible:

$$R(C) = \frac{\log |C|}{n \log |\Sigma|}.$$

The (minimum) *distance* of a code C, $d(C)$, is the minimum Hamming distance between any two of its codewords,

$$d(C) = \min_{x,y \in C, x \neq y} \Delta(x, y).$$

We sometimes like to normalize the distance of a code to its block length, as we do with rate. To do so, we define the *relative distance* of a code C, denoted $\delta(C)$, as the ratio of its distance to its block length:

$$\delta(C) = \frac{d(C)}{n}.$$

Given an $n \times k$ *generator matrix* over $\mathbb{F}_q$ of full rank G, we defined a *linear code* C

$$C = \{Gx \mid x \in \mathbb{F}_q^k\}.$$

We say that C is an $[n, k, d]_q$ code, where we may omit d and/or q if their values are understood or not relevant to the discussion.

1 The $[7, 4, 3]_2$ Hamming Code and its Parity Check Matrix

We defined the $C_0 = [7, 4, 3]_2$ Hamming code using generator matrix

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

Lemma 1.1. $C_0 = \{x \in \mathbb{F}_2^7 \mid Hx = 0\}$ *for*

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

Proof. Let $D = \{x \in \mathbb{F}_2^7 \mid Hx = 0\}$. We will show that $C_0 \subseteq D$ and that $D \subseteq C_0$.

Observe that $HG = 0$. Then for $c \in C_0$, $Hc = HGx$ for some $x \in \mathbb{F}_2^4$ and $HGx = (HG)x = 0x = 0$. Thus $C_0 \subseteq D$.

In linear algebraic terms, D is simply the null space of H. For the reverse direction, note that H has full rank 3 and that its null space must therefore have dimension 4. However, G is in the null space of H and has dimension 4 and thus spans D, showing that $D \subseteq C_0$. □

In general, an $[n, k]$ linear code $C \subseteq \mathbb{F}^n$ can be described as

$$C = \{c \in \mathbb{F}^n \mid Hc = 0\}$$

for an $(n - k) \times n$ matrix H of rank $n - k$. Such an H is called a *parity check matrix*.

Correcting One Bit Flip in the Hamming code: Suppose that y is a noisy transmission of codeword $c \in C_0$, that is y is c with a single bit flipped. We have not yet discussed any practical ways to recover c from y. A simple method with which we could do this would be to flip each bit of y and see if the resulting vector is in the null space of H.

We can represent y as $c + e_i$, where e_i is the column vector of all zeros except a single 1 in the ith position. The method above recovers i by brute force, requiring up to n matrix-vector multiplications. A more clever way to correct y is to simply calculate

$$Hy = H(c + e_i) = Hc + He_i = He_i = \text{the } i\text{th column of } H.$$

The ith column of H is the binary representation of i, and thus this method can recover the value of i with only a single multiplication.

2 Generalized Hamming Codes

Define H_r to be the $r \times (2^r - 1)$ matrix where column i of H_r is the binary representation of i. This matrix must contain e_1 through e_r, which are the binary representations of all powers of two from 1 to 2^{r-1}, and thus has full rank.

Now we can define

$$C_r^{\text{Ham}} = \{c \in \mathbb{F}_2^{2^r-1} \mid H_r c = 0\}.$$

C_r^{Ham} is an $[n_r = 2^r - 1, k_r = 2^r - 1 - r]$ binary linear code. We would like to know what the distance of this code is, but in order to compute that, we will need the following lemma.

Lemma 2.1. *The minimum distance of a linear code is the minimum Hamming weight of a nonzero codeword.*

Proof. Let C be a linear code with minimum distance d. Then there exist $x, y \in C$ such that $\Delta(x, y) = d$. Since C is linear, $x - y$ is a (nonzero) codeword and $wt(x - y) = \Delta(x, y) = d$. Thus d is at least as large as the minimum Hamming weight of a nonzero codeword.

However, C must contain the zero codeword, and thus d is by definition as small as the minimum over all codewords c of $\Delta(c, 0) = wt(c)$. □

Then to know the distance of the Hamming code, we must determine the minimum weight of a nonzero codeword. Since the codewords are defined as the vectors c such that $H_r c = 0$, this is equivalent to determining the size of the smallest linearly dependent set of columns of H_r. H_r does not contain a zero column, nor does it contain two equal columns, so this set must contain at least 3 elements. However, the first three columns of H_r represent 1, 2, and 3 in binary and sum (mod 2) to zero. This shows that the Hamming code in fact has distance 3.

Fact 2.2. *The Hamming code is the best possible code (in terms of k) with distance 3. In other words, the Hamming code has optimal rate.*

Proof. Consider a distance 3 code of block length n with 2^k codewords. Within distance 1 of each codeword there are $n + 1$ different points, n corresponding to the n possible bit flips and one corresponding to flipping no bits. There are only 2^n points in the entire space, and thus n and k must satisfy

$$2^k(n + 1) \leq 2^n,$$

implying that

$$k \leq n - \log_2(n + 1).$$

The Hamming codes C_r^{Ham} with $k_r = 2^r - 1 - r$ and $n_r = 2^r - 1$ achieve equality of this bound. □

Note that in general, the volume of a Hamming ball of radius d is

$$\sum_{i=0}^{d} \binom{n}{i},$$

and that using this in conjunction with the above logic we obtain the following lemma, called the *Volume bound* or the *Sphere-Packing boun* or the *Hamming bound*:

Lemma 2.3. *A code of block length n and distance d has at most*

$$\frac{2^n}{\sum_{i=0}^{\lfloor \frac{d-1}{2} \rfloor} \binom{n}{i}}$$

codewords.

Codes which meet this bound are called *perfect codes*. It has been shown that for binary linear codes, the Hamming codes, the trivial code ($[n, 1, n]_2$ code with codewords 0^n and 1^n for odd n), and the Golay code [23,12] are the only perfect codes.

3 The Dual

The *dual* of a code C, denoted $C^\perp$, is

$$C^\perp = \{c' \mid c' \cdot c = 0 \text{ for all } c \in C\}.$$

For a linear code C with parity check matrix H, $C^\perp$ is the linear code with generator matrix H^T. If C is an $[n, k]$ code, $C^\perp$ is an $[n, n-k]$ code.

The dual of the Hamming code is a linear code with parameters $[2^r - 1, r]$ with a generator matrix whose rows are all the nonzero r-bit vectors. This is the Simplex code. If we include the zero row, we obtain the Hadamard code $[2^r, r]$. The Hadamard code is the most redundant possible linear code in which no codeword bit repeats in every codeword.

We saw earlier that the Hamming code has optimal rate, but its relative distance is $\frac{3}{2^r}$. The Hadamard and Simplex codes have the awful rate $\frac{r}{2^r}$, which goes to zero as r increases, but they make up for this by having a very large distance:

Fact 3.1. *The Hadamard and Simplex codes have distance* $\frac{n}{2}$.

Proof. Let a be any nonzero element of $\mathbb{F}_2^r$ and i be an index such that $a_i \neq 0$. For $x \in \mathbb{F}_2^r$, $a \cdot x + a \cdot (x + e_i) = a \cdot (x + x + e_i) = a \cdot e_i = 1$, implying that $a \cdot x \neq a \cdot (x + e_i)$. Thus if we partition the 2^r numbers into 2^{r-1} pairs $(x, x + e_i)$, then for each pair one of $a \cdot x$ and $a \cdot (x + e_i)$ equals 1. This implies that exactly half of the bits of the encoding of a, $H_r^T a$, will be 1, and that for any nonzero a, $wt(H_r^T a) = 2^{r-1}$. Therefore every nonzero codeword of the Hadamard/Simplex codes has weight 2^{r-1} and by an earlier lemma, these codes have distance $2^{r-1} = \frac{n}{2}$. □

We will see more about these codes and their special properties relating to distance in a future lecture.

Conclusion

The comparison between the Hamming and Hadamard/Simplex codes show the two endpoints of the spectrum implied by the Hamming bound. The Hamming code has optimal rate but low relative distance, and the Hadamard/Simplex codes have poor rate but optimal distance. An interesting question is

Are there codes that have good rate and relative distance?

More specifically, is there an infinite family of $[n_i, k_i, d_i]$ binary codes of increasing block lengths n_i where

$$\frac{k_i}{n_i} \geq R \text{ and } \frac{d_i}{n_i} \geq \delta$$

for some $(R, \delta) > 0$? If so, what pairs (R, δ) are possible? Finally, from a practical standpoint, are any of these codes nicely structured, i.e., do we know what they are and how to describe and implement them efficiently?