1 Introduction

Coding theory was introduced by Hamming in 1948, and by Shannon in 1950. Coding theory addresses the following problem: Consider two parties, the sender (Alice) and the receiver (Bob). Alice picks up a message $m \in \{0, 1\}^k$, encodes the message as $E(m) \in \{0, 1\}^n$ and transmits $E(m)$ over a noisy channel. Bob receives $E(m) + \text{noise} = C'$, and decodes $C'$ to obtain $m$.

In this course we look for answers for the following questions:

1. What is possible to achieve?
2. What kinds of codes exist?

Plan for this course:

1. Results from Hamming’s work.
2. Results from Shannon’s work.
3. Efficient coding algorithms.
4. Applications of codes in complexity theory and in cryptography.

2 Hamming Codes

Let the message space be $\{0, 1\}^k$ and let $m \in \{0, 1\}^k$ be a message. We write $m$ as $m_1 m_2 \ldots m_k$, where $m_i \in \{0, 1\}$. Let us assume the channel introduces a 1-bit error. Send three copies of the message $m$ to the receiver. The receiver reconstructs the message by taking the majority of each bit. These codes are called repetition codes. These codes can tolerate 1-bit of error correction. The size of the code is three times the size of the message. The redundancy is large. Can we do something better?

If we add parity of the message to the message, then this provides error detection but does not provide error correction.

Consider $m \in \{0, 1\}^4$. Construct a $7 \times 4$ matrix as

$$G = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 1
\end{pmatrix}.$$
We define \( y = E(m) = Gm \). For such a \( G \), we can find a \( H \) such that \( HG = 0 \).

**Error Correction:** Given a vector \( y' \) where \( y' = y + e_i \), where \( e_i = (0, \ldots, 1, \ldots, 0) \), how do we find \( i \), the location of the error?

**Hamming's approach:**

\[
H y' = H(y + e_i) \\
= Hy + He_i \\
= HGm + He_i \\
= 0 + He_i
\]

where \( i \) denotes the index of the error location. \( H \) can be chosen as

\[
H = \begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1
\end{pmatrix}
\]

Then \( He_i \) represents the binary representation of \( i \).

A naive way is to try to flip the \( i \)th bit of \( y' \), compute \( Hy' \) and check if \( Hy' = 0 \).

### 3 Definitions and Notations

**Hamming Distance:** If \( x, y \in \Sigma^n \), the Hamming distance between \( x \) and \( y \) is given by

\[
\Delta(x, y) = |\{i : x_i \neq y_i\}|.
\]

**Hamming Weight:** For \( x \in \Sigma^n \), the Hamming weight of \( x \) is given by the number of non-zero coordinates of \( x \). That is,

\[
wt(x) = |\{i : x_i \neq 0\}|,
\]

\[
wt(x - y) = \Delta(x, y).
\]

**Notation:** An error correction code \( C \subseteq \Sigma^n \) for some positive \( n \). The parameter associated with code \( C \) is

- \( n \)—block length
- \( k \)—message length
- \( d \)—minimum distance, defined as
  \[
d(C) = \min_{x, y \in C} \Delta(x, y)
\]

and \( \frac{d(C)}{n} \) denotes the relative distance.

- \( \Sigma \)—alphabet set
- \( q \)—alphabet size \( |\Sigma| \)

**Lemma 3.1.** If \( d(C) = 2t + 1 \), then one can correct up to \( t \) errors.
**Linear Code:** C is called a linear code if it is a vector space over some field $F$.

**Generator Matrix:** Let $C \subseteq F_q^n$ be a linear space of dimension $k$. Let $x_1, \ldots, x_k \subseteq F_q^n$ be a basis of $C$. That is,

$$C = \{Gm | m \in F_q^k\}$$

where $G \in F_n^{k \times n}$ is a generator matrix of $C$.

**Parity-check matrix:** Since $C$ is a subspace of dimension $k$, we can find a subspace $C^\perp$ of dimension $n-k$ such that $\forall x \in C$ and $y \in C^\perp$, $<x, y> = 0$. Let $H^\top \in F_q^{(n-k) \times n}$ be the generator matrix for $C^\perp$. $H$ is called the parity-check matrix for $C$. We can write $C = \{y | Hy = 0\}$.

**Lemma 3.2.**

(a) $C$ has distance $\geq 3$.

(b) $C$ has block length $2^r - 1$.

(c) $C$ has information rate $2^r - 1 - r$.

Proof of (a): We first show that the minimum distance of a linear code is the minimum weight of a non-zero codeword. That is, $d(C) = \min_{x \in C} \text{wt}(x)$. Since $0 \in C$, $\Delta(x, 0) = \text{wt}(x)$.

$$d(C) \leq \min_{x \in C} \text{wt}(x)$$

If $x$ and $y$ are two codes, then $x - y$ is a code. Hence, $d(C) = \min_{x \in C} \text{wt}(x)$. Let $\Sigma = \{0, 1\}$. Since $C = \{y | yH = 0\}$, if the matrix $H$ is such that $yH \neq 0$ for any vector of weight 1 or 2 then the code will have distance $\geq 3$. If all the columns $h_i$ of $H$ are non-zero then dist $\geq 2$. If $h_i \neq h_j$, $i \neq j$ then $yH = h_i \neq h_j \neq 0$. For such a $H$, dist$(C) \geq 3$.

The following lemma is called volume bound or sphere-packing bound or the Hamming bound.

**Lemma 3.3.** If $C$ has block length $n$, and distance $2d + 1$, then $|C| \leq \frac{2^n}{\sum_{i=0}^{d} \binom{n}{i}}$.

**Proof:** If $d(C) = 2d + 1$ then at most $d$ errors made during the transmissions of the codeword can be corrected. Let $B(y, d)$ denote a ball of radius $d$ centered at $y$. Let $Vol(d, n) = \sum_{i=0}^{d} \binom{n}{i}$ be the volume of $B(y, d)$.

If $x$ and $y$ are two codewords of $C$, then $B(x, d)$ and $B(y, d)$ are disjoint since $\cup_{x \in C} B(x, d) \subseteq \{0, 1\}^n$,

$$\sum_{x \in C} |B(x, d)| \leq 2^n,$$

$$\implies |C| \cdot Vol(d, n) \leq 2^n$$

$$\implies |C| \leq \frac{2^n}{\sum_{i=0}^{d} \binom{n}{i}}$$
**Perfect Codes:** Codes that satisfy the Hamming Bound exactly are called perfect codes.

**Dual code:** Dual of a code $C$ is given by

$$C^\perp = \{ x \mid < x, y >= 0 \forall y \in C \}$$

If $C$ is a linear code with parity check matrix $H$ then $C^\perp$ is a linear code with generator matrix $H^\perp$.

The dual of the Hamming code is the Simplex code with block length $2^r - 1$ and message length $r$, whose generator matrix are all nonzero $r$-bit vectors.

We obtain the Hadamard code by adding the zero vector to the generator matrix of the Simplex code. It has distance $n/2$, blocklength $2^r$ and information rate $r$-bit messages. As we can see, the redundancy increases in both the Simplex code as well as the Hadamard code.