6.440: Essential Coding Theory Spring 2013

Lecture 1 — February 6, 2013

Prof. Madhu Sudan *Scribe: Badih Ghazi*

1 Administrative

1. Instructor: Prof. Madhu Sudan; Email: madhu@mit.edu
2. Course website: http://people.csail.mit.edu/madhu/ST13/
3. Get added to the mailing list.
4. Course Requirements:
   (a) Hand in 3-4 problem sets.
   (b) Scribe at least one lecture.
   (c) Complete a course project.

2 History

Two seminal papers in the field of coding theory are those of Claude Shannon and Richard Hamming, namely:

1. The 1948 paper "A mathematical theory of communication" [Sha48] of Shannon where he defined the notion of information.
2. The 1950 paper "Error detecting and error correcting codes" [Ham50] of Hamming.

The two papers are very interelated but they have different prespectives on what is an error and how to correct it. In the Shannon model, the errors are drawn from a probability distribution whereas they are chosen adversarially in the Hamming model. In the rest of this lecture, we will introduce the Haming model. First, we give an example with a concrete setting of parameters before giving the formal definitions.

3 Example

Assume that we have a storage medium that can store a block of 63 bits and that in the course of one week, one arbitrary (adversarially chosen) bit gets flipped. We would like to have a mechanism that determines which bit got flipped and figures out what the original piece of information was.

3.1 The repetition code

In order to encode a message, the repetition code repeats each symbol three times. For instance, the message $[a\ b\ c\ d]$ is encoded by $[a\ a\ a\ b\ b\ b\ c\ c\ c\ d\ d\ d]$. Now, assume that an arbitrary bit of the codeword gets flipped. Then, by taking a majority vote in each of the 4 triplets, we can recover the original message. Note that for every bit of information, we store 3 bits on the storage device, so the rate of this code is 1/3. Note that the repetition code can recover some patterns of 2 bit flips (in fact, it can recover all patterns that have at most one bit flipped in each block) but it cannot recover *all* patterns of 2 bit flips. Before the works of Hamming and Shannon, the common belief was that the repetition code was roughly the best code that one can construct. In his 1950 paper, Hamming gave 2 schemes that improve on the repetition code and that we describe next.

3.2 A better code due to Hamming

We divide the storage space of 63 bits into 9 blocks of 7 bits each. Each block of 7 bits will encode a seperate 4-bit message as follows:

$$[a\ b\ c\ d] \begin{bmatrix} 1&0&0&0&1&1&0 \\ 0&1&0&0&1&0&1 \\ 0&0&1&0&0&1&1 \\ 0&0&0&1&1&1&1 \end{bmatrix} = [a\ b\ c\ d\ (a \oplus b \oplus d)\ (a \oplus c \oplus d)\ (b \oplus c \oplus d)]$$

Note that if we take 2 different bit patterns (a, b, c, d) and (a', b', c', d'), the corresponding 7-bit blocks will differ in at least 3 bits. So this code can correct one bit flip. Moreover, it uses 7 bits of storage for every 4 bits of information so its rate is 4/7, which is an improvement over the repetition code. Is this the highest rate code that can correct one bit flip? It turned out that there is a better code that we describe next.

3.3 The Hamming code with block length 63

Hamming constructed a matrix $G \in \mathbb{F}_2^{57\times 63}$ s.t. for all $x, y \in \mathbb{F}_2^{57\times 63}$, the distance between xG and yG is at least 3. Thus, this code can correct one bit flip. Moreover, its rate is 57/63 which is an improvement over both previous codes. Remarkably, Hamming proved that this code is optimal in the sense that we cannot encode more than 2^{57} messages using 63 bits of space while requiring that every 2 of the codewords differ by at least 3 bits.
Moreover, Hamming constructed a matrix $H \in \mathbb{F}_2^{63\times 6}$ s.t. for all $y \in \mathbb{F}_2^{63}$, $yH = 0$ if and only if y is a codeword. It turns out that the matrix H has a special form; its 63 rows consist of all non-zero binary strings of length 6. Thus, the matrix H has the following two nice features:

1. H can be used to determine whether a given element y of $\mathbb{F}_2^{63}$ is a codeword or not. We just multiply y by H and check whether the product is 0 or not.
2. If only the ith bit is flipped, then the product will be equal to the ith row of H which is the binary representation of the integer i.

The matrix G is usually called the generator matrix and the matrix H is called the parity check matrix.

4 Formal definitions

We now formalize the notions introduced in the previous section.

4.1 Codes, distance and error correction

Definition 1. *(Error correcting code)*
Let n be a positive integer and Σ a finite set. An error correcting code C of block length n over the alphabet Σ is a subset of Σ^n. Σ^n is usually called the "ambient space".

Definition 2. *(Hamming distance)*
Given $x, y \in \{0,1\}^n$, the Hamming distance between x and y is given by $\Delta(x,y) = |\{i \in [n] : x_i \neq y_i\}|$.

Proposition 3. *The Haming distance Δ is a metric, namely it satisfies:*

1. *Non-negativity: $\Delta(x,y) \geq 0$ for all $x, y \in \{0,1\}^n$.*
2. *Symmetry: $\Delta(x,y) = \Delta(y,x)$ for all $x, y \in \{0,1\}^n$.*
3. *Triangle inequality: $\Delta(x,y) + \Delta(y,z) \geq \Delta(x,z)$ for all $x, y, z \in \{0,1\}^n$.*
4. *$\Delta(x,y) = 0$ if and only if $x = y$.*

Thus, the ambient space Σ^n together with the Hamming distance form a metric space. This observation allows us to think about codes not only as combinatorial objects but also as geometric objects. This geometric picture will be very useful when we derive bounds on codes. We next define the notion of distance of a code.

Definition 4. *(Distance of a code)*
The distance of a code C is defined by $\Delta(C) = \min_{\substack{x,y \in C, \\ x \neq y}} \Delta(x,y)$.

Definition 5. *($(n,k,d)_q$-code)*
Let n, k, d, q be positive integers. A $(n,k,d)_q$-code C is a subset of Σ^n where

1. *Σ is a finite set called the alphabet of the code and $|\Sigma| = q$.*
2. *$\Delta(C) \geq d$.*
3. *$|C| \geq q^k$.*

Note 6. *We would like a (n,k,d,q)-code to have small n, large k and large d. Strictly speaking, there is no requirement on q. However, given a code with a certain alphabet size, we are usually able to construct codes with the same performance but with larger alphabet sizes. This leads us to try to decrease q.*

Notation 7. *For every $x \in \{0,1\}^n$ and $t \in \{0,1,\ldots,n\}$, the Hamming ball $B(x,t)$ centered at x and of radius t is given by*

$$B(x,t) = \{y \in \{0,1\}^n : \Delta(x,y) \leq t\}$$

Definition 8. *(Error correction)*
Let $e \in \{0, 1, \ldots, n\}$. A code C is said to correct e errors if for all $y \in \{0,1\}^n$, $|B(y, e) \cap C| \leq 1$.

Proposition 9. *A code C of distance d corrects at least $\lfloor (d-1)/2 \rfloor$ errors. Conversely, a code that corrects e errors has distance at least $2e + 1$.*

Proof. Since C has distance d, the Hamming balls $B(x, \lfloor (d-1)/2 \rfloor)$ and $B(y, \lfloor (d-1)/2 \rfloor)$ are disjoint for any $x, y \in C$. Thus, for all $y \in \{0,1\}^n$, $|B(y, e) \cap C| \leq 1$. Conversely, if a code corrects e errors, then for any $x, y \in C$, $B(x, e)$ and $B(y, e)$ are disjoint, which implies that $\Delta(x, y) \geq 2e + 1$. □

Most of the codes that we will study in this course will be linear codes, which we now define.

Definition 10. *(Linear codes)*
A code $C \subseteq \Sigma^n$ is said to be linear if $\Sigma = \mathbb{F}_q$ and C is an $\mathbb{F}_q$-vector space.

Notation 11. *We say that C is an $[n, k, d]_q$ code to mean that C is an $(n, k, d)_q$ linear code.*

Proposition 12. *If C is an $[n, k, d]_q$ code, then*

1. *There exist linearly independent vectors $b_1, \ldots, b_k \in \mathbb{F}_q^n$ s.t $C = \{xG \mid x \in \mathbb{F}_q^k\}$ where*
$$G = \begin{bmatrix} b_1^T \\ b_2^T \\ \cdots \\ b_k^T \end{bmatrix} \in \mathbb{F}_q^{k \times n}$$
is said to be a generator matrix of C.
2. *$\Delta(C) = \min_{\substack{x \in C \\ x \neq 0}} wt(x)$ where $wt(x)$ is the number of non-zero coordinates of x.*
3. *There exists a matrix $H \in \mathbb{F}_q^{n \times (n-k)}$, called a parity check matrix of C, s.t. for all $x \in C$, $xH = 0$.*

Proof. The first part follows from the fact that C is a k-dimensional $\mathbb{F}_q$-subspace of $\mathbb{F}_q^n$. Since C is a linear code, $0 \in C$ and if $x, y \in C$, then $x - y \in C$. Moreover, $\Delta(x, y) = wt(x - y)$. This proves the second part. To prove the third part, note that for any $G \in \mathbb{F}_q^{k \times n}$, there exists $H \in \mathbb{F}_q^{n \times (n-k)}$ of rank $n - k$ s.t. $GH = 0$. □

4.2 The Hamming bound

Hamming proved the following relation between the parameters of any code.

Theorem 13. *(The Hamming bound)*
Any $(n, k, d)_q$ code should satisfy
$$q^k Vol(n, \lfloor (d-1)/2 \rfloor) \leq q^n \tag{1}$$
where $Vol(n, \lfloor (d-1)/2 \rfloor)$ is the volume of a ball of radius $\lfloor (d-1)/2 \rfloor$ in Σ^n.

Proof. Consider the q^k balls of radius $\lfloor (d-1)/2 \rfloor$ centered at each of the q^k codewords of the code in the ambient space Σ^n (where $|\Sigma| = q$). Since the code has distance d, the balls are all disjoint. Thus, the theorem follows. □

4.3 The general Hamming code

Hamming codes are $[2^m - 1, 2^m - m - 1, 3]_2$ codes that are constructed by setting the rows of the parity check matrix H to all non-zero binary strings of length m.

Note 14.

1. *The Hamming code is an optimal code of distance* 3 *in the sense that it satisfies the Hamming bound with equality. This follows from equation (1) with* $q = 2$ *and by noting that the volume of any ball in* Σ^n *of radius* 1 *is* $n + 1$.
2. *Hamming's construction can be used to obtain codes of distance* 4.

4.4 Asymptotics of codes

Definition 15. *(Rate and relative distance of a family of codes)*
The rate R *and relative distance* δ *of a family of* $(n, k, d)_q$*-codes are given by*

$$R = \liminf_{n\to\infty} \frac{k}{n}$$

$$\delta = \liminf_{n\to\infty} \frac{d}{n}$$

Note that $\delta, R \in [0, 1]$. A main component of the course will be concerned with determining the best tradeoffs between R and δ. Note that we can represent any family of codes by the point (δ, R) in the plane. We are interested in the set of all points (δ, R) for which there exists a family of codes with rate R and relative distance δ. In later lectures, we will prove upper bounds on R in terms of δ, which will rule out certain regions of the plane as "forbidden". For instance, we will show that for any family of codes, $R + \delta \leq 1$. Moreover, we will prove existential results which assert that certain points of the plane are "achievable".

Note 16.

1. *Our knowledge of the achievable region is not complete. Closing the gap between the achievable and forbidden regions is a main open problem in coding theory.*
2. *We study families that are not on the boundary of the achievable region (in the sense that there exists a family with a larger rate and a larger relative distance) since they may provide algorithmic advantages in terms of the encoding and decoding algorithms.*

References

[Ham50] R.W. Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[Sha48] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.