CS 860 Topics in Coding Theory Sept. 9, 2025

Lecture 2

Lecturer: Elena Grigorescu *Scribe: Alfred Mikhael*

Recall the following definitions and theorems from last lecture.

Definition 1 *$B_n(d) := \{z \in \Sigma^n | wt(x) \leq d\}$ is the Hamming ball of radius d in n dimensions, where $wt(x)$ is the number of non-zero coordinates in x. It holds that $|B_n(d)| = \sum_{k=0}^{d} \binom{n}{k}$*

Lemma 2 (Hamming Bound) *For all codes C on an alphabet Σ with distance at least d we have that*

$$|C| \leq \frac{|\Sigma|^n}{B_n(\lfloor (d-1)/2 \rfloor)}$$

Definition 3 *Let H be the matrix with columns $\mathbb{F}_2^t \backslash \{\mathbf{0}\}$. The Hamming code $\mathrm{Ham} \subseteq \mathbb{F}_2^n$ is the kernel of H.*

In this lecture, we define linear codes, examine the Hamming bound for $d = \Omega(n)$, prove the Gilbert-Varshamov lower bound, and present the solution to last lecture's puzzle.

Definition 4 *A code $C \subseteq \mathbb{F}_q^n$ is linear if it is a linear subspace of F_q^n. Equivalently. $x, y \in C \implies x + y \in C$*

Linear codes can be defined in two ways: they can be defined as the kernel of a matrix, called a parity check matrix, or they can be defined as the image of a matrix, called a generator matrix.

The following property about the distance of a linear code is useful to know.

Claim 5 *Let C be a linear code, then the distance of C is the minimum weight of a non-zero codeword, i.e. $\Delta(C) = \min_{c \in C} wt(c)$*

Proof Let $\Delta(C) = d$, and let $x, y \in C$ such that $\Delta(x, y) = d$. Because C is linear, $x - y \in C$, and $\Delta(x, y) = wt(x - y) = d$. Therefore $\min_{c \in C} wt(c) \leq d$, and clearly $\min_{c \in C} wt(c) \geq d$. ■

The Hamming code is a linear code, so using this characterization of the distance we can prove the following properties of the Hamming code

Lemma 6 *Let Ham be the Hamming code of length $n = 2^t - 1$, then*

1. $|\mathrm{Ham}| = \frac{2^{\mathrm{n}}}{\mathrm{n+1}}$ *and*
2. $\Delta(\mathrm{Ham}) = 3$.

Proof First we will compute the size of Ham. Note that Ham is the kernel of a matrix of size $t \times 2^t - 1$, hence it has dimension at least $2^t - t - 1 = n - t$. Therefore, $|\text{Ham}| \geq 2^n/2^t = 2^n/(n+1)$. By the Hamming bound, it follows that $|C| = \frac{2^n}{n+1}$.

To see that $\Delta(\text{Ham}) = 3$ first note that $\Delta(\text{Ham}) > 1$ because all columns of H are non-zero vectors. We also have that $\Delta(\text{Ham}) > 2$ because the vectors of $\mathbb{F}_2^t$ are pairwise linearly independent. It is easy to find 3 vectors in F_2^t whose sum is 0, so $\Delta(\text{Ham}) = 3$. ■

The moral of the story is that the Hamming codes is tight for the Hamming bound, i.e. $|Ham|$ is exactly the max possible for $d = 3$. Such codes are called *perfect* codes. They have the property that the union of balls of radius 1 cover all the binary vectors of length n.

Hamming bound for large distances Usually, we would like the distance of a code to be as large as possible, in particular we would like $d = \Omega(n)$ to be a constant fraction of n. This would allow for error correction when sending messages over a noisy channel for a constant fraction of errors.

Let us examine the Hamming bound for $d = \delta n$, with $\delta \in (0, 1)$. For a code C with distance $d = \delta n$, we have that

$$\begin{aligned} |C| &\leq 2^n / B_n\left(\left\lfloor \frac{\delta n - 1}{2} \right\rfloor\right) \\ &\lesssim 2^n / B_n(\delta n/2) \\ &= 2^n / \sum_{k=0}^{\frac{1}{2}\delta n} \binom{n}{k} \end{aligned}$$

We can approximate $\binom{n}{\delta n}$ by using Sterling approximation. Let the binary entropy function be

$$H(\delta) = \delta \log_2 \frac{1}{\delta} + (1-\delta)\log_2 \frac{1}{1-\delta}$$

Fact 7

$$\frac{2^{H(\delta)n}}{\text{poly}\log n} \leq \binom{n}{\delta n} \leq 2^{H(\delta)n}$$

$$2^{n(H(\delta)-o(1))} \leq |B_n(\delta n)| \leq 2^{n(H(\delta)+o(1))}$$

Using the approximation, we have that $|C| \leq 2^n/2^{n(H(\delta/2)-o(1))}$ and so for $n \to \infty$ the rate of C satisfies $R = \frac{\log|C|}{n} \leq 1 - H(\delta/2)$.

How good is the Hamming upper bound in general.

We saw that Hamming codes achieve the Hamming bound for $d = 3$ (in fact it is tight for $d = \Theta(1)$), but in general are there codes achieving this bound? We'll look into this question by comparing it with other bounds.

Now we consider lower bounds for codes.

Gilbert-Varshamov The Gilbert-Varshamov bound gives us a non-explicit code for each distance d, with reasonable rate.

Theorem 8 (Gilbert-Varmashov Bound) *For each distance d, there exists a code $C \subseteq \Sigma^n$ of distance d satisfying $|C| \geq \frac{|\Sigma^n|}{B_n(d-1)}$.*

Proof The proof is by greedily constructing a code according to the following very simple algorithm.

1. $C \leftarrow \emptyset$
2. While $\exists x \in \Sigma^n$ such that $\Delta(x, C) > d - 1$, let $C \leftarrow C \cup \{x\}$

Clearly the resulting code has distance at least d. Observe that at each step, we remove at most $B_n(d-1)$ many points from Σ^n, and so $|C||B_n(d-1)| \geq |\Sigma^n|$ ■

In particular, when $\Sigma = F_2$ and $d = \delta n$ we get that $|C| \geq \frac{2^n}{2^{H(\delta)n+o(n)}}$ and so the rate of C is at least $1 - H(\delta) - o(1)$. Note that the code constructed in the GV bound takes exponential time to construct.

One of the big questions in coding theory is about the gap between the GV bound and the Hamming bound. Precisely, are there explicit codes which achieve, or even do better than the GV bound?

There are *algebraic geometry* codes, but they require high alphabet sizes (e.g. $|\Sigma| = 49$). In 2017, there was a breakthrough work by Ta-Shma who constructs binary codes with distance (1/2 - ϵ), which approach the GV bound, but the decoding time was exponential. In 2020, there was another breakthrough by Jeronimo, Quintana, Srivastava, and Tulsiani, who designed codes approaching the GV bound with almost-linear time decoding algorithms.

We will see that there are also linear codes which achieve the GV bound.

Lemma 9 *(GV bound for linear codes) Given $\delta \in (0,1), \forall \epsilon \in (0, 1-H(\delta))$ there is a linear code $C \subseteq \mathbb{F}_2^n$ such that $\Delta(C) \geq \delta n$ and $R(C) \geq 1 - H(\delta) - \epsilon$.*

Proof Let $k = (1-H(\delta)-\epsilon)n$, and pick a uniformly random generator matrix $G \in F_2^{n\times k}$. Let C be the code generated by G. We will show that the probability of C satisfying the conditions in the lemma is ¿0, hence the statement follows by the probabilistic method.

Fix a non-zero $m \in F_2^k$, and note that $Gm \sim \mathrm{Unif(F_2^n)}$ is a uniformly random vector in F_2^n. Then

$$\Pr(wt(Gm) < \delta n - 1) = \frac{B_n(\delta n - 1)}{2^n} \leq 2^{n(H(\delta)+o(1)-1)} = 2^{-k-\epsilon n+o(1)}$$

Taking a union bound over all non-zero values of m we get

$$\Pr(\Delta(C) \leq \delta n - 1) \leq (2^k - 1)2^{-k-\epsilon n+o(1)} \leq 2^{o(1)-\epsilon n}.$$

Therefore for large enough n, the probability of obtaining a vector of weight $\leq \delta n - 1$ is exponentially small, and so $\Pr(\Delta(C) \geq \delta n) \geq 1 - 2^{-\epsilon n} > 0$.

Now let G be a matrix that generates C with $\Delta(C) \geq \delta n$. It remains to show that the rate of C is $\geq k = (1 - H(\delta) - \epsilon)$, that is, to show that the dimension of C is $\geq k$. For this we need to show that G has full rank. Indeed, if G is not of full rank, then $\exists m \neq 0$ s.t. $Gm = 0$, and therefore there exists a non-zero message $m \in \mathbb{F}_2^k$ s.t. the codeword Gm has $wt(Gm) = 0 < \delta n$, which contradicts the fact that G generates a codes of distance $\geq \delta n$. ■

Solution to the puzzle from last lecture. Recall the statement of the problem. There are n students, each is given a black or white sticker on their forehead independently and uniformly at random. Students are not allowed to communicate with each other at all once the stickers have been given. Each student can either guess the color of their sticker, or pass. If at least one student guesses, and all guesses are correct, than students win, otherwise students lose (if nobody guesses or if at least one guess is incorrect).

There is a strategy which wins 50% of the time: dedicate one student to be the guesser and have them guess randomly. However, it is possible to win with probability $1 - \frac{1}{n+1}$, provided that $n = 2^k - 1$.

- Choose an ordering of students
- Each student considers the string of 1s and 0s formed by the stickers of all the other students $x_1 \cdots x_n$. If student i sees that $x_1 \cdots x_{i-1} 0 x_{i+1} \cdots x_n$ or $x_1 \cdots x_{i-1} 1 x_{i+1} \cdots x_n$ is a codeword, they should guess 1 or 0 respectively, so that if their guess is b, the string $x_1 \cdots x_{i-1} b x_i \cdots x_n$ is *not a codeword.*
- If a student sees that the string of 1s and 0s formed by the stickers of all other students cannot be completed to a codeword, then they should pass.

For example, suppose that there are 3 students. There are only 2 codewords, (0, 0, 0) and (1, 1, 1). Suppose that the students are assigned (0, 1, 0). Then student 1 should pass, student 2 should guess 1, and student 3 should pass. Suppose that students are assigned (1, 1, 1). Then all students should guess 0.

It is clear that if students are assigned a codeword they will all guess wrong and fail. If students are not assigned a codeword then, because every string is distance 1 away from a codeword, there will be exactly 1 student who will guess, and they will guess correctly. Note that there cannot be more than one student who guesses in this case, because a string cannot match two codewords to $n - 1$ bits (this would imply the distance between them is 2). Therefore, students win if they are not assigned a codeword. This happens with probability $1 - \frac{1}{n+1}$, as a $\frac{1}{n+1}$ fraction of the space is codewords.