Lecture 2: Linear and Random Codes

Topics in Error-Correcting Codes (Fall 2022)
University of Toronto
Swastik Kopparty
Scribe: Andrew Feng and Yuchong Zhang

1 Linear Codes

Previously, a code is just a subset without special properties. It turns out that codes with vector space structure admit nice properties such as efficient encoding and decoding.

In the following, we identify $\{0,1\}^n$ with $(\mathbb{F}_2)^n$ and work with the latter. In general, the results in this lecture also holds for fields whose sizes are prime powers. A *linear code* is a linear subspace of $(\mathbb{F}_2)^n$. From linear algebra, we know that a k-dimensional subspace of $(\mathbb{F}_2)^n$ has size 2^k.

The following fact shows that linear codes can achieve the GV-bound.

Fact 1. *For any $d \leq n$, there is a linear code C of distance d such that $|C| \geq \frac{2^n}{B_n(d-1)}$.*

Proof. The proof is almost the same as that of theorem 6 from previous lecture (GV-bound). Except when we add a vector y, we will update $C := C \cup y + C$. We need to show that if C is a linear code with distance d and y is at least distance d away from any element of C, then $C \cup y + C$ is also distance d.

Take $u, v \in C \cup y + C$. If $u \in C$ and $v \in C$, then $\Delta(u,v) > d$ because C has distance d. Similarly for $u, v \in y + C$. Now suppose $u \in C$ and $v = y + v' \in y + C$. Then

$$\Delta(u,v) = \Delta(u, y + v') = \Delta(y, u - v') \geq d.$$

□

Given a linear code C of dimension k, an efficient way to encode and decode messages from $(\mathbb{F}_2)^k$ is by using a $k \times n$ matrix G where the rows form a basis for C. The encoding map is then given by $x \mapsto G^T x$ and the decoding map is given by $y \mapsto G(G^T G)^{-1} y$. The matrix G is called the *generator matrix* of C. Notice that we can get G to have the form $\begin{bmatrix} I_k \mid A \end{bmatrix}$ by row and column operations (therefore possibly altering the code).

A second way to specify C is via an $n-k \times n$ matrix H such that $C = \ker H$; H is called the *parity check matrix*.

Recall that we have not specified what it means to construct a code. In the case of linear codes, "constructing" a code of distance d means to have an algorithm that runs in time $poly(n)$ and outputs a generator matrix for some C where C is a linear code of distance d.

2 Random Construction of Codes

Pick $x_1, x_2, \ldots, x_K$ uniformly and independently at random from $(\mathbb{F}_2)^n$. We hope to obtain a code that has distance d.

Fix i, j distinct, we have $\Pr[\Delta(x_i, x_j) < d] = \frac{B_n(d-1)}{2^n}$. Applying union bound gives

$$\Pr[\exists i, j : \ \Delta(x_i, x_j) < d] = \binom{K}{2} \frac{B_n(d-1)}{2^n}.$$

We can set $K \approx \sqrt{\frac{2^n}{B_n(d-1)}}$ to make the probability < 1, which shows the existence of a code with distance d and size K. Replacing d with δn for $\delta < 1/2$, we see that the generated code would have rate $\log(K)/n \to (1 - H(\delta))/2$ as $n \to \infty$, which is half of the GV-bound.

In the above, we could also set K so that the probability $< 1/100$. This attributes to a constant factor $1/10$ in front of K. In $(\log K)/n$, this factor becomes an additive constant, which $\to 0$ as $n \to \infty$. This means by decreasing K by a factor of 10, we get that $99/100$ of the randomly generated codes have distance δn while still having rate equal to half of the GV-bound.

Alternatively, let us compute the expected number of pairs of vectors in $x_1, \ldots, x_K$ that are too close and remove from each pair one vector.

Define $Z_{ij} = 1$ if $\Delta(x_i, x_j) < d$ else 0. Then, $\mathbb{E}[Z_{ij}] = \frac{B_n(d-1)}{2^n}$. By linearity of expectation, we get

$$\mathbb{E}\left[\sum_{i,j} Z_{ij}\right] = \binom{K}{2} \frac{B_n(d-1)}{2^n}.$$

By setting $K \leq \frac{2^n}{100 B_n(d-1)}$, we get that $\mathbb{E}[\sum_{i,j} Z_{ij}] \leq K/100$. This means there is a set of size $K = \lfloor \frac{2^n}{100 B_n(d-1)} \rfloor$ with at most $K/100$ pairs of vectors being too close. Remove one from each pair to get a code C of size $99K/100$ with distance d. The size of C is only a constant factor smaller than $\frac{2^n}{B_n(d-1)}$, so the rate of codes obtained this way has rate equal to the GV-bound.

3 Random Construction of Linear Codes

Pick $x_1, \ldots, x_k$ uniformly and independently at random from $(\mathbb{F}_2)^n$, we compute the probability that $\mathrm{span}(x_1, \ldots, x_k)$ has distance d.

Fact 2. *Let C be a linear code. Then C has distance d if and only if $wt(y) = \Delta(0, y) \geq d$ for all nonzero $y \in C$.*

The proof of this fact is the simple observation that $\Delta(x, y) = \Delta(0, y - x)$.

For $v \in (\mathbb{F}_2)^k$, define event $B_v =$ "$wt(\sum_{i=1}^k v_i x_i) < d$". Clearly, the above fact shows C is a linear code of distance d and and only if B_v does not happen for any nonzero v.

We have $\Pr[B_v] = \frac{B_n(d-1)}{2^n}$ since $\sum_{i=1}^{k} v_i x_i$ is uniformly distributed on $(\mathbb{F}_2)^n$. By union bound, we get

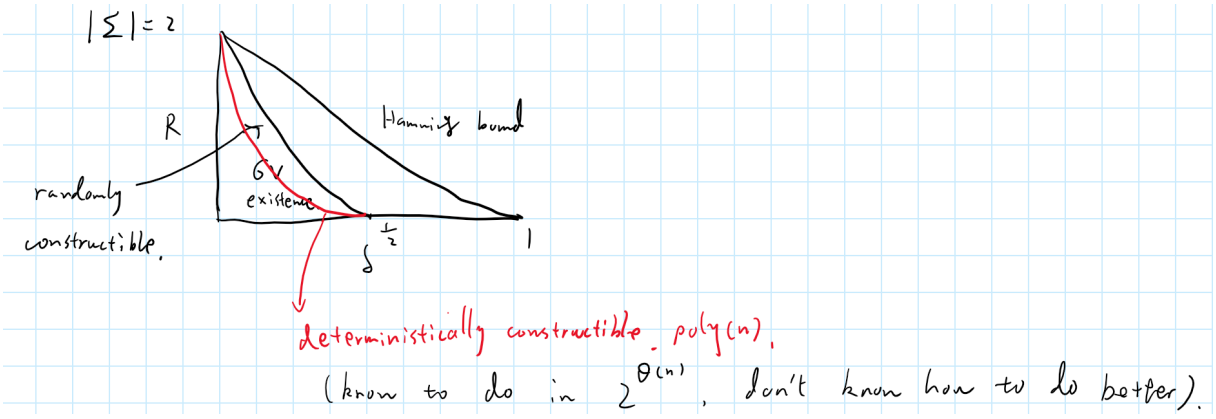
$$\Pr[\exists v \neq 0 : \ B_v] = (2^k - 1)\frac{B_n(d-1)}{2^n}.$$

If we set k so that $2^k < \frac{2^n}{B_n(d-1)}$, then there exists a linear code of dimension k (dimension k because $wt(\sum_{i=1}^{k} v_i x_i) \geq d$ for all nonzero v implies linear independence) and distance d. Even better, if we decrease k by 10, then at least 1023/1024 of codes constructed this way are of dimension k and distance d.

The moral of Random Linear Codes is the following:

1. Most linear codes whose dimension is not too large meet the GV bound.
2. Constructing a code meeting the GV bound is a derandomization process.

Let us consider the rate-distance tradeoff curve again:

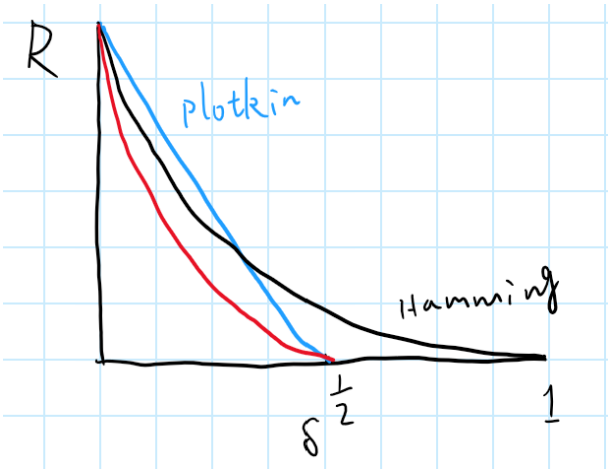


The algorithm for random linear codes can construct codes that satisfy the GV bound, but we do not know a deterministic algorithm for this that runs in $poly(n)$ time. In fact, we know how to deterministally construct codes that satisfy the GV bound in $2^{\Theta(n)}$ time, but we don't know how to do better.

4 Plotkin Bound

Plotkin provides new insights to the relation between rate and relative distance, he also shows that codes with large distance cannot contain too many code words. The main statements are as follows:

1. $d \geq n/2 \implies$ any code with distance d has $\leq O(n)$ code words.
2. For rate R and relative distance d, one has $R + 2\delta < 1 + o(1)$.



Statement 2 provides a new rate-distance curve:

We first prove statement 1.

Proof. Consider the map $\phi : \mathbb{F}_2^n \to \mathbb{R}^n$ defined by

$$\phi(x_1, \ldots, x_n) = ((-1)^{x_1}, \ldots, (-1)^{x_n})$$

Clearly ϕ is injective, so $|\phi(C)| = |C|$. $\Delta(x, y)$ and the inner product on $\mathbb{R}^n$ satisfy the following relation:

$$\begin{aligned}\langle \phi(x), \phi(y) \rangle &= \sum_{i=1}^{n} (-1)^{x_i + y_i} \\ &= (\text{number of agreements} - \text{number of disagreements})(x, y) \\ &= n - 2\Delta(x, y)\end{aligned}$$

Thus, if C has distance $d \geq n/2$, then $\forall x, y \in C$ such that $x \neq y$, we have $\langle \phi(x), \phi(y) \rangle < 0$. The statement then follows directly from the following claim:

Claim 3. *For any collection of nonzero vectors in $\mathbb{R}^n$, if any two vectors has ≤ 0 inner product, then the collection has size at most $2n$.*

We will prove the claim by induction on n. For $\mathbb{R}^n$, apply appropriate scaling and rotation, denoted T, to the collection so that some vector is mapped to $e_1 = (1, 0, \ldots, 0)$. Since rotation and scaling preserve the sign of inner products, all other vectors after applying the transformation T have the form $(a, \vec{b})$ where $a \leq 0$ and $\vec{b} \in \mathbb{R}^{n-1}$. For any distinct pair of such vectors, we have:

$$\langle (a_1, \vec{b_1}), (a_2, \vec{b_2}) \rangle = a_1 a_2 + \langle \vec{b_1}, \vec{b_2} \rangle \leq 0$$

Since $a_1, a_2 \leq 0$, this implies $\langle \vec{b_1}, \vec{b_2} \rangle \leq 0$. Because we know that after applying T the collection contains e_1, it follows that there can be in total at most two vectors with the form $(a, \vec{0})$. So after removing at most these two vectors, the rest of $\vec{b_i}$'s are in $\mathbb{R}^{n-1}$, nonzero, and satisfy $\langle \vec{b_i}, \vec{b_j} \rangle \leq 0$. By induction hypothesis, there can by at most $2n - 2$ $\vec{b_i}$'s, and thus the original collection has at

most $2n-2$ vectors of the form $(a, \vec{b})$ satisfying $\vec{b} \neq 0$. Such vectors together with the at most two of the form $(a, \vec{0})$ constitute the collection, so there are at most $2n$ vectors in the collection.

□