CS395T Pseudorandomness and Combinatorial Constructions
Lecture 8, February 8, 2001
Lecturer: Prof. David Zuckerman, Scribe: Michael Cyr

1 Error-Correcting Codes

An error-correcting code encodes a message by a somewhat longer string, in such a way that even if some bits of the longer string are corrupted the original message can be recovered. We begin defining codes and giving an example.

1.1 Example: Reed-Solomon Codes

The message bits of a Reed-Solomon code define a polynomial p of degree less than k over $\mathbb{F}_q = \{0, a_1, ..., a_{q-1}\}$, $q \geq n$. The encoding of p is the vector $(p(0), p(a_1), \ldots, p(a_{q-1})) \in \mathbb{F}_q^n$. The point is that to recover p we only need p evaluated at k points.

Instead of viewing a code as a function, we usually view a code as the range of the function, i.e., the set of possible encodings.

1.2 Definitions

Definition 1 *A code is a subset $C \subseteq \mathbb{F}_q^n$.*

An important type of code is a linear code.

Definition 2 *A linear code is a subspace of the vector space $\mathbb{F}_q^n$, i.e.,*

$$c_1, c_2 \in C \Rightarrow c_1 + c_2 \in C$$
$$\lambda \in \mathbb{F}_q, c \in C \Rightarrow \lambda c \in C$$

Proposition 1 *Reed-Solomon codes are linear.*

This is because adding two polynomials or multiplying a polynomial by a scalar doesn't increase the degree of the polynomial(s).

Definition 3 *The Hamming distance* $\Delta(c_1, c_2) = |\{i : (c_1)_i \neq (c_2)_i\}|$

Definition 4 $\text{distance}(C) = \min_{c_1, c_2 \in C, c_1 \neq c_2} \Delta(c_1, c_2)$

A linear code over $\mathbb{F}_q$ of length n, dimension k, and distance d is often called an $[n, k, d]_q$ code. If the distance is not known it is often called an $[n, k]_q$ code. The rate of a code is $\frac{k}{n}$.

The distance of a Reed-Solomon code is $\text{distance}(RS[n, k]_q) = n - (k - 1)$.

Proposition 2 *If the distance of a code = d, then it can be used to:*

1. *correct $d - 1$ erasures (i.e. missing symbols)*
2. *detect $d - 1$ errors (i.e. changed symbols)*
3. *correct $\frac{d-1}{2}$ errors*

1.3 A bound on the distance

Lemma 1 *For any* $[n,k,d]_q$ *code,* $k+d \leq n+1$

Proof: Start with any $[n,k,d]_q$ code. Puncture it, i.e., for some i, remove the i^{th} symbol from each code word. This gives an $[n-1,k,\geq d-1]_q$ code. Repeat until we get an $[n-d+1,k,\geq 1]_q$ code. Since the dimension is at most the length, $n-d+1 \geq k$. □

Note that we can convert a Reed-Solomon code into a binary code by taking q to be a power of 2, say $q = 2^m$, and convert each symbol in $\mathbb{F}_q$ to m binary symbols. However, the relationship between distance and dimension is not ideal. The new binary code also has q^k codewords, giving dimension dimension $\log_2 q^k = mk$, and hence is an $[mn, mk, \geq d]_2$ code. Notice that the rate $(\frac{mk}{mn})$ stays the same, but the relative distance $= \frac{d}{mn}$ drops.

2 Asymptotically Good Codes

We can achieve much better binary codes, which are called asymptotically good codes. Such codes have

$$\frac{k}{n} = \Omega(1)$$
$$\frac{d}{n} = \Omega(1)$$

We will show just such an existence argument below, but first we need the following definitions and lemma:

Definition 5 *The weight of a word* $\mathrm{wt}(c) = |\{i : c_i \neq 0\}|$. *The minimum weight of a code is* $\min-\mathrm{wt}(C) = \min_{c\in C, c\neq 0}\{\mathrm{wt}(c)\}$.

Lemma 2 *For linear codes* C, $\mathrm{distance}(C) = \min - \mathrm{wt}(C)$.

Proof: $(\leq)$: If c is a minimum weight codeword, then $\mathrm{distance}(c, 0) = \min - \mathrm{wt}(c)$.
$(\geq)$: If $\mathrm{distance}(v_i, v_j) = \min - \mathrm{wt}(c)$, then $v_i - v_j \in C$ and $\mathrm{wt}(v_i - v_j) = \mathrm{distance}(C)$. □

Proposition 3 *There exists asymptotically good binary codes.*

Proof: For arbitrary codes (not necessarily linear): choose $v_1, ..., v_K \in \{0,1\}^n$, where $K = 2^k$. By using one of the Chernoff bounds:

$$\Pr\left[\Delta(v_i, v_j) \leq (\frac{1}{2} - \alpha)n\right] \leq e^{-2\alpha^2 n}$$
$$\Pr\left[(\exists i \neq j : \Delta(v_i, v_j) \leq (\frac{1}{2} - \alpha)n\right] \leq \binom{K}{2} e^{-2\alpha^2 n}$$

The code will exist if the above probability is < 1, or in other words, if $\binom{K}{2} < e^{2\alpha^2 n}$. □

It is somewhat surprising that we can achieve better constants by insisting that the code be linear.

Proposition 4 *There exists asymptotically good binary codes that are linear, with slightly better parameters.*

Proof: Choose $v_1, ..., v_k \in \mathbb{F}_2^n$ randomly, which gives us a linear code $C = \{\lambda_1 v_1 + ... + \lambda_k v_k : \lambda_i \in \{0, 1\}\}$. Using Lemma 2, for any $\lambda_1, ..., \lambda_k$ not all 0:

$$\Pr\left[\mathrm{wt}(\lambda_1 v_1 + ... + \lambda_k v_k) \leq (\frac{1}{2} - \alpha)n\right] \leq e^{-2\alpha^2 n}$$

By the union bound,

$$\Pr\left[\exists c \in C, \mathrm{wt}(c) \leq (\frac{1}{2} - \alpha)n\right] \leq (2^k - 1)e^{-2\alpha^2 n} < Ke^{-2\alpha^2 n}$$

So we can take $K = e^{2\alpha^2 n}$ and thus $k = \log K = 2\alpha^2(\log e)n$, and if we let $\alpha = \frac{1}{4}$, we get $d = \frac{1}{4}n$, $k = \Omega(n)$.

$\square$

Note: We can greedily construct asymptotically good codes in time $2^{O(n)}$.

3 Concatenated Codes

We will see how to use concatenated codes to explicitly construct asymptotically good codes. A concatenated code is a new code based on an outer code and an inner code.

Outer code: $[n, k, d]_q$ code C, $q = 2^{k'}$.

Inner code: $[n', k', d']$ code C'. $C' \subseteq \mathbb{F}_2^{n'}$, but think of $C' : \mathbb{F}_2^{k'} \Rightarrow \mathbb{F}_2^{n'}$

Elements of C concat C' are $< C'(c_1), C'(c_2), ..., C'(c_n) >$, where $< c_1, c_2, ..., c_n > \in C$

Now choose C to be a Reed-Solomon code: an $RS[n, k = \frac{n+1}{2}, d = \frac{n+1}{2}]_{q=n=2^{k'}}$ code. C' is an $[n' = 2k', k', d' = \Omega(n')]_2$ code constructed greedily. Because $n' = O(\log n)$ this construction takes polynomial time.

Then C concat C' is an $[n \cdot n', \underbrace{k \cdot k'}_{\simeq \frac{1}{4}(n \cdot n')}, \underbrace{d \cdot d'}_{\simeq \Omega(n \cdot n')}]_2$ code, and hence is asymptotically good.