6.897 Algorithmic Introduction to Coding Theory September 16, 2002

Lecture 4

Lecturer: Madhu Sudan *Scribe: Amit J. Deshpande*

1 Introduction

The main topics included in today's lecture are:

- Singleton Bound
- Reed-Solomon Codes
- Multivariate Polynomial Codes

2 Singleton Bound

So far we have discussed Shannon's theory, Hamming's metric and Hamming codes, and Hadamard codes. We are looking for aymptotically good codes to correct some constant fraction of errors while still trasmitting the information through the noisy channel at a positive rate. In our usual notation of $[n, k, d]_q$-codes, Hamming's construction gives a $[n, n - \log_2 n, 3]_2$-code while Hadamard codes were $[n, \log_2 n, \frac{n}{2}]_2$-codes. But we are looking for codes where $\frac{k}{n}$ and $\frac{d}{n}$ have a lower bound independent of n and q is not growing to ∞. In this scenario, we discuss the following simple impossibility result:

Theorem 1 *For a code $C : \Sigma^k \to \Sigma^n$ with minimum distance d, $n \geq k + d - 1$.*

Proof In other words, we want to prove that $d \leq n - (k - 1)$. Just project all the codewords on the first $(k - 1)$ coordinates. Since there are q^k different codewords, by pigeon-hole principle at least two of them should agree on these $(k - 1)$ coordinates. But these then disagree on at most the remaining $n - (k - 1)$ coordinates. And hence the minimum distance of the code C, $d \leq n - (k - 1)$. ■

3 Reed-Solomon Codes

Definition 2 *Let $\Sigma = \mathbb{F}_q$ a finite field and $\alpha_1, \ldots, \alpha_n$ be distinct elements of $\mathbb{F}_q$. Given n, k and $\mathbb{F}_q$ such that $k \leq n \leq q$, we define the encoding function for Reed-Solomon codes as: $C : \Sigma^k \to \Sigma^n$ which on message $m = (m_0, m_1, \ldots, m_{(k-1)})$ consider the polynomial $p(X) = \sum_{i=0}^{(k-1)} m_i X^i$ and $C(m) = \langle p(\alpha_1), p(\alpha_2), \ldots, p(\alpha_n) \rangle$.*

Theorem 3 *Reed-Solomon code matches the singleton bound. i.e. it's a $[n, k, n-(k-1)]_q$-code.*

Proof The proof is based only on the simple fact that a non-zero polynomial of degree l over a field can have at most l zeroes.

For Reed-Solomon code, two codewords (with corresponding polynomials p_1 and p_2) agree at i-th coordinate iff $(p_1 - p_2)(\alpha_i) = 0$. But $(p_1 - p_2)$, from the above fact, can have at most $(k-1)$ zeros which means that the minimum distance $d \geq n - (k-1)$. ■

Reed-Solomon codes are linear. This can be easily verified by the fact that the polynomials of degree $\leq (k-1)$ form a vector space (i.e. if p_1, p_2 are polynomials of degree $\leq (k-1)$ then similarly are βp_1 and $p_1 + p_2$). Since the polynomials $p(X) \equiv 1, p(X) = X, \ldots, p(X) = X^{(k-1)}$ form the basis for this vector space, we can also find a generator matrix for Reed-Solomon codes.

$$G = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \ldots & \alpha_n \\ : & : & : & : \\ {\alpha_1}^{(k-1)} & {\alpha_2}^{(k-1)} & \ldots & {\alpha_n}^{(k-1)} \end{bmatrix}$$

One can also prove the theorem about the minimum distance of Reed-Solomon codes by using the fact that any k columns of G are linearly independent (because α_i's are distinct and thus the Vandermonde matrix formed by the k columns is non-singular).

Using Reed-Solomon codes with $k = \frac{n}{2}$ we can get a $[n, \frac{n}{2}, \frac{(n+2)}{2}]_q$-code which means that we can correct much more erros than before. Typically Reed-Solomon codes are used for storage of information in CD's because they are robust against bursty errors that come in contiguous manner, unlike the random error model studied by Shannon. Also if some information is erased in the corrupted encoding, we can still retrieve the original message by interpolating the polynomial on the remaining values we get.

A way to visualize Reed-Solomon code on binary alphabet is to consider $q = n = 2^m$. This gives $C : \Sigma^k = {\mathbb{F}_2}^{k\log_2 n} \rightarrow \Sigma^n = {\mathbb{F}_2}^{n\log_2 n}$ a $[n\log_2 n, k\log_2 n, n-(k-1)]_2$-code. And if you put $n\log_2 n = N$ and $n - (k-1) = 5$ then this gives a $[N, N - 4\log_2 N, 5]_2$-code. As we have already seen, Hamming's construction gave a $[N, N - \log_2 N, 3]_2$-code and Hamming's impossibility result said that for a $[N, k, 2t+1]_2$-code, $k \geq N - t\log_2 N$. Reed-Solomon code don't achieve this but give a fairly closer $k \geq N - 2t\log N$ (We will discuss later about BCH codes which match this bound).

4 Multivariate Polynomial Codes

Here instead of considering polynomials over one variable (like in Reed-Solomon codes), we will consider multivariate polynomials. For example, let's see the following encoding similar to Reed-Solomon codes but using bivariate polynomials.

Definition 4 *Let $\Sigma = \mathbb{F}_q$ and let $k = l^2$ and $n = q^2$. A message is typically $m = (m_{00}, m_{01}, \ldots, m_{ll})$ and is treated as the coefficients of a bivariate polynomial $p(X, Y) = \sum_{i=0}^{l} \sum_{j=0}^{l} m_{ij} X^i Y^j$ which has degree l in each variable. The encoding is just evaluating the polynomial over all the elements of $\mathbb{F}_q \times \mathbb{F}_q$. i.e. $C(m) = \langle p(x,y) \rangle_{(x,y) \in \mathbb{F}_q \times \mathbb{F}_q}$.*

But what is the minimum distance of this code ? We will use the following lemma to find it.

Lemma 5 (Schwartz-Zippel lemma) *A multivariate polynomial $Q(X_1, \ldots, X_m)$ (not identically 0) of total degree L is non-zero on at least $(1 - \frac{L}{|S|})$ fraction of points in S^m, where $S \subseteq \mathbb{F}_q$.*

Proof of Lemma : We will prove it by induction on the number of variables. For $m = 1$, it's easy as we know that $Q(X_1, \ldots, X_m)$ can have at most L zeros over the field $\mathbb{F}_q$. Now assume that the induction hypothesis is true for the a multivariate polynomial with upto $(m-1)$ variables, for $m > 1$. Consider

$$Q(X_1, \ldots, X_m) = \sum_{i=0}^{t} {X_1}^i Q_i(X_2, \ldots, X_m)$$

where $t \leq L$ is the largest exponent of X_1 in $Q(X_1, \ldots, X_m)$. So the total degree of $Q_t(X_2, \ldots, X_m)$ is at most $(L - t)$.

By induction hypothesis it implies that $Q_t(X_2, \ldots, X_m) \neq 0$ on at least $(1 - \frac{(L-t)}{|S|})$ points in $S^{(m-1)}$. But suppose $Q_t(s_2, \ldots, s_m) \neq 0$ then $Q(X_1, s_2, \ldots, s_m)$ is a not-identically-zero polynomial of degree t in X_1, and therefore is non-zero on at least $(1 - \frac{t}{|S|})$ fraction of choices for X_1.

So putting it all together, $Q(X_1, \ldots, X_m)$ is non-zero on at least $(1 - \frac{(L-t)}{|S|})(1 - \frac{t}{|S|}) \geq (1 - \frac{L}{|S|})$ points in S^m. ■

Using this, for $m = 2$ and $S = \mathbb{F}_q$, we get that the above bivariate polynomial code is a $[q^2, l^2, (1 - \frac{2l}{q})q^2]_q$-code.

Some interesting cases of multivariate polynomial codes include Reed-Solomon code ($l = k$ and $m = 1$), Bivariate polynomial code ($l = \sqrt{k}$ and $m = 2$) and Hadamard code ($l = 1$, $m = k$ and $q = 2$) !