6.895 Essential Coding Theory September 22, 2004

Lecture 5

Lecturer: Madhu Sudan *Scribe: Swastik Kopparty*

1 Algebraic Codes

In this lecture we will study combinatorial properties of several algebraic codes. In particular, we will introduce:

- Reed-Solomon Codes based on univariate polynomials over finite fields.
- Reed-Muller Codes based on multivariate polynomials over finite fields.
- Hadamard Codes as an interesting special case of Reed Muller Codes.
- The Plotkin bound that shows that Hadamard codes are optimal for their distance.

2 Projection Bound Revisited and Reed Solomon Codes

Recall that the projection bound for an $(n, k, d)_q$ code said that $d \leq n - k + 1$. This was proved using the following argument: Project the q^k codewords to the first $k - 1$ coordinates. Since there are only q^{k-1} possible $k - 1$ tuples, there must be two codewords that projected to the same point. At worst, those two codewords can differ in only the remaining $n - (k - 1)$ places. Thus the minimum distance is at least $n - (k - 1) = n - k + 1$.

Althought the argument used above was very loose, this bound is tight! To show this, we will produce a code $\mathcal{C} : \mathbb{F}_q^k \to \mathbb{F}_q^n$ that has the following extremely strong property: $\forall S \subset [n], |S| = k$, the restriction map $\mathcal{C}|_S : \mathbb{F}_q^k \to \mathbb{F}_q^k$ is a bijection.

Such a map is in fact well known. It is the polynomial evaluation map based on the following polynomial interpolation theorem:

Theorem 1 Polynomial Interpolation Theorem: *Let F be a field. For any $\alpha_1, \alpha_2, \ldots \alpha_{l+1} \in F$, pairwise distinct, and any $y_1, y_2, \ldots, y_{l+1} \in F, \exists$ unique polynomial $p(x)$ with degree $\leq l$ such that $p(\alpha_i) = y_i, \forall i \in [l + 1]$.*

Proof Idea One treats the $l + 1$ coefficients of the polynomial as unknowns. The $l + 1$ conditions $p(\alpha_i) = y_i$ are linear constraints on them. Show that the linear system has a unique solution by proving the constraint matrix is invertible(it is a Vandermonde matrix). ■

Thus if we pick distinct $\alpha_1, \ldots \alpha_{l+1} \in F$, we get a bijection:
$\{$Coefficients of p, a degree l polynomial $\} \Leftrightarrow \{ \langle p(\alpha_1), \ldots p(\alpha_{l+1}) \rangle \}$

As it is, this map is no good: we merely map $l + 1$-tuples to $l + 1$-tuples. But a slight variation on Theorem 1 (which follows directly from it) does the trick.

Theorem 2 Useful Polynomial Interpolation Theorem: *Let F be a field. We are given $\alpha_1, \alpha_2, \ldots \alpha_n \in F$, pairwise distinct. Then for any distinct $i_1, i_2, \ldots i_{l+1} \in [n]$ and any $y_1, y_2, \ldots, y_{l+1} \in F, \exists$ unique polynomial $p(x)$ with degree $\leq l$ such that $p(\alpha_{i_t}) = y_i, \forall t \in [l + 1]$.*

This motivates the Reed Solomon Code (1960ish) as defined below:

- Given $n, q, k, \alpha_1, \ldots \alpha_n$, with α_i distinct elements of $\mathbb{F}_q$.
- For $c = (c_0, c_1, \ldots c_{k-1})$, define polynomial $p_c(x) = c_0 + c_1 x + \ldots c_{k-1} x^{k-1}$

- The code is specified by the encoding function RS : $(c_0, c_1, \ldots c_{k-1}) := \langle p_c(\alpha_1), \ldots, p_c(\alpha_n) \rangle$

If any two polynomials agree on at least k points, by the theorem above they are in fact the same polynomial. So for any two distinct polynomials, on at least $n - (k-1)$ points their evaluations differ. Thus $d \geq n - k + 1$. By the projection bound, $d = n - k + 1$. Observe that this is a linear code because a linear combination of polynomials of degree at most d is again a polynomial of degree at most d.

A couple of remarks on the alphabet size q. $q \geq n$ since the α_i have to be distinct. Thus we don't get binary codes. Further, the field size grows with the code size. Thus one has to take the great asymptotic parameters of this family of codes with a pinch of salt.

Thus for any $k \leq n \leq q$,with q a prime power, $\exists [n, k, n-k+1]_q$ code.

There has been a school of thought that uses $\alpha_i = \omega^i$, where ω is a generator of the multiplicative group $\mathbb{F}_q^*$. However, today we don't bother with that.

Any code that meets the projection bound (with $d = n - k + 1$) is called a Maximum Distance Separable (MDS) code.

2.1 Linear Codes and their Duals

Recall that for any linear code $\mathcal{C}$, we had 2 matrices associated with it: the $k \times n$ generator matrix G and the $n \times (n-k)$parity check matrix H. Define its dual $\mathcal{C}^\perp$ to be the code with H^T as the generator matrix. This is a code with $n - k$ information symbols and n code length. Since $GH = 0$, we have $H^T G^T = 0$ and thus G^T is the parity check matrix for $\mathcal{C}^\perp$. Clearly, $\left(\mathcal{C}^\perp\right)^\perp = \mathcal{C}$

It has empirically been found that very often the duals of "interesting" codes are "interesting". This is why we study duals. For instance, the dual of a Reed Solomon code is a Reed Solomon code. We will not prove that here. Neither will we prove the other interesting fact that the dual of an MDS code is MDS. The latter is an easy exercise. Specifically the result is: if $\mathcal{C}$ is $[n, k, n-k+1]_q$, then $\mathcal{C}^\perp$ is $[n, n-k, k+1]_q$.

3 Multivariate Polynomials and Reed Muller Codes

The large alphabet size of Reed Solomon codes makes it not as nice as we would have liked. There is a natural generalization of these codes to multivariate polynomials which mitigates this problem to some extent.

Without further ado, we define the multivariate polynomial evaluation map, the Reed Muller code (1955-1956):

- Given m, l, q, our code will be over $P_m^l := \{\text{polynomials over } \mathbb{F}_q \text{ in } m \text{ variables of degree}^1\ l\}$
- RM : $P_m^l \rightarrow \mathbb{F}_q^m$
- RM$(p) = \left\langle p(x) : x \in \mathbb{F}_q^m \right\rangle$, the evaluation of p on each point of $\mathbb{F}_q^m$.

Note that this too is a linear code. To determine the parameters of the code, we need a little bit of work.

- $n = q^m$
- $k = \binom{l+m}{l}$ when $l \leq q - 1$. This is merely the number of monomials of degree $\leq l$, and since the monomials form a basis for P_m^l, it is the dimension of our code. If $l \geq q$ it gets messy because of identities like $x^q = x, \forall x \in \mathbb{F}_q$, and thus we will ignore this case forever.

[1] The degree of a multivariate polynomial $\sum_{i_1, i_2, \ldots i_m} a_{i_1 i_2 \ldots i_m} x_1^{i_1} x_2^{i_2} \ldots x_m^{i_m}$ is defined to be the largest d for which there is a non-zero $a_{i_1 i_2 \ldots i_m}$ with $i_1 + i_2 + \ldots i_m = d$.

- $d = \left(1 - \frac{l}{q}\right) n$, because of the following result which will be proved later: Any non-zero polynomial on $\mathbb{F}_q^m$ of degree $\leq l$ is zero on at most $\frac{l}{q} q^m$ points. We already know this result for $m = 1$ (and indeed used it to prove the distance of the RS code).
- $q = q$

Let us just pick some parameters arbitrarily to get a feel for the kinds of codes that we get.

- $m = 2$
- $l = \frac{1}{3} q$
- So $n = q^2$, $k = \binom{\frac{1}{3}q+2}{2} \approx \frac{1}{18} q^2$, $d = \frac{2}{3} n = \frac{2}{3} q^2$

So $\exists [q^2, \frac{1}{18} q^2, \frac{2}{3} q^2]_q$ codes (note the constant rate and constant relative distance). Here $q \approx \sqrt{n}$, which is already much better than the RS code. With a slightly different choice of parameters, for any constant m, there is a family of $[q^m, \frac{1}{(2m)^m} q^m, \frac{1}{2} q^m]_q$ codes with good distance and $q \approx n^{1/m}$.

4 The Plotkin Bound and Hadamard Codes

Let us see what kind of binary RM codes exist. If we put $q = 2$, since $l \leq q - 1$, we are forced to put $l = 1$. Thus we are dealing with linear multivariate polynomials over $\mathbb{F}_2$. The only parameter we can vary is m. We get $n = 2^m$, $k = \binom{m+1}{1} = m + 1$ and $d = 1/2 \cdot 2^m$. Now the rate of this code is horrendous, $m + 1$ bits get encoded as 2^m bits. However, we get great relative distance ($= 1/2$). This code is called the Hadamard code.

Last lecture we saw the unnerving phenomenon that there cannot exist a binary code with more than 2 codewords with relative distance $> 2/3$ [2] The Plotkin bound extends this idea to codes with relative distance 1/2 and shows that the Hadamard codes are optimal for this distance.

Theorem 3 Plotkin Bound: *If there exists a $(n, k, n/2)_2$ code, then $k \leq \log_2(2n)$.*

Sketch of Proof Suppose the code consists of words $c_1, c_2, \ldots c_K \in 0, 1^n$. Create vectors $v_1, v_2 \ldots v_K \in 1, -1^n$ by $(v_i)_j = (-1)^{(c_i)_j}$. Now, the condition that c_i and c_j have distance at least 1/2 is equivalent to the condition that v_i and v_j have inner product $\langle v_i, v_j \rangle \leq 0$. As we will see tomorrow, in n dimensions, there can be at most $2n$ vectors that have this property. Thus $K \leq 2n$ and $k = \log_2 K \leq \log_2(2n)$.

■

5 Notes

Reed Solomon codes are ubiquitous in computer hardware today. The large alphabet size is not much of a problem since for practical applications, something like a $n = 256$ code is about as large as we want to handle, thus not requiring a field size of more than 256 elements (conveniently chosen so that 1 byte can handle it all). By judicious choice of interleaving strategy, data on a disk can be spread out so that any local catastrophe (spatially close bits are destroyed) only affects a few bytes of lots of codewords (all of which can be recovered) as opposed to many bytes of a single codeword.

[2] To recap, let x, y, z be codewords in a code with relative distance $2/3 + \epsilon$. Look at $x - y$ and $x - z$. They are both of relative hamming weight $2/3 + \epsilon$. Thus they agree in at least $1/3 + 2\epsilon$ of the positions. So the relative distance between y and z is at most $2/3 - 2\epsilon$.

5.1 Proof of the statement about zeroes of a multivariate polynomial

We want to show that for a degree l polynomial p in m variables over $\mathbb{F}_q$, with $S \subset F$

$$Pr_{\alpha \in_R \S^m}\left[p(\alpha) = 0\right] \leq l/q$$

We already know that for a univariate polynomial of degree l, the number of 0s in a set S can be no more than $l/|S|$. We proceed by induction, assuming the result for all polynomials of degree l in $< m$ variables.

Write the polynomial $p(x_1, \ldots, x_m) = \sum_{i=0}^{l_1} p_i(x_1, \ldots, x_{m-1}) x_m^i$, with $p_{l_1}(x_1, \ldots, x_{m-1})$ not identically 0. Observe that degree $p_{l_1} \leq l - l_1$. Then

$$Pr_{x \in_R S^m}\left[p(x) = 0\right] \leq Pr_{x \in_R S^m}\left[p(x) = 0 | p_{l_1}(x_1, \ldots x_{m-1}) \neq 0\right] + Pr_{x \in_R S^m}\left[p_{l_1}(x_1, \ldots x_{m-1}) = 0\right]$$

$\leq \frac{l-l_1}{|S|} + \frac{l_1}{|S|} = \frac{l}{|S|}$.

Note that this also proves (by putting $q = 2$, $l = 1$, $S = \mathbb{F}_2$) the fact that the inner product of a nonzero vector with a purely random vector gives a purely random bit. This result shows up in many other contexts with very different proofs.