CSE 533: Error-Correcting Codes **(Autumn 2006)**

Lecture 7: Explicit Linear Code Constructions

October 20, 2006

Lecturer: Venkatesan Guruswami *Scribe: Patrick Tague*

1 Introduction

To this point, we have focused on the derivation of asymptotic bounds for the existence/non-existence of $(n, k, d)_q$ codes with rate $R = \frac{k}{n}$ and relative distance $\delta = \frac{d}{n}$. We are now interested in explicit constructions of linear codes in an attempt to achieve or approach the previously derived bounds.

2 Reed-Solomon Codes

A Reed-Solomon (RS) code is an $[n, k, d]_q$ linear code with $\Sigma = \mathbb{F}_q$ and $q \geq n$ described in terms of the encoding function. The encoding function

$$Enc : \Sigma^k \to \Sigma^n$$

maps a k-symbol message $[m_0, \ldots, m_{k-1}]$ to an n-symbol codeword $[M(\alpha_0), \ldots, M(\alpha_{n-1})]$ where $M(x)$ is the polynomial

$$M(x) = \sum_{i=0}^{k-1} m_i x^i, \tag{1}$$

and $\alpha_0, \ldots, \alpha_{n-1}$ are distinct elements in $\mathbb{F}_q$. Typically, $q = n$ and each α_i is an element of $\mathbb{F}_q$, or $n = q - 1$ and each α_i is a non-zero element of $\mathbb{F}_q$.

The linearity of an RS code $\mathcal{C}$ can be easily verified by checking the conditions for closure under addition and scalar multiplication. Let $c, c' \in \mathcal{C}$ be codewords corresponding to the messages $m = [m_0, \ldots, m_{k-1}]$ and $m' = [m'_0, \ldots, m'_{k-1}]$, respectively. Then the i^{th} entry $M(\alpha_i) + M'(\alpha_i)$ of $c + c'$ is given by

$$\begin{aligned} M(\alpha_i) + M'(\alpha_i) &= \sum_{i=0}^{k-1} m_i x^i + \sum_{i=0}^{k-1} m'_i x^i \\ &= \sum_{i=0}^{k-1} (m_i + m'_i) x^i, \end{aligned} \tag{2}$$

so $c + c' \in \mathcal{C}$ is the image of the element $m + m' \in \mathbb{F}_q^k$ under the encoding function Enc. Similarly, for any constant $\alpha \in \mathbb{F}$, αc is the image of the element $\alpha m \in \mathbb{F}_q^k$ under the encoding function Enc.

Alternatively, the linearity of an RS code follows from the fact that $(f+g)(x) = f(x) + g(x)$ and $(\alpha f)(x) = \alpha f(x)$ for polynomials over a field $\mathbb{F}$, i.e. the space of polynomials of $\mathbb{F}$ is a linear space of functions.

An RS code $\mathcal{C}$ can thus be described using the $n \times k$ generator matrix G. From the encoding function Enc defined using (1), it is clear that G is the Vandermonde matrix

$$G = \begin{pmatrix} 1 & \alpha_0 & \dots & \alpha_0^{k-1} \\ 1 & \alpha_1 & \dots & \alpha_1^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_{n-1} & \dots & \alpha_{n-1}^{k-1} \end{pmatrix}. \tag{3}$$

The minimum distance d of an RS code $\mathcal{C}$ can be computed algebraicly using Lemma 2.1.

Lemma 2.1. *A polynomial of degree D over a field $\mathbb{F}$ has at most D roots (counting multiplicity).*

Proof. Let β_i be roots of the polynomial $f(x)$ of degree D with coefficients in $\mathbb{F}$, and let m_i be the multiplicity of each root β_i. Since $f(\beta_i) = 0$ for each β_i, the product $\prod_i (x - \beta_i)^{m_i}$ necessarily divides $f(x)$ and has degree $\sum_i m_i$. Since no factor of $f(x)$ can have degree greater than that of $f(x)$, it follows that the number of roots, given by the sum of the multiplicities of the β_i is bounded as $\sum_i m_i \leq D$.[1] □

Since the degree of the encoded polynomial in (1) is $k-1$, a codeword c can have at most $k-1$ elements $M(\alpha_i)$ equal to zero. The minimum distance d, equal to the minimum weight of any codeword in $\mathcal{C}$, is bounded as $d \geq n-k+1$. The Singleton bound (proven in Lecture 5) provides a bound of $d \leq n-k+1$ for any code. Hence, the minimum distance of the RS code $\mathcal{C}$ is $d = n-k+1$. This result can be demonstrated by constructing a codeword with exactly $d = n-k+1$ non-zero entries. Let $M(x) = (x-\alpha_0)(x-\alpha_1)\dots(x-\alpha_{k-2})$ be the encoding polynomial as in (1). Since the degree of $M(x)$ is $k-1$, there exists a message $m = [m_0, \dots, m_{k-1}]$ which corresponds to the polynomial $M(x)$, simply by mathinc coefficients in (1). Hence, evaluating $M(x)$ for all $\alpha_i, i = 0, \dots, q-1$ yields a codeword with $k-1$ zeros followed by $n-k+1$ non-zero entries.

RS codes can be used to achieve a relative distance of $\delta = \frac{d}{n} = \frac{n-k+1}{n} = 1 - R + o(1)$ for any rate $R = \frac{k}{n}$. However, the alphabet size q scales as $q = \mathcal{O}(n)$, which is not practical. The fundamental ideas involved in the construction of RS codes can be further investigated through generalization.

3 Reed-Muller Codes

In what follows, a generalization is provided for the RS codes described in Section 2 by expanding the polynomial encoding in (1) to multivariate polynomials. The resulting codes are hereafter referred to as Reed-Muller (RM) codes.[2]

[1] This result also requires that the polynomials of interest are uniquely factorizable, but this is beyond the scope of the lecture. Unique factorizations exist in this case because the space of polynomials of degree at most D forms an extension field of $\mathbb{F}$.

[2] An alternate definition of Reed-Muller codes is common, but Prof. Guruswami claims the multivariate polynomial interpretation is more clear.

3.1 Bivariate RM Codes

We begin with the simplest extension, from univariate to bivariate polynomials. Let m be the matrix $[m_{ij}]$ for $0 \leq i \leq \ell - 1$ and $0 \leq j \leq \ell - 1$ denoting a message of $k = \ell^2$ symbols in $\mathbb{F}_q$. The encoding function

$$Enc : \mathbb{F}_q^{\ell \times \ell} \to \mathbb{F}_q^{q \times q}$$

is given by mapping a message m to a codeword c given by the matrix $[M(\hat{\imath}, \hat{\jmath})]$ for $0 \leq \hat{\imath} \leq q - 1$ and $0 \leq \hat{\jmath} \leq q - 1$, where $M(x, y)$ is given by

$$M(x, y) = \sum_{i=0}^{\ell-1} \sum_{j=0}^{\ell-1} m_{ij} x^i y^j. \tag{4}$$

The resulting RM code is a $[q^2, \ell^2, d]_q$ linear code. Linearity can be verified as in Section 2. The minimum distance d of the RM code can be computed using the following result.

Lemma 3.1. *The tensor product of two $[q, \ell, d]_q$ RS codes $\mathcal{C}_1$ and $\mathcal{C}_2$ is the $[q^2, \ell^2, d^2]_q$ (bivariate) RM code $\mathcal{C}$.*

Proof. The tensor product of two codes $\mathcal{C}_1$ and $\mathcal{C}_2$ is defined as the code $\mathcal{C} = \mathcal{C}_1 \otimes \mathcal{C}_2$ given by

$$\mathcal{C}_1 \otimes \mathcal{C}_2 = \left\{ G_1 m G_2^T \mid m \in \{0, 1\}^{\ell \times \ell} \right\},$$

where G_1 and G_2 are the generator matrices for $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. Since both $\mathcal{C}_1$ and $\mathcal{C}_2$ are RS codes, the matrices G_1 and G_2 are both equal to the RS generator matrix G given in (3). Hence, a message m is mapped to the codeword $M = GmG^T \in \mathcal{C}$. The entry $M(\alpha_x, \alpha_y)$ in row x and column y of the codeword M is given by the product $g_x m g_y^T$, where g_x denotes the row $[1, \alpha_x, \ldots, \alpha_x^{\ell-1}]$ of G, for $0 \leq x \leq q - 1$. Hence, the product code is such that

$$M(\alpha_x, \alpha_y) = \sum_{i=0}^{\ell-1} \sum_{j=0}^{\ell-1} m_{ij} \alpha_x^i \alpha_y^j,$$

which is consistent with the definition of the bivariate Reed-Muller code $\mathcal{C}$ in (4) with x and y replaced with α_x and α_y. □

The use of tensor product codes and the result of Lemma 3.1 implies that the $[q^2, \ell^2, d]_q$ Reed-Muller code has distance $d = (q - \ell + 1)^2 = q^2 - 2q(\ell - 1) + (\ell - 1)^2$ and rate $R = \frac{\ell^2}{q^2}$. Note that the distance $d = (q - \ell + 1)^2$ no longer achieves equality in the Singleton bound $d \leq q^2 - \ell^2 + 1$. However, the alphabet size q in this case scales as $q = \mathcal{O}(\sqrt{n})$. This demonstrates the trade-off between optimal distance and practical alphabet size characteristic of RM codes over RS codes.

3.2 Multivariate RM Codes

The bivariate extension of Section 3.1 generalizes in the natural way to multivariate polynomials. A multivariate RM code $\mathcal{C}$ with v variables $x_1, \ldots, x_v$ can be interpreted as the tensor product code of v RS codes $\mathcal{C}_1, \ldots, \mathcal{C}_v$. The encoding function

$$Enc : \mathbb{F}_q^{\ell_1 \times \cdots \times \ell_v} \to \mathbb{F}_q^{q \times \cdots \times q}$$

maps a message $m = [m_{i_1 \ldots i_v}]$ to a codeword $M(x_1, \ldots, x_v)$ as

$$M(x_1, \ldots, x_v) = \sum_{i_1=0}^{\ell_1 - 1} \cdots \sum_{i_v=0}^{\ell_v - 1} m_{i_1 \ldots i_v} \prod_{j=1}^{v} x_j^{i_j}. \tag{5}$$

The resulting RM code is a $\left[q^v, \prod_{j=1}^{v} \ell_j, \prod_{j=1}^{v} d_j\right]_q$ linear code. Linearity can be verified using an identical method to that of Section 2. The minimum distance d of the multivariate RM code can be proven using the multivariate extension to Lemma 3.1 or using the following result.

Lemma 3.2. *A non-zero polynomial $P(x_1, \ldots, x_v)$ over a field $\mathbb{F}$ with maximum degree d_i for the variable x_i is non-zero in at least $\prod_{i=1}^{v}(q - d_i)$ points in $\mathbb{F}^v$.*

Proof. Fix $x_1, \ldots, x_{v-1}$ and express $P(x_1, \ldots, x_v)$ as

$$P(x_1, \ldots, x_v) = R_{d_v}(x_1, \ldots, x_{v-1}) x_v^{d_v} + \ldots + R_0(x_1, \ldots, x_{v-1}),$$

which is a polynomial of degree d_v in the variable x_v. By Lemma 2.1, there are at least $q - d_v$ values of x_v for which $P(x_1, \ldots, x_v)$ is non-zero. For each of the (at least $q - d_v$) values of x_v which yield non-zero $P(x_1, \ldots, x_v)$, repeat the argument resursively for fixed $x_1, \ldots, x_{v-2}$ and bound the number of values of x_{v-1} for which $P(x_1, \ldots, x_v)$ is non-zero. The desired result is obtained as a lower bound because there may be more than $q - d_i$ non-roots at a given recursion level. □

The following construction demonstrates how equality is achieved in the bound provided by Lemma 3.2. Since the bound results from fewer than $q - d_i$ roots for any given x_i, equality is achieved whenever there are exactly $q - d_i$ roots for each x_i. Hence, let $M_i(x_i)$ be the product $(x_i - \alpha_{i,1}) \ldots (x_i - \alpha_{i,\ell_i - 1})$, where the $\alpha_{i,j}$ are distinct, and let $M(x_1, \ldots, x_v) = \prod_{i=1}^{v} M_i(x_i)$.

3.3 Variant on Multivariate Reed-Muller Codes

We next relax the condition on multivariate RM codes independently bounding the maximum degree of each variable x_i and allow for codeword polynomials $M(x_1, \ldots, x_v)$ with total degree at most ℓ. The encoding function is similar to that in Section 3.2 with the encoding polynomial M given by

$$M(x_1, \ldots, x_v) = \sum_{\substack{i_1, \ldots, i_v \geq 0, \\ i_1 + \ldots + i_v \leq \ell}} m_{i_1 \ldots i_v} \prod_{j=1}^{v} x_j^{i_j}. \tag{6}$$

The resulting code $\mathcal{C}$ is a $[q^v, k, d]_q$ linear code, where q^k is the total number of messages $m_{i_1 \ldots i_v}$ such that $i_1, \ldots, i_v \geq 0$ and $i_1 + \ldots + i_v \leq \ell$. The values of k and d are computed using the following results.

Observation 3.3. *The value k for the given code $\mathcal{C}$ is approximately $\binom{v+l}{v}$ (stated without proof).*

Lemma 3.4. *A non-zero polynomial $P(x_1, \ldots, x_v)$ of total degree at most ℓ over $\mathbb{F}_q$ is zero on at most a fraction $\frac{\ell}{q}$ of points in $\mathbb{F}_q^v$.*

Proof. The statement is proved via induction. The case $v = 1$ states that a univariate polynomial of degree ℓ has at most ℓ roots and is proved using Lemma 2.1. We next note that such a polynomial can be written as

$$P(x_1, \ldots, x_v) = R_{\ell_1}(x_1, \ldots, x_{v-1}) x_v^{\ell_1} + \ldots + R_0(x_1, \ldots, x_{v-1}).$$

The probability that $P(\alpha_1, \ldots, \alpha_v) = 0$ is computed using conditional probability as

$$\begin{aligned}
\Pr[P(\alpha_1, \ldots, \alpha_v) = 0] = & \Pr[P(\alpha_1, \ldots, \alpha_v) = 0 \mid R_{\ell_1}(\alpha_1, \ldots, \alpha_{v-1}) = 0] \\
& \times \Pr[R_{\ell_1}(\alpha_1, \ldots, \alpha_{v-1}) = 0] \\
= & \Pr[P(\alpha_1, \ldots, \alpha_v) = 0 \mid R_{\ell_1}(\alpha_1, \ldots, \alpha_{v-1}) \neq 0] \\
& \times \Pr[R_{\ell_1}(\alpha_1, \ldots, \alpha_{v-1}) \neq 0] \\
\leq & 1 \times \frac{\ell - \ell_1}{q} + 1 \times \frac{\ell_1}{q} = \frac{\ell}{q} \qquad (7)
\end{aligned}$$

where the inequality is due to the trivial inequality that the corresponding probabilities are at most 1 and the induction step. □

The result of Lemma 3.4 can then be used to yield the result that (assuming $\ell \leq q$) the distance of the code $\mathcal{C}$ can be bounded as $d \geq \left(1 - \frac{\ell}{q}\right) q^v$. This suggests that RM codes do not provide $R, \delta > 0$ for constant q, i.e. q increases with n.

4 Binary Reed-Muller Codes

We now shift our attention to the binary version of Reed-Muller codes (defined by Muller 1954 with decoding due to Reed 1954) in which $q = 2$ and codes are defined over $\mathbb{F}_2 = \{0, 1\}$. The binary RM code $\mathcal{C}$ thus results from encoding each point in $\mathbb{F}_2^v$ using the *multilinear* encoding polynomial M given by

$$M(x_1, \ldots, x_v) = \sum_{S \,:\, |S| \leq \ell} c_S \prod_{i \in S} x_i,$$

where the coefficient c_S depends on the message m. The binary RM code $\mathcal{C}$ is thus a $[2^v, \sum_{i=0}^{\ell} \binom{v}{i}, d]_2$ linear code. The distance d is given by the following lemma.

Lemma 4.1. *The minimum distance d of the binary RM code described above is $d = 2^{v-\ell}$.*

Proof. Consider the encoding polynomial $M(x_1, \ldots, x_v) = \prod_{i=1}^{\ell} x_i$ resulting from the message leading to the coefficient $c_S = 1$ if and only if $S = \{1, \ldots, \ell\}$. There are exactly $2^{v-\ell}$ choices for $(x_1, \ldots, x_v)$ that make M non-zero, namely those with $x_1 = \ldots = x_\ell = 1$. The distance d is thus bounded as $d \leq 2^{v-\ell}$. Next, consider the non-zero polynomial $M(x_1, \ldots, x_v)$ and let $\prod_{i=1}^{r} x_i$ be the maximal monomial of M, i.e. reorder the indices $\{1, \ldots, v\}$ such that

$$M(x_1, \ldots, x_v) = \prod_{i=1}^{r} x_i + R(x_1, \ldots, x_v)$$

where there is no monomial term in $R(x_1, \ldots, x_v)$ with more than r variables. There are 2^{v-r} ways to choose the variables $x_{r+1}, \ldots, x_v$, but none of them can cause the maximal monomial to be cancelled. This leads to the bound $d \geq 2^{v-r}$, which implies $d \geq 2^{v-\ell}$ since $r \leq \ell$ by the definition of M. □

5 Summary

Two families of linear codes, Reed-Solomon and Reed-Muller, were presented and analyzed using various algebraic properties. Though the Reed-Solomon codes can be used to achieve $R, \delta > 0$, this can only be done if the alphabet size q increases linearly in the block length, i.e. $q = \mathcal{O}(n)$. The use of Reed-Muller codes effectively generalize the Reed-Solomon codes, offering a reduced alphabet size of $q = \mathcal{O}(\sqrt[v]{n})$ for an asymptotically decaying relative distance δ.