IERG6120 Coding for Distributed Storage Systems	Lecture 4 - 22/09/2016

The Singleton bound and Reed-Solomon Code

Lecturer: Kenneth Shum *Scribe: Bowen Zhang*

1 Minimum Hamming distance and minimum Hamming weight

Notation 1. *A q-ary linear code of block length n and dimension k will be referred to as an $[n,k]_q$-code. Further, if the code has minimum distance d, it will be referred to as an $[n,k,d]_q$-code. The alphabet is a finite field $GF(q)$ of size q.*

Remark: We showed in lecture 3 how to perform calculations in a prime field. In this lecture we can take q as a prime number. Anyway, all materials in this lecture note holds true for a finite field in general.

Here we give the definition of the Hamming distance and Hamming weight for a linear code $\mathcal{C}$.

Definition 1. *The minimum Hamming distance d between codewords in a linear code $\mathcal{C}$ is defined as:*

$$d_H(\mathcal{C}) \triangleq min\,\{d_H(\boldsymbol{u},\boldsymbol{v}) : \boldsymbol{u},\boldsymbol{v} \in \mathcal{C}, \boldsymbol{u} \neq \boldsymbol{v}\} \tag{1}$$

The minimum weight w of a linear code $\mathcal{C}$ is defined as:

$$w_H(\mathcal{C}) \triangleq min\,\{w_H(\boldsymbol{u}) : \boldsymbol{u} \in \mathcal{C}, \boldsymbol{u} \neq \boldsymbol{0}\} \tag{2}$$

Now for theorem about Hamming distance.

Theorem 1. *The minimum Hamming distance d between codewords in a linear code $\mathcal{C}$ is equal to the minimum number of non-zero symbols in a code vector $\boldsymbol{u} \neq \boldsymbol{0}$ of $\mathcal{C}$, which is $d_H(\mathcal{C}) = w_H(\mathcal{C})$.*

Proof Let $\mathbf{u}$ and $\mathbf{v}$ be distinct code vectors with minimum distance $d_H(\mathbf{u},\mathbf{v}) = d$. Then $\mathbf{a} = \mathbf{u} - \mathbf{v}$ is also a code vector in $\mathcal{C}$, the number of non-zero symbols in $\mathbf{a}$ is equal to d. This shows that the minimum number of non-zero symbols in a code is less than or equal to d, i.e., $d_H(\mathcal{C}) \geq w_H(\mathcal{C})$. To prove the opposite inequality, note that $\mathbf{0}$ is a code vector in a linear code. Let $\mathbf{u}$ be a nonzero codeword in $\mathcal{C}$ whose Hamming weight is the smallest among the nonzero codewords. We have $d_H(\mathbf{u},\mathbf{0})$ equal to the number of non-zero symbols in $\mathbf{u}$, so $d_H(\mathcal{C}) \leq d_H(\mathbf{u},\mathbf{0}) = w_H(\mathcal{C})$. □

Example 1 Consider a linear code over $\mathbb{Z}_3$ whose codewords are in the row span of matrix

$$\begin{bmatrix} 1 & 1 & 2 & 0 \\ 2 & 0 & 1 & 1 \end{bmatrix}.$$

The codewords are listed as follows:

0 0 0 0
1 1 2 0
2 2 1 0
2 0 1 1
0 1 0 1
1 2 2 1
1 0 2 2
2 1 1 2
0 2 0 2

It's easy to verify $d_H(\mathcal{C}) = w_H(\mathcal{C})$.

2 Minimal distance and parity-check matrix

Recall that a $k \times n$ matrix $\mathbf{G}$ is a generator matrix of an $[n,k]_q$-code $\mathcal{C}$ if the rows of $\mathbf{G}$ form a basis of $\mathcal{C}$. An $(n-k) \times n$ matrix $\mathbf{H}$ is a parity-check matrix of $\mathcal{C}$ if the rows of $\mathbf{H}$ form a basis of the dual code $\mathcal{C}^\perp := \{\mathbf{u} : \mathbf{u} \cdot \mathbf{c} = 0, \forall \mathbf{c} \in \mathcal{C}\}$. The matrix $\mathbf{G}$ and $\mathbf{H}$ satisfy $\mathbf{G}\mathbf{H}^T = \mathbf{0}$.

We can characterize the minimum distance of a code in terms of a generator matrix or a parity-check matrix.

Theorem 2. *Suppose that $\mathbf{H}$ is a parity-check matrix of a linear $[n,k]_q$-code $\mathcal{C}$. Then $\mathcal{C}$ has minimum distance d if and only if any $d-1$ or less columns of $\mathbf{H}$ are linearly independent, but we can find d columns of $\mathbf{H}$ that are linearly dependent.*

Proof For $i = 1, 2, \ldots, n$, let $\mathbf{h}_i$ denote the i-th column of the parity-matrix $\mathbf{H}$.

Consider a general vector $(a_1, a_2, \ldots, a_n)$ of length n. The vector $(a_1, a_2, \ldots, a_n)$ is a codeword if and only if $(a_1, a_2, \ldots, a_n) \cdot \mathbf{H}^T = \mathbf{0}$. Denote the indices of the *nonzero* components of $(a_1, a_2, \ldots, a_n)$ by $i_1 < i_2 < \ldots < i_w$, so that $a_{i_1}, a_{i_2}, \ldots a_{i_w}$ are all nonzero. A nonzero vector $(a_1, a_2, \ldots, a_n)$ with nonzero components precisely at $i_1, i_2, \ldots, i_w$ is a codeword if and only if

$$a_{i_1}\mathbf{h}_{i_1} + a_{i_2}\mathbf{h}_{i_2} + \cdots + a_{i_w}\mathbf{h}_{i_w} = \mathbf{0}.$$

By Theorem 1, a linear code $\mathcal{C}$ has minimum distance d if and only if there is a codeword of weight d, but no codeword of weight strictly less than d, except the zero codeword. Hence, $d_H(\mathcal{C}) = d$ if and only if there are d column indices, say $j_1 < j_2 < \cdots < j_d$, such that the columns of $\mathbf{H}$ with indices $j_1, j_2, \ldots, j_d$ are linearly dependent, but any $d-1$ or less columns of $\mathbf{H}$ are linearly independent. $\square$

Theorem 3. *Suppose $\mathbf{G}$ is a generator matrix of a linear $[n,k]_q$-code $\mathcal{C}$. Then $\mathcal{C}$ has minimum distance d if and only if any $n-d+1$ or more columns of $\mathbf{G}$ form a matrix with rank k, but we can find a $k \times (n-d)$ submatrix of $\mathbf{G}$ which has rank strictly less than k.*

Proof Consider a nonzero vector $(a_1, a_2, \ldots, a_n)$ of length n. It is a codeword in $\mathcal{C}$ if and only if there is a nonzero message vector $(m_1, \ldots, m_k)$ such that

$$(m_1, \ldots, m_k) \cdot \mathbf{G} = (a_1, a_2, \ldots, a_n).$$

Denote the indices of the *zero* components of $(a_1, a_2, \ldots, a_n)$ by $i_1 < i_2 < \ldots < i_{n-w}$, so that $a_{i_1}, a_{i_2}, \ldots a_{i_{n-w}}$ are all zero. A nonzero vector $(a_1, a_2, \ldots, a_n)$ with zero components precisely at positions $i_1, i_2, \ldots, i_{n-w}$ is a codeword if and only if the rows of the $k \times (n-w)$ submatrix of $\mathbf{G}$, with column indices $i_1, i_2, \ldots i_{n-w}$, has a nontrivial linear combination that is equal to zero, i.e., the submatrix has rank strictly less than k.

By Theorem 1, a linear code $\mathcal{C}$ has minimum distance d if and only if there is a codeword of weight d, but no codeword of weight strictly less than d, except the zero codeword. Hence, $d_H(\mathcal{C}) = d$ if and only if there are $n-d$ column indices, say $j_1 < j_2 < \ldots, j_{n-w}$, such that the submatrix of $\mathbf{G}$ obtained by retaining columns $j_1, j_2, \ldots, j_{n-w}$ has rank strictly less than k, but any $n-d+1$ or more columns of $\mathbf{G}$ form a matrix with rank k. $\square$

As an example consider the code in Example 1. The matrix

$$\mathbf{H} = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix}$$

is a parity-check matrix. The vector $(0,1,0,1)$ is a codeword with minimum weight 2, and the non-zero components of $(0,1,0,1)$ are located at position 2 and 4. Columns 2 and 4 of $\mathbf{H}$,

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

are scalar multiple of each other. However, each column of $\mathbf{H}$ is non-zero, and thus linearly independent. We can also check that any 3 or more columns of the generator matrix form a matrix of rank k, but the submatrix of $\mathbf{G}$ obtained by retaining columns 1 and 3, namely

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

only has rank 1.

Theorem 4 (Singleton bound)**.** *The minimum distance of an linear $[n,k]_q$-code satisfies*

$$d \le n - k + 1.$$

Proof We can prove the Singleton bound by Theorem 2 or 3.

(Proof 1) Since there are $n-k$ rows in a parity-check matrix of $\mathcal{C}$, any $n-k+1$ columns of a parity-matrix are linearly dependent. By Theorem 2, we obtain $d \le n-k+1$.

(Proof 2) Pick any $k-1$ columns of a generator matrix of $\mathcal{C}$. They form a matrix with rank strictly less than k. By Theorem 3, we have $n-d \ge k-1$. □

Definition 5. *Any code satisfying the Singleton bound with equality is called a* maximal-distance separable (MDS) *code.*

In the next section, we give an important family of MDS codes.

3 Reed-Solomon Codes

All linear codes offer the advantages that they are easy to encode, and the minimum Hamming distance reduces to a simpler concept, the weight. However, the weight of an arbitrary linear code is still not easy to determine. Also, the mere fact that a code is linear doesn't help all that much in applying the standard decoding algorithm. In order to compute the weight an decode more easily, we need to use linear codes with special properties. Usually the special properties are based on algebra, in which case the code is called an *algebraic code*. The Reed-Solomon codes that we will define are examples of algebraic codes.

Definition 6. *Let p be a prime number and let $k \le n \le p$. The* Reed-Solomon code *over the field $GF(p)$ with dimension k and length n is defined as follows. Let $\mathcal{A} = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$ be a set of n distinct element in $GF(p)$. Let $\mathcal{P}(k)$ be the set of polynomials*

$$\mathcal{P}(k) \triangleq \left\{ c_0 + c_1 x^1 + \ldots + c_{k-1} x^{k-1} : c_i \in GF(p)) \right\} \tag{3}$$

with degree strictly less than k. $\mathcal{P}(k)$ is a vector space on $GF(p)$. The cardinality of $\mathcal{P}(k)$ is p^k and the dimension of $\mathcal{P}(k)$ over $GF(p)$ is k. The coefficients c_i's are the message symbol to be encoded. The code vector corresponding to the message vector $(c_0, \dots, c_{k-1})$ is obtained by evaluating the polynomial $f(x) = c_0 + c_1 x^1 + \ldots + c_{k-1} x^{k-1}$ at the points in $\mathcal{A}$,

$$(f(\alpha_1), f(\alpha_2), \dots, f(\alpha_n)).$$

We now show that the encoding function mapping a polynomial f in $\mathcal{P}(k)$ to the codeword vector $(f(\alpha_1), f(\alpha_2), \dots, f(\alpha_n))$ is one-to-one. This amounts to saying that if two message polynomials $P_1(x)$ and $P_2(x)$ give rise to the the same code vector, that is, if

$$P_1(\alpha_i) = P_2(\alpha_i) \text{ for } i = 1, 2, \dots, n, \tag{4}$$

then we must have $P_1(x) = P_2(x)$. To see this, observe that $P_1(x) - P_2(x)$ is a polynomial of degree less than k, and it has n distinct roots $\alpha_1, \dots, \alpha_n$. Since $k \le n$, the results follows from

Theorem 7. *Let F be a field and $P(x)$ be a non-zero polynomial of degree e with coefficients in F. Then, $P(x)$ has no more than e distinct roots in F.*

Proof If degree $e = 0$, then the polynomial is a nonzero constant, which has no root. So, the number of roots is zero and the degree is also zero.

Suppose the theorem holds for all degree strictly less than e, and consider a polynomial $P(x)$ with degree e. If $P(x)$ has no root in F, then the theorem holds for $P(x)$. Otherwise, if α is a root of $P(x)$, then by performing long division, we can factor $P(x)$ as $(x-\alpha)f(x)$ for some polynomial $f(x)$ with degree $e-1$. Any root of $p(x)$, say β, which is not equal to α must be a root of $f(x)$, because

$$(\beta - \alpha)f(\beta) = P(\beta) = 0 \Rightarrow f(\beta) = 0.$$

By applying the induction hypothesis, the polynomial $f(x)$ has no more than $e-1$ distinct roots, and hence the polynomial $P(x)$ has no more than e distinct roots. $\square$

Theorem 8. *A Reed-Solomon code is a linear code with length n, dimension k and generator matrix*

$$\mathbf{G} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \alpha_1 & \alpha_2 & \dots & \alpha_n \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{k-1} & \alpha_2^{k-1} & \dots & \alpha_n^{k-1} \end{bmatrix} \tag{5}$$

Theorem 9. *The minimum distance d of an $[n, k]$ Reed-Solomon code is equal to $n - k + 1$.*

Proof (proof 1) Any $k \times k$ submatrix of the generator matrix in (5) is a nonsingular Vandermonde matrix. Hence any k or more columns of (5) form a matrix with rank k. By Theorem 3, we have $n - d + 1 \leq k$, implying $d \geq n - k + 1$. Together with the Singleton bound $d \leq n - k + 1$, we conclude that $d = n - k + 1$.

(proof 2) The components of a codeword is obtained by evaluating a polynomial of degree less than or equal to $k-1$ at n distinct points. By Theorem (7), there are at most $k-1$ points which are evaluated to zero. Therefore, the number of nonzero components is at least $n - k + 1$. $\square$

Example 2 Take $GF(7)$ as the alphabet. A Reed-Solomon code of length 7 and dimension 3 has minimum distance 5. If we associate the columns of a generator with finite field elements $0, 1, \dots, 6$, we can write down a generator matrix as

$$\mathbf{G} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 1 & 2^2 & 3^2 & 4^2 & 5^2 & 6^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 1 & 4 & 2 & 2 & 4 & 1 \end{bmatrix}$$

We note that any 3×3 submatrix of $\mathbf{G}$ is nonsingular.

Exercise: Show that

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 1 & 2^2 & 3^2 & 4^2 & 5^2 & 6^2 \\ 0 & 1 & 2^3 & 3^3 & 4^3 & 5^3 & 6^3 \end{bmatrix}$$

is a parity-check matrix of the Reed-Solomon matrix in Example 2.