Pries: M460 - Information and Coding Theory, Spring 2019
Homework 8: Polynomials over finite fields. Due Friday 4/12.

1. Find the three monic polynomials g_1, g_2, g_3 in $\mathbb{Z}/3[x]$ which have degree 2 and are irreducible over $\mathbb{Z}/3$. Compute the product $(x^3 - x)g_1g_2g_3$ modulo 3.

2. (a) Find the roots of $x^2 - 1$ in $\mathbb{Z}/8$. Why is the answer surprising?
   (b) Find two truly different factorizations of $x^2 - 1$ in $\mathbb{Z}/8[x]$.

3. This problem is about the old Reed Solomon code RS$(3, 7)$ with $q = 7$ and $k = 3$.
   (a) Use the basis $\{1, x, x^2\}$ of L_2 to find the generator matrix of RS$(3, 7)$.
   (b) Find the minimal distance d for RS$(3, 7)$ and find a codeword of weight d.
   (c) Encode the data $(c_0, c_1, c_2) = (3, 2, 1)$ into a codeword of RS$(3, 7)$.
   (d) You receive the codeword $(4, 1, 0, 1, 4, 2)$. Decode to find the data (c_0, c_1, c_2).
   (e) You receive the codeword $(1, 0, 0, 4, 5, 3)$. If one error occurred and you have high confidence in the digits which are zero, what is the corrected codeword and what is the data?

1 Week 10: Version 1 of Reed Solomon

Reed-Solomon codes were invented in 1960 at MIT Lincoln Lab. At that time, technology was too weak to implement them. Their first uses (in the early 1980s) were for digital photos for Voyager space probe (info rate 44.8 kbits/sec) and compact disks.

Example 1.1. Let $q = 5$ and $k = 3$. Let $\alpha_1 = 1, \ldots, \alpha_4 = 4$. A basis for L_2 is $\{1, x, x^2\}$. Let $f_1(x) = 1$, let $f_2(x) = x$, and $f_3(x) = x^2$. Then $f_1(\vec{\alpha}) = (1, 1, 1, 1)$, and $f_2(\vec{\alpha}) = (1, 2, 3, 4)$, and $f_3(\vec{\alpha}) = (1, 4, 4, 1)$. These three vectors are a basis for $\mathrm{RS}(2, 5)$. This is a code over $\mathbb{Z}/5$ with length 4 and dimension 3 (info rate 3/4).

Example 1.2. Suppose $q = 5$ and $k = 3$. Then we can transmit three pieces of data, let's say (c_0, c_1, c_2). First we create a function $f(x) = c_0 + c_1x + c_2x^2$. Then the code word is $(f(1), f(2), f(3), f(4))$. Specifically, if the data is $(4, 0, 1)$ then $f(x) = x^2 + 4$ and the code word is $(0, 3, 3, 0)$.

This can be implemented easily using linear algebra.

Example 1.3. $d = 2$. Have already found codeword of weight 2 and a non-zero polynomial with degree 2 has at most 2 roots.

This reaches Singleton bound.

Notation 1.4. Let $q = p^n$ be a prime power and label the non-zero elements of $\mathbb{F}_q$ as $\alpha_1, \ldots, \alpha_{q-1}$. Let k be such that $1 \le k \le q - 1$. Let L_{k-1} be the set of polynomials $g(x)$ with $\deg(g) \le k - 1$. Given $f \in L_{k-1}$, let $f(\vec{\alpha}) = (f(\alpha_1), \ldots, f(\alpha_{q-1}))$.

Definition 1.5. The Reed-Solomon code $\mathrm{RS}(k, q)$ is the subset of $\mathbb{F}_q^{q-1}$ consisting of all vectors $f(\vec{\alpha})$ for $f \in L_{k-1}$.

Example 1.6. If $k = q - 1$, then $\mathrm{RS}(q - 1, q) = \mathbb{F}_q^{q-1}$. If $k = 1$, then $\mathrm{RS}(1, q)$ is the $q - 1$ repetition code.

Lemma 1.7. 1. *L_{k-1} is a vector space over $\mathbb{F}_q$ of dimension k, with basis $1, x, \ldots, x^{k-1}$.*

2. *$\mathrm{RS}(k, q)$ is a linear code over $\mathbb{F}_q$ of length $q - 1$ and dimension k.*

3. *The number of codewords is q^k.*

Proof. 1. A basis for L_{k-1} over $\mathbb{F}_q$ is $\{1, x, x^2, \ldots, x^{k-1}\}$.

2. Suppose $f_1(\vec{\alpha})$ and $f_2(\vec{\alpha})$ are in $\mathrm{RS}(k, q)$. Then $f_1(x)$ and $f_2(x)$ are in L_{k-1}. By part (i), L_{k-1} is a vector space over $\mathbb{F}_q$. So if $c \in \mathbb{F}_q$, then $cf_1 + f_2 \in L_{k-1}$. Then $cf_1(\vec{\alpha}) + cf_2(\vec{\alpha}) = (cf_1 + f_2)(\vec{\alpha}) \in \mathrm{RS}(k, q)$. Thus $\mathrm{RS}(k, q)$ is a linear code over $\mathbb{F}_q$.

   Consider the k vectors $\vec{v}_i = (\alpha_1^i, \ldots, \alpha_{q-1}^i)$ $0 \le i \le k - 1$ (these are found by evaluating the basis functions for L_{k-1}. By definition, they span $\mathrm{RS}(k, q)$. To finish, we need to show they are linearly independent.

   Remark: Fermat's little Theorem evaluating the function x^{p-1} on the non-zero entries of $\mathbb{Z}/p$ gives the constant vector $\vec{1}$. Luckily this is excluded by our bounds on k.

Method 1: Show there exist functions f_i for $1 \leq i \leq k$, such that $f_i(\alpha_j) = \delta_{ij}$ for $1 \leq j \leq k$. This shows that $\mathrm{RS}(k, q)$ contains k codewords beginning with $(1, 0, 0, 0, ...)$, $(0, 1, 0, 0, ...)$, so its dimension is at least k.

Example: for $k = 2$. Given $\alpha_1 \neq \alpha_2 \in \mathbb{F}_q$, find f_1, f_2 polynomials of degree 1 such that

$$f_1(\alpha_1) = 1, \qquad f_2(\alpha_1) = 0,$$

and

$$f_1(\alpha_2) = 0, \qquad f_2(\alpha_2) = 1.$$

□

Example 1.8. Given distinct values $\alpha_1, \alpha_2, \alpha_3$, find f_1 a polynomial of degree 2 such that

$$f_1(\alpha_1) = 1, \qquad f_1(\alpha_2) = 0, \qquad f_1(\alpha_3) = 0.$$

Example 1.9. Let's investigate the Reed-Solomon code $\mathrm{RS}(2, 5)$.

1. What is a basis for the codewords? How many codewords are there? What is the matrix $M_{2,5}$?

2. If the data is $\vec{c} = (1, 3)$, what is the codeword? Answer $(4, 2, 0, 3)$.

3. If the codeword is $\vec{b} = (1, 0, 4, 3)$, what is the data? Answer $(2, 4)$.

4. If the codeword is $\vec{b} = (1, 0, 2, 4)$, show that an error occurred in transmission. What is the best guess for the data vector $\vec{c}$?

   Linear system: $g = c_0 + c_1 x$. Then

   $$c_0 + c_1 = 1, \; c_0 + 2c_1 = 0, \; c_0 + 3c_1 = 2, \; c_0 + 4c_1 = 4.$$

   This is inconsistent. It turns out that the first digit is incorrect. Taking $c_0 = 1$ and $c_1 = 2$ corrects to $(3, 0, 2, 4)$.

5. What is the distance invariant for $\mathrm{RS}(2, 5)$? How many errors can this code detect? How many errors can this code correct?

Remark 1.10. Notice that L_{k-1} is the vector space of functions on $\mathbb{P}^1$ whose only poles are at ∞ and such that the order of the pole at ∞ is at most $k - 1$. The entries of a codeword are the values of the function at the (non-zero) points of $\mathbb{P}^1$.

Transmitting data One interesting thing about the Reed-Solomon code is that all the entries are treated equally. Unlike the ISBN code or the $(7, 4)$ code, there are no 'check digits'. The data does not appear as part of the transmitted message. So how does it work?

Transmitting data:

Definition 1.11. If $\vec{c} = (c_0, \ldots, c_{k-1})$ is the data, then let $f_{\vec{c}}(x) = \sum_{i=0}^{k-1} c_i x^i$ and the code word is $f_{\vec{c}}(\vec{\alpha})$.

Definition 1.12. The generator matrix $M_{k,q}$ for $\mathrm{RS}(k, q)$ is the matrix with $q - 1$ columns and k rows constructed as follows: let $f_1(x) = 1, f_2(x) = x, \ldots, f_k(x) = x^{k-1}$. The jth row of $M_{k,q}$ is the codeword $f_k(\vec{\alpha})$. Given a data vector $\vec{c} = (c_0, \ldots, c_{k-1})$, the code word is $\vec{c}M_{k,q}$.

Example 1.13.

$$M_{3,5} = \begin{matrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 4 & 1 \end{matrix}$$

Interpreting data: Suppose the codeword is $\vec{b} = (b_1, \ldots, b_{q-1})$. We need to find a function $f(x)$ with degree at most $k - 1$ such that $f(\vec{\alpha}) = \vec{b}$, i.e., such that $f(\alpha_i) = b_i$. Then the data is given by the coefficients of $f(x)$.

Lemma 1.14. *The data vector $\vec{c}$ is the solution to the linear system $\vec{c}M_{k,q} = \vec{b}$.*

Specifically, we can find the data vector $\vec{c}$ using row reduction.

Lemma 1.15. *If the linear system is inconsistent, then there has been an error in transmission.*

This was improved by Berlekamp and Massey in 1969.

Remark 1.16. Problem - encoding and decoding is slow, uses $k(n-k)\log_2(n)$ operations. Will later improve this using low density parity check codes which also have high rate and short blocks. Another problem - need to choose k/n before transmission, can't improve on the fly.

The minimal distance of a Reed-Solomon code Recall that the minimal distance of a code is the smallest non-zero Hamming distance between two codewords. This measures how many entries are different between the codewords. It determines the number of errors in transmission that can be detected and corrected. The distance of a Reed-Solomon code is optimal given the fixed length $q - 1$ and dimension k.

Example 1.17. There is no vector in $\mathrm{RS}(5, 3)$ with weight 1. To see this, suppose $f(\vec{\alpha})$ is a vector in $\mathrm{RS}(5, 3)$ with length 4 and weight 1; this means exactly three entries are zero, e.g., $(0, 0, 0, 2)$. Then $f(x)$ has three roots in $\mathbb{Z}/5$. This is impossible since $\deg(f) \leq 2$.

Theorem 1.18. *The Reed-Solomon code* $\mathrm{RS}(k, q)$ *has distance $d = q - k$.*

Proof. By the Singleton bound, $d \leq n - k + 1$ where n is the length of the codewords. Since $n = q - 1$, this implies $d \leq q - k$.

To prove that $d \geq q - k$, recall that the minimal distance of a linear code is the same as the minimal weight. The weight of a codeword is the number of its entries which are non-zero. So we need to show that if $f(\vec{\alpha})$ is a non-zero codeword, then the number of its entries which are non-zero is at least $q-k$. In other words, we need to show that the number of roots of $f(x)$ is at most $(q-1)-(q-k)=k-1$. This is true since $f(x)$ has degree at most $k-1$. □

The Reed-Solomon codes are very good codes, especially when used in concatenated codes. Good for erasure correction - if any k signals received then can decode. On digital CD's can correct burst errors up to 4000 bits. (The number of consecutive errors goes down dramatically in outer code).

For fixed q and k, the distance of the code $\mathrm{RS}(k,q)$ is optimal. One drawback of the Reed-Solomon codes is that, once q is fixed, the length of the codewords is fixed at $q-1$ and the dimension is bounded by $q-1$. Next time we will look for codes where the distance and dimension can be large relative to q.