University of California, Berkeley Instructor: Venkatesan Guruswami
CS 294-226: Advances in Error-Correcting Codes Term: Fall 2022

PROBLEM SET 2

Due by 11.59pm PDT Friday, October 14

INSTRUCTIONS

- You are allowed to collaborate with up to two students taking the class in solving problem sets. But here are some rules concerning such collaboration:

  1. You should think about each problem by yourself for at least 30 minutes before commencing any collaboration.
  2. Collaboration is defined as discussion of the lecture material and solution approaches to the problems. Please note that *you are not allowed to share any written material and you must write up solutions on your own*. You must clearly acknowledge your collaborator(s) in the write-up of your solutions.
  3. Of course, if you prefer, you can also work alone.

- Solutions typeset in LaTeX are encouraged, but not required. If you are submitting handwritten solutions, please write clearly and legibly (you might want to first write the solution sketch in rough, before transferring it to the version you turn in which should be a "finished product").
- Solutions should be turned in via gradescope.
- You may consult any of the class notes or videos that were assigned for initial preparation, and the "Essential Coding Theory" text linked from the course page, but please refrain from consulting other sources. In particular, you should not search for solutions on the web. Do not hesitate to contact the GSI or instructor should you have any questions.
- Please start work on the problem set early. The problems vary in difficulty and length and it will be a good idea to give yourself enough time for the trickier problems.

A. **(This problem need not be turned in)**

(a) Let C_1 be an $[n_1, k_1, d_1]_q$ linear code, and C_2 an $[n_2, k_2, d_2]_q$ linear code. Define their tensor product code $C \subseteq \mathbb{F}_2^{n_1 \times n_2}$, denoted $C = C_1 \otimes C_2$, to be the subset of $n_1 \times n_2$ matrices whose columns belong to C_1 and whose rows belong to C_2.

Prove that C is an $[n_1 n_2, k_1 k_2, d_1 d_2]_q$ linear code.

(b) Consider the bivariate version of the Reed-Solomon code, which encodes a polynomial $f \in \mathbb{F}_q[X, Y]$ with degree less than k in *both* X and Y by its evaluations at all q^2 points $(\alpha, \beta) \in \mathbb{F}_q \times \mathbb{F}_q$. What are the block length, dimension, and minimum distance of this code? Clearly justify your answer. ((The natural) hint: Relate to Part (a) of this question.)

B. **(This problem need not be turned in)** For a field $\mathbb{F}$ with $|\mathbb{F}| \geq n$, an n-tuple $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ of n *distinct* elements of $\mathbb{F}$, and a vector $\mathbf{v} = (v_1, v_2, \ldots, v_n) \in (\mathbb{F}^*)^n$ of n (not necessarily distinct) nonzero elements from $\mathbb{F}$, the *Generalized Reed-Solomon code* $\mathrm{GRS}_{\mathbb{F}}(\vec{\alpha}, \mathbf{v}, k)$ is defined as follows:

$$\mathrm{GRS}_{\mathbb{F}}(\vec{\alpha}, \mathbf{v}, k) = \{(v_1 \cdot p(\alpha_1), v_2 \cdot p(\alpha_2), \ldots, v_n \cdot p(\alpha_n)) \mid p(X) \in \mathbb{F}[X] \text{ has degree} < k\} .$$

(a) Check that $\mathrm{GRS}_{\mathbb{F}}(\vec{\alpha}, \mathbf{v}, k)$ is an $[n, k, n-k+1]_{\mathbb{F}}$ linear code.

(b) Prove that the dual code of $\mathrm{GRS}_{\mathbb{F}}(\vec{\alpha}, \mathbf{v}, k)$ is

$$\mathrm{GRS}_{\mathbb{F}}(\vec{\alpha}, \mathbf{v}, k)^{\perp} = \mathrm{GRS}_{\mathbb{F}}(\vec{\alpha}, \mathbf{u}, n-k)$$

for $\mathbf{u} = (u_1, u_2, \ldots, u_n) \in (\mathbb{F}^*)^n$ where for $i = 1, 2, \ldots, n$,

$$u_i = \frac{1}{v_i \prod_{j \neq i}(\alpha_i - \alpha_j)} .$$

---

1. In this problem, you will prove that a certain "ultimate" form of Reed-Solomon decoding is NP-hard. You may assume that the following problem is NP-hard.

   Instance: A set $S = \{\alpha_1, \ldots, \alpha_n\} \subseteq \mathbb{F}_{2^m}$, an element $\beta \in \mathbb{F}_{2^m}$, and an integer $1 \leq k < n$.
   Question: Is there a nonempty subset $T \subseteq \{1, 2, \ldots, n\}$ with $|T| = k+1$ such that $\sum_{i \in T} \alpha_i = \beta$?

   Consider the $[n, k]$ Reed-Solomon code C_{RS} over $\mathbb{F}_{2^m}$ obtained by evaluating polynomials of degree at most $k-1$ at points in S. Define $y \in (\mathbb{F}_{2^m})^n$ as follows: $y_i = \alpha_i^{k+1} - \beta \alpha_i^k$ for $i = 1, 2, \ldots, n$.

   Prove that there is a codeword of C_{RS} at Hamming distance at most $n-k-1$ from y if and only if there is a set T as above of size $k+1$ satisfying $\sum_{i \in T} \alpha_i = \beta$.

   Conclude that finding the nearest codeword in a Reed-Solomon code over exponentially large fields is NP-hard.

2. In this problem, we will consider the number-theoretic counterpart of Reed-Solomon codes. Let $1 \leq k < n$ be integers and let $p_1 < p_2 < \cdots < p_n$ be n distinct primes. Denote $K = \prod_{i=1}^{k} p_i$ and $N = \prod_{i=1}^{n} p_i$. The notation $\mathbb{Z}_M$ stands for integers modulo M, i.e., the set $\{0, 1, \ldots, M-1\}$. Consider the *Chinese Remainder code* defined by the encoding map $E : \mathbb{Z}_K \to \mathbb{Z}_{p_1} \times \mathbb{Z}_{p_2} \times \cdots \times \mathbb{Z}_{p_n}$ defined by:

   $$E(m) = (m \mod p_1,\ m \mod p_2,\ \cdots,\ m \mod p_n) .$$

   (Note that this is not a code in the usual sense we have been studying since the symbols at different positions belong to different alphabets. Still notions such as distance of this code make sense and are studied in the questions below.)

   (a) Suppose that $m_1 \neq m_2$. For $1 \leq i \leq n$, define the indicator variable $b_i = 1$ if $E(m_1)_i \neq E(m_2)_i$ and $b_i = 0$ otherwise. Prove that $\prod_{i=1}^{n} p_i^{b_i} > N/K$.
   Use the above to deduce that when $m_1 \neq m_2$, the encodings $E(m_1)$ and $E(m_2)$ differ in at least $n-k+1$ locations.

   (b) This exercise examines how the idea behind the Welch-Berlekamp decoder can be used to decode these codes.
   Suppose $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ is the received word where $r_i \in \mathbb{Z}_{p_i}$. By Part (a), we know there can be at most one $m \in \mathbb{Z}_K$ such that

   $$\prod_{i : E(m)_i \neq r_i} p_i^{b_i} \leq \sqrt{N/K} . \tag{1}$$

   (Be sure you see why this is the case.) The exercises below develop a method to find the unique such m, assuming one exists.
   In what follows, let r be the unique integer in $\mathbb{Z}_N$ such that $r \mod p_i = r_i$ for every $i = 1, 2, \ldots, n$ (note that the Chinese Remainder theorem guarantees that there is a unique such r).

i. Assuming an m satisfying (1) exists, prove that there exist integers y, z with $0 \le y < \sqrt{NK}$ and $1 \le z \le \sqrt{N/K}$ such that $y \equiv rz \pmod{N}$.

ii. Prove also that if y, z are any integers satisfying the above conditions, then in fact $m = y/z$.

(**Remark:** A pair of integers (y, z) satisfying above can be found by solving the integer linear program with integer variables y, z, t and linear constraints: $0 < z \le \sqrt{N/K}$; and $0 \le z \cdot r - t \cdot N < \sqrt{NK}$. This is an integer program in a fixed number of dimensions and can be solved in polynomial time. Faster, easier methods are also known for this special problem.)

(c) **(Optional extra credit)** Instead of condition (1) what if we want to decode under the more natural condition for Hamming metric, that is $|\{i : E(m)_i \neq r_i\}| \le \frac{n-k}{2}$? Using ideas similar to GMD decoding, show how this can be done by calling the above decoder many times, by erasing the last i symbols for each choice of $1 \le i \le n$.

3. In this exercise, we will study random binary linear codes with sparse parity check matrices. Let r, n be integers where we think of r as a constant and n as growing. We will assume for simplicity that r divides n. Consider an $n/r \times n$ matrix A where the i'th row has 1's in the i'th block of r columns, i.e., in the columns $(i-1)r+1, (i-1)r+2, \cdots, ir$, and 0's elsewhere.

For a permutation σ on $\{1, 2, \ldots, n\}$, we define A^σ to be the matrix obtained by permuting the columns according to σ.

For an integer parameter $s < r$, pick s independent random permutations $\sigma_1, \sigma_2, \ldots, \sigma_s$ and define $H_i = A^{\sigma_i}$. Let H be the $sn/r \times n$ matrix obtained by "stacking" the H_i's on top of each other.

Note that H has exactly r 1's in each row, and s 1's in each column, and is thus sparse. Let C be the binary linear code with parity check matrix H. This is called a random low-density parity-check (LDPC) code.

(a) Argue that the rate of C is at least $R := 1 - s/r$.

(b) Let $\lambda \in (0, 1/2)$. Fix a vector $z \in \mathbb{F}_2^n$ of weight λn. Prove that, for every i, $1 \le i \le n/r$, the probability that $H_i z = 0$ is at most

$$O_\lambda(\sqrt{n}) \cdot \left(\frac{1 + (1-2\lambda)^r}{2}\right)^{n/r}.$$

(Here $O_\lambda(\cdot)$ hides a constant factor depending on λ.)

Hint: For a fixed z, the probability that $H_i z = 0$ is the same as the probability that a random x of weight λn satisfied $Ax = 0$ (Why?). Now compute the latter probability. It might be convenient to pass to sampling the coordinates of x i.i.d from Bernoulli(λ).

(c) Prove that the probability that C contains some codeword of weight *exactly* λn is at most

$$2^{h(\lambda)n} n^{O(s)} \left(\frac{1 + (1-2\lambda)^r}{2}\right)^{sn/r}.$$

(d) Prove that for every $\gamma > 0$, there exists $r = r(\gamma)$ such that with high probability C has no codewords of weight exactly $h^{-1}(1 - R - \gamma)n$.

(e) Does the above show that C approaches the Gilbert-Varshamov bound with high probability?

4. Consider the Tanner code $X(H, C_0)$ considered in lecture, where $H = (L, R, E)$ is a r-regular $n \times n$ ϵ-pseudorandom bipartite graph and $C_0 \in \mathbb{F}_2^r$ is a linear code of distance $\delta_0 r$. We proved that $X(H, C_0)$ has relative distance at least $\delta_0(\delta_0 - \epsilon)$, and gave an iterative algorithm to correct a fraction

$\approx \frac{\delta_0}{4}(\delta_0 - 2\epsilon)$ of errors. The goal of this exercise is to improve the number of corrected errors to almost half the (designed) distance.

Consider the following modification to the iterative algorithm based on alternate rounds of left and right side decoding discussed in lecture. For each threshold $t \in \{0, 1, 2, \ldots, \delta_0 d/2\}$, run the following decoding algorithm on a received word $y \in \{0,1\}^E$: for any node $u \in L$ such that $y|_{E(u)}$ is not within distance t of some codeword of C_0, declare erasures on all edges in $E(u)$, and decode the remaining left vertices $u \in L$ to the closest codeword in C_0. This gives us a string $z \in \{0, 1, ?\}^E$. Then on the right side, run an errors-and-erasure decoder at each right vertex $v \in R$, replacing $z|_{E(v)}$ with the codeword in C_0 with smallest Hamming distance on the unerased positions (breaking ties arbitrarily). This gives a string $w \in \{0,1\}^E$. Now run the iterative decoding algorithm we discussed in lecture on the string w for $c \log n$ rounds for a suitable constant c.

Prove that, for any desired $\gamma > 0$, the above algorithm, for a large enough choice of the constant $c = c(\gamma)$, corrects up to a fraction $(1-\gamma)\frac{\delta_0}{2}(\delta_0 - 3\epsilon)$ of errors.

5. In this exercise, we will use expanders to construct codes that almost match the Singleton bound, using a refinement of the ABNNR construction presented in class. We will combine the following ingredients:

   - Let $E_0 : \mathbb{F}_q^k \to (\mathbb{F}_q^b)^n$ be (the encoding function of) a code C_0 of rate R_0 and relative distance δ_0.
   - Let $G = (L, R, E)$ be a r-regular $n \times n$ ϵ-pseudorandom bipartite graph, namely for every $S \subseteq L$ and $T \subseteq R$, we have $\left|E(S,T)| - \frac{r|S||T|}{n}\right| \leq \epsilon r\sqrt{|S||T|}$ (recall we can have such a graph with degree $r \leq O(1/\epsilon^2)$).
   - Let $E_1 : \mathbb{F}_q^b \to \mathbb{F}_q^r$ be the encoding map of a code C_1 of rate R_1 and relative distance δ_1. We denote by $\widetilde{E}_1 : (\mathbb{F}_q^b)^n \to (\mathbb{F}_q^r)^n$ the coordinate-wise application of E_1.

   We use G in a natural way to define a bijection $\varphi_G : (\mathbb{F}_q^r)^n \to (\mathbb{F}_q^r)^n$ as follows (we identify L, R with $\{1, 2, \ldots, n\}$ under some fixed labeling): for any $j \in \{1, 2, \ldots, r\}$ and $i \in \{1, 2, \ldots, n\}$, in the i'th block of the input, the j'th symbol of the block is "sent" to to the j'th neighbor of the left vertex i (for some fixed numbering of the neighbors of i by $\{1, 2, \ldots, r\}$). Now, each right vertex $\ell \in R$ "collects" r elements of $\mathbb{F}_q$ from its neighbors on the left. These are bundled into an element of $\mathbb{F}_q^r$, which then forms the ℓ'th component of the output of φ_G.

   We are now ready to define our final code C^* given by the encoding $E : \mathbb{F}_q^n \to (\mathbb{F}_q^r)^n$ defined as

   $$E := \varphi_G \circ \widetilde{E}_1 \circ E_0 \, .$$

   Now to your problems.

   (a) What is the rate of C^*?

   (b) Prove that the relative distance of C^* is at least $(\delta_1 - \epsilon/\sqrt{\delta_0})$.
   Hint: Apply ϵ-pseudorandomness to S being indices of nonzero blocks after encoding by $\widetilde{E}_1 \circ E_0$ and T being nonzero blocks after applying φ_G.

   (c) Show that for any desired rate $R \in (0,1)$ and $\gamma > 0$, there are suitable parameter choices for C_0, C_1 so that the constructed code C^* has rate R, relative distance at least $1 - R - \gamma$, and alphabet size bounded by a constant depending only on γ.