Carnegie Mellon University Instructor: Venkatesan Guruswami
15-859Y: Coding Theory Term: Fall 2014

PROBLEM SET 2
Due by lecture Wednesday, October 22

INSTRUCTIONS

- You are allowed to collaborate with up to two other 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 (see the last bullet item for some "credit" for doing so).
- Solutions typeset in LaTeX are encouraged. If this is not possible, please write legibly.
- You should not search for solutions on the web. More generally, try and solve the problems without consulting any reference material other than the course notes and what we cover in class. However, if needed you may use references to brush up on the underlying math skills needed to solve some of the problems, such as linear algebra, matrix theory, number theory, finite fields, etc.
- Please start work on the problem set early. The problem set has **seven** problems for a total of 120 points, but we will consider your score to be out of 100 (treating any excess score above 100 as bonus points). Thus, if you prefer, you can just pick a subset of about 100 points to attempt.

---

1. (15 points) Consider the binary expander code based on an unbalanced bipartite $(n, m, D, \gamma, D(1-\epsilon))$-expander as defined in lecture (i.e., the code whose parity check matrix is the bipartite adjacency matrix of the expander) for some $\epsilon < 1/20$. Recall that in an $(n, m, D, \gamma, D(1-\epsilon))$-expander, every subset S of up to γn nodes on the left has at least $D(1-\epsilon)|S|$ neighbors on the right. In this exercise you are asked to analyze the following parallel iterative decoder.

   > For $c \log n$ rounds (for a constant c chosen large enough), do the following in parallel for each variable node: If the variable is in at least $2D/3$ unsatisfied checks, flip its value.

   Prove that the above algorithm corrects any pattern of $\gamma(1-3\epsilon)n$ errors.

2. (20 points) We mentioned the MRRW bound that the rate of binary codes of relative distance $1/2-\epsilon$ is at most $O(\epsilon^2 \log(1/\epsilon))$. In this problem, you will prove this for the special case of ϵ-biased codes, which are binary codes in which every pair of distinct codewords have relative Hamming distance in the range $[1/2-\epsilon, 1/2+\epsilon]$ (note such codes have relative distance at least $1/2-\epsilon$, but in addition no two codewords differ in more than $(1/2+\epsilon)$ fraction of positions).

   Consider the following claim:

   Claim: *Let A be an $m \times m$ real symmetric matrix with 1's on the diagonal, and all off-diagonal elements at most ϵ in absolute value. Assume $\epsilon \in (0, 1/4)$ and m is sufficiently large. Then,* $\mathsf{rank}(A) \geq \Omega\left(\frac{\log m}{\epsilon^2 \log(1/\epsilon)}\right)$.

(a) Show why the claim implies that an ϵ-biased code of block length n can have rate at most $O(\epsilon^2 \log(1/\epsilon))$.

(b) Towards proving the above claim, first prove the following fact when the off-diagonal entries are really small:

> Let B be a real symmetric $m \times m$ matrix with 1's on the diagonal. If all off-diagonal entries of B are at most $1/\sqrt{m}$ in absolute value, then $\mathsf{rank}(B) \geq m/2$.

Suggestion: Use the fact that $\mathsf{Trace}(B) = \lambda_1 + \cdots + \lambda_m$ if λ_i are the eigenvalues of B, and relate $\mathsf{Trace}(B^2)$ and $\mathsf{Trace}(B)$ via Cauchy-Schwarz.

(c) Prove the claim above using the fact from Part (b).

Hint: Given A, consider the matrix B who entries are t'th powers of the entries of A for some large t, and argue that $\mathsf{rank}(B)$ is at most $\binom{\mathsf{rank}(A)+t}{t}$.

3. (20 points) For integers $1 \leq k \leq n$, call a (multi)set $S \subseteq \{0,1\}^n$ to be k-wise independent if for every $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ and $(a_1, a_2, \ldots, a_k) \in \{0,1\}^k$

$$\mathrm{Prob}_{x \in S}[x_{i_1} = a_1 \wedge x_{i_2} = a_2 \wedge \cdots \wedge x_{i_k} = a_k] = \frac{1}{2^k}$$

where the probability is over an element x chosen uniformly at random from S. Small sample spaces of k-wise independent sets are of fundamental importance in derandomization. In this problem, you will see how codes can be used to construct k-wise independent sets of near-optimal size.

(a) Using BCH codes and Problem 5 of Problem set 1, show how one can construct a $2t$-wise independent subset of $\{0,1\}^n$ of size at most $(n+1)^t$ when n is of the form $2^m - 1$, and a $(2t+1)$-wise independent subset of $\{0,1\}^n$ of size at most $2n^t$ when n is a power of 2.

(b) Prove an almost matching lower bound, namely any k-wise independent set $S \subseteq \{0,1\}^n$ satisfies

$$|S| \geq \sum_{i=0}^{\lfloor \frac{k}{2} \rfloor} \binom{n}{i} . \tag{1}$$

Suggestion: Find a set of linearly independent vectors in $\mathbb{R}^{|S|}$ of cardinality at least the R.H.S of (1). Specifically, for $T \subseteq \{1, 2, \ldots, n\}$ of size $\leq \lfloor k/2 \rfloor$, consider the $\langle \chi_T(x) \rangle_{x \in S}$ where $\chi_T(x) = (-1)^{\sum_{i \in T} x_i}$.

4. (20 points) For the noise model where one bit of the codeword gets erased (and we know which location got erased), the parity check code gives a simple solution to recover the missing bit, with just one bit redundancy. Now, consider the harsher model where one bit gets *deleted* and we don't know the position of the missing bit.

(a) Suppose $C \subseteq \{0,1\}^n$ is a binary code capable of recovering from deletion of one bit. Prove that $|C| \leq O(2^n/n)$. Thus about $\log n$ bits of redundancy are needed in such a code.

(b) For integers n, ℓ, $0 \leq \ell \leq n$, consider the code

$$C_\ell = \{(x_1, x_2, \ldots, x_n) \in \{0,1\}^n \mid x_1 + 2x_2 + 3x_3 + \cdots + nx_n \equiv \ell \pmod{(n+1)}\} ,$$

where the sum above is over integers.

Prove that each C_ℓ is capable of correctly recovering a deleted bit in its codewords. Deduce the existence of a code of size $\geq 2^n/(n+1)$ that can correct a single deletion.

(c) **Bonus question:** Can you construct an explicit code of size $2^n/\mathrm{poly}(n)$ than can recover from deletion of *two* bits?

5. (10 points) For this problem, assume the NP-hardness of the following problem (this can be shown via a reduction from Subset Sum):

   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. (Proving this for polynomial sized fields remains an embarrassing open question.)

6. (15 points) Let $\mathbb{F}_q$ be the field with q elements, and let $\alpha \in \mathbb{F}_q$. Prove that the polynomial $X^{q-1} - \alpha$ is irreducible over $\mathbb{F}_q$ **if and only if** α is a primitive element of $\mathbb{F}_q$.

7. (20 points) 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

$$\prod_{i : E(m)_i \neq r_i} p_i^{b_i} \leq \sqrt{N/K}\,. \tag{2}$$

   (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 (2) exists, prove that there exist integers y, z with $0 \leq y < \sqrt{NK}$ and $1 \leq z \leq \sqrt{N/K}$ 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: $0 < z \leq \sqrt{N/K}$; and $0 \leq z \cdot r - t \cdot N < \sqrt{NK}$. 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) Instead of condition (2) what if we want to decode under the more natural condition for Hamming metric, that is $|\{i : E(m)_i \neq r_i\}| \leq \frac{n-k}{2}$? Show how this can be done by calling the above decoder many times, by erasing the last i symbols for each choice of $1 \leq i \leq n$.