15-859V: Introduction to coding theory — Spring 2010
Carnegie Mellon University — Venkatesan Guruswami

PROBLEM SET 3
Due by Friday, April 23

INSTRUCTIONS

- This problem sets can be turned in groups of two people; i.e., a single write-up for each two person team suffices. If you prefer, you can also work alone (see the last bullet item for some "credit" for doing so). Solutions typeset in LaTeX are preferred.
- You are *strongly urged* to try and solve the problems without consulting *any* reference material other than the course notes and what we cover in class. If for some reason you feel the need to consult some source (such as a textbook, research paper, or sources on the web), *you must acknowledge the source* and also articulate the difficulty you couldn't overcome before consulting the source and how it helped you overcome that difficulty. Alternatively, before consulting any such material, I encourage you to ask me for a hint, preferably by posting a comment on the blog post dedicated to this problem set, so all students can take advantage of any hints.
- Please use the comments section of the blog for any questions or clarifications about the problems.
- Please start work on the problem set early. The problem set has **seven** problems and is worth a total of 100 points. As a rough estimate, a score of 70 (or 60 if you work and turn in solutions solo), suffices for an A on this problem set.

---

1. (12 points) In this exercise, you are asked to show that good codes can be used to construct expanders of logarithmic degree.

   Let $G \in \mathbb{F}_2^{n \times k}$ be the generator matrix of an $[n, k]_2$ binary linear code with the property that every nonzero codeword of C has Hamming weight in the range $[(1/2 - \epsilon)n, (1/2 + \epsilon)n]$ (such a code is usually called ϵ-biased). Define a graph $H = (\mathbb{F}_2^k, E)$ where $(x, y) \in E$ if $x + y$ equals one of the n rows of G (if some row occurs multiple times, place an edge of appropriate multiplicity between x, y).

   Prove that the second largest eigenvalue of the adjacency matrix of H in absolute value is at most 2ϵ. Use this to deduce the existence of spectral expanders (with second largest eigenvalue in absolute value bounded away from the degree) of logarithmic degree.

2. (10 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.

3. (15 points) Consider the Tanner code $T(H, C_0)$ considered in lecture, where $H = (L, R, E)$ is a d-regular $n \times n$ bipartite expander with second largest eigenvalue λ, and $C_0 \in \mathbb{F}_2^d$ is a linear code of distance $\delta_0 d$. We proved that $T(H, C_0)$ has relative distance at least $\delta_0(\delta_0 - \lambda/d)$, and gave an iterative algorithm to correct a fraction $(1-\epsilon)\frac{\delta_0}{4}(\delta_0 - \lambda/d)$ errors. The goal of this exercise is to improve the number of corrected errors to 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 algorithm. In the first round of left side decoding with received word y, 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 rest to the closest codeword in C_0. This gives a string $z \in \{0, 1, ?\}^E$. Then on the right side, run an errors-an-erasure decoder at each node $v \in R$ replacing $z_{|E(v)}$ with the codeword 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 $\epsilon > 0$, the above algorithm, for a large enough choice of the constant c, corrects up to a fraction $(1-\epsilon)\frac{\delta_0}{2}\left(\delta_0 - \frac{2\lambda}{d}\right)$ of errors.

4. (10 points) In this exercise, the goal is to show that a variant of the expander-based Tanner codes $T(H, C_0)$ can be be used to achieve the capacity on the binary symmetric channel.

   For a $n \times n$ d-regular expander $H = (L, R, E)$ as above, and binary linear codes C_0, C_1 of block length d, define the code $T(H, C_0, C_1)$ to consist of those strings $z \in \{0,1\}^E$ such that $z_{|E(u)} \in C_0$ for each $u \in L$ and $z_{|E(v)} \in C_1$ for each $v \in R$.

   Show that for any desired $p \in (0, 1/2)$ and $\epsilon > 0$, with suitable choice of d and the codes C_0, C_1, one can construct a family of codes $T(H, C_0, C_1)$ with rate $1 - h(p) - \epsilon$ together with an $O(N \log N)$ time decoding algorithm for communicating on BSC_p with error probability at most $2^{-c_{p,\epsilon} N}$ for some constant $c_{p,\epsilon} > 0$.

5. (10 points)

   (a) (Due to Salil Vadhan) Let $\epsilon > 0$ be a sufficiently small real. Suppose that $C \subseteq \{0,1\}^n$ is a code of relative distance at least $1/3$ and rate at most $a\epsilon^2$ for some $a > 0$. Suppose a codeword $c \in C$ is transmitted on BSC_p for $p = 1/2 - \epsilon$, and we receive $r \in \{0,1\}^n$. Prove that if a is a small enough constant (independent of n, ϵ), then with all but exponentially small probability over the errors, c will be the unique codeword within Hamming distance $(1-\epsilon)n/2$ from r.

   (b) **Extra credit question:** *(For your fun only; No need to turn anything in, and I think the question might still be open.)* Can one deduce the same conclusion without assuming the upper bound on rate, but instead based on the hypothesis C is list-decodable up to a fraction $(1/2 - \epsilon/3)$ of errors (with lists of size $\mathrm{poly}(1/\epsilon)$, say), and has relative distance at least $(1/2 - \epsilon/3)$?

6. ($4 + 8 + 8 = 20$ points) In this problem, we will consider the number-theoretic counterpart of Reed-Solomon codes. Let $1 \leqslant k < n$ 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 $1 \leqslant i \leqslant n$, 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

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

(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 $0 \leqslant y < \sqrt{NK}$ and $1 \leqslant z \leqslant \sqrt{N/K}$ 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: $0 < z \leqslant \sqrt{N/K}$; and $0 \leqslant z \cdot r - t \cdot N < \sqrt{NK}$. 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 (1) what if we want to decode under the more natural condition for Hamming metric, that is $|\{i : E(m)_i \neq r_i\}| \leqslant \frac{n-k}{2}$? 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 $1 \leqslant i \leqslant n$.

7. ($6+5+4+3+3+2 = 23$ points) We have mentioned objects called algebraic-geometric codes, that generalize Reed-Solomon codes and have some amazing properties, a couple of times in the course. The objective of this exercise is to construct one such AG code, and establish its rate vs distance trade-off.

Let p be a prime and $q = p^2$. Consider the equation

$$Y^p + Y = X^{p+1} \tag{2}$$

over $\mathbb{F}_q$.

(a) Prove that there are exactly p^3 solutions in $\mathbb{F}_q \times \mathbb{F}_q$ to (2). That is, if $S \subseteq \mathbb{F}_q^2$ is defined as

$$S = \{(\alpha, \beta) \in \mathbb{F}_q^2 \mid \beta^p + \beta = \alpha^{p+1}\}$$

then $|S| = p^3$.

(b) Prove that the polynomial $p(X, Y) = Y^p + Y - X^{p+1}$ is irreducible over $\mathbb{F}_q$.
(**Suggestion**: One approach is to use the Eisenstein criterion, considering p as a polynomial in X over $\mathbb{F}_q[Y]$.)

(c) Let $n = p^3$. Consider the evaluation map $\mathrm{ev} : \mathbb{F}_q[X, Y] \to \mathbb{F}_q^n$ defined by

$$\mathrm{ev}(f) = (f(\alpha, \beta) : (\alpha, \beta) \in S) \ .$$

Argue that if $f \neq 0$ and is not divisible by $Y^p + Y - X^{p+1}$, then $\mathrm{ev}(f)$ has Hamming weight at least $n - \deg(f)(p+1)$, where $\deg(f)$ denotes the *total* degree of f.
(**Hint**: You are allowed to make use of ***Bézout's theorem***, which states that if $f, g \in \mathbb{F}_q[X, Y]$ are nonzero polynomials *with no common factors*, then they have at most $\deg(f)\deg(g)$ common zeroes.)

(d) For an integer parameter $\ell \geqslant 1$, consider the set $\mathcal{F}_\ell$ of bivariate polynomials

$$\mathcal{F}_\ell = \{f \in \mathbb{F}_q[X, Y] \mid \deg(f) \leqslant \ell, \deg_X(f) \leqslant p\}$$

where $\deg_X(f)$ denotes the degree of f in X.

Argue that $\mathcal{F}_\ell$ is an $\mathbb{F}_q$-linear space of dimension $(\ell+1)(p+1) - \frac{p(p+1)}{2}$.

(e) Consider the code $C \subseteq \mathbb{F}_q^n$ for $n = p^3$ defined by

$$C = \{\mathrm{ev}(f) \mid f \in \mathcal{F}_\ell\} \ .$$

Prove that C is a linear code with minimum distance at least $n - \ell(p+1)$.

(f) Deduce a construction of an $[n, k]_q$ code with distance $d \geqslant n - k + 1 - p(p-1)/2$.
(**Remark:** Reed-Solomon codes have $d = n - k + 1$, whereas these codes are off by $p(p-1)/2$ from the Singleton bound. However they are much longer than RS codes, with a block length of $n = q^{3/2}$, and the deficiency from the Singleton bound is only $o(n)$.)