Problem Set 3

- Due Date: **9 Dec, 2022**
- Turn in your problem sets electronically (LaTeX, pdf or text file) by email. If you submit handwritten solutions, start each problem on a fresh page.
- Collaboration is encouraged, but all writeups must be done individually and must include names of all collaborators.
- Refering sources other than the text book and class notes is strongly discouraged. But if you do use an external source (eg., other text books, lecture notes, or any material available online), ACKNOWLEDGE all your sources (including collaborators) in your writeup. This will not affect your grades. However, not acknowledging will be treated as a serious case of academic dishonesty.
- The points for each problem are indicated on the side. The total for this set is 80.
- Be clear in your writing.
- Problems 3 and 7 are due to Venkat Guruswami while Problem 5 is due to Mrinal Kumar.

1. **[Kakeya Sets]** (4+2+2+2)

   Let $\mathbb{F}$ be a finite field of size q. A *Kakeya set* in $\mathbb{F}^m$ is a set $K \subseteq \mathbb{F}^n$ such that K contains a line in every direction. More precisely, K is a Kakeya set if for every $y \in \mathbb{F}^m$ there exists a $z \in \mathbb{F}^m$ such that the line

   $$L_{z,y} = \{z + t \cdot y | t \in \mathbb{F}\}$$

   is contained in K.

   A trivial upper bound on th size of K is q^m and this can be improved to $q^m/2^{m-1}$. In this problem, we will use the polynomial method to show a lower bound of $q^m/m!$. More precisely, we will show that

   $$K \geq \binom{q+m-1}{m}.$$

   Suppose, for contradiction that this is not the case.

   (a) Show that there exists a m-variate non-zero polynomial g of degree $d \leq q-1$ such that $g(x) = 0$ for all $x \in K$.

   Let g_d be the homogenous part of degree d of g so that g_d is non-zero and homogenous.

   For any $y \in \mathbb{F}^m$, we know that there exists a $z \in \mathbb{F}^m$ such that the line $L_{z,y}$ is contained in K. Consider the following univariate polynomial

   $$P_{y,z}(t) := g(z + t \cdot y).$$

   (b) Argue that $P_{y,z}$ is identically zero and hence the coefficient of t^d in $P_{y,z}(t)$ is zero.

   (c) Show that the coefficient of t^d in $P_{y,z}(t)$ is exactly $g_d(y)$.

   (d) Conclude that g_d is identically zero, a contradiction.

2. **[Linear LRCs]** (15)

(a) First, we will prove a general structural result about linear codes. Let $\mathcal{C} \subseteq \mathbb{F}_q^n$ be a linear code. Let $i \in [n]$. Prove that one of the following has to hold:

1. There exists $v \in \mathcal{C}^\perp$ such that $i \in supp(v)$, ordering
2. $$\mathcal{C} = \{(c_1, c_2, \ldots, c_{i-1}, \alpha, c_{i+1}, \ldots, c_n) : \alpha \in \mathbb{F}_q, (c_1, c_2, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n) \in \mathcal{C}|_{[n]\setminus i}\}$$

(b) Let $\mathcal{C}$ be an (r, d) LRC that is an $[n, k, d]_q$ code. Then, argue that for every $i \in [n]$, there exists a dual codeword $v \in \mathcal{C}^\perp$ with $i \in supp(v)$ with $supp(v) \subseteq R_i \cup \{i\}$. (Recall that R_i is the set of at most r values that c_i can be recovered from)

(c) Using the previous part or otherwise, argue that any c_i for any codeword $(c_1, \ldots, c_n)$ can be recovered as a linear combination of values in c_{R_i}

3. **[20 Questions]** (15)

In the game of 20 questions, an oracle has an arbitrary secret $s \in \{0,1\}^n$ and the aim is to determine the secret by asking the oracle as few yes/no questions about s as possible. It is easy to see that n questions are necessary and sufficient. Here we consider the variant where the oracle has two secrets s_1 and s_2 in $\{0,1\}^n$ and can adversarially decide to answer each question according to either s_1 or s_2. That is, for a question $f : \{0,1\}^n \to \{0,1\}$, the oracle may answer with either $f(s_1)$ or $f(s_2)$. Here it turns out to be impossible to pin down either of the secrets with certainty, no matter how many questions we ask, but we can hope to compute a small list L of secrets such that $|L \cap \{s_1, s_2\}| \neq \emptyset$. (In fact, $|L|$ can be made as small as 2.) This variant of twenty questions was apparently motivated by questions about internet traffic routing.

(a) Let $\mathrm{Enc} : \{0,1\}^n \to \{0,1\}^{\hat{n}}$ be a code such that every two codewords in Enc *agree* in at least a $1/2 - \varepsilon$ fraction of position, and that Enc has a polynomial-time $(1/4 + \varepsilon, \ell)$- list decoding algorithm. Show how to solve the above problem in polynomial time by asking the $\hat{n}$ questions $\{f_i\}$ defined by $f_i(x) = \mathrm{Enc}(x)$.

(b) Recall (???) that if a q-ary code $\mathcal{C} \subseteq \Sigma^n$ of rate ρ is (δ, L)-list decodable, then $\rho \leq 1 - H_q(\delta, n) + log_q(L)/n$. Taking Enc to be such a code, deduce that $\hat{n} = \mathrm{poly}(n)$ questions suffices.

4. **[List decodable Codes vs Extractors]** (15)

Given a code $\mathcal{C}\colon \{0,1\}^n \to \Sigma^D$ where $|\Sigma| = q$, we set $N := 2^n, M = q \cdot D$ and associate the elements of $\{0,1\}^n$ with that of $[N]$ and the elements of $[D] \times \Sigma$ with that of $[M]$. Define the following function: $\Gamma\colon [N] \times [D] \to [M]$ as follows:

$$\Gamma(x, i) = (i, \mathcal{C}(x)_i).$$

We think of the function Γ as specifying a D-left-regular bipartite graph $([N], [M], E)$ where the i^{th} neighbour of $x \in [N]$ is given by $\Gamma(x, i)$.

For any set $T \subset [M]$ and $\varepsilon \in [0, 1)$, define

$$\mathrm{List}_\Gamma(T, \varepsilon) := \left\{ x \in [N] \colon \Pr_{i \in [D]} [\Gamma(x, i) \in T] > \varepsilon \right\}.$$

(a) Prove that $\mathcal{C}$ is $(1 - 1/q - \varepsilon, L)$-list-decodable iff for every $r \in \Sigma^D$, we have

$$|\mathrm{List}_\Gamma(T_r, 1/q + \varepsilon)| \leq L, \tag{1}$$

where $T_r := \{(i, r_i) \colon i \in [D]\}$.

(b) A bipartite graph $([N], [M], E)$ is said the be (k, ε)-extractor if for every set $X \subseteq [N]$ of size at least 2^k, we have that the distribution $\Gamma(U_X, U_{[D]})$ obtained on $[M]$ by picking a uniformly random element x of $[N]$ and independently a uniformly random element i of $[D]$ and outputting $\Gamma(x, i)$ is ε-close to the uniform distribution[1].

Suppose the code $\mathcal{C}$ satisfied that for all $T \subset [M]$, we have

$$|\mathrm{List}_\Gamma(T, \mu(T) + \varepsilon)| \leq L, \tag{2}$$

where $\mu(T) := |T|/M$ ((1) is exactly (2) but restricted to sets T of the form T_r for some $r \in \Sigma^D$. Note $\mu(T_r) = 1/q$). Then show that the corresponding bipartite graph given by Γ is a $(\log_2 L + \log_2(1/\varepsilon), 2\varepsilon)$-extractor.

5. **[Additive-folded Reed-Solomon Codes]** (15)

In this question, we will see a list decoding algorithm for codes which are closely related to Folded Reed-Solomon codes and multiplicity codes. We have parameters n, k, s and we work over a field $\mathbb{F}$ such that $|\mathbb{F}|$ is a prime number larger than sn. The message space is again the space of univariate polynomials of degree less than k over $\mathbb{F}$. The encoding of a polynomial $f \in \mathbb{F}[x]$ is given by the function $\mathsf{Enc} : \mathbb{F}[x] \to (\mathbb{F}^s)^n$, defined as follows:

$$\mathsf{Enc}(f) = (f(s \cdot i), f(s \cdot i + 1), \ldots, f(s \cdot i + s - 1))_{i=1}^n \,.$$

In other words, the encoding outputs an n length vector where each coordinate is an s-tuple of rational numbers, and the i^{th} coordinate contains the evaluation of f and on inputs $si, si + 1, \ldots, si + (s - 1)$.

We will now see a version of list decoding for these codes very closely related to the algorithm that we saw for multiplicity codes. The main difference will be that we choose the polynomial Q such that the degree $\deg_{Y_i}(Q) \leq 1$ (this is the key difference between the Guruswami-Wang and Kopparty algorithms for list-decoding).

As an input, we have a received word $\mathbf{b} \in (\mathbb{F}^s)^n$, where for every $i \in \{1, 2, \ldots, n\}$ the ith coordinate of $\mathbf{b}$ is denoted by $b_i = (b_{i,0}, \ldots, b_{i,s-1})$. Ideally, we would like to recover all polynomials $f \in \mathbb{F}_{<k}[x]$ such that $\mathsf{Enc}(f)$ and $\mathbf{b}$ have large agreement. But here, we will just output a linear space of small dimension containing all such polynomials f.

(a) What is the minimum distance of this code, as a function of n, k, s, i.e.

$$\min_{f,g \in \mathbb{F}[x], \deg(f), \deg(g) < k, f \neq g} \Delta(\mathsf{Enc}(f), \mathsf{Enc}(g)) \,.$$

?

(b) Let $m < s$ be a parameter. As a first step of the decoding algorithm, show that there is a non-zero polynomial $Q(x, y_0, y_1, \ldots, y_{m-1})$ of the form $Q := Q_0(x)y_0 + \ldots + Q_{m-1}(x)y_{m-1}$ such that the following conditions hold.

- Degree of Q is at most $D + 1$, where $D = \frac{n(s-m+1)}{m}$
- For every $i \in \{1, 2, \ldots, n\}$, $j \in \{0, 1, \ldots, s - m\}$,

$$Q_0(si + j)b_{i,j} + \ldots + Q_{m-1}(si + j)b_{i,m-1+j} = 0 \,.$$

[1] Two distributions P and Q on the set $[N]$ are said to be ε-close if for all $T \subset [M]$, we have

$$|\Pr_{m \sim P}[m \in T] - \Pr_{m \sim Q}[m \in T]| \leq \varepsilon.$$

(c) If f is a polynomial in $\mathbb{F}[x]$ of degree less than k such that there exists an $i \in \{1, 2, \ldots, n\}$, with

$$b_i = (f(s \cdot i), f(s \cdot i + 1), \ldots, f(s \cdot i + s - 1))$$

then, conclude that the univariate polynomial

$$R(x) := Q\left(x, f(x), f(x+1), \ldots, f(x+m-1)\right)$$

has at least $s - m + 1$ distinct roots in $\mathbb{F}$.

(d) Conclude that if the number of coordinates i, where $\mathbf{b}$ and $\mathsf{Enc}(f)$ agree is at least $\frac{D+k-1}{s-m}$, then $R(x)$ must be identically zero.

All that remains now for algorithmic list decoding of these codes is to be able to extract all polynomials f of degree less than k from the polynomial Q satisfying

$$Q_0(x)f(x) + Q_1(x)f(x+1) + \cdots + Q_{m-1}f(x+m-1) = 0.$$

Observe first that the set of solutions forms a linear space.

(e) Argue that the dimension of the space of solutions, i.e., polynomials $f(x)$ of degree less than k such that the polynomial

$$R(x) = Q\left(x, f(x), f(x+1), \ldots, f(x+m-1)\right)$$

is identically zero can be upper bounded by $m-1$. (Hint: an appropriate change of basis might help in the dimension counting).s

6. **[Exponential lower bounds for 2-query linear LDCs]** (3+4+3+5)

In this problem, we will prove an exponential lower bound for 2-query linear locally decodable codes.

Recall that a code $\mathcal{C} : \{0,1\}^k \to \{0,1\}^n$ is said to be (q, δ, ε)-locally decodable if there exists a (probabilistic) decoder D such that on oracle access to any $\mathbf{y} \in \{0,1\}^n$ that satisfies $\Delta(\mathbf{y}, \mathcal{C}(\mathbf{x})) \leq \delta n$, we have

- $\forall i \in [k], \Pr\left[D^{\mathbf{y}}(i) = \mathbf{x}_i\right] \geq \frac{1}{2} + \varepsilon$.
- D makes at most q probes into $\mathbf{y}$ on any input i and internal random coins.

For fixed $c \in \mathbb{R}$, $\varepsilon \in (0,1)$ and integer 2, we say that $\mathcal{C} : \{0,1\}^k \to \{0,1\}^n$ is a (q, c, ε)-smooth code if there exists a probabilistic oracle machine A such that:

- In every invocation, A makes at most q queries non-adaptively.
- For every $\mathbf{x} \in \{0,1\}^k$ and and for every $i \in [k]$, we have

$$\Pr[A^{\mathcal{C}(\mathbf{x})}(i) = \mathbf{x}_i] \geq \frac{1}{2} + \varepsilon.$$

- For every $i \in [k]$ and $j \in [n]$, the probability that on input i the oracle machine A queries index j is at most $\frac{c}{m}$.

(a) Show that if $\mathcal{C} : \{0,1\}^k \to \{0,1\}^n$ is a (q, δ, ε)-locally decodable code, then $\mathcal{C}$ is also a $(q, q/\delta, \varepsilon)$-smooth code.

Let $\mathcal{C} : \{0,1\}^k \to \{0,1\}^n$ be a linear code. Since $\mathcal{C}$ is linear, we might wlog. assume that there exist $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \{0,1\}^k$, such that for all $\mathbf{x} \in \{0,1\}^k$ and $j \in [n]$, we have

$\mathcal{C}(\mathbf{x})_j = \langle \mathbf{a}_j, \mathbf{x} \rangle$. For simplicity, let us assume that all the $\mathbf{a}_i$'s are distinct. Suppose $\mathcal{C}$ is a $(2, \delta, \varepsilon)$-locally decodable for some $\delta, \varepsilon \in (0, 1)$. Let us further make a simplifying assumption that the local D (corresponding to $\mathcal{C}$) makes exactly 2 probes every time and uses both the probes. It follows from 6a that $\mathcal{C}$ is $(2, 2/\delta, \varepsilon)$-smooth.

Construct recovery graphs $\{G_i = ([n], E_i)\}_{i=1}^k$ based on the smooth decoder A for $\mathcal{C}$ as follows: the vertices of all the k graphs G_i's are $[n]$. Two vertices $j, j' \in [n]$ are connected in G_i if

$$\Pr[A^{\mathcal{C}(\mathbf{x})}(i) = \mathbf{x}_i | A \text{ queries } \mathcal{C}(\mathbf{x}) \text{ at indices } j \text{ and } j'] > \frac{1}{2}.$$

(b) If G is $(2, c, \varepsilon)$-smooth, show that for each $i \in [k]$, the graph G_i has a matching $M_i \subseteq E_i$ of size at least $\varepsilon n/c$.

(c) Argue that for each $i \in [k]$, if $(j, j') \in E_i$ then $\mathbf{e}_i \in \text{span}\{\mathbf{a}_j, \mathbf{a}_{j'}\}$. It then follows from our assumption ("the local D makes exactly 2 probes every time and uses both the probes") that $\mathbf{a}_j + \mathbf{a}_{j'} = \mathbf{e}_i$.

[For extra credit, do not make this simplifying assumption and modify the following part suitably to still yield an exponential lower bound.]

We can thus identify the vertices $[n]$ with the set $A = \{\mathbf{a}_j | j \in [m]\}$, a subset of the vertices of the hypercube $\{0, 1\}^k$ and the edges (j, j') with the corresponding edges in the hypercube. Consider the graph $G = ([n], E_1 \cup \cdots \cup E_k)$. From the above identification, we get that G is a subgraph of the hypercube. Furthermore, from 6c, we get that the k edge-sets E_i are all distinct. Hence, from 6b, we have $|E(A, A)| \geq \sum_{i=1}^k |E_i| \geq k \cdot (\varepsilon n/c) = \varepsilon \delta k n/2$. Here, $E(A, A)$ refers to the edges in G both of whose endpoints in is A.

(d) Since G is a subgraph of the hypercube, use the upper bound on $E(A, A)$ to conclude that $n \geq 2^{\varepsilon \delta k}$.

This proves an exponential lower bound on the size of any 2-query linear LDC (provided all the codeword bits are distinct, ie. $\mathbf{a}'_j s$ are distinct). For extra credit, see if you can remove this assumption of distinctness.

7. **[Not for submission][Algrebraic-Geometric codes]** (15)

We have mentioned objects called algebraic-geometric codes, that generalize Reed-Solomon codes and have some amazing properties, a few times in the course. The objective of this exercise is to construct one such AG code and establish its rate-distance tradeoff.

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

$$Y^p + Y = X^{p+1}$$

over $\mathbb{F}_q$.

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

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

then $|S| = p^3$

(b) Prove that the polynomial $f(X, Y) = Y^p + Y - X^{p+1}$ is irreducible over $\mathbb{F}_q$. **Suggestion:** One approach is to use the Eisenstein criterion (feel free to look this up), considering $f(X, Y)$ to be a polynomial in X with coefficients from $\mathbb{F}_q(Y)$.

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

$$\text{ev}(f(X, Y)) = (f(\alpha, \beta) \ : \ \alpha, \beta \in S)$$

where S is defined as in part (a). Argue that if $f \not\equiv 0$, 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 $\deg(f)$ is the *total* degree of f. **Hint:** You are allowed to use *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 $\ell \geq 1$, consider the set $\mathcal{F}_\ell$ of bivariate polynomials

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

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

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

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 $d \geq n - k + 1 - p(p-1)/2$.

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 $o(n)$.