ASSIGNMENT 4

Exercise 1 (List decodability of linear codes)**.** Show that with high probability, a random (binary) linear code obtained by choosing an $nR \times n$ generator matrix uniformly at random is (p, L)-list decodable as long as

$$R \leq 1 - H(p) - \frac{1}{\lceil \log_2(L+1) \rceil}.$$

Hint: Argue that any set of $L+1$ vectors in $\mathbb{F}_2^k$ contains at least $\lceil \log_2(L+1) \rceil$ linearly independent vectors. If two messages are linearly independent, then what can you say about the corresponding codewords of the random linear code?

Exercise 2 (List decoding from erasures)**.** We say that a code is (p, L)-erasure list-decodable if for any vector $\mathbf{y} \in \{0, 1, *\}^n$ (where $*$ denotes the erasure symbol) with at most pn erasures, there are at most L codewords that agree with $\mathbf{y}$ in the unerased positions. For any vector $\mathbf{c}$ and $T \subset [n]$, let $\mathbf{c}_T$ denote the restriction of $\mathbf{c}$ to T, i.e., it is the $|T|$-length vector $(c_i : i \in T)$. Formally, a code $\mathcal{C} \subset \mathbb{F}_2^n$ is (p, L)-erasure list-decodable if for every $T \subset [n]$ with $|T| \geq (1-p)n$, and $\mathbf{y}' \in \{0,1\}^{|T|}$, we have

$$|\{\mathbf{c} \in \mathcal{C} : \mathbf{c}_T = \mathbf{y}'\}| \leq L.$$

Prove the following:

1. If $\mathcal{C}$ has minimum distance d, then it is $\left(\frac{d-1}{n}, 1\right)$-list decodable.
2. For every $\epsilon > 0$, there exists a (p, L)-erasure list decodable code of rate

$$R \geq \frac{L}{L+1}(1-p) - \frac{1}{L} - \epsilon$$

   Hint: Use random codes. For a fixed $T, \mathbf{y}'$, compute the probability that the codeword for a fixed message is equal to $\mathbf{y}$ when restricted to T. Do this for $L+1$ messages. Then take a union bound over messages, $\mathbf{y}$', and T.

3. For every $\epsilon > 0$, there exists a linear (p, L)-erasure list-decodable code of rate

$$R \geq \frac{J-1}{J}(1-p) - \frac{1}{J-1} - \epsilon$$

   where $J = \lceil \log_2(L+1) \rceil$.

4. Show that if a code of rate $1 - p + \epsilon$ is (p, L)-erasure list-decodable, then $L = 2^{\Omega(n)}$.

Exercise 3 (Binary symmetric channel)**.** Let us examine the performance of linear codes against random errors. The binary symmetric channel with crossover probability $p < 1/2$ is defined by the following process: Given a codeword $\mathbf{c} \in \mathbb{F}_2^n$, we generate a random vector $\mathbf{y}$ where y_i is obtained by flipping c_i with probability p, independently of everything else. Equivalently,

$$\mathbf{y} = \mathbf{c} + \mathbf{z},$$

where $\mathbf{z}$ is a random vector whose components are independent and follow a Bernoulli(p) distribution. Here $\mathbf{y}$ is called the received vector, and $\mathbf{z}$ the noise vector.

We will measure the performance of a code $\mathcal{C} \subset \mathbb{F}_2^n$ of size 2^{nR} using the *average probability of error* under a minimum distance decoder $\mathrm{DEC}(\mathbf{y}) = \arg\min_{\mathbf{c}\in\mathcal{C}} d(\mathbf{y}, \mathbf{c})$:

$$\begin{aligned} P_e(\mathcal{C}) &= \frac{1}{2^{nR}} \sum_{\mathbf{c}\in\mathcal{C}} \Pr_{\mathbf{z}}[\exists \mathbf{c}' \in \mathcal{C}\backslash\{\mathbf{c}\} : \mathrm{DEC}(\mathbf{y}) = \mathbf{c}'] \\ &= \frac{1}{2^{nR}} \sum_{\mathbf{c}\in\mathcal{C}} \Pr_{\mathbf{z}}[\exists \mathbf{c}' \in \mathcal{C}\backslash\{\mathbf{c}\} : d(\mathbf{y}, \mathbf{c}') \le d(\mathbf{y}, \mathbf{c})], \end{aligned}$$

where $d(\cdot, \cdot)$ denotes Hamming distance. This is the average probability that there exists a codeword different from $\mathbf{c}$, that is closer to the received vector.

The goal of this and the next exercise is to show that for every $\epsilon > 0$ there exist linear codes of rate $R = 1 - H(p) - \epsilon$ whose probability of error is $2^{-\Omega(n)}$.

1. First, show that the Hamming distance between $\mathbf{y}$ and $\mathbf{c}$ is approximately np:

   $$\Pr[d(\mathbf{c}, \mathbf{y}) > np(1 + \epsilon/2)] \le 2^{-\Omega(n)}$$

   Hint: Find the probability that $\mathbf{z}$ has Hamming weight greater than $np(1 + \epsilon/2)$. You can use Chernoff bound, or directly compute the probability and then use Stirling's approximation.

2. Next, show that the probability of error can be bounded from above as $P_e(\mathcal{C}) \le P_e^{(1)} + P_e^{(2)}$, where

   $$P_e^{(1)} = \frac{1}{2^{nR}} \sum_{\mathbf{c}\in\mathcal{C}} \Pr_{\mathbf{z}}[\exists \mathbf{c}' \in \mathcal{C}\backslash\{\mathbf{c}\} : d(\mathbf{y}, \mathbf{c}') \le np(1 + \epsilon/2)]$$

   and

   $$P_e^{(2)} = \Pr[d(\mathbf{c}, \mathbf{y}) > np(1 + \epsilon/2)] \le 2^{-\Omega(n)}$$

3. Let us now find the probability of error for a random linear code obtained by choosing a generator matrix G uniformly. Show that for any two nonzero message vectors $\mathbf{u}_1 \neq \mathbf{u}_2$, the corresponding codeword $\mathbf{u}_1 G$ and $\mathbf{u}_2 G$ are statistically independent.

4. For fixed messages $\mathbf{u}_1 \neq \mathbf{u}_2$, show that

   $$\Pr_{G,\mathbf{z}}\Big[d(\mathbf{u}_1 G, \mathbf{u}_2 G + \mathbf{z}) < np(1 + \epsilon/2)\Big] \le 2^{-n(H(p(1+\epsilon/2))+o(1))}$$

   Hint: First compute $\Pr_G\Big[d(\mathbf{u}_1 G, \mathbf{x}) < np(1 + \epsilon/2)\Big]$ for a fixed $\mathbf{x} \in \mathbb{F}_2^n$. Then average over $\mathbf{z}$.

5. Use part 4 to show that if $R < 1 - H(p) - \epsilon$, then $P_e^{(2)} = 2^{-\Omega(n)}$.

6. Combine everything to prove that there exists a linear code with rate $R \ge 1 - H(p) - \epsilon$ and $P_e = o(1)$.

Exercise 4 (Concatenated codes for BSC)**.** We now want to show that a rate of $1 - H(p) - \epsilon$ and probability of error $2^{-\Omega(n)}$ can be achieved with polynomial encoding and decoding complexity.

Show that concatenating an inner random linear code of blocklength $O(\log n)$ and rate close to $1 - H(p)$ with an outer Reed-Solomon code over a field of size $2^{O(\log n)}$ and rate close to 1 helps us achieve our objective.

Hint: Compute the probability that k inner codewords are in error. The outer code is guaranteed to correct a certain fraction of erroneous inner codewords. Choose your parameters so that the number of inner codewords in error is well within the error correction capability of the outer code.