## Homework 3
## Due Monday, April 4, 2011 in class

You can collaborate in groups of up to 3. However, the write-ups must be done individually, that is, your group might have arrived at the solution of a problem together but everyone in the group has to write up the solution in their own words. Further, you **must** state at the beginning of your homework solution the names of your collaborators. Just to be sure that there is no confusion, the group that you pick has to be for all problems [i.e. you cannot pick different groups for different problems :-)]

If you are not typesetting your homework, please make sure that your handwriting is legible. Illegible handwriting will most probably lose you points.

Unless stated otherwise, for all homeworks, you are **only** allowed to use notes from the course: this includes any notes that you might have taken in class or any scribed notes from Fall 07 or Spring 09 version or the current version of the course. Doing otherwise will be considered cheating. Note that if your collaborator cheats and you use his solution, then you have cheated too (ignorance is not a valid excuse).

Please use the comment section of the post on HW 3 on the blog if you have any questions and/or you need any clarification.

You might find Problem 2 in HW 0 useful for this homework. You can use any statement from HW 0 without proof.

In total you can use at most **seven** pages for this homework.

I encourage you to start thinking on the problems **early**.

1. (**Alternate definition of codes**) ($10 + 5 = 15$ points)

   (a) We have defined Reed-Solomon in class. In this problem you will prove that a certain alternate definition also suffices.

   Consider the Reed-Solomon code over a field $\mathbb{F}$ of size $q$ and block length $n = q - 1$ defined as

   $$C_1 = \{(p(1), p(\alpha), \ldots, p(\alpha^{n-1})) \mid p(X) \in \mathbb{F}[X] \text{ has degree } \leq k-1\}$$

   where $\alpha$ is the generator of the multiplicative group $\mathbb{F}^*$ of $\mathbb{F}$.[1] Note that $C_1$ is the

[1] This means that $\mathbb{F}^* = \mathbb{F} \setminus \{0\} = \{1, \alpha, \ldots, \alpha^{q-2}\}$, i.e. for any $\gamma \in \mathbb{F}^*$, $\gamma = \alpha^i$ for some $0 \leq i \leq q-2$. Further, $\alpha^{q-1} = 1$. For example, 2 is a generator for $\mathbb{F}_5$ as $1 = 2^0 \mod 5$, $2 = 2^1 \mod 5$, $3 = 2^3 \mod 5$ and $4 = 2^2 \mod 5$. However, 2 is *not* a generator for $\mathbb{F}_7$ as e.g. there is no $0 \leq i \leq 5$, such that $3 = 2^i \mod 7$.

$\mathrm{RS}_{\mathbb{F}}[n, k, n-k+1]$ code as we defined in class. Define

$$\begin{aligned} C_2 \;\; = \;\; & \{(c_0, c_1, \ldots, c_{n-1}) \in \mathbb{F}^n \mid c(\alpha^\ell) = 0 \text{ for } 1 \le \ell \le n-k \ , \\ & \text{where } c(X) = c_0 + c_1 X + \cdots + c_{n-1} X^{n-1}\} \ . \end{aligned} \tag{1}$$

- Prove that $C_1 \subseteq C_2$. (*Hint*: Prove that the identity $\sum_{i=0}^{n-1} \alpha^{ji} = 0$ holds for all $j$, $1 \le j \le n-1$, and then make use of it.)
- Prove that $C_1 = C_2$. (*Hint:* Use the previous part and the dimensions of $C_1$ and $C_2$ to argue equality.)

(b) Recall that the $[2^r, r, 2^{r-1}]_2$ Hadamard code is generated by the $r \times 2^r$ matrix whose $i$th (for $0 \le i \le 2^r - 1$) column is the binary representation of $i$. Briefly argue that the Hadamard codeword for the message $(m_1, m_2, \ldots, m_r) \in \{0,1\}^r$ is the evaluation of the multivariate polynomial[2] $m_1 X_1 + m_2 X_2 + \cdots + m_r X_r$ (where $X_1, \ldots, X_r$ are the $r$ variables) over the points in $\{0,1\}^r$.

2. (**Shannon's Capacity theorem for** $\mathrm{BSC}_p$) ($4+6 = 10$ points) In class, we proved Shannon's capacity theorem by choosing general random codes. I mentioned that a similar result can be proved using random linear codes. Also, we saw that a code with relative distance slightly more than $2p$ can have reliable communication over $\mathrm{BSC}_p$. It turns out that the converse needs to be true. We revisit these two issues in this problem.

(a) Briefly argue (full proof not required) why the proof of Shannon's theorem for the binary symmetric channel that we did in class holds even if the encoding function $E$ is restricted to be linear.

(*Hint:* The proof for the linear case does not need the expurgation part of the proof for the general random code case. Argue why this is the case and then make use of it.)

(b) Prove that for communication on $\mathrm{BSC}_p$, if an encoding function $E$ achieves a maximum decoding error probability (taken over all messages) that is exponentially small, i.e., at most $2^{-\gamma n}$ for some $\gamma > 0$, then there exists a $\delta = \delta(\gamma, p) > 0$ such that the code defined by $E$ has relative distance at least $\delta$. In other words, good distance is *necessary* for exponentially small maximum decoding error probability.

(*Hint:* Analyze the probability that the $\mathrm{BSC}_p$ noise converts one codeword into another.)

3. (**Shannon's Capacity theorem for Erasure Channels**) ($6+4+4+1 = 15$ points) The binary erasure channel with erasure probability $\alpha$ has capacity $1 - \alpha$. In this problem, you will prove this result (and its generalization to larger alphabets) via a sequence of smaller results.

(a) For positive integers $k \le n$, show that less than a fraction $q^{k-n}$ of the $k \times n$ matrices $G$ over $\mathbb{F}_q$ fail to generate a linear code of block length $n$ and dimension $k$. (Or equivalently, except with probability less than $q^{k-n}$, the rank of a random $k \times n$ matrix $G$ over $\mathbb{F}_q$ is $k$.)

(*Hint:* Try out the obvious greedy algorithm to construct a $k \times n$ matrix of rank $k$. You will see that you will have many choices every step: from this compute (a lower bound on) the number of full rank matrices that can be generated by this algorithm.)

[2] E.g. $f(X_1, X_2, X_3) = X_1 + X_3$ is a multivariate polynomial and $f$ evaluated at $(1,0,0)$ is $1 + 0 = 1$.

(b) Consider the $q$-ary erasure channel with erasure probability $\alpha$ ($q\mathrm{EC}_\alpha$, for some $\alpha$, $0 \le \alpha \le 1$): the input to this channel is a field element $x \in \mathbb{F}_q$, and the output is $x$ with probability $1-\alpha$, and an erasure '?' with probability $\alpha$. For a linear code $C$ generated by an $k \times n$ matrix $G$ over $\mathbb{F}_q$, let $D : (\mathbb{F}_q \cup \{?\})^n \to C \cup \{\mathsf{fail}\}$ be the following decoder:

$$D(y) = \begin{cases} c & \text{if } y \text{ agrees with exactly one } c \in C \text{ on the unerased entries in } \mathbb{F}_q \\ \mathsf{fail} & \text{otherwise} \end{cases}$$

For a set $J \subseteq \{1, 2, \ldots, n\}$, let $P_{\mathrm{err}}(G|J)$ be the probability (over the channel noise and choice of a random message) that $D$ outputs fail conditioned on the erasures being indexed by $J$. Prove that the average value of $P_{\mathrm{err}}(G|J)$ taken over all $G \in \mathbb{F}_q^{k \times n}$ is less than $q^{k-n+|J|}$.

(c) Let $P_{\mathrm{err}}(G)$ be the decoding error probability of the decoder $D$ for communication using the code generated by $G$ on the $q\mathrm{EC}_\alpha$. Show that when $k = Rn$ for $R < 1-\alpha$, the average value of $P_{\mathrm{err}}(G)$ over all $k \times n$ matrices $G$ over $\mathbb{F}_q$ is exponentially small in $n$.

(d) Conclude that one can reliably communicate on the $q\mathrm{EC}_\alpha$ at any rate less than $1-\alpha$ using a linear code.

(*Note:* Even if you cannot prove a sub-problem, you can use it as a given for the subsequent sub-problems.)

4. (**Intractability of Maximum Likelihood Decoding**) I have mentioned a few times in class that MLD is a notoriously hard to implement any faster than exponential time. In this problem we will show that doing MLD for linear codes in general is NP-hard.
(*This problem is for your cognitive pleasure only; no need to turn this problem in*)

Given an undirected graph $G = (V, E)$, consider the binary code $C_G \subseteq \{0,1\}^{|E|}$, where every codeword in $C_G$ corresponds to a cut in $G$. More precisely, every position in any vector in $\{0,1\}^{|E|}$ is associated with an edge in $E$. Let $\mathbf{c} \in C_G$ be a codeword. Let $E_{\mathbf{c}} = \{i \in E | c_i = 1\}$. Then $E_{\mathbf{c}}$ must correspond to exactly the edges in some cut of $G$.

(a) Prove that $C_G$ is a linear code.

(b) Prove that if one can do MLD on $G$ in polynomial time then one can solve the Max-Cut problem[3] on $G$ in polynomial time. Conclude that solving the MLD problem on linear codes in general is NP-hard.
(*Hint*: Try to think of a vector $\mathbf{y} \in \{0,1\}^{|E|}$ such that solving MLD with received word $\mathbf{y}$ for $C_G$ is equivalent to solving the Max-Cut problem on $G$.)

[3] Given a graph $G = (V, E)$, a cut is a partition of the vertices into sets $S \subseteq V$ and $\overline{S} = V \setminus S$. The size of the cut is the number of edges that have exactly one end-point in $S$ and the other in $\overline{S}$. The Max-Cut of $G$ is a cut with the maximum possible size. Max-Cut is a well known NP-hard problem.