Error Correcting Codes: Combinatorics, Algorithms and Applications Fall 2007

HOMEWORK

Due Wednesday October 31, 2007 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. You are strongly urged to try and solve the problems without consulting any reference material besides what we cover in class (such as any textbooks or material on the web where the solution may appear either fully or in part). If for some reason you feel the need to consult some source, please acknowledge the source and try to explain what difficulty you couldn't overcome before consulting the source and how it helped you overcome that difficulty.
You will also be busy with your project proposals which are due on October 19, so I encourage you to start thinking on the problems early. I should mention that the problems below are not meant to *directly* test what we have covered in the class. Rather, the philosophy behind these problems is to cover topics that were either covered fleetingly in class or have not been covered at all (but are closely related to topics that have been covered in some detail).

---

1. (**Systematic Codes**) In the class I mentioned that there is a way to easily obtain the parity check matrix of a linear code from its generator matrix. In this problem, we will look at this "conversion" procedure.

   (a) Prove that any generator matrix $\mathbf{G}$ of an $[n,k]_2$ code C (recall that $\mathbf{G}$ is a $k \times n$ matrix) can be converted into another equivalent generator matrix of the form $\mathbf{G}' = [\mathbf{I}_k|\mathbf{A}]$, where $\mathbf{I}_k$ is the $k \times k$ identity matrix and $\mathbf{A}$ is some $k \times (n-k)$ matrix. By equivalent, I mean that the code generated by $\mathbf{G}'$ is the same as C, except some positions in (all of) the codewords might be permuted (in some fixed order).

   Note that the code generated by $\mathbf{G}'$ has the message bits as its first k bits in the corresponding codeword. Such codes are called *systematic codes*. In other words, every linear code can be converted into a systematic code.

   (b) Given an $k \times n$ generator matrix of the form $[\mathbf{I}_k|\mathbf{A}]$, give a corresponding $(n-k) \times n$ parity check matrix. Briefly justify why your construction of the parity check matrix is correct.
   (*Hint*: Try to think of a parity check matrix that can be decomposed into two submatrices: one will be closely related to $\mathbf{A}$ and the other will be an identity matrix, though the latter might not be a $k \times k$ matrix).

2. (**Operations on Codes**) In class we have seen some examples of how one can modify one code to get another code with interesting properties (for example, the construction of the Hadamard code from the Simplex code and the construction of codes with smaller block lengths in the proof of the Singleton bound). In this problem you will need to come up with various ways of constructing new codes from existing ones.

Prove the following statements (recall that the notation $(n, k, d)_q$ code is used for general codes with q^k codewords where k need not be an integer, whereas the notation $[n, k, d]_q$ code stands for a *linear code* of dimension k):

(a) If there exists an $(n, k, d)_q$ code, then there also exists an $(n-1, k, d' \geq d-1)_q$ code.

(b) If there exists an $(n, k, d)_2$ code with d odd, then there also exists an $(n+1, k, d+1)_2$ code.

(c) If there exists an $(n, k, d)_{2^m}$ code, then there also exists an $(nm, km, d' \geq d)_2$ code.

(d) If there exists an $[n, k, d]_{2^m}$ code, then there also exists an $[nm, km, d' \geq d]_2$ code.

(e) If there exists an $[n, k, d]_q$ code, then there also exists an $[n-d, k-1, d' \geq \lceil d/q \rceil]_q$ code.

3. (**Alternate Proof of the Singleton Bound**) Use 2(a) above to prove the Singleton bound. (*Hint*: Apply 2(a) repetitively till you "cannot" apply it anymore. What can you say about the resulting code at that point?)

4. (**Shannon's Capacity theorem for** BSC_p) 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 p can have reliable communication over BSC_p. I also mentioned 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.

   (b) Prove that for communication on 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 δ. In other words, good distance is *necessary* for exponentially small maximum decoding error probability.

5. (**Shannon's Capacity theorem for Erasure Channels**) In class I mentioned that the binary erasure channel with erasure probability α 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 $k \leq n$, 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 G is k.)

   (b) Consider the q-ary erasure channel with erasure probability α ($q\mathrm{EC}_\alpha$, for some α, $0 \leq \alpha \leq 1$): 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 α. 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 $P_{\text{err}}(G|J)$ be the probability (over the channel noise and choice of a random message) that D outputs $\textsf{fail}$ conditioned on the erasures being indexed by J. Prove that the average value of $P_{\text{err}}(G|J)$ taken over all $G \in \mathbb{F}_q^{k \times n}$ is less than $q^{k-n+|J|}$.

(c) Let $P_{\text{err}}(G)$ be the decoding error probability of the decoder D for communication using the code generated by G on the $q\text{EC}_\alpha$. Show that when $k = Rn$ for $R < 1 - \alpha$, the average value of $P_{\text{err}}(G)$ 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 $q\text{EC}_\alpha$ at any rate less than $1 - \alpha$ using a linear code.

6. (**Alternate definitions of codes**) We have defined Reed-Solomon and Hadamard codes in class. In this problem you will prove that certain alternate definitions also suffice.

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

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

where α is the generator of the multiplicative group $\mathbb{F}^*$ of $\mathbb{F}$.[1]

Prove that

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

(*Hint*: Prove that the identity $\sum_{i=0}^{n-1} \alpha^{ji} = 0$ holds for all j, $1 \leq j \leq n-1$, and then make use of it.)

(b) Recall that the $[2^r, r, 2^{r-1}]_2$ Hadamard code is generated by the $r \times 2^r$ matrix whose ith (for $0 \leq i \leq 2^r - 1$) 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 $m_1 X_1 + m_2 X_2 + \cdots + M_r X_r$ (where $X_1, \ldots, X_r$ are the r variables) over all the possible assignments to the variables $(X_1, \ldots, X_r)$ from $\{0,1\}^r$.

Using the definition of Hadamard codes above prove the fact that the code has distance 2^{r-1}.

(*Hint*: First prove that for fixed vector $(m_1, \ldots, m_r) \in \{0,1\}^r$, The probability that $\sum_{i=1}^r m_i X_i = 0$ is exactly $\frac{1}{2}$ (where the probability is over random assignment of values to $X_1, \ldots, X_r$) and then make use of it.)

7. (**BCH codes**) In this problem you will look at a very important class of codes called BCH codes[2], which unfortunately we did not have time to study in class.

Let $\mathbb{F} = \mathbb{F}_{2^m}$. Consider the binary code C_{BCH} defined as $\text{RS}_{\mathbb{F}}[n, k, n-k+1] \cap \mathbb{F}_2^n$.

(a) Prove that C_{BCH} is a binary linear code of distance at least $d = n - k + 1$ and dimension at least $n - (d-1)\log_2(n+1)$.
(*Hint*: Use the characterization (1) of the Reed-Solomon code.)

[1] This means that $\mathbb{F}^* = \mathbb{F} \setminus \{0\} = \{1, \alpha, \ldots, \alpha^{n-1}\}$. Further, $\alpha^n = 1$.

[2] The acronym BCH stands for Bose-Chaudhuri-Hocquenghem, the discoverers of this family of codes.

(b) Prove a better lower bound of $n - \lceil \frac{d-1}{2} \rceil \log_2(n+1)$ on the dimension of C_{BCH}.
(*Hint*: Try to find redundant checks amongst the "natural" parity checks defining C_{BCH}).

(c) For $d = 3$, C_{BCH} is the same as another code we have seen. What is that code?

(d) (*For your cognitive pleasure only; no need to turn this part in*) For constant d (and growing n), prove that C_{BCH} have nearly optimal dimension for distance d, in that the dimension cannot be $n - t\log_2(n+1)$ for $t < \frac{d-1}{2}$.

8. (**Rate of linear list-decodable codes**) For $0 < p < 1$ and a positive integer L, call a code $C \subset \Sigma^n$ to be (p, L)-list decodable if every Hamming ball of radius pn (in the space Σ^n) has at most L codewords of C. Prove that for every finite field $\mathbb{F}_q$, $0 < p < 1 - 1/q$, integer $L \geq 1$, and large enough n, there is a (p, L)-list decodable linear code $C \subseteq \mathbb{F}_q^n$ that has rate at least $1 - H_q(p) - \frac{1}{\log_q(L+1)} - o(1)$.
(*Hint*: Apply the usual random coding method of picking a generator matrix at random. In estimating the probability that L nonzero messages all get mapped into a ball of radius pn, these L events are not all independent (and this is the difference compared to picking a general random code). But at least how many of these events are independent of one another?)

9. (**Intractability of Maximum Likelihood Decoding**) I have mentioned a few times in class that MLD is a notoriously hard decoding function 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.