Error Correcting Codes: Combinatorics, Algorithms and Applications Spring 2009

HOMEWORK
Due Monday March 23, 2009 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 :-)]

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 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).

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

Also note that March 23 is the **deadline** to let me know which Wikipedia entry you will create/update. A one page report is due a week after March 23.

---

1. (**Systematic Codes**) ($4 + 6 = 10$ points) In the class I alluded to the fact 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 $\mathbf{G}'$ has a linear bijective map to C.

   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**) ($1+2+2+1+4 = 10$ points) 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. (*Note:* This one is a bit tricky!)

3. (**Distance of General Random Codes**) (5 points) In class, we saw Varshamov's proof that random *linear* code meets the GV bound. It is natural to as the question for general random codes (i.e., the random codes considered by Shannon in his proof for the capacity of BSC_p). We will do so in this problem.

   (a) Prove that a random binary code with rate $R > 0$ with high probability has relative distance $\delta \geq H^{-1}(1 - 2R - \epsilon)$.[1] Note that this is worse than the bound we proved in class for random linear codes.

   (*Hint:* Proceed with the proof as in the random linear case: what events do you now need to take care of in the union bound?)

   (b) (*For your cognitive pleasure only; no need to turn this part in in*)
   Prove that with high probability the relative distance of a random code of rate R is at most $H^{-1}(1 - 2R) + \epsilon$. In other words, general random codes are worse than random linear codes in terms of their distance.

4. (**Toeplitz Matrix**) ($1 + 10 + 2 + 2 = 15$ points) In class I mentioned that the Varshamov proof can be converted (using standard derandomization tricks) into a $q^{O(n)}$ deterministic algorithm to construct a q-ary code that lies on the GV bound. In this problem, we will look at another way to arrive at an exponential time construction.

   (a) (*A warmup*) Argue that Varshamov's proof gives a $q^{O(kn)}$ time algorithm that constructs an $[n, k]_q$ code on the GV bound. (Thus, the goal of this problem is to "shave" off a factor of k from the exponent.)

   (b) A $k \times n$ *Toeplitz Matrix* $A = \{A_{i,j}\}_{i=1,\, j=1}^{k,\, n}$ satisfies the property that $A_{i,j} = A_{i-1,j-1}$. In other words, any diagonal has the same value. For example, the following is a 4×6 Toeplitz matrix:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 7 & 1 & 2 & 3 & 4 & 5 \\ 8 & 7 & 1 & 2 & 3 & 4 \\ 9 & 8 & 7 & 1 & 2 & 3 \end{pmatrix}$$

[1] I forgot to mention this in class but generally $H(\cdot)$ is used to denote $H_2(\cdot)$. Further the "inverse" function $H^{-1}(\cdot)$ is defined as follows. For every $y \in [0, 1]$, $H^{-1}(y) = x$ iff $H(x) = y$ and $x \in [0, 1/2]$.

A *random* $k \times n$ Toeplitz matrix $T \in \mathbb{F}_q^{k \times n}$ is chosen by picking the entries in the first row and column uniformly (and independently) at random.

Prove the following claim: For any non-zero $\mathbf{m} \in \mathbb{F}_q^k$, the vector $\mathbf{m} \cdot T$ is uniformly distributed over $\mathbb{F}_q^n$.

(*Hint:* Write down the expression for the value at each of the n positions in the vector $\mathbf{m} \cdot T$ in terms of the values in the first row and column of T. Think of the values in the first row and column as variables. Then divide these variables into two sets (this "division" will depend on $\mathbf{m}$) say S and $\overline{S}$. Then argue the following: for every fixed $\mathbf{y} \in \mathbb{F}_q^n$ and for every fixed assignment to variables in S, there is a unique assignment to variables in $\overline{S}$ such that $\mathbf{m}T = \mathbf{y}$.)

(c) Briefly argue why the claim in part (b) implies that a random code defined by picking its generator matrix as a random Toeplitz matrix with high probability lies on the GV bound.

(d) Conclude that an $[n, k]_q$ code on the GV bound can be constructed in time $2^{O(k+n)}$.

5. (**Shannon's Capacity theorem for** 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 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 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.

   (*Hint:* Analyze the probability that the BSC_p noise converts one codeword into another.)

6. (**Shannon's Capacity theorem for Erasure Channels**) (6 + 4 + 4 + 1 = 15 points) 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 a random $k \times n$ matrix G over $\mathbb{F}_q$ 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 $P_{\mathrm{err}}(G|J)$ 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.

7. (**Alternate definition of RS codes**) (10 points) 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

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

where α is the generator of the multiplicative group $\mathbb{F}^*$ of $\mathbb{F}$.[2]

Prove that

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

(*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.)

(*Hint on hint:* Convince yourself that the formula for geometric series also holds over finite fields.)

8. (**Rate of linear list-decodable codes**) (6 + 4 = 10 points)

(a) Let $\mathbf{v}_1, \ldots, \mathbf{v}_\ell \in \mathbb{F}_q^a$ be *linearly-independent* vectors. Then prove that for a random $a \times b$ matrix $M \in \mathbb{F}_q^{a \times b}$, the vectors $\mathbf{v}_i \cdot M$ $(1 \leq i \leq \ell)$ are random *independent* vectors in $\mathbb{F}_q^b$.

(b) 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? Part (a) will be useful in answering this question.)

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

9. (**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.