Essential Coding Theory Madhu Sudan
6.896
Due: Wednesday, October 30, 2002

Problem Set 3

Instructions: See PS1.

1. Asymptotics of codes: Given $\epsilon > 0$ express the rate of the best family of binary codes of relative distance $\frac{1}{2} - \epsilon$, you can (a) construct, and (b) show the existence of. Express the rate in big-Oh notation (i.e., $O(\epsilon^d)$ implies there exist constants c and ϵ_0 such that for all $\epsilon < \epsilon_0$, the rate of the code of relative distance $\frac{1}{2} - \epsilon$ is at least $c\epsilon^d$.) How constructive are your codes in Part (a)?

    Obviously one should be doing Part (b) first. The Gilbert-Varshamov bound gives a rate of $R = 1 - H(\frac{1}{2} - \epsilon)$. Using $H(\frac{1}{2} - \epsilon) = 1 - \Theta(\epsilon^2)$, we get $R = \Theta(\epsilon^2)$ for a randomly chosen code. As mentioned in lectures, this is shown to be optimal by the Linear Programming bound, so can't be improved.

    For constructive results, we have two options:

    Option 1 - Concatenation of Reed-Solomon code with "greedily chosen code": Here for an appropriate integer ℓ, we pick an outer RS code of relative distance $1 - \epsilon/2$ and block length 2^ℓ over an alphabet of size $q = 2^\ell$. The rate of this code is $\epsilon/2$. We then concatenate it with a greedily chosen binary code of relative distance $\frac{1}{2} - \frac{\epsilon}{2}$, message length ℓ and block length $\ell/\Theta(\epsilon^2)$. The concatenated code has block length $n = q^2/\Theta(\epsilon^2)$, rate $\Theta(\epsilon^3)$ and relative distance $(1 - \epsilon/2)(\frac{1}{2} - \frac{\epsilon}{2}) \geq \frac{1}{2} - \epsilon$. The construction time of this code is dominated by the time taken to construct the inner code which is $2^{\ell/\Theta(\epsilon^2)} = n^{\Theta(1/\epsilon^2)}$.

    Option 2 - Concatenation of AG codes with Hadamard code: In this case we pick an appropriately large constant q so that an outer AG code of relative distance $1 - 2\epsilon$ and rate ϵ of arbitrarily long block length ℓ can be found. Note that this requires $q = O(\epsilon^2)$ (since an AG code needs $1 - R - \delta \geq \frac{1}{\sqrt{q}-1}$). Now concatenate this code with an inner Hadamard code with q messages and block length roughly q. The rate of this code is $\log q/q = \Theta(\epsilon^{-2} \log \frac{1}{\epsilon})$. Concatenating the two gives a code of block length $n = \ell\epsilon^{-2}$, rate $\Omega(\epsilon^{-3} \log \frac{1}{\epsilon})$ and relative distance $\frac{1}{2} - \epsilon$. So the code has slightly better rate (by a logarithmic factor. Furthermore it can be constructed in time that is a fixed polynomial in n ($\Theta(n^2)$ as per latest results).

    The literature has a third code with such rate and distance, due to Alon, Bruck, Naor, Naor and Roth. We may encounter this code later.

2. Variants of RS codes: The two parts of this question consider variants of Reed-Solomon codes over $\mathbb{F}_q$, obtained by evaluations of polynomials at n distinct points $\alpha_1, \ldots, \alpha_n \in \mathbb{F}_q$. The message will be speified by a sequence of coefficients $c_0, \ldots, c_{k-1} \in \mathbb{F}_q$ and its encoding will be the evaluation of a polynomial $p(x)$ at $\alpha_1, \ldots, \alpha_n$. What will be different is the definition of $p(x)$ given $c_0, \ldots, c_{k-1}$. Give exact bounds on the distance of the resulting code. (Note, the distance may be a function of the set $\{\alpha_1, \ldots . \alpha_n\}$.)

(a) $p(x) = \sum_{i=0}^{k-1} c_i x^{i+\ell}$, where ℓ is some non-negative integer.

(b) $p(x) = \sum_{i=0}^{k-1} c_i x^{2i}$.

(a) For a message $c_0, \ldots, c_{k-1}$, let $C(x) = \sum_{i=0}^{k-1} c_i x^i$ be the standard Reed-Solomon polynomial associate with this message. Then in the first part, the encoding of this message is the sequence $\langle \alpha_i^\ell C(\alpha_i) \rangle_{i=1}^n$. If $\ell = 0$, this is just the Reed-Solomon code and its distance equals $n - k + 1$. Else consider a non-zero codeword of minimum weight. For every i s.t. $\alpha_i^\ell C(\alpha_i) = 0$, either $\alpha_i = 0$ or $C(\alpha_i) = 0$. There are at most $k - 1$ locations for which the latter can hold (for a non-zero message) and at most one coordinate for which the former holds. Thus the code has distance at least $n - k$. Furthermore equality holds iff there exists an i s.t. $\alpha_i = 0$, else it distance is $n - k + 1$.

(b) Let C be as above. Then here the encoding is the evaluation of C at $\beta_1, \ldots, \beta_n$ where $\beta_i = \alpha_i^2$. This is essentially just a RS code, except we don't know that the β_i's are distinct. Indeed every β_i could potentially appear twice in the sequence. Let S be the set of β_i's that appear twice, and let T be the set of β_i's that appear exactly once. Let $s = |S|$ and $t = |T|$ (so $2s+t = n$). Then the minimum distance of the code is at most $n - (k - 1 + \min\{s, k-1\})$. (Consider C which is zero on all of S and $k - 1 - s$ points of T if $s < k - 1$ or on $k - 1$ points of S is $s \geq k - 1$ - this achieves the bound.)
Now to get some bounds on s for given q and n. If q is a power of two, then $s = 0$, since the map $\alpha \mapsto \alpha^2$ is a bijection with the map $\beta \mapsto \beta^{q/2}$ inverting it. For other fields there are at exactly $q/2$ distinct squares in the field. So this implies $n - q/2 \leq s \leq n/2$ and any s in this range is achievable.

3. Hadamard matrices: Recall that an $n \times n$ matrix H all of whose entries are from $\{+1, -1\}$ is a Hadamard matrix if $H \cdot H^T = n \cdot I$ where the matrix product is over the reals and I is the $n \times n$ identity matrix.

(a) Show that if there is an $n \times n$ Hadamard matrix then n is either 1 or 2 or a multiple of 4.

Let $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ be three distinct rows of a Hadamard matrix. (So we are assuming $n \geq 3$.) For $i, j \in \{1, -1\}$, let $S_{i,j} = \{k | a_k = i \cdot b_k \text{ and } a_k = j \cdot c_k\}$. Let $\alpha = |S_{1,1}|$, $\beta = |S_{1,-1}|$, $\gamma = |S_{-1,1}|$, and $\delta = |S_{-1,-1}|$. Then $\alpha + \beta$ counts the number of coordinates where $\mathbf{a}$ equals $\mathbf{b}$ and so $\alpha + \beta = n/2$. Similarly $\alpha + \gamma$ counts the number of coordinates where $\mathbf{a}$ equals $\mathbf{c}$ and so $\alpha + \gamma = n/2$. Finally, $\alpha + \delta$ counts the number of coordinates where $\mathbf{b}$ equals $\mathbf{c}$ and so $\alpha + \delta = n/2$. Oh, and of course, $\alpha + \beta + \gamma + \delta = n$. Solving the 4×4 linear system above, we get $\alpha = \beta = \gamma = \delta = n/4$. Since each is an integer, we have n must be a multiple of 4.

(b) Given an $n \times n$ Hadamard matrix H_n and an $m \times m$ Hadamard matrix H_m, construct an $(nm) \times (nm)$ Hadamard matrix.

Let $\mathbb{F}$ be any field (say rationals, for this problem). For vectors $\mathbf{a} \in \mathbb{F}^n$ and $\mathbf{b}\mathbb{F}^m$, let $\mathbf{a} \otimes \mathbf{b} \in \mathbb{F}^{nm}$ denote their outer product (aka tensor product), namely the vector whose ij-th coordinate is $a_i \cdot b_j$. Note that if $\mathbf{a}$, $\mathbf{b}$ are $+1/-1$ vectors, then so is

$\mathbf{a} \otimes \mathbf{b}$. Furthermore, if $\mathbf{a}, \mathbf{c} \in \mathbb{F}^n$ and $\mathbf{b}, \mathbf{d} \in \mathbb{F}^m$,

$$\langle \mathbf{a} \otimes \mathbf{b}, \mathbf{c} \otimes \mathbf{d} \rangle = \sum_{ij} a_i b_j c_i d_j = (\sum_i a_i c_i)(\sum_j b_j d_j) = \langle \mathbf{a}, \mathbf{c} \rangle \cdot \langle \mathbf{b}, \mathbf{d} \rangle.$$

We show how to use tensor products to build a big Hadamard matrix from two smaller ones.

Let $\mathbf{u}_1 \ldots, \mathbf{u}_n$ be the rows of H_n and let $\mathbf{v}_1 \ldots, \mathbf{v}_m$ be the rows of H_m. By the condition $H_n H_n^T = nI$, we have $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$ if $i \neq j$. (Similarly for the $\mathbf{v}_i$'s.) Let H_{nm} be the matrix whose rows are $\mathbf{u}_i \otimes \mathbf{v}_j$ for all $i \in [n]$, $j \in [m]$. As noted above, this is a $+1/-1$ matrix. Thus the diagonal entries of $H_{nm} H_{nm}^T$ are all nm as required. Now consider the off-diagonal entry $(H_{nm} H_{nm}^T)_{(ij),(kl)} = \langle \mathbf{u}_i \otimes \mathbf{v}_j, \mathbf{u}_k \otimes \mathbf{v}_l \rangle = \langle \mathbf{u}_i, \mathbf{u}_k \rangle \cdot \langle \mathbf{v}_j, \mathbf{v}_l \rangle$. Since at least one of the conditions $i \neq k$ or $j \neq l$ holds, we have the above inner product is zero. This proves the off-diagonal entries are zero as required.

(c) (Not to be turned in) Let q be a prime power equivalent to 3 modulo 4. Let $H = \{h_{ij}\}$ be the $(q+1) \times (q+1)$ matrix with $h_{ij} = 1$ if $i = 1$ or $j = 1$ or $i = j$, and $h_{ij} = (j-i)^{(q-1)/2}$ otherwise. Verify that H is a Hadamard matrix. (The purpose of this exercise is point out that Hadamard matrices of many size, and not just powers of 2, exist.)

Turns out there were some typoes in the above question so it was not easy to "verify". Will post revised question + answer shortly. (Some you nevertheless responded with "verified"!)

4. Let C be an infinite family of binary codes obtained by concatenation of two infinite families of codes C_1 and C_2. (The ith code of C is obtained by concatenating the ith code of C_1 with the ith code in C_2. The block lengths of the codes in C_1 and C_2 tend to infinity as $i \to \infty$.) Give an upper bound on the rate of C as a function of its minimum distance.

Informally, If C_1 has rate R_1 and rel. distance δ_1 then by the Singleton bound $R_1 + \delta_1 \leq 1$. Similarly if C_2 has rate R_2 and rel. distance δ_1, then $R_2 + 2\delta_2 \leq 1$ (by the Plotkin bound, since C_2 is binary). The concatenated code has rel. distance $\delta = \delta_1 \delta_2$ and rate $R \leq (1 - \delta_1)(1 - 2\delta_2)$. Setting $\delta_2 = \delta/\delta_1$ and maximizing over δ_1 we get $R \leq \max_{\delta \leq \delta_1 \leq 1}\{1 - \delta_1 - 2\delta/\delta_1 + 2\delta\} = 1 - 2\sqrt{2\delta} + 2\delta$. A "Mathematica" plot reveals this is not as good as the GV bound. Some analytic confirmation of this fact can be found when δ approaches 0 - where the GV bound gives a rate of $1 - O(\delta \log \frac{1}{\delta})$, while this bound is $1 - \Omega(\sqrt{\delta})$. On the other hand, when $\delta = \frac{1}{2} - \epsilon$ and $\epsilon \to 0$, this bound does not beat the GV bound asymptotically. (Both are $\Theta(\epsilon^2)$.) However replacing the Plotkin bound above with the LP bound will allow us to prove an upper bound on the rate of concatenated codes of $O(\epsilon^3)$ which is again worse than the GV bound.

To be more formal with any of the above, we need to take into account the fact that specific members of these families are not subject exactly to the Singleton bound/Plotkin bound, and furthermore their rates may vary. To deal with all this formally, let $R_{i,j}$ (resp. $\delta_{i,j}$) denote the rate (resp. relative distance) of the jth code in the ith family. Since the block lengths of the code C_1 tends to infinity, we have: For every $\epsilon > 0$, there exists a j_0 such that for every $j \geq j_0$, it is the case that $R_{1,j} + \delta_{1,j} \leq 1 + \epsilon$ (Singleton bound using $n_j \geq \frac{1}{\epsilon}$). Now since the length of the codes C_2 tend to infinity, once again we have: For every $\epsilon > 0$, there exists a j_0' such that for all $j \geq j_0'$, $R_{2,j} + 2\delta_{2,j} \leq 1 + 2\epsilon$. Combining the two, we have for every $j \geq \max\{j_0, j_0'\}$ the rate of the jth concatenated code $R_j = R_{1,j} \cdot R_{2,j} \leq (1 - \delta_{1,j})(1 - 2\delta_{2,j}) + O(\epsilon)$. Using $\delta_{1,j} \cdot \delta_{2,j} \geq \delta$ gives $R_j \leq 1 - 2\sqrt{2\delta} + 2\delta + O(\epsilon)$. Letting $\epsilon \to 0$ gives the bound of the first para, formally.

5. Consider the following simple edit distance between strings: $x \in \Sigma^n$ is at distance d from $y \in \Sigma^m$ if y can be obtained from x by first deleting upto d coordinates of x and getting an intermediate string $z \in \Sigma^\ell$ where $\ell \geq n - d$, and then inserting up to d characters into z (at arbitrary locations) to get y. What are the analogs of the Singleton bound, the Hamming (packing) bound on codes, and the Gilbert-Varshamov bounds for this measure of distance?

   Both the Singleton and Hamming bounds obviously hold for edit distance codes as well (since the edit distance is upper bounded by the Hamming distance). The issue is: Are they tight?

   The Singleton bound is essentially tight. Consider an $[n, k, n-k+1]$ RS code C over a q-ary alphabet with $q \gg n$. Now consider the code C' over the alphabet $\Sigma = \mathbb{F}_q \times [n]$, whose codewords are strings of the form $(c'_1, \ldots, c'_n)$ where $c'_i = (i, c_i)$ and $(c_1, \ldots, c_n)$ is a codeword of C. This code has $q^k = |\Sigma|^{k'}$ codewords with edit distance $n - k + 1$, where $k' = k \log q / (\log q + \log n)$. As $q \to \infty$, this quantity approaches k, indicating that these codes are achieving the Singleton bound ($d \geq n - k + 1$).

   As for the Hamming bound, obviously it is not tight since it is not tight even for Hamming distance. However I don't believe it is tight even for fixed d, unlike the case of the Hamming distance. I don't have a proof. Attempts welcome.

   Finally the G-V bound in this case would be obtained by upper bounding the volume of the ball of radius d. An easy bound is roughly $\left(\binom{n}{d}\right)^2 2^d$ (for a binary alphabet). This would lead to the asymptotic bound saying there exists codes of rate R and distance δ for $R = 1 - \delta - 2H(\delta)$.

   Again, my guess would be that the G-V bound is close to being right (and in particular, one can't have codes of relative distance $\delta = .4$, say.) Proofs/counterexamples welcome.