Phdopen lectures, University of Warsaw
Algorithmic coding theory: Some recent advances

Venkatesan Guruswami
Carnegie Mellon University

November 2012

1 Problem set

Please attempt *at least* **three** of the five problems in this section, and turn in your solutions as a pdf (preferably typeset in LaTeX). You are *urged* to try and solve the problems without consulting any reference material other than material covered in lectures and the accompanying lecture notes posted on the website. If for some reason you end up consulting some external source (such as a textbook or sources on the web), *please acknowledge the source.*

For any clarifications about the questions, do not hesitate to email me at `guruswami@cmu.edu`.

1. (Singly exponential time algorithm to construct codes meeting GV bound) Let n, k, d be positive integers satisfying $n \geq k, d$ and

$$2^k < \frac{2^n}{\sum_{j=0}^{d-2} \binom{n-1}{j}} .$$

   Prove that in such a case there exists an $[n, k, d]$ binary linear code (i.e., a linear code of block length n, with 2^k codewords, and minimum distance d). Also give an algorithm running in time $2^{n-k}\text{poly}(n)$ to construct the parity check matrix of such a code.

   Hint: Use the characterization of distance in terms of minimal sized linear dependence of columns of a parity check matrix.

2. (d-wise independent sets) For integers $1 \leq d \leq n$, call a subset $S \subseteq \{0,1\}^n$ to be d-wise independent if for every $1 \leq i_1 < i_2 < \cdots < i_d \leq n$ and $(a_1, a_2, \ldots, a_d) \in \{0,1\}^d$

$$\text{Prob}_{x \in S}[x_{i_1} = a_1 \wedge x_{i_2} = a_2 \wedge \cdots \wedge x_{i_d} = a_d] = \frac{1}{2^d}$$

   where the probability is over an element x chosen uniformly at random from S.

   Small sample spaces of d-wise independent sets are of fundamental importance in derandomization. The goal of this problem is to show how codes can be used to construct d-wise independent sets.

   Let $H \in \mathbb{F}_2^{m \times n}$ be the parity check matrix of an $[n, n-m, D]_2$ binary linear code of distance $D \geq d+1$. Define $S = \{x^T H \mid x \in \mathbb{F}_2^m\}$. Prove that S is a d-wise independent set of $\{0,1\}^n$ of size 2^m.

3. (Codes and graph cuts) Let $G = (V, E)$ be a connected undirected graph. For each $U \subset V$, define by $\chi_U \in \{0,1\}^E$ the indicator vector χ_U for the edge cut $(U : V \setminus U)$

$$\chi_U(e) = \begin{cases} 1 & \text{if } e = \{u, v\} \text{ with } u \in U,\ v \in V \setminus U \\ 0 & \text{otherwise} \end{cases}$$

(a) Prove the collection of vectors χ_U, $U \subseteq V$, is a binary linear code (call it $\mathsf{cuts}(G)$) of block length $|E|$.

(b) What is the minimum distance of $\mathsf{cuts}(G)$? (Express the answer in terms of a basic quantity concerning the graph G.)

(c) What is the rate of $\mathsf{cuts}(G)$? (Hint: you answer should only involve $|V|$ and $|E|$.)

(d) Can there be a family of graphs $\{G_n : n \geq 1\}$ such that the rate and relative distance of $\mathsf{cuts}(G_n)$ are both bounded away from 0 as $n \to \infty$? (In other words, can these "cut codes" be asymptotically good?)

(e) What is the Hamming distance of the closest codeword in $\mathsf{cuts}(G)$ to 1^E, the all-ones vector?

(f) Can you argue why given $x \in \{0,1\}^E$, finding the codeword of $\mathsf{cuts}(G)$ that is closest to x in Hamming distance is an NP-hard problem? (Hint: use part (e) above.)

(g) *(Bonus; no need to turn in unless you want to)* Can you describe the dual code of $\mathsf{cuts}(G)$? What are its rate and minimum distance? Can you describe a basis for the dual code?

4. (Products of codes) Let C_1 be an $[n_1, k_1, d_1]_2$ binary linear code, and C_2 an $[n_2, k_2, d_2]$ binary linear code. Let $C \subseteq \mathbb{F}_2^{n_1 \times n_2}$ be the subset of $n_1 \times n_2$ matrices whose columns belong to C_1 and whose rows belong to C_2.

   Prove that C is an $[n_1 n_2, k_1 k_2, d_1 d_2]_2$ binary linear code.

5. (Chinese remainder theorem with errors) In this problem, we will consider the number-theoretic counterpart of Reed-Solomon codes. Let $1 \leq k < n$ be integers and let $p_1 < p_2 < \cdots < p_n$ be n distinct primes. Denote $K = \prod_{i=1}^{k} p_i$ and $N = \prod_{i=1}^{n} p_i$. The notation $\mathbb{Z}_M$ stands for integers modulo M, i.e., the set $\{0, 1, \ldots, M-1\}$. Consider the *Chinese Remainder code* defined by the encoding map $E : \mathbb{Z}_K \to \mathbb{Z}_{p_1} \times \mathbb{Z}_{p_2} \times \cdots \times \mathbb{Z}_{p_n}$ defined by:

$$E(m) = (m \mod p_1,\ m \mod p_2,\ \cdots,\ m \mod p_n)\ .$$

   (Note that this is not a code in the usual sense we have been studying since the symbols at different positions belong to different alphabets. Still notions such as distance of this code make sense and are studied in the questions below.)

   (a) Suppose that $m_1 \neq m_2$. For $1 \leq i \leq n$, define the indicator variable $b_i = 1$ if $E(m_1)_i \neq E(m_2)_i$ and $b_i = 0$ otherwise. Prove that $\prod_{i=1}^{n} p_i^{b_i} > N/K$.

   Use the above to deduce that when $m_1 \neq m_2$, the encodings $E(m_1)$ and $E(m_2)$ differ in at least $n - k + 1$ locations.

   (b) This exercise examines how the idea behind the Welch-Berlekamp decoder for Reed-Solomon codes, which we saw in lecture, can be used to decode these codes.

   Suppose $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ is the received word where $r_i \in \mathbb{Z}_{p_i}$. By Part (a), we know there can be at most one $m \in \mathbb{Z}_K$ such that

$$\prod_{i : E(m)_i \neq r_i} p_i^{b_i} \leq \sqrt{N/K}\ . \tag{1}$$

   (Be sure you see why this is the case.) The exercises below develop a method to find the unique such m, assuming one exists.

In what follows, let r be the unique integer in $\mathbb{Z}_N$ such that $r \mod p_i = r_i$ for every $i = 1, 2, \ldots, n$ (note that the Chinese Remainder theorem guarantees that there is a unique such r).

i. Assuming an m satisfying (1) exists, prove that there exist integers y, z with $0 \leq y < \sqrt{NK}$ and $1 \leq z \leq \sqrt{N/K}$ such that $y \equiv rz \pmod{N}$.

ii. Prove also that if y, z are any integers satisfying the above conditions, then in fact $m = y/z$.

(**An algorithmic side remark:** A pair of integers (y, z) satisfying above can be found by solving the integer linear program with integer variables y, z, t and linear constraints: $0 < z \leq \sqrt{N/K}$; and $0 \leq z \cdot r - t \cdot N < \sqrt{NK}$. This is an integer program in a fixed number of dimensions and can be solved in polynomial time. Faster, easier methods are also known for this special problem.)

2 Some "extra" problems

All problems in this section are **optional**. They are merely meant to provide additional challenges to those interested. You are welcome to attempt and email back solutions to any of these, but this is not required.

1. (NP-hardness of Reed-Solomon decoding over large fields) In one of the above problems, we saw that finding the closest codeword in the "code of graph cuts" is NP-hard. In this problem, you will prove that finding the closest codeword in Hamming metric for a certain Reed-Solomon code is NP-hard.

   You may assume that the following problem is NP-hard.

   Instance: A set $S = \{\alpha_1, \ldots, \alpha_n\} \subseteq \mathbb{F}_{2^m}$, an element $\beta \in \mathbb{F}_{2^m}$, and an integer $1 \leq k < n$. (Here $\mathbb{F}_{2^m}$ denotes the field with 2^m elements, which is a field extension of degree m over the field $\mathbb{F}_2$ with two elements.)

   Question: Is there a nonempty subset $T \subseteq \{1, 2, \ldots, n\}$ with $|T| = k+1$ such that $\sum_{i \in T} \alpha_i = \beta$?

   Consider the $[n, k]$ Reed-Solomon code C_{RS} over $\mathbb{F}_{2^m}$ obtained by evaluating polynomials of degree at most $k-1$ at points in S. Define $y \in (\mathbb{F}_{2^m})^n$ as follows: $y_i = \alpha_i^{k+1} - \beta \alpha_i^k$ for $i = 1, 2, \ldots, n$.

   Prove that there is a codeword of C_{RS} at Hamming distance at most $n - k - 1$ from y if and only if there is a set T as above of size $k+1$ satisfying $\sum_{i \in T} \alpha_i = \beta$.

   Conclude that finding the nearest codeword in a Reed-Solomon code over exponentially large fields is NP-hard.

   Remark: Proving NP-hardness for Reed-Solomon code over polynomially sized fields remains an important open problem!

2. (Lower bound on alphabet size for optimal rate list decoding) Let $\epsilon > 0$ be a positive constant, and Σ be a fixed finite alphabet. Suppose we have an infinite family of codes of increasing block lengths over alphabet Σ, each having rate R and list decodable up to a fraction $1 - R - \epsilon$ of errors with a list size bounded from above by n^{1/ϵ^2} where n is the block length of the code. Prove the following lower bound on the alphabet size: $|\Sigma| \geq 2^{\Omega(1/\epsilon)}$.

3. (Optimal size d-wise independent sets) In this problem we will present a generalization of Hamming codes to larger distance, and obtain an implied construction of d-wise independent sample spaces via the connection from one of the earlier problems. We will then prove the optimality of the bound achieved by these codes.

   (a) Let $D = 2t + 1$ be an odd integer. Let $n = 2^m - 1$, and let α be a primitive element of the extension field $\mathbb{F}_{2^m}$. Define the following subset of $\mathbb{F}_2^n$:

   $$\{(f_0, f_1, \ldots, f_{n-1}) \in \mathbb{F}_2^n \mid f(\alpha) = f(\alpha^2) = f(\alpha^3) = \cdots = f(\alpha^{2t}) = 0 \text{ for } f(X) = f_0 + f_1 X + \cdots + f_{n-1} X^{n-1} \in \mathbb{F}_2[X]\} .$$

   Prove that the above is an $[n, k, d]$ binary linear code for $k \geq n - t \log_2(n+1)$ and $d \geq D$.
   Hint: The distance bound is based on non-singularity of Vandermonde matrices. For the lower bound on k, the identity $f(\gamma)^2 = f(\gamma^2)$ for polynomials $f \in \mathbb{F}_2[X]$ is handy.

   (b) Using the above (and problem 2 of Section 1), show how one can construct a $2t$-wise independent subset of $\{0,1\}^n$ of size at most $(n+1)^t$ when n is of the form $2^m - 1$. Deduce a construction of size at most $(2n)^t$ for any n.

   (c) Prove an almost matching lower bound, namely any $2t$-wise independent set $S \subseteq \{0,1\}^n$ satisfies

   $$|S| \geq \sum_{i=0}^{t} \binom{n}{i} . \tag{2}$$

   Suggestion: Use the "linear algebra" method. That is, find an orthonormal set of vectors in $\mathbb{R}^{|S|}$ of cardinality at least the R.H.S of (2).

4. (Covering codes) For $\tau \in [0, 1/2]$, define a binary code C of block length n to be *τ-covering* if every $\mathbf{r} \in \{0,1\}^n$ is within Hamming distance τn from some codeword of C.

   (a) Prove that the rate of a τ-covering code must be at least $1 - h(\tau)$.

   (b) Prove the following characterization for when a binary linear code is τ-covering:
   If H is a parity check matrix for an $[n, k]_2$ linear code C, then C is τ-covering if and only if for every $\mathbf{s} \in \mathbb{F}_2^{n-k}$, there is a set of at most τn columns of H which sum up to $\mathbf{s}$ (over $\mathbb{F}_2$).

   (c) Prove that there exist τ-covering binary **linear** codes C of rate $1 - h(\tau) + o(1)$.
   (Hint: (a) First prove that a random linear code of rate $1 - h(\tau) + o(1)$ τ-covers *most* of the points in $\mathbb{F}_2^n$. This step will rely on pairwise independence of the nonzero codewords in a random linear code, and Chebyshev's tail inequality. (b) Then prove that some $O(\log n)$ translates (cosets) of such a linear code suffice to τ-cover the whole space.)