KTH Computer Science and Communication

Algebraic Gems in TCS: Problem Set 2

Due: Sunday Dec 14, 2014, at 23:59. Submit your solutions as a PDF file by e-mail to `jakobn at kth dot se` with the subject line `Problem set 2: ⟨your full name⟩`. Name the PDF file `PS2_⟨YourFullName⟩.pdf` (with your name coded in ASCII without national characters), and also state your name and e-mail address at the top of the first page. Solutions should be written in LaTeX or some other math-aware typesetting system. Please try to be precise and to the point in your solutions and refrain from vague statements. *Write so that a fellow student of yours can read, understand, and verify your solutions.* In addition to what is stated below, the general rules stated on the course webpage always apply.

Collaboration: Discussions of ideas in groups of two people are allowed—and indeed, encouraged—but you should write down your own solution individually and understand all aspects of it fully. You should also acknowledge any collaboration. State at the beginning of the problem set if you have been collaborating with someone and if so with whom. (Note that collaboration is on a per problem set basis, so you should not discuss different problems on the same problem set with different people.)

Reference material: Some of the problems are "classic" and hence it might be easy to find solutions on the Internet, in textbooks or in research papers. It is not allowed to use such material in any way unless explicitly stated otherwise. Anything said during the lectures or in the lecture notes should be fair game, though, unless you are specifically asked to show something that we claimed without proof in class. It is hard to pin down 100% formal rules on what all this means—when in doubt, ask the lecturer.

About the problems: Some of the problems are meant to be quite challenging and you are not necessarily expected to solve all of them. A total score of around 60 points should be enough for grade E, 90 points for grade D, 120 points for grade C, 150 points for grade B, and 180 points for grade A on this problem set. Any corrections or clarifications will be given at `piazza.com/kth.se/fall2014/dd2442/` and any revised versions will be posted on the course webpage `www.csc.kth.se/DD2442/semteo14/`.

1 (20 p) In Lecture 12, we saw that the Hadamard code could be defined so as to encode a binary string $x \in \{0,1\}^k$ as the value of all linear functions evaluated on x, i.e., as the sequence $(\ell(x) : \ell \in \mathcal{L})$, where $\mathcal{L} = \{\sum_{i=1}^{k} c_i x_i \mid c_i \in \{0,1\}\}$ denotes the set of all linear functions from $\{0,1\}^k$ to $\{0,1\}$. (This is equivalent to the definition we had before our final tweak to double the number of codewords).

In the same way, one can define the *long code* of $x \in \{0,1\}^k$ as the value of *all functions* (linear or not) evaluated on x, i.e., as the sequence $(f(x) : f \in \mathcal{F})$, where we write $\mathcal{F} = \{f \mid f : \{0,1\}^k \to \{0,1\}\}$ to denote the set of all functions from $\{0,1\}^k$ to $\{0,1\}$. Determine the block length and distance of this code. Is it linear?

Hint: It might be helpful to observe that a function $f : \{0,1\}^k \to \{0,1\}$ can be represented as a bitstring $\{0,1\}^{2^k}$. Identifying an integer x with the n-bit binary representation of x, we can then think of $f(x)$ as the xth bit in this bitstring.

2 (20 p) An $(n, k, d)_q$ code is called *systematic* if the message is encoded into a codeword consisting of the message as the k first symbols followed by $n - k$ check symbols. Linear codes can always be made systematic by transforming the generator matrix G to the form $[I_k \mid A]$ (possibly permuting the positions in the codeword).

Write the generator matrix for the $\left[2^\ell - 1, 2^\ell - \ell - 1, 3\right]_2$ Hamming code in systematic form. Can you explain the meaning of the check bits?

3 (20 p) Suppose that we have a Reed-Solomon code with parameters[1] $[n, n/2, n/2]_n$ for $n = 2^m$ and concatenate this code with the trivial binary code that is identity (i.e., has block length and message length m and distance 1). We claimed in class that this yields an $[mn, mn/2, n/2]_2$-code, but is it really true that the distance can be as bad as only $n/2$? Either prove that the minimum distance will in fact be significantly better than $n/2$ or provide an explicit example of a code where the distance is close to $n/2$.

4 (30 p) Recall that the *binary entropy function* $H(p)$ is defined by $H(0) = H(1) = 0$ and $H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$ for $0 < p < 1$. In Lecture 12, we claimed that for the volume of the Hamming ball of radius pn in $\mathbb{F}_2^n$ it holds that $\mathrm{Vol}_2(pn, n) = \sum_{0 \le i \le pn} \binom{n}{i} \approx 2^{nH(p)}$. We now want to make this claim formal.

4a Prove that for $0 \le p \le \frac{1}{2}$ it holds that $\sum_{0 \le i \le pn} \binom{n}{i} \le 2^{nH(p)}$.

Hint: Write $1 = (p + (1-p))^n$ and expand.

4b Prove that for $0 \le p \le \frac{1}{2}$ it holds that

$$\lim_{n \to \infty} \frac{\log_2 \left(\sum_{0 \le i \le pn} \binom{n}{i} \right)}{n} = H(p) \ .$$

Hint: Since you will want to round pn to an integer in order for the binomial coefficients to typecheck, you might also want to use the corollary of the mean value theorem saying that

$$|f(x) - f(y)| \le |x - y| \max_{\xi \in (x,y)} |f'(\xi)| \ .$$

5 (50 p) In this problem we want to investigate various aspects of the Schwartz-Zippel lemma.

5a In Lecture 9, we proved the version of the Schwartz-Zippel lemma that says that any polynomial $f \in \mathbb{F}_q[x_1, x_2, \ldots, x_n]$ of total degree d is zero on at most a fraction $\frac{d}{q}$ of the points in $\mathbb{F}_q^n$. In our proof we wrote the polynomial as $f(x_1, x_2, \ldots, x_n) = \sum_{i=0}^{d} x_1^i f_i(x_2, \ldots, x_n)$, picked the largest i^* such that $f_{i^*}(x_2, \ldots, x_n) \not\equiv 0$, and then argued by induction.

A question that was raised after this lecture was whether we really needed to pick i^* maximal, or whether any i such that $f_i(x_2, \ldots, x_n) \not\equiv 0$ would work equally well. Answer this question. That is, either show that the argument that we had goes through for any i or point out where it fails.

[1] Too avoid clutter, we are a bit sloppy here and ignore rounding and off-by-one errors in the message length-distance relation.

5b There is also a version of the Schwartz-Zippel lemma with bounds on individual degrees that says the following: If $f \in \mathbb{F}_q[x_1, x_2, \ldots, x_n]$ is a non-zero polynomial with individual degrees $\deg_{x_i}(f) \leq d_i$ for $i = 1, \ldots, n$, then f is non-zero on at least a fraction $\frac{\prod_{i=1}^{n}(q-d_i)}{q^n}$ of the points in $\mathbb{F}_q^n$. Prove this lemma.

Hint: Use the fact that we can think of f as a univariate polynomial in x_n with coefficients in the ring $\mathbb{F}_q[x_1, x_2, \ldots, x_{n-1}]$.

5c Prove the version of the Schwartz-Zippel lemma that we needed in Lecture 9 for our analysis of Reed-Muller codes where the polynomials have total degree larger than the field size. That is, prove that if $f \in \mathbb{F}_q[x_1, x_2, \ldots, x_n]$ is a non-zero polynomial with individual degrees $\deg_{x_i}(f) \leq s$ for $i = 1, \ldots, n$ and total degree $\deg(f) = d = sk + r$, then f is non-zero on at least a fraction $\left(1 - \frac{s}{q}\right)^k \left(1 - \frac{r}{q}\right)$ of the points in $\mathbb{F}_q^n$.

Hint: Combine the other Schwartz-Zippel lemmas we have seen.

6 (30 p) In one of the guest lectures on polynomial identity testing for depth-3 powering circuits, or $\Sigma\Lambda\Sigma$ circuits, Michael Forbes considered a lexicographic ordering on monomials and showed that small $\Sigma\Lambda\Sigma$ circuits have to compute polynomials with small leading monomials with respect to this ordering. After some further reasoning, this led to the construction of hitting sets of size roughly $s^{\log s}$ for size-s $\Sigma\Lambda\Sigma$ circuits.

Recall that in the lexicographic ordering there is some order $x_1 > x_2 > \cdots > x_n$ on the variables, and we have that monomials containing x_1 always win over monomials not containing x_1. More formally, if we denote $\vec{x}^{\vec{a}} = \prod x_i^{a_i}$ then $\vec{x}^{\vec{a}} > \vec{x}^{\vec{b}}$ if the first non-zero entry in the vector $\vec{a} - \vec{b}$ is positive, so that, for instance, $x_1x_2 > x_1 > x_2x_3 > x_3x_4^2x_5$ holds. Because of his background in proof complexity, Jakob likes much better the degree lexicographic ordering (also known as the graded lexicographic ordering), in which monomials of larger total degree always win over monomials of smaller total degree and lexicographic ordering is only used to split ties between monomials of the same degree. That is, for degree-lexicographic ordering we would have, for example, $x_3x_4^2x_5 > x_1x_2 > x_2x_3 > x_1$.

Jakob spent a fair chunk of the lecture thinking about why Michael did not choose this latter, clearly more pleasing, ordering instead, and whether all the proofs would still work in this setting or whether the arguments would break down at some critical junction. Please help Jakob to figure this out. That is, either show that everyhing Michael did on the board would still work, and explain why the critical steps would still go through (not necessarily repeating the whole argument verbatim, though), or point out where the proofs would break and why.

7 (40 p) It follows from the Schwartz-Zippel lemma that for any field $\mathbb{F}_q$ and any subset $S = \{\alpha_0, \alpha_1, \ldots, \alpha_d\} \subseteq \mathbb{F}_q$ of size $\left|S\right| = d + 1 \leq q$ it holds that $S^n = \{(\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_n}) \mid 0 \leq i_j \leq d\}$ is a hitting set for polynomials in $\mathbb{F}_q[x_1, x_2, \ldots, x_n]$ of total degree at most d. Prove that it is in fact possible to find hitting sets of size $\binom{n+d}{d} \ll \left|S^n\right| = (d+1)^n$ for such polynomials.
Hint: Consider the set $S' = \{(\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_n}) \mid i_j \geq 0, \sum_{j=1}^{n} i_j \leq d\}$ and analyze the proof of the Schwartz-Zippel lemma.

8 (60 p) We say that an undirected graph $G = (V, E)$ is an (s, d, c)-*vertex expander* if G is d-regular and for every $S \subseteq V(G)$ of size $|S| \leq s$ it holds that $|N(S)| \geq c|S|$, where we write $N(S) = \{u \mid \exists (u, v) \in E(G) \text{ for } v \in S\}$ to denote the set of neighbours of S. The goal of this problem is to establish a connection between vertex expansion and spectral expansion.

8a Prove that if $\mathbf{p}$ is a probability vector, then $\|\mathbf{p}\|_2^2$ is the probability that if $i, j \in [n]$ are chosen independently distributed according to $\mathbf{p}$, we get $i = j$.

8b Prove that if $\mathbf{s}$ is the probability vector denoting the uniform distribution over some subset $S \subseteq V(G)$ of a graph G with random-walk matrix A, then $\|A\mathbf{s}\|_2^2 \geq 1/|N(S)|$.

8c Prove that if G is an (n, d, λ)-spectral expander and $S \subseteq V(G)$ has size $|S| = \epsilon n$, then

$$|N(S)| \geq \frac{|S|}{(1-\epsilon)\lambda^2 + \epsilon} \ .$$

Hint: Show that for $\mathbf{s}$ being the uniform distribution over S it holds that $\|A\mathbf{s}\|_2^2 \leq \|A\mathbf{1}/n\|_2^2 + \lambda^2 \|\mathbf{s} - \mathbf{1}/n\|_2^2$.

Remark: This shows that a Ramanujan graph (with second eigenvalue roughly $2/\sqrt{d}$) has vertex expansion roughly $d/4$ for small enough vertex sets. This can actually be improved to $d/2$, which is tight. Random d-regular graphs, however, have vertex expansion $(1 - \mathrm{o}(1))d$ almost surely.

9 (50 p) ***For this problem, and for this problem only, please feel free to use textbooks, search in the research literature, or roam the internet to find helpful information.***

Let $p \in \mathbb{N}^+$ be a prime number and for $x \in \mathbb{F}_p$ define

$$x^* = \begin{cases} 0 & \text{if } x = 0, \\ x^{-1} & \text{otherwise.} \end{cases}$$

Let $G_p = \big(V_p, E_p\big)$ be the undirected, 3-regular graph with vertex set $V_p = \mathbb{F}_p$ and edges $(x, x+1)$, $(x, x-1)$, and (x, x^*), for each $x \in V_p$.

We claimed in Lecture 5 that such graphs G_p are good expanders. Compute exactly for $p = 7, 11, 13, 17, 19, \ldots$ and as far up as you can go the exact edge expansion

$$h\big(G_p\big) = \min_{\substack{S \subseteq V(G_p) \\ 0 < |S| < p/2}} \frac{|E(S, \overline{S})|}{|S|}$$

for these graphs. Does the expansion seem to converge to some value? Can you find any theoretic lower (or upper) bounds on the expansion in the literature? How do such theoretic bounds compare with the exact values that you can obtain?

In addition to answering the above questions, please explain briefly how you coded up the program used to solve this problem and what running times you observed (your time-out limit should be at least 30 minutes on a reasonably powerful hardware platform as explained below). Please do not submit any code, however, but instead describe how it works. Place the actual code in a directory in the AFS file system where `jakobn` has reading and listing permission `rl` (as shown by `fs la .`). Note that permission `l` is needed for the whole path leading to the directory. Make sure your code works in the KTH CSC Ubuntu Linux environment. Include a Makefile in the directory, or a shellscript `make` that will compile your code. If there are problems with any of the above, contact the lecturer to agree on some other technical solution.

Before starting to do serious computations, please ask (e.g., by contacting the lecturer via Piazza) for an account on one of our workstations in the TCS group where you can run more heavy-duty tasks, or make sure to use other hardware with comparable specifications.