A Theorist's Toolkit **CMU 15-751*, Fall 2016**

PROBLEM SET 8

Due: Noon, Monday Dec 5, email the pdf to toolkit2016homework@gmail.com

---

Homework policy: Exactly the same as last time. Solve any **6** out of 8 problems.

---

1. Given a code $C_{\text{out}} \subseteq \Sigma^n$ over a large alphabet Σ (such as the Reed-Solomon code), and a binary code $C_{\text{in}} \subseteq \{0,1\}^m$ with $|\Sigma|$ codewords and an encoding function $E_{\text{in}} : \Sigma \to \{0,1\}^m$, one can combine them (this operation is called "concatenation," though composition is perhaps a better name) to produce a code $C \subseteq \{0,1\}^{nm}$ as follows:

$$C = \{(E_{\text{in}}(c_1), E_{\text{in}}(c_2), \ldots E_{\text{in}}(c_n)) \mid (c_1, c_2, \ldots, c_n) \in C_{\text{out}}\} .$$

   (a) Prove that the distance of C is at least $d_{\text{out}} \cdot d_{\text{in}}$ where d_{out} and d_{in} are distances of C_{out} and C_{in} respectively.

   (b) Does the distance of C always equal $d_{\text{out}} \cdot d_{\text{in}}$? Justify your answer.

   (c) Show that if $\Sigma = \mathbb{F}_{2^t}$ and C_{out} and C_{in} are both $\mathbb{F}_2$-linear, then C is a binary linear code.

2. Let $\sigma = \langle a_1, a_2, \ldots, a_m \rangle$ be a stream of m distinct items from the universe $\{1, 2 \ldots, n\}$. Assume m is odd and $m < n/2$. The rank of an item a_i is defined to be 1 plus the number of items in σ that are less than a_i. The median of σ is a number of rank $\lceil \frac{m}{2} \rceil$.

   (a) Consider the problem of computing the median of σ exactly in the single pass streaming model. Either prove that any deterministic streaming algorithm that solves this problem must use $\Omega(m \log(n/m))$ bits in the worst case, or give a deterministic streaming algorithm that solves Median exactly using a sub-linear number of bits. If you give an algorithm, you should also prove its correctness and analyze the number of bits of storage it uses.

   (b) Suppose we are allowed to return an element of σ with rank between $(1/2 - \varepsilon)m$ and $(1/2 + \varepsilon)m$ for some small $\varepsilon > 0$, and further we are allowed failure probability $\delta > 0$. Show a single pass randomized streaming algorithm with space $O(\varepsilon^{-2} \log n \log(1/\delta))$ to solve this problem.

   (c) Now suppose we do not know the length m of the stream in advance (but are promised that $m < n/2$). Can you adapt your streaming algorithm from part (b) to work in this setting, using similar amount of space?

3. (a) Suppose we have a randomized single-pass streaming algorithm, which on input a stream consisting of the edges of an undirected graph in arbitrary order, correctly decides whether or not the graph is connected with probability at least $3/4$. Prove that the algorithm must use $\Omega(n)$ space in the worst case, where n is the number of vertices in the graph.

   Hint: There is an easy reduction from one of the communication problems we saw.

(b) Suppose we have a randomized single-pass streaming algorithm, which on input a stream consisting of the edges of an undirected graph in arbitrary order, correctly decides whether or not the graph has a perfect matching with probability at least $3/4$. Prove that the algorithm must use $\Omega(n^2)$ space in the worst case, where n is the number of vertices in the graph.

Hint: Reduction from the INDEX communication problem. If it helps, you might want to first prove an $\Omega(n)$ lower bound.

4. Prove that any t-wise independent set $S \subseteq \{0,1\}^n$ satisfies

$$|S| \geq \sum_{i=0}^{\lfloor \frac{t}{2} \rfloor} \binom{n}{i}. \tag{1}$$

Suggestion: Try and find a set of pairwise orthogonal vectors in $\mathbb{R}^{|S|}$ of cardinality at least the R.H.S. of (1).

5. A binary code is said to be ε-biased if every pair of distinct codewords have relative Hamming distance in the range $[1/2 - \varepsilon, 1/2 + \varepsilon]$ (note such codes have relative distance at least $1/2 - \varepsilon$, but in addition no two codewords differ in more than $(1/2 + \varepsilon)$ fraction of positions). We are interested in a family of ε-biased codes of increasing block length.

Consider the following claim:

Claim: *Let A be an $m \times m$ real symmetric matrix with 1's on the diagonal, and all off-diagonal elements at most ϵ in absolute value. Assume $\epsilon \in (0, 1/4)$ and m is sufficiently large. Then,* $\mathsf{rank}(A) \geq \Omega\left(\frac{\log m}{\epsilon^2 \log(1/\epsilon)}\right)$.

(a) Show why the claim implies that an ε-biased code family can have rate at most $O(\varepsilon^2 \log(1/\varepsilon))$.

(b) Towards proving the above claim, first prove the following fact when the off-diagonal entries are really small:

> Let B be a real symmetric $m \times m$ matrix with 1's on the diagonal. If all off-diagonal entries of B are at most $1/\sqrt{m}$ in absolute value, then $\mathsf{rank}(B) \geq m/2$.

Suggestion: Use the fact that $\mathsf{Trace}(B) = \lambda_1 + \cdots + \lambda_m$ if λ_i are the eigenvalues of B, and relate $\mathsf{Trace}(B^2)$ and $\mathsf{Trace}(B)$ via Cauchy-Schwarz.

(c) Prove the claim above using the fact from Part (b).
Hint: Given A, consider the matrix B who entries are t'th powers of the entries of A for some large t, and argue that $\mathsf{rank}(B)$ is at most $\binom{\mathsf{rank}(A)+t}{t}$.

6. 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* given 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 Reed-Solomon decoder described in class can be adapted 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

$$\prod_{i:E(m)_i \neq r_i} p_i \leq \sqrt{N/K}\,. \tag{2}$$

(Make 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 $\mathbb{Z}_N$ (guarenteed by the Chinese Remainder Theorem) such that $r \mod p_i = r_i$ for every $i = 1, 2, \ldots, n$.

i. Assuming an m satisfying (2) 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 $m = y/z$.

(**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.)

(c) Instead of condition (2) what if we want to decode under the more natural condition for Hamming metric, that is $|\{i : E(m)_i \neq r_i\}| \leq \frac{n-k}{2}$? Show how this can be done by calling the above decoder many times, on the first j symbols for each choice of $k \leq j \leq n$, where for a particular j we try to find an m such that

$$\prod_{\substack{i:i \leq j \\ E(m)_i = r_i}} p_i \geq \sqrt{KN_j}\,. \tag{3}$$

where $N_j = \prod_{i=1}^{j} p_i$.

7. Cloud storage of massive amounts of data has recently motivated the study of codes where every codeword symbol can be recovered from few other codeword symbols, so as to enable recovery from the failure of any single node in a distributed system.

More formally, we are interested in codes that produce an n-symbol codeword from k information symbols and, for any symbol of the codeword, there exist at most ℓ other symbols such that the value of the symbol can be recovered from them. Here we think of $\ell \ll n$.

(i) Prove that the rate of such a code is at most $\frac{\ell}{\ell+1}$.

(ii) Prove that the minimum distance of such a code is at most $n - k - \lceil \frac{k}{\ell} \rceil + 2$. Which coding bound mentioned in class does this generalize?
Suggestion: One approach is to use the fact that if a code $C \subset \Sigma^n$ has distance d, then $n - d$ equals the largest size of a subset $T \subseteq \{1, 2, \ldots, n\}$ such that $|C_T| < |C|$, where $C_T \subset \Sigma^T$ is the code C projected onto coordinates in T.

8. Let $\boldsymbol{X}$ be a random variable. Often we are concerned with *concentration*: showing that $\boldsymbol{X}$ is usually close to its mean. But sometimes we want the opposite, *anticoncentration*: showing that $\boldsymbol{X}$ is somewhat often somewhat far from its mean. We can get a pretty good concentration bound (using Chebyshev's inequality) if we can control $\mathbf{E}[\boldsymbol{X}^2]$; as this problem shows, we can get a pretty good anticoncentration bound if we can control $\mathbf{E}[\boldsymbol{X}^4]$.

   (a) Let $\boldsymbol{Y}$ be a nonnegative random variable (with $0 < \mathbf{E}[\boldsymbol{Y}^2] < \infty$). Show the following generalization of Homework 2 Problem 5(b): for any parameter $0 \le \theta \le 1$,

   $$\mathbf{Pr}[\boldsymbol{Y} > \theta\,\mathbf{E}[\boldsymbol{Y}]] \ge (1-\theta)^2 \frac{\mathbf{E}[\boldsymbol{Y}]^2}{\mathbf{E}[\boldsymbol{Y}^2]}.$$

   (Hint: as with HW2-5(b), start with $\mathbf{E}[\boldsymbol{Y}] = \mathbf{E}[\boldsymbol{Y} \cdot 1_{\{\boldsymbol{Y} \le \theta\,\mathbf{E}[\boldsymbol{Y}]\}}] + \mathbf{E}[\boldsymbol{Y} \cdot 1_{\{\boldsymbol{Y} > \theta\,\mathbf{E}[\boldsymbol{Y}]\}}]$.)

   (b) Let $\boldsymbol{X}$ be any random variable (with $0 < \mathbf{E}[\boldsymbol{X}^4] < \infty$). Assume for simplicity that $\mathbf{E}[\boldsymbol{X}] = 0$. Show the following anticoncentration result, which works better the smaller $\mathbf{E}[\boldsymbol{X}^4]$ is: for any $0 \le t \le 1$,

   $$\mathbf{Pr}[|\boldsymbol{X}| \ge t\sigma] \ge (1-t^2)^2 \frac{\sigma^4}{\mathbf{E}[\boldsymbol{X}^4]}.$$

   (c) Often in algorithmic applications (e.g., spectral graph theory, optimization) you are interested in computing the largest eigenvalue λ of an $n \times n$ PSD matrix A with $m \ge n$ nonzeros. If you are not too greedy about getting it exactly, you can use the "Power Method" to find a number $\lambda' \ge (1-\varepsilon)\lambda$ and a vector ϕ' such that $A\phi' = \lambda'\phi'$ in $O(m \cdot \log n \cdot (1/\varepsilon))$ time. Very briefly, the "Power Method" first picks a vector $\boldsymbol{x} \in \{-1,1\}^n$ uniformly at random, and then it outputs $A^k\boldsymbol{x}$ for $k = O(\frac{\log n}{\varepsilon})$. We won't analyze the whole algorithm here, but just a part of it.

   The only reason $\boldsymbol{x}$ is chosen at *random* is that the analysis requires the following: whatever the (unit-length) eigenvector ϕ associated to λ is, there is a good chance that $|\langle \boldsymbol{x}, \phi \rangle| \ge 1/2$. Prove this property: If $u \in \mathbb{R}^n$ is any unit vector, $\boldsymbol{x} \sim \{-1,1\}^n$ is chosen uniformly at random, and $\boldsymbol{S} = u_1\boldsymbol{x}_1 + \cdots + u_n\boldsymbol{x}_n$, then $\mathbf{Pr}[|\boldsymbol{S}| \ge \frac{1}{2}] \ge \frac{3}{16}$.

   (d) Describe a polynomial-time *deterministic* algorithm that (without knowing u), outputs a bunch of strings x with the guarantee that at least $\frac{3}{16}$ of them have $|\langle u, x \rangle| \ge \frac{1}{2}$.