Homework 2 – Due Thursday, February 6 *before* 11am on Gradescope

Instructions

- Solutions written in LaTeX are strongly preferred, but you can upload any pdf files, including scanned hand-written solutions. Template latex files are on the course webpage.
- Collaboration is allowed and encouraged. However, each of you should think about a problem before discussing it with others and write up your solution independently. You may consult books and online sources to get information about well-known theorems, such as the Chernoff bound. But you are not allowed to look up solutions directly in papers or any other sources. And you *must* list all collaborators and sources! (See full details in the General Information Handout.)
- Correctness, clarity, and succinctness of the solution will determine your score.

Problems

1. **(Short questions)** This is a collection of questions with short answers (at most several sentences per question).

   (a) Recall that the relative Hamming distance between two strings is the fraction of character positions on which they differ. Give an algorithm for estimating the relative Hamming distance between two strings of the same length within additive error ϵ. Your algorithm should give a good estimate with probability at least 2/3. What's the running time of your algorithm?

   (b) We define the *property* as a set $\mathcal{P}$ of objects (intuitively, the collection of objects that satisfy the property). For example, it can be the set of monotone functions of the form $f : \{0,1\}^d \to \{0,1\}$. Recall that an ϵ-tester for $\mathcal{P}$ has to, with probability at least 2/3,

   - accept objects in $\mathcal{P}$;
   - reject objects that are ϵ-far from $\mathcal{P}$.

   Consider properties $\mathcal{P}_1$ and $\mathcal{P}_2$ such that $\mathcal{P}_1 \subseteq \mathcal{P}_2$. Let $q(\epsilon)$ be some function that represents query complexity. E.g., $q(\epsilon)$ could be $1/\epsilon$ or $1/\epsilon^2$.

   Prove or disprove:

   i. If $\mathcal{P}_1$ has an ϵ-tester that makes $O(q(\epsilon))$ queries then so does $\mathcal{P}_2$.
   ii. If $\mathcal{P}_2$ has an ϵ-tester that makes $O(q(\epsilon))$ queries then so does $\mathcal{P}_1$.
   iii. If $\mathcal{P}_1$ has an ϵ-tester that makes $O(q(\epsilon))$ queries then so does $\overline{\mathcal{P}_1}$.

2. **(Testing connectedness)** In class, we saw a tester for connectedness that made $O(\frac{1}{\epsilon^2 d})$ queries. Give a tester for connectedness that makes $O(\frac{1}{\epsilon}\text{polylog}\frac{1}{\epsilon d})$ queries, where polylog m means that there is a constant c such that the expression is $\log^c m$.

   Hint: In class, we proved that if a graph is ϵ-far from connected, it has many small connected components. ("Many" was $\geq \frac{\epsilon dn}{4}$ and "small" was of size $\leq \frac{4}{\epsilon d}$.) Try to do a more careful accounting by considering small components of different sizes separately. I.e., break components into *buckets* according to their size (1, 2 to 3, 4 to 7, etc.) and prove that at least one of the buckets contains "enough" components. Modify the test accordingly.

3. **(A lower bound for unateness, 20 points).** A Boolean function $f : \{0,1\}^d \to \{0,1\}$ is *unate* if, along each coordinate, the function is either nondecreasing or nonincreasing. For example, the XOR of bits x_1 and x_2 is not unate, while the function $x_1 \cdot \overline{x_2}$ is unate. The goal of this problem is to prove that every nonadaptive, 1-sided error unateness tester for functions $f : \{0,1\}^d \to \{0,1\}$ with the distance parameter $\epsilon \le \frac{1}{8}$ must make $\Omega(\frac{d}{\log d})$ queries.

   (a) We will use Yao's principle. The hard distribution $\mathcal{D}$ is defined as follows: pick three different dimensions $a, b, c \in [d]$ uniformly at random and return $f_{a,b,c}(x) = x_a \cdot x_b + (1 - x_a) \cdot x_c$. We call a, b, c the *influential dimensions,* since the value of the function depends only on them. The coordinate x_a determines if $f_{a,b,c}(x)$ should be set to x_b or x_c. If $x_a = 1$, then $f_{a,b,c}(x) = x_b$; otherwise, $f_{a,b,c}(x) = x_c$.

   Prove that every function $f_{a,b,c}$ in the support of $\mathcal{D}$ is $\frac{1}{8}$-far from unate.

   (b) Fix a nonadaptive, 1-sided error tester, and let Q be the set of queries it makes. Let $f_{|Q}$ denote the restriction of the function f to the points in Q. We say that $f_{|Q}$ is *extendable* to a unate function if there exists a unate function g such that $g_{|Q} = f_{|Q}$. Argue (in 1 sentence) that the tester should accept if $f_{|Q}$ is extendable to a unate function.

   (c) Next, we define a conjunctive normal form (CNF) Boolean formula $\phi(f_{|Q})$. Intuitively, each pair (x, y) of domain points on which f differs imposes a constraint on f (assuming f is unate). Specifically, at least one of the dimensions on which x and y differ must be consistent (i.e., nondecreasing or nonincreasing) with the change of the function value between x and y. This constraint is formalized in the definition of $\phi(f_{|Q})$ as follows. For each dimension i, we have a variable z_i which is true if f is nondecreasing along the dimension i, and false if it is nonincreasing along that dimension. For each $x, y \in Q$ such that $f(x) = 1$ and $f(y) = 0$, create a clause (think of x, y as sets where $i \in x$ iff $x_i = 1$):

   $$c_{x,y} = \bigvee_{i \in x \setminus y} z_i \vee \bigvee_{i \in y \setminus x} \overline{z_i}.$$

   Set $\phi(f_{|Q}) = \bigwedge_{x,y \in Q : f(x)=1, f(y)=0} c_{x,y}$. Briefly argue that $f_{|Q}$ is extendable to a unate function iff $\phi(f_{|Q})$ is satisfiable.

   (d) The *width* of a clause is the number of literals in it; the *width* of a CNF formula is the minimum width of a clause in it. Use probabilistic method to argue that every CNF formula that has width at least $3 \log d$ and at most d^2 clauses is satisfiable.[1]

   (e) Now suppose $|Q| \le \frac{d}{30 \log d}$. Prove that with probability at least $\frac{2}{3}$ over $f \sim \mathcal{D}$, the width of $\phi(f_{|Q})$ is at least $3 \log d$.

   Hint: Let $x \Delta y$ denote the symmetric difference between x and y. Consider a graph with vertex set Q and edges corresponding to $x, y \in Q$ such that $|x \Delta y| \le 3 \log d$.

   Hint 2: Spanning forest.

   (f) Put it all together to prove the desired lower bound.

[1] We use log to denote $\log_2$.