Here is the official list of COMS 6998-3 problems. If you haven’t done so yet, please turn in a
detailed solution of one of the problems below to Rocco and Clement (via email) by the day of our last
lecture, Wed April 30.

Your writeup should be clear and self-contained, giving precise definitions for any terms that are
not clearly defined in the statement of the problem; part of the point of the exercise is to go from a
perhaps somewhat vague or underspecified statement to a clear statement (and correct proof).

You are welcome to discuss the HW problems with other students as long as the
discussions are “high level” ones – you should not write anything down in the course of
your discussions, and everything you turn in must be entirely your own work. Consulting
outside sources is not allowed. Please contact me or Clement if you have any questions
about this policy.

Problem 1. We showed in class that there is an $O\left(\frac{1}{\epsilon} \log N\right)$-query algorithm that can test whether
any list $\bar{a} = (a_1, \ldots, a_N)$ of integers (to which it has random access) is sorted versus $\epsilon$-far from sorted.

Show that there is an $O\left(\frac{1}{\epsilon}\right)$-query algorithm that can test whether any list $\bar{a} = (a_1, \ldots, a_N)$ of bits – so now $\bar{a}$ is guaranteed to lie in $\{0, 1\}^n$ – is sorted versus $\epsilon$-far from sorted.

Problem 2. Give a detailed proof of the following lemma which was stated in class: “There is an
algorithm which, given access to an MQ($f$) oracle for $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$, a value $0 \leq k \leq n$, a set
$S_1 \subseteq \{1, \ldots, k\}$, and values $\tau, \delta > 0$, with probability at least $1 - \delta$ outputs a value $v$ such that

$$|v - \mathbb{E}[(f_{k,S_1})^2]| \leq \tau$$

and runs in time poly($n, 1/\tau, \log(1/\delta)$).” You may use any results we proved in class along with standard
Chernoff/Hoeffding bounds.

Problem 3. Give a detailed proof of the performance guarantee that we stated in class for the KM
algorithm: briefly, that given a parameter $\theta > 0$ and MQ($f$) access, it outputs a collection $S$ such that

- if $S \in S$ then $|\hat{f}(S)| \geq \theta/2$, and
- if $S \notin S$ then $|\hat{f}(S)| \leq 2\theta$.

You may use any results proved in class, standard Chernoff/Hoeffding bounds, and the result of Problem
2 above.

Problem 4. Show that there is an algorithm which, given MQ($f$) access to $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$
and $\epsilon, \delta > 0$, with probability $1 - \delta$ outputs an $h : \{-1, 1\}^n \rightarrow \{-1, 1\}$ such that $\text{dist}(f, h) \leq \epsilon$ and
runs in time \(\text{poly}(n, L_1(f), 1/\epsilon, \log(1/\delta))\). Note that the algorithm is not given the value \(L = L_1(f)\) or an upper bound on this value.

**Problem 5.** Let \(D_1\) and \(D_2\) be two probability distributions over a finite set \(\Omega\). Recall that the total variation distance between \(D_1\) and \(D_2\) is

\[
d_{TV}(D_1, D_2) = \frac{1}{2} \sum_{\omega \in \Omega} |D_1(\omega) - D_2(\omega)| = \max_{S \subseteq \Omega} (D_1(S) - D_2(S)),
\]

where \(D(S)\) denotes \(\sum_{\omega \in S} D(\omega)\).

Let \(A\) be any algorithm (deterministic or randomized) which, on input an element \(\omega \in \Omega\), either outputs “accept” or “reject.” Prove that

\[
\Pr_{\omega \in D_1} [A(\omega) \text{ outputs “accept”}] - \Pr_{\omega \in D_2} [A(\omega) \text{ outputs “accept”}] \leq d_{TV}(D_1, D_2).
\]

**Problem 6.** Prove the slightly more relaxed version of Yao’s principle (yielding lower bounds for nonadaptive property testing) mentioned in class, that allows both the \(D_Y\) and \(D_N\) distributions to be almost entirely supported on functions of the correct types – i.e. a random \(f\) drawn from \(D_Y\) may now fail to have property \(P\) with probability at most (say) 1/100, and likewise a random \(f\) drawn from \(D_N\) may be \(\epsilon\)-close to \(P\) with probability up to 1/100. (Give a clear and precise statement of what you are proving before you prove it.) You may use the statement of Yao’s principle proved in class in your solution.

**Problem 7.** Let \(m = n/10\) and let \(f : \{-1, 1\}^n \to \{-1, 1\}\) be defined as

\[
f(x) = \text{sign} \left( -(x_1 + \cdots + x_m) + \frac{7}{3} (x_{m+1} + \cdots + x_n) \right).
\]

Show that for \(i \in [m]\) we have \(\inf_i(f) = \Omega(1/\sqrt{n})\). You may freely use any standard properties of binomial coefficients, Stirling’s approximation, etc. in your solution.

**Problem 8.** Show that the “Main Lemma” about \(k\)-part juntas from class implies that the algorithm \textbf{Junta-Test} from class uses \(O(k/\epsilon + k \log k)\) queries and is a 1-sided \(\epsilon\)-tester for whether \(f\) is a \(k\)-junta. (Recall the “Main Lemma” says that if \(I\) is a random partition of \([n]\) into \(s = \text{poly}(k, 1/\epsilon)\) parts and \(\text{dist}(f, J_k) > \epsilon\), then with probability at least 5/6 over \(I\), the function \(f\) \(\epsilon/2\)-violates being a \(k\)-part junta with respect to \(I\).)

**Problem 9.** Recall from our analysis of the junta test in class that the random variable \(X_j\), for \(j \in I\), is defined to be \(\inf_{\ell}^{\leq 2k}(j)\) if \(j \in I_1 \setminus H\) and 0 otherwise, where \(H\) is the set of variables \(i\) that have \(\inf_{\ell}^{\leq 2k}(i) > \theta\) and \(\theta = \epsilon^2 \log(k/\epsilon)/(10^9k^4)\). The randomness here is over the random assignment of variables in \([n]\) to elements \(I_1, \ldots, I_s\) of the partition \(I\). Show that

\[
\mathbb{E}_I \left[ \sum_{j \in [n]} X_j \right] \leq \frac{\epsilon}{8k}
\]

and that

\[
\Pr_I \left[ \sum_j X_j \geq \mathbb{E} \left[ \sum_{j \in [n]} X_j \right] + \frac{\epsilon}{8k} \right] \leq \frac{1}{18s}.
\]
It is a good idea to use the following Chernoff bound: for independent random variables $Z_1, \ldots, Z_n$ where $Z_i$ always takes values in $[a_i, b_i]$, we have

$$
\Pr \left[ \sum_{j=1}^{n} Z_i > \mathbf{E} \left[ \sum_{i} Z_i \right] + t \right] \leq \exp \left( -\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).
$$

**Problem 10.** Let $f$ be a random Boolean function over variables $x_1, \ldots, x_{k+1}$. Show that for $k > C$ where $C$ is some universal constant, we have that $f$ is $0.49$-far from every $k$-junta with probability at least $0.999$. 