

Homework 4

Lecturer: Ronitt Rubinfeld

Due Date: January 18, 2010

Turn in your solution to each problem on a separate sheet of paper, with your name on each one.

1. This problem concerns testing closeness to a distribution that is entirely known to the algorithm. In the following, assume that p and q are distributions over D . The algorithm is given access to samples of p , and knows an exact description of the distribution q in advance – the query complexity of the algorithm is only the number of samples from p . Assume that $|D| = n$.

- (a) (warmup) Show that $\|p\|_2^2 - \|U_D\|_2^2 = \|p - U_D\|_2^2$.
- (b) Define $\text{Bucket}(q, D, \epsilon)$ as a partition $\{D_0, D_1, \dots, D_k\}$ of D with $k = (2/\log(1 + \epsilon)) \cdot \log |D|$, $D_0 = \{i \mid q(i) < 1/(|D| \log |D|)\}$, and for all i in $[k]$,

$$D_i = \left\{ j \in D \mid \frac{(1 + \epsilon)^{i-1}}{|D| \log |D|} \leq q(j) < \frac{(1 + \epsilon)^i}{|D| \log |D|} \right\}.$$

Show that if one considers the restriction of q to any of the buckets D_i , then the distribution is close to uniform: i.e., Show that if q is a distribution over D and $\{D_0, \dots, D_k\} = \text{Bucket}(q, D, \epsilon)$, then for $i \in [k]$ we have $|q|_{D_i} - U_{D_i}|_1 \leq \epsilon$, $\|q|_{D_i} - U_{D_i}\|_2^2 \leq \epsilon^2/|D_i|$, and $q(D_0) \leq 1/\log |D|$.

Hint: it may be helpful to remember that $1/(1 + \epsilon) > 1 - \epsilon$.

- (c) Let $(D_0, \dots, D_k) = \text{Bucket}(q, [n], \epsilon)$. For each i in $[k]$, if $\|p|_{D_i}\|_2^2 \leq (1 + \epsilon^2)/|D_i|$ then $|p|_{D_i} - U_{D_i}|_1 \leq \epsilon$ and $|p|_{D_i} - q|_{D_i}|_1 \leq 2\epsilon$.
- (d) Show that for any fixed q , there is an $\tilde{O}(\sqrt{n} \text{poly}(1/\epsilon))$ query algorithm \mathcal{A} with the following behavior:
Given access to samples of a distribution p over domain D , and an error parameter ϵ ,
- if $p = q$, then \mathcal{A} outputs “pass” with probability at least $2/3$.
 - if $|p - q|_1 > \epsilon$, then \mathcal{A} outputs “fail” with probability at least $2/3$.
- (e) (Don't turn in) Note that the last problem part generalizes uniformity testing. As a sanity check, what does the algorithm do in the case that $q = U_D$? Also, it is open whether the time complexity of the algorithm can also be made to be $\tilde{O}(\sqrt{n} \text{poly}(1/\epsilon))$ (assume that q is given as an array, in which accessing q_i requires one time step).

2. The goal of this problem is to carefully prove a lower bound on testing whether a distribution is uniform.

- (a) For a distribution p over $[n]$ and a permutation π on $[n]$, define $\pi(p)$ to be the distribution such that for all i , $\pi(p)_{\pi(i)} = p_i$.

Let \mathcal{A} be an algorithm that takes samples from a black-box distribution over $[n]$ as input. We say that \mathcal{A} is *symmetric* if, once the distribution is fixed, the output distribution of \mathcal{A} is identical for any permutation of the distribution.

Show the following: Let \mathcal{A} be an arbitrary testing algorithm for uniformity. Suppose \mathcal{A} has sample complexity at most $s(n)$, where n is the domain size of the distributions. Then, there exists a symmetric algorithm that tests uniformity with sample complexity at most $s(n)$.

- (b) Define a *fingerprint* of a sample as follows: Let S be a multiset of at most s samples taken from a distribution p over $[n]$. Let the random variable C_i , for $0 \leq i \leq s$, denote the number of elements that appear exactly i times in S . The collection of values that the random variables $\{C_i\}_{0 \leq i \leq s}$ take is called the *fingerprint* of the sample.

For example, let $D = \{1..7\}$ and the sample set be $S = \{5, 7, 3, 3, 4\}$. Then, $C_0 = 3$ (elements 1, 2 and 6), $C_1 = 3$ (elements 4, 5 and 7), $C_2 = 1$ (element 3), and $C_i = 0$ for all $i > 2$.

Show the following: If there exists a symmetric algorithm \mathcal{A} for testing uniformity, then there exist an algorithm for testing uniformity that gets as input only the fingerprint of the sample that \mathcal{A} takes.

- (c) Show that any algorithm making $o(\sqrt{n})$ queries cannot have the following behavior when given error parameter ϵ and access to samples of a distribution p over a domain D of size n :
- if $p = U_D$, then \mathcal{A} outputs “pass” with probability at least $2/3$.
 - if $|p - U_D|_1 > \epsilon$, then \mathcal{A} outputs “fail” with probability at least $2/3$
3. Suppose an algorithm has the following behavior when given error parameter ϵ and access to samples of a distribution p over a domain $D = \{1, \dots, n\}$:
- if p is monotone, then \mathcal{A} outputs “pass” with probability at least $2/3$.
 - if for all monotone distributions q over D , $|p - q|_1 > \epsilon$, then \mathcal{A} outputs “fail” with probability at least $2/3$

Show that this algorithm must make $\Omega(\sqrt{n})$ queries.

Hint: Reduce from the problem of testing uniformity.