Sub-linear Algorithms April 20, 2017

Homework 8

Lecturer: Ronitt Rubinfeld *Due Date: May 2, 2017*

Homework guidelines: You may work with other students, as long as (1) they have not yet solved the problem, (2) you write down the names of all other students with which you discussed the problem, and (3) you write up the solution on your own. No points will be deducted, no matter how many people you talk to, as long as you are honest. It's ok to look up famous sums and inequalities that help you to solve the problem, but don't look up an entire solution.

1. (Property testing of the clusterability of a set of points.) Given a set X of points in any metric space. Assume that one can compute the distance between any pair of points in one step. Say that X is (k, b)-diameter clusterable if X can be partitioned into k subsets (clusters) such that the maximum distance between any pair of points in a cluster is b. Say that X is ϵ-far from (k, b)-diameter clusterable if at least $\epsilon|X|$ points must be deleted from X in order to make it (k, b)-diameter clusterable.

   Show how to distinguish the case when X is (k, b)-diameter clusterable from the case when X is ϵ-far from $(k, 2b)$-diameter clusterable. Your algorithm should use polynomial in $k, 1/\epsilon$ queries. It is possible to get an algorithm which uses $O((k^2 \log k)/\epsilon)$ queries.

2. In the Run Length Encoding (RLE) compression scheme, the data is encoded as follows: each run, or a sequence of consecutive occurrences of the same character, is stored as a pair containing the character in the first location and the length of the run in the second location.[1] For example, the string 11111101000 would be stored as $(1, 6)(0, 1)(1, 1)(0, 3)$. The *cost of the run-length encoding*, denoted by $C(w)$, is the sum over all runs of $\log(\ell + 1) + \log|\Sigma|$ (where ℓ is the run-length).

   Assume that the alphabet characters are all in the set $\{0, 1\}$, i.e., that the alphabet Σ is of size 2.

   (a) Give an algorithm that, given a parameter ϵ, outputs an ϵn-additive estimate[2] to $C(w)$ with high probability and makes $\text{poly}(1/\epsilon, \log n)$ queries.

   (b) Show that there is a distribution on inputs such any that any deterministic approximation algorithm for $C(w)$ making an expected number of queries that is $o(\frac{n}{A^2 \log n})$ must fail to output an A-multiplicative approximation with probability at least 1/3. (Here the expectation in the number of queries is over the choice of an input from the distribution). (It's also ok to give a lower bound for deterministic algorithms by showing that for each algorithm there is an input that causes it to fail).

[1] Run-length encoding is used to compress black and white images, faxes, and other simple graphic images, such as icons and line drawings, which usually contain many long runs.

[2] For a function f, algorithm A outputs an ϵn-additive estimate if on any input x, $f(x) - \epsilon n \leq A(x) \leq f(x) + \epsilon n$.