EXAM

ADS - RKMVERI - June 2021

1 Instructions

1. The final exam due date will be announced by Swati Prabhu MJ.
2. Please write clearly.
3. Cite all the resources that you have used.
4. Please **DO NOT** collaborate with anybody. Treat this as a take-home exam.
5. If a question/problem isn't clear, please make assumptions, state them and solve the problem accordingly.
6. Solve as many problems as you can.
7. Most of the exercises are from my Notes on Algorithm Design `https://people.scs.carleton.ca/~maheshwa/Notes/DAA/notes.pdf`. Also Corollary, Observations, and Lemma numbers are with reference to my notes.

2 Problems

1. Consider the following bipartite graph $G = (V = L \cup R, E)$ where $L = \{l_1, \ldots, l_n\}$, $R = \{r_1, \ldots, r_n\}$, and $E = \{(l_i, r_i) | 1 \leq i \leq n\} \cup \{(l_i, r_j) | \frac{n}{2} + 1 \leq i \leq n \text{ and } 1 \leq j \leq \frac{n}{2}\}$. Assume that the vertices in L are known in advance and the vertices in R come in increasing order of their indices. The online algorithm (called GREEDY RANDOM) matches the next vertex $r_j \in R$ to any of its unmatched neighbors in L (if there is any) uniformly at random. Show that the expected size of the matching computed by GREEDY RANDOM is $\frac{n}{2} + \log n$. (Hint: For $1 \leq j \leq \frac{n}{2}$, show that with probability at most $\frac{1}{\frac{n}{2}-j+1}$ the vertex r_j will be matched to l_j.)
2. Suppose there are only two possible actions $\{\uparrow, \downarrow\}$ of Dow Jones Index (DJI) at the end of each day. Answer the following questions for the different scenarios.

(a) Each morning the algorithm chooses the actions based on some smart scheme. If the algorithm chooses $\uparrow$ with probability $\geq \frac{1}{2}$, the adversary assigns the reward of -1 for choosing the action $\uparrow$ and a reward of $+1$ for choosing $\downarrow$. If the algorithm chooses $\downarrow$ with probability $\geq \frac{1}{2}$, the adversary assigns a reward of $+1$ to the action $\uparrow$ and a reward of -1 to $\downarrow$. Over a run of T days, show that the expected reward of the algorithm is at most 0. How does this compares with the reward of the adversary if it somehow choose an optimal action for each day? (Remark: This exercise shows that the algorithm (even a randomized scheme) has no match for the adversary that chooses an optimal action on each day.)

(b) Each morning the algorithm chooses one of the two actions by following some deterministic strategy. If the choice of our action for that day matches DJI we get a reward of $+1$, otherwise we get a reward of 0. Show that in a run of T days, an adversary can design the outcomes for each day in such a way that the reward that our algorithm gets is 0, whereas there is a fixed action if chosen for all the days will generate a revenue of at least $\frac{T}{2}$. If there are n actions, show that the algorithms revenue can be zero, whereas there is a fixed action that can generate a revenue of at least $T(1 - \frac{1}{n})$. Conclude that no deterministic algorithm can ensure a positive reward. (Note that the problem is that we don't know which action will generate that kind of revenue till we have observed the behaviour of all the actions for T days.)

3. The Section number refers to my notes. In Subsection 11.5.4, m_i^t's were the losses of experts on day t. They can take any values in the interval $[-1, 1]$. Instead of thinking of m_i^t's as the loss of expert i on day t, assume that it is the gain of the expert. In that section we wanted to establish that our online strategy doesn't incur significantly more loss than the best expert. Show what changes you need to make in the multiplicative weight update method if m_i^t's are gains. Show that the expected gain of the algorithm is $\sum_{t=1}^{T} M^t \geq \sum_{t=1}^{T} m_i^t - \frac{\ln n}{\eta} - \eta \sum_{t=1}^{T} |m_i^t|$, where $\sum_{t=1}^{T} m_i^t$ is the gain of the best expert.
(Hint: Can we think of the loss vector as $-m^t$ and use the same algorithm as in Subsection 11.5.4?)

4. Assume that we have a stream consisting of numbers from the set $\{-1, 0, +1\}$ and we are interested in maintaining the sum of last N bits of the stream. In this exercise we will show that it will require $\Omega(N)$ bits to maintain an approximate sum that is within a constant factor of the exact sum. Suppose we have an algorithm $\mathcal{A}$ that maintains the approximate sum. Assume that we have a bit string consisting of $\frac{N}{2}$-bits composed of 0s and 1s. We replace each 0-bit by a pair of bits $(1, -1)$ and each 1-bit by the pair $(-1, 1)$. Now this sequence of N-bits is presented to our algorithm $\mathcal{A}$ that maintains the approximate sum within a constant factor. Note that the exact sum of these N-bits is 0. In addition to these N bits, the next set of N bits that will be received in the stream are only 0-bits. Answer the following:

(a) Show that if the next bit (i.e. the $(N+1)$-st bit) in the stream is 0, the output to the sum query on receiving this bit will be $+1$ (respectively -1) if and only if the 1st bit in the stream was a 1 (respectively, -1).

(b) For a positive integer $i < \frac{N}{2}$, show that after receiving the $(N+2i-1)$-th 0 bit, the output to the sum query will be $+1$ (respectively -1) if and only if the i-th bit in the stream was a 1 (respectively, -1).

(c) Show that after receiving the $2N$-th 0 bit, we would have completely recovered the first N-bits of the stream.

(d) Conclude that to estimate the approximate sum within a constant factor in a sliding window of size N in a stream of (positive and negative) numbers we need to store $\Theta(n)$ bits.

5. This problem is about the power of medians of means. Assume that we want to compute a value $\mathcal{X}$ using a randomized algorithm. In the analysis of our algorithm we use a random variable X that estimates $\mathcal{X}$, i.e. $E[X] = \mathcal{X}$. To have a good estimation, we take $k \times s$ independent random variables that have identical distribution as that of X, where $s = O(\log \frac{1}{\epsilon})$ and $k = \frac{cVar[X]}{\gamma^2 E[X]^2}$ for some positive constants c, γ, and ϵ. We denote them by $\{X_{11}, \ldots, X_{1k}, X_{21}, \ldots, X_{2k}, \ldots, X_{s1}, \ldots, X_{sk}\}$. Define $Y_i = \frac{1}{k}\sum_{j=1}^{k} X_{ij}$, $1 \leq i \leq s$, and Z as the median value of $\{Y_1, \ldots, Y_s\}$. Show the following.

   (a) For $i \in \{1, \ldots, s\}$, $E[Y_i] = \mathcal{X}$.

   (b) $E[Z] = \mathcal{X}$.

   (c) $Var[Y_i] = \frac{1}{k} Var[X]$.

   (d) Using Chebyshev's inequality show that $Pr(|Y_i - \mathcal{X}| \geq \gamma\mathcal{X}) \leq \frac{1}{c}$.

   (e) Using the ideas from Observation 10.3.7 (see my notes) and the Chernoff bounds, show that $Pr(|Z - \mathcal{X}| \geq \gamma\mathcal{X}) \leq \epsilon$.

6. Let $S = \{x_1, \ldots, x_n\}$ be a set of n distinct numbers. We are interested in finding an approximate median element of S. Define the rank of an element $y \in S$ as the number of elements in S that are $\leq y$, i.e. $rank(y) = |\{x \in S | x \leq y\}|$. An element $y \in S$ is an approximate median of S, if $\frac{n}{2} - \epsilon n \leq rank(y) \leq \frac{n}{2} + \epsilon n$ for some $\epsilon \leq \frac{1}{6}$. We employ the following strategy to find an approximate median element. We sample s elements from S, each independently and uniformly at random with replacement. Let $S' \subset S$ be the set of sampled elements. We set y to be the median of the sampled elements. Define the three subsets of S as follows.

$$L = \{x \in S : rank(x) < \frac{n}{2} - \epsilon n\}$$
$$U = \{x \in S : rank(x) > \frac{n}{2} + \epsilon n\}$$
$$M = \{x \in S : \frac{n}{2} - \epsilon n \leq rank(x) \leq \frac{n}{2} + \epsilon n\}$$

Answer the following.

(a) Show that the probability that a sampled element is from the set L is $\frac{1}{2} - \epsilon$.

(b) Let $X = |L \cap S'|$. Show that $E[X] = (\frac{1}{2} - \epsilon)s$.

(c) Show that if $|L \cap S'| > \frac{s}{2}$, then y is not an approximate median. Same holds if $|R \cap S'| > \frac{s}{2}$.

(d) Show that $Pr(X > \frac{s}{2}) \leq Pr(X \geq (1+\epsilon)E[X])$.

(e) Using Chernoff bounds and by setting $s = \frac{9}{\epsilon^2} \log \frac{2}{\delta}$ show that $Pr(X \geq (1+\epsilon)E[X]) \leq exp(-\frac{\epsilon^2}{3}E[X]) \leq \frac{\delta}{2}$.

(f) Show that if $|L \cap S'| \leq \frac{s}{2}$ and $|R \cap S'| \leq \frac{s}{2}$, then y is an approximate median.

(g) Show that if we draw $s = \frac{9}{\epsilon^2} \log \frac{2}{\delta}$ samples, $Pr(\frac{n}{2} - \epsilon n \leq rank(y) \leq \frac{n}{2} + \epsilon n)) \geq 1 - \delta$.

(h) How many samples we need to draw if $\epsilon = 0.1$ and we want to succeed with probability at least 3/4?

7. For the locality-sensitive hashing technique with respect to signatures of sets, we partitioned the signature matrix in b bands, each band consisting of r rows, and analyzed that the probability that the two sets with Jaccard similarity of s, will be reported similar with probability $f(s) = 1 - (1 - s^r)^b$, using the so called AND-OR construction. This analysis was based on estimating the probability that signatures for the two sets should match in all rows (constituting the AND-family) of at least one of the bands (the OR-family). Suppose, we alter our strategy, and use OR-AND construction. To be more precise, we have the same partitioning in terms of b bands and r rows, but now we say that the signatures match in a band, if they match in at least one of the rows in that band, but we declare the two sets to be similar if their signatures match in all the bands. Estimate what will be the probability that the two sets are reported similar whose Jaccard similarity is s using the OR-AND strategy. Call this estimate $f'(s)$. Furthermore, compare the two estimates, $f(s)$ and $f'(s)$, for various values of s, (you may fix $b = 20$ and $r = 5$ or to any other values).

8. Assume we have a set $P = \{p_1, \ldots, p_n\}$ of n distinct points in plane, where $p_i = (x(p_i), x(y_i))$ and $x(p_i)$ and $y(p_i)$ refers to x and y coordinates of p_i, respectively. For any point z in plane, define the function $\Delta(z, P)$ as

$$\Delta(z, P) \quad = \quad \sum_{i=1}^{n} ||p_i - z||_2^2 \tag{1}$$

Note that $||p_i - z||_2^2$ refers to the square of the Euclidean distance between p_i and z. Answer the following

(a) Let the point $z^* = (\frac{\sum_{i=1}^{n} x(p_i)}{n}, \frac{\sum_{i=1}^{n} y(p_i)}{n})$. Show that for any arbitrary point z in plane, $\Delta(z, P) = \Delta(z^*, P) + n||z - z^*||_2^2$

(b) Show that among all points in the plane, the point z^* minimizes Equation 1

(c) Choose a point uniformly at random from P. Let the chosen point be $p \in P$. Show that $E[\Delta(p, P)] \leq 2\Delta(z^*, P)$

9. Show that any metric space $\langle X, d \rangle$ on n-points, can be embedded in $O(\log^2 n)$-dimensional space with a distortion factor of $O(\log^2 n)$, where the distances are measured with respect to L_1-metric (i.e. the Manhattan metric). See Corollary 12.4.10. You need to provide missing details in the proof.

10. How can you approximate Euclidean Minimum Spanning Tree for a set of n points P in $\mathbb{R}^d$ using the ideas from the Locality-Sensitive ordering paper/lecture? EMST is the minimum spanning tree of the complete graph on P. Weight of each edge $e = (uv)$ is the Euclidean distance between the points u and v of P. What is the approximation factor? What is approximately the running time?

11. Consider the utility matrix $M = \begin{bmatrix} 4 & 0 & 0 \\ 5 & 1 & 0 \\ 0 & 1 & 4 \\ 0 & 0 & 5 \end{bmatrix}$

It represents 4 users as rows, 3 items as columns, and each entry is the item's ranking by a user. Answer the following questions:

(a) Provide a best rank 2 approximation of M using the Singular-Value Decomposition. (You may use some package to compute SVD). Let M' represents the rank-2 approximation of M.

(b) Compute the Loss in Energy when we approximate M by M'

(c) Use M' to map all the users to the concept space

(d) For the following users $q_1 = [3, 0, 0]$, $q_2 = [0, 3, 0]$, and $q_3 = [0, 0, 3]$, what are the items you will recommend? Provide some justification for your choice. (Note $q_1 = [3, 0, 0]$ refers to that the user q_1 gives the rank of 3 to Item 1, but has not given any rankings to Items 2 and 3.)