Randomized Algorithms 2025A – Final (Take-Home Exam)

Robert Krauthgamer and Moni Naor

February 18, 2025
Due within 72 hours

General instructions. The exam has 2 parts.

Policy: You may consult textbooks and the class material (lecture notes and homework), but no other sources (like web search). You should work on these problems and write up the solutions by yourself with no help from others.

You may use without proof theorems stated in class, provided you state the appropriate theorem that you are using. As usual, assume n (or $|V|$) is large enough.

Part I (25 points)

Answer 2 of the following 3 questions. Give short answers, that sketch the proof or provide a convincing justification in 2-5 sentences, even for true/false questions.

A. Suppose $S \subset \mathbb{R}^d$ is an ε-coreset for $X \subset \mathbb{R}^d$, and similarly S' for X'. Is it true that the union of the coresets (i.e., $S \cup S'$), is an ε-coreset for the union of the data sets (i.e., $X \cup X'$)?

We refer here to strong coresets for the k-median problem, and union is as multisets (equivalent to summing the weight of each point).

B. Is there a Las Vegas version (i.e., randomized with error probability zero) of the JL Lemma? This would mean that for every d, ε, n there is a randomized matrix $L \in \mathbb{R}^{k \times d}$ for $k = \text{poly}(\varepsilon^{-1} \log n)$, such that for all $x_1, \ldots, x_n \in \mathbb{R}^d$, with probability 1,

$$\forall i, j \in [n], \qquad \|Lx_i - Lx_j\|_2 \in (1 \pm \varepsilon)\|x_i - x_j\|_2.$$

C. You want to prevent a user from visiting a collection of suspicious URLs, but this should all be done on the user's browser, without further consultation. The list is of size n and the URLs may be long, but you are allowing innocent URLs to be censored, provided that the probability of censorship is no more than δ. Suggest a succinct data structure that would not involve downloading too many bit for doing it. What is the best result you can hope for?

Part II (75 points)

Answer 3 of the following 4 questions.

1. Given $x \in \{\pm 1\}^n$ that is viewed as a 2-coloring of $[n]$, together with a set system $S_1, \ldots, S_m \subseteq [n]$, we define the *imbalance of* S_i as $b_i(x) := |\sum_{j \in S_i} x_j|$, and the *imbalance of the set system* as $b^*(x) := \max_{i \in [m]} b_i(x)$.

Design an algorithm that, given such a set system and a threshold $t > 0$, estimates within factor $1 \pm \epsilon$ the number of colorings x for which $b^*(x) \geq t$. The algorithm should run in time $\text{poly}(nm/\varepsilon)$.

Hint: For each $i \in [m]$ define $X_i := \{x : \ b_i(x) \geq t\}$, and estimate $|\cup_i X_i|$.

2. Call a set B of integers *sum free* if it does not contain any three elements x, y, z such that $x + y = z$. For example, if K is an integer then the set consisting of the integer interval $[K + 1, \ldots, 2K + 1]$ is sum free. The goal is to prove that for every finite set A of positive integers, there is a subset $B \subset A$ of size at least $|A|/3$ that is sum free. Let $P = 3K + 2$ be a prime larger than the largest element in A (there are infinitely many such primes).

   (a) Argue that if for a set $B \subset A$ there is an integer r such that for all triples $x, y, z \in B$ we have that $rx + ry \neq rz \pmod P$ then B is sum free.

   (b) Show that a sum free set of size at least $|A|/3$ exists.

   Hint: For an integer $r > 0$, let $B_r \subset A$ contain all $x \in A$ with $K + 1 \leq rx \pmod P \leq 2K + 1$ (i.e., do the multiplication in Z_P^*, treat the result as an integer, and check if it falls in the integer interval $[K + 1, \ldots, 2K + 1]$). Argue that B_r is sum free, and explain how to choose r so that $|B_r|$ is large.

   (c) Suggest a randomized or deterministic algorithm for finding a sum free subset of size at least $|A|/3 - 1$. The running time should be polynomial in $|A|$ and in the *representation size* of the elements of A, i.e., polynomial in $\sum_{a \in A} \log |a|$.

3. For a d-regular graph $G = (V, E)$ consider an assignment of labels to nodes ("colors") from the set $[1, \ldots, c]$. Call the assignment *c-colorful* if for every node $v \in V$ and every color $b \in [1, \ldots, c]$, the node v has a neighbor colored b. Show that there is an $\alpha > 0$ s.t. for all d and for $c = \lfloor \alpha \cdot d / \log d \rfloor$, every d-regular graph G has a c-colorful assignment.

   Hint: Use the Lovasz Local Lemma.

4. The *total distance* within a set $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ is defined as

$$TD(X) := \sum_{i<j} \|x_i - x_j\|_2.$$

   Design an algorithm that, given an input X and $\varepsilon \in (0, 1)$, estimates $TD(X)$ within factor $1 \pm \varepsilon$ in time $\tilde{O}(\varepsilon^{-2} nd)$.

   Hint: Consider a random vector $g \in \mathbb{R}^d$, where each coordinate g_i has a standard Gaussian distribution $N(0, 1)$, and let $Y = \{y_1, \ldots, y_n\} \subset \mathbb{R}$ where each $y_i = \langle g, x_i \rangle$. Show that $TD(Y)$ can be computed exactly in $O(n \log n)$ time, and how it can be used to estimate $TD(X)$.

   You may use the following facts. First, if $g_0 \sim N(0, 1)$ then $\mathbb{E}[|g_0|] = \sqrt{2/\pi}$. Second, if two random variables X, Y (not necessarily independent) have standard deviation σ_X, σ_Y respectively, then $\sigma_{X+Y} \leq \sigma_X + \sigma_Y$.

Good Luck.

THE END.