Modern Algorithm Design (Monsoon 2024)
Homework 3

Deadline: 23rd November, 2024, 11 pm (IST)

Release Date: 12th November, 2024

1. **Yet Another Randomized Quicksort.** Given an array A of n numbers (which we assume are distinct for simplicity), the algorithm picks a pivot x uniformly at random from A and computes the rank of x. If the rank of x is between $n/4$ and $3n/4$ (call such a pivot a good pivot), it behaves like the normal QuickSort in partitioning the array A and recursing on both sides. If the rank of x does not satisfy the desired property (the pivot picked is not good), the algorithm simply repeats the process of picking a pivot until it finds a good one. Note that in principle the algorithm may never terminate!

   (a) Prove that the expected runtime of the above algorithm is $\mathcal{O}(n \log n)$ in expectation.

   (4 points) Solution Sketch. Firstly, observe that the pivot picking process for any subproblem is repeated till you get a 'good' pivot. This means that once you finally select a pivot, any subproblem of size n will be broken in to two subproblems of sizes, each of which is at least of size $n/4$ and consequently at most $3n/4$ (*+1 for this point*). Thus, if the work done outside the recursive calls can be bounded by $O(n)$, we are done by the Master's Theorem or whatever. Note that once you select a candidate pivot, you will need $O(n)$ to figure out whether it is a good or a bad one (you can essentially do the paritioning simultaneously) (*+1 for this part*). So it all boils down to showing that in expectation, you do not need to repeat the random pivot picking too many times. This is easy to argue. What is the chance that a uniform random selection gives you a good pivot. Well, it has to be any element that is between the rank $n/4$th number and rank $3n/4$th number in the 'sorted' version of the subarray. The chance of hitting this range is exactly $1/2$ (*+1 for this part*). Hence, the expected number of times you need to repeat is 2 (*+1 for this point*) (we have discussed this several times).

   (b) Prove that the runtime of the above algorithm is $\mathcal{O}(n \log n)$ with probability at least $1 - 1/n$

   (6 points) Solution Sketch. This is slightly harder to prove but we follow the same steps. Let us consider the jth level in the recursion tree, where $j = 0, 1, 2, \cdots c \log n$, where c is a constant. Note that the number of levels is $c \log n$ by the argument in the previous section (*+1 for this argument*). Now, there are 2^j subproblems at level j and the size of any subproblem is at most $(3/4)^j n$ (*+1 for this argument*). Let us focus on the ith subproblem and let X_i denote the random variable that counts the number of times you repeat the random picking of pivot till you hit a 'good' one for this subproblem. Also, let Y_j denote the random variable which denotes the total number of times across all subproblems at level j for which the random selection is repeated. Clearly $Y = \sum_{i=1}^{\eta_j} X_i$, where $\eta_j = 2^j$. Now, from the argument in part (a), $\mathbf{E}[X_i] = 2$, for all $i = 1, 2, \cdots \eta_j$. Hence, by linearity of expectation $Y_j = 2 \cdot 2^j = 2^{j+1}$ (*+1 for correctly defining random variables*). Also, it is not hard to see that $\mathbf{Var}[X_i] = 2$, for all $i = 1, 2, \cdots, \eta_j$ (this needs some calculations but you should be matured enough to handle this). Further, since the pivot selection is independent for each subproblem at level j, $\mathbf{Var}[Y_j] = 2^{j+1}$ (*+1 for showing the variances*). Now, by Chebyshev's inequality (*+1 for applying the bound*),

$$\mathbf{Pr}[|Y_j - \mathbf{E}[Y_j]| > 2n] \leq \frac{\sqrt{\mathbf{Var}[Y_j]}}{4n^2} = 2^{(j+1)/2}/4n^2$$

   Note the above inequality holds for a fixed j. Taking union bound over all $j = 0, 1, 2, \cdots c \log n$, we get that the probability that Y_j exceeds n for any level is at most $\sum_{j=0}^{c \log n} \frac{2^{(j+1)/2}}{4n^2} = \frac{1}{2n}$. The last ineqaulity follows since the numerators add up as a geometric sum to at most $2n$ (I am being sloppy here but cannot be more than that) (*+1 for this union bound computation*). Thus, with a probability of $1 - 1/2n$, the total runtime is $O(n \log n)$.

2. **Nearly Orthonormal Vectors.** We call a set of unit vectors "near-orthonormal" if the inner product of any two of them is close to zero. In this problem we will show that while there are at most d orthonormal vectors in

$\mathbb{R}^d$, there can be exponentially more near-orthonormal vectors. For vectors $x, y \in \mathbb{R}^d$, we use $\langle x, y\rangle = \sum_{i=1}^{d} x_i y_i$ to denote the inner product.

Acknowledgement : The following solution has been written by Rachit Arora.

(a) Let $x = (x_1, x_2, \cdots, x_d)$ and $y = (y_1, y_2, \cdots, y_d)$ be two independently and uniformly chosen vectors in $\{-1, 1\}^d$. (I.e., each bit x_i and y_i in each vector is independently and uniformly chosen from $\{-1, 1\}$.) Show that

$$\mathbb{P}r[|\langle x, y\rangle| \geq \varepsilon d] \leq 2\exp(-\varepsilon^2 d/6)$$

Solution. Let us calculate a lower bound on $\mathbb{P}r[\langle x, y\rangle \geq \varepsilon d]$. We have,

$$\begin{aligned}
&\mathbb{P}r[\langle x, y\rangle \geq \varepsilon d] \\
&= \mathbb{P}r[x_1y_1 + x_2y_2 + \cdots + x_dy_d \geq \varepsilon d] \\
&= \mathbb{P}r[(x_1y_1 + 1) + (x_2y_2 + 1) + \cdots + (x_dy_d + 1) \geq (1 + \varepsilon)d] \\
&= \mathbb{P}r[\frac{(x_1y_1 + 1)}{2} + \frac{(x_2y_2 + 1)}{2} + \cdots + \frac{(x_dy_d + 1)}{2} \geq (1 + \varepsilon)\frac{d}{2}]
\end{aligned}$$

Define $X_i = \frac{x_i y_i + 1}{2}$.
X_i can only take values 0 and 1 due to constraints on x_i and y_i, and $E[X_i] = \frac{1}{2}$.
Now, let $X = \sum_{i=1}^{d} X_i$. Since all X_i and X_j are pairwise independent (since all entries of vectors are drawn independently), and $\mu = E[X] = E[\sum_{i=1}^{d} X_i] = \frac{d}{2}$, we can apply the Chernoff-Hoeffding bound for $0 < \varepsilon < 1$:

$$\begin{aligned}
\mathbb{P}r[\frac{(x_1y_1 + 1)}{2} + \frac{(x_2y_2 + 1)}{2} + \cdots + \frac{(x_dy_d + 1)}{2} \geq (1 + \varepsilon)\frac{d}{2}] &= \mathbb{P}r[\sum_{i=1}^{d} X_i \geq (1 + \varepsilon)\frac{d}{2}] \\
&= \mathbb{P}r[X \geq (1 + \varepsilon)\frac{d}{2}] \\
&\leq \exp(-\varepsilon^2\mu/3) \\
&\leq \exp(-\varepsilon^2 d/6)
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{P}r[\frac{(x_1y_1 + 1)}{2} + \frac{(x_2y_2 + 1)}{2} + \cdots + \frac{(x_dy_d + 1)}{2} \geq (1 + \varepsilon)\frac{d}{2}] &\leq \exp(-\varepsilon^2 d/6) \\
\mathbb{P}r[\frac{(x_1y_1)}{2} + \frac{(x_2y_2)}{2} + \cdots + \frac{(x_dy_d)}{2} \geq \varepsilon\frac{d}{2}] &\leq \exp(-\varepsilon^2 d/6) \\
\mathbb{P}r[(x_1y_1) + (x_2y_2) + \cdots + (x_dy_d) \geq \varepsilon d] &\leq \exp(-\varepsilon^2 d/6) \\
\mathbb{P}r[\langle x, y\rangle \geq \varepsilon d] &\leq \exp(-\varepsilon^2 d/6)
\end{aligned}$$

Since $\langle x, y\rangle$ and -$\langle x, y\rangle$ have the same probability distribution,

$$\begin{aligned}
\mathbb{P}r[\langle x, y\rangle \leq -\varepsilon d] &= \mathbb{P}r[\langle x, y\rangle \geq \varepsilon d] \\
&\leq \exp(-\varepsilon^2 d/6)
\end{aligned}$$

This gives us,

$$\begin{aligned}
\mathbb{P}r[|\langle x, y\rangle| \geq \varepsilon d] &= \mathbb{P}r[\langle x, y\rangle \leq -\varepsilon d] + \mathbb{P}r[\langle x, y\rangle \geq \varepsilon d] \text{(both events are disjoint)} \\
&\leq 2\exp(-\varepsilon^2 d/6)
\end{aligned}$$

(b) Given any constant $\varepsilon > 0$, a set S of unit vectors is called ε-*orthonormal* if for all $\vec{x}, \vec{y} \in S$,

$$|\langle \vec{x}, \vec{y} \rangle| \leq \varepsilon.$$

Show that there exist constants $c, d_0 > 0$ (possibly depending on ε) such that for any $d \geq d_0$, if you sample $N := exp(c\varepsilon^2 d)$ random vectors independently and uniformly from the set $\left\{-\frac{1}{\sqrt{d}}, +\frac{1}{\sqrt{d}}\right\}^d$, this sampled set is ε-orthonormal with probability at least $1/2$.

Solution. The entries of x and y are sampled uniformly at random from $\left\{-\frac{1}{\sqrt{d}}, +\frac{1}{\sqrt{d}}\right\}$.

Let $x' = \sqrt{d}x$ and $y' = \sqrt{d}y$. Since x' and y' are now vectors in $\mathbb{R}^d$ where entries are sampled uniformly at random from $\{-1, 1\}$,

$$\mathbb{P}r[|\langle x', y' \rangle| \geq \varepsilon d] \leq 2\exp(-\varepsilon^2 d/6)$$

Substituting in x and y,

$$\begin{aligned}
\mathbb{P}r[|\langle \sqrt{d}x, \sqrt{d}y \rangle| \geq \varepsilon d] &\leq 2\exp(-\varepsilon^2 d/6) \\
\mathbb{P}r[\sqrt{d} \cdot \sqrt{d}|\langle x, y \rangle| \geq \varepsilon d] &\leq 2\exp(-\varepsilon^2 d/6) \\
\mathbb{P}r[d|\langle x, y \rangle| \geq \varepsilon d] &\leq 2\exp(-\varepsilon^2 d/6) \\
\mathbb{P}r[|\langle x, y \rangle| > \varepsilon] &\leq 2\exp(-\varepsilon^2 d/6)
\end{aligned}$$

Thus, the probability that x and y are **not** ε-orthonormal is bounded by $2\exp(-\varepsilon^2 d/6)$.

Suppose $x_1, x_2, \cdots, x_N$ are N random vectors sampled independently and uniformly from the set $\left\{-\frac{1}{\sqrt{d}}, +\frac{1}{\sqrt{d}}\right\}^d$.

Let $O_{i,j}$ be the event "x_i and x_j are **not** ε-orthonormal".

The probability of all of $x_1, x_2, \cdots, x_N$ being pairwise ε-orthonormal can be expressed as:

$$1 - \mathbb{P}r\Big[\bigcup_{\substack{i<j \\ 1\leq i,j\leq N \\ i,j\in\mathbb{N}}} O_{i,j}\Big]$$

By union bound,

$$1 - \mathbb{P}r\Big[\bigcup_{\substack{i<j \\ 1\leq i,j\leq N \\ i,j\in\mathbb{N}}} O_{i,j}\Big] \geq 1 - \sum_{\substack{i<j \\ 1\leq i,j\leq N \\ i,j\in\mathbb{N}}} \mathbb{P}r[O_{i,j}]$$

For this expression to be $\geq 1/2$, we want

$$\sum_{\substack{i<j \\ 1\leq i,j\leq N \\ i,j\in\mathbb{N}}} \mathbb{P}r[O_{i,j}] \leq 1/2$$

Vectors x_i and x_j not being ε-orthonormal does not depend on any other pair not being ε-orthonormal. And all x_i are i.i.d, all $\mathbb{P}r[O_{i,j}]$ are equal for any $i < j$.

wlog, let us substitute all $\mathbb{P}r[O_{i,j}]$ terms as $\mathbb{P}r[O_{1,2}]$.

$$\begin{aligned}
\sum_{\substack{i<j \\ 1\leq i,j\leq N \\ i,j\in\mathbb{N}}} \mathbb{P}r[O_{i,j}] &= \sum_{\substack{i<j \\ 1\leq i,j\leq N \\ i,j\in\mathbb{N}}} \mathbb{P}r[O_{1,2}] \\
&= \binom{N}{2}\mathbb{P}r[O_{1,2}] \\
&\leq \frac{N^2}{2}\mathbb{P}r[O_{1,2}]
\end{aligned}$$

We know that $N = \exp(c\varepsilon^2 d)$ and $\mathbb{P}r[O_{1,2}] \leq 2\exp(-\varepsilon^2 d/6)$.
Therefore,

$$\begin{aligned}\frac{N^2}{2}\mathbb{P}r[O_{1,2}] &\leq \exp(2c\varepsilon^2 d) \cdot 1/2 \cdot 2 \cdot \exp(-\varepsilon^2 d/6) \\ &= \exp(2c\varepsilon^2 d - \frac{\varepsilon^2 d}{6})\end{aligned}$$

We want to find choices for c and d_0 such that $\exp(2c\varepsilon^2 d - \frac{\varepsilon^2 d}{6}) \leq 1/2$.
Let us set c to a constant $< 1/12$, say $1/16$.

$$\begin{aligned}\exp(2c\varepsilon^2 d - \frac{\varepsilon^2 d}{6}) &= \exp((\frac{1}{8} - \frac{1}{6})\varepsilon^2 d) \\ &= \exp(-\frac{\varepsilon^2 d}{24})\end{aligned}$$

Now we want to find the constraint on d such that this quantity is $\leq 1/2$.

$$\begin{aligned}\exp(-\frac{\varepsilon^2 d}{24}) &\leq \frac{1}{2} \\ -\frac{\varepsilon^2 d}{24} &\leq \ln\frac{1}{2} \\ \frac{\varepsilon^2 d}{24} &\geq \ln 2 \\ d &\geq \frac{24\ln 2}{\varepsilon^2} = d_0\end{aligned}$$

Thus, $c = \frac{1}{16}$ and $d_0 = \frac{24\ln 2}{\varepsilon^2}$ is a possible assignment of c and d_0 such that for any $d \geq d_0$, if we sample $\exp(c\varepsilon^2 d)$ random vectors independently and uniformly from the set $\left\{-\frac{1}{\sqrt{d}}, +\frac{1}{\sqrt{d}}\right\}^d$, this sampled set is ε-orthonormal with probability at least $1/2$.