COMS 4995 (Randomized Algorithms): Exercise Set #6

For the week of October 7–11, 2019

Instructions:

(1) *Do not turn anything in.*

(2) The course staff is happy to discuss the solutions of these exercises with you in office hours or in the course discussion forum.

(3) While these exercises are certainly not trivial, you should be able to complete them on your own (perhaps after consulting with the course staff or a friend for hints).

Exercise 24

Recall that in Lecture #10 we proved the following version of the Chernoff bound. Let $X_1, \ldots, X_n$ be i.i.d. (i.e., independent and identically distributed) Bernouilli random variables with parameter p (i.e., equal to 1 with probability p and 0 with probability $1-p$). Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbf{E}[X] = np$. Then

(a) For every $\gamma > 0$,

$$\mathbf{Pr}[X > (1+\gamma)\mu] \leq \left(\frac{e^\gamma}{(1+\gamma)^{(1+\gamma)}}\right)^\mu.$$

(b) For every $\gamma \in (0, 1]$,

$$\mathbf{Pr}[X > (1+\gamma)\mu] \leq e^{-\mu\gamma^2/3}.$$

Verify that these exact same inequalities hold (with the exact same proof) when $X_1, \ldots, X_n$ are arbitrary independent Bernouilli random variables (i.e., with possibly different parameters p_i).

Exercise 25

Derive from the inequality (a) in Exercise 24 the following inequality, which can be particularly easy to apply:

(c) For every $R \geq 6\mu$,

$$\mathbf{Pr}[X > R] \leq 2^{-R}.$$

Exercise 26

In our application of Chernoff bounds to the expected maximum search time of hashing with chaining (Lecture #10), we claimed that for $\gamma = 3\frac{\ln n}{\ln \ln n}$,

$$\frac{e^\gamma}{(1+\gamma)^{1+\gamma}} \leq \frac{1}{n^2}.$$

Verify this inequality.

Exercise 27

In our analysis of the expected maximum search time of hashing with chaining, we assumed that the hash function h was a random oracle. Show that if we only assume that h is chosen uniformly at random from a universal family of hash functions, then with $m = n$ (i.e., size of data set equal to number of buckets), the expected maximum search time might be as large as $\Omega(\sqrt{n})$.

[Hint: Revisit Exercise #21.]

Exercise 28

Modify the variant of the count-min sketch data structure from Problem 9 on Problem Set #2 and its analysis to achieve the same correctness guarantee while reducing the dependence on the failure probability δ from $1/\delta$ to $\log(1/\delta)$ (while keeping the dependence on all other parameters the same).

[Hint: Proceed as in our application of amplifying the correctness probability of randomized algorithms with two-sided error (Lecture #10).]

Exercise 29

In Lecture #11 we mentioned that, because a JL map preserves (approximately) interpoint distances, it also preserves (approximately) angles between points (which in turn is relevant for computing k-nearest neighbors with the cosine similarity function). Here's what we mean.

(a) Prove that the inner product $\langle p, q \rangle$ between two vectors can be expressed purely as a function of the norms of p, q, and $p - q$. Thus, norm preservation implies inner product preservation.[1]

[Hint: look up the "polarization identity."]

(b) Prove that the angle between two vectors can be expressed purely as a function of the norms of p, q, and $p - q$. Thus, norm preservation implies angle preservation.

Analogous statements apply for the approximate preservation of distances, inner products, and angles (optional: work this out carefully).

Exercise 30

In Lecture #11 we only proved a bound on the upper tail of the sum of chi-squared random variables. Use a slight variation on that proof to prove an analogous bound for the lower tail. Specifically, for i.i.d. standard Gaussians $X_1, \ldots, X_m$ and $\alpha = (1 - \epsilon)^2 m$, prove that

$$\mathbf{Pr}\left[\sum_{i=1}^{m} X_i^2 < \alpha\right] \leq e^{-c \cdot m \epsilon^2},$$

where $c > 0$ is a constant (independent of m and ϵ).

[1] Remember that preserving norms is a special case of preserving interpoint distances—the case where one of the two points is the origin.