CS265, Winter 2022

Class 9: Agenda, Questions, and Links

1 Announcements

- HW5 is out, due next Wednesday.

2 Lecture Recap and Questions?

Questions from minilectures and pre-class quiz? (Compressed sensing; RIP; Gaussian matrices have the RIP with high probability.)

3 More matrices with the RIP whp

Group Work

1. Let $\varepsilon \in (0, 1/4)$. Suppose that $A \in \mathbb{R}^{m\times n}$ is a distribution on matrices so that, for some constant c:

$$\forall x \in \mathbb{R}^n, \Pr\{|\|Ax\|_2 - \|x\|_2| \geq \varepsilon\|x\|_2\} \leq 2\exp(-cm\varepsilon^2). \quad (1)$$

   (a) Is it the case that A is a good JL transform (aka, for any set $S \subseteq \mathbb{R}^n$ of size N, $\|A(x-y)\|_2 = (1 \pm \varepsilon)\|x-y\|_2$ with high probability), with $m = O(\varepsilon^{-2}\log N)$?

   (b) Is it the case that, with high probability, A has the (k, ε)-RIP with $m = O(\varepsilon^{-2}k\log n)$?

2. Let $A \in (\pm 1)^{m\times n}$ be a matrix where every entry is independently selected to be either $+1$ or -1. In this question, you'll show that for a cleverly chosen constant s, the matrix sA satisfies (1). (Notice that sA is much easier to generate than a random Gaussian matrix, and is also nicer to compute with).

   (a) What should s be as a function of m and n, so that for any vector $x \in \mathbb{R}^n$, $\mathbb{E}\|sAx\|_2^2 = \|x\|_2^2$?

   (b) For a vector $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$, let Z denote the random variable representing the inner product of x a row of matrix A. Namely $Z = \sum_{i=1}^n Y_i x_i$ where Y_i is independently chosen to be ± 1 with probability $1/2$ each, and x_i denotes the ith coordinate of x. The following bound on the moment generating function of Z^2 is not too hard (but a bit tedious) to prove: for any $t \in (0, 1/3)$,

$$E[e^{tZ^2}] \leq 1 + t + 12t^2.$$

Using this bound on the moment generating function of Z^2, prove that

$$\Pr[\|sAx\|_2^2 \geq (1+\varepsilon)] = \Pr[Z_1^2 + Z_2^2 + \ldots + Z_m^2 \geq (1+\varepsilon)m] \leq e^{-\varepsilon^2 m/100},$$

where the Z_i's represent independent realizations of the random variable Z.
Hint: *Proceed as in the proof of Chernoff bounds by multiplying both sides by t, exponentiating, applying Markov's inequality, then using the given bounds on the moment-generating function of each Z_i.*
Hint: *In the final step, you'll want to plug in an 'optimal' value of t. Try something like $t = \varepsilon/24$ to get the math to work out cleanly.*

(c) Have we proved that sA satisfies property (1)? If not, what is missing?

3. Here are a few "challenge" questions to think about:

(a) What other distributions on a matrix A can you come up with (other than i.i.d. Gaussians and i.i.d. $\pm 1/\sqrt{m}$ entries) that are (a) natural and (b) seem like they'd satisfy (1)? For example, what about any matrix with i.i.d. mean-zero entries? What about any matrix with i.i.d. mean-zero *bounded* entries? (i.e., the entries should never be larger than 100).

(b) Suppose that A has the RIP. Consider a matrix $A \cdot D$, where D is a diagonal matrix with i.i.d. mean-zero ± 1 entries on the diagonal. Show that AD satisfies (1), up to log factors.
Hint: *This is pretty tricky to do quantitatively, but you may be able to come up with some intuition for why it should be true qualitatively.*
Hint: *For a complete solution, check out this paper:* https://arxiv.org/pdf/1009.0744.pdf

4 Connected Components in Random Graphs

As one final super-cool application of Chernoff bounds, in this problem we will prove a really cool property of the sizes of the connected components in a natural random graph model.

Let $G_{n,p}$ denote the Erdos-Renyi random graph model, where each edge exists (independently) with probability $p = c/n$ for some constant c that does not vary with n.

Theorem 1. *Let G be drawn from $G_{n,p}$, with $p = c/n$ for some constant c :*

- *If $c < 1$, with probability tending to 1 as $n \to \infty$, the largest connected component of G has size $O(\log n)$.*
- *If $c > 1$, with probability tending to 1 as $n \to \infty$, the largest connected component of G has size $f(c)n \pm o(n)$, where $f(c)$ is a constant, independent of n, satisfying $f(c) \in (0,1)$ for all $c > 1$.[1], and the second-largest connected component of G has size $O(\log n)$.*

[1] $f(c)$ can actually be defined fairly cleanly: Suppose on day 1, we start with 1 rabbit. On each day, each existing rabbit will have a number of offspring drawn from (independent) Poisson random variables of

Group Work

1. Spend one minute pondering why we shouldn't expect any "medium" sized connected components. What is the intuition behind the above theorem?
2. In this problem we prove the $c < 1$ case of the above theorem.
   (a) For a given vertex v, prove that

   $$\Pr[v \text{ in connected component of size } \geq k] \leq \Pr\left[X \geq k-1\right],$$

   where X is distributed according to $Binomial[k \cdot n, c/n]$. [Hint: consider doing a breadth-first search of the neighborhood of v in the graph.]

   (b) Assuming the above, using a union bound over Chernoff bounds, prove that

   $$\Pr[\text{there is a connected component of size } \geq \frac{10 \log n}{(1-c)^2}] \leq 1/n.$$

3. In this problem, we prove the $c > 1$ case of the above theorem.
   (a) Given a random node v in the graph, prove that for any k satisfying $\frac{100c \log n}{(c-1)^2} \leq k \leq n^{3/4}$, the probability that the connected component of v has size k is no more than n^{-10}.
   Hint*: Consider a sort of breadth-first search that starts with a set that contains only v, then "marks" v and adds all the neighbors of v to the set , and then iteratively continues by "marking" an unmarked node of the set and adding all its neighbors to the set. Suppose we have "marked" k nodes, what is the chance that there are no more "unmarked" nodes in our set? Based on this, prove that, with high probability, if the connected component of v has size at least k, it will have size at least $k+1$. Be mindful of the way you condition events.*

   (b) Prove that we do not expect any connected components to have size in the interval $[\frac{100c \log n}{(c-1)^2}, n^{3/4}]$.

   (c) Prove that with probability tending to 1 as $n \to \infty$, there is at most one connected component of size $\geq n^{3/4}$. [Hint: conditioned on the neighborhood of both v and u having size at least $n^{3/4}$, show that the probability that they are not connected is tiny, then union bound over the at most n such neighborhoods.]

   (d) Challenge: Show that the size of the large component is within $o(n)$ of its expectation with probability tending to 1 as $n \to \infty$. [Hint: bound the variance of the number of nodes that are in "small" components of size at most $\frac{100c \log n}{(c-1)^2}$, then use Chebyshev's inequality.]

expectation c and then the original rabbits die, leaving only the offspring. $f(c)$ is the probability that this process never dies out (ie one minus the probability that there is some day with no more rabbits).