CS 388R: Randomized Algorithms, Fall 2021 November 16

Lecture 23: Randomized Numerical Algebra I

Prof. Eric Price *Scribe: Rojin Rezvan*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Overview

In last lecture we discussed Network Coding, which solves the problem of transmitting a message from a source vertex s to a target vertex t in some graph G.

In this lecture we will discuss the problem of solving the linear equation $Ax = b$, or more accurately approximately solving it. Solving this equation exactly can be specially hard when A is a tall matrix. We can think of it as a $n \times d$ matrix describing data for a learning algorithm in which d is the number of features and n is the number of users.

2 Problem Definition

Given a matrix $A_{n\times d}$ and a vector $b_{n\times 1}$, the goal is to find $x^*_{d\times 1}$ such that $x^* = \arg\min_x ||Ax - b||_2$.

Note that if A is a full column rank matrix, then $x^* = A^T b = (A^T A)^{-1} A^T b$. In order to find x^* exactly, we need to compute:

1. $A^T A$: a $d \times d$ matrix, takes $O(d^2 n)$ time, or $(d^{1.38} n)$ time with some improvements.
2. $(A^T A)^{-1}$: takes $O(d^3)$, or $O(d^{2.38})$ with improvements.
3. $(A^T A)^{-1}(A^T b)$: takes $O(nd)$ time.

This results in $O(nd^2)$ time for computation. In the case where $n >> d$, this is a long time. We are interested in finding a randomized algorithm that works in time $\tilde{O}(nd + poly(d))$. To this end, we will compromise on x^*, in that we will change our goal to finding: $\hat{x}$ such that

$$||A\hat{x} - b||_2 \leq (1 + \epsilon)||Ax^* - b||_2$$

One idea is to use conjugate gradients. This solution depends on A and its condition number, or $\kappa(A^T A)$ and will give run time of $O(nd \log \frac{n}{\epsilon}).\sqrt{\kappa(A^T A)}$.

3 Algorithm: Sketch and solve framework

We will achieve a run time of $\tilde{O}(nd poly(\frac{1}{\epsilon}) + d^3 poly(\frac{1}{\epsilon}))$. The idea is that we do not want to deal with huge number of rows. Rather than solving $\min_x ||Ax - b||_2$, pick "sketch" matrix $S \in \mathbb{R}^{m\times n}$

with $m \sim \frac{d}{\epsilon^2}$, and solve $\hat{x} = \arg\min_x ||SAx - Sb||_2$. Then we solve exactly. Ideally, we would want to have:

1. $||SA\hat{x} - Sb|| = (1 \pm \epsilon)||A\hat{x} - b||_2$
2. $||SAx^* - Sb||_2 = (1 \pm \epsilon)||Ax^* - b||_2$

With these two, we get:

$$||A\hat{x} - b||_2 \leq \frac{1}{1-\epsilon}||SA\hat{x} - Sb||_2 \leq \frac{1}{1-\epsilon}||SAx^8 - Sb||_2 \leq \frac{1+\epsilon}{1-\epsilon}||Ax^* - b||_2$$

3.1 Finding S

Suppose S is an iid Gaussian matrix: $S_{ij} \sim N(0, \frac{1}{m})$. Then we have $(Sx) \sim N(0, I_m.\frac{||x||_2^2}{m})$. Since $\mathbb{E}[||Sx||_2^2] = ||x||_2^2$, using a concentration inequality we get:

$$Pr[|\frac{||Sx||_2^2}{||x||_2^2} - 1| \geq \epsilon] \leq \exp\{-\Omega(\epsilon^2 m)\}$$

Then it suffices to set $m = O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$ for $\epsilon-$approximation with probability $1-\delta$

Problem? We cannot just double this number, because $\hat{x}$ depends on the whole subspace, unlike x^*. In other words, m cannot be less than d, because then $\hat{x}$ will have many answers and it will be a null space which we don't have any information from. In next section, we will address this issue.

4 Embedding

Definition 1. *S is a **subspace embedding** for space X if*

$$||Sx||_2^2 = (1 \pm \epsilon)||x||_2^2 \textit{ for all } x \in X$$

Definition 2. *S is a $(\epsilon, d)-$dim$-d$ **oblivious subspace embedding**, or **OSE**, if for any $d-$dim subspace $Y = \{y \in Ax | x \in \mathbb{R}^d\}$ such that $A \in \mathbb{R}^{n \times d}$, S is subspace embedding for Y with probability $1-\delta$.*

Lemma 3. *If S is (ϵ, δ) $d+1-$dim OSE, then "Sketch-and-solve" gives $(1 \pm O(\epsilon))$ accuracy with probability $1-\delta$.*

Proof. We can think of $Ax - b$ as the multiplication of $A|b|$ and $x, -1$ where the former is A with b added as its last column, and the latter is x with -1 added as its last row. (Note that the number of columns in $A|b$ and the number of rows in $x, -1$ is both $d+1$.) Now note that set of all x, X, has a dimension of at most $d+1$. If S is OSE, then with probability $1-\delta$ we have:

$$||S(Ax - b)||^2 = (1 \pm \epsilon)||Ax - b||^2, \quad \forall x \in X$$

□

Definition 4. *For a space X, say $N \subseteq X$ is a $\epsilon-$net if $\forall x \in X$, there exists $y \in N$ such that $||x - y|| \leq \epsilon$.*

Lemma 5. *The $d-$dimensional unit sphere has $\epsilon-$net of size at most $(1 + \frac{2}{\epsilon})^d$.*

Proof. Consider a greedy approach: Put a point in N for every point: if its distance is more than ϵ from current members, add it. The greedy net produces N points $x_1, \ldots, x_n$ with minimum distance $||x_i - x_j|| \geq \epsilon$. Then $B(x_i, \frac{\epsilon}{2})$ balls are disjoint for $i = 1, \ldots, n$. So :

$$\cup_i B(x_i, \frac{\epsilon}{2}) \subseteq B(0, 1 + \frac{\epsilon}{2})$$

So:

$$\begin{aligned} & Vol(\cup_i B(x_i, \frac{\epsilon}{2})) \leq Vol(B(0, 1 + \frac{\epsilon}{2})) \\ & \rightarrow N.Vol(B(x_i, \frac{\epsilon}{2})) \leq Vol(B(0, 1 + \frac{\epsilon}{2})) \\ & \rightarrow N.c_d(\frac{\epsilon}{2})^d \leq c_d(1 + \frac{\epsilon}{2})^d \\ & \rightarrow N \leq (\frac{1 + \frac{\epsilon}{2}}{\frac{\epsilon}{2}})^d = (1 + \frac{2}{\epsilon})^d \end{aligned} \tag{1}$$

□

Corollary: The same holds for $\{y = Ax | ||y||_2 = 1\}$ when $A \in \mathbb{R}^{n \times n}$ is full rank.

Definition 6. *S is $(\epsilon - \delta)$ distributional Johson-Lindenstrauss if:*

$$\forall x \in \mathbb{R}^n, ||Sx||_2^2 = (1 \pm \epsilon)_2^2 \ wp\ 1 - \delta$$

Example: If $S \in \mathbb{R}^{m \times n}$ is iid Gaussian, $m = O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$, then:

$$\forall x, y, \langle Sx, Sy \rangle = \langle x, y \rangle \pm \epsilon ||x||.||y|| \text{ wp } 1 - 2\delta$$

Proof.

$$\begin{aligned} ||x + y||_2^2 - ||x - y||_2^2 = 4\langle x, y \rangle & \\ & = ||S(x + y)||_2^2 - ||S(x - y)||_2^2 \pm \epsilon ||X \pm y||_2^2 \pm ||x - y||_2^2 \\ & = 4\langle Sx, Sy \rangle \pm \epsilon(||x + y||_2^2 + ||x - y||_2^2) \end{aligned} \tag{2}$$

If $||x|| = ||y|| = 1$, then $\langle x, y \rangle = \langle Sx, Sy \rangle \pm \epsilon$. This is true for all norms of x and y because we can scale. □

Lemma 7. *If S is $(\epsilon, \delta 25^{-d})-$distributional JL, then S is a $(4\epsilon, \delta)$ OSE of dimension d.*

Proof. Take a $\frac{1}{2}-$ net of the space $\{y|y \in Y, ||y||_2 = 1\}$. Then:

$$N \leq (1 + \frac{2}{\epsilon})^d = (1 + \frac{2}{1/2})^d = 5^d$$

Now consider the net $y_1, \dots, y_N$. For each pair y_i, y_j in the net, we know that"

$$\langle Sy_i, Sy_j \rangle = \langle y_i, y_j \rangle \pm \epsilon \text{ wp } 1 - \frac{\delta}{25^d}$$

We have $\binom{N}{2}$ such pairs. Since $25^d = N^2$, then using union bound, the above holds for every i, j with probability $1 - \delta$.

Now, for all $y \in Y$ we can write

$$y = y^0 + r^1$$

where $||r^1|| \leq \frac{1}{2}$ and y^0 is the point that is closest to y in the $\epsilon-$net. Equivalently, we may assume that $||r^1|| = 1$ and write:

$$y = y^0 + \epsilon^0 r^1, \quad \epsilon^1 \leq \frac{1}{2}$$

Now we can continue this expansion for $y^0, y^1, \dots$. Then we get:

$$y = y^0 + \epsilon^1 y^1 + \epsilon^2 y^2 + \dots, \quad y^i \in N, \epsilon^i \leq 2^{-i}$$

Now we will use this expansion to figure out $||Sy||_2^2$.

$$\begin{aligned}
||Sy||_2^2 &= \langle \sum_i \epsilon^i Sy^i, \sum_i \epsilon^i Sy^i \rangle \\
&= \sum_i \epsilon_i^2 ||Sy_i||^2 + \sum_{i<j} 2\epsilon_i \epsilon_j \langle Sy^i, Sy^j \rangle \\
&= \sum_i {\epsilon_i}^2 (||y^i||_2^2 \pm \epsilon) + \sum_{i<j} 2\epsilon_i \epsilon_j (\langle y^i, y^j \rangle \pm \epsilon) \\
&= ||y||_2^2 \pm \sum_i \epsilon \epsilon_i^2 \pm \sum_{i<j} \epsilon_i \epsilon_j \epsilon \\
&= ||y||_2^2 \pm \epsilon (\sum_i \epsilon_i^2 + \sum_{i<j} \epsilon_i \epsilon_j) \\
&\leq ||y||_2^2 \pm \epsilon (\sum_i 2^{-2i} + \sum_{i<j} 2^{-i-j}) \\
&\leq ||y||_2^2 \pm 4\epsilon
\end{aligned} \tag{3}$$

So we get OSE with: $O(\frac{1}{\epsilon^2} \log(\frac{25^d}{\delta})) = O(\frac{d}{\epsilon^2} + \log \frac{1/d}{\epsilon^2})$

$\square$

Problem? We still need to compute AS and this can be inefficient. There are ways to overcome this. For example, we may write S in the form $S = PHD$ such that P is a sub-sample matrix that can be computed in time $O(\frac{d}{\epsilon \log^2 d})$ and H is the Fourier matrix. With this format, AS computation can be done in time $O(nd \log n)$.