COMS E6998-15: Algorithms for Massive Data (Fall'23) — Sep 25, 2023

Lecture 6: Numerical Linear Algebra

Instructor: *Alex Andoni* — Scribes: *Tal Zussman*

1 Recap: Least Squares Regression

We want to find

$$x^\star = \operatorname*{argmin}_x \|Ax - b\|_2^2$$

where A, x, and b are as in Figure 1.

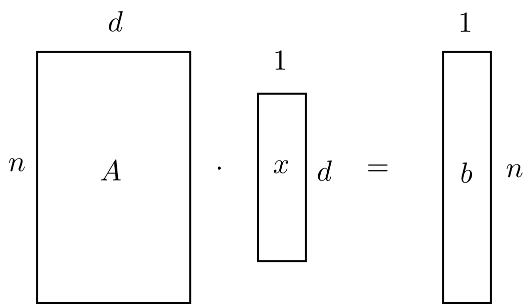

Figure 1: Least Squares Regression (dimensions)

Without FMM (Fast Matrix Multiplication), finding a solution takes time $O(nd^2)$.

2 Approximate LSR

We want to try to solve this faster, which we can do using an approximate version of LSR:

$$\hat{x} \text{ s.t. } \|A\hat{x} - b\|_2^2 \le (1+\epsilon)\|Ax^\star - b\|_2^2$$

We can use DJL to do this.

Theorem 1 (Distributional Johnson-Lindenstrauss)**.** *Given map $\varphi : \mathbb{R}^n \to \mathbb{R}^k$ and $\varphi(x) = G \cdot x$, where G is a random Gaussian matrix,*

$$\forall z \in \mathbb{R}^n, \;\; \Pr_{\varphi}\left[\frac{\|\varphi(z)\|_2^2}{\|z\|_2^2} = (1 \pm \epsilon)\right] \ge 1 - \delta$$

for $k = \Theta\left(\frac{\log \frac{1}{\delta}}{\epsilon^2}\right)$.

We treat $Ax - b$ as the vector upon which we do dimension reduction. Take S s.t. $\varphi(x) = S \cdot x$. Consider $\|S(Ax-b)\|_2^2$. By DJL, this is a $(1+\epsilon)$-approximation to $\|Ax - b\|_2^2$ with probability $\ge 1 - \delta$.

$Ax - b$ is an n-dimensional vector, and S reduces it to dimension k. We look at

$$x' = \text{argmin} \, \|S(Ax-b)\|_2^2 = \text{argmin} \, \|SAx - Sb\|_2^2$$

SA is a $k \times d$ matrix, and Sb is a vector of dimension k. We've effectively reduced this to LSR with $n = k$. Solving then takes time $O(kd^2)$. This is the main hope for the improvement in runtime.

2.1 Correctness

We want to verify the following:

$$\min \|Ax - b\|_2^2 \approx \min \|SAx - Sb\|$$

We want to be able to say that x' is an approximate minimizer. That is,

$$\|Ax' - b\|_2^2 \le (1+\epsilon)\|Ax^\star - b\|_2^2$$

with probability $\ge 1 - \delta$.

Effectively, we want to ensure that the approximate value we compute using x' doesn't vary too much from the actual minimum. This is similar to saying that we want the two curves in Figure 2 to not differ too much from each other.

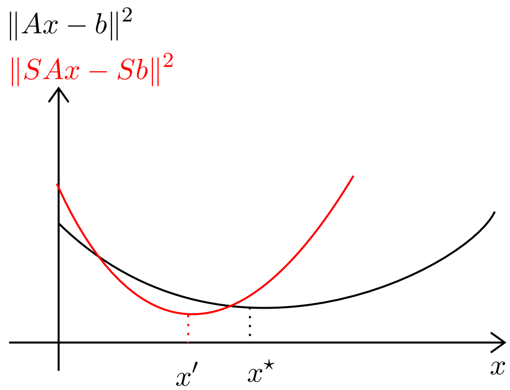

Figure 2: Minimization curves

1. From DJL, we know that

   $$\|SAx^\star - Sb\|_2^2 \le (1+\epsilon)\|Ax^\star - b\|_2^2$$

   with probability $\ge 1 - \delta$.

2. We want to show that

   $$\forall x, \ \|SAx - Sb\|_2^2 \ge (1-\epsilon)\|Ax - b\|_2^2$$

   We'd like to be able to obtain this from DJL, but unfortunately we can't, due to the following issues:

   - We can't apply DJL to all $x \in \mathbb{R}^n$

- We can't apply DJL to x' since x' is a function of S, which is random

3 Oblivious Subspace Embedding

Definition 2 (Oblivious Subspace Embedding (OSE))**.** *Fix n, d, ϵ, δ. $\forall$ subspace U of dimension d,*

$$\Pr_{\varphi}\left[\forall z \in U : \frac{\|\varphi(z)\|_2^2}{\|z\|_2^2} = (1 \pm \epsilon)\right] \geq 1 - \delta$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}^k$.

Theorem 3. *Suppose $\varphi : \mathbb{R}^n \to \mathbb{R}^k$ satisfies DFL, with $k = k(\epsilon, \delta)$. Then φ is an OSE with $k = k(\epsilon, \exp(-d))$, with $\delta = \exp(-d)$.*

This is similar to DJL, but instead of a fixed point we talk about fixed subspaces.

Corollary 4. *OSE with $\delta = \exp(-d)$ leads to $k = O\left(\frac{d}{\epsilon^2}\right)$.*

Proof omitted, have to prove in HW2.

3.1 Correctness

We'd like to solve LSR using an OSE. We take x' as before and S s.t. $\varphi(x) = Sx$. We want to show the same thing as in Section 2.1.

1. Holds as before with $\Pr \geq 1 - e^{-d}$.

2. We want to use the new property from OSE to say that the curves in Figure 2 are close to each other over all x.

$$\begin{aligned} U &\triangleq \{Ax + x_{d+1}b \mid x \in \mathbb{R}^{d+1}\} \\ &= \text{span}\{\text{columns of } A, b\} \end{aligned}$$

   $\dim U = d + 1$, and $\{Ax - b \mid x \in \mathbb{R}^d\} \subseteq U$.

   Using the OSE property on U, we see

$$\forall x,\ \|SAx - Sb\|_2^2 = (1 \pm \epsilon)\|Ax - b\|_2^2$$

   holds with probability $\geq 1 - e^{-d}$.

3.2 Runtime

We reduced our original problem to dimension reduction + LSR on a $k \times d$ problem, with $k = O\left(\frac{d}{\epsilon^2}\right)$.

Solving the new LSR problem takes $O(kd^2)$ for non-FMM. Computing SA takes $O(knd)$, and computing Sb takes $O(nd)$, so the total time is

$$O(kd^2) + O(knd) + O(nd) = O\left(\frac{nd^2}{\epsilon^2}\right)$$

Note that this is larger than our original runtime. In particular, computing SA takes longer than running the original LSR. We can clearly see that our dimension reduction is too slow!

4 Fast Dimension Reduction

Theorem 5 (Ailon-Chazelle'06)**.** *We can design a linear map* $\varphi : \mathbb{R}^n \to \mathbb{R}^k$ *with* $k = O\left(\frac{\log \frac{1}{\delta}}{\epsilon^2} \cdot \log \frac{n}{\delta}\right)$ *such that:*

1. *It satisfies DJL.*
2. *We can compute* $\varphi(x)$ *for fixed* x *in* $O(n \log n + k)$ *time.*

This is in contrast to standard DJL, which takes time $O(kn)$.

Corollary 6. *Using [AC'06] + OSE Theorem + the LSR approximation approach, the time to compute LSR is as follows.*

$$\begin{aligned} k &= O\left(\frac{d}{\epsilon^2} \cdot \log(ne^d)\right) \\ &= O\left(\frac{d}{\epsilon^2}(\log n + d)\right) \\ &\approx O\left(\frac{d^2}{\epsilon^2}\right) \end{aligned}$$

The last step follows from assuming that $d > \log n$.

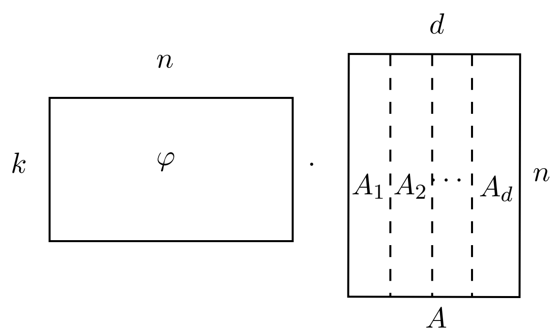

Figure 3: Computation of SA

Take $\varphi(x) = Sx$. *Since we need to perform dimension reduction on all* d *columns of* A *to compute* SA *(*$SA = SA_1 + SA_2 + \cdots + SA_d$*, see Figure 3), computing* SA *takes time*

$$d \cdot (n \log n + \frac{d^3}{\epsilon^2} + \frac{d^4}{\epsilon^2}) = nd \log n + \frac{d^3}{\epsilon^2}$$

The total time is then $O\left(nd \log n + \frac{d^3}{\epsilon^2} + \frac{d^4}{\epsilon^2}\right) < nd^2$*, when* $n \gg d$ *and* $d \gg \log n$.

Note that our run time can't be smaller than nd because A is of size nd, and our run time can't be less than our input size. Recent improvements have led to $O\left(nnz(A) + \left(\frac{d}{\epsilon}\right)^{O(1)}\right)$, where $z(A)$ is the number of non-zeros in A and the $O(1)$ factor ≤ 3.