Lecture 10 — Februray 6^{th}

Lecturer: Mert Pilanci *Scribe: Sakshi Namdeo & Ayan Mandal*

10.1 Faster Least Squares Optimization

In this lecture, we study the Least Squares problem and see how we can use random projections to reduce the time complexity of the least squares problem. More specifically, we will study about the basic inequality method which can be used to bound the norm of the error between the actual least squares solution and the randomized least squares solution.

Consider the least squares problem where A ϵ $\mathbb{R}^{n\times d}$ and b ϵ $\mathbb{R}^n$-

$$\min_{x\epsilon\mathbb{R}^d} ||Ax-b||_2^2 \tag{10.1.1}$$

The exact solution to the least squares problem is given as $x_{LS} = A^+.b$. Using classical methods the solution to this problem is $O(nd^2)$.

Using sketching we can reduce the the dimension of A and therefore the complexity of the least squares problem. Using a Random left sketching matrix $S\,\epsilon\,\mathbb{R}^{m\times n}$, where m¡¡n, form SA and Sb. For an approximate solution we can solve the smaller problem-

$$\min_{x\epsilon\mathbb{R}^d} ||SAx-Sb||_2^2 \tag{10.1.2}$$

The solution to the above least squares problem is given as $\tilde{x} = (SA)^+.b$. Using classical methods the solution to this problem is $O(md^2)$.

10.1.1 Basic Inequality Method

Given the two least squares problems we want to estimate how close $\tilde{x}$ is to x_{LS}. For this we define the quantity $\Delta = \tilde{x} - x_{LS}$. Since the least squares problem minimizes the cost function we can establish two optimality equations as below -

$$||Ax_{LS}-b||_2^2 \le ||Ax^{'}-b||_2^2 \text{ for any } x^{'} \text{ , i.e. } A^T(Ax_{LS}-b)=0 \tag{10.1.3}$$

$$||S(A\tilde{x}-b)||_2^2 \le ||S(Ax_{LS}-b)||_2^2 \tag{10.1.4}$$

Substituting $\tilde{x} = x_{LS} + \Delta$, we can rewrite the second equation as -

$$||S(A(x_{LS}+\Delta)-b)||_2^2 \le ||S(Ax_{LS}-b)||_2^2 \tag{10.1.5}$$

$$\implies ||S(Ax_{LS}-b)||_2^2 + ||SA\Delta||_2^2 - 2b^{\perp^T}(S^TS-I)A\Delta \le ||S(Ax_{LS}-b)||_2^2; \text{where } b^\perp = Ax_{LS}-b$$

$$\implies ||SA\Delta||_2^2 \le 2b^{\perp^T}(S^TS-I)A\Delta \tag{10.1.6}$$

The above equation is called the basic inequality. We want to use it to say that $||A\Delta||_2^2$ is small. Note that if $S^TS = I$ then $\Delta = 0$. We know that for embeddings that satisfy approximate matrix multiplication $S^TS \approx I$.

To argue the closeness of $\tilde{x}$ and x_{LS} first consider the following metric which measures the worst case distortion caused by S -

$$\max_{\Delta} \left| \frac{||SA\Delta||_2^2}{||A\Delta||_2^2} - 1 \right| ; A = U\Sigma V^T \implies A\Delta = U\Sigma V^T\Delta = Uz \tag{10.1.7}$$

$$\begin{aligned} \max_{\Delta} \left| \frac{||SA\Delta||_2^2}{||A\Delta||_2^2} - 1 \right| &= \max_{z} \left| \frac{||SUz||_2^2}{||z||_2^2} - 1 \right| \\ &= \max_{z} \left| \frac{z^TU^TS^TSUz}{z^TU^TUz} - 1 \right| \\ &= \max_{z} \left| \frac{z^T}{||z||_2}(U^TS^TSU - I)\frac{z}{||z||_2} \right| \\ &= \sigma_{max}^2(U^TS^TSU - I) \end{aligned} \tag{10.1.8}$$

We know that the frobenius norm $||.||_F = \sqrt{\Sigma\sigma_i^2}$. Therefore for any Q, $\sigma_{max}(Q) \leq ||Q||_F$. Using the above and approximate matrix multiplication we can write

$$\sigma_{max}(U^TS^TSU - I) \leq ||U^TS^TSU - I||_F \leq \epsilon||U^TU||_F^2, \tag{10.1.9}$$

where the second inequality follows from AMM. This is called a Subspace Embedding and σ_{max} can be controlled by choosing an appropriate m (Sampling size in S). In other words we can rescale ϵ to get $\sigma_{max}(U^TS^TSU - I) \leq \epsilon^{'}$ for some m.

$$\begin{aligned} &\implies \max_{\Delta} \left| \frac{||SA\Delta||_2^2}{||A\Delta||_2^2} - 1 \right| \leq \epsilon^{'} \\ &\implies -\epsilon^{'}||A\Delta||_2^2 \leq ||SA\Delta||_2^2 - ||A\Delta||_2^2 \leq \epsilon^{'}||A\Delta||_2^2 \\ &\implies (1 - \epsilon^{'})||A\Delta||_2^2 \leq ||SA\Delta||_2^2 \end{aligned} \tag{10.1.10}$$

We have now derived the following inequalities -

$$||SA\Delta||_2^2 \leq 2b^{\perp^T}(S^TS - I)A\Delta \tag{10.1.11}$$

$$(1 - \epsilon^{'})||A\Delta||_2^2 \leq ||SA\Delta||_2^2 \tag{10.1.12}$$

Consider the RHS of the first inequality and $A = U\Sigma V^T \implies UU^TA = UU^TU\Sigma V^T = U\Sigma V^T = A$.

$$\begin{aligned} 2b^{\perp^T}(S^TS - I)A\Delta &= 2b^{\perp^T}(S^TS - I)UU^TA\Delta \\ &\leq 2||b^{\perp^T}(S^TS - I)UU^T||_2||A\Delta|| \\ &= 2||b^{\perp^T}S^TSUU^T - b^{\perp^T}UU^T||_2||A\Delta|| \\ &\leq \frac{\epsilon}{\sqrt{m}}||b^{\perp}||_F||UU^T||_F||A\Delta|| \text{ ; using Approximate Matrix Multiplication} \\ &\leq \frac{\epsilon}{\sqrt{m}}f(x_{LS})\sqrt{d}||A\Delta||_2 \text{ ; as } ||b^{\perp}||_F = ||Ax_{LS} - b||_2 = f(x_{LS}) \text{ and } ||UU^T||_F = \sqrt{d} \end{aligned} \tag{10.1.13}$$

From equations 10.1.6, 10.1.12 and 10.1.13 we can write -

$$(1-\epsilon^{'})||A\Delta||_2^2 \leq ||SA\Delta||_2^2 \leq 2b^{\perp^T}(S^TS-I)A\Delta \leq \frac{\epsilon}{\sqrt{m}}f(x_{LS})\sqrt{d}||A\Delta||_2 \tag{10.1.14}$$

$$\implies ||A\Delta||_2 \leq \frac{\epsilon^{'}}{1-\epsilon}f(x_{LS} \tag{10.1.15}$$

Special Case: Consider a linear system $Ax^* = b$ and an approximate linear system $SA\tilde{x} = Sb$ where x^* is the unique solution without sketching and $\tilde{x}$ is the least squares solution with a sketching matrix. Then,

$$Ax^* - A\tilde{x} = A\Delta \tag{10.1.16}$$

Also consider,

$$\begin{gathered} SA\tilde{x} = Sb \ \& SAx^* = Sb \\ \implies SA\Delta = 0 \end{gathered} \tag{10.1.17}$$

This also comes from the fact that if we sample enough number of times, then the error probability $\rightarrow 0. From Equation 10.1.14, we get \|A\Delta\|_2^2 \leq \frac{1}{1-\epsilon^{'}}\|SA\Delta\|_2^2 (10.1.18) Plugging$ the LHS in Equation 10.1.15, we get

$$\|A\Delta\|_2^2 = \frac{\epsilon}{1-\epsilon^{'}}f(x_{LS})\frac{\sqrt{d}}{\sqrt{m}} \tag{10.1.19}$$

under the assumption $U^TU = U^TS^TSU$.

10.2 Leverage Scores and Basic Inequality

Leverage scores are an important statistical measure to determine correlation between singular vectors of a matrix and the standard basis. They give a feel of how much influence or leverage a row has on the best least squares fit. This also gives information to the extent to which a data point is an outlier or errors. This method finds relevance especially while working in higher dimensional space to get high quality numerical implementation of randomized matrix algorithms.

Consider the Approximate Matrix Multiplication problem for U^TU where

$$\|U^TS^TSU - U^TU\|_F = \|U^TS^TSU - I\|_F \leq \epsilon \tag{10.2.1}$$

for some $\epsilon \in \{0, 1\}$ implies a Least Squares approximation. The S is chosen as a randomized sampling matrix so that the or L-2 norms are preserved with an ϵ-error with high probability.

10.2.1 Importance Sampling

Importance Sampling is a method of sampling from a distribution that over-weights the important regions and hence, helps in variance reduction. We can use importance sampling for for sensitivity analysis and as foundation for some methods of computing normalizing constants of probability densities. In our construction of Importance Sampling, the weights are proportional to the row norms of the matrix U for a data matrix $A = U\Sigma^T$.

The leverage scores are calculated as:

$$\ell_i = \|u_i\|_2^2 \tag{10.2.2}$$

where $\|u_i\|_2^2$ are the norms of row of U. Another way of looking at leverage scores is

$$\sum_i \ell_i = \sum_i U^T U \implies trI_d = d \tag{10.2.3}$$

when A is full column rank.

Thus, the sampling probabilities associated with which row can be computed by

$$p_i = \frac{1}{d}\|u_i\|_2^2 \tag{10.2.4}$$

which is the sampling probability distribution defined with replacement and $\sum_i p_i = 1$. This can be defined for a matrix A which maybe non uniform as well.

Note: If A is defined as $A = [I, 0]$ and $m = cd\log d$ and the row probabilities are proportional to the row norms $p_i = \frac{1}{d}\|u_i\|_2^2$ then,

$$\|A\tilde{x} - Ax^*|_2^2 \leq \epsilon \tag{10.2.5}$$

and is related to a cross validation set of hold out size = 1

Note: Hat Matrix is a projection matrix that maps the vector of response values to the vector of the predicted values. Consider a generalised linear model

$$A(A^TA)^{-1}A^T = UU^T$$
$$\implies \ell_i^T UU^T \ell_i = \|u_i\|_2^2$$

The computational complexity of this is in the order of $\mathcal{O}(nd^2)$ for computing the leverage scores ℓ_i via SVD.

10.3 Fast Johnson Lindenstrauss Transform

Given $0 < \epsilon < 1$, a set X of m points in $\mathbb{R}^N$, and a number $n > 8\ln(m)/\epsilon^2$, there is a linear map $f : \mathbb{R}^N \to \mathbb{R}^n$ such that

$$(1-\epsilon)\|u-v\|_2^2\| \leq \|f(u) - f(v)\|_2^2 \leq (1+\epsilon)\|u-v\|_2^2 \tag{10.3.1}$$

for all $u, v \in X$.

An alternative statement for any $0 < \epsilon, \delta < 1/2$ and positive integer d, there exists a

distribution $\mathcal{R}^{n\times d}$ from which the matric A is drawn such that $n = \mathcal{O}(\log(1/\delta)/\epsilon^2)$ anf for any unit length vector $x \in \mathcal{R}^d$ then

$$\mathbb{P}(\|\|Ax\|_2^2 - 1\| > \epsilon) < \delta \tag{10.3.2}$$

This can be obtained from the distributional version by setting $x = (u-v)/\|u-v\|_2$ and $\delta < 1/n^2$ for some pair $u, v \in X$.

The Fast JL Transform promises of computing matrix vector products in less than $\mathcal{O}(nd)$ by deriving distributions across the matrix under consideration.

Hadamard Matrix Let H be a $n \times n$ Hadamard Martix which is given as

$$H_n = \begin{bmatrix} H_{n-1} & H_{n-1} \\ H_{n-1} & -H_{n-1} \end{bmatrix}$$

and $H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. The rows and columns of a Hadamard matrix are orthogonal to each other. and finds application in image signal processing, optical multiplexing, error correcting coding and design & analysis of statistics.

10.3.1 Constructing FJLT Sketching Matrix

To construct a FJLT Sketching Matrix S

Step 1: Construct $H = n \times n$ Hadamard Matrix

Step 2: Generate matrix $D = \mathcal{R}^{n\times n}$ diagonal matrix of random ± 1 sampled from a uniform distribution; known as Rademacher matrix

Step 3: Consider $P = \mathcal{R}^{m\times n} \in \{0, 1\}$ a uniform sub-sampling matrix scaled by $\sqrt{n}/\sqrt{m}$.

In the problem construction, let $\mathbb{E}[S^T S] = I$. Then,

$$S = PHD \tag{10.3.3}$$

Thus, for $Sx = PHDx$, the P matrix samples the frequencies and the HDx computation can be performed in $\mathcal{O}(n \log n)$ time using Fast Hadamard Transform.

Note: If the Hadamard Matrix H is replaced with a Discrete Fourier Transform Matrix F, then our construction of the sketching matrix becomes $S = PFD$. It works well on matrices with probability density concentrated in a small region as follows from the Heisenberg's principle.