EE 270: Large Scale Matrix Computation and Optimization **Winter 2020**

Lecture 11 — Tuestday, Febuary 11

Lecturer: Mert Pilanci *Scribe: Chris Baca*

11.1 Overview of Matrix Approximation Methods

For approximating a matrix product ($A^T B \approx A^T S^T S B$), we generally have two options:

1) Sampling based methods (unifom sampling, row norm scores (which minimizes the variance of the Frobenius norm), leverage scores (this is relatively new to us)

a) Leverage scores take $O(nd^2)$ time to calculate exactly (by calculating SVD), and then the problem is solved, so we need an aproximate method. We do so by making leverage scores "nicer", i.e. nearly uniform, with some transform, Q (can be thought of as a rotation) then sampling uniformly

$$A' = QA$$
$$PA' = PQA$$

Where P is just a uniform sampling matrix, i.e. a 1 in each row and at most one 1 in each column, randomly chosen, normalized by a factor of $\sqrt{\frac{n}{m}}$ to make $E[P^T P] = I$. This provides us with a Leverage score sketching matrix, $S = PQ$. For example, with $Q = HD$, $S = PHD$, this is the Fast JL Transform. This method is useful in gradient descent, Newton's method, and others

2) Projection based methods (orthogonal and otherwise): Gaussian, Rademacher, Haar (uniform orthogonal), random projection, sparse J-L (Count-Sketch), Fast J-L (randomized Hadamard) All of these, except the first two, can actually be interpreted in the same $S = PQ$ form we used for leverage score-weighted sampling. For example, for Haar matrices we could continue the process of choosing orthogonal rows to creat an $n \times n$ matrix, an orthogonal rotation (Q), then randomly chose a subset of these orthogonal basis vectors with P, giving the Haar transform.

As we showed with the basic inequality method, row norm sampling doesn't work well for least squares. Random rotation "protects information" with high probability, by distributing the important information across all columns.

11.2 Spectral Interpretation of Leverage Scores

Let $U\Sigma V^T = A$ by a (compact) svd and S the leverage score sketching matrix. Then our usual bound for matrix multiplication gives us

$$||U^T S^T S U - U^T U||_F = ||U^T S^T S U - I||_F < \epsilon$$

As we showed on homework 1, the Frobenius norm is the square root of the sum of the squares of the singular values, so this implies

$$\sigma_{max}(U^TS^TSU - I) < \epsilon$$

Since this is true for σ_{max}, its true for all singular values. Since S is symmetric, its singular values are the absolute values of its eigenvalues, so we know

$$|\lambda_i(U^TS^TSU - I)| < \epsilon$$

But specifically because I is the identity, it just shifts the eigenvalues by 1, so we know

$$|\lambda_i(U^TS^TSU)| \in (1 - \epsilon, 1 + \epsilon)$$

This is important, because it means that U^TS^TSU is invertible with very high probability (note that this isn't true for uniform sampling). So when $(A^TA)^{-1}$ exists, $(A^TS^TSA)^{-1}$ exists (see below for proof), so we can solve sketched least squares with the pseudo-inverse of SA.

$$(A^TS^TSA)^{-1} = (V\Sigma U^TS^TSU\Sigma V^T)^{-1} = (\Sigma V^T)^{-1}(U^TS^TSU)^{-1}(V\Sigma)^{-1}$$

Where the invertibility of ΣV^T is invertible whenever $A^TA = V\Sigma^2V^T$ is invertible.

11.3 Leverage Scores and Spectral Properties

Since we have bounds on $\sigma_{max}(U^TS^TSU)$, we know that, for any x',

$$\frac{x'^T}{||x'||}(U^TS^TSU - I)\frac{x'}{||x'||} < \epsilon$$

This true if and only if

$$(1 - \epsilon)||x'|| < ||SUx'|| < (1 + \epsilon)||x'||$$

For any $x \in \mathbb{R}^d$, take $x' = \Sigma V^Tx$. Then the inequalities above become

$$(1 - \epsilon)||Ax|| < ||SAx|| < (1 + \epsilon)||Ax||$$

This is remarkable. We get, with high probability, conservation of euclidian norm for all $x \in \mathcal{R}(A)$, not just a specified set of x_i. This is called subspace embedding.

Weyls inequality allows us to work this subspace embedding property further, it tells us:

$$|\lambda_i(M) - \lambda_i(M')| \leq \sigma_{max}(M - M')$$

This inequality has its origins in perturbation theory in physics. This tells us that the spectrum is nearly conserved under leverage-score sketching:

$$|\lambda_i(A^TS^TSA) - \lambda_i(A^TA) <= \epsilon$$

A quick example of the usefulness of this bound can be seen in spectral graph theory. If we consider a graph with topography represented by A, which a row for each edge and a column for each node. If edge k goes from node i to node j, then $A[k, i] = 1$, $A[k, j] =$

-1. Disconnected components will correspond to 0-eigenvalues of A^TA, small eigenvalues represent "weakly connected coponents". Since leverage score sampling nearly preserves spectral integrity, its it tends to preserve those "important" edges which keep the eigenvalues non-zeros. However, note that we need epsilon smaller than "small eigenvalues" for this to work.

11.4 Leverage Scores and Generalization Error

To see the continued significance of leverage scores, we'll work through a slightly contrived example of the idea of generalized error, which will represent the sensitivity of the square loss to each term in its sum. Let $f(x) = ||Ax-b||_2^2$,$f_i(x) = (a_i^Tx - b_i)^2$, so that $f(x) = \sum_{i=1}^n f_i$. Our contrived step is that we assume there exists and exactly solution, i.e. $b = Ax^*$.

We define the worst case ratio of each term in the sum to the total:

$$\max_x R_j(x) = \max_x \frac{f_i(x)}{f(x)} = \max_x \frac{(a_i^Tx - b_i)^2}{||Ax-b||_2^2} = \max_x \frac{(a_i^T(x-x^*))^2}{\sum_{j=1}^n (a_j^T(x-x^*))^2}$$

.

Letting $x' = x - x^*$, this is

$$\max_{x'} \frac{(a_i^Tx')^2}{||Ax'||_2^2} = \max_{x'} \frac{||e_i^TAx'||_2^2}{||Ax'||_2^2} = \max_{x'} \frac{||e_i^TU\Sigma V^Tx'||^2}{||U\Sigma V^Tx'||^2}$$

If we let $\tilde{x} = \Sigma V^T x'$, then $||U\Sigma V^Tx'||^2 = ||\tilde{x}||^2$, because $U^TU = I$. So we get

$$\max_x R_j(x) = \max_{\tilde{x}} \frac{||e_i^TU\tilde{x}||^2}{||\tilde{x}||^2} = \max_{\tilde{x}} \frac{||u_i^T\tilde{x}||^2}{||\tilde{x}||^2}$$

From the Cauchy-Schwartz, inequality, with equality for parallel vectors, we know that this is exactly $||u_i||_2^2$, i.e. the leverage score l_i. So we can see, albeit in a contrived example, a heuristic for the fact that leverage scores correspond to importance of each row in least squares. So when we downsample rows we should pick higher leverage score terms with higher probability. This is leverage score sampling.

11.5 Fast J-L Transform Analysis

The Hadamard Transform is defined recursively as

$$H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$H_{2^{(n+1)}} = \begin{pmatrix} H_n & H_n \\ H_n & -H_n \end{pmatrix}$$

If we let D be an $n \times n$ diagonal matrix with random ± 1 along the diagonal, P be a random subsampling matrix scaled with $\frac{\sqrt{m}}{\sqrt{n}}$. Then $S = \frac{1}{\sqrt{n}}PHD$, so that $E[S^TS] = I$ since $DH^THD = nI$ and $E[P^TP] = I$.

We'll write the leverage scores of a matrix in slightly different form, noting that $A(A^TA)^{-1}A^T = UU^T$:

$$l_i = ||U^Te_i|| = e_i^TUU^Te_i = e_i^TA(A^TA)^{-1}A^Te_i$$

So the leverage scores of $\frac{1}{\sqrt{n}}HDA$, denoted $\tilde{l_i}$, can be written as

$$\begin{aligned}\tilde{l_i} &= e_i^THDA(A^TDH^THDA)^{-1}A^TDH^Te_i \\ &= \frac{1}{n}e_i^THD(A(A^TA)^{-1}A^T)DH^Te_i \\ &= \frac{1}{n}e_i^THD(UU^T)DH^Te_i \\ &= \frac{1}{n}h_i^TDUU^TDh_i\end{aligned}$$

where h_i is the i-row of H. In the first second equality we used $DH^THD = nI$. Each h_i is simply a vector of ±1, so Dh_i is i.i.d. ±1, after random sign flips.

So l_i is distributed as $\frac{1}{n}r^TUU^Tr$ where r is i.i.d. ±1. So $E[l_i] = E[\frac{1}{n}tr(UU^Trr^T)] = \frac{1}{n}tr(UU^TI) = \frac{d}{n}$.

We use the Chernoff bound,

$$P[\frac{1}{n}h_i^TDu_j \geq t] \leq 2e^{-t^2n/2}$$

Applying a union bound (didn't quite follow this, end of lecture was a bit rushed),

$$l_i = \frac{1}{n}h_i^TDUU^TDh_i \leq cost\frac{dlog(nd)}{n}$$

with high probability. Since leverage scores are near uniform, we can uniformly sample with P. So in summary, we apply HD to A, making leverage scores near uniform with high probability, then uniformly sample with P, giving a subspace embedding, $S = \frac{1}{\sqrt{n}}PHD$