CS369M: Algorithms for Modern Massive Data Set Analysis **Lecture 4 - 10/05/2009**

Norms of Random Matrices & Low-Rank via Sampling

Lecturer: Michael Mahoney *Scribes: Jacob Bien and Noah Youngs*

**Unedited Notes*

1 Overview

Looked at so far:

-Randomized algorithms for (fast) low-rank approximation
-Johnson Lindenstrauss Lemma → Random Projections
-Approximate multiplication → Random Sampling Algorithms
Both have additive error

Today:

Wigner's Semicircle Law
Random Matrix Theorum → Element-wise sampling algorithm

Next Time:

Approximate L2 Regression → $(1 + \epsilon)$ Approximation

2 Wigner's Semicircle Law

What does a 10^6 dimensional data-set look like?
-What would the "null hypothesis" of a truly random data set look like?

Wigner's Semicircle Law:

Let $A = (a_{ij})$ be a symmetric matrix ($a_{ij} = a_{ji}$) such that

1. $E(a_{ij}) = 0$
2. $Var(a_{ij}) = \sigma^2$
3. $|a_{ij}| \leq K$

Let $\lambda_1 \geq \lambda_2 \geq ...\lambda_n \in \mathbb{R}$ be the eigenvalues of A
Let $W_n(x) =$ the empirical distribution of the eigenvalues.

Then $\lim_{n\to+\infty} W_n(x * 2\sigma\sqrt{n}) = W(x)$ in probability

Where $W(x) = \begin{cases} \frac{2}{\pi}(1-x^2)^{1/2} & \text{for } |x| \leq 1 \\ 0 & \text{for } |x| > 1 \end{cases}$

This function is a semicircle centered at the origin with radius 1

Proof Idea: Compare the moments of W(x) with the moments of $\frac{2}{\pi}(1-x^2)^{1/2}$

3 Extensions

1. If $E(a_{ij}) = \mu \neq 0$
   Then $\lambda_1(A) \sim N(n\mu + \frac{\sigma^2}{\mu}, 2\sigma^2)$
   And the rest of the eigenvalues will follow the semicircle law

2. Apply Wigner's Law to the adjacency matrix of a random graph

   If G = (V,E), and $\forall ij \in V \times V$, $ij \in E$ with probability p
   Then $A_{ij} = \begin{cases} 1 & \text{with probability p} \\ 0 & \text{with probability (1-p)} \end{cases}$
   Therefore the entries of this matrix have non-zero mean, but fixed variance

   How sparse can this matrix be before the results break down?
   Let d = np be the average degree of the matrix
   Then the "trace argument" only works if $p \geq \frac{log(n)^4}{n}$
   In other words, if the above relation holds, then $d_{empirical} \approx d_{expected}$

   What if $p = \frac{2}{n}$?
   What if you want a random graph with expected degrees?

4 Bounds on the largest eigenvalue

We want: For a fixed n, to be able to make some statement about the largest eigenvalue

Fact: If G is a random n x d matrix, $d \leq n$ with entries $N(0, \sigma^2)$, i.i.d, and for a fixed ϵ, k
then w.p. $\geq \frac{1}{poly(k,\epsilon)}$
$||G||_2 = ||G_n||_2 \approx (2+\epsilon)\sigma\sqrt{n}$
$||G||_F \leq (2+\epsilon)\sigma\sqrt{nk}$

Is this scale good or bad?

Let D be the trivial rank-k approximation to G obtained by keeping the first k columns.
Then $||D||_F \approx \sigma\sqrt{nk}$
Rank(D) $\leq k$, by definition, so $||D||_2 \geq \frac{||D||_F}{\sqrt{k}} \approx \sigma\sqrt{n}$

This generalizes to the where the distribution of the entries in G has:
-Mean zero
-Bounded entries
-Independence

The following results about the largest eigenvalue of a random symmetric matrix can be found in the references:

From Furedi and Komlos:
Let A be a random symmetric matrix with: $|a_{ij}| \leq k$, $E(a_{ij}) = 0$, and $var(a_{ij}) = \sigma^2$
Then there exists a constant $C = C(\sigma, k)$ such that with high probability:
$2\sigma\sqrt{n} - Cn^{1/3}log(n) \leq \lambda_1(A) \leq 2\sigma\sqrt{n} + Cn^{1/3}log(n)$

This theorem is extended in Alon, Krivelevich, and Vu:

Then given A,$\forall$ t, $P[|\lambda_1(A) - E(\lambda_1(A))| > ct] \leq 4e^{\frac{-t^2}{32}}$

5 "How can we use these ideas to do something low rank without causing too much damage?"

Suppose $A \in R^{m \times n}$ and N is a noise matrix (i.e. 0 mean, etc.) such that $A + N \approx A$. Then we will try to add a data-dependent noise matrix to get a speed up (without messing things up too much).

Lemma 5.1. *Let A and N be such that $\hat{A} = A + N$. Then*

1) $||A - A_k||_2 \leq ||A - \hat{A}_k||_2 \leq ||A - A_k||_2 + 2||N_k||_2$

2) $||A - A_k||_F \leq ||A - \hat{A}_k||_F \leq ||A - A_k||_F + ||N_k||_F + 2\sqrt{||N_k||_F||A_k||_F}$

(Note: A_k, $\hat{A}_k$, and N_k denote the best rank k approximations to A, $\hat{A}$ and N respectively.)

Proof of Lemma. The first inequality of both (1) and (2) follows from the definition of A_k as the best rank k approximation to A. Now, let B be an any matrix (e.g. $A + N$).

1)

$$\begin{aligned} ||A - B_k||_2 &\leq ||A - B||_2 + ||B - B_k||_2 \text{ by the triangle inequality} \\ &\leq ||A - B||_2 + ||B - A_k||_2 \text{ since } B_k \text{ is the best rank } k \text{ approximation to } B \\ &\leq ||A - B||_2 + ||B - A||_2 + ||A - A_k||_2 \text{ by the triangle inequality} \\ &= ||A - A_k||_2 + 2||B - A||_2 \\ &= ||A - A_k||_2 + 2||(B - A)_k||_2 \end{aligned}$$

where the last equality holds since $B - A$ and $(B - A)_k$ have the same largest eigenvalue. Part (1) of the lemma follows letting $B = \hat{A} = A + N$.

2) Let P_M denote the projection onto the column space of the matrix M.

Claim 5.2. $||P_{A_k} A||_F \leq ||P_{B_k} A||_F + 2||(A - B)_k||_F$

Proof of Claim.

$$\begin{aligned} ||P_{A_k} A||_F &\leq ||P_{A_k}(A - B)||_F + ||P_{A_k} B||_F \text{ by the triangle inequality} \\ &\leq ||P_{A_k}(A - B)||_F + ||P_{B_k} B||_F \text{ since of all } B_k \text{ is the best rank } k \text{ approximation to } B \\ &\leq ||P_{A_k}(A - B)||_F + ||P_{B_k}(B - A)||_F + ||P_{B_k} A||_F \\ &\leq ||P_{B_k} A||_F + 2||P_{(A-B)_k}(A - B)||_F \text{ since } (A - B)_k \text{ is the best rank } k \text{ approx. to } A - B \\ &= ||P_{B_k} A||_F + 2||(A - B)_k||_F \text{ since } P_{(A-B)_k}(A - B) = (A - B)_k \end{aligned}$$

which proves the claim. □

Now, from the claim it follows that

$$\begin{aligned} ||P_{B_k} A||_F^2 &\geq (||P_{A_k} A||_F - 2||(A - B)_k||_F)^2 \\ &= ||P_{A_k} A||_F^2 - 4||P_{A_k} A|| \cdot ||(A - B)_k||_F + 4||(A - B)_k||_F^2. \end{aligned}$$

Thus, we have that

$$||P_{B_k}A||_F^2 \geq ||P_{A_k}A||_F^2 - 4||P_{A_k}A|| \cdot ||(A-B)_k||_F, \tag{1}$$

which we shall use shortly. Next, observe that

$$\begin{aligned} ||A-B_k||_F &\leq ||A-P_{B_k}A||_F + ||P_{B_k}A - B_k||_F \text{ by the triangle inequality} \\ &\leq ||A-P_{B_k}A||_F + ||P_{B_k}(A-B)||_F \text{ since } P_{B_k}B = B_k \\ &\leq ||A-P_{B_k}A||_F + ||P_{(A-B)_k}(A-B)||_F \text{ since } (A-B)_k \text{ is the best rank } k \text{ approx. to } A-B. \end{aligned}$$

Thus,

$$||A-B_k||_F \leq ||A-P_{B_k}A||_F + ||(A-B)_k||_F. \tag{2}$$

Finally, we make use of Equations 1 and 2:

$$\begin{aligned} ||A-P_{B_k}A||_F &\leq (||A||_F^2 - ||P_{B_k}A||_F^2||)^{1/2} \text{ since } (A-P_{B_k}A) \perp P_{B_k}A \\ &\leq (||A||_F^2 - ||P_{A_k}A||_F^2 + 4||P_{A_k}A|| \cdot ||(A-B)_k||_F)^{1/2} \text{ by Equation 1} \\ &= (||A-A_k||_F^2 + 4||A_k|| \cdot ||(A-B)_k||_F)^{1/2} \\ &\leq ||A-A_k||_F + 2(||A_k|| \cdot ||(A-B)_k||_F)^{1/2} \end{aligned}$$

using that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$. And by Equation 2,

$$\begin{aligned} ||A-B_k||_F &\leq ||A-P_{B_k}A||_F + ||(A-B)_k||_F \\ &\leq ||A-A_k||_F + 2(||A_k|| \cdot ||(A-B)_k||_F)^{1/2} + ||(A-B)_k||_F, \end{aligned}$$

Taking $B = \hat{A} = A + N$, this gives

$$||A-\hat{A}_k||_F \leq ||A-A_k||_F + 2(||A_k|| \cdot ||N_k||_F)^{1/2} + ||N_k||_F, \tag{3}$$

which completes the proof of the second part of the lemma.

□

6 Applications of the above lemma

The lemma above establishes that a rank k approximation of the perturbed matrix may not be too much worse. Now, we give two examples where perturbing can help in terms of memory and speed:

1. In representing A, each a_{ij} takes 32 or 64 bits.
2. Iterative eigensolvers depend on the number of non-zero entries of A.

6.1 Quantize the data

Given A, let $b = \max_{ij} |A_{ij}|$ and define

$$\hat{A}_{ij} = \begin{cases} +b & wp.\ 1/2 + A_{ij}/(2b) \\ -b & wp.\ 1/2 - A_{ij}/(2b) \end{cases} \tag{4}$$

Now, $E[\hat{A}_{ij}] = A_{ij}$ and $Var[\hat{A}_{ij}] = b^2 - A_{ij}$ and all the $\hat{A}_{ij}$ are independent. It follows from Theorem 3.1 of the Achlioptas and McSherry paper, that with high probability, $||A-\hat{A}||_{F \text{ or } 2}$ is not too large. By the lemmas, a low rank approximation to this quantized version of A will not be too bad.

6.2 Sparsify the data

Let $p = (\frac{8 \log n}{n})^4$. Here, we sample (independently) elementwise :

$$\hat{A}_{ij} = \begin{cases} A_{ij}/p & wp.\ p \\ 0 & wp.\ 1-p \end{cases} \tag{5}$$

So $E[\hat{A}_{ij}] = A_{ij}$ and $Var[\hat{A}_{ij}] = A_{ij}^2(1/p - 1) \leq b^2/p$. And with high probability $||(A - \hat{A})_k|| \leq$ "$2(1+\epsilon)\sigma\sqrt{n}$)" $\leq 4b\sqrt{n/p}$.

We can actually do better by non-uniform sampling: Let $p_{ij} = pA_{ij}^2/b^2$ and

$$\hat{A}_{ij} = \begin{cases} A_{ij}/p_{ij} & wp.\ p_{ij} \\ 0 & wp.\ 1-p_{ij} \end{cases} \tag{6}$$

We still have $E[\hat{A}_{ij}] = A_{ij}$ and $Var[\hat{A}_{ij}] = A_{ij}^2(1/p_{ij} - 1)$ and the error bounds still hold. The expected number of non-zero elements is in this case

$$E[\text{Num. non-zeros}] = \sum p_{ij} = p||A||_F^2/b^2 = \frac{pmn}{b^2}\frac{||A||_F^2}{mn}. \tag{7}$$

(Alternatively, we can use $p_{ij} \sim A_{ij}$ for small entries to keep from violating the bound constraint.)

7 References:

D. Achlioptas, F. Mcsherry, *Fast computation of low-rank matrix approximations*, Journal of ACM 54.2(2007).

N. Alon, M. Krivelevich, and V. Vu, *On the concentration of eigenvalues of random symmetric matrices*, Israel Journal of Mathematics 131.2(2002), 259 - 267.

Z. Furedi and J. Komlos, *The eigenvalues of random symmetric matrices*, Combinatorica1 (1981), 233 - 241.