CS 229r: Algorithms for Big Data Fall 2015

Lecture 15 — October 22, 2015

Prof. Jelani Nelson *Scribe: Brabeeba Wang*

1 Overview

We are going to focus on large sclae linear algebra and today it is on approximation matrix multiplication

1. $A \in \mathbb{R}^{n\times d}$, $B \in \mathbb{R}^{d\times p}$
2. Want to compute $A^T B$

Straight forward algorithm have $O(ndp)$ for loop. Alternatively, we can break out A^T, B into several $d \times d$ blocks and multiply it block by block. So we can use fast square matrix multiplication in $O(d^\omega)$.

1. $\omega < \log_2 7$ (Strassen)
2. $\omega < 2.376$ (Coppersmith, Winograd)
3. $\omega < 2.374$ (Stothevs)
4. $\omega < 2.3728642$ (Vassilevke-Williams)
5. $\omega < 2.3728639$ (Le Gell)

We can also reduce to rectangular matrix multiplication like we can multiply $r \times r^\alpha$ by $r^\alpha \times r$ in $r^{\alpha+o(1)}$ where $\alpha > 0.30298$ (Le Gall). But today we are going to settle for computing $C \in \mathbb{R}^{d\times p}$ such that $\|A^T B - C\|_X$ small. For example $\|\|_X = \|\|_F$.
Two approaches:

1. Sampling (Drineas, Kannan, Mahoney SIJC'06)[1]
2. JL (Sarlos FOCS'06)[2]

2 Sampling

1. $A^T B = \sum_{k=1}^n a_k b_k^T$
2. $C = (\Pi A)^T (\Pi B), \Pi \in \mathbb{R}^{m\times n}$
3. Π is a sampling matrix with rows $\Pi_1, \cdots, \Pi_m$
4. Π_t are independent across t

5. $\Pi_t = e_i/\sqrt{mp_i}$ with probability p_i proportional to $\|a_i\|_2 \cdot \|b_i\|_2$
6. $C = \sum_{t=1}^{m} 1/m a_{k_t} b_{k_t}^T / p_{k_t}$
7. Define $Z_t = 1/m a_{k_t} b_{k_t}^T / p_{k_t}$

Claim 1. $\mathbb{E}C = A^T B$

Proof. It is trivial by linear expectation since $\mathbb{E}Z_t = A^t B/m$ □

Claim 2. *If $m > 1/\epsilon^2\delta$, then*

$$\mathbb{P}(\|A^T B - (\Pi A)^T(\Pi B)\|_F > \epsilon\|A\|_F\|B\|_F) < \delta \tag{1}$$

Proof. By Markov,

$$\mathbb{P}(\|A^T B - (\Pi A)^T(\Pi B)\|_F > \epsilon\|A\|_F\|B\|_F) < \mathbb{E}(\|A^T B - C\|_F^2)/\epsilon^2\|A\|_F^2\|B\|_F^2 \tag{2}$$

1. WLOG, $\|A\|_F = \|B\|_F = 1$
2. $A^T B - C = \sum_{t=1}^{m}(Z_t - \mathbb{E}(Z_t))$
3. $\mathbb{E}(A^T B - C) = \sum_{i,j} \mathbb{E}(\sum_{t=1}^{m}(Z_t - \mathbb{E}Z_t)_{ij})^2 = \sum_{i,j} Var[\sum_{t=1}^{m}(Z_t)_{ij}] = \sum_{i,j} \sum Var[(Z_t)_{ij}] = m\sum_{i,j} Var[Z_{ij}]$
4. $Z_i j = 1/m \sum_{k=1}^{n} \rho_k/p_k \cdot a_{k_i} b_{k_j}$ where $\rho_\alpha = 1$ if tth row of Π sampled row k.
5. $Var[Z_{ij}] \leq \mathbb{E}(Z_{ij})^2 = 1/m^2 \cdot \sum_{k=1}^{n} \mathbb{E}(\rho_k)/p_k^2 \cdot a_{k_i}^2 b_{k_j}^2$
6. $\mathbb{E}\|A^T B - C\|_F^2 \leq 1/m \cdot \sum_{k=1}^{n} 1/p_k (\sum_{i,j} a_{k_i}^2 b_{k_j}^2) = 1/m \cdot \sum_{k=1}^{n} 1/p_k \|a_k\|_2^2 \|b_k\|_2^2 = 1/m(\sum_{k=1}^{n} \|a_k\|_2^2\|b_k\|_2^2)^2 \leq 1/m(\sum_{k=1}^{n} \|a_k\|_2^2)(\sum_{k=1}^{n} \|b_k\|_2^2)$.
7. This gives us $\mathbb{P}(\|A^T B - (\Pi A)^T(\Pi B)\|_F > \epsilon\|A\|_F\|B\|_F) < 1/\epsilon^2 m < \delta$

□

We want to improve the runtime to $\log(1/\delta)$. Following is the trick (Clarkson, Woodruff STOC'09)[3].

1. set $r = \Theta(\log(1/\delta))$
2. Compute $C_1, \cdots, C_r$ as before each with failure probability 1/10
3. Know by Chernoff $> 2/3$ of the C_i give low error $(\epsilon/4\|A\|_F\|B\|_F)$
4. Check which one is good:

for $i = 1$ to r:
    if $\|A^T B - C_i\|_F \leq \epsilon/4\|A\|_F\|B\|_F$:
        return C_i

5. Trick:
   for $i = 1$ to r:
       $ctr \leftarrow 0$
       for $j = 1$ to r:
           if $\|C_i - C_j\|_F < \epsilon/2\|A\|_F\|B\|_F$:
               $ctr \leftarrow ctr + 1$
           if $ctr > r/2$
               return C_i

Analysis:

1. # of good i is $> 2r/3$

2. if i is good, then for good j we have $\|C_i - C_j\|_F \leq \|A^TB - C_i\|_F + \|A^TB - C_j\|_F \leq \epsilon/4 + \epsilon/4 \leq \epsilon/2\|A\|_F\|B\|_F$

3 JL Approach

Definition 3. $\Pi \in \mathbb{R}^{m\times n}$ *and* D *is a distribution over* Π *satisfies the* $(\epsilon, \delta, p)-$*JL moment property if for any* $x \in S^{n-1}$ *we have* $\mathbb{E}_{\Pi\sim D}|\|\Pi x\|_2^2 - 1|^p < \epsilon^p\delta$

Example 4.
1. $\Pi_{ij} = \pm 1/\sqrt{m}$. *This induces* $(\epsilon, \delta, 2)-$ *JL moment property with* $m \geq 1/\epsilon^2\delta$ *and* $(\epsilon, \delta, \log(1/\delta))-$ *JL moment property with* $m \geq \log(1/\delta)/\epsilon^2$

2. *Based on the pset 1 problem 4, we have* $(\epsilon, \delta, 2)-$ *JL moment property with* $m \geq 1/\epsilon^2\delta$

Claim 5. *Suppose* Π *comes from* $(\epsilon, \delta, p)-$ *JL moment property for some* $p \geq 2$. *Then for any* A, B *with* n *rows, we have*

$$\mathbb{P}_{\Pi\sim D}(\|A^TB - (\Pi A)^T(\Pi B)\|_F > \epsilon\|A\|_F\|B\|_F) < \delta \tag{3}$$

Proof.
1. By Markov, $\mathbb{P}_{\Pi\sim D}(\|A^TB-(\Pi A)^T(\Pi B)\|_F > \epsilon\|A\|_F\|B\|_F) < \mathbb{E}\|A^TB-(\Pi A)^T(\Pi B)\|_F^p/\epsilon^p\|A\|_F^p\|B\|$

2. So we want to bound $\|\|A^TB - (\Pi A)^T(\Pi B)\|_F^2\|_{p/2}^{1/2}$

3. $\|A^TB - (\Pi A)^T(\Pi B)\|_F^2 = \sum_{i,j}\|a_i\|_2^2\|b_j\|_2^2X_{i,j}^2$

4. $X_{i,j} = \langle \Pi a_i/\|a_i\|_2, \Pi b_j\|b_j\|_2\rangle - \langle a_i/\|a_i\|_2, b_j\|b_j\|_2\rangle$

5. $\|\|A^TB - (\Pi A)^T(\Pi B)\|_F^2\|_{p/2} \leq \sum_{i,j}\|a_i\|_2^2\|b_j\|_2^2\|X_{ij}^2\|_{p/2} \leq \max_{i,j}\|a_i\|_2^2\|b_j\|_2^2\|X_{ij}^2\|_{p/2}$

6. Fix i, j we have $X_{i,j} = \langle \Pi x, \Pi y\rangle - \langle x, y\rangle$ where $\|x\|_2 = \|y\|_2 = 1$

7. $\|X_{ij}^2\|_{p/2} = \|X_{ij}^2\|_p$

8. $X_{ij} = 1/2[(\|\Pi(x-y)\|_2^2) + (\|\Pi x\|_2^2 - 1) + (\|\Pi y\|_2^2 - 1)]$

9. $\|X_{ij}\|_p \leq 1/2[\|x-y\|_2^2\|\frac{\|\Pi(x-y)\|^2}{\|x-y\|^2} - 1\|_p] + \|\|\Pi x\|_2^2 - 1\|_p + \|\|\Pi y\|_2^2 - 1\|_p \leq 1/2\epsilon\delta^{1/p}[\|x-y\|_2^2 + 1 + 1] \leq 3\epsilon\delta^{1/p}$

10. And this gives us $\mathbb{P}_{\Pi \sim D}(\|A^T B - (\Pi A)^T (\Pi B)\|_F > \epsilon \|A\|_F \|B\|_F) \leq \|A\|_F^2 \|B\|_F^2 9\epsilon^2 \delta^{2/p} / (3\epsilon \|A\|_F \|B\|_F)^p = \delta$

$\square$

4 Next class

We are going to get some results on operator bound. $\|(\Pi A)^T (\Pi B) - A^T B\| < \epsilon \|A\| \|B\|$

1. WLOG $\|A\| = \|B\| = 1$
2. WLOG $A = B$ because consider m as the justaposition of A^T and B. We can easily see that $\|m\| = 1$ And this gives us $\|m \frac{x}{y}\|_2^2 = \|Ax\|^2 + \|By\|^2 \leq \|x\|^2 + \|y\|^2$

And we will get something stronger. From now on $A = B$ and we want for any x, $\|\Pi A x\|_2^2 = (1 \pm \epsilon)\|Ax\|_2^2$. This is stronger because $\|(\Pi A)^T (\Pi A) - A^T A\| = \sup_{\|x\|=1} |\|\Pi A x\|_2^2 - \|Ax\|_2^2| < \epsilon \|Ax\|_2^2$

References

[1] Petros Drineas, Ravi Kannan, Michael Mahoney. Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication. *SIAM J. Comput* 36(1):132–157, 2006.

[2] Tamas Sarlos. Improved Approximation Algorithms for Large Matrices via Random Projections. *FOCS* 2006.

[3] Kenneth Clarkson, David Woodruff. Numerical Linear Algebra in the Streaming Model. *STOC*, 205–214, 2009.