CMPUT 675: Algorithms for Streaming and Big Data **Fall 2019**

Lecture 19 (Nov 18, 2019): Faster Matrix Multiplications

Lecturer: Mohammad R. Salavatipour *Scribe: Ramin Mousavi*

19.1 Introduction

Computing the product of two matrices are appeared in many algorithms implicitly or explicitly. Let A be a $p \times n$ matrix and B be a $n \times q$ matrix. Then, the running time of computing AB exactly is $p \cdot n \cdot q$ (basic 3 "For" loops algorithm). We can boost the running time by breaking up these matrices into smaller $k \times k$ block matrices and multiplying these block matrices together using Strassen's algorithm with $O(k^{2.8})$ arithmetic operations [S69].

There are two main approaches using randomized algorithm to estimate AB, namely, the sampling approach and Johnson-Lindenstrauss (JL) approach. In this lecture we talk about two examples of such approaches.

19.2 Frobenius and Spectral Norms

For a matrix $A \in \mathbb{R}^{m \times n}$, The Frobenius norm of A is defined as

$$\|A\|_F = \big(\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2\big)^{\frac{1}{2}},$$

the spectral norm of A defined as

$$\|A\|_2 = \sup_{x \in \mathbb{R}^n : \|x\|=1} \|Ax\|_2 .$$

Let $A^{(i)}$ be the i-th column of A and $A_{(j)}$ be the j-th row of A.

19.3 Sampling Approach

The key observation in this approach is that we can write the product of two matrices as the sum of outer products. More precisely, let $A \in \mathbb{R}^{p \times n}$ and $B \in \mathbb{R}^{n \times q}$, then $AB = \sum_{i=1}^{n} A^{(i)} B_{(i)}$. Then, we can sample columns of A and rows of B based on a distribution and then outer product of the sampled columns and rows give a good estimate of AB.

The following algorithm is given by [DKM06]. Let p_i's for $1 \leq i \leq n$ be a probability distribution which will be determined later.

Approximate Matrix multiplication (Sampling approach)
Input: $A \in \mathbb{R}^{p\times n}$, $B \in \mathbb{R}^{n\times q}$, and probability distribution p_i's for $1 \le i \le n$, and $t \in \mathbb{R}_+$.
Output: Matrix C that its Frobenius distance to AB is "small".

1. For $k = 1$ to t do
   - Pick index $j_k \in [n]$ with probability p_{j_k} independently and with replacement.
   - Set $C_k = \frac{A^{(j_k)}B_{(j_k)}}{p_k}$
2. Return $C = \frac{1}{t}\sum_{k=1}^{t} C_k$.

First we show that in expectation the above algorithm returns AB.

Lemma 1 $\mathbb{E}[C] = AB$.

Proof. In k-th iteration, the algorithm picks the j-th column of A and j-th row of B with probability p_j so $\mathbb{E}[C_k] = \sum_{j=1}^{n} p_k(\frac{A^{(j_k)}B_{(j_k)}}{p_{j_k}}) = \sum_{j=1}^{n} A^{(j)}B_{(j)} = AB$. So we have $\mathbb{E}[C] = \frac{1}{t}\sum_{k=1}^{t}\mathbb{E}[C_k] = AB$, as desired. ■

In order to bound the deviation from the expectation, we give an upper bound on $\mathbb{E}[\|AB - C\|_F^2]$ and then apply the Markov's inequality.

Lemma 2 $\mathbb{E}[\|AB - C\|_F^2] = \frac{1}{t}\big(\sum_{l=1}^{n}\frac{1}{p_l}\left\|A^{(l)}\right\|_2^2\left\|B_{(l)}\right\|_2^2\big) - \frac{1}{t}\|AB\|_F^2$.

Proof.

$$\mathbb{E}[\|AB - C\|_F^2] = \sum_{i=1}^{p}\sum_{j=1}^{q}\mathbb{E}[(AB - C)_{ij}^2] \tag{19.1}$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\mathbb{E}\left[\big(\mathbb{E}[C] - C\big)_{ij}^2\right] \tag{19.2}$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\mathrm{Var}[(C)_{ij}]. \tag{19.3}$$

Next we compute $\mathrm{Var}[(C)_{ij}]$.

Claim 1 $\mathrm{Var}[(C)_{ij}] = \frac{1}{t}\sum_{l=1}^{n}\frac{A_{il}^2B_{lj}^2}{p_k} - \frac{1}{t}(AB)_{ij}^2$.

Proof of Claim 1. Let C_k be the matrices defined in the algorithm. Then, $\mathbb{E}[(C_k)_{ij}^2] = \sum_{l=1}^{n} p_l\big(\frac{A_{il}B_{lj}}{p_l}\big)^2 = \sum_{l=1}^{n}\frac{A_{il}^2B_{lj}^2}{p_l}$.

Note that from Lemma 1 we conclude that $\mathbb{E}[(C_k)_{ij}]^2 = (AB)_{ij}^2$, together with the above line, we have $\mathrm{Var}[(C_k)_{ij}] = \sum_{l=1}^{n}\frac{A_{il}^2B_{lj}^2}{p_l} - (AB)_{ij}^2$.

Since C_k's are independent, we have $\mathrm{Var}[(C)_{ij}] = \frac{1}{t^2}\sum_{k=1}^{t}\mathrm{Var}[(C_k)_{ij}]$, and the claim follows. ■

Now the lemma follows from (19.3) and the above claim.

■

Since the smaller $\mathbb{E}[\|AB-C\|_F^2]$ is the better is the approximation, we have to find a probability distribution that minimizes this term. In the paper [DKM06], it was shown that the following probability distribution minimizes $\mathbb{E}[\|AB-C\|_F^2]$: For $1 \le l \le n$ define

$$p_k = \frac{\left\|A^{(k)}\right\|_2 \left\|B_{(k)}\right\|_2}{\sum_{l=1}^{n}\left\|A^{(l)}\right\|_2 \left\|B_{(l)}\right\|_2}. \tag{19.4}$$

So if we plug in (19.4) into Lemma 2, we get that

$$\begin{aligned}
\mathbb{E}[\|AB-C\|_F^2] &= \frac{1}{t}\Big(\sum_{l=1}^{n}\left\|A^{(l)}\right\|_2 \left\|B_{(l)}\right\|_2\Big)^2 - \frac{1}{t}\|AB\|_F^2 \\
&\le \frac{1}{t}\sum_{l=1}^{n}\left\|A^{(l)}\right\|_2^2 \sum_{l=1}^{n}\left\|B_{(l)}\right\|_2^2 - \frac{1}{t}\|AB\|_F^2 \\
&= \frac{1}{t}\|A\|_F^2\|B\|_F^2 - \frac{1}{t}\|AB\|_F^2 \\
&\le \frac{1}{t}\|A\|_F^2\|B\|_F^2,
\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz inequality.

Now we can apply Markov's inequality to bound the deviation probability.

$$\Pr\left[\,\|AB-C\|_F \ge \epsilon\|A\|_F\|B\|_F\,\right] = \Pr\left[\,\|AB-C\|_F^2 \ge \epsilon^2\|A\|_F^2\|B\|_F^2\,\right] \le \frac{\mathbb{E}\left[\,\|AB-C\|_F^2\,\right]}{\epsilon^2\|A\|_F^2\|B\|_F^2} \le \frac{1}{t\cdot\epsilon^2}.$$

So for the input of the algorithm, we pick p_j's as 19.4 describes and we set t to be $O(\frac{1}{\epsilon^2})$. Then, the algorithm gives a good estimation of the product of two matrices with high probability.

Some comments:

1. This algorithm runs with two passes over columns and rows of A and B respectively. In the first pass, we compute the sampling probabilities and in the second pass we construct the matrix C.

2. In $O(n)$ additional space we can keep $\left\|A^{(l)}\right\|_2$ and $\left\|B_{(l)}\right\|_2$ for $1 \le l \le n$. Also with $O(n)$ arithmetic operations, we can compute p_js. Then, in the second pass, in each iteration we sample an index which takes $O(n)$ time and computing the outer product of a vector of size p by a vector of size q which takes $O(p\cdot q)$ arithmetic operations. So in total, in the second pass we require additional $O(t(n+p\cdot q))$ time.

3. The algorithm might do poorly in the case of uniform sampling. Similar to the case of sum of bunch of numbers with big range, then uniform sampling does poorly. The intuition says we have to pick the largest number with higher probability.

19.4 JL Approach

Another approach to approximate the product of two matrices is based on Johnson-Lindenstrauss dimensionality reduction [S06] and [KN14].

Recall the (JL) theorem from Lecture 5. Given n vectors $v_1, ..., v_n$ in $\mathbb{R}^d$, there is a linear mapping $f : \mathbb{R}^d \to \mathbb{R}^k$ where $k = O(\frac{\log n}{\epsilon^2})$ such that for all pairs v_i, v_j we have

$$\Pr\left[(1-\epsilon)\left\|v_i - v_j\right\|_2 \leq \left\|f(v_i) - f(v_j)\right\|_2 \leq (1+\epsilon)\left\|v_i - v_j\right\|_2\right] \geq 1 - \delta.$$

Furthermore, f can be constructed with randomized algorithm in polytime.

Given a matrices $A \in \mathbb{R}^{n\times p}$ and $B \in \mathbb{R}^{n\times q}$, the idea is to using f to reduce the dimensions of A and B to $k \times p$ and $k \times q$, respectively. Then, we show that the product of these projected matrices is close to $A^T B$ in Frobenius norm.

Before we start, we need the following definition:

Definition 1 ((ϵ, δ, p)-JL moment) *Distribution $\mathcal{D}$ over $\mathbb{R}^{k\times n}$ has (ϵ, δ, p)-JL moment if for all $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$,*

$$\mathop{\mathbb{E}}_{\Pi\sim\mathcal{D}}\left[\left|\left\|\Pi x\right\|_2^2 - 1\right|^p\right] \leq \epsilon^p \cdot \delta.$$

The following is an useful property of such distributions with JL moment property.

Lemma 3 *If Π comes from an (ϵ, δ, l)-JL moment distribution $\mathcal{D}$ wherer $l \geq 2$, then for all unit vectors $x, y \in \mathbb{R}^n$, we have*

$$\left\|\langle \Pi x, \Pi y\rangle - \langle x, y\rangle\right\|_l \leq 3\epsilon\delta^{\frac{1}{l}},$$

where $\|X\|_l := (\mathbb{E}\,|X|^l)^{\frac{1}{l}}$ for $l \geq 1$. Note that by Minkowski's inequality this defines a norm.

Proof. We can write

$$\langle x, y\rangle = \frac{1}{2}\left(\left\|x\right\|_2^2 \left\|y\right\|_2^2 - \left\|x - y\right\|_2^2\right) \tag{19.5}$$

and similarly

$$\langle \Pi x, \Pi y\rangle = \frac{1}{2}\left(\left\|\Pi x\right\|_2^2 \left\|\Pi y\right\|_2^2 - \left\|\Pi(x - y)\right\|_2^2\right) \tag{19.6}$$

Subtracting (19.5) from (19.6), we get

$$\left|\langle \Pi x, \Pi y\rangle - \langle x, y\rangle\right| = \left|\frac{1}{2}\left((\left\|\Pi x\right\|_2^2 - 1) + (\left\|\Pi y\right\|_2^2 - 1) - (\left\|\Pi(x - y)\right\|_2^2 - \left\|x - y\right\|_2^2)\right)\right| \tag{19.7}$$

$$\leq \frac{1}{2}\left|\left\|\Pi x\right\|_2^2 - 1\right| + \frac{1}{2}\left|\left\|\Pi y\right\|_2^2 - 1\right| + \frac{1}{2}\left|\left\|\Pi(x - y)\right\|_2^2 - \left\|x - y\right\|_2^2\right)\right|, \tag{19.8}$$

where the equality comes from the fact that x and y are unit norm vectors. Taking $\|.\|_l$ from both sides in (19.8) yields

$$\begin{aligned}\left\|\left|\langle \Pi x, \Pi y\rangle - \langle x, y\rangle\right|\right\|_l &\leq \frac{1}{2}\left\|\left|\left\|\Pi x\right\|_2^2 - 1\right|\right\|_l + \frac{1}{2}\left\|\left|\left\|\Pi y\right\|_2^2 - 1\right|\right\|_l + \frac{1}{2}\left\|\left|\left\|\Pi(x - y)\right\|_2^2 - \left\|x - y\right\|_2^2\right)\right|\right\|_l \\ &\leq \frac{1}{2}(\epsilon \cdot \delta^{\frac{1}{l}} + \epsilon \cdot \delta^{\frac{1}{l}} + \left\|x - y\right\|_2^2 \epsilon \cdot \delta^{\frac{1}{l}}) \\ &\leq 3\epsilon\delta^{\frac{1}{l}},\end{aligned}$$

where the second inequality follows from the the definition of (ϵ, δ, l)-JL moment property of $\mathcal{D}$ and definition of $\|.\|_l$, and the last inequality follows from the fact that x and y are unit norm vectors so $\|x-y\|_2^2 \leq 2^2 = 4$. ■

Now we can prove the main result of this section.

Theorem 1 *Given two matrices $A \in \mathbb{R}^{n\times p}$, $B \in \mathbb{R}^{n\times q}$, and $\epsilon, \delta \in (0, \frac{1}{2})$. Let $\mathcal{D}$ be any distribution over matrices with $k = O(\frac{1}{\epsilon^2}\cdot\frac{1}{\delta})$ rows with (ϵ, δ, l)-JL moment property for $l \geq 2$. Then,*

$$\Pr_{\Pi\sim\mathcal{D}}\left[\left\|(\Pi A)^T(\Pi B)\right\|_F > 3\epsilon \|A\|_F \|B\|_F\right] < \delta.$$

Proof. Recall that $\|X\|_{\frac{l}{2}} = \mathbb{E}[X^{\frac{l}{2}}]^{\frac{2}{l}}$ and it is a norm for $l \geq 2$. Then, we can write

$$\begin{aligned}
\left\|\left\|(\Pi A)^T(\Pi B) - A^TB\right\|_F^2\right\|_{\frac{l}{2}} &= \left\|\sum_{i,j=1}^{n}\left(\langle \Pi A^{(i)}, \Pi B^{(j)}\rangle - \langle A^{(i)}, B^{(j)}\rangle\right)^2\right\|_{\frac{l}{2}} && (19.9)\\
&= \left\|\sum_{i,j}\left\|A^{(i)}\right\|_2^2\left\|B^{(j)}\right\|_2^2\left(\langle \Pi\frac{A^{(i)}}{\left\|A^{(i)}\right\|}, \Pi\frac{B^{(j)}}{\left\|B^{(j)}\right\|}\rangle - \langle\frac{A^{(i)}}{\left\|A^{(i)}\right\|}, \frac{B^{(j)}}{\left\|B^{(j)}\right\|}\rangle\right)^2\right\|_{\frac{l}{2}} && (19.10)\\
&\leq \sum_{i,j}\left\|A^{(i)}\right\|_2^2\left\|B^{(j)}\right\|_2^2\left\|\left(\langle \Pi\frac{A^{(i)}}{\left\|A^{(i)}\right\|}, \Pi\frac{B^{(j)}}{\left\|B^{(j)}\right\|}\rangle - \langle\frac{A^{(i)}}{\left\|A^{(i)}\right\|}, \frac{B^{(j)}}{\left\|B^{(j)}\right\|}\rangle\right)^2\right\|_{\frac{l}{2}} && (19.11)\\
&= \sum_{i,j}\left\|A^{(i)}\right\|_2^2\left\|B^{(j)}\right\|_2^2\left\|\langle \Pi\frac{A^{(i)}}{\left\|A^{(i)}\right\|}, \Pi\frac{B^{(j)}}{\left\|B^{(j)}\right\|}\rangle - \langle\frac{A^{(i)}}{\left\|A^{(i)}\right\|}, \frac{B^{(j)}}{\left\|B^{(j)}\right\|}\rangle\right\|_l^2 && (19.12)\\
&\leq (3\epsilon\delta^{\frac{1}{l}})^2\sum_{i,j}\left\|A^{(i)}\right\|_2^2\left\|B^{(j)}\right\|_2^2 && (19.13)\\
&= (3\epsilon\delta^{\frac{1}{l}})^2\|A\|_F^2\|B\|_F^2, && (19.14)
\end{aligned}$$

where (19.11) follows from Minkowski's inequality, (19.12) follows from the fact that $\left\|X^2\right\|_{\frac{l}{2}} = \|X\|_l^2$, and (19.13) follows from Lemma 3.

By definition of $\|.\|_l$, and (19.14) we get that $\mathbb{E}[\left\|(\Pi A)^T(\Pi B) - A^TB\right\|_F^l]^{\frac{2}{l}} \leq (3\epsilon\delta^{\frac{1}{l}})^2\|A\|_F^2\|B\|_F^2$ which implies

$$\mathbb{E}[\left\|(\Pi A)^T(\Pi B) - A^TB\right\|_F^l] \leq (3\epsilon)^l\delta\|A\|_F^l\|B\|_F^l. \quad (19.15)$$

Now we can apply Markov's inequality:

$$\begin{aligned}
\Pr_{\Pi\sim\mathcal{D}}\left[\left\|(\Pi A)^T(\Pi B) - A^TB\right\|_F > 3\epsilon\|A\|_F\|B\|_F\right] &= \Pr_{\Pi\sim\mathcal{D}}\left[\left\|(\Pi A)^T(\Pi B) - A^TB\right\|_F^l > (3\epsilon)^l\|A\|_F^l\|B\|_F^l\right]\\
&\leq \frac{\mathbb{E}[\left\|(\Pi A)^T(\Pi B) - A^TB\right\|_F^l]}{(3\epsilon)^l\|A\|_F^l\|B\|_F^l}\\
&\leq \frac{(3\epsilon)^l\delta\|A\|_F^l\|B\|_F^l}{(3\epsilon)^l\|A\|_F^l\|B\|_F^l}\\
&= \delta,
\end{aligned}$$

where the last inequality follows from (19.15). ■

Some comments:

1. This algorithm only requires one pass over A and B.

2. We can construct distributions with $(\epsilon, \delta, 2)$-JL moment such that matrices in this distribution has $k = O(\frac{1}{\epsilon^2} \cdot \frac{1}{\delta})$ rows [TZ12], i.e., $\Pi \in \mathbb{R}^{k \times n}$. So the running time of the algorithm is based on computing ΠA, ΠB, and $(\Pi A)^T(\Pi B)$ so the running time is $O(knp + knq + kpq)$.

References

S69 V. Strassen, Gaussian elimination is not optimal. *Numerische mathematik*, 13.4 (1969): 354-356.

DKM06 P. Drineas, R. Kannan, and M. W. Mahoney, Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, 36.1 (2006): 132-157.

S06 T. Sarlos, Improved approximation algorithms for large matrices via random projections. *FOCS'06*, 2006.

KN14 D. M. Kane and J. Nelson, Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61.1 (2014): 4.

TZ12 M. Thorup, and Y. Zhang, Tabulation-based 5-independent hashing with applications to linear probing and second moment estimation. *SIAM Journal on Computing*, 41.2 (2012): 293-331.