Special Topics in Theoretical Computer Science March 11, 2019

Lecture 15: Matrix Product using JL, Subspace Embedding

Lecturer: Huy Lê Nguyễn *Scribe: Xuangui Huang*

Last time we saw how to approximately compute matrix product using sampling. We also started discussion on using JLMP to approximately compute matrix product. In this lecture we will see its complete proof, and we will see how to construct distributions of sparse embedding matrices satisfying JLMP. Besides, we will consider a variant of sparse embedding matrices for subspaces, and use it to solve least square regression.

1 Approximate Matrix Product using JLMP

Definition 1. *Let D be a distribution over matrices $\Pi \in \mathbb{R}^{m\times n}$. We say that D satisfies (ε,δ,p)* Johnson-Lindenstrauss Moment Property *((ε,δ,p)*-JLMP*) if for any unit vector x we have*

$$\mathbb{E}_{\Pi\sim D}\left[\left|\|\Pi x\|_2^2-1\right|^p\right]\le\varepsilon^p\delta.$$

Last time we proved that applying Π with JLMP to vectors will keep their inner products approximately.

Lemma 1. *Suppose Π comes from D with (ε,δ,p)-JLMP for $p\ge 1$, then for any unit vectors x and y we have $\mathbb{E}_{\Pi\sim D}[|\langle\Pi x,\Pi y\rangle-\langle x,y\rangle|^p]\le(2\varepsilon)^p\,\delta$.*

Now we are going to prove that matrix product is approximately preserved with high probability after applying Π with JLMP. Therefore instead of calculating $A^\top B$ we can calculate $(\Pi A)^\top\,\Pi B$, which will speed up our calculation since the dimensions of ΠA and ΠB can be much smaller than those of A and B.

Theorem 1. *Suppose Π comes from D with (ε,δ,p)-JLMP for $p\ge 2$, then for any matrices $A\in\mathbb{R}^{n\times a}$ and $B\in\mathbb{R}^{n\times b}$, we have*

$$\Pr_{\Pi\sim D}\left[\left\|A^\top B-(\Pi A)^\top\,\Pi B\right\|_F\ge 2\varepsilon\,\|A\|_F\,\|B\|_F\right]\le\delta.$$

Proof. Let a_i be the i-th column of A for $i\in\{1,\dots,a\}$, and b_j be the j-th column of B for $j\in\{1,\dots,b\}$. Let $M=A^\top B-(\Pi A)^\top\,\Pi B$, then the entry $M_{i,j}$ on the i-th row, j-th column is

$$\begin{aligned}M_{i,j}&=\langle a_i,b_j\rangle-\langle\Pi a_i,\Pi b_j\rangle\\&=\|a_i\|\,\|b_j\|\left(\left\langle\frac{a_i}{\|a_i\|},\frac{b_j}{\|b_j\|}\right\rangle-\left\langle\Pi\frac{a_i}{\|a_i\|},\Pi\frac{b_j}{\|b_j\|}\right\rangle\right)\\&=\|a_i\|\,\|b_j\|\,X_{i,j},\end{aligned}$$

where we define $X_{i,j} = \left\langle \frac{a_i}{\|a_i\|}, \frac{b_j}{\|b_j\|} \right\rangle - \left\langle \Pi \frac{a_i}{\|a_i\|}, \Pi \frac{b_j}{\|b_j\|} \right\rangle$. Note that $\frac{a_i}{\|a_i\|}$ and $\frac{b_j}{\|b_j\|}$ are unit vectors so we can apply the above lemma to them.

By Markov inequality, we have

$$\begin{aligned}\Pr_{\Pi}\left[\left\|A^{\top}B - (\Pi A)^{\top}\Pi B\right\|_F \geq 2\varepsilon \|A\|_F \|B\|_F\right] &\leq \Pr_{\Pi}[\|M\|_F^p \geq (2\varepsilon \|A\|_F \|B\|_F)^p] \\ &\leq \frac{\mathbb{E}_{\Pi}[\|M\|_F^p]}{(2\varepsilon)^p \|A\|_F^p \|B\|_F^p}. \qquad (1)\end{aligned}$$

To bound $\mathbb{E}_{\Pi}[\|M\|_F^p]$, notice that it is $\mathbb{E}_{\Pi}\left[\left(\sum_{i,j} M_{i,j}^2\right)^{\frac{p}{2}}\right]$, then we raise it to the power of $\frac{2}{p}$:

$$\begin{aligned}\left(\mathbb{E}_{\Pi}\left[\left(\sum_{i,j} M_{i,j}^2\right)^{\frac{p}{2}}\right]\right)^{\frac{2}{p}} &\leq \sum_{i,j}\left(\mathbb{E}_{\Pi}\left[(M_{i,j}^2)^{\frac{p}{2}}\right]\right)^{\frac{2}{p}} && \text{(triangle inequality of } \tfrac{p}{2}\text{-norm)} \\ &= \sum_{i,j}\left(\|a_i\|^p \|b_j\|^p |X_{i,j}|^p\right)^{\frac{2}{p}} \\ &= \sum_{i,j}\|a_i\|^2 \|b_j\|^2 \left(\mathbb{E}_{\Pi}[|X_{i,j}|^p]\right)^{\frac{2}{p}} \\ &\leq \sum_{i,j}\|a_i\|^2 \|b_j\|^2 \left((2\varepsilon)^p\delta\right)^{\frac{2}{p}} && \text{(by Lemma 1)} \\ &= (2\varepsilon)^2\delta^{\frac{2}{p}} \|A\|_F^2 \|B\|_F^2 .\end{aligned}$$

Note here we need $p \geq 2$ so we have $\frac{p}{2}$-norm and can use the triangle inequality in the first step.

Therefore we have $\mathbb{E}_{\Pi}[\|M\|_F^p] = \mathbb{E}_{\Pi}\left[\left(\sum_{i,j} M_{i,j}^2\right)^{\frac{p}{2}}\right] \leq (2\varepsilon)^p\delta \|A\|_F^p \|B\|_F^p$, applying it to Equation (1) we get the result. □

2 Sparse Embedding Matrix for JLMP

Last time we mentioned that random matrices with i.i.d. Gaussian entries satisfies $(\varepsilon, \delta, \log\frac{1}{\delta})$-JLMP with $m = \Theta\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$. As such matrices are dense, the total running time for calculating $(\Pi A)^{\top}\Pi B$ might be as large as $O\left(ab\frac{1}{\varepsilon^2}\log\frac{1}{\delta} + (a+b)n\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ for $A \in \mathbb{R}^{n\times a}$ and $B \in \mathbb{R}^{n\times b}$. The first term is basically unavoidable for multiplying a dense $a \times m$ matrix with a dense $m \times b$ matrix. Since $n \gg m$, we expect ΠA and ΠB to be dense. But for the second term, we can improve the running time of calculating ΠA and ΠB by constructing sparse Π.

Consider a distribution of sparse embedding matrices $\Pi \in \mathbb{R}^{m\times n}$ in which each column has a single non-zero entry, generated by a pair-wise independent hash function $h\colon [n] \to [m]$ and a random function $\sigma\colon [n] \to \{\pm 1\}$, where for each $i \in [n]$, $h(i)$

is the row of the non-zero element of the i-th column and $\sigma(i)$ is the value of that element.

We will show that for $m = \Theta\left(\frac{1}{\varepsilon^2\delta}\right)$, this distribution satisfies $(\varepsilon, \delta, 2)$-JLMP. Note that Π is basically a linear sketch matrix, so the calculation of ΠA only takes time $O(\mathrm{nnz}(A))$, linear in the number of non-zeroes in A.

Claim 1. *For the distribution D of matrices Π we describe above with $m = \Theta\left(\frac{1}{\varepsilon^2\delta}\right)$, for any unit vector x we have*

$$\mathbb{E}_{\Pi\sim D}\left[\left|\|\Pi x\|^2 - 1\right|^2\right] \le \varepsilon^2\delta.$$

Proof. Note that $\mathbb{E}_{\Pi}\left[\left|\|\Pi x\|^2 - 1\right|^2\right] = \mathbb{E}_{\Pi}\left[\|\Pi x\|^4\right] - 2\mathbb{E}_{\Pi}\left[\|\Pi x\|^2\right] + 1$. We can bound the first and second terms similarly as what we've done before.

$$\begin{aligned}
\mathbb{E}_{\Pi}\left[\|\Pi x\|^2\right] &= \sum_{i=1}^{m} \mathbb{E}_{\Pi}\left[\left(\sum_{j:h(j)=i} x_j\sigma(j)\right)^2\right] \\
&= \sum_{i=1}^{m} \sum_{j_1,j_2:h(j_1)=h(j_2)=i} \mathbb{E}_{\Pi}[x_{j_1}x_{j_2}\sigma(j_1)\sigma(j_2)] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{1}{m}x_j^2 \\
&= \|x\|^2 \\
&= 1,
\end{aligned}$$

where the first step comes from linearity of expectation, and the third step comes from the fact that when $j_1 \neq j_2$ we have $\sigma(j_1)$ independent of $\sigma(j_2)$ so the expectation would be 0, thus the remaining term is $\sum_{j:h(j)=i} \mathbb{E}\left[x_j^2\sigma(j)^2\right] = \sum_{j=1}^{n} \frac{1}{m}x_j^2$.

$$\begin{aligned}
\mathbb{E}_{\Pi}\left[\|\Pi x\|^4\right] &= \sum_{i=1}^{m} \mathbb{E}_{\Pi}\left[\left(\sum_{j:h(j)=i} x_j\sigma(j)\right)^4\right] \\
&= \sum_{i=1}^{m} \sum_{\substack{j_1,j_2,j_3,j_4: \\ h(j_k)=i,\forall k\in[4]}} \mathbb{E}[x_{j_1}x_{j_2}x_{j_3}x_{j_4}\sigma(j_1)\sigma(j_2)\sigma(j_3)\sigma(j_4)] \\
&= \sum_{i=1}^{m}\left(\sum_{j:h(j)=i} \mathbb{E}\left[x_j^4\right] + 3\sum_{\substack{j_1\neq j_2: \\ h(j_1)=h(j_2)=i}} \mathbb{E}\left[x_{j_1}^2 x_{j_2}^2\right]\right) \\
&= \sum_{i=1}^{m}\left(\sum_{j=1}^{n}\frac{1}{m}x_j^4 + \sum_{j_1\neq j_2}\frac{3}{m^2}x_{j_1}^2x_{j_2}^2\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{m} \sum_{j=1}^{n} x_j^2 \sum_{k=1}^{n} x_k^2 \left(\frac{1}{m} + \frac{2}{m^2}\right) \\
&= 1 + \frac{2}{m},
\end{aligned}
$$

where the third step comes from the observation that whenever an element occurs an odd number of times, the expectation is 0.

Therefore we have $\mathbb{E}_\Pi\left[\left|\|\Pi x\|^2 - 1\right|^2\right] \leq 1 + \frac{2}{m} - 2 + 1 = \frac{2}{m} = \varepsilon^2\delta$ if we set $m = \Theta\left(\frac{1}{\varepsilon^2\delta}\right)$. □

3 Subspace Embedding

Now we look at a slightly different problem, where we only require Π to preserve the length of vectors in a specific linear subspace.

Definition 2. *For a linear subspace $E \subseteq \mathbb{R}^n$, we say Π is ε-*subspace embedding *for E if for any unit vector $x \in E$, we have $\left|\|\Pi x\|^2 - 1\right| \leq \varepsilon$.*

Suppose $\dim(E) = d$. Let U be an orthonormal basis of E, i.e. $U^\top U = I$, $U \in \mathbb{R}^{n\times d}$, and $E = \{x | x = Uz, z \in \mathbb{R}^d\}$. Then for any unit vector $x \in E$ we have $\|\Pi x\|^2 = \|(\Pi U)\, z\|^2$ for some vector unit vector $z \in \mathbb{R}^d$. We have the following fact.

Fact 2. *Suppose $\sigma_1 \geq \sigma_2 \geq \ldots$ are singular values of the matrix M, then $\|M\|_2 = \sigma_1 = \max_{z:\|z\|=1} z^\top M z$, and $\|M\|_F = \sqrt{\sum_i \sigma_i^2}$.*

Therefore we have

$$
\begin{aligned}
\max_{x\in E:\|x\|=1} \left|\|\Pi x\|^2 - 1\right| &= \max_{z:\|z\|=1} \left|\|(\Pi U)\, z\|^2 - 1\right| \\
&= \max_{z:\|z\|=1} \left|(\Pi u z)^\top \Pi U z - 1\right| \\
&= \max_{z:\|z\|=1} z^\top \left((\Pi U)^\top \Pi U - I\right) z \\
&= \left\|(\Pi U)^\top \Pi U - I\right\|_2,
\end{aligned}
$$

thus the condition in the above definition is equivalent to $\left\|(\Pi U)^\top \Pi U - I\right\|_2 \leq \varepsilon$. To find such matrix Π, it is sufficient to find Π such that $\left\|(\Pi U)^\top \Pi U - I\right\|_F \leq \varepsilon$.

From the previous sections we know that there is a distribution D of matrices $\Pi \in \mathbb{R}^{m\times n}$ with $(\varepsilon', \delta, 2)$-JLMP for m and ε' to be determined, such that

$$
\Pr\left[\left\|(\Pi U)^\top \Pi U - U^\top U\right\|_F \geq 2\varepsilon' \|U\|_F^2\right] \leq \delta,
$$

which is equivalent to $\Pr\left[\left\|(\Pi U)^\top \Pi U - I\right\|_F \geq 2\varepsilon' d\right] \leq \delta$. By setting $\varepsilon' = \varepsilon/2d$ and $m = \Theta(d^2/\varepsilon^2\delta)$, we get what we want: a distributional version of ε-subspace embedding. Note that the choice D only depends on d (in addition to ε and δ), i.e. it works for all linear subspaces E of the same dimension d.

Our Π comes from the previous section, so multiply it with matrix A only takes time $O(\mathrm{nnz}(A))$. An i.i.d. Gaussian Π will make the number of rows m only linear in d, but as it is dense it would be slower to multiply it with other matrices.

4 Ordinary Least Square Regression

In the least square regression problem, we are given an $X \in \mathbb{R}^{n\times d}$ representing n examples with d features, and a vector $y \in \mathbb{R}^n$. Usually we will have $n \gg d$. The goal is find the best parameter $\beta^{LS} = \mathrm{argmin}_{\beta\in\mathbb{R}^d} \|X\beta - y\|_2^2$.

There is an analytical solution for this problem: suppose $X^\top X$ is invertible, then $\beta^{LS} = \left(X^\top X\right)^{-1} X^\top y$. This method takes $O(nd^2)$ time, and $O(nd^{\omega-1})$ if we use fast matrix multiplication, which is too slow in practice.

Let E be the span of columns of X and y. Then we have $\dim(E) \leq d+1$. We can use our previous result for subspace embedding to speed up the calculation.

Claim 2. *If Π is an ε-subspace embedding for E, then with $\widetilde{\beta} = \mathrm{argmin}_\beta \|(\Pi X)\beta - \Pi y\|_2^2$, we have*

$$\left\|X\widetilde{\beta} - y\right\|_2^2 \leq \frac{1+\varepsilon}{1-\varepsilon} \left\|X\beta^{LS} - y\right\|_2^2.$$

Proof. By optimality of $\widetilde{\beta}$ and ε-subspace embedding property, we have

$$\left\|\Pi X\widetilde{\beta} - \Pi y\right\|^2 \leq \left\|\Pi X\beta^{LS} - \Pi y\right\|^2 \leq (1+\varepsilon)\left\|X\beta^{LS} - y\right\|^2.$$

On the other hand we also have $\left\|\Pi X\widetilde{\beta} - \Pi y\right\|^2 \geq (1-\varepsilon)\left\|X\widetilde{\beta} - y\right\|^2$. □

Therefore our strategy is to calculate ΠX and Πy first then solve the least square regression of reduced size using the analytical solution. This takes time $O(d{\cdot}\mathrm{nnz}(A)/\varepsilon + \mathrm{poly}(d, 1/\varepsilon))$ if we use Gaussian Π, and only $O(\mathrm{nnz}(A) + \mathrm{poly}(d, 1/\varepsilon))$ if we use the Π from the last section (but the $\mathrm{poly}(d, 1/\varepsilon)$ factor would be larger).