Matrix Multiplication using sampling.

Suppose A is $m \times n$ matrix
B $n \times p$ matrix.

Goal: Find AB.

running time $= O(mnp)$.

Q: Can we do faster?

$\Rightarrow$ we will use sampling to get an approximate product faster than the traditional multiplication.

Let $A(:,k)$ be kth column of A.
$\rightsquigarrow$ $m \times 1$ matrix.

$B(k,:)$ be kth row of B.
$\rightsquigarrow$ $1 \times p$ matrix.

Then,

$$AB = \sum_{k=1}^{n} A(:,k)\,B(k,:) \qquad (\text{Why?})$$

Observe that for each k,
$A(:,k)B(k,:)$ is an $m \times p$ matrix
each element of it is a single product of elements of A and B.

~~Define a random variable Z that takes values in $\{1,2,\dots,n\}$. and $\mathbb{P}(Z=k) =: p_k$~~

~~will choose p_k later.~~

$$\sum_{k=1}^{n} p_k = 1$$

$$\sum_{k=1}^{n} p_k = 1$$

Define a random (matrix) variable

$$X = \frac{1}{p_k} A(:,k)B(k,:) \quad \text{with probability } \overset{\text{will choose later.}}{p_k}.$$

Fact: $\mathbb{E}[X] = AB$.

proof. $\mathbb{E}[X] = \sum_{k=1}^{n} p_k \cdot \frac{1}{p_k} A(:,k)B(k,:) = AB$.

We are interested in bounding

$\mathrm{Var}[X] = \mathbb{E}(\|AB - X\|_F^2)$.

$$\text{Lemma: } \mathrm{Var}[X] = \sum_{i=1}^{m}\sum_{j=1}^{p} \mathrm{Var}[\overset{ij\text{th entry of matrix } X.}{X_{ij}}] = \sum_k \frac{1}{p_k}\|A(:,k)\|_2^2 \|B(k,:)\|_2^2 - \|AB\|_F^2.$$

proof: Exercise!

What is the best choice of p_k to minimize $\mathrm{Var}[X]$, i.e. to minimize $\sum_k \frac{1}{p_k}\|A(:,k)\|_2^2\,\|B(k,:)\|_2^2$ as $\|AB\|_F^2$ is fixed?

Exercise: Suppose $c_1, c_2, \ldots, c_n$ are nonnegative. Show that the minimum of $\sum_{k=1}^{n} \frac{c_k}{p_k}$ subject to the constraints $p_k \geq 0$ and $\sum_k p_k = 1$ is attained when p_k is proportional to $\sqrt{c_k}$.

Length squared sampling techniques.

$$\text{pick} \quad p_k = \frac{\|A(:,k)\|_2^2}{\|A\|_F^2}$$

$$\mathbb{E}\left[\|AB - X\|_F^2 \right] = Var[X] \leq \|A\|_F^2 \sum_k \|B(k,:)\|_2^2 = \|A\|_F^2 \|B\|_F^2$$

$$\cancel{\Rightarrow} \; Var[X] = \mathbb{E}\left[\|AB - X\|_F^2 \right] \leq \|A\|_F^2 \|B\|_F^2 .$$

How to reduce the variance?

$\Rightarrow$ mean of estimators.

Consider s independent trials of X, says $X_1, X_2, \dots, X_s$.

and take $\frac{1}{s} \sum_{i=1}^{s} X_i$ as our estimate of AB.

$$\Rightarrow \quad Var\left[\frac{1}{s} \sum_{i=1}^{s} X_i \right] = \frac{1}{s} Var[X] \leq \frac{1}{s} \|A\|_F^2 \|B\|_F^2 .$$

Let ~~X~~ $k_1, k_2, \dots, k_s$ be the ~~column~~ index k's chosen in each trial. Then.

$$\frac{1}{s} \sum_{i=1}^{s} X_i = \frac{1}{s} \left(\frac{A(:,k_1) B(k_1,:)}{p_{k_1}} + \frac{A(:,k_2) B(k_2,:)}{p_{k_2}} + \dots + \frac{A(:,k_s) B(k_s,:)}{p_{k_s}} \right)$$

Let $C = \left[\frac{A(:,k_1)}{\sqrt{s\, p_{k_1}}} \quad \frac{A(:,k_2)}{\sqrt{s\, p_{k_2}}} \quad \dots \quad \frac{A(:,k_s)}{\sqrt{s\, p_{k_s}}} \right]$

the $m \times s$ matrix consisting of columns which are scaled versions of the chosen columns of A.

~~Ca~~ we can show that ~~$\mathbb{E}[C] =$~~

$$\mathbb{E}[C C^T] = A A^T .$$

Let $R = \begin{bmatrix} — \dfrac{B(k_1,:)}{\sqrt{s p_{k_1}}} — \\ — \dfrac{B(k_2,:)}{\sqrt{s p_{k_2}}} — \\ \vdots \\ — \dfrac{B(k_s,:)}{\sqrt{s p_{k_s}}} — \end{bmatrix}$

we can also show that $\mathbb{E}[R^T R] = B^T B$.

$$\Rightarrow \quad \frac{1}{s}\sum_{i=1}^{s} X_i = CR.$$

$$\underbrace{\begin{bmatrix} \dfrac{A_k(:,k_1)}{\sqrt{s p_{k_1}}} & \dfrac{A(:,k_2)}{\sqrt{s p_{k_2}}} & \cdots & \dfrac{A(:,k_s)}{\sqrt{s p_{k_s}}} \end{bmatrix}}_{C} \underbrace{\begin{bmatrix} — \dfrac{B(k_1,:)}{\sqrt{s p_{k_1}}} — \\ — \dfrac{B(k_2,:)}{\sqrt{s p_{k_2}}} — \\ \vdots \\ — \dfrac{B(k_s,:)}{\sqrt{s p_{k_s}}} — \end{bmatrix}}_{R} \approx AB.$$

Thm: Suppose A is an $m \times n$ matrix and B $n \times p$ matrix. Then AB can be estimated by CR. The error is bounded by

$$(*) \qquad \mathbb{E}\left(\|AB - CR\|_F^2 \right) \leq \frac{\|A\|_F^2 \, \|B\|_F^2}{s}.$$

Thus, to ensure $\mathbb{E}\left(\|AB - CR\|_F^2\right) \leq \varepsilon^2 \|A\|_F^2 \|B\|_F^2$, take $s \geqslant \dfrac{1}{\varepsilon^2}$.

Note: the multiplication CR can be carried out in time $O(mp)$ if ε is large enough.

Q: When is the error bound (*) good and when is it not? (still an open problem ~~for~~ in general.)

Let's consider $B = A^T$.

1) $A = I$, then the (*) is not very good.

In this case,

$$\|AB\|_F^2 = \|AA^T\|_F^2 = \|II^T\|_F^2 = \|I\|_F^2 = n$$

But RHS of (*) $= \dfrac{\|A\|_F^2 \ \|B\|_F^2}{s} = \dfrac{\|I\|_F^2 \ \|I^T\|_F^2}{s} = \dfrac{n^2}{s}$.

$\Rightarrow$ We would need $s > n$ for the bound to be better than approximating AB by the zero matrix.

2) For general A.

Suppose that $\sigma_1, \sigma_2, \ldots$ are singular values of A.

$\Rightarrow \quad \sigma_1^2, \sigma_2^2, \ldots$ $\qquad \underset{AB}{\underbrace{AA^T}}$.

and $\|A\|_F^2 = \sum_i \sigma_i^2$ $\qquad \|AA^T\|_F^2 = \sum_i \sigma_i^4$.

$\|A^T\|_F = \sum_i \sigma_i^2$

$$(*) \Rightarrow \mathbb{E}\left(\|AA^T - CR\|_F^2\right) \leq \frac{\|A\|_F^2 \ \|A^T\|_F^2}{s} = \frac{\left(\sum_i \sigma_i^2\right)^2}{s}.$$

if $s \geq \dfrac{(\sigma_1^2 + \sigma_2^2 + \ldots)^2}{\sigma_1^4 + \sigma_2^4 + \ldots} = \dfrac{\|A\|_F^2 \ \|A^T\|_F^2}{\|AA^T\|_F^2}$, then

$$\mathbb{E}\left(\|AA^T - CR\|_F^2\right) \leq \|AA^T\|_F^2.$$

If $\text{rank}(A) = r$, then by Cauchy-Schwartz inequality.

$$(\sigma_1^2 + \sigma_2^2 + \dots + \sigma_r^2)^2 \leq (\sigma_1^4 + \sigma_2^4 + \dots + \sigma_r^4) \cdot r$$

$$\Rightarrow \frac{(\sigma_1^2 + \dots + \sigma_r^2)^2}{(\sigma_1^4 + \dots + \sigma_r^4)} \leq r.$$

Hence, in general s need to be at least r.

If A is full rank, this means samping will not gain us anything over taking the whole matrix!

(Matrix)

* Elementwise sampling.

Goal: Given a matrix A, find another matrix B such that $\|A-B\|$ is small and that B is much sparser than A.

"sparse matrix" = matrix with a lot of zero entries.

Consider any $m\times n$ matrix A.
Let $A_{i,j}$ be the $m\times n$ matrix whose entries are all zeros except entry (i,j) which is set to $a_{i,j}$.

E.g. $A=\begin{bmatrix}1&2&3\\4&5&6\end{bmatrix}$.

$$A_{11}=\begin{bmatrix}1&0&0\\0&0&0\end{bmatrix} \qquad A_{12}=\begin{bmatrix}0&2&0\\0&0&0\end{bmatrix}.$$

$A=A_{11}+A_{12}+A_{13}+A_{21}+A_{22}+A_{23}$.

$\Rightarrow$ For any $m\times n$ matrix

$$A=\sum_{i,j}A_{ij}.$$

Let's ~~defi~~ define a matrix valued random variables $\cancel{B}$ as follows.

$$\cancel{B}=\frac{1}{p_{ij}}A_{ij} \quad \text{with probability } p_{ij}.$$

then

$$E[\cancel{B}]=\sum_{i,j}\underbrace{P\left(\cancel{B}=\frac{1}{p_{ij}}A_{ij}\right)}_{p_{ij}}\cdot\frac{1}{p_{ij}}A_{ij}=\sum_{i,j}A_{ij}=A.$$

Since we cannot hope to approximate A with a matrix with only 1 non zero.

$$A = \begin{bmatrix} x & x & x & x & x \\ x & x & x & x & x \\ x & x & x & x & x \\ x & x & x & x & x \end{bmatrix} \rightsquigarrow \text{Find } B = \begin{bmatrix} & & x & \\ x & & x & x \\ & x & x & \\ x & & x & \end{bmatrix}$$

Consider $B_1, \dots, B_s$ independent copies of X.

let $B = \frac{1}{s} \sum_{k=1}^{s} B_k$

Then $\mathbb{E}[B] = \frac{1}{s} \sum_{k=1}^{s} \mathbb{E}[B_k] = \frac{1}{s} \sum_{k=1}^{s} \mathbb{E}[X] = \mathbb{E}[X] = A.$

$$\Rightarrow \quad \mathbb{E}[B] = A.$$

We would like to show that B is close to ~~its~~ mean A with high probability.

Generalized version of Chernoff's

Lemma: (Matrix Bernstein inequality).

Let $X_1, \dots, X_s$ be independent $m \times n$ matrix valued random variables such that

$$\mathbb{E}[X_k] = 0 \quad \text{and} \quad \|X_k\|_2 \leq R \quad \text{for } k = 1, \dots, s.$$

Set $\sigma^2 = \max\left\{ \left\| \sum_{k=1}^{s} \mathbb{E}[X_k X_k^T] \right\|_2, \left\| \sum_{k=1}^{s} \mathbb{E}[X_k^T X_k] \right\|_2 \right\}$. Then

$$\mathbb{P}\left(\left\| \sum_{k=1}^{s} X_k \right\|_2 > t \right) \leq (m+n)\, e^{-\frac{t^2}{\sigma^2 + Rt/3}}.$$

(Candès – Recht '12)

Tropp '15.

To use the lemma, we need to write $B-A$ in term of a sum of mean zero matrices.

$$B - A = \frac{1}{s}\sum_{k=1}^{s} B_k - A = \sum_{k=1}^{s} \frac{(B_k - A)}{s}.$$

Let $X_k = \frac{B_k - A}{s}$.

Then $\mathbb{E}[X_k] = \mathbb{E}\left[\frac{B_k - A}{s}\right] = \frac{\mathbb{E}[B_k] - A}{s} = 0$.

~~Let~~ Denote $|A|_1 = \sum_{(i,j)} |a_{i,j}|$ sum of absolute value of all entries of A.

Set $p_{i,j} = \frac{|a_{i,j}|}{|A|_1}$. $\Rightarrow \sum_{(i,j)} p_{i,j} = \sum_{(i,j)} \frac{|a_{i,j}|}{|A|_1} = 1$.

Let's bound $R = \max_k \|X_k\|_2$

$$\max_k \|X_k\|_2 = \max_k \left\|\left(\frac{A_{i,j}}{p_{i,j}} - A\right)/s\right\|_2$$

for any $m \times n$ matrix M and N, $\|M+N\|_2 \le \|M\|_2 + \|N\|_2$

$$\le \max_k \left\|\frac{A_{i,j}}{p_{i,j}}\right\|_2 \cdot \frac{1}{s} + \frac{\|A\|_2}{s}$$

$$= \max_k \left\|\frac{A_{i,j}}{|a_{i,j}|} \cdot |A|_1\right\|_2 \cdot \frac{1}{s} + \frac{\|A\|_2}{s}$$

$$\frac{A_{i,j}}{|a_{i,j}|} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & \frac{a_{i,j}}{|a_{i,j}|} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$\le \frac{|A|_1}{s} + \frac{\|A\|_2}{s}$$

$$\Rightarrow \cancel{R \le} \max_k \|X_k\|_2 \le \frac{|A|_1}{s} + \frac{\|A\|_2}{s}$$

Now we compute σ^2.

$$\Big\|\sum_k \mathbb{E}[X_k X_k^T]\Big\|_2 = \Big\| \sum_k \mathbb{E}\Big[\frac{(B_k - A)(B_k - A)^T}{s^2}\Big]\Big\|_2$$

$$= \Big\| \sum_k \mathbb{E}\, \frac{B_k B_k^T - B_k A^T - A B_k^T + AA^T}{s^2}\Big\|_2 .$$

$\mathbb{E}[B_k A^T] = \mathbb{E}[B_k] A^T = AA^T.$

$\mathbb{E}[A B_k^T] = AA^T.$

$$= \Big\| \sum_k \frac{\mathbb{E}[B_k B_k^T] \mp AA^T}{s^2}\Big\|_2$$

$$= \Big\| \frac{\mathbb{E}[B_k B_k^T] - AA^T}{s}\Big\|_2 .$$

$$\le \frac{\|\mathbb{E}[B_k B_k^T]\|_2}{s} + \frac{\|AA^T\|_2}{s} .$$

$$= \frac{\|\mathbb{E}[B_k B_k^T]\|_2}{s} + \frac{\|A\|_2^2}{s}$$

To compute $\mathbb{E}[B_k B_k^T]$, we observe that

Recall that $B_k = \dfrac{1}{p_{i,j}} A_{i,j}$ with probability $p_{i,j}$.

the (i,j) entry is $a_{i,j}$, all other entries are zero.

$$\Rightarrow \quad B_k B_k^T = \frac{1}{p_{i,j}^2} A_{i,j} A_{i,j}^T \quad \text{with probability } p_{i,j} .$$

$$= \frac{1}{p_{i,j}^2} \begin{bmatrix} & 0 & \\ 0 & a_{i,j} & 0 \\ & 0 & \end{bmatrix} \begin{bmatrix} & 0 & \\ 0 & a_{i,j} & 0 \\ & 0 & \end{bmatrix}$$

$$= |A|_1^2 \, \Delta \begin{bmatrix} & 0 & \\ 0 & 1 & 0 \\ & 0 & \end{bmatrix} \begin{bmatrix} & 0 & \\ 0 & 1 & 0 \\ & 0 & \end{bmatrix} .$$

$$\Rightarrow \quad B_k B_k^T = |A|_1^2 \, E_{i,i} \quad \text{with prob. } p_{i,j}.$$

where $E_{i,i}$ is a matrix such that (i,i) entry is 1 and other entries are 0.

$$\therefore \quad \mathbb{E}[B_k B_k^T] = \sum_{i,j} \mathbb{P}\left(B_k B_k^T = |A|_1^2 E_{i,i}\right) \cdot |A|_1^2 E_{i,i}$$

$$= \sum_{i,j} p_{i,j} \, |A|_1^2 \, E_{i,i}$$

$$= \sum_{i,j} \frac{a_{i,j}}{|A|_1} \, |A|_1^2 \, E_{i,i}$$

$$= |A|_1 \sum_{i,j} a_{i,j} E_{i,i}$$

$$\Rightarrow \quad \| \mathbb{E}[B_k B_k^T] \|_2 = |A|_1 \left\| \sum_{i,j} a_{i,j} E_{i,i} \right\|_2$$

$$= |A|_1 \left\| \sum_i \left(\sum_j a_{i,j} \right) E_{i,i} \right\|_2$$

$$\leq |A|_1 \max_i \sum_j |a_{i,j}|$$

$$= |A|_1 \, \|A\|_1 .$$

$$\therefore \quad \mathbb{P}\left(\|A - B\| > t \right) \leq (m+n) \, e^{-\frac{s t^2}{|A|_1 \|A\|_1 + \|A\|_2^2 + |A|_1 t/3 + \|A\|_2 t/3}}$$

Set $t = \varepsilon \|A\|$, and demand a failure probability of at most δ we get

$$s \geq \frac{\log((m+n)/\delta)}{\varepsilon^2} \left(\frac{|A|_1 \, \|A\|_1}{\|A\|_2^2} + 1 + \frac{\varepsilon \|A\|_1}{3 \|A\|_2} + \frac{\varepsilon}{3} \right).$$