Course Project for EE381V Advanced Probability (Fall 2015)

Low-Rank Approximation of Matrix Product in One Pass

Shanshan Wu (sw33323)

Dec, 6th, 2015

1 Problem

We consider the following problem: given two matrices A and B, find a rank-r approximation of their product A^TB. This type of linear algebra problem has many applications in the machine learning and statistics domain. For example, if $A = B$, then this general problem reduces to the well-known problem of finding principal components of a given data matrix. Another example is the low-rank approximation of a co-occurrence matrix A^TB, where A may be a user-by-query matrix and B may be a user-by-ad matrix, so A^TB computes the joint counts for each query-ad pair. As a third example, A^TB can be regarded as a cross-correlation matrix between two sets of variables (e.g., two different genomic datasets). A low-rank approximation of their correlation matrix A^TB can be used as a tool for understanding the association between different variables (e.g., the canonical-correlation analysis [3]).

We assume that the input matrices A and B are very large and stored in disk. Since disk I/O usually dominates the runtime, we prefer pass-efficient algorithms, i.e., algorithms that only require a few constant number of passes over the input matrices. A naive approach for solving the given problem is to first compute the matrix product A^TB, and then perform truncated SVD to get a rank-r approximation. However, this naive approach is quite inefficient because A^TB could be still quite large and cannot fit into memory, and performing standard SVD using iterative methods (e.g., Lanczos algorithm and Arnoldi's algorithm [6]) need at least r passes over A^TB.

A two-pass algorithm for directly computing a low-rank approximation without explicitly computing the product matrix A^TB is proposed by [2], which works as follows. Given $A \in \mathbb{R}^{d\times n}$ and $B \in \mathbb{R}^{d\times n}$, in the first pass, we compute the norm of every column vector of A and B, respectively. We then independently sample each entry (i, j) of A^TB with probability $\min\{1, q_{ij}\}$ defined as

$$q_{ij} := m \cdot (\frac{||A_i||^2}{||A||_F^2} + \frac{||B_j||^2}{||B||_F^2}),$$

where m denotes sampling complexity, i.e., the number of samples that are expected to get, and $||A_i||$ (or $||B_j||$) is the norm of the i-th (or j-th) column vector of A (or B). Let $\Omega \subset [n] \times [n]$ be the set of sample locations obtained from this biased sampling probability. In the second pass, we compute the corresponding sample value $A_i^TB_j$ for every location $(i, j) \in \Omega$. After two passes, we get an incomplete matrix of A^TB with known values only at locations specified by Ω. The next step is to run weighted Alternating Least Square (ALS) procedure on this incomplete matrix to get the low-rank factor matrices. The weighted ALS procedure belongs to matrix completion algorithms, and we will not describe its details in this report as it is irrelevant to our main content.

In this project, our goal is to modify this two-pass algorithm into a one-pass algorithm. The motivation is that if we can compute low-rank approximation of $A^T B$ in one-pass, then the input matrices A and B need not to be stored in disk beforehand, instead, they can possibly come from live data streams. This streaming model of computation as well as the related online algorithms is quite attractive in the big data era. The rest of this report has the following structure. In Section 2, the intuition behind our one-pass algorithm is explained in details. In Section 3 and 4, we will present the main results as well as the proof sketch.

2 Key Idea

Recall that in the two-pass algorithm proposed by [2], the main job of the second pass is to calculate the inner product $A_i^T B_j$ for each location (i,j) specified by Ω. If we can obtain good approximation of these inner products in the first pass, then we would save the second pass. This can be done by using Johnson-Lindenstrauss (JL) transform, which, with high probability, preserves inner products within certain additive error. More formally, we follow the definition from [9]:

Definition 1. *A random matrix $\Pi \in \mathbb{R}^{k\times d}$ forms a Johnson-Lindenstrauss transform with parameters ϵ, δ, f or JLT(ϵ, δ, f) for short, if with probability at least $1-\delta$, for any f-element subset $V \subset \mathbb{R}^d$, for all $v, v' \in V$ it holds that $|\langle \Pi v, \Pi v'\rangle - \langle v, v'\rangle| \leq \epsilon ||v|| \cdot ||v'||$.*

Our one-pass algorithm works as follows. In the first pass, we perform JL transform on the input matrices A and B to get the sketches ΠA and ΠB (assume that the sketches are small enough to fit into memory). We then follow the same procedures as in the two-pass algorithm, except that instead of using a second pass to calculate the inner product $A_i^T B_j$, we now approximate it by $(\Pi A_i)^T(\Pi B_j)$.

There are many ways to construct a JL transform matrix as well as its variations such as fast JL transform [1] and sparse JL transform [5]. In this project, we focus on the simplest construction, whose entries are i.i.d. Gaussian random variables. The following lemma is again taken from [9]:

Lemma 1. *Let $0 < \epsilon, \delta < 1$, and $\Pi \in \mathbb{R}^{k\times d}$ be a random matrix where the entries Π_{ij} are i.i.d. $N(0, 1/k)$ random variables. If $k = \Omega(\log(f/\delta)\epsilon^{-2})$, then Π is a JLT(ϵ, δ, f).*

As indicated by the above lemma, a crucial problem is to determine the sketching dimension k. On one hand, we want to keep a small k so that the sketched matrices can fit into memory. On the other hand, the parameter k controls how much information is lost during sketching, and hence a larger k gives a higher accuracy in approximating the inner products. In the following two sections, we will present the main theorem that we currently have derived. It characterizes the interaction between the sketching dimension k, the sampling complexity m, and the accuracy of the output.

3 Main Result

Let $\Pi \in \mathbb{R}^{k\times d}$ be a random matrix with each entry being i.i.d. $N(0, 1/k)$. Denote $\widetilde{A} = \Pi A$ and $\widetilde{B} = \Pi B$. We independently sample each entry $(i,j) \in [n] \times [n]$ of $\widetilde{A}^T \widetilde{B}$ with probability $\hat{q}_{ij} = \min\{1, q_{ij}\}$ where

$$q_{ij} := m \cdot \left(\frac{||\widetilde{A}_i||^2}{n||\widetilde{A}||_F^2} + \frac{||\widetilde{B}_j||^2}{n||\widetilde{B}||_F^2}\right). \tag{1}$$

Let the output of weighted ALS procedure be $\widehat{A^TB_r}$. Then the following theorem provides a bound for the spectral-norm error $||A^TB - \widehat{A^TB_r}||$ in terms of the number of samples m, the dimension of the random matrix k, and the number of iterations performed T.

Theorem 1. *Let the number of samples m be*

$$m = \frac{C_1}{\gamma} \cdot \left(\frac{||\widetilde{A}||_F^2 + ||\widetilde{B}||_F^2}{||\widetilde{A}^T\widetilde{B}||_F} \right)^2 \cdot \frac{nr^3}{\epsilon^2} \cdot \left(\frac{\widetilde{\sigma}_1}{\widetilde{\sigma}_r} \right)^2 \log(n) \log^2 \left(\frac{||\widetilde{A}||_F + ||\widetilde{B}||_F}{\zeta} \right),$$

where C_1 is some global constant independent of A and B, and $\tilde{\sigma}_i$ is the i-th singular values of $\widetilde{A}^T\widetilde{B}$. Let the number of iterations be $T = \log(\frac{||\widetilde{A}||_F + ||\widetilde{B}||_F}{\zeta})$. Define the maximum stable rank as $\tilde{r} := \max\{\frac{||A||_F^2}{||A||^2}, \frac{||B||_F^2}{||B||^2}\}$. Suppose the random Gaussian matrix $\Pi \in \mathbb{R}^{k \times d}$ has dimension

$$k \geq C_2 \cdot \frac{\max\{\tilde{r}, \log(2n)\} + \log(3/\gamma)}{\epsilon^2}, \tag{2}$$

where C_2 is some global constant independent of A and B, Then with probability at least $1 - \gamma$, the output $\widehat{A^TB_r}$ satisfies

$$||A^TB - \widehat{A^TB_r}|| \leq ||A^TB - (A^TB)_r|| + \epsilon||A^TB - (A^TB)_r||_F + \zeta + \eta||A||||B||,$$

where $\eta := \max\{2\epsilon + \epsilon^2(\tilde{r} + \sqrt{r}),\ \ 2\epsilon + \epsilon^2\tilde{r} + \epsilon^{1.5}\sqrt{2\tilde{r}C_{AB}(\sqrt{r} + \tilde{r})}\}$, and $C_{AB} := \frac{||A^TB||_F}{||A||_F||B||_F} \leq 1$.

4 Proof

We first introduce three lemmas that connect $\widetilde{A}$ and $\widetilde{B}$ with A and B.

Lemma 2. *Let $k = \Omega(\frac{\log(2n/\delta)}{\epsilon^2})$, then with probability at least $1 - \delta$,*

$$(1-\epsilon)||A||_F^2 \leq ||\widetilde{A}||_F^2 \leq (1+\epsilon)||A||_F^2, \quad (1-\epsilon)||B||_F^2 \leq ||\widetilde{B}||_F^2 \leq (1+\epsilon)||B||_F^2,$$
$$||\widetilde{A}^T\widetilde{B} - A^TB||_F \leq \epsilon||A||_F||B||_F.$$

Proof. This is a standard result of JL transformation, e.g., see Definition 2.3 and Theorem 2.1 of [9] and Lemma 6 of [7] . □

Lemma 3. *Let $k = \Theta(\frac{\tilde{r} + \log(1/\delta)}{\epsilon^2})$, where $\tilde{r} = \max\{\frac{||A||_F^2}{||A||^2}, \frac{||B||_F^2}{||B||^2}\}$ is the maximum stable rank, then with probability at least $1 - \delta$,*

$$||\widetilde{A}^T\widetilde{B} - A^TB|| \leq \epsilon||A||||B||.$$

Proof. This follows from a recent paper [4]. □

Lemma 4. *Let $\widetilde{\sigma}_i$ and $\widetilde{\sigma}_i$ be the i-th singular values of $\widetilde{A}^T\widetilde{B}$ and A^TB, respectively. For any integer r such that $1 \leq r \leq n$, we have*

$$\left| \sqrt{\sum_{i=1}^r \sigma_i^2} - \sqrt{\sum_{i=1}^r \widetilde{\sigma}_i^2} \right| \leq \sqrt{r}||\widetilde{A}^T\widetilde{B} - A^TB||.$$

Proof. Since $\widetilde{A}^T\widetilde{B} = A^TB + \widetilde{A}^T\widetilde{B} - A^TB$, we can apply Weyl's inequality [8][1] to get

$$\widetilde{\sigma}_i \le \sigma_i + ||\widetilde{A}^T\widetilde{B} - A^TB||, \quad \forall 1 \le i \le n.$$

Taking squares of both sides and summing over $1 \le i \le r$ gives

$$\sum_{i=1}^{r} \widetilde{\sigma}_i^2 \le \sum_{i=1}^{r} \sigma_i^2 + 2||\widetilde{A}^T\widetilde{B} - A^TB|| \left(\sum_{i=1}^{r} \sigma_i\right) + r||\widetilde{A}^T\widetilde{B} - A^TB||^2 \tag{3}$$

$$\le \sum_{i=1}^{r} \sigma_i^2 + 2||\widetilde{A}^T\widetilde{B} - A^TB||\sqrt{r}\sqrt{\sum_{i=1}^{r} \sigma_i^2} + r||\widetilde{A}^T\widetilde{B} - A^TB||^2 \tag{4}$$

$$\le \left(\sqrt{\sum_{i=1}^{r} \sigma_i^2} + \sqrt{r}||\widetilde{A}^T\widetilde{B} - A^TB||\right)^2. \tag{5}$$

Therefore, we get $\sqrt{\sum_{i=1}^{r} \widetilde{\sigma}_i^2} - \sqrt{\sum_{i=1}^{r} \sigma_i^2} \le \sqrt{r}||\widetilde{A}^T\widetilde{B} - A^TB||$. In the above analysis, the role of $\widetilde{A}^T\widetilde{B}$ and A^TB is interchangeable, so it is also true that $\sqrt{\sum_{i=1}^{r} \sigma_i^2} - \sqrt{\sum_{i=1}^{r} \widetilde{\sigma}_i^2} \le \sqrt{r}||\widetilde{A}^T\widetilde{B} - A^TB||$, and hence the lemma follows. □

Now we are ready to prove the main theorem stated in the previous section.

Proof. Suppose Π is fixed. Because the sampling probability in Eq. (1) depends only on $\widetilde{A}$ and $\widetilde{B}$, the algorithm works as if we were computing a low-rank approximation for $\widetilde{A}^T\widetilde{B}$. Therefore, the original theoretical bound (e.g., Theorem 3.4 of [2]) holds, which is formally stated as follows.

Let the number samples m be

$$m = \frac{C_1}{\gamma} \cdot \left(\frac{||\widetilde{A}||_F^2 + ||\widetilde{B}||_F^2}{||\widetilde{A}^T\widetilde{B}||_F}\right)^2 \cdot \frac{nr^3}{\epsilon^2} \cdot \left(\frac{\widetilde{\sigma}_1}{\widetilde{\sigma}_r}\right)^2 \log(n) \log^2\left(\frac{||\widetilde{A}||_F + ||\widetilde{B}||_F}{\zeta}\right), \tag{6}$$

where $\widetilde{\sigma}_i$ is the i-th singular values of $\widetilde{A}^T\widetilde{B}$. Let $T = \log(\frac{||\widetilde{A}||_F + ||\widetilde{B}||_F}{\zeta})$, then the output $\widehat{A^TB}_r$ satisfies (w.p. $\ge 1 - \gamma/3$):

$$||\widetilde{A}^T\widetilde{B} - \widehat{A^TB}_r|| \le ||\widetilde{A}^T\widetilde{B} - (\widetilde{A}^T\widetilde{B})_r|| + \epsilon||\widetilde{A}^T\widetilde{B} - (\widetilde{A}^T\widetilde{B})_r||_F + \zeta. \tag{7}$$

To get a bound on $||A^TB - \widehat{A^TB}_r||$, we first apply triangle inequality to (7):

$$||A^TB - \widehat{A^TB}_r|| \le ||A^TB - \widetilde{A}^T\widetilde{B}|| + ||\widetilde{A}^T\widetilde{B} - (\widetilde{A}^T\widetilde{B})_r|| + \epsilon||\widetilde{A}^T\widetilde{B} - (\widetilde{A}^T\widetilde{B})_r||_F + \zeta. \tag{8}$$

Note that since Eq. (8) holds with probability at least $1 - \gamma/3$ for any fixed Π, now suppose Π is chosen randomly according to certain distribution, Eq. (8) should also hold with probability at least $1 - \gamma/3$. We will use the randomness over Π in the following proof.

Next we bound $||\widetilde{A}^T\widetilde{B} - (\widetilde{A}^T\widetilde{B})_r||$ and $||\widetilde{A}^T\widetilde{B} - (\widetilde{A}^T\widetilde{B})_r||_F$ in terms of $||A^TB - \widetilde{A}^T\widetilde{B}||$ and $||A^TB - \widetilde{A}^T\widetilde{B}||_F$.

[1] The original Weyl's inequality applies to Hermitian matrices. For a general matrix M, we can construct a Hermitian matrix $\begin{bmatrix} 0 & M \\ M^* & 0 \end{bmatrix}$ which has eigenvalues $\pm\sigma_1(M), \pm\sigma_2(M), ..., \pm\sigma_n(M)$. More results about Weyl's inequality can be found from [8].

To bound $||\widetilde{A}^T\widetilde{B} - (\widetilde{A}^T\widetilde{B})_r||$, observe that $\widetilde{A}^T\widetilde{B}$ can be written as $\widetilde{A}^T\widetilde{B} - A^TB + A^TB$, so applying Weyl's inequality [8] gives

$$||\widetilde{A}^T\widetilde{B} - (\widetilde{A}^T\widetilde{B})_r|| \leq ||A^TB - (A^TB)_r|| + ||\widetilde{A}^T\widetilde{B} - A^TB||. \tag{9}$$

To bound $||\widetilde{A}^T\widetilde{B} - (\widetilde{A}^T\widetilde{B})_r||_F$, observe that it is equivalent to bounding

$$\sum_{i=r+1}^{n} \widetilde{\sigma}_i^2 = \sum_{i=1}^{n} \widetilde{\sigma}_i^2 - \sum_{i=1}^{r} \widetilde{\sigma}_i^2 = ||\widetilde{A}^T\widetilde{B}||_F^2 - ||(\widetilde{A}^T\widetilde{B})_r||_F^2. \tag{10}$$

We next bound $||\widetilde{A}^T\widetilde{B}||_F$ and $||(\widetilde{A}^T\widetilde{B})_r||_F$ as

$$||\widetilde{A}^T\widetilde{B}||_F \leq ||A^TB||_F + ||\widetilde{A}^T\widetilde{B} - A^TB||_F \tag{11}$$

$$||(\widetilde{A}^T\widetilde{B})_r||_F \geq ||(A^TB)_r||_F - \sqrt{r}||\widetilde{A}^T\widetilde{B} - A^TB||, \tag{12}$$

where Eq. (11) follows from the triangle inequality, and Eq. (12) follows from Lemma 4. However, the lower bound in (12) is useful only when its right-hand side is greater than 0, which may not be true. Hence, we divide the following analysis into two cases.

Note that the following analysis will make use of Lemma 2-3 as well as Eq. (7). Each of them holds with probability at least $1 - \gamma/3$ (assuming that k is chosen according to Eq. (2)), so the following analysis fails with probability at most γ.

Case 1. $||(A^TB)_r||_F \leq \sqrt{r}||\widetilde{A}^T\widetilde{B} - A^TB||$.

Recall that our current goal is to bound $||\widetilde{A}^T\widetilde{B} - (\widetilde{A}^T\widetilde{B})_r||_F$.

$$\begin{aligned}
||\widetilde{A}^T\widetilde{B} - (\widetilde{A}^T\widetilde{B})_r||_F &\leq ||\widetilde{A}^T\widetilde{B}||_F \leq ||A^TB||_F + ||\widetilde{A}^T\widetilde{B} - A^TB||_F &(13)\\
&\leq ||A^TB - (A^TB)_r||_F + ||(A^TB)_r||_F + ||\widetilde{A}^T\widetilde{B} - A^TB||_F &(14)\\
&\leq ||A^TB - (A^TB)_r||_F + \sqrt{r}||\widetilde{A}^T\widetilde{B} - A^TB|| + ||\widetilde{A}^T\widetilde{B} - A^TB||_F, &(15)
\end{aligned}$$

where the last inequality follows from the assumption of Case 1. Substituting Eqs. (15) and (9) into (8) gives that

$$\begin{aligned}
&||A^TB - \widehat{A^TB}_r|| - \big(||A^TB - (A^TB)_r|| + \epsilon||A^TB - (A^TB)_r||_F + \zeta\big) &(16)\\
&\leq (2 + \epsilon\sqrt{r})||\widetilde{A}^T\widetilde{B} - A^TB|| + \epsilon||\widetilde{A}^T\widetilde{B} - A^TB||_F &(17)\\
&\leq (2 + \epsilon\sqrt{r})\epsilon||A||||B|| + \epsilon^2||A||_F||B||_F &(18)\\
&\leq (2 + \epsilon\sqrt{r})\epsilon||A||||B|| + \epsilon^2\tilde{r}||A||||B|| &(19)\\
&= \big(2\epsilon + \epsilon^2(\tilde{r} + \sqrt{r})\big)\, ||A||||B|| &(20)
\end{aligned}$$

Eq.(18) follows from Lemma 2 and 3, while Eq. (19) follows from the definition of the maximum stable rank $\tilde{r} := \max\{\frac{||A||_F^2}{||A||^2}, \frac{||B||_F^2}{||B||^2}\}$.

Case 2. $||(A^TB)_r||_F > \sqrt{r}||\widetilde{A}^T\widetilde{B} - A^TB||$.

In this case, we have a nontrivial lower bound for $||(\widetilde{A}^T\widetilde{B})_r||_F$ (provided in Eq. (12)). Therefore,

$$
\begin{aligned}
||\widetilde{A}^T\widetilde{B}-(\widetilde{A}^T\widetilde{B})_r||_F^2 &= ||\widetilde{A}^T\widetilde{B}||_F^2-||(\widetilde{A}^T\widetilde{B})_r||_F^2 && (21)\\
&\le \left(||A^TB||_F+||\widetilde{A}^T\widetilde{B}-A^TB||_F\right)^2-\left(||(A^TB)_r||_F-\sqrt{r}||\widetilde{A}^T\widetilde{B}-A^TB||\right)^2 && (22)\\
&\le ||A^TB-(A^TB)_r||_F^2+||\widetilde{A}^T\widetilde{B}-A^TB||_F^2+2||A^TB||_F||\widetilde{A}^T\widetilde{B}-A^TB||_F\\
&\quad+2||(A^TB)_r||_F\sqrt{r}||\widetilde{A}^T\widetilde{B}-A^TB|| && (23)\\
&\le ||A^TB-(A^TB)_r||_F^2+||\widetilde{A}^T\widetilde{B}-A^TB||_F^2+2||A^TB||_F||\widetilde{A}^T\widetilde{B}-A^TB||_F\\
&\quad+2\sqrt{r}||A^TB||_F||\widetilde{A}^T\widetilde{B}-A^TB||, && (24)
\end{aligned}
$$

where the first inequality follows from Eqs. (11) and (12), the last inequality follows from $||(A^TB)_r||_F \le ||A^TB||_F$. Taking square root over both sides of Eq. (24) gives the following inequality

$$
\begin{aligned}
&||\widetilde{A}^T\widetilde{B}-(\widetilde{A}^T\widetilde{B})_r||_F-||A^TB-(A^TB)_r||_F && (25)\\
&\le ||\widetilde{A}^T\widetilde{B}-A^TB||_F+\sqrt{2\sqrt{r}||A^TB||_F||\widetilde{A}^T\widetilde{B}-A^TB||+2||A^TB||_F||\widetilde{A}^T\widetilde{B}-A^TB||_F} && (26)\\
&\le \epsilon||A||_F||B||_F+\sqrt{2\sqrt{r}C_{AB}||A||_F||B||_F\epsilon||A||||B||+2C_{AB}||A||_F||B||_F\epsilon||A||_F||B||_F} && (27)\\
&\le \left(\epsilon\tilde{r}+\epsilon^{0.5}\sqrt{2\tilde{r}C_{AB}(\sqrt{r}+\tilde{r})}\right)||A||||B||. && (28)
\end{aligned}
$$

Eq.(27) follows from Lemma 2 and 3, and the definition of $C_{AB} := \frac{||A^TB||_F}{||A||_F||B||_F} \le 1$. Eq. (28) follows from the definition of the maximum stable rank $\tilde{r}$.

Substituting Eqs. (28) and (9) into (8) gives us a bound on $||A^TB-\widehat{A^TB_r}||$:

$$
\begin{aligned}
&||A^TB-\widehat{A^TB_r}||-\left(||A^TB-(A^TB)_r||+\epsilon||A^TB-(A^TB)_r||_F+\zeta\right) && (29)\\
&\le 2||\widetilde{A}^T\widetilde{B}-A^TB||+\epsilon\left(\epsilon\tilde{r}+\epsilon^{0.5}\sqrt{2\tilde{r}C_{AB}(\sqrt{r}+\tilde{r})}\right)||A||||B|| && (30)\\
&\le \left(2\epsilon+\epsilon^2\tilde{r}+\epsilon^{1.5}\sqrt{2\tilde{r}C_{AB}(\sqrt{r}+\tilde{r})}\right)||A||||B||, && (31)
\end{aligned}
$$

where the last inequality follows from Lemma 3.

By combining results of the above two cases (Eqs. (20) and (31)), we can get the desired bound for $||A^TB-\widehat{A^TB_r}||$.

$$||A^TB-\widehat{A^TB_r}|| \le ||A^TB-(A^TB)_r||+\epsilon||A^TB-(A^TB)_r||_F+\zeta+\eta||A||||B||, \tag{32}$$

where $\eta := \max\{2\epsilon+\epsilon^2(\tilde{r}+\sqrt{r}),\ \ 2\epsilon+\epsilon^2\tilde{r}+\epsilon^{1.5}\sqrt{2\tilde{r}C_{AB}(\sqrt{r}+\tilde{r})}\}$.

As mentioned earlier in our proof, the above analysis uses results from Lemma 2-3 and Eq. (7), and each of them fails with probability less than $\gamma/3$ (provided that the values of m, T, and k satisfy the assumptions in the theorem). Therefore, Eq. (32) holds with probability at least $1-\gamma$ for the chosen m, T, and k. □

References

[1] N. Ailon and B. Chazelle. The fast johnson-lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

[2] S. Bhojanapalli, P. Jain, and S. Sanghavi. Tighter low-rank approximation via sampling the leveraged element. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 902–920. SIAM, 2015.

[3] X. Chen, H. Liu, and J. G. Carbonell. Structured sparse canonical correlation analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 199–207, 2012.

[4] M. B. Cohen, J. Nelson, and D. P. Woodruff. Optimal approximate matrix product in terms of stable rank. *arXiv preprint arXiv:1507.02268*, 2015.

[5] A. Dasgupta, R. Kumar, and T. Sarlós. A sparse johnson: Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 341–350. ACM, 2010.

[6] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*, volume 6. SIAM, 1998.

[7] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 143–152. IEEE, 2006.

[8] T. Tao. 254a, notes 3a: Eigenvalues and sums of hermitian matrices. *Terence Tao's blog*, 2010.

[9] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014.