Co-Occurring Directions Sketching for Approximate Matrix Multiply

Youssef Mroueh **Etienne Marcheret** **Vaibhava Goel**
IBM T.J Watson Research Center, NY, USA.

Abstract

We introduce *co-occurring directions* sketching, a deterministic algorithm for approximate matrix product (AMM), in the streaming model. We show that co-occurring directions achieves a better error bound for AMM than other randomized and deterministic approaches for AMM. Co-occurring directions gives a $(1+\varepsilon)$-approximation of the optimal low rank approximation of a matrix product. Empirically our algorithm outperforms competing methods for AMM, for a small sketch size. We validate empirically our theoretical findings and algorithms.

1 Introduction

The vast and continuously growing amount of multimodal content poses some challenges with respect to the collection and the mining of this data. Multimodal datasets are often viewed as multiple large matrices describing the same content with different modality representations (multiple views) such as images and their textual descriptions. The product of large multimodal matrices is of practical interest as it models the correlation between different modalities. Methods such as Partial Least Squares (PLS) [Weg00], Canonical Correlation Analysis (CCA)[Hot36], Spectral Co-Clustering [Dhi01], exploit the low rank structure of the correlation matrix to mine the hidden joint factors, by computing the truncated singular value decomposition of a matrix product.

The data streaming paradigm assumes a single pass over the data and a small memory footprint, resulting in a space/accuracy tradeoff. Multimodal data can occupy a large amount of memory or may be generated sequentially, hence it is important for the streaming model to capture the data correlation .

Approximate Matrix Multiplication (AMM), is gaining an increasing interest in streaming applications (See the recent monograph [Woo14] for more details). In AMM we are given matrices X,Y, with a large number of columns n, and the goal is to compute matrices B_X, B_Y, with smaller number of columns ℓ, such that $||XY^\top - B_X B_Y^\top||_Z$ is small for some norm $\|.\|_Z$. In streaming AMM, columns of B_X, B_Y, need to be updated as the data arrives sequentially. We refer to B_X and B_Y as sketches of X and Y.

Randomized approaches for AMM were pioneered by the work of [DKM06]. The approach of [DKM06] is based on the *sampling* of ℓ columns of X and Y. [DKM06] shows that by choosing an appropriate sampling matrix $\Pi \in \mathbb{R}^{n\times \ell}$, we obtain a Frobenius error guarantee ($\|.\|_Z = \|.\|_F$):

$$\left\|XY^\top - X\Pi(Y\Pi)^\top\right\|_F \leq \varepsilon \left\|X\right\|_F \left\|Y\right\|_F, \qquad (1)$$

for $\ell = \Omega(1/\varepsilon^2)$, with high probability. The same guarantee of Eq. (1) was achieved in [Sar06], by using a *random projection* $\Pi \in \mathbb{R}^{n\times\ell}$ that satisfies the guarantees of a Johnson- Lindenstrauss (JL) transform ($\forall x \in \mathbb{R}^n \left\|\Pi x\right\|^2 \sim (1\pm\varepsilon)\left\|x\right\|^2$, with probability $1-\delta$), where $\ell = O(1/\varepsilon^2 \log(1/\delta))$. Other randomized approaches focused on error guarantees given in spectral norm ($\|.\|_Z = \|.\|$) , such as JL embeddings or efficient subspace embeddings [Sar06, MZ11, ATKZ14, CNW15] that can be applied to any type of matrices X in input sparisty time [CW13]. [CNW15] showed that using a subspace embedding $\Pi \in \mathbb{R}^{n\times\ell}$ we have with a probability $1-\delta$:

$$\left\|XY^\top - X\Pi(Y\Pi)^\top\right\| \leq \varepsilon \left\|X\right\| \left\|Y\right\|, \qquad (2)$$

for $\ell = O((sr(X) + sr(Y) + \log(1/\delta))/\varepsilon^2)$, where $sr(X) = \frac{\|X\|_F^2}{\|X\|^2}$ is the stable rank of X. Note that $sr(X) \leq rank(X)$, hence results stated in term of stable rank are sharper and more robust than the one stated with the rank [Sar06, MZ11, ATKZ14].

Covariance sketching refers to AMM for $X = Y$. An elegant deterministic approach for covariance sketching called *frequent directions* was introduced recently

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

in [Lib13, GLPW15], drawing the connection between *covariance matrix sketching*, and the classic problem of estimation of *frequent items* [MG82]. Another approach for AMM, consists of concatenating matrices X and Y, and of applying a covariance sketch technique on the resulting matrix, this approach results in a looser guarantee; The right hand side in Equations (1),(2) is replaced by $\varepsilon(\|X\|_F^2 + \|Y\|_F^2)$. Based on this observation, [YLZ16] proposed to use the *frequent directions* algorithm of [Lib13] to perform AMM in a deterministic way, we refer to this approach as FD-AMM. FD-AMM [YLZ16] outputs B_X, B_Y such that

$$\left\|XY^\top - B_X B_Y^\top\right\| \le \varepsilon(\|X\|_F^2 + \|Y\|_F^2), \qquad (3)$$

for $\ell = \lceil \frac{1}{\varepsilon} \rceil$. The sketch length ℓ dependency on ε in randomized methods is quadratic, FD-AMM improves this dependency to linear.

In this paper we introduce *co-occurring directions*, a deterministic algorithm for AMM. Our algorithm is inspired by *frequent directions* and enables similar guarantees to (2) in spectral norm, but with a linear dependency of ℓ on ε as in FD-AMM. Given with stable ranks, co-occurring direction achieves the guarantee of (2) for $\ell = O(\sqrt{sr(X)sr(Y)}/\varepsilon)$.

The paper is organized as follows: In Section 2 we review *frequent directions*, introduce our *co-occurring directions* sketching algorithm, and give error bounds analysis in AMM and in low rank approximation of a matrix product. We state our proofs in Section 3. In section 2.2.2 and Section 4 we discuss error bounds, space and time requirements, and compare our approach to related work on AMM and low rank approximation. Finally we validate the empirical performance of *co-occurring directions* in Section 5, on both synthetic and real world multimodal datasets.

Notation. We note by $C = U\Sigma V^\top$, the thin svd of C, and by $\sigma_{\max}(C)$ the maximum singular value, Tr refers to the trace. σ_j are the singular values that are assumed to be given in decreasing order. Note that for $C \in \mathbb{R}^{m_x \times m_y}$ the spectral norm is defined as follows $\|C\| = \max_{u,v,\|u\|=\|v\|=1} \left|u^\top C v\right| = \sigma_{\max}(C)$. The nuclear norm (known also as trace or 1− schatten norm) is defined as follows: $\|C\|_* = Tr(\Sigma)$. The stable rank of C is $sr(C) = \frac{\|C\|_F^2}{\|C\|^2}$. Assume C and D have the same number of column, $[C; D]$ denotes their concatenation on their row dimensions. For $n \in \mathbb{N}, [n] = \{1, \ldots n\}$.

2 Sketching from Covariance to Correlation

In this section we review covariance sketching with the *frequent directions* algorithm of [Lib13] and state its theoretical guarantees [Lib13, GLPW15]. We then introduce correlation sketching and present and analyze our *co-occurring directions* algorithm.

2.1 Covariance Sketching: *Frequent Directions*

Let $X \in \mathbb{R}^{m_x \times n}$, where n is the number of samples and m_x the dimension. We assume that $n > m_x$. The goal of covariance sketching is to find a small matrix $D_X \in \mathbb{R}^{m_x \times \ell}$, where $\ell << n$ (ℓ is assumed to be an even number), such that $XX^\top \approx D_X D_X^\top$. Frequent directions algorithm introduced in [Lib13] (Algorithm 1) achieves this goal. Intuitively *frequent directions* algorithm sets a noise level using the median of the spectrum of the covariance of the sketch D_X. It then discards directions below that level and replaces them with fresh samples. This results in the updated covariance estimate. This process is repeated as the data is streaming.

Algorithm 1 Frequent Directions

```
procedure FD($X \in \mathbb{R}^{m_x \times n}$)
    $D_X \leftarrow 0 \in \mathbb{R}^{m_x \times \ell}$ .
    for $i \in [n]$ do
        Insert column $X_i$ into a zero column of $D_X$
        if $D_X$ has no zero valued column then
            $[U, \Sigma, V] \leftarrow \mathrm{SVD}(D_X)$
            $\delta \leftarrow \sigma_{\ell/2}^2$                ▷ median value of $\Sigma^2$
            $\tilde{\Sigma} \leftarrow \sqrt{\max(\Sigma^2 - \delta I_\ell, 0)}$     ▷ shrinkage
            $D_X \leftarrow U\tilde{\Sigma}$
        end if
    end for
    return $D_X$
end procedure
```

Theorem 1 ([Lib13]) *D_X the output of algorithm 1 satisfies:*

$$\left\|XX^\top - D_X D_X^\top\right\| \le \frac{2\,\|X\|_F^2}{\ell}. \qquad (4)$$

2.2 Correlation Sketching: *Co-occurring Directions*

We start by defining correlation sketching:

Definition 1 (Correlation Sketching/AMM) *Let $X \in \mathbb{R}^{m_x \times n}$, $Y \in \mathbb{R}^{m_y \times n}$, where $n > \max(m_x, m_y)$. Let $B_X \in \mathbb{R}^{m_x \times \ell}$ and $B_Y \in \mathbb{R}^{m_y \times \ell}$ $(\ell < n, \ell \le \min(m_x, m_y))$. Let $\eta > 0$. The matrix pair (B_X, B_Y) is called an η-correlation sketch of (X, Y) if it satisfies in spectral norm:*

$$\left\|XY^\top - B_X B_Y^\top\right\| \le \eta.$$

We now present our *co-occurring directions* algorithm (Algorithm 2). Intuitively Algorithm 2 sets a noise level using the median of the singular values of the correlation matrix of the sketch $B_X B_Y^\top$. The SVD of $B_X B_Y^\top$ is computed efficiently in lines 8,9 and 10 of Algorithm 2 using QR decomposition. Left and right singular vectors below this noise threshold are replaced by fresh samples from X and Y, correlation sketches are updated and the process continues. Theorem 2 shows that our *co-occurring directions* algorithm outputs (B_X, B_Y) a correlation sketch of (X, Y) as defined above in Definition 1.

Algorithm 2 Co-occurring Directions

```
 1: procedure Co-D($X \in \mathbb{R}^{m_x \times n}, Y \in \mathbb{R}^{m_y \times n}$)
 2:     $B_X \leftarrow 0 \in \mathbb{R}^{m_x \times \ell}$ .
 3:     $B_Y \leftarrow 0 \in \mathbb{R}^{m_y \times \ell}$ .
 4:     for $i \in [n]$ do
 5:         Insert a column $X_i$ into a zero valued column of $B_X$
 6:         Insert a column $Y_i$ into a zero valued column of $B_Y$
 7:         if $B_X, B_Y$ have no zero valued column then
 8:             $[Q_x, R_x] \leftarrow \mathrm{QR}(B_X)$
 9:             $[Q_y, R_y] \leftarrow \mathrm{QR}(B_Y)$
10:             $[U, \Sigma, V] \leftarrow \mathrm{SVD}(R_x R_y^\top)$
11:                         ▷ $Q_x \in \mathbb{R}^{m_x \times \ell}, R_x \in \mathbb{R}^{\ell \times \ell}$,
12:             ▷ $Q_y \in \mathbb{R}^{m_y \times \ell}, R_y \in \mathbb{R}^{\ell \times \ell}$, $U, \Sigma, V \in \mathbb{R}^{\ell \times \ell}$.
13:             $C_x \leftarrow Q_x U \sqrt{\Sigma}$
14:             $C_y \leftarrow Q_y V \sqrt{\Sigma}$
15:                         ▷ $C_x, C_y$ not computed
16:             $\delta \leftarrow \sigma_{\ell/2}(\Sigma)$    ▷ the median value of $\Sigma$
17:             $\tilde{\Sigma} \leftarrow \max(\Sigma - \delta I_\ell, 0)$    ▷ shrinkage
18:             $B_X \leftarrow Q_x U \sqrt{\tilde{\Sigma}}$
19:             $B_Y \leftarrow Q_y V \sqrt{\tilde{\Sigma}}$
20:                 ▷ at least last $\ell/2$ columns are zero
21:         end if
22:     end for
23:     return $B_X, B_Y$
24: end procedure
```

It is important to see that while frequent directions shrinks Σ^2, co-occurring directions filters Σ. We prove in the following an approximation bound in spectral norm for co-occurring directions.

2.2.1 Main Results

We give in the following our main results, on the approximation error of co-occurring direction in AMM (Theorem 2), and in the $k-$th rank approximation of a matrix product (Theorem 3). Proofs are given in Section 3.

Theorem 2 (AMM) *The output of* co-occurring *directions (Algorithm 2) gives a correlation sketch (B_X, B_Y) of (X, Y), for $\ell \leq \min(m_x, m_y)$ satisfying:*

For a correlation sketch of length ℓ, we have:

$$\left\| XY^\top - B_X B_Y^\top \right\| \leq \frac{2 \left\| X \right\|_F \left\| Y \right\|_F}{\ell}.$$

2) Algorithm 2 runs in $O(n(m_x + m_y + \ell)\ell)$ time and requires a space of $O((m_x + m_y + \ell)\ell)$.

Theorem 3 (Low Rank Product Approximation) *Let (B_X, B_Y) be the output of Algorithm 2. Let $k \leq \ell$. Let U_k, V_k be the matrices whose columns are the k-th largest left and right singular vectors of $B_X B_Y^\top$. Let $\pi_U^k(X) = U_k U_k^\top X, \pi_V^k(Y) = V_k V_k^\top Y$. Let $\varepsilon > 0$, for $\ell \geq 8 \frac{\sqrt{sr(X) sr(Y)}}{\varepsilon} \frac{||X||||Y||}{\sigma_{k+1}(XY^\top)}$ we have: $\left\| XY^\top - \pi_U^k(X) \pi_V^k(Y)^\top \right\| \leq \sigma_{k+1}(XY^\top)(1 + \varepsilon)$.*

2.2.2 Discussion of Main Results

For $\ell = \lceil \frac{1}{\varepsilon} \rceil, \varepsilon \in [\frac{1}{\min(m_x, m_y)}, 1]$ from Theorem 2 we see that (B_X, B_Y) produced by Algorithm 2 is an η-correlation sketch of (X, Y) for $\eta = 2\varepsilon \left\| X \right\|_F \left\| Y \right\|_F$. In AMM, bounds are usually stated in term of the product of spectral norms of X and Y as in Equation (2). Let $sr(X) = \frac{\left\| X \right\|_F^2}{\left\| X \right\|^2}$ be the stable rank of X. It is easy to see that co-occurring directions for $\ell = \frac{2\sqrt{sr(X) sr(Y)}}{\varepsilon}$, gives an error bound of $\varepsilon \left\| X \right\| \left\| Y \right\|$. While in randomized methods the error is $O(1/\sqrt{\ell})$, co-occurring direction's error is $O(1/\ell)$. Moreover the dependency on stable ranks in co-occurring directions is $2\sqrt{sr(X) sr(Y)} \leq sr(X) + sr(Y)$, the latter appears in subspace embedding based AMM [CNW15, MZ11, ATKZ14]. For $X = Y$ *co-occurring directions* reduces to *frequent directions* of [Lib13], and Theorem 2 recovers Theorem 1 of [Lib13].

Stronger bounds for frequent directions were given in [GLPW15] where the bound in Equation (4) is improved, for $\ell > 2k$, for any k:

$$\left\| XX^\top - D_X D_X^\top \right\| \leq \frac{2}{\ell - 2k} \left\| X - X_k \right\|_F^2,$$

where X_k is the $k-$th rank approximation of X (with $X_0 = 0$). Hence by defining $Z = [X; Y] \in \mathbb{R}^{(m_x + m_y) \times n}$ and applying frequent directions to Z (FD-AMM [YLZ16]), we obtain B_X, B_Y satisfying: $\left\| XY^\top - B_X B_Y^\top \right\| \leq \frac{2}{\ell - 2k} \left\| Z - Z_k \right\|_F^2$, hence the perfomance of FD-AMM depends on the low rank structure of Z. A sharper analysis for co-occurring directions remains an open question, but the following discussion of Theorem 3 will shed some light on the advantages of co-occurring directions on FD-AMM [YLZ16].

Theorem 3 shows that co-occurring directions sketching gives a $(1+\varepsilon)$- approximation of the optimal low rank approximation of the matrix product $XY^\top$. Note that $\sigma_{k+1}(XY^\top) \leq \frac{\|XY^\top\|_*}{k+1}$. Hence for $\ell \geq 8(k+1)/\varepsilon$, we obtain a $1+\varepsilon$- approximation of the optimal k rank approximation of $XY^\top$. This highlights the relation between the sketch length in co-occurring directions ℓ and the rank of $XY^\top$. Note that the maximum rank of $XY^\top$ is $\min(rank(X), rank(Y))$. When using FD-AMM, based on the covariance sketch of the concatenation of X and Y, the sketch length ℓ is related to the rank of $Z = [X; Y]$. Note that the maximum rank of the concatenation (Z) is bounded by $rank(X) + rank(Y)$. Hence we see that co-occurring directions guarantees a $1+\varepsilon$ approximation of the optimal k-rank approximation of $XY^\top$ for a smaller sketch size then FD-AMM ($\min(rank(X), rank(Y))$ for *co-occurring directions* versus $rank(X) + rank(Y)$ for FD-AMM).

In the following we comment on the running time of co-occurring directions.

2.2.3 Running Time Analysis and Parralelization.

Running Time. We compare the space and the running time of our sketch to a naive implementation of the correlation sketch.
1) Naive Correlation Sketch: In the if statement of Algorithm 2, compute the ℓ thin svd $\text{SVD}(B_X B_Y^\top) = [U, \Sigma, V]$, $B_X \leftarrow U\sqrt{\tilde{\Sigma}}, B_Y \leftarrow V\sqrt{\tilde{\Sigma}}$. We need a space $O(m_x m_y)$ to store $B_X B_y^\top$. The running time is dominated by computing an ℓ thin svd $O(m_x m_y \ell)$ each $\frac{n}{\ell/2}$ that is $O(n m_x m_y)$, hence no gain with respect to brute force.
2) Co-occurring Directions: Algorithm 2 avoids computing $B_X B_Y^\top$ by using the QR decomposition of B_X and B_Y. The space needed is $O(\ell(m_x+m_y+\ell))$. We have a computation done every $\frac{n}{\ell/2}$, that is dominated by computing QR factorization and svd : $O((m_x+m_y+\ell)\ell^2)$ (computing $R_x R_y^\top$ requires $O(\ell^3)$ operations). This results in a total running time : $O(n(m_x+m_y+\ell)\ell)$. There is a computational and memory advantage when $\ell < \frac{m_x m_y}{m_x+m_y}$.

Parallelization of Co-occurring Directions (Sketches of Sketches). Similarly to the frequent directions [Lib13], co-occurring directions algorithm is simply parallelizable. Let $X = [X_1, X_2] \in \mathbb{R}^{m_x \times (n_1+n_2)}$, and $Y = [Y_1, Y_2] \in \mathbb{R}^{m_x \times (n_1+n_2)}$. Let (B_X^1, B_Y^1) be the correlation sketch of (X_1, Y_1), and (B_X^2, B_Y^2) be the correlation sketch of (X_2, Y_2). Then the correlation sketch (C_X, C_Y) of $([B_X^1, B_X^2], [B_Y^1, B_Y^2])$ is a correlation sketch of (X, Y), and is as good as (B_X, B_Y) the correlation sketch of (X, Y). Hence we can sketch the data in M-independent chunks on M machines then merge by concatenating the sketches and performing another sketch on the concatenation, by doing so we divide the running time by M.

3 Proofs

In this Section we give proofs of our main results:

Proof 1 (Proof of Theorem 2) *By construction we have:*

$$\begin{aligned} C_x C_y^\top &= \left(Q_x U\sqrt{\Sigma}\right)\left(Q_y V\sqrt{\Sigma}\right)^\top \\ &= Q_x\left(U\Sigma V^\top\right)Q_y^\top = Q_x\left(R_x R_y^\top\right)Q_y^\top \\ &= (Q_x R_x)(Q_y R_y)^\top . \end{aligned}$$

Hence the algorithm is computing a form of R-SVD of $B_X B_Y^\top$, followed by a shrinkage of the correlation matrix. Let $B_x^i, B_y^i, C_x^i, C_y^i, \Sigma^i, \tilde{\Sigma}^i, \delta_i$, the values of $B_X, B_Y, C_x, C_y, \Sigma, \tilde{\Sigma}, \delta$ after the execution of the main loop.if we don't enter the if statement $\delta_i = 0$ ($B_x^i = C_x^i$ and $B_y^i = C_y^i$ in that case).
Hence we have at an iteration i:

$$C_x^i C_y^{i,\top} = B_x^{i-1} B_y^{i-1,\top} + X_i Y_i^\top .$$

Note that:

$$\begin{aligned} &XY^\top - B_X B_Y^\top = XY^\top - B_x^n B_y^{n,\top} \\ &= \sum_{i=1}^n \left(X_i Y_i^\top + B_x^{i-1} B_y^{i-1,\top} - B_x^i B_y^{i,\top}\right) \\ &= \sum_{i=1}^n \left(C_x^i C_y^{i,\top} - B_x^i B_y^{i,\top}\right). \end{aligned}$$

By the triangle inequality we can bound the spectral norm:

$$\left\|XY^\top - B_X B_Y^\top\right\| \leq \sum_{i=1}^n \left\|C_x^i C_y^{i,\top} - B_x^i B_y^{i,\top}\right\|.$$

We are left with bounding $\left\|C_x^i C_y^{i,\top} - B_x^i B_y^{i,\top}\right\|$: $C_x^i C_y^{i,\top} = \left(Q_x^i U^i\right)\Sigma^i\left(Q_y^i V^i\right)^\top, B_x^i B_y^{i,\top} = \left(Q_x^i U^i\right)\tilde{\Sigma}^i\left(Q_y^i V^i\right)^\top$. Note that:

$$\begin{aligned} \left\|C_x^i C_y^{i,\top} - B_x^i B_y^{i,\top}\right\| &= \left\|(Q_x^i U^i)(\Sigma^i - \tilde{\Sigma}^i)(Q_y^i V^i)^\top\right\| \\ &= \left\|\Sigma^i - \tilde{\Sigma}^i\right\| \\ &\leq \delta_i, \end{aligned}$$

where the first equality follows from the fact that, $Q_x^i U^i, Q_y^i V^i$, are orthonormal. And $\Sigma^i - \tilde{\Sigma}^i$ is a diagonal matrix with at least $\ell/2$ entries equal δ_i or 0, and *the other entries are less than δ_i. It follows that we have in spectral norm:*

$$\left\|XY^\top - B_X B_Y^\top\right\| \le \sum_{i=1}^n \delta_i. \qquad (5)$$

Now we want to relate $\sum_{i=1}^n \delta_i$ to ℓ, and propreties of X, Y.
Let $\|.\|_*$, the $1-$ schatten norm. For a matrix A of rank r, and singular values σ_i : $\|A\|_* = \sum_{i=1}^r \sigma_i(A)$. We have:

$$\begin{aligned}
\left\|B_X B_Y^\top\right\|_* &= \left\|B_x^n B_y^{n,\top}\right\|_* \\
&= \sum_{i=1}^n \left\|B_x^i B_y^{i,\top}\right\|_* - \left\|B_x^{i-1} B_y^{i-1,\top}\right\|_* \\
&= \sum_{i=1}^n \left(\left\|C_x^i C_y^{i,\top}\right\|_* - \left\|B_x^{i-1} B_y^{i-1,\top}\right\|_*\right) \\
&- \sum_{i=1}^n \left(\left\|C_x^i C_y^{i,\top}\right\|_* - \left\|B_x^i B_y^{i,\top}\right\|_*\right) \qquad (6)
\end{aligned}$$

We have at an iteration i, the R-SVD of $C_x^i C_y^{i,\top}$ and $B_x^i B_y^{i,\top}$:

$$\left\|C_x^i C_y^{i,\top}\right\|_* = Tr(\Sigma^i) \text{ and } \left\|B_x^i B_y^{i,\top}\right\|_* = Tr(\tilde{\Sigma}^i).$$

Hence we have by the definition of the shrinking operation:

$$\left\|C_x^i C_y^{i,\top}\right\|_* - \left\|B_x^i B_y^{i,\top}\right\|_* = Tr(\Sigma^i - \tilde{\Sigma}^i) = \sum_{j=1}^{\ell} \sigma_j^i - \tilde{\sigma}_j^i$$

$$= \sum_{j, \sigma_j^i > \delta_i} \delta_i + \sum_{j, \sigma_j^i \le \delta_i} \sigma_j^i \ge \frac{\ell}{2}\delta_i. \qquad (7)$$

On the other hand using the reverse triangle inequality for the $1-$ schatten norm we have:

$$\left\|C_x^i C_y^{i,\top}\right\|_* - \left\|B_x^{i-1} B_y^{i-1,\top}\right\|_* \le \left\|C_x^i C_y^{i,\top} - B_x^{i-1} B_y^{i-1,\top}\right\|_*$$

Recall that: $C_x^i C_y^{i,\top} = B_x^{i-1} B_y^{i-1,\top} + X_i Y_i^\top$, hence we have:

$$\left\|C_x^i C_y^{i,\top}\right\|_* - \left\|B_x^{i-1} B_y^{i-1,\top}\right\|_* \le \left\|X_i Y_i^\top\right\|_* = \|X_i\|_2 \|Y_i\|_2, \qquad (8)$$

since $X_i Y_i^\top$ is rank one. Finally putting together Equations (6), (7),(8), *we have:*

$$\left\|B_X B_Y^\top\right\|_* \le \sum_{i=1}^n \|X_i\|_2 \|Y_i\|_2 - \frac{\ell}{2}\sum_{i=1}^n \delta_i. \qquad (9)$$

It follows from Equation (9) *that:*

$$\begin{aligned}
\sum_{i=1}^n \delta_i &\le \frac{2}{\ell}\left(\sum_{i=1}^n \|X_i\|_2 \|Y_i\|_2 - \left\|B_X B_Y^\top\right\|_*\right) \\
&\le \frac{2}{\ell}\left(\sqrt{\sum_{i=1}^n \|X_i\|_2^2}\sqrt{\sum_{i=1}^n \|Y_i\|_2^2}\right) \\
&= \frac{2}{\ell}\|X\|_F \|Y\|_F, \qquad (10)
\end{aligned}$$

where in the last inequality we used the Cauchy-Schwarz inequality. Putting together Equations (5) *and* (10) *we have finally:*

$$\left\|XY^\top - B_X B_Y^\top\right\| \le \frac{2}{\ell}\|X\|_F \|Y\|_F. \qquad (11)$$

2) Refer to Section 2.2.3.

Proof 2 (Proof of Theorem 3) *Let $\pi_U^k(X) = U_k U_k^\top X, \pi_V^k(Y) = V_k V_k^\top Y$. Let $\mathcal{H}_k^x$ be the span of $\{u_1, \dots u_k\}$, and $\mathcal{H}_{m_x-k}^x$ be the orthogonal of $\mathcal{H}_k^x$. Similarly define $\mathcal{H}_k^y$ the span of $\{v_1, \dots v_k\}$, and $\mathcal{H}_{m_y-k}^y$ its orthogonal. For all $u \in \mathbb{R}^{m_x}, \|u\| = 1$, there exits $a_x, b_x \in \mathbb{R}$, $a_x^2 + b_x^2 = 1$, such that $u = a_x w_x + b_x z_x$, where $w_x \in \mathcal{H}_k^x, ||w_x|| = 1$ and $z_x \in \mathcal{H}_{m_x-k}^x, ||z_x|| = 1$. Similarly for $v \in \mathbb{R}^{m_y}, \|v\| = 1$ there exits $a_y, b_y \in \mathbb{R}$, $a_y^2 + b_y^2 = 1$, such that $v = a_y w_y + b_y z_y$, where $w_y \in \mathcal{H}_k^y, ||w_y|| = 1$ and $z_y \in \mathcal{H}_{m_y-k}^y, ||v_y|| = 1$.*

Let $\Delta = XY^\top - \pi_U^k(X)\pi_V^k(Y)^\top$, we have $\|\Delta\| = \max_{u \in \mathbb{R}^{m_x}, v \in \mathbb{R}^{m_y}, ||u||=||v||=1} |u^\top \Delta v|$

$$\begin{aligned}
|u^\top \Delta v| &= |(a_x w_x + b_x z_x)^\top \Delta (a_y w_y + b_y z_y)| \\
&\le |a_x a_y||w_x^\top \Delta w_y| + |b_x b_y||z_x^\top \Delta z_y| \\
&+ |a_x b_y||w_x^\top \Delta z_y| + |b_x a_y||z_x^\top \Delta w_y|
\end{aligned}$$

Since $w_x \in \mathcal{H}_k^x, w_y \in \mathcal{H}_k^y$, we have $w_x^\top \Delta w_y = 0$. Since $z_x \in \mathcal{H}_{m_x-k}^x, z_y \in \mathcal{H}_{m_y-k}^y, z_x^\top \Delta z_y = z_x^\top XY^\top z_y$. Similarly $w_x^\top \Delta z_y = w_x^\top XY^\top z_y$, and $z_x^\top \Delta w_y = z_x^\top XY^\top w_y$. Note that $|a_x|, |b_x|, |a_y|, |b_y|$ are bounded by 1. Hence we have (maximum is taken on each appropriate set defined above, all vectors are unit norm):

$$\begin{aligned}
\max_{u,v} |u^\top \Delta v| &\le \max_{z_x, z_y} |z_x^\top XY^\top z_y| + \max_{w_x, z_y} |w_x^\top XY^\top z_y| \\
&+ \max_{z_x, w_y} |z_x^\top XY^\top w_y|
\end{aligned}$$

For $z_x \in \mathcal{H}_{m_x-k}^x, z_y \in \mathcal{H}_{m_y-k}^y$ we have:

$$\begin{aligned}
|z_x^\top XY^\top z_y| &\le |z_x^\top (XY^\top - B_X B_Y^\top) z_y| + |z_x^\top B_X B_Y^\top z_y| \\
&\le \left\|XY^\top - B_X B_Y^\top\right\| + \sigma_{k+1}(B_X B_Y^\top) \\
&\le 2\left\|XY^\top - B_X B_Y^\top\right\| + \sigma_{k+1}(XY^\top),
\end{aligned}$$

where we used that $\sigma_{k+1}(B_X B_Y^\top) = \max_{z_x \in \mathcal{H}_{m_x-k}^x, z_y \in \mathcal{H}_{m_y-k}^y} |z_x^\top B_X B_Y^\top z_y|$, by definition of σ_{k+1}. The last inequality follows from weyl inequality $|\sigma_{k+1}(B_X B_Y^\top) - \sigma_{k+1}(XY^\top)| \le \left\|XY^\top - B_X B_Y^\top\right\|$.

Note that for $w_x \in \mathcal{H}_k^x$ and $z_y \in \mathcal{H}_{m_y-k}^y$ we have $w_x^\top B_X B_Y^\top z_y = 0$. To see that, note that $w_x \in span\{u_1, \dots u_k\}$, $z_y \in span\{v_{k+1}, \dots v_\ell\}$. There exists β_j, such that $z_y = \sum_{j=k+1}^{\ell} \beta_j v_j$, hence $B_X B_Y^\top z_y = \sum_{i=1}^{\ell}\sum_{j=k+1}^{\ell} \sigma_i \beta_j u_i v_i^\top v_j = \sum_{j=k+1}^{\ell} \sigma_j \beta_j u_j \perp w_x$. Hence we have:

$$\begin{aligned}
|w_x^\top XY^\top z_y| &= |w_x^\top (XY^\top - B_X B_Y^\top) z_y| \\
&\le \left\|XY^\top - B_X B_Y^\top\right\|.
\end{aligned}$$

Similarly for for $z_x \in \mathcal{H}^x_{m_x-k}$ and $w_y \in \mathcal{H}^y_k$ we conclude that: $|w_x^\top XY^\top z_y| \leq \|XY^\top - B_X B_Y^\top\|$. Finally we have:

$$\begin{aligned}\|\Delta\| &\leq 4\left\|XY^\top - B_X B_Y^\top\right\| + \sigma_{k+1}(XY^\top) \\ &\leq \frac{8\left\|X\right\|_F\left\|Y\right\|_F}{\ell} + \sigma_{k+1}(XY^\top) \\ &\leq \sigma_{k+1}(XY^\top)(1 + 8\frac{\sqrt{sr(X)sr(Y)}}{\ell}\frac{||X||||Y||}{\sigma_{k+1}(XY^\top)})\end{aligned}$$

For $\ell \geq 8\frac{\sqrt{sr(X)sr(Y)}}{\varepsilon}\frac{||X||||Y||}{\sigma_{k+1}(XY^\top)}$, we have: $\|\Delta\| \leq \sigma_{k+1}(XY^\top)(1+\varepsilon)$.

4 Previous Work on Approximate Matrix Multilply

We list here a catalog of baselines for AMM:

Brute Force. We keep a running correlation $C \leftarrow C + X_i Y_i^\top$. We perform an ℓ thin svd at the end of the stream. Space $O(m_x m_y)$, running time: $O(nm_x m_y) + O(m_x m_y \ell)$, the cost of the sketch update and the ℓ thin svd.

Sampling [DKM06]. We define a distribution over $[n]$, $p_i = \frac{\|X_i\|\|Y_i\|}{S}$, where $S = \sum_{i=1}^n \|X_i\|\,\|Y_i\|$. Form B_X and B_Y by taking ℓ iids samples (column indices), using p_i. In the streaming model, since S is not known, we use ℓ independent reservoir samples. Hence the space needed is $O(\ell(m_x + m_y))$, the running time is $O(\ell(m_x + m_y)n)$.

Random Projection [Sar06]. B_X, B_Y are of the form $X\Pi$ and $Y\Pi$, where $\Pi \in \mathbb{R}^{n\times\ell}$, and $\Pi_{ij} \in \{-1/\sqrt{\ell}, 1/\sqrt{\ell}\}$, uniformly. This is easily implemented in the streaming model and requires $O(\ell(m_x + m_y))$ space and $O(\ell(m_x + m_y)n)$ time.

Hashing [CW13]. Let $h : [n] \to [\ell]$, and $s : [n] \to \{-1, 1\}$ be perfect hash functions. We initialize B_X, B_Y to all zeros matrices. When processing columns of X and Y we update columns of B_X and B_Y as follows: $B_{X,h(i)} \leftarrow B_{X,h(i)} + s(i)X_i$, $B_{Y,h(i)} \leftarrow B_{Y,h(i)} + s(i)Y_i$. Hashing requires $O(\ell(m_x + m_y))$ space and $O(n(m_x + m_y))$ time.

FD-AMM [YLZ16]. Let $Z = [X; Y] \in \mathbb{R}^{(m_x+my)\times n}$, let D_Z be the output of frequent directions (Algoritm 1). We partition $D_Z = [B_X; B_Y]$, and use B_X and B_Y in AMM. This requires $O(\ell(m_x + m_y))$ space and $O(n(m_x + m_y)\ell)$ time.

5 Experiments

AMM of Low Rank Matrices. We consider $X \in \mathbb{R}^{m_x\times n}$ and $Y \in \mathbb{R}^{m_y\times n}$, generated using a non-noisy low rank model [GLPW15] as follows: $X = V_x S_x U_x^\top$, where $U_x \in \mathbb{R}^{n\times k_x}$, $(U_x)_{i,j} \sim \mathcal{N}(0, 1)$, $S_x \in \mathbb{R}^{k_x\times k_x}$ is a diagonal matrix with $(S_x)_{jj} = 1 - (j - 1)/k_x$, and $V_x \in \mathbb{R}^{m_x\times k_x}$ is such that $V_x^\top V_x = I_{k_x}$. Similarly we generate $Y = V_y S_y U_y^\top, U_y \in \mathbb{R}^{n\times k_y}, S_y \in \mathbb{R}^{k_y\times k_y}, V_y \in \mathbb{R}^{m_y\times k_y}$. Hence X and Y are at most rank k_x, and k_y respectively. We consider $n = 10000$, $m_x = 1000$, $m_y = 2000$, and three regimes: both matrices have a large rank ($k_x = 400, k_y = 400$), one matrix has a smaller rank then the other ($k_x = 400, k_y = 40$), and both matrices have a small rank ($k_x = 40, k_y = 40$). We compare the performance of co-occurring directions to baselines given in Section 4 in those three regimes. For randomized baselines we run each experiments 50 times and report mean and standard deviations of performances. Experiments were conducted on a single core Intel Xeon CPU E5-2667, 3.30GHz, with 265 GB of RAM and 25.6 MB of cache.

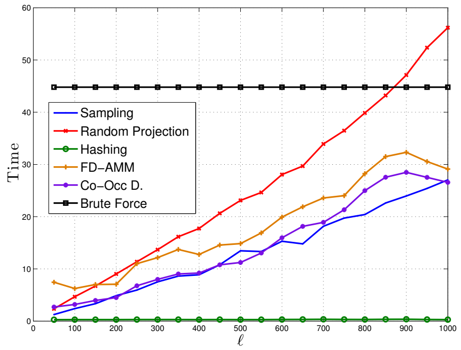
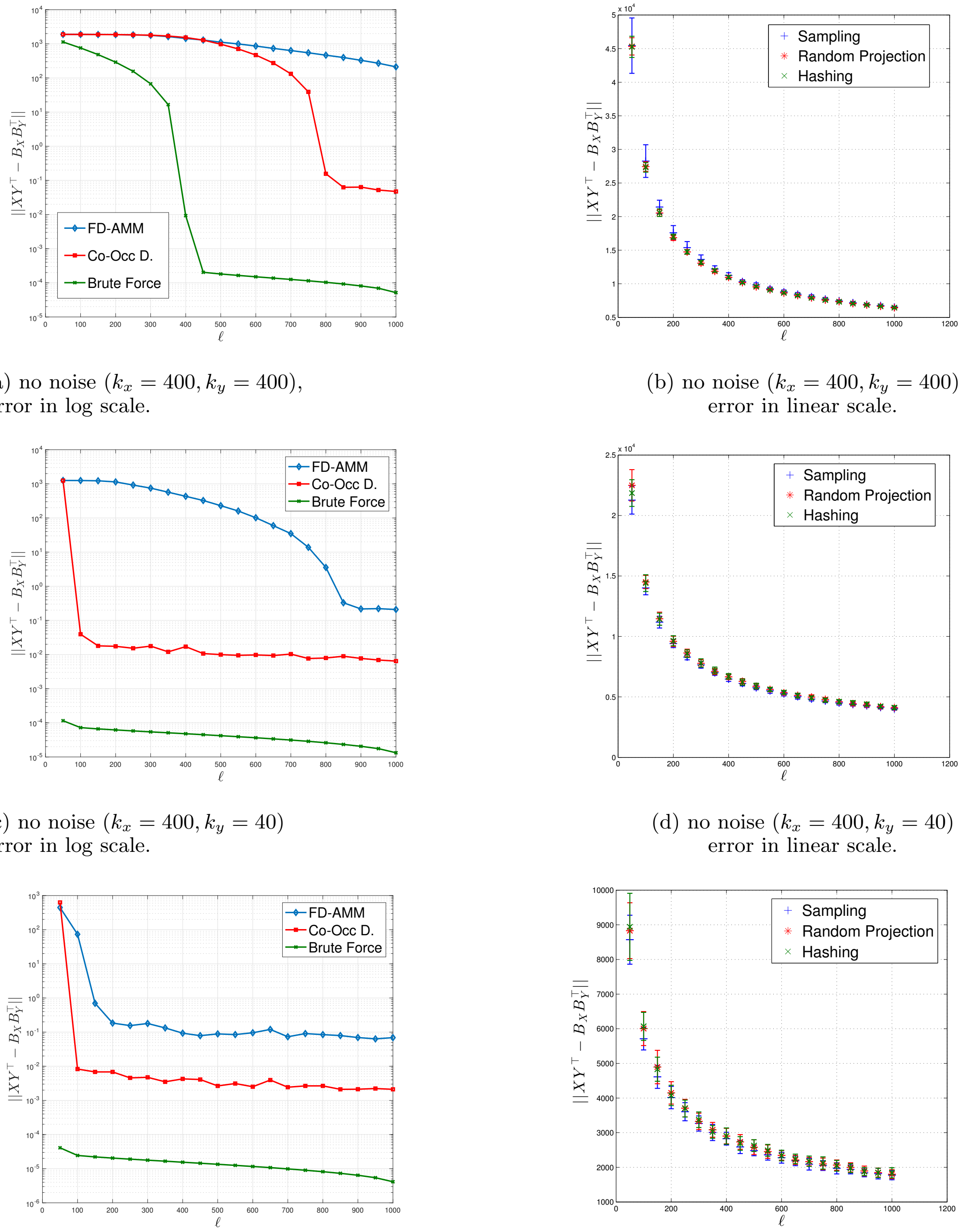


Figure 1: Time given in seconds versus sketch length ℓ.

We see in Figure 1, that hashing timing is, as expected, independent from the sketch length. Random projection requires the most amount of time. Co-occurring directions timing is on par with sampling and slightly better than FD-AMM. From Figure 2 [1] we see that the deterministic baselines (a,c,e) consistently outperform the randomized baselines (b,d,f) in all three regimes. As discussed previously randomized methods error bound are of the order of $O(1/\sqrt{\ell})$, while both co-occurring directions and FD-AMM have an error bound order $O(1/\ell)$. Note that the brute force error becomes zero (up to machine precision) when ℓ exceeds $\min(rank(X), rank(Y))$. When comparing co-occurring direction to FD-AMM we see a clear phase transition for co-occurring direction as ℓ exceeds $O(\min(rank(X), rank(Y)))$. For FD-AMM the phase transition happens when ℓ exceeds $O(rank(X) + rank(Y))$. The phase transition happens earlier for co-occurring directions and hence co-occurring directions outperforms FD-AMM for a smaller sketch size. This is in line with our discussion in Section 2.2.2. For instance plot (c) illustrates this

[1] Better seen in color.



(a) no noise ($k_x = 400, k_y = 400$), error in log scale.

(b) no noise ($k_x = 400, k_y = 400$) error in linear scale.

(c) no noise ($k_x = 400, k_y = 40$) error in log scale.

(d) no noise ($k_x = 400, k_y = 40$) error in linear scale.

(e) no noise ($k_x = 40, k_y = 40$) error in log scale.

(f) no noise ($k_x = 40, k_y = 40$) error in linear scale.

Figure 2: (a),(c),(e)Error of co-occurring directions versus the deterministic baseline FD-AMM, for clarity the error is given in log scale. (b)(d)(f) Error of co-occurring directions versus randomized baselines (sampling, random projection and hashing), for clarity the error is given in linear scale.

effect, $k_x = 400, k_y = 40$, as ℓ exceeds 50, the error of co-occurring directions sharply decreases , while FD-AMM error is still high. The latter starts a steep decreasing tendency when ℓ exceeds 400. We give plots for the low rank approximation as given in Theorem 3 for $k = \min(k_x, k_y)$ in the appendix, we see a similar trend in the approximation error.

AMM of Noisy Low Rank Matrices (Robustness). We consider the same model as before but we add a gaussian noise to the low rank matrices, i.e $X = V_x S_x U_x^\top + N_x/\zeta_x$, where $\zeta_x > 0$, and $N_x \in \mathbb{R}^{m_x \times n}$, $(N_x)_{i,j} \sim \mathcal{N}(0,1)$. Similarly for $Y = V_y S_y U_y^\top + N_y/\zeta_y$. In this scenario X and Y have still decaying singular values but with non zeros tails. We consider $\zeta_x = 1000$, and $\zeta_y = 100$. We compare here deterministic baselines in Figures 3,4, and 5, in the three scenarios we see that co-occurring directions still outperforms FD-AMM, but the gap between the two approaches becomes smaller in the low rank regimes (Figures 4, and 5), this hints to a weakness in the shrinking of singular values in both algorithms getting affected by the noise (Step 17 in Alg. 2). We give plots for the low rank approximation in the appendix.

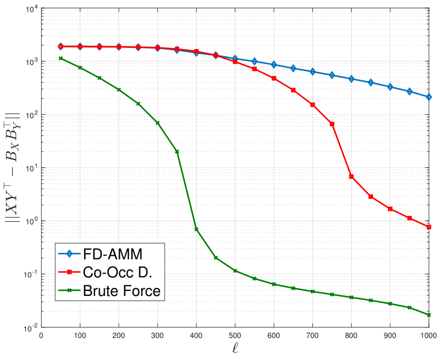


Figure 3: Noisy ($k_x = 400, k_y = 400$). log scale.

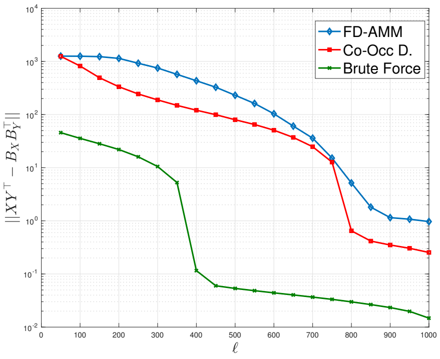


Figure 4: Noisy($k_x = 400, k_y = 40$). Error in log scale.

Multimodal Data Experiments. In this section we study the empirical performance of co-occurring directions in approximating correlation between images and captions. We consider Microsoft COCO [LMB+14] dataset. For visual features we use the residual CNN Resnet101, [HZRS16]. The last layer of Resnet results in a feature vector of dimension $m_x = 2048$. For text we use the Hierarchical Kernel Sentence Embedding HSKE of [MMG16] that results in a feature vector of dimension $m_y = 3000$. The training set size is $n = 113287$. We see in Fig. 6 that co-occurring directions outperforms FD-AMM in this case as well (timing experiment is given in the appendix).

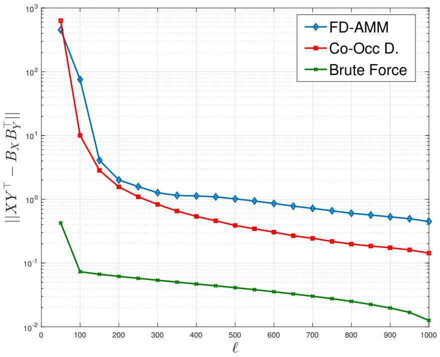


Figure 5: Noisy ($k_x = 40, k_y = 40$). Error in log scale.

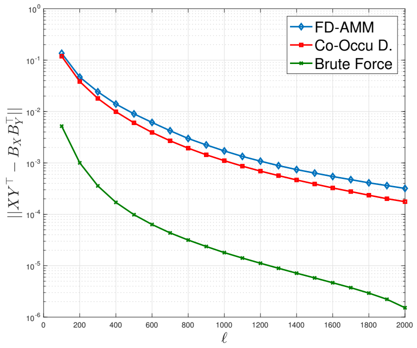


Figure 6: AMM error on MS-COCO.

6 Conclusion

In this paper we introduced a deterministic sketching algorithm for AMM that we termed co-occurring directions . We showed its error bounds (in spectral norm) for AMM and the low rank approximation of a product. We showed empirically that co-occurring directions outperforms deterministic and randomized baselines in the streaming model. Indeed co-occurring direction has the best error/space tradeoff among known baselines with errors given in spectral norm in the streaming model. We are left with two open questions. First, whether guarantees of Theorem 2 can be improved akin to the improved guarantees for *frequent directions* given [GLPW15]. This would give an explicit link of the sketch length ℓ, to the low rank structure of the matrix product $XY^\top$, and/or the low rank structure of the individual matrices. Second, whether robustness of co-occurring directions can be improved using robust shrinkage operators as in [GDP14].

References

[ATKZ14] Michail Vlachos Anastasios T. Kyrillidis and Anastasios Zouzias. Approximate matrix multiplication with application to linear embeddings. In *Corr*, 2014.

[CNW15] Michael B. Cohen, Jelani Nelson, and David P. Woodruff. Optimal approximate matrix product in terms of stable rank. *CoRR*, 2015.

[CW13] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *STOC*, 2013.

[Dhi01] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD*, 2001.

[DKM06] Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast monte carlo algorithms for matrices i: Approximating matrix multiplication. *SIAM J. Comput.*, 2006.

[GDP14] Mina Ghashami, Amey Desai, and Jeff M. Phillips. *Improved Practical Matrix Sketching with Guarantees.* 2014.

[GLPW15] Mina Ghashami, Edo Liberty, Jeff M. Phillips, and David P. Woodruff. Frequent directions : Simple and deterministic matrix sketching. *CoRR*, 2015.

[Hot36] Harold Hotteling. Relations between two sets of variates. *Biometrika*, 1936.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Lib13] Edo Liberty. Simple and deterministic matrix sketching. In *KDD*. ACM, 2013.

[LMB$^+$14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *EECV*, 2014.

[MG82] J. Misra and David Gries. Finding repeated elements. *Science of Computer Programming*, 1982.

[MMG16] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Multimodal retrieval with asymmetrically weighted CCA and hierarchical kernel sentence embedding. *ArXiv*, 2016.

[MZ11] Avner Magen and Anastasios Zouzias. Low rank matrix-valued chernoff bounds and approximate matrix multiplication. In *SODA*, 2011.

[Sar06] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. 2006.

[Weg00] Jacob A. Wegelin. A survey of partial least squares (pls) methods, with emphasis on the two-block case. Technical report, 2000.

[Woo14] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 2014.

[YLZ16] Qiaomin Ye, Luo Luo, and Zhihua Zhang. Frequent direction algorithms for approximate matrix multiplication with applications in CCA. In *IJCAI*, 2016.