Analysis of a randomized approximation scheme for matrix multiplication

Daniel Hsu¹, Sham M. Kakade¹, and Tong Zhang²

¹Microsoft Research, New England
²Department of Statistics, Rutgers University

August 27, 2018

Abstract

This note gives a simple analysis of a randomized approximation scheme for matrix multiplication proposed by [Sar06] based on a random rotation followed by uniform column sampling. The result follows from a matrix version of Bernstein’s inequality and a tail inequality for quadratic forms in subgaussian random vectors.

1 Introduction

Let $A := [a_1|a_2|\cdots|a_m] \in \mathbb{R}^{d_A \times m}$ and $B := [b_1|b_2|\cdots|b_m] \in \mathbb{R}^{d_B \times m}$ be fixed matrices, each with $m$ columns. If $m$ is very large, then the straightforward computation of the matrix product $AB^\top$ (with $\Omega(d_Ad_Bm)$ operations) can be prohibitive.

We can instead approximate the product using the following randomized scheme. Let $\Theta \in \mathbb{R}^{m \times m}$ be a random orthogonal matrix; the distribution of $\Theta$ will be specified later in Theorem 1, but a key property of $\Theta$ will be that the matrix products

$$\tilde{A} := A\Theta \quad \text{and} \quad \tilde{B} := B\Theta$$

can be computed with $O((d_A + d_B)m \log m)$ operations. Given the products $\tilde{A} = [\tilde{a}_1|\tilde{a}_2|\cdots|\tilde{a}_m]$ and $\tilde{B} = [\tilde{b}_1|\tilde{b}_2|\cdots|\tilde{b}_m]$, we take a small uniform random sample of pairs of their columns (drawn with replacement)

$$(\tilde{a}_{i_1}, \tilde{b}_{i_1}), (\tilde{a}_{i_2}, \tilde{b}_{i_2}), \ldots, (\tilde{a}_{i_n}, \tilde{b}_{i_n}),$$

and then compute the sum of outer products

$$\tilde{A}\tilde{B}^\top := \frac{m}{n} \sum_{j=1}^{n} \tilde{a}_{i_j}\tilde{b}_{i_j}^\top.$$

It is easy to check that $\tilde{A}\tilde{B}^\top$ is an unbiased estimator of $AB^\top$. The sum can be computed from $\tilde{A}$ and $\tilde{B}$ with $O(d_Ad_Bn)$ operations, so overall, the matrix $\tilde{A}\tilde{B}^\top$ can be computed with $O(d_Ad_Bn + \ldots$
(d_A + d_B)m \log m) operations. (In fact, the log m can be replaced by \log n \lfloor n \rfloor [AL08].) Therefore, we would like \( n \) to be as small as possible so that, with high probability, \( \| \hat{A}B^\top - AB^\top \| \leq \varepsilon \| A \| \| B \| \) for some error \( \varepsilon > 0 \), where \( \| \cdot \| \) denotes the spectral norm. As shown in Theorem 1, it suffices to have

\[
n = \Omega \left( \frac{(k + \log(m)) \log(k)}{\varepsilon^2} \right),
\]

where \( k := \max \{ \text{tr}(A^\top A) / \| A \|^2, \ \text{tr}(B^\top B) / \| B \|^2 \} \leq \max \{ \text{rank}(A), \ \text{rank}(B) \} \).

A flawed analysis of a different scheme based on non-uniform column sampling (without a random rotation \( \Theta \)) was given in [HKZ12a]; that analysis gave an incorrect bound on \( \| E[X^2] \| \) for a certain random symmetric matrix \( X \). A different analysis of this non-uniform sampling scheme can be found in [MZ11], but that analysis has some deficiencies as pointed out in [HKZ12a]. The scheme studied in the present work, which employs a certain random rotation followed by uniform column sampling, was proposed by [Sar06], and is based on the Fast Johnson-Lindenstrauss Transform of [AC09]. The analysis given in [Sar06] bounds the Frobenius norm error; in this work, we bound the spectral norm error. A similar but slightly looser analysis of spectral norm error was very recently provided in [ABTZ12].

## 2 Analysis

Let \( [m] := \{1, 2, \ldots, m\} \).

**Theorem 1.** Pick any \( \delta \in (0, 1/3) \), and let \( k := \max \{ \text{tr}(AA^\top) / \| A \|^2, \ \text{tr}(BB^\top) / \| B \|^2 \} \) (note that \( k \leq \max \{ \text{rank}(A), \ \text{rank}(B) \} \)). Assume \( \Theta = \frac{1}{\sqrt{m}} DH \), where \( D = \text{diag}(\varepsilon) \), \( \varepsilon \in \{ \pm 1 \}^m \) is a vector of independent Rademacher random variables, and \( H \in \{ \pm 1 \}^{m \times m} \) is a Hadamard matrix. With probability at least \( 1 - \delta \),

\[
\| \hat{A}B^\top - AB^\top \| \leq \| A \| \| B \| \left( \sqrt{\frac{4(k + 2\sqrt{k \ln(3m/\delta)} + 2\ln(3m/\delta) + 1) \ln(6k/\delta)}{n}} + \frac{2(k + 2\sqrt{k \ln(3m/\delta)} + 2\ln(3m/\delta) + 1) \ln(6k/\delta)}{3n} \right).
\]

The proof of Theorem 1 is a consequence of the following lemmas, combined with a union bound. Lemma 1 bounds the error in terms of a certain quantity \( \mu \) which depends on the random orthogonal matrix \( \Theta \) (and \( A \) and \( B \)). Lemma 2 gives a bound on \( \mu \) that holds with high probability over the random choice of \( \Theta \).

**Lemma 1.** Define \( Q = [q_1 q_2 \cdots q_m] := [A]^{-1} A \Theta, \ \ R = [r_1 r_2 \cdots r_m] := [B]^{-1} B \Theta, \ k_A := \text{tr}(QQ^\top) = \text{tr}(AA^\top) / \| A \|^2, \ k_B := \text{tr}(RR^\top) = \text{tr}(BB^\top) / \| B \|^2, \) and

\[
\mu := m \max \left( \{ \| q_i \|^2 : i \in [m] \} \cup \{ \| r_i \|^2 : i \in [m] \} \right).
\]

Then

\[
\Pr \left[ \| \hat{A}B^\top - AB^\top \| > \| A \| \| B \| \left( \sqrt{\frac{2(\mu + 1)t}{n}} + \frac{(\mu + 1)t}{3n} \right) \right] \leq 2\sqrt{k_A k_B} \cdot \frac{t}{e^t - t - 1}.
\]
Proof. Observe that because $\Theta$ is orthogonal,

$$\|\overline{AB^\top} - AB^\top\| = \|AB\| \left\| \frac{m}{n} \sum_{j=1}^{n} q_j r_i^\top - QR^\top \right\|.$$ 

We now derive a high probability bound for the last term on the right-hand side. Define a random symmetric matrix $X$ with

$$\Pr \left[ X = m \begin{bmatrix} 0 & q_r^\top \\ r_i q_i^\top & 0 \end{bmatrix} \right] = \frac{1}{m}, \quad i \in [m],$$

and let $X_1, X_2, \ldots, X_n$ be independent copies of $X$. Define

$$\hat{M} := \frac{1}{n} \sum_{j=1}^{n} X_j \quad \text{and} \quad M := \begin{bmatrix} 0 & QR^\top \\ RQ^\top & 0 \end{bmatrix}.$$

Then

$$\|\hat{M} - M\| = \left\| \frac{1}{n} \sum_{j=1}^{n} (X_j - M) \right\| \text{ distribution } = \left\| \frac{m}{n} \sum_{j=1}^{n} q_j r_i^\top - QR^\top \right\|.$$ 

Observe that $\mathbb{E}[X - M] = 0$ and $\|X - M\| \leq \|X\| + \|M\| \leq \mu + 1$. Moreover,

$$\mathbb{E}[X]^2 = M^2 = \begin{bmatrix} QR^\top QR^\top & 0 \\ 0 & RQ^\top QR^\top \end{bmatrix},$$

$$\mathbb{E}[X^2] = \sum_{i=1}^{m} m \begin{bmatrix} \|r_i\|^2 q_i q_i^\top \\ 0 \\ 0 \\ \|q_i\|^2 r_i r_i^\top \end{bmatrix} = m \begin{bmatrix} \sum_{i=1}^{m} \|r_i\|^2 q_i q_i^\top \\ 0 \\ 0 \\ \sum_{i=1}^{m} \|q_i\|^2 r_i r_i^\top \end{bmatrix},$$

$$\text{tr}(\mathbb{E}[X^2]) = 2m \sum_{i=1}^{m} \|r_i\|^2 \leq 2\mu \sum_{i=1}^{m} \|q_i\| \|r_i\| \leq 2\mu \sqrt{k_A k_B},$$

$$\|\mathbb{E}[X^2]\| \leq m \max \left\{ \left\| \sum_{i=1}^{m} \|r_i\|^2 q_i q_i^\top \right\|, \left\| \sum_{i=1}^{m} \|q_i\|^2 r_i r_i^\top \right\| \right\} \leq \mu \max \left\{ \|QQ^\top\|, \|RR^\top\| \right\} = \mu,$$

$$\|\mathbb{E}[(X - M)^2]\| = \|\mathbb{E}[X^2] - M^2\| \leq \mu + 1.$$ 

Therefore, by the matrix Bernstein inequality from [HKZ12a],

$$\Pr \left[ \|\hat{M} - M\| > \sqrt{\frac{2(\mu + 1)t}{n}} + \frac{(\mu + 1)t}{3n} \right] \leq 2\sqrt{k_A k_B} \cdot \frac{t}{e^t - t - 1}.$$ 

The lemma follows. \hfill $\Box$

The following lemma is a special case of a result found in [HKZ11].

**Lemma 2.** Assume $\Theta = \frac{1}{\sqrt{m}} DH$, where $D = \text{diag}(\epsilon)$, $\epsilon \in \{-1\}^m$ is a vector of independent Rademacher random variables, and $H \in \{-1\}^{m \times m}$ is a Hadamard matrix. Let $Z \in \mathbb{R}^{m \times d}$ be a matrix with $\|Z\| \leq 1$, and set $k_Z := \text{tr}(ZZ^\top)$. Then

$$\Pr \left[ \max \left\{ \|Z^\top \Theta e_i \|^2 : i \in [m] \right\} > \frac{k_Z + 2\sqrt{k_Z (\ln(m) + t)} + 2(\ln(m) + t)}{m} \right] \leq e^{-t}$$

where $e_i \in \{0,1\}^m$ is the $i$-th coordinate axis vector in $\mathbb{R}^m$. 

3
Proof. Observe that for each $i \in [m]$, the random vector $\sqrt{m}\Theta e_i$ has the same distribution as $\epsilon$. Moreover, $\epsilon$ is a subgaussian random vector in the sense that $\mathbb{E}[\exp(\alpha^T \epsilon)] \leq \exp(\|\alpha\|^2/2)$ for any vector $\alpha \in \mathbb{R}^m$. Therefore, we may apply a tail bound for quadratic forms in subgaussian random vectors (e.g., [HKZ12b]) to obtain

$$\Pr \left[ \|\sqrt{m}Z^T \Theta e_i\|^2 > \text{tr}(ZZ^T) + 2\sqrt{\text{tr}((ZZ^T)^2)}\tau + 2\|ZZ^T\|\tau \right] \leq e^{-\tau}$$

for each $i \in [m]$ and any $\tau > 0$. The lemma follows by observing that $\|ZZ^T\| \leq 1$ and $\text{tr}((ZZ^T)^2) \leq \text{tr}(ZZ^T)\|ZZ^T\| \leq kZ$, and applying a union bound over all $i \in [m]$.

We note that Lemma 2 holds for many other distributions of orthonormal matrices (with possibly worse constants). All that is required is that $\sqrt{m}\Theta e_i$ be a subgaussian random vector for each $i \in [m]$. See [HKZ11] for more discussion.

Proof of Theorem 1. We apply Lemma 2 with both $Z = A/\|A\|$ and $Z = B/\|B\|$, and combine the implied probability bounds with a union bound to obtain

$$\Pr \left[ \mu > k + 2\sqrt{k \log(3m/\delta) + 2\ln(3m/\delta)}} \right] \leq 2\delta/3,$$

where $\mu$ is defined in the statement of Lemma 1, and the probability is taken with respect to the random choice of $\Theta$. Now we apply Lemma 1, together with the bound $t/(e^t - t - 1) \leq e^{-t/2}$ for $t \geq 2.6$, and substitute $t := 2\ln(6k/\delta)$ to obtain

$$\Pr \left[ \|\hat{A}B^T - AB^T\| > \|A\|\|B\| \left( \sqrt{\frac{4(\mu + 1)\ln(6k/\delta)}{n}} + \frac{2(\mu + 1)\ln(6k/\delta)}{3n} \right) \right] \leq \delta/3.$$  

Combining the two probability bounds with a union bound implies the claim.

Acknowledgements

We thank Joel Tropp for pointing out the error in the analysis from [HKZ12a].

References


