

# A Note on Low-rank Matrix Decompositions via the Subsampled Randomized Hadamard Transform

Christos Boutsidis  
Computer Science Department  
Rensselaer Polytechnic Institute  
boutsc@cs.rpi.edu

May 10, 2011

## Abstract

We comment on two randomized approximation algorithms for constructing a decomposition of low rank to a given matrix. Both algorithms employ the so-called Subsampled Randomized Hadamard Transform. The first algorithm presented by Halko, Martinsson, and Tropp in [8]; here, we describe a new analysis that strengthens the corresponding approximation bound. A preliminary version of the second algorithm presented by Drineas, Mahoney, and Muthukrishnan in [7]; here, we describe a modification of this algorithm that achieves nearly the same approximation bound but reduces the corresponding computational cost.

## 1 Introduction

Low-rank decompositions to a matrix are ubiquitous in science and engineering. The setting is as follows. Fix  $A \in \mathbb{R}^{m \times n}$  of rank  $\rho$ . It is well known that there exist matrices  $X \in \mathbb{R}^{m \times \rho}$  and  $Y \in \mathbb{R}^{\rho \times n}$  such that  $A = XY$ . These matrices can be computed, for example, via the Singular Value Decomposition (SVD) in  $O(mn \min\{m, n\})$  time. Now, fix target rank  $k < \rho$ . A rank  $k$  decomposition to  $A$  is  $\hat{A} = \hat{X}\hat{Y}$ , where  $\hat{X} \in \mathbb{R}^{m \times k}$ ,  $\hat{Y} \in \mathbb{R}^{k \times n}$ , and  $\hat{A}$  has rank at most  $k$ . The rank  $k$  decomposition that minimizes  $\|A - \hat{A}\|_F$  over all  $\hat{A}$  is obtained via the SVD as well. We denote this best rank  $k$  matrix with  $A_k \in \mathbb{R}^{m \times n}$ .

In this note, we comment on two randomized approximation algorithms [8, 7] for constructing low rank matrix decompositions. Both algorithms employ the so-called Subsampled Randomized Hadamard Transform (see Section 2.2), are faster than the SVD, but achieve nearly the same error.

The first algorithm appeared recently in [8] (see the proto-algorithm and Theorem 11.2 in [8]). Fix  $A \in \mathbb{R}^{m \times n}$  and target rank  $k$ . For some  $r \geq 4 \left( \sqrt{k} + \sqrt{8 \log(kn)} \right)^2 \log(k)$ , let  $\Theta \in \mathbb{R}^{r \times n}$  be a Subsampled Randomized Hadamard Transform matrix (see Definition 6). Now, first construct  $B = A\Theta^T$  (in  $O(mn \log(r))$  time, from Lemma 7); then, orthonormalize the columns of  $B$  to obtain  $Q \in \mathbb{R}^{m \times r}$  (in  $O(mr^2)$  time, using the Gram-Schmidt algorithm).  $Q$  is an approximate orthonormal basis for the column space of  $A$ . Theorem 11.2 in [8] proved that, with probability  $1 - O(1/k)$ ,

$$\|A - QQ^T A\|_F^2 \leq \left(1 + \frac{7n}{r}\right) \|A - A_k\|_F^2.$$

Notice that, for any  $\epsilon > 0$ , obtaining a  $(1 + \epsilon)$ -error  $\|A - QQ^T A\|_F^2 \leq (1 + \epsilon) \|A - A_k\|_F^2$  requires  $r = 7n/\epsilon$ . Since the computational cost to obtain  $Q$  depends on  $r$ , obtaining a better dependence between  $n$  and  $r$  in the approximation bound would be useful.

We provide a new analysis of the algorithm of [8] which yields that, for  $r \geq 8k \log(40k) \log(40kn)$ , with constant probability, the approximation bound is

$$\|A - QQ^T A\|_F^2 \leq \left(1 + O\left(\frac{k \log(k) \log(kn)}{r}\right)\right) \|A - A_k\|_F^2.$$

(See Theorem 10 for a precise statement of our result. <sup>1</sup>) We should note here that  $r$  is given as input by the user and trades the approximation bound with the running time of the algorithm. The new analysis has the same starting point as the analysis in [8], which is Lemma 1 that we present in Section 2. It continues though by manipulating the right hand side in the equation of that Lemma in a different manner. Using a matrix multiplication type result from [5], we are able to obtain the improved bound, which indicates that a Hadamard Transform of size roughly  $r = O(k \log(k)/\epsilon)$  gives, with constant probability, a relative-error approximation. So, in roughly  $O(mn \log(k/\epsilon))$  time one can construct an orthonormal matrix  $Q \in \mathbb{R}^{m \times r}$  such that the rank  $r$  decomposition  $QQ^T A$  is almost as good as the rank  $k$  decomposition obtained via the SVD.

Our second algorithm replaces the matrix  $Q$  of the first algorithm with a matrix  $C$  containing columns of  $A$ . Column-based low rank decompositions have found numerous applications in linear algebra and data analysis (see the discussion in [7]). Drineas et al [7] presented a randomized algorithm constructing such a decomposition. The algorithm of [7] is as follows. Given  $A$  and the target rank  $k$ , construct the matrix  $V_k \in \mathbb{R}^{n \times k}$  with the top  $k$  right singular vectors of  $A$ . Now, construct a probability distribution  $p_1, p_2, \dots, p_n$  over the columns of  $A$  ( $i = 1, \dots, n$ ):

$$p_i = \frac{\|(V_k)_{(i)}\|_2^2}{\|V_k\|_F^2}.$$

( $(V_k)_{(i)}$  denotes the  $i$ -th row of  $V_k$ .) Finally, for some  $r = \Omega(k \log(k))$ , sample  $r$  columns of  $A$  with the corresponding probabilities. Theorem 3 in [7] (see also the discussion in Section 3.6.5 in [7]) proved that, with constant probability,

$$\|A - CC^+ A\|_F \leq \left(1 + O\left(\sqrt{\frac{k \log(k)}{r}}\right)\right) \|A - A_k\|_F,$$

which implies that a  $(1+\epsilon)$ -error  $\|A - CC^+ A\|_F \leq (1+\epsilon) \|A - A_k\|_F$ , requires  $r = O(k \log(k)/\epsilon^2)$  columns. The obvious problem with this approach is that it necessitates the computation of the matrix  $V_k$  from the SVD, which is expensive.

In this note, by employing our first algorithm, we show that a  $(1+\epsilon)$ -error can be achieved by selecting roughly  $O(k \log(k)/\epsilon^2)$  columns that are sampled with probabilities that can be computed in roughly  $O(mn \log(k/\epsilon))$  time (See Theorem 11). More specifically, we use probabilities

$$p_i = \frac{\|Q_{(i)}\|_2^2}{\|Q\|_F^2}.$$

We should note that Sarlos [14] made the same observation that approximate probabilities suffice for column sampling; he gave a similar method with complexity  $O(mnk \log k)$ , while ours is roughly  $O(mn \log k)$ . In the above,  $Q$  is obtained by applying our first algorithm to  $A^T$ . In more details, first, construct  $B^T = \Theta A$ ; then, orthonormalize the rows of  $B^T$  to obtain  $Q^T$ . Although we prove that the matrix  $C$  constructed with the new probabilities is as good as the matrix  $C$  constructed with the probabilities that require the SVD, we do not prove that our sampling probabilities approximate the original probabilities per se. In fact, although progress has been done for short-fat matrices [10] or vertex-by-edge incidence matrices of graphs [15], approximating these probabilities in sub-SVD time appears to be impossible for general matrices [9].

---

<sup>1</sup>A comparable approximation bound to ours was obtained via different arguments in [11].

## 2 Preliminaries

**Basic Notation.** We use  $A, B, \dots$  to denote matrices;  $\mathbf{a}, \mathbf{b}, \dots$  to denote column vectors.  $A = [\mathbf{a}_1, \dots, \mathbf{a}_n] \in \mathbb{R}^{m \times n}$  represents a matrix with columns  $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^m$ .  $I_n$  is the  $n \times n$  identity matrix;  $\mathbf{0}_{m \times n}$  is the  $m \times n$  matrix of zeros;  $\mathbf{e}_i$  is the standard basis (whose dimensionality will be clear from the context).  $A_{(i)}$  denotes the  $i$ -th row of  $A$ ;  $A^{(j)}$  denotes the  $j$ -th column of  $A$ ;  $A_{ij}$  denotes the  $(i, j)$ -th element of  $A$ . Logarithms are base two. We abbreviate “independent identically distributed” to “i.i.d” and “with probability” to “w.p”.

**Sampling Matrices.** Let  $A = [\mathbf{a}_1, \dots, \mathbf{a}_n]$  and  $C = [\mathbf{a}_{i_1}, \dots, \mathbf{a}_{i_r}]$  be  $r$  columns of  $A$ . We can equivalently write  $C = A\Omega$ , where the *sampling matrix* is  $\Omega = [\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_r}]$  and  $\mathbf{e}_i$  are standard basis vectors in  $\mathbb{R}^n$ . Let  $S$  denote an  $r \times r$  diagonal *rescaling matrix* with non-zero entries; then,  $C = A\Omega S$  contains  $r$  columns from  $A$  rescaled with the corresponding diagonal elements of  $S$ . Notice that  $A\Omega(A\Omega)^+ = A\Omega S(A\Omega S)^+$ , because rescaling  $C$  does not change the subspace spanned by its columns.

**Matrix norms.** We use the Frobenius and the spectral norm of a matrix:  $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$  and  $\|A\|_2 = \max_{\mathbf{x}: \|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2$ , respectively. For any two matrices  $A$  and  $B$  of appropriate dimensions,  $\|A\|_2 \leq \|A\|_F \leq \sqrt{\text{rank}(A)}\|A\|_2$ ,  $\|AB\|_F \leq \|A\|_F\|B\|_2$ , and  $\|AB\|_F \leq \|A\|_2\|B\|_F$ . The latter two properties are stronger versions of the standard submultiplicativity property:  $\|AB\|_\xi \leq \|A\|_\xi\|B\|_\xi$ . We will refer to these two stronger versions as spectral submultiplicativity. The notation  $\|A\|_\xi$  indicates that an expression holds for both  $\xi = 2$  and  $\xi = F$ .

**Singular Value Decomposition.** The Singular Value Decomposition (SVD) of the matrix  $A \in \mathbb{R}^{m \times n}$  with  $\text{rank}(A) = \rho$  is:

$$A = \underbrace{\begin{pmatrix} U_k & U_{\rho-k} \end{pmatrix}}_{U_A \in \mathbb{R}^{m \times \rho}} \underbrace{\begin{pmatrix} \Sigma_k & \mathbf{0} \\ \mathbf{0} & \Sigma_{\rho-k} \end{pmatrix}}_{\Sigma_A \in \mathbb{R}^{\rho \times \rho}} \underbrace{\begin{pmatrix} V_k^T \\ V_{\rho-k}^T \end{pmatrix}}_{V_A^T \in \mathbb{R}^{\rho \times n}},$$

with singular values  $\sigma_1 \geq \dots \geq \sigma_k \geq \sigma_{k+1} \geq \dots \geq \sigma_\rho > 0$ . We will use  $\sigma_i(A)$  to denote the  $i$ -th singular value of  $A$ . The matrices  $U_k \in \mathbb{R}^{m \times k}$  and  $U_{\rho-k} \in \mathbb{R}^{m \times (\rho-k)}$  contain the left singular vectors of  $A$ ; and, similarly, the matrices  $V_k \in \mathbb{R}^{n \times k}$  and  $V_{\rho-k} \in \mathbb{R}^{n \times (\rho-k)}$  contain the right singular vectors of  $A$ . It is well-known that  $A_k = U_k \Sigma_k V_k^T \in \mathbb{R}^{m \times n}$  minimizes  $\|A - X\|_\xi$  over all matrices  $X \in \mathbb{R}^{m \times n}$  of rank at most  $k$ . We use  $A_{\rho-k} \in \mathbb{R}^{m \times n}$  to denote the matrix  $A - A_k = U_{\rho-k} \Sigma_{\rho-k} V_{\rho-k}^T \in \mathbb{R}^{m \times n}$ . Also,  $\|A\|_F = \sqrt{\sum_{i=1}^\rho \sigma_i^2(A)}$  and  $\|A\|_2 = \sigma_1(A)$ . The best rank  $k$  approximation to  $A$  satisfies  $\|A - A_k\|_2 = \sigma_{k+1}(A)$  and  $\|A - A_k\|_F = \sqrt{\sum_{i=k+1}^\rho \sigma_i^2(A)}$ . Finally,  $A^+ = V_A \Sigma_A^{-1} U_A^T \in \mathbb{R}^{n \times m}$  denotes the Moore-Penrose pseudo-inverse of  $A \in \mathbb{R}^{m \times n}$  ( $\Sigma_A^{-1}$  is the inverse of  $\Sigma_A$ ). By the SVD of  $A$  and  $A^+$ , it is easy to verify that, for all  $i = 1, \dots, \rho = \text{rank}(A) = \text{rank}(A^+)$ ,  $\sigma_i(A^+) = 1/\sigma_{\rho-i+1}(A)$ .

**Deterministic Result for Low-rank Matrix Reconstruction.** The following lemma appeared recently in [3]. A preliminary version of this lemma appeared before in [4, 8].

**Lemma 1** ([3]). *Fix  $A \in \mathbb{R}^{m \times n}$  and integer  $\hat{k}$ . For some  $Z \in \mathbb{R}^{m \times \hat{k}}$ , let  $Z^T Z = I_{\hat{k}}$ . Let  $W \in \mathbb{R}^{n \times r}$  be any matrix with  $r \geq \hat{k}$  such that  $\text{rank}(Z^T W) = \hat{k} = \text{rank}(Z)$ . Let  $C = AW \in \mathbb{R}^{m \times r}$ . Then,*

$$\|A - CC^+ A\|_F^2 \leq \|A - AZZ^T\|_F^2 + \|(A - AZZ^T)W(Z^T W)^+\|_F^2.$$

## 2.1 Randomized Sampling

**Definition 2** (Random Sampling with Replacement [13]). Let  $\mathbf{X} \in \mathbb{R}^{n \times k}$  with  $n > k$ ,  $\mathbf{x}_i^T \in \mathbb{R}^{1 \times k}$  denotes the  $i$ -th row of  $\mathbf{X}$ , and  $0 < \beta \leq 1$ . For  $i = 1, \dots, n$ , if  $\beta = 1$ , then  $p_i = (\mathbf{x}_i^T \mathbf{x}_i) / \|\mathbf{X}\|_F^2$ , otherwise compute some  $p_i \geq \beta(\mathbf{x}_i^T \mathbf{x}_i) / \|\mathbf{X}\|_F^2$  with  $\sum_{i=1}^n p_i = 1$ . Let  $r$  be an integer with  $1 \leq r \leq n$ . Construct a sampling matrix  $\Omega \in \mathbb{R}^{n \times r}$  and a rescaling matrix  $\mathbf{S} \in \mathbb{R}^{r \times r}$  as follows. Initially,  $\Omega = \mathbf{0}_{n \times r}$  and  $\mathbf{S} = \mathbf{0}_{r \times r}$ . Then, for every column  $j = 1, \dots, r$  of  $\Omega, \mathbf{S}$ , independently, pick an index  $i$  from the set  $\{1, 2, \dots, n\}$  with probability  $p_i$  and set  $\Omega_{ij} = 1$  and  $\mathbf{S}_{jj} = 1/\sqrt{p_i r}$ . To denote this  $O(nk + r \log(r))$  time randomized algorithm we will write,

$$[\Omega, \mathbf{S}] = \text{RandomizedSampling}(\mathbf{X}, \beta, r).$$

(It takes  $O(nk)$  to compute the probabilities and  $O(r \log(r))$  to sample  $r$  indices with replacement.) Now consider applying this randomized algorithm to an orthonormal matrix.

**Lemma 3** (Originally proved in [13] with an unspecified constant). Let  $\mathbf{V} \in \mathbb{R}^{n \times k}$  with  $n > k$  and  $\mathbf{V}^T \mathbf{V} = \mathbf{I}_k$ . Let  $0 < \beta \leq 1$ ,  $0 < \delta \leq 1$ , and  $4k \ln(2k/\delta)/\beta < r \leq n$ . Let  $[\Omega, \mathbf{S}] = \text{RandomizedSampling}(\mathbf{V}, \beta, r)$ . Then, for all  $i = 1, \dots, k$ , w.p. at least  $1 - \delta$ :

$$1 - \sqrt{\frac{4k \ln(2k/\delta)}{r\beta}} \leq \sigma_i^2(\mathbf{V}^T \Omega \mathbf{S}) \leq 1 + \sqrt{\frac{4k \ln(2k/\delta)}{r\beta}}.$$

*Proof.* In Theorem 2 of [10], set  $\mathbf{S} = \mathbf{I}$  (the identity matrix of appropriate dimension) and replace  $\epsilon$  in terms of  $r, \beta, d$ . The lemma is proved; one should be careful to fit this into our notation. ■

**Lemma 4.** For any  $\beta, r, \mathbf{X} \in \mathbb{R}^{n \times k}$ , and  $\mathbf{Y} \in \mathbb{R}^{m \times n}$ , let  $[\Omega, \mathbf{S}] = \text{RandomizedSampling}(\mathbf{X}, \beta, r)$ ; then, w.p.  $1 - \delta$ :  $\|\mathbf{Y} \Omega \mathbf{S}\|_F^2 \leq \frac{1}{\delta} \|\mathbf{Y}\|_F^2$ .

*Proof.* Let  $x = \|\mathbf{Y} \Omega \mathbf{S}\|_F^2$  be a random variable with nonnegative values. Assume that the following equation is true:  $\mathbf{E} [\|\mathbf{Y} \Omega \mathbf{S}\|_F^2] = \|\mathbf{Y}\|_F^2$ . Applying Markov's inequality to this equation gives the bound in the lemma. All that it remains to prove now is the above assumption. Let  $\mathbf{X} = \mathbf{Y} \Omega \mathbf{S} \in \mathbb{R}^{m \times r}$ , and for  $t = 1, \dots, r$ , let  $\mathbf{X}^{(t)}$  denotes the  $t$ -th column of  $\mathbf{X} = \mathbf{Y} \Omega \mathbf{S}$ . We manipulate the term  $\mathbf{E} [\|\mathbf{Y} \Omega \mathbf{S}\|_F^2]$  as follows:

$$\mathbf{E} [\|\mathbf{Y} \Omega \mathbf{S}\|_F^2] \stackrel{(a)}{=} \mathbf{E} \left[ \sum_{t=1}^r \|\mathbf{X}^{(t)}\|_2^2 \right] \stackrel{(b)}{=} \sum_{t=1}^r \mathbf{E} \left[ \|\mathbf{X}^{(t)}\|_2^2 \right] \stackrel{(c)}{=} \sum_{t=1}^r \sum_{j=1}^n p_j \frac{\|\mathbf{Y}^{(j)}\|_2^2}{r p_j} \stackrel{(d)}{=} \frac{1}{r} \sum_{t=1}^r \|\mathbf{Y}\|_F^2 = \|\mathbf{Y}\|_F^2.$$

(a) follows by the definition of the Frobenius norm of  $\mathbf{X}$ . (b) follows by the linearity of expectation. (c) follows by our construction of  $\Omega, \mathbf{S}$ . (d) follows by the definition of the Frobenius norm of  $\mathbf{Y}$ . ■

## 2.2 Subsampled Randomized Hadamard Transform

We give the definitions of the ‘‘Normalized Walsh-Hadamard’’ and the ‘‘Subsampled Randomized Hadamard Transform’’ matrices as well as a few basic facts for computations with such matrices. This form of structured dimension reduction was introduced in [1] and subsequently refined/used in [14, 6, 12, 2, 11, 16].

**Definition 5** (Normalized Walsh-Hadamard Matrix). Fix an integer  $m = 2^p$ , for  $p = 1, 2, 3, \dots$ . The (non-normalized)  $m \times m$  matrix of the Hadamard-Walsh transform is defined recursively as:

$$\mathbf{H}_m = \begin{bmatrix} \mathbf{H}_{m/2} & \mathbf{H}_{m/2} \\ \mathbf{H}_{m/2} & -\mathbf{H}_{m/2} \end{bmatrix}, \quad \text{with} \quad \mathbf{H}_2 = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix}.$$

The  $m \times m$  normalized matrix of the Walsh-Hadamard transform is equal to  $\mathbf{H} = m^{-\frac{1}{2}} \mathbf{H}_m$ .

**Definition 6** (Subsampled Randomized Hadamard Transform (SRHT) matrix). *Fix integers  $r$  and  $m = 2^p$  with  $r < m$  and  $p = 1, 2, 3, \dots$ . A SRHT matrix is an  $r \times m$  matrix of the form*

$$\Theta = S^T \Omega^T D H;$$

- $H \in \mathbb{R}^{m \times m}$  is a normalized Walsh-Hadamard matrix.
- $D \in \mathbb{R}^{m \times m}$  is a diagonal matrix constructed as follows: each diagonal element is a random variable taking values  $\{+1, -1\}$  with equal probability.
- $\Omega \in \mathbb{R}^{m \times r}$  is a sampling matrix constructed as follows: for  $j = 1, 2, \dots, r$  i.i.d random trials pick a vector  $\mathbf{e}_j$  from the standard basis of  $\mathbb{R}^m$  with probability  $\frac{1}{m}$  (uniform sampling) and set the  $j$ -th column of  $\Omega$  equal to that vector.
- $S \in \mathbb{R}^{r \times r}$  is a rescaling (diagonal) matrix containing the value  $\sqrt{\frac{m}{r}}$ .

**Proposition 7** (Fast Matrix-Vector Multiplication, Theorem 2.1 in [2]). *Given  $\mathbf{x} \in \mathbb{R}^m$  and  $r < n$ , one can construct  $\Theta \in \mathbb{R}^{r \times m}$  and compute  $\Theta \mathbf{x}$  with at most  $2m \log(r + 1)$  operations.*

**Lemma 8** ([1], Lemma 3 in [6]). *Let  $U \in \mathbb{R}^{m \times k}$  has orthonormal columns. Let  $(DHU)_{(i)}$  denotes the  $i$ -th row of the matrix  $DHU \in \mathbb{R}^{m \times k}$  and  $\mathcal{E}_i$  denotes the probabilistic event that  $\|(DHU)_{(i)}\|_2^2 \leq \frac{2k \log(40mk)}{m}$  (over the randomness of  $D$ ):  $\Pr[\mathcal{E}_1 \cup \mathcal{E}_2 \dots \cup \mathcal{E}_m] \leq 0.95$ .*

The above lemma was essentially proved in [1]. We chose to state a form that appeared in [6]. A mild improvement of this result is in Lemma 3.3 in [16]. Now, we consider multiplying an orthonormal matrix with a SRHT matrix. Let  $U \in \mathbb{R}^{m \times k}$  has orthonormal columns and  $m \gg k$ . Lemma 9 studies the singular values of the matrix  $\Theta U$ . Note also that, by construction,  $HH^T = I_m$  and  $HDD^T H^T = DHH^T D = I_m$ .

**Lemma 9.** *Let  $U \in \mathbb{R}^{m \times k}$  with  $m > k$  and  $U^T U = I_k$ . Let  $\beta = \frac{1}{2 \log(40km)}$ , and  $4k \ln(2k/\delta)/\beta < r \leq m$ . Let  $[\Omega, S] = \text{RandomizedSampling}(DHU, \frac{1}{2 \log(40km)}, r)$  and  $p_i = 1/m$  in Definition 2. Then, for all  $i = 1, \dots, k$ , w.p. at least  $0.95 - \delta$ :*

$$1 - \sqrt{\frac{8k \ln(2k/\delta) \log(40km)}{r}} \leq \sigma_i^2(U^T D^T H^T \Omega S) = \sigma_i^2(\Theta U) \leq 1 + \sqrt{\frac{8k \ln(2k/\delta) \log(40km)}{r}}.$$

*Proof.* The result in the lemma is very similar with the result of Lemma 3 with the only difference being the fact that before applying the *RandomizedSampling* method on the rows of  $U$  we pre-multiply it with a randomized Hadamard Transform, i.e.

$$[\Omega, S] = \text{RandomizedSampling}(DHU, \frac{1}{2 \log(40km)}, r).$$

Since  $\beta < 1$ , we need to specify the sampling probabilities  $p_i$ 's in Definition 2:

$$p_i = \frac{1}{m} \geq \frac{1}{2 \log(40km)} \frac{\|(DHU)_{(i)}\|_2^2}{k} = \beta \frac{\|(DHU)_{(i)}\|_2^2}{k}.$$

The inequality in this derivation is from Lemma 8. Using the bounds of Lemma 3 concludes the proof. The failure probability follows by a simple union bound.  $\blacksquare$

This result appeared also in [6] but with a slightly larger constant than ours, which is 8. A mild *asymptotic* improvement on the bounds of this lemma can be found in [16].

### 3 Results and Discussion

We start by analyzing the proto-algorithm of [8] implemented as indicated in Theorem 11.2 of [8]. Theorem 10 below gives the details of this algorithm and the new approximation bound. We state the bound with respect to a parameter  $\epsilon > 0$ . The proof of this theorem is in the Appendix.

**Theorem 10.** *Fix  $A \in \mathbb{R}^{m \times n}$  of rank  $\rho$ , target rank  $k < \rho$ , and parameter  $0 < \epsilon < 1/2$ . Construct an orthonormal matrix  $Q \in \mathbb{R}^{m \times r}$  as follows.*

- 1: Let  $r = 200 \cdot k \cdot \ln(40k) \cdot \log(40kn)/\epsilon$ .
- 2: Using definition 6 construct a SRHT matrix  $\Theta \in \mathbb{R}^{r \times n}$ .
- 3: Construct the matrix  $B = A\Theta^T$ .
- 4: Orthonormalize the columns in  $B$  to obtain  $Q \in \mathbb{R}^{m \times r}$ .

Then, with probability at least 0.7:

$$\|A - QQ^T A\|_F^2 \leq (1 + \epsilon) \|A - A_k\|_F^2.$$

The matrix  $Q$  can be constructed in  $O(mn \log(k \log(kn))/\epsilon) + mk^2 \ln^2(k) \log^2(kn)/\epsilon^2$  time.

Theorem 11.2 of [8] showed that a  $(1 + \epsilon)$ -error requires a Hadamard matrix of size  $O(n/\epsilon)$ , which implies a corresponding running time of the order roughly  $O(mn \log(n/\epsilon))$ .

We continue by describing a modification of the algorithm of Theorem 3 in [7]. This algorithm samples columns from the input matrix with probabilities that are computed via the SVD. We prove that “approximate” probabilities can do almost as well. We will compute these “approximate” probabilities via the Subsampled Randomized Hadamard Transform matrix, i.e. by employing the algorithm of Theorem 10. The proof of Theorem 11 is in the Appendix as well.

**Theorem 11.** *Fix  $A \in \mathbb{R}^{m \times n}$  of rank  $\rho$ , target rank  $k < \rho$ , and parameter  $0 < \epsilon < 1/2$ . Construct a matrix  $C \in \mathbb{R}^{m \times r}$  with columns of  $A$  as indicated below. The number of columns is*

$$r = 104 \cdot 200 \cdot k \cdot \ln(40k) \cdot \log(40km) \cdot \ln(40 \cdot 200 \cdot k \cdot \ln(40k) \cdot \log(40km)/\epsilon)/\epsilon^2.$$

- 1: Let  $\hat{r} = 200 \cdot k \cdot \ln(40k) \cdot \log(40km)/\epsilon$ .
- 2: Using definition 6 construct a SRHT matrix  $\Theta \in \mathbb{R}^{\hat{r} \times m}$ .
- 3: Construct the matrix  $B^T = \Theta A \in \mathbb{R}^{\hat{r} \times n}$ .
- 4: Orthonormalize the rows in  $B^T$  to obtain  $Q^T \in \mathbb{R}^{\hat{r} \times n}$ .
- 5: Let  $r = 104 \cdot \hat{r} \cdot \ln(40\hat{r})/\epsilon$ .
- 6: Using definition 2, let  $[\Omega, S] = \text{RandomizedSampling}(Q, 1, r)$ .
- 7: Return  $C = A\Omega S$  with  $r$  (rescaled) columns of  $A$ .

Then, with probability at least 0.5:

$$\|A - CC^+ A\|_F \leq (1 + \epsilon) \|A - A_k\|_F.$$

$C$  can be constructed in  $O(mn \log(k \log(km))/\epsilon) + nk^2 \ln^2(k) \log^2(km)/\epsilon^2 + r \log(r)$  time.

The running time of this algorithm is remarkable, which is essentially  $O(mn \log(k/\epsilon))$ . The number of sampled columns thought is somewhat larger than what we would like to be. In particular, it would be nice to get rid of the constant 20,800 and the term  $O(\log(km))$ . A straightforward modification of this algorithm that replaces  $Q \in \mathbb{R}^{n \times \hat{r}}$  with the  $n \times k$  matrix of the right singular vectors of  $AQQ^T$  achieves a  $(1 + \epsilon)$ -error with  $r = 104k \ln(40k)/\epsilon$  columns in  $O(mnk/\epsilon)$  time. Also, it is possible to use Lemma 3.3 of [16] in place of Lemma 8 to asymptotically reduce the number of columns (which will still be  $\omega(k \log(k))$ ) while keeping the cost to  $O(mn \log(k/\epsilon))$ . Finally, using adaptive sampling as in Theorem 5 of [3] gives a  $(1 + \epsilon)$ -error with  $r = 4k \ln(40k) + O(k/\epsilon)$  columns in  $O(mnk/\epsilon)$  time. We leave it as an open question of whether it is possible to achieve a  $(1 + \epsilon)$ -error with  $O(k \log(k)/\epsilon)$  columns in  $O(mn \log(k/\epsilon))$  time.

**Acknowledgement.** I would like to thank Joel Tropp for encouraging me to study if there is a better analysis of the Frobenius norm error term in Theorem 11.2 in [8].

## References

- [1] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, 2006.
- [2] N. Ailon and E. Liberty. Fast dimension reduction using rademacher series on dual bch codes. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2008.
- [3] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near optimal column based matrix reconstruction. *arXiv:1103.0995*, 2011.
- [4] C. Boutsidis, M. W. Mahoney, and P. Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2009.
- [5] P. Drineas, R. Kannan, and M. Mahoney. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal of Computing*, 36(1):132–157, 2006.
- [6] P. Drineas, M. Mahoney, S. Muthukrishnan, and T. Sarlos. Faster least squares approximation. *arXiv:0710.1435*, 2007.
- [7] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal Matrix Analysis and Applications*, 30(2):844–881, 2008.
- [8] N. Halko, P. Martinsson, and J. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, to appear.
- [9] M. Magdon Ismail. Personal Communication.
- [10] M. Magdon-Ismail. Row Sampling for Matrix Algorithms via a Non-Commutative Bernstein Bound. *arXiv:1008.0587*, 2010.
- [11] N. Nguyen, T. Do, and T. Tran. A fast and efficient algorithm for low-rank approximation of a matrix. In *Proceedings of the ACM symposium on Theory of computing (STOC)*, 2009.
- [12] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. In *Proceedings of the National Academy of Sciences*, 105(36):13212, 2008.
- [13] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, 54, 2007.
- [14] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS)*, 2006.
- [15] N. Srivastava and D. Spielman. Graph sparsifications by effective resistances. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, 2008.
- [16] J. Tropp. Improved analysis of the subsampled randomized hadamard transform. *Adv. Adapt. Data Anal., special issue, "Sparse Representation of Data and Images*, 2011.

## A Proof of Theorem 10

We first comment on the running time. Step 3 takes  $O(mn \log(r))$  (Lemma 7). Step 4 takes  $O(mr^2)$ . Our choice of  $r$  gives the overall running time. We continue by manipulating the term  $\|A - QQ^T A\|_F^2$ . We would like to apply Lemma 1 with  $Z = V_k \in \mathbb{R}^{n \times k}$  and  $W = \Theta^T \in \mathbb{R}^{n \times r}$ . First, notice that Lemma 9 gives:

$$1 - \sqrt{\frac{8k \ln(2k/\delta) \log(40kn)}{r}} \leq \sigma_i^2(\Theta V_k) \leq 1 + \sqrt{\frac{8k \ln(2k/\delta) \log(40kn)}{r}}.$$

Now, our choice of  $r$  with  $\delta = 0.05$  imply that w.p. at least 0.9:

$$1 - \frac{\sqrt{\epsilon}}{\sqrt{25}} \leq \sigma_i^2(\Theta V_k) \leq 1 + \frac{\sqrt{\epsilon}}{\sqrt{25}}.$$

The assumption on  $\epsilon < 1/2$  and the left hand side of this inequality imply that with probability 0.9:  $\text{rank}(V_k^T \Theta^T) = k = \text{rank}(V_k)$ ; so, we can apply Lemma 1 (with a failure probability 0.1):

$$\|A - QQ^T A\|_F^2 \leq \|A - A_k\|_F^2 + \|(A - A_k)\Theta^T(V_k^T \Theta^T)^+\|_F^2.$$

We will return to this generic equation later. First, we prove three results of independent interest.

**A bound for the singular values.** Recall that, by Lemma 9 and our choice of  $r$ , for all  $i = 1, \dots, k$  and w.p. 0.9:  $1 - \frac{\sqrt{\epsilon}}{\sqrt{25}} \leq \sigma_i^2(V_k^T \Theta^T) \leq 1 + \frac{\sqrt{\epsilon}}{\sqrt{25}}$ . Let  $X = V_k^T \Theta^T \in \mathbb{R}^{k \times r}$  with SVD:  $X = U_X \Sigma_X V_X^T$ . Here,  $U_X \in \mathbb{R}^{k \times k}$ ,  $\Sigma_X \in \mathbb{R}^{k \times k}$ , and  $V_X \in \mathbb{R}^{r \times k}$ . By taking the SVD of  $X^+$ ,  $X^T$ :

$$\|(V_k^T \Theta^T)^+ - (V_k^T \Theta^T)^T\|_2 = \|V_X \Sigma_X^{-1} U_X^T - V_X \Sigma_X U_X^T\|_2 = \|\Sigma_X^{-1} - \Sigma_X\|_2,$$

since  $V_X$  and  $U_X^T$  can be dropped without changing any unitarily invariant norm. Let  $Y = \Sigma_X^{-1} - \Sigma_X \in \mathbb{R}^{k \times k}$  be diagonal. Assuming that, for all  $i = 1, \dots, k$ ,  $\tau_i(Y)$  denotes the  $i$ -th diagonal element of  $Y$ :  $\tau_i(Y) = \frac{1 - \sigma_i^2(X)}{\sigma_i(X)}$ . Since  $Y$  is a diagonal matrix:

$$\|Y\|_2 = \max_{1 \leq i \leq k} |\tau_i(Y)| = \max_{1 \leq i \leq k} \frac{|1 - \sigma_i^2(X)|}{\sigma_i(X)} \leq \frac{\frac{\sqrt{\epsilon}}{\sqrt{25}}}{\sqrt{1 - \frac{\sqrt{\epsilon}}{\sqrt{25}}}}.$$

The inequality follows by using the bounds for  $\sigma_i^2(X)$  from above. The failure probability is 0.1 because the bounds for  $\sigma_i^2(X)$  fail with this probability. Overall, we proved that w.p. 0.9,

$$\|(V_k^T \Theta^T)^+ - (V_k^T \Theta^T)^T\|_2 \leq \frac{\frac{\sqrt{\epsilon}}{\sqrt{25}}}{\sqrt{1 - \frac{\sqrt{\epsilon}}{\sqrt{25}}}}.$$

**A matrix-multiplication-type bound.** Consider the term:  $\|(A - A_k)\Theta^T \Theta V_k\|_F^2$ . We would like to upper bound this term. Recall that  $\Theta^T = HD\Omega S \in \mathbb{R}^{n \times r}$ . Eqn. (4) of Lemma 4 of [5] gives a result for the above matrix-multiplication-type term and any set of probabilities  $p_1, p_2, \dots, p_n$  (for notational convenience, let  $X = (A - A_k)HD$  and  $Y = D^T H^T V_k$ ):

$$\mathbf{E} [\|(A - A_k)HDD^T H^T V_k - (A - A_k)HD\Omega S S^T \Omega^T D^T H^T V_k\|_F^2] \leq \sum_{i=1}^n \frac{\|X^{(i)}\|_2^2 \|Y_{(i)}\|_2^2}{r p_i} - \frac{1}{r} \|XY\|_F^2.$$

First, notice that  $XY = \mathbf{0}_{m \times k}$ . Our choice of  $p_i$ 's is:

$$p_i = \frac{1}{n} \geq \frac{1}{2 \log(40kn)} \frac{\|(\text{DHV}_k)_{(i)}\|_2^2}{k}.$$

By using this inequality and rearranging:

$$\mathbf{E} [\|(A - A_k)\Theta^T \Theta V_k\|_F^2] \leq \frac{2k \log(40kn)}{r} \|(A - A_k)\text{HD}\|_F^2 = \frac{2k \log(40kn)}{r} \|A - A_k\|_F^2,$$

since HD can be dropped without changing the Frobenius norm. Finally, apply Markov's inequality to the random variable  $x = \|(A - A_k)\Theta^T \Theta V_k\|_F^2$  to get that with probability 0.9

$$\|(A - A_k)\Theta^T \Theta V_k\|_F^2 \leq \frac{20k \log(40kn)}{r} \|A - A_k\|_F^2.$$

**A Frobenius norm bound.** We would like to compute an upper bound for the term  $\|(A - A_k)\Theta^T\|_F^2$ . Replace  $\Theta^T = \text{HD}\Omega \in \mathbb{R}^{n \times r}$ . Then, Lemma 4 on the random variable  $x = \|(A - A_k)\Theta^T\|_F^2$  implies that with probability 0.9:

$$\|(A - A_k)\Theta^T\|_F^2 \leq 10 \|(A - A_k)\text{HD}\|_F^2.$$

Notice that HD can be dropped without changing the Frobenius norm; so, w.p. 0.9:

$$\|(A - A_k)\Theta^T\|_F^2 \leq 10 \|A - A_k\|_F^2.$$

**Concluding the proof .** Equipped with the above bounds, we are ready to conclude the proof of the theorem. We continue as follows:  $\|A - \text{QQ}^T A\|_F^2 \leq$

$$\begin{aligned} &\leq \|A - A_k\|_F^2 + \|(A - A_k)\Theta^T (\mathbf{V}_k^T \Theta^T)^+\|_F^2 \\ &\leq \|A - A_k\|_F^2 + 2\|(A - A_k)\Theta^T \Theta V_k\|_F^2 + 2\|(A - A_k)\Theta^T ((\mathbf{V}_k^T \Theta^T)^+ - (\mathbf{V}_k^T \Theta^T)^T)\|_F^2 \\ &\leq \|A - A_k\|_F^2 + 2\|(A - A_k)\Theta^T \Theta V_k\|_F^2 + 2\|(A - A_k)\Theta^T\|_F^2 \|((\mathbf{V}_k^T \Theta^T)^+ - (\mathbf{V}_k^T \Theta^T)^T)\|_F^2 \\ &\leq \|A - A_k\|_F^2 + 2 \frac{20k \log(40kn)}{r} \|A - A_k\|_F^2 + 2 \cdot 10 \|A - A_k\|_F^2 \left(\frac{\epsilon}{25}\right) \left(1 - \frac{\sqrt{\epsilon}}{\sqrt{25}}\right)^{-1} \\ &\leq \|A - A_k\|_F^2 + \left(\frac{40}{200 \ln(40)} + \frac{20/25}{1 - 1/\sqrt{2 \cdot 25}}\right) \epsilon \|A - A_k\|_F^2 \leq (1 + 0.986 \cdot \epsilon) \|A - A_k\|_F^2 \end{aligned}$$

The failure probability follows by a union bound on all the probabilistic events involved in the proof of the theorem.

## B Proof of Theorem 11

We first comment on the running time. Steps 1-4 correspond to the algorithm of Theorem 10 applied on  $A^T$ . The running time of the remaining steps are from Lemma 3. Our choice of  $r$  and  $\hat{r}$  gives the overall running time. Notice that  $\|A - \text{AQQ}^T\|_F^2 = \|A^T - \text{QQ}^T A^T\|_F^2$ . The later term is bounded with  $(1 + \epsilon)\|A - A_k\|_F^2$  (from Theorem 10). Now, we continue by manipulating the term  $\|A - \text{CC}^+ A\|_F^2$ . We would like to apply Lemma 1 with  $Z = Q \in \mathbb{R}^{n \times \hat{r}}$  and  $W = \Omega \in \mathbb{R}^{n \times r}$ . First, notice that Lemma 3 gives w.p.  $1 - \delta$ :

$$1 - \sqrt{\frac{4\hat{r} \ln(2\hat{r}/\delta)}{r}} \leq \sigma_i^2(Q^T \Omega) \leq 1 + \sqrt{\frac{4\hat{r} \ln(2\hat{r}/\delta)}{r}}.$$

Now, our choice of  $r$  with  $\delta = 0.05$  imply that w.p. at least 0.95 for all  $i = 1, \dots, \hat{r}$ :

$$1 - \frac{\sqrt{\epsilon}}{\sqrt{26}} \leq \sigma_i^2(\mathbf{Q}^T \Omega \mathbf{S}) \leq 1 + \frac{\sqrt{\epsilon}}{\sqrt{26}}.$$

The assumption on  $\epsilon < 1/2$  and the left hand side of this inequality imply that w.p. at least 0.95  $\text{rank}(\mathbf{Q}^T \Omega \mathbf{S}) = \hat{r} = \text{rank}(\mathbf{Q}^T)$ ; so, we can apply Lemma 1:

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^+ \mathbf{A}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}\mathbf{Q}\mathbf{Q}^T\|_F^2 + \|(\mathbf{A} - \mathbf{A}\mathbf{Q}\mathbf{Q}^T)\Omega\mathbf{S}(\mathbf{Q}^T \Omega \mathbf{S})^+\|_F^2.$$

**A bound for the singular values.** Recall that by Lemma 3 and our choice of  $r$ , for all  $i = 1, \dots, \hat{r}$  and w.p. 0.95:  $1 - \frac{\sqrt{\epsilon}}{\sqrt{26}} \leq \sigma_i^2(\mathbf{Q}^T \Omega \mathbf{S}) \leq 1 + \frac{\sqrt{\epsilon}}{\sqrt{26}}$ . Let  $\mathbf{X} = \mathbf{Q}^T \Omega \mathbf{S} \in \mathbb{R}^{\hat{r} \times r}$  with SVD:  $\mathbf{X} = \mathbf{U}_X \Sigma_X \mathbf{V}_X^T$ . Here,  $\mathbf{U}_X \in \mathbb{R}^{\hat{r} \times \hat{r}}$ ,  $\Sigma_X \in \mathbb{R}^{\hat{r} \times \hat{r}}$ , and  $\mathbf{V}_X \in \mathbb{R}^{r \times \hat{r}}$ . By taking the SVD of  $\mathbf{X}^+$ ,  $\mathbf{X}^T$ :

$$\|(\mathbf{Q}^T \Omega \mathbf{S})^+ - (\mathbf{Q}^T \Omega \mathbf{S})^T\|_2 = \|\mathbf{V}_X \Sigma_X^{-1} \mathbf{U}_X^T - \mathbf{V}_X \Sigma_X \mathbf{U}_X^T\|_2 = \|\Sigma_X^{-1} - \Sigma_X\|_2,$$

since  $\mathbf{V}_X$  and  $\mathbf{U}_X^T$  can be dropped without changing any unitarily invariant norm. Let  $\mathbf{Y} = \Sigma_X^{-1} - \Sigma_X \in \mathbb{R}^{\hat{r} \times \hat{r}}$  be diagonal; Assuming that, for all  $i = 1, \dots, \hat{r}$ ,  $\tau_i(\mathbf{Y})$  denotes the  $i$ -th diagonal element of  $\mathbf{Y}$ :  $\tau_i(\mathbf{Y}) = \frac{1 - \sigma_i^2(\mathbf{X})}{\sigma_i(\mathbf{X})}$ . Since  $\mathbf{Y}$  is a diagonal matrix:

$$\|\mathbf{Y}\|_2 = \max_{1 \leq i \leq \hat{r}} |\tau_i(\mathbf{Y})| = \max_{1 \leq i \leq \hat{r}} \frac{|1 - \sigma_i^2(\mathbf{X})|}{\sigma_i(\mathbf{X})} \leq \frac{\frac{\sqrt{\epsilon}}{\sqrt{26}}}{\sqrt{1 - \frac{\sqrt{\epsilon}}{\sqrt{26}}}}.$$

The inequality follows by using the bounds for  $\sigma_i^2(\mathbf{X})$  from above. The failure probability is 0.05 because the bounds for  $\sigma_i^2(\mathbf{X})$  fail with this probability. Overall, we proved that w.p. 0.95

$$\|(\mathbf{Q}^T \Omega \mathbf{S})^+ - (\mathbf{Q}^T \Omega \mathbf{S})^T\|_2 \leq \frac{\frac{\sqrt{\epsilon}}{\sqrt{26}}}{\sqrt{1 - \frac{\sqrt{\epsilon}}{\sqrt{26}}}}.$$

**A matrix-multiplication-type bound.** Consider the term:  $\|(\mathbf{A} - \mathbf{A}\mathbf{Q}\mathbf{Q}^T)\Omega\mathbf{S}\mathbf{S}^T\Omega^T\mathbf{Q}\|_F^2$ . We would like to upper bound this term. Eqn. (4) of Lemma 4 of [5] gives a result for the above matrix-multiplication-type term and any set of probabilities  $p_1, p_2, \dots, p_n$  (for notational convenience, let  $\mathbf{E} = \mathbf{A} - \mathbf{A}\mathbf{Q}\mathbf{Q}^T$ ,  $\mathbf{Z} = \mathbf{Q}$ ):

$$\mathbf{E} [\|(\mathbf{A} - \mathbf{A}\mathbf{Q}\mathbf{Q}^T)\mathbf{Q} - (\mathbf{A} - \mathbf{A}\mathbf{Q}\mathbf{Q}^T)\Omega\mathbf{S}\mathbf{S}^T\Omega^T\mathbf{Q}\|_F^2] \leq \sum_{i=1}^n \frac{\|\mathbf{E}^{(i)}\|_2^2 \|\mathbf{Z}_{(i)}\|_2^2}{r p_i} - \frac{1}{r} \|\mathbf{E}\mathbf{Z}\|_F^2.$$

First, notice that  $\mathbf{E}\mathbf{Z} = \mathbf{0}_{m \times \hat{r}}$ . Our choice of  $p_i$ 's is:

$$p_i = \frac{\|\mathbf{Q}_{(i)}\|_2^2}{\|\mathbf{Q}\|_F^2}.$$

By using this inequality and rearranging:

$$\mathbf{E} [\|(\mathbf{A} - \mathbf{A}\mathbf{Q}\mathbf{Q}^T)\Omega\mathbf{S}\mathbf{S}^T\Omega^T\mathbf{Q}\|_F^2] \leq \frac{\hat{r}}{r} \|\mathbf{A} - \mathbf{A}\mathbf{Q}\mathbf{Q}^T\|_F^2.$$

Finally, apply Markov's inequality to the random variable  $x = \|(\mathbf{A} - \mathbf{A}\mathbf{Q}\mathbf{Q}^T)\Omega\mathbf{S}\mathbf{S}^T\Omega^T\mathbf{Q}\|_F^2$  to get that with probability 0.95

$$\|(\mathbf{A} - \mathbf{A}\mathbf{Q}\mathbf{Q}^T)\Omega\mathbf{S}\mathbf{S}^T\Omega^T\mathbf{Q}\|_F^2 \leq \frac{20\hat{r}}{r} \|\mathbf{A} - \mathbf{A}\mathbf{Q}\mathbf{Q}^T\|_F^2.$$

**A Frobenius norm bound.** We would like to compute an upper bound for the term  $\|(A - AQQ^T)\Omega S\|_F^2$ . Lemma 4 on the random variable  $x = \|(A - AQQ^T)\Omega S\|_F^2$  implies that with probability 0.9:

$$\|(A - AQQ^T)\Omega S\|_F^2 \leq 10\|A - AQQ^T\|_F^2.$$

**Concluding the proof.** Equipped with the above bounds, we are ready to conclude the proof of the theorem. We continue by manipulating the term  $\|A - CC^+A\|_F^2$  as follows  $\|A - CC^+A\|_F^2 \leq$

$$\begin{aligned} &\leq \|A - AQQ^T\|_F^2 + \|(A - AQQ^T)\Omega S(Q^T\Omega S)^+\|_F^2 \\ &\leq \|A - AQQ^T\|_F^2 + 2\|(A - AQQ^T)\Omega SS^T\Omega^T Q\|_F^2 + 2\|(A - AQQ^T)\Omega S((Q^T\Omega S)^+ - (Q^T\Omega S)^T)\|_F^2 \\ &\leq \|A - AQQ^T\|_F^2 + 2\|(A - AQQ^T)\Omega SS^T\Omega^T Q\|_F^2 + 2\|(A - AQQ^T)\Omega S\|_F^2 \|((Q^T\Omega S)^+ - (Q^T\Omega S)^T)\|_2^2 \\ &\leq \|A - AQQ^T\|_F^2 + 2\frac{20\hat{r}}{r}\|A - AQQ^T\|_F^2 + 2 \cdot 10\|A - AQQ^T\|_F^2 \frac{\epsilon}{26} \left(1 - \frac{\sqrt{\epsilon}}{\sqrt{26}}\right)^{-1} \\ &\leq \|A - AQQ^T\|_F^2 + \left(\frac{40}{104 \ln(40)} + \frac{20/26}{1 - 1/\sqrt{2 \cdot 26}}\right) \epsilon \|A - AQQ^T\|_F^2 \\ &\leq (1 + 0.998 \cdot \epsilon)\|A - AQQ^T\|_F^2 \leq (1 + 0.998 \cdot \epsilon)(1 + \epsilon)\|A - A_k\|_F^2 \leq (1 + \epsilon)^2\|A - A_k\|_F^2 \end{aligned}$$

Taking square roots on both sides concludes the proof. The failure probability follows by a union bound on all the probabilistic events involved in the proof of the theorem