Low Rank Approximation using Error Correcting Coding Matrices

Shashanka Ubaru UBARU001@UMN.EDU
Arya Mazumdar ARYA@UMN.EDU
Yousef Saad SAAD@CS.UMN.EDU
University of Minnesota-Twin Cities, MN USA

Abstract

Low-rank matrix approximation is an integral component of tools such as principal component analysis (PCA), as well as is an important instrument used in applications like web search, text mining and computer vision, e.g., face recognition. Recently, randomized algorithms were proposed to effectively construct low rank approximations of large matrices. In this paper, we show how matrices from error correcting codes can be used to find such low rank approximations.

The benefits of using these code matrices are the following: (i) They are easy to generate and they reduce randomness significantly. (ii) Code matrices have low coherence and have a better chance of preserving the geometry of an entire subspace of vectors; (iii) Unlike Fourier transforms or Hadamard matrices, which require sampling $O(k \log k)$ columns for a rank-k approximation, the log factor is not necessary in the case of code matrices. (iv) Under certain conditions, the approximation errors can be better and the singular values obtained can be more accurate, than those obtained using Gaussian random matrices and other structured random matrices.

1. Introduction

Many scientific computations, data analysis and machine learning applications (Halko et al., 2011; Drineas et al., 2006) lead to large dimensional matrices which can be well approximated by a low dimensional basis. It is more efficient to solve such computational problems by first transforming these large matrices into a low dimensional space, while preserving the invariant subspace that captures the essential information of the matrix. Several algorithms have been proposed in the literature for finding such low rank approximations of a matrix (Ye, 2005; Haeffele et al., 2014; Papailiopoulos et al., 2013). Recently, research focussed on developing techniques which use randomization for computing low rank approximations and matrix decompositions of such matrices. It is found that randomness provides an effective way to construct low dimensional bases with high reliability and computational efficiency.

The randomization techniques for matrix approximations (Halko et al., 2011; Martinsson et al., 2006; Liberty et al., 2007) aim to compute a basis that approximately spans the input matrix A, by sampling the matrix using Gaussian random matrices. This task is accomplished by first forming the matrix-matrix product $Y = A\Omega$, where Ω is a random matrix of smaller dimension, and then computing the orthonormal basis of $Y = QR$ that identifies the range of the reduced matrix Y. It can be shown that $A \approx QQ^*A$ with high probability. Recently, it has been observed that structured random matrices, like subsampled random Fourier transform (SRFT) and Hadamard transform (SRHT) matrices can also be used in place of Gaussian random matrices (Liberty, 2009; Woolfe et al., 2008; Tropp, 2011). This paper demonstrates how error correcting coding matrices can be a good choice for computing low rank approximations.

The input matrices whose low rank approximation is to be computed, usually have very large dimensions (e.g., in the order of $10^6 - 10^8$). In order to form a Gaussian random matrix which samples the input matrix in randomized algorithms, we need to generate a large number of random numbers. This could be a serious practical issue, (in terms of time complexity and storage). This issue can be addressed by using the structured random matrices, like SRFT and SRHT matrices. However, mixing of columns might not be as uniform, and there is potential loss in the accuracy. Other practical issues arise such as: the Fourier Transform matrices require handling complex numbers and the Hadamard matrices exist only for the sizes which are in powers of 2. These drawbacks can be overcome if the code matrices presented in this paper are used for sampling input matrices.

In digital communication, information is encoded by adding redundancy into (predominantly binary) vectors or

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

codewords, that are then transmitted over a noisy channel (Cover & Thomas, 2012). These codewords are required to be far apart in terms of some distance metric for noise-resilience. Coding schemes usually generate codewords that maintain a fixed minimum Hamming distance between each other, hence they are widespread and act like random numbers. We can define probability measures for matrices formed by stacking up these codewords, (see section 2.2 for details). The idea is to use subsampled versions of these code matrices as sampling matrices in the randomized algorithms for matrix approximations. Section 5.2 shows that subsampled code matrices have low coherence and have a better chance of preserving the geometry of an entire subspace of vectors. In some cases, it is possible to compute the matrix-matrix product faster with code matrices because of their structure. Importantly, contrary to SRFT/SRHT matrices, certain subsampled code matrices do not require the log factor, thus achieving the order optimal $O(k)$ in the number of samples with deterministic matrices, see sec. 5.4 for an explanation.

2. Preliminaries

First, we present some of the notation used and give a brief description of error correcting coding techniques that are used in communication systems and information theory.

2.1. Notation and Problem Formulation

Throughout the paper, $\|\cdot\|$ refers to the ℓ_2 norm. We use $\|\cdot\|_F$ for the Frobenius norm. The singular value decomposition (SVD) of a matrix A is denoted by $A = U\Sigma V^*$ and the singular values by $\sigma_j(A)$. We use e_j for the jth standard basis vector. Given a random subset T of indices in $\{1, \dots, 2^r\}$ with size n and $r \geq \lceil \log_2 n \rceil$, we define a restriction (sampling) operator $S_T : \mathbb{R}^{2^r} \rightarrow \mathbb{R}^T$ given by

$$(S_T \boldsymbol{x})(j) = x_j,\ j \in T.$$

A Rademacher random variable takes values ± 1 with equal probability. We write ε for a Rademacher variable.

In low rank approximation methods, we compute an orthonormal basis that approximately spans the range of an $m \times n$ input matrix A. That is, a matrix Q having orthonormal columns such that $A \approx QQ^\top A$. The basis matrix Q must contain as few columns as possible, but it needs to be an accurate approximation of the input matrix. I.e., we seek a matrix Q with k orthonormal columns such that,

$$\|A - QQ^\top A\|_\xi \leq \epsilon, \tag{1}$$

for a positive error tolerance ϵ and an integer $\xi \geq 2$. The best rank-k approximation of A with respect to both Frobenius and spectral norm is given by the Eckart-Young theorem (Eckart & Young, 1936), and it is $\hat{A}_k = U_k \Sigma_k V_k^\top$, where U_k and V_k are the k-dominant left and right singular vectors of A, respectively and diagonal Σ_k contains the top k singular values of A. So, the optimal Q in (1) will be U_k for $\xi \in \{2, F\}$.

2.2. Error Correcting Codes

In communication systems, data are transmitted from a source (transmitter) to a destination (receiver) through physical channels. These channels are usually noisy, causing errors in the data received. In order to facilitate the ability to detect and correct these errors in the receiver, error-correcting codes are used (MacWilliams & Sloane, 1977). A block of information (data) symbols are encoded in to a binary vector[1], also called a codeword, by the encoding error-correcting code. Error-correcting coding methods check the correctness of the codeword received. The set of codewords corresponding to a set of data-vectors (or symbols) that can possibly be transmitted, is called the *code*. Hence, a code $\mathcal{C}$ is a subset of $\mathbb{F}_2^\ell$, ℓ being an integer.

A code is said to be linear when adding two codewords of the code coordinate-wise using modulo-2 arithmetic results in a third codeword of the code. Usually a linear code $\mathcal{C}$ is represented by the tuple $[\ell, r]$, where ℓ represents the codeword length and $r = \log_2 |\mathcal{C}|$ is the number of information bits that can be encoded by the code. There are $\ell - r$ redundant bits in the codeword, which are sometimes called parity check bits, generated from messages using an appropriate rule. It is not necessary for a codeword to have the information bits as r of its coordinates, but the information must be uniquely recoverable from the codeword.

It is perhaps obvious that a linear code $\mathcal{C}$ is a linear subspace of dimension r in the vector space $\mathbb{F}_2^\ell$. The basis of $\mathcal{C}$ can be written as the rows of a matrix, which is known as the generator matrix of the code. The size of the generator matrix G is $r \times \ell$, and for any information vector $\boldsymbol{m} \in \mathbb{F}_2^r$, the corresponding codeword is found by the linear map:

$$\boldsymbol{c} = \boldsymbol{m}G.$$

Note that all the arithmetic operations above are over the binary field $\mathbb{F}_2$. To encode r bits, we must have 2^r unique codewords. Then, we may form a matrix of size $2^r \times \ell$ by stacking up all codewords that are formed by the generator matrix of a given linear coding scheme,

$$\underbrace{C}_{2^r \times \ell} = \underbrace{M}_{2^r \times r} \underbrace{G}_{r \times \ell}. \tag{2}$$

For a given tuple $[\ell, r]$, different error correcting coding schemes have different generator matrices and the resulting codes have different properties. For example, for any

[1] Here, and in the rest of the text, we are considering only binary codes. Codes over larger alphabets are also quite common.

two integers t and q, a BCH code (Bose & Ray-Chaudhuri, 1960) has length $\ell = 2^q - 1$ and dimension $r = 2^q - 1 - tq$. Any two codewords in this BCH code maintain a minimum (Hamming) distance of at least $2t + 1$ between them. The pairwise minimum distance is an important parameter of a code and is called just the minimum distance of the code. As a linear code C is a subspace of a vector space, the null-space $C^\perp$ of the code is another well-defined subspace. This is called the *dual* of the code. The dual of the $[2^q - 1, 2^q - 1 - tq]$-BCH code is a code with length $2^q - 1$, dimension tq and minimum distance at least $2^{q-1} - (t-1)2^{q/2}$. The minimum distance of the dual code is called the dual distance of the code.

Depending on the coding schemes used, the codeword matrix C will have a variety of favorable properties, e.g., low coherence which is useful in compressed sensing (Barg et al., 2015). Since the codewords need to be far apart, they show some properties of random vectors. We can define probability measures for codes generated from a given coding scheme. If $\mathcal{C} \subset \{0,1\}^\ell$ is an $\mathbb{F}_2$-linear code whose dual $\mathcal{C}^\perp$ has a minimum distance above k (dual distance $> k$), then the code matrix is an *orthogonal array* of strength k (Delsarte & Levenshtein, 1998). This means, in such a code $\mathcal{C}$, for any k entries of each codeword $\boldsymbol{c}$ say $\mathbf{c}' = \{c_{i_1}, c_{i_2}, \ldots, c_{i_k}\}$ and for any k bit binary string α, we have

$$\mathbf{Pr}[\mathbf{c}' = \alpha] = 2^{-k}.$$

This is called the k-wise independence property of codes. We will use this property of codes in our theoretical analysis (see section 5 for details).

TThe codeword matrix C has 2^r codewords each of length ℓ (a $2^r \times \ell$ matrix), i.e., a set of 2^r vectors in $\{0,1\}^\ell$. Given a codeword $\mathbf{c} \in \mathcal{C}$, let us map it to a vector $\phi \in \mathbb{R}^\ell$ by setting $1 \longrightarrow \frac{-1}{\sqrt{2^r}}$ and $0 \longrightarrow \frac{1}{\sqrt{2^r}}$. In this way, a binary code $\mathcal{C}$ gives rise to a code matrix $\Phi = (\phi_1, \ldots, \phi_{2^r})^\top$. Such a mapping is called binary phase-shift keying (BPSK) and appeared in the context of sparse recovery (e.g., p. 66 (Mazumdar, 2011)). For codes with dual distance ≥ 3, this code matrix Φ will have orthonormal columns. In section 5.2, we will show that these code matrices with certain mild properties can preserve the geometry of vector subspaces with high probability. Hence, in the randomized techniques for matrix approximations, we can use a subsampled and scaled version of this matrix Φ to sample a given input matrix and find the active subspaces of the matrix.

3. Construction of Subsampled Code Matrix

For an input matrix A of size $m \times n$, and a target rank k, we choose $r \geq \lceil \log_2 n \rceil$ as the dimension of the code (length of the message vector) and $\ell > k$, as the length of the code. The value of ℓ will depend on the coding scheme used, particularly on the dual distance of of the code, (details in section 5.2). We consider an $[\ell, r]$-linear coding scheme and form the sampling matrix as follows: We draw the sampling test matrix say Ω as

$$\Omega = \sqrt{\frac{2^r}{\ell}} DS\Phi, \qquad (3)$$

where

- D is a random $n \times n$ diagonal matrix whose entries are independent random signs, i.e., random variables uniformly distributed on $\{\pm 1\}$.
- S is the uniformly random downsampler, an $n \times 2^r$matrix whose n rows are randomly selected from a $2^r \times 2^r$ identity matrix.
- Φ is the $2^r \times \ell$ code matrix, generated using an $[\ell, r]$-linear coding scheme, with BPSK mapping and scaled by $2^{-r/2}$ such that all columns have unit norm.

Intuition The design of a subsampled code matrix is similar to the design of SRFT and SRHT matrices. The intuition for using such a design is well established in (Tropp, 2011; Halko et al., 2011). The matrix Φ has entries with magnitude $\pm 2^{-r/2}$ and has orthonormal columns when a coding scheme with dual distance of the codes is ≥ 3 is used. The scaling $\sqrt{\frac{2^r}{\ell}}$ is used to make the energy of the sampling matrix equal to unity, i.e., to make the rows of Ω unit vectors. The purpose of multiplying by D is to flatten out input vectors. We refer to (Tropp, 2011) for further details. For a fixed unit vector $\boldsymbol{x}$, the first component of $\boldsymbol{x}^* DS\Phi$ is given by $(\boldsymbol{x}^\top DS\Phi)_1 = \sum_{j=1}^n x_j \varepsilon_j \phi_{j'1}$, where ϕ_{ij} are components of the code matrix Φ, the index j' depends on the downsampler S and ε_j is the Rademacher variable from D. This sum clearly has zero mean and since entries of Φ have magnitude $2^{-r/2}$, the variance of the sum is 2^{-r}. The Hoeffding inequality (Hoeffding, 1963) shows that

$$\mathbb{P}\{|(\boldsymbol{x}^* DS\Phi)_1| \geq \tilde{t}\} \leq 2e^{-2^r \tilde{t}^2/2}.$$

That is, the magnitude of the first component of $\boldsymbol{x}^* DS\Phi$ is about $2^{-r/2}$. Similarly, the argument holds for the remaining entries. Therefore, it is unlikely that any one of the ℓ components of $\boldsymbol{x}^* DS\Phi$ is larger than $\sqrt{2\log(2\ell)/2^r}$, (the failure probability is ℓ^{-1}).

4. Algorithm

We use the same prototype algorithm as discussed in (Halko et al., 2011) for the low rank approximation and decomposition of input matrix A. The subsampled code matrices given in (3), generated from a chosen coding scheme is used as the sampling test matrix. The algorithm is as follows:

Algorithm 1 Prototype Algorithm

```
**Input:** An $m \times n$ matrix $A$, a target rank $k$.
**Output:** Rank-$k$ factors $U, \Sigma$, and $V$ in an approximate SVD $A \approx U\Sigma V^*$.
**1.** Form an $n \times \ell$ subsampled code matrix $\Omega$, as described in Section 3 and (3), using an $[\ell, r]-$linear coding scheme, where $\ell > k$ and $r \geq \lceil \log_2 n \rceil$.
**2.** Form the $m \times \ell$ sample matrix $Y = A\Omega$.
**3.** Form an $m \times \ell$ orthonormal matrix $Q$ such that $Y = QR$.
**4.** Form the $\ell \times n$ matrix $B = Q^*A$.
**5.** Compute the SVD of the small matrix $B = \hat{U}\Sigma V^*$.
**6.** Form the matrix $U = Q\hat{U}$.
```

5. Analysis

This section discusses the performance analysis of the subsampled code matrices as sampling matrices in algorithm 1. First, we give the deterministic error bound for the algorithm for a given sampling matrix Ω. Then, we show how code matrices preserve the geometry of an entire subspace of vectors by establishing connection to Johnson Lindenstrauss Transforms (JLT) and sign matrices, via the k-wise independence property of codes. Finally, we give the bounds for the approximation error and the singular values obtained from the algorithm.

Setup Let A be an $m \times n$ input matrix with a singular value decomposition given by $A = U\Sigma V^*$, and partition its SVD as follows

$$A = U \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \end{array} \begin{array}{cc} n & \\ \begin{bmatrix} V_1^* \\ V_2^* \end{bmatrix} & \begin{array}{c} k \\ n-k \end{array} \end{array} . \qquad (4)$$

Let Ω be the $n \times \ell$ test (sampling) matrix, where ℓ is the number of samples. Consider the matrices

$$\Omega_1 = V_1^*\Omega \quad \text{and} \quad \Omega_2 = V_2^*\Omega. \qquad (5)$$

The objective of any low rank approximation algorithm is to try and approximate the subspace which spans the top k left singular vectors of A. The test matrix Ω is said to preserve the geometry of an entire subspace of vectors, if for any orthonormal matrix V, a matrix of the form $V^*\Omega$ is well conditioned (Halko et al., 2011).

5.1. Deterministic Error bounds

Algorithm 1 constructs an orthonormal basis Q for the range of Y, and the goal is to quantify how well this basis captures the action of the input matrix A. Let $QQ^* = P_Y$ where P_Y is the unique orthogonal projector with range(P_Y)=range(Y). If Y is full rank, we can express the projector as : $P_Y = Y(Y^*Y)^{-1}Y^*$. We seek to find an upper bound for the approximation error given by, for $\xi \in \{2, F\}$

$$\|A - QQ^*A\|_\xi = \|(I - P_Y)A\|_\xi.$$

The deterministic upper bound for the approximation error for Algorithm 1 is given in (Halko et al., 2011). We restate theorem 9.1 in (Halko et al., 2011) below:

Theorem 1 (Deterministic error bound) *Let A be $m \times n$ matrix with singular value decomposition given by $A = U\Sigma V^*$, and fix $k \geq 0$. Choose a test matrix Ω and construct the sample matrix $Y = A\Omega$. Partition Σ as in (4), and define Ω_1 and Ω_2 via (5). Assuming that Ω_1 is full row rank, the approximation error satisfies for $\xi \in \{2, F\}$*

$$\|(I - P_Y)A\|_\xi^2 \leq \|\Sigma_2\|_\xi^2 + \|\Sigma_2\Omega_2\Omega_1^\dagger\|_\xi^2. \qquad (6)$$

An elaborate proof for the above theorem can be found in (Halko et al., 2011). Using the submultiplicative property of the spectral and Frobenius norms, and the Eckart-Young theorem, equation (6) can be simplified to

$$\|A - QQ^*A\|_\xi \leq \|A - \hat{A}_k\|_\xi \sqrt{1 + \|\Omega_2\|^2\|\Omega_1^\dagger\|^2}. \qquad (7)$$

Recently Ming Gu (Gu, 2015), developed deterministic lower bounds for the singular values obtained from randomization algorithms, particularly for the power method (Halko et al., 2011), which is one of the alternatives of randomized algorithms. Given below is the modified version of Theorem 4.3 in (Gu, 2015) for algorithm 1.

Theorem 2 (Deterministic singular value bounds) *Let $A = U\Sigma V^*$ be the SVD of A, for a fix k, and let $V^*\Omega$ be partitioned as in (5). Assuming that Ω_1 is full row rank, then Algorithm 1 must satisfy for $j = 1, \ldots, k$:*

$$\sigma_j \geq \sigma_j(A_k) \geq \frac{\sigma_j}{\sqrt{1 + \|\Omega_2\|^2\|\Omega_1^\dagger\|^2 \left(\frac{\sigma_{k+1}}{\sigma_j}\right)^2}} \qquad (8)$$

where σ_j are the jth singular value of A and A_k is the rank-k approximation obtained by our algorithm.

The proof for the above theorem can be seen in (Gu, 2015). For a given sampling matrix Ω, the major challenge is to show that Ω_1 is indeed full rank. That is, we need to show that for any orthonormal matrix V, with high probability, $V^*\Omega$ is well conditioned.

5.2. Subsampled Code Matrices Preserve Geometry

Recall from section 3 the construction of the 'tall and thin' $n \times \ell$ subsampled error correcting code matrices Ω. One of the critical facts to show is that these matrices approximately preserve the geometry of an entire subspace of vectors. This will imply that Ω_1 will be full rank and we can use the deterministic bounds for analysis. To prove this, we establish connections between the properties of code matrices and two important results existing in the literature. The first connection is to the well known the Johnson-Lindenstrauss Transform (JLT) (Johnson & Lindenstrauss, 1984) and the second is with the random sign matrices. Both these connections depend on the k-wise independence property of the code matrices.

5.2.1. Connection to Johnson-Lindenstrauss Transform

One of the primary results developed in the randomized matrix algorithms literature was establishing the relation between the Johnson-Lindenstrauss Transform (JLT) and preserving the geometry of subspaces (Sarlos, 2006). We first give the definition of JLT and then state this important result. We will then show that code matrices under certain mild conditions satisfy JLT.

Definition 1 *A matrix $\Omega \in \mathbb{R}^{n\times\ell}$ is Johnson-Lindenstrauss Transform with parameters ϵ, δ or JLT(ϵ, δ) for any $0 < \epsilon, \delta < 1$, if for any vector $v \in \mathbb{R}^n$, it holds*

$$(1-\epsilon)\|v\|^2 \leq \|v^*\Omega\|^2 \leq (1+\epsilon)\|v\|^2$$

with probability $1-\delta$, under certain conditions on ℓ, which will depend on ϵ, δ and the reduced dimension desired.

So, if the sampling matrix Ω is JLT, it preserves the distance of any vector v whose dimensionality reduction we seek. Sarlos (Sarlos, 2006) gave the important relation between JLT and random matrix sampling (also known as subspace embedding). The following lemma, which is corollary 11 in (Sarlos, 2006) gives this relation.

Lemma 3 *Let $0 < \epsilon, \delta < 1$ and f be some function. If Ω is a JLT from $\mathbb{R}^n$ to $O(k\log(k/\epsilon)/\epsilon^2.f(\delta))$, then for an orthonormal matrix $V \in \mathbb{R}^{n\times k}, n \geq k$ we have*

$$\mathbf{Pr}(\forall \in [1..k] : |1-\sigma_i(V^*\Omega)| \leq \epsilon) \geq 1-\delta$$

The above lemma shows that, if the sampling matrix Ω is JLT and $\ell = O(k\log(k/\epsilon))$, (choosing $f(\delta)$ close to ϵ^2,) then the singular values of $V^*\Omega$ are bounded, i.e., $V^*\Omega$ is well conditioned with high probability. So, if our subsampled code matrix is a JLT then, it will preserve the geometry of V with high probability.

Next, we give two results that show that code matrices with certain mild properties satisfy JLT property. The first result is by Ailon and Liberty (Ailon & Liberty, 2009), where they show a matrix Ω which is 4-wise independent will satisfy JLT. Interestingly, they give 2 error correcting dual BCH codes as examples and show how fast multiplication can be achieved with code matrices. A small drawback here is that the maximum entries of A need to be restricted.

The second result is by Clarkson and Woodruff (Clarkson & Woodruff, 2009) (see Theorem 2.2), where they show if Ω is a $4\lceil\log(\sqrt(2)/\delta)\rceil$-wise independent matrix, then Ω will satisfy JLT property. We know that a code matrix with dual distance $> k$ is k-wise independent. Thus, any error correcting code matrix with a dual distance > 4 (more than 2 error correcting ability) will preserve the geometry of and entire subspace of vectors (i.e., Ω_1 is full rank) with high probability.

5.2.2. Code matrices as random sign matrices

Any code matrix with a dual distance > 4 will preserve the geometry of V. However, we need the number of samples to be $\ell = O(k\log(k/\epsilon))$, which is similar to a subsampled Fourier or Hadamard matrix. Next, we show that $O(k)$ can be achieved in the number of samples required for code matrices, if the codes satisfy certain conditions.

We know that code matrices act as random matrices as the distance of the code increases. We can treat code matrices as random sign matrices having certain probabilistic distributions. Indeed a code with dual distance above k supports k-wise independent probability measure. This property of code matrices helps us to use the following lemma given in (Clarkson & Woodruff, 2009) (Lemma 3.4) which states,

Lemma 4 *Given an integer k and $\epsilon, \delta > 0$. If Ω is $\rho(k + \log(1/\delta)$-wise independent with an absolute constant $\rho > 1$, then for an orthonormal matrix $V \in \mathbb{R}^{n\times k}$ and $\ell = O(k\log(1/\delta)/\epsilon)$, with probability at least $1-\delta$ we have*

$$\|V^*\Omega\Omega^*V - I\| \leq \epsilon.$$

Thus, a sampling matrix Ω which is $\lceil k + \log(1/\delta)\rceil$-wise independent preserves the geometry of V with number of samples (length) $\ell = O(k/\epsilon)$. Hence, a code matrix with dual distance $> \lceil k + \log(1/\delta)\rceil$ will preserve the geometry of V with $\ell = O(k)$.

Therefore, any code matrix with dual distance > 4 will preserve the geometry of V with $\ell = O(k\log(k/\epsilon))$ and if the dual distance is $> k$, then the code matrix can preserve the geometry of V with $\ell = O(k/\epsilon)$.

5.3. Error Bounds

The following theorem gives the approximation error bounds when the subsampled code matrix is used as test matrix Ω in Theorem 1. The upper and lower bounds for the singular values obtained are also given.

Theorem 5 (Error bounds for code matrix) *Let A be $m \times n$ matrix with singular values $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots$. Generate a subsampled code matrix Ω from a desired coding scheme as in (3) with $r \geq \lceil \log_2(n) \rceil$ as the length of the message vector. For any code matrix Ω with* **dual distance** > 4 **and length** $\ell = O(k \log(k/\epsilon)/\epsilon^2.f(\delta))$*, the approximation error for algorithm 1 satisfies, for $\xi \in \{2, F\}$*

$$\|A - QQ^*A\|_\xi \leq \|A - A_k\|_\xi \sqrt{1 + \frac{(1+\eta)n}{(1-\epsilon)^2\ell}} \tag{9}$$

for a small constant $\eta > 0$ with failure probability δ and for any code matrix Ω with **dual distance** $\geq (k + \log(1/\delta))$ **and length** $\ell = O(k \log(1/\delta)/\epsilon)$*, the approximation error satisfies*

$$\|A - QQ^*A\|_F \leq \|A - A_k\|_F(1 + \epsilon) \tag{10}$$

with failure probability δ. The bounds for the singular values obtained are:

$$\sigma_j \geq \sigma_j(A_k) \geq \frac{\sigma_j}{\sqrt{1 + \left(\frac{(1+\eta)n}{(1-\epsilon)^2\ell}\right)\left(\frac{\sigma_{k+1}}{\sigma_j}\right)^2}} \tag{11}$$

The proof of the theorem follows from the deterministic bounds given earlier. Equation (9) and (13) are derived with the help of two lemmas (given in appendix) which show subsampling a code matrix with dual distance ≥ 3 is well conditioned. Equation (10) is straight forward from Teorem 4.2 in (Clarkson & Woodruff, 2009). The detailed proof of the theorem in given in the appendix.

Differences in the construction An important difference between the construction of subsampled code matrices given in (3) and the construction of SRHT or SRFT given in (Halko et al., 2011; Tropp, 2011) is in the way these matrices are subsampled. In the case of SRHT, a Hadamard matrix of size $n \times n$ is applied to input matrix A and ℓ out of n columns are sampled at random (n must be a power of 2). In contrast, in the case of subsampled code matrices, a $2^r \times \ell$ code matrix generated from an $[\ell, r]$-linear coding scheme is considered, and n out of 2^r codewords are chosen. The subsampling will not affect the k-wise independent property of the code matrix (or the distinctness of rows) when uniformly subsampled. This need not be true in the case of SRHT. The importance of the distinctness of rows is discussed next.

5.4. Logarithmic factor

A crucial advantage of the code matrices is that they have very low *coherence*. Coherence is defined as the maximum inner product between any two rows. This is in particular true when the minimum distance of the code is close to half the length. If the minimum distance of the code is d then the code matrix generated from an $[\ell, r]$-code has coherence equal to $\frac{\ell - 2d}{2^r}$. For example, if we consider dual BCH code (see sec. 2.2) the coherence is $\frac{2(t-1)\sqrt{\ell+1}-1}{2^r}$. Low coherence ensures near orthogonality of rows. This is a desirable property in many applications such as compressed sensing and sparse recovery.

For a rank-k approximation using subsampled Fourier or Hadamard matrices, we need to sample $O(k \log k)$ columns. This logarithmic factor emerges as a necessary condition in the theoretical proof (given in (Tropp, 2011)) that shows that these matrices approximately preserve the geometry of an entire subspace of input vectors. The log factor is also necessary to tackle the worst case input matrices. The discussions in sec. 11 of (Halko et al., 2011) and sec. 3.3 of (Tropp, 2011) give more details. In the case of subsampled code matrices, the log factor does not seem necessary to tackle the worst case input matrices. To see why this is true, let us consider the worst case example for orthonormal matrix V described in Remark 11.2 of (Halko et al., 2011).

An infinite family of worst case examples of the matrix V is as follows. For a fixed integer k, let $n = k^2$. Form an $n \times k$ orthonormal matrix V by regular decimation of the $n \times n$ identity matrix. That is, V is a matrix whose jth row has a unit entry in column $(j-1)/k$ when $j \equiv 1 \pmod{k}$ and is zero otherwise. This type of matrix is troublesome when DFT or Hadamard matrices are used for sampling.

Suppose that we apply $\Omega = DFR^*$ to the matrix V^*, where D is same as in (3), F is an $n \times n$ DFT or Hadamard matrix and R is $\ell \times n$ matrix that samples ℓ coordinates from n uniformly at random. We obtain a matrix $X = V^*\Omega = WR^*$, which consists of ℓ random columns sampled from $W = V^*DF$. Up to scaling and modulation of columns, W consists of k copies of a $k \times k$ DFT or Hadamard matrix concatenated horizontally. To ensure that X is well conditioned (preserve geometry), we need $\sigma_k(X) > 0$. That is, we must pick at least one copy of each of the k distinct columns of W. This is the coupon collector's problem (Motwani & Raghavan, 1995) in disguise and to obtain a complete set of k columns with non-negligible probability, we must draw at least $k \log(k)$ columns.

In the case of code matrices, we apply a subsampled code matrix $\Omega = DS\Phi$ to the matrix V^*. We obtain $X = V^*\Omega = V^*DS\Phi$, which consists of k randomly selected rows of the code matrix Φ. That is, X consists of k distinct codewords of length ℓ. The code matrix has low coherence and all rows are distinct. If we use a code matrix with dual distance $> k$, then X contains k rows which are k-wise independent (near orthonormal) and $\sigma_k(X) > 0$; as a result the geometry of V is preserved and the log factor is not necessary. Thus, for the worst case scenarios we have an $O(\log k)$ factor improvement over other structured matrices. More importantly, this shows that the order optimal can be achieved with the immediate lower bound of $O(k)$ in the number of samples required with deterministic matrices.

5.5. Choice of error-correcting code

The requirement of k-wise independence of codewords translates to the dual distance of the code being greater than k. Since a smaller code (less number of codewords, i.e., smaller r) leads to less randomness in sampling, we would like to use the smallest code with dual distance $\geq k$.

One of the choices of the code can be the family of dual BCH codes. As mentioned earlier, this family has length ℓ, dimension $t\log(\ell+1)$ and dual distance at least $2t+1$. Hence, to guarantee dual distance at least k, the size of the code must be $2^{\frac{k\log(\ell+1)}{2}} = (\ell+1)^{k/2}$. We can choose n vectors of length $\frac{k\log(\ell+1)}{2}$ and form the codewords by simply multiplying these with the generator matrix (over $\mathbb{F}_2$) to form the subsampled code matrix. Therefore, forming these code matrices will be much faster than generating $n \times \ell$ i.i.d Gaussian random matrices or random sign matrices which have k-wise independent rows.

If the log factor is not an issue (for smaller k), then we can choose any code matrix with dual distance > 4 and $r = \lceil \log_2 n \rceil$. These code matrices are almost deterministic and unlike SRFT/SRHT, subsampling of columns is not required. In fact, Hadamard matrices are also a class of linear codes. In practice, code matrices generated by any linear coding scheme can be used in place of Gaussian random matrices. As there are many available classes of algebraic and combinatorial codes, we have a large pool of candidate matrices. In this paper we chose dual BCH codes for our numerical experiments as they particularly have low coherence, and turn out to perform quite well in practice.

6. Numerical Experiments

The following experiments will illustrate the performance of subsampled code matrices as sampling matrices in Algorithm 1. Our first experiment is with a 4770×4770 matrix named Kohonen from the Pajek network (a directed graph's matrix representation), available from the UFL Sparse Matrix Collection (Davis & Hu, 2011). Such graph Laplacian matrices are commonly encountered in machine learning and image processing applications. The performance of the dual BCH code matrix, Gaussian matrix, subsampled Fourier transform (SRFT) and Hadamard (SRHT) matrices are compared as sampling matrices Ω in Algorithm 1. For SRHT, we had to subsample the rows as well (similar to code matrices), since the input size is not a power of 2. All experiments were implemented in matlab v8.1.

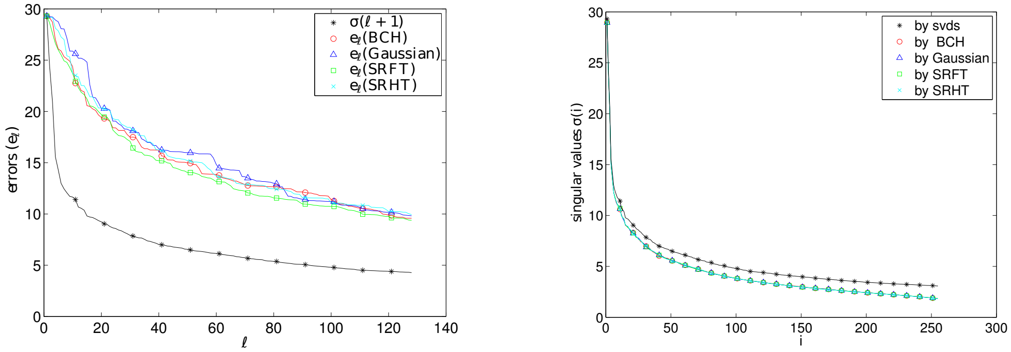
Table 1. Comparison of errors

MATRIX	DUAL BCH	GAUSSIAN	SRFT	$\sigma_{\ell+1}$
LPICERIA3D $\ell = 63$	15.4865	18.3882	16.3619	6.4625
DETER3 $\ell = 127$	9.2602	9.2658	9.2984	5.7499
S80PI $\ell = 63$	3.8148	3.8492	3.7975	1.9996
DELAUNAY $\ell = 63$	6.3864	6.3988	6.3829	5.8469
EPA $\ell = 255$	5.5518	5.5872	5.4096	2.5655
EPA $\ell = 511$	3.2171	3.2003	3.1752	1.3697
KOHONEN $\ell = 511$	4.2977	4.2934	4.2610	2.0239
KOHONEN ℓ =1023	2.4581	2.4199	2.4718	1.0236

Figure 1(A) gives the actual error $e_\ell = \|A - Q^{(\ell)}(Q^{(\ell)})^\top A\|$ for each ℓ number of samples when a subsampled dual BCH code matrix, a Gaussian matrix, SRFT and SRHT matrices are used as sampling matrices in algorithm 1, respectively. The best rank-ℓ approximation error $\sigma_{\ell+1}$ is also given. Figure 1(B) plots the singular values obtained from algorithm 1, for $\ell = 255$ and different sampling matrices Ω used. The top 255 exact singular values of the matrix (available in the UFL database) are also plotted. We observe that, in practice, the performance of all four sampling matrices are similar.

Table 1 compares the errors e_ℓ for ℓ number of samples, obtained for a variety of input matrices from different applications when subsampled dual BCH code, Gaussian and SRFT matrices were used. It also provides the theoretical minimum $\sigma_{\ell+1}$ value for each input matrices. All matrices were obtained from the UFL database. Matrices lpi_ceria3d (4400×3576) and deter3 (21777×7647) are from linear programming problems. S80PI_n1 (4028×4028) is from an eigenvalue/model reduction problem. Delaunay (4096×4096), EPA (4772×4772) and Kohonen are graph Laplacian matrices. We see in the first four examples, for small ℓ, the error performance of the code matrices is better than that of the Gaussian matrices. For higher ℓ, the error remains similar to the error for Gaussian matrices. Therefore, in practice, we can use code matrices in place of other sampling matrices due to their advantages.

Eigenfaces: Eigenfaces is a popular method for face recognition that is based on Principal Component Analysis (PCA) (Turk & Pentland, 1991; Sirovich & Meytlis, 2009). In this experiment (chosen as a verifiable comparison with results in (Gu, 2015)), we demonstrate the performance of randomized algorithm with different sampling matrices on face recognition. The face dataset is obtained from the AT&T Labs Cambridge database of faces (Samaria & Harter, 1994). There are ten different images of each of 40



(A) (B)

Figure 1. (A) The theoretical minimum $\sigma_{\ell+1}$ and approximate error as a function of the number of random samples ℓ using dual BCH code, Gaussian, SRFT and SRHT matrices as sampling matrix in algorithm 1 for input matrix Kohonen. (B) Estimates for top 255 singular values computed by algorithm 1 and the exact singular values by svds function.

Table 2. Comparison of the Number of Incorrect Matches

Rank	Dual BCH		Gaussian		SRFT		T-SVD
	p		p		p		
k	10	20	10	20	10	20	
10	18	13	19	15	21	18	26
20	14	11	14	12	16	12	13
30	10	08	13	08	12	09	10
40	09	08	08	07	08	10	06

distinct subjects. The size of each image is 92×112 pixels, with 256 gray levels per pixel. 200 of these faces, 5 from each individual are used as training images and the remaining 200 as test images to classify.

In the first step, we compute the principal components (dimensionality reduction) of mean shifted training image dataset using Algorithm 1, with different sampling matrix Ω and different p values. Next, we project the mean-shifted images into the singular vector space using the singular vectors obtained from the first step. The projections are called feature vectors and are used to train the classifier. To classify a new face, we mean-shift the image and project it onto the singular vector space obtained in the first step, obtaining a new feature vector. The new feature vector is classified using a classifier which is trained on the feature vectors from the training images. We used the in-built MATLAB function **classify** for feature training and classification. We compare the performance of the dual BCH code matrix, Gaussian matrix and SRFT matrix against exact truncated SVD (T-SVD). The results are summarized in Table 2. For $p = 10$ dual BCH code matrices give results that are similar to those of truncated SVD, and for rank $k < 40, p = 20$ our results are superior.

7. Conclusion

This paper advocated the use of matrices generated by error correcting codes as an alternative to random Gaussian or subsampled Fourier/Hadamard matrices for computing low rank matrix approximations. Among the attractive properties of the proposed approach are the numerous choices of parameters available, the orthogonality of columns and the near-orthogonality of rows. We showed that any code matrix with dual distance > 4 preserves the geometry of an entire subspace of vectors. Indeed if the dual distance of the code matrix is $> k$, then the length of the code (sampling complexity) required is in $O(k)$, thus leading to an order optimal in the worst-case guaranteed sampling complexity, an improvement by a factor of $O(\log k)$ over other known structured matrices. This is significant when the expected rank k is large and/or when the input matrix is sparse.

It is known that Gaussian matrices perform much better in practice compared to their theoretical analysis (Halko et al., 2011). Our code matrices (a) are almost deterministic, and (b) have ± 1 entries. Still, they perform equally well (as illustrated by experiments) compared to random real Gaussian matrices and complex Fourier matrices. Because of the availability of different families of classical codes in the rich literature of coding theory, many possible choices of code matrices are at hand. One of the contributions of this paper is to open up these options for use as structured sampling operators in low-rank approximations. Decoding of many, if not most, structured codes can be performed by the Fast Fourier Transform (Blahut, 1979). Hence, we can compute matrix-matrix products with code matrices substantially faster due to the availability of these fast transform techniques using the method described in (Ailon & Liberty, 2009). Interesting future work includes extending the framework of code matrices to other random sampling applications.

Acknowledgements: This work was supported by NSF under grant NSF/CCF-1318597 (S. Ubaru and Y. Saad) and NSF/CCF-1318093 (A. Mazumdar).

Bibliography

Ailon, Nir and Liberty, Edo. Fast dimension reduction using rademacher series on dual bch codes. *Discrete & Computational Geometry*, 42(4):615–630, 2009.

Barg, Alexander, Mazumdar, Arya, and Wang, Rongrong. Restricted isometry property of random subdictionaries. *IEEE Transactions on Information Theory*, 61(8), 2015.

Blahut, Richard E. Transform techniques for error control codes. *IBM Journal of Research and development*, 23(3):299–315, 1979.

Bose, Raj Chandra and Ray-Chaudhuri, Dwijendra K. On a class of error correcting binary group codes. *Information and control*, 3(1):68–79, 1960.

Clarkson, Kenneth L and Woodruff, David P. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 205–214. ACM, 2009.

Cover, Thomas M and Thomas, Joy A. *Elements of information theory*. John Wiley & Sons, 2012.

Davis, Timothy A and Hu, Yifan. The University of Florida sparse matrix collection. *ACM Transactions on Mathematical Software (TOMS)*, 38(1):1, 2011.

Delsarte, Philippe and Levenshtein, Vladimir I. Association schemes and coding theory. *Information Theory, IEEE Transactions on*, 44(6):2477–2504, 1998.

Drineas, Petros, Kannan, Ravi, and Mahoney, Michael W. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36 (1):158–183, 2006.

Eckart, Carl and Young, Gale. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Gu, M. Subspace iteration randomization and singular value problems. *SIAM Journal on Scientific Computing*, 37(3):A1139–A1173, 2015. doi: 10.1137/130938700. URL `http://dx.doi.org/10.1137/130938700`.

Haeffele, Benjamin, Young, Eric, and Vidal, Rene. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 2007–2015, 2014.

Halko, N., Martinsson, P., and Tropp, J. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806.

Hoeffding, Wassily. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Johnson, William B and Lindenstrauss, Joram. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Liberty, Edo. *Accelerated Dense Random Projections*. PhD thesis, Yale University, 2009.

Liberty, Edo, Woolfe, Franco, Martinsson, Per-Gunnar, Rokhlin, Vladimir, and Tygert, Mark. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.

MacWilliams, Florence Jessie and Sloane, Neil James Alexander. *The theory of error-correcting codes*, volume 16. Elsevier, 1977.

Martinsson, Per-Gunnar, Rockhlin, Vladimir, and Tygert, Mark. A randomized algorithm for the approximation of matrices. Technical report, DTIC Document, 2006.

Mazumdar, Arya. *Combinatorial methods in coding theory*. PhD thesis, University of Maryland, 2011.

Motwani, R. and Raghavan, P. *Randomized Algorithms*. Cambridge International Series on Parallel Computation. Cambridge University Press, 1995. ISBN 9780521474658.

Papailiopoulos, Dimitris, Dimakis, Alexandros, and Korokythakis, Stavros. Sparse PCA through Low-rank Approximations. In *Proceedings of The 30th International Conference on Machine Learning*, pp. 747–755, 2013.

Samaria, Ferdinando S and Harter, Andy C. Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pp. 138–142. IEEE, 1994.

Sarlos, Tamas. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pp. 143–152. IEEE, 2006.

Sirovich, Lawrence and Meytlis, Marsha. Symmetry, probability, and recognition in face space. *Proceedings of the National Academy of Sciences*, 106(17):6895–6899, 2009.

Tropp, Joel A. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3 (01n02):115–126, 2011.

Turk, Matthew and Pentland, Alex. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

Woolfe, Franco, Liberty, Edo, Rokhlin, Vladimir, and Tygert, Mark. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25 (3):335–366, 2008.

Ye, Jieping. Generalized Low Rank Approximations of Matrices. *Machine Learning*, 61(1-3):167–191, 2005. ISSN 0885-6125. doi: 10.1007/s10994-005-3561-6.

A. Proof of Theorem 5

The proof for the theorem is as follows. For the approximate error bounds given in (9), we start from equation (7) in Theorem 1. The terms that depend on the choice of test matrix Ω are $\|\Omega_2\|^2$ and $\|\Omega_1^\dagger\|^2$. We saw that the code matrix Ω preserves the geometry of the entire subspace of vectors and this also ensures that the spectral norm of $\Omega_1^\dagger$ is under control. From Lemma 3 and Lemma 3.6 in (Liberty et al., 2007), we have

$$\|\Omega_1^\dagger\|^2 = \frac{1}{\sigma_k^2(\Omega_1)} \leq \frac{1}{(1-\epsilon)^2}.$$

We bound the spectral norm of Ω_2 as follows $\|\Omega_2\|^2 = \|V_2^*\Omega\|^2 \leq \|V_2\|^2\|\Omega\|^2 = \|\Omega\|^2 = \sigma_1^2(\Omega)$, since V_2 is an orthonormal matrix. The following two lemmas give the upper bound for the singular values of Ω. The first lemma shows that if a code has dual distance ≥ 3, the resulting code matrix Φ has orthonormal columns.

Lemma 6 (Code matrix with orthonormal columns) *A code matrix Φ, generated by a coding scheme which results in codes that have dual distance between the codewords ≥ 3, has orthonormal columns.*

Proof. If a code has dual distance 3, then the corresponding code matrix (stacked up codewords as rows) is an orthogonal array of strength 2 (Delsarte & Levenshtein, 1998). This means all the tuples of bits, i.e., $\{0,0\}, \{0,1\}, \{1,0\}, \{1,1\}$, appear with equal frequencies in any two columns of the codeword matrix C. As a result the Hamming distance between any two columns of C is exactly 2^{r-1} (half the length of the column). This means after the BPSK mapping, the inner product between any two codewords will be zero. It is easy to see that the columns are unit norm as well.

This fact helps us use Lemma 3.4 from (Tropp, 2011) which shows that randomly sampling the rows of such a code matrix results in a well-conditioned matrix and gives bounds for the singular values.

Lemma 7 (Row sampling) *Let Φ be an $2^r \times \ell$ code matrix (with orthonormal columns), and let $M = 2^r . \max_{j=1,\dots,2^r} \|e_j^*\Phi\|^2$. For a positive parameter α, select the sample size*

$$n \geq \alpha M \log(\ell).$$

Draw a random subset T from $\{1,\dots,2^r\}$ by sampling n coordinates without replacement. Then,

$$\sqrt{\frac{(1-\nu)n}{2^r}} \leq \sigma_\ell(S_T\Phi) \text{ and } \sigma_1(S_T\Phi) \leq \sqrt{\frac{(1+\eta)n}{2^r}} \tag{12}$$

with failure probability at most

$$\ell.\left[\frac{e^{-\nu}}{(1-\nu)^{(1-\nu)}}\right]^{\alpha\log(\ell)} + \ell.\left[\frac{e^{\eta}}{(1+\eta)^{(1+\eta)}}\right]^{\alpha\log(\ell)}$$

where $\nu \in [0,1)$ and $\eta > 0$.

Since n is fixed and $M = \ell$ for a code matrix (all the entries of the matrix are $\pm 2^{-r/2}$), we get the condition $n \geq \alpha\ell\log(\ell)$. The parameters α, ν and η are chosen based on the inputs ℓ and n and the failure probability accepted. The bounds on the singular values of the above lemma are proved in (Tropp, 2011) using Matrix Chernoff Bounds. Since we use the scaling $\sqrt{\frac{2^r}{\ell}}$, the bounds on the singular values of the subsampled code matrix Ω will be

$$\sqrt{\frac{(1-\nu)n}{\ell}} \leq \sigma_\ell(\Omega) \text{ and } \sigma_1(\Omega) \leq \sqrt{\frac{(1+\eta)n}{\ell}}. \tag{13}$$

We substitute the above values for $\|\Omega_2\|^2$ and $\|\Omega_1^\dagger\|^2$ in (7) to get the error bounds in (9) and substitute these values in (8) of theorem 2 to get the bounds on singular values (11).

Clarkson and Woodruff (Clarkson & Woodruff, 2009) also give the Frobenius norm error bound for low rank approximation using k-wise independent sampling matrices. The error bound in (10) is straight from the following lemma which is a modification of theorem 4.2 in (Clarkson & Woodruff, 2009).

Lemma 8 *If $\Omega \in \mathbb{R}^{n\times\ell}$ is a $\rho(k+\log(1/\delta)$-wise independent sampling matrix, then for $\ell = O(k\log(1/\delta)/\epsilon)$ with probability at least $1-\delta$, we have*

$$\|A - QQ^*A\|_F \leq \|A - A_k\|_F(1+\epsilon) \tag{14}$$

Proof of this lemma is clear from the proof of theorem 4.2 in (Clarkson & Woodruff, 2009). This completes the proof of theorem 5.