

Improving CUR Matrix Decomposition and the Nyström Approximation via Adaptive Sampling

Shusen Wang

*College of Computer Science and Technology
Zhejiang University
Hangzhou, Zhejiang 310027, China*

WSS@ZJU.EDU.CN

Zhihua Zhang*

*Department of Computer Science and Engineering
Shanghai Jiao Tong University
800 Dong Chuan Road, Shanghai, China 200240*

ZHIHUA@SJTU.EDU.CN

Editor: Mehryar Mohri

Abstract

The CUR matrix decomposition and the Nyström approximation are two important low-rank matrix approximation techniques. The Nyström method approximates a symmetric positive semidefinite matrix in terms of a small number of its columns, while CUR approximates an arbitrary data matrix by a small number of its columns and rows. Thus, CUR decomposition can be regarded as an extension of the Nyström approximation.

In this paper we establish a more general error bound for the adaptive column/row sampling algorithm, based on which we propose more accurate CUR and Nyström algorithms with expected relative-error bounds. The proposed CUR and Nyström algorithms also have low time complexity and can avoid maintaining the whole data matrix in RAM. In addition, we give theoretical analysis for the lower error bounds of the standard Nyström method and the ensemble Nyström method. The main theoretical results established in this paper are novel, and our analysis makes no special assumption on the data matrices.

Keywords: large-scale matrix computation, CUR matrix decomposition, the Nyström method, randomized algorithms, adaptive sampling

1. Introduction

Large-scale matrices emerging from stocks, genomes, web documents, web images and videos everyday bring new challenges in modern data analysis. Most efforts have been focused on manipulating, understanding and interpreting large-scale data matrices. In many cases, matrix factorization methods are employed for constructing parsimonious and informative representations to facilitate computation and interpretation. A principled approach is the truncated singular value decomposition (SVD) which finds the best low-rank approximation of a data matrix. Applications of SVD such as eigenfaces (Sirovich and Kirby, 1987; Turk and Pentland, 1991) and latent semantic analysis (Deerwester et al., 1990) have been illustrated to be very successful.

*. Corresponding author.

However, using SVD to find basis vectors and low-rank approximations has its limitations. As pointed out by Berry et al. (2005), it is often useful to find a low-rank matrix approximation which possesses additional structures such as sparsity or nonnegativity. Since SVD or the standard QR decomposition for sparse matrices does not preserve sparsity in general, when the sparse matrix is large, computing or even storing such decompositions becomes challenging. Therefore it is useful to compute a low-rank matrix decomposition which preserves such structural properties of the original data matrix.

Another limitation of SVD is that the basis vectors resulting from SVD have little concrete meaning, which makes it very difficult for us to understand and interpret the data in question. An example of Drineas et al. (2008) and Mahoney and Drineas (2009) has well shown this viewpoint; that is, the vector $[(1/2)\text{age} - (1/\sqrt{2})\text{height} + (1/2)\text{income}]$, the sum of the significant uncorrelated features from a data set of people's features, is not particularly informative. Kuruvilla et al. (2002) have also claimed: "it would be interesting to try to find basis vectors for all experiment vectors, using actual experiment vectors and not artificial bases that offer little insight." Therefore, it is of great interest to represent a data matrix in terms of a small number of actual columns and/or actual rows of the matrix. *Matrix column selection* and the *CUR matrix decomposition* provide such techniques.

1.1 Matrix Column Selection

Column selection has been extensively studied in the theoretical computer science (TCS) and numerical linear algebra (NLA) communities. The work in TCS mainly focuses on choosing good columns by randomized algorithms with provable error bounds (Frieze et al., 2004; Deshpande et al., 2006; Drineas et al., 2008; Deshpande and Rademacher, 2010; Boutsidis et al., 2011; Guruswami and Sinop, 2012). The focus in NLA is then on deterministic algorithms, especially the rank-revealing QR factorizations, that select columns by pivoting rules (Foster, 1986; Chan, 1987; Stewart, 1999; Bischof and Hansen, 1991; Hong and Pan, 1992; Chandrasekaran and Ipsen, 1994; Gu and Eisenstat, 1996; Berry et al., 2005). In this paper we focus on randomized algorithms for column selection.

Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, column selection algorithms aim to choose c columns of \mathbf{A} to construct a matrix $\mathbf{C} \in \mathbb{R}^{m \times c}$ such that $\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger\mathbf{A}\|_\xi$ achieves the minimum. Here " $\xi = 2$," " $\xi = F$," and " $\xi = *$ " respectively represent the matrix spectral norm, the matrix Frobenius norm, and the matrix nuclear norm, and \mathbf{C}^\dagger denotes the Moore-Penrose inverse of \mathbf{C} . Since there are $\binom{n}{c}$ possible choices of constructing \mathbf{C} , selecting the best subset is a hard problem.

In recent years, many polynomial-time approximate algorithms have been proposed. Among them we are especially interested in those algorithms with *multiplicative upper bounds*; that is, there exists a polynomial function $f(m, n, k, c)$ such that with $c (\geq k)$ columns selected from \mathbf{A} the following inequality holds

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger\mathbf{A}\|_\xi \leq f(m, n, k, c) \|\mathbf{A} - \mathbf{A}_k\|_\xi$$

with high probability (w.h.p.) or in expectation w.r.t. \mathbf{C} . We call f the *approximation factor*. The bounds are strong when $f = 1 + \epsilon$ for an error parameter ϵ —they are known as *relative-error bounds*. Particularly, the bounds are called *constant-factor bounds* when f does not depend on m and n (Mahoney, 2011). The relative-error bounds and constant-factor bounds of the CUR matrix decomposition and the Nyström approximation are similarly defined.

However, the column selection method, also known as the $\mathbf{A} \approx \mathbf{CX}$ decomposition in some applications, has its limitations. For a large sparse matrix \mathbf{A} , its submatrix \mathbf{C} is sparse, but the coefficient matrix $\mathbf{X} \in \mathbb{R}^{c \times n}$ is not sparse in general. The \mathbf{CX} decomposition suffices when $m \gg n$, because \mathbf{X} is small in size. However, when m and n are near equal, computing and storing the dense matrix \mathbf{X} in RAM becomes infeasible. In such an occasion the CUR matrix decomposition is a very useful alternative.

1.2 The CUR Matrix Decomposition

The CUR matrix decomposition problem has been widely discussed in the literature (Goreinov et al., 1997a,b; Stewart, 1999; Tyrtshnikov, 2000; Berry et al., 2005; Drineas and Mahoney, 2005; Mahoney et al., 2008; Bien et al., 2010), and it has been shown to be very useful in high dimensional data analysis. Particularly, a CUR decomposition algorithm seeks to find a subset of c columns of \mathbf{A} to form a matrix $\mathbf{C} \in \mathbb{R}^{m \times c}$, a subset of r rows to form a matrix $\mathbf{R} \in \mathbb{R}^{r \times n}$, and an intersection matrix $\mathbf{U} \in \mathbb{R}^{c \times r}$ such that $\|\mathbf{A} - \mathbf{CUR}\|_{\xi}$ is small. Accordingly, we use $\tilde{\mathbf{A}} = \mathbf{CUR}$ to approximate \mathbf{A} .

Drineas et al. (2006) proposed a CUR algorithm with additive-error bound. Later on, Drineas et al. (2008) devised a randomized CUR algorithm which has relative-error bound w.h.p. if sufficiently many columns and rows are sampled. Mackey et al. (2011) established a divide-and-conquer method which solves the CUR problem in parallel. The CUR algorithms guaranteed by relative-error bounds are of great interest.

Unfortunately, the existing CUR algorithms usually require a large number of columns and rows to be chosen. For example, for an $m \times n$ matrix \mathbf{A} and a target rank $k \ll \min\{m, n\}$, the *subspace sampling algorithm* (Drineas et al., 2008)—a classical CUR algorithm—requires $O(k\epsilon^{-2} \log k)$ columns and $O(k\epsilon^{-4} \log^2 k)$ rows to achieve relative-error bound w.h.p. The subspace sampling algorithm selects columns/rows according to the statistical leverage scores, so the computational cost of this algorithm is at least equal to the cost of the truncated SVD of \mathbf{A} , that is, $O(mnk)$ in general. However, maintaining a large scale matrix in RAM is often impractical, not to mention performing SVD. Recently, Drineas et al. (2012) devised fast approximation to statistical leverage scores which can be used to speedup the subspace sampling algorithm heuristically—yet no theoretical results have been reported that the leverage scores approximation can give provably efficient subspace sampling algorithm.

The CUR matrix decomposition problem has a close connection with the column selection problem. Especially, most CUR algorithms such as those of Drineas and Kannan (2003); Drineas et al. (2006, 2008) work in a two-stage manner where the first stage is a standard column selection procedure. Despite their strong resemblance, CUR is a harder problem than column selection because “one can get good columns or rows separately” does not mean that one can get good columns and rows together. If the second stage is naïvely solved by a column selection algorithm on \mathbf{A}^T , then the approximation factor will trivially be $\sqrt{2}f^1$ (Mahoney and Drineas, 2009). Thus, more sophisticated error analysis techniques for the second stage are indispensable in order to achieve relative-error bound.

1. It is because $\|\mathbf{A} - \mathbf{CUR}\|_F^2 = \|\mathbf{A} - \mathbf{CC}^\dagger \mathbf{A} + \mathbf{CC}^\dagger \mathbf{A} - \mathbf{CC}^\dagger \mathbf{AR}^\dagger \mathbf{R}\|_F^2 = \|(\mathbf{I} - \mathbf{CC}^\dagger) \mathbf{A}\|_F^2 + \|\mathbf{CC}^\dagger (\mathbf{A} - \mathbf{AR}^\dagger \mathbf{R})\|_F^2 \leq \|\mathbf{A} - \mathbf{CC}^\dagger \mathbf{A}\|_F^2 + \|\mathbf{A} - \mathbf{AR}^\dagger \mathbf{R}\|_F^2 \leq 2f^2 \|\mathbf{A} - \mathbf{A}_k\|_F^2$, where the second equality follows from $(\mathbf{I} - \mathbf{CC}^\dagger)^T \mathbf{CC}^\dagger = 0$.

1.3 The Nyström Methods

The Nyström approximation is closely related to CUR, and it can potentially benefit from the advances in CUR techniques. Different from CUR, the Nyström methods are used for approximating symmetric positive semidefinite (SPSD) matrices. The methods approximate an SPSP matrix only using a subset of its columns, so they can alleviate computation and storage costs when the SPSP matrix in question is large in size. In fact, the Nyström methods have been extensively used in the machine learning community. For example, they have been applied to Gaussian processes (Williams and Seeger, 2001), kernel SVMs (Zhang et al., 2008), spectral clustering (Fowlkes et al., 2004), kernel PCA (Talwalkar et al., 2008; Zhang et al., 2008; Zhang and Kwok, 2010), etc.

The Nyström methods approximate any SPSP matrix in terms of a subset of its columns. Specifically, given an $m \times m$ SPSP matrix \mathbf{A} , they require sampling c ($< m$) columns of \mathbf{A} to construct an $m \times c$ matrix \mathbf{C} . Since there exists an $m \times m$ permutation matrix Π such that $\Pi\mathbf{C}$ consists of the first c columns of $\Pi\mathbf{A}\Pi^T$, we always assume that \mathbf{C} consists of the first c columns of \mathbf{A} without loss of generality. We partition \mathbf{A} and \mathbf{C} as

$$\mathbf{A} = \begin{bmatrix} \mathbf{W} & \mathbf{A}_{21}^T \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} \mathbf{W} \\ \mathbf{A}_{21} \end{bmatrix},$$

where \mathbf{W} and \mathbf{A}_{21} are of sizes $c \times c$ and $(m-c) \times c$, respectively. There are three models which are defined as follows.

- **The Standard Nyström Method.** The standard Nyström approximation to \mathbf{A} is

$$\tilde{\mathbf{A}}_c^{\text{nys}} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T = \begin{bmatrix} \mathbf{W} & \mathbf{A}_{21}^T \\ \mathbf{A}_{21} & \mathbf{A}_{21}\mathbf{W}^\dagger\mathbf{A}_{21}^T \end{bmatrix}. \quad (1)$$

Here \mathbf{W}^\dagger is called the *intersection matrix*. The matrix $(\mathbf{W}_k)^\dagger$, where $k \leq c$ and \mathbf{W}_k is the best k -rank approximation to \mathbf{W} , is also used as an intersection matrix for constructing approximations with even lower rank. But using \mathbf{W}^\dagger results in a tighter approximation than using $(\mathbf{W}_k)^\dagger$ usually.

- **The Ensemble Nyström Method** (Kumar et al., 2009). It selects a collection of t samples, each sample $\mathbf{C}^{(i)}$, ($i = 1, \dots, t$), containing c columns of \mathbf{A} . Then the ensemble method combines the samples to construct an approximation in the form of

$$\tilde{\mathbf{A}}_{t,c}^{\text{ens}} = \sum_{i=1}^t \mu^{(i)} \mathbf{C}^{(i)} \mathbf{W}^{(i)\dagger} \mathbf{C}^{(i)T}, \quad (2)$$

where $\mu^{(i)}$ are the weights of the samples. Typically, the ensemble Nyström method seeks to find out the weights by minimizing $\|\mathbf{A} - \tilde{\mathbf{A}}_{t,c}^{\text{ens}}\|_F$ or $\|\mathbf{A} - \tilde{\mathbf{A}}_{t,c}^{\text{ens}}\|_2$. A simple but effective strategy is to set the weights as $\mu^{(1)} = \dots = \mu^{(t)} = \frac{1}{t}$.

- **The Modified Nyström Method** (proposed in this paper). It is defined as

$$\tilde{\mathbf{A}}_c^{\text{imp}} = \mathbf{C}(\mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T)\mathbf{C}^T.$$

This model is not strictly the Nyström method because it uses a quite different intersection matrix $\mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T$. It costs $O(mc^2)$ time to compute the Moore-Penrose inverse \mathbf{C}^\dagger and m^2c

flops to compute matrix multiplications. The matrix multiplications can be executed very efficiently in multi-processor environment, so ideally computing the intersection matrix costs time only linear in m . This model is more accurate (which will be justified in Section 4.3 and 4.4) but more costly than the conventional ones, so there is a trade-off between time and accuracy when deciding which model to use.

Here and later, we call those which use intersection matrix \mathbf{W}^\dagger or $(\mathbf{W}_k)^\dagger$ *the conventional Nyström methods*, including the standard Nyström and the ensemble Nyström.

To generate effective approximations, much work has been built on the upper error bounds of the sampling techniques for the Nyström method. Most of the work, for example, Drineas and Mahoney (2005), Li et al. (2010), Kumar et al. (2009), Jin et al. (2011), and Kumar et al. (2012), studied the additive-error bound. With assumptions on matrix coherence, better additive-error bounds were obtained by Talwalkar and Rostamizadeh (2010), Jin et al. (2011), and Mackey et al. (2011). However, as stated by Mahoney (2011), additive-error bounds are less compelling than relative-error bounds. In one recent work, Gittens and Mahoney (2013) provided a relative-error bound for the first time, where the bound is in nuclear norm.

However, the error bounds of the previous Nyström methods are much weaker than those of the existing CUR algorithms, especially the relative-error bounds in which we are more interested (Mahoney, 2011). Actually, as will be proved in this paper, the lower error bounds of the standard Nyström method and the ensemble Nyström method are even much worse than the upper bounds of some existing CUR algorithms. This motivates us to improve the Nyström method by borrowing the techniques in CUR matrix decomposition.

1.4 Contributions and Outline

The main technical contribution of this work is the adaptive sampling bound in Theorem 5, which is an extension of Theorem 2.1 of Deshpande et al. (2006). Theorem 2.1 of Deshpande et al. (2006) bounds the error incurred by projection onto column or row space, while our Theorem 5 bounds the error incurred by the projection simultaneously onto column space and row space. We also show that Theorem 2.1 of Deshpande et al. (2006) can be regarded as a special case of Theorem 5.

More importantly, our adaptive sampling bound provides an approach for improving CUR and the Nyström approximation: no matter which relative-error column selection algorithm is employed, Theorem 5 ensures relative-error bounds for CUR and the Nyström approximation. We present the results in Corollary 7.

Based on the adaptive sampling bound in Theorem 5 and its corollary 7, we provide a concrete CUR algorithm which beats the best existing algorithm—the subspace sampling algorithm—both theoretically and empirically. The CUR algorithm is described in Algorithm 2 and analyzed in Theorem 8. In Table 1 we present a comparison between our proposed CUR algorithm and the subspace sampling algorithm. As we see, our algorithm requires much fewer columns and rows to achieve relative-error bound. Our method is more scalable for it works on only a few columns or rows of the data matrix in question; in contrast, the subspace sampling algorithm maintains the whole data matrix in RAM to implement SVD.

Another important application of the adaptive sampling bound is to yield an algorithm for the modified Nyström method. The algorithm has a strong relative-error upper bound: for a target rank k , by sampling $\frac{2k}{\epsilon^2}(1+o(1))$ columns it achieves relative-error bound in expectation. The results are shown in Theorem 10.

	#column (c)	#row (r)	time	space
Adaptive	$\frac{2k}{\epsilon}(1 + o(1))$	$\frac{c}{\epsilon}(1 + \epsilon)$	Roughly $O(nk^2\epsilon^{-4}) + T_{\text{Multiply}}(mnk\epsilon^{-1})$	$O(\max\{mc, nr\})$
Subspace	$O\left(\frac{k \log k}{\epsilon^2}\right)$	$O\left(\frac{c \log c}{\epsilon^2}\right)$	$O(mnk)$	$O(mn)$

Table 1: Comparisons between our *adaptive sampling* based CUR algorithm and the best existing algorithm—the *subspace sampling* algorithm of Drineas et al. (2008).

	$\frac{\ \mathbf{A}-\tilde{\mathbf{A}}\ _F}{\max_{i,j} a_{ij} }$	$\frac{\ \mathbf{A}-\tilde{\mathbf{A}}\ _2}{\max_{i,j} a_{ij} }$	$\frac{\ \mathbf{A}-\tilde{\mathbf{A}}\ _*}{\max_{i,j} a_{ij} }$	$\frac{\ \mathbf{A}-\tilde{\mathbf{A}}\ _F}{\ \mathbf{A}-\mathbf{A}_k\ _F}$	$\frac{\ \mathbf{A}-\tilde{\mathbf{A}}\ _2}{\ \mathbf{A}-\mathbf{A}_k\ _2}$	$\frac{\ \mathbf{A}-\tilde{\mathbf{A}}\ _*}{\ \mathbf{A}-\mathbf{A}_k\ _*}$
Standard	$\Omega\left(\frac{m\sqrt{k}}{c}\right)$	$\Omega\left(\frac{m}{c}\right)$	$\Omega(m-c)$	$\Omega\left(\sqrt{1 + \frac{mk}{c^2}}\right)$	$\Omega\left(\frac{m}{c}\right)$	$\Omega\left(1 + \frac{k}{c}\right)$
Ensemble	$\Omega\left(\frac{m\sqrt{k}}{c}\right)$	–	$\Omega(m-c)$	$\Omega\left(\sqrt{1 + \frac{mk}{c^2}}\right)$	–	$\Omega\left(1 + \frac{k}{c}\right)$

Table 2: Lower bounds of the standard Nyström method and the ensemble Nyström method. The blanks indicate the lower bounds are unknown to us. Here m denotes the column/row number of the SPSD matrix, c denotes the number of selected columns, and k denotes the target rank.

Finally, we establish a collection of lower error bounds of the standard Nyström and the ensemble Nyström that use \mathbf{W}^\dagger as the intersection matrix. We show the lower bounds in Theorem 12 and Table 3; here Table 2 briefly summarizes the lower bounds in Table 3. From the table we can see that the upper error bound of our adaptive sampling algorithm for the modified Nyström method is even better than the lower bounds of the conventional Nyström methods.²

The remainder of the paper is organized as follows. In Section 2 we give the notation that will be used in this paper. In Section 3 we survey the previous work on the randomized column selection, CUR matrix decomposition, and Nyström approximation. In Section 4 we present our theoretical results and corresponding algorithms. In Section 5 we empirically evaluate our proposed CUR and Nyström algorithms. Finally, we conclude our work in Section 6. All proofs are deferred to the appendices.

2. Notation

First of all, we present the notation and notion that are used here and later. We let \mathbf{I}_m denote the $m \times m$ identity matrix, $\mathbf{1}_m$ denote the $m \times 1$ vector of ones, and $\mathbf{0}$ denote a zero vector or matrix with appropriate size. For a matrix $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{m \times n}$, we let $\mathbf{a}^{(i)}$ be its i -th row, \mathbf{a}_j be its j -th column, and $\mathbf{A}_{i:j}$ be a submatrix consisting of its i to j -th columns ($i \leq j$).

Let $\rho = \text{rank}(\mathbf{A}) \leq \min\{m, n\}$ and $k \leq \rho$. The singular value decomposition (SVD) of \mathbf{A} can be written as

$$\mathbf{A} = \sum_{i=1}^{\rho} \sigma_{\mathbf{A},i} \mathbf{u}_{\mathbf{A},i} \mathbf{v}_{\mathbf{A},i}^T = \mathbf{U}_{\mathbf{A}} \Sigma_{\mathbf{A}} \mathbf{V}_{\mathbf{A}}^T = \begin{bmatrix} \mathbf{U}_{\mathbf{A},k} & \mathbf{U}_{\mathbf{A},k\perp} \end{bmatrix} \begin{bmatrix} \Sigma_{\mathbf{A},k} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\mathbf{A},k\perp} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{\mathbf{A},k}^T \\ \mathbf{V}_{\mathbf{A},k\perp}^T \end{bmatrix},$$

2. This can be valid because the lower bounds in Table 2 do not hold when the intersection matrix is not \mathbf{W}^\dagger .

where $\mathbf{U}_{A,k}$ ($m \times k$), $\Sigma_{A,k}$ ($k \times k$), and $\mathbf{V}_{A,k}$ ($n \times k$) correspond to the top k singular values. We denote $\mathbf{A}_k = \mathbf{U}_{A,k} \Sigma_{A,k} \mathbf{V}_{A,k}^T$ which is the best (or closest) rank- k approximation to \mathbf{A} . We also use $\sigma_i(\mathbf{A}) = \sigma_{A,i}$ to denote the i -th largest singular value. When \mathbf{A} is SPSD, the SVD is identical to the eigenvalue decomposition, in which case we have $\mathbf{U}_A = \mathbf{V}_A$.

We define the matrix norms as follows. Let $\|\mathbf{A}\|_1 = \sum_{i,j} |a_{ij}|$ be the ℓ_1 -norm, $\|\mathbf{A}\|_F = (\sum_{i,j} a_{ij}^2)^{1/2} = (\sum_i \sigma_{A,i}^2)^{1/2}$ be the Frobenius norm, $\|\mathbf{A}\|_2 = \max_{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \sigma_{A,1}$ be the spectral norm, and $\|\mathbf{A}\|_* = \sum_i \sigma_{A,i}$ be the nuclear norm. We always use $\|\cdot\|_{\xi}$ to represent $\|\cdot\|_2$, $\|\cdot\|_F$, or $\|\cdot\|_*$.

Based on SVD, the *statistical leverage scores* of the columns of \mathbf{A} relative to the best rank- k approximation to \mathbf{A} is defined as

$$\ell_j^{[k]} = \|\mathbf{v}_{A,k}^{(j)}\|_2^2, \quad j = 1, \dots, n. \quad (3)$$

We have that $\sum_{j=1}^n \ell_j^{[k]} = k$. The leverage scores of the rows of \mathbf{A} are defined according to $\mathbf{U}_{A,k}$. The leverage scores play an important role in low-rank matrix approximation. Informally speaking, the columns (or rows) with high leverage scores have greater influence in rank- k approximation than those with low leverage scores.

Additionally, let $\mathbf{A}^\dagger = \mathbf{V}_{A,\rho} \Sigma_{A,\rho}^{-1} \mathbf{U}_{A,\rho}^T$ be the Moore-Penrose inverse of \mathbf{A} (Ben-Israel and Greville, 2003). When \mathbf{A} is nonsingular, the Moore-Penrose inverse is identical to the matrix inverse. Given matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{X} \in \mathbb{R}^{m \times p}$, and $\mathbf{Y} \in \mathbb{R}^{q \times n}$, $\mathbf{X}\mathbf{X}^\dagger \mathbf{A} = \mathbf{U}_X \mathbf{U}_X^T \mathbf{A} \in \mathbb{R}^{m \times n}$ is the projection of \mathbf{A} onto the column space of \mathbf{X} , and $\mathbf{A}\mathbf{Y}^\dagger \mathbf{Y} = \mathbf{A}\mathbf{V}_Y \mathbf{V}_Y^T \in \mathbb{R}^{m \times n}$ is the projection of \mathbf{A} onto the row space of \mathbf{Y} .

Finally, we discuss the computational costs of the matrix operations mentioned above. For an $m \times n$ general matrix \mathbf{A} (assume $m \geq n$), it takes $O(mn^2)$ flops to compute the full SVD and $O(mnk)$ flops to compute the truncated SVD of rank k ($< n$). The computation of \mathbf{A}^\dagger also takes $O(mn^2)$ flops. It is worth mentioning that, although multiplying an $m \times n$ matrix by an $n \times p$ matrix runs in mnp flops, it can be easily performed in parallel (Halko et al., 2011). In contrast, implementing operations like SVD and QR decomposition in parallel is much more difficult. So we denote the time complexity of such a matrix multiplication by $T_{\text{Multiply}}(mnp)$, which can be tremendously smaller than $O(mnp)$ in practice.

3. Previous Work

In Section 3.1 we present an adaptive sampling algorithm and its relative-error bound established by Deshpande et al. (2006). In Section 3.2 we highlight the near-optimal column selection algorithm of Boutsidis et al. (2011) which we will use in our CUR and Nyström algorithms for column/row sampling. In Section 3.3 we introduce two important CUR algorithms. In Section 3.4 we introduce the only known relative-error algorithm for the standard Nyström method.

3.1 The Adaptive Sampling Algorithm

Adaptive sampling is an effective and efficient column sampling algorithm for reducing the error incurred by the first round of sampling. After one has selected a small subset of columns (denoted \mathbf{C}_1), an adaptive sampling method is used to further select a proportion of columns according to the residual of the first round, that is, $\mathbf{A} - \mathbf{C}_1 \mathbf{C}_1^\dagger \mathbf{A}$. The approximation error is guaranteed to be decreasing by a factor after the adaptive sampling (Deshpande et al., 2006). We show the result of Deshpande et al. (2006) in the following lemma.

Lemma 1 (The Adaptive Sampling Algorithm) (Deshpande et al., 2006) *Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we let $\mathbf{C}_1 \in \mathbb{R}^{m \times c_1}$ consist of c_1 columns of \mathbf{A} , and define the residual $\mathbf{B} = \mathbf{A} - \mathbf{C}_1 \mathbf{C}_1^\dagger \mathbf{A}$. Additionally, for $i = 1, \dots, n$, we define*

$$p_i = \|\mathbf{b}_i\|_2^2 / \|\mathbf{B}\|_F^2.$$

We further sample c_2 columns i.i.d. from \mathbf{A} , in each trial of which the i -th column is chosen with probability p_i . Let $\mathbf{C}_2 \in \mathbb{R}^{m \times c_2}$ contain the c_2 sampled columns and let $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2] \in \mathbb{R}^{m \times (c_1 + c_2)}$. Then, for any integer $k > 0$, the following inequality holds:

$$\mathbb{E} \|\mathbf{A} - \mathbf{C} \mathbf{C}^\dagger \mathbf{A}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \frac{k}{c_2} \|\mathbf{A} - \mathbf{C}_1 \mathbf{C}_1^\dagger \mathbf{A}\|_F^2,$$

where the expectation is taken w.r.t. \mathbf{C}_2 .

We will establish in Theorem 5 a more general and more useful error bound for this adaptive sampling algorithm. It can be shown that Lemma 1 is a special case of Theorem 5.

3.2 The Near-Optimal Column Selection Algorithm

Boutsidis et al. (2011) proposed a relative-error column selection algorithm which requires only $c = 2k\epsilon^{-1}(1+o(1))$ columns get selected. Boutsidis et al. (2011) also proved the lower bound of the column selection problem which shows that no column selection algorithm can achieve relative-error bound by selecting less than $c = k\epsilon^{-1}$ columns. Thus this algorithm is near optimal. Though an optimal algorithm recently proposed by Guruswami and Sinop (2012) attains the the lower bound, this algorithm is quite inefficient in comparison with the near-optimal algorithm. So we prefer to use the near-optimal algorithm in our CUR and Nyström algorithms for column/row sampling.

The near-optimal algorithm consists of three steps: the approximate SVD via random projection (Boutsidis et al., 2011; Halko et al., 2011), the dual set sparsification algorithm (Boutsidis et al., 2011), and the adaptive sampling algorithm (Deshpande et al., 2006). We describe the near-optimal algorithm in Algorithm 1 and present the theoretical analysis in Lemma 2.

Lemma 2 (The Near-Optimal Column Selection Algorithm) *Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ of rank ρ , a target rank k ($2 \leq k < \rho$), and $0 < \epsilon < 1$. Algorithm 1 selects*

$$c = \frac{2k}{\epsilon} (1 + o(1))$$

columns of \mathbf{A} to form a matrix $\mathbf{C} \in \mathbb{R}^{m \times c}$, then the following inequality holds:

$$\mathbb{E} \|\mathbf{A} - \mathbf{C} \mathbf{C}^\dagger \mathbf{A}\|_F^2 \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2,$$

where the expectation is taken w.r.t. \mathbf{C} . Furthermore, the matrix \mathbf{C} can be obtained in $O(mk^2\epsilon^{-4/3} + nk^3\epsilon^{-2/3}) + T_{\text{Multiply}}(mnk\epsilon^{-2/3})$ time.

This algorithm has the merits of low time complexity and space complexity. None of the three steps—the randomized SVD, the dual set sparsification algorithm, and the adaptive sampling—requires loading the whole of \mathbf{A} into RAM. All of the three steps can work on only a small subset of the columns of \mathbf{A} . Though a relative-error algorithm recently proposed by Guruswami and Sinop (2012) requires even fewer columns, it is less efficient than the near-optimal algorithm.

Algorithm 1 The Near-Optimal Column Selection Algorithm of Boutsidis et al. (2011).

- 1: **Input:** a real matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, target rank k , error parameter $\varepsilon \in (0, 1]$, target column number $c = \frac{2k}{\varepsilon}(1 + o(1))$;
 - 2: Compute approximate truncated SVD via random projection such that $\mathbf{A}_k \approx \tilde{\mathbf{U}}_k \tilde{\Sigma}_k \tilde{\mathbf{V}}_k$;
 - 3: Construct $\mathcal{U} \leftarrow$ columns of $(\mathbf{A} - \tilde{\mathbf{U}}_k \tilde{\Sigma}_k \tilde{\mathbf{V}}_k)$; $\mathcal{V} \leftarrow$ columns of $\tilde{\mathbf{V}}_k^T$;
 - 4: Compute $\mathbf{s} \leftarrow$ Dual Set Spectral-Frobenius Sparsification Algorithm ($\mathcal{U}, \mathcal{V}, c - 2k/\varepsilon$);
 - 5: Construct $\mathbf{C}_1 \leftarrow \mathbf{A} \text{Diag}(\mathbf{s})$, and then delete the all-zero columns;
 - 6: Residual matrix $\mathbf{D} \leftarrow \mathbf{A} - \mathbf{C}_1 \mathbf{C}_1^\dagger \mathbf{A}$;
 - 7: Compute sampling probabilities: $p_i = \|\mathbf{d}_i\|_2^2 / \|\mathbf{D}\|_F^2, i = 1, \dots, n$;
 - 8: Sampling $c_2 = 2k/\varepsilon$ columns from \mathbf{A} with probability $\{p_1, \dots, p_n\}$ to construct \mathbf{C}_2 ;
 - 9: **return** $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2]$.
-

3.3 Previous Work in CUR Matrix Decomposition

We introduce in this section two highly effective CUR algorithms: one is deterministic and the other is randomized.

3.3.1 THE SPARSE COLUMN-ROW APPROXIMATION (SCRA)

Stewart (1999) proposed a deterministic CUR algorithm and called it the sparse column-row approximation (SCRA). SCRA is based on the truncated pivoted QR decomposition via a quasi Gram-Schmidt algorithm. Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the truncated pivoted QR decomposition procedure deterministically finds a set of columns $\mathbf{C} \in \mathbb{R}^{m \times c}$ by column pivoting, whose span approximates the column space of \mathbf{A} , and computes an upper triangular matrix $\mathbf{T}_C \in \mathbb{R}^{c \times c}$ that orthogonalizes those columns. SCRA runs the same procedure again on \mathbf{A}^T to select a set of rows $\mathbf{R} \in \mathbb{R}^{r \times n}$ and computes the corresponding upper triangular matrix $\mathbf{T}_R \in \mathbb{R}^{r \times r}$. Let $\mathbf{C} = \mathbf{Q}_C \mathbf{T}_C$ and $\mathbf{R}^T = \mathbf{Q}_R \mathbf{T}_R$ denote the resulting truncated pivoted QR decomposition. The intersection matrix is computed by $\mathbf{U} = (\mathbf{T}_C^T \mathbf{T}_C)^{-1} \mathbf{C}^T \mathbf{A} \mathbf{R}^T (\mathbf{T}_R^T \mathbf{T}_R)^{-1}$. According to our experiments, this algorithm is quite effective but very time expensive, especially when c and r are large. Moreover, this algorithm does not have data-independent error bound.

3.3.2 THE SUBSPACE SAMPLING CUR ALGORITHM

Drineas et al. (2008) proposed a two-stage randomized CUR algorithm which has a relative-error bound with high probability (w.h.p.). In the first stage the algorithm samples c columns of \mathbf{A} to construct \mathbf{C} , and in the second stage it samples r rows from \mathbf{A} and \mathbf{C} simultaneously to construct \mathbf{R} and \mathbf{W} and let $\mathbf{U} = \mathbf{W}^\dagger$. The sampling probabilities in the two stages are proportional to the leverage scores of \mathbf{A} and \mathbf{C} , respectively. That is, in the first stage the sampling probabilities are proportional to the squared ℓ_2 -norm of the rows of $\mathbf{V}_{A,k}$; in the second stage the sampling probabilities are proportional to the squared ℓ_2 -norm of the rows of \mathbf{U}_C . That is why it is called the *subspace sampling algorithm*. Here we show the main results of the subspace sampling algorithm in the following lemma.

Lemma 3 (Subspace Sampling for CUR) *Given an $m \times n$ matrix \mathbf{A} and a target rank $k \ll \min\{m, n\}$, the subspace sampling algorithm selects $c = O(k\varepsilon^{-2} \log k \log(1/\delta))$ columns and $r = O(c\varepsilon^{-2} \log c \log(1/\delta))$ rows without replacement. Then*

$$\|\mathbf{A} - \mathbf{CUR}\|_F = \|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger \mathbf{R}\|_F \leq (1 + \varepsilon) \|\mathbf{A} - \mathbf{A}_k\|_F,$$

holds with probability at least $1 - \delta$, where \mathbf{W} contains the rows of \mathbf{C} with scaling. The running time is dominated by the truncated SVD of \mathbf{A} , that is, $O(mnk)$.

3.4 Previous Work in the Nyström Approximation

In a very recent work, Gittens and Mahoney (2013) established a framework for analyzing errors incurred by the standard Nyström method. Especially, the authors provided the first and the only known relative-error (in nuclear norm) algorithm for the standard Nyström method. The algorithm is described as follows and, its bound is shown in Lemma 4.

Like the CUR algorithm in Section 3.3.2, the Nyström algorithm also samples columns by the subspace sampling of Drineas et al. (2008). Each column is selected with probability $p_j = \frac{1}{k} \ell_j^{[k]}$ with replacement, where $\ell_1^{[k]}, \dots, \ell_m^{[k]}$ are leverage scores defined in (3). After column sampling, \mathbf{C} and \mathbf{W} are obtained by scaling the selected columns, that is,

$$\mathbf{C} = \mathbf{A}(\mathbf{SD}) \quad \text{and} \quad \mathbf{W} = (\mathbf{SD})^T \mathbf{A}(\mathbf{SD}).$$

Here $\mathbf{S} \in \mathbb{R}^{m \times c}$ is a column selection matrix that $s_{ij} = 1$ if the i -th column of \mathbf{A} is the j -th column selected, and $\mathbf{D} \in \mathbb{R}^{c \times c}$ is a diagonal scaling matrix satisfying $d_{jj} = \frac{1}{\sqrt{cp_i}}$ if $s_{ij} = 1$.

Lemma 4 (Subspace Sampling for the Nyström Approximation) *Given an $m \times m$ SPSD matrix \mathbf{A} and a target rank $k \ll m$, the subspace sampling algorithm selects*

$$c = 3200\epsilon^{-1}k \log(16k/\delta)$$

columns without replacement and constructs \mathbf{C} and \mathbf{W} by scaling the selected columns. Then the inequality

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger \mathbf{C}^T\|_* \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_*,$$

holds with probability at least $0.6 - \delta$.

4. Main Results

We now present our main results. We establish a new error bound for the adaptive sampling algorithm in Section 4.1. We apply adaptive sampling to the CUR and modified Nyström problems, obtaining effective and efficient CUR and Nyström algorithms in Section 4.2 and Section 4.3 respectively. In Section 4.4 we study lower bounds of the conventional Nyström methods to demonstrate the advantages of our approach. Finally, in Section 4.5 we show that our expected bounds can extend to with high probability (w.h.p.) bounds.

4.1 Adaptive Sampling

The relative-error adaptive sampling algorithm is originally established in Theorem 2.1 of Deshpande et al. (2006) (see also Lemma 1 in Section 3.1). The algorithm is based on the following idea: after selecting a proportion of columns from \mathbf{A} to form \mathbf{C}_1 by an arbitrary algorithm, the algorithm randomly samples additional c_2 columns according to the residual $\mathbf{A} - \mathbf{C}_1 \mathbf{C}_1^\dagger \mathbf{A}$. Here we prove a new and more general error bound for the same adaptive sampling algorithm.

Theorem 5 (The Adaptive Sampling Algorithm) *Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a matrix $\mathbf{C} \in \mathbb{R}^{m \times c}$ such that $\text{rank}(\mathbf{C}) = \text{rank}(\mathbf{C}\mathbf{C}^\dagger\mathbf{A}) = \rho$ ($\rho \leq c \leq n$). We let $\mathbf{R}_1 \in \mathbb{R}^{r_1 \times n}$ consist of r_1 rows of \mathbf{A} , and define the residual $\mathbf{B} = \mathbf{A} - \mathbf{A}\mathbf{R}_1^\dagger\mathbf{R}_1$. Additionally, for $i = 1, \dots, m$, we define*

$$p_i = \|\mathbf{b}^{(i)}\|_2^2 / \|\mathbf{B}\|_F^2.$$

We further sample r_2 rows i.i.d. from \mathbf{A} , in each trial of which the i -th row is chosen with probability p_i . Let $\mathbf{R}_2 \in \mathbb{R}^{r_2 \times n}$ contain the r_2 sampled rows and let $\mathbf{R} = [\mathbf{R}_1^T, \mathbf{R}_2^T]^T \in \mathbb{R}^{(r_1+r_2) \times n}$. Then we have

$$\mathbb{E}\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger\mathbf{A}\mathbf{R}^\dagger\mathbf{R}\|_F^2 \leq \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger\mathbf{A}\|_F^2 + \frac{\rho}{r_2}\|\mathbf{A} - \mathbf{A}\mathbf{R}_1^\dagger\mathbf{R}_1\|_F^2,$$

where the expectation is taken w.r.t. \mathbf{R}_2 .

Remark 6 *This theorem shows a more general bound for adaptive sampling than the original one in Theorem 2.1 of Deshpande et al. (2006). The original one bounds the error incurred by projection onto the column space of \mathbf{C} , while Theorem 5 bounds the error incurred by projection onto the column space of \mathbf{C} and row space of \mathbf{R} simultaneously—such situation rises in problems such as CUR and the Nyström approximation. It is worth pointing out that Theorem 2.1 of Deshpande et al. (2006) is a direct corollary of this theorem when $\mathbf{C} = \mathbf{A}_k$ (i.e., $c = n$, $\rho = k$, and $\mathbf{C}\mathbf{C}^\dagger\mathbf{A} = \mathbf{A}_k$).*

As discussed in Section 1.2, selecting good columns or rows separately does not ensure good columns and rows together for CUR and the Nyström approximation. Theorem 5 is thereby important for it guarantees the combined effect column and row selection. Guaranteed by Theorem 5, any column selection algorithm with relative-error bound can be applied to CUR and the Nyström approximation. We show the result in the following corollary.

Corollary 7 (Adaptive Sampling for CUR and the Nyström Approximation) *Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, a target rank $k \ll m, n$, and a column selection algorithm \mathcal{A}_{col} which achieves relative-error upper bound by selecting $c \geq C(k, \epsilon)$ columns. Then we have the following results for CUR and the Nyström approximation.*

- (1) *By selecting $c \geq C(k, \epsilon)$ columns of \mathbf{A} to construct \mathbf{C} and $r_1 = c$ rows to construct \mathbf{R}_1 , both using algorithm \mathcal{A}_{col} , followed by selecting additional $r_2 = c/\epsilon$ rows using the adaptive sampling algorithm to construct \mathbf{R}_2 , the CUR matrix decomposition achieves relative-error upper bound in expectation:*

$$\mathbb{E}\|\mathbf{A} - \mathbf{C}\mathbf{U}\mathbf{R}\|_F \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F,$$

where $\mathbf{R} = [\mathbf{R}_1^T, \mathbf{R}_2^T]^T$ and $\mathbf{U} = \mathbf{C}^\dagger\mathbf{A}\mathbf{R}^\dagger$.

- (2) *Suppose \mathbf{A} is an $m \times m$ symmetric matrix. By selecting $c_1 \geq C(k, \epsilon)$ columns of \mathbf{A} to construct \mathbf{C}_1 using \mathcal{A}_{col} and selecting $c_2 = c_1/\epsilon$ columns of \mathbf{A} to construct \mathbf{C}_2 using the adaptive sampling algorithm, the modified Nyström method achieves relative-error upper bound in expectation:*

$$\mathbb{E}\|\mathbf{A} - \mathbf{C}\mathbf{U}\mathbf{C}^T\|_F \leq (1 + \epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F,$$

where $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2]$ and $\mathbf{U} = \mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T$.

Based on Corollary 7, we attempt to solve CUR and the Nyström by adaptive sampling algorithms. We present concrete algorithms in Section 4.2 and 4.3.

Algorithm 2 Adaptive Sampling for CUR.

- 1: **Input:** a real matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, target rank k , $\varepsilon \in (0, 1]$, target column number $c = \frac{2k}{\varepsilon}(1 + o(1))$, target row number $r = \frac{c}{\varepsilon}(1 + \varepsilon)$;
 - 2: Select $c = \frac{2k}{\varepsilon}(1 + o(1))$ columns of \mathbf{A} to construct $\mathbf{C} \in \mathbb{R}^{m \times c}$ using Algorithm 1;
 - 3: Select $r_1 = c$ rows of \mathbf{A} to construct $\mathbf{R}_1 \in \mathbb{R}^{r_1 \times n}$ using Algorithm 1;
 - 4: Adaptively sample $r_2 = c/\varepsilon$ rows from \mathbf{A} according to the residual $\mathbf{A} - \mathbf{A}\mathbf{R}_1^\dagger\mathbf{R}_1$;
 - 5: **return** \mathbf{C} , $\mathbf{R} = [\mathbf{R}_1^T, \mathbf{R}_2^T]^T$, and $\mathbf{U} = \mathbf{C}^\dagger\mathbf{A}\mathbf{R}^\dagger$.
-

4.2 Adaptive Sampling for CUR Matrix Decomposition

Guaranteed by the novel adaptive sampling bound in Theorem 5, we combine the near-optimal column selection algorithm of Boutsidis et al. (2011) and the adaptive sampling algorithm for solving the CUR problem, giving rise to an algorithm with a much tighter theoretical bound than existing algorithms. The algorithm is described in Algorithm 2 and its analysis is given in Theorem 8. Theorem 8 follows immediately from Lemma 2 and Corollary 7.

Theorem 8 (Adaptive Sampling for CUR) *Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a positive integer $k \ll \min\{m, n\}$, the CUR algorithm described in Algorithm 2 randomly selects $c = \frac{2k}{\varepsilon}(1 + o(1))$ columns of \mathbf{A} to construct $\mathbf{C} \in \mathbb{R}^{m \times c}$, and then selects $r = \frac{c}{\varepsilon}(1 + \varepsilon)$ rows of \mathbf{A} to construct $\mathbf{R} \in \mathbb{R}^{r \times n}$. Then we have*

$$\mathbb{E}\|\mathbf{A} - \mathbf{CUR}\|_F = \mathbb{E}\|\mathbf{A} - \mathbf{C}(\mathbf{C}^\dagger\mathbf{A}\mathbf{R}^\dagger)\mathbf{R}\|_F \leq (1 + \varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_F.$$

The algorithm costs time $O((m + n)k^3\varepsilon^{-2/3} + mk^2\varepsilon^{-2} + nk^2\varepsilon^{-4}) + T_{\text{Multiply}}(mnk\varepsilon^{-1})$ to compute matrices \mathbf{C} , \mathbf{U} and \mathbf{R} .

When the algorithm is executed in a single-core processor, the time complexity of the CUR algorithm is linear in mn ; when executed in multi-processor environment where matrix multiplication is performed in parallel, ideally the algorithm costs time only linear in $m+n$. Another advantage of this algorithm is that it avoids loading the whole $m \times n$ data matrix \mathbf{A} into RAM. Neither the near-optimal column selection algorithm nor the adaptive sampling algorithm requires loading the whole of \mathbf{A} into RAM. The most space-expensive operation throughout this algorithm is computation of the Moore-Penrose inverses of \mathbf{C} and \mathbf{R} , which requires maintaining an $m \times c$ matrix or an $r \times n$ matrix in RAM. To compute the intersection matrix $\mathbf{C}^\dagger\mathbf{A}\mathbf{R}^\dagger$, the algorithm needs to visit each entry of \mathbf{A} , but it is not RAM expensive because the multiplication can be done by computing $\mathbf{C}^\dagger\mathbf{a}_j$ for $j = 1, \dots, n$ separately. The above analysis is also valid for the Nyström algorithm in Theorem 10.

Remark 9 *If we replace the near-optimal column selection algorithm in Theorem 8 by the optimal algorithm of Guruswami and Sinop (2012), it suffices to select $c = k\varepsilon^{-1}(1 + o(1))$ columns and $r = c\varepsilon^{-1}(1 + \varepsilon)$ rows totally. But the optimal algorithm is less efficient than the near-optimal algorithm.*

4.3 Adaptive Sampling for the Nyström Approximation

Theorem 5 provides an approach for bounding the approximation errors incurred by projection simultaneously onto column space and row space. Thus this approach can be applied to solve the modified Nyström method. The following theorem follows directly from Lemma 2 and Corollary 7.

	$\frac{\ \mathbf{A}-\tilde{\mathbf{A}}\ _F}{\max_{i,j} a_{ij} }$	$\frac{\ \mathbf{A}-\tilde{\mathbf{A}}\ _2}{\max_{i,j} a_{ij} }$	$\frac{\ \mathbf{A}-\tilde{\mathbf{A}}\ _*}{\max_{i,j} a_{ij} }$
<i>Standard</i>	$0.99\sqrt{m-c-k+k\left(\frac{m+99k}{c+99k}\right)^2}$	$\frac{0.99(m+99)}{c+99}$	$0.99(m-c)\left(1+\frac{k}{c+99k}\right)$
<i>Ensemble</i>	$0.99\sqrt{\left(m-2c+\frac{c}{t}-k\right)+k\left(\frac{m-c+\frac{c}{t}+99k}{c+99k}\right)^2}$	–	$0.99(m-c)\left(1+\frac{k}{c+99k}\right)$

	$\frac{\ \mathbf{A}-\tilde{\mathbf{A}}\ _F}{\ \mathbf{A}-\mathbf{A}_k\ _F}$	$\frac{\ \mathbf{A}-\tilde{\mathbf{A}}\ _2}{\ \mathbf{A}-\mathbf{A}_k\ _2}$	$\frac{\ \mathbf{A}-\tilde{\mathbf{A}}\ _*}{\ \mathbf{A}-\mathbf{A}_k\ _*}$
<i>Standard</i>	$\sqrt{1+\frac{m^2k-c^3}{c^2(m-k)}}$	$\frac{m}{c}$	$\frac{m-c}{m-k}\left(1+\frac{k}{c}\right)$
<i>Ensemble</i>	$\sqrt{\frac{m-2c+c/t-k}{m-k}\left(1+\frac{k(m-2c+c/t)}{c^2}\right)}$	–	$\frac{m-c}{m-k}\left(1+\frac{k}{c}\right)$

Table 3: Lower bounds of the standard Nyström method and the ensemble Nyström method. The blanks indicate the lower bounds are unknown to us. Here m denotes the column/row number of the SPSD matrix, c denotes the number of selected columns, and k denotes the target rank.

Theorem 10 (Adaptive Sampling for the Modified Nyström Method) *Given a symmetric matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ and a target rank k , with $c_1 = \frac{2k}{\varepsilon}(1 + o(1))$ columns sampled by Algorithm 1 and $c_2 = c_1/\varepsilon$ columns sampled by the adaptive sampling algorithm, that is, with totally $c = \frac{2k}{\varepsilon^2}(1 + o(1))$ columns being sampled, the approximation error incurred by the modified Nyström method is upper bounded by*

$$\mathbb{E}\|\mathbf{A} - \mathbf{C}\mathbf{U}\mathbf{C}^T\|_F \leq \mathbb{E}\left\|\mathbf{A} - \mathbf{C}\left(\mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T\right)\mathbf{C}^T\right\|_F \leq (1 + \varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_F.$$

The algorithm costs time $O(mk^2\varepsilon^{-4} + mk^3\varepsilon^{-2/3}) + T_{\text{Multiply}}(m^2k\varepsilon^{-2})$ in computing \mathbf{C} and \mathbf{U} .

Remark 11 *The error bound in Theorem 10 is the only Frobenius norm relative-error bound for the Nyström approximation at present, and it is also a constant-factor bound. If one uses the optimal column selection algorithm of Guruswami and Sinop (2012), which is less efficient, the error bound is further improved: only $c = \frac{k}{\varepsilon^2}(1 + o(1))$ columns are required. Furthermore, the theorem requires the matrix \mathbf{A} to be symmetric, which is milder than the SPSD requirement made in the previous work.*

This is yet the strongest result for the Nyström approximation problem—much stronger than the best possible algorithms for the conventional Nyström method. We will illustrate this point by revealing the lower error bounds of the conventional Nyström methods.

4.4 Lower Error Bounds of the Conventional Nyström Methods

We now demonstrate to what an extent our modified Nyström method is superior over the conventional Nyström methods (namely the standard Nyström defined in (1) and the ensemble Nyström in (2)) by showing the lower error bounds of the conventional Nyström methods. The conventional Nyström methods work no better than the lower error bounds unless additional assumptions are made on the original matrix \mathbf{A} . We show in Theorem 12 the lower error bounds of the conventional Nyström methods; the results are briefly summarized previously in Table 2.

To derive lower error bounds, we construct two adversarial cases for the Nyström methods. To derive the spectral norm lower bounds, we use an SPSD matrix \mathbf{B} whose diagonal entries equal to 1

and off-diagonal entries equal to $\alpha \in [0, 1)$. For the Frobenius norm and nuclear norm bounds, we construct an $m \times m$ block diagonal matrix \mathbf{A} which has k diagonal blocks, each of which is $\frac{m}{k} \times \frac{m}{k}$ in size and constructed in the same way as \mathbf{B} . For the lower bounds on $\frac{\|\mathbf{A} - \tilde{\mathbf{A}}\|_{\xi}}{\max_{i,j} |a_{ij}|}$, α is set to be constant; for the bounds on $\frac{\|\mathbf{A} - \tilde{\mathbf{A}}\|_{\xi}}{\|\mathbf{A} - \mathbf{A}_k\|_{\xi}}$, α is set to be $\alpha \rightarrow 1$. The detailed proof of Theorem 12 is deferred to Appendix C.

Theorem 12 (Lower Error Bounds of the Nyström Methods) *Assume we are given an SPSSD matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ and a target rank k . Let \mathbf{A}_k denote the best rank- k approximation to \mathbf{A} . Let $\tilde{\mathbf{A}}$ denote either the rank- c approximation to \mathbf{A} constructed by the standard Nyström method in (1), or the approximation constructed by the ensemble Nyström method in (2) with t non-overlapping samples, each of which contains c columns of \mathbf{A} . Then there exists an SPSSD matrix such that for any sampling strategy the approximation errors of the conventional Nyström methods, that is, $\|\mathbf{A} - \tilde{\mathbf{A}}\|_{\xi}$, ($\xi = 2, F$, or “*”), are lower bounded by some factors which are shown in Table 3.*

Remark 13 *The lower bounds in Table 3 (or Table 2) show the conventional Nyström methods can be sometimes very ineffective. The spectral norm and Frobenius norm bounds even depend on m , so such bounds are not constant-factor bounds. Notice that the lower error bounds do not meet if \mathbf{W}^{\dagger} is replaced by $\mathbf{C}^{\dagger} \mathbf{A} (\mathbf{C}^{\dagger})^T$, so our modified Nyström method is not limited by such lower bounds.*

4.5 Discussions of the Expected Relative-Error Bounds

The upper error bounds established in this paper all hold in expectation. Now we show that the expected error bounds immediately extend to w.h.p. bounds using Markov’s inequality. Let the random variable $X = \|\mathbf{A} - \tilde{\mathbf{A}}\|_F / \|\mathbf{A} - \mathbf{A}_k\|_F$ denote the error ratio, where

$$\tilde{\mathbf{A}} = \mathbf{C} \mathbf{U} \mathbf{R} \text{ or } \mathbf{C} \mathbf{U} \mathbf{C}^T.$$

Then we have $\mathbb{E}(X) \leq 1 + \epsilon$ by the preceding theorems. By applying Markov’s inequality we have that

$$\mathbb{P}(X > 1 + s\epsilon) < \frac{\mathbb{E}(X)}{1 + s\epsilon} < \frac{1 + \epsilon}{1 + s\epsilon},$$

where s is an arbitrary constant greater than 1. Repeating the sampling procedure for t times and letting $X_{(i)}$ correspond to the error ratio of the i -th sample, we obtain an upper bound on the failure probability:

$$\mathbb{P}\left(\min_i \{X_{(i)}\} > 1 + s\epsilon\right) = \mathbb{P}\left(X_{(i)} > 1 + s\epsilon \forall i = 1, \dots, t\right) < \left(\frac{1 + \epsilon}{1 + s\epsilon}\right)^t \triangleq \delta, \quad (4)$$

which decays exponentially with t . Therefore, by repeating the sampling procedure multiple times and choosing the best sample, our CUR and Nyström algorithms are also guaranteed with w.h.p. relative-error bounds. It follows directly from (4) that, by repeating the sampling procedure for

$$t \geq \frac{1 + \epsilon}{(s - 1)\epsilon} \log\left(\frac{1}{\delta}\right)$$

times, the inequality

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_F \leq (1 + s\epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F$$

holds with probability at least $1 - \delta$.

For instance, we let $s = 1 + \log(1/\delta)$, then by repeating the sampling procedure for $t \geq 1 + 1/\epsilon$ times, the inequality

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_F \leq \left(1 + \epsilon + \epsilon \log(1/\delta)\right) \|\mathbf{A} - \mathbf{A}_k\|_F$$

holds with probability at least $1 - \delta$.

For another instance, we let $s = 2$, then by repeating the sampling procedure for $t \geq (1 + 1/\epsilon) \log(1/\delta)$ times, the inequality

$$\|\mathbf{A} - \tilde{\mathbf{A}}\|_F \leq (1 + 2\epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F$$

holds with probability at least $1 - \delta$.

5. Empirical Analysis

In Section 5.1 we empirically evaluate our CUR algorithms in comparison with the algorithms introduced in Section 3.3. In Section 5.2 we conduct empirical comparisons between the standard Nyström and our modified Nyström, and comparisons among three sampling algorithms. We report the approximation error incurred by each algorithm on each data set. The error ratio is defined by

$$\text{Error Ratio} = \frac{\|\mathbf{A} - \tilde{\mathbf{A}}\|_F}{\|\mathbf{A} - \mathbf{A}_k\|_F},$$

where $\tilde{\mathbf{A}} = \mathbf{C}\mathbf{U}\mathbf{R}$ for the CUR matrix decomposition, $\tilde{\mathbf{A}} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$ for the standard Nyström method, and $\tilde{\mathbf{A}} = \mathbf{C}(\mathbf{C}^\dagger\mathbf{A}(\mathbf{C}^\dagger)^T)\mathbf{C}^T$ for the modified Nyström method.

We conduct experiments on a workstation with two Intel Xeon 2.40GHz CPUs, 24GB RAM, and 64bit Windows Server 2008 system. We implement the algorithms in MATLAB R2011b, and use the MATLAB function ‘svds’ for truncated SVD. To compare the running time, all the computations are carried out in a single thread by setting ‘maxNumCompThreads(1)’ in MATLAB.

5.1 Comparison among the CUR Algorithms

In this section we empirically compare our adaptive sampling based CUR algorithm (Algorithm 2) with the subspace sampling algorithm of Drineas et al. (2008) and the deterministic sparse column-row approximation (SCRA) algorithm of Stewart (1999). For SCRA, we use the MATLAB code released by Stewart (1999). As for the subspace sampling algorithm, we compute the leverages scores exactly via the truncated SVD. Although the fast approximation to leverage scores (Drineas et al., 2012) can significantly speedup subspace sampling, we do not use it because the approximation has no theoretical guarantee when applied to subspace sampling.

Data Set	Type	Size	#Nonzero Entries	Source
Enron Emails	text	39,861 × 28,102	3,710,420	Bag-of-words, UCI
Dexter	text	20,000 × 2,600	248,616	Guyon et al. (2004)
Farm Ads	text	54,877 × 4,143	821,284	Mesterharm and Pazzani (2011)
Gisette	handwritten digit	13,500 × 5,000	8,770,559	Guyon et al. (2004)

Table 4: A summary of the data sets for CUR matrix decomposition.

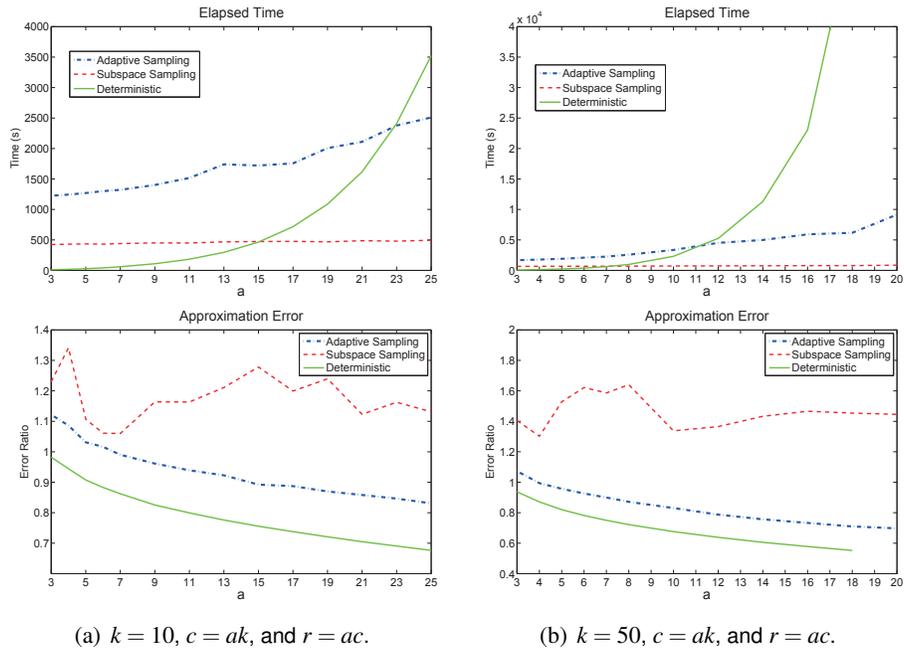


Figure 1: Results of the CUR algorithms on the Enron data set.

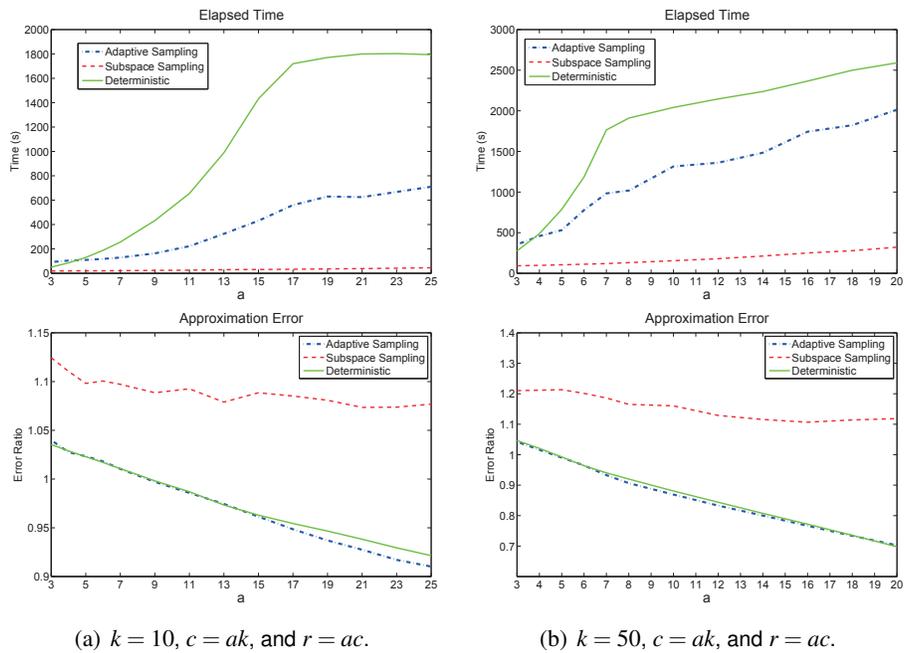


Figure 2: Results of the CUR algorithms on the Dexter data set.

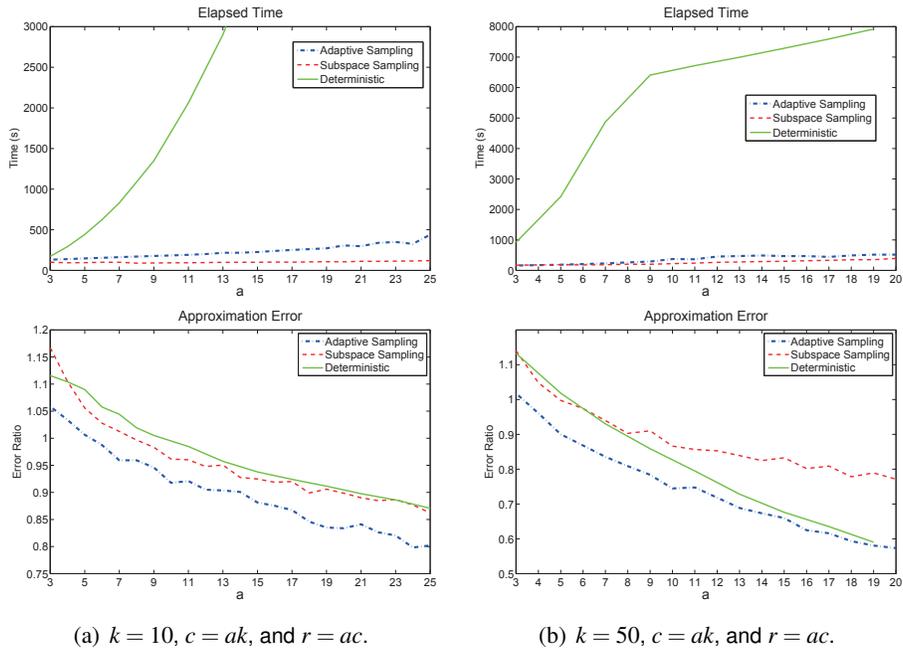


Figure 3: Results of the CUR algorithms on the Farm Ads data set.

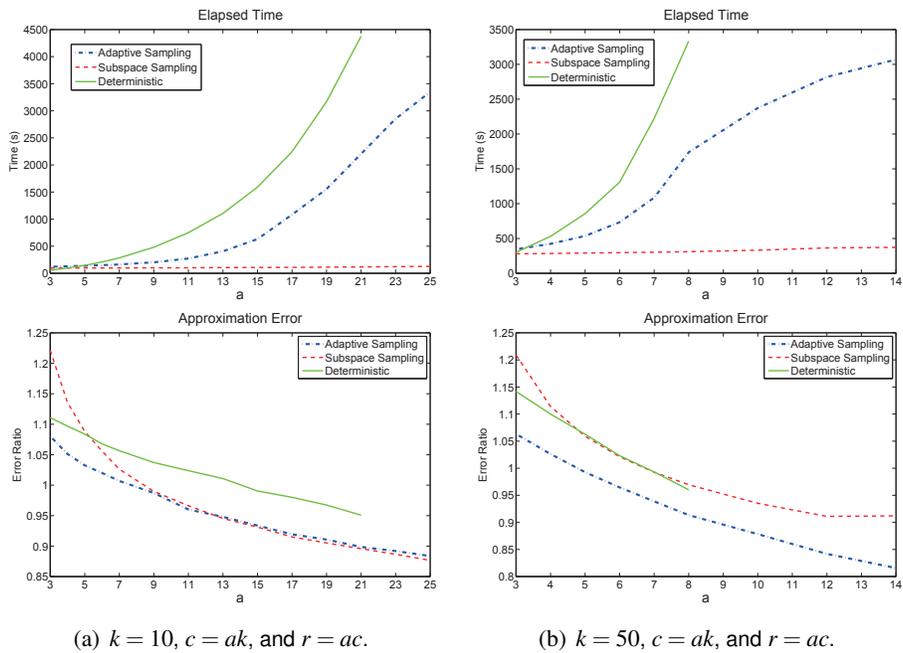


Figure 4: Results of the CUR algorithms on the Gisette data set.

We conduct experiments on four UCI data sets (Frank and Asuncion, 2010) which are summarized in Table 4. Each data set is represented as a data matrix, upon which we apply the CUR algorithms. According to our analysis, the target rank k should be far less than m and n , and the column number c and row number r should be strictly greater than k . For each data set and each algorithm, we set $k = 10$ or 50 , and $c = ak$, $r = ac$, where a ranges in each set of experiments. We repeat each of the two randomized algorithms 10 times, and report the minimum error ratio and the total elapsed time of the 10 rounds. We depict the error ratios and the elapsed time of the three CUR matrix decomposition algorithms in Figures 1, 2, 3, and 4.

We can see from Figures 1, 2, 3, and 4 that our adaptive sampling based CUR algorithm has much lower approximation error than the subspace sampling algorithm in all cases. Our adaptive sampling based algorithm is better than the deterministic SCRA on the Farm Ads data set and the Gisette data set, worse than SCRA on the Enron data set, and comparable to SCRA on the Dexter data set. In addition, the experimental results match our theoretical analysis in Section 4 very well. The empirical results all obey the theoretical relative-error upper bound

$$\frac{\|\mathbf{A} - \mathbf{CUR}\|_F}{\|\mathbf{A} - \mathbf{A}_k\|_F} \leq 1 + \frac{2k}{c}(1 + o(1)) = 1 + \frac{2}{a}(1 + o(1)).$$

As for the running time, the subspace sampling algorithm and our adaptive sampling based algorithm are much more efficient than SCRA, especially when c and r are large. Our adaptive sampling based algorithm is comparable to the subspace sampling algorithm when c and r are small; however, our algorithm becomes less efficient when c and r are large. This is due to the following reasons. First, the computational cost of the subspace sampling algorithm is dominated by the truncated SVD of \mathbf{A} , which is determined by the target rank k and the size and sparsity of the data matrix. However, the cost of our algorithm grows with c and r . Thus, our algorithm becomes less efficient when c and r are large. Second, the truncated SVD operation in MATLAB, that is, the ‘svds’ function, gains from sparsity, but our algorithm does not. The four data sets are all very sparse, so the subspace sampling algorithm has advantages. Third, the truncated SVD functions are very well implemented by MATLAB (not in MATLAB language but in Fortran/C). In contrast, our algorithm is implemented in MATLAB language, which is usually less efficient than Fortran/C.

5.2 Comparison among the Nyström Algorithms

In this section we empirically compare our adaptive sampling algorithm (in Theorem 10) with some other sampling algorithms including the subspace sampling of Drineas et al. (2008) and the uniform sampling, both without replacement. We also conduct comparison between the standard Nyström and our modified Nyström, both use the three sampling algorithms to select columns.

We test the algorithms on three data sets which are summarized in Table 5. The experiment setting follows Gittens and Mahoney (2013). For each data set we generate a radial basis function (RBF) kernel matrix \mathbf{A} which is defined by

$$a_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right),$$

where \mathbf{x}_i and \mathbf{x}_j are data instances and σ is a scale parameter. Notice that the RBF kernel is dense in general. We set $\sigma = 0.2$ or 1 in our experiments. For each data set with different settings of σ , we fix a target rank $k = 10, 20$ or 50 and vary c in a very large range. We will discuss the choice

Data Set	#Instances	#Attributes	Source
Abalone	4,177	8	UCI (Frank and Asuncion, 2010)
Wine Quality	4,898	12	UCI (Cortez et al., 2009)
Letters	5,000	16	Statlog (Michie et al., 1994)

	$\ \mathbf{A} - \mathbf{A}_k\ _F / \ \mathbf{A}\ _F$			$\frac{m}{k} \text{std}(\ell^{[k]})$		
	$k = 10$	$k = 20$	$k = 50$	$k = 10$	$k = 20$	$k = 50$
Abalone ($\sigma = 0.2$)	0.4689	0.3144	0.1812	0.8194	0.6717	0.4894
Abalone ($\sigma = 1.0$)	0.0387	0.0122	0.0023	0.5879	0.8415	1.3830
Wine Quality ($\sigma = 0.2$)	0.8463	0.7930	0.7086	1.8703	1.6490	1.3715
Wine Quality ($\sigma = 1.0$)	0.0504	0.0245	0.0084	0.3052	0.5124	0.8067
Letters ($\sigma = 0.2$)	0.9546	0.9324	0.8877	5.4929	3.9346	2.6210
Letters ($\sigma = 1.0$)	0.1254	0.0735	0.0319	0.2481	0.2938	0.3833

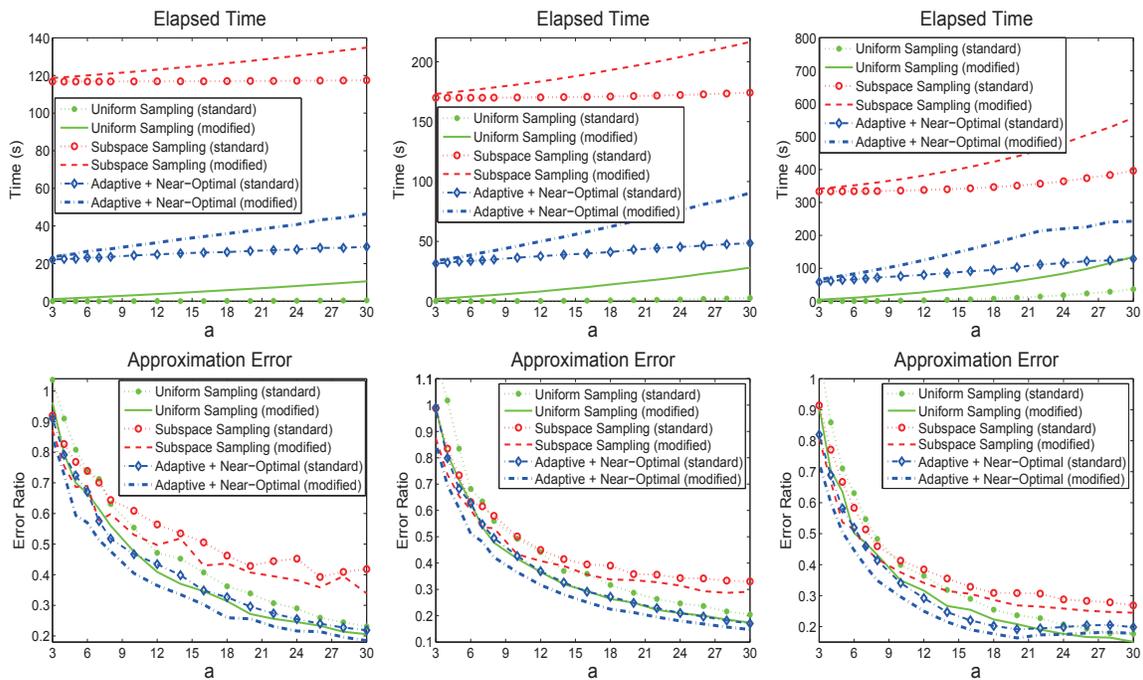
Table 5: A summary of the data sets for the Nyström approximation. In the second tabular $\text{std}(\ell^{[k]})$ denotes the standard deviation of the statistical leverage scores of \mathbf{A} relative to the best rank- k approximation to \mathbf{A} . We use the normalization factor $\frac{m}{k}$ because $\frac{m}{k} \text{mean}(\ell^{[k]}) = 1$.

of σ and k in the following two paragraphs. We run each algorithm for 10 times, and report the the minimum error ratio as well as the total elapsed time of the 10 repeats. The results are shown in Figures 5, 6, and 7.

Table 5 provides useful implications on choosing the target rank k . In Table 5, $\frac{\|\mathbf{A} - \mathbf{A}_k\|_F}{\|\mathbf{A}\|_F}$ denotes ratio that is not captured by the best rank- k approximation to the RBF kernel, and the parameter σ has an influence on the ratio $\|\mathbf{A} - \mathbf{A}_k\|_F / \|\mathbf{A}\|_F$. When σ is large, the RBF kernel can be well approximated by a low-rank matrix, which implies that (i) a small k suffices when σ is large, and (ii) k should be set large when σ is small. So the settings ($\sigma = 1, k = 10$) and ($\sigma = 0.2, k = 50$) are more reasonable than the rest. Let us take the RBF kernel in the Abalone data set as an example. When $\sigma = 1$, the rank-10 approximation well captures the kernel, so k can be safely set as small as 10; when $\sigma = 0.2$, the target rank k should be set large, say larger than 50, otherwise the approximation is rough.

The standard deviation of the leverage scores reflects whether the advanced importance sampling techniques such as the subspace sampling and adaptive sampling are useful. Figures 5, 6, and 7 show that the advantage of the subspace sampling and adaptive sampling over the uniform sampling is significant whenever the standard deviation of the leverage scores is large (see Table 5), and vice versa. Actually, as reflected in Table 5, the parameter σ influences the homogeneity/heterogeneity of the leverage scores. Usually, when σ is small, the leverage scores become heterogeneous, and the effect of choosing “good” columns is significant.

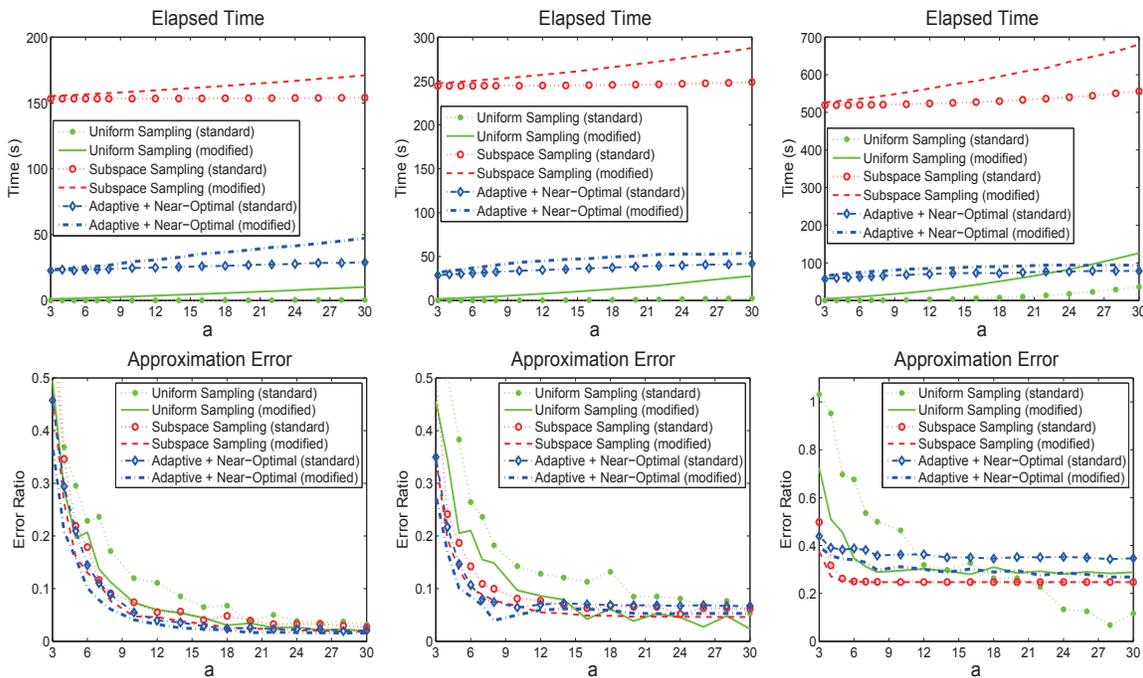
The experimental results also show that the subspace sampling and adaptive sampling algorithms significantly outperform the uniform sampling when c is reasonably small, say $c < 10k$. This indicates that the subspace sampling and adaptive sampling algorithms are good at choosing “good” columns as basis vectors. The effect is especially evident on the RBF kernel with the scale parameter $\sigma = 0.2$, where the leverage scores are heterogeneous. In most cases our adaptive sampling algorithm achieves the lowest approximation error among the three algorithms. The error ratios of our adaptive sampling for the modified Nyström are in accordance with the theoretical bound in



(a) $\sigma = 0.2, k = 10, \text{ and } c = ak.$

(b) $\sigma = 0.2, k = 20, \text{ and } c = ak.$

(c) $\sigma = 0.2, k = 50, \text{ and } c = ak.$

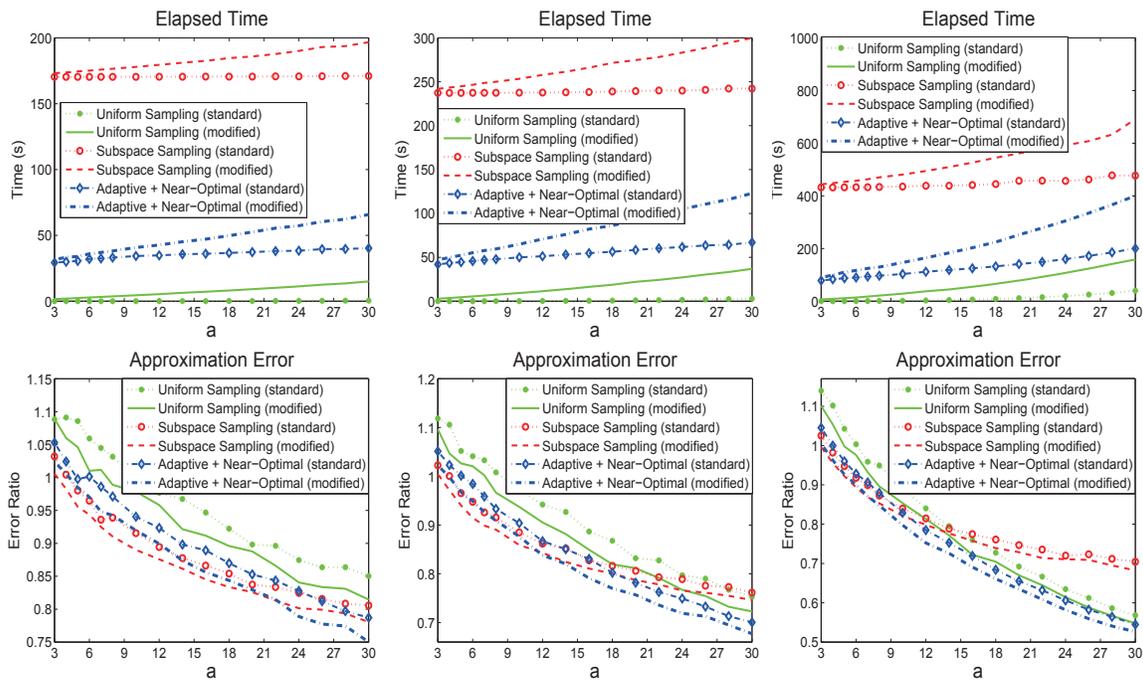


(d) $\sigma = 1, k = 10, \text{ and } c = ak.$

(e) $\sigma = 1, k = 20, \text{ and } c = ak.$

(f) $\sigma = 1, k = 50, \text{ and } c = ak.$

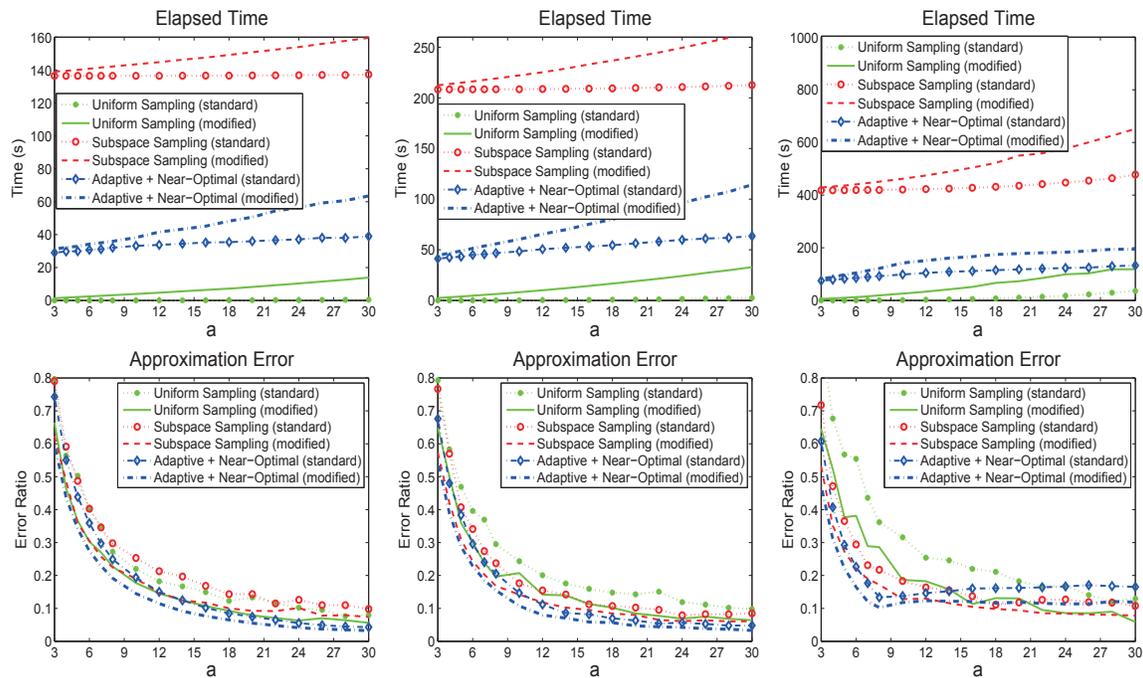
Figure 5: Results of the Nyström algorithms on the RBF kernel in the Abalone data set.



(a) $\sigma = 0.2, k = 10, \text{ and } c = ak.$

(b) $\sigma = 0.2, k = 20, \text{ and } c = ak.$

(c) $\sigma = 0.2, k = 50, \text{ and } c = ak.$



(d) $\sigma = 1, k = 10, \text{ and } c = ak.$

(e) $\sigma = 1, k = 20, \text{ and } c = ak.$

(f) $\sigma = 1, k = 50, \text{ and } c = ak.$

Figure 6: Results of the Nyström algorithms on the RBF kernel in the Wine Quality data set.

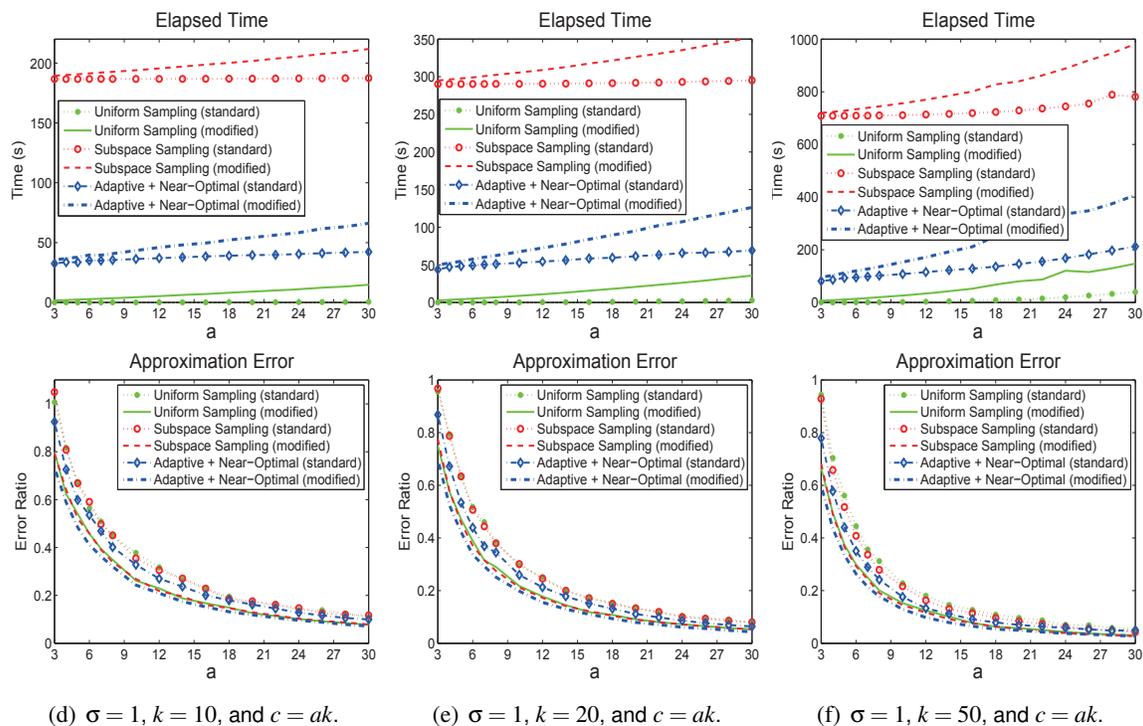
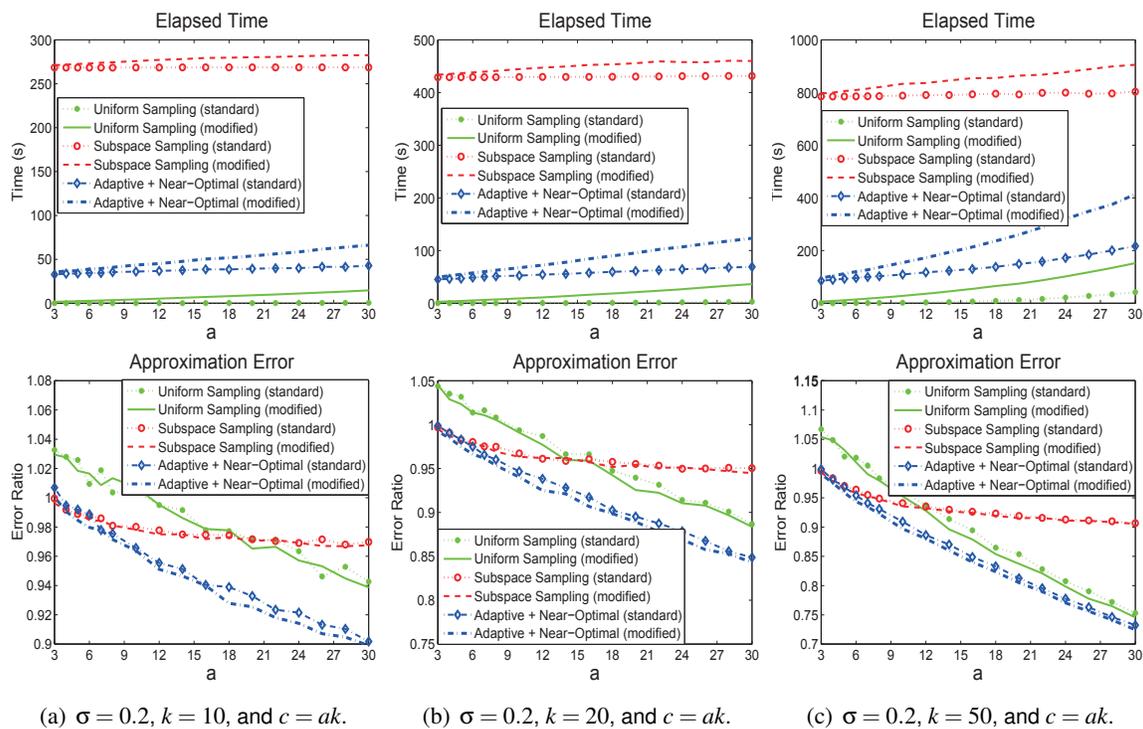


Figure 7: Results of the Nyström algorithms on the RBF kernel in the Letters data set.

Theorem 10; that is,

$$\frac{\|\mathbf{A} - \mathbf{C}\mathbf{U}\mathbf{C}^T\|_F}{\|\mathbf{A} - \mathbf{A}_k\|_F} \leq 1 + \sqrt{\frac{2k}{c}(1 + o(1))} = 1 + \sqrt{\frac{2}{a}(1 + o(1))}.$$

As for the running time, our adaptive sampling algorithm is more efficient than the subspace sampling algorithm. This is partly because the RBF kernel matrix is dense, and hence the subspace sampling algorithm costs $O(m^2k)$ time to compute the truncated SVD.

Furthermore, the experimental results show that using $\mathbf{U} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$ as the intersection matrix (denoted by “modified” in the figures) always leads to much lower error than using $\mathbf{U} = \mathbf{W}^\dagger$ (denoted by “standard”). However, our modified Nyström method costs more time to compute the intersection matrix than the standard Nyström method costs. Recall that the standard Nyström costs $O(c^3)$ time to compute $\mathbf{U} = \mathbf{W}^\dagger$ and that the modified Nyström costs $O(mc^2) + T_{\text{Multiply}}(m^2c)$ time to compute $\mathbf{U} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$. So the users should make a trade-off between time and accuracy and decide whether it is worthwhile to sacrifice extra computational overhead for the improvement in accuracy by using the modified Nyström method.

6. Conclusion

In this paper we have built a novel and more general relative-error bound for the adaptive sampling algorithm. Accordingly, we have devised novel CUR matrix decomposition and Nyström approximation algorithms which demonstrate significant improvement over the classical counterparts. Our relative-error CUR algorithm requires only $c = 2k\epsilon^{-1}(1 + o(1))$ columns and $r = c\epsilon^{-1}(1 + \epsilon)$ rows selected from the original matrix. To achieve relative-error bound, the best previous algorithm—the subspace sampling algorithm—requires $c = O(k\epsilon^{-2} \log k)$ columns and $r = O(c\epsilon^{-2} \log c)$ rows. Our modified Nyström method is different from the conventional Nyström methods in that it uses a different intersection matrix. We have shown that our adaptive sampling algorithm for the modified Nyström achieves relative-error upper bound by sampling only $c = 2k\epsilon^{-2}(1 + o(1))$ columns, which even beats the lower error bounds of the standard Nyström and the ensemble Nyström. Our proposed CUR and Nyström algorithms are scalable because they need only to maintain a small fraction of columns or rows in RAM, and their time complexities are low provided that matrix multiplication can be highly efficiently executed. Finally, the empirical comparison has also demonstrated the effectiveness and efficiency of our algorithms.

Acknowledgments

This work has been supported in part by the Natural Science Foundations of China (No. 61070239) and the Scholarship Award for Excellent Doctoral Student granted by Chinese Ministry of Education.

Appendix A. The Dual Set Sparsification Algorithm

For the sake of self-contained, we attach the dual set sparsification algorithm and describe some implementation details. The deterministic dual set sparsification algorithm is established by Boutsidis et al. (2011) and serves as an important step in the near-optimal column selection algorithm (described in Lemma 2 and Algorithm 1 in this paper). We show the dual set sparsification algorithm

Algorithm 3 Deterministic Dual Set Spectral-Frobenius Sparsification Algorithm.

- 1: **Input:** $\mathcal{U} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^l$, ($l < n$); $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^k \subset \mathbb{R}^k$, with $\sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^T = \mathbf{I}_k$ ($k < n$); $k < r < n$;
- 2: **Initialize:** $\mathbf{s}_0 = \mathbf{0}$, $\mathbf{A}_0 = \mathbf{0}$;
- 3: Compute $\|\mathbf{x}_i\|_2^2$ for $i = 1, \dots, n$, and then compute $\delta_U = \frac{\sum_{i=1}^n \|\mathbf{x}_i\|_2^2}{1 - \sqrt{k/r}}$;
- 4: **for** $\tau = 0$ to $r - 1$ **do**
- 5: Compute the eigenvalue decomposition of \mathbf{A}_τ ;
- 6: Find any index j in $\{1, \dots, n\}$ and compute a weight $t > 0$ such that

$$\delta_U^{-1} \|\mathbf{x}_j\|_2^2 \leq t^{-1} \leq \frac{\mathbf{v}_j^T (\mathbf{A}_\tau - (L_\tau + 1)\mathbf{I}_k)^{-2} \mathbf{v}_j}{\phi(L_\tau + 1, \mathbf{A}_\tau) - \phi(L_\tau, \mathbf{A}_\tau)} - \mathbf{v}_j^T (\mathbf{A}_\tau - (L_\tau + 1)\mathbf{I}_k)^{-1} \mathbf{v}_j;$$

where

$$\phi(L, \mathbf{A}) = \sum_{i=1}^k (\lambda_i(\mathbf{A}) - L)^{-1}, \quad L_\tau = \tau - \sqrt{rk};$$

- 7: Update the j -th component of \mathbf{s}_τ and \mathbf{A}_τ : $\mathbf{s}_{\tau+1}[j] = \mathbf{s}_\tau[j] + t$, $\mathbf{A}_{\tau+1} = \mathbf{A}_\tau + t \mathbf{v}_j \mathbf{v}_j^T$;
 - 8: **end for**
 - 9: **return** $\mathbf{s} = \frac{1 - \sqrt{k/r}}{r} \mathbf{s}_r$.
-

algorithm in Algorithm 3 and its bounds in Lemma 14, and we also analyze the time complexity using our defined notation.

Lemma 14 (Dual Set Spectral-Frobenius Sparsification) *Let $\mathcal{U} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^l$ ($l < n$) contain the columns of an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{l \times n}$. Let $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_k\} \subset \mathbb{R}^k$ ($k < n$) be a decomposition of the identity, that is, $\sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^T = \mathbf{I}_k$. Given an integer r with $k < r < n$, Algorithm 3 deterministically computes a set of weights $s_i \geq 0$ ($i = 1, \dots, n$) at most r of which are non-zero, such that*

$$\lambda_k \left(\sum_{i=1}^n s_i \mathbf{v}_i \mathbf{v}_i^T \right) \geq \left(1 - \sqrt{\frac{k}{r}} \right)^2 \quad \text{and} \quad \text{tr} \left(\sum_{i=1}^n s_i \mathbf{x}_i \mathbf{x}_i^T \right) \leq \|\mathbf{X}\|_F^2.$$

The weights s_i can be computed deterministically in $O(rnk^2) + T_{\text{Multiply}}(nl)$ time.

Here we mention some implementation issues of Algorithm 3 which were not described in detail by Boutsidis et al. (2011). In each iteration the algorithm performs once eigenvalue decomposition: $\mathbf{A}_\tau = \mathbf{W} \mathbf{\Lambda} \mathbf{W}^T$. Here \mathbf{A}_τ is guaranteed to be SPSD in each iteration. Since

$$(\mathbf{A}_\tau - \alpha \mathbf{I}_k)^q = \mathbf{W} \text{Diag} \left((\lambda_1 - \alpha)^q, \dots, (\lambda_k - \alpha)^q \right) \mathbf{W}^T,$$

$(\mathbf{A}_\tau - (L_\tau + 1)\mathbf{I}_k)^q$ can be efficiently computed based on the eigenvalue decomposition of \mathbf{A}_τ . With the eigenvalues at hand, $\phi(L, \mathbf{A}_\tau)$ can also be computed directly.

The algorithm runs in r iterations. In each iteration, the eigenvalue decomposition of \mathbf{A}_τ requires $O(k^3)$, and the n comparisons in Line 6 each requires $O(k^2)$. Moreover, computing $\|\mathbf{x}_i\|_2^2$ for each \mathbf{x}_i requires $T_{\text{Multiply}}(nl)$. Overall, the running time of Algorithm 3 is at most $O(rk^3) + O(rnk^2) + T_{\text{Multiply}}(nl) = O(rnk^2) + T_{\text{Multiply}}(nl)$.

The near-optimal column selection algorithm described in Lemma 2 has three steps: randomized SVD via random projection which costs $O(mk^2 \varepsilon^{-4/3}) + T_{\text{Multiply}}(mnk \varepsilon^{-2/3})$ time, the dual

set sparsification algorithm which costs $O(nk^3\varepsilon^{-2/3}) + T_{\text{Multiply}}(mn)$ time, and the adaptive sampling algorithm which costs $O(mk^2\varepsilon^{-4/3}) + T_{\text{Multiply}}(mnk\varepsilon^{-2/3})$ time. Therefore, the near-optimal column selection algorithm costs totally $O(mk^2\varepsilon^{-4/3} + nk^3\varepsilon^{-2/3}) + T_{\text{Multiply}}(mnk\varepsilon^{-2/3})$ time.

Appendix B. Proofs of the Adaptive Sampling Bounds

We present the proofs of Theorem 5, Corollary 7, Theorem 8, and Theorem 10 in Appendices B.1, B.2, B.3, and B.4, respectively.

B.1 The Proof of Theorem 5

Theorem 5 can be equivalently expressed in Theorem 15. In order to stick to the column space convention throughout this paper, we prove Theorem 15 instead of Theorem 5.

Theorem 15 (The Adaptive Sampling Algorithm) *Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a matrix $\mathbf{R} \in \mathbb{R}^{r \times n}$ such that $\text{rank}(\mathbf{R}) = \text{rank}(\mathbf{A}\mathbf{R}^\dagger\mathbf{R}) = \rho$ ($\rho \leq r \leq m$), let $\mathbf{C}_1 \in \mathbb{R}^{m \times c_1}$ consist of c_1 columns of \mathbf{A} , and define the residual $\mathbf{B} = \mathbf{A} - \mathbf{C}_1\mathbf{C}_1^\dagger\mathbf{A}$. For $i = 1, \dots, n$, let*

$$p_i = \|\mathbf{b}_i\|_2^2 / \|\mathbf{B}\|_F^2,$$

where \mathbf{b}_i is the i -th column of the matrix \mathbf{B} . Sample further c_2 columns from \mathbf{A} in c_2 i.i.d. trials, where in each trial the i -th column is chosen with probability p_i . Let $\mathbf{C}_2 \in \mathbb{R}^{m \times c_2}$ contain the c_2 sampled columns and $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2] \in \mathbb{R}^{m \times (c_1+c_2)}$ contain the columns of both \mathbf{C}_1 and \mathbf{C}_2 , all of which are columns of \mathbf{A} . Then the following inequality holds:

$$\mathbb{E}\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger\mathbf{A}\mathbf{R}^\dagger\mathbf{R}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}\mathbf{R}^\dagger\mathbf{R}\|_F^2 + \frac{\rho}{c_2}\|\mathbf{A} - \mathbf{C}_1\mathbf{C}_1^\dagger\mathbf{A}\|_F^2.$$

where the expectation is taken w.r.t. \mathbf{C}_2 .

Proof With a little abuse of symbols, we use bold uppercase letters to denote random matrices and bold lowercase to denote random vectors, without distinguishing between random matrices/vectors and non-random matrices/vectors.

We denote the j -th column of $\mathbf{V}_{\mathbf{A}\mathbf{R}^\dagger\mathbf{R}, \rho} \in \mathbb{R}^{n \times \rho}$ as \mathbf{v}_j , and the (i, j) -th entry of $\mathbf{V}_{\mathbf{A}\mathbf{R}^\dagger\mathbf{R}, \rho}$ as v_{ij} . Define random vectors $\mathbf{x}_{j,(l)} \in \mathbb{R}^m$ such that for $j = 1, \dots, n$ and $l = 1, \dots, c_2$,

$$\mathbf{x}_{j,(l)} = \frac{v_{ij}}{p_i}\mathbf{b}_i = \frac{v_{ij}}{p_i}(\mathbf{a}_i - \mathbf{C}_1\mathbf{C}_1^\dagger\mathbf{a}_i) \quad \text{with probability } p_i, \quad \text{for } i = 1, \dots, n,$$

Notice that $\mathbf{x}_{j,(l)}$ is a linear function of a column of \mathbf{A} sampled from the above defined distribution. We have that

$$\begin{aligned} \mathbb{E}[\mathbf{x}_{j,(l)}] &= \sum_{i=1}^n p_i \frac{v_{ij}}{p_i} \mathbf{b}_i = \mathbf{B}\mathbf{v}_j, \\ \mathbb{E}\|\mathbf{x}_{j,(l)}\|_2^2 &= \sum_{i=1}^n p_i \frac{v_{ij}^2}{p_i^2} \|\mathbf{b}_i\|_2^2 = \sum_{i=1}^n \frac{v_{ij}^2}{\|\mathbf{b}_i\|_2^2 / \|\mathbf{B}\|_F^2} \|\mathbf{b}_i\|_2^2 = \|\mathbf{B}\|_F^2. \end{aligned}$$

Then we let $\mathbf{x}_j = \frac{1}{c_2} \sum_{l=1}^{c_2} \mathbf{x}_{j,(l)}$, we have

$$\begin{aligned} \mathbb{E}[\mathbf{x}_j] &= \mathbb{E}[\mathbf{x}_{j,(l)}] = \mathbf{B}\mathbf{v}_j, \\ \mathbb{E}\|\mathbf{x}_j - \mathbf{B}\mathbf{v}_j\|_2^2 &= \mathbb{E}\left\|\mathbf{x}_j - \mathbb{E}[\mathbf{x}_j]\right\|_2^2 = \frac{1}{c_2} \mathbb{E}\left\|\mathbf{x}_{j,(l)} - \mathbb{E}[\mathbf{x}_{j,(l)}]\right\|_2^2 = \frac{1}{c_2} \mathbb{E}\|\mathbf{x}_{j,(l)} - \mathbf{B}\mathbf{v}_j\|_2^2. \end{aligned}$$

According to the construction of $\mathbf{x}_1, \dots, \mathbf{x}_\rho$, we define the c_2 columns of \mathbf{A} to be $\mathbf{C}_2 \in \mathbb{R}^{m \times c_2}$. Note that all the random vectors $\mathbf{x}_1, \dots, \mathbf{x}_\rho$ lie in the subspace $\text{span}(\mathbf{C}_1) + \text{span}(\mathbf{C}_2)$. We define random vectors

$$\mathbf{w}_j = \mathbf{C}_1 \mathbf{C}_1^\dagger \mathbf{A} \mathbf{R}^\dagger \mathbf{R} \mathbf{v}_j + \mathbf{x}_j = \mathbf{C}_1 \mathbf{C}_1^\dagger \mathbf{A} \mathbf{v}_j + \mathbf{x}_j, \quad \text{for } j = 1, \dots, \rho,$$

where the second equality follows from Lemma 16; that is, $\mathbf{A} \mathbf{R}^\dagger \mathbf{R} \mathbf{v}_j = \mathbf{A} \mathbf{v}_j$ if \mathbf{v}_j is one of the top ρ right singular vectors of $\mathbf{A} \mathbf{R}^\dagger \mathbf{R}$. Then we have that any set of random vectors $\{\mathbf{w}_1, \dots, \mathbf{w}_\rho\}$ lies in $\text{span}(\mathbf{C}) = \text{span}(\mathbf{C}_1) + \text{span}(\mathbf{C}_2)$. Let $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_\rho]$ be a random matrix, we have that $\text{span}(\mathbf{W}) \subset \text{span}(\mathbf{C})$. The expectation of \mathbf{w}_j is

$$\mathbb{E}[\mathbf{w}_j] = \mathbf{C}_1 \mathbf{C}_1^\dagger \mathbf{A} \mathbf{v}_j + \mathbb{E}[\mathbf{x}_j] = \mathbf{C}_1 \mathbf{C}_1^\dagger \mathbf{A} \mathbf{v}_j + \mathbf{B}\mathbf{v}_j = \mathbf{A} \mathbf{v}_j,$$

therefore we have that

$$\mathbf{w}_j - \mathbf{A} \mathbf{v}_j = \mathbf{x}_j - \mathbf{B}\mathbf{v}_j.$$

The expectation of $\|\mathbf{w}_j - \mathbf{A} \mathbf{v}_j\|_2^2$ is

$$\begin{aligned} \mathbb{E}\|\mathbf{w}_j - \mathbf{A} \mathbf{v}_j\|_2^2 &= \mathbb{E}\|\mathbf{x}_j - \mathbf{B}\mathbf{v}_j\|_2^2 = \frac{1}{c_2} \mathbb{E}\|\mathbf{x}_{j,(l)} - \mathbf{B}\mathbf{v}_j\|_2^2 \\ &= \frac{1}{c_2} \mathbb{E}\|\mathbf{x}_{j,(l)}\|_2^2 - \frac{2}{c_2} (\mathbf{B}\mathbf{v}_j)^T \mathbb{E}[\mathbf{x}_{j,(l)}] + \frac{1}{c_2} \|\mathbf{B}\mathbf{v}_j\|_2^2 \\ &= \frac{1}{c_2} \mathbb{E}\|\mathbf{x}_{j,(l)}\|_2^2 - \frac{1}{c_2} \|\mathbf{B}\mathbf{v}_j\|_2^2 = \frac{1}{c_2} \|\mathbf{B}\|_F^2 - \frac{1}{c_2} \|\mathbf{B}\mathbf{v}_j\|_2^2 \\ &\leq \frac{1}{c_2} \|\mathbf{B}\|_F^2. \end{aligned} \tag{5}$$

To complete the proof, we denote

$$\mathbf{F} = \left(\sum_{q=1}^{\rho} \sigma_q^{-1} \mathbf{w}_q \mathbf{u}_q^T \right) \mathbf{A} \mathbf{R}^\dagger \mathbf{R},$$

where σ_q is the q -th largest singular value of $\mathbf{A} \mathbf{R}^\dagger \mathbf{R}$ and \mathbf{u}_q is the corresponding left singular vector of $\mathbf{A} \mathbf{R}^\dagger \mathbf{R}$. The column space of \mathbf{F} is contained in $\text{span}(\mathbf{W}) (\subset \text{span}(\mathbf{C}))$, and thus

$$\|\mathbf{A} \mathbf{R}^\dagger \mathbf{R} - \mathbf{C} \mathbf{C}^\dagger \mathbf{A} \mathbf{R}^\dagger \mathbf{R}\|_F^2 \leq \|\mathbf{A} \mathbf{R}^\dagger \mathbf{R} - \mathbf{W} \mathbf{W}^\dagger \mathbf{A} \mathbf{R}^\dagger \mathbf{R}\|_F^2 \leq \|\mathbf{A} \mathbf{R}^\dagger \mathbf{R} - \mathbf{F}\|_F^2.$$

We use \mathbf{F} to bound the error $\|\mathbf{A} \mathbf{R}^\dagger \mathbf{R} - \mathbf{C} \mathbf{C}^\dagger \mathbf{A} \mathbf{R}^\dagger \mathbf{R}\|_F^2$. That is,

$$\begin{aligned} \mathbb{E}\|\mathbf{A} - \mathbf{C} \mathbf{C}^\dagger \mathbf{A} \mathbf{R}^\dagger \mathbf{R}\|_F^2 &= \mathbb{E}\|\mathbf{A} - \mathbf{A} \mathbf{R}^\dagger \mathbf{R} + \mathbf{A} \mathbf{R}^\dagger \mathbf{R} - \mathbf{C} \mathbf{C}^\dagger \mathbf{A} \mathbf{R}^\dagger \mathbf{R}\|_F^2 \\ &= \mathbb{E}\left[\|\mathbf{A} - \mathbf{A} \mathbf{R}^\dagger \mathbf{R}\|_F^2 + \|\mathbf{A} \mathbf{R}^\dagger \mathbf{R} - \mathbf{C} \mathbf{C}^\dagger \mathbf{A} \mathbf{R}^\dagger \mathbf{R}\|_F^2 \right] \\ &\leq \|\mathbf{A} - \mathbf{A} \mathbf{R}^\dagger \mathbf{R}\|_F^2 + \mathbb{E}\|\mathbf{A} \mathbf{R}^\dagger \mathbf{R} - \mathbf{F}\|_F^2, \end{aligned} \tag{6}$$

where (6) is due to that $\mathbf{A}(\mathbf{I} - \mathbf{R}^\dagger \mathbf{R})$ is orthogonal to $(\mathbf{I} - \mathbf{C}\mathbf{C}^\dagger)\mathbf{A}\mathbf{R}^\dagger \mathbf{R}$. Since $\mathbf{A}\mathbf{R}^\dagger \mathbf{R}$ and \mathbf{F} both lie on the space spanned by the right singular vectors of $\mathbf{A}\mathbf{R}^\dagger \mathbf{R}$ (i.e., $\{\mathbf{v}_j\}_{j=1}^\rho$), we decompose $\mathbf{A}\mathbf{R}^\dagger \mathbf{R} - \mathbf{F}$ along $\{\mathbf{v}_j\}_{j=1}^\rho$, obtaining that

$$\begin{aligned} \mathbb{E}\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}\mathbf{R}^\dagger \mathbf{R}\|_F^2 &\leq \|\mathbf{A} - \mathbf{A}\mathbf{R}^\dagger \mathbf{R}\|_F^2 + \mathbb{E}\|\mathbf{A}\mathbf{R}^\dagger \mathbf{R} - \mathbf{F}\|_F^2, \\ &= \|\mathbf{A} - \mathbf{A}\mathbf{R}^\dagger \mathbf{R}\|_F^2 + \sum_{j=1}^\rho \mathbb{E}\left\|\left(\mathbf{A}\mathbf{R}^\dagger \mathbf{R} - \mathbf{F}\right)\mathbf{v}_j\right\|_2^2 \\ &= \|\mathbf{A} - \mathbf{A}\mathbf{R}^\dagger \mathbf{R}\|_F^2 + \sum_{j=1}^\rho \mathbb{E}\left\|\mathbf{A}\mathbf{R}^\dagger \mathbf{R}\mathbf{v}_j - \left(\sum_{q=1}^\rho \sigma_q^{-1} \mathbf{w}_q \mathbf{u}_q^T\right)\sigma_j \mathbf{u}_j\right\|_2^2 \\ &= \|\mathbf{A} - \mathbf{A}\mathbf{R}^\dagger \mathbf{R}\|_F^2 + \sum_{j=1}^\rho \mathbb{E}\left\|\mathbf{A}\mathbf{R}^\dagger \mathbf{R}\mathbf{v}_j - \mathbf{w}_j\right\|_2^2 \\ &= \|\mathbf{A} - \mathbf{A}\mathbf{R}^\dagger \mathbf{R}\|_F^2 + \sum_{j=1}^\rho \mathbb{E}\|\mathbf{A}\mathbf{v}_j - \mathbf{w}_j\|_2^2 \end{aligned} \quad (7)$$

$$\leq \|\mathbf{A} - \mathbf{A}\mathbf{R}^\dagger \mathbf{R}\|_F^2 + \frac{\rho}{c_2} \|\mathbf{B}\|_F^2, \quad (8)$$

where (7) follows from Lemma 16 and (8) follows from (5). \blacksquare

Lemma 16 *We are given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a matrix $\mathbf{R} \in \mathbb{R}^{r \times n}$ such that $\text{rank}(\mathbf{A}\mathbf{R}^\dagger \mathbf{R}) = \text{rank}(\mathbf{R}) = \rho$ ($\rho \leq r \leq m$). Letting $\mathbf{v}_j \in \mathbb{R}^n$ be the j -th top right singular vector of $\mathbf{A}\mathbf{R}^\dagger \mathbf{R}$, we have that*

$$\mathbf{A}\mathbf{R}^\dagger \mathbf{R}\mathbf{v}_j = \mathbf{A}\mathbf{v}_j, \quad \text{for } j = 1, \dots, \rho.$$

Proof First let $\mathbf{V}_{\mathbf{R},\rho} \in \mathbb{R}^{n \times \rho}$ contain the top ρ right singular vectors of \mathbf{R} . Then the projection of \mathbf{A} onto the row space of \mathbf{R} is $\mathbf{A}\mathbf{R}^\dagger \mathbf{R} = \mathbf{A}\mathbf{V}_{\mathbf{R},\rho} \mathbf{V}_{\mathbf{R},\rho}^T$. Let the thin SVD of $\mathbf{A}\mathbf{V}_{\mathbf{R},\rho} \in \mathbb{R}^{m \times \rho}$ be $\tilde{\mathbf{U}}\tilde{\Sigma}\tilde{\mathbf{V}}^T$, where $\tilde{\mathbf{V}} \in \mathbb{R}^{\rho \times \rho}$. Then the compact SVD of $\mathbf{A}\mathbf{R}^\dagger \mathbf{R}$ is

$$\mathbf{A}\mathbf{R}^\dagger \mathbf{R} = \mathbf{A}\mathbf{V}_{\mathbf{R},\rho} \mathbf{V}_{\mathbf{R},\rho}^T = \tilde{\mathbf{U}}\tilde{\Sigma}\tilde{\mathbf{V}}^T \mathbf{V}_{\mathbf{R},\rho}^T.$$

According to the definition, \mathbf{v}_j is the j -th column of $(\mathbf{V}_{\mathbf{R},\rho} \tilde{\mathbf{V}}) \in \mathbb{R}^{n \times \rho}$. Thus \mathbf{v}_j lies on the column space of $\mathbf{V}_{\mathbf{R},\rho}$, and \mathbf{v}_j is orthogonal to $\mathbf{V}_{\mathbf{R},\rho}^\perp$. Finally, since $\mathbf{A} - \mathbf{A}\mathbf{R}^\dagger \mathbf{R} = \mathbf{A}\mathbf{V}_{\mathbf{R},\rho}^\perp \mathbf{V}_{\mathbf{R},\rho}^T$, we have that \mathbf{v}_j is orthogonal to $\mathbf{A} - \mathbf{A}\mathbf{R}^\dagger \mathbf{R}$, that is, $(\mathbf{A} - \mathbf{A}\mathbf{R}^\dagger \mathbf{R})\mathbf{v}_j = \mathbf{0}$, which directly proves the lemma. \blacksquare

B.2 The Proof of Corollary 7

Since \mathbf{C} is constructed by columns of \mathbf{A} and the column space of \mathbf{C} is contained in the column space of \mathbf{A} , we have $\text{rank}(\mathbf{C}\mathbf{C}^\dagger \mathbf{A}) = \text{rank}(\mathbf{C}) = \rho \leq c$. Consequently, the assumptions of Theorem 5 are satisfied. The assumptions in turn imply

$$\begin{aligned} \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}\|_F &\leq (1 + \varepsilon) \|\mathbf{A} - \mathbf{A}_k\|_F, \\ \|\mathbf{A} - \mathbf{A}\mathbf{R}_1^\dagger \mathbf{R}_1\|_F &\leq (1 + \varepsilon) \|\mathbf{A} - \mathbf{A}_k\|_F, \end{aligned}$$

and $c/r_2 = \varepsilon$. It then follows from Theorem 5 that

$$\begin{aligned} \mathbb{E}_{\mathbf{R}} \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}\mathbf{R}^\dagger \mathbf{R}\|_F^2 &= \mathbb{E}_{\mathbf{R}_1} \left[\mathbb{E}_{\mathbf{R}_2} \left[\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}\mathbf{R}^\dagger \mathbf{R}\|_F^2 \middle| \mathbf{R}_1 \right] \right] \\ &\leq \mathbb{E}_{\mathbf{R}_1} \left[\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}\|_F^2 + \frac{\rho}{r_2} \|\mathbf{A} - \mathbf{A}\mathbf{R}_1^\dagger \mathbf{R}_1\|_F^2 \right] \\ &\leq \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}\|_F^2 + \frac{c}{r_2} (1 + \varepsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2 \\ &= \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}\|_F^2 + \varepsilon(1 + \varepsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2. \end{aligned}$$

Furthermore, we have that

$$\begin{aligned} \left[\mathbb{E} \|\mathbf{A} - \mathbf{C}\mathbf{U}\mathbf{R}\|_F \right]^2 &\leq \mathbb{E} \|\mathbf{A} - \mathbf{C}\mathbf{U}\mathbf{R}\|_F^2 = \mathbb{E}_{\mathbf{C}, \mathbf{R}} \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}\mathbf{R}^\dagger \mathbf{R}\|_F^2 \\ &= \mathbb{E}_{\mathbf{C}} \left[\mathbb{E}_{\mathbf{R}} \left[\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}\mathbf{R}^\dagger \mathbf{R}\|_F^2 \middle| \mathbf{C} \right] \right] \\ &\leq \mathbb{E}_{\mathbf{C}} \left[\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}\|_F^2 + \varepsilon(1 + \varepsilon) \|\mathbf{A} - \mathbf{A}_k\|_F^2 \right] \\ &\leq (1 + \varepsilon)^2 \|\mathbf{A} - \mathbf{A}_k\|_k^2, \end{aligned}$$

which yields the error bound for CUR matrix decomposition.

When the matrix \mathbf{A} is symmetric, the matrix \mathbf{C}_1^T consists of the rows \mathbf{A} , and thus we can use Theorem 15 (which is identical to Theorem 5) to prove the error bound for the Nyström approximation. By replacing \mathbf{R} in Theorem 15 by \mathbf{C}_1^T , we have that

$$\begin{aligned} \mathbb{E} \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}(\mathbf{C}_1^\dagger)^T \mathbf{C}_1^T\|_F^2 &\leq \|\mathbf{A} - \mathbf{A}(\mathbf{C}_1^\dagger)^T \mathbf{C}_1^T\|_F^2 + \frac{c_1}{c_2} \|\mathbf{A} - \mathbf{C}_1 \mathbf{C}_1^\dagger \mathbf{A}\|_F^2 \\ &= \left(1 + \frac{c_1}{c_2} \right) \|\mathbf{A} - \mathbf{C}_1 \mathbf{C}_1^\dagger \mathbf{A}\|_F^2, \end{aligned}$$

where the expectation is taken w.r.t. \mathbf{C}_2 . Together with the inequality

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}(\mathbf{C}^\dagger)^T \mathbf{C}^T\|_F^2 \leq \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}(\mathbf{C}_1^\dagger)^T \mathbf{C}_1^T\|_F^2$$

given by Lemma 17, we have that

$$\begin{aligned} \mathbb{E}_{\mathbf{C}_1, \mathbf{C}_2} \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}(\mathbf{C}^\dagger)^T \mathbf{C}^T\|_F^2 &\leq \mathbb{E}_{\mathbf{C}_1, \mathbf{C}_2} \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}(\mathbf{C}_1^\dagger)^T \mathbf{C}_1^T\|_F^2 \\ &= \left(1 + \frac{c_1}{c_2} \right) \mathbb{E}_{\mathbf{C}_1} \|\mathbf{A} - \mathbf{C}_1 \mathbf{C}_1^\dagger \mathbf{A}\|_F^2 \\ &= (1 + \varepsilon)^2 \|\mathbf{A} - \mathbf{A}_k\|_F^2. \end{aligned}$$

Hence $\mathbb{E} \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}(\mathbf{C}^\dagger)^T \mathbf{C}^T\|_F \leq \left[\mathbb{E} \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}(\mathbf{C}^\dagger)^T \mathbf{C}^T\|_F^2 \right]^{-\frac{1}{2}} \leq (1 + \varepsilon) \|\mathbf{A} - \mathbf{A}_k\|_F$.

Lemma 17 *Given an $m \times m$ matrix \mathbf{A} and an $m \times c$ matrix $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2]$, the following inequality holds:*

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}(\mathbf{C}^\dagger)^T \mathbf{C}^T\|_F^2 \leq \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}(\mathbf{C}_1^\dagger)^T \mathbf{C}_1^T\|_F^2.$$

Proof Let $\mathcal{P}_C \mathbf{A} = \mathbf{C} \mathbf{C}^\dagger \mathbf{A}$ denote the projection of \mathbf{A} onto the column space of \mathbf{C} , and $\bar{\mathcal{P}}_C = \mathbf{I}_m - \mathbf{C} \mathbf{C}^\dagger$ denote the projector onto the space orthogonal to the column space of \mathbf{C} . It has been shown by Halko et al. (2011) that, for any matrix \mathbf{A} , if $\text{span}(\mathbf{M}) \subset \text{span}(\mathbf{N})$, then the following inequalities hold:

$$\|\mathcal{P}_M \mathbf{A}\|_\xi \leq \|\mathcal{P}_N \mathbf{A}\|_\xi \quad \text{and} \quad \|\bar{\mathcal{P}}_M \mathbf{A}\|_\xi \geq \|\bar{\mathcal{P}}_N \mathbf{A}\|_\xi.$$

Accordingly, $\mathbf{A} \mathcal{P}_{\mathbf{R}^T}^T = \mathbf{A} \mathbf{R}^\dagger \mathbf{R}$ is the projection of \mathbf{A} onto the row space of $\mathbf{R} \in \mathbb{R}^{r \times n}$. We further have that

$$\begin{aligned} \|\mathbf{A} - \mathcal{P}_C \mathbf{A} \mathcal{P}_C^T\|_F^2 &= \|\mathbf{A} - \mathcal{P}_C \mathbf{A} + \mathcal{P}_C \mathbf{A} - \mathcal{P}_C \mathbf{A} \mathcal{P}_C^T\|_F^2 \\ &= \|\bar{\mathcal{P}}_C \mathbf{A} + \mathcal{P}_C \mathbf{A} \bar{\mathcal{P}}_C^T\|_F^2 = \|\bar{\mathcal{P}}_C \mathbf{A}\|_F^2 + \|\mathcal{P}_C \mathbf{A} \bar{\mathcal{P}}_C^T\|_F^2 \end{aligned}$$

and

$$\begin{aligned} \|\mathbf{A} - \mathcal{P}_C \mathbf{A} \mathcal{P}_{\mathbf{C}_1}^T\|_F^2 &= \|\mathbf{A} - \mathcal{P}_C \mathbf{A} + \mathcal{P}_C \mathbf{A} - \mathcal{P}_C \mathbf{A} \mathcal{P}_{\mathbf{C}_1}^T\|_F^2 \\ &= \|\bar{\mathcal{P}}_C \mathbf{A} + \mathcal{P}_C \mathbf{A} \bar{\mathcal{P}}_{\mathbf{C}_1}^T\|_F^2 = \|\bar{\mathcal{P}}_C \mathbf{A}\|_F^2 + \|\mathcal{P}_C \mathbf{A} \bar{\mathcal{P}}_{\mathbf{C}_1}^T\|_F^2, \end{aligned}$$

where the last equalities follow from $\mathcal{P}_C \perp \bar{\mathcal{P}}_C$. Since $\text{span}(\mathbf{C}_1) \subset \text{span}(\mathbf{C})$, we have $\|\mathcal{P}_C \mathbf{A} \bar{\mathcal{P}}_{\mathbf{C}_1}^T\|_F^2 \geq \|\mathcal{P}_C \mathbf{A} \bar{\mathcal{P}}_C^T\|_F^2$, which proves the lemma. \blacksquare

B.3 The Proof of Theorem 8

The error bound follows directly from Lemma 2 and Corollary 7. The near-optimal column selection algorithm costs $O(mk^2 \varepsilon^{-4/3} + nk^3 \varepsilon^{-2/3}) + T_{\text{Multiply}}(mnk \varepsilon^{-2/3})$ time to construct \mathbf{C} and $O(nk^2 \varepsilon^{-4/3} + mk^3 \varepsilon^{-2/3}) + T_{\text{Multiply}}(mnk \varepsilon^{-2/3})$ time to construct \mathbf{R}_1 . Then the adaptive sampling algorithm costs $O(nk^2 \varepsilon^{-2}) + T_{\text{Multiply}}(mnk \varepsilon^{-1})$ time to construct \mathbf{R}_2 . Computing the Moore-Penrose inverses of \mathbf{C} and \mathbf{R} costs $O(mc^2) + O(nr^2) = O(mk^2 \varepsilon^{-2} + nk^2 \varepsilon^{-4})$ time. The multiplication of $\mathbf{C}^\dagger \mathbf{A} \mathbf{R}^\dagger$ costs $T_{\text{Multiply}}(mnc) = T_{\text{Multiply}}(mnk \varepsilon^{-1})$ time. So the total time complexity is $O((m+n)k^3 \varepsilon^{-2/3} + mk^2 \varepsilon^{-2} + nk^2 \varepsilon^{-4}) + T_{\text{Multiply}}(mnk \varepsilon^{-1})$.

B.4 The Proof of Theorem 10

The error bound follows immediately from Lemma 2 and Corollary 7. The near-optimal column selection algorithm costs $O(mk^2 \varepsilon^{-4/3} + mk^3 \varepsilon^{-2/3}) + T_{\text{Multiply}}(m^2 k \varepsilon^{-2/3})$ time to select $c_1 = O(k \varepsilon^{-1})$ columns of \mathbf{A} construct \mathbf{C}_1 . Then the adaptive sampling algorithm costs $O(mk^2 \varepsilon^{-2}) + T_{\text{Multiply}}(m^2 k \varepsilon^{-1})$ time to select $c_2 = O(k \varepsilon^{-2})$ columns construct \mathbf{C}_2 . Finally it costs $O(mc^2) + T_{\text{Multiply}}(m^2 c) = O(mk^2 \varepsilon^{-4}) + T_{\text{Multiply}}(m^2 k \varepsilon^{-2})$ time to construct the intersection matrix $\mathbf{U} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$. So the total time complexity is $O(mk^2 \varepsilon^{-4} + mk^3 \varepsilon^{-2/3}) + T_{\text{Multiply}}(m^2 k \varepsilon^{-2})$.

Appendix C. Proofs of the Lower Error Bounds

In Appendix C.1 we construct two adversarial cases which will be used throughout this appendix. In Appendix C.2 we prove the lower bounds of the standard Nyström method. In Appendix C.3 we prove the lower bounds of the ensemble Nyström method. Theorems 20, 21, 22, 24, and 25 are used for proving Theorem 12.

C.1 Construction of the Adversarial Cases

We now consider the construction of adversarial cases for the spectral norm bounds and the Frobenius norm and nuclear norm bounds, respectively.

C.1.1 THE ADVERSARIAL CASE FOR THE SPECTRAL NORM BOUND

We construct an $m \times m$ positive definite matrix \mathbf{B} as follows:

$$\mathbf{B} = (1 - \alpha)\mathbf{I}_m + \alpha\mathbf{1}_m\mathbf{1}_m^T = \begin{bmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{W} & \mathbf{B}_{21}^T \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}, \quad (9)$$

where $\alpha \in [0, 1)$. It is easy to verify $\mathbf{x}^T \mathbf{B} \mathbf{x} > 0$ for any nonzero $\mathbf{x} \in \mathbb{R}^m$. We show some properties of \mathbf{B} in Lemma 18.

Lemma 18 *Let \mathbf{B}_k be the best rank- k approximation to the matrix \mathbf{B} defined in (9). Then we have that*

$$\begin{aligned} \|\mathbf{B}\|_F &= \sqrt{m^2\alpha^2 + m(1 - \alpha^2)}, & \|\mathbf{B} - \mathbf{B}_k\|_F &= \sqrt{m - k}(1 - \alpha), \\ \|\mathbf{B}\|_2 &= 1 + m\alpha - \alpha, & \|\mathbf{B} - \mathbf{B}_k\|_2 &= 1 - \alpha, \\ \|\mathbf{B}\|_* &= m, & \|\mathbf{B} - \mathbf{B}_k\|_* &= (m - k)(1 - \alpha), \end{aligned}$$

where $1 \leq k \leq m - 1$.

Proof The squared Frobenius norm of \mathbf{B} is

$$\|\mathbf{B}\|_F^2 = \sum_{i,j} b_{ij}^2 = m + (m^2 - m)\alpha^2.$$

Then we study the singular values of \mathbf{B} . Since \mathbf{B} is SPSD, here we do not distinguish between its singular values and eigenvalues.

The spectral norm, that is, the largest singular value, of \mathbf{B} is

$$\|\mathbf{B}\|_2 = \sigma_1 = \lambda_1 = \max_{\|\mathbf{x}\|_2 \leq 1} \mathbf{x}^T \mathbf{B} \mathbf{x} = \max_{\|\mathbf{x}\|_2 \leq 1} (1 - \alpha)\|\mathbf{x}\|_2^2 + \alpha(\mathbf{1}_m^T \mathbf{x})^2 = 1 - \alpha + m\alpha,$$

where the maximum is attained when $\mathbf{x} = \frac{1}{\sqrt{m}}\mathbf{1}_m$. Thus $\mathbf{u}_1 = \frac{1}{\sqrt{m}}\mathbf{1}_m$ is the top singular vector of \mathbf{B} . Then the projection of \mathbf{B} onto the subspace orthogonal to \mathbf{u}_1 is

$$\mathbf{B}_{1\perp} \triangleq \mathbf{B} - \mathbf{B}_1 = \mathbf{B} - \sigma_1 \mathbf{u}_1 \mathbf{u}_1^T = \frac{1 - \alpha}{m} (m\mathbf{I}_m - \mathbf{1}_m \mathbf{1}_m^T).$$

Then for all $j > 1$, the j -th top eigenvalue σ_j and eigenvector \mathbf{u}_j , that is, the singular value and singular vector, of \mathbf{B} satisfy

$$\sigma_j \mathbf{u}_j = \mathbf{B} \mathbf{u}_j = \mathbf{B}_{1\perp} \mathbf{u}_j = \frac{1 - \alpha}{m} (m\mathbf{u}_j - (\mathbf{1}_m^T \mathbf{u}_j)\mathbf{1}_m) = \frac{1 - \alpha}{m} (m\mathbf{u}_j - \mathbf{0}),$$

where the last equality follows from $\mathbf{u}_j \perp \mathbf{u}_1$, that is, $\mathbf{1}_m^T \mathbf{u}_j = 0$. Thus $\sigma_j = 1 - \alpha$, and

$$\|\mathbf{B} - \mathbf{B}_k\|_2 = \sigma_{k+1} = 1 - \alpha$$

for all $1 \leq k < m$. Finally we have that

$$\begin{aligned} \|\mathbf{B} - \mathbf{B}_k\|_F^2 &= \|\mathbf{B}\|_F^2 - \sum_{i=1}^k \sigma_i^2 = (m-k)(1-\alpha)^2, \\ \|\mathbf{B} - \mathbf{B}_k\|_* &= (m-k)\sigma_2 = (m-k)(1-\alpha), \\ \|\mathbf{B}\|_* &= \sum_{i=1}^m \sigma_i = (1+m\alpha-\alpha) + (m-1)(1-\alpha) = m, \end{aligned}$$

which complete our proofs. ■

C.1.2 THE ADVERSARIAL CASE FOR THE FROBENIUS NORM AND NUCLEAR NORM BOUNDS

Then we construct another adversarial case for proving the Frobenius norm and nuclear norm bounds. Let \mathbf{B} be a $p \times p$ matrix with diagonal entries equal to one and off-diagonal entries equal to α . Let $m = kp$ and we construct an $m \times m$ block diagonal matrix \mathbf{A} as follows:

$$\mathbf{A} = \text{BlkDiag}(\underbrace{\mathbf{B}, \dots, \mathbf{B}}_{k \text{ blocks}}) = \begin{bmatrix} \mathbf{B} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B} \end{bmatrix}. \quad (10)$$

Lemma 19 *Let \mathbf{A}_k be the best rank- k approximation to the matrix \mathbf{A} defined in (10). Then we have that*

$$\begin{aligned} \sigma_1(\mathbf{A}) &= \cdots = \sigma_k(\mathbf{A}) = 1 + p\alpha - \alpha, \\ \sigma_{k+1}(\mathbf{A}) &= \cdots = \sigma_m(\mathbf{A}) = 1 - \alpha, \\ \|\mathbf{A} - \mathbf{A}_k\|_F &= (1 - \alpha)\sqrt{m - k}, \\ \|\mathbf{A} - \mathbf{A}_k\|_* &= (1 - \alpha)(m - k). \end{aligned}$$

Lemma 19 can be easily proved using Lemma 18.

C.2 Lower Bounds of the Standard Nyström Method

Theorem 20 *For an $m \times m$ matrix \mathbf{B} with diagonal entries equal to one and off-diagonal entries equal to $\alpha \in [0, 1)$, the approximation error incurred by the standard Nyström method is lower bounded by*

$$\begin{aligned} \|\mathbf{B} - \tilde{\mathbf{B}}_c^{nys}\|_F &\geq (1 - \alpha)\sqrt{(m - c)\left(1 + \frac{m + c + \frac{2}{\alpha} - 2}{(c + \frac{1-\alpha}{\alpha})^2}\right)}, \\ \|\mathbf{B} - \tilde{\mathbf{B}}_c^{nys}\|_2 &\geq \frac{(1 - \alpha)\left(m + \frac{1-\alpha}{\alpha}\right)}{c + \frac{1-\alpha}{\alpha}}, \\ \|\mathbf{B} - \tilde{\mathbf{B}}_c^{nys}\|_* &\geq (m - c)(1 - \alpha)\frac{1 + c\alpha}{1 + c\alpha - \alpha}. \end{aligned}$$

Furthermore, the matrix $(\mathbf{B} - \tilde{\mathbf{B}}_c^{nys})$ is SPSD.

Proof The matrix \mathbf{B} is partitioned as in (9). The residual of the Nyström approximation is

$$\|\mathbf{B} - \tilde{\mathbf{B}}_c^{nys}\|_\xi = \|\mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{W}^\dagger\mathbf{B}_{21}^T\|_\xi, \quad (11)$$

where $\xi = 2, F, \text{ or } *$. Since $\mathbf{W} = (1 - \alpha)\mathbf{I}_c + \alpha\mathbf{1}_c\mathbf{1}_c^T$ is nonsingular when $\alpha \in [0, 1)$, so $\mathbf{W}^\dagger = \mathbf{W}^{-1}$. We apply the Sherman-Morrison-Woodbury formula

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}$$

to compute \mathbf{W}^\dagger , yielding

$$\mathbf{W}^\dagger = \frac{1}{1 - \alpha}\mathbf{I}_c - \frac{\alpha}{(1 - \alpha)(1 - \alpha + c\alpha)}\mathbf{1}_c\mathbf{1}_c^T.$$

According to the construction, \mathbf{B}_{21} is an $(m - c) \times c$ matrix with all entries equal to α , it follows that $\mathbf{B}_{21}\mathbf{W}^\dagger\mathbf{B}_{21}^T$ is an $(m - c) \times (m - c)$ matrix with all entries equal to

$$\eta \triangleq \alpha^2\mathbf{1}_c^T\mathbf{W}^\dagger\mathbf{1}_c = \frac{c\alpha^2}{1 - \alpha + c\alpha}. \quad (12)$$

Then we obtain that

$$\mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{W}^\dagger\mathbf{B}_{21}^T = (1 - \alpha)\mathbf{I}_{m-c} + (\alpha - \eta)\mathbf{1}_{m-c}\mathbf{1}_{m-c}^T. \quad (13)$$

It is easy to check that $\eta \leq \alpha \leq 1$, thus the matrix $(1 - \alpha)\mathbf{I}_{m-c} + (\alpha - \eta)\mathbf{1}_{m-c}\mathbf{1}_{m-c}^T$ is SPSD, and so is $(\mathbf{B} - \tilde{\mathbf{B}}_c^{nys})$.

Combining (11) and (13), we have that

$$\begin{aligned} \|\mathbf{B} - \tilde{\mathbf{B}}_c^{nys}\|_F^2 &= \|(1 - \alpha)\mathbf{I}_{m-c} + (\alpha - \eta)\mathbf{1}_{m-c}\mathbf{1}_{m-c}^T\|_F^2 \\ &= (m - c)(1 - \eta)^2 + \left((m - c)^2 - (m - c)\right)(\alpha - \eta)^2 \\ &= (m - c)(1 - \alpha)^2 \left(1 + \frac{\alpha^2(m + c) + 2(\alpha - \alpha^2)}{(1 - \alpha + c\alpha)^2}\right) \\ &= (m - c)(1 - \alpha)^2 \left(1 + \frac{m + c + \frac{2}{\alpha} - 2}{\left(c + \frac{1 - \alpha}{\alpha}\right)^2}\right), \end{aligned} \quad (14)$$

which proves the Frobenius norm of the residual.

Now we compute the spectral norm of the residual. Based on the results above we have that

$$\|\mathbf{B} - \tilde{\mathbf{B}}_c^{nys}\|_2 = \|(1 - \alpha)\mathbf{I}_{m-c} + (\alpha - \eta)\mathbf{1}_{m-c}\mathbf{1}_{m-c}^T\|_2.$$

Similar to the proof of Lemma 18, it is easily obtained that $\frac{1}{\sqrt{m - c}}\mathbf{1}_{m-c}$ is the top singular vector of the SPSD matrix $(1 - \alpha)\mathbf{I}_{m-c} + (\alpha - \eta)\mathbf{1}_{m-c}\mathbf{1}_{m-c}^T$, so the top singular value is

$$\sigma_1(\mathbf{B} - \tilde{\mathbf{B}}_c^{nys}) = (m - c)(\alpha - \eta) + 1 - \alpha = \frac{(1 - \alpha)\left(m + \frac{1 - \alpha}{\alpha}\right)}{c + \frac{1 - \alpha}{\alpha}}, \quad (15)$$

which proves the spectral norm bound because $\|\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}}\|_2 = \sigma_1(\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}})$.

It is also easy to show the rest singular values obey

$$\begin{aligned} \sigma_2(\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}}) &= \dots = \sigma_{m-c}(\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}}) \geq 0, \\ \sigma_{m-c+1}(\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}}) &= \dots = \sigma_m(\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}}) = 0. \end{aligned}$$

Thus we have, for $i = 2, \dots, m - c$,

$$\sigma_i^2(\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}}) = \frac{\|\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}}\|_F^2 - \sigma_1^2(\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}})}{m - c - 1} = (1 - \alpha)^2.$$

The nuclear norm of the residual $(\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}})$ is

$$\begin{aligned} \|\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}}\|_* &= \sum_{i=1}^m \sigma(\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}}) \\ &= \sigma_1(\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}}) + (m - c - 1)\sigma_2(\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{nys}}) \\ &= (m - c)(1 - \eta) \\ &= (m - c)(1 - \alpha) \left(1 + \frac{1}{c + \frac{1-\alpha}{\alpha}}\right). \end{aligned} \tag{16}$$

The theorem follows from equalities (14), (15), and (16). ■

Now we use the matrix \mathbf{A} constructed in (10) to show the Frobenius norm and nuclear norm lower bound. The bound is stronger than the one in Theorem 20 by a factor of k .

Theorem 21 *For the $m \times m$ SPSD matrix \mathbf{A} defined in (10), the approximation error incurred by the standard Nyström method is lower bounded by*

$$\begin{aligned} \|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_F &\geq (1 - \alpha) \sqrt{m - c - k + \frac{k(m + \frac{1-\alpha}{\alpha}k)^2}{(c + \frac{1-\alpha}{\alpha}k)^2}}, \\ \|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_* &\geq (1 - \alpha)(m - c) \left(1 + \frac{k}{c + \frac{1-\alpha}{\alpha}k}\right), \end{aligned}$$

where $k < m$ is an arbitrary positive integer.

Proof Let \mathbf{C} consist of c column sampled from \mathbf{A} and $\hat{\mathbf{C}}_i$ consist of c_i columns sampled from the i -th block diagonal matrix in \mathbf{A} . Without loss of generality, we assume $\hat{\mathbf{C}}_i$ consists of the first c_i columns of \mathbf{B} , and accordingly $\hat{\mathbf{W}}_i$ consists of the top left $c_i \times c_i$ block of \mathbf{B} . Thus $\mathbf{C} = \text{BlkDiag}(\hat{\mathbf{C}}_1, \dots, \hat{\mathbf{C}}_k)$

and $\mathbf{W} = \text{BlkDiag}(\hat{\mathbf{W}}_1, \dots, \hat{\mathbf{W}}_k)$.

$$\begin{aligned}
 \tilde{\mathbf{A}}_c^{\text{nys}} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C} &= \begin{bmatrix} \hat{\mathbf{C}}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\mathbf{C}}_k \end{bmatrix} \begin{bmatrix} \hat{\mathbf{W}}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\mathbf{W}}_k \end{bmatrix}^\dagger \begin{bmatrix} \hat{\mathbf{C}}_1^T & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\mathbf{C}}_k^T \end{bmatrix} \\
 &= \begin{bmatrix} \hat{\mathbf{C}}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\mathbf{C}}_k \end{bmatrix} \begin{bmatrix} \hat{\mathbf{W}}_1^\dagger & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\mathbf{W}}_k^\dagger \end{bmatrix} \begin{bmatrix} \hat{\mathbf{C}}_1^T & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\mathbf{C}}_k^T \end{bmatrix} \\
 &= \begin{bmatrix} \hat{\mathbf{C}}_1 \hat{\mathbf{W}}_1^\dagger \hat{\mathbf{C}}_1^T & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\mathbf{C}}_k \hat{\mathbf{W}}_k^\dagger \hat{\mathbf{C}}_k^T \end{bmatrix}. \tag{17}
 \end{aligned}$$

Then it follows from Theorem 20 that

$$\begin{aligned}
 \|\mathbf{A} - \tilde{\mathbf{A}}_c^{\text{nys}}\|_F^2 &= \sum_{i=1}^k \|\mathbf{B} - \hat{\mathbf{C}}_i \hat{\mathbf{W}}_i^\dagger \hat{\mathbf{C}}_i^T\|_F^2 \\
 &= \sum_{i=1}^k (p - c_i)(1 - \alpha)^2 \left(1 + \frac{p + c_i + 2\frac{1-\alpha}{\alpha}}{(c_i + \frac{1-\alpha}{\alpha})^2}\right) \\
 &= (1 - \alpha)^2 \sum_{i=1}^k (\hat{p} - \hat{c}_i) \left(1 + \frac{\hat{p} + \hat{c}_i}{\hat{c}_i^2}\right) \\
 &= (1 - \alpha)^2 \left(m - c - k + \hat{p}^2 \sum_{i=1}^k \hat{c}_i^{-2}\right),
 \end{aligned}$$

where $\hat{p} = p + \frac{1-\alpha}{\alpha}$ and $\hat{c}_i = c_i + \frac{1-\alpha}{\alpha}$. Since $\sum_{i=1}^k \hat{c}_i = c + \frac{1-\alpha}{\alpha}k \triangleq \hat{c}$, the term $\sum_{i=1}^k \hat{c}_i^{-2}$ is minimized when $\hat{c}_1 = \dots = \hat{c}_k$. Thus $\sum_{i=1}^k \hat{c}_i^{-2} = k\frac{k^2}{\hat{c}^2} = k^3\hat{c}^{-2}$. Finally we have that

$$\begin{aligned}
 \|\mathbf{A} - \tilde{\mathbf{A}}_c^{\text{nys}}\|_F^2 &= (1 - \alpha)^2 \left(m - c - k + \hat{p}^2 \sum_{i=1}^k \hat{c}_i^{-2}\right) \\
 &\geq (1 - \alpha)^2 \left(m - c - k + \frac{k(m + \frac{1-\alpha}{\alpha}k)^2}{(c + \frac{1-\alpha}{\alpha}k)^2}\right),
 \end{aligned}$$

by which the Frobenius norm bound follows.

Since the matrices $\mathbf{B} - \hat{\mathbf{C}}_i \hat{\mathbf{W}}_i^\dagger \hat{\mathbf{C}}_i^T$ are all SPSD by Theorem 20, so the matrix $(\mathbf{A} - \tilde{\mathbf{A}}_c^{\text{nys}})$ is also SPSD. We have that

$$\begin{aligned}
 \|\mathbf{A} - \tilde{\mathbf{A}}_c^{\text{nys}}\|_* &= \sum_{i=1}^k \|\mathbf{B} - \hat{\mathbf{C}}_i \hat{\mathbf{W}}_i^\dagger \hat{\mathbf{C}}_i^T\|_* \\
 &\geq (1 - \alpha) \sum_{i=1}^k (p - c_i) \left(1 + \frac{1}{c_i + \frac{1-\alpha}{\alpha}}\right) \\
 &\geq (1 - \alpha) k \left(\frac{m}{k} - \frac{c}{k}\right) \left(1 + \frac{1}{c/k + \frac{1-\alpha}{\alpha}}\right) \\
 &= (1 - \alpha)(m - c) \left(1 + \frac{k}{c + \frac{1-\alpha}{\alpha}k}\right),
 \end{aligned}$$

where the former inequality follows from Theorem 20, and the latter inequality follows by minimizing w.r.t. c_1, \dots, c_k subjecting to $c_1 + \dots + c_k = c$. ■

Theorem 22 *There exists an $m \times m$ SPSP matrix \mathbf{A} such that the approximation error incurred by the standard Nyström method is lower bounded by*

$$\begin{aligned} \frac{\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_F}{\|\mathbf{A} - \mathbf{A}_k\|_F} &\geq \sqrt{1 + \frac{m^2k - c^3}{c^2(m-k)}}, \\ \frac{\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_2}{\|\mathbf{A} - \mathbf{A}_k\|_2} &\geq \frac{m}{c}, \\ \frac{\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_*}{\|\mathbf{A} - \mathbf{A}_k\|_*} &\geq \frac{m-c}{m-k} \left(1 + \frac{k}{c}\right), \end{aligned}$$

where $k < m$ is an arbitrary positive integer.

Proof For the spectral norm bound we use the matrix \mathbf{A} constructed in (9) and set $\alpha \rightarrow 1$, then it follows directly from Lemma 18 and Theorem 20. For the Frobenius norm and nuclear norm bounds, we use the matrix \mathbf{A} constructed in (10) and set $\alpha \rightarrow 1$, then it follows directly from Lemma 19 and Theorem 21. ■

C.3 Lower Bounds of the Ensemble Nyström Method

The ensemble Nyström method (Kumar et al., 2009) is previously defined in (2). To derive lower bounds of the ensemble Nyström method, we assume that the t samples are non-overlapping. According to the construction of the matrix \mathbf{B} in (9), each of the t non-overlapping samples are equally “important”, so without loss of generality we set the t samples with equal weights: $\mu^{(1)} = \dots = \mu^{(t)} = \frac{1}{t}$.

Lemma 23 *Assume that the ensemble Nyström method selects a collection of t samples, each sample $\mathbf{C}^{(i)}$ ($i = 1, \dots, t$) contains c columns of \mathbf{B} without overlapping. For an $m \times m$ matrix \mathbf{B} with all diagonal entries equal to one and off-diagonal entries equal to $\alpha \in [0, 1)$, the approximation error incurred by the ensemble Nyström method is lower bounded by*

$$\begin{aligned} \|\mathbf{B} - \tilde{\mathbf{B}}_{t,c}^{ens}\|_F &\geq (1 - \alpha) \sqrt{\left(m - 2c + \frac{c}{t}\right) \left(1 + \frac{m + \frac{c}{t} + \frac{2}{\alpha} - 2}{\left(c + \frac{1-\alpha}{\alpha}\right)^2}\right)}, \\ \|\mathbf{B} - \tilde{\mathbf{B}}_{t,c}^{ens}\|_* &\geq (1 - \alpha)(m - c) \frac{c + \frac{1}{\alpha}}{c + \frac{1-\alpha}{\alpha}}. \end{aligned}$$

where $\tilde{\mathbf{B}}_{t,c}^{ens} = \frac{1}{t} \sum_{i=1}^t \mathbf{C}^{(i)} \mathbf{W}^{(i)\dagger} \mathbf{C}^{(i)T}$. Furthermore, the matrix $(\mathbf{B} - \tilde{\mathbf{B}}_{t,c}^{ens})$ is SPSP.

Proof We use the matrix \mathbf{B} constructed in (9). It is easy to check that $\mathbf{W}^{(1)} = \dots = \mathbf{W}^{(t)}$, so we use the notation \mathbf{W} instead. We assume that the samples contain the first tc columns of \mathbf{B} and each sample contains neighboring columns, that is,

$$\mathbf{B} = [\mathbf{C}^{(1)}, \dots, \mathbf{C}^{(t)}, \mathbf{B}_{(tc+1):m}].$$

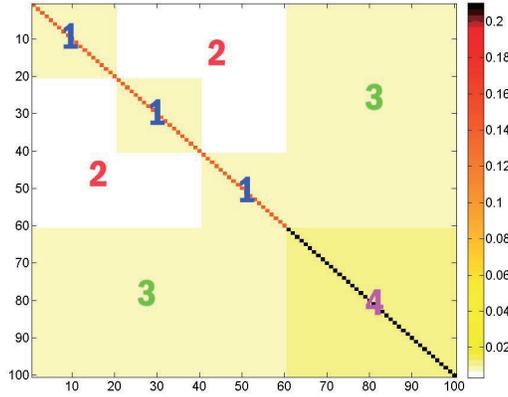


Figure 8: An illustration of the matrix $\mathbf{B} - \mathbf{B}_{t,c}^{\text{ens}}$ for the ensemble Nyström method where \mathbf{B} is defined in (9). Here we set $m = 100$, $c = 20$, $\alpha = 0.8$, and $t = 3$. For the ensemble Nyström method without overlapping, the matrix $\mathbf{B} - \mathbf{B}_{t,c}^{\text{ens}}$ can always be partitioned into four regions as annotated.

If a sample \mathbf{C} contains the first c columns of \mathbf{B} , then

$$\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T = \begin{bmatrix} \mathbf{W} & \mathbf{B}_{21}^T \\ \mathbf{B}_{21}\mathbf{W}^\dagger & \mathbf{B}_{21}^T\mathbf{B}_{21}^T \end{bmatrix} \quad \text{and} \quad \mathbf{B} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{W}^\dagger\mathbf{B}_{21}^T \end{bmatrix};$$

otherwise, after permuting the rows and columns of $\mathbf{B} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$, we get the same result:

$$\Pi(\mathbf{B} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T)\Pi^T = \mathbf{B} - \Pi(\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T)\Pi^T = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{22} - \mathbf{B}_{21}\mathbf{W}^\dagger\mathbf{B}_{21}^T \end{bmatrix},$$

where Π is a permutation matrix. As was shown in Equation (12), $\mathbf{B}_{21}\mathbf{W}^\dagger\mathbf{B}_{21}^T$ is an $(m-c) \times (m-c)$ matrix with all entries equal to

$$\eta = \frac{c\alpha^2}{1 - \alpha + c\alpha}.$$

Based on the properties of the matrix $\mathbf{B} - \mathbf{C}^{(i)}\mathbf{W}^{(i)\dagger}\mathbf{C}^{(i)T}$, we study the values of the entries of $\mathbf{B} - \tilde{\mathbf{B}}_{t,c}^{\text{ens}}$. We can express it as

$$\mathbf{B} - \tilde{\mathbf{B}}_{t,c}^{\text{ens}} = \mathbf{B} - \frac{1}{t} \sum_{i=1}^t \mathbf{C}^{(i)}\mathbf{W}^{(i)\dagger}\mathbf{C}^{(i)T} = \frac{1}{t} \sum_{i=1}^t \left(\mathbf{B} - \mathbf{C}^{(i)}\mathbf{W}^\dagger\mathbf{C}^{(i)T} \right), \quad (18)$$

and then a discreet examination reveals that $\mathbf{B} - \tilde{\mathbf{B}}_{t,c}^{\text{ens}}$ can be partitioned into four kinds of regions as illustrated in Figure 8. We annotate the regions in the figure and summarize the values of entries in each region in the table below. (Region 1 and 4 are further partitioned into diagonal entries and off-diagonal entries.)

Region	1 (diag)	1 (off-diag)	2	3	4 (diag)	4 (off-diag)
#Entries	tc	$tc^2 - tc$	$(tc)^2 - tc^2$	$2tc(m - tc)$	$m - tc$	$(m - tc)^2 - (m - tc)$
Value	$\frac{t-1}{t}(1 - \eta)$	$\frac{t-1}{t}(\alpha - \eta)$	$\frac{t-2}{t}(\alpha - \eta)$	$\frac{t-1}{t}(\alpha - \eta)$	$1 - \eta$	$\alpha - \eta$

Now we do summation over the entries of $\mathbf{B} - \tilde{\mathbf{B}}_{t,c}^{\text{ens}}$ to compute its squared Frobenius norm:

$$\begin{aligned} \|\mathbf{B} - \tilde{\mathbf{B}}_{t,c}^{\text{ens}}\|_F^2 &= tc \left[\frac{t-1}{t} (1-\eta) \right]^2 + \cdots + [(m-tc)^2 - (m-tc)](\alpha-\eta)^2 \\ &= (1-\alpha)(1+\alpha-2\eta)(m-2c+\frac{c}{t}) + (\alpha-\eta)^2 \left(4c^2 - 4cm + m^2 + \frac{2cm-3c^2}{t} \right) \\ &= (1-\alpha)^2 \left(m-2c+\frac{c}{t} \right) + \frac{(1-\alpha)^2}{(c+\frac{1-\alpha}{\alpha})^2} \left[(m-2c+\frac{c}{t}) \left(\frac{2}{\alpha} - 2 + m \right) + \frac{c(m-c)}{t} \right] \\ &\geq (1-\alpha)^2 \left(m-2c+\frac{c}{t} \right) \left(1 + \frac{m+\frac{c}{t}+\frac{2}{\alpha}-2}{(c+\frac{1-\alpha}{\alpha})^2} \right), \end{aligned}$$

where the last inequality follows from $\frac{c(m-c)}{t} = \frac{c}{t} \left((m-2c+\frac{c}{t}) + (c-\frac{c}{t}) \right) \geq \frac{c}{t} \left(m-2c+\frac{c}{t} \right)$.

Furthermore, since the matrices $\mathbf{B} - \mathbf{C}^{(i)} \mathbf{W}^\dagger \mathbf{C}^{(i)T}$ are all SPSD by Theorem 20, so their sum is also SPSD. Then the SPSD property of $(\mathbf{B} - \tilde{\mathbf{B}}_{t,c}^{\text{ens}})$ follows from (18). Therefore, the nuclear norm of $(\mathbf{B} - \tilde{\mathbf{B}}_{t,c}^{\text{ens}})$ equals to the matrix trace, that is,

$$\begin{aligned} \|\mathbf{B} - \tilde{\mathbf{B}}_{t,c}^{\text{ens}}\|_* &= \text{tr}(\mathbf{B} - \tilde{\mathbf{B}}_{t,c}^{\text{ens}}) \\ &= tc \cdot \frac{t-1}{t} (1-\eta) + (m-tc) \cdot (1-\eta) \\ &= (1-\alpha)(m-c) \frac{c+\frac{1}{\alpha}}{c+\frac{1-\alpha}{\alpha}}, \end{aligned}$$

which proves the nuclear norm bound in the lemma. ■

Theorem 24 Assume that the ensemble Nyström method selects a collection of t samples, each sample $\mathbf{C}^{(i)}$ ($i = 1, \dots, t$) contains c columns of \mathbf{A} without overlapping. For a the matrix \mathbf{A} defined in (10), the approximation error incurred by the ensemble Nyström method is lower bounded by

$$\begin{aligned} \|\mathbf{A} - \tilde{\mathbf{A}}_{t,c}^{\text{ens}}\|_F &\geq (1-\alpha) \sqrt{\left(m-2c+\frac{c}{t} - k \right) + k \left(\frac{m-c+\frac{c}{t}+k\frac{1-\alpha}{\alpha}}{c+k\frac{1-\alpha}{\alpha}} \right)^2}, \\ \|\mathbf{A} - \tilde{\mathbf{A}}_{t,c}^{\text{ens}}\|_* &\geq (1-\alpha)(m-c) \frac{c+\frac{1}{\alpha}k}{c+\frac{1-\alpha}{\alpha}k}, \end{aligned}$$

where $\tilde{\mathbf{A}}_{t,c}^{\text{ens}} = \frac{1}{t} \sum_{i=1}^t \mathbf{C}^{(i)} \mathbf{W}^{(i)\dagger} \mathbf{C}^{(i)T}$.

Proof According to the construction of \mathbf{A} in (10), the i -th sample $\mathbf{C}^{(i)}$ is also block diagonal. We denote it by $\mathbf{C}^{(i)} = \text{BlkDiag}(\hat{\mathbf{C}}_1^{(i)}, \dots, \hat{\mathbf{C}}_k^{(i)})$. Akin to (17), we have

$$\tilde{\mathbf{A}}_{t,c}^{\text{ens}} = \begin{bmatrix} \frac{1}{t} \sum_{i=1}^t \hat{\mathbf{C}}_1^{(i)} \hat{\mathbf{W}}_1^\dagger (\hat{\mathbf{C}}_1^{(i)})^T & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{t} \sum_{i=1}^t \hat{\mathbf{C}}_k^{(i)} \hat{\mathbf{W}}_k^\dagger (\hat{\mathbf{C}}_k^{(i)})^T \end{bmatrix}.$$

Thus the approximation error of the ensemble Nyström method is

$$\begin{aligned} \|\mathbf{A} - \tilde{\mathbf{A}}_{t,c}^{\text{ens}}\|_F^2 &= \sum_{j=1}^k \left\| \mathbf{B} - \frac{1}{t} \sum_{i=1}^t \hat{\mathbf{C}}_j^{(i)} \hat{\mathbf{W}}_j^\dagger (\hat{\mathbf{C}}_j^{(i)})^T \right\|_F^2 \\ &\geq (1-\alpha)^2 \sum_{j=1}^k \left(p - 2c_j + \frac{c_j}{t} \right) \left(1 + \frac{p + \frac{c_j}{t} + \frac{2}{\alpha} - 2}{(c_j + \frac{1-\alpha}{\alpha})^2} \right) \\ &= (1-\alpha)^2 \left[\left(m - 2c + \frac{c}{t} \right) + \sum_{j=1}^k \left(p - 2c_j + \frac{c_j}{t} \right) \frac{p + \frac{c_j}{t} + \frac{2(1-\alpha)}{\alpha}}{(c_j + \frac{1-\alpha}{\alpha})^2} \right], \end{aligned}$$

where the inequality follows from Lemma 23, and the last equality follows from $\sum_{j=1}^k c_j = c$ and $kp = m$. The summation in the last equality equals to

$$\begin{aligned} &\sum_{j=1}^k \left[\left(p + \frac{c_j}{t} + \frac{2(1-\alpha)}{\alpha} \right) - 2 \left(c_j + \frac{1-\alpha}{\alpha} \right) \right] \frac{p + \frac{c_j}{t} + \frac{2(1-\alpha)}{\alpha}}{(c_j + \frac{1-\alpha}{\alpha})^2} \\ &= -k + \sum_{j=1}^k \left(\frac{p + \frac{c_j}{t} + \frac{2(1-\alpha)}{\alpha}}{c_j + \frac{1-\alpha}{\alpha}} - 1 \right)^2 \\ &\geq -k + k \left(\frac{m - c + \frac{c}{t} + k \frac{1-\alpha}{\alpha}}{c + k \frac{1-\alpha}{\alpha}} \right)^2. \end{aligned}$$

Here the inequality holds because the function is minimized when $c_1 = \dots = c_k = c/k$. Finally we have that

$$\|\mathbf{A} - \tilde{\mathbf{A}}_{t,c}^{\text{ens}}\|_F^2 \geq (1-\alpha)^2 \left[\left(m - 2c + \frac{c}{t} - k \right) + k \left(\frac{m - c + \frac{c}{t} + k \frac{1-\alpha}{\alpha}}{c + k \frac{1-\alpha}{\alpha}} \right)^2 \right],$$

which proves the Frobenius norm bound in the theorem.

Furthermore, since the matrix $\mathbf{B} - \frac{1}{t} \sum_{i=1}^t \hat{\mathbf{C}}_j^{(i)} \hat{\mathbf{W}}_j^\dagger (\hat{\mathbf{C}}_j^{(i)})^T$ is SPSD by Lemma 23, so the block diagonal matrix $(\mathbf{A} - \tilde{\mathbf{A}}_{t,c}^{\text{ens}})$ is also SPSD. Thus we have

$$\|\mathbf{A} - \tilde{\mathbf{A}}_{t,c}^{\text{ens}}\|_* = (1-\alpha) \sum_{i=1}^t (p - c_i) \frac{c_i + \frac{1}{\alpha}}{c_i + \frac{1-\alpha}{\alpha}} \geq (1-\alpha)(m - c) \left(1 + \frac{k}{c + \frac{1-\alpha}{\alpha}k} \right),$$

which proves the nuclear norm bound in the theorem. ■

Theorem 25 *Assume that the ensemble Nyström method selects a collection of t samples, each sample $\mathbf{C}^{(i)}$ ($i = 1, \dots, t$) contains c columns of \mathbf{A} without overlapping. Then there exists an $m \times m$ SPSD matrix \mathbf{A} such that the relative-error ratio of the ensemble Nyström method is lower bounded by*

$$\begin{aligned} \frac{\|\mathbf{A} - \tilde{\mathbf{A}}_{t,c}^{\text{ens}}\|_F}{\|\mathbf{A} - \mathbf{A}_k\|_F} &\geq \sqrt{\frac{m - 2c + c/t - k}{m - k} \left(1 + \frac{k(m - 2c + c/t)}{c^2} \right)}, \\ \frac{\|\mathbf{A} - \tilde{\mathbf{A}}_{t,c}^{\text{ens}}\|_*}{\|\mathbf{A} - \mathbf{A}_k\|_*} &\geq \frac{m - c}{m - k} \left(1 + \frac{k}{c} \right), \end{aligned}$$

where $\tilde{\mathbf{A}}_{t,c}^{\text{ens}} = \frac{1}{t} \sum_{i=1}^t \mathbf{C}^{(i)} \mathbf{W}^{(i)\dagger} \mathbf{C}^{(i)T}$.

Proof The theorem follows directly from Theorem 24 and Lemma 19 by setting $\alpha \rightarrow 1$. ■

References

- A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications. Second Edition*. Springer, 2003.
- M. W. Berry, S. A. Pulatova, and G. W. Stewart. Algorithm 844: computing sparse reduced-rank approximations to sparse matrices. *ACM Transactions on Mathematical Software*, 31(2):252–269, 2005.
- J. Bien, Y. Xu, and M. W. Mahoney. CUR from a sparse optimization viewpoint. In *Advances in Neural Information Processing Systems (NIPS)*. 2010.
- C. H. Bischof and P. C. Hansen. Structure-preserving and rank-revealing QR-factorizations. *SIAM Journal on Scientific and Statistical Computing*, 12(6):1332–1350, 1991.
- C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near-optimal column-based matrix reconstruction. *CoRR*, abs/1103.0995, 2011.
- T. F. Chan. Rank revealing QR factorizations. *Linear Algebra and Its Applications*, 88:67–82, 1987.
- S. Chandrasekaran and I. C. F. Ipsen. On rank-revealing factorisations. *SIAM Journal on Matrix Analysis and Applications*, 15(2):592–622, 1994.
- P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.
- S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of The American Society for Information Science*, 41(6):391–407, 1990.
- A. Deshpande and L. Rademacher. Efficient volume sampling for row/column subset selection. In *Proceedings of the 51st IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 329–338, 2010.
- A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2(2006):225–247, 2006.
- P. Drineas and R. Kannan. Pass-efficient algorithms for approximating large matrices. In *Proceeding of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 223–232, 2003.
- P. Drineas and M. W. Mahoney. On the Nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.
- P. Drineas, R. Kannan, and M. W. Mahoney. Fast Monte Carlo algorithms for matrices III: computing a compressed approximate matrix decomposition. *SIAM Journal on Computing*, 36(1):184–206, 2006.

- P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, September 2008.
- P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. In *International Conference on Machine Learning (ICML)*, 2012.
- L. V. Foster. Rank and null space calculations using matrix decomposition without column interchanges. *Linear Algebra and its Applications*, 74:47–71, 1986.
- C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.
- A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL <http://archive.ics.uci.edu/ml>.
- A. Frieze, R. Kannan, and S. Vempala. Fast Monte Carlo algorithms for finding low-rank approximations. *Journal of the ACM*, 51(6):1025–1041, November 2004. ISSN 0004-5411.
- A. Gittens and M. W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. *arXiv preprint arXiv:1303.1849*, 2013.
- S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin. A theory of pseudoskeleton approximations. *Linear Algebra and Its Applications*, 261:1–21, 1997a.
- S. A. Goreinov, N. L. Zamarashkin, and E. E. Tyrtyshnikov. Pseudo-skeleton approximations by matrices of maximal volume. *Mathematical Notes*, 62(4):619–623, 1997b.
- M. Gu and S. C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.
- V. Guruswami and A. K. Sinop. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2012.
- I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the NIPS 2003 feature selection challenge. *Advances in Neural Information Processing Systems (NIPS)*, 2004.
- N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.
- Y. P. Hong and C. T. Pan. Rank-revealing QR factorizations and the singular value decomposition. *Mathematics of Computation*, 58(197):213–232, 1992.
- R. Jin, T. Yang, and M. Mahdavi. Improved bound for the Nyström method and its application to kernel classification. *CoRR*, abs/1111.2262, 2011.
- S. Kumar, M. Mohri, and A. Talwalkar. Ensemble Nyström method. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

- S. Kumar, M. Mohri, and A. Talwalkar. Sampling methods for the Nyström method. *Journal of Machine Learning Research*, 13:981–1006, 2012.
- F. G. Kuruvilla, P. J. Park, and S. L. Schreiber. Vector algebra in the analysis of genome-wide expression data. *Genome Biology*, 3:research0011–research0011.1, 2002.
- M. Li, J. T. Kwok, and B.-L. Lu. Making large-scale Nyström approximation possible. In *International Conference on Machine Learning (ICML)*, 2010.
- L. Mackey, A. Talwalkar, and M. I. Jordan. Divide-and-conquer matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*. 2011.
- M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.
- M. W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.
- M. W. Mahoney, M. Maggioni, and P. Drineas. Tensor-CUR decompositions for tensor-based data. *SIAM Journal on Matrix Analysis and Applications*, 30(3):957–987, 2008.
- C. Mesterharm and M. J. Pazzani. Active learning using on-line algorithms. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.
- D. Michie, D. J. Spiegelhalter, and C. C. Taylor. Machine learning, neural and statistical classification. 1994.
- L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A*, 4(3):519–524, Mar 1987.
- G. W. Stewart. Four algorithms for the efficient computation of truncated pivoted QR approximations to a sparse matrix. *Numerische Mathematik*, 83(2):313–323, 1999.
- A. Talwalkar and A. Rostamizadeh. Matrix coherence and the Nyström method. *arXiv preprint arXiv:1004.2008*, 2010.
- A. Talwalkar, S. Kumar, and H. Rowley. Large-scale manifold learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.
- E. E. Tyrtyshnikov. Incomplete cross approximation in the mosaic-skeleton method. *Computing*, 64:367–380, 2000.
- C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2001.
- K. Zhang and J. T. Kwok. Clustered Nyström method for large scale manifold learning and dimension reduction. *IEEE Transactions on Neural Networks*, 21(10):1576–1587, 2010.
- K. Zhang, I. W. Tsang, and J. T. Kwok. Improved Nyström low-rank approximation and error analysis. In *International Conference on Machine Learning (ICML)*, 2008.