THE SPECTRAL NORM ERROR OF A SIMPLE CUR DECOMPOSITION FOR POSITIVE SEMIDEFINITE MATRICES

ALEX GITTENS

ABSTRACT. The simple Nyström extension forms a low-rank approximation to a positive-semidefinite matrix by uniformly randomly sampling from its columns. This paper provides the first relative-error bound on the spectral norm error incurred in this process. This bound follows from a natural connection between the Nyström extension and the column subset selection problem. The main tool is a matrix Chernoff bound for sampling without replacement.

1. INTRODUCTION

The truncated singular value decomposition (SVD) is the classical tool for low-rank approximation of matrices. Given a square matrix of size n, a rank-k truncated SVD costs $\mathrm{O}(kn^2)$ operations to compute. Due to the large size of modern datasets, much interest has been expressed in finding less expensive low-rank approximation schemes that offer approximation guarantees comparable to that of the truncated SVD.

In one promising class of approximations, the matrix is approximated by the product $\boldsymbol{CUR}$, where $\boldsymbol{C}$ and $\boldsymbol{R}$ are respectively small subsets of the columns and rows of the matrix and $\boldsymbol{U}$, the coupling matrix, is a generalized inverse of the overlap of $\boldsymbol{C}$ and $\boldsymbol{R}$. Accordingly, these schemes are known as CUR decompositions. In general, $\boldsymbol{C}$ and $\boldsymbol{R}$ can be chosen independently, e.g. both could be chosen uniformly at random.

However, it may be advantageous to require $\boldsymbol{C}$ and $\boldsymbol{R}$ to be related. In particular, if the matrix is positive semidefinite (PSD), a low-rank approximation which is also PSD can be formed by selecting the same rows as columns (i.e., by taking $\boldsymbol{R} = \boldsymbol{C}^t$). This subset of CUR decompositions is known as Nyström extensions, as they can alternately be interpreted as linear algebraic analogs of the Nyström extensions familiar to numerical analysts.

We consider the simple Nyström extension, a particular scheme in which the columns are sampled uniformly without replacement. By exploiting a natural connection between Nyström extensions and the column subset selection problem, we find the first relative-error spectral norm guarantees for the simple Nyström extension.

1.1. **Efficacy of the simple Nyström extension.** Perhaps surprisingly, given that one uses no information about the matrix itself to make the column selections, the simple Nyström extension is effective in practice. Because of its data agnosticism and empirical accuracy, the simple Nyström extension is a natural choice for any application involving a positive semidefinite matrix where one wishes to avoid the cost of examining (or even constructing) the entire matrix before approximation.

The simple Nyström extension has proven to be particularly useful in image-processing applications, which typically involve computations with large dense matrices [FBCM04, WDT+09, BF]. In spectral image segmentation, for example, one constructs a matrix of pairwise pixel affinities by comparing neighborhoods of each pair of pixels. Several leading eigenvectors of this matrix are then used to segment the image. The affinity matrix of an $N \times N$ image has dimension $N^2 \times N^2$, so it is challenging to construct and hold the affinity matrix in memory even for images of a moderate

size. Similarly, the density and size of the affinity matrix makes it challenging to compute the leading eigenvectors. [FBCM04] proposes using the simple Nyström extension to approximate the eigenvectors of the affinity matrix. Doing so allows one to work with much larger images, because it is only necessary to compute a fraction of the columns of the affinity matrix.

1.2. **Structure of the Nyström extension.** Let A be a PSD matrix of size n. Select $\ell \ll n$ columns of A to constitute the columns of a matrix C. Let W be the $\ell \times \ell$ matrix formed by the intersection of the columns in C and the corresponding rows in A. The matrix $CW^\dagger C^t$ is then a Nyström extension of A (see Figure 1). Here $(\cdot)^\dagger$ denotes Moore-Penrose pseudoinversion. Since W is a principal submatrix of A, it is positive-semidefinite, and hence the Nyström extension is also positive-semidefinite.
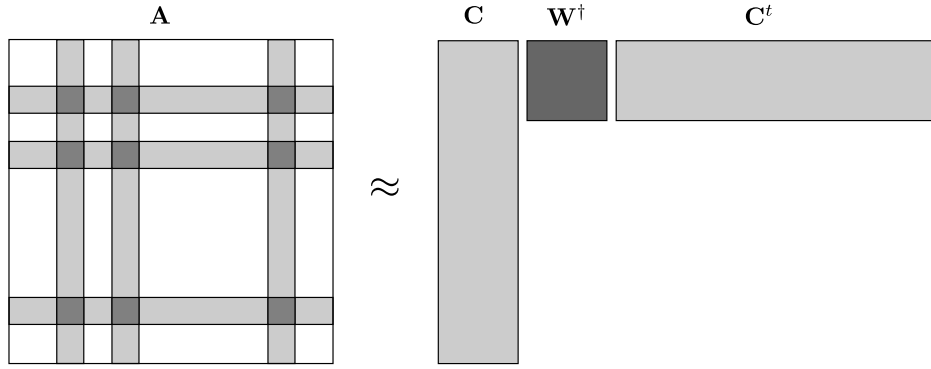

FIGURE 1. The Nyström extension procedure

The manner in which the columns are sampled and $W^\dagger$ is calculated or approximated determines the type of the Nyström extension. Various sampling schemes have been proposed, ranging from the simple scheme in which the columns are selected uniformly at random without replacement to more sophisticated and calculation-intensive schemes that involve sampling from a distribution determined by the determinants of principal submatrices of A [BW09]. In practice the simple scheme represents a favorable trade-off between speed and accuracy [KMT09b].

1.3. **Nyström approximation of invariant subspaces.** In many applications, including the image-processing example taken from [FBCM04], the Nyström extension is used to obtain approximations to the dominant invariant subspace of a PSD matrix A, rather than a low-rank approximation [HM]. Through the Davis–Kahan sinΘ theorem, the spectral norm approximation error provides information on the quality of the approximate invariant subspace obtained via Nyström extensions [Bha97, Section VII.3].

To be more precise, let $\tilde{A}$ be a Nyström approximation to A and assume both A and $\tilde{A}$ have unique dominant k-dimensional invariant subspaces. Let U_k and $\tilde{U}_k$ have orthogonal columns and span, respectively, the dominant k-dimensional invariant subspace of A and that of $\tilde{A}$. Recall one natural definition for the distance between the dominant k-dimensional invariant subspaces of A and $\tilde{A}$,

$$\mathrm{dist}(U_k, \tilde{U}_k) = \|P_{U_k} - P_{\tilde{U}_k}\|_2.$$

Denote the kth-largest eigenvalue of a matrix M by $\lambda_k(M)$, so that $\lambda_1(M) \geq \lambda_2(M) \geq \ldots$. It follows from the Davis–Kahan sinΘ theorem that if $\lambda_k(A) - \lambda_{k+1}(\tilde{A}) > 0$, then

$$\mathrm{dist}(U_k, \tilde{U}_k) \leq \frac{\|A - \tilde{A}\|_2}{\lambda_k(A) - \lambda_{k+1}(\tilde{A})}. \tag{1}$$

Assume that we have a relative-error spectral norm bound of the form

$$\|\boldsymbol{A} - \tilde{\boldsymbol{A}}\|_2 \leq \mathrm{C}\lambda_j(\boldsymbol{A}),$$

for some $j \geq k$. Then equation (1) becomes

$$\mathrm{dist}(\boldsymbol{U}_k, \tilde{\boldsymbol{U}}_k) \leq \frac{\mathrm{C}\lambda_j(\boldsymbol{A})}{\lambda_k(\boldsymbol{A}) - \lambda_{k+1}(\boldsymbol{A}) - \mathrm{C}\lambda_j(\boldsymbol{A})}.$$

Thus, we conclude that if $\mathrm{C}\lambda_j(\boldsymbol{A})$ is sufficiently smaller than the eigengap $\lambda_k(\boldsymbol{A}) - \lambda_{k+1}(\boldsymbol{A})$, the Nyström extension yields a quality approximation to the dominant k-dimensional invariant subspace of $\boldsymbol{A}$.

This paper presents a simple framework for the analysis of Nyström schemes that yields a state-of-the-art spectral norm error bound in the case of the simple Nyström extension scheme. Specifically, it generalizes the coherence-based exact recovery result in [TR10] to also guarantee small relative error in the case of a matrix with a fast-decaying spectrum. This is the first truly relative-error spectral norm bound available for any Nyström extension method. When the eigengap $\lambda_k(\boldsymbol{A}) - \lambda_{k+1}(\boldsymbol{A})$ is sufficiently large, our result sanctions the use of the simple Nyström extension for the approximation of the dominant k-dimensional invariant subspace of $\boldsymbol{A}$.

1.4. **Our relative-error spectral norm bound.** The efficacy of the simple Nyström extension is of course dependent on the data set to which it is applied. Intuitively, the extension should perform better if the information is spread evenly throughout the columns of the matrix. The *coherence* of the invariant subspaces of $\boldsymbol{A}$ provides a quantitative measure of the informativity of the columns. Let $\mathcal{S}$ be a k-dimensional subspace of $\mathbb{R}^n$ and $\boldsymbol{P}_\mathcal{S}$ denote the projection onto $\mathcal{S}$. Then the coherence of $\mathcal{S}$ is

$$\mu_0(\mathcal{S}) = \frac{n}{k} \max_i (\boldsymbol{P}_\mathcal{S})_{ii}.$$

Corollary 1 is a condensed version of our main result, Theorem 2, and uses the notion of coherence to provide a bound on the error of the simple Nyström extension.

Corollary 1. *Let $\boldsymbol{A}$ be a real PSD matrix of size n. Given an integer $k \leq n$, let τ denote the coherence of a dominant k-dimensional invariant subspace of $\boldsymbol{A}$. Fix a nonzero failure probability δ. If $\ell \geq 8\tau k \log(k/\delta)$ columns of $\boldsymbol{A}$ are chosen uniformly at random without replacement, then*

$$\|\boldsymbol{A} - \boldsymbol{C}\boldsymbol{W}^\dagger \boldsymbol{C}^t\|_2 \leq \lambda_{k+1}(\boldsymbol{A}) \left(1 + \frac{2n}{\ell}\right)$$

with probability exceeding $1 - \delta$.

For Corollary 1 to provide a meaningful estimate of ℓ, the required number of column samples, the coherence τ must be small enough that $\ell \ll n$. This requirement reflects our intuition that approximations formed using a small number of columns will not be accurate if a small number of columns are significantly more influential than the others. In the best-case scenario of $\tau = 1$, the columns are equally informative and we find that the error of a simple Nyström extension formed using just $\mathrm{C}k \log k$ columns is close to that of the optimal rank-k approximation.

1.5. **Relevant literature.** We briefly review the literature on CUR decompositions and Nyström extensions, focusing on the simple Nyström scheme. In this section $\boldsymbol{A}$ is a PSD matrix, $\boldsymbol{A}_k$ is a rank-k approximation to $\boldsymbol{A}$ that is optimal in the spectral norm, and ℓ is the number of columns used to construct a Nyström extension of $\boldsymbol{A}$.

In [TGZ97], Goreinov, Tyrtyshnikov, and Zamarashkin introduce deterministic CUR decompositions under the name pseudo-skeleton approximations. They establish that there exists at least one set of exactly k columns and k rows for which the corresponding pseudo-skeleton approximation has a spectral norm error smaller than $\|\boldsymbol{A} - \boldsymbol{A}_k\|_2(1 + 4\sqrt{kn})$. Drineas, Kannan, and Mahoney [DKM06] introduce randomized CUR decompositions and show that with appropriately chosen

Table 1. Simple Nyström extension error guarantees

Summary of prior guarantees on the performance of the simple Nyström extension. Given an $n \times n$ positive semidefinite matrix $\boldsymbol{A}$, we compare the error of the simple Nyström extension to the error of the optimal rank-k approximation in the spectral and Frobenius norms. Let the diagonal matrix $\boldsymbol{\Sigma}_2$ contain the bottom $n - k$ eigenvalues of $\boldsymbol{A}$. Then these optimal errors are given by $\|\boldsymbol{\Sigma}_2\|_2$ and $\|\boldsymbol{\Sigma}_2\|_F$ respectively. Notation: $(\cdot)$ denotes the difference between $\boldsymbol{A}$ and the Nyström extension, ℓ is the number of column samples used, the columns of $\boldsymbol{U}_1$ comprise the top k eigenvectors of $\boldsymbol{A}$, and $\sum_{i=1}^{\ell} a_{ii}^{\downarrow}$ is the sum of the ℓ largest entries on the diagonal of $\boldsymbol{A}$.

Source	Guarantees	ℓ	Remarks
[KMT09b]	$\mathbb{E}\|\cdot\|_F \leq \|\boldsymbol{\Sigma}_2\|_F + \varepsilon\left(\left(\frac{n}{\ell}\sum_{i=1}^{\ell} a_{ii}^{\downarrow}\right)\sqrt{n\sum_{i=1}^{n} a_{ii}^2}\right)^{1/2}$	$64\varepsilon^{-4}k$	sampling without replacement
[TR10]	$\|\cdot\|_2 = \|\cdot\|_F = 0$	$\mathrm{O}(n\max_{ij}|(\boldsymbol{U}_1)_{ij}|^2 k\log(k))$	sampling with replacement, constant probability of success, requires $\mathrm{rank}(\boldsymbol{A}) = k$
[LKL10]	$\mathbb{E}\|\cdot\|_2 \leq \Omega(1)\|\boldsymbol{\Sigma}_2\|_2 + \Omega\left(\frac{n}{\sqrt{\ell}}\right)\max_i a_{ii}$	$\mathrm{O}(k)$	large scale variant, with replacement
	$\mathbb{E}\|\cdot\|_F \leq \Omega(\sqrt{k})\|\boldsymbol{\Sigma}_2\|_F + \Omega(n)\max_i a_{ii}$	$\mathrm{O}(k)$	
[MTJ]	$\|\cdot\|_F \leq (1+\varepsilon)\|\boldsymbol{\Sigma}_2\|_F$	$\mathrm{O}(\varepsilon^{-2}n\max_i|(\boldsymbol{U}_1)_i|^2\log(k))$	sampling without replacement, constant probability of success

non-uniform sampling distributions, the spectral and Frobenius norm errors of the approximations are with high probability within an additive factor of $\varepsilon\|\boldsymbol{A}\|_F$ of the optimal errors. In [DMM08], Drineas, Mahoney, and Muthukrishnan show that using "subspace sampling" probabilites gives a $(1 + \epsilon)$ relative Frobenius norm error guarantee on the quality of the CUR approximations with constant probability, but computing these probabilities seems to be as expensive as computing the rank-k truncated SVD.

Most recently, in a simultaneously completed work, Chiu and Demanet [CD] consider CUR decomposition of a rectangular matrix $\boldsymbol{A}$ under the assumption that the singular vectors are incoherent. A novel addition to their CUR decomposition algorithm involves truncating (the SVD of) the coupling matrix so that it is well conditioned. They establish that *if the truncation level is chosen appropriately* then, with high probability, the error of a CUR decomposition where the columns and rows are chosen uniformly at random stays within a multiplicative $\mathrm{O}(n/\ell)$ factor of the optimal rank-k approximation error. Thus, the results of [CD] almost imply our main result, Theorem 2. However, we note that our argument is simpler and does not require knowledge of the correct truncation level. Further, we establish an explicit connection between the randomized column subset selection problem and the quality of Nyström extensions.

Williams and Seeger introduce the Nyström extension in [WS01], based upon a similar method used in numerical integral equation solvers, as a heuristic method for efficiently approximating the eigendecomposition of kernel matrices. Drineas and Mahoney provide the first rigorous analysis of a Nyström extension in [DM05]; in the scheme they consider, columns are sampled with probability proportional to the square of the diagonal entries of $\boldsymbol{A}$. In addition to probabilistic schemes, many adaptive sampling schemes have been proposed. These attempt to progressively choose the columns to decrease the approximation error. For an introduction to this body of literature, we refer the interested reader to the discussion in [FGK11].

Kumar, Mohri, and Talwalkar attempt the first analysis of the simple Nyström extension in [KMT09b], resulting in bounds for the Frobenius norm error. Their analysis proceeds by bounding the expectation and variance of the error then applying a concentration of measure argument. A simplified yet representative statement of their bound is that

$$\|\boldsymbol{A} - \boldsymbol{C}\boldsymbol{W}^{\dagger}\boldsymbol{C}^t\|_F \leq \|\boldsymbol{A} - \boldsymbol{A}_k\|_F + \varepsilon n \cdot \max_i (\boldsymbol{A})_{ii}$$

with constant probability when $\ell \geq Ck/\varepsilon^4$.

In [KMT09a], Kumar, Mohri, and Talwalkar establish that if $\mathrm{rank}(\boldsymbol{W}) = \mathrm{rank}(\boldsymbol{A}) = r$, then $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{W}^{\dagger}\boldsymbol{C}^t$. Talwalkar and Rostamizadeh prove this implies that, if $\ell \geq \mathrm{C}r\mu\log(r/\delta)$, simple Nyström extension results in exact recovery with constant probability [TR10]. Here μ is a measure of the coherence of the column space of $\boldsymbol{A}$ that differs slightly from the definition used in this paper. Their key observation is that if no columns of $\boldsymbol{A}$ are singularly influential, then $\boldsymbol{W}$ will have maximal rank when ℓ is slightly larger than the rank of $\boldsymbol{A}$. Thus, the number of samples required

for exact recovery is determined by the rank of $\boldsymbol{A}$ and the coherence, μ, of its range space. They use a standard result from the compressed sensing literature to quantify this phenomenon and obtain an estimate for ℓ [CR07].

In [LKL10], Kwok, Li, and Lu propose replacing $\boldsymbol{W}$ with a low-rank approximation $\tilde{\boldsymbol{W}}$ to facilitate the pseudoinversion operation. This large-scale variant of the simple Nyström extension allows a larger number of column samples to be drawn and leads to smaller empirical approximation errors. The approximation $\tilde{\boldsymbol{W}}$ is constructed using the randomized methodology espoused in [HMT11]. The analysis of the error combines bounds provided in [HMT11] with a matrix sparsification argument. In addition to ℓ and k, the Nyström algorithm presented in [LKL10] depends on two additional parameters that control the creation of $\tilde{\boldsymbol{W}}$: an oversampling factor p and the number of iterations of the power method, respectively p and q. After taking $p = \ell - k$ and $q = 1$, the results of [LKL10] provide error bounds for the simple Nyström extension:

$$\mathbb{E}\|\boldsymbol{A} - \boldsymbol{C}\boldsymbol{W}^{\dagger}\boldsymbol{C}^t\|_2 \leq \mathrm{C}\left(\|\boldsymbol{A} - \boldsymbol{A}_k\|_2 + \frac{n}{\sqrt{l}}\max_i(\boldsymbol{A})_{ii}\right) \tag{2}$$

$$\mathbb{E}\|\boldsymbol{A} - \boldsymbol{C}\boldsymbol{W}^{\dagger}\boldsymbol{C}^t\|_F \leq \mathrm{C}\sqrt{\ell}\left(\|\boldsymbol{A} - \boldsymbol{A}_k\|_F + \frac{n}{\sqrt{\ell}}\max_i(\boldsymbol{A})_{ii}\right).$$

Of the works mentioned, only [LKL10] provides a bound on the spectral error of the Nyström method for $\boldsymbol{A}$ of arbitrary rank. Unfortunately, the quantity $\max_i(\boldsymbol{A})_{ii}$ is, for a general $\boldsymbol{A} \succeq 0$, bounded only by $\lambda_1(\boldsymbol{A})$. Thus equation (2) does not provide a relative-error bound. In fact, the spectral norm error bound provided in this paper is always tighter than the bound provided in [LKL10], for any choice of p and q.

Our work presents an intuitive and simple approach to the analysis of Nyström extensions through their connection to the randomized column subset selection problem. This allows us to obtain the first truly relative-error guarantee on the spectral norm error. This paper analyzes the simple sampling scheme but we believe that the framework given is flexible enough to be fruitfully applied to the analysis of other Nyström extension schemes including, in particular, the large-scale variant introduced in [LKL10].

1.6. **Outline.** In Section 2 we introduce our notation and review some algebraic preliminaries. In Section 3 we establish a connection between the Nyström extension procedure and the column subset selection problem. We exploit this connection and a result from [HMT11] to provide a general error bound for any Nyström extension scheme. In Section 4 we specialize this result to the case of the simple Nyström extension.

2. NOTATION

We work exclusively with real matrices and order the eigenvalues of a PSD matrix $\boldsymbol{A}$ so that $\lambda_1(\boldsymbol{A}) \geq \lambda_2(\boldsymbol{A}) \geq \cdots \geq \lambda_n(\boldsymbol{A})$. Each PSD matrix $\boldsymbol{A}$ has a unique square root $\boldsymbol{A}^{1/2}$ that is also positive-semidefinite, has the same eigenspaces as $\boldsymbol{A}$, and satisfies $\boldsymbol{A} = \left(\boldsymbol{A}^{1/2}\right)^2$.

The projection onto the column space of a matrix $\boldsymbol{M}$ is written $\boldsymbol{P_M}$ and satisfies

$$\boldsymbol{P_M} = \boldsymbol{M}\boldsymbol{M}^{\dagger} = \boldsymbol{M}(\boldsymbol{M}^t\boldsymbol{M})^{\dagger}\boldsymbol{M}^t.$$

The notation $(\boldsymbol{x})_j$ refers to the jth entry of the vector $\boldsymbol{x}$, and $(\boldsymbol{M})_i$ refers to the ith column of the matrix $\boldsymbol{M}$. Likewise, $(\boldsymbol{M})_{ij}$ refers to the (i, j) entry of $\boldsymbol{M}$.

The *coherence* of a matrix $\boldsymbol{U} \in \mathbb{R}^{n \times k}$ with orthonormal columns is, up to a scaling factor, the maximum of the squared Euclidean norms of its rows:

$$\mu_0(\boldsymbol{U}) := \frac{n}{k}\max_i\|(\boldsymbol{U}^t)_i\|^2 = \frac{n}{k}\max_i(\boldsymbol{U}\boldsymbol{U}^t)_{ii} = \frac{n}{k}\max_i(\boldsymbol{P_U})_{ii}.$$

From the last equality, we see that the coherence is in fact an intrinsic property of the subspace spanned by $\boldsymbol{U}$. Thus we refer to the coherence of a subspace without first choosing a particular orthogonal basis $\boldsymbol{U}$.

3. The connection to the column subset selection problem

In this section we establish a fruitful connection between the performance of the Nyström extension and the performance of randomized *column subset selection.*

Given a matrix $\boldsymbol{M}$, the goal of column selection is to choose a small but informative subset $\boldsymbol{C}$ of the columns of $\boldsymbol{M}$ so that, after approximating $\boldsymbol{M}$ with the matrix obtained by projecting $\boldsymbol{M}$ onto the span of $\boldsymbol{C}$, the residual $(\boldsymbol{I} - \boldsymbol{P_C})\boldsymbol{M}$ is small in some norm. In randomized column subset selection, the columns $\boldsymbol{C}$ are chosen randomly, either uniformly or according to some data-dependent distribution. Column subset selection has important applications in statistical data analysis and has been investigated by both the numerical linear algebra and the theoretical computer science communities. For an introduction to the column subset selection literature, biased towards approaches involving randomization, we refer the interested reader to the surveys [Maha, Mahb].

Our first theorem establishes that the Nyström extension of $\boldsymbol{A}$ is intimately related to the randomized column subset selection problem for $\boldsymbol{A}^{1/2}$. We model the column sampling operation as follows: let $\boldsymbol{S}$ be a random matrix with ℓ columns, each of which has exactly one nonzero element. Then right multiplication by $\boldsymbol{S}$ selects ℓ columns from $\boldsymbol{A}$:

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{S} \quad \text{and} \quad \boldsymbol{W} = \boldsymbol{S}^t \boldsymbol{A}\boldsymbol{S}.$$

The distribution of $\boldsymbol{S}$ reflects the type of sampling being performed. In the case of the simple Nyström extension, $\boldsymbol{S}$ is distributed as the first ℓ columns of a uniformly random permutation matrix.

We use the following partitioning of the eigenvalue decomposition of $\boldsymbol{A}$ to state our results:

$$\boldsymbol{A} = \overset{k \quad n-k}{\begin{bmatrix} \boldsymbol{U}_1 \ \boldsymbol{U}_2 \end{bmatrix}} \overset{k \qquad n-k}{\begin{bmatrix} \boldsymbol{\Sigma}_1 & \\ & \boldsymbol{\Sigma}_2 \end{bmatrix}} \begin{bmatrix} \boldsymbol{U}_1^t \\ \boldsymbol{U}_2^t \end{bmatrix} \tag{3}$$

The columns of $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ respectively span a dominant k-dimensional invariant subspace of $\boldsymbol{A}$ and the corresponding bottom $(n-k)$-dimensional invariant subspace of $\boldsymbol{A}$. The interaction of the column sampling matrix $\boldsymbol{S}$ with the invariant subspaces spanned by $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ is captured by the matrices

$$\boldsymbol{\Omega}_1 = \boldsymbol{U}_1^t \boldsymbol{S}, \quad \boldsymbol{\Omega}_2 = \boldsymbol{U}_2^t \boldsymbol{S}. \tag{4}$$

Theorem 1. *Let $\boldsymbol{A}$ be a PSD matrix of size n and let $\boldsymbol{S}$ be an $n \times \ell$ matrix. Partition $\boldsymbol{A}$ as in equation (3) and define $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ as in equation (4).*

Assume $\boldsymbol{\Omega}_1$ has full row rank. Then the spectral approximation error of the Nyström extension of $\boldsymbol{A}$ using $\boldsymbol{S}$ as the column sampling matrix satisfies

$$\|\boldsymbol{A} - \boldsymbol{C}\boldsymbol{W}^\dagger \boldsymbol{C}^t\|_2 = \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{S}})\boldsymbol{A}^{1/2}\|_2^2 \leq \|\boldsymbol{\Sigma}_2\|_2 \left(1 + \|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\|_2^2\right). \tag{5}$$

Prior analyses of Nyström extensions have used the Cholesky decomposition of $\boldsymbol{A}$. By instead using the square-root, Theorem 1 establishes an equivalence between the column subset selection problem and the Nyström extension procedure and gives a deterministic relative-error bound on the performance of Nyström extensions.

To establish Theorem 1, we use the following bound on the error incurred by projecting a matrix onto a random subspace of its range ([HMT11, Theorem 9.1]).

Proposition 1. *Let $\boldsymbol{M}$ be a PSD matrix of size n. Fix integers k and ℓ satisfying $1 \leq k \leq l \leq n$.*

Let $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ be matrices with orthogonal columns spanning, respectively, a dominant k-dimensional invariant subspace of $\boldsymbol{M}$ and the corresponding bottom $(n-k)$-dimensional invariant subspace of $\boldsymbol{M}$. Let $\boldsymbol{\Sigma}_2$ be the diagonal matrix of eigenvalues corresponding to the bottom $(n-k)$-dimensional invariant subspace of $\boldsymbol{M}$.

Given a matrix $\boldsymbol{S}$ of size $n \times \ell$, define $\boldsymbol{\Omega_1} = \boldsymbol{U}_1^t \boldsymbol{S}$ and $\boldsymbol{\Omega_2} = \boldsymbol{U}_2^t \boldsymbol{S}$. Then, assuming that $\boldsymbol{\Omega_1}$ has full row rank,

$$\|(\boldsymbol{I} - \boldsymbol{P_{MS}})\boldsymbol{M}\|_2^2 \leq \|\boldsymbol{\Sigma}_2\|_2^2 + \left\|\boldsymbol{\Sigma}_2 \boldsymbol{\Omega}_2 \boldsymbol{\Omega}_1^\dagger\right\|_2^2.$$

Proof of Theorem 1. We write the Nyström extension in terms of the square root of $\boldsymbol{A}$ and a projection onto the space spanned by $\boldsymbol{A}^{1/2}\boldsymbol{S}$:

$$\begin{aligned}
\boldsymbol{CW}^\dagger \boldsymbol{C}^t &= \boldsymbol{AS}(\boldsymbol{S}^t \boldsymbol{AS})^\dagger \boldsymbol{S}^t \boldsymbol{A} \\
&= \boldsymbol{A}^{1/2}[\boldsymbol{A}^{1/2}\boldsymbol{S}(\boldsymbol{S}^t \boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}\boldsymbol{S})^\dagger \boldsymbol{S}^t \boldsymbol{A}^{1/2}]\boldsymbol{A}^{1/2} \\
&= \boldsymbol{A}^{1/2}\boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{S}}\boldsymbol{A}^{1/2}.
\end{aligned}$$

It follows that the spectral error of the Nyström extension satisfies

$$\begin{aligned}
\left\|\boldsymbol{A} - \boldsymbol{CW}^\dagger \boldsymbol{C}^t\right\|_2 &= \left\|\boldsymbol{A}^{1/2}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{S}})\boldsymbol{A}^{1/2}\right\|_2 = \left\|\boldsymbol{A}^{1/2}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{S}})^2 \boldsymbol{A}^{1/2}\right\|_2 \\
&= \left\|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}^{1/2}\boldsymbol{S}})\boldsymbol{A}^{1/2}\right\|_2^2.
\end{aligned}$$

The second equality holds because of the idempotency of projections. The third follows from the fact that $\|\boldsymbol{AA}^t\|_2 = \|\boldsymbol{A}\|_2^2$ for any matrix $\boldsymbol{A}$. Partition $\boldsymbol{A}$ as in equation (3). Equation (5) now follows immediately from Proposition 1 with $\boldsymbol{M} = \boldsymbol{A}^{1/2}$. $\qquad\square$

4. Error bounds for simple Nyström extension

In this section, we provide a bound on the spectral norm approximation error of simple Nyström extensions. For convenience, we recall the following partitioning of the eigenvalue decomposition of a PSD matrix of size n:

$$\boldsymbol{A} = \begin{bmatrix} \overset{k}{} \ \overset{n-k}{} \\ \boldsymbol{U}_1 \ \boldsymbol{U}_2 \end{bmatrix} \begin{bmatrix} \overset{k}{\boldsymbol{\Sigma}_1} & \overset{n-k}{} \\ & \boldsymbol{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_1^t \\ \boldsymbol{U}_2^t \end{bmatrix}, \tag{6}$$

where the columns of $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ respectively span a dominant k-dimensional invariant subspace of $\boldsymbol{A}$ and the corresponding bottom $(n-k)$-dimensional invariant subspace of $\boldsymbol{A}$. We also recall the matrices

$$\boldsymbol{\Omega}_1 = \boldsymbol{U}_1^t \boldsymbol{S}, \quad \boldsymbol{\Omega}_2 = \boldsymbol{U}_2^t \boldsymbol{S} \tag{7}$$

that capture the interaction of the column sampling operation with the invariant subspaces of $\boldsymbol{A}$. Theorem 2 establishes that, if the spectrum of $\boldsymbol{A}$ decays sufficiently and an appropriate number of columns are sampled, then the error incurred by the simple Nyström extension process is small.

Theorem 2. *Let $\boldsymbol{A}$ be a PSD matrix of size n. Given an integer $k \leq n$, partition $\boldsymbol{A}$ as in equation (6). Let τ denote the coherence of $\boldsymbol{U}_1$,*

$$\tau = \mu_0(\boldsymbol{U}_1).$$

Fix a failure probability $\delta \in (0,1)$. For any $\varepsilon \in (0,1)$, if

$$\ell \geq \frac{2\tau k \log\left(\frac{k}{\delta}\right)}{(1 - \varepsilon)^2}$$

columns of $\boldsymbol{A}$ are chosen uniformly at random and used to form a Nyström extension, the spectral norm error of the approximation satisfies

$$\|\boldsymbol{A} - \boldsymbol{CW}^\dagger \boldsymbol{C}^t\|_2 \leq \lambda_{k+1}(\boldsymbol{A})\left(1 + \frac{n}{\varepsilon\ell}\right)$$

with probability exceeding $1 - \delta$.

Remark 1. The coherence of $\boldsymbol{U}_1$ is a measure of how much comparative influence the individual columns of $\boldsymbol{A}$ have over the dominant k-dimensional invariant subspace of $\boldsymbol{A}$ spanned by $\boldsymbol{U}_1$: if τ is small, then all columns have essentially the same influence; if τ is large, then it is possible that there is a single column in $\boldsymbol{A}$ which alone determines one of the top k eigenvectors of $\boldsymbol{A}$.

For illustrative purposes, we point out that the coherence of a random $k \times n$ orthogonal matrix, i.e. a matrix distributed uniformly on the Stiefel manifold, is $\mathrm{O}(\max(k, \log n)/k)$ with high probability [CR09]. The coherence of an arbitrary $\boldsymbol{U}_1$ is no smaller than 1, and may be as large as $\frac{n}{k}$.

Remark 2. Theorem 2, like the main result of [TR10], promises exact recovery when $\boldsymbol{A}$ is exactly rank k and has small coherence, with a sample of $\mathrm{O}(k \log k)$ columns. Unlike the result in [TR10], Theorem 2 is applicable in the case that $\boldsymbol{A}$ is full-rank but has a sufficiently fastly decaying spectrum.

Remark 3. We might ask where attempts at sharpening the analysis of the simple Nyström extension should be aimed: toward more refined linear algebra bounds on column selection (Proposition 1), or toward a deeper analysis of the randomness (Theorem 2)?

It is known that, for uniform sampling, the quantity $\left\|\boldsymbol{\Omega}_1^\dagger\right\|_2^2$ remains $\Omega(n/\ell)$ once $\ell \gtrsim \tau k \log k$ [Rud99]. Likewise, in the regime of low τ and $k \ll n$, $\boldsymbol{\Omega}_2$ is likely to be an almost isometric embedding. This follows from the fact that the columns of $\boldsymbol{\Omega}_2$ contain $n - k$ of the n entries of the corresponding columns of $\boldsymbol{U}_2$, so they are likely to be almost orthogonal and linearly independent. Together, these observations suggest that $\left\|\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^\dagger\right\|_2$ remains $\Omega(n/\ell)$ also. Thus, we expect that no bounds much sharper than Theorem 2 on the error of the simple Nyström extension can be derived using the algebraic results in Proposition 1 as the starting point.

To obtain Theorem 2, we use Theorem 1 in conjunction with a bound on $\left\|\boldsymbol{\Omega}_1^\dagger\right\|_2^2$ provided by the following lemma.

Lemma 1. *Let $\boldsymbol{U}$ be an $n \times k$ matrix with orthonormal columns. Take τ to be the coherence of $\boldsymbol{U}$,*

$$\tau = \mu_0(\boldsymbol{U}).$$

Select $\varepsilon \in (0, 1)$ and a nonzero failure probability δ. Let $\boldsymbol{S}$ be a random matrix distributed as the first ℓ columns of a uniformly random permutation matrix of size n, where

$$\ell \geq \frac{2\tau}{(1 - \varepsilon)^2} k \log \frac{k}{\delta}.$$

Then with probability exceeding $1 - \delta$, the matrix $\boldsymbol{U}^t\boldsymbol{S}$ has full row rank and satisfies

$$\left\|(\boldsymbol{U}^t\boldsymbol{S})^\dagger\right\|_2^2 \leq \frac{n}{\varepsilon l}.$$

We now proceed with the proof of Theorem 2.

Proof of Theorem 2. Because we are using uniform sampling without replacement, the sampling matrix $\boldsymbol{S}$ is formed by taking the first ℓ columns of a uniformly sampled random permutation matrix. Note that by Lemma 1, $\boldsymbol{\Omega}_1$ has full row rank, so the bounds in Theorem 1 are applicable.

Applying Lemma 1, we see that $\left\|\boldsymbol{\Omega}_1^\dagger\right\|_2^2 \leq \frac{n}{\varepsilon\ell}$ with probability exceeding $1 - \delta$. From Theorem 1, we conclude that

$$\|\boldsymbol{A} - \boldsymbol{C}\boldsymbol{W}^\dagger\boldsymbol{C}^t\|_2 \leq \left\|\boldsymbol{\Sigma}_2\right\|_2 \left(1 + \left\|\boldsymbol{\Omega}_2\right\|_2^2\left\|\boldsymbol{\Omega}_1^\dagger\right\|_2^2\right) \leq \lambda_{k+1}(\boldsymbol{A})\left(1 + \frac{n}{\varepsilon\ell}\right)$$

with at least the same probability. To obtain the second inequality, we used the fact that $\left\|\boldsymbol{\Omega}_2\right\|_2 \leq \left\|\boldsymbol{U}_2\right\|_2\left\|\boldsymbol{\Omega}\right\|_2 \leq 1$.

$\square$

One potential source of difficulty in the proof of Lemma 1 is the fact that the columns are sampled without replacement, which introduces dependencies among the entries of the sampling matrix $\boldsymbol{S}$. The following matrix Chernoff bound, a standard simplification of the lower Chernoff bound developed in [Tro11, Theorem 2.2], allows us to gloss over these dependencies.

Proposition 2. *Let $\mathcal{X}$ be a finite set of PSD matrices with dimension k, and suppose that*

$$\max_{\boldsymbol{X} \in \mathcal{X}} \lambda_1(\boldsymbol{X}) \leq B.$$

Sample $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\ell\}$ uniformly at random from $\mathcal{X}$ without replacement. Compute

$$\mu_{min} = \ell \cdot \lambda_k(\mathbb{E}\boldsymbol{X}_1).$$

Then

$$\mathbb{P}\left\{\lambda_k\left(\sum_i \boldsymbol{X}_i\right) \leq \varepsilon \mu_{min}\right\} \leq k \cdot \mathrm{e}^{-(1-\varepsilon)^2 \mu_{min}/(2B)} \quad \text{for } \varepsilon \in [0,1].$$

Proof of Lemma 1. Note that $\boldsymbol{U}^t \boldsymbol{S}$ has full row rank if $\lambda_k(\boldsymbol{U}^t \boldsymbol{S} \boldsymbol{S}^t \boldsymbol{U}) > 0$. Furthermore,

$$\left\|(\boldsymbol{U}^t \boldsymbol{S})^\dagger\right\|_2^2 = \lambda_k^{-1}(\boldsymbol{U}^t \boldsymbol{S} \boldsymbol{S}^t \boldsymbol{U}).$$

Thus to obtain both conclusions of the lemma, it is sufficient to verify that

$$\lambda_k(\boldsymbol{U}^t \boldsymbol{S} \boldsymbol{S}^t \boldsymbol{U}) \geq \frac{\varepsilon \ell}{n}$$

when ℓ is as stated.

We apply Proposition 2 to bound the probability that this inequality is not satisfied. Let $\boldsymbol{u}_i$ denote the ith column of $\boldsymbol{U}^t$. Then

$$\lambda_k(\boldsymbol{U}^t \boldsymbol{S} \boldsymbol{S}^t \boldsymbol{U}) = \lambda_k\left(\sum_{i=1}^\ell \boldsymbol{X}_i\right),$$

where the $\boldsymbol{X}_i$ are chosen uniformly at random, without replacement, from the set $\mathcal{X} = \{\boldsymbol{u}_i \boldsymbol{u}_i^t\}_{i=1,\ldots,n}$. Clearly

$$B = \max_i \|\boldsymbol{u}_i\|^2 = \frac{k}{n}\tau \quad \text{and} \quad \mu_{\min} = \ell \cdot \lambda_k(\mathbb{E}\boldsymbol{X}_1) = \frac{\ell}{n}\lambda_k(\boldsymbol{U}^t \boldsymbol{U}) = \frac{\ell}{n}.$$

Proposition 2 yields

$$\mathbb{P}\left\{\lambda_k\left(\boldsymbol{U}^t \boldsymbol{S} \boldsymbol{S}^t \boldsymbol{U}\right) \leq \varepsilon \frac{\ell}{n}\right\} \leq k \cdot \mathrm{e}^{-(1-\varepsilon)^2 \ell/(2k\tau)}.$$

We require enough samples that

$$\lambda_k(\boldsymbol{U}^t \boldsymbol{S} \boldsymbol{S}^t \boldsymbol{U}) \geq \varepsilon \frac{\ell}{n}$$

with probability greater than $1 - \delta$, so we set

$$k \cdot \mathrm{e}^{-(1-\varepsilon)^2 \ell/(2k\tau)} \leq \delta$$

and solve for ℓ, finding

$$\ell \geq \frac{2\tau}{(1-\varepsilon)^2} k \log \frac{k}{\delta}.$$

Thus, for values of ℓ satisfying this inequality, we achieve the stated spectral error bound and ensure that $\boldsymbol{U}^t \boldsymbol{S}$ has full row rank. $\qquad \square$

References

[BF]   A. L. Bertozzi and A. Flenner, *Diffuse interface models on graphs for classification of high dimensional data*, submitted 2011.

[Bha97]   R. Bhatia, *Matrix Analysis*, Springer-Verlag, 1997.

[BW09]   M. Belabbas and P. J. Wolfe, *Spectral methods in machine learning and new strategies for very large datasets*, Proc. Nat. Acad. Sci. **106** (2009), 369–374.

[CD]   J. Chiu and L. Demanet, *Sublinear randomized algorithms for skeleton decompositions*, Preprint, arXiv:1110.4193, October 2011.

[CR07]   E. Candés and J. Romberg, *Sparsity and incoherence in compressive sampling.*, Inverse Problems **23** (2007), 969–986.

[CR09]   E. Candés and B. Recht, *Exact Matrix Completion via Convex Optimization*, Found. Comput. Math. **9** (2009), 717–772.

[DKM06]   P. Drineas, R. Kannan, and M. W. Mahoney, *Fast Monte Carlo Algorithms for Matrices III: Computing an Efficient Approximate Decomposition of a Matrix*, SIAM J. Comput. **36** (2006), 184–206.

[DM05]   P. Drineas and M. W. Mahoney, *On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning*, J. Mach. Learn. Res. **6** (2005), 2153–2175.

[DMM08]   P. Drineas, M. W. Mahoney, and S. Muthukrishnan, *Relative-error CUR Matrix Decompositions*, SIAM J. Matrix Anal. Appl. **30** (2008), 844–881.

[FBCM04]   C. Fowlkes, S. Belongie, F. Chung, and J. Malik, *Spectral Grouping Using the Nyström Method*, IEEE Trans. Pattern Anal. Mach. Intell. **26** (2004), 214–225.

[FGK11]   A. K. Farahat, A. Ghodsi, and M. S. Kamel, *A novel greedy algorithm for Nyström approximation*, Proc. 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011), 2011.

[HM]   D. Homrighausen and D. J. McDonald, *Spectral approximations in machine learning*, Preprint arXiv:1107.4340. July 2011.

[HMT11]   N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding Structure With Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions*, SIAM Rev. **53** (2011), 217–288.

[KMT09a]   S. Kumar, M. Mohri, and A. Talwalkar, *On sampling-based approximate spectral decomposition*, Proc. 26th International Conference on Machine Learning (ICML 2009), 2009.

[KMT09b]   ———, *Sampling Techniques for the Nyström Method*, Conference on Artificial Intelligence and Statistics, 2009, pp. 304–311.

[LKL10]   M. Li, J. T. Kwok, and B. Lu, *Making Large-Scale Nyström Approximation Possible*, Proc. 27th International Conference on Machine Learning (ICML 2010), 2010, pp. 1097–1104.

[Maha]   M. Mahoney, *Algorithmic and Statistical Perspectives on Large-Scale Data Analysis*, Preprint, arXiv:1010.1609, 2010. To appear in Combinatorial Scientific Computing, U. Naumann and O. Schenk, eds., 2011.

[Mahb]   ———, *Randomized algorithms for matrices and data*, Preprint, arXiv:1104.5557, 2011.

[MTJ]   L. Mackey, A. Talwalkar, and M. I. Jordan, *Divide-and-Conquer Matrix Factorization*, Preprint arXiV:1107.0789, August 2011.

[Rud99]   M. Rudelson, *Random Vectors in the Isotropic Position*, J. Funct. Anal. **164** (1999), no. 1, 60–72.

[TGZ97]   E. E. Tyrtyshnikov, S. A. Goreinov, and N. L. Zamarashkin, *A theory of pseudoskeleton approximations*, Linear Algebra Appl. **261** (1997), 1–21.

[TR10]   A. Talwalkar and A. Rostamizadeh, *Matrix Coherence and the Nyström Method*, Proc. 26th Conference on Uncertainty in Artifical Intelligence (UAI 2010), 2010.

[Tro11]   J. Tropp, *Improved analysis of the subsampled randomized Hadamard transform*, Adv. Adapt. Data Anal. **3** (2011), 115–126.

[WDT$^+$09] J. Wang, Y. Dong, X. Tong, Z. Lin, and B. Guo, *Kernel Nyström method for light transport*, ACM Trans. Graph. **28** (2009), 29:1–29:10.

[WS01]   C. Williams and M. Seeger, *Using the Nyström method to speed up kernel machines*, Advances in Neural Information Processing Systems 13, 2001.