THE SPECTRAL ERROR OF THE SIMPLE NYSTRÖM EXTENSION 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 shows that the spectral norm error incurred during this process is within a multiplicative factor of the error of the optimal rank-k approximant. The dependence of the multiplicative factor on the dimension of the matrix and the number of columns sampled is optimal. Another bound guarantees smaller error when spectral decay is present. These results flow from a natural connection between Nyström extension and the column subset selection problem. Two regularized versions of the simple Nyström algorithm are introduced to eliminate instabilities caused by the potential pseudoinversion of ill-conditioned matrices. When the regularization parameters are chosen appropriately, the errors of the regularized Nyström approximations are at most a small multiple of that of the simple Nyström approximation.

1. Introduction

The truncated singular value decomposition (SVD) is the classical tool for low-rank approximation of matrices. A rank-k truncated SVD of a square matrix can be computed from a rank-revealing QR decomposition, but the initial QR decomposition costs $\mathrm{O}(n^3)$ operations [CH92]. For small k and large n, Krylov space methods can potentially provide truncated SVDs in much less time. In practice, the number of operations required varies considerably depending upon the specifics of the method and the spectral properties of the matrix, but since one must perform at least k dense matrix-vector multiplies, computing a rank-k truncated SVD using a Krylov method requires at least $\Omega(kn^2)$ operations. Methods based on low-rank approximation have become popular because they capture the low-dimensional structure implicit in massive high-dimensional modern datasets. Low-rank approximations are also used for their noise-elimination and regularization properties [Han90]. Among many applications, we mention PCA [HTF08], multidimensional scaling [CC00], collaborative filtering [SAJ10], manifold learning [HLMS04], and latent semantic indexing [DDF$^+$90]. Due to the large size of modern datasets, much interest has been expressed in finding $\mathrm{o}(kn^2)$ 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 $\mathbf{CUR}$, where $\mathbf{C}$ and $\mathbf{R}$ are respectively small subsets of the columns and rows of the matrix and $\mathbf{U}$, the *coupling matrix*, is a generalized inverse of the overlap of $\mathbf{C}$ and $\mathbf{R}$ or the restriction of $\mathbf{A}$ to the column span of $\mathbf{C}$ and the row span of $\mathbf{R}$ [DKM06]. Accordingly, these schemes are known as CUR decompositions. If the matrix is sparse, then so are $\mathbf{C}$ and $\mathbf{R}$, so CUR decompositions have found applications in areas where the preservation of sparsity is important [SXZF07]. Because they represent the matrix in terms of linear combinations of only a few of its rows and columns, CUR decompositions are preferable to other low-rank approximation techniques in settings where interpretability of the decomposition is important [DM09, HMT08].

In general, $\mathbf{C}$ and $\mathbf{R}$ can be chosen independently, e.g. both could be chosen uniformly at random. However, it may be advantageous to require $\mathbf{C}$ and $\mathbf{R}$ to be related. In particular, if the matrix is

Research supported, under the auspice of Joel Tropp, by ONR awards N00014-08-1-0883 and N00014-11-1-0025, AFOSR award FA9550-09-1-0643, and a Sloan Fellowship. In addition to his financial support, the author thanks Joel for fruitful discussions on the interpretation of these error bounds.

symmetric positive semidefinite (SPSD), a low-rank approximation that is also SPSD can be formed by selecting the same rows as columns (i.e., by taking $\mathbf{R} = \mathbf{C}^T$). 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 [WS01]. Nyström extensions are often used to increase the speed of kernelized machine learning algorithms or to make it feasible to apply such algorithms to massive datasets [CMT10, WS01, FBCM04, TKR08, KPSH07].

We consider the simple Nyström extension, a particular scheme in which the columns are sampled uniformly at random without replacement. This paper presents a simple framework for the analysis of Nyström schemes based upon a natural connection between Nyström extensions and the column subset selection problem. This framework yields state-of-the-art spectral norm error bounds in the case of the simple Nyström extension scheme, as well as an understanding of what form such bounds *should* take. Our first result is the first truly relative-error spectral norm bound available for any Nyström extension method—in that it bounds the error of the Nyström extension to within a multiple of that of the optimal rank-k approximation—, and generalizes the coherence-based exact recovery result in [TR10]. Our second result guarantees small error in the case of a matrix with fast spectral decay. Because the simple Nyström extension cannot be computed stably in general, we also introduce two stable regularized Nyström approximation algorithms.

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, BF12]. 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 $\mathbf{A}$ be a real SPSD matrix of size n. Select $\ell \ll n$ columns of $\mathbf{A}$ to constitute the columns of a matrix $\mathbf{C}$. Let $\mathbf{W}$ be the $\ell \times \ell$ coupling matrix formed by the intersection of the columns in $\mathbf{C}$ and the corresponding rows in $\mathbf{A}$. The matrix $\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$ is then a Nyström extension of $\mathbf{A}$ (see Figure 1). Here $(\cdot)^\dagger$ denotes Moore-Penrose pseudoinversion. Since $\mathbf{W}$ is a principal submatrix of $\mathbf{A}$, it is positive-semidefinite, and hence the Nyström extension is also positive-semidefinite.

The manner in which the columns are sampled and $\mathbf{W}^\dagger$ 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 $\mathbf{A}$ [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 eigenspaces of a real SPSD matrix $\mathbf{A}$, rather than a low-rank approximation [HM]. In these applications the spectral norm approximation error is of interest, as it

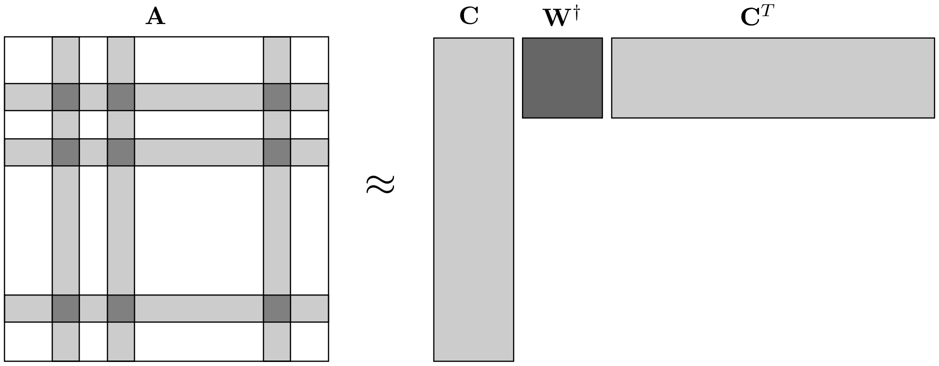


Figure 1. The Nyström extension procedure

provides information on the quality of the approximate dominant eigenspaces obtained via Nyström extensions.

To be more precise, denote the ith-largest eigenvalue of an SPSD matrix $\mathbf{M}$ by $\lambda_i(\mathbf{M})$, so that $\lambda_1(\mathbf{M}) \geq \lambda_2(\mathbf{M}) \geq \ldots$. Let $\tilde{\mathbf{A}}$ be a Nyström approximation to $\mathbf{A}$ and assume both $\mathbf{A}$ and $\tilde{\mathbf{A}}$ have well-defined dominant k-dimensional eigenspaces; that is, assume that $\lambda_k(\mathbf{A}) > \lambda_{k+1}(\mathbf{A})$ and likewise for $\tilde{\mathbf{A}}$. Let $\mathbf{U}_k$ and $\tilde{\mathbf{U}}_k$ have orthonormal columns and span, respectively, the dominant k-dimensional eigenspace of $\mathbf{A}$ and that of $\tilde{\mathbf{A}}$. Recall one natural definition for the distance between these subspaces [GV96, Section 2.6.3],

$$\operatorname{dist}(\mathbf{U}_k, \tilde{\mathbf{U}}_k) = \left\|\mathbf{P}_{\mathbf{U}_k} - \mathbf{P}_{\tilde{\mathbf{U}}_k}\right\|_2 = \left\|\mathbf{U}_k(\tilde{\mathbf{U}_k}^{\perp})^T\right\|_2.$$

Here $\tilde{\mathbf{U}_k}^{\perp}$ is a matrix with orthonormal columns spanning the bottom $(n-k)$-dimensional eigenspace of $\tilde{\mathbf{A}}$ and $\mathbf{P}_{\mathbf{U}_k}$ ($\mathbf{P}_{\tilde{\mathbf{U}}_k}$) denotes the projection onto the span of $\mathbf{U}_k$ ($\tilde{\mathbf{U}}_k$).

Since $\mathbf{A}$ and $\tilde{\mathbf{A}}$ are both normal, it follows from [Bha97, Theorem VII.3.1] that

$$\operatorname{dist}(\mathbf{U}_k, \tilde{\mathbf{U}}_k) \leq \frac{\left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_2}{\lambda_k(\mathbf{A}) - \lambda_{k+1}(\tilde{\mathbf{A}})} \leq \frac{\left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_2}{\lambda_k(\mathbf{A}) - \lambda_{k+1}(\mathbf{A}) - \left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_2} \tag{1}$$

when $\lambda_k(\mathbf{A}) - \lambda_{k+1}(\mathbf{A}) - \left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_2$ is positive. Hence, if $\left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_2$ is smaller than the eigengap $\lambda_k(\mathbf{A}) - \lambda_{k+1}(\mathbf{A})$, we can estimate the error in approximating the dominant k-dimensional eigenspace of $\mathbf{A}$ with that of $\tilde{\mathbf{A}}$. This observation provides an additional motivation for seeking tighter bounds on the spectral error $\left\|\mathbf{A} - \tilde{\mathbf{A}}\right\|_2$ of the simple Nyström extension.

1.4. **Our error bounds.** 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. We use the concept of *coherence*, taken from the matrix completion literature [CR09], to provide a quantitative measure of the informativity of the columns of $\mathbf{A}$. Let $\mathcal{S}$ be a k-dimensional subspace of $\mathbb{R}^n$ and $\mathbf{P}_{\mathcal{S}}$ denote the projection onto $\mathcal{S}$. Then the coherence of $\mathcal{S}$ is

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

The coherence of the dominant k-dimensional eigenspace of $\mathbf{A}$ is a measure of how much comparative influence the individual columns of $\mathbf{A}$ have on this subspace: if μ is small, then all columns have essentially the same influence; if μ is large, then it is possible that there is a single column in $\mathbf{A}$ which alone determines one of the top k eigenvectors of $\mathbf{A}$. We mention that if $\mathbf{A}$ is rank-k, then the quantities $(\mathbf{P}_{\mathcal{S}})_{ii}$ are known to statisticians as the leverage scores of the columns of $\mathbf{A}$ [DM10].

For illustrative purposes, we point out that the coherence of a random $n \times k$ orthonormal matrix, i.e. a matrix distributed uniformly on the Stiefel manifold, is $\mathrm{O}(\max(k, \ln n)/k)$ with high probability [CR09, Lemma 2.2]. The coherence of an arbitrary k-dimensional subspace is no smaller than 1, and may be as large as n/k.

Corollary 1 is a condensed version of our main result, Theorem 3, and uses the notion of coherence to provide a bound on the error of the simple Nyström extension.

Corollary 1. *Let* $\mathbf{A}$ *be a real SPSD matrix of size* n. *Given an integer* $k \leq n$, *let* μ *denote the coherence of a dominant* k*-dimensional invariant subspace of* $\mathbf{A}$. *Fix a nonzero failure probability* $\delta > 0$. *If* $\ell \geq 8\mu k \ln(k/\delta)$ *columns of* $\mathbf{A}$ *are chosen uniformly at random without replacement, then*

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

with probability exceeding $1-\delta$ *and*

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right\|_2 \leq \lambda_{k+1}(\mathbf{A}) + \frac{2}{\delta} \cdot \sum\nolimits_{i>k} \lambda_i(\mathbf{A})$$

with probability exceeding $1-2\delta$. *If, additionally,* $k \geq \operatorname{rank}(\mathbf{A})$, *then*

$$\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$$

with probability exceeding $1-\delta$.

The dependence of the relative error bound provided in Corollary 1 is optimal on n and ℓ; this is established in Section 4.

For Corollary 1 to provide a meaningful estimate of ℓ, the required number of column samples, the coherence 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 $\mu = 1$, the columns are equally informative and we find that the error of a simple Nyström extension formed using just $8k\ln(k/\delta)$ columns is a bounded multiple of the error of the optimal rank-k approximation. Furthermore, we see that the efficacy of the simple Nyström extension increases in the presence of spectral decay. If the eigenvalues of $\mathbf{A}$ decay sufficiently then the error of the simple Nyström extension may even be bounded by a small multiple of the error of the optimal rank-k approximation.

1.5. **Stable algorithms.** In general, the computation of the simple Nyström extension is unstable because it requires one to take the pseudoinverse of a matrix $\mathbf{W}$ which may be nearly singular. The product $\mathbf{W}^\dagger\mathbf{C}^T$ can be computed stably if the condition number $\kappa(\mathbf{W}) = \lambda_1(\mathbf{W})/\lambda_\ell(\mathbf{W})$ is small (here $\mathbf{W}$ is an $\ell \times \ell$ matrix) [GV96]. The maximum eigenvalue of $\mathbf{W}$ is smaller[1] than that of $\mathbf{A}$, so when the minimum eigenvalue of $\mathbf{W}$ is sufficiently bounded away from zero the condition number of $\mathbf{W}$ is small and the simple Nyström extension can be stably computed. If we do not have *a priori* knowledge that the condition number of $\mathbf{W}$ is small, one should instead compute an approximation to the simple Nyström extension using an algorithm that ensures that $\lambda_\ell(\mathbf{W})$ is not small. We propose two such algorithms.

Algorithm 1 returns a simple Nyström extension of $\mathbf{A}_\rho := \mathbf{A} + \rho\mathbf{I}$ for some small positive constant ρ. Since $\mathbf{A}_\rho$ has a minimum eigenvalue bounded away from zero by at least ρ, the computation of the simple Nyström extension is stable: one can use, for instance, a Cholesky decomposition to compute the required inversion stably.

Algorithm 2 returns a Nyström extension where the matrix $\mathbf{C}$ is taken from $\mathbf{A}$, but the coupling matrix $\mathbf{W}_\rho$ is taken from $\mathbf{A}_\rho$. Intuitively, for small ρ, the approximation returned by Algorithm 2, $\mathbf{C}\mathbf{W}_\rho^{-1}\mathbf{C}^T$, should be a better approximation to $\mathbf{A}$ than that returned by Algorithm 1, $\mathbf{C}_\rho\mathbf{W}_\rho^{-1}\mathbf{C}_\rho^T$,

[1] Matrix concentration inequalities can be used to estimate $\lambda_1(\mathbf{W})$ more precisely; see e.g. [Tro12, MJC$^+$].

Algorithm 1: Nyström extension of a regularized $\mathbf{A}$

```
Input: $\mathbf{A}$, an $n \times n$ SPSD matrix; an integer $k \in [1, n]$; a regularization parameter $\rho > 0$; and an integer $\ell \geq k$.
Output: $\mathbf{C}_\rho$, an $n \times \ell$ matrix; and $\mathbf{W}_\rho$, an $\ell \times \ell$ SPSD matrix.

Let $\mathbf{A}_\rho = \mathbf{A} + \rho\mathbf{I}$.
Uniformly select, without replacement, $\ell$ indices in $\{1, \ldots, n\}$ and let $\mathbf{S} \in \mathbb{R}^{n \times \ell}$ be the corresponding column selector matrix. That is, let $(\mathbf{S})_{i,j} = 1$ if $i$ was the $j$th index selected, and $(\mathbf{S})_{i,j} = 0$ otherwise.
Form the matrix $\mathbf{C}_\rho = \mathbf{A}_\rho\mathbf{S}$.
Form the matrix $\mathbf{W}_\rho = \mathbf{S}^T\mathbf{A}_\rho\mathbf{S}$.
```

because it uses more information from $\mathbf{A}$. However, it is not clear that this intuition should hold for larger ρ: the added accuracy from using $\mathbf{C}$ may be negated by the fact that $\mathbf{C}$ and $\mathbf{W}_\rho$ are from two different matrices, so $\mathbf{C}\mathbf{W}_\rho^{-1}\mathbf{C}^T$ can no longer be thought of as a Nyström extension of a matrix close to $\mathbf{A}$, whereas $\mathbf{C}_\rho\mathbf{W}_\rho^{-1}\mathbf{C}_\rho^T$ is always a Nyström extension of a matrix that is close to $\mathbf{A}$.

Algorithm 2: Nyström extension using a regularized coupling matrix

```
Input: $\mathbf{A}$, an $n \times n$ SPSD matrix; an integer $k \in [1, n]$; a tolerance $\rho > 0$; and an integer $\ell \geq k$.
Output: $\mathbf{C}$, an $n \times \ell$ matrix; and $\mathbf{W}_\rho$, an $\ell \times \ell$ SPSD matrix.

Uniformly select, without replacement, $\ell$ indices in $\{1, \ldots, n\}$ and let $\mathbf{S} \in \mathbb{R}^{n \times \ell}$ be the corresponding column selector matrix. That is, let $(\mathbf{S})_{i,j} = 1$ if $i$ was the $j$th index selected, and $(\mathbf{S})_{i,j} = 0$ otherwise.
Form the matrix $\mathbf{C} = \mathbf{A}\mathbf{S}$.
Form the matrix $\mathbf{W} = \mathbf{S}^T\mathbf{A}\mathbf{S}$.
If $\lambda_{\min}(\mathbf{W}) \geq \rho$, then take $\mathbf{W}_\rho = \mathbf{W}$,
    otherwise take $\mathbf{W}_\rho = \mathbf{W} + \rho\mathbf{I}$.
```

The cost of forming the Nyström extension returned by each of these algorithms is $\mathrm{O}(\ell^3 + \ell^2 n)$. We have the following accuracy guarantees.

Theorem 1. *Let $\mathbf{A}$ be a real SPSD matrix of size n. Given an integer $k \leq n$ and a tolerance $\rho > 0$, let μ denote the coherence of a dominant k-dimensional invariant subspace of $\mathbf{A}$. Fix a nonzero failure probability δ.*

If $\ell \geq 8\mu k \ln(k/\delta)$ then the Nyström extension returned by Algorithm 1 satisfies

$$\left\|\mathbf{A} - \mathbf{C}_\rho\mathbf{W}_\rho^{-1}\mathbf{C}_\rho^T\right\|_2 \leq \lambda_{k+1}(\mathbf{A})\left(1 + \frac{2n}{\ell}\right) + \frac{2\rho n}{\ell} + \rho.$$

with probability at least $1 - \delta$ and

$$\left\|\mathbf{A} - \mathbf{C}_\rho\mathbf{W}_\rho^{-1}\mathbf{C}_\rho^T\right\|_2 \leq \lambda_{k+1}(\mathbf{A}) + \frac{2}{\delta}\sum\nolimits_{i=k+1}^{n} \lambda_i(\mathbf{A}) + \frac{2\rho(n-k)}{\delta} + 2\rho$$

with probability at least $1 - 2\delta$.

Likewise, if $\ell \geq 8\mu k \ln(k/\delta)$ then the Nyström extension returned by Algorithm 2 satisfies

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}_\rho^{-1}\mathbf{C}^T\right\|_2 \leq \lambda_{k+1}(\mathbf{A})\left(3 + \frac{2n}{\ell}\right) + \frac{2\rho n}{\ell} + 4\rho\sqrt{\frac{n}{\ell}} + 5\rho$$

with probability at least $1-\delta$ *and*

$$\left\|\mathbf{A}-\mathbf{C}\mathbf{W}_{\rho}^{-1}\mathbf{C}^{T}\right\|_{2} \leq 3\lambda_{k+1}(\mathbf{A})+\frac{2}{\delta}\sum\nolimits_{i=k+1}^{n}\lambda_{i}(\mathbf{A})+\frac{2\rho(n-k)}{\delta}+4\rho\sqrt{\frac{n}{\ell}}+5\rho$$

with probability at least $1-2\delta$.

1.6. **Relevant literature.** We survey the literature on CUR decompositions and Nyström extensions, focusing on the simple Nyström scheme. Discussion of the column subset selection literature is deferred to Section 3, after the connection between the Nyström extension procedure and the column subset selection problem is made explicit. In this section $\mathbf{A}$ is an $n\times n$ SPSD matrix, $\mathbf{A}_k$ is a rank-k approximation to $\mathbf{A}$ that is optimal in the spectral norm, ℓ is the number of columns used to construct a Nyström extension of $\mathbf{A}$, $\delta \in (0,1)$ is a failure probability, and $\varepsilon > 0$ is an accuracy parameter.

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 $\left\|\mathbf{A}-\mathbf{A}_k\right\|_2(1+4\sqrt{kn})$. Finding these columns and rows is connected to the problem of finding a $k\times k$ submatrix of $\mathbf{A}$ whose determinant has the maximal modulus [GZT97].

Drineas, Kannan, and Mahoney [DKM06] show that if appropriate probability distributions are used to select the columns and rows, the accuracy of the resulting CUR approximations does not depend on the dimensions k and n. Specifically, if the columns and rows are sampled with probability proportional to their Euclidean norms, then the spectral and Frobenius norm errors of the approximations are within an additive factor of $\varepsilon\|\mathbf{A}\|_F$ of the optimal errors. Let c and r denote the number of columns and rows sampled to form $\mathbf{C}$ and $\mathbf{R}$ respectively, then

$$\begin{aligned}\left\|\mathbf{A}-\mathbf{CUR}\right\|_2 &\leq \left\|\mathbf{A}-\mathbf{A}_k\right\|_2+\varepsilon\left\|\mathbf{A}\right\|_{\mathrm{F}}\\ \left\|\mathbf{A}-\mathbf{CUR}\right\|_{\mathrm{F}} &\leq \left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}}+\varepsilon\left\|\mathbf{A}\right\|_{\mathrm{F}}\end{aligned}$$

with probability at least $1-\delta$ when $c=\Omega(k\varepsilon^{-4}\ln\delta^{-1})$ and $r=\Omega(k\varepsilon^{-2}\delta^{-2})$. The coupling matrix $\mathbf{U}$ is computed from $\mathbf{C}$ and $\mathbf{R}$.

In [DMM08], Drineas, Mahoney, and Muthukrishnan propose forming CUR decompositions by using "subspace sampling" probabilities that capture the influence of the columns of $\mathbf{A}$ on the top k-dimensional eigenspace of $\mathbf{A}$: if $\mathbf{P}$ is the projection onto this eigenspace, then the probability of selecting the ith column is taken to be proportional to $(\mathbf{P})_{ii}$. Later works refer to these probabilities as leverage score probabilities to emphasize their connection to the leverage scores encountered in linear regression problems [MD09, DM10]. [DMM08] gives a $(1+\epsilon)$ Frobenius norm relative-error guarantee on the quality of the CUR approximations: if $c=\Omega(k^2\varepsilon^{-2}\ln\delta^{-1})$ columns are sampled from $\mathbf{A}$ with replacement according to the leverage score distribution of $\mathbf{A}$ and $r=\Omega(c^2\varepsilon^{-2}\ln\delta^{-1})$ rows are then sampled from $\mathbf{A}$ according to the leverage score distribution of $\mathbf{C}^T$, then

$$\left\|\mathbf{A}-\mathbf{CUR}\right\|_{\mathrm{F}} \leq (1+\varepsilon)\left\|\mathbf{A}-\mathbf{A}_k\right\|_{\mathrm{F}}$$

with probability at least $1-\delta$. Here the coupling matrix $\mathbf{U}$ is the pseudoinverse of the intersection between $\mathbf{R}$ and $\mathbf{C}$.

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 $\mathbf{A}$. They show that if $\mathbf{C}$ is constructed by sampling $\Omega(k\varepsilon^{-4}\ln\delta^{-1})$ columns with replacement from this distribution, then

$$\|\mathbf{A}-\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_2 \leq \|\mathbf{A}-\mathbf{A}_k\|_2 + \varepsilon \sum\nolimits_{k=1}^{n} (\mathbf{A})_{ii}^2$$
$$\|\mathbf{A}-\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_{\mathrm{F}} \leq \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}} + \varepsilon \sum\nolimits_{k=1}^{n} (\mathbf{A})_{ii}^2$$

with probability at least $1-\delta$ [DM05]. 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

$$\|\mathbf{A}-\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_{\mathrm{F}} \leq \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}} + \varepsilon n \cdot \max_i\,(\mathbf{A})_{ii}$$

with probability at least $1-\delta$ when $\ell = \Omega(k\varepsilon^{-4}\ln\delta^{-1})$.

In [KMT09a], Kumar, Mohri, and Talwalkar establish that if $\operatorname{rank}(\mathbf{W}) = \operatorname{rank}(\mathbf{A}) = k$, then $\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$. Talwalkar and Rostamizadeh prove this implies that, if $\ell = \Omega(k\tau\ln(k/\delta))$, simple Nyström extension results in exact recovery with probability at least $1-\delta$ [TR10]. That is, if ℓ is sufficiently large, then the Nyström extension is exactly $\mathbf{A}$. Here τ is a different measure of the coherence of the top k-dimensional eigenspace of $\mathbf{A}$ than is used in this paper: let the columns of $\mathbf{U}_1$ be the normalized top k eigenvectors of $\mathbf{A}$, then $\tau = n\max_{i,j}\left(\mathbf{U}_1\right)_{ij}^2$. Their key observation is that if no columns of $\mathbf{A}$ are singularly influential, then $\mathbf{W}$ will have maximal rank when ℓ is slightly larger than the rank of $\mathbf{A}$. Thus, the number of samples required for exact recovery is determined by the rank of $\mathbf{A}$ and the coherence, τ, 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 $\mathbf{W}$ with a low-rank approximation $\tilde{\mathbf{W}}$ 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 $\tilde{\mathbf{W}}$ is constructed using the randomized methodology espoused in [HMT11]. Namely, to construct $\tilde{\mathbf{W}}$, one first forms the product $\mathbf{Y} = \mathbf{W}^q\mathbf{\Omega}$ by applying $\mathbf{W}$ multiple times to a random $\ell \times (k+p)$ matrix $\mathbf{\Omega}$ whose entries are standard Gaussian random variables. Intuitively, the range space of $\mathbf{Y}$ captures the top k-dimensional eigenspace of $\mathbf{W}$, so to form $\tilde{\mathbf{W}}$ one lets $\mathbf{Q}$ be an orthonormal basis for the range space of $\mathbf{Y}$ and takes $\tilde{\mathbf{W}} = \mathbf{Q}\mathbf{Q}^T\mathbf{W}\mathbf{Q}\mathbf{Q}^T$ to be the restriction of $\mathbf{W}$ to the range space of $\mathbf{Y}$. The pseudoinverse of $\tilde{\mathbf{W}}$ is potentially much cheaper to compute than that of $\mathbf{W}$, since one only has to invert a smaller $k+p$ matrix: $\tilde{\mathbf{W}}^\dagger = \mathbf{Q}(\mathbf{Q}^T\mathbf{W}\mathbf{Q})^\dagger\mathbf{Q}^T$.

The analysis of Kwok et al.'s scheme 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 $\tilde{\mathbf{W}}$: the oversampling factor p and the number of iterations q. The results of [LKL10] provide error bounds for the simple Nyström extension (take $p = \ell - k$ and $q = 1$):

$$\mathbb{E}\|\mathbf{A}-\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_2 \leq \mathrm{C}\left(\|\mathbf{A}-\mathbf{A}_k\|_2 + \frac{n}{\sqrt{\ell}}\max_i(\mathbf{A})_{ii}\right) \tag{2}$$
$$\mathbb{E}\|\mathbf{A}-\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_F \leq \mathrm{C}\big(\sqrt{\ell}\cdot\|\mathbf{A}-\mathbf{A}_k\|_F + n\cdot\max_i(\mathbf{A})_{ii}\big).$$

Here the constant C depends on the relationship between k and ℓ; if, e.g., $\ell = k\ln k$ and $k > 6$ then C can be taken to be 8.

Of the works mentioned, only [LKL10] provides a bound on the spectral error of the simple Nyström extension for $\mathbf{A}$ of arbitrary rank. Unfortunately, the quantity $\max_i(\mathbf{A})_{ii}$ is, for a general

SPSD $\mathbf{A}$, bounded only by $\lambda_1(\mathbf{A})$. 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.

Most recently, in a simultaneously completed work, Chiu and Demanet consider CUR decompositions of rectangular matrices that are formed by sampling uniformly at random with replacement from the columns and rows of the matrix [CD]. Using a different technique, they establish bounds similar to those given in Theorem 1. Their proposed CUR algorithm, applied to SPSD matrices, sets the eigenvalues of $\mathbf{W}$ that are smaller than ρ to zero, thereby ensuring that $\left\|\mathbf{W}^\dagger\right\|_2 \leq \rho^{-1}$ and allowing the Nyström extension to be computed stably (see Section 1.5).

The results in [CD] imply that if $\rho = \lambda_{k+1}(\mathbf{A})$ then, with high probability, the error of a Nyström extension constructed using this algorithm stays within a multiplicative $\mathrm{O}(n/\ell)$ factor of the optimal rank-k approximation error. We note that Theorem 1 gives the same guarantee for Algorithms 1 and 2 for the same choice of ρ. However, in general one does not know $\lambda_{k+1}(\mathbf{A})$, so it is instructive to compare the bounds from [CD] with our bounds, for arbitrary regularization parameters ρ. According to the first bound in Theorem 1, the error of the extensions calculated by Algorithms 1 and 2 is on the order of

$$(\lambda_{k+1}(\mathbf{A}) + \rho)\frac{n}{\ell},$$

while the corresponding bound for the algorithm proposed in [CD] has an error on the order of

$$\left(\lambda_{k+1}(\mathbf{A}) + \rho + \frac{\lambda_{k+1}(\mathbf{A})^2}{\rho}\right)\frac{n}{\ell},$$

which may be much larger when $\rho < \lambda_{k+1}(\mathbf{A})$.

In contrast to prior work, this work (concurrently with [CD]) supplies the first truly relative error spectral norm approximation guarantee for the simple Nyström approximation, and establishes an explicit connection between the coherence of the dominant k-dimensional eigenspace of the matrix and the number of column samples needed to obtain this guarantee. A spectral norm error bound is also provided that, in the presence of spectral decay, gives smaller additive error than available in the current literature. We show the optimality of the dependence of the relative error bound on the dimension of the matrix and the number of column samples used in forming the Nyström extension.

Our results are obtained by establishing an explicit connection between the randomized column subset selection problem and the quality of Nyström extensions. This connection is of independent interest, as it implies that any guarantee on the error of randomized column subset selection directly implies a matching guarantee on the error of Nyström extension, and vice versa. While this paper considers only the spectral norm error of the simple Nyström extension scheme, we believe the framework given is flexible enough to be fruitfully applied to the analysis of errors measured in other norms and of other Nyström schemes including, in particular, the large-scale variant introduced in [LKL10].

1.7. **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. We conclude with an empirical investigation of Algorithms 1 and 2 in Section 6.

2. Notation

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

The projection onto the column space of a matrix $\mathbf{M}$ is written $\mathbf{P}_\mathbf{M}$ and satisfies

$$\mathbf{P}_\mathbf{M} = \mathbf{M}\mathbf{M}^\dagger = \mathbf{M}(\mathbf{M}^T\mathbf{M})^\dagger\mathbf{M}^T.$$

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

The coherence of a matrix $\mathbf{U} \in \mathbb{R}^{n\times k}$ with orthonormal columns is the coherence of the subspace $\mathcal{S}$ which it spans:

$$\mu(\mathbf{U}) := \mu(\mathcal{S}) = \frac{n}{k}\max\nolimits_i(\mathbf{P}_\mathcal{S})_{ii} = \frac{n}{k}\max\nolimits_i(\mathbf{U}\mathbf{U}^T)_{ii}.$$

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 $\mathbf{M}$, the goal of column selection is to choose a small but informative subset $\mathbf{C}$ of the columns of $\mathbf{M}$. Informativity can be defined in many ways; in our context, $\mathbf{C}$ is informative if, after approximating $\mathbf{M}$ with the matrix obtained by projecting $\mathbf{M}$ onto the span of $\mathbf{C}$, the residual $(\mathbf{I}-\mathbf{P}_\mathbf{C})\mathbf{M}$ is small in some norm. In randomized column subset selection, the columns $\mathbf{C}$ 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 [Mah12, Mah11].

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

$$\mathbf{C} = \mathbf{A}\mathbf{S} \quad \text{and} \quad \mathbf{W} = \mathbf{S}^T\mathbf{A}\mathbf{S}.$$

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

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

$$\mathbf{A} = \begin{array}{c} \begin{array}{cc} \scriptstyle k & \scriptstyle n-k \end{array} \\ \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{cc} \scriptstyle k & \scriptstyle n-k \end{array} \\ \begin{bmatrix} \mathbf{\Sigma}_1 & \\ & \mathbf{\Sigma}_2 \end{bmatrix} \end{array} \begin{bmatrix} \mathbf{U}_1^T \\ \mathbf{U}_2^T \end{bmatrix} \tag{3}$$

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

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

Theorem 2. *Let $\mathbf{A}$ be an SPSD matrix of size n and let $\mathbf{S}$ be an $n\times\ell$ matrix. Partition $\mathbf{A}$ as in equation* (3) *and define $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ as in equation* (4)*.*

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

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

If $\mathrm{rank}(\mathbf{A}) < k$*, then in fact*

$$\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T.$$

Remark 1. We emphasize that the first relation in (5) is an equality. That is, the spectral norm error of approximating $\mathbf{A}$ with a Nyström extension is exactly the square of the spectral norm error of approximating $\mathbf{A}^{1/2}$ with a specific corresponding low-rank approximant.

In fact, this equality holds when $\mathbf{A}^{1/2}$ is replaced with any generalized Cholesky factorization of $\mathbf{A}$: by appropriately modifying the proof of Theorem 2, it can be seen that if $\mathbf{P}\mathbf{A}\mathbf{P}^T = \mathbf{B}^T\mathbf{B}$ where $\mathbf{P}$ is a permutation matrix, then

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right\|_2 = \left\|(\mathbf{I} - \mathbf{P}_{\mathbf{BPS}})\mathbf{B}\right\|_2^2$$

as well. We take $\mathbf{P} = \mathbf{I}$ and $\mathbf{B} = \mathbf{A}^{1/2}$ in this paper.

To establish Theorem 2, 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* $\mathbf{M}$ *be an SPSD matrix of size* n*. Fix integers* k *and* ℓ *satisfying* $1 \leq k \leq \ell \leq n$*.*

Let $\mathbf{U}_1$ *and* $\mathbf{U}_2$ *be matrices with orthonormal columns spanning, respectively, a dominant* k*-dimensional eigenspace of* $\mathbf{M}$ *and the corresponding bottom* $(n-k)$*-dimensional eigenspace of* $\mathbf{M}$*. Let* $\mathbf{\Sigma}_1$ *and* $\mathbf{\Sigma}_2$ *be the diagonal matrices of eigenvalues corresponding, respectively, to the dominant* k*-dimensional eigenspace of* $\mathbf{M}$ *and the bottom* $(n-k)$*-dimensional eigenspace of* $\mathbf{M}$*.*

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

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

If, additionally, $\mathbf{\Sigma}_1$ *is singular, then*

$$\mathbf{M} = \mathbf{P}_{\mathbf{MS}}\mathbf{M}.$$

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

$$\begin{aligned}\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T &= \mathbf{A}\mathbf{S}(\mathbf{S}^T\mathbf{A}\mathbf{S})^\dagger\mathbf{S}^T\mathbf{A} \\ &= \mathbf{A}^{1/2}[\mathbf{A}^{1/2}\mathbf{S}(\mathbf{S}^T\mathbf{A}^{1/2}\mathbf{A}^{1/2}\mathbf{S})^\dagger\mathbf{S}^T\mathbf{A}^{1/2}]\mathbf{A}^{1/2} \\ &= \mathbf{A}^{1/2}\mathbf{P}_{\mathbf{A}^{1/2}\mathbf{S}}\mathbf{A}^{1/2}.\end{aligned}$$

It follows that the spectral error of the Nyström extension satisfies

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

The second equality holds because of the idempotency of projections. The third follows from the fact that $\|\mathbf{A}\mathbf{A}^T\|_2 = \|\mathbf{A}\|_2^2$ for any matrix $\mathbf{A}$. Partition $\mathbf{A}$ as in equation (3). Equation (5) and the following assertion hold by Proposition 1 with $\mathbf{M} = \mathbf{A}^{1/2}$. □

3.1. Nyström extensions from RRQR factorizations.

Since we have seen that the Nyström extension is closely connected to the column subset selection problem, it is natural to consider the implications of prior results on column subset selection.

Early subset selection methods relied upon the SVD or column pivoted QR decompositions [GKS76, GV96], but rank-revealing QR (RRQR) factorizations have become quite popular, as they provably reveal the numerical rank of matrices like SVDS, but are cheaper to compute [CH92]. Let $\mathbf{M}$ be an $m \times n$ real matrix, with $m \geq n$; a QR factorization

$$\mathbf{M}\mathbf{\Pi} = \mathbf{Q}\mathbf{R} = \mathbf{Q}\begin{bmatrix}\mathbf{R}_{11} & \mathbf{R}_{12} \\ & \mathbf{R}_{22}\end{bmatrix}$$

where $\mathbf{\Pi}$ is a permutation matrix and $\mathbf{R}_{11}$ is a $k \times k$ matrix is called rank-revealing if the conditions

$$\sigma_k(\mathbf{R}_{11}) \geq \frac{\sigma_k(\mathbf{M})}{p(k,n)} \quad \text{and} \quad \sigma_1(\mathbf{R}_{22}) \leq p(k,n)\sigma_{k+1}(\mathbf{M}) \tag{6}$$

are satisfied, where $p(k,n)$ is some low-degree polynomial in k and n. Golub and Businger introduced the first algorithm for computing RRQR factorization; notable subsequent improved algorithms were given by Hong and Pan [HP92], Chandrasekan and Ipsen [CI94], and Pan and Tang [PT99]. Gu and Eisenstat [GE96] supplies an algorithm for computing a RRQR factorization with the most favorable properties to date. The Gu–Eisenstat algorithm provides an RRQR factorization with $p(k,n) = \sqrt{1 + f^2 k(n-k)}$ that can be computed in $\mathrm{O}((m + n \log_f n)n^2)$ time; here $f > 1$ is a tolerance parameter.

The relevance of RRQR factorizations to the column subset selection problem comes from the fact that conditions (6) imply that the first k columns of $\mathbf{Q}$ and thus the first k columns of $\mathbf{M\Pi}$ span the dominant k-dimensional invariant subspace of $\mathbf{M}$. This suggests that, if we take $\mathbf{C}$ to be the first k columns of $\mathbf{M\Pi}$, then $\mathbf{P_C M}$ is an almost optimal rank-k approximation to $\mathbf{M}$ [CH92]. Indeed,

$$\mathbf{P_C M} = \mathbf{Q} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ 0 & 0 \end{bmatrix} \mathbf{\Pi}^T,$$

so

$$\left\|(\mathbf{I} - \mathbf{P_C})\mathbf{M}\right\|_2 = \left\|\mathbf{R}_{22}\right\|_2 \leq p(k,n)\sigma_{k+1}(\mathbf{M}).$$

In conjunction with Theorem 2, this suggests a scheme for forming accurate Nyström approximations: take $\mathbf{S}$ to be the first k columns of the permutation matrix returned from an RRQR factorization of $\mathbf{A}^{1/2}$ (or some other generalized Cholesky factor of $\mathbf{A}$). With this choice of $\mathbf{S}$, the minimum eigenvalue of $\mathbf{W}$ is bounded: $\lambda_{\min}(\mathbf{W}) = \sigma_k(\mathbf{A}^{1/2}\mathbf{S})^2 \geq \lambda_k(\mathbf{A})/p(k,n)^2$. Therefore, such a Nyström extension is numerically stable (see Section 1.5), is formed from exactly k columns, and satisfies

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right\|_2 \leq p(k,n)^2 \lambda_{k+1}(\mathbf{A})$$

deterministically. However, forming this Nyström extension is quite expensive, as it requires calculating both a square root (or Cholesky factorization) and a RRQR factorization.

4. Error bounds for simple Nyström extension

In this section, we bound the spectral norm approximation error of simple Nyström extensions. To obtain our results, we use Theorem 2 in conjunction with the estimate of $\left\|\mathbf{\Omega}_1^\dagger\right\|_2^2$ provided by the following lemma.

Lemma 1. *Let* $\mathbf{U}$ *be an* $n \times k$ *matrix with orthonormal columns. Take* μ *to be the coherence of* $\mathbf{U}$*. Select* $\varepsilon \in (0,1)$ *and a nonzero failure probability* δ*. Let* $\mathbf{S}$ *be a random matrix distributed as the first* ℓ *columns of a uniformly random permutation matrix of size* n*, where*

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

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

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

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 $\mathbf{S}$. 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 collection of at least ℓ SPSD matrices of size $k \times k$ (some matrices may be repeated) and that*

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

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

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

Then

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

We note that Proposition 2 is very related to the "Operator Law of Large Numbers" developed by Rudelson and Vershynin in [RV07], in that both provide tail bounds for the eigenvalues of a sum of random matrices. Proposition 2 has the advantages of providing explicit constants and applying to matrices which are sampled without replacement.

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

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

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

$$\lambda_k(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}) \geq \frac{\varepsilon \ell}{n}$$

when ℓ is as stated.

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

$$\lambda_k(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}) = \lambda_k\left(\sum\nolimits_{i=1}^{\ell} \mathbf{X}_i\right),$$

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

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

Proposition 2 yields

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

We require enough samples that

$$\lambda_k(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}) \geq \varepsilon \frac{\ell}{n}$$

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

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

and solve for ℓ, finding

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

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

Theorem 3 establishes that the error incurred by the simple Nyström extension process is small when an appropriate number of columns is sampled. The first error bound in Theorem 3 compares the spectral norm error to the smallest error achievable when approximating $\mathbf{A}$ with a rank-k matrix. It does not use any information about the spectrum of $\mathbf{A}$ other than the value of the $(k+1)$st eigenvalue. The second bound uses information about the entire tail of the spectrum of $\mathbf{A}$. If the spectrum of $\mathbf{A}$ decays fast, the second bound is much tighter than the first. If the spectrum of $\mathbf{A}$ is flat, then the first bound is tighter.

Theorem 3. *Let* $\mathbf{A}$ *be an SPSD matrix of size* n. *Given an integer* $k \leq n$, *partition* $\mathbf{A}$ *as in equation* (3). *Let* μ *denote the coherence of* $\mathbf{U}_1$. *Fix a failure probability* $\delta \in (0,1)$.

For any $\varepsilon \in (0,1)$, *if*

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

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

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

with probability exceeding $1-\delta$ *and*

$$\left\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\right\|_2 \leq \lambda_{k+1}(\mathbf{A}) + \frac{1}{\delta\varepsilon}\sum\nolimits_{i=k+1}^{n} \lambda_i(\mathbf{A})$$

with probability exceeding $1-2\delta$. *If, additionally,* $k \geq \operatorname{rank}(\mathbf{A})$, *then*

$$\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$$

with probability exceeding $1-\delta$.

Theorem 3, like the main result of [TR10], promises exact recovery with probability at least $1-\delta$ when $\mathbf{A}$ is exactly rank k and has small coherence, with a sample of $\mathrm{O}(k\ln(k/\delta))$ columns. Unlike the result in [TR10], Theorem 3 is applicable in the case that $\mathbf{A}$ is full-rank but has a sufficiently fastly decaying spectrum.

Proof of Theorem 3. Because we are using uniform sampling without replacement, the sampling matrix $\mathbf{S}$ is formed by taking the first ℓ columns of a uniformly sampled random permutation matrix. Recall that $\mathbf{\Omega}_1 := \mathbf{U}_1^T\mathbf{S}$ and $\mathbf{\Omega_2} := \mathbf{U}_2^T\mathbf{S}$. By Lemma 1, $\mathbf{\Omega}_1$ has full row rank with probability at least $1-\delta$, so the bounds in Theorem 2 are applicable. In particular, if $k > \operatorname{rank}(\mathbf{A})$ then $\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$.

Applying Lemma 1, we see that $\left\|\mathbf{\Omega}_1^\dagger\right\|_2^2 \leq n/(\varepsilon\ell)$ with probability exceeding $1-\delta$. Hence Theorem 2 gives that

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_2 \leq \left\|\mathbf{\Sigma}_2\right\|_2 + \frac{n}{\varepsilon\ell}\cdot\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\right\|_2^2 \tag{7}$$

with at least the same probability. The two bounds follow by estimating the term $\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\right\|_2^2$ in different ways. The first follows from the observation that $\left\|\mathbf{\Omega}_2\right\|_2 \leq \left\|\mathbf{U}_2\right\|_2\left\|\mathbf{S}\right\|_2 \leq 1$, which implies that

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_2 \leq \left\|\mathbf{\Sigma}_2\right\|_2\left(1 + \frac{n}{\varepsilon\ell}\right)$$

with probability at least $1-\delta$.

The second bound is a consequence of Markov's inequality:

$$\mathbb{P}\left\{\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\right\|_2^2 \geq \frac{1}{\delta}\mathbb{E}\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\right\|_\mathrm{F}^2\right\} \leq \mathbb{P}\left\{\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\right\|_2^2 \geq \frac{1}{\delta}\mathbb{E}\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\right\|_2^2\right\} \leq \delta.$$

The expectation of the Frobenius norm is easily calculated;

$$\mathbb{E}\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\right\|_\mathrm{F}^2 = \mathbb{E}\left\|\sum\nolimits_{i=1}^{\ell}\mathbf{a}_i\mathbf{e}_i^T\right\|_\mathrm{F}^2 = \sum\nolimits_{i=1}^{\ell}\mathbb{E}\|\mathbf{a}_i\|_2^2 = \frac{\ell}{n}\left\|\mathbf{\Sigma}_2^{1/2}\right\|_\mathrm{F}^2 = \frac{\ell}{n}\operatorname{tr}\mathbf{\Sigma}_2,$$

where the vectors $\{\mathbf{a}_i\}$ are chosen randomly without replacement from the columns of $\mathbf{\Sigma}_2^{1/2}\mathbf{U}_2^T$ and $\{\mathbf{e}_i\}$ is the standard basis for $\mathbb{R}^\ell$. Consequently,

$$\left\|\mathbf{\Sigma}_2^{1/2}\mathbf{\Omega}_2\right\|_2^2 \leq \frac{\ell}{\delta n}\operatorname{tr}\mathbf{\Sigma}_2$$

with probability at least $1-\delta$. This observation together with (7) gives that the desired bound

$$\|\mathbf{A}-\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T}\|_2 \leq \left\|\mathbf{\Sigma}_2\right\|_2 + \frac{1}{\delta\varepsilon}\cdot \operatorname{tr}\mathbf{\Sigma}_2$$

holds with probability at least $1-2\delta$. □

The multiplicative factor in the relative error bound asserted in Theorem 3 is optimal in terms of its dependence on n and ℓ. This fact follows from the connection between Nyström extensions and the column subset selection problem.

Indeed, [BDMI11] establishes the following lower bound for the column subset selection problem: for any $\alpha > 0$ there exist matrices $\mathbf{M}_\alpha$ such that any $k \geq 1$ and any $\ell \geq 1$, the error of approximating $\mathbf{M}_\alpha$ with $\mathbf{P}_{\mathbf{D}}\mathbf{M}_\alpha$, where $\mathbf{D}$ may be *any* subset of ℓ columns of $\mathbf{M}_\alpha$, satisfies

$$\left\|\mathbf{M}_\alpha - \mathbf{P}_{\mathbf{D}}\mathbf{M}_\alpha\right\|_2 \geq \sqrt{\frac{n+\alpha^2}{\ell+\alpha^2}}\cdot\left\|\mathbf{M}_\alpha - (\mathbf{M}_\alpha)_k\right\|_2,$$

where $(\mathbf{M}_\alpha)_k$ is the rank-k matrix that best approximates $\mathbf{M}_\alpha$ in the spectral norm. We get a lower bound on the error of the simple Nyström extension by taking $\mathbf{A}_\alpha = \mathbf{M}_\alpha^T\mathbf{M}_\alpha$: it follows from the remark following Theorem 2 that for any $k \geq 1$ and $\ell \geq 1$, any Nyström extension formed using $\mathbf{C} = \mathbf{A}_\alpha\mathbf{S}$ consisting of ℓ columns sampled from $\mathbf{A}_\alpha$ satisfies

$$\left\|\mathbf{A}_\alpha - \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_2 \geq \frac{n+\alpha^2}{\ell+\alpha^2}\cdot\lambda_{k+1}(\mathbf{A}_\alpha).$$

5. Error bounds for the stable approximation algorithms

The error bounds we have produced for the simple Nyström extension assume that the calculations are carried out with infinite precision. In reality, numerical pseudoinversion is not a stable procedure, so a direct application of the Nyström procedure may not produce a result that is close to a valid Nyström approximation. Specifically, if $\mathbf{W}$ is ill-conditioned, the product $\mathbf{W}^{\dagger}\mathbf{C}^T$ may not be computed accurately. To address this concern, we note that the maximum eigenvalue of $\mathbf{W}$ is bounded above by that of $\mathbf{A}$, thus if the minimum eigenvalue of $\mathbf{W}$ is sufficiently bounded away from zero, then $\mathbf{W}^{\dagger} = \mathbf{W}^{-1}$ has a bounded condition number. Accordingly, the approximation $\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T$ can be formed stably, e.g. using a pivoted Cholesky factorization of $\mathbf{W}$. This observation suggests the two stable algorithms for Nyström extension given in Section 1.5. In this section, we prove Theorem 1, which bounds the error of the extensions calculated by Algorithms 1 and 2.

Our main tool in the argument bounding the error of Algorithm 2 is [CD, Lemma 2.2]. The result in [CD, Lemma 2.2] is stated for general matrices; the following proof exploits the fact that we are sampling from an SPSD matrix to get a sharper inequality.

Proposition 3. *Let* $\mathbf{A}$ *be an SPSD matrix of size* n*. Given an integer* $k \leq n$*, partition* $\mathbf{A}$ *as in equation* (3)*. Let* $\ell > k$ *and let* $\mathbf{S}$ *be the corresponding column selector matrix. Let* $\mathbf{C} = \mathbf{A}\mathbf{S}$ *and* $\mathbf{W} = \mathbf{S}^T\mathbf{A}\mathbf{S}$*. Assume that* $\mathbf{\Omega}_1 = \mathbf{U}_1^T\mathbf{S}$ *has full row rank and* $\mathbf{W}$ *is nonsingular, then*

$$\left\|\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_2 \leq \left\|\mathbf{\Omega}_1^{\dagger}\right\|_2 + \left\|\mathbf{W}^{\dagger}\right\|_2\left\|\mathbf{\Sigma}_2\right\|_2.$$

Proof. Since $\mathbf{\Omega}_1\mathbf{\Omega_1}^{\dagger} = \mathbf{I}$,

$$\begin{aligned}\left\|\mathbf{W}^{\dagger}\mathbf{C}^T\right\|_2 &\leq \left\|\mathbf{W}^{\dagger}\mathbf{C}^T\mathbf{U}_1\right\|_2 + \left\|\mathbf{W}^{\dagger}\mathbf{C}^T\mathbf{U}_2\right\|_2\\ &\leq \left\|(\mathbf{S}^T\mathbf{A}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{U}_1\mathbf{\Sigma}_1\right\|_2 + \left\|\mathbf{W}^{\dagger}\right\|_2\left\|\mathbf{S^T}\mathbf{U}_2\mathbf{\Sigma}_2\right\|_2\\ &\leq \left\|(\mathbf{S}^T\mathbf{A}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{\Omega}_1\mathbf{\Omega}_1^{\dagger}\right\|_2 + \left\|\mathbf{W}^{\dagger}\right\|_2\left\|\mathbf{\Sigma}_2\right\|_2\\ &\leq \left\|(\mathbf{S}^T\mathbf{A}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{U}_1^T\mathbf{S}\right\|_2\left\|\mathbf{\Omega}_1^{\dagger}\right\|_2 + \left\|\mathbf{W}^{\dagger}\right\|_2\left\|\mathbf{\Sigma}_2\right\|_2.\end{aligned}$$

Now we observe that $\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{U}_1^T \preceq \mathbf{A}$, so $\mathbf{S}^T\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{U}_1^T\mathbf{S} \preceq \mathbf{S}^T\mathbf{A}\mathbf{S}$. This in turn implies that

$$\left\|(\mathbf{S}^T\mathbf{A}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{U}_1^T\mathbf{S}\right\|_2 \leq 1,$$

and we have the desired inequality. □

Proof of Theorem 1. First we address the performance of Algorithm 1. The triangle inequality gives

$$\|\mathbf{A} - \mathbf{C}_\rho \mathbf{W}_\rho^{-1} \mathbf{C}_\rho^T\|_2 \leq \|\mathbf{A}_\rho - \mathbf{C}_\rho \mathbf{W}_\rho^{-1} \mathbf{C}_\rho^T\|_2 + \|\mathbf{A}_\rho - \mathbf{A}\|_2 \leq \|\mathbf{A}_\rho - \mathbf{C}_\rho \mathbf{W}_\rho^{-1} \mathbf{C}_\rho^T\|_2 + \rho,$$

so it remains only to estimate the error in approximating $\mathbf{A}_\rho$ with the simple Nyström extension $\mathbf{C}_\rho \mathbf{W}_\rho^{-1} \mathbf{C}_\rho^T$. Since $\mathbf{A}$ and $\mathbf{A}_\rho$ have the same eigenspaces, the coherences of their dominant k-dimensional eigenspaces are identical. Furthermore, $\lambda_i(\mathbf{A}_\rho) = \lambda_i(\mathbf{A}) + \rho$ for $i = 1, \ldots, n$. The error bounds and failure probabilities stated for Algorithm 1 then follow from Corollary 1.

Now we address the performance of Algorithm 2. If the algorithm does not regularize $\mathbf{W}$, so that $\mathbf{W}_\rho = \mathbf{W} = \mathbf{S}^T \mathbf{A} \mathbf{S}$, then Algorithm 2 returns a simple Nyström extension. In this case the error bounds of Corollary 1 directly apply, giving sharper bounds than those stated in Theorem 1. Therefore, we consider only the case where $\mathbf{W}_\rho = \mathbf{W} + \rho \mathbf{I}$. To do so, we use a perturbation argument in the same vein as that used in [CD].

Let $\mathbf{C}_\rho = \mathbf{A}_\rho \mathbf{S}$. Since we sample without replacement, $\mathbf{S}^T \mathbf{S} = \mathbf{I}$, so in fact $\mathbf{W}_\rho = \mathbf{S}^T \mathbf{A}_\rho \mathbf{S}$, and

$$e_\rho := \|\mathbf{A}_\rho - \mathbf{C}_\rho \mathbf{W}_\rho^{-1} \mathbf{C}_\rho^T\|_2$$

is the error of a simple Nystrom extension. It is also evident that

$$\|\mathbf{C} - \mathbf{C}_\rho\|_2 \leq \rho \quad \text{and} \quad \|\mathbf{W}_\rho^{-1}\|_2 \leq \rho^{-1}.$$

We use these facts and a perturbation argument to relate the error of our regularized Nyström approximation to $\mathbf{A}$ to e_ρ. Namely, we observe that

$$\begin{aligned}
\|\mathbf{A} - \mathbf{C} \mathbf{W}_\rho^{-1} \mathbf{C}^T\|_2 &\leq \|\mathbf{A} - \mathbf{A}_\rho\|_2 + \|\mathbf{A}_\rho - \mathbf{C}_\rho \mathbf{W}_\rho^{-1} \mathbf{C}_\rho^T\|_2 + \|\mathbf{C}_\rho \mathbf{W}_\rho^{-1} \mathbf{C}_\rho^T - \mathbf{C} \mathbf{W}_\rho^{-1} \mathbf{C}^T\|_2 \\
&\leq \rho + e_\rho + \|\mathbf{C}_\rho \mathbf{W}_\rho^{-1} \mathbf{C}_\rho^T - \mathbf{C}_\rho \mathbf{W}_\rho^{-1} \mathbf{C}^T\|_2 + \|\mathbf{C}_\rho \mathbf{W}_\rho^{-1} \mathbf{C}^T - \mathbf{C} \mathbf{W}_\rho^{-1} \mathbf{C}^T\|_2 \\
&\leq \rho + e_\rho + \rho \|\mathbf{C}_\rho \mathbf{W}_\rho^{-1}\|_2 + \rho \|\mathbf{W}_\rho^{-1} \mathbf{C}^T\|_2 \\
&\leq \rho + e_\rho + \rho \|\mathbf{C}_\rho \mathbf{W}_\rho^{-1}\|_2 + \rho (\|\mathbf{W}_\rho^{-1} \mathbf{C}_\rho^T\|_2 + \|\mathbf{W}_\rho^{-1} (\mathbf{C}^T - \mathbf{C}_\rho^T)\|_2) \\
&\leq \rho + e_\rho + 2\rho \|\mathbf{W}_\rho^{-1} \mathbf{C}_\rho^T\|_2 + \rho^2 \|\mathbf{W}_\rho^{-1}\|_2 \\
&\leq e_\rho + 2\rho (1 + \|\mathbf{W}_\rho^{-1} \mathbf{C}_\rho^T\|_2).
\end{aligned}$$

Let $\mathbf{\Omega}_1 = \mathbf{U}_1^T \mathbf{S}$. By Proposition 3 applied to $\mathbf{A}_\rho$, we have the inequality

$$\|\mathbf{W}_\rho^{-1} \mathbf{C}_\rho^T\|_2 \leq \|\mathbf{\Omega}_1^\dagger\|_2 + \rho^{-1} (\|\mathbf{\Sigma}_2\|_2 + \rho)$$

which holds whenever $\mathbf{\Omega}_1$ has full row-rank. Thus

$$\|\mathbf{A} - \mathbf{C} \mathbf{W}_\rho^{-1} \mathbf{C}^T\|_2 \leq e_\rho + 2\rho \|\mathbf{\Omega}_1^\dagger\|_2 + 2\|\mathbf{\Sigma}_2\|_2 + 4\rho, \tag{8}$$

whenever $\mathbf{\Omega}_1$ has full row-rank.

Now we note from Lemma 1 that, with probability at least $1 - \delta$, $\mathbf{\Omega}_1$ has full row-rank and satisfies $\|\mathbf{\Omega}_1^\dagger\|_2 \leq \sqrt{2n/\ell}$ when $\ell \geq 8\mu k \ln(k/\delta)$. We also note that when $\mathbf{\Omega}_1$ satisfies these conditions, the bounds on e_ρ provided in Corollary 1 hold. Note that since e_ρ is the error of a Nyström extension of $\mathbf{A}_\rho$, it contains factors involving ρ. Identify these factors to conclude that the extensions produced by Algorithm 2 satisfy the bounds stated in Theorem 1. □

6. Computational Investigations

In this section we demonstrate the tightness of the bound provided for the simple Nystrom extension in Corollary 1 and compare Algorithms 1 and 2 to the algorithm introduced in [CD] for a regularized CUR decomposition of rectangular matrices.

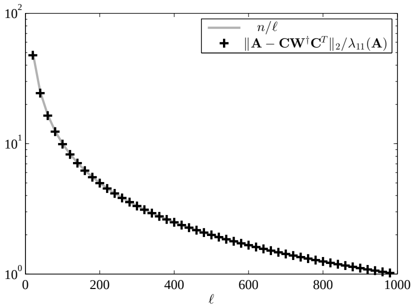


FIGURE 2. The empirical spectral norm error of simple Nyström extensions of $\mathbf{A}$, the matrix defined in (9), relative to the spectral norm error of the optimal rank-10 approximation to $\mathbf{A}$. Each point is the worse relative error observed in 60 trials. The ratio n/ℓ is plotted; this is the dependence on n and ℓ of the bound given in Theorem 3.

Optimality. In the first experiment, a matrix introduced in [BDMI11] demonstrates that, in the worst case, the dependences on n and ℓ in the relative error bounds provided in Corollary 1 and Theorem 3 are optimal. Let $\mathbf{A} \in \mathbb{R}^{1000\times 1000}$ be defined by

$$\mathbf{A} = \mathbf{M}^T\mathbf{M} \quad \text{where} \quad \mathbf{M} = [\mathbf{e}_2 + \mathbf{e}_1, \quad \mathbf{e}_3 + \mathbf{e}_1, \quad \cdots, \quad \mathbf{e}_{1001} + \mathbf{e}_1]; \tag{9}$$

here $\mathbf{e}_i$ denotes the ith standard basis vector in $\mathbb{R}^{1001}$. By construction $\lambda_{\min}(\mathbf{A}) = 1$, so the simple Nyström algorithm can be stably applied to $\mathbf{A}$. Figure 2 plots the ratio of the spectral norm error of approximating $\mathbf{A}$ using simple Nyström extensions against $\lambda_{k+1}(\mathbf{A})$ for $k = 10$. The ratio n/ℓ is provided for comparison. It is clear that the n/ℓ term present in the error bounds are necessary.

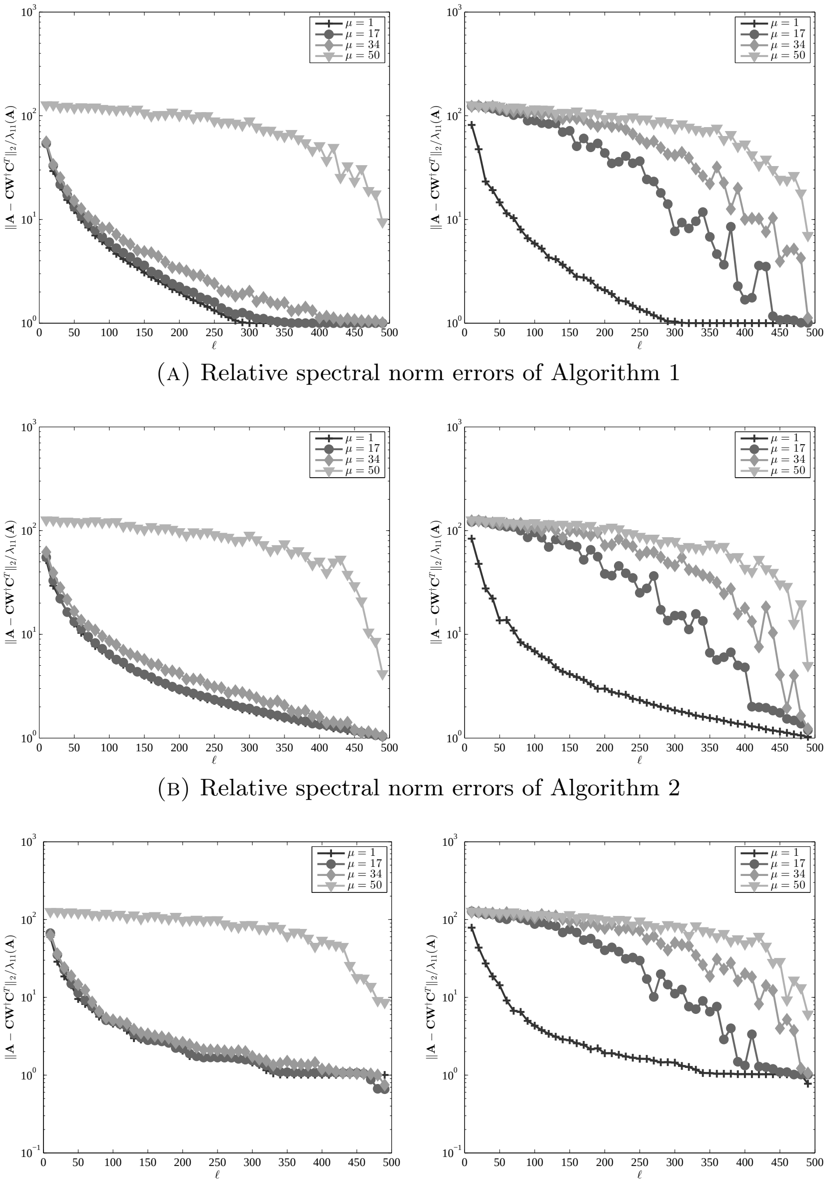
6.1. **Dependence on coherence.** In the following experiments, we use 500×500 matrices $\mathbf{A}$ with eigendecompositions of the form

$$\mathbf{A} = [\mathbf{U}_1 \quad \mathbf{U}_2] \begin{bmatrix} \mathbf{\Sigma}_1 & \\ & \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^T \\ \mathbf{U}_2^T \end{bmatrix} \tag{10}$$

where $\mathbf{U}_1$ is a 500×10 matrix with orthonormal columns and specified coherence and the matrix $\mathbf{U}_2$ is chosen so that $[\mathbf{U}_1 \quad \mathbf{U}_2]$ is an orthogonal matrix. The 20 largest eigenvalues of $\mathbf{A}$ range logarithmically from 10 to 10^{-3}, and the remaining eigenvalues are identically 10^{-15}. Routines from the *kappaSQ* Matlab package introduced in [IW] are used to generate $\mathbf{U}_1$ with specified coherences. For each value of coherence, we consider two types of $\mathbf{U}_1$ achieving this coherence: dense $\mathbf{U}_1$, in which many rows of $\mathbf{U}_1$ are nonzero, and sparse $\mathbf{U}_1$, in which many rows of $\mathbf{U}_1$ are zero. Dense $\mathbf{U}_1$ are generated using the `mtxGenMethod1` routine, and sparse $\mathbf{U}_1$ are generated using the `mtxGenMethod3` routine.

In addition to Algorithms 1 and 2, we consider the performance of the regularized CUR algorithm given in [CD] for general matrices. In comparison to Algorithm 2 which regularizes $\mathbf{W}$ by adding a small multiple of the identity, $\rho \cdot \mathbf{I}$, the algorithm of [CD] regularizes $\mathbf{W}$ by zeroing all singular values of $\mathbf{W}$ which fall below the threshold ρ. In Figure 3 we plot the approximation errors of the three Nyström extensions over the approximation error of the optimal rank-10 approximant as the coherence and sparsity of $\mathbf{U}_1$ vary. The regularization parameter ρ is assigned the value $\lambda_{11}(\mathbf{A})$.

All three algorithms obey a basic principle implied by Theorem 3: as the coherence of the dominant eigenspace increases, the number of samples needed to obtain a small relative error increases. Additionally, Figure 3 shows that the structure of the eigenvectors is as important as the coherence of the eigenspace: when the eigenvectors are dense, the number of samples needed to obtain a small relative error is much less sensitive to the coherence than when the eigenvectors



(A) Relative spectral norm errors of Algorithm 1

(B) Relative spectral norm errors of Algorithm 2

(C) Relative spectral norm errors of the algorithm of [CD]

FIGURE 3. The relative spectral norm errors of Nyström extensions of $\mathbf{A}$, the matrix defined in (10), generated using Algorithms 1 and 2 and the algorithm of [CD] as a function of the coherence of the dominant 10-dimensional eigenspace. The errors are measured relative to the error of the optimal rank-10 approximation, and averaged over 60 runs for each value of ℓ. The eigenvectors spanning the dominant eigenspace of the matrices used in the experiments on the left-hand side are dense, and the corresponding eigenvectors of the matrices used in the experiments on the right-hand side are sparse. The coherences range from the minimum possible, 1, to the maximum of 50.

are sparse. That is, for a fixed coherence and number of column samples, the Nyström extensions give lower errors when the eigenvectors are dense than they do when the eigenvectors are sparse.

Dependence on ρ. Both Algorithms 1 and 2 and the algorithm of [CD] require the choice of a regularization parameter ρ. In Figure 4, we observe the effect of the regularization parameter ρ on the errors of the Nyström extensions. Here the matrix $\mathbf{A}$ is again a 500×500 matrix with eigendecomposition

$$\mathbf{A} = [\mathbf{U}_1 \quad \mathbf{U}_2] \begin{bmatrix} \mathbf{\Sigma}_1 & \\ & \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{U}_1^T \\ \mathbf{U}_2^T \end{bmatrix}, \tag{11}$$

where $\mathbf{U}_1$ is a 500×20 matrix with orthonormal columns and $\mathbf{U}_2$ is chosen so that $[\mathbf{U}_1 \quad \mathbf{U}_2]$ is an orthogonal matrix. The 40 dominant eigenvalues of $\mathbf{A}$ range logarithmically from 1 to 10^{-10} and all remaining eigenvalues are identically 10^{-10}. The `mtxGenMethod1` routine is used to construct $\mathbf{U}_1$ with coherence 1.

Figure 4 shows, as a function of ρ, the approximation errors of the simple Nyström extension and the extensions computed by Algorithm 1, Algorithm 2, and the algorithm of [CD] to the approximation error of the optimal rank-20 approximant of this $\mathbf{A}$. The number of columns used to form the extensions is fixed at $\ell = 200$. We see that all three regularized algorithms exhibit the same behavior. For large values of ρ, they have higher error than the simple Nyström extension; as ρ decreases, their errors become orders of magnitude smaller than that of the simple Nyström extension, and as ρ continues to decrease, their errors once again approach that of the simple Nyström extension. This behavior highlights the importance of choosing an appropriate regularization parameter: if ρ is too small then there is no benefit gained from the regularization, and if it is too large then the regularization has a deliterious effect.

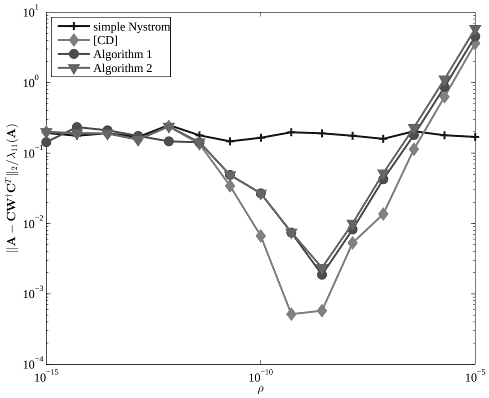


Figure 4. For the matrix $\mathbf{A}$ defined in (11), the spectral norm error of the simple Nyström extension and the extensions generated using Algorithms 1 and 2 and the algorithm of [CD] as a function of the regularization parameter ρ. The errors are averaged over 60 runs for each value of ρ and plotted relative to the spectral norm error of the optimal rank-10 approximation.

References

[BDMI11] C. Boutsidis, P. Drineas, and M. Magdon-Ismail, *Near optimal column-based matrix reconstruction*, Proc. 52nd IEEE Annual Symposium on Foundations of Computer Science (FOCS), 2011, IEEE, 2011, pp. 305–314.

[BF12] A. L. Bertozzi and A. Flenner, *Diffuse interface models on graphs for classification of high dimensional data*, Multiscale Model. Simul. **10** (2012), 1090–1118.

[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.

[CC00] T. F. Cox and M. A. A. Cox, *Multidimensional scaling*, 2nd ed., Chapman and Hall/CRC, 2000.

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

[CH92] T. F. Chan and P. C. Hansen, *Some Applications of the Rank Revealing QR Factorization*, SIAM J. Sci. Comput. **13** (1992), 727–741.

[CI94] S. Chandrasekaran and I. C. F. Ipsen, *On Rank-Revealing Factorizations*, SIAM J. Matrix Anal. Appl. **15** (1994), 592–622.

[CMT10] C. Cortes, M. Mohri, and A. Talwalkar, *On the Impact of Kernel Approximation on Learning Accuracy*, Proc. 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010), 2010.

[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.

[DDF$^+$90] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, *Indexing by latent semantic analysis*, J. Am. Soc. Inf. Sci. Technol. **41** (1990), 391–407.

[DKM06] P. Drineas, R. Kannan, and M. W. Mahoney, *Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition*, 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.

[DM09] ———, *CUR matrix decompositions for improved data analysis*, Proc. Nat. Acad. Sci. USA **106** (2009), 697–702.

[DM10] ———, *Effective Resistances, Statistical Leverage, and Applications to Linear Equation Solving*, Tech. Report arXiv:1005.3097, Mathematics Department, Stanford University, 2010.

[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.

[GE96] M. Gu and S. C. Eisenstat, *Efficient Algorithms for Computing a Strong Rank-Revealing QR Factorization*, SIAM J. Sci. Comput. **17** (1996), 848–869.

[GKS76] G. H. Golub, V. Klema, and G. W. Stewart, *Rank Degeneracy and Least Squares Problems*, Tech. Report TR-456, Department of Computer Science, University of Maryland, 1976.

[GV96] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, 1996.

[GZT97] S. A. Goreinov, N. L. Zamarashkin, and E. E. Tyrtyshnikov, *Pseudo-Skeleton Approximations by Matrices of Maximal Volume*, Math. Notes **62** (1997), 515–519.

[Han90] P. C. Hansen, *Truncated Singular Value Decomposition Solutions to Discrete Ill-Posed Problems with Ill-Determined Numerical Rank*, SIAM J. Sci. Comput. **11** (1990), 503–518.

[HLMS04] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, *A kernel view of the dimensionality reduction of manifolds*, Proc. 21th International Conference on Machine Learning (ICML 2004), 2004.

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

[HMT08] S. Hyvönen, P. Miettinen, and E. Terzi, *Interpretable nonnegative matrix decompositions*, Proc. 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD08), 2008.

[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.

[HP92] Y. P. Hong and C.-T. Pan, *Rank-Revealing QR Factorizations and the Singular Value Decomposition*, Math. Comp. **58** (1992), 213–232.

[HTF08] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2008.

[IW] I. C. F. Ipsen and T. Wentworth, *The effect of coherence on sampling from matrices with orthonormal columns, and preconditioned least squares problems*, Preprint, arXiv:1203.4809, March 2012.

[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.

[KPSH07] P. Karsmakers, K. Pelckmans, J. Suykens, and J. V. Hamme, *Fixed-Size Kernel Logistic Regression for Phoneme Classification*, Proc. Interspeech 2007, 2007.

[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.

[Mah11] M. Mahoney, *Randomized algorithms for matrices and data*, Foundations and Trends in Machine Learning **3** (2011), 123–224, Preprint, arXiv:1104.5557, 2011.

[Mah12] ———, *Combinatorial scientific computing*, ch. Algorithmic and Statistical Perspectives on Large-Scale Data Analysis, Chapman and Hall/CRC, 2012.

[MD09] M. W. Mahoney and P. Drineas, *CUR matrix decompositions for improved data analysis*, Proc. Nat. Acad. Sci. USA **106** (2009), 697–702.

[MJC$^+$] Lester Mackey, Michael I. Jordan, Richard Y. Chen, Brendan Farrell, and Joel A. Tropp, *Matrix Concentration Inequalities via the Method of Exchangeable Pairs*, Preprint, arXiv:1201.6002. January 2012.

[PT99] C.-T. Pan and P. T. P. Tang, *Bounds on Singular Values Revealed by QR Factorizations*, BIT. Numerical Mathematics **39** (1999), 740–756.

[RV07] M. Rudelson and R. Vershynin, *Sampling from large matrices: An approach through geometric functional analysis*, J. ACM **54** (2007), 21.

[SAJ10] N. Srebro, N. Alon, and T. S. Jaakkola, *Generalization Error Bounds for Collaborative Prediction with Low-Rank Matrices*, Proc. 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010), 2010.

[SXZF07] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos, *Less is More: Compact Matrix Decomposition for Large Sparse Graphs*, Proc. 2007 SIAM International Conference on Data Mining (SDM07), 2007.

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

[TKR08] A. Talwalkar, S. Kumar, and H. Rowley, *Large-scale manifold learning*, Proc. IEEE Conference on Computer Vision and Pattern Recognition 2008 (CVPR 2008), 2008.

[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.

[Tro12] J. A. Tropp, *User-friendly tail bounds for sums of random matrices*, Found. Comput. Math. **12** (2012), 389–434.

[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.