RANDOMIZED LOW-RANK APPROXIMATION FOR SYMMETRIC INDEFINITE MATRICES*

YUJI NAKATSUKASA† AND TAEJUN PARK†

Abstract. The Nyström method is a popular choice for finding a low-rank approximation to a symmetric positive semi-definite matrix. The method can fail when applied to symmetric indefinite matrices, for which the error can be unboundedly large. In this work, we first identify the main challenges in finding a Nyström approximation to symmetric indefinite matrices. We then prove the existence of a variant that overcomes the instability, and establish relative-error nuclear norm bounds of the resulting approximation that hold when the singular values decay rapidly. The analysis naturally leads to a practical algorithm, whose robustness is illustrated with experiments.

Key words. Symmetric matrices, Nyström method, Low-rank approximation, Randomized linear algebra

AMS subject classifications. 15A23, 65F55

1. Introduction.

Low-rank structures are ubiquitous in the computational sciences. They appear frequently as matrices having low numerical rank [35]. A low-rank approximation to a matrix provides an efficient way to store and process the matrix when the dimension is large. The Nyström method [14, 24, 38] has been a popular choice for finding low-rank approximations to symmetric positive semi-definite (SPSD) matrices, especially in the machine learning community for kernel-based methods.

Let $A \in \mathbb{R}^{n\times n}$ be a SPSD matrix and let the positive integer r be the target rank. Then the Nyström method is given by $A_{nys}^{(s)} = CW^{\dagger}C^T$ where $C := AX \in \mathbb{R}^{n\times s}$ and $W := X^TAX \in \mathbb{R}^{s\times s}$ with $r \le s < n$ and $X \in \mathbb{R}^{n\times s}$ is a sketching matrix. The positive integer s is called the sketch size, and typically $r < s \ll n$. Traditionally, X is chosen to be a column sampling matrix, which has exactly one non-zero entry equal to 1 in each column [14, 38]. In this case, C is a subset of s columns of A and W is an $s \times s$ principal submatrix of A. There are different sampling schemes for column sampling, including uniform sampling, leverage score sampling [14, 19, 38, 39, 21] and k-means++ sampling [25]. In recent years, other choices for X have been shown to be practical, including Gaussian matrices, subsampled randomized trigonometric transforms (SRTTs) and sparse maps [15, 20]. These are *random embeddings*, which are the focus of this paper, and unlike column sampling, they mix up the coordinates of a vector when applied [20].

In this paper, we investigate the effect of using $A_{nys}^{(s)}$ and its rank-restricted variants for symmetric matrices that are possibly *indefinite*. Low-rank approximation of symmetric indefinite matrices arises in many applications, such as learning in reproducing kernel Krein̆ spaces [26], natural language processing [8, 27] and non-metric proximity transformations [12], which has applications in bioinformatics and social networks. The original matrix A does not have to be SPSD for one to form the Nyström approximation $A_{nys}^{(s)}$. However, the theory does not translate directly to symmetric indefinite matrices because it uses the fact that the original matrix is SPSD [13, 14, 36]. Indeed, the Nyström approximation can be very poor for indefinite A, as we illustrate below. In this work, we show that a judiciously constructed

*Date: May 31, 2023

Funding: TP was supported by the Heilbronn Institute for Mathematical Research.

†Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK, (nakatsukasa@maths.ox.ac.uk, park@maths.ox.ac.uk).

rank-restricted variant of the Nyström approximation, when used with random embeddings, is robust even for symmetric indefinite matrices, which often outperforms other existing methods as we show for synthetic datasets (Figure 5) and real datasets (Figure 6) in Section 4. We also show in Section 3 that there exists a projection for the core matrix W such that the Nyström approximation gives a good low-rank approximation to *any* symmetric matrix when the singular values decay sufficiently fast.

1.1. Nyström methods and related work. There are several variants of the Nyström method for SPSD matrices. There are two rank-restricted versions that give a rank-r approximation to $A_{nys}^{(s)}$ where $r < s$. The first version, which is more traditional, is defined by $A_{nys}^{(s,r)} = C[\![W]\!]_r^\dagger C^T$ [9, 14, 18] where $[\![W]\!]_r$ denotes the best rank-r approximation to the matrix W using the truncated SVD. The second version is given by $[\![A_{nys}^{(s)}]\!]_r = [\![CW^\dagger C^T]\!]_r$ [28, 32, 36], which was suggested more recently. The difference between the two methods is that $A_{nys}^{(s,r)}$ performs rank-truncation in the core matrix, W, which makes this method cheaper to compute, while $[\![A_{nys}^{(s)}]\!]_r$ performs rank-truncation in the Nyström approximation $A_{nys}^{(s)}$, which makes this method take advantage of the full Nyström approximation, C and W, when performing the rank-truncation. There are also other variants of the Nyström method, including one for rectangular matrices [22, 33] and one that guarantees numerical stability [22]. This paper will mostly focus on $A_{nys}^{(s,r)}$.

It is known that for SPSD matrices, $A_{nys}^{(s)}$ [14] and $[\![A_{nys}^{(s)}]\!]_r$ [36] satisfy relative-error bounds in the nuclear norm. This means that if $\hat{A}$ is a low-rank approximation to A (in this case, $A_{nys}^{(s)}$ or $[\![A_{nys}^{(s)}]\!]_r$) and $\epsilon > 0$ then

$$\left\| A - \hat{A} \right\|_* \leq (1+\epsilon) \left\| A - [\![A]\!]_r \right\|_* \tag{1.1}$$

holds with high probability under some conditions on the sketch X and the sketch size $s > r$ where $\|\cdot\|_*$ is the nuclear norm (the sum of the singular values). The details are in the relevant papers [14, 36]. On the other hand, it is not known whether $A_{nys}^{(s,r)}$ satisfies a relative-error norm bound mentioned above [36]. In [28], an example of a 3×3 SPSD matrix is given, showing the downside of using $A_{nys}^{(s,r)}$ for kernel approximations which commonly uses a column sampling matrix. The authors propose $[\![A_{nys}^{(s)}]\!]_r$[1] as an alternative, for which later Wang, Gittens and Mahoney derived a relative-error norm bound [36]. For this example, the problem persists even if we use random embeddings. However, this is a small example that can yield results with high variability, and random embeddings do give a smaller expected relative-error in the nuclear norm and a smaller variance result than column sampling, especially when the dimension of the matrix is large. This hints that random embeddings can be more robust and reliable than column sampling. This type of phenomena have been discussed before, for example in [20] where the authors point out that column sampling is less reliable than random embeddings due to their relatively high variance results.

For symmetric indefinite matrices, which are the focus of this paper, not much has been shown. It is however known that the problem is rather difficult. We can easily see that the plain Nyström approximation, $A_{nys}^{(s)}$ can behave poorly for sym-

[1] As in [28], for SPSD matrices, it should be noted that $\left\| A - [\![A_{nys}^{(s)}]\!]_r \right\| \leq \left\| A - A_{nys}^{(s,r)} \right\|$ will hold in the spectral norm and the Frobenius norm.

metric indefinite matrices.[2] In Figure 1, the two plots were generated using 100×100 symmetric indefinite matrices with Haar distributed eigenvectors. In the left plot, the eigenvalues decay geometrically from 1 to 10^{-8} with random signs, and in the right plot, the first 20 eigenvalues are equal to ± 1 and the other 80 eigenvalues are equal to $\pm 10^{-10}$ where the signs were applied randomly with equal probability. We apply the plain Nyström approximation $A_{nys}^{(r)}$ using the Gaussian sketch to A. We can see that the plain Nyström approximation can be unstable. This type of issue has also

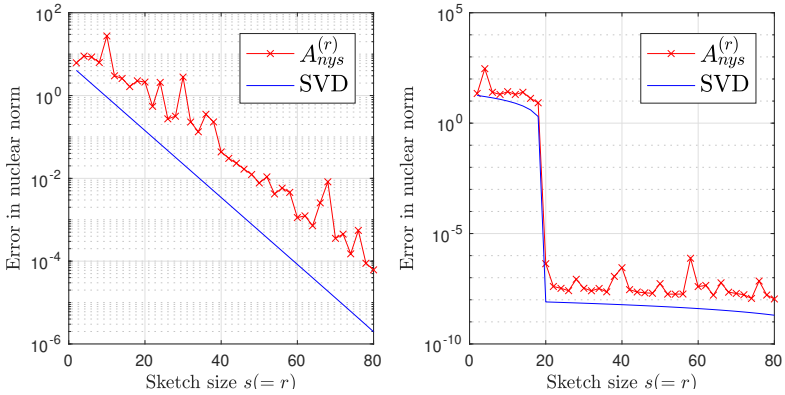


FIG. 1. *Plain Nyström approximation $A_{nys}^{(r)}$ using the Gaussian sketch to 100×100 symmetric indefinite matrices. We can see that $A_{nys}^{(r)}$ can be unstable.*

been observed in a different context for CUR approximations of rectangular matrices [6]. Essentially, the issue arises from the possible presence of an eigenvalue of X^TAX much smaller than $\sigma_r(A)$ or even $\sigma_{\min}(A)$—a phenomenon that is absent when A is SPSD. This blows up the norm of the core matrix $(X^TAX)^\dagger$, causing instability.

Contributions. Our first contribution is to identify the main challenges in finding a good Nyström approximation to symmetric indefinite matrices. We find that the accuracy of the Nyström method is related to controlling the singular values of the core matrix $W = X^TAX$, and show that the accuracy can be lost even if the singular values of W are sufficiently larger than the unit roundoff if W severely underestimates the leading eigenvalues of A. We then perform an analysis in Section 3 that overcomes the challenges. The analysis shows that a certain truncation in the core matrix can give a reliable Nyström approximation that guarantees (1.1) to symmetric indefinite matrices when the singular values decay sufficiently quickly. To our knowledge, this is the first relative-error norm bound for the Nyström method concerning general symmetric matrices that are possibly indefinite.

Our second contribution is providing a practical algorithm (Algorithm 2.1) that gives a Nyström approximation to symmetric indefinite matrices. We show its robustness by comparing the algorithm to some of the existing methods in Section 4 and show that the algorithm performs robustly for symmetric indefinite matrices even in the presence of small singular values in the core matrix, whereas the other algorithms can fail. This algorithm is not new in the context of the Nyström method for SPSD matrices. However, to our knowledge, it has not been suggested or studied before for symmetric indefinite matrices.

[2]For a contrived 2×2 example, where the plain Nyström approximation fails to succeed because the matrix is indefinite, see the arXiv version https://arxiv.org/abs/2212.01127.

Existing methods. We review three existing ideas for using the Nyström method for indefinite matrices, among others. Cai, Nagy and Xi [3] derive an error bound for the Nyström method, $A_{nys}^{(s)}$ for symmetric indefinite matrices that arise from a symmetric function. This bound depends on how close the function values of the sampled points are, which is not an attractive dependence and may not be very useful in more general or practical situations. They suggest the plain Nyström method $A_{nys}^{(r)}$, which can be unstable. They also suggest $AX(X^TAX)_\epsilon^\dagger(AX)^T$ for the Nyström approximation motivated by [22] with the aim of improving the stability. This version truncates the core matrix $W = X^TAX$ so that $\sigma_{\min}((X^TAX)_\epsilon) > \epsilon$ where ϵ is of the order of the unit roundoff. However, this version can give worse approximations than $A_{nys}^{(s)}$ [3] and does not always improve the stability of the Nyström approximation. Second, Ray *et al.* [29] suggest submatrix-shifted (SMS) Nyström to provide an efficient algorithm that deals with symmetric matrices that have only few negative eigenvalues. This method uses an eigenvalue shift based on the minimum eigenvalue of a small principal submatrix before applying the plain Nyström method $A_{nys}^{(r)}$. The downside of this method is that the eigenvalue shift can have serious negative impact on the approximation quality. Lastly, the authors in [12, 26] devise strategies to form the Nyström approximation to symmetric indefinite matrices. However, these methods use eigenvalue information of the original matrix, which is expensive to compute. The three existing methods described above use column sampling matrices for X, which is different from random embeddings. In the final section (Section 5), we will revisit their differences in relation to our method and discuss the implications.

Non-Nyström approaches. In [15], a low-rank approximation for symmetric matrices in the form of the randomized SVD is given. This approximation is given by QQ^TAQQ^T where $Q \in \mathbb{R}^{n\times s}$ is the orthonormal matrix in the thin QR decomposition of AX and is known to satisfy a relative-error norm bound. The dominant cost is $O(n^2s)$ flops for forming Q^TA (assuming A is dense), which becomes prohibitive when n, s are large. Wang, Luo and Zhang derived in [37] a relative-error norm bound to any symmetric matrices (possibly indefinite) for the prototype model. This model computes the low-rank approximation by first forming the sketch $C = AX$ and then approximating A by CXC^T where $X = C^\dagger A(C^\dagger)^T$. The authors show that if C contains $s = O(k/\epsilon)$ columns of A chosen by adaptive sampling then the prototype model has relative-error of at most $(1+\epsilon)$. The dominant costs for the algorithm in [37] are $O(n^2r\log r)$ for computing C and $O(n^2r)$ for computing $C^\dagger A$, which becomes very costly with large n.

Non-symmetric approaches. We can use non-symmetric low-rank approximation to symmetric indefinite matrices. Examples are the randomized SVD [15], which is given by QQ^TA using the notation in the previous paragraph and the generalized Nyström method [4, 22, 33] given by $AX(Y^TAX)^\dagger Y^TA$ where X and Y are independent random embeddings of different dimensions. The details can be found in the relevant papers. For both methods, since their representation is not symmetric, if we want to force symmetry in their representations (e.g. by taking the symmetric part $(M^T + M)/2$), we may risk doubling the rank in the approximation. In addition, as mentioned in the previous paragraph, the randomized SVD has the cost of computing Q^TA, which becomes prohibitive when n, s are large. For generalized Nyström, we approximately double the number of matrix-vector multiplications needed as A needs to be multiplied by two independent random embeddings X and Y and this, in turn doubles the storage requirement (in fact, more than double because Y (or X) is recommended to be larger [22]). In this paper, we focus on symmetric low-rank

approximations.

Notation. Throughout, we use $\|\cdot\|_2$ for the spectral norm or the vector-ℓ_2 norm, $\|\cdot\|_*$ for the nuclear norm (sum of singular values) and $\|\cdot\|_{\mathrm{F}}$ for the Frobenius norm. We use dagger † to denote the pseudoinverse of a matrix and $[\![A]\!]_r$ to denote the best rank-r approximation to A in any unitarily invariant norm, i.e., the approximation derived from truncated SVD [16]. Unless specified otherwise, $\sigma_i(A)$ denotes the ith largest singular value of the matrix A and $\lambda_i(A)$ the ith largest eigenvalue in magnitude. Lastly, we use MATLAB style notation for matrices and vectors. For example, for the kth to $(k+j)$th columns of a matrix A we write $A(:, k:k+j)$.

2. Proposed method. When we use the Nyström method on symmetric indefinite matrices, it can lead to problems. The main concern is in the core matrix $W = X^TAX$ because the positive and negative eigenvalues of A can 'cancel' each other out when forming W, making the eigenvalues of W much smaller than $\sigma_r(A)$. This causes inaccuracies and instabilities when computing the pseudo-inverse of W. More specifically, if we use column sampling then W would be a principal submatrix of A. By Cauchy's interlacing theorem, the spectrum of W is contained in the interval $[\lambda_{\min}(A), \lambda_{\max}(A)]$ which contains both positive and negative values since A is indefinite. Therefore the magnitude of the eigenvalues of W can be significantly smaller in magnitude from those of A, resulting in the matrix $W^\dagger$ blowing up. In addition, the computation of the pseudo-inverse of W can be numerically unstable if $\sigma_{\min}(W) < u$ where u is the unit roundoff. Thus, the main challenge is to ensure that $W^\dagger$ does not ruin the Nyström approximation quality. One approach is to introduce a potentially large shift to make A SPSD, but this can severely affect the approximation quality unless A is nearly definite, that is, the negative eigenvalues of A are very small in magnitude, for example, on the order of machine precision. This idea is used for SPSD matrices where a small shift is introduced to gain numerical stability, however the shift here needs to be small enough to ensure that accuracy is still high [17, 32].

In light of these observations, we propose

$$A_{indef}^{(c,r)} = AX[\![X^TAX]\!]_r^\dagger (AX)^T$$

for symmetric *indefinite* matrices $A \in \mathbb{R}^{n\times n}$ where $X \in \mathbb{R}^{n\times cr}$ is a random embedding, $c > 1$ is a modest constant, say $c = 1.5$ or $c = 2$, and r is the target rank. When A is SPSD and the sketch size s is proportional to the target rank, $A_{indef}^{(c,r)}$ is equivalent to $A_{nys}^{(cr,r)}$. This rank-restricted version truncates the bottom $(c-1)r$ singular values of $W \in \mathbb{R}^{cr\times cr}$, which can potentially be harmful even if they are sufficiently larger than the unit roundoff. This is different to the truncation used in [3] as they use truncation based on the magnitudes of the singular values of W, whereas for our method, the number of bottom singular values we truncate is proportional to the target rank. This intuition is justified by Andoni and Nguyên [1], who prove that the largest eigenvalues (whose proportional to the sketch size) of symmetric matrices with rapidly decaying singular values are approximately preserved under conjugation by a Gaussian sketch with an appropriate normalization factor.

Now, let us define a quantity that will measure how well the singular values are preserved in the core matrix W of the Nyström method. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, a target rank r and a sketch size $s \geq r$, define

$$\kappa_W(A, r, s) := \frac{\max_{1\leq i\leq r} \sigma_i(X^TAX)/\sigma_i(A)}{\min_{1\leq j\leq r} \sigma_j(X^TAX)/\sigma_j(A)} = \max_{1\leq i\leq r} \max_{1\leq j\leq r} \frac{\sigma_i(X^TAX)}{\sigma_j(X^TAX)} \frac{\sigma_j(A)}{\sigma_i(A)} \tag{2.1}$$

where $X \in \mathbb{R}^{n\times s}$ is a Gaussian embedding matrix. This quantity measures the ratio between the worst over-approximation and the worst under-approximation of the leading singular values of A using the singular values in the core matrix W. $\kappa_W(A, r, s)$ will help us see how much the singular values of W have deviated from the leading singular values of A, which directly affects the Nyström approximation quality as we illustrate below.

In Figure 2, we show how important it is to ensure that the spectrum of W does not ruin the approximation quality. In this experiment[3], $A \in \mathbb{R}^{1000\times 1000}$ is a symmetric indefinite matrix constructed as in the left plot of Figure 1. The smallest singular value in the core matrix was larger than 10^{-7} throughout this experiment. For the truncated cases, $A_{indef}^{(1.5,r)}$ and $A_{nys}^{(r+5,r)}$, the approximation is robust as seen in Figure 2a. This robustness we see is illustrated in Figure 2b where the singular values of $W = X^TAX$ behaves well in the sense that there is no wild fluctuations in $\kappa_W(A, r, r+5)$ and $\kappa_W(A, r, 1.5r)$. However, when the sketch size is not proportional to the target rank ($s = r + 5$), the relative approximation error for $A_{nys}^{(r+5,r)}$ (when compared with the truncated SVD) and $\kappa_W(A, r, r+5)$ grow as we increase the target rank. This problem can become worse and the approximation can become unstable when we use SRTT matrices for efficiency with the sketch size $s = r + 5$ (See Figure 3 and Subsection 2.1). When the sketch size is proportional to the target rank, $\kappa_W(A, r, 1.5r)$ and the relative approximation error for $A_{indef}^{(1.5,r)}$ are approximately a constant, which motivates us to choose the oversample size to be proportional to the target rank. On the other hand, without the truncation in the core matrix we see that $\kappa_W(A, r, r)$ behaves wildly. This indicates that the singular values of W inaccurately approximates the leading singular values of A. As a result, the Nyström approximations $A_{nys}^{(1.5r)}$ and $[\![A_{nys}^{(1.5r)}]\!]_r$ can yield unstable results. Empirically, this provides a reason to favour $A_{indef}^{(c,r)}$ over other variants of the Nyström method for symmetric indefinite matrices.

2.1. Random embeddings. A subspace embedding [30] is a linear map which preserves the 2-norm of every vector in a given subspace, that is, $S \in \mathbb{R}^{s\times n}$ is a subspace embedding for the span of $A \in \mathbb{R}^{n\times n}$ with distortion $\epsilon \in (0, 1)$ if

$$(1-\epsilon)\,\|Ax\|_2 \le \|SAx\|_2 \le (1+\epsilon)\,\|Ax\|_2 \tag{2.2}$$

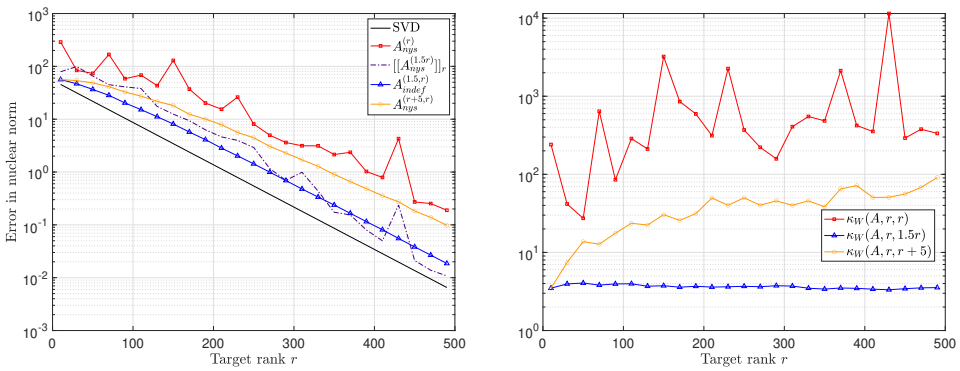
for every $x \in \mathbb{R}^n$. A random embedding is a subspace embedding drawn at random that satisfy Equation (2.2) with high probability.

Random embeddings have more attractive properties than column sampling matrices [11, 20], one of which is that the results obtained using random embeddings generally have smaller variance than the results obtained using column sampling. Below are few important examples of random embeddings.

2.1.1. Gaussian matrices. A Gaussian embedding is a random matrix $G \in \mathbb{R}^{s\times n}$ with i.i.d. entries $G_{ij} \sim N(0, 1/s)$. The scaling ensures that $\mathbb{E}[\|Gx\|_2^2] = \|x\|_2^2$ for every $x \in \mathbb{R}^n$. Gaussian embedding is the most widely used random embedding for theoretical analysis[4] and often has optimal guarantees [15, 20]. The cost of applying a Gaussian embedding to an $n \times n$ matrix is $O(n^2 s)$. This becomes prohibitive for large n, so a more structured random embeddings are often used in practice.

[3] All experiments were performed in MATLAB version 2021a using double precision arithmetic.

[4] Other random embeddings often lack strong theoretical guarantees, however they behave similarly to a Gaussian embedding in practice. For this reason, Gaussian theory is often used to provide a rule of thumb for the general behavior [20].



(a) (b)

FIG. 2. *Accuracy of the Nyström approximations* $A_{indef}^{(1.5,r)}$, $A_{nys}^{(r)}$, $A_{nys}^{(r+5,r)}$ *and* $[\![A_{nys}^{(1.5r)}]\!]_r$ *to a symmetric indefinite matrix* $A \in \mathbb{R}^{1000\times 1000}$. *Figure 2a shows the Nyström error in the nuclear norm and Figure 2b shows the accuracy of the singular values of* $W = X^TAX$ *when compared with the singular values of* A. *We observe that the truncation in the core matrix* W *can significantly increase the robustness and the accuracy of the Nyström approximation.*

2.1.2. SRTTs. A subsampled randomized trigonometric transform (SRTT) matrix is an $n \times s$ matrix with $n \geq s$ of the form

$$S = \sqrt{\frac{n}{s}} DFR^T \tag{2.3}$$

where $D \in \mathbb{R}^{n\times n}$ is a random diagonal matrix whose entries are independent and take ± 1 with equal probability, $F \in \mathbb{C}^{n\times n}$ is a unitary trigonometric transform and $R \in \mathbb{R}^{s\times n}$ is a random restriction. In the complex case, F is the unitary discrete Fourier transform (DFT) and in the real case, F is commonly the discrete cosine transform (DCT). The sketch size needs to be $s = O(r \log r)$ for theoretical guarantees [31], but in practice $s = O(r)$ often suffices[5] [15, 20]. The cost of applying SRTT to an $n \times n$ matrix is $O(n^2 \log r)$ [2] using the subsampled FFT algorithm [40].

2.1.3. Sparse maps. Sparse maps are sparse matrices with nonzero entries that are random signs [4, 20, 23, 39]. They are particularly useful for sparse data and they take the form

$$S = \frac{1}{\sqrt{s}}[s_1, ..., s_n] \in \mathbb{R}^{s\times n} \tag{2.4}$$

where the columns of S, the s_i's are statistically independent and has exactly ξ nonzero entries that take ± 1 with equal probability, placed uniformly at random coordinates. We need the sketch size to be $s = O(r \log r)$ and the sparsity parameter to be $\xi = O(\log r)$ for theoretical guarantees [5]. In [34], $\xi = \min\{s, 8\}$ was recommended in practice. The cost of applying sparse maps to a matrix A is $O(\xi \cdot nnz(A))$ where $nnz(A)$ is the number of nonzero entries of A if sparse data structures and arithmetic are available.

[5] For difficult examples, say a coherent example, the $\log r$ factor is necessary. (See Figure 3)

2.2. Suggested algorithm. For a general symmetric matrix $A \in \mathbb{R}^{n\times n}$ with the target rank r, we suggest

$$A_{indef}^{(c,r)} = AX[\![X^TAX]\!]_r^\dagger (AX)^T = C[\![W]\!]_r^\dagger C^T \tag{2.5}$$

where $X \in \mathbb{R}^{n\times s}$ is a random embedding with the sketch size $s = cr$ where $c > 1$ is a modest constant. The algorithm is given in Algorithm 2.1. For the choice of random embeddings, if A is sparse then we suggest sparse maps with sparsity $\xi = \min\{cr, 8\}$ and when A is dense we suggest SRTT matrices. The recommended sketch size is $s = 1.5r$ for efficiency, but if one wants a better approximation quality guarantee then the sketch size can be increased to, for example, $s = 2r$ or $s = 4r$. Note that the truncation is performed irrespectively of the singular values of W (unlike previous studies, e.g. [3]); our analysis in Section 3 suggests that it is important that the number of singular values to be truncated $(s - r) = (c - 1)r$ is proportional to r.

Algorithm 2.1 Judiciously truncated Nyström approximation for indefinite matrices

Require: Symmetric matrix $A \in \mathbb{R}^{n\times n}$, target rank $r < n$, sketch size $r < s < n$ (rec. $s = 1.5r$)
Ensure: $C \in \mathbb{R}^{n\times s}$ and $W_r^\dagger \in \mathbb{R}^{s\times s}$ with $\mathrm{rank}(W) \le r$ as in (2.5)

1: Draw a random embedding $X \in \mathbb{R}^{n\times s}$ ▷ Sparsity $\xi = \min\{s, 8\}$ for sparse maps
2: $C \leftarrow AX$
3: $W \leftarrow X^TC$
4: $[V, \Lambda] = \mathrm{eig}(W)$, eigendecomposition of W
5: $W_r^\dagger = V(:, 1:r)\Lambda(1:r, 1:r)^\dagger V(:, 1:r)^T$, pseudoinverse of the best rank-r approximation of W
6: Output $C \in \mathbb{R}^{n\times s}$ and $W_r^\dagger \in \mathbb{R}^{s\times s}$

Complexity. When a sparse map is used, the cost of Algorithm 2.1 is $O(\xi \cdot nnz(A) + r^3)$ which consists of $O(\xi \cdot nnz(A))$ flops for forming the sketch and $O(r^3)$ flops for the eigendecomposition. With an SRTT sketch, the total cost is $O(n^2 \log r + r^3)$, where $O(n^2 \log r)$ is needed for forming the sketch and $O(r^3)$ for computing the eigendecomposition.[6]

Eigendecomposition of $A_{indef}^{(c,r)}$. Algorithm 2.1 as presented does not output the eigendecomposition of $A_{indef}^{(c,r)}$. To do this, we require an extra $O(nr^2 + r^3)$ flops. We need $O(nr^2)$ flops to compute the thin QR decomposition of $C = QR$, $O(r^3)$ flops to form and compute the eigendecomposition of $R[\![W]\!]_r^\dagger R^T = U\Sigma U^T$ and $O(nr^2)$ flops to form $U_1 = QU$ giving us the eigendecomposition, $A_{indef}^{(c,r)} = U_1\Sigma U_1^T$.

In Figure 3, we illustrate Algorithm 2.1 for the SRFT sketch and the sparse map. The experiment was conducted with synthetic 2000×2000 symmetric indefinite matrices. The top two plots have eigenvalues that decay geometrically from 1 to 10^{-12} each assigned a random sign with equal probability and the eigenvectors are in a 2×2 block diagonal form, $\mathrm{diag}(I_{200}, U)$ where I_{200} is the 200×200 identity matrix and $U \in \mathbb{R}^{1800\times 1800}$ is a Haar distributed orthogonal matrix. This eigenvector matrix is a more coherent example than our previous examples and is known to be a difficult example for SRTT matrices [2] (when the eigenvectors are Haar distributed, SRTT (or essentially any sketch) behaves the same as a Gaussian sketch, giving good results). The bottom two plots were generated using the same eigenvector matrix, but with eigenvalues equal to ± 1 for the first 100, $\pm 10^{-4}$ for the next 100, $\pm 10^{-8}$ for the 100 eigenvalues after that and $\pm 10^{-16}$ for the last 1700 eigenvalues each assigned a

[6] Since we are using random embeddings for robustness, Algorithm 2.1 is strictly more expensive than classical Nyström methods (column subsampling) if the columns can be sampled quickly.

random sign with equal probability. In the two left plots, we see that the SRFT sketch can fail if the sketch size is not large enough. This instability in the approximation can be fixed by enlarging the sketch size. We see that $s = r + 5$ does not do well, but when $s = 4r$ the approximation becomes more accurate and robust. In the right plot, we see that the SRFT sketch with the sketch size $s = r \log r$, which comes with theoretical guarantees has excellent approximation quality. Finally, we see that the sparse map with sparsity $\xi = 8$ gives a robust approximation throughout, which can be improved by enlarging the sketch size.

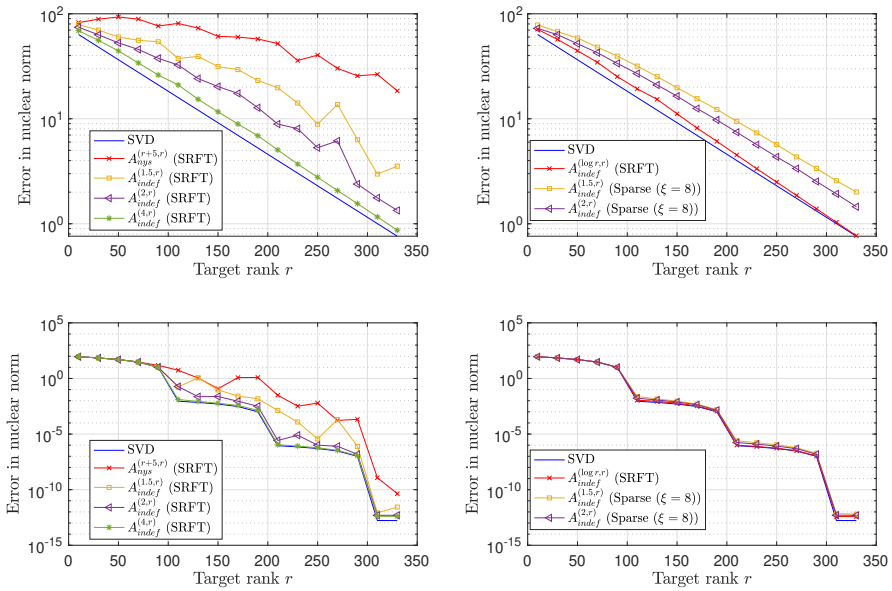


FIG. 3. *Algorithm 2.1: A difficult (coherent) example for the SRFT sketch. The approximation can be unstable if the sketch size is too small for the SRFT sketch (left plots). This problem can be fixed by enlarging the sketch size. The right plots show that sparse maps have no issue with this example and the approximation is robust.*

3. Analysis. For a general symmetric matrix $A \in \mathbb{R}^{n\times n}$, there are no known relative-error norm bounds for the Nyström method. Here we show that for general symmetric matrices, the Nyström method when used with a Gaussian sketch satisfies in expectation a relative-error nuclear norm bound under some orthogonal projection in the core matrix, when the singular values decay sufficiently fast. The analysis that follows establishes the accuracy not of Algorithm 2.1, but of a closely related variant of the Nyström method. The last paragraph of this section discusses this in more detail.

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix and let the eigendecomposition of A be

$$A = V\Lambda V^T = [V_1, V_2, V_3] \begin{bmatrix} \Lambda_1 & 0 & 0 \\ 0 & \Lambda_2 & 0 \\ 0 & 0 & \Lambda_3 \end{bmatrix} [V_1, V_2, V_3]^T \tag{3.1}$$

where $V \in \mathbb{R}^{n\times n}$ is the orthogonal eigenvector matrix of A and $\Lambda \in \mathbb{R}^{n\times n}$ is a diagonal matrix containing the eigenvalues of A. The matrices with subscript 1 have r columns,

those with subscript 2 have $(c_1-1)r$ columns and subscript 3 have $(n-c_1r)$ columns where $r < c_1r < n$ and $c_1 > 1$ is a constant such that c_1r is a positive integer. The eigenvalues are ordered in non-increasing order with respect to their magnitude, so we have $\sigma_i(A) = |\lambda_i(A)|$ for all i.

Now we state our main theorem, and discuss the three key facts that will accompany our proof before getting to the proof immediately.

THEOREM 3.1. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix as in (3.1) and assume that $\lambda_r(A) \neq 0$. Let c_1 and c_2 be constants with $1 < c_1 < c_2 < \frac{n}{r} - 1$ such that c_1r and c_2r are positive integers. Define $X_i := V_i^T X$ for $i = 1, 2, 3$ where $X \in \mathbb{R}^{n\times c_2r}$ is a Gaussian matrix, and set $B = X_3Q_\perp(X_1Q_\perp)^\dagger \in \mathbb{R}^{(n-c_1r)\times r}$ where $Q_\perp \in \mathbb{R}^{c_2r\times(c_2-c_1+1)r}$ is an orthogonal complement of $X_2^T \in \mathbb{R}^{c_2r\times(c_1-1)r}$. Let $(X_1Q_\perp)^\dagger = \hat{Q}\hat{R}$ be the thin QR decomposition of $(X_1Q_\perp)^\dagger$ and set $U := Q_\perp\hat{Q} \in \mathbb{R}^{c_2r\times r}$. Then the orthogonal projector $P = UU^T \in \mathbb{R}^{c_2r\times c_2r}$ satisfies*

$$\mathbb{E}\left[\|E\|_* \,|\Omega_F\right] \leq (1+\epsilon_{r,A})\,\|A - [\![A]\!]_r\|_* \tag{3.2}$$

where

$$E := A - AX(PX^TAXP)^\dagger X^TA \tag{3.3}$$

is the associated Nyström error, Ω_F is an event defined as

$$\Omega_F := \left\{ \left\| |\Lambda_3|^{1/2}B \right\|_F^2 \leq 0.5|\lambda_r(A)| \right\} \tag{3.4}$$

where $|\Lambda_3|$ is defined element-wise and

$$\epsilon_{r,A} := 2b\sqrt{r}\left(1 + \frac{|\lambda_{c_1r+1}(A)|}{|\lambda_r(A)|} + \frac{2}{\sqrt{b}}\right)\frac{\|\Lambda_3\|_*}{\|\Lambda_2\|_* + \|\Lambda_3\|_*} \tag{3.5}$$

where $b = \frac{r}{(c_2-c_1)r-1}$.

In the above theorem, c_1 and c_2 are oversampling factors which are of modest size, say $c_1 = 1.5$ and $c_2 = 2$. We need two factors because we need $X_1Q_\perp \in \mathbb{R}^{r\times(c_2-c_1+1)r}$ and $X_3Q_\perp \in \mathbb{R}^{(n-c_1r)\times(c_2-c_1+1)r}$ to be *rectangular* Gaussian matrices, which makes them well-conditioned with high probability [7]. We can view c_1 as c in Algorithm 2.1 and c_2 to be the oversampling factor introduced to make the analysis possible. By making c_1, c_2 and $(c_2 - c_1)$ larger, we can improve the bound in the above Theorem. The orthogonal projector $P = UU^T$ truncates the core matrix $W = X^TAX$ by removing the largest 'unwanted' eigenvalues of A, i.e. the eigenvalues in Λ_2, using $X_\perp$ factor in U. This helps the core matrix to not be corrupted by the interaction between the target and the large 'unwanted' singular values and singular vectors of A, which can happen when forming X^TAX. Lastly, the $\epsilon_{r,A}$ in the theorem plays a similar role to the distortion ϵ in Equation (1.1) and Ω_F is roughly the event that the eigenvalues of A decay rapidly enough. If we assume that A has a low-rank structure, for example, $|\lambda_r(A)| \gg |\lambda_{c_1r+1}(A)|$, then Ω_F would hold with high probability and $\epsilon_{r,A}$ would be a moderately-sized constant, which tells us that the relative-error nuclear norm bound in (3.2) is good.

We now introduce three key facts that will be useful for our proof. The first fact follows closely the analysis in [22]. Let $\mathcal{P} := \Lambda V^TX(PX^TAXP)^\dagger X^TV$ be an oblique projector. Then we can rewrite the associated Nyström error as

$$E = V(I - \mathcal{P})\Lambda V^T. \tag{3.6}$$

As shown in [22], it is straightforward to see that we can rewrite the associated Nyström error as

$$V^T E V = (I - \mathcal{P})\Lambda = (I - \mathcal{P})\Lambda(I - V^T X U M) \tag{3.7}$$

for any $M \in \mathbb{R}^{r\times n}$. Let $V_r = [I_r, 0]^T \in \mathbb{R}^{n\times r}$ and set $M = (V_r^T V^T X U)^\dagger V_r^T$ then we get

$$V^T E V = (I - \mathcal{P})\Lambda(I - V_r V_r^T)(I - V^T X U (V_r^T V^T X U)^\dagger V_r^T). \tag{3.8}$$

This modification of the associated Nyström error will be important for our proof.

The second fact is the following. Let $f(x)$ be convex in the interval $[x_1, x_2]$ with $x_1 < x_2$. Define $g(x)$ on $[x_1, x_2]$ to be the linear function joining the endpoints of f on $[x_1, x_2]$, that is, $g(x) = \frac{f(x_2)-f(x_1)}{x_2-x_1} x + \frac{f(x_1)x_2 - f(x_2)x_1}{x_2-x_1}$. Then $f(x) \le g(x)$ on $[x_1, x_2]$. Let Y be a random variable with $Y \in [x_1, x_2]$ almost surely. Then $f(Y) \le g(Y)$ almost surely. Furthermore, if $Y \in [x_1, x_2]$ conditional on an event Ω, then conditional on Ω we get

$$f(Y) \le g(Y). \tag{3.9}$$

The last fact is based on expected norm bounds for Gaussian matrices from [15, App. A]. We can deduce the following lemma.

LEMMA 3.2. *Let B be the matrix as in Theorem 3.1 and let S be a fixed real matrix such that SB is defined. Then*

$$\mathbb{E}\,\|SB\|_F^2 = b\,\|S\|_F^2 \tag{3.10}$$

where $b = \frac{r}{(c_2-c_1)r-1}$ as in Theorem 3.1.

Proof. Since conditional on X_2, $X_3 Q_\perp$ and $X_1 Q_\perp$ are two independent Gaussian matrices, we have

$$\begin{aligned}
\mathbb{E}_{X_1,Q_\perp,X_3}\,\|SB\|_F^2 &= \mathbb{E}_{X_1,Q_\perp}\left[\mathbb{E}_{X_3}\left[\left\|SX_3Q_\perp(X_1Q_\perp)^\dagger\right\|_F^2 \,\middle|\, X_1, X_2\right]\right] \\
&= \|S\|_F^2\, E_{X_1,Q_\perp}\left\|(X_1Q_\perp)^\dagger\right\|_F^2 \\
&= \frac{r}{(c_2-c_1)r-1}\,\|S\|_F^2
\end{aligned}$$

using the tower property and the propositions in [15, App. A]. ∎

Now using these three key facts we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. Since U is an orthonormal matrix we have

$$\begin{aligned}
AX(PX^TAXP)^\dagger X^T A &= AX(UU^TX^TAXUU^T)^\dagger X^T A \\
&= AXU(U^TX^TAXU)^\dagger U^T X^T A.
\end{aligned}$$

Now since $X_1 Q_\perp \in \mathbb{R}^{r\times(c_2-c_1+1)r}$ is a fat rectangular Gaussian matrix, hence full rank with probability 1, we have $X_1Q_\perp(X_1Q_\perp)^\dagger = I_r$. Therefore

$$X_1 Q_\perp \hat{Q} = \hat{R}^{-1} \tag{3.11}$$

and we get

$$V^T X U = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} Q_\perp \hat{Q} = \begin{bmatrix} X_1 Q_\perp \hat{Q} \\ 0 \\ X_3 Q_\perp \hat{Q} \end{bmatrix} = \begin{bmatrix} \hat{R}^{-1} \\ 0 \\ B\hat{R}^{-1} \end{bmatrix} \tag{3.12}$$

where $B = X_3 Q_\perp (X_1 Q_\perp)^\dagger \in \mathbb{R}^{(n-c_1 r)\times r}$.

Now we use the first key fact (Equation (3.8)) and get

$$V^T E V = (I - \mathcal{P})\Lambda(I - V_r V_r^T)(I - V^T X U (V_r^T V^T X U)^\dagger V_r^T) \tag{3.13}$$

where $\mathcal{P} = \Lambda V^T X (P X^T A X P)^\dagger X^T V$ and V_r is as below. Using

$$V^T X U = \begin{bmatrix} \hat{R}^{-1} \\ 0 \\ B\hat{R}^{-1} \end{bmatrix}, \Lambda = \begin{bmatrix} \Lambda_1 & 0 & 0 \\ 0 & \Lambda_2 & 0 \\ 0 & 0 & \Lambda_3 \end{bmatrix}, V_r = \begin{bmatrix} I_r \\ 0 \end{bmatrix} \tag{3.14}$$

we get

$$\begin{aligned} V^T E V &= (I - \mathcal{P})\Lambda \begin{bmatrix} 0 \\ I_{n-r} \end{bmatrix} [0, I_{n-r}] \left(I - \begin{bmatrix} \hat{R}^{-1} \\ 0 \\ B\hat{R}^{-1} \end{bmatrix} \left(\hat{R}^{-1} \right)^\dagger [I_r, 0] \right) \\ &= (I - \mathcal{P}) \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Lambda_2 & 0 \\ -\Lambda_3 B & 0 & \Lambda_3 \end{bmatrix}. \end{aligned}$$

We also get

$$\begin{aligned} \mathcal{P} &= \begin{bmatrix} \Lambda_1 \hat{R}^{-1} \\ 0 \\ B\hat{R}^{-1} \end{bmatrix} \left(\begin{bmatrix} \hat{R}^{-1} \\ 0 \\ B\hat{R}^{-1} \end{bmatrix}^T \begin{bmatrix} \Lambda_1 \hat{R}^{-1} \\ 0 \\ \Lambda_3 B \hat{R}^{-1} \end{bmatrix} \right)^\dagger \begin{bmatrix} \hat{R}^{-1} \\ 0 \\ B\hat{R}^{-1} \end{bmatrix}^T \\ &= \begin{bmatrix} \Lambda_1 \hat{R}^{-1} \\ 0 \\ \Lambda_3 B \hat{R}^{-1} \end{bmatrix} \left(\hat{R}^{-T} \left(\Lambda_1 + B^T \Lambda_3 B \right) \hat{R}^{-1} \right)^\dagger \begin{bmatrix} \hat{R}^{-1} \\ 0 \\ B\hat{R}^{-1} \end{bmatrix}^T \\ &= \begin{bmatrix} \Lambda_1 \\ 0 \\ \Lambda_3 B \end{bmatrix} \left(\Lambda_1 + B^T \Lambda_3 B \right)^\dagger [I_r, 0, B^T] \end{aligned}$$

by taking out a factor of $\hat{R}^{-1}$ and $\hat{R}^{-T}$ from the pseudo-inverse. This is possible because if we condition on Ω_F then $(\Lambda_1 + B^T \Lambda_3 B)$ is a non-singular $r \times r$ matrix. Now for shorthand let $S := \Lambda_1 + B^T \Lambda_3 B$. Then

$$I - \mathcal{P} = \begin{bmatrix} I_r - \Lambda_1 S^\dagger & 0 & -\Lambda_1 S^\dagger B^T \\ 0 & I_{(c_1-1)r} & 0 \\ -\Lambda_3 B S^\dagger & 0 & I_{n-c_1 r} - \Lambda_3 B S^\dagger B^T \end{bmatrix}.$$

Therefore

$$
\begin{aligned}
V^T E V &= (I - \mathcal{P}) \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Lambda_2 & 0 \\ -\Lambda_3 B & 0 & \Lambda_3 \end{bmatrix} \\
&= \begin{bmatrix} \Lambda_1 S^\dagger B^T \Lambda_3 B & 0 & -\Lambda_1 S^\dagger B^T \Lambda_3 \\ 0 & \Lambda_2 & 0 \\ -\Lambda_3 B + \Lambda_3 B S^\dagger B^T \Lambda_3 B & 0 & \Lambda_3 - \Lambda_3 B S^\dagger B^T \Lambda_3 \end{bmatrix} \\
&= \begin{bmatrix} \Lambda_1 S^\dagger B^T \Lambda_3 B & 0 & -\Lambda_1 S^\dagger B^T \Lambda_3 \\ 0 & \Lambda_2 & 0 \\ -\Lambda_3 B S^\dagger \Lambda_1 & 0 & \Lambda_3 - \Lambda_3 B S^\dagger B^T \Lambda_3 \end{bmatrix}.
\end{aligned}
$$

We now bound E in the nuclear norm. For shorthand, define the following

$$
\begin{aligned}
a_1 &= \Lambda_1 S^\dagger B^T \Lambda_3 B \\
a_2 &= \Lambda_1 S^\dagger B^T \Lambda_3 \\
a_3 &= \Lambda_3 - \Lambda_3 B S^\dagger B^T \Lambda_3.
\end{aligned}
$$

We then have

$$
\|E\|_* \leq \|\Lambda_2\|_* + \|a_1\|_* + 2\|a_2\|_* + \|a_3\|_*.
$$

Let us note

$$
\Lambda_1 S^\dagger = \Lambda_1 \left(\Lambda_1 + B^T \Lambda_3 B\right)^\dagger = \left(I_r + B^T \Lambda_3 B \Lambda_1^{-1}\right)^\dagger \tag{3.15}
$$

conditional on Ω_F since $\lambda_r(A) \neq 0$ and

$$
\left\|B^T \Lambda_3 B\right\|_F = \left\|B^T |\Lambda_3|^{1/2} \operatorname{sgn}(\Lambda_3) |\Lambda_3|^{1/2} B\right\|_F \leq \left\||\Lambda_3|^{1/2} B\right\|_F^2 \tag{3.16}
$$

where $|\Lambda_3|$ and $\operatorname{sgn}(\Lambda_3)$ are defined element-wise.

We now bound $\mathbb{E}[\|a_1\|_* |\Omega_F]$, $\mathbb{E}[\|a_2\|_* |\Omega_F]$ and $\mathbb{E}[\|a_3\|_* |\Omega_F]$ using the second (Equation (3.9)) and the third (Lemma 3.2) key fact. We start with a_1. Conditional on Ω_F, we have

$$
\begin{aligned}
\|a_1\|_* &\leq \sqrt{r} \left\|\Lambda_1 S^\dagger B^T \Lambda_3 B\right\|_F \\
&\leq \sqrt{r} \left\|\left(I_r + B^T \Lambda_3 B \Lambda_1^{-1}\right)^\dagger\right\|_2 \left\|B^T \Lambda_3 B\right\|_F \\
&\leq \sqrt{r} \frac{\left\|B^T \Lambda_3 B\right\|_F}{1 - \left\|B^T \Lambda_3 B\right\|_F \left\|\Lambda_1^{-1}\right\|_2} \\
&\leq \frac{\sqrt{r}}{\left\|\Lambda_1^{-1}\right\|_2} \frac{\left\||\Lambda_3|^{1/2} B\right\|_F^2 \left\|\Lambda_1^{-1}\right\|_2}{1 - \left\||\Lambda_3|^{1/2} B\right\|_F^2 \left\|\Lambda_1^{-1}\right\|_2} \\
&\leq \frac{\sqrt{r}}{\left\|\Lambda_1^{-1}\right\|_2} \left(2 \left\||\Lambda_3|^{1/2} B\right\|_F^2 \left\|\Lambda_1^{-1}\right\|_2\right)
\end{aligned}
$$

where the last inequality was obtained using the second fact with $Y = \left\||\Lambda_3|^{1/2} B\right\|_F^2 \left\|\Lambda_1^{-1}\right\|_2$, the event Ω_F, the interval $[0, 0.5]$, $f(x) = \frac{x}{1-x}$ which is convex on $[0, 0.5]$ and $g(x) = 2x$. Now taking conditional expectation and using the third

fact (Lemma 3.2) we get

$$\begin{aligned}\mathbb{E}\left[\|a_1\|_* |\Omega_F\right] &\le 2\sqrt{r}\mathbb{E}\left[\left\||\Lambda_3|^{1/2}B\right\|_F^2 \middle| \Omega_F\right] \\ &= 2\sqrt{r}b\left\||\Lambda_3|^{1/2}\right\|_F^2 \\ &= 2\sqrt{r}b\|\Lambda_3\|_*.\end{aligned}$$

For a_2, it is similar to a_1. Conditional on Ω_F we have

$$\begin{aligned}\|a_2\|_* &\le \sqrt{r}\left\|\Lambda_1(\Lambda_1 + B^T\Lambda_3 B)^\dagger B^T\Lambda_3\right\|_F \\ &\le \sqrt{r}\left\|(I_r + B^T\Lambda_3 B\Lambda_1^{-1})^\dagger\right\|_2 \left\||\Lambda_3|^{1/2}B\right\|_F \left\||\Lambda_3|^{1/2}\right\|_F \\ &\le \frac{\sqrt{r}\sqrt{\|\Lambda_3\|_*}}{\sqrt{\left\|\Lambda_1^{-1}\right\|_2}} \frac{\left\||\Lambda_3|^{1/2}B\right\|_F \sqrt{\left\|\Lambda_1^{-1}\right\|_2}}{1 - \left\||\Lambda_3|^{1/2}B\right\|_F^2 \left\|\Lambda_1^{-1}\right\|_2} \\ &\le \frac{\sqrt{r}\sqrt{\|\Lambda_3\|_*}}{\sqrt{\left\|\Lambda_1^{-1}\right\|_2}} 2\left\||\Lambda_3|^{1/2}B\right\|_F \sqrt{\left\|\Lambda_1^{-1}\right\|_2}\end{aligned}$$

where we used the second fact for the last inequality with $Y = \left\||\Lambda_3|^{1/2}B\right\|_F \sqrt{\left\|\Lambda_1^{-1}\right\|_2}$, the interval $[0, \sqrt{0.5}]$, $f(x) = \frac{x}{1-x^2}$ and $g(x) = 2x$. Therefore we get

$$\begin{aligned}\mathbb{E}[\|a_2\|_* |\Omega_F] &\le 2\sqrt{r}\sqrt{\|\Lambda_3\|_*}\mathbb{E}\left[\left\||\Lambda_3|^{1/2}B\right\|_F \middle| \Omega_F\right] \\ &\le 2\sqrt{r}\sqrt{\|\Lambda_3\|_*}\sqrt{\mathbb{E}\left[\left\||\Lambda_3|^{1/2}B\right\|_F^2 \middle| \Omega_F\right]} \\ &\le 2\sqrt{rb}\|\Lambda_3\|_*\end{aligned}$$

using Lemma 3.2.

Finally for a_3, we get

$$\|a_3\|_* \le \|\Lambda_3\|_* + \sqrt{r}\left\||\Lambda_3|^{1/2}\right\|_2^2 \left\||\Lambda_3|^{1/2}B\right\|_F^2 \left\|(\Lambda_1 + B^T\Lambda_3 B)^\dagger\right\|_2$$

in a similar manner, and conditional on Ω_F we have

$$\begin{aligned}\left\||\Lambda_3|^{1/2}B\right\|_F^2 \left\|(\Lambda_1 + B^T\Lambda_3 B)^\dagger\right\|_2 &\le \left\||\Lambda_3|^{1/2}B\right\|_F^2 \left\|\Lambda_1^{-1}\right\|_2 \left\|\Lambda_1(\Lambda_1 + B^T\Lambda_3 B)^\dagger\right\|_2 \\ &\le \frac{\left\||\Lambda_3|^{1/2}B\right\|_F^2 \left\|\Lambda_1^{-1}\right\|_2}{1 - \left\||\Lambda_3|^{1/2}B\right\|_F^2 \left\|\Lambda_1^{-1}\right\|_2} \\ &\le 2\left\||\Lambda_3|^{1/2}B\right\|_F^2 \left\|\Lambda_1^{-1}\right\|_2\end{aligned}$$

using the second fact with the same values as the a_1 case. Therefore

$$\begin{aligned}\mathbb{E}[\|a_3\|_* |\Omega_2] &\le \|\Lambda_3\|_* + 2\sqrt{r}\|\Lambda_3\|_2 \left\|\Lambda_1^{-1}\right\|_2 \mathbb{E}\left[\left\||\Lambda_3|^{1/2}B\right\|_F^2 \middle| \Omega_F\right] \\ &\le \|\Lambda_3\|_* + 2\sqrt{rb}\|\Lambda_3\|_2 \left\|\Lambda_1^{-1}\right\|_2 \|\Lambda_3\|_*.\end{aligned}$$

Finally, combining everything together we get

$$\mathbb{E}\left[\|E\|_* \,|\Omega_F\right] \leq \|\Lambda_2\|_* + \|\Lambda_3\|_* + 2b\sqrt{r}\left(1 + \frac{|\lambda_{c_1 r+1}(A)|}{|\lambda_r(A)|} + \frac{2}{\sqrt{b}}\right)\|\Lambda_3\|_*. \tag{3.17}$$

Therefore

$$\mathbb{E}\left[\|E\|_* \,|\Omega_F\right] \leq (1+\epsilon_{r,A})\left(\|\Lambda_2\|_* + \|\Lambda_3\|_*\right) = (1+\epsilon_{r,A})\,\|A - [\![A]\!]_r\|_* \tag{3.18}$$

with

$$\epsilon_{r,A} = 2b\sqrt{r}\left(1 + \frac{|\lambda_{c_1 r+1}(A)|}{|\lambda_r(A)|} + \frac{2}{\sqrt{b}}\right)\frac{\|\Lambda_3\|_*}{\|\Lambda_2\|_* + \|\Lambda_3\|_*}. \tag{3.19}$$

∎

Remark 3.3.

1. The relative-error nuclear norm bound is informative if $\epsilon_{r,A}$ is small. Now since $b \approx (c_2 - c_1)^{-1} = O(1)$, we have

$$\epsilon_{r,A} = O\left(\frac{\sqrt{r}\,\|\Lambda_3\|_*}{\|\Lambda_2\|_* + \|\Lambda_3\|_*}\right). \tag{3.20}$$

   Therefore the relative-error nuclear norm bound is good if

$$\sqrt{r}\sum_{j=c_1 r+1}^{n}|\lambda_j(A)| = \sqrt{r}\,\|\Lambda_3\|_* \lesssim \|\Lambda_2\|_* = \sum_{j=r+1}^{c_1 r}|\lambda_j(A)|. \tag{3.21}$$

2. Using a similar proof technique we can obtain mixed norm bounds. The 2-norm version of Theorem 3.1 would give

$$\mathbb{E}\left[\|E\|_2 \,|\Omega_F\right] \leq \|A - [\![A]\!]_r\|_2 + \frac{\epsilon_{r,A}}{\sqrt{r}}\,\|A - [\![A]\!]_r\|_* \tag{3.22}$$

   and the Frobenius norm version would give

$$\mathbb{E}\left[\|E\|_F \,|\Omega_F\right] \leq \|A - [\![A]\!]_r\|_F + \frac{\epsilon_{r,A}}{\sqrt{r}}\,\|A - [\![A]\!]_r\|_* \tag{3.23}$$

   where $\epsilon_{r,A}$ is as in Theorem 3.1. Therefore the constant in front of the best rank-r nuclear norm error improves to $\epsilon_{r,A}/\sqrt{r} = O(1)$ using the second remark (3.20). This type of mixed norm bounds along with the relative-error nuclear norm bound in Theorem 3.1 are fairly consistent with the SPSD versions in Table 1 of [14].
3. We can relax the condition Ω_F to $\Omega_2 := \left\{\left\||\Lambda_3|^{1/2}B\right\|_2^2 \leq 0.5|\lambda_r(A)|\right\}$ at the cost of a slightly worse bound in Equation (3.2). It is easy to show that the bound in Equation (3.2) then changes to

$$\mathbb{E}\left[\|E\|_* \,|\Omega_2\right] \leq (1 + \sqrt{r}\epsilon_{r,A})\,\|A - [\![A]\!]_r\|_*. \tag{3.24}$$

Probability of Ω_F. The probability of the event Ω_F happening can be computed by following the proof of Theorem 10.8 in [15] using $k = r$ and $p = (c_2 - c_1)r$ and Lemma 3.2. We get

$$\mathbb{P}\left(\left\||\Lambda_3|^{1/2}B\right\|_F \leq \sqrt{\|\Lambda_3\|_*}\sqrt{\frac{3r}{(c_2-c_1)r+1}}t + \sqrt{\|\Lambda_3\|_2}\frac{e\sqrt{(c_2-c_1+1)r}}{(c_2-c_1)r+1}tu\right) \geq 1 - 2t^{-(c_2-c_1)r} - e^{-u^2/2}$$

for $u, t > 0$. Now using $(x+y)^2 \le 2(x^2+y^2)$, we get

$$\left(\sqrt{\|\Lambda_3\|_*}\sqrt{\frac{3r}{(c_2-c_1)r+1}}t + \sqrt{\|\Lambda_3\|_2}\frac{e\sqrt{(c_2-c_1+1)r}}{(c_2-c_1)r+1}tu\right)^2$$
$$\le 2t^2\left(\|\Lambda_3\|_*\frac{3r}{(c_2-c_1)r+1} + \|\Lambda_3\|_2\frac{e^2(c_2-c_1+1)r}{((c_2-c_1)r+1)^2}u^2\right).$$

Therefore

$$\mathbb{P}(\Omega_F) = \mathbb{P}\left(\left\||\Lambda_3|^{1/2}B\right\|_F^2 \le 0.5|\lambda_r(A)|\right) \ge 1 - 2t^{-(c_2-c_1)r} - e^{-u^2/2} \tag{3.25}$$

if

$$0.5|\lambda_r(A)| \ge 2t^2\left(\|\Lambda_3\|_*\frac{3r}{(c_2-c_1)r+1} + \|\Lambda_3\|_2\frac{e^2(c_2-c_1+1)r}{((c_2-c_1)r+1)^2}u^2\right), \tag{3.26}$$

i.e., Ω_F holds with high probability when the tail singular values of A decay rapidly. A similar result can also be derived for Ω_2 by following the same results in [15].

Mixed norm bounds. We can obtain mixed norm bounds for Theorem 3.1. The 2-norm version of Theorem 3.1 would give

$$\mathbb{E}\left[\|E\|_2 \,|\Omega_F\right] \le \|A - [\![A]\!]_r\|_2 + \frac{\epsilon_{r,A}}{\sqrt{r}}\|A - [\![A]\!]_r\|_* \tag{3.27}$$

and the Frobenius norm version would give

$$\mathbb{E}\left[\|E\|_F \,|\Omega_F\right] \le \|A - [\![A]\!]_r\|_F + \frac{\epsilon_{r,A}}{\sqrt{r}}\|A - [\![A]\!]_r\|_* \tag{3.28}$$

where $\epsilon_{r,A}$ is as in Theorem 3.1. This improves the constant in front of the best rank-r nuclear norm error to $\epsilon_{r,A}/\sqrt{r} = O(1)$ using the first remark (3.20) in Remark 3.3. The proof for the two mixed norm bounds above can be obtained by following the proof of Theorem 3.1. More specifically, the proof for the mixed norm bounds stay the same until we bound a_1, a_2 and a_3. To get the mixed norm bound, we use the appropriate norms to bound a_1, a_2 and a_3. For example, to bound $\|a_1\|_F$, we start similarly as in the nuclear norm case by conditioning on Ω_F to obtain

$$\|a_1\|_F \le \left\|\left(I_r + B^T\Lambda_3 B\Lambda_1^{-1}\right)^\dagger\right\|_2 \left\|B^T\Lambda_3 B\right\|_F$$
$$\le \frac{\left\|B^T\Lambda_3 B\right\|_F}{1 - \|B^T\Lambda_3 B\|_F \left\|\Lambda_1^{-1}\right\|_2}$$
$$\le \frac{1}{\left\|\Lambda_1^{-1}\right\|_2}\left(2\left\||\Lambda_3|^{1/2}B\right\|_F^2\left\|\Lambda_1^{-1}\right\|_2\right).$$

We then get

$$\mathbb{E}\left[\|a_1\|_* \,|\Omega_F\right] \le 2\mathbb{E}\left[\left\||\Lambda_3|^{1/2}B\right\|_F^2 \middle| \Omega_F\right] = 2b\|\Lambda_3\|_*.$$

The bound for $\|a_2\|_F, \|a_3\|_F, \|a_1\|_2, \|a_2\|_2$ and $\|a_3\|_2$ follows similarly. The mixed norm bounds (3.27) and (3.28) along with the relative-error nuclear norm bound in Theorem 3.1 are fairly consistent with the SPSD versions in Table 1 of [14].

Theorem 3.1 and its proof cannot simply be translated into an algorithm because the proof relies on the eigendecomposition of A, which is too expensive to compute. However, the proof naturally suggests Algorithm 2.1. From the proof of Theorem 3.1, under the condition that the matrix has a low-rank structure discussed in this section, for example in the paragraph after the statement of Theorem 3.1 or in the remark above, we have that a projection is desired in the core matrix. This projection gets rid of the large 'unwanted' eigenvalues of A, i.e. the eigenvalues in Λ_2. In the Nyström method, a natural analogue is to truncate the smallest few singular values in the core matrix $W = X^T AX$ to achieve the target rank r, which is what has been done in Algorithm 2.1. The theorem also suggests that the sketch size should be proportional to the target rank r, which is what we suggest in Algorithm 2.1. Despite Algorithm 2.1 lacking complete theory (even for the SPSD case), we suggest it because the algorithm does seem to work well in practice as we illustrate below.

4. Numerical illustration. We first illustrate Theorem 3.1 and Algorithm 2.1 through experiments. In Figure 4, we show a priori and a posteriori error in Theorem 3.1, and Algorithm 2.1 using 1000×1000 symmetric indefinite matrices. In the left plot, the matrix A has eigenvalues that decay geometrically from 1 to 10^{-12} each assigned a random sign with equal probability. In the right plot, A has eigenvalues equal to ± 1 for the first 100 eigenvalues and $\pm 10^{-10}$ for the other 900 eigenvalues each assigned a random sign with equal probability; this example illustrates the performance when there is a gap in the singular values. The eigenvectors for both plots are in a 2×2 block diagonal form, $\mathrm{diag}(I_{100}, U)$ where I_{100} is the 100×100 identity matrix and $U \in \mathbb{R}^{900\times 900}$ is a Haar distributed orthogonal matrix. Both the algorithm and the theorem were constructed using the Gaussian sketch with the sketch size $1.5r$ for the algorithm and $c_1 r = 1.5r$ and $c_2 r = 2r$ for the theorem. We see that Ω_F holds whenever there is a rapid decay of eigenvalues, i.e., when $|\lambda_r| \gg |\lambda_{c_1 r+1}|$. But more importantly, we see that the bound holds when the event Ω_F occurs (circles) and frequently holds even if the event Ω_F did not occur (crosses). The theorem does extremely well when Ω_F has occurred. We see that the algorithm gives a good robust approximation that is a modest factor worse than the best approximation given by the SVD. Although the theorem does better than the algorithm when Ω_F holds, the theorem can give unstable approximation when Ω_F does not hold. This illustrates that the algorithm, which arose from the theorem, works well in practice.

In experiments not shown here, we compared Algorithm 2.1 with randomized SVD [15] and the generalized Nyström method [4, 22, 33, 40], which are applicable to nonsymmetric (and rectangular) matrices and do not preserve symmetry. We observe that Algorithm 2.1 tends to obtain a slightly better approximant for a fixed rank r.

4.1. Synthetic examples. We now compare some of the existing algorithms against Algorithm 2.1 using different kernel functions and synthetic dataset. We illustrate the following algorithms

1. Algorithm 2.1 with the SRFT sketch and the sketch size $s = 2r$,
2. Algorithm 2.1 with uniform column sampling and the sketch size $s = 2r$,
3. Algorithm 2.1 with leverage score column sampling and the sketch size $s = 2r$,
4. Submatrix-Shifted (SMS) Nyström [29] with uniform column sampling and $s_1 = r$, $s_2 = 2r$ and $\alpha = 1.5$,
5. Submatrix-Shifted (SMS) Nyström [29] with the Gaussian sketch and $s_1 = r$, $s_2 = 2r$ and $\alpha = 1.5$,
6. Stabilized Nyström [3] with the SRFT sketch, $s = r$ and $\epsilon = 10^{-14}$

where r is the target rank and the parameters for SMS Nyström and Stabilized

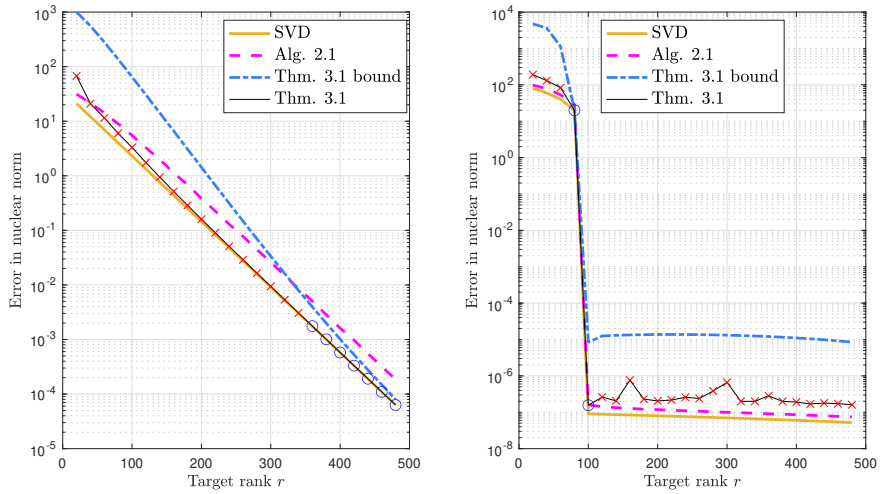


FIG. 4. *Two plots showing the empirical results for Theorem 3.1 and Algorithm 2.1. Algorithm 2.1 is robust with the approximation being a modest factor worse than the best approximation. Theorem 3.1 bound holds when Ω_F has occurred (circles on Theorem 3.1) and also frequently holds even it Ω_F has not occurred (crosses on Theorem 3.1). Theorem 3.1 does extremely well when Ω_F has occurred.*

Nyström are as recommended in their original papers.[7] For SMS Nyström method, the Gaussian sketch was not used in the original paper [29]. We use the following kernel functions

1. Epanechnikov kernel: $k_1(x,y) = \max\{1 - \|x-y\|^2, 0\}$
2. Multiquadric kernel: $k_2(x,y) = \sqrt{1 + \|x-y\|^2}$
3. Thin plate spline: $k_3(x,y) = \|x-y\|^2 \ln\left(\|x-y\|^2\right)$

to generate the kernel matrices. The kernel matrices $K^{(1)}, K^{(2)}$ and $K^{(3)}$ corresponding to the kernel functions k_1, k_2 and k_3 were generated by sampling 1000 random numbers $\{x_i\}_{i=1}^{1000}$ from the standard normal distribution, i.e., $K_{ij}^{(\ell)} = k_\ell(x_i, x_j)$. All the kernel matrices are symmetric indefinite.

In Figure 5, we illustrate the results. The eigenvalue histogram is shown in the left plots. The right plots show the approximation. We see that SMS Nyström performs poorly in all 3 examples except the Gaussian case for the multiquadric kernel. This is possibly because the extreme eigenvalues are large in magnitude so the large shift is ruining the approximation quality. The stabilized Nyström works well for the multiquadric kernel and the thin plate spline, but the approximation is very unstable for the Epanechnikov kernel. This is possibly because the number of positive and the negative eigenvalues are about the same with similar magnitudes for the Epanechnikov kernel, which can increase the chance of instability in the core matrix.[8] This also tells

[7] For stabilized Nyström method, $s = r$ was chosen to ensure that all approximations in the experiment have rank at most r and $\epsilon = 10^{-14}$ as suggested in the original paper was chosen to try diminish the error that might come from taking the pseudo-inverse of the core matrix W.

[8] To our knowledge, the numerical behavior of stabilized Nyström method is an open problem; the stability analysis in [22] applies only to an algorithm where A is sketched from both sides using independent sketches of different dimensions.

us that the truncation in the core matrix should not depend on the magnitude of the singular values of W, but the truncation should always happen proportional to the target rank. Algorithm 2.1 using uniform column sampling and leverage score column sampling are both unstable for all 3 examples, which shows the unreliability of using column sampling matrices. On the other hand, Algorithm 2.1 using the SRFT sketch works well in all cases.

4.2. Dataset examples. We now compare the three different methods using two different high-dimensional datasets, the Covertype and the Anuran Calls (MFCC) from the UC Irvine Machine Learning Repository [10]. We illustrate the following algorithms

1. Algorithm 2.1 with the SRFT sketch and the sketch size $s = 2r$,
2. Algorithm 2.1 with k-means++ samples and the sketch size $s = 2r$,
3. Algorithm 2.1 with uniform column sampling and the sketch size $s = 2r$,
4. Stabilized Nyström with k-means++ samples, the sketch size $s = r$ and $\epsilon = 10^{-14}$

where r is the target rank. We use the following kernel functions

1. Thin plate spline kernel: $\|x-y\|^2 \log\left(\|x-y\|^2\right)$
2. Sigmoid kernel: $\tanh\left(1+\|x-y\|^2\right)$
3. Multiquadric kernel: $\sqrt{1+\|x-y\|^2}$

with the datasets

1. Covertype ($n = 581012$) with dimension $d = 54$,
2. Anuran Calls (MFCC) ($n = 7195$) with dimension $d = 22$.

For each dataset, we sample $n = 4000$ data uniformly at random and then center the mean and normalize all features to have variance 1.

The results are illustrated in Figure 6. We observe that the cause of instability in the Nyström approximation for symmetric indefinite matrices is not necessarily coming from the core matrix W having very small singular values as Stabilized Nyström can give unstable approximations as seen in Figure 6. Also, although Algorithm 2.1 using k-means++ samples is more accurate than uniform column sampling, they both do not give robust low-rank approximations. This shows that it is difficult to find a column sampling scheme that guarantees stable Nyström approximation for symmetric indefinite matrices. On the other hand, Algorithm 2.1 using the SRFT sketch gives robust approximation throughout the experiment and sometimes outperforms the other methods in this experiment such as in Figure 6a and 6e.

5. Discussion. Much of the literature on approximating symmetric matrices using any of the variants of the Nyström method is based on column sampling. In this work, we used random embeddings for our algorithm (Algorithm 2.1) and a special class of random embeddings for the analysis, namely Gaussian embeddings. Random embeddings were used as they are more robust than column sampling, and Gaussian embeddings were used for analysis because we can leverage their rich theoretical properties. The general behaviour when we use the Nyström method with column sampling matrices on symmetric indefinite matrices is unknown. In Figure 5, we see that the two frequently used column sampling schemes, uniform sampling and leverage score sampling can be unstable. It appears to be difficult to find a column sampling scheme that guarantees robust Nyström approximation for symmetric indefinite matrices and, to our knowledge, is an open problem. We hope that our results would shed light on the development of a robust indefinite Nyström method based on

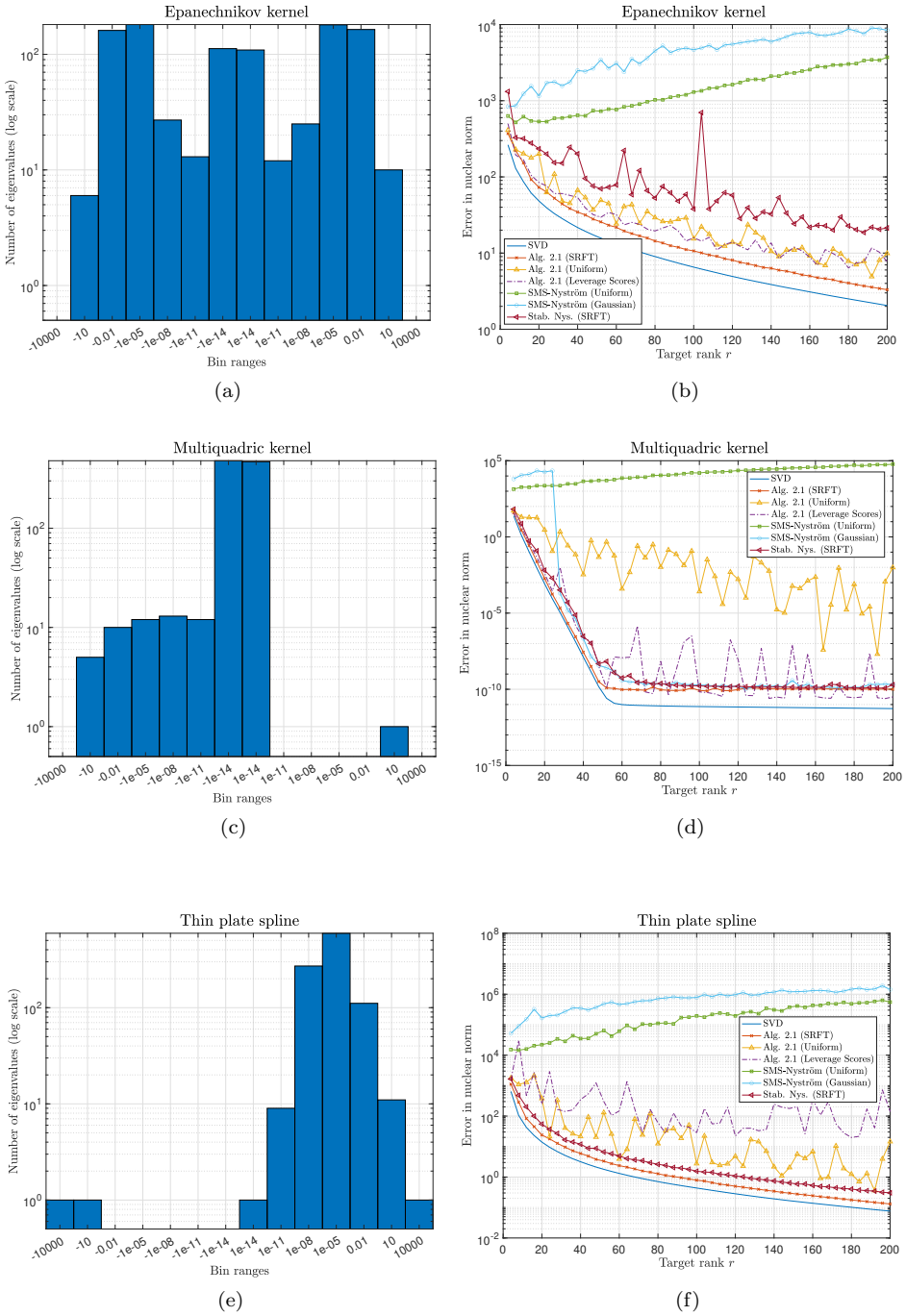


FIG. 5. *Comparison of different methods for symmetric indefinite matrices: SMS-Nyström [29], stabilized Nyström [3] and Algorithm 2.1. The first two methods and Algorithm 2.1 using uniform column sampling and leverage score column sampling can fail on some kernels while Algorithm 2.1 using the SRFT sketch (random embedding) works well for all the kernels in the experiment.*

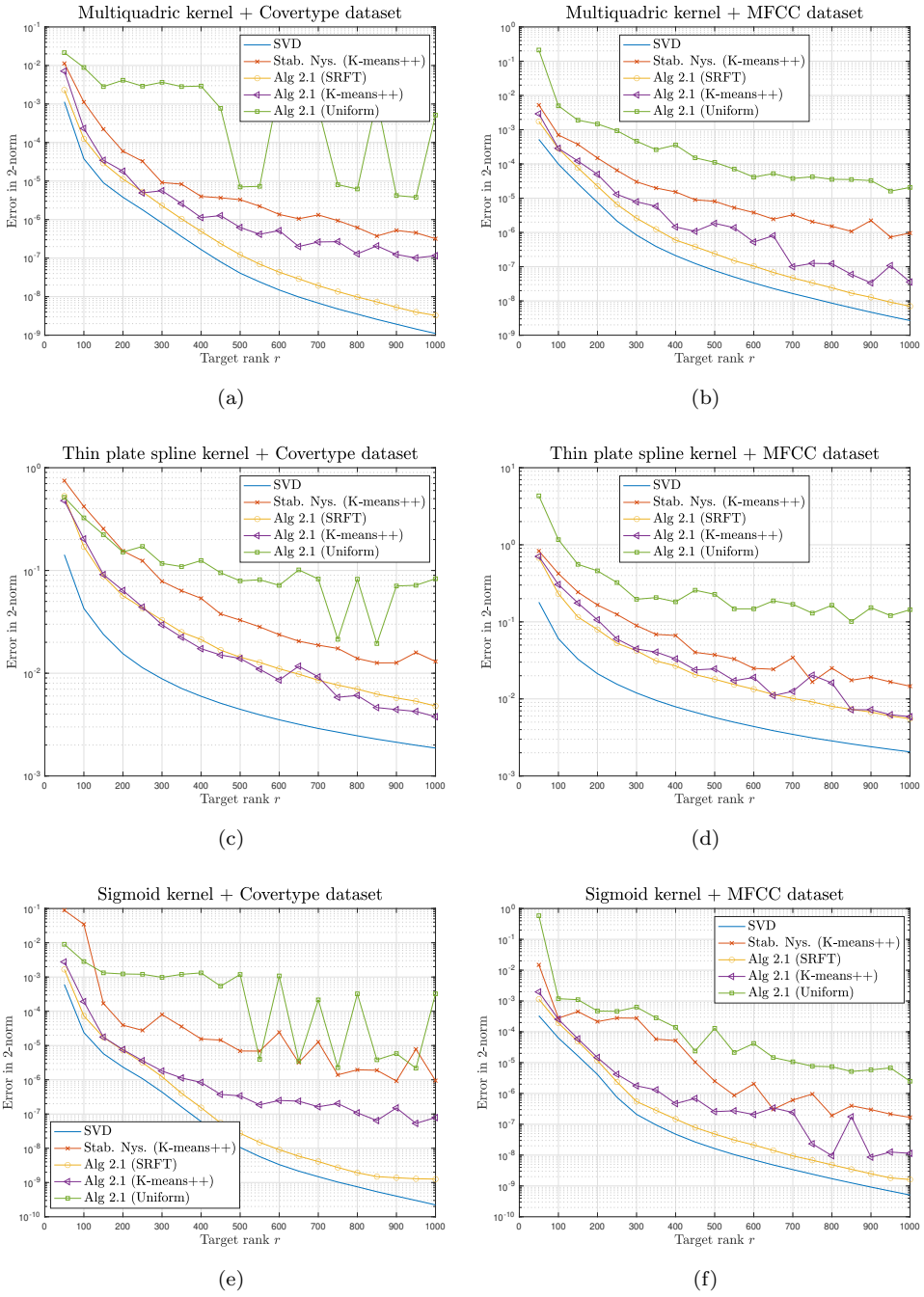


FIG. 6. *Comparison of stabilized Nyström [3] and Algorithm 2.1 for symmetric indefinite matrices using three different indefinite kernels and two different datasets. Stabilized Nyström method and Algorithm 2.1 using k-means++ samples and uniform column sampling can give unstable low-rank approximation while Algorithm 2.1 using the SRFT sketch (random embedding) gives robust approximation throughout the experiment.*

column subsampling.

Acknowledgements. We thank the anonymous referees and the editor for their many insightful comments and suggestions, which helped us to improve the quality of the paper.

REFERENCES

[1] A. Andoni and H. L. Nguyên, *Eigenvalues of a matrix in the streaming model*, in Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, 2013, pp. 1729–1737, https://doi.org/10.1137/1.9781611973105.124.

[2] C. Boutsidis and A. Gittens, *Improved matrix algorithms via the subsampled randomized Hadamard transform*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1301–1340, https://doi.org/10.1137/120874540.

[3] D. Cai, J. Nagy, and Y. Xi, *Fast deterministic approximation of symmetric indefinite kernel matrices with high dimensional datasets*, SIAM J. Matrix Anal. Appl., 43 (2022), pp. 1003–1028, https://doi.org/10.1137/21M1424627.

[4] K. L. Clarkson and D. P. Woodruff, *Low-rank approximation and regression in input sparsity time*, J. ACM, 63 (2017), pp. 1–45, https://doi.org/10.1145/3019134, https://doi.org/10.1145/3019134.

[5] M. B. Cohen, *Nearly tight oblivious subspace embeddings by trace inequalities*, in Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, 2016, pp. 278–287, https://doi.org/10.1137/1.9781611974331.ch21.

[6] A. Cortinovis and D. Kressner, *Low-rank approximation in the Frobenius norm by column and row subset selection*, SIAM J. Matrix Anal. Appl., 41 (2020), pp. 1651–1673.

[7] K. R. Davidson and S. J. Szarek, *Local operator theory, random matrices and Banach spaces*, in Handbook of the Geometry of Banach Spaces, W. Johnson and J. Lindenstrauss, eds., vol. 1, Elsevier, 2001, pp. 317–366, https://doi.org/https://doi.org/10.1016/S1874-5849(01)80010-3.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, Association for Computational Linguistics, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423.

[9] 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), pp. 2153–2175, http://jmlr.org/papers/v6/drineas05a.html.

[10] D. Dua and C. Graff, *UCI machine learning repository*, 2017, http://archive.ics.uci.edu/ml.

[11] Z. Frangella, J. A. Tropp, and M. Udell, *Randomized Nyström preconditioning*, arXiv preprint arXiv:2110.02820, (2021), https://doi.org/10.48550/ARXIV.2110.02820.

[12] A. Gisbrecht and F.-M. Schleif, *Metric and non-metric proximity transformations at linear costs*, Neurocomputing, 167 (2015), pp. 643–657, https://doi.org/https://doi.org/10.1016/j.neucom.2015.04.017.

[13] A. Gittens, *The spectral norm error of the naïve Nyström extension*, arXiv preprint arXiv:1110.5305, (2011), https://doi.org/10.48550/ARXIV.1110.5305.

[14] A. Gittens and M. W. Mahoney, *Revisiting the Nyström method for improved large-scale machine learning*, J. Mach. Learn. Res., 17 (2016), p. 3977–4041.

[15] N. Halko, P.-G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), p. 217–288, https://doi.org/10.1137/090771806.

[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 2 ed., 2012, https://doi.org/10.1017/9781139020411.

[17] H. Li, G. C. Linderman, A. Szlam, K. P. Stanton, Y. Kluger, and M. Tygert, *Algorithm 971: An implementation of a randomized algorithm for principal component analysis*, ACM Trans. Math. Softw., 43 (2017), https://doi.org/10.1145/3004053.

[18] M. Li, W. Bi, J. T. Kwok, and B.-L. Lu, *Large-scale Nyström kernel matrix approximation using randomized SVD*, IEEE Trans. Neural Netw. Learn. Syst., 26 (2015), pp. 152–164, https://doi.org/10.1109/TNNLS.2014.2359798.

[19] M. W. Mahoney and P. Drineas, *CUR matrix decompositions for improved data analysis*, Proceedings of the National Academy of Sciences, 106 (2009), pp. 697–702, https://doi.

org/10.1073/pnas.0803205106.
[20] P.-G. Martinsson and J. A. Tropp, *Randomized numerical linear algebra: Foundations and algorithms*, Acta Numer., 29 (2020), p. 403–572, https://doi.org/10.1017/s0962492920000021.
[21] C. Musco and C. Musco, *Recursive sampling for the Nyström method*, in Adv. Neural Inf. Process. Syst., I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper_files/paper/2017/file/a03fa30821986dff10fc66647c84c9c3-Paper.pdf.
[22] Y. Nakatsukasa, *Fast and stable randomized low-rank matrix approximation*, arXiv preprint arXiv:2009.11392, (2020), https://arxiv.org/abs/2009.11392.
[23] J. Nelson and H. L. Nguyên, *OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings*, in Proc. IEEE 54th Annu. Symp. Found. Comput. Sci., 2013, pp. 117–126, https://doi.org/10.1109/FOCS.2013.21.
[24] E. J. Nyström, *Über die praktische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben*, Acta Math., 54 (1930), pp. 185 – 204, https://doi.org/10.1007/BF02547521, https://doi.org/10.1007/BF02547521.
[25] D. Oglic and T. Gärtner, *Nyström method with kernel k-means++ samples as landmarks*, in Proceedings of the 34th International Conference on Machine Learning, D. Precup and Y. W. Teh, eds., vol. 70 of Proceedings of Machine Learning Research, PMLR, 06–11 Aug 2017, pp. 2652–2660, https://proceedings.mlr.press/v70/oglic17a.html.
[26] D. Oglic and T. Gärtner, *Scalable learning in reproducing kernel Kreĭn spaces*, in International Conference on Machine Learning, PMLR, 2019, pp. 4912–4921.
[27] B. Piccoli and F. Rossi, *Generalized Wasserstein Distance and its Application to Transport Equations with Source*, Archive for Rational Mechanics and Analysis, 211 (2014), pp. 335–358, https://doi.org/10.1007/s00205-013-0669-x, https://arxiv.org/abs/1206.3219.
[28] F. Pourkamali-Anaraki, S. Becker, and M. Wakin, *Randomized clustered Nyström for large-scale kernel machines*, Proceedings of the AAAI Conference on Artificial Intelligence, 32 (2018), pp. 3960–3967, https://doi.org/10.1609/aaai.v32i1.11614.
[29] A. Ray, N. Monath, A. McCallum, and C. Musco, *Sublinear time approximation of text similarity matrices*, Proceedings of the AAAI Conference on Artificial Intelligence, 36 (2022), pp. 8072–8080, https://doi.org/10.1609/aaai.v36i7.20779.
[30] T. Sarlos, *Improved approximation algorithms for large matrices via random projections*, in Proc. IEEE 47th Annu. Symp. Found. Comput. Sci., 2006, p. 143–152, https://doi.org/10.1109/FOCS.2006.37.
[31] J. A. Tropp, *Improved analysis of the subsampled randomized Hadamard transform*, Advances in Adaptive Data Analysis, 03 (2011), pp. 115–126, https://doi.org/10.1142/S1793536911000787.
[32] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, *Fixed-rank approximation of a positive-semidefinite matrix from streaming data*, in Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, p. 1225–1234.
[33] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, *Practical sketching algorithms for low-rank matrix approximation*, SIAM J. Matrix Anal. Appl., 38 (2017), p. 1454–1485, https://doi.org/10.1137/17m1111590.
[34] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, *Streaming low-rank matrix approximation with an application to scientific simulation*, SIAM J. Sci. Comp., 41 (2019), pp. A2430–A2463, https://doi.org/10.1137/18M1201068.
[35] M. Udell and A. Townsend, *Why are big data matrices approximately low rank?*, SIAM Journal on Mathematics of Data Science, 1 (2019), pp. 144–160, https://doi.org/10.1137/18M1183480.
[36] S. Wang, A. Gittens, and M. W. Mahoney, *Scalable kernel k-means clustering with Nyström approximation: Relative-error bounds*, J. Mach. Learn. Res., 20 (2019), p. 431–479.
[37] S. Wang, L. Luo, and Z. Zhang, *SPSD matrix approximation vis column selection: Theories, algorithms, and extensions*, J. Mach. Learn. Res., 17 (2014), pp. 49:1–49:49.
[38] C. Williams and M. Seeger, *Using the Nyström method to speed up kernel machines*, in Advances in Neural Information Processing Systems, T. Leen, T. Dietterich, and V. Tresp, eds., vol. 13, MIT Press, 2000, https://proceedings.neurips.cc/paper/2000/file/19de10adbaa1b2ee13f77f679fa1483a-Paper.pdf.
[39] D. P. Woodruff, *Sketching as a tool for numerical linear algebra*, Found. Trends Theor. Comput. Sci., 10 (2014), p. 1–157, https://doi.org/10.1561/0400000060.
[40] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert, *A fast randomized algorithm for the approximation of matrices*, Appl. Comput. Harmon. Anal., 25 (2008), pp. 335–366, https://doi.org/https://doi.org/10.1016/j.acha.2007.12.002.