

Improved fixed-rank Nyström approximation via QR decomposition: Practical and theoretical aspects

Farhad Pourkamali-Anaraki^{a,*}, Stephen Becker^b

^a Department of Computer Science, University of Massachusetts Lowell, MA, USA

^b Department of Applied Mathematics, University of Colorado Boulder, CO, USA

ARTICLE INFO

Article history:

Received 7 December 2018

Revised 13 April 2019

Accepted 25 June 2019

Available online 20 July 2019

Communicated by Dr. Yiming Ying

Keywords:

Kernel methods

Nyström approximation

Matrix factorization

ABSTRACT

The Nyström method is a popular technique that uses a small number of landmark points to compute a fixed-rank approximation of large kernel matrices that arise in machine learning problems. In practice, to ensure high quality approximations, the number of landmark points is chosen to be greater than the target rank. However, for simplicity the standard Nyström method uses a sub-optimal procedure for rank reduction. In this paper, we examine the drawbacks of the standard Nyström method in terms of poor performance and lack of theoretical guarantees. To address these issues, we present an efficient modification for generating improved fixed-rank Nyström approximations. Theoretical analysis and numerical experiments are provided to demonstrate the advantages of the modified method over the standard Nyström method. Overall, the aim of this paper is to convince researchers to use the modified method, as it has nearly identical computational complexity, is easy to code, has greatly improved accuracy in many cases, and is optimal in a sense that we make precise.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Kernel methods are widely used in various machine learning problems. Well-known examples include support vector machines [1,2], kernel clustering [3–6], and kernel ridge regression [7–10]. The main idea behind kernel-based learning is to map the input data points into a feature space, where all pairwise inner products of the mapped data points can be computed via a nonlinear kernel function that satisfies Mercer's condition [11]. Thus, kernel methods allow one to use linear algorithms in the feature space which correspond to nonlinear algorithms in the original space. For this reason, kernel machines have received much attention as an effective tool to tackle problems with complex and nonlinear structures.

Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be a set of n data points in \mathbb{R}^p . The inner products in feature space are calculated using a nonlinear kernel function $\kappa(\cdot, \cdot)$:

$$K_{ij} \stackrel{\text{def}}{=} \kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle, \quad \forall i, j \in \{1, \dots, n\}, \quad (1)$$

where $\Phi: \mathbf{x} \mapsto \Phi(\mathbf{x})$ is the kernel-induced feature map. A popular choice is the Gaussian kernel function $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/c)$, with the parameter $c > 0$. In kernel machines, the pairwise inner products are stored in the symmetric positive semidefinite

(SPSD) kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$. However, it takes $\mathcal{O}(n^2)$ memory to store the full kernel matrix and subsequent processing of \mathbf{K} within the learning process is quite expensive or prohibitive for large data sets.

A popular approach to tackle these challenges is to use the best rank- r approximation $\llbracket \mathbf{K} \rrbracket_r = \mathbf{U}_r \mathbf{\Lambda}_r \mathbf{U}_r^T$, obtained via the eigenvalue decomposition of \mathbf{K} , for $r \leq \text{rank}(\mathbf{K})$. Here, the columns of $\mathbf{U}_r \in \mathbb{R}^{n \times r}$ span the top r -dimensional eigenspace of \mathbf{K} , and the diagonal matrix $\mathbf{\Lambda}_r \in \mathbb{R}^{r \times r}$ contains the top r eigenvalues. Since the kernel matrix is SPSP, we have:

$$\mathbf{K} \approx \llbracket \mathbf{K} \rrbracket_r = \mathbf{U}_r \mathbf{\Lambda}_r \mathbf{U}_r^T = \mathbf{L} \mathbf{L}^T, \quad (2)$$

where $\mathbf{L} \stackrel{\text{def}}{=} \mathbf{U}_r \mathbf{\Lambda}_r^{1/2} \in \mathbb{R}^{n \times r}$.

When the target rank r is small and chosen independently of n (e.g., r is chosen according to the degrees of freedom in the learning problem [12]), the benefits of the rank- r approximation in (2) are twofold. First, it takes $\mathcal{O}(nr)$ to store the matrix \mathbf{L} which is only linear in the number of samples n . Second, the rank- r approximation leads to substantial computational savings within the learning process. For example, approximating \mathbf{K} with $\mathbf{L} \mathbf{L}^T$ means the matrix inversion $(\mathbf{K} + \lambda \mathbf{I}_{n \times n})^{-1}$ in kernel ridge regression can be calculated using the Sherman–Morrison–Woodbury formula in $\mathcal{O}(nr^2 + r^3)$ time compared to $\mathcal{O}(n^3)$ if done naïvely. Other examples are kernel K-means clustering, which is performed on the columns of the matrix $\mathbf{L}^T \in \mathbb{R}^{r \times n}$, and so each step of the K-means algorithm runs in time proportional to r .

* Corresponding author.

E-mail address: farhad_pourkamali@uml.edu (F. Pourkamali-Anaraki).

Although it has been shown that the fixed-rank approximation of kernel matrices is a promising approach to trade-off accuracy for scalability [13–16], the eigenvalue decomposition of \mathbf{K} has at least quadratic time complexity and takes $\mathcal{O}(n^2)$ space. To address this issue, one line of prior work is centered around efficient techniques for approximating the best rank- r approximation when we have ready access to \mathbf{K} ; see [17–19] for a survey. However, \mathbf{K} is typically unknown in kernel methods and the cost to form \mathbf{K} using standard kernel functions is $\mathcal{O}(pn^2)$, which is extremely expensive for large high-dimensional data sets. For this reason, the Nyström method [20] has been a popular technique for computing fixed-rank approximations, which eliminates the need to access every entry of the full kernel matrix. The Nyström method works by selecting a small set of vectors, referred to as landmark points, and computes the kernel similarities between the input data points and landmark points.

To be formal, the standard Nyström method generates a rank- r approximation of \mathbf{K} using m landmark points $\mathbf{z}_1, \dots, \mathbf{z}_m$ in \mathbb{R}^p . In practice, it is common to choose m greater than r for obtaining higher quality rank- r approximations [21,22], since the accuracy of the Nyström method depends on the number of selected landmark points and the selection procedure. The landmark points can be sampled with respect to a uniform or nonuniform distribution from the set of n input data points [23,24]. Moreover, some recent techniques utilize out-of-sample landmark points for generating improved Nyström approximations, e.g., centroids found from K-means clustering on the input data points [25–28]. For a fixed set of landmark points, let $\mathbf{C} \in \mathbb{R}^{n \times m}$ and $\mathbf{W} \in \mathbb{R}^{m \times m}$ be two matrices with the (i, j) th entries $C_{ij} = \kappa(\mathbf{x}_i, \mathbf{z}_j)$ and $W_{ij} = \kappa(\mathbf{z}_i, \mathbf{z}_j)$. Then, the rank- m Nyström approximation has the form $\mathbf{G} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$, where \mathbf{W}^\dagger is the pseudo-inverse of \mathbf{W} . For the fixed-rank case, the standard Nyström method restricts the rank of the $m \times m$ inner matrix \mathbf{W} and computes its best rank- r approximation $[\mathbf{W}]_r$ to obtain $\mathbf{G}_{(r)}^{\text{Nys}} = \mathbf{C}[\mathbf{W}]_r^\dagger\mathbf{C}^T$, which has rank no greater than r . The manner of the rank- m to rank- r reduction may appear *ad hoc*, and *improving this reduction is the topic of the paper*.

Although the rank reduction process in the standard Nyström method is simple, it disregards the structure of \mathbf{C} . This method generates the rank- r approximation $\mathbf{G}_{(r)}^{\text{Nys}}$ solely based on filtering \mathbf{W} because of its smaller size compared to the matrix \mathbf{C} of size $n \times m$. As a result, the selection of more distinct landmark points in the standard Nyström method does not guarantee improved rank- r approximations of kernel matrices. For example, our experimental results in Section 7 reveal that the increase in the number of landmark points may even produce less accurate rank- r approximations due to the poor rank reduction process, cf. Remarks 7 and 9.

This paper considers the fundamental problem of rank reduction in the Nyström method. In particular, we present an efficient technique for computing a rank- r approximation in the form of $\mathbf{G}_{(r)}^{\text{opt}} = [\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T]_r$, which runs in time comparable with the standard Nyström method. The modified method utilizes the thin QR decomposition of the matrix \mathbf{C} for computing a more accurate rank- r approximation of $\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$ compared to $\mathbf{G}_{(r)}^{\text{Nys}}$. Moreover, unlike the standard Nyström method, our results show that both theoretically and empirically, modified Nyström produces more accurate rank- r approximations as the number of distinct landmark points increases.

1.1. Contributions

In this work, we make the following contributions:

1. In Algorithm 2, we present an efficient method for generating improved rank- r Nyström approximations. The modified method computes the best rank- r approximation of $\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$, i.e., $\mathbf{G}_{(r)}^{\text{opt}} = [\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T]_r$, in linear time with respect to the sample size

n . In Theorem 3, it is shown that $\mathbf{G}_{(r)}^{\text{opt}}$ always produces a more accurate rank- r approximation of \mathbf{K} compared to $\mathbf{G}_{(r)}^{\text{Nys}}$ with respect to the trace norm, when m is greater than r and landmark points are selected from the input data set. Remark 5 shows this is not necessarily true in the Frobenius norm, although it is rarely seen in practice.

2. Theorem 6 proves that the accuracy of the modified rank- r Nyström approximation always improves (with respect to the trace norm) as more distinct landmark points are selected from the input data set.
3. We provide counter-examples in Remarks 7 and 9 showing that an equivalent of Theorem 6 cannot hold for the standard Nyström method. Example 2 shows a situation where the modified Nyström method is arbitrarily better than the standard method, with respect to the trace and Frobenius norms. Remark 4 gives insight into when we expect the standard and modified methods to differ.
4. Theorem 8 shows that, under certain conditions, our theoretical results are also applicable to more recent selection techniques based on out-of-sample extensions of the input data, such as centroids found from K-means clustering.
5. Finally, we provide experimental results to demonstrate the superior performance and advantages of modified Nyström.

To our knowledge, the modified Nyström method was not discussed in the literature until our preliminary preprint [29], though its derivation is straightforward and so we suspect it may have been previously derived in unpublished work; our main contribution is the mathematical analysis. Due to the importance of rank reduction in the Nyström method, there are two recent works [30,31] that independently study the approximation error of $[\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T]_r$, when landmark points are selected from the input data set. However, there are two principal differences between this work and the aforementioned references. First, the main focus of this paper is to directly compare the standard and modified Nyström methods, and provide both theoretical and experimental evidences on the effectiveness of modified Nyström, while [30,31] do not provide results comparing the two methods. Second, we present theoretical results for the important class of out-of-sample landmark points, which often lead to accurate Nyström approximations.

1.2. Paper organization

In Section 2, we present the notation and give a brief review of some matrix decomposition and low-rank approximation techniques. Section 3 reviews the standard Nyström method for computing rank- r approximations and we explain the process of obtaining approximate eigenvalues and eigenvectors. In Section 4, we present an efficient modified method for computing improved rank- r approximations of kernel matrices. The main theoretical results are given in Sections 5 and 6, and we present experimental results comparing the modified and standard Nyström methods in Section 7. Section 8 provides a brief conclusion.

2. Notation and preliminaries

We denote column vectors with lower-case bold letters and matrices with upper-case bold letters. $\mathbf{I}_{n \times n}$ is the identity matrix of size $n \times n$; $\mathbf{0}_{n \times m}$ is the $n \times m$ matrix of zeros. For a vector $\mathbf{x} \in \mathbb{R}^p$, let $\|\mathbf{x}\|_2$ denote the Euclidean norm, and $\text{diag}(\mathbf{x})$ represents a diagonal matrix with the elements of \mathbf{x} on the main diagonal. The (i, j) th entry of \mathbf{A} is denoted by A_{ij} , \mathbf{A}^T is the transpose of \mathbf{A} , and $\text{tr}(\cdot)$ is the trace operator. We assume scalars, vectors and matrices are real-valued, though many of the results extend to complex numbers.

Each $n \times m$ matrix \mathbf{A} with $\rho = \text{rank}(\mathbf{A}) \leq \min\{n, m\}$ admits a factorization in the form of $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{n \times \rho}$ and $\mathbf{V} \in \mathbb{R}^{m \times \rho}$ are orthonormal matrices known as the left singular vectors and right singular vectors, respectively. The diagonal matrix $\Sigma = \text{diag}([\sigma_1(\mathbf{A}), \dots, \sigma_\rho(\mathbf{A})])$ contains the singular values of \mathbf{A} in descending order, i.e., $\sigma_1(\mathbf{A}) \geq \dots \geq \sigma_\rho(\mathbf{A}) > 0$. This factorization is known as the *thin* singular value decomposition (SVD).

Throughout the paper, we use several standard matrix norms. The Frobenius norm of \mathbf{A} is defined as $\|\mathbf{A}\|_F^2 \stackrel{\text{def}}{=} \sum_{i=1}^{\rho} \sigma_i(\mathbf{A})^2 = \text{tr}(\mathbf{A}^T\mathbf{A})$ and $\|\mathbf{A}\|_* \stackrel{\text{def}}{=} \sum_{i=1}^{\rho} \sigma_i(\mathbf{A}) = \text{tr}(\sqrt{\mathbf{A}^T\mathbf{A}})$ denotes the trace norm (or nuclear norm) of \mathbf{A} . The spectral norm of \mathbf{A} is the largest singular value of \mathbf{A} , i.e., $\|\mathbf{A}\|_2 \stackrel{\text{def}}{=} \sigma_1(\mathbf{A})$. It is straightforward to show that $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \|\mathbf{A}\|_*$.

When \mathbf{K} is a *kernel matrix*, meaning it is generated via (1), we assume the kernel function κ satisfies Mercer's condition and therefore \mathbf{K} is symmetric positive semidefinite (SPSD) [11,32]. Let $\mathbf{K} \in \mathbb{R}^{n \times n}$ be any SPSP matrix with $\rho = \text{rank}(\mathbf{K}) \leq n$. Similar to the SVD, the matrix \mathbf{K} can be factorized as $\mathbf{K} = \mathbf{U}\Lambda\mathbf{U}^T$, where $\mathbf{U} \in \mathbb{R}^{n \times \rho}$ contains the orthonormal eigenvectors, i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{\rho \times \rho}$, and $\Lambda = \text{diag}([\lambda_1(\mathbf{K}), \dots, \lambda_\rho(\mathbf{K})])$ is a diagonal matrix which contains the nonzero eigenvalues of \mathbf{K} in descending order. This factorization is known as the *thin* eigenvalue decomposition (EVD). The matrices \mathbf{U} and Λ can be partitioned for a target rank r ($r \leq \rho$) in the form of $\mathbf{K} = \mathbf{U}_r\Lambda_r\mathbf{U}_r^T + \mathbf{U}_{\rho-r}\Lambda_{\rho-r}\mathbf{U}_{\rho-r}^T$, where $\mathbf{U}_r \in \mathbb{R}^{n \times r}$ contains the r leading eigenvalues and the columns of $\mathbf{U}_r \in \mathbb{R}^{n \times r}$ span the top r -dimensional eigenspace, and $\Lambda_{\rho-r} \in \mathbb{R}^{(\rho-r) \times (\rho-r)}$ and $\mathbf{U}_{\rho-r} \in \mathbb{R}^{n \times (\rho-r)}$ contain the remaining $(\rho - r)$ eigenvalues and eigenvectors. It is well-known that $[\mathbf{K}]_r \stackrel{\text{def}}{=} \mathbf{U}_r\Lambda_r\mathbf{U}_r^T$ is the “best rank- r approximation” to \mathbf{K} in the sense that $[\mathbf{K}]_r$ minimizes $\|\mathbf{K} - \mathbf{K}'\|_F$ and $\|\mathbf{K} - \mathbf{K}'\|_*$ over all matrices $\mathbf{K}' \in \mathbb{R}^{n \times n}$ of rank at most r . If $\lambda_r(\mathbf{K}) = \lambda_{r+1}(\mathbf{K})$, then $[\mathbf{K}]_r$ is not unique. The Moore-Penrose pseudo-inverse of \mathbf{K} can be obtained from the EVD as $\mathbf{K}^\dagger = \mathbf{U}\Lambda^{-1}\mathbf{U}^T$. When \mathbf{K} is full rank, we have $\mathbf{K}^\dagger = \mathbf{K}^{-1}$.

Another matrix factorization technique that we use is the QR decomposition. An $n \times m$ matrix \mathbf{A} , with $n \geq m$, can be decomposed as $\mathbf{A} = \mathbf{Q}\mathbf{R}$, where $\mathbf{Q} \in \mathbb{R}^{n \times m}$ has m orthonormal columns, i.e., $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_{m \times m}$, and $\mathbf{R} \in \mathbb{R}^{m \times m}$ is an upper triangular matrix. Sometimes this is called the *thin* QR decomposition, to distinguish it from a *full* QR decomposition which finds $\mathbf{Q} \in \mathbb{R}^{n \times n}$ and zero-pads \mathbf{R} accordingly.

Finally, we state a standard result on the rank- r approximation of a matrix expressed as a product of two matrices. The proof of this result can be found in [33].

Lemma 1. Consider the matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ and let $\mathbf{Q} \in \mathbb{R}^{n \times m}$ be a matrix that has $m < n$ orthonormal columns. For any positive integer $r \leq m$, we have:

$$\|\mathbf{Q}^T\mathbf{K}\|_r = \arg \min_{\mathbf{T}: \text{rank}(\mathbf{T}) \leq r} \|\mathbf{K} - \mathbf{Q}\mathbf{T}\|_F^2. \quad (3)$$

3. The standard Nyström method

The Nyström method generates a fixed-rank approximation of the SPSP kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ by selecting a small set of vectors referred to as “landmark points”. The simplest selection technique is uniform sampling without replacement [20,21], where each data point is sampled with the same probability, i.e., $p_i = \frac{1}{n}$, for $i = 1, \dots, n$. The advantage of this technique is the low computational complexity associated with sampling landmark points. However, uniform sampling does not take into account the nonuniform structure of many data sets and the resulting kernel matrices. Therefore, sampling mechanisms with respect to nonuniform distributions have been proposed to address this problem. This line of work requires the computation of statistical leverage scores of \mathbf{K} , which is more expensive than uniform sampling [34–36]. In addition, leverage score sampling often requires computing the entire

kernel matrix \mathbf{K} , which negates one of the principal benefits of the Nyström method. A comprehensive review and comparison of uniform and nonuniform landmark selection techniques can be found in [21,37].

More recently, generating landmark points using out-of-sample extensions of input data has been shown to be effective for high quality Nyström approximations. This line of research originates from the work of Zhang et al. [25,26], and it is based on the observation that the Nyström approximation error depends on the quantization error of encoding the data set with the landmark points. Hence, the landmark points are selected to be the centroids found from K-means clustering. In machine learning and pattern recognition, K-means clustering is a well-established technique to partition a data set into clusters by trying to minimize the total sum of the squared Euclidean distances of each point to the closest cluster center [38].

In general, assume that a set of $m \ll n$ landmark points in \mathbb{R}^p , denoted by $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_m] \in \mathbb{R}^{p \times m}$, are given. Let us consider two matrices $\mathbf{C} \in \mathbb{R}^{n \times m}$ and $\mathbf{W} \in \mathbb{R}^{m \times m}$, where $C_{ij} = \kappa(\mathbf{x}_i, \mathbf{z}_j)$ and $W_{ij} = \kappa(\mathbf{z}_i, \mathbf{z}_j)$. The Nyström method uses both \mathbf{C} and \mathbf{W} to construct an approximation of the kernel matrix \mathbf{K} in the form of $\mathbf{K} \approx \mathbf{G} = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$, which has rank at most m . For the fixed-rank case, the Nyström method generates a rank- r approximation of the kernel matrix, $r \leq m$, by computing the best rank- r approximation of the inner matrix \mathbf{W} [22,23,39], which results in $\mathbf{G}_{(r)}^{nys} = \mathbf{C}[\mathbf{W}]_r^\dagger\mathbf{C}^T$, where $[\mathbf{W}]_r^\dagger$ represents the pseudo-inverse of $[\mathbf{W}]_r$. Thus, the EVD of the matrix $\mathbf{W} = \mathbf{V}\Sigma\mathbf{V}^T$ should be computed to find the top r eigenvalues $\Sigma_r \in \mathbb{R}^{r \times r}$ and corresponding eigenvectors $\mathbf{V}_r \in \mathbb{R}^{m \times r}$:

$$\mathbf{G}_{(r)}^{nys} = \mathbf{L}^{nys}(\mathbf{L}^{nys})^T, \quad \mathbf{L}^{nys} = \mathbf{C}\mathbf{V}_r(\Sigma_r^{-1})^{1/2}. \quad (4)$$

The time complexity of the Nyström method to form \mathbf{L}^{nys} is $\mathcal{O}(pnm + m^2r + nmr)$, where it takes $\mathcal{O}(pnm)$ to construct \mathbf{C} and \mathbf{W} . It takes $\mathcal{O}(m^2r)$ time to perform the partial EVD of \mathbf{W} and $\mathcal{O}(nmr)$ represents the cost of the matrix multiplication $\mathbf{C}\mathbf{V}_r$. Thus, for $r \leq m \ll n$, the computation cost to form the rank- r approximation of the kernel matrix is only linear in the data set size n . The eigenvalues and eigenvectors of \mathbf{K} can be estimated by using the rank- r approximation in (4), and in fact this approach provides the exact eigenvalue decomposition of $\mathbf{G}_{(r)}^{nys}$. The first step is to find the EVD of the $r \times r$ matrix: $(\mathbf{L}^{nys})^T\mathbf{L}^{nys} = \tilde{\mathbf{V}}\tilde{\Sigma}\tilde{\mathbf{V}}^T$, where $\tilde{\mathbf{V}}, \tilde{\Sigma} \in \mathbb{R}^{r \times r}$. Then, the estimates of r leading eigenvalues and eigenvectors of \mathbf{K} are obtained as: $\hat{\mathbf{U}}_r^{nys} = \mathbf{L}^{nys}\tilde{\mathbf{V}}(\tilde{\Sigma}^{-1})^{1/2}$ and $\hat{\Lambda}_r^{nys} = \tilde{\Sigma}$. The procedure to estimate the r leading eigenvalues/eigenvectors is summarized in Algorithm 1.

Algorithm 1 Standard Nyström.

Input: data set \mathbf{X} , m landmark points \mathbf{Z} , kernel function κ , target rank r

Output: estimates of r leading eigenvectors and eigenvalues of the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$: $\hat{\mathbf{U}}_r^{nys} \in \mathbb{R}^{n \times r}$, $\hat{\Lambda}_r^{nys} \in \mathbb{R}^{r \times r}$

- 1: Form \mathbf{C} and \mathbf{W} : $C_{ij} = \kappa(\mathbf{x}_i, \mathbf{z}_j)$, $W_{ij} = \kappa(\mathbf{z}_i, \mathbf{z}_j)$
 - 2: Compute EVD: $\mathbf{W} = \mathbf{V}\Sigma\mathbf{V}^T$
 - 3: Form the matrix: $\mathbf{L}^{nys} = \mathbf{C}\mathbf{V}_r(\Sigma_r^{-1})^{1/2}$
 - 4: Compute EVD: $(\mathbf{L}^{nys})^T\mathbf{L}^{nys} = \tilde{\mathbf{V}}\tilde{\Sigma}\tilde{\mathbf{V}}^T$
 - 5: $\hat{\mathbf{U}}_r^{nys} = \mathbf{L}^{nys}\tilde{\mathbf{V}}(\tilde{\Sigma}^{-1})^{1/2}$ and $\hat{\Lambda}_r^{nys} = \tilde{\Sigma}$
-

4. Improved Nyström approximation via QR decomposition

In the previous section, we explained the Nyström method for computing rank- r approximations of SPSP kernel matrices based on selecting a small set of landmark points. Although the final goal

is to find an approximation that has rank no greater than r , it is often preferred to select $m > r$ landmark points and then restrict the resultant approximation to have rank at most r . The main intuition is that selecting $m > r$ landmark points and then restricting the approximation to a lower rank- r space has a regularization effect which can lead to more accurate approximations [21,23]. For example, when landmark points are chosen to be centroids from K-means clustering, more landmark points lead to smaller quantization error of the data set, and thus higher quality Nyström approximations.

In the standard Nyström method presented in Algorithm 1, the rank of the matrix \mathbf{G} is restricted by computing the best rank- r approximation of the inner matrix \mathbf{W} : $\mathbf{G}_{(r)}^{nys} = \mathbf{C}[\mathbf{W}]_1^{\dagger} \mathbf{C}^T$. Since the inner matrix in the representation of $\mathbf{G}_{(r)}^{nys}$ has rank no greater than r , it follows that $\mathbf{G}_{(r)}^{nys}$ has rank at most r . The main benefit of this technique is the low computational cost of performing an exact eigenvalue decomposition on a relatively small matrix of size $m \times m$. However, the standard Nyström method ignores the structure of the matrix \mathbf{C} in the rank reduction process. In fact, since the rank- r approximation $\mathbf{G}_{(r)}^{nys}$ does not utilize the full knowledge of \mathbf{C} , the selection of more landmark points does not guarantee an improved low-rank approximation in the standard Nyström method, cf. Remarks 7 and 9.

To solve this problem, we present an efficient method to compute the best rank- r approximation of $\mathbf{G} = \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T$, for given matrices $\mathbf{C} \in \mathbb{R}^{n \times m}$ and $\mathbf{W} \in \mathbb{R}^{m \times m}$. In contrast with the standard Nyström method, the modified approach takes advantage of both matrices \mathbf{C} and \mathbf{W} . To begin, let us consider the best rank- r approximation of \mathbf{G} in any unitarily invariant norm $\|\cdot\|$, such as the Frobenius norm or trace norm:

$$\begin{aligned} \mathbf{G}_{(r)}^{opt} &\stackrel{\text{def}}{=} \arg \min_{\mathbf{G}': \text{rank}(\mathbf{G}') \leq r} \|\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T - \mathbf{G}'\| \\ &\stackrel{(i)}{=} \arg \min_{\mathbf{G}': \text{rank}(\mathbf{G}') \leq r} \|\underbrace{\mathbf{Q}\mathbf{R}\mathbf{W}^{\dagger}\mathbf{R}^T}_{m \times m} \mathbf{Q}^T - \mathbf{G}'\| \\ &\stackrel{(ii)}{=} \arg \min_{\mathbf{G}': \text{rank}(\mathbf{G}') \leq r} \|(\mathbf{Q}\mathbf{V}')^T \Sigma' (\mathbf{Q}\mathbf{V}') - \mathbf{G}'\| = (\mathbf{Q}\mathbf{V}')_r \Sigma'_r (\mathbf{Q}\mathbf{V}')_r^T, \end{aligned} \quad (5)$$

where (i) follows from the QR decomposition of $\mathbf{C} \in \mathbb{R}^{n \times m}$; $\mathbf{C} = \mathbf{Q}\mathbf{R}$, where $\mathbf{Q} \in \mathbb{R}^{n \times m}$ and $\mathbf{R} \in \mathbb{R}^{m \times m}$. To get (ii), the EVD of the $m \times m$ matrix $\mathbf{R}\mathbf{W}^{\dagger}\mathbf{R}^T$ is computed: $\mathbf{R}\mathbf{W}^{\dagger}\mathbf{R}^T = \mathbf{V}'\Sigma'\mathbf{V}'^T$, where the diagonal matrix $\Sigma' \in \mathbb{R}^{m \times m}$ contains m eigenvalues in descending order on the main diagonal and the columns of $\mathbf{V}' \in \mathbb{R}^{m \times m}$ are the corresponding eigenvectors. Moreover, we note that the columns of $\mathbf{Q}\mathbf{V}' \in \mathbb{R}^{n \times m}$ are orthonormal because both \mathbf{Q} and \mathbf{V}' have orthonormal columns. Thus, the decomposition $(\mathbf{Q}\mathbf{V}')^T \Sigma' (\mathbf{Q}\mathbf{V}')^T$ contains the m eigenvalues and orthonormal eigenvectors of the Nyström approximation $\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T$. Hence, the best rank- r approximation of $\mathbf{G} = \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T$ is then computed using the r leading eigenvalues $\Sigma'_r \in \mathbb{R}^{r \times r}$ and corresponding eigenvectors $\mathbf{Q}\mathbf{V}'_r \in \mathbb{R}^{n \times r}$, as given (5). Thus, the estimates of the top r eigenvalues and eigenvectors of the kernel matrix \mathbf{K} from the Nyström approximation $\mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T$ are obtained as: $\hat{\mathbf{U}}_r^{opt} = \mathbf{Q}\mathbf{V}'_r$ and $\hat{\Lambda}_r^{opt} = \Sigma'_r$.

The modified method for estimating the r leading eigenvalues/eigenvectors of the kernel matrix is presented in Algorithm 2. The time complexity is $\mathcal{O}(pnm + nm^2 + m^3 + nmr)$, where $\mathcal{O}(pnm)$ represents the cost to form \mathbf{C} and \mathbf{W} . The complexity of the QR decomposition is $\mathcal{O}(nm^2)$ and it takes $\mathcal{O}(m^3)$ time to compute the EVD of $\mathbf{R}\mathbf{W}^{\dagger}\mathbf{R}^T$. Finally, the cost to compute the matrix multiplication $\mathbf{Q}\mathbf{V}'_r$ is $\mathcal{O}(nmr)$.

We can compare the computational complexity of Nyström via QR decomposition with that of the standard Nyström method. Since our focus in this paper is on large-scale data sets with n large, we only consider terms involving n which lead to dominant computation costs. Based on our previous discussion, it takes

Algorithm 2 Nyström via QR Decomposition (“modified” Nyström).

Input: data set \mathbf{X} , m landmark points \mathbf{Z} , kernel function κ , target rank r

Output: estimates of r leading eigenvectors and eigenvalues of the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$: $\hat{\mathbf{U}}_r^{opt} \in \mathbb{R}^{n \times r}$, $\hat{\Lambda}_r^{opt} \in \mathbb{R}^{r \times r}$

- 1: Form \mathbf{C} and \mathbf{W} : $C_{ij} = \kappa(\mathbf{x}_i, \mathbf{z}_j)$, $W_{ij} = \kappa(\mathbf{z}_i, \mathbf{z}_j)$
- 2: Perform the thin QR decomposition: $\mathbf{C} = \mathbf{Q}\mathbf{R}$
- 3: Compute EVD: $\mathbf{R}\mathbf{W}^{\dagger}\mathbf{R}^T = \mathbf{V}'\Sigma'\mathbf{V}'^T$
- 4: $\hat{\mathbf{U}}_r^{opt} = \mathbf{Q}\mathbf{V}'_r$ and $\hat{\Lambda}_r^{opt} = \Sigma'_r$

$C_{nys} = \mathcal{O}(pnm + nmr + nr^2)$ time to compute the eigenvalue decomposition using the standard Nyström method, while the cost of the modified technique is $C_{opt} = \mathcal{O}(pnm + nmr + nm^2)$. Thus, for data of even moderate dimension with $p \gtrsim m$, the dominant term in both C_{nys} and C_{opt} is $\mathcal{O}(pnm)$. Hence, there is no significant increase in cost, as is the case in our runtime example shown in Fig. 4.

Example 2. In the rest of this section, we present a simple example to gain some intuition on the superior performance of the modified technique.

Let us consider a small kernel matrix of size 3×3 :

$$\mathbf{K} = \begin{bmatrix} 1 & 0 & 10 \\ 0 & 1.01 & 0 \\ 10 & 0 & 100 \end{bmatrix}. \quad (6)$$

One can find, for example, a data matrix \mathbf{X} that generates this kernel matrix as $\mathbf{K} = \mathbf{X}^T\mathbf{X}$. Here, the goal is to compute the rank $r = 1$ approximation of \mathbf{K} . Sample $m = 2$ columns of \mathbf{K} , and suppose we choose the first and second columns:

$$\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1.01 \\ 10 & 0 \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} 1 & 0 \\ 0 & 1.01 \end{bmatrix}. \quad (7)$$

In the standard Nyström method, the best rank-1 approximation of the matrix \mathbf{W} is first computed. Then, the rank-1 approximation of \mathbf{K} using standard Nyström is:

$$\mathbf{G}_{(1)}^{nys} = \mathbf{C}[\mathbf{W}]_1^{\dagger} \mathbf{C}^T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1.01 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \quad (8)$$

The normalized approximation error in terms of the Frobenius norm and trace norm is large: $\|\mathbf{K} - \mathbf{G}_{(1)}^{nys}\|_F / \|\mathbf{K}\|_F = 0.99$ and $\|\mathbf{K} - \mathbf{G}_{(1)}^{nys}\|_* / \|\mathbf{K}\|_* = 0.99$. On the other hand, the modified method first computes the QR decomposition of $\mathbf{C} = \mathbf{Q}\mathbf{R}$:

$$\mathbf{Q} = \begin{bmatrix} \frac{1}{\sqrt{101}} & 0 \\ 0 & 1 \\ \frac{10}{\sqrt{101}} & 0 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \sqrt{101} & 0 \\ 0 & 1.01 \end{bmatrix}. \quad (9)$$

Then, the product of three matrices $\mathbf{R}\mathbf{W}^{\dagger}\mathbf{R}^T$ is computed to find its EVD:

$$\begin{aligned} \mathbf{R}\mathbf{W}^{\dagger}\mathbf{R}^T &= \begin{bmatrix} \sqrt{101} & 0 \\ 0 & 1.01 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{1.01} \end{bmatrix} \begin{bmatrix} \sqrt{101} & 0 \\ 0 & 1.01 \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\mathbf{V}'} \underbrace{\begin{bmatrix} 101 & 0 \\ 0 & 1.01 \end{bmatrix}}_{\Sigma'} \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\mathbf{V}'^T}. \end{aligned} \quad (10)$$

Finally, the rank-1 approximation of the kernel matrix in the modified method is:

$$\mathbf{G}_{(1)}^{opt} = \begin{bmatrix} 1 & 0 \\ \sqrt{101} & 1 \\ 0 & 0 \\ 10 & 0 \\ \sqrt{101} & 0 \end{bmatrix} \begin{bmatrix} 101 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 10 \\ \sqrt{101} & 0 & \sqrt{101} \\ 0 & 1 & 0 \end{bmatrix} \\ = \begin{bmatrix} 1 & 0 & 10 \\ 0 & 0 & 0 \\ 10 & 0 & 100 \end{bmatrix}, \quad (11)$$

where $\|\mathbf{K} - \mathbf{G}_{(1)}^{opt}\|_F / \|\mathbf{K}\|_F = 0.01$ and $\|\mathbf{K} - \mathbf{G}_{(1)}^{opt}\|_* / \|\mathbf{K}\|_* = 0.01$. In fact, one can show that our approximation is the same as the best rank-1 approximation of \mathbf{K} , i.e., $\mathbf{G}_{(1)}^{opt} = [\mathbf{K}]_1$. Furthermore, by taking $K_{22} \searrow 1$ in (6), we can make the improvement of the modified method over the standard method arbitrarily large.

5. Main theoretical results

In order to compare the accuracy of modified Nyström with standard Nyström, we first provide an alternative formulation of these two methods. We assume the landmark points $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_m]$ are *in-sample*, meaning they are selected in any fashion (deterministic or random) from among the set of input data points, so that the matrix $\mathbf{C} \in \mathbb{R}^{n \times m}$ contains m columns of the kernel matrix \mathbf{K} . This column selection process can be viewed as forming a *sampling matrix* $\mathbf{P} \in \mathbb{R}^{n \times m}$ that has exactly one nonzero entry in each column, where its location corresponds to the index of the selected landmark point. Then, the matrix product $\mathbf{C} = \mathbf{K}\mathbf{P} \in \mathbb{R}^{n \times m}$ contains m columns sampled from the kernel matrix \mathbf{K} and $\mathbf{W} = \mathbf{P}^T \mathbf{K} \mathbf{P} \in \mathbb{R}^{m \times m}$ is the intersection of the m columns with the corresponding m rows of \mathbf{K} .

Let us define $\mathbf{D} \stackrel{\text{def}}{=} \mathbf{K}^{1/2} \mathbf{P} \in \mathbb{R}^{n \times m}$, which means that $\mathbf{C} = \mathbf{K}\mathbf{P} = \mathbf{K}^{1/2} \mathbf{D}$ and $\mathbf{W} = \mathbf{P}^T \mathbf{K} \mathbf{P} = \mathbf{D}^T \mathbf{D}$. Moreover, we consider the SVD of $\mathbf{D} = \mathbf{F} \mathbf{S} \mathbf{N}^T$, where the columns of $\mathbf{F} \in \mathbb{R}^{n \times m}$ are the left singular vectors of \mathbf{D} , and we have $\mathbf{S}, \mathbf{N} \in \mathbb{R}^{m \times m}$. Thus, we get the EVD of the matrix $\mathbf{W} = \mathbf{D}^T \mathbf{D} = \mathbf{N} \mathbf{S}^2 \mathbf{N}^T$. For simplicity of presentation, we assume \mathbf{D} and \mathbf{W} have full rank, though the results still hold as long as they have rank greater than or equal to r .

The rank- r approximation in standard Nyström is $\mathbf{G}_{(r)}^{nys} = \mathbf{L}^{nys} (\mathbf{L}^{nys})^T$:

$$\mathbf{L}^{nys} = \mathbf{C} (\mathbf{W}^\dagger)^{1/2} = (\mathbf{K}^{1/2} \mathbf{F} \mathbf{S} \mathbf{N}^T) (\mathbf{N} \mathbf{S}_r^{-2} \mathbf{N}_r^T)^{1/2} = \mathbf{K}^{1/2} \mathbf{F}_r \mathbf{N}_r^T, \quad (12)$$

where we have used $(\mathbf{N}_r \mathbf{S}_r^{-2} \mathbf{N}_r^T)^{1/2} = \mathbf{N}_r \mathbf{S}_r^{-1} \mathbf{N}_r^T$, and the following two properties:

$$\mathbf{N}_r^T \mathbf{N}_r = \begin{bmatrix} \mathbf{I}_{r \times r} \\ \mathbf{0}_{(m-r) \times r} \end{bmatrix}, \quad \mathbf{F} \begin{bmatrix} \mathbf{N}_r^T \\ \mathbf{0}_{(m-r) \times m} \end{bmatrix} = \mathbf{F}_r \mathbf{N}_r^T. \quad (13)$$

Since the columns of \mathbf{N}_r are orthonormal, i.e., $\mathbf{N}_r^T \mathbf{N}_r = \mathbf{I}_{r \times r}$, the rank- r approximation of the kernel matrix \mathbf{K} in the standard Nyström method is given by:

$$\mathbf{G}_{(r)}^{nys} = \mathbf{L}^{nys} (\mathbf{L}^{nys})^T = \mathbf{K}^{1/2} \mathbf{F}_r \mathbf{F}_r^T \mathbf{K}^{1/2}. \quad (14)$$

Next, we present an alternative formulation of the rank- r approximation $\mathbf{G}_{(r)}^{opt}$ in terms of the left singular vectors of \mathbf{D} . The modified Nyström method finds the best rank- r approximation of $\mathbf{C} \mathbf{W}^\dagger \mathbf{C}^T$, and observe that:

$$\mathbf{C} (\mathbf{W}^\dagger)^{1/2} = (\mathbf{K}^{1/2} \mathbf{F} \mathbf{S} \mathbf{N}^T) (\mathbf{N} \mathbf{S}^{-1} \mathbf{N}^T) = \mathbf{K}^{1/2} \mathbf{F} \mathbf{N}^T. \quad (15)$$

Thus, we get $\mathbf{C} \mathbf{W}^\dagger \mathbf{C}^T = \mathbf{K}^{1/2} \mathbf{F} \mathbf{F}^T \mathbf{K}^{1/2}$, and the best rank- r approximation is:

$$\mathbf{G}_{(r)}^{opt} = [\mathbf{C} \mathbf{W}^\dagger \mathbf{C}^T]_r = [\mathbf{K}^{1/2} \mathbf{F}]_r \mathbf{F}_r^T [\mathbf{F}^T \mathbf{K}^{1/2}]_r, \quad (16)$$

where we used $\mathbf{F}^T \mathbf{F} = \mathbf{I}_{m \times m}$. Based on [30, Lemma 6], let $\mathbf{H} \in \mathbb{R}^{n \times r}$ be the orthonormal bases of the rank- r matrix $[\mathbf{F}^T \mathbf{K}^{1/2}]_r \in \mathbb{R}^{n \times n}$. Then, we have $[\mathbf{F}^T \mathbf{K}^{1/2}]_r = \mathbf{H} \mathbf{H}^T \mathbf{K}^{1/2}$, which allows us to simplify (16):

$$\mathbf{G}_{(r)}^{opt} = \mathbf{K}^{1/2} \mathbf{H} \mathbf{H}^T \mathbf{K}^{1/2}. \quad (17)$$

In the following, we present a theorem which shows that the modified Nyström method generates improved rank- r approximation of \mathbf{K} compared to standard Nyström.

Theorem 3 (Modified Nyström is more accurate than standard Nyström). *Let $\mathbf{K} \in \mathbb{R}^{n \times n}$ be an SPSPD kernel matrix, and r be the target rank. Let \mathbf{P} be any $n \times m$ matrix, with $m \geq r$, such that $\mathbf{C} = \mathbf{K}\mathbf{P} \in \mathbb{R}^{n \times m}$ and $\mathbf{W} = \mathbf{P}^T \mathbf{K} \mathbf{P} \in \mathbb{R}^{m \times m}$. Then, we have:*

$$\|\mathbf{K} - \mathbf{G}_{(r)}^{opt}\|_* \leq \|\mathbf{K} - \mathbf{G}_{(r)}^{nys}\|_*, \quad (18)$$

where the Nyström method via QR decomposition generates $\mathbf{G}_{(r)}^{opt} = [\mathbf{C} \mathbf{W}^\dagger \mathbf{C}^T]_r$, and the standard Nyström method produces $\mathbf{G}_{(r)}^{nys} = \mathbf{C} [\mathbf{W}]_r^\dagger \mathbf{C}^T$.

Proof. We start with the alternative formulation of $\mathbf{G}_{(r)}^{opt} = \mathbf{K}^{1/2} \mathbf{H} \mathbf{H}^T \mathbf{K}^{1/2}$, where the columns of $\mathbf{H} \in \mathbb{R}^{n \times r}$ are orthonormal. Note that $\mathbf{K} - \mathbf{G}_{(r)}^{opt}$ is an SPSPD matrix, and thus its trace norm is equal to the trace of this matrix:

$$\|\mathbf{K} - \mathbf{G}_{(r)}^{opt}\|_* = \text{tr}(\mathbf{K}^{1/2} (\mathbf{I}_{n \times n} - \mathbf{H} \mathbf{H}^T) \mathbf{K}^{1/2}) \\ \stackrel{(i)}{=} \text{tr}(\mathbf{K}^{1/2} (\mathbf{I}_{n \times n} - \mathbf{H} \mathbf{H}^T)^2 \mathbf{K}^{1/2}) \\ = \|(\mathbf{I}_{n \times n} - \mathbf{H} \mathbf{H}^T) \mathbf{K}^{1/2}\|_F^2 \\ \stackrel{(ii)}{=} \|\mathbf{K}^{1/2} - \mathbf{F} [\mathbf{F}^T \mathbf{K}^{1/2}]_r\|_F^2 \stackrel{(iii)}{=} \min_{\mathbf{T}: \text{rank}(\mathbf{T}) \leq r} \|\mathbf{K}^{1/2} - \mathbf{F} \mathbf{T}\|_F^2, \quad (19)$$

where (i) follows from $(\mathbf{I}_{n \times n} - \mathbf{H} \mathbf{H}^T)^2 = (\mathbf{I}_{n \times n} - \mathbf{H} \mathbf{H}^T)$, (ii) is based on the observation $\mathbf{H} \mathbf{H}^T \mathbf{K}^{1/2} = \mathbf{F} [\mathbf{F}^T \mathbf{K}^{1/2}]_r$, and (iii) is based on Lemma 1. Let us define the function $f(\mathbf{T}) \stackrel{\text{def}}{=} \|\mathbf{K}^{1/2} - \mathbf{F} \mathbf{T}\|_F^2$, and consider the matrix $\mathbf{T}' \in \mathbb{R}^{m \times n}$ with rank no greater than r : $\mathbf{T}' = \begin{bmatrix} \mathbf{F}_r^T \\ \mathbf{0}_{(m-r) \times n} \end{bmatrix} \mathbf{K}^{1/2}$. Then, we see that:

$$\|\mathbf{K} - \mathbf{G}_{(r)}^{opt}\|_* \leq f(\mathbf{T}') \\ = \|\mathbf{K}^{1/2} - \mathbf{F}_r \mathbf{F}_r^T \mathbf{K}^{1/2}\|_F^2 \\ = \text{tr}(\mathbf{K}^{1/2} (\mathbf{I}_{n \times n} - \mathbf{F}_r \mathbf{F}_r^T)^2 \mathbf{K}^{1/2}) = \|\mathbf{K} - \mathbf{G}_{(r)}^{nys}\|_*, \quad (20)$$

where we used $\mathbf{F} \mathbf{T}' = \mathbf{F}_r \mathbf{F}_r^T \mathbf{K}^{1/2}$, $(\mathbf{I}_{n \times n} - \mathbf{F}_r \mathbf{F}_r^T)^2 = (\mathbf{I}_{n \times n} - \mathbf{F}_r \mathbf{F}_r^T)$, the alternative formulation of $\mathbf{G}_{(r)}^{nys} = \mathbf{K}^{1/2} \mathbf{F}_r \mathbf{F}_r^T \mathbf{K}^{1/2}$, cf. (14), and that $\mathbf{K} - \mathbf{G}_{(r)}^{nys}$ is SPSPD. \square

In Example 2, we showed that the modified method outperforms the standard Nyström method. The essential structural feature of the example was the presence of a large-magnitude block of the kernel matrix, denoted \mathbf{K}_{21} below. The following remark shows that when this block is zero, the two methods perform the same.

Remark 4. Let $\mathbf{P} \in \mathbb{R}^{n \times m}$ be the sampling matrix, where m columns of the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ are sampled according to any distribution. Without loss of generality, the matrices \mathbf{C} and \mathbf{K} can be permuted as follows:

$$\mathbf{K} = \begin{bmatrix} \mathbf{W} & \mathbf{K}_{21}^T \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{W} \\ \mathbf{K}_{21} \end{bmatrix}, \quad (21)$$

where $\mathbf{K}_{21} \in \mathbb{R}^{(n-m) \times m}$ and $\mathbf{K}_{22} \in \mathbb{R}^{(n-m) \times (n-m)}$. If $\mathbf{K}_{21} = \mathbf{0}_{(n-m) \times m}$, then Nyström via QR decomposition and the standard Nyström

method generate the same rank- r approximation of the kernel matrix \mathbf{K} , i.e., $\mathbf{G}_{(r)}^{opt} = \mathbf{G}_{(r)}^{nys}$.

Proof. Given that $\mathbf{K}_{21} = \mathbf{0}_{(n-m) \times m}$, we have $\mathbf{K}^{1/2} = \begin{bmatrix} \mathbf{W}^{1/2} & \mathbf{0}_{m \times (n-m)} \\ \mathbf{0}_{(n-m) \times m} & \mathbf{K}_{22}^{1/2} \end{bmatrix}$ and $\mathbf{D} = \mathbf{K}^{1/2} \mathbf{P} = \begin{bmatrix} \mathbf{W}^{1/2} \\ \mathbf{0}_{(n-m) \times m} \end{bmatrix}$. Let us consider the EVD of $\mathbf{W} = \mathbf{V} \mathbf{\Sigma} \mathbf{V}^T$, where $\mathbf{V}, \mathbf{\Sigma} \in \mathbb{R}^{m \times m}$. Then, the left singular vectors of \mathbf{D} have the following form: $\mathbf{F} = \begin{bmatrix} \mathbf{V} \\ \mathbf{0}_{(n-m) \times m} \end{bmatrix}$. Thus, we get $\mathbf{F}^T \mathbf{K}^{1/2} = [\mathbf{V}^T \mathbf{W}^{1/2}, \mathbf{0}_{m \times (n-m)}]$. Note that $\mathbf{V}^T \mathbf{W}^{1/2} = \mathbf{\Sigma}^{1/2} \mathbf{V}^T$, and the best rank- r approximation of $\mathbf{F}^T \mathbf{K}^{1/2}$ can be written as:

$$\llbracket \mathbf{F}^T \mathbf{K}^{1/2} \rrbracket_r = [\mathbf{Y} \mathbf{\Sigma}_r^{1/2} \mathbf{V}_r^T, \mathbf{0}_{m \times (n-m)}], \quad (22)$$

where $\mathbf{Y} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{I}_{r \times r} \\ \mathbf{0}_{(m-r) \times r} \end{bmatrix} \in \mathbb{R}^{m \times r}$. Next, we compute the matrix $\mathbf{T}' = \begin{bmatrix} \mathbf{F}_r^T \\ \mathbf{0}_{(m-r) \times n} \end{bmatrix} \mathbf{K}^{1/2}$ in the proof of Theorem 3 by first simplifying $\mathbf{F}_r^T \mathbf{K}^{1/2} = [\mathbf{V}_r^T \mathbf{W}^{1/2}, \mathbf{0}_{r \times (n-m)}]$. Since $\mathbf{V}_r^T \mathbf{W}^{1/2} = \mathbf{\Sigma}_r^{1/2} \mathbf{V}_r^T$, we have $\mathbf{T}' = \begin{bmatrix} \mathbf{F}_r^T \mathbf{K}^{1/2} \\ \mathbf{0}_{(m-r) \times n} \end{bmatrix} = [\mathbf{Y} \mathbf{\Sigma}_r^{1/2} \mathbf{V}_r^T, \mathbf{0}_{m \times (n-m)}]$. Thus, we get $\mathbf{T}' = \llbracket \mathbf{F}^T \mathbf{K}^{1/2} \rrbracket_r$ and this completes the proof. \square

Remark 5. In Theorem 3, we showed that Nyström via QR decomposition generates improved rank- r approximation of kernel matrices with respect to the trace norm. However, this property is not always satisfied in terms of the Frobenius norm. For example, consider the following 4×4 SPSP matrix:

$$\mathbf{K} = \begin{bmatrix} 1.0 & 0.7 & 0.9 & 0.4 \\ 0.7 & 1.0 & 0.6 & 0.6 \\ 0.9 & 0.6 & 1.0 & 0.6 \\ 0.4 & 0.6 & 0.6 & 1.0 \end{bmatrix}. \quad (23)$$

If we sample the first and second column of \mathbf{K} to form $\mathbf{C} \in \mathbb{R}^{4 \times 2}$, i.e., $m = 2$, then we get $\|\mathbf{K} - \mathbf{G}_{(1)}^{nys}\|_* = 1.3441$ and $\|\mathbf{K} - \mathbf{G}_{(1)}^{opt}\|_* = 1.3299$. Thus, we have $\|\mathbf{K} - \mathbf{G}_{(1)}^{opt}\|_* \leq \|\mathbf{K} - \mathbf{G}_{(1)}^{nys}\|_*$, as expected by Theorem 3. If we compare these two error terms based on the Frobenius norm, then we see that $\|\mathbf{K} - \mathbf{G}_{(1)}^{nys}\|_F = 0.9397$ and $\|\mathbf{K} - \mathbf{G}_{(1)}^{opt}\|_F = 0.9409$. Thus, in this example, the standard Nyström method has slightly better performance in terms of the Frobenius norm. To explain this observation, let us define $g(\mathbf{T}) \stackrel{\text{def}}{=} \|(\mathbf{K}^{1/2} - \mathbf{F}\mathbf{T})(\mathbf{K}^{1/2} - \mathbf{F}\mathbf{T})^T\|_F^2$ for an input matrix $\mathbf{T} \in \mathbb{R}^{m \times n}$ and fixed matrices \mathbf{K} and \mathbf{F} . Based on the proof of Theorem 3, it is straightforward to show that $\|\mathbf{K} - \mathbf{G}_{(r)}^{nys}\|_F^2 = g(\mathbf{T}^{nys})$, where $\mathbf{T}^{nys} = \begin{bmatrix} \mathbf{F}_r^T \\ \mathbf{0}_{(m-r) \times n} \end{bmatrix} \mathbf{K}^{1/2}$. Also, we have $\|\mathbf{K} - \mathbf{G}_{(r)}^{opt}\|_F^2 = g(\mathbf{T}^{opt})$, $\mathbf{T}^{opt} = \llbracket \mathbf{F}^T \mathbf{K}^{1/2} \rrbracket_r$. Additionally, we can express the error term $g(\mathbf{T})$ as the sum of the following three terms:

$$\begin{aligned} g(\mathbf{T}) &= \left\| \left(\mathbf{K}^{1/2} - \begin{bmatrix} \mathbf{F} & \mathbf{F}^\perp \end{bmatrix} \begin{bmatrix} \mathbf{T} \\ \mathbf{0} \end{bmatrix} \right) \left(\mathbf{K}^{1/2} - \begin{bmatrix} \mathbf{F} & \mathbf{F}^\perp \end{bmatrix} \begin{bmatrix} \mathbf{T} \\ \mathbf{0} \end{bmatrix} \right)^T \right\|_F^2 \\ &= \underbrace{\|(\mathbf{F}^T \mathbf{K}^{1/2} - \mathbf{T})(\mathbf{F}^T \mathbf{K}^{1/2} - \mathbf{T})^T\|_F^2}_{=g_1(\mathbf{T})} \\ &\quad + 2 \underbrace{\|(\mathbf{F}^T \mathbf{K}^{1/2} - \mathbf{T})(\mathbf{K}^{1/2} \mathbf{F}^\perp)\|_F^2}_{=g_2(\mathbf{T})} + g_3, \end{aligned} \quad (24)$$

where $\mathbf{F}^\perp \in \mathbb{R}^{n \times (n-m)}$ is the orthogonal complement of \mathbf{F} , we used the unitary invariance of the Frobenius norm, and $g_3 = \|(\mathbf{F}^\perp)^T \mathbf{K} \mathbf{F}^\perp\|_F^2$ is independent of \mathbf{T} . For the given kernel matrix in (23), we have $g_1(\mathbf{T}^{opt}) = 0.2524$, $g_2(\mathbf{T}^{opt}) = 0.0274$, $g_1(\mathbf{T}^{nys}) = 0.2669$, $g_2(\mathbf{T}^{nys}) = 0.0191$, and $g_3 = 0.5780$. Since $g_1(\mathbf{T}^{opt}) < g_1(\mathbf{T}^{nys})$ and $g_2(\mathbf{T}^{opt}) > g_2(\mathbf{T}^{nys})$ in this example, it is clear that the modified Nyström method does not necessarily lead to more accurate approximations with respect to the Frobenius norm.

Even though it possible that the standard Nyström can outperform Nyström via QR decomposition in the Frobenius norm, our substantial numerical experiments on real-world data sets in Section 7 show that this does not happen in practice.

Next, we present an important theoretical result on the quality of rank- r Nyström approximations when the number of landmark points are increased. Specifically, let us first sample $m_1 \geq r$ landmark points from the set of input data points to generate the rank- r approximation using the modified Nyström method, namely $\mathbf{G}_{(r)}^{opt}$. If we sample $(m_2 - m_1) \in \mathbb{N}$ additional landmark points to form the new rank- r approximation $\tilde{\mathbf{G}}_{(r)}^{opt}$ using the total of m_2 landmark points, the following result states that $\|\mathbf{K} - \tilde{\mathbf{G}}_{(r)}^{opt}\|_* \leq \|\mathbf{K} - \mathbf{G}_{(r)}^{opt}\|_*$. Therefore, increasing the number of distinct landmark points in the modified Nyström method leads to improved rank- r approximation.

Theorem 6 (More landmark points help). *Let $\mathbf{K} \in \mathbb{R}^{n \times n}$ be an SPSP kernel matrix, and r be the target rank. Consider a sampling matrix $\mathbf{P} \in \mathbb{R}^{n \times m_1}$, with $m_1 \geq r$, such that $\mathbf{C} = \mathbf{K}\mathbf{P} \in \mathbb{R}^{n \times m_1}$ and $\mathbf{W} = \mathbf{P}^T \mathbf{K} \mathbf{P} \in \mathbb{R}^{m_1 \times m_1}$. Then, the modified Nyström method generates the rank- r approximation of \mathbf{K} as $\mathbf{G}_{(r)}^{opt} = \llbracket \mathbf{C} \mathbf{W}^\dagger \mathbf{C}^T \rrbracket_r$. Increase the number of distinct landmark points by concatenating the matrix \mathbf{P} with $\mathbf{P}^{new} \in \mathbb{R}^{n \times (m_2 - m_1)}$ with $m_2 > m_1$, i.e., $\tilde{\mathbf{P}} = [\mathbf{P}, \mathbf{P}^{new}] \in \mathbb{R}^{n \times m_2}$. The resulting matrix $\tilde{\mathbf{P}}$ can be used to form $\tilde{\mathbf{C}} = \mathbf{K}\tilde{\mathbf{P}} \in \mathbb{R}^{n \times m_2}$ and $\tilde{\mathbf{W}} = \tilde{\mathbf{P}}^T \mathbf{K} \tilde{\mathbf{P}} \in \mathbb{R}^{m_2 \times m_2}$, and the modified Nyström method generates $\tilde{\mathbf{G}}_{(r)}^{opt} = \llbracket \tilde{\mathbf{C}} \tilde{\mathbf{W}}^\dagger \tilde{\mathbf{C}}^T \rrbracket_r$. Then this new approximation is better in the sense that:*

$$\|\mathbf{K} - \tilde{\mathbf{G}}_{(r)}^{opt}\|_* \leq \|\mathbf{K} - \mathbf{G}_{(r)}^{opt}\|_*. \quad (25)$$

Proof. Let $\mathbf{F} \in \mathbb{R}^{n \times m_1}$ and $\tilde{\mathbf{F}} \in \mathbb{R}^{n \times m_2}$ be the left singular vectors of $\mathbf{K}^{1/2} \mathbf{P} \in \mathbb{R}^{n \times m_1}$ and $\mathbf{K}^{1/2} \tilde{\mathbf{P}} = [\mathbf{K}^{1/2} \mathbf{P}, \mathbf{K}^{1/2} \mathbf{P}^{new}] \in \mathbb{R}^{n \times m_2}$, respectively. Then, we get:

$$\begin{aligned} \|\mathbf{K} - \tilde{\mathbf{G}}_{(r)}^{opt}\|_* &= \min_{\tilde{\mathbf{T}}: \text{rank}(\tilde{\mathbf{T}}) \leq r} \|\mathbf{K}^{1/2} - \tilde{\mathbf{F}} \tilde{\mathbf{T}}\|_F^2 \\ &\leq \min_{\mathbf{T}: \text{rank}(\mathbf{T}) \leq r} \|\mathbf{K}^{1/2} - \mathbf{F} \mathbf{T}\|_F^2 = \|\mathbf{K} - \mathbf{G}_{(r)}^{opt}\|_* \end{aligned} \quad (26)$$

where both equalities follow from (19) and the inequality follows from the fact that $\text{Range}(\mathbf{F}) \subset \text{Range}(\tilde{\mathbf{F}})$. \square

Remark 7. Theorem 6 is not true for the standard Nyström method. Consider the kernel matrix from Example 2. By sampling the first two columns, the standard Nyström method gave relative errors of 0.99 in both the trace and Frobenius norms. Had we sampled just the first column, the standard Nyström method would have 0.01 relative error in these norms, meaning that adding additional landmark points leads to a worse approximation. See also Remark 9 for experiments.

6. Extension to out-of-sample landmark points

The main component in our theoretical results is the existence of $\mathbf{P} \in \mathbb{R}^{n \times m}$ such that \mathbf{C} and \mathbf{W} can be written as: $\mathbf{C} = \mathbf{K}\mathbf{P}$ and $\mathbf{W} = \mathbf{P}^T \mathbf{K} \mathbf{P}$. As mentioned, this assumption holds true if the landmark points are selected (randomly or arbitrarily) from the set of input data points, since then \mathbf{P} is a sampling matrix consisting of columns of the identity matrix. However, some recent selection techniques utilize out-of-sample extensions of the input data to improve the accuracy of the Nyström method, e.g., centroids found from K-means clustering. In this case, the matrix $\mathbf{C} \in \mathbb{R}^{n \times m}$, where $C_{ij} = \kappa(\mathbf{x}_i, \mathbf{z}_j)$, does not necessarily contain the columns of \mathbf{K} . Thus, we cannot hope for a sampling matrix \mathbf{P} that satisfies $\mathbf{C} = \mathbf{K}\mathbf{P}$. In this section, we present two techniques to show that our theoretical results hold for the case of out-of-sample landmark points.

6.1. Approach 1: full-rank kernel matrices

We show that, under certain conditions, our theoretical results in Section 5 are applicable to the case of out-of-sample landmark points. To be formal, consider a set of n distinct data points in \mathbb{R}^p , i.e., $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$, and the Gaussian kernel of the form $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/c)$, $c > 0$, which leads to the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$. Let $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_m] \in \mathbb{R}^{p \times m}$ be m cluster centroids from K-means clustering on $\mathbf{x}_1, \dots, \mathbf{x}_n$, and we form $\mathbf{C} \in \mathbb{R}^{n \times m}$ and $\mathbf{W} \in \mathbb{R}^{m \times m}$ with $C_{ij} = \kappa(\mathbf{x}_i, \mathbf{z}_j)$ and $W_{ij} = \kappa(\mathbf{z}_i, \mathbf{z}_j)$.

Based on [11, Theorem 2.18], the kernel matrix \mathbf{K} defined on a set of n distinct data points using the Gaussian kernel function has full rank. Thus by defining $\mathbf{P} = \mathbf{K}^{-1}\mathbf{C} \in \mathbb{R}^{n \times m}$, we can write $\mathbf{C} = \mathbf{K}\mathbf{P}$, but because this \mathbf{P} is not a sampling matrix, it does not follow that $\mathbf{W} = \mathbf{P}^T\mathbf{K}\mathbf{P}$, so our aim is to show that $\mathbf{W} \approx \mathbf{P}^T\mathbf{K}\mathbf{P}$. Let us consider the empirical kernel map Φ_e , defined on the set of input data points:

$$\Phi_e(\mathbf{z}) : \mathbf{z} \mapsto \mathbf{K}^{-1/2}[\kappa(\mathbf{x}_1, \mathbf{z}), \dots, \kappa(\mathbf{x}_n, \mathbf{z})]^T \in \mathbb{R}^n. \quad (27)$$

This map approximates the kernel-induced map Φ for out-of-sample data points $\mathbf{z}_1, \dots, \mathbf{z}_m$ such that $\langle \Phi_e(\mathbf{z}_i), \Phi_e(\mathbf{z}_j) \rangle \approx \langle \Phi(\mathbf{z}_i), \Phi(\mathbf{z}_j) \rangle = \kappa(\mathbf{z}_i, \mathbf{z}_j)$ [11]. Since $\mathbf{P}^T = \mathbf{C}^T\mathbf{K}^{-1}$ and the j th column of \mathbf{C} is $[\kappa(\mathbf{x}_1, \mathbf{z}_j), \dots, \kappa(\mathbf{x}_n, \mathbf{z}_j)]^T$, we have:

$$\begin{aligned} \mathbf{P}^T\mathbf{K}\mathbf{P} &= (\mathbf{C}^T\mathbf{K}^{-1/2})(\mathbf{K}^{-1/2}\mathbf{C}) \\ &= [\Phi_e(\mathbf{z}_1), \dots, \Phi_e(\mathbf{z}_m)]^T [\Phi_e(\mathbf{z}_1), \dots, \Phi_e(\mathbf{z}_m)] \stackrel{\text{def}}{=} \mathbf{W}_e \in \mathbb{R}^{m \times m}. \end{aligned} \quad (28)$$

Therefore, if we use out-of-sample landmark points with the Gaussian kernel function, there exists a matrix \mathbf{P} that satisfies $\mathbf{C} = \mathbf{K}\mathbf{P}$ and $\mathbf{W} = \mathbf{W}_e + \mathbf{E} = \mathbf{P}^T\mathbf{K}\mathbf{P} + \mathbf{E}$, where $\mathbf{E} \in \mathbb{R}^{m \times m}$ represents the approximation error. It is known that when the relative amount of error is small (e.g., with respect to the spectral norm), \mathbf{W} and \mathbf{W}_e are close to one another and their eigenvalues and eigenvectors are perturbed proportional to the relative error [40,41]. However, in this work, our goal is to prove that the approximations $\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$ and $\mathbf{C}\mathbf{W}_e^\dagger\mathbf{C}^T$ are close to one another when the relative amount of error is small. To demonstrate the importance of this result, note that for any invertible matrices \mathbf{M} and \mathbf{M}' , we have the identity $\mathbf{M}'^{-1} - \mathbf{M}^{-1} = -\mathbf{M}'^{-1}(\mathbf{M}' - \mathbf{M})\mathbf{M}^{-1}$. Thus, the small norm of $\mathbf{M}' - \mathbf{M}$ cannot be directly used to conclude \mathbf{M}'^{-1} and \mathbf{M}^{-1} are close to one another. In the following, we present an error bound for the difference between the Nyström approximations, i.e., $\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T - \mathbf{C}\mathbf{W}_e^\dagger\mathbf{C}^T$, in terms of the relative amount of error caused by the empirical kernel map.

Theorem 8. Consider a set of n distinct data points $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$, and the Gaussian kernel function of the form $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/c)$, $c > 0$, which leads to the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$. Let $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_m] \in \mathbb{R}^{p \times m}$ be arbitrary (e.g., m distinct cluster centroids from K-means clustering on $\mathbf{x}_1, \dots, \mathbf{x}_n$), and we form $\mathbf{C} \in \mathbb{R}^{n \times m}$ and $\mathbf{W} \in \mathbb{R}^{m \times m}$ with $C_{ij} = \kappa(\mathbf{x}_i, \mathbf{z}_j)$ and $W_{ij} = \kappa(\mathbf{z}_i, \mathbf{z}_j)$. Then, there exists a matrix $\mathbf{P} \in \mathbb{R}^{n \times m}$ such that $\mathbf{C} = \mathbf{K}\mathbf{P}$ and $\mathbf{W} = \mathbf{W}_e + \mathbf{E}$, where $\mathbf{W}_e = \mathbf{P}^T\mathbf{K}\mathbf{P}$ and $\mathbf{E} \in \mathbb{R}^{m \times m}$ represents the approximation error of the empirical kernel map defined in (27). Assuming that $\eta \stackrel{\text{def}}{=} \|\mathbf{W}_e^{-1/2}\mathbf{E}\mathbf{W}_e^{-1/2}\|_2 < 1$ for the positive definite matrix \mathbf{W}_e , then:

$$\frac{\|\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T - \mathbf{C}\mathbf{W}_e^\dagger\mathbf{C}^T\|_2}{\|\mathbf{K}\|_2} \leq \frac{\eta}{1 - \eta}. \quad (29)$$

Proof. The kernel matrix \mathbf{K} using the Gaussian kernel function has full rank. Thus, there exists a matrix \mathbf{P} such that $\mathbf{C} = \mathbf{K}\mathbf{P}$ and the empirical kernel map Φ_e can be defined as in (27). Recall the SVD of $\mathbf{D} = \mathbf{K}^{1/2}\mathbf{P} = \mathbf{F}\mathbf{S}\mathbf{N}^T$, and the EVD of $\mathbf{W}_e = \mathbf{P}^T\mathbf{K}\mathbf{P} = \mathbf{N}\mathbf{S}^2\mathbf{N}^T$. Let $\mathbf{W} = \tilde{\mathbf{N}}\tilde{\mathbf{S}}^2\tilde{\mathbf{N}}^T$ be the EVD of \mathbf{W} (since \mathbf{W} is also SPSD). Moreover, let

us define:

$$\tilde{\mathbf{E}} \stackrel{\text{def}}{=} \mathbf{S}\mathbf{N}^T\tilde{\mathbf{N}}\tilde{\mathbf{S}}^{-2}\tilde{\mathbf{N}}^T\mathbf{N}\mathbf{S} - \mathbf{I}_{m \times m} \in \mathbb{R}^{m \times m}. \quad (30)$$

If we have $\mathbf{E} = \mathbf{0}_{m \times m}$, i.e., the approximate kernel map Φ_e is equal to the kernel-induced map, then $\tilde{\mathbf{N}} = \mathbf{N}$, $\tilde{\mathbf{S}} = \mathbf{S}$, and $\tilde{\mathbf{E}} = \mathbf{0}_{m \times m}$. Next, we find an upper bound for $\|\tilde{\mathbf{E}}\|_2$ in terms of the relative error η . Consider the EVD of $\mathbf{W} = \mathbf{W}_e + \mathbf{E}$:

$$\tilde{\mathbf{N}}\tilde{\mathbf{S}}^2\tilde{\mathbf{N}}^T = \mathbf{N}\mathbf{S}^2\mathbf{N}^T + \mathbf{E} = \mathbf{N}\mathbf{S}(\mathbf{I}_{m \times m} + \mathbf{O})\mathbf{S}\mathbf{N}^T, \quad (31)$$

where $\mathbf{O} \stackrel{\text{def}}{=} \mathbf{S}^{-1}\mathbf{N}^T\mathbf{E}\mathbf{N}\mathbf{S}^{-1} \in \mathbb{R}^{m \times m}$ is a symmetric matrix. Note that $\|\mathbf{O}\|_2 = \|\mathbf{N}\mathbf{O}\mathbf{N}^T\|_2 = \eta$, because of the unitary invariance of the spectral norm. If we multiply (31) on the left by \mathbf{N}^T and on the right by $\tilde{\mathbf{N}}$, we get:

$$\mathbf{N}^T\tilde{\mathbf{N}}\tilde{\mathbf{S}}^2 = \mathbf{S}(\mathbf{I}_{m \times m} + \mathbf{O})\mathbf{S}\mathbf{N}^T\tilde{\mathbf{N}}. \quad (32)$$

Next, we multiply (32) on the left by $\tilde{\mathbf{N}}^T\mathbf{N}$, and we see that:

$$\tilde{\mathbf{S}}^2 = \tilde{\mathbf{N}}^T\mathbf{N}\mathbf{S}^2\mathbf{N}^T\tilde{\mathbf{N}} + \tilde{\mathbf{N}}^T\mathbf{N}\mathbf{S}\mathbf{O}\mathbf{S}\mathbf{N}^T\tilde{\mathbf{N}}. \quad (33)$$

Finally, we multiply (33) on the left and right by $\tilde{\mathbf{S}}^{-1}$:

$$\mathbf{I}_{m \times m} = (\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{N}}^T\mathbf{N}\mathbf{S})(\mathbf{I}_{m \times m} + \mathbf{O})(\mathbf{S}\mathbf{N}^T\tilde{\mathbf{N}}\tilde{\mathbf{S}}^{-1}). \quad (34)$$

Thus, we observe that $\mathbf{S}\mathbf{N}^T\tilde{\mathbf{N}}\tilde{\mathbf{S}}^{-1} = (\mathbf{I}_{m \times m} + \mathbf{O})^{-1/2}\mathbf{T}$, where $\mathbf{T} \in \mathbb{R}^{m \times m}$ is an orthogonal matrix. Thus, we have:

$$\tilde{\mathbf{E}} = (\mathbf{I}_{m \times m} + \mathbf{O})^{-1} - \mathbf{I}_{m \times m}. \quad (35)$$

To find an upper bound for the spectral norm of $\tilde{\mathbf{E}}$, we simplify (35) by using the Neumann series $(\mathbf{I}_{m \times m} + \mathbf{O})^{-1} = \sum_{n=0}^{\infty} (-1)^n \mathbf{O}^n$, where $\|\mathbf{O}\|_2 = \eta < 1$ by assumption. Hence, we get the following upper bound for the spectral norm of $\tilde{\mathbf{E}} = \sum_{n=1}^{\infty} (-1)^n \mathbf{O}^n$:

$$\|\tilde{\mathbf{E}}\|_2 \leq \sum_{n=1}^{\infty} \|\mathbf{O}^n\|_2 \leq \sum_{n=1}^{\infty} \|\mathbf{O}\|_2^n \leq \sum_{n=1}^{\infty} \eta^n = \frac{\eta}{1 - \eta}, \quad (36)$$

where we have used the convergence of the Neumann series and the continuity of norms in the first inequality and the submultiplicativity property of the spectral norm in the second. To finish, we relate the difference between the two Nyström approximations $\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$ and $\mathbf{C}\mathbf{W}_e^\dagger\mathbf{C}^T$ to the norm of $\tilde{\mathbf{E}}$:

$$\mathbf{C}(\mathbf{W}^\dagger)^{1/2} = \mathbf{K}^{1/2}\mathbf{D}(\tilde{\mathbf{N}}\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{N}}^T) = \mathbf{K}^{1/2}\mathbf{F}\mathbf{S}\mathbf{N}^T\tilde{\mathbf{N}}\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{N}}^T. \quad (37)$$

Then, given the definition of $\tilde{\mathbf{E}}$ in (30), we observe that:

$$\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T - \mathbf{C}\mathbf{W}_e^\dagger\mathbf{C}^T = \mathbf{K}^{1/2}\tilde{\mathbf{E}}\mathbf{F}^T\mathbf{K}^{1/2}. \quad (38)$$

Thus, using the submultiplicativity property of the spectral norm, we have:

$$\|\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T - \mathbf{C}\mathbf{W}_e^\dagger\mathbf{C}^T\|_2 \leq \|\mathbf{K}^{1/2}\|_2^2 \|\mathbf{F}\|_2^2 \|\tilde{\mathbf{E}}\|_2. \quad (39)$$

Note that $\|\mathbf{K}^{1/2}\|_2^2 = \|\mathbf{K}\|_2$, and $\|\mathbf{F}\|_2 = 1$ since \mathbf{F} has orthonormal columns. \square

To gain some intuition for Theorem 8, we present a numerical experiment on the pendigits data set ($p = 16$ and $n = 10,992$) used in Section 7. Here, the Gaussian kernel function is employed with the parameter c chosen as the averaged squared distances between all the data points and sample mean. The standard K-means clustering algorithm is performed on the input data points to select the landmark points $\mathbf{z}_1, \dots, \mathbf{z}_m$ for various values of $m = 2, \dots, 10$. For each value of m , we form two matrices \mathbf{C} and \mathbf{W} . Also, we compute $\mathbf{W}_e = \mathbf{P}^T\mathbf{K}\mathbf{P}$ and $\eta = \|\mathbf{W}_e^{-1/2}\mathbf{E}\mathbf{W}_e^{-1/2}\|_2$, where $\mathbf{P} = \mathbf{K}^{-1}\mathbf{C}$ and $\mathbf{E} = \mathbf{W} - \mathbf{W}_e$; calculating \mathbf{W}_e is impractical for larger data sets and we do so only to support our theorem. Fig. 1 reports the mean of $\|\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T - \mathbf{C}\mathbf{W}_e^\dagger\mathbf{C}^T\|_2 / \|\mathbf{K}\|_2$ over 50 trials for varying number of landmark points. The figure also plots the mean of our theoretical bound in Theorem 8, i.e., $\eta/(1 - \eta)$. We observe that $\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$ and $\mathbf{C}\mathbf{W}_e^\dagger\mathbf{C}^T$ provide very similar Nyström approximations of \mathbf{K} , such

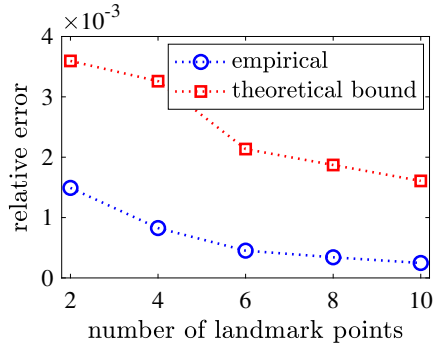


Fig. 1. Mean of the relative error and the theoretical error bound $\eta/(1-\eta)$.

that the relative error with respect to the spectral norm is less than 0.002. Furthermore, it is clear that Theorem 8 provides a meaningful upper bound for the relative error of the Nyström approximations.

Based on Theorem 8, the closeness of $\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$ and $\mathbf{C}\mathbf{W}_e^\dagger\mathbf{C}^T$ with respect to the spectral norm is a function of the quantity η . Note that η measures the relative amount of perturbation of the eigenvalues and eigenvectors of \mathbf{W}_e , and we have $\eta = \|\mathbf{W}_e^{-1/2}\mathbf{E}\mathbf{W}_e^{-1/2}\|_2 \leq \|\mathbf{W}_e^{-1}\mathbf{E}\|_2 \leq \|\mathbf{W}_e^{-1}\|_2\|\mathbf{E}\|_2$ [41, Lemma 2.2]. Therefore, when $\|\mathbf{E}\|_2$ is small, $\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$ and $\mathbf{C}\mathbf{W}_e^\dagger\mathbf{C}^T$ lead to similar low-rank approximations of the kernel matrix. In particular, as $\|\mathbf{E}\|_2$ goes to zero, $\mathbf{C}\mathbf{W}_e^\dagger\mathbf{C}^T$ converges to $\mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T$. Hence, we expect our theoretical results on the rank- r Nyström approximations to be valid for the case out-of-sample landmark points for small values of η .

6.2. Approach 2: clustered Nyström

We present an alternative to our previous approach that does not require full-rank kernel matrices. This technique is motivated by the clustered Nyström method of Zhang et al. [25,26], which shows that the Nyström error can be bounded by the quantization error of the input data using landmark points. Assume that the kernel function κ satisfies: $(\kappa(\mathbf{a}, \mathbf{b}) - \kappa(\mathbf{c}, \mathbf{d}))^2 \leq \eta(\|\mathbf{a} - \mathbf{c}\|_2^2 + \|\mathbf{b} - \mathbf{d}\|_2^2)$, where η is a constant depending on κ . It is shown that for a number of widely used kernel functions, such as Gaussian and polynomial kernels, this property is satisfied. Then, the Nyström approximation error is upper bounded: $\|\mathbf{K} - \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T\|_F \leq \eta_1\sqrt{\phi} + \eta_2\phi$, where $\phi \stackrel{\text{def}}{=} 1/n \cdot \sum_{i=1}^n \|\mathbf{x}_i - \mu(\mathbf{x}_i)\|_2^2$, $\mu(\mathbf{x}_i) \in \mathbb{R}^p$ is the closest landmark point to each data point \mathbf{x}_i , and η_1, η_2 are two constants.

Given out-of-sample landmark points, this approach approximates the landmark points with in-sample points, and pays just a factor of 2 on the quantization error. To prove this, without loss of generality and to simplify notation, we consider a single cluster since we can work cluster-by-cluster. If the cluster assignments change after selecting the new landmark points, that can only reduce the quantization error further. Thus, let \mathbf{s} be the (possibly out-of-sample) center of mass for the points $\mathbf{x}_i, i = 1, \dots, n$, i.e., $\mathbf{s} = 1/n \cdot \sum_{i=1}^n \mathbf{x}_i$. As observed in [42, Lemma 2.1], for any vector $\mathbf{t} \in \mathbb{R}^p$:

$$\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{t}\|_2^2 = \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{s}\|_2^2 + \|\mathbf{t} - \mathbf{s}\|_2^2. \quad (40)$$

We will choose a new in-sample cluster center: $\mathbf{t} = \mathbf{x}_j$ for $j \in \arg \min_{i \in \{1, \dots, n\}} \|\mathbf{x}_i - \mathbf{s}\|_2^2$. Let $\hat{\phi}$ denote the quantization error with this new in-sample center, and ϕ be the original quantization error

using \mathbf{s} as the cluster center. Then

$$\begin{aligned} \hat{\phi} &\stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{t}\|_2^2 = \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{s}\|_2^2 + \|\mathbf{t} - \mathbf{s}\|_2^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{s}\|_2^2 + \sum_{i=1}^n \frac{1}{n} \|\mathbf{x}_i - \mathbf{s}\|_2^2 = 2\phi, \end{aligned} \quad (41)$$

where the first line follows from (40). Note that the computational cost is small, since choosing an in-sample cluster center is $\mathcal{O}(np)$.

7. Experimental results

We present experimental results on the fixed-rank approximation of kernel matrices using the standard and modified Nyström methods, and show evidence to (1) corroborate our theory of Section 5, (2) suggest that our theory of Section 6 still holds even if the assumptions are not exactly satisfied (e.g., out-of-sample landmark points), and (3) highlight the benefits of the modified Nyström method. In order to illustrate the effectiveness of modified Nyström, we compare its accuracy to that of the standard Nyström method for the target rank $r = 2$ and varying number of landmark points $m = r, \dots, 5r$. To provide a baseline for the comparison, we report the accuracy of the best rank-2 approximation obtained via the eigenvalue decomposition (EVD), which requires the computation and storage of full kernel matrices and hence is impractical for very large data sets.

7.1. Fixed-rank approximation error

Experiments are conducted on four data sets from the LIB-SVM archive [43]: (1) *pendigits* ($p = 16$ and $n = 10,992$); (2) *satimage* ($p = 36$ and $n = 6435$); (3) *w6a* ($p = 300$ and $n = 13,267$); and (4) *E2006-tfidf* ($p = 150,360$ and $n = 3000$). In all experiments, the Gaussian kernel $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/c)$ is used with the parameter c chosen as the averaged squared distances between all the data points and sample mean. We consider two landmark selection techniques: (1) uniform sampling, where m landmark points are selected uniformly at random without replacement from n data points; and (2) out-of-sample landmark points obtained via K-means clustering on the original data set, as in [26]. To perform the K-means clustering algorithm, we use MATLAB's built-in function `kmeans` and the maximum number of iterations is set to 10. A MATLAB implementation of modified and standard Nyström is available at <https://github.com/pourkamali/RandomizedClusteredNyström>.

We measure the quality of fixed-rank approximations using the relative error with respect to the trace norm, i.e., $\|\mathbf{K} - \mathbf{G}_{(r)}^{\text{nys}}\|_* / \|\mathbf{K}\|_*$ vs. $\|\mathbf{K} - \mathbf{G}_{(r)}^{\text{opt}}\|_* / \|\mathbf{K}\|_*$, and thus in this metric, no method can ever outperform the EVD baseline. In Fig. 2, the mean and standard deviation of relative error over 50 trials are reported for all four data sets and varying number of landmark points $m = r, \dots, 5r$. Modified Nyström and the standard Nyström method have identical performance for $m = r$ because no rank restriction step is involved in this case. As the number of landmark points increases beyond the rank parameter $r = 2$, Nyström via QR decomposition always generates better rank-2 approximations of the kernel matrix than the standard method does. This observation is consistent with Theorem 3, which states that for the same landmark points, $\|\mathbf{K} - \mathbf{G}_{(r)}^{\text{opt}}\|_* \leq \|\mathbf{K} - \mathbf{G}_{(r)}^{\text{nys}}\|_*$ (one can divide both sides by constant $\|\mathbf{K}\|_*$ for the relative error). Even with out-of-sample landmark points, the modified method can be drastically better. In the right plot of Fig. 2(b), the modified method achieves a mean error of 0.47 (compared to the 0.45 EVD baseline) from just $m = 2r$ landmark points, while the standard method has a higher mean error

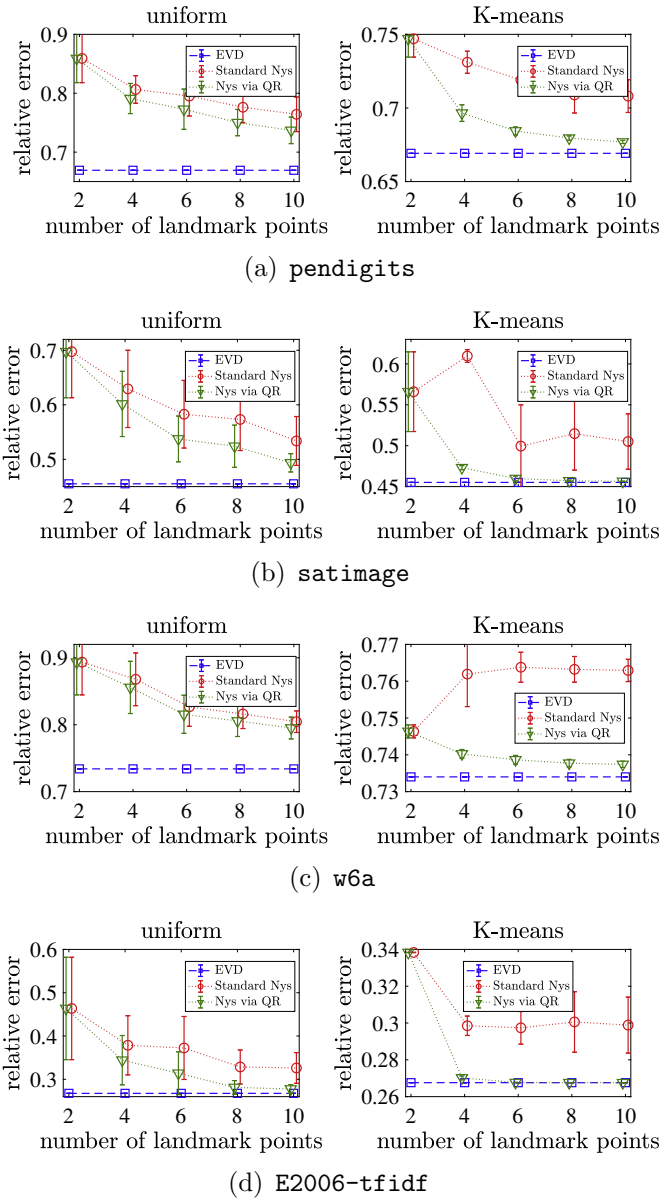


Fig. 2. Mean and standard deviation of the relative error with respect to the trace norm. For each case, the left plot uses in-sample landmark points while the right plot uses out-of-sample points.

(0.50) even using $m = 5r$ landmark points. This exemplifies the importance and effectiveness of the precise rank restriction step in the Nyström via QR decomposition method. Fig. 3 is generated the same as Fig. 2 but the error is reported in the Frobenius norm. The pattern of behavior is very similar to that in Fig. 2 even though we lack theoretical guarantees. In fact, neither method dominates the other for all kernel matrices (cf. the adversarial example in Remark 5), but in these practical data sets, the modified method always performs better.

Remark 9. Remark 7 showed that in both the trace and Frobenius norms, the standard Nyström method can perform worse when we sample additional landmark points. Figs. 2 and 3 show that a similar effect happens with the standard Nyström method when we use out-of-sample landmark points selected via K-means (in this case, as we increase m , we do not necessarily include the landmark points selected for smaller m). For example, according to Fig. 2(b)

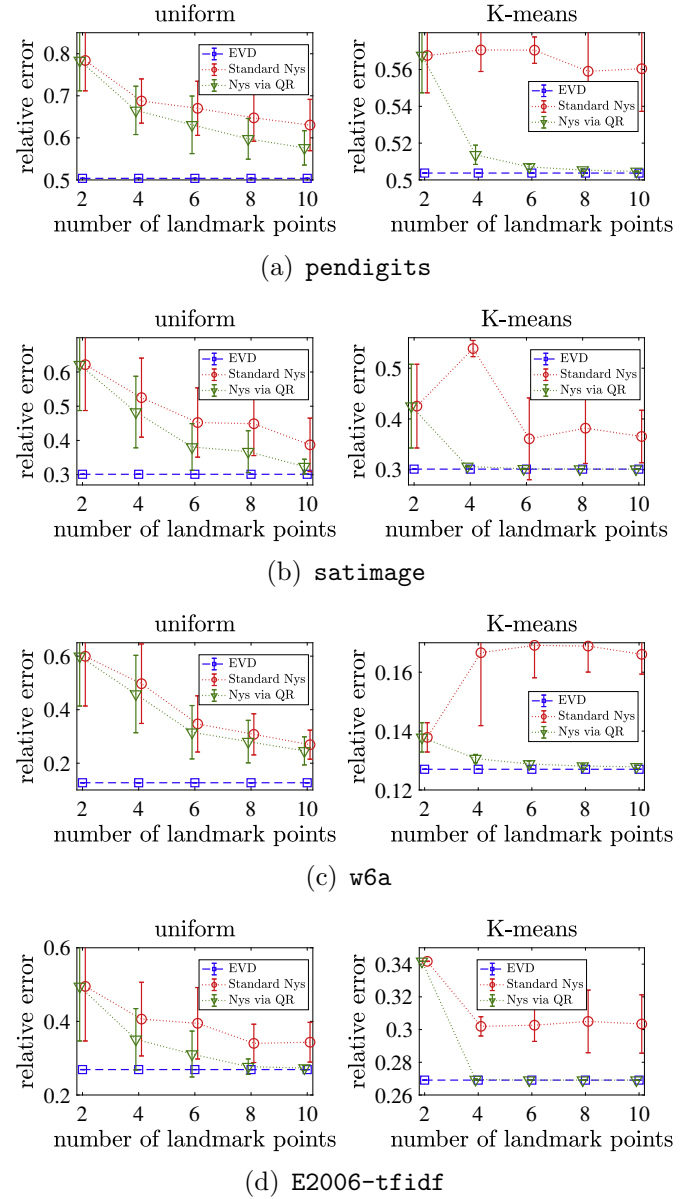


Fig. 3. Mean and standard deviation of the relative error with respect to the Frobenius norm.

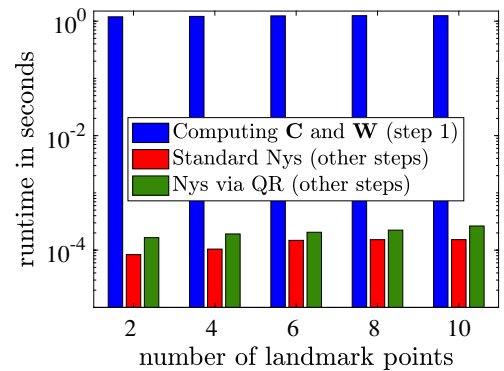


Fig. 4. Running time results for standard Nyström and the modified technique.

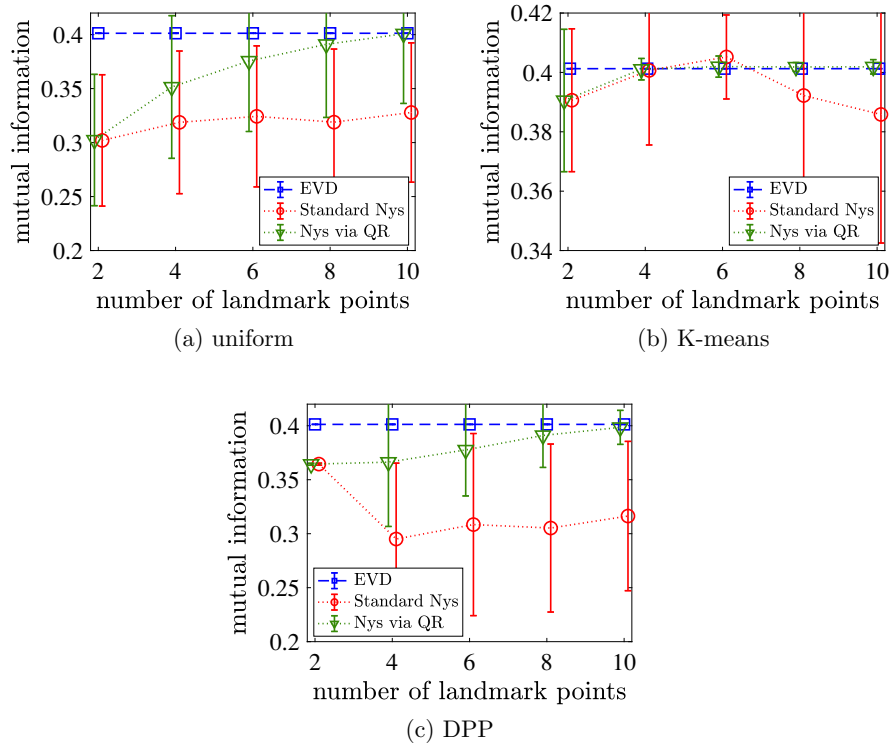


Fig. 5. Normalized mutual information using three landmark selection techniques.

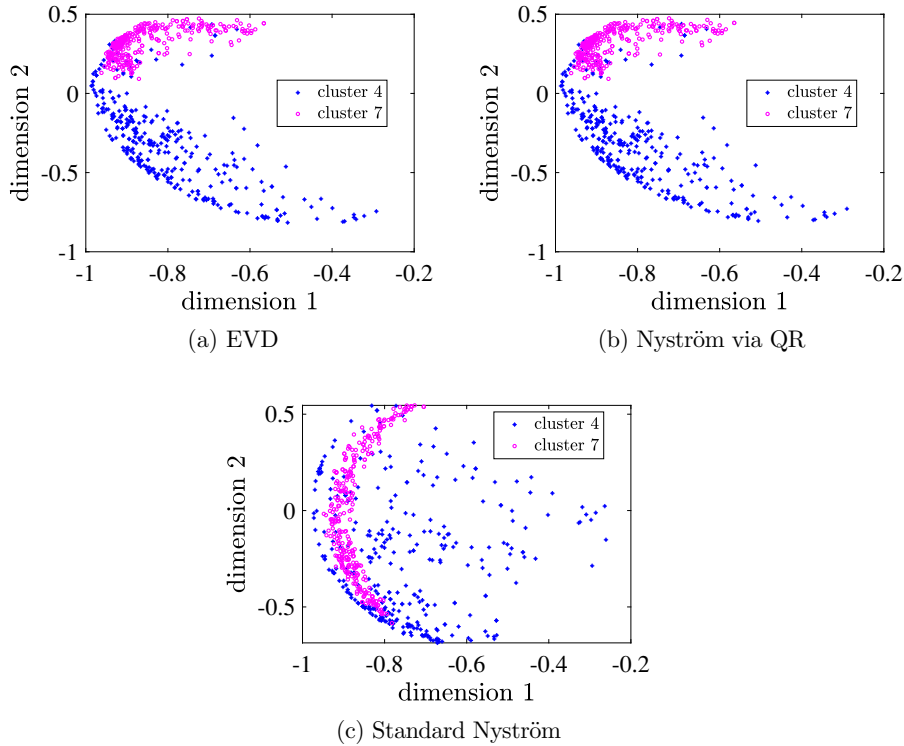


Fig. 6. Visualization of the columns of \mathbf{L}^T that correspond to clusters 4 and 7.

(right), the mean relative error of standard Nyström is increased from 0.56 to 0.61 when we increase from $m = 2$ to $m = 4$ landmark points selected via K-means centroids.

This counter-intuitive effect of decreased accuracy even with more landmark points (Remarks 7 and 9) is due to the sub-optimal

restriction procedure of standard Nyström. Theorem 6 proves that the modified Nyström method does not suffer from the same effect in terms of the trace norm and in-sample landmark points, and Figs. 2 and 3 do not show any evidence of this effect even if we switch to the Frobenius norm or consider out-of-sample landmark points.

We also demonstrate the efficiency of Nyström via QR decomposition by plotting the averaged running time on a logarithmic scale for E2006-tfidf. The running time results are omitted for the remaining data sets because the average running time was less than one second. Fig. 4 shows that the dominant computational cost is related to constructing \mathbf{C} and \mathbf{W} in step 1 of both standard and modified Nyström methods. Moreover, given these two matrices, the cost of finding the best rank- r approximation in our proposed method is only slightly higher than the standard Nyström method, as explained in Section 4. Thus, the overall computational cost of modified method is almost identical to the standard Nyström method and our method achieves superior performance as demonstrated in Figs. 2(d) and 3(d).

7.2. Application to kernel clustering

In the last experiment, we demonstrate the performance of modified Nyström on a kernel K-means clustering task. We use the segment data set ($p = 19$ and $n = 2,310$) from the LIBSVM archive that consists of 7 clusters. We choose the homogeneous polynomial kernel function of order 2, i.e., $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$. Previously, we considered Gaussian kernel functions and we would like to show the performance of our method for polynomial kernels as well [44]. In addition to the two landmark selection techniques used in Section 7.1, we consider another recent technique based on Determinantal Point Processes (DPP) that was introduced in [45]. Fig. 5 reports the mean and standard deviation of normalized mutual information (NMI) [46] over 200 trials for fixed rank $r = 2$ and various number of landmark points, where the standard K-means algorithm is performed on the columns of $\mathbf{L}^T \in \mathbb{R}^{r \times n}$ after computing the fixed-rank approximation $\mathbf{K} \approx \mathbf{L}\mathbf{L}^T$, cf. (2). NMI is a popular clustering quality metric which ranges from 0 to 1, and larger values of NMI indicate the higher quality of clustering.

As we see in Fig. 5, the modified method outperforms standard Nyström when $m > r$, which is necessary for obtaining accuracies that are close to the baseline EVD. Furthermore, it is observed that modified Nyström leads to higher accuracy results as the number of distinct landmark points increases. In particular, when landmark points are K-means centroids in Fig. 5(b), the mean NMI of our method reaches EVD with very small standard deviation. However, for all three landmark selection techniques, the standard Nyström method does not necessarily provide better results when more landmark points are used.

To further illustrate the importance of modified method, we plot the columns of \mathbf{L}^T (known as virtual samples [15]) that correspond to two of the 7 clusters in Fig. 6. In this case, $m = 10$ landmark points are sampled uniformly at random from the input data. We see that the EVD and our modified method have almost identical results. However, the standard Nyström method spreads out the virtual samples, which justifies its lower accuracy compared to the modified Nyström method and the best rank-2 approximation obtained via EVD.

8. Conclusion

In this paper, we have presented a modified technique for the important process of rank reduction in the Nyström method. Theoretical analysis shows that: (1) the modified method provides improved fixed-rank approximations compared to standard Nyström with respect to the trace norm; and (2) the quality of fixed-rank approximations generated via the modified method improves as the number of distinct landmark points increases. Our theoretical results are accompanied by illustrative numerical experiments comparing the modified method with standard Nyström. We also

showed that the modified method has almost the same computational complexity as standard Nyström, which makes it suitable for large-scale kernel machines.

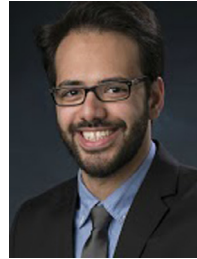
Declaration of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] C. Cortes, V. Vapnik, Support-vector networks, *Mach. Learn.* 20 (3) (1995) 273–297.
- [2] J. Suykens, J. Vandewalle, Least squares support vector machine classifiers, *Neural Process. Lett.* 9 (3) (1999) 293–300.
- [3] M. Girolami, Mercer kernel-based clustering in feature space, *IEEE Trans. Neural Netw.* 13 (3) (2002) 780–784.
- [4] R. Chitta, R. Jin, T. Havens, A. Jain, Approximate kernel k-means: solution to large scale kernel clustering, in: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 895–903.
- [5] F. Pourkamali-Anaraki, S. Becker, A randomized approach to efficient kernel clustering, in: *Proceedings of the IEEE Global Conference on Signal and Information Processing*, 2016, pp. 207–211.
- [6] R. Langone, J. Suykens, Fast kernel spectral clustering, *Neurocomputing* 268 (2017) 27–33.
- [7] C. Saunders, A. Gammerman, V. Vovk, Ridge regression learning algorithm in dual variables, in: *Proceedings of the International Conference on Machine Learning*, 1998, pp. 515–521.
- [8] C. Hsieh, S. Si, I. Dhillon, Fast prediction for large-scale kernel machines, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2014, pp. 3689–3697.
- [9] A. Alaoui, M. Mahoney, Fast randomized kernel ridge regression with statistical guarantees, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2015, pp. 775–783.
- [10] Y. Yang, M. Pilianci, M. Wainwright, Randomized sketches for kernels: fast and optimal nonparametric regression, *Ann. Stat.* 45 (3) (2017) 991–1023.
- [11] B. Schölkopf, A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT press, 2001.
- [12] F. Bach, Sharp analysis of low-rank kernel matrix approximations, in: *Proceedings of the Conference on Learning Theory*, 2013, pp. 185–209.
- [13] C. Cortes, M. Mohri, A. Talwalkar, On the impact of kernel approximation on learning accuracy, in: *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2010, pp. 113–120.
- [14] K. Zhang, L. Lan, Z. Wang, F. Moerchen, Scaling up kernel SVM on limited resources: a low-rank linearization approach, in: *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2012, pp. 1425–1434.
- [15] A. Golts, M. Elad, Linearized kernel dictionary learning, *IEEE J. Sel. Top. Signal Process.* 10 (4) (2016) 726–739.
- [16] S. Wang, Z. Zhang, T. Zhang, Towards more efficient SPSP matrix approximation and CUR matrix decomposition, *J. Mach. Learn. Res.* 17 (210) (2016) 1–49.
- [17] N. Halko, P. Martinsson, J. Tropp, Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions, *SIAM Rev.* 53 (2) (2011) 217–288.
- [18] F. Pourkamali-Anaraki, S. Becker, Preconditioned data sparsification for big data with applications to PCA and K-means, *IEEE Trans. Inf. Theory* 63 (5) (2017) 2954–2974.
- [19] J. Tropp, A. Yurtsever, M. Udell, V. Cevher, Practical sketching algorithms for low-rank matrix approximation, *SIAM J. Matrix Anal. Appl.* 38 (4) (2017) 1454–1485.
- [20] C. Williams, M. Seeger, Using the Nyström method to speed up kernel machines, in: *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, 2001, pp. 682–688.
- [21] S. Kumar, M. Mohri, A. Talwalkar, Sampling methods for the Nyström method, *J. Mach. Learn. Res.* 13 (2012) 981–1006.
- [22] M. Li, W. Bi, J. Kwok, B. Lu, Large-scale Nyström kernel matrix approximation using randomized SVD, *IEEE Trans. Neural Netw. Learn. Syst.* 26 (1) (2015) 152–164.
- [23] A. Gittens, M. Mahoney, Revisiting the Nyström method for improved large-scale machine learning, *J. Mach. Learn. Res.* 17 (1) (2016) 3977–4041.
- [24] C. Musco, C. Musco, Recursive sampling for the Nyström method, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2017, pp. 3836–3848.
- [25] K. Zhang, I. Tsang, J. Kwok, Improved Nyström low-rank approximation and error analysis, in: *Proceedings of the International Conference on Machine Learning*, 2008, pp. 1232–1239.
- [26] K. Zhang, J. Kwok, Clustered Nyström method for large scale manifold learning and dimension reduction, *IEEE Trans. Neural Netw.* 21 (10) (2010) 1576–1587.
- [27] D. Oglic, T. Gärtner, Nyström method with kernel k-means++ samples as landmarks, in: *Proceedings of the International Conference on Machine Learning*, 2017, pp. 2652–2660.

- [28] F. Pourkamali-Anaraki, S. Becker, M. Wakin, Randomized clustered Nyström for large-scale kernel machines, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 3960–3967.
- [29] F. Pourkamali-Anaraki, S. Becker, Randomized clustered Nyström for large-scale kernel machines, *arXiv:1612.06470v1* (2016).
- [30] S. Wang, A. Gittens, M. Mahoney, Scalable kernel k-means clustering with Nyström approximation: Relative-error bounds, *J. Mach. Learn. Res.* 20 (2019) 1–49.
- [31] J. Tropp, A. Yurtsever, M. Udell, V. Cevher, Fixed-rank approximation of a positive-semidefinite matrix from streaming data, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2017, pp. 1225–1234.
- [32] N. Aronszajn, Theory of reproducing kernels, *Trans. Am. Math. Soc.* 68 (1950) 337–404.
- [33] C. Boutsidis, P. Drineas, M. Magdon-Ismail, Near-optimal column-based matrix reconstruction, *SIAM J. Comput.* 43 (2) (2014) 687–717.
- [34] P. Drineas, M. Mahoney, On the Nyström method for approximating a gram matrix for improved kernel-based learning, *J. Mach. Learn. Res.* (2005) 2153–2175.
- [35] M. Mahoney, P. Drineas, CUR matrix decompositions for improved data analysis, *Proc. Natl. Acad. Sci.* 106 (3) (2009) 697–702.
- [36] P. Drineas, M. Magdon-Ismail, M. Mahoney, D. Woodruff, Fast approximation of matrix coherence and statistical leverage, *J. Mach. Learn. Res.* 13 (2012) 3475–3506.
- [37] S. Sun, J. Zhao, J. Zhu, A review of Nyström methods for large-scale machine learning, *Inf. Fusion* 26 (2015) 36–48.
- [38] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- [39] J. Lu, S. Hoi, J. Wang, P. Zhao, Z. Liu, Large scale online kernel learning, *J. Mach. Learn. Res.* 17 (47) (2016) 1–43.
- [40] R. Mathias, K. Veselić, A relative perturbation bound for positive definite matrices, *Linear Algebr. Appl.* 270 (1–3) (1998) 315–321.
- [41] F. Dopico, J. Moro, J. Molera, Weyl-type relative perturbation bounds for eigen-systems of hermitian matrices, *Linear Algebr. Appl.* 309 (1–3) (2000) 3–18.
- [42] D. Arthur, S. Vassilvitskii, k-means++: The advantages of careful seeding, in: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007, pp. 1027–1035.
- [43] C. Chang, C. Lin, LIBSVM: a library for support vector machines, *ACM Trans. Intell. Syst. Technol.* 2 (3) (2011) 27:1–27:27.
- [44] R. Hamid, Y. Xiao, A. Gittens, D. DeCoste, Compact random feature maps, in: *Proceedings of the International Conference on Machine Learning*, 2014, pp. 19–27.
- [45] C. Li, S. Jegelka, S. Sra, Fast DPP sampling for Nyström with application to kernel methods, in: *Proceedings of the International Conference on Machine Learning*, 2016, pp. 2061–2070.
- [46] H. Oh Song, S. Jegelka, V. Rathod, K. Murphy, Deep metric learning via facility location, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5382–5390.



Farhad Pourkamali-Anaraki is an assistant professor in the department of Computer Science at the University of Massachusetts Lowell. He received his PhD in Electrical Engineering from the University of Colorado Boulder in 2017. His research is at the intersection of statistics, optimization, and machine learning on the one hand and the intersection of theory, applications, and implementations on the other side. In particular, he works on extending the use of machine learning algorithms to other domains, including reliability analysis of critical infrastructures in Civil Engineering and advanced manufacturing in Mechanical Engineering.



Stephen Becker is an assistant professor in the department of Applied Mathematics at the University of Colorado at Boulder. He was the recipient of a postdoctoral fellowship to study at Paris 6, and the Goldstine postdoctoral fellowship at IBM Research. He completed his PhD in 2011 in Applied and Computational Mathematics from the California Institute of Technology, and his bachelors in Mathematics and Physics from Wesleyan University. His work on the TFOCS algorithm won the triennial Beale-Orchard-Hays Prize from the Mathematical Optimization Society. His current research is centered around deriving efficient optimization methods for large-scale data analysis, with emphasis on techniques for sparse and low-rank models, as well as numerical linear algebra for statistical techniques such as randomized tensor compression.