A Perturbation-Based Kernel Approximation Framework

Roy Mitz   roymitz@mail.tau.ac.il
Yoel Shkolnisky   yoelsh@tauex.tau.ac.il
School of Mathematical Sciences
Tel-Aviv University
Tel-Aviv, Israel

Abstract

Kernel methods are powerful tools in various data analysis tasks. Yet, in many cases, their time and space complexity render them impractical for large datasets. Various kernel approximation methods were proposed to overcome this issue, with the most prominent method being the Nyström method. In this paper, we derive a perturbation-based kernel approximation framework building upon results from classical perturbation theory. We provide an error analysis for this framework, and prove that in fact, it generalizes the Nyström method and several of its variants. Furthermore, we show that our framework gives rise to new kernel approximation schemes, that can be tuned to take advantage of the structure of the approximated kernel matrix. We support our theoretical results numerically and demonstrate the advantages of our approximation framework on both synthetic and real-world data.

Keywords: perturbation theory, kernel approximation, kernel-based non-linear dimensionality reduction, Nyström method

1. Introduction

In the last decades, kernel methods became widely used in various data analysis tasks. Two notable examples that are widely used for machine learning are kernel SVM (Cortes and Vapnik, 1995) and kernel ridge regression (Saunders et al., 1998). Another common application of kernel methods is non-linear dimensionality reduction, where the lower-dimensional embedding of the data is derived from the eigenvectors of some data-dependent kernel matrix. Examples of such dimensionality reduction algorithms include Laplacian eigenmaps (Belkin and Niyogi, 2003), LLE (Roweis and Saul, 2000), Isomap (Balasubramanian et al., 2002), MDS (Buja et al., 2008), Spectral clustering (Shi and Malik, 2000; Ng et al., 2002), and more. In all of the aforementioned kernel-based methods, the dimension of the kernel matrix grows linearly with the number of data points n. This makes kernel-related algorithms impractical for large n, both in terms of memory and runtime. To overcome this issue, several methods for kernel approximation were proposed.

The most common form of kernel approximation is low-rank kernel approximation, where we seek to find a low-rank representation of the kernel matrix. It is known that the best rank-m approximation of a matrix in the spectral and Frobenius norms is given by truncating its SVD decomposition. However, since the dimension of the kernel matrix grows with the number of data points, the computation of the full or even the partial eigendecomposition

of a large kernel matrix is impractical due to its runtime and space requirements. For example, algorithms for partial eigendecomposition such as the Lanczos algorithm and some variants of SVD require $O(n^2 m)$ floating point operations, where n is the dimension of the matrix (number of data points) and m is the number of eigenvectors calculated. Randomized algorithms (Halko et al., 2011a,b) use random projections of the data to reduce the time complexity of the decomposition to $O(n^2 \log m)$, which is still impractical for large n. Moreover, all eigendecomposition algorithms require to store the $n \times n$ kernel matrix either in RAM or on disk.

Since the exact calculation of the low-rank decomposition of a kernel matrix is unfeasible for large datasets, several approximated methods were proposed. The most prominent low-rank kernel approximation method is the Nyström method (Williams and Seeger, 2001), that will be described in detail in the next section. Some variants of the Nyström method were proposed in literature: the Randomized SVD Nyström method (Li et al., 2014), which uses efficient randomized eigensolvers that enable it to use more samples in order to improve performance; the ensemble Nyström method (Kumar et al., 2009), which averages several Nyström approximations in order to improve accuracy (we note that this is in fact not a low-rank approximation); the spectral shifted Nyström method (Wang et al., 2014), which provides superior performance in cases where the spectrum of the matrix decays slowly; and the modified Nyström method (Wang and Zhang, 2013). The latter usually outperforms the classical Nyström method, but is impractical for large datasets due to its time and memory requirements. To alleviate these requirements, Wang and Zhang (2014) suggested a faster version of the modified Nyström method.

When the approximated kernel is not low-rank, the kernel approximation methods mentioned above may result in a poor approximation. More recently, works that use the structure of the kernel matrix were proposed. For example, the MEKA algorithm (Si et al., 2017) provides superior kernel approximation for kernels that admit a block-diagonal structure, and are not necessarily low-rank.

A problem related to low-rank approximation that is relevant to this paper is updating a known eigendecomposition of a matrix following a "small" perturbation, without calculating the entire decomposition from scratch. Classical perturbation results (Stewart, 1990; Byron and Fuller, 2012) exist for a general symmetric perturbation, and will be described in detail in the next section. Other related works consider perturbations that have some structure; see for example, Bunch et al. (1978); Mitz et al. (2019) for the case where the perturbation is of rank one, and Oh and Hu (2018); Brand (2006) for a general low-rank perturbation. Other approaches of updating a given eigendecomposition include restarting the power method (Langville and Meyer, 2006) or the inverse iteration algorithm (Trefethen and Bau III, 1997), both require applying the updated matrix several times until convergence, which may be expensive if the matrix is large.

The contribution of the current paper is threefold. First, we derive eigendecomposition perturbation formulas accompanied by error bounds for matrices that only part of their spectrum is known. Second, we use these perturbation formulas to derive a new framework for kernel approximation. Unlike some of the existing methods, we present explicit error bounds for the individual approximated eigenvectors rather than only to the kernel approximation. Third, we prove that the Nyström method and its variants are in fact special cases of our framework. This reveals the essence behind existing Nyström methods, allows

to analyze their accuracy, and provides means to derive new Nyström-type approximations that take advantage of the structure of the kernel matrix.

The rest of this paper is organized as follows. In Section 2, we describe classical perturbation results, along with the Nyström method and some of its variants. In Section 3, we extend the classical perturbation formulas to the case where only part of the spectrum of the perturbed matrix is known, and derive their error bounds. In Section 4, we use these formulas to develop a perturbation-based kernel approximation framework. In Section 5, we prove that our kernel approximation framework generalizes the Nyström method. In Section 6, we suggest several types of kernesl approximations based on our framework, and prove that some of them are related to variants of the Nyström method. In Section 7, we provide numerical results to support our theory and show the advantages of our kernel approximation framework. In Section 8, we provide some concluding remarks and discuss future research.

2. Preliminaries

In this section, we describe two methods for approximating the eigendecomposition of a matrix that are relevant to our work. We first describe the Nyström method in Section 2.1, and then describe the perturbation method in Section 2.2.

2.1 The Nyström method and its variants

In this section, we describe in detail some of the Nyström-type methods discussed in the Introduction. We begin with describing the classical Nyström method, continue with some of its variants that are particularly relevant to our work, and finish by discussing some results regarding the error bounds of the Nyström method.

2.1.1 THE CLASSICAL NYSTRÖM METHOD

Let $K \in \mathbb{R}^{n \times n}$ be a symmetric positive-definite matrix. We wish to find the k leading eigenpairs $\{(\lambda_i, u_i)\}_{i=1}^k$ of K. The Nyström method (Sun et al., 2015; Williams and Seeger, 2001) finds an approximation $\{(\widetilde{\lambda}_i, \widetilde{u}_i)\}_{i=1}^k$ to these eigenpairs as follows. First, we select randomly k columns of K (typically uniformly without replacement). We assume, without loss of generality, that the columns and rows of K are rearranged so that the first k columns of K are sampled. Denote by K' the $k \times k$ upper-left submatrix of K and by C the $n \times k$ matrix consisting of the first k columns of K. Then, we calculate the k eigenpairs of K' and denote them by $\{(\lambda_i', u_i')\}_{i=1}^k$. Finally, the Nyström method approximates the k leading eigenvectors of K by

$$\widetilde{u}_i = \sqrt{\frac{k}{n}} \frac{1}{\lambda_i'} C u_i', \quad i = 1, \ldots, k. \tag{1}$$

Moreover, the k leading eigenvalues of K are approximated by

$$\widetilde{\lambda}_i = \frac{n}{k} \lambda_i', \quad i = 1, \ldots, k. \tag{2}$$

3

Finally, the Nyström approximation of K is

$$\widetilde{K}_{\text{nys}} = \sum_{i=1}^{k} \widetilde{\lambda}_i \widetilde{u}_i^T \widetilde{u}_i. \tag{3}$$

The runtime complexity of the Nyström method (Formula (1)) is $O(nk^2 + k^3)$.

The Nyström method requires sampling a representative subset of the data and hence many methods for sampling this subset were proposed, see Kumar et al. (2012); Sun et al. (2015) and the references therein. Our results derived below are independent of the methodology used to obtain the subset of samples, and hence we do not discuss this issue in detail.

2.1.2 Variants of the Nyström method

One generalization of the Nyström method is the Randomized SVD Nyström method (Li et al., 2014). It introduces a parameter $l \geq k$ and chooses K' to be the $l \times l$ top-left submatrix of K. Then, the k leading eigenpairs of K' are calculated via some efficient randomized SVD method, followed by the use of (1) and (2) for the approximation. This form of approximation is a generalization of the Nyström method, since choosing $l = k$ is equivalent to the Nyström method. On the other hand, if we choose $l = n$, we get the exact eigenvectors of K. Intuitively, the larger l is, the better the approximation is likely to be, at the cost of a greater computational complexity. The runtime complexity of this method is $O(nk^2 + lk^2)$.

Since the Nyström approximation (3) of the kernel matrix K is a low-rank approximation, it may provide poor results when K is not low-rank. This might occur, for example, when the spectrum of K decays slowly. A possible approach to overcome this problem is the spectrum shifted Nyström method (Wang et al., 2014). This method essentially applies the classical Nyström method on a shifted kernel matrix, that is, applies the Nyström method on

$$K_{\text{shift}} = K - \mu I, \tag{4}$$

for some $\mu \geq 0$. The updated eigenvalues (2) are then shifted-back by μ. If we denote by $\widetilde{K}_{\text{shift}}$ the kernel approximation obtained by using the shifted Nyström method, it is shown in Wang et al. (2014) that

$$\left\| K - \widetilde{K}_{\text{shift}} \right\|_F \leq \left\| K - \widetilde{K}_{\text{nys}} \right\|_F.$$

Alternatively, the kernel K may admit a block-diagonal structure. As demonstrated in Si et al. (2017), this may happen for some kernel functions when the data consist of several clusters. In this case, the MEKA algorithm (Si et al., 2017) essentially performs a Nyström approximation on each cluster of the data. Each such approximation corresponds to a block on the diagonal of the kernel matrix, and the resulting approximation is block-diagonal.

An approach related to the MEKA algorithm for improving the Nyström approximation (3) is the ensemble Nyström method (Kumar et al., 2009). The idea behind this method is to perform q independent Nyström kernel approximations on random subsets of the data, and then average them. Formally, given q independent Nyström approximations $\{\widetilde{K}_i\}_{i=1}^q$,

the ensemble Nyström approximation is given by

$$\widetilde{K}_{\text{ens}} = \sum_{i=1}^{q} \mu_i \widetilde{K}_i,$$

for some weights $\{\mu_i\}_{i=1}^{q}$. It is suggested in Kumar et al. (2009) to use $\mu_i = \frac{1}{q}$ for $1 \le i \le q$. Better error bounds for this method compared to the classical Nyström method are proven in Kumar et al. (2009).

The difference between the ensemble Nyström method and the MEKA algorithm is that in the former, the individual Nyström approximations are chosen at random rather than by clusters, and the resulting approximation is their average rather than their concatenation in a block-diagonal matrix. We note that in both the ensemble Nyström method and the MEKA algorithm, the resulting approximation is not low-rank, and is generally of a rank greater than k.

2.1.3 Error bounds for the Nyström method

The error bounds of the Nyström method and its variants have been of great interest. For a comprehensive review we refer the readers to Sun et al. (2015), Wang and Zhang (2013) and Gittens and Mahoney (2013). If we denote by K_m the best rank-m approximation of a kernel matrix K, then all error bounds obtained in the literature are of the forms

$$\left\| K - \widetilde{K}_{\text{nys}} \right\| \le \alpha \| K - K_m \| \quad \text{or} \quad \left\| K - \widetilde{K}_{\text{nys}} \right\| \le \| K - K_m \| + \beta,$$

for some $\alpha, \beta \ge 0$, where $\|\cdot\|$ is usually the Frobenius norm or the 2-norm. To the best of our knowledge, no error bounds were obtained for the individual Nyström-approximated eigenvectors, though such bounds may be of interest when using the Nyström method as part of a dimensionality reduction algorithm.

2.2 Perturbation of eigenvalues and eigenvectors

Let $A' \in \mathbb{R}^{n \times n}$ be a real symmetric positive-definite matrix with distinct eigenvalues $\{t_i\}_{i=1}^{n}$ and their corresponding orthonormal eigenvectors $\{v_i\}_{i=1}^{n}$. Assume that $t_1 > t_2 > \cdots > t_n$. Let $E \in \mathbb{R}^{n \times n}$ a real symmetric matrix. Consider a perturbation A of A' given by $A = A' + E$, with the eigenpairs of A denoted by $\{(s_i, w_i)\}_{i=1}^{n}$, so that $s_1 > s_2 > \cdots > s_m$. We wish to find an approximation $\{(\widetilde{s}_i, \widetilde{w}_i)\}_{i=1}^{n}$ to the eigenpairs of A. The classical perturbation solution to this problem (Stewart, 1990; Byron and Fuller, 2012) is as follows. The approximated eigenvectors of A are given by

$$\widetilde{w}_i = v_i + \sum_{k=1, k \ne i}^{n} \frac{\langle Ev_i, v_k \rangle}{t_i - t_k} v_k + O(\|E\|_2^2), \quad 1 \le i \le n, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ is the standard dot product between vectors, and the approximated eigenvalues of A are given by

$$\widetilde{s}_i = t_i + v_i^T E v_i + O(\|E\|_2^2), \quad 1 \le i \le n. \tag{6}$$

We note that the eigenvalues update formula (6) depends only on the eigenvalue we wish to update and its corresponding eigenvector, whereas the eigenvectors update formula (5) depends on all eigenvalues and eigenvectors of A'.

There exist perturbation results for matrices with non-simple eigenvalues (Byron and Fuller, 2012). However, since non-simple eigenvalues are highly unlikely in data-dependent matrices, we do not discuss this case and leave it for a future work.

3. Truncating the perturbation formulas

In this section, we consider a variant of the problem presented in Section 2.2, in which only the m leading eigenpairs $\{(t_i, v_i)\}_{i=1}^m$ of the unperturbed matrix A' are known, and we wish to approximate the m leading eigenpairs of A. To this end, we introduce a parameter $\mu \in \mathbb{R}$ whose purpose is to approximate the unknown eigenvalues $\{t_i\}_{i=m+1}^n$ of A'. We denote by $V^{(m)}$ the $n \times m$ matrix consisting of the m leading eigenvectors of A', and define

$$r_i = \left(I - V^{(m)}V^{(m)T}\right)Ev_i, \tag{7}$$

for $1 \le i \le m$.

We derive two approximation formulas based on the classical perturbation formula (5). These two approximation formulas differ in their order of approximation and in their computational complexity. The first formula, which we refer to as the first-order truncated perturbation formula, provides a first-order approximation to the eigenvectors of A, whereas the second formula, which we refer to as the second-order truncated perturbation formula, provides a second-order approximation to the eigenvectors of A. We describe these approximations in the following propositions.

Proposition 1 (First-order approximation) *Let $1 \le i \le m$. Let $\mu \in \mathbb{R}$ a parameter. Using the notation of Section 2.2, the first-order approximation to w_i is given by*

$$\widetilde{w}_i^{(1)} = v_i + \sum_{k=1, k \neq i}^m \frac{\langle Ev_i, v_k \rangle}{t_i - t_k} v_k + \frac{1}{t_i - \mu} r_i, \tag{8}$$

with an error satisfying

$$\left\| w_i - \widetilde{w}_i^{(1)} \right\|_2 \le \frac{\sum_{k=m+1}^n |t_k - \mu|}{|t_i - t_{m+1}||t_i - \mu|} \|E\|_2 + O\big(\|E\|_2^2\big). \tag{9}$$

Proposition 2 (Second-order approximation) *Let $1 \le i \le m$. Let $\mu \in \mathbb{R}$ a parameter. Using the notation of Section 2.2, the second-order approximation to w_i is given by*

$$\widetilde{w}_i^{(2)} = v_i + \sum_{k=1, k \neq i}^m \frac{\langle Ev_i, v_k \rangle}{t_i - t_k} v_k + \frac{1}{t_i - \mu} r_i - \frac{\mu}{(t_i - \mu)^2} r_i + \frac{1}{(t_i - \mu)^2} A' r_i, \tag{10}$$

with an error satisfying

$$\left\| w_i - \widetilde{w}_i^{(2)} \right\|_2 \le \frac{\sum_{k=m+1}^n |t_k - \mu|^2}{|t_i - t_{m+1}||t_i - \mu|^2} \|E\|_2 + O\big(\|E\|_2^2\big). \tag{11}$$

The proofs of Proposition 1 and Proposition 2 are given in Appendix A and Appendix B, respectively. We discuss the runtime and memory requirements of formulas (8) and (10) in Appendix C.

The update formulas (8) and (10) depend on a parameter μ, whose choice is discussed in Mitz et al. (2019). To be concrete, if A' is known to be low-rank, we use $\mu = 0$. When A' is not low-rank, and especially if its spectrum is known to decay slowly, we follow Mitz et al. (2019) and suggest to use

$$\mu_{\text{mean}} = \frac{\text{trace}(A') - \sum_{i=1}^{m} t_i}{n - m}, \tag{12}$$

which is the mean of the unknown eigenvalues of the perturbed matrix A'. μ_{mean} can be easily computed since the trace of A' and its m leading eigenvalues are known.

We conclude this section by proving that under a certain assumption on A', the first-order truncated perturbation formula (8) and the second-order truncated perturbation formula (10) coincide. Furthermore, in this case, the $O(\|E\|_2)$ term in the error bound of both approximations cancels out, as stated by the following proposition.

Proposition 3 *Let $\delta \geq 0$ and assume that A' can be written in the form of a low-rank matrix plus a spectrum shift, that is $A' = V^{(m)} T V^{(m)T} + \delta I$ for some diagonal matrix $T \in \mathbb{R}^{m \times m}$. Then, for $\mu = \delta$, the first-order truncated perturbation formula (8) and the second-order truncated perturbation formula (10) coincide, that is*

$$\widetilde{w}_i^{(1)} = \widetilde{w}_i^{(2)},$$

and the approximation errors satisfy

$$\left\| w_i - \widetilde{w}_i^{(1)} \right\|_2 = \left\| w_i - \widetilde{w}_i^{(2)} \right\|_2 = O(\|E\|_2^2),$$

for all $1 \leq i \leq m$.

The proof of Proposition 3 is given in Appendix D.

Corollary 4 *If A' is of rank m, and $\mu = 0$ in (8) and (10), then the first-order and second-order truncated perturbation formulas give rise to the same approximation. The error of this approximation is $O(\|E\|_2^2)$.*

4. Perturbation-based kernel approximation framework

In this section, we derive our perturbation-based kernel approximation framework based on Proposition 1. Let $K \in \mathbb{R}^{n \times n}$ be a symmetric positive-definite matrix with m distinct leading eigenpairs that are denoted by $\{(\lambda_i, u_i)\}_{i=1}^{m}$, ordered in descending order. Let $K^s \in \mathbb{R}^{n \times n}$ be a symmetric matrix consisting of any subset of entries of K, with the rest of its entries being 0, as illustrated in Figure 1. Our kernel approximation framework approximates the eigenvectors of K using the eigenvectors of any such K^s, as follows.

Let $\{(\lambda_i^s, u_i^s)\}_{i=1}^{m}$ be the leading eigenpairs of K^s ordered in a descending order, and assume that λ_i^s are distinct. Let $U^{s(m)} \in \mathbb{R}^{n \times m}$ be the matrix whose columns are the m

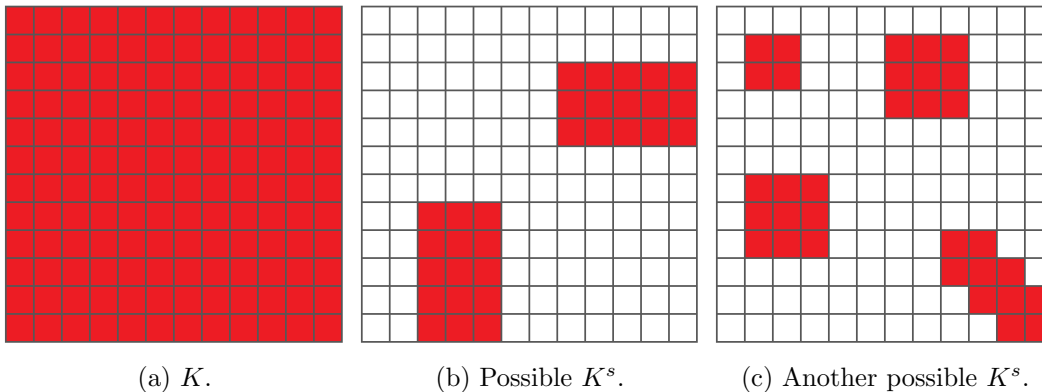(a) K. (b) Possible K^s. (c) Another possible K^s.

Figure 1: Illustration of the submatrix K^s. Blank entries indicate 0.

eigenvectors of K^s corresponding to the m largest eigenvalues of K^s, and let $\mu \geq 0$ be a parameter. Let $1 \leq i \leq m$. By the first-order approximation in Proposition 1, the eigenvector u_i is approximated by

$$\widetilde{u}_i = u_i^s + \sum_{k=1,k\neq i}^{m} \frac{\langle (K - K^s)u_i^s, u_k^s \rangle}{\lambda_i^s - \lambda_k^s} u_k^s + \frac{1}{\lambda_i^s - \mu}\big(I_m - U^{s(m)}U^{s(m)T}\big)(K - K^s)u_i^s, \quad (13)$$

with an error satisfying

$$\|u_i - \widetilde{u}_i\|_2 \leq \frac{\sum_{k=m+1}^{n}\big|\lambda_k^s - \mu\big|}{\big|\lambda_i^s - \lambda_m^s\big|\big|\lambda_i^s - \mu\big|}\|K - K^s\|_2 + O\big(\|K - K^s\|_2^2\big). \quad (14)$$

Furthermore, by (6), the eigenvalue λ_i is approximated by

$$\widetilde{\lambda}_i = \lambda_i^s + u_i^{sT}(K - K^s)u_i^s, \quad (15)$$

with an error of magnitude $\big|\lambda_i - \widetilde{\lambda}_i\big| = O(\|K - K^s\|_2^2)$.

We refer to (13) and (15) as our perturbation approximation for the eigenvectors and eigenvalues, respectively. Finally, we define our perturbation kernel approximation by

$$\widetilde{K}_{\mathrm{pert}} = \sum_{i=1}^{m} \widetilde{\lambda}_i \widetilde{u}_i^T \widetilde{u}_i. \quad (16)$$

Note that this framework is quite general, and can be applied to any symmetric submatrix of K. We will propose and discuss several choices for K^s in Section 6.

A kernel approximation framework analogous to (13) that is based on the second-order approximation in Proposition 2 can also be obtained in a similar way, but is typically impractical for large n due to its time and space requirements.

5. Equivalence with the Nyström method

In this section, we prove that our perturbation-based kernel approximation framework (13) is in fact a generalization of the Nyström method described in Section 2.1, by showing that

the Nyström method arises from our kernel approximation framework by a specific choice of K^s.

Let $K \in \mathbb{R}^{n \times n}$ be a kernel matrix, and let $m < n$. Assume without loss of generality, that we apply the classical Nystrom method, given in (1) and (2), to the $m \times m$ upper-left submatrix of K, and denote by $\{(\hat{\lambda}_i, \hat{u}_i)\}_{i=1}^m$ the resulting approximate eigenpairs of K. The following proposition states that for a specific choice of the matrix K^s, the perturbation approximation $\widetilde{K}_{\text{pert}}$ of (16) is exactly the classical Nyström approximation $\widetilde{K}_{\text{nys}}$ of (3).

Proposition 5 *Using the above notation, let K^s be the $n \times n$ matrix whose top left $m \times m$ submatrix is the top left $m \times m$ submatrix of K, and the rest of its entries are 0. Denote by $\{(\lambda_i^s, u_i^s)\}_{i=1}^m$ the eigenpairs of K^s. Set μ in (13) to be 0, and denote by $\{(\widetilde{\lambda}_i, \widetilde{u}_i)\}_{i=1}^m$ the perturbation approximation of the eigenpairs $\{(\lambda_i^s, u_i^s)\}_{i=1}^m$ of K^s to the eigenpairs of K. Denote by K' the top left $m \times m$ submatrix of K and by $\{(\lambda_i', u_i')\}_{i=1}^m$ its eigenpairs. Denote by $\{(\hat{\lambda}_i, \hat{u}_i)\}_{i=1}^m$ the Nyström approximation of $\{(\lambda_i', u_i')\}_{i=1}^m$ (see (1) and (2)). Then,*

$$\hat{u}_i = \sqrt{\frac{m}{n}}\widetilde{u}_i \quad and \quad \hat{\lambda}_i = \frac{n}{m}\widetilde{\lambda}_i \tag{17}$$

for all $1 \leq i \leq m$. In particular,

$$\widetilde{K}_{nys} = \widetilde{K}_{pert}.$$

The proof of Proposition 5 is given in Appendix E.

Formulating the Nyström method as a perturbation-based approximation using Proposition 5 enables us to provide an error bound for the Nyström method based on Propositions 1 and 2. Contrary to previous works that only provide error bounds for the approximated kernel resulting from the Nyström method (3), our error bound is for the individual approximated eigenvectors, as stated in the following proposition.

Proposition 6 (Error bound for the Nyström method) *Using the above notation, the error induced by the Nyström method satisfies*

$$\|u_i - \hat{u}_i\|_2 = O\big(\|K - K^s\|_2^2\big), \quad 1 \leq i \leq m.$$

The proof follows directly from the equivalence stated in Proposition 5 by noting that in the Nyström method setting, the requirements of Corollary 4 hold.

6. New kernel approximation schemes based on the perturbation framework

The perturbation-based kernel approximation framework derived in Section 4 is very flexible, and allows for various approximations that depend on the choice of the matrix K^s. There are two main considerations in choosing K^s. First, since calculating the eigendecomposition of K^s is the most expensive part of the perturbation-based kernel approximation, we wish to choose a matrix K^s whose eigendecomposition is "easy" to compute. This will make our framework computationally attractive. Second, we would like to take advantage of the flexibility of our framework and choose a matrix K^s that "captures" as much of a given

kernel K as possible (that is, to minimize the $\|K - K_s\|_2$ term in (13)). This will allow for better approximation results compared to classical Nyström-type methods.

In this section, we propose several approximation schemes corresponding to different choices of K^s. Furthermore, we prove that several of the Nystöom method variants described in Section 2 also arise from our general approximation framework for suitable choices of K^s. The list below is not by all means comprehensive, and users might come up with different approximation schemes that are more suitable to their problems' settings.

6.1 l-block kernel approximation

We define the l-block kernel approximation by choosing the matrix K^s to be the top left $l \times l$ submatrix of K padded with zeros to size $n \times n$ (see Figure 2), with $l \geq m$ being a parameter. The difference of this approach from the Nyström method is that while we still calculate m eigenpairs, we do so on a larger block, which "captures" more of K. This comes at the price of a greater computational cost.

Proposition 7 *Using the notation of this section, the eigenpairs calculated by the randomized SVD Nyström method (Li et al., 2014) and the l-block kernel approximation method are equal.*

The proof easily follows from the definitions and Proposition 5.

6.2 μ-shifted kernel approximation

We define the μ-shifted kernel approximation by choosing the matrix K^s to be the top left $m \times m$ submatrix of K padded with zeros to size $n \times n$, similarly to the Nyström method (see Figure 2). The difference from the Nyström method lies in the parameter μ of (13). In Proposition 5, we used the value of the parameter μ to be $\mu = 0$. This is a reasonable choice when the kernel matrix K is low-rank, or when its spectrum decays fast. When that is not the case, it might be beneficial to choose a parameter μ that approximates the unknown eigenvalues of K. A reasonable choice for μ in such a case is μ_{mean} of (12).

We now prove that given a parameter $\mu \geq 0$, the spectral shifted Nyström method (Wang et al., 2014) with parameter μ coincides with the perturbation approximation method with the same μ, as detailed in the following proposition.

Proposition 8 *Using the notation of this section, the eigenpairs calculated by the spectrum shifted Nyström method (Wang et al., 2014) and the μ-shifted approximation method are equal.*

The proof of Proposition 8 is given in Appendix F.

6.3 Block-diagonal kernel approximation

We define the block-diagonal kernel approximation by choosing the matrix K^s to be a block diagonal matrix (see Figure 2). The block sizes can be arbitrary, but for simplicity of notation, we choose k blocks of an identical size $l \geq m$. For each block, we pad the block with zeros to obtain an $n \times n$ matrix, and calculate its m leading eigenpairs. We

then approximate the eigenpairs of K using (13). Denote by $\{(\widetilde{\lambda}_i^{(j)}, \widetilde{u}_i^{(j)})\}_{i=1}^m$ the extended eigenpairs of block j, and by $\widetilde{K}_j \in \mathbb{R}^{n \times n}$ the resulting kernel approximation. To combine the k approximations $\{\widetilde{K}_j\}_{j=1}^k$ to an approximation of K, we set

$$\widetilde{K} = \frac{1}{k} \sum_{i=1}^k \widetilde{K}_i.$$

We note that the kernel matrix approximation obtained by this method generally won't be low-rank.

Proposition 9 *Using the notation of this section, the eigenpairs calculated by the ensemble Nyström method (Kumar et al., 2009) and the block diagonal kernel approximation method are equal.*

The proof easily follows from the definitions and Proposition 5.

6.4 p-band kernel approximation

We define the p-band kernel approximation by choosing the matrix K^s to be a band matrix of width p (see Figure 2). This approximation may provide superior results when the kernel K has most of its energy concentrated along the diagonal. Such a kernel may arise naturally for sequential data, where adjacent entries are more "similar" to each other. In such a case, the p-band kernel approximation will capture most of K. Existing Nyström-type approximations, on the other hand, are not able to do so since they are limited to block matrices. We demonstrate the advantage of this kernel approximation method in Section 7.2.1.

6.5 Sparse kernel approximation

We define the sparse kernel approximation by choosing the matrix K^s to be some sparse submatrix of K, as illustrated in Figure 3. More concretely, in the sparse approximation framework, we denote by $\mathrm{nnz}(K)$ the number of non-zero entries of K, and define K^s by choosing $q \cdot \mathrm{nnz}(K)$ entries of K, for some $0 < q \leq 1$. While this approximation is valid for any (symmetric) subset of elements of K, motivated by the $\|E\|_2$ term in the error bounds (9) and (11), we suggest choosing the $q \cdot \mathrm{nnz}(K)$ largest entries of K in their absolute value. We demonstrate the advantage of this kernel approximation method in Section 7.2.2.

7. Numerical examples

In this section, we demonstrate numerically the results obtained in the previous sections. We start by demonstrating numerically the error bounds derived in Section 3. Then, we demonstrate the advantages of the kernel approximation methods proposed in Section 6 using both real and synthetic datasets.

The MATLAB code to reproduce the graphs in this section is found in
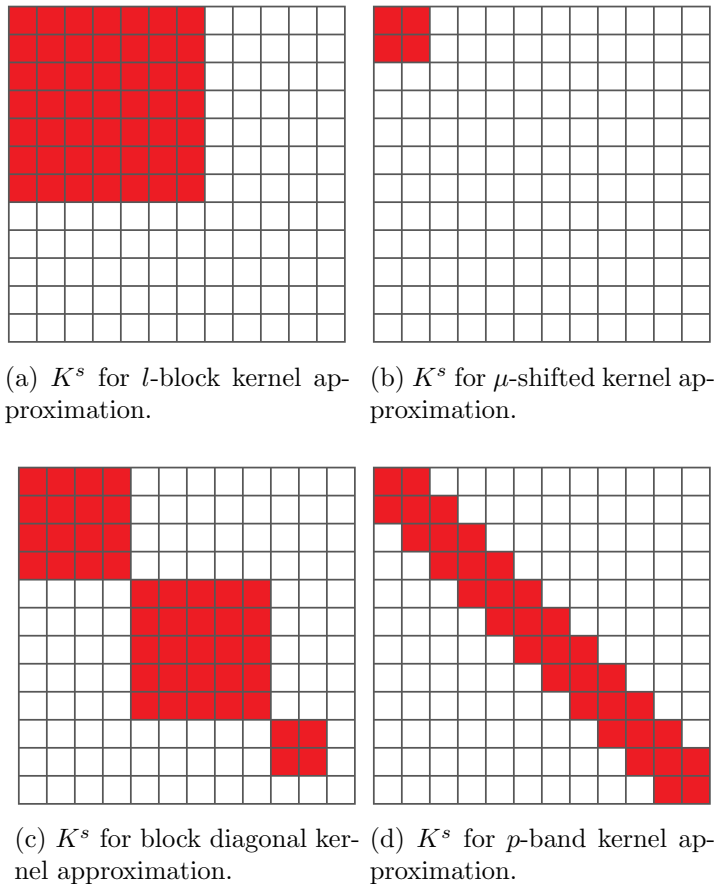`github.com/roymitz/perturbation_kernel_approximation`.

(a) K^s for l-block kernel approximation.

(b) K^s for μ-shifted kernel approximation.

(c) K^s for block diagonal kernel approximation.

(d) K^s for p-band kernel approximation.

Figure 2: Illustration of the submatrix K^s for each of the discussed kernel approximations. Blank entries indicate 0.

7.1 Perturbation error bounds

In this section, we demonstrate numerically the behavior of the error bounds (9) and (11) in Propositions 1 and 2, respectively. In our first example, we demonstrate the predicted linear dependence of the errors $\left\|w_i - \widetilde{w}_i^{(1)}\right\|_2$ and $\left\|w_i - \widetilde{w}_i^{(2)}\right\|_2$ on $\|E\|_2$. We also show the quadratic dependence of these errors on $\|E\|_2$ for a proper choice of μ. To that end, we generate a random symmetric matrix $A' \in \mathbb{R}^{1000 \times 1000}$ whose 10 leading eigenvalues are between 1 and 2, and the rest are exactly 0.5. We then generate a random symmetric matrix E, and normalize it to have a unit spectral norm. Then, for various values of c, we approximate the 10 leading eigenpairs of $A_c = A' + cE$ by the first and second-order approximations (8) and (10) using $\mu = 0$ and μ_{mean}. Denote by v_c the leading eigenvector of A_c, and by u_c^1 and u_c^2 its approximations by (8) and (10) using $\mu = 0$, respectively. Denote by w_c^1 and w_c^2 the respective approximations using μ_{mean}. For each c, we measure the errors $\left\|v_c - u_c^1\right\|_2, \left\|v_c - u_c^2\right\|_2, \left\|v_c - w_c^1\right\|_2$ and $\left\|v_c - w_c^2\right\|_2$. In Figure 4a, we plot all the $\log_{10}$-errors versus $\log_{10}\|cE\|_2$. As predicted by theory, there is a linear dependence between the error in

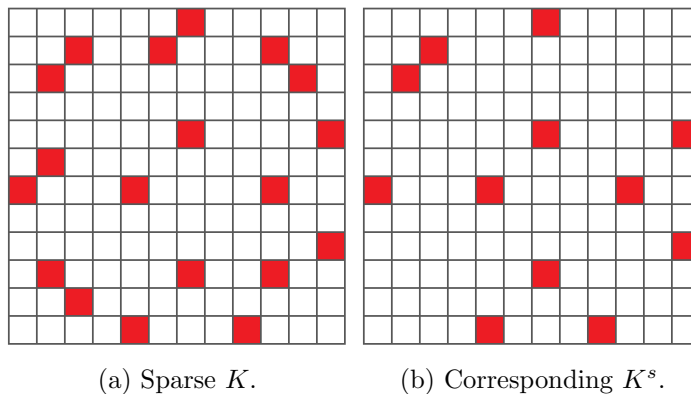(a) Sparse K.　　　　　(b) Corresponding K^s.

Figure 3: Illustration of a sparse kernel approximation. Blank entries indicate 0.

the eigenvector approximation and the norm of the perturbation matrix when using $\mu = 0$. Furthermore, we see that when using μ_{mean}, both formulas coincide and give rise to the same second-order approximation, as predicted by Proposition 3.

In our second example, we demonstrate the linear and quadratic dependence of the error on $\sum_{j=m+1}^{n} |\lambda_j - \mu|$, that is, on the unknown eigenvalues of the perturbed matrix. For various values of $c \in \mathbb{R}$, we generate matrices $A'_c \in \mathbb{R}^{1000 \times 1000}$ as follows. Their 10 leading eigenvalues are between 1 and 2, and are the same for all values of c. The rest of their eigenvalues are exactly c. Then, we generate a random symmetric matrix $E \in \mathbb{R}^{1000 \times 1000}$ and normalize it to have a norm of 10^{-6}. We choose $\|E\|_2$ to be relatively small, so that its contribution to the error will not mask the effect of $\sum_{j=m+1}^{n} |\lambda_j - \mu|$. We approximate the 10 leading eigenpairs of $A_c = A'_c + E$ by the first and second-order approximations (8) and (10) using $\mu = 0$, and measure the error in the same way as in the previous example. In Figure 4b, we plot $\log\|v_c - u_c^1\|_2$ and $\log\|v_c - u_c^2\|_2$ versus $\log|\lambda_j - \mu| = \log|\lambda_j| = \log c$. As predicted by Propositions 1 and 2, there is a linear dependence between the error in the eigenvector approximation and c for the first-order approximation, and a quadratic dependence for the second-order formula. We note that using μ_{mean} in this example will cancel out the $\sum_{j=m+1}^{n} |\lambda_j - \mu|$ term, since in such a case $\mu_{\mathrm{mean}} = c$.

7.2 Perturbation-based approximation for synthetic and real data

In this section, we compare the various kernel approximation methods proposed in Section 6 for both synthetic and real data. We demonstrate that the various approximation schemes perform differently, depending on the structure of the kernel matrix.

As our metric for comparing the performance of the various methods we use the kernel reconstruction error, i.e., if K_m is the best rank-m approximation of the kernel matrix of the entire data (obtained by SVD), and $\widetilde{K}$ is its approximation, we define

$$\mathrm{err} = \frac{\left\|K_m - \widetilde{K}\right\|_2}{\|K_m\|_2}.$$

We note that other metrics we tested performed qualitatively similarly. The metrics we tested were the principal angle (Knyazev and Zhu, 2012) between the subspace spanned by

13
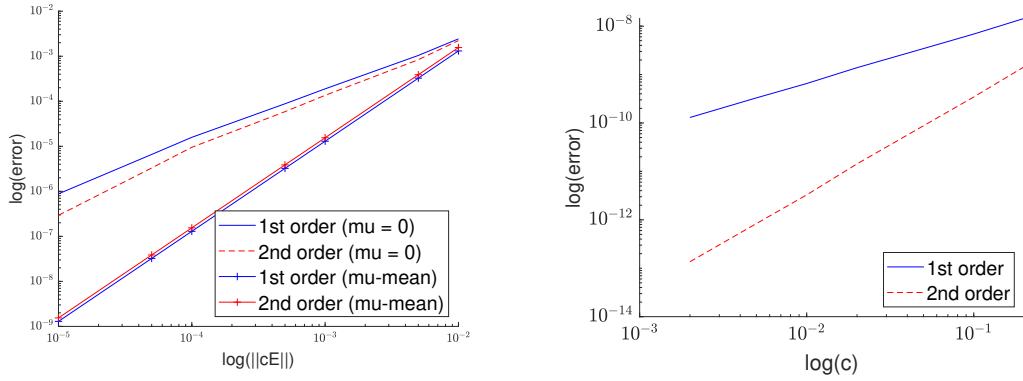
(a) The dependence of the first and second-order formulas on the perturbation matrix norm $\|E\|_2$.

(b) The dependence of the first and second-order formulas on the unknown eigenvalues of the perturbed matrix $\sum_{j=m+1}^{n}\big|\lambda_j - \mu\big|$.

Figure 4: Numerical demonstration of the error terms in the approximations (8) and (10). (a) $\log$ (error) vs. $\log\|cE\|_2$. The slope of both curves corresponding to $\mu = 0$ is ≈ 1, demonstrating the linear dependence of the error terms (9) and (11) on $\|E\|_2$ for both the first and second-order approximations. On the other hand, both curves corresponding to μ_{mean} coincide with slope ≈ 2, demonstrating Proposition 3. (b) $\log$ (error) vs. $\log\big|\lambda_j - \mu\big|$. The slope of the line corresponding to the first-order approximation is ≈ 1, whereas the slope of the line corresponding to the second-order approximation is ≈ 2, demonstrating the linear and quadratic dependence of the first and second-order error terms on $\sum_{j=m+1}^{n}\big|\lambda_j - \mu\big|$, respectively.

14

the kernel's top eigenvectors and the subspace spanned by their approximations, and the subspace projection error $\frac{\|UU^T - VV^T\|_2}{\|UU^T\|_2}$, where U and V are the ground truth subspace and its approximation, respectively. The results for all tested metrics are given in Appendix G.

In all the numerical examples to follow, we use $\mu = 0$ for all approximation schemes. We note that we typically witness a marginal difference between the performance of $\mu = 0$ and $\mu = \mu_{\text{mean}}$ for real-world data, as such data are usually close to being low-rank. Thus, we do not include the μ-shifted kernel approximation in the experiments of this section. We also set $n = 1000$ in all experiments. We choose m to be the number of components that account for 90% of the energy of K, with a maximum value of 5. For the block-diagonal kernel approximation, we always use two blocks of the same size.

Each experiment is performed as follows. For each kernel type, we generate several kernels of that type, each with a different kernel parameter (to be explained later for each example). Then, we approximate each of the kernels using each of the approximation methods, where in any case the matrix K^s used in the approximation consists of 20% of the entries of the approximated kernel. For each such kernel approximation, we measure the approximation error versus the Hoyer score (Hurley and Rickard, 2009) of the kernel matrix K written as a long vector. The Hoyer score of a vector $v \in \mathbb{R}^n$ is defined by

$$\text{Hoyer}(v) = \frac{\sqrt{n} - \frac{\|v\|_1}{\|v\|_2}}{\sqrt{n} - 1}.$$

The Hoyer score is a number between 0 and 1, with a higher score corresponding to a sparser vector. We repeat this procedure 20 times for each approximation scheme and kernel parameter, each time with a different random subset of the data. Finally, we plot the mean approximation error for each kernel approximation scheme versus the Hoyer score of the approximated kernel K. We also add to the plot the standard deviation of the error of each scheme, presented as a shaded plot around the mean.

7.2.1 Kernels concentrated along the diagonal

In this example, we demonstrate the performance of each of the kernel approximation schemes of Section 6 using kernel matrices whose energy is concentrated along their diagonal. We build these kernels as follows. We generate a matrix K whose (i, j) and (j, i) entries are $|i - j|^{-\alpha} + X$, where X are i.i.d samples from a normal distribution with mean zero and standard deviation 0.0001. Larger values of α correspond to a more concentrated matrix, whereas smaller values of α correspond to a more spread matrix.

We execute the experiment described at the beginning of Section 7.2. The results are presented in Figure 5. We can see that for higher Hoyer scores, which correspond to sparser matrices that are concentrated along the diagonal, the error graphs of the p-band and sparse kernel approximation schemes are lower than the error graphs of the other kernel approximation schemes. We conclude that when the kernel is concentrated along its diagonal, the p-band kernel approximation and the sparse kernel approximation have superior performance. The performance of the sparse kernel approximation is comparable to the performance of the p-band kernel approximation.
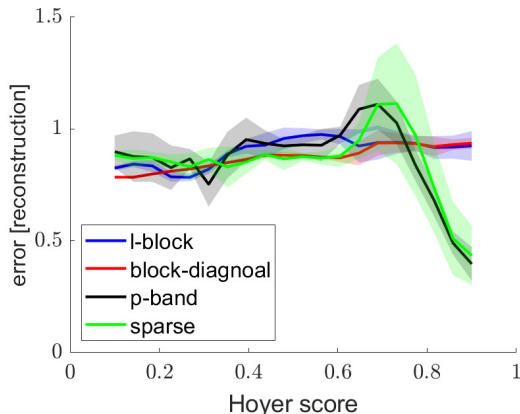
15

Figure 5: The error of the various kernel approximation schemes for a kernel concentrated along its diagonal. When the kernel is concentrated along the diagonal, and hence sparser, the p-band kernel approximation outperforms the other schemes. The performance of the sparse kernel approximation is comparable to the performance of the p-band kernel approximation.

7.2.2 Sparse kernels

In this example, we wish to demonstrate the performance of each kernel approximation scheme using kernel matrices that are sparse. As our kernel matrix, we use the symmetric normalized graph Laplacian matrix, used in Laplacian eigenmaps dimensionality reduction (Belkin and Niyogi, 2003). The (i, j) entry of the graph Laplacian matrix is given by $\exp(-\left\|x_i - x_j\right\|_2^2/\sigma)$, followed by some data-dependant normalization. The reason we use this kernel is that for small values of σ, this kernel is essentially sparse.

In this subsection, we use real-world datasets taken from the UCI Machine Learning Repository (Dua and Graff, 2017), as described in Table 1. For each dataset, we repeat the experiment described in the introduction of Section 7.2 multiple times, each time on a random subset of 1000 points from the dataset, and for several values of σ, reflecting the transition between a sparse matrix (small σ) and a dense matrix (large σ). The results of this experiment are presented in Figure 6. We can see that for higher Hoyer scores, which correspond to sparser matrices, the error graphs of the sparse kernel approximation scheme are lower than the error graphs of the other kernel approximation schemes. We conclude that when the kernel admits a sparse structure (typically, Hoyer score > 0.75), the sparse kernel approximation scheme has superior performance. For lower Hoyer scores, however, the kernel is usually no longer sparse and the performance of the sparse kernel approximation scheme is no longer superior to the other kernel approximation schemes. We notice that none of the methods performs well for dense matrices, whereas for sparse matrices, the use of the sparse extension may be the difference between a nearly meaningless result and an informative one.

Name	Dimension	Description
MNIST	784	Each sample is a grey scale image of a handwritten digit between zero and nine.
Superconductivity	81	Each sample contains 81 features extracted from one of 21263 superconductors.
Poker	10	Each sample is a hand consisting of five playing cards drawn from a standard deck of 52 cards. Each card is described using two attributes (suit and rank).
Wine quality	11	Each sample corresponds to a variant of a Portuguese wine, where the 11 attributes are numerical characteristics of the wine such as acidity, pH, residual sugar etc.

Table 1: Real-world datasets used.

(a) MNIST

(b) Superconductivity
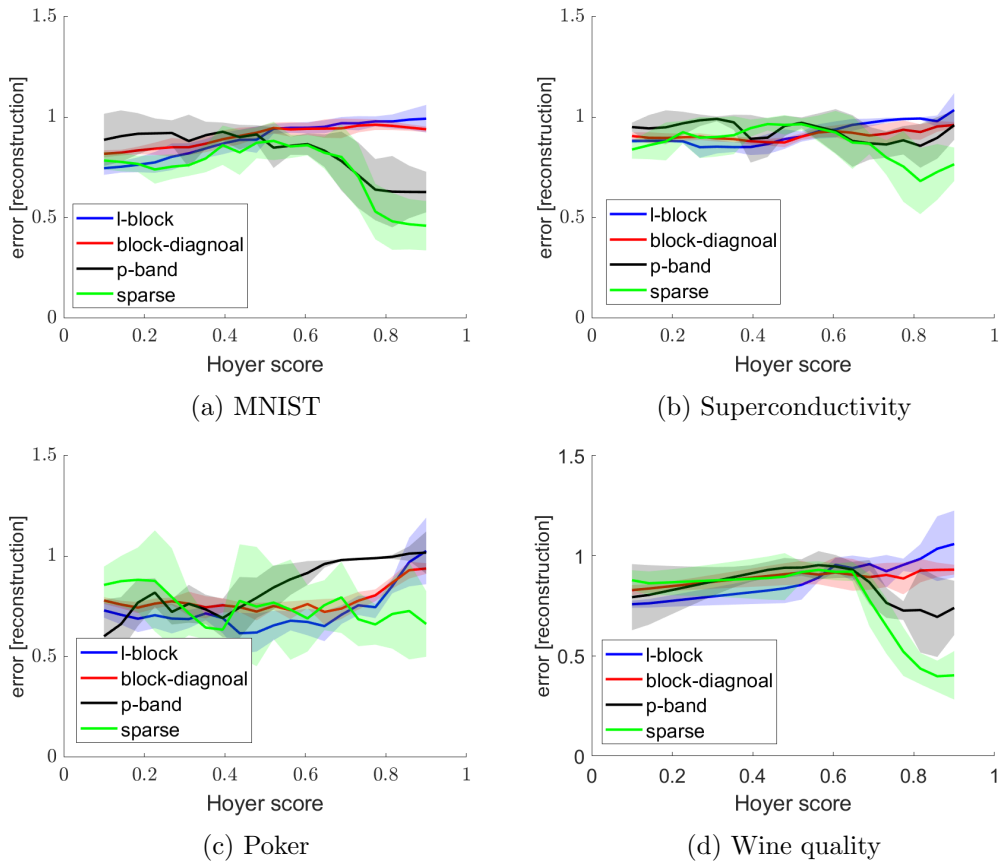
(c) Poker

(d) Wine quality

Figure 6: Error of the graph Laplacian approximation for various datasets and kernel approximation schemes as a function of the approximated kernel Hoyer score. We can see that for sparse kernels (higher Hoyer score, typically > 0.75), the sparse approximation is superior.

8. Summary and future work

In this paper, we propose a kernel approximation framework that is based on perturbation theory. We prove that this framework is a generalization of the popular Nyström method and some of its variants. Furthermore, contrary to existing error bounds for the Nyström method, our framework provides error bounds for the individual eigenvectors. This is useful when the approximation is used as part of a dimensionality reduction procedure. Our kernel approximation framework is very flexible, and can thus take advantage of the structure of the kernel matrix. We demonstrate our theoretical derivations numerically for kernel matrices that are either sparse or concentrated along their diagonal. Currently, our framework does not handle kernels with degenerate eigenvalues, nor crossovers between eigenvalues. The latter might cause a "change of order" of the eigenvectors.

For a future work, one might consider schemes to construct kernel matrices with structure that can take advantage of our framework. For example, the MEKA algorithm orders the data in clusters, resulting in a block-diagonal kernel.

Acknowledgments

Appendix A. Proof of Proposition 1

Let $1 \leq i \leq m$. Ignoring the $O(\|E\|_2^2)$ term, we split (5) into the known and unknown terms, resulting in

$$\widetilde{w}_i = v_i + \sum_{k=1, k \neq i}^{m} \frac{\langle Ev_i, v_k \rangle}{t_i - t_k} v_k + \sum_{k=m+1}^{n} \frac{\langle Ev_i, v_k \rangle}{t_i - t_k} v_k. \tag{18}$$

As the second term in (18) is unknown, we approximate it by replacing the unknown eigenvalues with a parameter μ, resulting in

$$\sum_{k=m+1}^{n} \frac{\langle Ev_i, v_k \rangle}{t_i - \mu} v_k = \frac{1}{t_i - \mu} \sum_{k=m+1}^{n} \langle Ev_i, v_k \rangle v_k = \frac{1}{t_i - \mu} (Ev_i - V^{(m)} V^{(m)T} Ev_i) = \frac{1}{t_i - \mu} r_i, \tag{19}$$

where r_i is defined in (7). Formula (8) follows by replacing the second term in (18) with the rightmost term in (19). We denote the approximation error introduced into the approximation (8) by e_i, that is

$$e_i = \left\| \sum_{k=m+1}^{n} \frac{\langle Ev_i, v_k \rangle}{t_i - t_k} v_k - \frac{1}{t_i - \mu} r_i \right\|_2.$$

By using the identity

$$\frac{1}{t_i - t_k} = \frac{1}{t_i - \mu} + \frac{t_k - \mu}{(t_i - t_k)(t_i - \mu)},$$

we get

$$e_i = \left\| \sum_{k=m+1}^{n} \frac{t_k - \mu}{(t_i - t_k)(t_i - \mu)} \langle Ev_i, v_k \rangle v_k \right\|_2.$$

By the triangle inequality and the Cauchy-Schwarz inequality we get

$$e_i \leq \sum_{k=m+1}^{n} \frac{|t_k - \mu|}{|t_i - t_k||t_i - \mu|} |\langle Ev_i, v_k \rangle| \leq \frac{\|E\|_2}{|t_i - t_{m+1}||t_i - \mu|} \sum_{k=m+1}^{n} |t_k - \mu|.$$

Recalling that the original perturbation approximation (5) induces an error of $O(\|E\|_2^2)$ concludes the proof. ∎

Appendix B. Proof of Proposition 2

Let $1 \leq i \leq m$. Ignoring the $O(\|E\|_2^2)$ term, we split (5) into the known and unknown terms, resulting in

$$\widetilde{w}_i = v_i + \sum_{k=1, k \neq i}^{m} \frac{\langle Ev_i, v_k \rangle}{t_i - t_k} v_k + \sum_{k=m+1}^{n} \frac{\langle Ev_i, v_k \rangle}{t_i - t_k} v_k. \tag{20}$$

To obtain (10), we expand the unknown (second) term in (20) by using the identity

$$\frac{1}{t_i - t_k} = \frac{1}{t_i - \mu} + \frac{t_k - \mu}{(t_i - \mu)^2} + \frac{(t_k - \mu)^2}{(t_i - t_k)(t_i - \mu)^2}. \tag{21}$$

We note that

$$\sum_{k=m+1}^{n} \langle Ev_i, v_k \rangle (t_k - \mu) v_k = \sum_{k=m+1}^{n} \langle Ev_i, v_k \rangle t_k v_k - \mu \sum_{k=m+1}^{n} \langle Ev_i, v_k \rangle v_k \qquad (22)$$

$$= \sum_{k=m+1}^{n} \langle Ev_i, v_k \rangle A' v_k - \mu \sum_{k=m+1}^{n} \langle Ev_i, v_k \rangle v_k \qquad (23)$$

$$= A' r_i - \mu r_i, \qquad (24)$$

where r_i is defined in (7). Using (24) and (19), we get that the unknown term in (20) can be written as

$$\sum_{k=m+1}^{n} \frac{\langle Ev_i, v_k \rangle}{t_i - t_k} v_k = \frac{1}{t_i - \mu} r_i - \frac{\mu}{(t_i - \mu)^2} r_i + \frac{1}{(t_i - \mu)^2} A' r_i + \sum_{k=m+1}^{n} \frac{(t_k - \mu)^2}{(t_i - t_k)(t_i - \mu)^2} \langle Ev_i, v_k \rangle v_k.$$
$$(25)$$

The second-order formula (10) now follows by discarding the last term in (25). Denoting the approximation error of equation (10) by e_i, we have

$$e_i = \left\| \sum_{k=m+1}^{n} \frac{\langle Ev_i, v_k \rangle}{t_i - t_k} v_k - \left(\frac{1}{t_i - \mu} r_i - \frac{\mu}{(t_i - \mu)^2} r_i + \frac{1}{(t_i - \mu)^2} A' r_i \right) \right\|_2.$$

Using (21) and the triangle and the Cauchy-Schwarz inequalities, we obtain

$$e_i = \left\| \sum_{k=m+1}^{n} \frac{(t_k - \mu)^2}{(t_i - t_k)(t_i - \mu)^2} \langle Ev_i, v_k \rangle v_k \right\|_2$$

$$\leq \sum_{k=m+1}^{n} \frac{|t_k - \mu|^2}{|t_i - t_k||t_i - \mu|^2} \left| \langle Ev_i, v_k \rangle \right|$$

$$\leq \frac{\|E\|_2}{|t_i - t_{m+1}||t_i - \mu|^2} \sum_{k=m+1}^{n} |t_k - \mu|^2 .$$

Recalling that the original perturbation approximation (5) induces an error of $O(\|E\|_2^2)$ concludes the proof. ∎

Appendix C. Runtime and space complexity

In this section, we discuss the runtime and space complexities of formulas (8) and (10). We start with the first-order formula (8). The space complexity of the first-order formula is $O(mn)$, since it needs to store in memory the m leading eigenvectors of A'. As of the runtime of the first-order formula, the computation of r_i of (7) involves the calculation of Ev_i that requires $O(\mathrm{nnz}(E))$ operations. The result is then multiplied by $V^{(m)T}$, which requires $O(mn)$ operations, and then by $V^{(m)}$, which also requires $O(mn)$ operations. Thus, the total runtime complexity for computing all $\{r_i\}_{i=1}^{m}$ is $O(m \cdot \mathrm{nnz}(E) + m^2 n)$ operations. The first-order formula also requires to compute $O(m^2)$ terms of the form $\langle Ev_i, v_k \rangle v_k$ for

$1 \leq i \neq k \leq m$. Each such term requires $O(\mathrm{nnz}(E) + n)$ operations, resulting in a total of $O(m^2 \cdot \mathrm{nnz}(E) + m^2 n)$ operations for all eigenvectors. We conclude that evaluating the first-order formula requires a total of $O(m^2 \cdot \mathrm{nnz}(E) + m^2 n)$ operations.

The analysis of the second-order formula (10) is similar, except that it requires also to store A', which requires $O(\mathrm{nnz}(A'))$ memory, and to compute $A' r_i$, which requires additional $O(m \cdot \mathrm{nnz}(A'))$ operations. We conclude that evaluating the second-order formula requires a total of $O(m^2 \cdot \mathrm{nnz}(E) + m \cdot \mathrm{nnz}(A') + m^2 n)$ operations.

Appendix D. Proof of Proposition 3

Let $i \in \{1, \ldots, m\}$. Since $\mu = \delta$, we get that if $A' r_i = \delta r_i$ then the last two terms in (10) cancel out and (10) reduces to the first-order formula (8). Thus, in order to prove that $w_i^{(1)} = w_i^{(2)}$, it is sufficient to prove that under the settings of the proposition $A' r_i = \delta r_i$. Indeed,

$$
\begin{aligned}
A' r_i &= A' \Big(I - V^{(m)} V^{(m)T} \Big) E v_i = \Big(V^{(m)} T V^{(m)T} + \delta I \Big) \Big(I - V^{(m)} V^{(m)T} \Big) E v_i \\
&= \Big(V^{(m)} T V^{(m)T} - V^{(m)} T V^{(m)T} + \delta I - \delta V^{(m)} V^{(m)T} \Big) E v_i \\
&= \delta \Big(I - V^{(m)} V^{(m)T} \Big) E v_i = \delta r_i.
\end{aligned}
$$

For the error, we note that the $n - m$ unknown eigenvalues of a matrix A' of the form $A' = V^{(m)} T V^{(m)T} + \delta I$ are exactly δ, and thus, when choosing $\mu = \delta$, the first term in (9) and (11) cancels out and we are left with only the $O(\|E\|_2^2)$ term. ∎

Appendix E. Proof of Proposition 5

Let $i \in \{1, \ldots, m\}$. By (13), for $\mu = 0$

$$
\widetilde{u}_i = u_i^s + \sum_{k=1, k\neq i}^m \frac{((K - K^s)u_i^s, u_k^s)}{\lambda_i^s - \lambda_k^s} u_k^s + \frac{1}{\lambda_i^s}\Big(I - U^{s(m)} U^{s(m)T} \Big)(K - K^s)u_i^s, \qquad (26)
$$

and by (15),

$$
\widetilde{\lambda}_i = \lambda_i^s + u_i^{sT}(K - K^s)u_i^s. \qquad (27)
$$

We first prove that $\hat{u}_i = \sqrt{\frac{m}{n}} \widetilde{u}_i$ (see (17)). We start by simplifying (26) based on the specific choice of K^s. We make the following observations. First, we note that for all $i = 1, \ldots, m$, the last $n - m$ entries of u_i^s are 0, and the top left $m \times m$ submatrix of $K - K^s$ is 0. This implies that the first m entries of $(K - K^s)u_i^s$ are 0, and so $\langle (K - K^s)u_i^s, u_k^s \rangle = 0$ for all $1 \leq i, k \leq m$. A direct consequence of the latter is that $U^{s(m)} U^{s(m)T}(K - K^s)u_i^s = 0$ for all $1 \leq i \leq m$. Thus, (26) reduces to

$$
\widetilde{u}_i = u_i^s + \frac{1}{\lambda_i^s}(K - K^s)u_i^s. \qquad (28)
$$

Next, we note that the first term in (28), u_i^s, is non-zero only on its first m entries, whereas the second term, $\frac{1}{\lambda_i^s}(K - K^s)u_i^s$ is non-zero only on its last $n - m$ entries. This means that

the first m entries of $\widetilde{u}_i$ are exactly the first m entries u_i^s, and the last $n-m$ entries of $\widetilde{u}_i$ are equal to the last $n-m$ entries of $\frac{1}{\lambda_i^s}(K - K^s)u_i^s$. We start by proving the equivalence between the first m entries of $\widetilde{u}_i$ and $\hat{u}_i$. Let $1 \le p \le m$. Denote by $A_{p\to}$ the p'th row of a matrix A. Denote by C the $n \times m$ matrix consisting of the first m columns of K. We note that for the Nyström method (see (1))

$$\hat{u}_{i,p} = \sqrt{\frac{m}{n}}\frac{1}{\lambda_i'}C_{p\to}u_i' = \sqrt{\frac{m}{n}}\frac{1}{\lambda_i'}K_{p\to}'u_i' = \sqrt{\frac{m}{n}}\frac{1}{\lambda_i'}\lambda_i'u_{i,p}' = \sqrt{\frac{m}{n}}u_{i,p}'.$$

Thus, the first m entries of the approximated vector in the Nyström method are merely a re-scaling of the vector u_i' by $\sqrt{\frac{m}{n}}$. Since K^s is equal to K' padded with zeros, we get that the first m entries of u_i^s are exactly u_i'. Thus, we conclude that the first m entries of $\hat{u}_i$ and $\sqrt{\frac{m}{n}}\widetilde{u}_i$ are identical.

Next, we prove the equivalence between the last $n-m$ entries of $\widetilde{u}_i$ and $\hat{u}_i$. Let $m+1 \le p \le n$. We have for the Nyström method (see (1))

$$\hat{u}_{i,p} = \sqrt{\frac{m}{n}}\frac{1}{\lambda_i'}C_{p\to}u_i', \tag{29}$$

and for the perturbation approximation, by using (28), and since the last $n-m$ entries of u_i^s are 0,

$$\widetilde{u}_{i,p} = \frac{1}{\lambda_i^s}(K - K^s)_{p\to}u_i^s = \frac{1}{\lambda_i^s}(K_{p\to}u_i^s - 0) = \frac{1}{\lambda_i^s}C_{p\to}u_i', \tag{30}$$

where the last equality follows since (as explained above) the first m entries of u_i^s are exactly u_i'. Finally, since $\lambda_i' = \lambda_i^s$ for $1 \le i \le m$, we conclude by (29) and (30) that $\hat{u}_{i,p} = \sqrt{\frac{m}{n}}\widetilde{u}_{i,p}$ for $m+1 \le p \le n$, that is, the last $n-m$ entries of $\hat{u}_i$ and $\sqrt{\frac{m}{n}}\widetilde{u}_i$ are also identical.

We now prove the equivalence of the eigenvalues. By the same arguments as above, we note that for all $1 \le i \le m$, $u_i^{sT}(K - K^s)u_i^s = 0$. Thus, by (27) we have $\widetilde{\lambda}_i = \lambda_i^s$ and consequently, using (2),

$$\hat{\lambda}_i = \frac{n}{m}\lambda_i' = \frac{n}{m}\lambda_i^s = \frac{n}{m}\widetilde{\lambda}_i,$$

as required. ∎

Appendix F. Proof of Proposition 8

Denote by $K_{\text{shift}}^s \in \mathbb{R}^{n \times n}$ the top left $m \times m$ submatrix of K_{shift} (see (4)) padded with zeros. By the equivalence of the Nyström method and the perturbation approximation proved in Proposition 5, the spectral shifted Nyström method is equivalent to the perturbation approximation of K_{shift}^s using $\mu = 0$. Thus, it suffices to prove that the perturbation approximation of the eigenpairs of K_{shift}^s to the eigenpairs of K_{shift} with $\mu = 0$ equals to the perturbation approximation of the eigenpairs of K^s to the eigenpairs of K with $\mu = \delta$.

Let the top m eigenpairs of K^s be $\{(\lambda_i^s, u_i^s)\}_{i=1}^m$. It follows that the top m eigenpairs of K_{shift}^s are $\{(\lambda_i^s - \delta, u_i^s)\}_{i=1}^m$. Let $1 \le i \le m$. By (28), the perturbation approximation (13) of the eigenvectors of K_{shift}^s to eigenvectors of K_{shift} with $\mu = 0$ reduces to

$$\widetilde{u}_i = u_i^s + \frac{1}{\lambda_i - \delta - 0}(K_{\text{shift}} - K_{\text{shift}}^s)u_i^s,$$

whereas the perturbation approximation of the eigenvectors of K^s to eigenvectors of K with $\mu = \delta$ reduces to

$$\hat{u}_i = u_i^s + \frac{1}{\lambda_i - \delta}(K - K^s)u_i^s.$$

But since the last $n - m$ entries of u_i^s are 0, and the entries of $K_{\text{shift}} - K_{\text{shift}}^s$ are equal to those of $K - K^s$ except for the last $n - m$ diagonal elements, we have that

$$(K_{\text{shift}} - K_{\text{shift}}^s)u_i^s = (K - K^s)u_i^s, \tag{31}$$

and we conclude that $\widetilde{u}_i = \hat{u}_i$.

For the eigenvalues, the perturbation approximation of the eigenvalues of K_{shift}^s to the eigenvalues of K_{shift} (see (15)) yields

$$\widetilde{\lambda}_i = (\lambda_i^s - \delta) + u_i^s(K_{\text{shift}} - K_{\text{shift}}^s)u_i^s.$$

We note that by (15), (31) and by the proof of this section for the eigenvectors, if $\{(\tau_i, v_i)\}_{i=1}^m$ are the approximated eigenpairs of a matrix A, then $\{(\tau_i + \delta, v_i)\}_{i=1}^m$ are the approximated eigenpairs of the matrix $A_{\text{shift}} = A + \delta I$. Thus, in order to recover the approximation of the eigenvalues of K, we need to shift the approximated eigenvalues $\{\widetilde{\lambda}_i\}_{i=1}^m$ back by δ, yielding

$$\widetilde{\lambda}_i' = \lambda_i^s + u_i^s(K_{\text{shift}} - K_{\text{shift}}^s)u_i^s.$$

On the other hand, the perturbation approximation of K^s yields

$$\hat{\lambda}_i = \lambda_i^s + u_i^s(K - K^s)u_i^s.$$

By (31), we have that $\widetilde{\lambda}_i' = \hat{\lambda}_i$. ∎

Appendix G. Performance of various error metrics

In this section, we provide the results of Section 7.2.2 for the MNIST and wine datasets for three different error measures, demonstrating that the performance of our approximation scheme is qualitatively similar in various reasonable error metrics. The results are presented in Figure 7 and in Figure 8 for the MNIST and wine datasets respectively.
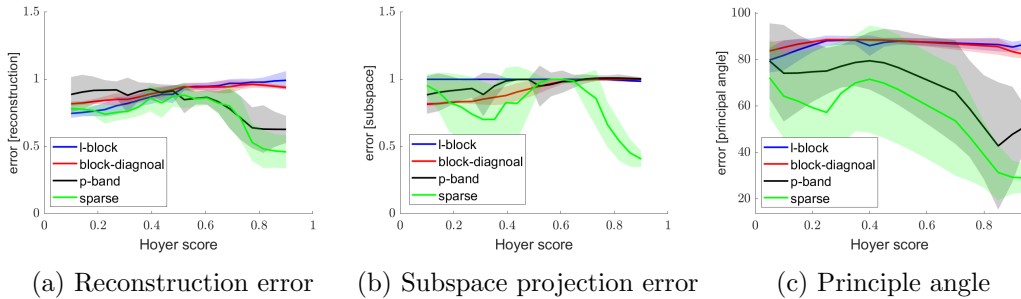

(a) Reconstruction error   (b) Subspace projection error   (c) Principle angle

Figure 7: Approximation error for the MNIST dataset in various error metrics.

(a) Reconstruction error   (b) Subspace projection error   (c) Principle angle
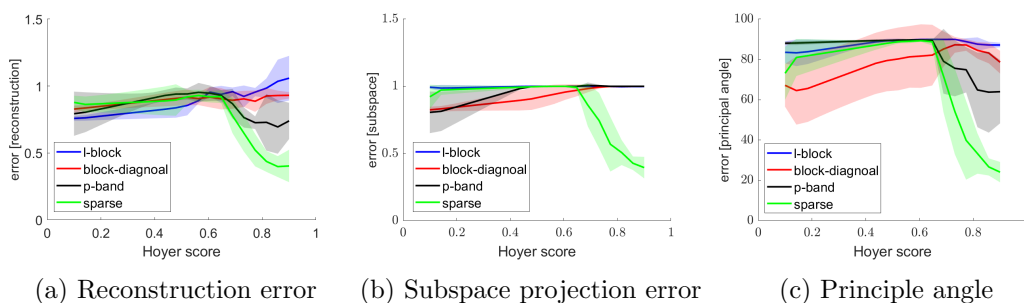
Figure 8: Approximation error for the wine dataset in various error metrics.

References

Mukund Balasubramanian, Eric L Schwartz, Joshua B Tenenbaum, Vin de Silva, and John C Langford. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Matthew Brand. Fast low-rank modifications of the thin singular value decomposition. *Linear algebra and its applications*, 415(1):20–30, 2006.

Andreas Buja, Deborah F Swayne, Michael L Littman, Nathaniel Dean, Heike Hofmann, and Lisha Chen. Data visualization with multidimensional scaling. *Journal of computational and graphical statistics*, 17(2):444–472, 2008.

James R Bunch, Christopher P Nielsen, and Danny C Sorensen. Rank-one modification of the symmetric eigenproblem. *Numerische Mathematik*, 31(1):31–48, 1978.

Frederick W Byron and Robert W Fuller. *Mathematics of classical and quantum physics*. Courier Corporation, 2012.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Alex Gittens and Michael Mahoney. Revisiting the Nyström method for improved large-scale machine learning. In *International Conference on Machine Learning*, pages 567–575. PMLR, 2013.

Nathan Halko, Per-Gunnar Martinsson, Yoel Shkolnisky, and Mark Tygert. An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific computing*, 33(5):2580–2594, 2011a.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011b.

Niall Hurley and Scott Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

Andrew V Knyazev and Peizhen Zhu. Principal angles between subspaces and their tangents. *arXiv preprint arXiv:1209.0523*, 2012.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Ensemble Nyström method. In *Advances in Neural Information Processing Systems*, pages 1060–1068, 2009.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the Nyström method. *Journal of Machine Learning Research*, 13(Apr):981–1006, 2012.

Amy N Langville and Carl D Meyer. Updating markov chains with an eye on google's pagerank. *SIAM journal on matrix analysis and applications*, 27(4):968–987, 2006.

Mu Li, Wei Bi, James T Kwok, and Bao-Liang Lu. Large-scale Nyström kernel matrix approximation using randomized svd. *IEEE transactions on neural networks and learning systems*, 26(1):152–164, 2014.

Roy Mitz, Nir Sharon, and Yoel Shkolnisky. Symmetric rank-one updates from partial spectrum with an application to out-of-sample extension. *SIAM Journal on Matrix Analysis and Applications*, 40(3):973–997, 2019.

Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

HyungSeon Oh and Zhe Hu. Multiple-rank modification of symmetric eigenvalue problem. *MethodsX*, 5:103–117, 2018.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. 1998.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Si Si, Cho-Jui Hsieh, and Inderjit S Dhillon. Memory efficient kernel approximation. *The Journal of Machine Learning Research*, 18(1):682–713, 2017.

Gilbert W Stewart. Matrix perturbation theory. 1990.

Shiliang Sun, Jing Zhao, and Jiang Zhu. A review of Nyström methods for large-scale machine learning. *Information Fusion*, 26:36–48, 2015.

Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

Shusen Wang and Zhihua Zhang. Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling. *The Journal of Machine Learning Research*, 14(1): 2729–2769, 2013.

Shusen Wang and Zhihua Zhang. Efficient algorithms and error analysis for the modified Nyström method. In *Artificial Intelligence and Statistics*, pages 996–1004. PMLR, 2014.

Shusen Wang, Chao Zhang, Hui Qian, and Zhihua Zhang. Improving the modified Nyström method using spectral shifting. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–620, 2014.

Christopher KI Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.