A perturbation based out-of-sample extension framework

Roy Mitz* 1 and Yoel Shkolnisky†1

1School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel

Abstract

Out-of-sample extension is an important task in various kernel based non-linear dimensionality reduction algorithms. In this paper, we derive a perturbation based extension framework by extending results from classical perturbation theory. We prove that our extension framework generalizes the well-known Nyström method as well as some of its variants. We provide an error analysis for our extension framework, and suggest new forms of extension under this framework that take advantage of the structure of the kernel matrix. We support our theoretical results numerically and demonstrate the advantages of our extension framework both on synthetic and real data.

Key Words. perturbation theory, out-of-sample extension, kernel approximation, Nyström method.

1 Introduction

Non-linear dimensionality reduction algorithms are a powerful tool in data analysis and manifold learning. Many of these algorithms require the calculation of the eigendecomposition of some data-dependent kernel matrix. Examples of such algorithms include Laplacian eigenmaps [3], LLE [21], Isomap [2], MDS [6], Spectral clustering [22], and more. One of the key issues with these non-linear methods is how to map new unseen data to the previously learnt embedding. This process is called out-of-sample extension.

Naively, one could repeat the eigendecomposition calculation on the entire data from scratch. However, since in most of these methods the dimension 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 components calculated. Randomized algorithms [12, 13] 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 the RAM or on the disc.

Various methods for out-of-sample extension were proposed (see for example [26, 14, 11, 19]), with the most prominent one being the Nyström method [4, 9]. We will describe the Nyström method in detail in

*roymitz@mail.tau.ac.il
†yoelsh@tauex.tau.ac.il
the next section. Due to the prominence and relevance of the Nyström method to this work, we wish to further discuss some of its aspects.

An important application of the Nyström method is kernel approximation. In this task, we are mainly interested in approximating the kernel matrix of the entire data itself, and the approximation of the individual eigenvalues and eigenvectors of the kernel plays a second role. In this case, we use the eigendecomposition approximation obtained by the Nyström method to produce a low-rank approximation of the kernel. This allows to speed-up kernel related calculations [31].

The error analysis of the Nyström method and its variants is widely investigated, see [30, 11] and the references therein. To the best of our knowledge, all error bounds obtained in the literature focus on the error of the kernel matrix approximation, rather than on the error of the individual extended eigenvectors. These bounds may be less useful when the Nyström method is used for out-of-sample extension (usually as part of dimensionality reduction), where the individual eigenvectors of the kernel are of great importance.

The performance of the Nyström method depends on sampling a subset of the data points and many methods for sampling this subset were proposed, see [17, 27] and the references therein. Our results are independent of the methodology used to obtain the subset of samples, and hence we do not discuss this issue in detail.

Further improvements of the Nyström method that are not sampling-related were also proposed in literature. The most notable ones include the ensemble Nyström method [16], that averages several Nyström extensions in order to improve performance, the spectral shifted Nyström method [29] that provides superior performance in cases where the spectrum of the matrix decays slowly, and the modified Nyström method [30]. More recently, works that use the structure of the kernel matrix were proposed. For example, the MEKA algorithm [23] provides superior kernel approximation for kernels the admit a block-diagonal structure. We will describe some of these methods in detail in the next section.

A problem related to eigendecomposition 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 [24] 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, [7, 19] for the case where the perturbation is of rank one, and [20, 5] for a general low-rank perturbation. Other approaches of updating a known eigendecomposition include restarting the power method [18] or the inverse iteration algorithm [28], 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 out-of-sample extension of eigenvectors. Unlike some of the existing extension methods, we show explicit error bounds for our approach for the individual extended eigenvectors. Third, we prove that the Nyström method and its generalizations 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 extensions that utilize the structure of the kernel matrix. It also allows for our approach to be used for kernel approximation, analogously to the Nyström method.

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 term. In Section 4 we use these formulas to develop a perturbation based extension framework. In Section 5, we prove that our extension framework generalizes the Nyström method. In Section 6, we suggest methods to improve the accuracy of our extension, 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 extension framework. In Section 8, we summarize our work.

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 perturbation method in Section 2.1, and then describe the Nyström method in Section 2.2.

2.1 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 \( \{(s_i, w_i)\}_{i=1}^n \). We wish to find an approximation to the eigenpairs of \( A \). The classical perturbation solution to this problem \[25\] is as follows. The approximated eigenvectors of \( A \) are given by

\[
\tilde{w}_i = v_i + \sum_{k=1, k \neq i}^n \frac{(Ev_i, v_k)}{t_i - t_k} v_k + O(\|E\|_2^2), \quad 1 \leq i \leq n, \tag{1}
\]

and the approximated eigenvalues of \( A \) are given by

\[
\tilde{s}_i = t_i + v_i^T Ev_i + O(\|E\|_2^2), \quad 1 \leq i \leq n. \tag{2}
\]

We note that the eigenvalues update formula (2) depends only on the updated eigenvalue and its corresponding eigenvector, whereas the eigenvectors update formula (1) depends on all eigenvalues and eigenvectors of \( A' \).

**Remark 2.1.** There exist perturbation results for matrices with non-simple eigenvalues \[8\]. 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.

2.2 The Nyström method and its variants

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 \[27, 31\] finds an approximation \( \{(	ilde{\lambda}_i, \tilde{u}_i)\}_{i=1}^k \) to these eigenpairs as follows. First, \( k \) columns of \( K \) are sampled (typically uniformly at random without replacement). We assume, without loss of generality, that the columns and rows of \( K \) were rearranged so that the first \( k \) columns of \( K \) were sampled. Denote by \( K' \) the \( k \times k \) matrix consisting of the first \( k \) rows and columns
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 extension approximates the $k$ leading eigenvectors of $K$ by

$$\tilde{u}_i = \sqrt{\frac{k}{n}} \lambda'_i u'_i, \quad i = 1, \ldots, k.$$  \hfill (3)

Moreover, the $n$ leading eigenvalues of $K$ are approximated by

$$\tilde{\lambda}_i = \frac{n}{k} \lambda'_i, \quad i = 1, \ldots, k.$$  \hfill (4)

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

In some applications, we are interested in an approximation of $K$ itself rather than its $k$ leading eigenpairs. In this case, the Nyström approximation of $K$ is

$$\tilde{K}_{\text{nys}} = \sum_{i=1}^k \tilde{\lambda}_i \tilde{u}_i \tilde{u}_i^T.$$  \hfill (5)

A straightforward generalization of the Nyström method is the following. Let $l \geq k$ and choose $K'$ to be the $l \times l$ top-left submatrix of $K$. We calculate the $k$ leading eigenpairs of $K'$, and then extend them using (3) and (4). This form of extension is a generalization of the Nyström method, since choosing $l = k$ is equivalent to the Nyström method. If we choose $l = n$, we get the exact eigenvectors of $K$. Intuitively, the larger $l$ is, the better the approximation will be, at the cost of a greater computational complexity. The runtime complexity of this method is $O(nk^2 + lk^2)$. We will use this generalization of the Nyström method in the numerical experiments in Section 7.

Since the Nyström approximation (5) 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 its spectrum decays slowly. A possible approach to overcome this problem is the spectrum shifted Nyström extension [29]. This method essentially applies the classical Nyström extension on a shifted kernel matrix, i.e., applies the Nyström extension on

$$K_{\text{shift}} = K - \mu I,$$  \hfill (6)

for some $\mu \geq 0$. The updated eigenvalues (4) are then shifted-back by $\mu$. It is suggested in [29] to set $\mu$ to be the mean of the smallest $n-k$ eigenvalues of $K$, that is

$$\mu = \frac{\sum_{j=k+1}^n \lambda_j}{n-k} = \frac{\text{trace}(K) - \sum_{j=1}^k \lambda_j}{n-k}.$$  \hfill (7)

If we denote by $\tilde{K}_{\text{shift}}$ the kernel approximation (5) obtained using the shifted Nyström extension, it is shown in [29] that

$$\|K - \tilde{K}_{\text{shift}}\|_F \leq \|K - \tilde{K}_{\text{nys}}\|_F.$$  \hfill (8)

Alternatively, the kernel $K$ may admit a block-diagonal structure. As demonstrated in [23], this may happen for some kernel functions when the data consist of several clusters. In this case, the MEKA algorithm [23] essentially performs a Nyström approximation on each cluster of data. Each such approximation corresponds to a block on the diagonal of the kernel matrix, and the resulting approximation is
block-diagonal.

A related approach to the MEKA algorithm for improving the Nyström approximation \cite{5} is the ensemble Nyström method \cite{16}. 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 \( \{ \tilde{K}_i \}_{i=1}^q \), the ensemble Nyström approximation is given by

\[
\tilde{K}_{\text{ens}} = \frac{1}{q} \sum_{i=1}^q \mu_i \tilde{K}_i,
\]

for some weights \( \{ \mu_i \}_{i=1}^q \). It is suggested in \cite{16} to use \( \mu_i = \frac{1}{q} \) for \( 1 \leq i \leq q \). Better error bounds for this method compared to the classical Nyström method are proven in \cite{16}.

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.

\section{Truncating the perturbation formulas}

In this section, we consider a variant of the problem presented in Section 2.1 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 derive two approximation formulas based on the classical perturbation formula \cite{1}. These two approximation formulas differ in their order of approximation as well as 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 \( A \), as detailed in the following proposition.

\begin{proposition}
Let \( A' \in \mathbb{R}^{n \times n} \) be a real symmetric matrix with \( m \) leading eigenpairs \( \{(t_i, v_i)\}_{i=1}^m \). Assume that \( t_1 > t_2 > \cdots > t_m \). Let \( E \in \mathbb{R}^{n \times n} \) be a real symmetric matrix. Let \( A = A' + E \) be a perturbation of \( A' \), and denote the \( m \) leading eigenpairs of \( A \) by \( \{(s_i, w_i)\}_{i=1}^m \). Denote by \( V^{(m)} \) the \( n \times m \) matrix consisting of the \( m \) leading eigenvectors of \( A' \). Let \( \mu \in \mathbb{R} \) and \( 1 \leq i \leq m \). Denote

\[
r_i = \left( I - V^{(m)} V^{(m)T} \right) E v_i.
\]

Then, \( w_i \) is approximated by the first order truncated perturbation formula

\[
\tilde{w}_i^{(1)} = v_i + \sum_{k=1, k \neq i}^m \frac{(E v_i, v_k)}{t_i - t_k} v_k + \frac{1}{t_i - \mu} r_i,
\]

with an error satisfying

\[
\|w_i - \tilde{w}_i^{(1)}\| \leq \sum_{k=m+1}^n \frac{|t_k - \mu|}{|t_i - t_m| |t_i - \mu|} \|E\|_2 + O(\|E\|_2^2).
\]

The proof for Proposition 3.1 is given in Appendix A.
The second formula, which we refer to as the second order truncated perturbation formula, provides a second order approximation to the eigenvectors of $A$, as detailed in the following proposition.

**Proposition 3.2.** Let $A' \in \mathbb{R}^{n \times n}$ be a real symmetric matrix with $m$ leading eigenpairs $\{(t_i, v_i)\}_{i=1}^m$. Assume that $t_1 > t_2 > \cdots > t_m$. Let $E \in \mathbb{R}^{n \times n}$ be a real symmetric matrix. Let $A = A' + E$ be a perturbation of $A'$, and denote the $m$ leading eigenpairs of $A$ by $\{(s_i, w_i)\}_{i=1}^m$. Denote by $V^{(m)}$ the $n \times m$ matrix consisting of the $m$ leading eigenvectors of $A'$. Let $\mu \in \mathbb{R}$ and $1 \leq i \leq m$. Denote

$$r_i = \left(I - V^{(m)}V^{(m)T}\right)Ev_i. \quad (13)$$

Then, $w_i$ is approximated by the second order truncated perturbation formula

$$\tilde{w}_i^{(2)} = v_i + \sum_{k=1, k \neq i}^m \frac{(Ev_i, v_k)}{t_i - t_k}v_k + \frac{1}{t_i - \mu}r_i - \frac{\mu}{(t_i - \mu)^2}r_i + \frac{1}{(t_i - \mu)^2}A'r_i, \quad (14)$$

with an error satisfying

$$\left\| w_i - \tilde{w}_i^{(2)} \right\| \leq \sum_{k=m+1}^n \left| \frac{t_k - \mu}{t_i - t_m}\right|^2 \|E\|_2 + O(\|E\|_2^2). \quad (15)$$

The proof for Proposition 3.2 is given in Appendix B.

We note that formula (14) requires applying $A'$, and is computationally more expensive. We discuss in detail the runtime and memory requirements of formulas (11) and (14) in Appendix C.

The update formulas (11) and (14) depend on a parameter $\mu$, whose choice is discussed in [19]. 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 [19] and suggest to use

$$\mu_{\text{mean}} = \frac{\text{trace}(A') - \sum_{i=1}^m t_i}{n - m}, \quad (16)$$

which is the mean of the unknown eigenvalues.

We conclude this section by proving that under a certain assumption on $A'$, the first order truncated perturbation formula (11) and the second order truncated perturbation formula (14) are equal. Furthermore, in this case, the $O(\|E\|_2^2)$ term in the error bound of both approximations cancels out, as stated in the following proposition.

**Proposition 3.3.** 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 $A' = V^{(m)}TV^{(m)T} + \delta I$ for some diagonal matrix $T \in \mathbb{R}^{m \times m}$. Then, for $\mu = \delta$, the first order truncated perturbation formula (11) and the second order truncated perturbation formula (14) are equal, that is

$$\tilde{w}_i^{(1)} = \tilde{w}_i^{(2)}, \quad (17)$$

and the approximation errors satisfy

$$\left\| w_i - \tilde{w}_i^{(1)} \right\| = \left\| w_i - \tilde{w}_i^{(2)} \right\| = O(\|E\|_2^2), \quad (18)$$

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

6
The proof of Proposition 3.3 is given in Appendix D.

Corollary 3.4. If \( A' \) is of rank \( m \), and \( \mu = 0 \) in (11) and (14), 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 extension framework

In this section, we derive our perturbation based extension framework based on Proposition 3.1. Let \( K \in \mathbb{R}^{n \times n} \) be a symmetric positive semidefinite matrix whose \( m \) leading eigenpairs are denoted by \( \{(\lambda_i, u_i)\}_{i=1}^m \). 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 extension framework enables to "extend" the eigenvectors of any such \( K^s \) to the eigenvectors of \( K \), as follows.

Let \( \{(\lambda_i^s, u_i^s)\}_{i=1}^m \) be the leading eigenpairs of \( K^s \), let \( U^s(m) \in \mathbb{R}^{n \times m} \) be the matrix consisting of the \( m \) eigenvectors 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 3.1, the eigenvector \( u_i \) is approximated by

\[
\tilde{u}_i = u_i^s + \sum_{k=1,k \neq i}^m \frac{(K - K^s)u_i^s, u_k^s}{\lambda_i^s - \lambda_k^s} u_k^s + \frac{1}{\lambda_i^s - \mu} \left( I_m - U^s(m)U^s(m)^T \right) (K - K^s)u_i^s,
\]

with an error satisfying

\[
\|u_i - \tilde{u}_i\| \leq \sum_{k=m+1}^n \frac{|\lambda_k^s - \mu|}{|\lambda_i^s - \lambda_k^s|} |\lambda_i^s - \mu| \|K - K^s\|_2 + O(\|K - K^s\|_2^2).
\]

Furthermore, by (2), the eigenvalue \( \lambda_i \) is approximated by

\[
\tilde{\lambda}_i = \lambda_i^s + \lambda_i^s (K - K^s)u_i^s,
\]

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

Equations (19) and (21) are our perturbation based extension method. We will refer to this extension as the perturbation extension of the eigenpairs of \( K^s \) to the eigenpairs of \( K \). Note that this framework is quite general, and enables us to perform extensions over any symmetric sub-matrix of \( K \). We will propose and discuss several methods for choosing \( K^s \) in Section 6.

An extension framework analogous to (19) that is based on the second order approximation in Proposition 3.2 can also be obtained.

Our extension framework can also be used to obtain a low-rank approximation of the kernel matrix of the entire data \( K \). Analogously to the Nyström method [31], this low-rank kernel approximation is defined by

\[
\tilde{K}_{\text{pert}} = \sum_{i=1}^m \tilde{\lambda}_i \tilde{u}_i \tilde{u}_i^T.
\]
5 Equivalence with the Nyström method

In this section, we prove that our perturbation extension framework (19) is in fact a generalization of the Nyström method described in Section 2.2, by showing that the Nyström method arises from our extension 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 sample the first $m$ columns of $K$ to perform the Nyström extension, and denote by $\{(\hat{\lambda}_i, \hat{u}_i)\}_{i=1}^m$ the approximation to eigenpairs of $K$ obtained by the Nyström method as defined in (3) and (4). The following proposition states that a specific choice of the matrix $K^s$ for the perturbation extension (19) gives rise, up to a multiplicative constant, to the Nyström method eigenpairs defined above.

**Proposition 5.1.** 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 $\{(\hat{\lambda}_i, \hat{u}_i)\}_{i=1}^m$ the perturbation extension 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 $\{(\check{\lambda}_i, \check{u}_i)\}_{i=1}^m$ its eigenpairs. Denote by $\{(\tilde{\lambda}_i, \tilde{u}_i)\}_{i=1}^m$ the Nyström extension of $\{(\lambda_i^s, u_i^s))_{i=1}^m$ (see Section 2.2). Then,

$$\hat{u}_i = \sqrt{\frac{m}{n}} \check{u}_i \quad \text{and} \quad \hat{\lambda}_i = \frac{n}{m} \check{\lambda}_i \quad (23)$$

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

The proof of Proposition 5.1 is given in Appendix E.

The formulation of the Nyström method as a perturbation based extension using Proposition 5.1 enables us to provide an error analysis based on Propositions 3.1 and 3.2. Contrary to previous works that provide an error bound for the kernel approximation itself, our error analysis is for the individual approximated eigenvectors, as stated in the following proposition.

**Proposition 5.2 (Vector-wise error for the Nyström method).** Using the above notation, the error induced by the Nyström method satisfies

$$\|u_i - \hat{u}_i\| = O(\|K - K^s\|_F^2), \quad 1 \leq i \leq m. \quad (24)$$

Figure 1: Illustration of the submatrix $K^s$. Blank entries indicate 0.
Proof. Follows directly from the equivalence stated in Proposition 5.1 by noting that in the Nyström method settings, the requirements of Corollary 3.4 hold.

6 New extensions based on the perturbation framework

The perturbation extension framework derived in Section 4 allows for various extensions that depend on the choice of the matrix $K^s$. In this section, we propose several types of extensions corresponding to different choices of $K^s$. The key idea is to choose a matrix $K^s$ whose eigendecomposition is “easy” to compute, in order to make the extension computationally attractive. In addition, we prove that similarly to the classical Nyström method, the spectral shifted Nyström method and the ensemble Nyström method described in Section 2 are in fact special cases of our general extension framework, for suitable choices of $K^s$.

6.1 $\mu$-shifted extension

In this type of extension, we choose the matrix $K^s$ to be the top left $m \times m$ submatrix of $K$ padded with zeros, similarly to the Nyström method (see Figure 2). The difference from the Nyström method lies in the parameter $\mu$ of (19). In Proposition 5.1, we used the value of the parameter $\mu$ to be $\mu = 0$. This might be 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 $\mu$ that approximates the unknown eigenvalues of $K$. A reasonable choice for $\mu$ in such a case is $\mu_{\text{mean}}$ of (16).

We now prove that given a parameter $\mu \geq 0$, the spectral shifted Nyström method with parameter $\mu$ coincides with the perturbation extension method with the same $\mu$, as detailed in the following proposition.

Proposition 6.1. Using the notation of this section, the eigenpairs approximated by the spectrum shifted Nyström method and the $\mu$-shifted extension method are equal.

The proof of Proposition 6.1 is given in Appendix F.

The runtime complexity of the $\mu$-shifted extension is the same as of the Nyström method, that is $O(nm^2 + m^3)$.

6.2 Block diagonal extension

In this type of extension, we choose $K^s$ 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, and then extend them using (19). Denote by $\{(\tilde{\lambda}_i^{(j)}, \tilde{\mathbf{u}}_i^{(j)})\}_{i=1}^m$ the eigenpairs extension of block $j$, and by $\tilde{K}_j \in \mathbb{R}^{n \times n}$ the resulting kernel approximation. To combine the $k$ approximations $\{\tilde{K}_j\}_{j=1}^k$, one might use a weighted mean, that is

$$\tilde{K} = \sum_{i=1}^k \mu_i \tilde{K}_i.$$  \hfill (25)
It easily follows from Proposition 5.1 that the block diagonal extension is identical to the ensemble Nyström method. The runtime complexity of this method is $O(k(nm^2 + lm^2))$. However, the eigendecomposition of the blocks can be done in parallel.

### 6.3 $p$-band extension

In this type of extension, we choose $K^*$ to be a band matrix of width $p$ (see Figure 2). This extension may provide superior results when the kernel $K$ has most of its energy concentrated around the diagonal. We demonstrate the advantage of this extension method in Section 7.2.1.

![Figure 2: Illustration of the submatrix $K^*$ for each of the discussed extensions. Blank entries indicate 0.](image)

### 6.4 Sparse extension

In this type of extension, we assume that the kernel matrix $K$ is sparse, and choose $K^*$ to be some sparse submatrix of it, as illustrated in Figure 3. More concretely, denoting by $\text{nnz}(K)$ the number of non-zero entries of $K$, in the sparse extension framework, we need to choose $q \cdot \text{nnz}(K)$ entries of $K$ to define $K^*$, for some $0 < q \leq 1$. While this extension can be applied to any such subset, motivated by the $\|E\|_2$ term in the error bounds (12) and (15), we suggest to choose the $q \cdot \text{nnz}(K)$ largest entries of $K$. We demonstrate the advantage of this extension 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 Propositions 3.1 and 3.2. Then, we demonstrate the advantages of the extensions proposed in Section 6 for both real and synthetic datasets.

#### 7.1 Perturbation error bounds

In this section, we demonstrate numerically the behavior of the error bounds in Propositions 3.1 and 3.2. In our first example, we demonstrate the linear dependence of the error on the norm of the matrix $E$ in
Figure 3: Illustration of a sparse extension. Blank entries indicate 0.

Propositions 3.1 and 3.2. To that end, we generate a random symmetric matrix $A'$, normalize it to have a unit norm, and then calculate its 10 leading eigenpairs. We then generate a random symmetric matrix $E$ and normalize it to have a unit 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 (11) and (14) using $\mu = 0$. Denote by $v_c$ the leading eigenvector of $A_c$, and by $u^1_c$ and $u^2_c$ its approximations by (11) and (14), respectively. For each $c$, we measure the errors $\|v_c - u^1_c\|$ and $\|v_c - u^2_c\|$. In Figure 4a, we plot $\log \|v_c - u^1_c\|$ and $\log \|v_c - u^2_c\|$ versus $\log \|cE\|$. As predicted by theory, there is a linear dependence between the error in the eigenvector approximation and the norm of the perturbation matrix. Furthermore, the errors achieved by the first and second order formulas are comparable since the dominant term in (12) and (15) is the $\|E\|$ term.

In our second example, we demonstrate the linear and quadratic dependence of the error on the $\sum_{j=m+1}^n |\lambda_j - \mu|$ term. We generate a random symmetric matrix $A'$ of rank 10, so that its 10 leading eigenvalues are between 1 and 2. We then generate a random symmetric matrix $E$ 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 $\sum_{j=m+1}^n |\lambda_j - \mu|$. Then, for various values of $c$, we generate a matrix $A'_c$, whose leading 10 eigenvalues are the same as of $A'$, and the rest are exactly $c$. We approximate the 10 leading eigenpairs of $A_c = A'_c + E$ by the first and second order approximations (11) and (14) using $\mu = 0$, and measure the error in the same way as in the previous example. In Figure 4b we plot $\log \|v_c - u^1_c\|$ and $\log \|v_c - u^2_c\|$ versus $\log |\lambda_j - \mu| = \log c$. As predicted by theory, 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.

7.2 Perturbation extension for synthetic and real-world data

In this section, we demonstrate the advantages of the various extension methods proposed in Section 6 for both synthetic and real-world data. We do not include in this section the $\mu$-shifted extension and the block diagonal extension, as they were proven to be identical to variants of the Nyström method that were already discussed in the literature (see [16, 29]).

The kernel functions we use in this section are the Gaussian kernel, resulting in a kernel matrix whose
(a) Dependence on $\|E\|$.

(b) Dependence on $\sum_{j=m+1}^{n} |\lambda_j - \mu|$.

Figure 4: Numerical demonstration of the error terms in the approximations (11) and (14). (a) log (error) vs. log $\|cE\|$. The slope of both curves is 1, demonstrating the linear dependence of the error terms (12) and (15) on $\|E\|$ for both the first and second order approximations. (b) log (error) vs. log $c$. The slope of the linear curve of the first order approximation is 1, whereas the slope of the linear curve of the second order approximation is 2, demonstrating the linear and quadratic dependence of the first and second order error terms on $\sum_{j=m+1}^{n} |\lambda_j - \mu|$, respectively.

| 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.

$(i, j)$ entry is $\exp(-\gamma \|x_i - x_j\|_2^2)$ for some parameter $\gamma > 0$, and the polynomial kernel, resulting in the kernel matrix whose $(i, j)$ entry is $(1 + x_i^T x_j)^d$ for some integer $d$.

As our metric for comparing the performance of the various methods we use the principal angle [15] between the exact subspace spanned by the kernel’s top eigenvectors and the subspace spanned by their approximations.

The real-world datasets we use are taken from the UCI Machine Learning Repository [10] and are described in Table 1.

7.2.1 $p$-band extension

To demonstrate the advantage of the $p$-band extension method, we generate a matrix $K$ whose $(i, j)$ and $(j, i)$ entries are $X^{-\frac{1}{p+1}}$, where $X$ is drawn uniformly between 0 and 1 (see Figure 5a). We also set entries of $K$ that are smaller than $10^{-10}$ to 0. Then, for various values of $p$, we compute a $p$-
band extension with $m = 10$, and compare the approximation error of the $p$-band extension to the best rank 10 approximation of $K$ using the $\| \cdot \|_2$. For comparison, we also compute several Nyström extensions for several values of $l$ and measure their error in the same way. We repeat this experiment 20 times. In Figure 5b, we plot the approximation error of each repetition versus the percentage of non-zero entries selected in $K^s$ out of the total number of non-zero entries of $K$. We can see that the error graphs of the the $p$-band extension decay to 0 much faster than the error graphs of the partial-eigenspace extension. We conclude that when the kernel admits the structure discussed in this section, the $p$-band extension results in superior performance and converges to the optimal solution much faster than the Nyström extension.

### 7.2.2 Sparse extension

To demonstrate the advantage of the sparse extension, we randomly choose $n = 1000$ points from each tested dataset, and normalize each of its features to have 0 mean and unit variance. We then calculate the corresponding kernel matrices using a Gaussian kernel, a linear kernel, and a quadratic kernel. To sparsify the kernel, we set its 90% smallest entries to 0. We then compute several sparse extensions for various values of $q$ with $m$ depending on the dataset and kernel chosen. We compare the error of the sparse extensions to best $m$-rank approximation of $K$ using the $\| \cdot \|_2$. For comparison, we also compute several Nyström extensions for several values of $l$, and measure their error in the same way. We repeat this experiment 20 times. We then plot the approximation error of each extension method versus the percentage of non-zero entries selected in $K^s$ out of the total number of non-zero entries of $K$. The results of this procedure are shown in Figure 6 for the MNIST dataset, in Figure 7 for the superconductivity dataset, in Figure 8 for the poker dataset and in Figure 9 for the wine quality dataset. We can see that in most scenarios, the error graphs of the the sparse extension decay to 0 much faster than the error graphs of the partial-eigenspace extension in most repetitions of the experiments. Additionally, the error decay of the sparse extension has smaller variance. We conclude that when the kernel is sparse, the sparse extension usually provides superior performance, and convergences to the optimal solution much faster than the Nyström extension.
Figure 6: Extension of the MNIST dataset.

(a) Gaussian kernel
\( (\gamma = 0.01, m = 5) \)

(b) Linear kernel
\( (m = 20) \)

(c) Quadratic kernel
\( (m = 5) \)

Figure 7: Extension of the superconductivity dataset.

(a) Gaussian kernel
\( (\gamma = 0.1, m = 5) \)

(b) Linear kernel
\( (m = 10) \)

(c) Quadratic kernel
\( (m = 10) \)

Figure 8: Extension of the poker dataset.

(a) Gaussian kernel
\( (\gamma = 0.1, m = 5) \)

(b) Linear kernel
\( (m = 10) \)

(c) Quadratic kernel
\( (m = 5) \)
8 Summary

In this paper, we propose an eigenvectors extension 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 extension is used as part of a dimensionality reduction procedure. Our extension framework is quite 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 around the diagonal.

Acknowledgements

This research was supported by the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement 723991 - CRYOMATH).

References

[1] Yariv Aizenbud, Amit Bermanis, and Amir Averbuch. Pca-based out-of-sample extension for dimensionality reduction. arXiv preprint arXiv:1511.00831, 2015.

[2] 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.

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

[4] Yoshua Bengio, Jean-françois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas L Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In Advances in neural information processing systems, pages 177–184, 2004.

[5] Matthew Brand. Fast low-rank modifications of the thin singular value decomposition. Linear algebra and its applications, 415(1):20–30, 2006.
[6] 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.

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

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

[9] Petros Drineas and Michael W Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of machine learning research*, 6(Dec):2153–2175, 2005.

[10] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[11] Alex Gittens. The spectral norm error of the naive nystrom extension. *arXiv preprint arXiv:1110.5305*, 2011.

[12] 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, 2011.

[13] 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, 2011.

[14] Aren Jansen, Gregory Sell, and Vince Lyzinski. Scalable out-of-sample extension of graph embeddings using deep neural networks. *Pattern Recognition Letters*, 94:1–6, 2017.

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

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

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

[18] 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.

[19] 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.

[20] HyungSeon Oh and Zhe Hu. Multiple-rank modification of symmetric eigenvalue problem. *MethodsX*, 5:103–117, 2018.
Appendix A  Proof of Proposition 3.1

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

\[
\tilde{v}_i = v_i + \sum_{k=1, k \neq i}^{m} \frac{(Ev_i, v_k)}{t_i - t_k} v_k + \sum_{k=m+1}^{n} \frac{(Ev_i, v_k)}{t_i - t_k} v_k. \tag{26}
\]

As the second term in (26) is unknown, we approximate it by replacing the unknown eigenvalues with a parameter \( \mu \),

\[
\sum_{k=m+1}^{n} \frac{(Ev_i, v_k)}{t_i - \mu} v_k = \frac{1}{t_i - \mu} \sum_{k=m+1}^{n} (Ev_i, v_k) v_k = \frac{1}{t_i - \mu} (Ev_i - V^{(m)} V^{(m)T} E v_i) = \frac{1}{t_i - \mu} r_i. \tag{27}
\]
where \( r_i \) is defined in (10). Formula (11) follows by replacing the second term in (26) with (27). We denote the approximation error introduced into the approximation (11) by \( e_i \), that is

\[
e_i = \left\| \sum_{k=m+1}^{n} \left( \frac{(Ev_i, v_k)}{t_i-t_k}v_k - \frac{1}{t_i-\mu}r_i \right) \right\|.
\]

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

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

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

\[
e_i \leq \sum_{k=m+1}^{n} \frac{|t_k-\mu|}{|t_i-t_k||t_i-\mu|} |(Ev_i, v_k)| \leq \frac{\|E\|}{|t_i-t_m||t_i-\mu|} \sum_{k=m+1}^{n} |t_k-\mu|.
\]

Recalling that the original perturbation approximation (1) induces an error of \( O(\|E\|_2^2) \), concludes the proof.

\[\square\]

Appendix B  Proof of Proposition 3.2

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

\[
\tilde{v}_i = v_i + \sum_{k=1, k \neq i}^{m} \frac{(Ev_i, v_k)}{t_i-t_k}v_k + \sum_{k=m+1}^{n} \frac{(Ev_i, v_k)}{t_i-t_k}v_k.
\]

To obtain Equation (14), we replace the unknown (second) term in (29) by using the identity

\[
\frac{1}{t_i-t_k} = \frac{1}{t_i-\mu} + \frac{t_k-\mu}{(t_i-\mu)^2} + \frac{(t_k-\mu)^2}{(t_i-t_k)(t_i-\mu)^2},
\]

and noting that

\[
\sum_{k=m+1}^{n} (Ev_i, v_k)(t_k-\mu)v_k = \sum_{k=m+1}^{n} (Ev_i, v_k)t_kv_k - \mu \sum_{k=m+1}^{n} (Ev_i, v_k)v_k
\]

\[
= \sum_{k=m+1}^{n} (Ev_i, v_k)A'r_k - \mu \sum_{k=m+1}^{n} (Ev_i, v_k)v_k
\]

\[
= A'r_i - \mu r_i,
\]

where \( r_i \) is defined in (13).
We denote the approximation error introduced into the approximation (14) by $e_i$, that is

$$e_i = \| \sum_{k=m+1}^{n} \left( Ev_i, v_k \right) v_k - \left( \frac{1}{t_i - \mu} r_i - \frac{\mu}{(t_i - \mu)^2} r_i + \frac{1}{(t_i - \mu)^2} A' r_i \right) \|_.$$  \(31\)

Using (30) and the triangle and the Cauchy-Schwarz inequalities, we obtain

$$e_i = \left\| \sum_{k=m+1}^{n} \frac{(t_k - \mu)^2}{(t_i - t_k)(t_i - \mu)^2} (Ev_i, v_k) v_k \right\| \leq \sum_{k=m+1}^{n} \frac{|t_k - \mu|^2}{|t_i - t_k||t_i - \mu|^2} \left| (Ev_i, v_k) \right| \leq \frac{\|E\|}{|t_i - t_m||t_i - \mu|^2} \sum_{k=m+1}^{n} |t_k - \mu|^2. \quad (33)$$

Recalling that the original perturbation approximation (1) induces an error of $O(\|E\|_2^2)$ concludes the proof. \(\square\)

### Appendix C  Runtime and space complexity

In this section, we discuss the runtime and space complexity of the truncated perturbation formulas (11) and (14). The first order formula needs to store in memory the $m$ eigenvectors of $A'$ resulting in $O(mn)$ space. The computation of $r_i$ of (10) involves the calculation of $Ev_i$ that requires $O(nnz(E))$. 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 complexity for all $\{r_i\}_{i=1}^{m}$ is $O(m \cdot nnz(E) + m^2 n)$ operations. The first order formula also requires the computation of all $O(m^2)$ terms of the form $(Ev_i, v_k)v_k$ for $1 \leq i \neq k \leq m$. Each such term requires $O(nnz(E) + n)$ operations, resulting in a total of $O(m^2 \cdot nnz(E) + m^2 n)$ operations for all eigenvectors. We conclude that the first order formula requires a total of $O(m^2 \cdot nnz(E) + m^2 n)$ operations.

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

### Appendix D  Proof of Proposition 3.3

Let $i \in \{1, \ldots, m\}$. Since $\mu = \delta$, we note that if $A'r_i = \delta r_i$ then the last two terms in (14) cancel out and (14) is reduced to the first order formula (11). Thus, in order to prove that $w_i^{(1)} = w_i^{(2)}$, it is sufficient
to prove that under the proposition settings, $A'r_i = \delta r_i$. Indeed,

$$A'r_i = A'(I - V^{(m)}V^{(m)T})Ev_i = \left(V^{(m)}TV^{(m)T} + \delta I\right)(I - V^{(m)}V^{(m)T})Ev_i$$

$$= \left(V^{(m)}TV^{(m)T} - V^{(m)}TV^{(m)T} + \delta I - \delta V^{(m)}V^{(m)T}\right)Ev_i$$

$$= \delta(I - V^{(m)}V^{(m)T})Ev_i = \delta r_i.$$

For the error, we note that the $n - m$ unknown eigenvalues of $A'$ of the form $A' = V^{(m)}TV^{(m)T} + \delta I$ are exactly $\delta$, and thus, when choosing $\mu = \delta$ the first term in (12) and (15) cancels out and we are left with only the $O(\|E\|^2)$ term. \hfill \square

### Appendix E Proof of Proposition 5.1

Let $i \in \{1, \ldots, m\}$. By (19), the eigenvectors perturbation extension for $\mu = 0$ reads

$$\tilde{u}_i = u_i^s + \sum_{k=1, k\neq i}^{m} \frac{(K - K^s)u^s_k, u^s_k}{\lambda^s_i - \lambda^s_k}u^s_k + \frac{1}{\lambda^s_i}(I - U^{s(m)}U^{s(m)T})(K - K^s)u^s_i, \quad (35)$$

and by formula (21), the eigenvalues update reads

$$\tilde{\lambda}_i = \lambda^s_i + u^s_i(K - K^s)u^s_i. \quad (36)$$

We first prove that $\hat{u}_i = \sqrt{\frac{n}{m}} \tilde{u}_i$ (see (23)). We start by simplifying (35) 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 $u^s_i$ are 0, and the top left $m \times m$ submatrix of $K - K^s$ is 0. That implies that the first $m$ coordinates of $(K - K^s)u^s_i$ are 0, and so $((K - K^s)u^s_i, u^s_k) = 0$ for all $1 \leq i, k \leq m$. A direct consequence of the latter is that $U^{s(m)}U^{s(m)T}(K - K^s)u^s_i = 0$ for all $1 \leq i \leq m$. Thus, (35) reduces to

$$\tilde{u}_i = u_i^s + \frac{1}{\lambda^s_i}(K - K^s)u^s_i. \quad (37)$$

Next, we note that the first term in (37), $u_i^s$, is non-zero only on its first $m$ entries, whereas the second term, $\frac{1}{\lambda^s_i}(K - K^s)u^s_i$ is non-zero only on its last $n - m$ entries. This means that the first $m$ entries of $\tilde{u}_i$ are exactly the first $m$ entries $u^s_i$ and the last $n - m$ entries of $\tilde{u}_i$ are determined by $\frac{1}{\lambda^s_i}(K - K^s)u^s_i$. We start by proving the equivalence between the first $m$ coordinates of $\tilde{u}_i$ and $\hat{u}_i$. Let $1 \leq p \leq m$. Denote by $A_{p \rightarrow}$ 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 extension (see (3)),

$$\hat{u}_{i,p} = \sqrt{\frac{m}{n}} \frac{1}{\lambda_i} C_{p \rightarrow}u^s_i = \sqrt{\frac{m}{n}} \frac{1}{\lambda_i} K^p_{\rightarrow}u^s_i = \sqrt{\frac{m}{n}} \frac{1}{\lambda_i} \lambda_i u^s_i = \sqrt{\frac{m}{n}} u^s_i, \quad p = 1.$$  

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

Next, we prove the equivalence between the last $n - m$ entries of $\tilde{u}_i$ and $\hat{u}_i$. Let $m + 1 \leq p \leq n$. We
have for the Nyström extension (see \(3\))

\[ \hat{u}_{i,p} = \sqrt{\frac{m}{n}} \frac{1}{\lambda_i} C_{p \rightarrow} u_i', \]  

(38)

and for the perturbation extension, by using \(37\), and since the last \(n - m\) entries of \(u_i^s\) are 0,

\[ \tilde{u}_{i,p} = \frac{1}{\lambda_i} (K - K^s)_{p \rightarrow} u_i^s = \frac{1}{\lambda_i} (K_{p \rightarrow} u_i^s - 0) = \frac{1}{\lambda_i} C_{p \rightarrow} u_i', \]  

(39)

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 \(1 \leq i \leq m\), we conclude by \(38\) and \(39\) that \(\hat{u}_{i,p} = \sqrt{\frac{m}{n}} \tilde{u}_i\) for \(m + 1 \leq p \leq n\), and thus, the last \(n - m\) entries of \(\hat{u}_i\) and \(\sqrt{\frac{m}{n}} \tilde{u}_i\) are also identical.

We now prove the eigenvalues equivalence. By the same arguments as above, we note that for all \(1 \leq i \leq m\), \(u_i^sT (K - K^s) u_i^s = 0\). Thus, by \(36\) we have \(\tilde{\lambda}_i = \lambda_i^s\) and consequently, using \(4\),

\[ \hat{\lambda}_i = \frac{n}{m} \lambda_i' = \frac{n}{m} \lambda_i^s = \frac{n}{m} \tilde{\lambda}_i, \]  

(40)

as required.

\(\square\)

Appendix F  Proof of Proposition 6.1

Denote by \(K^s_{\text{shift}} \in \mathbb{R}^{n \times n}\) the top left \(m \times m\) submatrix of \(K_{\text{shift}}\) (see \(6\)) padded with zeros. By the equivalence of the Nyström extension and the perturbation extension proved in Proposition 5.1, the spectral shifted Nyström extension is equivalent to the perturbation extension of \(K^s_{\text{shift}}\) to the eigenpairs of \(K_{\text{shift}}\) with \(\mu = 0\). Thus, it suffices to prove that the perturbation extension of the eigenpairs of \(K^s_{\text{shift}}\) to the eigenpairs of \(K^s_{\text{shift}}\) with \(\mu = 0\) equals to the perturbation extension of the eigenpairs of \(K^s_{\text{shift}}\) to the eigenpairs of \(K_{\text{shift}}\) 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 \(K^s_{\text{shift}}\) are \(\{(\lambda_i^s - \delta, u_i^s)\}_{i=1}^m\). Let \(1 \leq i \leq m\). By \(37\), the perturbation extension \(19\) of the eigenvectors of \(K^s_{\text{shift}}\) to eigenvectors of \(K_{\text{shift}}\) with \(\mu = 0\) reduces to

\[ \tilde{u}_i = u_i^s + \frac{1}{\lambda_i - \delta - 0} (K_{\text{shift}} - K^s_{\text{shift}}) u_i^s, \]

whereas the perturbation extension 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 \(K_{\text{shift}} - K^s_{\text{shift}}\) are equal to those of \(K - K^s\) except for the last \(n - m\) diagonal elements, we have that

\[ (K_{\text{shift}} - K^s_{\text{shift}}) u_i^s = (K - K^s) u_i^s, \]  

(41)

and conclude that \(\tilde{u}_i = \hat{u}_i\).
For the eigenvalues, the perturbation extension of the eigenvalues of $K^s_{\text{shift}}$ to the eigenvalues of $K_{\text{shift}}$ (see [21]) gives

$$\tilde{\lambda}_i = (\lambda^s_i - \delta) + u^s_i (K_{\text{shift}} - K^s_{\text{shift}}) u^s_i.$$ 

Since the eigenvalues of $K_{\text{shift}}$ are shifted by $\delta$, in order recover the eigenvalues of $K$ we need to shift the extended eigenvalues $\{\tilde{\lambda}_i\}_{i=1}^m$ back by $\delta$ (similarly to the spectral shifted Nyström method, see [2.2]), and thus we obtain

$$\tilde{\lambda}_i = \lambda^s_i + u^s_i (K_{\text{shift}} - K^s_{\text{shift}}) u^s_i.$$ 

On the other hand, the perturbation extension of $K^s$ yields

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

By (41), we have that $\tilde{\lambda}_i = \hat{\lambda}_i$. 

$\square$