Large-scale eigenvector approximation via Hilbert Space Embedding Nyström

Ming Lin\textsuperscript{a}, Fei Wang\textsuperscript{b}, Changshui Zhang\textsuperscript{a,}\textsuperscript{*}

\textsuperscript{a}Department of Automation, Tsinghua University, FIT Building 3-120, Beijing 100084, PR China
\textsuperscript{b}IBM T.J. Watson Research Center, New York, USA

\textbf{A R T I C L E   I N F O}

Article history:
Received 31 July 2013
Received in revised form 12 September 2014
Accepted 24 November 2014
Available online 2 December 2014

Keywords:
Eigenvectors and eigenfunctions
Sampling methods
Spectral analysis

\textbf{A B S T R A C T}

The Nyström method approximates eigenvectors of a given kernel matrix by randomly sampling subset of data. Previous researches focus on good kernel approximation while the quality of eigenvector approximation is rarely explored. In online eigenvector approximation method, one can minimize the kernel approximation error to guarantee a good eigenvector approximation. However in this work, we paradoxically prove that for batch approximation methods like Nyström, it is no longer true. This unexpected discovery opens a question: What criterion should we use in Nyström to generate a decent eigenvector approximation? To address this problem, we propose a novel criterion named Hilbert Space Embedding (HSE) Nyström criterion which directly minimizes the eigenvector approximation error. The proposed HSE criterion provides a general framework to approximate eigenvectors within linear time and space complexity. We then show that we can rediscover many successful Nyström methods with the proposed criterion, including K-means Nyström and Density Nyström. To further demonstrate the power of our criterion, we actually design a novel algorithm to approximate eigenvectors of Laplacian matrices based on the proposed criterion with better accuracy among existing linear complexity methods. We demonstrate the efficiency and efficacy of our proposal in numerical experiments.

\textcopyright 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Computing eigenvectors of a given kernel matrix is a fundamental problem and has aroused considerable interests from machine learning and data mining fields in recent decades. Many algorithms utilize the eigenvalue decomposition as their core step, such as spectral clustering [1] and eigenmap embedding [2,3]. Though these methods are theoretically powerful, their high computational and storage cost in the spectral decomposition step limits their practical applications. Given an $N \times N$ matrix, its spectral decomposition needs up to $O(N^3)$ time complexity and $O(N^2)$ space complexity. Thus seeking solutions for spectral decomposition on large-scale datasets is a very challenging task. The Nyström method [4–6] is one of the most popular ways to address this problem effectively.

The Nyström method is first proposed in 1931 to compute eigenfunctions of integral equations [7]. Then it is introduced in machine learning to speed up kernel machine training in 2001 [8]. The motivation of the Nyström method is to approximate the eigenvectors of a given kernel matrix using a subset of data samples. By sampling, the Nyström methods could (1) reduce the time/space complexity, (2) find representative points for the original eigensystem.

While the former goal is obviously important, the latter goal is of equal importance if we also want to find a compact representation of original dataset. The standard Nyström method samples uniformly on each column. Drineas and Mahoney [9] propose to sample the columns under the probability proportional to column norm. Ouimit and Bengio [10] propose to sample the columns with a greedy strategy that minimizes the mean projection distance between the data and the subspace spanned by the sampled columns. Zhang et al. [11] choose K-means centers as sampling points in the Nyström method.

Although there are extensive works, most existing researches focus on the Frobenius norm or spectral norm of kernel matrix approximation error to measure the approximation quality of the Nyström methods [9–14]. The norm of kernel matrix approximation error is relevant to the generalization error of kernel machine [15]. Albeit these researches are important in efficient kernel learning, it is quite a different story if our target is just the eigenvectors itself. To our surprise, in the Nyström method, a small error in kernel matrix approximation does not imply small error in eigenvector approximation. Bengio et al. [16] propose a theorem which states that online minimizing the kernel approximation error is necessary and sufficient for approximating eigenvectors. Unfortunately, this theorem does not hold true in the Nyström methods. The reason for the invalidity of Proposition 3 in [16] is that the Nyström method is a batch approximation method rather than an online one. In fact, we found that in eigenmap embedding with Gaussian kernel function, we need...
to choose a different kernel width in the Nyström method to get an accurate eigenvector approximation. This unexpected discovery opens a question: What criterion should we use in the Nyström method to guarantee a decent eigenvector approximation? To address this question, we propose a novel criterion named Hilbert Space Embedding (HSE) Nyström criterion. The proposed criterion provides a general framework to design eigenvector approximation methods of linear time and space complexity. Many successful Nyström methods can be rediscovered by our criterion, including K-means Nyström and Density Nyström. In our criterion, both the average weights and kernel width are adaptively optimized. The proposed criterion is closely related to the kernel density estimation, in which the kernel width also need to be adaptively chosen based on the sampled data. To further demonstrate the power of our criterion, we innovate a novel algorithm named Hilbert Space Embedding Nyström based on our criterion, which achieves better eigenvector approximation accuracy within linear time complexity.

The rest of our paper is organized as follows. In Section 2 we introduce our notations. In Section 3 we rehearse Proposition 3 in [16] as Theorem 1, then demonstrate that the online eigenvector approximation based on this theorem cannot be applied on batch-based Nyström method. In Section 4, we propose our HSE Nyström criterion. Then we apply our criterion in eigenmap embedding problems to approximate eigenvectors of Laplacian matrices. We reveal the connection between our criterion and kernel density estimation at the end of the theoretical part. Experiments in Section 5 show that the method derived from our criterion outperforms the state-of-the-art methods with better accuracy within linear complexity.

2. Notations for the Nyström method

Let $\kappa(x, y)$ denote kernel function defined on $x, y \in X$. Given $N$ data points $X = \{x_1, x_2, ..., x_N\}$ drawn identically and independently distributed (i.i.d.) from probability measure $P(x)$, the Nyström method [7] numerically approximates eigenvectors of kernel matrix $K$

$$\frac{1}{N}Kf = \lambda f, \quad (Kf)_{ij} = \kappa(x_i, x_j),$$

(1)

where $(Kf)_{ij}$ is the $i$th row, $j$th column element of kernel matrix $K$, $f$ is an $N \times 1$ eigenvector with unit norm, and $\lambda$ is the corresponding eigenvalue. The standard Nyström method randomly draws $M$ samples $S_M = \{s_1, s_2, ..., s_M\}$ from $X$, then solves following equation instead:

$$\frac{1}{M}K_Mv_t = \lambda_Mv_t, \quad (K_Mv_t)_{ij} \approx \kappa(s_i, s_j),$$

(2)

where $t = 1, 2, ..., M$. $(v_t)_{i,j}^{M-1}$ denotes $M$ orthogonal eigenvectors. The non-normalized approximated eigenvectors of $K$ can be obtained via the Nyström extension formula:

$$\hat{f}_t = \frac{1}{M_{M,t}}K_Mv_t, \quad (K_Mv_t)_{ij} \approx \kappa(s_i, s_j).$$

(3)

The subscript $[x, s]$ indicates that the row elements of kernel matrix $K_{x,s}$ are from set $X$, the column elements are from set $S_M$. The corresponding normalized eigenvector is

$$\hat{f}_t = \frac{\hat{f}_t}{\|\hat{f}_t\|}, \quad \|\hat{f}_t\| \approx \sqrt{\sum_j |(f_j)_{i,j}^2| / N}.$$  

(4)

Typically $N$ is significantly larger than $M$, therefore the Nyström method can save considerable computational cost from $O(N^2)$ to $O(M^2)$, and save the storage space by only keeping part of kernel matrix in memory. The approximation error of eigenvector is $\varepsilon_{\hat{f}_t} = \|\hat{f} - f\|$. The kernel matrix $K$ is approximated by $\hat{K} = K_{x,s}K_{s,M}^T$, where $K_{x,s}^T$ denotes Moore-Penrose pseudo-inverse of $K_M$. The approximation error of kernel matrix is $\varepsilon_{\hat{K} - K} = \|K - \hat{K}\|^2$.

3. Online and batch eigenvector approximation

It is widely known that in online method, minimizing kernel approximation error is equal to minimizing eigenvector approximation error. This is first proved by [16]. However, in a batch-based method such as the Nyström method, this is no longer true. In this section, we first give a counter example to verify our argument, then prove our statement theoretically. For the online-based eigenvector approximation method, the following theorem has been known for a long time [16]:

**Theorem 1.** Assuming $[f(x)]_{i=1}^{M-1}$ are the $M-1$ eigenfunctions of $K$ with corresponding eigenvalues $\lambda_k$, and are already obtained, then $f_M(x)$ is the eigenfunction of $K$ corresponding to eigenvalue $\lambda_M$ if and only if it minimizes $\varepsilon_M = \int \int (f_M(x) - \lambda_Mf(x))dP(x)dP(y)$, where

$L(x, y) = \kappa(x, y) - \lambda_Mf(x)f(y) - \sum_{i=1}^{M-1} \lambda_i f_i(x)f_i(y).$  

The discrete form of this theorem by substituting integral with summation still holds.

To compute the $M$th eigenvector, Theorem 1 requires to provide the exact very first $M-1$ eigenvectors. But what we do in the Nyström method is a batch manner—Eq. (3) obtains $M$ vectors together. Thus there is no guarantee that any of these $M$ vectors is an exact eigenvector of $K$. Therefore Theorem 1 is not applicable to the Nyström methods. We give a counter example here, in which\footnote{Note that $f_i$ can be very large while keeping $\varepsilon_K = 0$. Let $X \in \mathbb{R}^d$, and all $x \in X$ lie in a two dimensional subspace $H = \{[x_1, x_2] : x_1 + x_2 = 0\}$. $x_i$’s are drawn from a mixture distribution of two Gaussian distributions of unit variance on $H$ with means $\mu_1 = \frac{\sqrt{2}}{2}0$, $-\frac{\sqrt{2}}{2}0$, $\mu_2 = \frac{\sqrt{2}}{2}0$, $\frac{\sqrt{2}}{2}0$. We sample 2000 points from the distribution, then choose two arbitrary linear independent points on $H$ as the sampling set $S_M = \{s_1, s_2\}$, for example, $s_1 = [3.5355, 5, -3.5355], s_2 = [3.5355, -5, -3.5355]$. We repeat the experiment 100 times. In this counter example, $\varepsilon_{K - K} = 0 \pm 0$ but $\varepsilon_{\hat{f}_t - f} \approx 0.83 \pm 0.3$. If $f$ is of unit norm.

We now give some theoretical analysis on this phenomenon. To make things clear, we emphasize some subtle concepts by the following definition:

**Definition 1.** $\hat{K}$ is called "reconstructed matrix". $\hat{f}$ is called "Nyström vector". $f_M$ is the eigenvector of $K$, and is called "reconstructed eigenvector". We say $f$ is "incompetent" if $\|f - f_\hat{}\| \geq 0.5$. Note that $f$ is NOT eigenvector of $K$.

In the Nyström method, we obtain "Nyström vector" $\hat{f}$ directly by Eq. (4) using $S_M$, and reconstruct $K$ from $f$. Decomposing "reconstructed matrix" $\hat{K}$ will obtain "reconstructed eigenvector" $f_M$. If $K = K$, their eigenvectors are equal, i.e., $f = f_M$. But with different $S_M$ sampled, we will reconstruct different $f$. Therefore we can find infinite groups of $f$, and all of them can reconstruct the same $K$ exactly equal to $K$. Part of these $f$ are incompetent Nyström vectors. This shows that $\hat{K} = K$ only implies the "reconstructed eigenvector" $f_M$. NOT Nyström vector $f$, is equal to the eigenvector of $K$. Thus neither $\|f_M - f\|$ nor $K - K$ in this case is meaningful when concerning $f$. We summarize our analysis in Proposition 1:

**Proposition 1.** Let $X = \{x_1, x_2, ..., x_N\} \in \mathbb{R}^d$, $\kappa(x, y) = \langle x, y \rangle = \langle x, y \rangle$ is an inner product kernel, $S_M = \{s_1, s_2, ..., s_M\}$ is sampled from $\mathbb{R}^d$ (not necessarily in $X$), and $H$ is the space spanned by $\{s_i\}_{i=1}^M$. $R$ is the
operator projecting \( x \in \mathcal{X} \) onto \( \mathcal{H} \). Then \( \epsilon_{K - \tilde{K}} \) is bounded by
\[
epsilon_{K - \tilde{K}} \leq \sum_{i=1}^{N} \| x_i - K x_i \|^2.
\] (5)

For a non-linear kernel, a similar statement holds in Reproducing Kernel Hilbert Space (RKHS) by Mercer’s Theorem.

**Proposition 1** states that \( \epsilon_{K - \tilde{K}} \) is upper bounded by the summation of distance between \( x \in \mathcal{X} \) and subspace \( H \). If all of \( \mathcal{X} \) lie in \( H \), we can reconstruct the kernel matrix exactly, i.e., \( \epsilon_{K - \tilde{K}} = 0 \), no matter how we choose samples \( \mathcal{S}_M \), for \( \mathcal{S}_M \) always lie in \( H \) apparently. However, the eigenvector reconstruction error \( \epsilon_{K - \tilde{K}} \) can be very large while keeping \( \epsilon_{K - \tilde{K}} = 0 \), which has been shown in the counter example.

### 4. Hilbert Space Embedding Nyström criterion and its application in Laplacian matrices

We propose a novel criterion to address the question raised in the introduction, named Hilbert Space Embedding Nyström criterion, and then design a new Nyström algorithm based on the criterion. Our criterion is inspired by following theorem in spectral theory [17]:

**Theorem 2.** Let \( \mathcal{T} \) be a compact bounded linear operator. If the operator array \( |\mathcal{T}|_{M} \) is a strongly stable approximation of \( \mathcal{T} \), for a large enough \( M \),
\[
\theta(\mathcal{T}, \mathcal{T}_M) = O(\|\mathcal{T} - \mathcal{T}_M\|),
\]
where \( \theta(\mathcal{T}, \mathcal{T}_M) \) is the gap between the subspaces spanned by the eigenfunctions of \( \mathcal{T} \) and \( \mathcal{T}_M, R \) is the projection to the eigenspace of \( \mathcal{T} \).

**Theorem 2** states that the difference between eigenfunctions of two linear operators \( \mathcal{T} \) and \( \mathcal{T}_M \) is bounded by the norm of \( \|\mathcal{T} - \mathcal{T}_M\| \) restricted on the eigenspace of \( \mathcal{T} \). Inspired by this observation, we shall base our criterion on minimizing \( \|\mathcal{T} - \mathcal{T}_M\| \). Let \( \Psi(x, y) \) denote a PSD kernel function that may not be the same as the kernel \( \kappa(x, y) \), and
\[
\mathcal{T}(f) = \frac{1}{N} \sum_{i=1}^{N} \kappa((x_i), f(x_i)) \quad \text{T}_M(f) = \sum_{i=1}^{M} \alpha_i \Psi(s_i, f(s_i)),
\]
are operator forms of Eq. \(3\), \( \alpha_i \)'s are indeterminate coefficients, and \( s_i \)'s are samplings. If we get the optimal \( s_i, \alpha_i \) and \( \Psi \) according to the proposed criterion, we can estimate eigenvectors by the following steps:

1. Solve eigenvalue decomposition problem of an \( M \times M \) matrix:
\[
\Psi_M \mathbf{A} = \lambda \mathbf{A}, \quad \text{where} \quad \{\Psi_M\}_{i,j} = \Psi(s_j, s_i), \quad \mathbf{A} \text{ is a diagonal matrix},
\]
where \( \{\Psi_M\}_{i,j} = \Psi(s_j, s_i) \).
2. Obtain normalized approximated eigenvectors \( \tilde{f}_t \) as
\[
\tilde{f}_t = \Psi \mathbf{A} \mathbf{f}_t, \quad \tilde{f}_t = \tilde{f}_t / \| \tilde{f}_t \| \quad t = 1, 2, ..., M,
\]
where \( \{\mathbf{f}_t\}_{i,j} = \Psi(s_j, s_t) \).

The next question is, how to choose an appropriate \( \mathcal{T}_M \) so that its eigenspace is close to that of \( \mathcal{T} \). According to **Theorem 2**, we can adopt the following criterion:
\[
\min_{\mathcal{T}_M} \epsilon_1 \leq \|\mathcal{T} - \mathcal{T}_M\|.
\] (8)

Eq. (8) is the most general form of our criterion. However it is impractical since \( \mathcal{H}_{\tilde{T}} \) is unknown target. Luckily, in certain cases, we still know some information about the eigenspace \( \mathcal{H}_{\tilde{T}} \). For example, to use Radamacher complexity to bound Eq. (8) for general matrices. Another simple case is that Laplacian matrices [2] in eigenmap embedding problems always has an eigenvector \( \mathbf{1}(x) \equiv 1 \). We estimate Eq. (8) by replacing \( R \) with \( \mathbf{1}(x) \) and propose the following Hilbert Space Embedding Nyström criterion:
\[
\min_{\mathcal{T}_M} \epsilon_2 \leq \|\mathcal{T} - \mathcal{T}_M\| / 1(\mathbf{1}(x)).
\] (9)

Or equivalently its discrete form:
\[
\min_{\mathcal{T}_M} \epsilon_2 \leq \sum_{j=1}^{N} \frac{1}{N} \sum_{i=1}^{N} \kappa(x_i, x_j) - \sum_{t=1}^{M} \alpha_t \Psi(s_t, s_j) / \| \mathbf{1}(x) \|.
\]

The operation \( \int \kappa(x, y) dy / dy \) is to embed \( f \) into a RKHS identified by both kernel function \( \kappa \) and probability measure \( P(y) \). Let \( \hat{\Phi}(x) = \sum_{i=1}^{N} \delta(x-x_i) / N, \hat{\Phi}(x) = \sum_{i=1}^{M} \alpha_t \delta(x-x_i) / \sum_{t=1}^{M} \alpha_t, \mathcal{T}(f) \) and \( \mathcal{T}_M(f) \) can also be viewed as embeddings from the following perspective:
\[
\mathcal{T}(f) = \int \Psi(x, y) dy / dy \quad \mathcal{T}_M(f) = \int \Psi(x, y) dy / dy \hat{\Phi}(y).
\]

This happens to be a Hilbert Space Embedding step of \( f \), a concept recently proposed by [18], which we name our criterion after. Siripuram et al. [19] prove that the embedding \( \int \kappa(x, y) dy / dy \) actually defines a coarse measure on the RKHS. So HSE Nyström criterion is in fact minimizing the distance of embeddings in RKHS of two different operators \( \mathcal{T} \) and \( \mathcal{T}_M \).

### Connections with kernel density estimation
The kernel density estimation approximates probabilistic density \( p(x) \) with its empirical kernel density estimator
\[
p_h(x) = \frac{1}{N} \sum_{i=1}^{N} K_h(x, x_i) \approx p(x),
\]
where \( K_h \) is a Gaussian kernel function with kernel width \( \sigma \). Our Hilbert Space embedding operator \( \mathcal{T}_M(f) \) is identical to the kernel density estimator if we set \( f(x)=1 \). It is widely known that in kernel density estimation, the kernel width should be adaptively chosen according to the sampling data. Thus our criterion is closely related to the kernel density estimation. That is, in order to get an accurate eigenvector approximation, we do not need to approximate the kernel matrix defined by the given kernel function, but also need to consider the probabilistic density of data distribution.

**Optimization:** We now use our criterion (9) to design a HSE Nyström method to approximate eigenvectors of Laplacian matrices in eigenmap embedding problems. In this paper, we focus on the Gaussian kernel \( \kappa(x, y) = \exp(-\|x-y\|^2 / (2\sigma^2)) \) for its popularity. The Laplacian matrix here is defined as \( \mathbf{I} - D^{-1} \mathbf{K} \). It is easy to check that Laplacian matrix always has a constant eigenvector, and the \( k \)th smallest eigenvector is just the \( k \)th largest eigenvector of \( D^{-1} \mathbf{K} \). The standard eigenmap embedding procedure [1] is as follows:

1. Given dataset \( \mathbf{X} \) of size \( N \) and kernel width \( \sigma \), calculate the kernel matrix \( \{K\}_{i,j} = \kappa(x_i, x_j, \sigma) \), and the degree matrix \( \{D\}_{i,j} = \sum_{j=1}^{N} \{K\}_{i,j} \).
2. Obtain the top \( k \) largest eigenvectors \( \mathbf{v} \) of \( D^{-1} \mathbf{K} \), and using \( \mathbf{v} \) as eigenmap embedding.

### Table 1

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Two-elbow</th>
<th>mnist01</th>
<th>mnist68</th>
<th>ijcnn</th>
<th>covtype</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>20,000</td>
<td>12,665</td>
<td>11,769</td>
<td>20,000</td>
<td>20,000</td>
</tr>
<tr>
<td>( t_{\text{ops}} )</td>
<td>137.3</td>
<td>210.2</td>
<td>181.7</td>
<td>37.7</td>
<td>145.8</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.05</td>
<td>3.49</td>
<td>3.52</td>
<td>0.5099</td>
<td>1042.5</td>
</tr>
<tr>
<td>( t_{\text{ops}} )</td>
<td>55.8</td>
<td>255.8</td>
<td>219.3</td>
<td>12.5</td>
<td>61.7</td>
</tr>
</tbody>
</table>

\( N \) is the number of instances in the dataset, \( t_{\text{ops}} \) is the time in seconds to compute the top 10 eigenvectors. \( \sigma \) is the optimal Gaussian kernel width. \( t_{\text{ops}} \) is the time-cost for computing row average vector \( \mathbf{m} \).
According to our criterion \((9)\), we need to choose \(\Psi\) first. In the following, we assume \(\Psi\) to be a Gaussian kernel with kernel width \(\sigma_0 \neq \sigma\). \(\Psi(\mathbf{x}, \mathbf{y}) \equiv \mathcal{G}(\mathbf{x}, \mathbf{y}, \sigma_0)\). We will discuss how to optimize \(\sigma_0\) and \(\alpha_i\)'s later by our HSE Nyström criterion. Note that here we choose \(\Psi' \neq K\), which is different from traditional Nyström methods.

For the sample set \(S_M\), we allow it to be drawn from the entire space rather than just from \(X\). Sampling densely around some fix point \(\mathbf{x}_0\) is not optimal because we can tune \(\alpha_i\) to achieve the same effect and save the redundant samplings. This suggests that it is better to spread \(S_i\) to the region where \(P(\mathbf{x}) > 0\), and a simple and efficient way

---

Fig. 1. The 2nd eigenvector of two-elbow dataset, \(M = 25\). (a) two-elbow, (b) true eig, (c) HN eig, and (d) DN eig.

---

Fig. 2. Eigenvector approximation error and Hilbert Space Embedding error. (a) mnist01 top5 eig err, (b) mnist01:HS-err, (c) mnist68 top5 eig err, (d) mnist68:HS-err, (e) ijcnn top5 eig err, (f) ijcnn:HS-err, (g) covtype top5 eig err, and (h) covtype:HS-err.
to achieve this goal is the K-means algorithm, as the centers of the clusters obtained in K-means are well separated from each other. This explains why K-means sampling is a good choice from a perspective different from [11]. In fact, we utilize early-stopping in the experiments. 20 iterations in K-means step are enough to generate far separated samplings, no matter it converges or not.

Once we get the samplings, we also need to select $\sigma_0$ and $\alpha_i$'s for a good approximation. We adopt simple grid search to determine $\sigma_0$, and get $\alpha_i$'s by solving following Quadratic Programming (QP) problem.

$$\begin{align}
\min_{\alpha} & \quad \sum_{j=1}^{n} \frac{1}{N} \left[ \sum_{i=1}^{N} s(x_i, x_j) - \alpha_i \right]^2 \\
\text{s.t.} & \quad \alpha_i \geq 0.
\end{align}$$

Eq. (10) can be efficiently solved by many optimization methods, e.g., FISTA algorithm [20]. The input column vector $m$ and the output column vector $\alpha$ is defined by $[m]_i = \frac{1}{N} \sum_{j=1}^{N} s(x_i, x_j), [\alpha]_i = \alpha_i$. So Eq. (10) can be rewritten in matrix form:

$$\begin{align}
\min_{\alpha} & \quad (m - \Psi \alpha)^2 \\
\text{s.t.} & \quad \alpha \geq 0.
\end{align}$$

$\geq$ denotes element-wise non-negativity constraint.

After solving $\alpha_i$, we obtain the eigen system of $T_{M}$, and we can use Eq. (7) in our Nyström method. Finally, we summarize our method as follows, and name it Hilbert Space Embedding Nyström:

1. Perform K-means on $X$ with cluster number $M$, and let $S_M$ be the set of K-means centers.
2. Compute vector $m$, the HSE of $T : [m]_i = (D_{M})_{ii}/N$.
3. Grid search $\sigma_0$ based on criterion (9). For each $\sigma_0$, solve $\alpha_i$ in Eq. (10) using FISTA algorithm.
4. For each parameter pair $(\sigma_0, \alpha)$, computing the objective function in Eq. (10), and select $(\sigma_0, \alpha)$ with the smallest objective function value.
5. Solve the generalized eigenvalue problem

$$D_{M}^{-1/2} \Psi_{M} A v = \lambda v$$

$D_{M}^{-1/2}$ is the top k largest eigenvector.
6. Approximate the eigenvector by $f = D^{-1} \Psi A v$ and finally normalize as $f_{l} = f_{l}/\|f_{l}\|$.

5. Experiments

We compare the proposed HSE Nyström (HN) to two state-of-the-art algorithms, K-means Nyström (KN) [11] and Density Nyström (DN) [21], together with Standard Nyström (SN). Our code is implemented in MATLAB, running on an Intel(R) Core(TM)2 Duo 6300, 1.86 GHz CPU, with 2 GB physical memory. For comparison, we obtain the “true eigenvector” of given Laplacian matrix by MATLAB command “eigs” on a 64-bit supercomputer with 76 GB physical memory.

5.1. Datasets

We test the four algorithms on five datasets: synthetic 2 dimension two-elbow data with unit variance; 784 dimension mnist data \(^1\) with digit pair “0:’1’(mnist01) and pair ’5:’6’ (mnist68), with each pixel scaled to range $[0, 1]$. 22 dimension ijcnn1 and 54 dimension covtype. \(^2\) The statistics of the datasets are summarized in Table 1. Following the settings in previous works [21], we gradually set the sample size $M$ from 5 to 55 at an interval of 10, and make use of the sampling set $S_M$ to approximate the top 5 largest eigenvectors. The grid search range for $\sigma_0$ is $1.1 \times \{0: 25: 100\} \times \sigma$. Numerical experiences suggest that $\sigma_0$ out of this range does not help improve the accuracy. For each $M$, we report performance averaged over 30 trials.

Note that we compute the “true eigenvector” of the given eigen system for comparison with $N \geq 20,000$, that distinguishes our experiments from most previous works on large-scale datasets. Computing $m$ is a standard step of computing Laplacian matrix, so the corresponding time cost $T_m$ is subtracted in the experiments. In our implementation, we only compute $m$ once and use it through all our experiments. Our method has $O(NM + M^2)$ time complexity and $O(NM)$ space complexity, the same as other Nyström methods.

5.2. Eigenvector approximation

In eigenmap embedding, we are interested in the top $k$ largest eigenvectors of matrix $D^{-1} K$ [2]. Fig. 1 demonstrates the second largest eigenvector of two-elbow dataset. Fig. 1(b) shows the true ground of the 2nd eigenvector using full data. Fig. 1(c) shows the approximated eigenvector by HSE Nyström at $M=25$. Fig. 1(d) shows the approximation by Density Nyström, which is quite different from the True ground.

Fig. 2 demonstrates the overall performance (average + standard deviation bars) of 4 Nyström methods on the 5 datasets. The left column of Fig. 2 is the average approximation error rate of top 5 largest eigenvectors of four methods. Our method outperforms the other three methods on all datasets, especially when $M$ is small. To ensure our improvement is statistically significant, we give a non-parametric pair-wised Wilcoxon test in Table 2. We compute the $p$-value of HN against each other method, and set the level of significance $\alpha=0.05$. Our method achieves significant improvements in most cases, except on covtype against DN at $M=45, 55$. In Fig. 2(g) that both methods achieve high accuracies thus there are no significant differences. The right column demonstrates the relationship between the eigenvector approximation error and Hilbert Space Embedding error defined in

\(^1\) http://www.cs.nyu.edu/~roweis/data/mnist_all.mat.
\(^2\) http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Table 2

<table>
<thead>
<tr>
<th>M</th>
<th>5</th>
<th>15</th>
<th>25</th>
<th>35</th>
<th>45</th>
<th>55</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>2e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>KN</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>DN</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>mnist01</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>SN</td>
<td>2e-06</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>KN</td>
<td>4e-05</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>DN</td>
<td>4e-09</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>mnist68</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>SN</td>
<td>5e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>KN</td>
<td>6e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>DN</td>
<td>7e-08</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>covtype</td>
<td>8e-07</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>SN</td>
<td>9e-06</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>KN</td>
<td>1e-05</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>DN</td>
<td>1e-09</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>covtype</td>
<td>1e-08</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>SN</td>
<td>1e-07</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>KN</td>
<td>1e-06</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>DN</td>
<td>1e-09</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>covtype</td>
<td>1e-08</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>SN</td>
<td>2e-03</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>KN</td>
<td>4e-03</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>DN</td>
<td>6e-09</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
<tr>
<td>covtype</td>
<td>8e-07</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
<td>3e-11</td>
</tr>
</tbody>
</table>

Each column corresponds to different M, and each row presents the $p$-value of Wilcoxon test of HN against other Nyström methods. We set the level of significance $\alpha=5\%$. $p$-values in bold means we accept the null hypothesis that there is no statistical significant difference between two methods.
Eq. (9). We can see from the figures that they are positively correlated, which verifies our theory.

Fig. 3 shows the 2D eigenmap embeddings of the true eigenvectors and the approximated eigenvectors by HSEN and DN at \( M=55 \). The 2D eigenmap embeddings are plotted with the 2nd and 3rd largest eigenvectors, since the largest eigenvector is a constant. The approximated embeddings by our method are almost the same as the ground truth.

To precisely show the time-accuracy trade-off of different methods, we list the average approximation error rate of top 5 eigenvectors together with time cost in Table 3. Each entry in the table is in format \( \varepsilon/t \), where \( \varepsilon \) is the average error rate and \( t \) is the time cost in seconds. Our method achieves better accuracy at the cost of extra linear running time. Our main purpose is to demonstrate that our criterion could lead to more accurate approximations within linear complexity. Different approaches of our criterion may lead to different time-accuracy trade-off methods. For example, if we omit the optimization steps of \( \sigma \) and \( \alpha \) in HSE Nyström criterion for speed, our method will degenerate to Density Nyström.

To show the linear time complexity \( O(N) \) of the proposed method, we plot time cost with different sampling size \( M \) in Fig. 4. Though at a fix \( M \) our method requires more computing time, our method achieves better accuracy than other methods. Increasing \( M \) may improve accuracy, but will result in \( O(M^2) \) space cost and \( O(M^3) \) time cost. Compare with time cost \( T_{\text{eigs}} \) of “eigs” command in Table 1, the reported time cost of the proposed method at \( M=55 \) is close to \( T_{\text{eigs}} \). The reason is that the time of Table 1 only compute 10 eigenvectors, while the proposed method...
produces $M=55$ eigenvectors. What’s more, “eigs” requires to store $O(N^2)$ kernel matrix in memory, which is impossible on a laptop with 2 GB memory, while the proposed method is able to.

5.3. Kernel matrix and eigenvector approximation error

To verify that a small kernel matrix approximation does not necessarily imply a small eigenvector approximation error, we show the relation of two errors in Fig. 5. The $x$-axis is the approximation error of top 5 eigenvectors. The $y$-axis is the approximation error of kernel matrix. We use red circle to present DN and blue circle to present DHSEN. On each dataset, we show the relation of the two errors under different sampling sizes. In the figure, although some DN experiments achieve much smaller kernel matrix approximation error, their eigenvector approximation error is still much larger than DHSEN. For the same method, we can see a positive correlation of the two errors, because a good eigenvector approximation is usually a good kernel matrix approximation.

5.4. Parameter tuning in HSEN

There are three main parameters in HSEN methods: the weights $\alpha$, the sampled points $S_M$, and the kernel parameter $\sigma_0$. All of these parameters are important to obtain the optimal approximation. In the following experiments, we alter each parameter in the HSEN algorithm to show their effects. Specifically, we compare the following cases:

1. Fix $\alpha_i = 1/M$ as the weights assigned in K-means Nyström.
2. Set $\alpha_i$ equal to $i$th cluster size as the weights assigned in Density Nyström.
3. Randomly choose sampling point $S_M$ as Standard Nyström.
4. Fix $\sigma_0 = \sigma$ as most previous Nyström methods.

Fig. 6 shows our results. If we set $\alpha_i$ equal to $1/M$ (‘1/M’ curve) or $i$th cluster size (‘CS’ curve), the accuracy is always not optimal. On some datasets ‘CS’ curve is better than ‘1/M’, and on the other datasets they are competitive. In the case we randomly choose $S_M$ (‘SR’ curve), it is better than ‘CS’ on mnist01 and is worse than ‘CS’ on other datasets. Fixing $\sigma_0 = \sigma$ (‘OS’ curve) achieves almost the same accuracy on mnist01 and mnist68, because these datasets are not densely sampled enough. The optimal $\sigma$ on these datasets is not too small. Therefore fixing $\sigma_0 = \sigma$ still work in these cases as we analyze above. On dataset icmnc and covtype, ‘OS’ achieves secondary accuracy. These empirical results suggest that all of the three parameters should be chosen correctly to obtain the optimal accuracy. Each step in HSEN is non-trivial.

6. Conclusion

In this paper, we first prove that the criterion used in online eigenvector approximation is not applicable to batch-based method such as the Nyström method, then we propose our Hilbert Space Embedding Nyström criterion to fill the gap between kernel
approximation and eigenvector approximation. Our work reveals that eigenvector approximation is fundamentally difficult than kernel approximation—A perfect kernel approximation cannot guarantee the quality of eigenvector approximation. We hope this work will draw the attention of research community to separate the eigenvector approximation from kernel approximation, since the former one are more difficult at least in the Nyström method. Potential extensions of our work may be how to find a faster approximation under our criterion, and how to generalize our criterion to the problems other than Laplacian matrices.

Conflict of interest

None declared.

Acknowledgment

This work is supported by 973 Program (2013CB329503), NSFC (Grant No. 91120301 and No. 61473167).
Appendix A. Proof of Proposition 1

Proposition 2. Let $X = \{x_1, x_2, \ldots, x_m\} \in \mathbb{R}^d$, $k(x, y) = \langle x, y \rangle = x^T y$ is an inner product kernel, $S_M = \{s_1, s_2, \ldots, s_M\}$ is sampled from $\mathbb{R}^d$ (not necessarily in $X$), and $H$ is the space spanned by $\{s_j\}_{j=1}^M$, $P$ is the operator projecting $x \in X$ to $H$. Then $\epsilon_{K-K} = \|K-PK\|^2$.

To prove proposition 1, we need following two lemmas. For the sake of simplicity, we always assume $S_M = \{s_1, s_2, \ldots, s_M\}$ is of full column rank. If $S_M$ is not of full column rank, we can crop the linear correlation components $s_i$’s in $S_M$ and use cropped $S_M$ instead.

Lemma 1. Let $P, H$ be the same as definitions in Proposition 1, and $Y = \{y_1, y_2, \ldots, y_M\}$ is the projected $x_i$’s onto $H$, i.e., $y_j = P x_i$. Then the kernel matrix $K$ reconstructed by the Nyström method Eq. (6) is equal to the inner product matrix of projected $X$, i.e., $K = Y^T Y$.

Proof. From linear algebra, $P$ can be expressed by a matrix:

$$P = S_M (S_M^T S_M)^{-1} S_M^T,$$

such that $y_j = P x_i \in \text{span}(x_i)$, where $\text{span}(x_i) = H$. Then we have:

$$Y = PX \quad Y^T Y = (PX)^T (PX) \quad = X^T P^T P X \quad = X^T (S_M S_M^T S_M^{-1} S_M^T)^{-1} (S_M S_M^T S_M^{-1} S_M^T) X \quad = (X^T S) (S_M^T S_M^{-1} S_M^T S_M S_M^T S_M^{-1} S_M^T X)^T \quad = K x_j K_M^{-1} K x_i \quad = K_R \quad \Box$$

And

Lemma 2. If $\forall x \in X, x \in H$, then $K = K_R$.

Proof. For $\forall x \in X, x \in H$, so there exist a coefficient matrix $A$, $X = SA$.

Then we have

$$K = K_R K_M^{-1} K x_i \quad = (X^T S) (S_M^T S_M^{-1} S_M^T S_M S_M^T S_M^{-1} S_M^T S_M S_M^{-1} S_M^T A)^T \quad = A^T (S_M S_M^T S_M S_M^{-1} S_M^T S_M A) \quad = A^T S_M S_M^T \quad = X^T X = K \quad \Box$$

Lemma 2 reveals that if all $x_i$’s lie on $H$, then the Nyström method can reconstruct the kernel matrix exactly.

Using Lemmas 1 and 2, we can prove Proposition 1.

Proof of Proposition 1. Let $X, Y, P$ be the same as above context. Define $z_i = x_i - y_i$, such that $z_i \perp H$, and $x_i = z_i + y_i$. Then

$$K = X^T X = (Z + Y)^T (Z + Y) \quad = Z^T Z + Y^T Y + Z^T Y + Y^T Z \quad = Z^T Z + Y^T Y \quad \text{(for } z_i \perp y_i) \quad = Z^T Z + K_R \quad \Box$$

So the kernel reconstruction error is bounded by

$$\|K - K_R\|^2 \leq \sum_{i,j} (z_i^T z_j)^2 \quad \leq \sum_{i,j} (z_i^T z_i)^2 (z_j^T z_j)^2 \quad = \sum_{i,j} |z_i|^2 |z_j|^2 \quad = \sum_{i,j} \|x_i - y_i\|^2 \quad \Box$$

References