Multi-scale Nyström Method

Woosang Lim\textsuperscript{1}  woosang.lim@cc.gatech.edu  
Rundong Du\textsuperscript{1}  rdu@gatech.edu  
Bo Dai\textsuperscript{1}  bodai@gatech.edu  
Kyomin Jung\textsuperscript{2}  kjung@snu.ac.kr  
Le Song\textsuperscript{1}  lsong@cc.gatech.edu  
Haesun Park\textsuperscript{1}  hpark@cc.gatech.edu  
\textsuperscript{1}Georgia Institute of Technology, Atlanta, GA, USA  
\textsuperscript{2}Seoul National University, Seoul, South Korea

Abstract

Kernel methods are powerful tools for modeling nonlinear data. However, the amount of computation and memory required for kernel methods becomes the bottleneck when dealing with large-scale problems. In this paper, we propose Nested Nyström Method (NNM) which achieves a delicate balance between the approximation accuracy and computational efficiency by exploiting the multilayer structure and multiple compressions. Even when the size of the kernel matrix is very large, NNM consistently decomposes very small matrices to update the eigen-decomposition of the kernel matrix. We theoretically show that NNM implicitly updates the principal subspace through the multiple layers, and also prove that its corresponding errors of rank-$k$ PSD matrix approximation and kernel PCA (KPCA) are decreased by using additional sublayers before the final layer. Finally, we empirically demonstrate the decreasing property of errors of NNM with the additional sublayers through the experiments on the constructed kernel matrices of real data sets, and show that NNM effectively controls the efficiency both for rank-$k$ PSD matrix approximation and KPCA.

1 Introduction

The scalability of kernel methods is the major bottleneck for applying them to large-scale problems due to the computational and memory cost caused by the large dense kernel matrices. Nyström method is one of the effective methods for accelerating the kernel methods by low-rank approximation of the kernel matrix, $K \in \mathbb{R}^{n \times n}$. There has been a large body of work that further improves the approximation quality and computational efficiency via adopting various sampling methods [5, 14, 4, 8, 12, 3, 10, 21, 9] and refining approximation formula [7, 5, 11, 19, 12, 17]. Especially, for rank-$k$ spectral decomposition of $K$, there are two basic rank-$k$ Nyström methods which are rank-$k$ Standard Nyström Method (SNM) [5] and orthogonal Nyström method (ONM) [7]. Recently, their efficient versions which are SNM using Randomized SVD (SNM+Rand.SVD) [11] and Double Nyström Method (DNM) [12] were proposed. All these four methods implicitly approximate the first $k$ principal directions $U_{Y,k}$ of $n$ mapped data points $Y$ in the feature space to compute the rank-$k$ spectral decomposition of $K = Y^\top Y$ with distinct schemes based on different motivations [12]. Rank-$k$ SNM [5] actually computes the first $k$ principal directions $U_{S,k}$ of $s$ sample mapped points $S$ in the feature space, and SNM+Rand.SVD [11] uses randomized SVD to improve efficiency for computing the principal directions of sample mapped points. That is, rank-$k$ SNM and SNM+Rand.SVD approximate $U_{Y,k}$ via $U_{S,k}$, which is computed by a particular form. However, it is known that both these two approximations are biased to the sample subspace which is $\text{range}(S)$. On the other hand, the ONM computes the best $k$ approximate principal orthogonal direction in the sample subspace $\text{range}(S)$ in the sense to minimize the KPCA reconstruction error [12]. However, such approximation requires extra computation, resulting higher time complexity $O(s^2n)$ compared to the time complexity of rank-$k$ SNM which is $O(ksn + k^3)$. To further accelerate ONM, DNM [12] uses ONM twice in different scales, so that to compress the sample subspace $\text{range}(S)$ for reducing the dimension of possible solution space for efficient computing of $U_{Y,k}$. Although the algorithm performs well in practice, there is no analysis about how its rank-$k$ approximation error varies after compression of sample subspace, and it is not clear whether the double scales are enough in terms of the balance between approximation accuracy and computation efficiency.

To achieve a better trade-off between these two factors, we extend DNM to a multi-scale Nyström method. Accelerating the algorithms by exploiting multi-scale structures has been studied for the various methods.
Nested Nyström Method (NNM)

Require: $n \times s$ matrix $C$ and $s \times s$ matrix $K_S$, where $C = Y^\top S$ and $K_S = S^\top S$, where $s \ll n$

Ensure: rank-$k$ spectral decomposition of $K$

1: **Subsampling part:**
   Subsampling indices from the index set $J$ of $S$ s.t. $J \supseteq J_1 \supseteq \ldots \supseteq J_t$, and corresponding $C \supseteq K_S \supseteq C_1 \supseteq K_{S_1} \ldots \supseteq C_t \supseteq K_{S_t}$, where $|J_i| = s_i$, $s \gg s_1 \gg \ldots \gg s_t$

2: **For $i$-th sublayer from the 1st to the $t$-th sublayer:**
   Rank-$s_t$ Nyström method: Compute $\tilde{V}_{S_{t-i}, s_t}$ of $K_{S_{t-i}}$ with $C_{t-(i-1)}'$ and $K_{S_{t-(i-1)}}'$ (optional use ONM)
   Compression: Compress sample matrices $C_{t-i}$ and $K_{S_{t-i}}$ as $C'_{t-i}$ and $K'_{S_{t-i}}$ (Eqn (2))

3: **Final layer:**
   Run ONM [7] with $C'$ and $K_S'$

Algorithm 1 Nested Nyström Method (NNM)

---

including FEM [6], Bayesian optimization [20] and neural network [1] to solve the nonlinear problems, and there are also a number of applications such as multi-scale stable kernel construction [16], manifold learning [18], dictionary learning [15], and object detection [2, 13]. Among them, feature pyramid networks [13] successfully achieves both efficient and accurate object detection.

Inspired by the multi-scale approximation, we propose a multi-scale Nyström method, Nested Nyström Method (NNM), for both efficient and accurate eigen-decomposition of PSD matrices. NNM has a multilayer structure which consists of $t$ sublayers and the final layer to efficiently and accurately updates the first $k$ principal direction $U_{Y, k}$ for computing a rank-$k$ spectral decomposition of $K$. We note that NNM is a general multi-scale framework which can be combined with any other column sampling, and our contribution is orthogonal to the column samplings. Interestingly, it contains $t$ fully connected neural networks in the structure of NNM for the compressions of sample subspaces as described in Fig 1.

# Nested Nyström Method

The multilayer architecture of NNM is described in Alg 1 and Fig 1, and it consists of the following three parts: subsampling part, rank-$s_t$ Nyström method part, and compression part.
Subsampling part: Given index set $\mathcal{J}$ of $s$ samples and the corresponding sample matrices $S$ and $K_S$, we construct a nested index sets $\mathcal{J} \supseteq \mathcal{J}_1 \supseteq \ldots \supseteq \mathcal{J}_t$ and the corresponding nested sequence of submatrices as Eqn (1).

$$S \supseteq S_1 \supseteq S_2 \supseteq \ldots \supseteq S_t, \quad C \supseteq K_S \supseteq C_1 \supseteq K_{S_1} \supseteq C_2 \supseteq K_{S_2} \supseteq \ldots \supseteq C_t \supseteq K_{S_t},$$

where $|\mathcal{J}_i| = s_i$, and $s \gg s_1 \gg \ldots \gg s_t$. Especially, we can understand $(s_{i-1}) \times s_i$ matrix $C_i$, and $s_i \times s_i$ matrix $K_{S_i}$ with implicit equations as $C_i = S_{i-1}^T S_i$ and $K_{S_i} = S_i^T S_i$, for $1 \leq i \leq t$, where $S_i$ is $d \times s_i$ and $S_0 = S$, and $C = Y^T S$. We will compress $S_i$, $C_i$ and $K_{S_i}$ as $S'_i$, $C'_i$ and $K'_{S_i}$ later.

Rank-$s_t$ Nyström method part: In this part, we compute the approximate eigenvectors $\tilde{V}_{S_i}$ of $K_{S_i}$ by using compressed submatrices $C'_{S_{i-1}}$ and $K'_{S_{i+1}}$, where $A_k = U_A, S A_k V_A^T$ denotes the rank-$k$ SVD of a general matrix $A$, and tilde means their approximations. From the 1st to the $(t-1)$-th sublayer: We compute the first $s_t$ approximate eigenvectors $V_{S_t,s_t}$ of $K_{S_t}$ by using compressed submatrices $C_t$ and $K_{S_{t+1}}$ on the $(t-i)$-th layer, where $i \in \{1, 2, \ldots, (t-1)\}$ and $C_t = C_i$ and $K'_{S_t} = K_{S_t}$. On the $t$-th sublayer: We compute the first $s_t$ approximate eigenvectors $V_{S_t,s_t}$ of $K_{S_t}$ by using $C'_t$ and $K'_{S_t}$, and select $V_{S_t,t}$ from $V_{S_{t+1},s_{t+1}}$, where $s_t \geq \ell \geq k$.

Compression part: In this part, we compress sample matrices by using the approximate eigenvectors. We compress sample matrices $C_i$ and $K_{S_i}$, by using $V_{S_i,s_i}$ as

$$C'_i = C_i \tilde{V}_{S_i,s_i}, \quad K'_{S_i} = (\tilde{V}_{S_i,s_i})^T K_{S_i} \tilde{V}_{S_i,s_i}, \quad C' = CV_{S_t,t}, \quad K' = (\tilde{V}_{S_t,t})^T K_{S_t} \tilde{V}_{S_t,t},$$

where $\tilde{V}_{S_i,s_i}$ is computed at $(t-i)$-th layer with $i \in \{1, 2, \ldots, (t-1)\}$, and we compress sample matrices $C$ and $K_S$ by using $V_{S_t,t}$ with $k \leq \ell \leq s_t$. We can connect the compression of sample matrices to the compression of sample subspace with implicit equations

$$C'_i = S_{i-1}^T S'_i, \quad C' = Y^T S', \quad K'_i = S_i^T S'_i, \quad K'_S = S'^T S',$$

where $S'_i = S_i \tilde{V}_{S_i,s_i}, S' = S \tilde{V}_{S_t,t}, i \in \{1, 2, \ldots, (t-1)\}$, and $S_0 = S$.

Based on Eqn (3), it can be interpreted that the sample subspace range$(S_i)$ is compressed into a smaller dimensional subspace range$(S'_i)$, where $i \in 0, 1, \ldots, (t-1)$ and $S_0 = S$. If we use ONM for rank-$s_t$ Nyström method part in NNM, and set the nested sequence of subsamples with $\sum_{j=1}^{t} s_j = O(s)$, where $s \gg s_1 \gg \ldots \gg s_t \geq \ell \geq k$. Then, the total time and space complexities of NNM are $O(\ell s_1 + s_1 s_1)$ and $O(s_1)$, respectively. A large portion of the total time complexity $O(\ell s_1 + s_1 s_1)$ is $O(\ell s_1)$ which corresponds to the simple matrix multiplications in the compression parts. Furthermore, by extending the multilayer structure of NNM, we can efficiently update the spectral decomposition with additional samples and data points.

2.1 Error Analysis of NNM

NNM efficiently and accurately updates the compressed sample matrix $S'$ so that range$(S'_i)$ closely approximates the true principal subspace based on Eqn (3) until the final layer. That is, we want to compute $S'$ s.t. range$(U_k) \subset$ range$(S'_i)$, and we can give the implicit representation of the principal subspace as range$(U_k) = \text{range}(U_k \Sigma_{Y,k}) = \text{range}(Y \tilde{V}_{Y,k})$. Similarly, we can give the implicit representations of compressed sample subspaces, and Lem 1 formally provide them.

Lemma 1 Given the multilayer Nyström structure of NNM with $t$ sublayers, NNM computes $S'_i = Y \tilde{V}_{Y,s_i}$ on the $(t-i)$-th layer, and $S' = Y \tilde{V}_{Y,t}$ with $(\tilde{V}_{Y,s_t})^T \tilde{V}_{Y,s_t} = I$ and $(\tilde{V}_{Y,t})^T \tilde{V}_{Y,t} = I$.

Now we provide Lem 2 which provides the differences between the optimal error and the error of NNM both for KPCA and rank-$k$ PSD kernel matrix approximation.

Lemma 2 Suppose that $S' = Y \tilde{V}_{Y,t}$ is the compressed samples as an input of the final layer of NNM, where $\tilde{V}_{Y,t} \tilde{V}_{Y,t} = I$ and $k \leq \ell \leq s_t$. Then, the differences between approximation error of NNM and the optimal errors for KPCA and rank-$k$ PSD matrix approximation are bounded by constant times of $c_2(\tilde{V}_{Y,k})$, where $\tilde{V}_{Y,k}$ is any submatrix consisting of $k$ columns of $\tilde{V}_{Y,t}$, and $c_2(\tilde{V}_{Y,k}) = \text{tr}(\tilde{V}_{Y,k}' K \tilde{V}_{Y,k}) - \text{tr}(\tilde{V}_{Y,t}' K \tilde{V}_{Y,k})$ is the sum of errors of eigenvalues from $\tilde{V}_{Y,k}$.
with NNM (ours). NNM is more efficient than other state-of-the-art Nyström methods within the same short time. Fig 3 shows that the errors of NNM both for sublayers regardless of data sets, and also shows that the errors of NNM is smaller than errors of other Nyström methods.

Lem 2 states that if \( \epsilon_2(\tilde{\mathbf{V}}_{\mathbf{Y},k}) \) goes to 0, then the approximation errors of NNM go to the optimal errors both for KPCA and rank-\( k \) kernel matrix approximation. Since reducing \( \epsilon_2(\tilde{\mathbf{V}}_{\mathbf{Y},k}) \) is important, we need to show how \( \epsilon_2(\tilde{\mathbf{V}}_{\mathbf{Y},k}) \) varies through the sublayers.

Now, we provide our main theoretical result Thm 1 which states that the quality of compressed input at the final layer is important, and we can increase accuracy by using more sublayers.

**Theorem 1** Suppose that we use ONM for the kernel PCA parts in the sublayers. As we use additional sublayers, \( \epsilon_2(\tilde{\mathbf{V}}_{\mathbf{Y},k}) \) decreases, and the upper error bounds of NNM in Lem 2 decrease.

### 3 Experiments

In this section, we present experimental results that demonstrate our theoretical work. We use 3 real data sets which are MNIST, MiniBooNE, and Covertype. We compare rank-\( k \) Nyström methods to the rank-\( k \) kernel matrix approximation and KPCA. The three error measures which we used are

- matrix reconstruction error (MRE(\( \mathbf{K}_k \)) = \( \| \mathbf{K} - \hat{\mathbf{K}}_k \|_F \)),
- relative matrix reconstruction error (rMRE(\( \mathbf{K}_k \)) = \( \frac{\| \mathbf{K} - \hat{\mathbf{K}}_k \|_F}{\| \mathbf{K} \|_F} \)),
- relative KPCA reconstruction error (rNRE(\( \mathbf{U}_k \)) = \( \frac{\text{NRE}(\mathbf{U}_k)}{\text{NRE}(\hat{\mathbf{U}}_k)} \)).

\( \mathbf{U}_k \) consists of the first \( k \) approximate principal directions which are implicitly computed by KPCA. \( \text{NRE}(\tilde{\mathbf{U}}_{\mathbf{Y},k}) = \frac{\| \mathbf{Y} - \tilde{\mathbf{U}}_{\mathbf{Y},k} \tilde{\mathbf{U}}_{\mathbf{Y},k}^\top \mathbf{Y} \|_F}{\| \mathbf{Y} \|_F} \) is the normalized reconstruction error (NRE) of KPCA, and \( \| \mathbf{K} - \hat{\mathbf{K}}_k \|_F \) is the optimum which comes from SVD.

To construct PSD matrix \( \mathbf{K} \), we use RBF kernel which is defined as \( \sigma(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\| \mathbf{x}_i - \mathbf{x}_j \|^2}{2\sigma^2}\right) \), where \( \sigma \) is a kernel parameter. We abbreviate NNM with \( i \) sublayers to NNM (\( t = i \)) for convenience, and DNM [12] is the same with NNM (\( t = 1 \)).

Fig 2 demonstrates that the error of NNM decreases and its efficiency can be improved as we use additional sublayers regardless of data sets, and also shows that the errors of NNM is smaller than errors of other state-of-the-art Nyström methods within the same short time. Fig 3 shows that the errors of NNM both for KPCA and rank-\( k \) kernel matrix approximation rapidly decrease compared to other rank-\( k \) Nyström methods.
References


