Double Nyström Method: An Efficient and Accurate Nyström Scheme for Large-Scale Data Sets

Woosang Lim QUASAR17@KAIST.AC.KR
School of Computing, KAIST, Daejeon, Korea

Minhwan Kim MINHWAN1.KIM@LGE.COM
LG Electronics, Seoul, Korea

Haesun Park HPARK@CC.GATECH.EDU
School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Kyomin Jung KJUNG@SNU.AC.KR
Department of Electrical and Computer Engineering, Seoul National University, Seoul, Korea

Abstract

The Nyström method has been one of the most effective techniques for kernel-based approach that scales well to large data sets. Since its introduction, there has been a large body of work that improves the approximation accuracy while maintaining computational efficiency. In this paper, we present a novel Nyström method that improves both accuracy and efficiency based on a new theoretical analysis. We first provide a generalized sampling scheme, CAPS, that minimizes a novel error bound based on the subspace distance. We then present our double Nyström method that reduces the size of the decomposition in two stages. We show that our method is highly efficient and accurate compared to other state-of-the-art Nyström methods by evaluating them on a number of real data sets.

1. Introduction

Low-rank matrix approximation is one of the core techniques to mitigate the space requirement that arises in large-scale machine learning and data mining. Consequently, many methods in machine learning involve a low-rank approximation of matrices that represent data, such as manifold learning (Fowlkes et al., 2004; Talwalkar et al., 2008), support vector machines (Fine & Scheinberg, 2002) and kernel principal component analysis (Zhang et al., 2008). These methods typically involve spectral decomposition of a symmetric positive semi-definite (SPSD) matrix, but its exact computation is prohibitively expensive for large data sets.

The standard Nyström method (Williams & Seeger, 2001) is one of the popular methods for approximate spectral decomposition of a large kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ (generally a SPSD matrix), due to its simplicity and efficiency (Kumar et al., 2009). One of the main characteristics of the Nyström method is that it uses samples to reduce the original problem of decomposing the given $n \times n$ kernel matrix to the problem of decomposing a $s \times s$ matrix, where s is the number of samples much smaller than n. The standard Nyström method has time complexity $O(ksn + s^3)$ for rank-k approximation and is indeed scalable. However, the accuracy is typically its weakness, and there have been many studies to improve its accuracy. Most of the recent work in this line of research can be roughly categorized into the following two types of approaches:

Refining Decomposition Exemplar works in this category are one-shot Nyström method (Fowlkes et al., 2004), modified Nyström method (Wang & Zhang, 2013), ensemble Nyström method (Kumar et al., 2012) and standard Nyström method using randomized SVD (Li et al., 2015). All of these methods redefine the *intersection* matrix that appears in the reconstructed form of the input matrix, of which we provide details in Section 2.

The motivation of the one-shot Nyström method (Fowlkes et al., 2004) is to obtain an orthonormal set of approximate eigenvectors of the given kernel matrix via diagonalization of the standard Nyström approximation in one-shot with $O(s^2 n)$ running time. Because of the orthonormality, the

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

one-shot Nyström is widely used for kernel PCA (Zhang et al., 2008), however there is no more elaborate analysis for it except the work of Fowlkes et al. (2004). The modified Nyström method (Wang & Zhang, 2013) involves multiplying both sides of kernel matrix by projection matrix consisting of orthonormal basis of subspace spanned by samples. It solves the problem of minimizing matrix reconstruction error, which is $\min_{\mathbf{U}} \|\mathbf{K} - \mathbf{C}\mathbf{U}\mathbf{C}^\top\|_F$, given the kernel matrix $\mathbf{K}$ and the $\mathbf{C}$, where $\mathbf{C}$ is a submatrix consisting of ℓ columns of $\mathbf{K}$. Although the modified Nyström approximation is more accurate than the standard Nyström approximation, it is more expensive to compute and is not able to compute a rank-k approximation. Its time complexity is $O(s^2n + sn^2)$, and the latter term $O(sn^2)$ comes from some matrix multiplications and dominates $O(s^2n)$. The ensemble Nyström method (Kumar et al., 2012) that takes a mixture of t (≥ 1) standard Nyström approximations, and is more accurate than the standard Nyström approximations in the empirical results. Its time complexity is $O(ksnt + s^3t + \mu)$, where μ is the cost of computing the mixture weights. To obtain more efficiency, we can adopt randomized SVD to approximate the pseudo inverse of the intersection matrix of the standard Nyström approximation (Li et al., 2015). Its time complexity is $O(ksn)$, but it needs larger samples than other Nyström methods due to adopting approximate SVD.

Improving Sampling The Nyström methods require column/row samples which heavily affects the accuracy. Among many sampling strategies for Nyström methods, uniform sampling without replacement is the most basic sampling strategy (Williams & Seeger, 2001), of which probabilistic error bounds for the standard Nyström method are recently derived (Kumar et al., 2012; Gittens & Mahoney, 2013).

Recent work includes the non-uniform samplings, which are square of diagonal sampling (Drineas & Mahoney, 2005), square of L_2 column norm sampling (Drineas & Mahoney, 2005), leverage score sampling (Mahoney & Drineas, 2009) and approximate leverage score sampling (Drineas et al., 2012; Gittens & Mahoney, 2013). The adaptive sampling strategies also have been studied, *e.g.* (Deshpande et al., 2006; Kumar et al., 2012). Some of the heuristic sampling strategies for Nyström methods are utilizing normal K-means algorithm (Zhang & Kwok, 2010) with $K = s$, and adopting pseudo centroids of normal (weighted) K-means (Hsieh et al., 2014). Normal K-means sampling clusters the original data which is not applied kernel function, and uses s centroids to generate matrices consisting of kernel function values among all data points and s centroids to perform Nyström methods. Although it has been shown to give good empirical accuracy, its proposed analysis is quite loose and does not show any connection to the minimum error of rank-k approximation.

1.1. Our Contributions

In this paper, we propose a novel Nyström method, *Double Nyström Method*, which tightly integrates the strengths of both types of approaches. Our contribution can be appreciated in three aspects: comprehensive analysis of the one-shot Nyström method, generalization of sampling methods, and integration of our two results. Summary of those are described as follows.

1.1.1. Comprehensive Analysis of the One-shot Nyström Method

In Section 3, we provide an analysis that one-shot Nyström is quite a good compromise between the standard Nyström method and the modified Nyström method, since it yields accurate rank-k approximations for $k < s$ (Thm 1). In addition, we show that it is robust (Proposition 1).

1.1.2. Generalization of Sampling Methods

In Section 4, we investigate how we can improve accuracies of Nyström methods. First, we present new upper error bounds both for the standard and one-shot Nyström methods (Thm 3), and provide a generalized view of sampling schemes which makes connection among some of the sampling schemes and minimization of our error bounds (Rem 1). Next, we propose *Capturing Approximate Principal Subspace* (CAPS) algorithm which minimizes our upper error bounds efficiently (Proposition 2).

1.1.3. The Double Nyström Method

In Section 5, we propose Double Nyström Method (Alg 3) that combines the advantages of CAPS sampling and the one-shot Nyström method. It reduces the size of the decomposition problem in twice, and consequently is much more efficient than the standard Nyström method for large data sets, but is as accurate as the one-shot Nyström method. Its time complexity is also comparable to the running time of standard Nyström method using randomized SVD, since it is $O(\ell sn + m^2 s)$ and linear for s, where $\ell \leq m \ll s \ll n$.

2. Preliminaries

Given the data set $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and its corresponding matrix $\mathbf{X} \in \mathbb{R}^{d_0 \times n}$, we define the kernel function without explicit feature mapping ϕ as $\kappa(\mathbf{x}_i, \mathbf{x}_j)$, where ϕ is a feature mapping such that $\phi : \mathcal{X} \to \mathbb{R}^d$. The corresponding *kernel matrix* $\mathbf{K} \in \mathbb{R}^{n \times n}$ is a positive semi-definite (PSD) matrix with elements $\kappa(\mathbf{x}_i, \mathbf{x}_j)$. Without loss of generality, let $\mathbf{\Phi} = \phi(\mathbf{X}) \in \mathbb{R}^{d \times n}$ be the matrix constructed by taking $\phi(\mathbf{x}_i)$ as the i-th column, so that $\mathbf{K} = \mathbf{\Phi}^\top \mathbf{\Phi}$ (Drineas & Mahoney, 2005). We assume that $\mathrm{rank}(\mathbf{\Phi}) = r$, which in

turn implies $\text{rank}(\mathbf{K}) = r$.

Consider the compact singular value decomposition (compact SVD) [1] $\mathbf{\Phi} = \mathbf{U}_{\Phi,r}\mathbf{\Sigma}_{\Phi,r}\mathbf{V}_{\Phi,r}^\top$, where $\mathbf{\Sigma}_{\Phi,r}$ is the diagonal matrix consisting of r nonzero singular values $(\sigma(\mathbf{\Phi})_1, \ldots, \sigma(\mathbf{\Phi})_r)$ of $\mathbf{\Phi}$ in decreasing order, and $\mathbf{U}_{\Phi,r} \in \mathbb{R}^{d\times r}$ and $\mathbf{V}_{\Phi,r} \in \mathbb{R}^{n\times r}$ are the matrices consisting of the left and right singular vectors, respectively. Especially, we simply denote compact SVD of $\mathbf{\Phi}$ as $\mathbf{\Phi} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top$, and $(\sigma(\mathbf{\Phi})_1, \ldots, \sigma(\mathbf{\Phi})_r)$ as $(\sigma_1, \ldots, \sigma_r)$ in this paper. Then, we can obtain the compact SVD $\mathbf{K} = \mathbf{V}_r\mathbf{\Sigma}_r^2\mathbf{V}_r^\top$ and its pseudo-inverse obtained by $\mathbf{K}^\dagger = \mathbf{V}_r\mathbf{\Sigma}_r^{-2}\mathbf{V}_r^\top$. The best rank-$k$ (with $k \le r$) approximation of $\mathbf{K}$ can be obtained from its SVD by

$$\mathbf{K}_k = \mathbf{V}_k\mathbf{\Sigma}_k^2\mathbf{V}_k^\top = \textstyle\sum_{i=1}^k \lambda_i(\mathbf{K})\mathbf{v}_i\mathbf{v}_i^\top,$$

where $\lambda_i(\mathbf{K}) = \sigma_i^2$ are the first k eigenvalues of $\mathbf{K}$. Especially we simply denote again $\lambda_i(\mathbf{K}) = \lambda_i$ in this paper, thus $\lambda_i = \sigma_i^2$.

Given a set $W = \{\mathbf{w}_1, \ldots, \mathbf{w}_s\}$ of s mapped samples of $\mathcal{X}$ (i.e. $\mathbf{w}_i = \phi(\mathbf{x}_j)$ for some $1 \le j \le n$), let $\mathbf{W} \in \mathbb{R}^{d\times s}$ denote the matrix consisting of $\mathbf{w}_i$ as the i-th column, and $\mathbf{C} = \mathbf{\Phi}^\top\mathbf{W} \in \mathbb{R}^{n\times s}$ the inner product matrix of the whole data instances and the samples. Since the kernel matrix $\mathbf{K}_W \in \mathbb{R}^{s\times s}$ for the subset W is $\mathbf{K}_W = \mathbf{W}^\top\mathbf{W}$, we can rearrange the rows and columns of $\mathbf{K}$ such that

$$\mathbf{K} = \begin{bmatrix}\mathbf{K}_W & \mathbf{K}_{21}^\top \\ \mathbf{K}_{21} & \mathbf{K}_{22}\end{bmatrix} \text{ and } \mathbf{C} = \begin{bmatrix}\mathbf{K}_W \\ \mathbf{K}_{21}\end{bmatrix}. \tag{1}$$

In the later part of the paper, we will generalize the samples $\mathbf{w}_i$ to arbitrary vectors of dimension d, not necessarily mapped vectors of some data instances.

2.1. The Standard Nyström Method

The standard Nyström method for approximating the kernel matrix $\mathbf{K}$ using the subset W of s sample data instances yields a rank-s' approximation matrix

$$\tilde{\mathbf{K}}^{nys} = \tilde{\mathbf{K}}_{s'}^{nys} = \mathbf{C}\mathbf{K}_W^\dagger\mathbf{C}^\top \approx \mathbf{K},$$

where $\mathbf{K}_W^\dagger$ is the pseudo-inverse of $\mathbf{K}_W$ and $s' = \text{rank}(\mathbf{W})$. The rank-$k$ approximation matrix (with $k \le s'$) is computed by

$$\tilde{\mathbf{K}}_k^{nys} = \mathbf{C}\mathbf{K}_{W,k}^\dagger\mathbf{C}^\top,$$

where $\mathbf{K}_{W,k}^\dagger$ is the pseudo-inverse of $\mathbf{K}_{W,k}$, the best rank-k approximation of $\mathbf{K}_W$, which can be computed from the SVD $\mathbf{K}_{W,k} = \mathbf{V}_{W,k}\mathbf{\Sigma}_{W,k}^2\mathbf{V}_{W,k}^\top$. The time complexity of computing $\tilde{\mathbf{K}}_k^{nys}$ is $O(ksn + s^3)$.

[1] In this paper, we use compact SVD instead of full SVD unless we give a particular mention.

The Nyström method is also used to compute the first k approximate eigenvalues $(\tilde{\mathbf{\Sigma}}_k^{nys})^2$ and the corresponding k approximate eigenvectors $(\tilde{\mathbf{V}}_k^{nys})$ of the kernel matrix $\mathbf{K}$. Using SVD $\mathbf{K}_{W,k} = \mathbf{V}_{W,k}\mathbf{\Sigma}_{W,k}^2\mathbf{V}_{W,k}^\top$, the eigenvalues and eigenvectors are computed by

$$(\tilde{\mathbf{\Sigma}}_k^{nys})^2 = \frac{n}{s}(\mathbf{\Sigma}_{W,k})^2 \text{ and } \tilde{\mathbf{V}}_k^{nys} = \sqrt{\frac{s}{n}}\mathbf{C}\mathbf{V}_{W,k}\mathbf{\Sigma}_{W,k}^{-2}, \tag{2}$$

However in general, $\tilde{\mathbf{K}}_k^{nys}$ is *not* the best rank-k approximation of $\tilde{\mathbf{K}}^{nys}$, *nor* the eigenvectors $\tilde{\mathbf{V}}_k^{nys}$ are orthogonal even though $\mathbf{K}$ is symmetric.

2.2. The One-Shot Nyström Method

A straightforward way to obtain the best rank-k approximation of $\tilde{\mathbf{K}}^{nys}$ would be a two-stage computation, where we first construct the full $\tilde{\mathbf{K}}^{nys}$ and then reduce its rank via SVD. This approach is costly, since its time complexity amounts to performing SVD on the original matrix $\mathbf{K}$.

The one-shot Nyström method (Fowlkes et al., 2004), shown in Alg 1, computes the SVD in a single pass, and thus possesses the following nice property:

Lemma 1 *(Fowlkes et al., 2004)* The Nyström method *using a sample set W of s vectors can be decomposed as*

$$\tilde{\mathbf{K}}^{nys} = \tilde{\mathbf{K}}_{s'}^{nys} = \mathbf{C}\mathbf{K}_W^\dagger\mathbf{C}^\top = \mathbf{G}\mathbf{G}^\top,$$

with $s' = \text{rank}(\mathbf{K}_W)$ and $\mathbf{G} = \mathbf{C}\mathbf{V}_W\mathbf{\Sigma}_W^{-1} \in \mathbb{R}^{n\times s'}$. The one-shot Nyström method *computes SVD $\mathbf{G}^\top\mathbf{G} = \mathbf{V}_G\mathbf{\Sigma}_G^2\mathbf{V}_G^\top$, and computes s' eigenvectors of $\tilde{\mathbf{K}}^{nys}$ by $\tilde{\mathbf{V}}_{s'}^{osn} = \mathbf{G}\mathbf{V}_G\mathbf{\Sigma}_G^{-1}$. Consequently, we obtain the same result as the Nyström method for the rank-s' approximation*

$$\tilde{\mathbf{K}}^{osn} = \tilde{\mathbf{V}}_{s'}^{osn}\mathbf{\Sigma}_G^2(\tilde{\mathbf{V}}_{s'}^{osn})^\top = \mathbf{G}\mathbf{G}^\top = \tilde{\mathbf{K}}^{nys},$$

but better yet, the best rank-k (with $k < s'$) approximation

$$\tilde{\mathbf{K}}_k^{osn} = \tilde{\mathbf{V}}_k^{osn}\mathbf{\Sigma}_{G,k}^2(\tilde{\mathbf{V}}_k^{osn})^\top = (\tilde{\mathbf{K}}^{nys})_k \ne \tilde{\mathbf{K}}_k^{nys}.$$

The time complexity of the one-shot Nyström method is $O(s^2n)$ if the sample set W is a mapped samples of the data set, i.e. $W = \{\mathbf{w}_i | \mathbf{w}_i = \phi(\mathbf{x}_j) \text{ for some } 1 \le j \le n\}$. This is the case when the sample selection matrix $\mathbf{P}$ is a binary matrix with a single one per column. In the remainder of the paper, we will use the notation $\tilde{\mathbf{\Sigma}}_k^{osn}$ for $\mathbf{\Sigma}_{G,k}$ to emphasize that it is obtained from the one-shot Nyström method.

3. The One-Shot Nyström Method: The Optimal Sample-based KPCA

As reviewed in the previous section, the one-shot Nyström method computes the best rank-k approximation

Algorithm 1 The One-shot Nyström method

Input: Matrix $\mathbf{P}_s \in \mathbb{R}^{n\times s}$ representing the composition of s sample points such that $\mathbf{W} = \mathbf{\Phi}\mathbf{P}_s \in \mathbb{R}^{d\times s}$ with $\mathrm{rank}(\mathbf{W}) = s'$, kernel function κ

Output: Approximate kernel matrix $\tilde{\mathbf{K}}_k^{osn}$, its singular vectors $\tilde{\mathbf{V}}_k^{osn}$ and singular values $(\tilde{\mathbf{\Sigma}}_k^{osn})^2$

1: Obtain $\mathbf{K}_W = \mathbf{W}^\top\mathbf{W}$
2: Perform compact SVD $\mathbf{K}_W = \mathbf{V}_W\mathbf{\Lambda}_W\mathbf{V}_W^\top = \mathbf{V}_W\mathbf{\Sigma}_W^2\mathbf{V}_W^\top$
3: Compute $\mathbf{G}^\top\mathbf{G}$, where $\mathbf{G} = \mathbf{C}\mathbf{V}_W\mathbf{\Sigma}_W^{-1}$
4: Compute $\mathbf{V}_{G,k}$ the first k singular vectors of $\mathbf{G}^\top\mathbf{G}$ and corresponding singular values $\mathbf{\Sigma}_{G,k}^2$
5: $\tilde{\mathbf{\Sigma}}_k^{osn} = \mathbf{\Sigma}_{G,k}$, $\tilde{\mathbf{V}}_k^{osn} = \mathbf{G}\mathbf{V}_{G,k}\mathbf{\Sigma}_{G,k}^{-1}$ and $\tilde{\mathbf{K}}_k^{osn} = \mathbf{G}\mathbf{V}_{G,k}\mathbf{V}_{G,k}^\top\mathbf{G}^\top$

of $\tilde{\mathbf{K}}^{nys}$ to obtain orthonormal approximate eigenvectors $\tilde{\mathbf{V}}_k$ of $\mathbf{K}$ for the given samples $\mathbf{W}$. Here, we suggest that the one-shot Nyström method can be used for computing optimal solutions to other closely related problems, such as kernel principal component analysis (KPCA). In this section, we make a formal statement that the one-shot Nyström method provides an optimal KPCA for the given $\mathbf{W}$, which we build on in later sections.

We start with the observation that the most scalable KPCA algorithms are based on a set of samples (Frieze et al., 1998; Williams & Seeger, 2001) can be seen as computing k approximate eigenvectors of $\mathbf{K}$, given by

$$\tilde{\mathbf{V}}_k = \mathbf{C}\mathbf{A}_k\tilde{\mathbf{\Sigma}}_k^{-1}, \tag{3}$$

where $\mathbf{C} = \mathbf{\Phi}^\top\mathbf{W} \in \mathbb{R}^{n\times s}$ is the inner product matrix among the whole data set and the sample vectors (Eqn (1)), $\mathbf{A}_k \in \mathbb{R}^{s\times k}$ is the algorithm-dependent coefficient matrix for each pair of sample vector and principal direction, and $\tilde{\mathbf{\Sigma}}_k$ is the diagonal matrix of the first k approximate singular values. This view generalizes (Kumar et al., 2009).

To facilitate the analysis, we first reformulate Eqn (3) as

$$\tilde{\mathbf{V}}_k = \mathbf{\Phi}^\top\tilde{\mathbf{U}}_k\tilde{\mathbf{\Sigma}}_k^{-1}, \tag{4}$$

where $\tilde{\mathbf{U}}_k = \mathbf{W}\mathbf{A}_k \in \mathbb{R}^{d\times k}$ denotes k approximate principal directions in the feature space. Using the reconstruction error (RE) and the normalized reconstruction error (NRE) for KPCA (Günter et al., 2007)

$$\mathrm{RE}(\tilde{\mathbf{U}}_k) = \|\mathbf{\Phi} - \tilde{\mathbf{U}}_k\tilde{\mathbf{U}}_k^\top\mathbf{\Phi}\|_F$$
$$\mathrm{NRE}(\tilde{\mathbf{U}}_k) = \frac{\|\mathbf{\Phi} - \tilde{\mathbf{U}}_k\tilde{\mathbf{U}}_k^\top\mathbf{\Phi}\|_F}{\|\mathbf{\Phi}\|_F}$$

as the objective functions, and observing that $\mathbf{A}_k$ determines the k approximate principal directions $\tilde{\mathbf{U}}_k$, we can formulate KPCA based on samples $\mathbf{W}$ as an optimization problem:

Definition 1 *For the given samples $\mathbf{W}$, sample-based KPCA problem is defined as*

$$\underset{\mathbf{A}_k}{\text{minimize}}\ \mathrm{NRE}(\tilde{\mathbf{U}}_k) \text{ subject to } \tilde{\mathbf{U}}_k^\top\tilde{\mathbf{U}}_k = \mathbf{I}_k, \tilde{\mathbf{U}}_k = \mathbf{W}\mathbf{A}_k. \tag{5}$$

The following lemma provides two types of the approximate principal directions $\tilde{\mathbf{U}}_k$, which are computed from the standard Nyström method and one-shot Nyström method.

Lemma 2 *In standard Nyström method, approximate principal directions are*

$$\tilde{\mathbf{U}}_k^{nys} = \mathbf{U}_{W,k}, \tag{6}$$

where $\mathbf{W} = \mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top$. In the one-shot Nyström method, approximate principal directions are

$$\tilde{\mathbf{U}}_k^{osn} = \mathbf{U}_W\mathbf{V}_{G,k}, \tag{7}$$

where $\mathbf{G} = \mathbf{\Phi}^\top\mathbf{W}\mathbf{V}_W\mathbf{\Sigma}_W^{-1} = \mathbf{\Phi}^\top\mathbf{U}_W$ and $\mathbf{G}^\top\mathbf{G} = \mathbf{V}_G\mathbf{\Sigma}_G\mathbf{V}_G^\top$.

The following theorem states that the one-shot Nyström method can be used to solve the optimization problem defined in Def 1.

Theorem 1 *Given the s samples $\mathbf{W} \in \mathbb{R}^{d\times s}$ with $\mathrm{rank}(\mathbf{W}) = s'$, KPCA using the one-shot Nyström method solves the optimization problem in Def 1.*

Given samples $\mathbf{W}$, we proved that the one-shot Nyström method minimizes the $\mathrm{NRE}(\tilde{\mathbf{U}}_k)$ in Eqn (5) in Def 1. To give more intuition for it, we introduce the sum of eigenvalue errors which is closely related with the $\mathrm{NRE}(\tilde{\mathbf{U}}_k)$.

Definition 2 *Given k approximate principal directions $\tilde{\mathbf{U}}_k \in \mathbb{R}^{d\times k}$ of $\mathbf{\Phi}$ such that $\tilde{\mathbf{U}}_k^\top\tilde{\mathbf{U}}_k = \mathbf{I}_k$, the sum of eigenvalue errors from $\tilde{\mathbf{U}}_k$ is defined as*

$$\epsilon_1(\tilde{\mathbf{U}}_k) = \mathrm{tr}(\mathbf{U}_k^\top\mathbf{\Phi}\mathbf{\Phi}^\top\mathbf{U}_k) - \mathrm{tr}(\tilde{\mathbf{U}}_k^\top\mathbf{\Phi}\mathbf{\Phi}^\top\tilde{\mathbf{U}}_k),$$

where $\mathbf{U}_k$ is the matrix consisting of true principal directions as columns.

With the notion in Def 2, we can directly give a following corollary.

Corollary 1 *Minimizing the $\mathrm{NRE}(\tilde{\mathbf{U}}_k)$ is equivalent to minimizing the $\epsilon_1(\tilde{\mathbf{U}}_k)$ defined in Def 2, thus*

$$\tilde{\mathbf{U}}_k^{osn} = \underset{\tilde{\mathbf{U}}_k}{\mathrm{argmin}}\, \epsilon_1(\tilde{\mathbf{U}}_k) \text{ s.t. } \tilde{\mathbf{U}}_k^\top\tilde{\mathbf{U}}_k = \mathbf{I}_k, \tilde{\mathbf{U}}_k = \mathbf{W}\mathbf{A}_k.$$

Additional to Thm 1 and Cor 1, we show that outputs of the one-shot Nyström method depend only on subspace spanned by input samples.

Proposition 1 *Let $\mathbf{W}_1$ and $\mathbf{W}_2$ be the matrix consisting of s_1 samples and s_2 samples respectively. If two column spaces $col(\mathbf{W}_1)$ and $col(\mathbf{W}_2)$ are the same, then the outputs of the one-shot Nyström method are also the same regardless of difference between set of samples.*

We note that the standard Nyström method does not satisfy the robustness discussed in Proposition 1.

4. A Generalized View of Sampling Schemes

Motivated by Proposition 1, in this section, we provide new upper error bounds of the Nyström method based on subspace distance, and suggest a generalized view of sampling schemes for the Nyström method.

4.1. Error Analysis based on Subspace Distance

To provide new upper error bounds of the Nyström methods, our motivation is using the measure called "subspace distance" which can evaluate the difference between two subspaces (Wang et al., 2006; Sun et al., 2007).

Basically, the subspace distance of two subspaces depends on the notion of projection error, hence we discuss the projection error first.

Definition 3 *Given two matrices $\mathbf{U}$ and $\mathbf{V}$ consisting of orthonormal vectors, i.e., $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$, the projection error of $\mathbf{U}$ onto $\mathrm{col}(\mathbf{V})$ is defined as*

$$\mathrm{PE}(\mathbf{U}, \mathbf{V}) = \|\mathbf{U} - \mathbf{V}\mathbf{V}^\top\mathbf{U}\|_F.$$

Since any linear subspace can be represented by its orthonormal basis, subspace distance can be defined by the projection error between set of orthonormal vectors.

Definition 4 *(Wang et al., 2006; Golub & Van Loan, 2012) Given k dimensional subspace S_1 and k dimensional subspace S_2, the subspace distance $\mathrm{d}(S_1, S_2)$ is defined as*

$$\mathrm{d}(S_1, S_2) = \mathrm{PE}(\mathbf{U}_k, \tilde{\mathbf{U}}_k) = \|\mathbf{U}_k - \tilde{\mathbf{U}}_k\tilde{\mathbf{U}}_k^\top\mathbf{U}_k\|_F,$$

where $\mathbf{U}_k$ is an orthonormal basis of S_1 and $\tilde{\mathbf{U}}_k$ is an orthonormal basis of S_2.

Lemma 3 *(Wang et al., 2006; Sun et al., 2007; Golub & Van Loan, 2012) The subspace distances defined in Def 4 are invariant to the choice of orthonormal basis.*

Remind that the standard Nyström and one-shot Nyström methods satisfy $\tilde{\mathbf{K}}_k = \tilde{\mathbf{V}}_k\tilde{\mathbf{\Sigma}}_k^2\tilde{\mathbf{V}}_k = \mathbf{\Phi}^\top\tilde{\mathbf{U}}_k\tilde{\mathbf{U}}_k^\top\mathbf{\Phi}$, and those are characterized by $\tilde{\mathbf{U}}_k^{nys}$ and $\tilde{\mathbf{U}}_k^{osn}$ as proved in Lem 2. Therefore, we provide a following error analysis based on the subspace distance between $\mathrm{col}(\mathbf{U}_k)$ and $\mathrm{col}(\tilde{\mathbf{U}}_k)$.

Theorem 2 *Let $\tilde{\mathbf{U}}_k$ be a matrix consisting of k approximate principal directions computed by the Nyström methods given the sample matrix $\mathbf{W} \in \mathbb{R}^{d\times s}$ with $\mathrm{rank}(\mathbf{W}) \geq k$. Suppose that $\epsilon_0(\tilde{\mathbf{U}}_k) = \mathrm{d}(\mathrm{col}(\mathbf{U}_k), \mathrm{col}(\tilde{\mathbf{U}}_k))$, then the* NRE *is bounded by*

$$\mathrm{NRE}(\tilde{\mathbf{U}}_k) \leq \mathrm{NRE}(\mathbf{U}_k) + \sqrt{2}\epsilon_0,$$

where $\mathrm{NRE}(\mathbf{U}_k)$ is the optimal NRE *for rank-k. The error of the approximate kernel matrix is bounded by*

$$\|\mathbf{K} - \tilde{\mathbf{K}}_k\|_F \leq \|\mathbf{K} - \mathbf{K}_k\|_F + \sqrt{2}\epsilon_0\,\mathrm{tr}(\mathbf{K}),$$

where $\|\mathbf{K} - \mathbf{K}_k\|_F$ is the optimal error for rank-k.

The suggested upper error bounds in Thm 2 are applicable both to the standard and one-shot Nyström methods.

We note that the $\epsilon_1(\tilde{\mathbf{U}}_k)$ is bounded on both sides by constant times of the subspace distance $\mathrm{d}(\mathrm{col}(\mathbf{U}_k), \mathrm{col}(\tilde{\mathbf{U}}_k))$.

Lemma 4 *Suppose that k-th eigengap is nonzero given Gram matrix $\mathbf{K}$, i.e., $\gamma_k = \lambda_k - \lambda_{k+1} > 0$. Then, given the $\tilde{\mathbf{U}}_k \in \mathbb{R}^{d\times k}$ and $\tilde{\mathbf{V}}_k \in \mathbb{R}^{n\times k}$ such that $\tilde{\mathbf{U}}_k^\top\tilde{\mathbf{U}}_k = \mathbf{I}_k$ and $\tilde{\mathbf{V}}_k^\top\tilde{\mathbf{V}}_k = \mathbf{I}_k$, the subspace distance is bounded by*

$$\sqrt{\frac{\epsilon_1(\tilde{\mathbf{U}}_k)}{\lambda_1}} \leq \mathrm{d}(\mathrm{col}(\mathbf{U}_k), \mathrm{col}(\tilde{\mathbf{U}}_k)) \leq \sqrt{\frac{\epsilon_1(\tilde{\mathbf{U}}_k)}{\gamma_k}},$$
$$\sqrt{\frac{\epsilon_2(\tilde{\mathbf{V}}_k)}{\lambda_1}} \leq \mathrm{d}(\mathrm{col}(\mathbf{V}_k), \mathrm{col}(\tilde{\mathbf{V}}_k)) \leq \sqrt{\frac{\epsilon_2(\tilde{\mathbf{V}}_k)}{\gamma_k}},$$

where $\epsilon_2(\tilde{\mathbf{V}}_k) = \mathrm{tr}(\mathbf{V}_k^\top\mathbf{\Phi}^\top\mathbf{\Phi}\mathbf{V}_k) - \mathrm{tr}(\tilde{\mathbf{V}}_k^\top\mathbf{\Phi}^\top\mathbf{\Phi}\tilde{\mathbf{V}}_k)$.

Lem 4 tells us that the subspace distance goes to zero as $\epsilon_1(\tilde{\mathbf{U}}_k)$ goes to zero, and the converse is also true, like the squeeze theorem. Thus, we can replace $\epsilon_0(\tilde{\mathbf{U}}_k)$ in Thm 2 with $\epsilon_1(\tilde{\mathbf{U}}_k)$.

We can also provide a connection between $\epsilon_1(\tilde{\mathbf{U}}_k)$ and $\epsilon_2(\tilde{\mathbf{V}}_k)$ when we set $\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_\ell$ for the Nyström methods.

Lemma 5 *Suppose that ℓ samples are columns of $\mathbf{\Phi}\tilde{\mathbf{V}}_\ell$, i.e.$\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_\ell$, and $\tilde{\mathbf{V}}_k$ is a submatrix consisting of k columns of $\tilde{\mathbf{V}}_\ell$, where $\tilde{\mathbf{V}}_\ell^\top\tilde{\mathbf{V}}_\ell = \mathbf{I}_\ell$. Then, for any $k \leq \mathrm{rank}(\mathbf{W})$, $\tilde{\mathbf{U}}_k^{nys}$ and $\tilde{\mathbf{U}}_k^{osn}$ satisfy*

$$\epsilon_1(\tilde{\mathbf{U}}_k^{osn}) \leq \epsilon_1(\tilde{\mathbf{U}}_k^{nys}) \leq \epsilon_2(\tilde{\mathbf{V}}_k) \leq \epsilon_2(\tilde{\mathbf{V}}_\ell),$$

where $\tilde{\mathbf{U}}_k^{nys}$ and $\tilde{\mathbf{U}}_k^{osn}$ are defined in Lem 2.

Based on Thm 2, Lem 4 and Lem 5, we provide Thm 3 and Rem 1, that tell us how we can get sample vectors for Nyström methods to reduce their approximation errors.

Theorem 3 *Suppose that the k-th eigengap γ_k is nonzero given $\mathbf{K}$. If we set $\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_\ell$ with $\tilde{\mathbf{V}}_\ell^\top \tilde{\mathbf{V}}_\ell = \mathbf{I}_\ell$, then by the standard and one-shot Nyström methods, the* NRE *and the matrix approximation error are bounded as follows:*

$$\mathrm{NRE}(\tilde{\mathbf{U}}_k) \leq \mathrm{NRE}(\mathbf{U}_k) + \sqrt{\frac{2\epsilon_2(\tilde{\mathbf{V}}_k)}{\gamma_k}} \tag{8}$$

$$\|\mathbf{K} - \tilde{\mathbf{K}}_k\|_F \leq \|\mathbf{K} - \mathbf{K}_k\|_F + \sqrt{\frac{2\epsilon_2(\tilde{\mathbf{V}}_k)}{\gamma_k}}\,\mathrm{tr}(\mathbf{K}) \tag{9}$$

where $\tilde{\mathbf{V}}_k$ is any submatrix consisting of k columns of $\tilde{\mathbf{V}}_\ell$.

Remark 1 *By Thm 3, we suggest two kinds of strategies:*

- *Since $\epsilon_2(\tilde{\mathbf{V}}_k) \leq \epsilon_2(\tilde{\mathbf{V}}_\ell)$, if we set $\mathbf{W}$ as $\mathbf{\Phi}\tilde{\mathbf{V}}_\ell$ which has small $\epsilon_2(\tilde{\mathbf{V}}_\ell)$ or $\min_{\tilde{\mathbf{V}}_k} \epsilon_2(\tilde{\mathbf{V}}_k)$ with constraint $\tilde{\mathbf{V}}_\ell^\top \tilde{\mathbf{V}}_\ell = \mathbf{I}_\ell$, then we could get a small error induced by Nyström methods due to a short subspace distance to the principal subspace. The objective of the kernel K-means is minimizing the $\epsilon_2(\tilde{\mathbf{V}}_\ell)$ with some constraints, which will be discussed in detail in the supplementary material.*
- *Also, if we set $\mathbf{W}$ as $\mathbf{\Phi}\tilde{\mathbf{V}}_\ell$ which has small $\mathrm{PE}(\mathbf{V}_k, \tilde{\mathbf{V}}_\ell)$ or $\min_{\tilde{\mathbf{V}}_k} \mathrm{PE}(\mathbf{V}_k, \tilde{\mathbf{V}}_k)$ with constraint $\tilde{\mathbf{V}}_\ell^\top \tilde{\mathbf{V}}_\ell = \mathbf{I}_\ell$, then we could get a small error induced by Nyström methods due to a short subspace distance to the principal subspace. The leverage score sampling reduces the expectation of $\min_{\tilde{\mathbf{V}}_k} \mathrm{PE}(\mathbf{V}_k, \tilde{\mathbf{V}}_k)$.*

4.2. Capturing Approximate Principal Subspace (CAPS)

As discussed in Rem 1, minimizing $\epsilon_2(\tilde{\mathbf{V}}_\ell)$ or $\min_{\tilde{\mathbf{V}}_k} \epsilon_2(\tilde{\mathbf{V}}_k)$ is a key of reducing subspace distance $\mathrm{d}\left(\mathrm{col}(\mathbf{U}_k), \mathrm{col}(\tilde{\mathbf{U}}_k)\right)$ and can be a good objective of sampling methods for Nyström methods. Thus, our goal of this section is suggesting an algorithm of minimizing $\epsilon_2(\tilde{\mathbf{V}}_\ell)$.

Suggesting an efficient algorithm for minimizing $\epsilon_2(\tilde{\mathbf{V}}_\ell)$, we introduce the notion of spanning set S defined in Def 5, which can be utilized to approximate linear combinations such that approximated eigenvectors lie in the $\mathrm{col}(\mathbf{S})$, *e.g.*

$$\tilde{\mathbf{v}}_j \approx \sum_{\phi(\mathbf{x}_i) \in S} b_{ij}\phi(\mathbf{x}_i) \text{ for } j \in \{1, ..., \ell\}, \tag{10}$$

where b_{ij} is a coefficient.

Definition 5 *Given n data points, let S be a spanning set consisting of s representative points in n data points for linear combination, $\mathbf{S}$ be a matrix which consists of s representative vectors as its columns, and $\mathbf{T}_S$ be a indicator matrix such that $\mathbf{S} = \mathbf{\Phi}\mathbf{T}_S$.*

Algorithm 2 Capturing Approximate Principal Subspace (CAPS)

Input: The number of representatives s, where $\ell \ll s \ll n$

Output: Spanning set S consisting of s representative points, $\tilde{\mathbf{V}}_\ell = \mathbf{T}_S\mathbf{V}_{S,\ell}$ (or $\tilde{\mathbf{V}}_\ell = \mathbf{T}_S\tilde{\mathbf{V}}_{S,\ell}$), ℓ samples $\mathbf{W} = \mathbf{S}\mathbf{V}_{S,\ell}$ (respectively,$\mathbf{W} = \mathbf{S}\tilde{\mathbf{V}}_{S,\ell}$)

1: Construct a spanning set S consisting of s representative points which are obtained by column index sampling (*e.g.*, uniform random or approximate leverage score etc.)
2: Obtain $\mathbf{K}_S = \mathbf{S}^\top\mathbf{S}$
3: Perform compact SVD $\mathbf{K}_S = \mathbf{V}_S\mathbf{\Sigma}_S^2\mathbf{V}_S^\top$ or approximate compact SVD (*e.g.* randomized SVD or the one-shot Nyström method)
4: Obtain $\tilde{\mathbf{V}}_\ell$ as $\mathbf{T}_S\mathbf{V}_{S,\ell}$ or $\mathbf{T}_S\tilde{\mathbf{V}}_{S,\ell}$, where $\mathbf{T}_S$ is the indicator matrix for the set S

If we express approximate ℓ eigenvectors as described in Eqn (10) and set $s \ll n$ for very large-scale data, then the time complexities of computing $\tilde{\mathbf{V}}_\ell$ will be reduced.

Here is our strategy which is called "Capturing Approximate Principal Subspace" (CAPS).

1. For the scalability, we construct and utilize a spanning set S defined in Def 5, and set a constraint as $\mathrm{col}(\mathbf{W}) \subset \mathrm{col}(\mathbf{S})$. Applying the constraint to our Rem 1, then we have
$$\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_\ell = \mathbf{S}\mathbf{A}_\ell \text{ with } \mathbf{A}_\ell^\top\mathbf{A}_\ell = \mathbf{I}_\ell. \tag{11}$$
2. Under the condition in Eqn (11), the solution of the problem of minimizing $\epsilon_2(\tilde{\mathbf{V}}_\ell)$ is $\mathbf{V}_{S,\ell}$ by the Proposition 2. Thus, we compute $\mathbf{V}_{S,\ell}$ via SVD of $\mathbf{K}_S$, or $\tilde{\mathbf{V}}_{S,\ell}$ through approximate SVD including randomized SVD or the one-shot Nyström method.

Consequently, CAPS aims to get a small $\epsilon_2(\tilde{\mathbf{V}}_\ell)$ more directly with just $O(sn)$ memory, where $s \ll n$. The time complexity varies depending on step 1 and step 3 in Alg 2. For decomposing $\mathbf{K}_S$ in step 3, the time complexity is $O(s^3)$ for SVD and $O(m^2 s)$ for the one-shot Nyström method, where $\ell \leq m \ll s$.

Proposition 2 *Given spanning set S consisting of s representative points, suppose that we set $\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_\ell$ and $\tilde{\mathbf{V}}_\ell^\top\tilde{\mathbf{V}}_\ell = \mathbf{I}_\ell$ with the constraint $\mathrm{col}(\mathbf{W}) \subset \mathrm{col}(\mathbf{S})$. Then, under that condition, the problem of minimizing $\epsilon_2(\mathbf{V}_\ell)$ can be equivalently expressed as*

$$\underset{\mathbf{A}_\ell}{\mathrm{minimize}}\, \epsilon_2(\tilde{\mathbf{V}}_\ell) \text{ subject to } \tilde{\mathbf{V}}_\ell = \mathbf{T}_S\mathbf{A}_\ell, \mathbf{A}_\ell^\top\mathbf{A}_\ell = \mathbf{I}_\ell,$$

and the output of step 3 in Alg 2 with rank-ℓ SVD minimizes $\epsilon_2(\mathbf{V}_\ell)$, i.e., $\mathbf{V}_{S,\ell} = \mathrm{argmin}_{\mathbf{A}_\ell} \epsilon_2(\tilde{\mathbf{V}}_\ell)$ subject to $\tilde{\mathbf{V}}_\ell = \mathbf{T}_S\mathbf{A}_\ell, \mathbf{A}_\ell^\top\mathbf{A}_\ell = \mathbf{I}_\ell$, where $\mathbf{K}_S = \mathbf{V}_S\mathbf{\Sigma}_S^2\mathbf{V}_S^\top$.

We directly provide Cor 2 which is a revised version of Thm 3 for CAPS sampling.

Corollary 2 *By standard and one-shot Nyström methods, any set of ℓ samples computed by CAPS sampling satisfying $\epsilon_2(\tilde{\mathbf{V}}_\ell)$ error is guaranteed to satisfy Eqn (8) and Eqn (9).*

Since our main concern is minimizing $\epsilon_2(\tilde{\mathbf{V}}_\ell)$, and the CAPS with rank-ℓ SVD gives the optimal solution for the given condition in Proposition 2 as $\tilde{\mathbf{V}}_\ell = \mathbf{T}_S \mathbf{V}_{S,\ell}$, we can approximate $\epsilon_2(\mathbf{T}_S \mathbf{V}_{S,\ell})$ as $\epsilon_2(\mathbf{T}_S \tilde{\mathbf{V}}_{S,\ell})$, where $\tilde{\mathbf{V}}_{S,\ell}$ can be computed by one-shot Nyström method due to its optimality discussed in Thm 1, or can be obtained from randomized SVD.

5. The Double Nyström Method

In this section, we propose a new framework of Nyström method based on the CAPS and the one-shot Nyström method, which is called "Double Nyström Method" and described in Alg 3. In brief, the Double Nyström method reduces the original problem of decomposing the given $n \times n$ kernel matrix to the problem of decomposing a $s \times s$ matrix, and again reduces it to the problem of decomposing a $\ell \times \ell$ matrix.

1. In the first part, we select the one-shot Nyström method for step 3 in Alg 2, and run CAPS using the one-shot Nyström method to compute $\mathbf{V}_{S,\ell}$. Because we can consider the problem of computing $\mathbf{V}_{S,\ell}$ as the KPCA problem, and the one-shot Nyström method solves sample-based KPCA defined in Def 1. Also, it has a small running time complexity $O(m^2 s)$, since $\ell \leq m \ll s \ll n$.

2. For the second part, we consider $\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_\ell = \mathbf{S}\tilde{\mathbf{V}}_{S,\ell}$, and run the one-shot Nyström method. Since computed $\tilde{\mathbf{V}}_\ell = \mathbf{T}_S \tilde{\mathbf{V}}_{S,\ell}$ in the first part induces a small $\epsilon_2(\tilde{\mathbf{V}}_\ell)$, we may have an accurate $\tilde{\mathbf{K}}_k$ after the second part.

Constructing a spanning set S in CAPS by using uniform random sampling, the total time complexity of double Nyström methods is $O(\ell s n + m^2 s)$, and $O(m^2 s)$ term is not considerable compared to $O(\ell s n)$, since $\ell \leq m \ll s \ll n$.

We note that computing spanning set S is also important for CAPS, consequently for Double Nyström method. In Rem 2, We provide an example how we can quickly compute approximate leverage scores and construct a spanning set S. Also, we summarize the time complexity of the Nyström methods in Tbl 1.

Algorithm 3 The Double Nyström Method

Input: Kernel function κ, and the parameters k, ℓ, m, s, where $k \leq \ell \leq m \ll s \ll n$

Output: Approximate kernel matrix and spectral decomposition

1: Run CAPS using the one-shot Nyström method with m subsamples of spanning set S, and obtain $\tilde{\mathbf{V}}_\ell$ and $\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_\ell = \mathbf{S}\tilde{\mathbf{V}}_{S,\ell}$

2: Compute $\mathbf{K}_W = (\tilde{\mathbf{V}}_{S,\ell})^\top \mathbf{K}_S \tilde{\mathbf{V}}_{S,\ell} \in \mathbb{R}^{\ell \times \ell}$ and $\mathbf{C} = \mathbf{C}_0 \tilde{\mathbf{V}}_{S,\ell} \in \mathbb{R}^{n \times \ell}$ by using $\mathbf{W}$ and $\mathbf{C}_0 = \mathbf{\Phi}^\top \mathbf{S}$, and run the one-shot Nyström method with parameters k and ℓ

Remark 2 *Since the computational complexity for computing the exact leverage scores is high, we can obtain approximate leverage scores by using double Nyström method or other methods.*

First, we obtain s_1 instances by uniform random sampling and construct a spanning set S_1, where $s_1 \leq s$. Next, we run double Nyström method with the spanning set S_1 and approximate leverage scores in time $O(\ell_1 s_1 n + m_1^2 s_1)$. We sample additional $(s - s_1)$ instances to complete constructing a spanning set S by using the computed scores, where $\ell_1 \leq m_1 \ll s_1 \leq s \ll n$ and $S_1 \subseteq S$. If we run double Nyström method with the computed spanning set S, then the total running time is $O(\ell s n + m^2 s)$ for $\ell_1 = \Theta(\ell)$ and $m_1 = \Theta(m)$.

6. Experiments

In this section, we present experimental results that demonstrate our theoretical work and algorithms. We conduct experiments with the measure called "relative approximation error" (Relative Error): Relative Error $= \|\mathbf{K} - \tilde{\mathbf{K}}_k\|_F / \|\mathbf{K}\|_F$. We report the running time as the sum of the sampling time and the Nyström approximation time. Every experimental instances are run on MATLAB R2012b with Intel Xeon 2.90GHz CPUs, 96GB RAM, and 64bit CentOS system.

We choose 5 real data sets for performance comparisons and summarize them in Tbl 2. To construct kernel matrix $\mathbf{K}$, we use radial basis function(RBF) and it is defined as follows: $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right)$, where σ is a kernel parameter. We set σ for 5 data sets as follows: $\sigma = 100.0$ for Dexter, $\sigma = 1.0$ for Letter, $\sigma = 5.0$ for MNIST, $\sigma = 0.3$ for MiniBooNE, and $\sigma = 1.0$ for Covertype. We select $k = 20$ and $k = 50$ for each data set.

We empirically compare the double Nyström method described in Alg 3 with three representative Nyström methods: the standard Nyström method (Williams & Seeger, 2001), the standard Nyström method using randomized SVD (Li et al., 2015), and the one-shot Nyström method

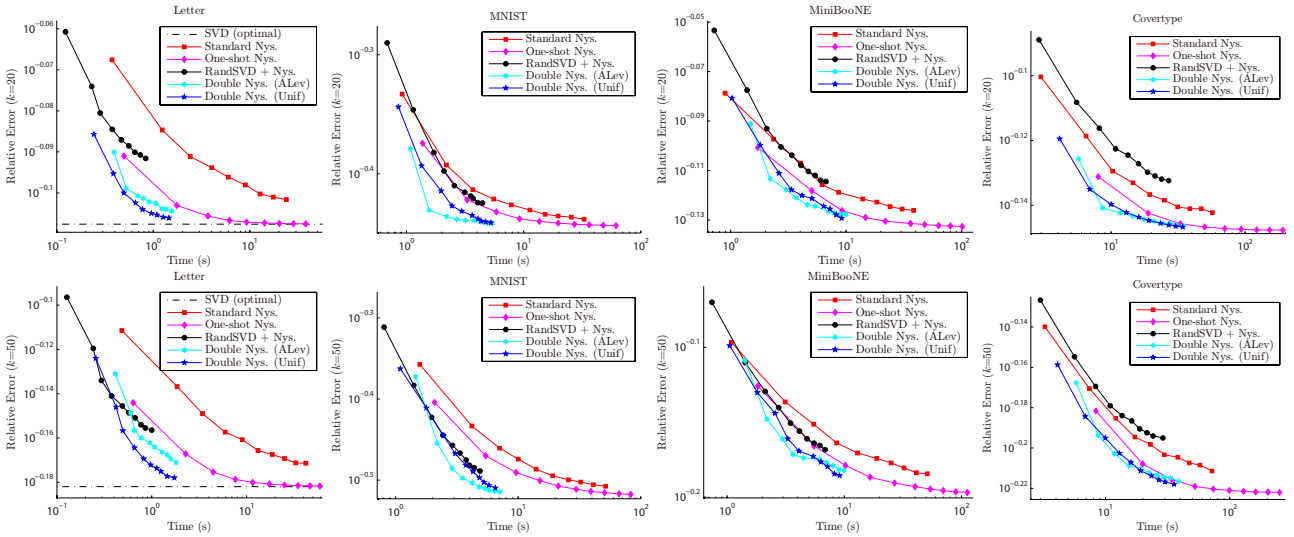

Figure 1. Performance comparison both for $k = 20$ and $k = 50$ among the four methods: the standard Nyström method (Williams & Seeger, 2001), the one-shot Nyström method (Fowlkes et al., 2004), the standard Nyström method using randomized SVD (Li et al., 2015), and the double Nyström method (ours). We gradually increase the number of samples s as 500, 1000, 1500,..., 5000, and there are corresponding 10 points on the each line. We perform SVD algorithm only on the Letter data set due to memory limit.

Table 1. Time complexities for the Nyström methods to obtain a rank-k approximation with spanning set S, where $\ell \leq m \ll s \ll n$ and CAPS(ALev) is described in Rem 2

The Sampling & Nyström Methods	time complexity	linearity for s	degree for s and n	#kernel elements for computation
Unif & The Standard	$O(ksn + s^3)$	No	cubic	$O(sn)$
Unif & The One-Shot	$O(s^2 n)$	No	cubic	$O(sn)$
Unif & Rand.SVD + The Standard	$O(ksn)$	**Yes**	**quadratic**	$O(sn)$
The Double (CAPS(Unif))	$O(\ell sn + m^2 s)$	**Yes**	**quadratic**	$O(sn)$
The Double (CAPS(ALev))	$O(\ell sn + m^2 s)$	**Yes**	**quadratic**	$O(sn)$

Table 2. The summary of 5 real data sets. n is the number of instances and d_0 is the dimension of the original data

data set	number of instances n	dimensionality d_0
Dexter	2600	20000
Letter	20000	16
MNIST	60000	784
MiniBooNE	130064	50
Covertype	581012	54

(Fowlkes et al., 2004). We run the double Nyström method with the spanning set S constructed by uniform random sampling (Unif) and approximate leverage scores (ALev). There are 10 episodes for each test, and 10 points on the each line in the figures. For example, we set $s = 500t$, $\ell = (140 + 5t)$, and $m = (250 + 50t)$ when $n \geq 20000$, where $t = 1, 2, ..., 10$. We display the experimental results in Fig 1 and 2. As shown in the experiments, the double Nyström method always shows better efficiency than other methods under the same condition of using $O(sn)$ kernel elements. In the experiment on the Letter data set, we can also notice that the error of the double Nyström method more rapidly decreases to the optimal error than the others.

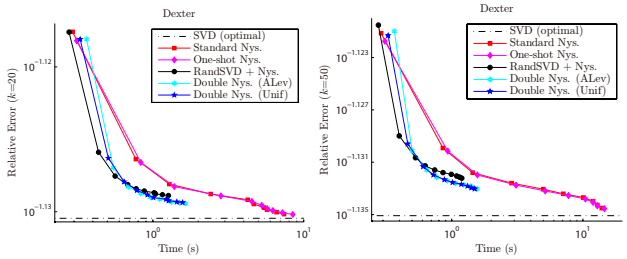

Figure 2. Additional experiments for high dimensional data set. We gradually increase s as 200, 400, 600,..., 2000.

7. Conclusion

In this paper, we provided a comprehensive analysis of the one-shot Nyström method and a generalized view of sampling strategy, and by integrating of these two results, we proposed the "Double Nyström Method" which reduces the size of the decomposition problem to a smaller size in two stages. Both theoretically and empirically, we demonstrated that the double Nyström method is much more efficient than the various Nyström methods, but is quite accurate. Thus, we recommend using the double Nyström method for large-scale data sets.

Acknowledgments

W. Lim acknowledges support from the KAIST Graduate Research Fellowship via Kim-Bo-Jung Fund. K. Jung acknowledges support from the Brain Korea 21 Plus Project in 2015.

References

Deshpande, Amit, Rademacher, Luis, Vempala, Santosh, and Wang, Grant. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2:225–247, 2006.

Drineas, Petros and Mahoney, Michael W. On the Nyström method for approximating a gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, 6:2153–2175, 2005.

Drineas, Petros, Magdon-Ismail, Malik, Mahoney, Michael W., and Woodruff, David P. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

Fine, Shai and Scheinberg, Katya. Efficient svm training using low-rank kernel representations. *The Journal of Machine Learning Research*, 2:243–264, 2002.

Fowlkes, Charless, Belongie, Serge, Chung, Fan, and Malik, Jitendra. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.

Frieze, Alan, Kannan, Ravi, and Vempala, Santosh. Fast monte-carlo algorithms for finding low-rank approximations. In *Proceedings of FOCS*, pp. 370–378. IEEE, 1998.

Gittens, Alex and Mahoney, Michael W. Revisiting the Nyström method for improved large-scale machine learning. In *Proceedings of ICML*, 2013.

Golub, Gene H and Van Loan, Charles F. *Matrix computations*, volume 3. JHU Press, 2012.

Günter, S., Schraudolph, N., and Vishwanathan, S.V.N. Fast iterative kernel principal component analysis. *Journal of Machine Learning Research*, 8, 2007.

Hsieh, Cho-Jui, Si, Si, and Dhillon, Inderjit S. Fast prediction for large-scale kernel machines. In *Proceeding of NIPS*, pp. 3689–3697, 2014.

Kumar, Sanjiv, Mohri, Mehryar, and Talwalkar, Ameet. On sampling-based approximate spectral decomposition. In *Proceedings of ICML*, pp. 553–560. ACM, 2009.

Kumar, Sanjiv, Mohri, Mehryar, and Talwalkar, Ameet. Sampling methods for the Nyström method. *The Journal of Machine Learning Research*, 98888:981–1006, 2012.

Li, Mu, Bi, Wei, Kwok, James T, and Lu, B-L. Large-scale Nyström kernel matrix approximation using randomized svd. *Neural Networks and Learning Systems, IEEE Transactions on*, 26(1):152–165, 2015.

Mahoney, Michael W. and Drineas, Petros. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

Sun, Xichen, Wang, Liwei, and Feng, Jufu. Further results on the subspace distance. *Pattern recognition*, 40 (1):328–329, 2007.

Talwalkar, Ameet, Kumar, Sanjiv, and Rowley, Henry. Large-scale manifold learning. In *Proceedings of CVPR*, pp. 1–8. IEEE, 2008.

Wang, Liwei, Wang, Xiao, and Feng, Jufu. Subspace distance analysis with application to adaptive bayesian algorithm for face recognition. *Pattern Recognition*, 39(3): 456–464, 2006.

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

Williams, Christopher and Seeger, Matthias. Using the Nyström method to speed up kernel machines. In *Proceedings of NIPS*, 2001.

Zhang, Kai and Kwok, James T. Clustered Nyström method for large scale manifold learning and dimension reduction. *IEEE Transactions on Neural Networks*, pp. 1576–1587, 2010.

Zhang, Kai, Tsang, Ivor W, and Kwok, James T. Improved Nyström low-rank approximation and error analysis. In *Proceedings of ICML*, pp. 1232–1239. ACM, 2008.