Double Nyström Method: An Efficient and Accurate Nyström Scheme for Large-Scale Data Sets (Supplementary Material)

Contents

Table 3: The summary of 7 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	σ	k
Dexter	2600	20000	100.0	20,50
WineZ	4898	11	1.0, 3.0	20
AbaloneZ	4177	8	1.0, 3.0	20
Letter	20000	16	1.0, 3.0	20,50
MNIST	60000	784	5.0, 10.0	20,50
MiniBooNE	130064	50	0.3, 1.0	20,50
Covertype	581012	54	1.0	20,50

A Additional Experiments

In this section, we present additional experimental results that demonstrate our theoretical work and algorithms. 7 real data sets, σ and rank-k, which are we used in the experiments, are summarized in Tbl 3. We adopt two measures for experiments: "relative approximation error" (Relative Error), and "normalized approximation error" (Normalized Error)

$$\text{Relative Error} = \|\mathbf{K} - \tilde{\mathbf{K}}_k\|_F / \|\mathbf{K}\|_F \tag{12}$$

$$\text{Normalized Error} = \|\mathbf{K} - \tilde{\mathbf{K}}_k\|_F / \|\mathbf{K} - \mathbf{K}_k\|_F, \tag{13}$$

where the minimum of the normalized error is 1.

A.1 Efficiency and Accuracy

In this section, we empirically compare the double Nyström method described in Alg 3 with four representative Nyström methods: the standard Nyström method (Williams & Seeger, 2001), the standard Nyström method using randomized SVD (Li et al., 2015), the one-shot Nyström method (Fowlkes et al., 2004), and the standard Nyström method using K-means sampling (Zhang & Kwok, 2010). 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 there are 10 points on the each line in the figures. We set $s = 500t$, $\ell = (140 + 5t)$, and $m = (250 + 50t)$ when $n \geq 20000$, where $t = 1, 2, ..., 10$. For Dexter data, we set $s = 200t$, $\ell = (100 + 5t)$, and $m = (150 + 20t)$, where $t = 1, 2, ..., 10$.

A.1.1 Different rank-k

For Nyström approximation, we select two different rank-k which are 20 and 50. We display the experimental results in Fig 3. Regardless of rank-k, Fig 3 shows that 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 approximation more rapidly decreases to the optimal error than the others.

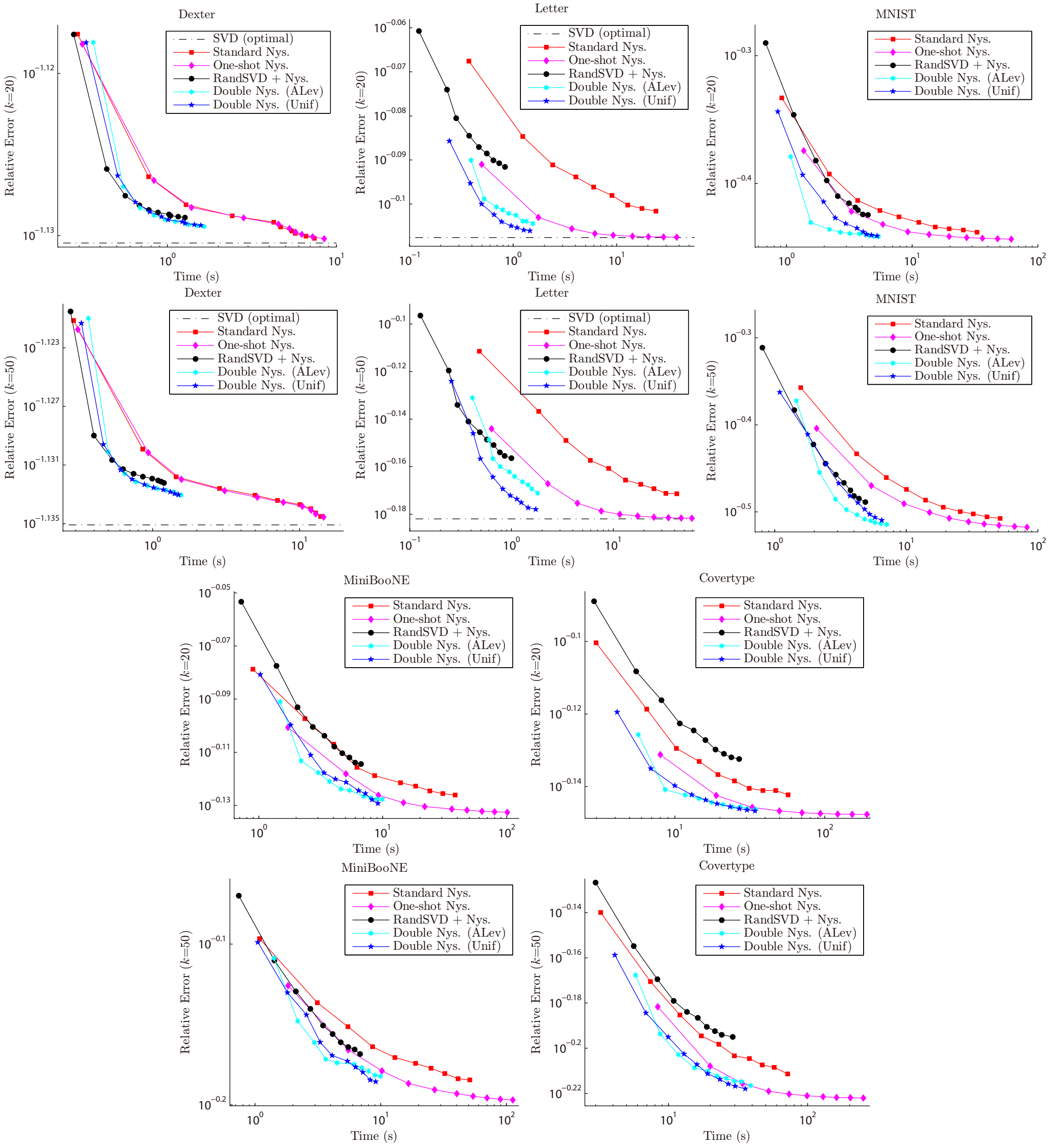

Figure 3: Performance comparison both for $k = 20$ and $k = 50$ among the four methods: the standard Nyström method (Williams & Seeger, 2001), the one-shot Nyström method (Fowlkes et al., 2004), the standard Nyström method using randomized SVD (Li et al., 2015), and the double Nyström method (ours). There are 10 episodes for each test, and there are 10 points on the each line in the figures. We perform SVD algorithm only on the Dexter and Letter data sets due to memory limit. In this experiment, 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.

A.1.2 Different σ

In this section, we report the results when we choose two different sigma for Gaussian kernel. Regardless of sigma, Fig 4 shows that our methods is both efficient and accurate compared to other methods.

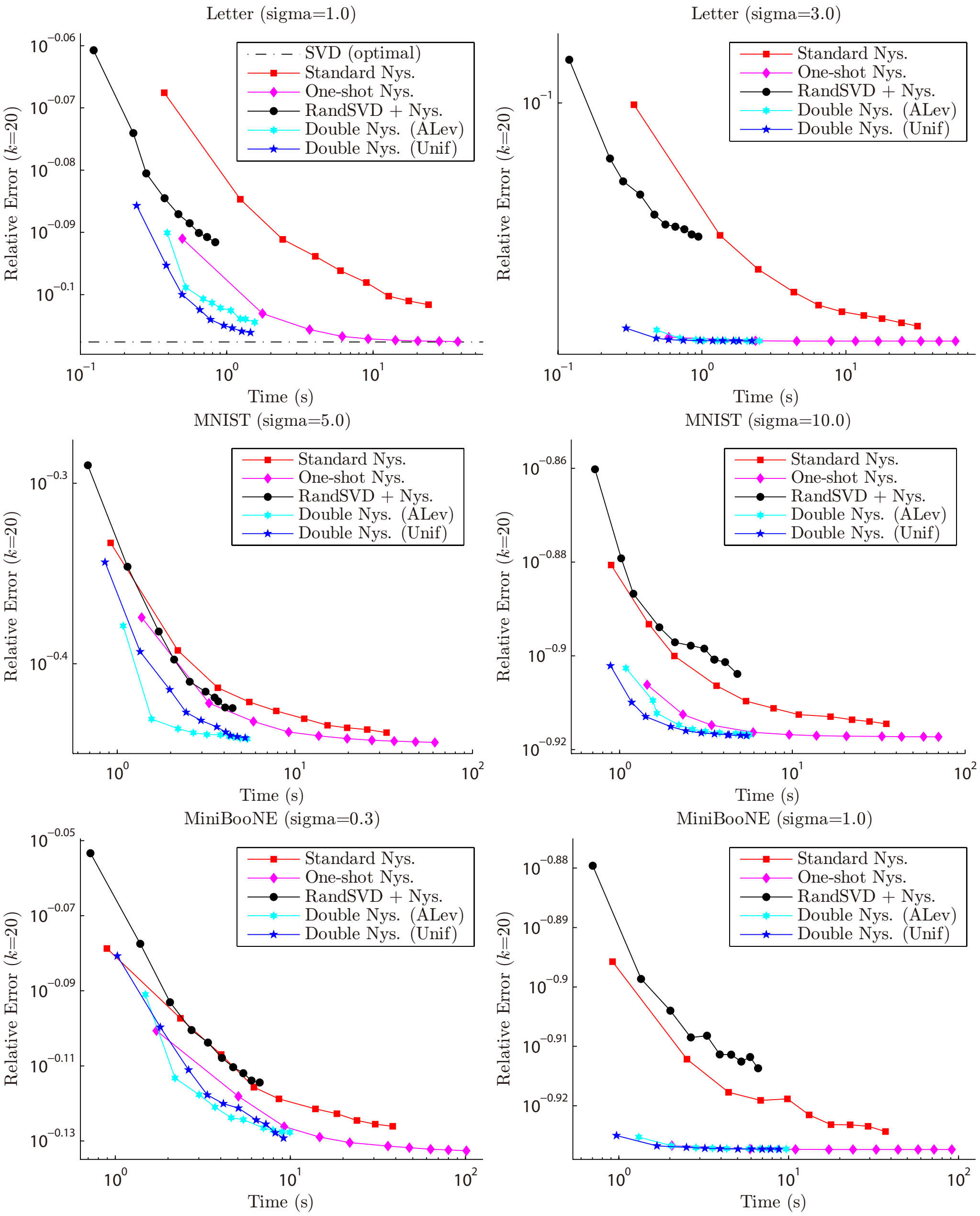

Figure 4: Experimental results for two different sigma on three data sets: Letter, MNIST, and MiniBooNE. We gradually increase the number of samples s as 500, 1000, 1500,..., 5000, and there are corresponding 10 points on the each line.

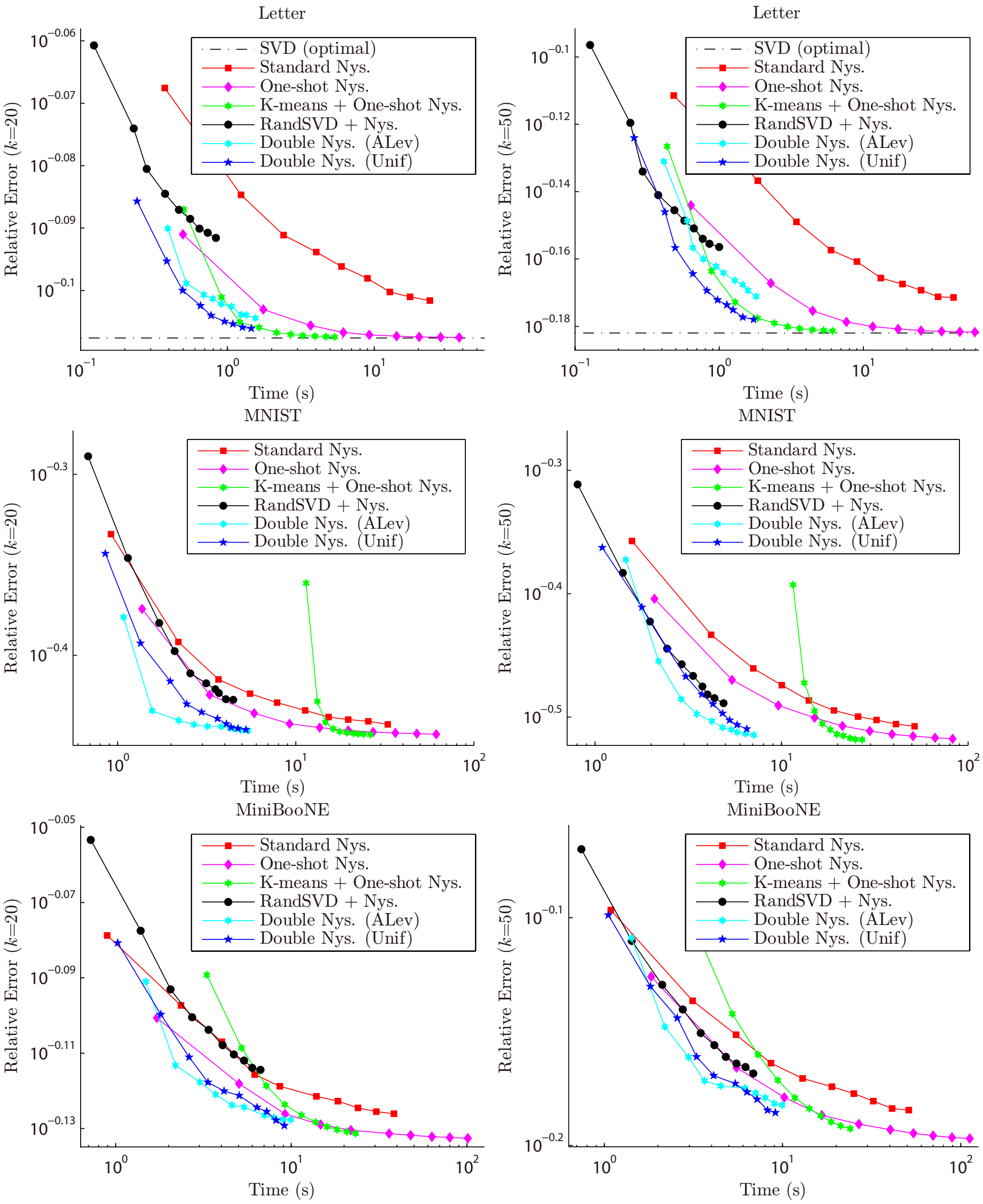

Figure 5: Comparison between the double Nyström method and the one-shot Nyström method using K-means sampling. We set $\sigma = 1.0$ for Letter, $\sigma = 5.0$ for MNIST, and $\sigma = 0.3$ for MiniBooNE in this experiment.

A.1.3 Double Nyström method vs K-means sampling + One-shot Nyström

One of the heuristic Nyström strategies is combining normal K-means sampling and the standard Nyström method (Zhang & Kwok, 2010). Thus, we also report experimental results of Nyström methods utilizing K-means sampling in this section. We adopted an efficient K-means algorithm, and limited the maximum iteration as 10. For K-means sampling, we gradually increase the number of clusters K as 100, 200, 300,..., 1000, thus there are corresponding 10 points on the line.

Fig 5 displays the corresponding results. We omit the experimental result of the standard Nyström method using K-means sampling (Zhang & Kwok, 2010), because it shows relatively poor efficiency and accuracy compared to other methods. The one-shot Nyström method using K-means sampling shows better accuracy and efficiency than the standard Nyström using K-means, however it is relatively slow than our methods and Nyström method using randomized SVD (Li et al., 2015) due to the running time of K-means. Especially, it is too slow for the experiments on the MNIST data set which has 784 original dimension. Whereas our methods shows a better efficiency even on the low-dimensional data sets which are Letter and MiniBooNE.

A.2 Leverage Scores, Kernel K-means and CAPS

Rem 1 implies the kernel K-means sampling, leverage score sampling, and CAPS sampling. Thus, we compare the three sampling methods for Nyström schemes in this section.

We introduced the notion of spanning set S and its corresponding matrix $\mathbf{S}$ in the main section. Let $\mathbf{C}_0$ be a $n \times s$ matrix such that $\mathbf{C}_0 = \mathbf{\Phi}^\top \mathbf{S}$. Then, we can also think that CAPS sampling extracts $\mathbf{W} \in \mathbb{R}^{d \times \ell}$ from $\mathbf{S} \in \mathbb{R}^{d \times s}$, and compress the $n \times s$ matrix $\mathbf{C}_0$ and $s \times s$ matrix $\mathbf{K}_S$ as a $n \times \ell$ matrix $\mathbf{C}$ and a $\ell \times \ell$ matrix $\mathbf{K}_W$ respectively. Consequently, we can understand that the three sampling methods compress $\mathbf{C}_0$ as $\mathbf{C}$, and test which sampling method extracts $n \times \ell$ matrix $\mathbf{C}$ inducing small approximation error. We denote ℓ as the number of columns of $\mathbf{C}$ in the experiments.

We apply each sampling method to the standard Nyström method (S.Nys) and one-shot Nyström method (O.S.Nys.). We select 7 sampling methods: uniform random sampling (Unif), adaptive-part sampling (Adapt-part) (Kumar et al., 2012), leverage-score sampling (lev) (Gittens & Mahoney, 2013), near-optimal sampling (NearOptimal) (Boutsidis et al., 2014), normal K-means sampling (Kmeans) (Zhang & Kwok, 2010), kernel K-means sampling (KKmeans), and CAPS sampling using one-shot Nyström (CAPS (unif), CAPS(ALev)) (ours). We set $s = n/10$ both for standard Nyström method using randomized SVD and double Nyström method, and assign $m = n/50$ in these experiments.

We adopted normalized approximation error (Normalized Error) which is defined in Eqn (13) in this experiment. Since the optimum value of the normalized approximation error is 1, we can easily interpret the results. Meanwhile, since we need to compute the optimal error $\|\mathbf{K} - \mathbf{K}_k\|_F$, thus we select 3 data sets which is not very large: Abalone, Wine, and Letter.

The experimental results are displayed in Fig 6 and Fig 7. Although the exact leverage score sampling utilizes the full kernel matrix $\mathbf{K}$, accuracy is not better than the normal K-means sampling. Since, column sampling methods including leverage score sampling select just one column index per a sample, whereas clustering samplings and CAPS samplings extract the samples as linear combination of instances. We note that kernel K-means sampling and CAPS samplings are superior to the other methods in terms of accuracy, and induce low errors closed to the optimum value 1 of the normalized error with a small ℓ. This means that the suggestions based on our analysis is correct. If we consider both running time and accuracy, then "CAPS sampling + Nyström methods" outperform all

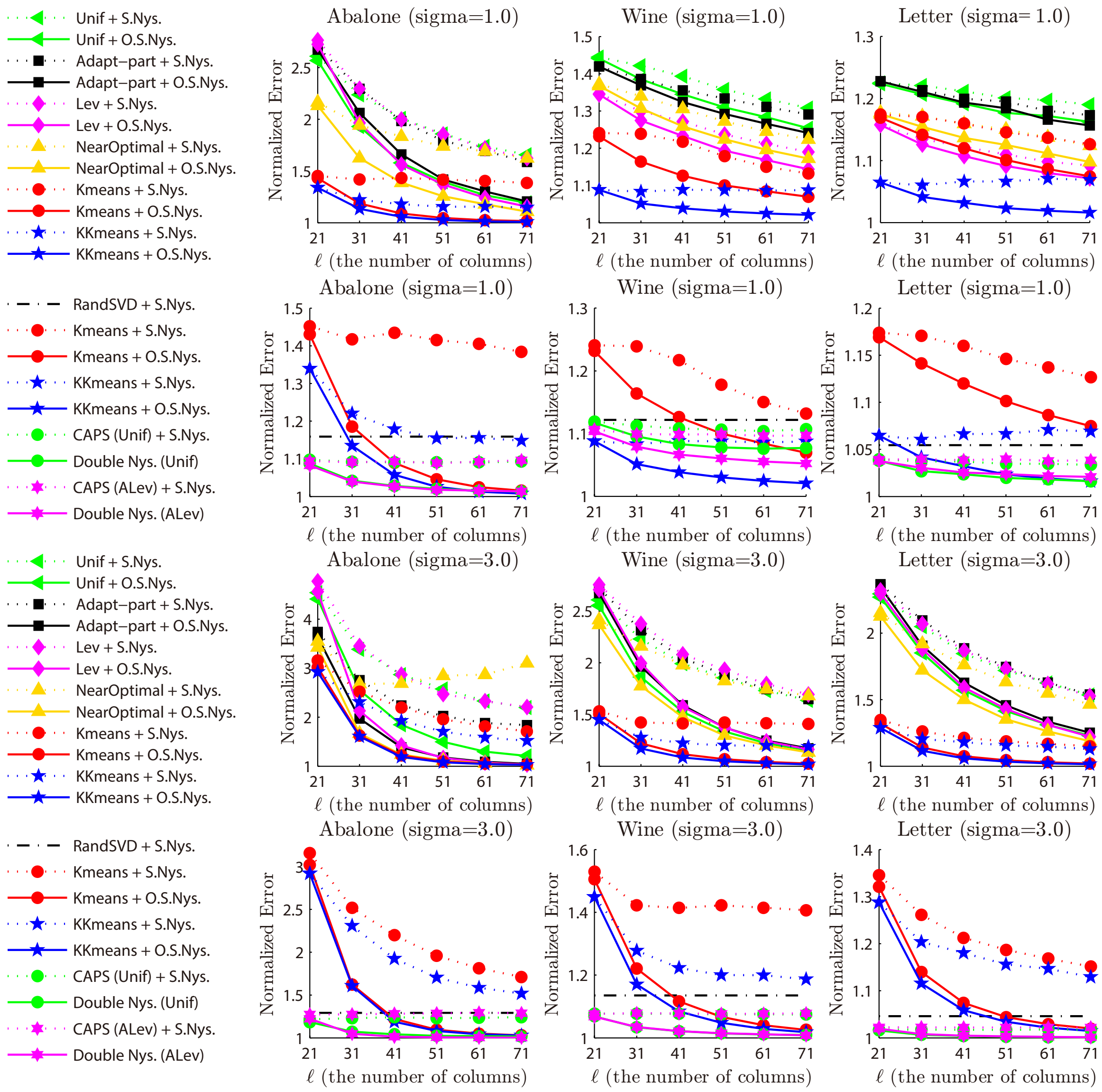

Figure 6: Comparison of the errors induced by sampling methods. The dotted line corresponds to the standard Nyström method, and the solid line corresponds to the one-shot Nyström method. Double Nyström denotes "CAPS + O.S.Nys.".

other strategies, especially the double Nyström method.

B Analysis in Section 3

B.1 The proof of Lemma 2

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

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

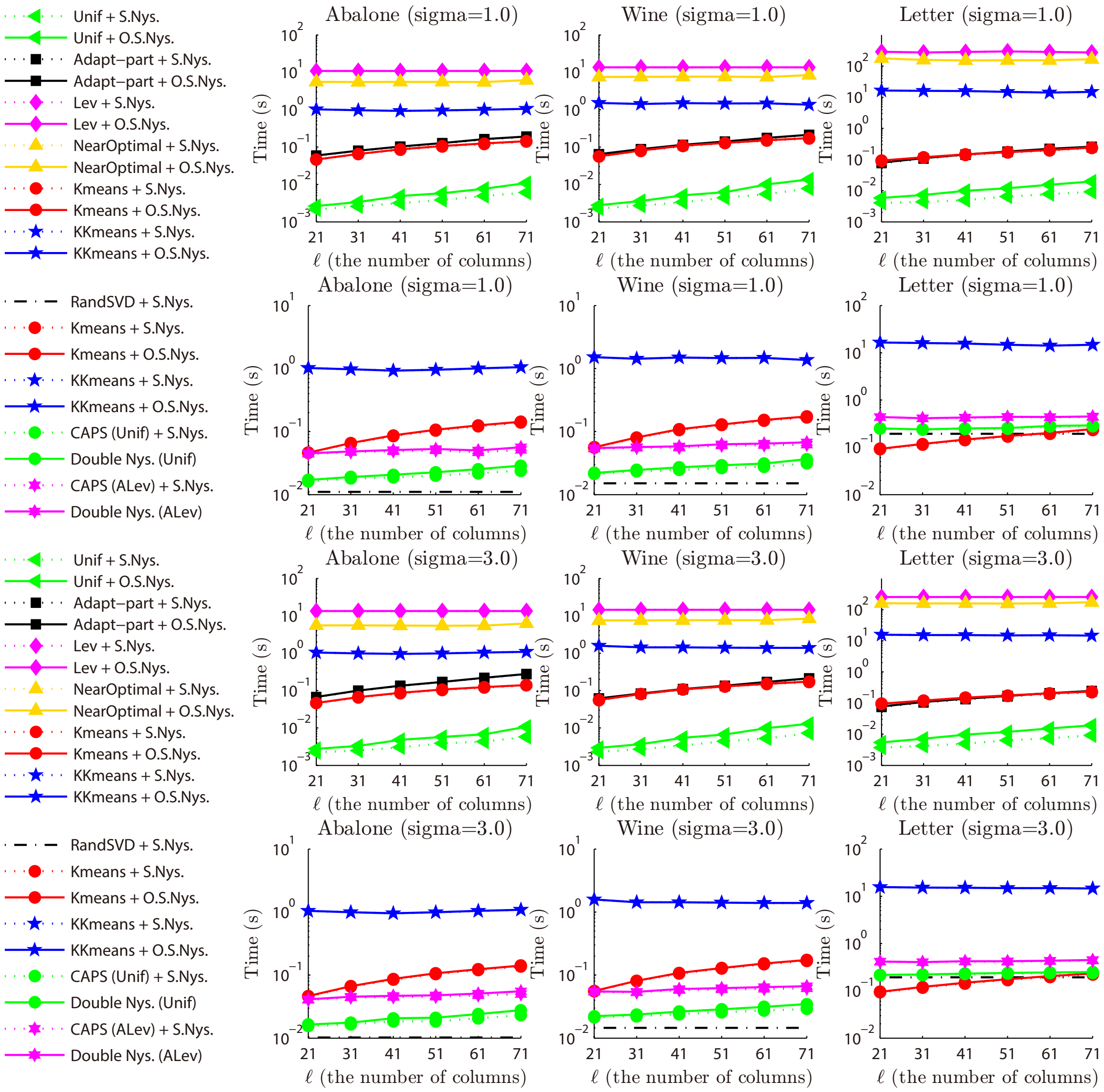

Figure 7: Comparison of the running times. Since the sizes of the data sets are small, the running times of two Nyström methods are relatively small compared to the sampling time. Thus, dotted are solid lines are close to each other.

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

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

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

Proof. Consider the reconstruction form $\mathbf{K}_k = \mathbf{V}_k \mathbf{\Sigma}_k^2 \mathbf{V}_k = \mathbf{\Phi}^\top \mathbf{U}_k \mathbf{U}_k^\top \mathbf{\Phi}$. Similarly, we can consider the reconstruction form both for the standard and one-shot Nyström approximations such as $\tilde{\mathbf{K}}_k = \tilde{\mathbf{V}}_k \tilde{\mathbf{\Sigma}}_k^2 \tilde{\mathbf{V}}_k = \mathbf{\Phi}^\top \tilde{\mathbf{U}}_k \tilde{\mathbf{U}}_k^\top \mathbf{\Phi}$. Then, we have a following equation for $\tilde{\mathbf{K}}_k^{nys}$

$$\tilde{\mathbf{K}}_k^{nys} = \mathbf{C} \mathbf{K}_{W,k}^\dagger \mathbf{C}^\top = \mathbf{\Phi}^\top \mathbf{W} \mathbf{V}_{W,k} \mathbf{\Sigma}_{W,k}^{-2} \mathbf{V}_{W,k}^\top \mathbf{W}^\top \mathbf{\Phi} = \mathbf{\Phi}^\top \mathbf{U}_{W,k} \mathbf{U}_{W,k}^\top \mathbf{\Phi},$$

and $\mathbf{U}_{W,k}^\top \mathbf{U}_{W,k} = \mathbf{I}_k$. Thus, $\tilde{\mathbf{U}}_k^{nys} = \mathbf{U}_{W,k}$. Second, we have a following equation for $\tilde{\mathbf{K}}_k^{osn}$

$$\tilde{\mathbf{K}}_k^{osn} = \mathbf{G} \mathbf{V}_{G,k} \mathbf{V}_{G,k}^\top \mathbf{G}^\top = \mathbf{\Phi}^\top \mathbf{U}_W \mathbf{V}_{G,k} \mathbf{V}_{G,k}^\top \mathbf{U}_W^\top \mathbf{\Phi},$$

and $(\mathbf{U}_W \mathbf{V}_{G,k})^\top \mathbf{U}_W \mathbf{V}_{G,k} = \mathbf{I}_k$. Consequently, $\tilde{\mathbf{U}}_k^{osn} = \mathbf{U}_W \mathbf{V}_{G,k}$. □

B.2 The proof of Theorem 1

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

Proof. Let $\tilde{\mathbf{U}}_k \in \mathbb{R}^{d \times k}$ be a matrix with orthonormal columns $\tilde{\mathbf{u}}_i$. The NRE can then be re-formulated as

$$\operatorname{NRE}(\tilde{\mathbf{U}}_k) = \sqrt{\frac{\operatorname{tr}((\mathbf{\Phi} - \tilde{\mathbf{U}}_k \tilde{\mathbf{U}}_k^\top \mathbf{\Phi})^\top (\mathbf{\Phi} - \tilde{\mathbf{U}}_k \tilde{\mathbf{U}}_k^\top \mathbf{\Phi}))}{\operatorname{tr}(\mathbf{\Phi}^\top \mathbf{\Phi})}} = \sqrt{1 - \frac{\operatorname{tr}(\tilde{\mathbf{U}}_k^\top \mathbf{\Phi} \mathbf{\Phi}^\top \tilde{\mathbf{U}}_k)}{\operatorname{tr}(\mathbf{K})}}.$$

where $\mathbf{K} = \mathbf{\Phi}^\top \mathbf{\Phi}$ is the full kernel matrix. Consequently, the optimization problem in Eqn (5) can be reformulated as

$$\underset{\mathbf{A}_k}{\text{maximize}} \ \operatorname{tr}(\tilde{\mathbf{U}}_k^\top \mathbf{\Phi} \mathbf{\Phi}^\top \tilde{\mathbf{U}}_k) \quad \text{subject to} \ \ \tilde{\mathbf{U}}_k^\top \tilde{\mathbf{U}}_k = \mathbf{I}_k, \tilde{\mathbf{U}}_k = \mathbf{W} \mathbf{A}_k. \tag{14}$$

Using compact SVD $\mathbf{W} = \mathbf{U}_W \mathbf{\Sigma}_W \mathbf{V}_W^\top$, we have

$$\tilde{\mathbf{U}}_k = \mathbf{W} \mathbf{A}_k = \mathbf{U}_W \mathbf{\Sigma}_W \mathbf{V}_W^\top \mathbf{A}_k = \mathbf{U}_W \mathbf{Z}_k,$$

by introducing matrix $\mathbf{Z}_k = \mathbf{\Sigma}_W \mathbf{V}_W^\top \mathbf{A}_k \in \mathbb{R}^{s' \times k}$. Since $\tilde{\mathbf{U}}_k^\top \tilde{\mathbf{U}}_k = \mathbf{Z}_k^\top \mathbf{Z}_k$, Eqn (14) becomes

$$\underset{\mathbf{Z}_k}{\text{maximize}} \ \operatorname{tr}(\mathbf{Z}_k^\top \mathbf{G}^\top \mathbf{G} \mathbf{Z}_k) \quad \text{subject to} \ \ \mathbf{Z}_k^\top \mathbf{Z}_k = \mathbf{I}_k, \tag{15}$$

where $\mathbf{G} = \mathbf{\Phi}^\top \mathbf{U}_W$. Note that this matrix is the same $\mathbf{G}$ in Alg 1 since $\mathbf{U}_W = \mathbf{W} \mathbf{V}_W \mathbf{\Sigma}_W^{-1}$ and $\mathbf{C} = \mathbf{\Phi}^\top \mathbf{W}$. By the work in Fan (1949), setting $\mathbf{Z}_k$ to the first k eigenvectors of the matrix $\mathbf{G}^\top \mathbf{G}$ solves the problem, which is $\mathbf{V}_{G,k}$ computed in Alg 1. Hence,

$$\tilde{\mathbf{U}}_k^{opt} = \mathbf{U}_W \mathbf{Z}_k^{opt} = \mathbf{U}_W \mathbf{V}_{G,k} = \mathbf{W} \mathbf{V}_W \mathbf{\Sigma}_W^{-1} \mathbf{V}_{G,k} = \tilde{\mathbf{U}}_k^{osn}. \tag{16}$$

Lastly, computing the projection of each data instance requires the eigenvectors of the kernel matrix, given in Eqn (4). Using $\tilde{\mathbf{U}}_k^{opt}$ in the above, the diagonal matrix of singular values $\tilde{\mathbf{\Sigma}}_k^{opt} = (\tilde{\mathbf{U}}_k^{opt})^\top \mathbf{\Phi}\mathbf{\Phi}^\top \tilde{\mathbf{U}}_k^{opt} = \mathbf{\Sigma}_{G,k}$ since $\mathrm{tr}(\mathbf{\Sigma}_{G,k})$ is the objective value at the optimum in Eqn (15), which is equal to that in Eqn (14). Finally, Eqn (4) with $\tilde{\mathbf{U}}_k^{opt}$ and $\tilde{\mathbf{\Sigma}}_k^{opt}$ yields

$$\tilde{\mathbf{V}}_k^{opt} = \mathbf{\Phi}^\top \tilde{\mathbf{U}}_k^{opt}(\tilde{\mathbf{\Sigma}}_k^{opt})^{-1} = \mathbf{\Phi}^\top \mathbf{W}\mathbf{V}_W \mathbf{\Sigma}_W^{-1}\mathbf{V}_{G,k}\mathbf{\Sigma}_{G,k}^{-1} = \mathbf{G}\mathbf{V}_{G,k}\mathbf{\Sigma}_{G,k}^{-1} = \tilde{\mathbf{V}}_k^{osn}.$$

Since $\tilde{\mathbf{V}}_k^{opt} = \tilde{\mathbf{V}}_k^{osn}$ and $\tilde{\mathbf{\Sigma}}_k^{opt} = \tilde{\mathbf{\Sigma}}_k^{osn}$, two approximate principal components are the same $\mathbf{\Phi}^\top \tilde{\mathbf{U}}_k^{opt} = \tilde{\mathbf{V}}_k^{opt}\tilde{\mathbf{\Sigma}}_k^{opt} = \tilde{\mathbf{V}}_k^{osn}\tilde{\mathbf{\Sigma}}_k^{osn} = \mathbf{\Phi}^\top \tilde{\mathbf{U}}_k^{osn}$. □

B.3 The Proof of Corollary 1

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

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

Proof. For any $k \leq \mathrm{rank}(\mathbf{K})$,

$$\mathrm{NRE}(\tilde{\mathbf{U}}_k) = \sqrt{\frac{c_k + \epsilon_1(\tilde{\mathbf{U}}_k)}{\mathrm{tr}(\mathbf{K})}},$$

where $c_k = \mathrm{tr}(\mathbf{K}) - \mathrm{tr}(\mathbf{K}_k)$. We finish the proof, since $\mathrm{tr}(\mathbf{K})$ and $\mathrm{tr}(\mathbf{K}_k)$ are constant given $\mathbf{K}$. □

B.4 The Proof of Proposition 1

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.*

Proof. As we proved in Lem 2, $\tilde{\mathbf{U}}_k^{nys} = \mathbf{U}_{W,k}$ and $\tilde{\mathbf{U}}_k^{osn} = \mathbf{U}_W\mathbf{V}_{G_W,k}$. $\tilde{\mathbf{U}}_k^{nys}$ can differ depending on scaling of $\mathbf{W}$, since the columns of $\mathbf{U}_{W,k}$ are the top k left singular vectors corresponding to the top k sigular values of $\mathbf{W}$. That is, even if two subspaces are the same, $\mathrm{col}(\mathbf{U}_{W_1,k})$ and $\mathrm{col}(\mathbf{U}_{W_2,k})$ may be different to each other. However, under the same condition, $\mathbf{U}_{W_1}$ can be represented as $\mathbf{U}_{W_2}\mathbf{B}_k$, where $\mathbf{B}_k$ is a $\mathrm{rank}(\mathbf{W}_1) \times k$ matrix such that $\mathbf{B}_k^\top \mathbf{B}_k = \mathbf{I}_k$. Using the $(\mathbf{B}_k\mathbf{Z}_k)^\top \mathbf{B}_k\mathbf{Z}_k = \mathbf{I}_k$ for $\mathbf{Z}_k^\top \mathbf{Z}_k = \mathbf{I}_k$, we can easily show that the sample based KPCA problem for $\mathbf{W}_1$ and for $\mathbf{W}_2$ are equivalent for $\mathrm{col}(\mathbf{W}_1) = \mathrm{col}(\mathbf{W}_2)$ as follows

$$\begin{aligned} &\underset{\mathbf{Z}_k}{\mathrm{maximize}}\ \mathrm{tr}(\mathbf{Z}_k^\top \mathbf{G}_1^\top \mathbf{G}_1 \mathbf{Z}_k) \quad \text{subject to } \ \mathbf{Z}_k^\top \mathbf{Z}_k = \mathbf{I}_k \\ \Longleftrightarrow &\underset{\mathbf{Z}_k}{\mathrm{maximize}}\ \mathrm{tr}(\mathbf{Z}_k^\top \mathbf{G}_2^\top \mathbf{G}_2 \mathbf{Z}_k) \quad \text{subject to } \ \mathbf{Z}_k^\top \mathbf{Z}_k = \mathbf{I}_k, \end{aligned}$$

where $\mathbf{G}_1 = \mathbf{\Phi}^\top \mathbf{U}_{W_1}$ and $\mathbf{G}_2 = \mathbf{\Phi}^\top \mathbf{U}_{W_2}$. □

Proposition 1 tells us that accuracies from the one-shot Nystrom method including the $\mathrm{NRE}(\tilde{\mathbf{U}}_k^{osn})$ and the $\epsilon_1(\tilde{\mathbf{U}}_k^{osn})$ are invariant to the scaling of samples, but depend on only subspace spanned by samples.

C Analysis in Section 4

C.1 The Proof of Theorem 2

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.

Proof. For any $\tilde{\mathbf{U}}_k$, we have following inequalities

$$\begin{aligned}
&\mathrm{NRE}(\tilde{\mathbf{U}}_k) - \mathrm{NRE}(\mathbf{U}_k) \\
&\qquad \leq \|(\mathbf{U}_k\mathbf{U}_k^\top - \tilde{\mathbf{U}}_k(\tilde{\mathbf{U}}_k)^\top)\mathbf{\Phi}\|_F / \|\mathbf{\Phi}\|_F \\
&\qquad \leq \|\mathbf{U}_k\mathbf{U}_k^\top - \tilde{\mathbf{U}}_k(\tilde{\mathbf{U}}_k)^\top\|_F.
\end{aligned}$$

The square of the last term in the above inequality is

$$\|\mathbf{U}_k\mathbf{U}_k^\top - \tilde{\mathbf{U}}_k(\tilde{\mathbf{U}}_k)^\top\|_F^2 = 2k - 2\,\mathrm{tr}(\mathbf{U}_k\mathbf{U}_k^\top\tilde{\mathbf{U}}_k(\tilde{\mathbf{U}}_k)^\top).$$

Since

$$\mathrm{PE}^2(\tilde{\mathbf{U}}_k, \mathbf{U}_k) = k - \mathrm{tr}(\mathbf{U}_k\mathbf{U}_k^\top\tilde{\mathbf{U}}_k(\tilde{\mathbf{U}}_k)^\top),$$

we finalize the proof with $\|\mathbf{U}_k\mathbf{U}_k^\top - \tilde{\mathbf{U}}_k(\tilde{\mathbf{U}}_k)^\top\|_F^2 = 2\mathrm{PE}^2(\tilde{\mathbf{U}}_k, \mathbf{U}_k)$.

Similarly,

$$\begin{aligned}
\|\mathbf{K} - \tilde{\mathbf{K}}_k\|_F - \|\mathbf{K} - \mathbf{K}_k\|_F &= \|\mathbf{\Phi}^\top\mathbf{\Phi} - \mathbf{\Phi}^\top\tilde{\mathbf{U}}_k(\tilde{\mathbf{U}}_k)^\top\mathbf{\Phi}\|_F - \|\mathbf{\Phi}^\top\mathbf{\Phi} - \mathbf{\Phi}^\top\mathbf{U}_k\mathbf{U}_k^\top\mathbf{\Phi}\|_F \\
&\leq \|(\mathbf{\Phi}^\top\mathbf{\Phi} - \mathbf{\Phi}^\top\tilde{\mathbf{U}}_k(\tilde{\mathbf{U}}_k)^\top\mathbf{\Phi}) - (\mathbf{\Phi}^\top\mathbf{\Phi} - \mathbf{\Phi}^\top\mathbf{U}_k\mathbf{U}_k^\top\mathbf{\Phi})\|_F \\
&= \|(\mathbf{\Phi}^\top(\mathbf{U}_k\mathbf{U}_k^\top - \tilde{\mathbf{U}}_k(\tilde{\mathbf{U}}_k)^\top)\mathbf{\Phi})\|_F \\
&\leq \|\mathbf{U}_k\mathbf{U}_k^\top - \tilde{\mathbf{U}}_k(\tilde{\mathbf{U}}_k)^\top\|_F\|\mathbf{\Phi}\|_F^2 \\
&= \sqrt{2}\mathrm{PE}(\tilde{\mathbf{U}}_k, \mathbf{U}_k)\,\mathrm{tr}(\mathbf{K}).
\end{aligned}$$

□

C.2 The Proof of Lemma 4

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

Proof. First, we prove an upper bound of subspace distance between $\mathrm{col}(\mathbf{U}_k)$ and $\mathrm{col}(\tilde{\mathbf{U}}_k)$. Without loss of generality, assume that $\tilde{\mathbf{U}}_k = \mathbf{U}_k\mathbf{P}_k + \mathbf{U}_{k+1}^d \mathbf{Q}_k$, where $\mathbf{U}_k$ be the matrix consisting of the first k principal directions which are eigenvectors of $\mathbf{\Phi}\mathbf{\Phi}^\top$, $\mathbf{U}_{k+1}^d$ be the matrix consisting of other $(d-k)$ eigenvectors of $\mathbf{\Phi}\mathbf{\Phi}^\top$, $\mathbf{\Lambda}_k$ and $\mathbf{\Lambda}_{k+1}^d$ are diagonal matrices consisting of corresponding eigenvalues. Then,

$$\begin{aligned}
\epsilon_1(\tilde{\mathbf{U}}_k) &= \mathrm{tr}(\mathbf{U}_k^\top \mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{U}_k) - \mathrm{tr}(\tilde{\mathbf{U}}_k^\top \mathbf{\Phi}\mathbf{\Phi}^\top \tilde{\mathbf{U}}_k) \\
&= \mathrm{tr}(\mathbf{\Lambda}_k) - \mathrm{tr}(\mathbf{\Lambda}_k \mathbf{P}_k \mathbf{P}_k^\top) - \mathrm{tr}(\mathbf{\Lambda}_{k+1}^d \mathbf{Q}_k \mathbf{Q}_k^\top) \\
&= \mathrm{tr}(\mathbf{\Lambda}_k(\mathbf{I}_k - \mathbf{P}_k\mathbf{P}_k^\top)) - \mathrm{tr}(\mathbf{\Lambda}_{k+1}^d \mathbf{Q}_k \mathbf{Q}_k^\top) \\
&\geq \lambda_k \,\mathrm{tr}(\mathbf{I}_k - \mathbf{P}_k\mathbf{P}_k^\top) - \mathrm{tr}(\mathbf{\Lambda}_{k+1}^d \mathbf{Q}_k \mathbf{Q}_k^\top).
\end{aligned}$$

Since $\mathrm{tr}(\mathbf{I}_k) = \mathrm{tr}(\mathbf{P}_k^\top \mathbf{P}_k + \mathbf{Q}_k^\top \mathbf{Q}_k)$, we have

$$\begin{aligned}
\epsilon_1(\tilde{\mathbf{U}}_k) &\geq \lambda_k \,\mathrm{tr}(\mathbf{I}_k - \mathbf{P}_k\mathbf{P}_k^\top) - \mathrm{tr}(\mathbf{\Lambda}_{k+1}^d \mathbf{Q}_k \mathbf{Q}_k^\top) \\
&= \lambda_k \,\mathrm{tr}(\mathbf{Q}_k\mathbf{Q}_k^\top) - \mathrm{tr}(\mathbf{\Lambda}_{k+1}^d \mathbf{Q}_k \mathbf{Q}_k^\top) \\
&= \mathrm{tr}((\lambda_k \mathbf{I}_{d-k} - \mathbf{\Lambda}_{k+1}^d)\mathbf{Q}_k \mathbf{Q}_k^\top) \\
&\geq (\lambda_k - \lambda_{k+1})\,\mathrm{tr}(\mathbf{Q}_k\mathbf{Q}_k^\top).
\end{aligned}$$

We finish this proof with $\mathrm{PE}^2(\mathbf{U}_k, \tilde{\mathbf{U}}_k) = \mathrm{tr}(\mathbf{Q}_k\mathbf{Q}_k^\top)$ and $\mathrm{PE}(\mathbf{U}_k, \tilde{\mathbf{U}}_k) = \mathrm{d}(\mathrm{col}(\mathbf{U}_k), \mathrm{col}(\tilde{\mathbf{U}}_k))$. Similarly, we can provide a lower bound of $\mathrm{d}(\mathrm{col}(\mathbf{U}_k), \mathrm{col}(\tilde{\mathbf{U}}_k))$.

$$\begin{aligned}
\epsilon_1(\tilde{\mathbf{U}}_k) &= \mathrm{tr}(\mathbf{U}_k^\top \mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{U}_k) - \mathrm{tr}(\tilde{\mathbf{U}}_k^\top \mathbf{\Phi}\mathbf{\Phi}^\top \tilde{\mathbf{U}}_k) \\
&= \mathrm{tr}(\mathbf{\Lambda}_k(\mathbf{I}_k - \mathbf{P}_k\mathbf{P}_k^\top)) - \mathrm{tr}(\mathbf{\Lambda}_{k+1}^d \mathbf{Q}_k \mathbf{Q}_k^\top) \\
&\leq \lambda_1 \,\mathrm{tr}(\mathbf{I}_k - \mathbf{P}_k\mathbf{P}_k^\top) - \mathrm{tr}(\mathbf{\Lambda}_{k+1}^d \mathbf{Q}_k \mathbf{Q}_k^\top) \\
&\leq \lambda_1 \,\mathrm{tr}(\mathbf{Q}_k\mathbf{Q}_k^\top).
\end{aligned}$$

Since $\mathrm{PE}^2(\mathbf{U}_k, \tilde{\mathbf{U}}_k) = \mathrm{tr}(\mathbf{Q}_k\mathbf{Q}_k^\top)$, we have $\sqrt{\frac{\epsilon_1(\tilde{\mathbf{U}}_k)}{\lambda_1}} \leq \mathrm{PE}(\mathbf{U}_k, \tilde{\mathbf{U}}_k)$.

Next, we prove upper and lower bound of $\mathrm{d}(\mathrm{col}(\mathbf{V}_k), \mathrm{col}(\tilde{\mathbf{V}}_k))$. Without loss of generality, suppose that $\tilde{\mathbf{V}}_k = \mathbf{V}_k\mathbf{P}_k + \mathbf{V}_{k+1}^n \mathbf{Q}_k$, where $\mathbf{V}_k$ be the matrix consisting of the first k eigenvectors of $\mathbf{K}$, $\mathbf{V}_{k+1}^n$ be the matrix consisting of other $(n-k)$ eigenvectors of $\mathbf{K}$, $\mathbf{\Lambda}_k$ and $\mathbf{\Lambda}_{k+1}^n$ are diagonal matrices consisting of corresponding eigenvalues. Then, the rest of the proof is similar to the case of $\mathrm{d}(\mathrm{col}(\mathbf{U}_k), \mathrm{col}(\tilde{\mathbf{U}}_k))$. □

C.3 The Proof of Lemma 5

Before showing the proof of Lem 5, we provide the following two lemmas.

Lemma 6. *$\mathbf{A} \in \mathbb{R}^{d\times n}$, $\tilde{\mathbf{U}}_k \in \mathbb{R}^{d\times k}$, $\tilde{\mathbf{V}}_k \in \mathbb{R}^{n\times k}$ be given, where $k \leq \min\{d, n\}$ and $\tilde{\mathbf{U}}_k$, $\tilde{\mathbf{V}}_k$ have orthonormal columns. Then $\sigma_i(\tilde{\mathbf{U}}_k^\top \mathbf{A}\tilde{\mathbf{V}}_k) \leq \sigma_i(\mathbf{A})$.*

Lem 7 comes directly from Lem 6.

Lemma 7. *Let $\tilde{\mathbf{V}}_k$ be a submatrix consisting of k columns of $\tilde{\mathbf{V}}_\ell$, and $\tilde{\mathbf{U}}_k$ be a submatrix consisting of k columns of $\tilde{\mathbf{U}}_\ell$, where $\tilde{\mathbf{U}}_\ell^\top \tilde{\mathbf{U}}_\ell = \mathbf{I}_\ell$ and $\tilde{\mathbf{V}}_\ell^\top \tilde{\mathbf{V}}_\ell = \mathbf{I}_\ell$. Then, given kernel matrix $\mathbf{K}$, we have*

$$\epsilon_1(\tilde{\mathbf{U}}_k) \leq \epsilon_2(\tilde{\mathbf{U}}_\ell)$$
$$\epsilon_2(\tilde{\mathbf{V}}_k) \leq \epsilon_2(\tilde{\mathbf{V}}_\ell).$$

Now, we provide Lem 5.

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.

Proof. Given $\tilde{\mathbf{V}}_\ell$, without loss of generality, we can arrange columns of $\tilde{\mathbf{V}}_\ell$ as descending order corresponding to the diagonal entry of $\mathbf{K}_W = \tilde{\mathbf{V}}_\ell^\top \mathbf{\Phi}^\top \mathbf{\Phi}\tilde{\mathbf{V}}_\ell$. Let $\tilde{\mathbf{V}}_k$ be the matrix consisting of the first k columns of $\tilde{\mathbf{V}}_\ell$, then by Lem 7,

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

The next goal is showing

$$\epsilon_1(\tilde{\mathbf{U}}_{W,k}) = \mathrm{tr}(\mathbf{U}_k^\top \mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{U}_k) - \mathrm{tr}((\mathbf{U}_{W,k})^\top \mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{U}_{W,k}) \leq \mathrm{tr}(\mathbf{V}_k^\top \mathbf{\Phi}^\top \mathbf{\Phi}\mathbf{V}_k) - \mathrm{tr}(\tilde{\mathbf{V}}_k^\top \mathbf{\Phi}^\top \mathbf{\Phi}\tilde{\mathbf{V}}_k),$$

where $\mathbf{U}_{W,k}$ consists of the first k columns of $\mathbf{U}_W$. Since $\mathrm{tr}(\mathbf{V}_k^\top \mathbf{\Phi}^\top \mathbf{\Phi}\mathbf{V}_k) = \mathrm{tr}(\mathbf{U}_k^\top \mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{U}_k)$, we will show that

$$\mathrm{tr}(\tilde{\mathbf{V}}_k^\top \mathbf{\Phi}^\top \mathbf{\Phi}\tilde{\mathbf{V}}_k) \leq \mathrm{tr}(\mathbf{U}_{W,k}^\top \mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{U}_{W,k}).$$

The matrix $\tilde{\mathbf{V}}_k^\top \mathbf{\Phi}^\top \mathbf{\Phi}\tilde{\mathbf{V}}_k$ is a principal submatrix of $\tilde{\mathbf{V}}_\ell^\top \mathbf{\Phi}^\top \mathbf{\Phi}\tilde{\mathbf{V}}_\ell$, thus by Lem 6,

$$\mathrm{tr}(\tilde{\mathbf{V}}_k^\top \mathbf{\Phi}^\top \mathbf{\Phi}\tilde{\mathbf{V}}_k) \leq \mathrm{tr}((\tilde{\mathbf{V}}_\ell^\top \mathbf{\Phi}^\top \mathbf{\Phi}\tilde{\mathbf{V}}_\ell)_k) = \mathrm{tr}(\mathbf{\Sigma}_{W,k}^2).$$

We give an alternative form for the same quantity of $\mathrm{tr}(\mathbf{\Sigma}_{W,k}^2)$ as follows

$$\begin{aligned}\mathrm{tr}(\mathbf{\Sigma}_{W,k}^2) &= \mathrm{tr}(\mathbf{V}_{W,k}\mathbf{\Sigma}_{W,k}^2\mathbf{V}_{W,k}^\top) \\ &= \mathrm{tr}(\mathbf{W}^\top \mathbf{U}_{W,k}\mathbf{U}_{W,k}^\top \mathbf{W}) \\ &= \mathrm{tr}(\tilde{\mathbf{V}}_\ell^\top \mathbf{\Phi}^\top \mathbf{U}_{W,k}\mathbf{U}_{W,k}^\top \mathbf{\Phi}\tilde{\mathbf{V}}_\ell).\end{aligned}$$

Since $\mathrm{tr}(\tilde{\mathbf{V}}_\ell^\top \mathbf{\Phi}^\top \mathbf{U}_{W,k}\mathbf{U}_{W,k}^\top \mathbf{\Phi}\tilde{\mathbf{V}}_\ell) \leq \mathrm{tr}(\mathbf{\Phi}^\top \mathbf{U}_{W,k}\mathbf{U}_{W,k}^\top \mathbf{\Phi})$, we finish the proof for the inequality $\mathrm{tr}(\tilde{\mathbf{V}}_k^\top \mathbf{\Phi}^\top \mathbf{\Phi}\tilde{\mathbf{V}}_k) \leq \mathrm{tr}(\mathbf{U}_{W,k}^\top \mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{U}_{W,k})$.

We proved $\epsilon_1(\mathbf{U}_{W,k}) = \mathrm{tr}(\mathbf{U}_k^\top \mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{U}_k) - \mathrm{tr}(\mathbf{U}_{W,k}^\top \mathbf{\Phi}\mathbf{\Phi}^\top \mathbf{U}_{W,k}) \leq \epsilon_2(\tilde{\mathbf{V}}_k)$. Meanwhile, $\tilde{\mathbf{U}}_k^{nys}$ is the $\mathbf{U}_{W,k}$ as we proved in Lem 2, and $\tilde{\mathbf{U}}_k^{osn}$ minimizes the ϵ_1 among $\tilde{\mathbf{U}}_k$ obtained from any sample-based KCPA methods including Nyström methods as we also proved in Cor 1. Thus,

$$\epsilon_1(\tilde{\mathbf{U}}_k^{osn}) \leq \epsilon_1(\tilde{\mathbf{U}}_k^{nys}) = \epsilon_1(\mathbf{U}_{W,k}) \leq \epsilon_2(\tilde{\mathbf{V}}_k).$$

□

C.4 The Proof of Theorem 3

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,* NRE *and* MRE *are bounded as follows:*

$$\begin{aligned}
\mathrm{NRE}(\tilde{\mathbf{U}}_k) &\leq \mathrm{NRE}(\mathbf{U}_k) + \sqrt{\frac{2\epsilon_2(\tilde{\mathbf{V}}_k)}{\gamma_k}} \\
&\leq \mathrm{NRE}(\mathbf{U}_k) + \sqrt{\frac{2\lambda_1}{\gamma_k}\mathrm{PE}(\mathbf{V}_k, \tilde{\mathbf{V}}_k)}, \\
\|\mathbf{K} - \tilde{\mathbf{K}}_k\|_F &\leq \|\mathbf{K} - \mathbf{K}_k\|_F + \sqrt{\frac{2\epsilon_2(\tilde{\mathbf{V}}_k)}{\gamma_k}}\, \mathrm{tr}(\mathbf{K}) \\
&\leq \|\mathbf{K} - \mathbf{K}_k\|_F + \sqrt{\frac{2\lambda_1}{\gamma_k}\mathrm{PE}(\mathbf{V}_k, \tilde{\mathbf{V}}_k)}\, \mathrm{tr}(\mathbf{K}),
\end{aligned}$$

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

Proof. The proof of this theorem comes directly from Thm 2, Lem 4 and Lem 5. □

C.5 The Proof of Proposition 2

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 the $\epsilon_2(\mathbf{V}_\ell)$ can be equivalently expressed as*

$$\underset{\mathbf{A}_\ell}{\mathrm{minimize}}\, \epsilon_2(\tilde{\mathbf{V}}_\ell) \text{ subject to } \tilde{\mathbf{V}}_\ell = \mathbf{T}_S\mathbf{A}_\ell, \mathbf{A}_\ell^\top \mathbf{A}_\ell = \mathbf{I}_\ell, \tag{17}$$

and the output of step 3 in Alg 2 with rank-ℓ SVD minimizes the $\epsilon_2(\mathbf{V}_\ell)$, i.e.,

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

where $\mathbf{K}_S = \mathbf{V}_S\mathbf{\Sigma}_S^2\mathbf{V}_S^\top$.

Proof. The condition is equivalently expressed as $\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_\ell = \mathbf{S}\mathbf{A}_\ell$ and $\tilde{\mathbf{V}}_\ell = \mathbf{T}_S\mathbf{A}_\ell$, where $\mathbf{A}_\ell^\top \mathbf{A}_\ell = \mathbf{I}_\ell$, and then the problem of minimizing the $\epsilon_2(\tilde{\mathbf{V}}_\ell)$ becomes Eqn (17). Next, we can directly show that

$$\begin{aligned} &\underset{\mathbf{A}_\ell}{\text{minimize}}\, \epsilon_2(\tilde{\mathbf{V}}_\ell) \;\text{ subject to }\; \tilde{\mathbf{V}}_\ell = \mathbf{T}_S\mathbf{A}_\ell \text{ and } \mathbf{A}_\ell^\top\mathbf{A}_\ell = \mathbf{I}_\ell \\ \Longleftrightarrow\, &\underset{\mathbf{A}_\ell}{\text{minimize}}\, \text{tr}(\mathbf{V}_k^\top \mathbf{\Phi}^\top \mathbf{\Phi}\mathbf{V}_k) - \text{tr}(\mathbf{A}_\ell^\top \mathbf{T}_S^\top \mathbf{\Phi}^\top \mathbf{\Phi}\mathbf{T}_S\mathbf{A}_\ell) \;\text{ subject to }\; \mathbf{A}_\ell^\top\mathbf{A}_\ell = \mathbf{I}_\ell. \end{aligned}$$

Since $\mathbf{S} = \mathbf{\Phi}\mathbf{T}_S$ and $\mathbf{K}_S = \mathbf{T}_S^\top \mathbf{\Phi}^\top \mathbf{\Phi}\mathbf{T}_S$,

$$\mathbf{V}_{S,\ell} = \underset{\mathbf{A}_\ell}{\text{argmin}}\, \text{tr}(\mathbf{V}_k^\top \mathbf{\Phi}^\top \mathbf{\Phi}\mathbf{V}_k) - \text{tr}(\mathbf{A}_\ell^\top \mathbf{T}_S^\top \mathbf{\Phi}^\top \mathbf{\Phi}\mathbf{T}_S\mathbf{A}_\ell) \;\text{ subject to }\; \mathbf{A}_\ell^\top\mathbf{A}_\ell = \mathbf{I}_\ell,$$

□

C.6 The Proof of Remark 1

C.6.1 The Kernel K-means Sampling

In this section, we insist that kernel K-means sampling for Nyström methods can induce more accurate approximation rather than normal K-means sampling (Zhang & Kwok, 2010). To show that, we will provide theoretical analysis of kernel K-means sampling for Nyström methods, and compare it with analysis of normal K-means sampling for standard Nyström method.

The normal K-means sampling (Zhang & Kwok, 2010) is heuristically good, but its motivation is weak due to ignoring kernel function $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ for computing landmark points in sampling step, even if the final goal is computing an approximate kernel matrix. In addition, suggested analysis displayed in Proposition 3 does not show any connection to the optimal error $\|\mathbf{K} - \mathbf{K}_k\|_F$ for rank-k.

Proposition 3. *(Zhang & Kwok, 2010) Given the original dataset $\mathcal{X}$ defined in Section 2, suppose that a clustering result from normal K-means on dataset $\mathcal{X}$ is represented by its centroids $\mathbf{z}_j = \sum_i \beta_{ij}\mathbf{x}_i = \mathbf{X}\boldsymbol{\beta}_j$, where $\boldsymbol{\beta}_j$ is the j-th L_1 cluster membership vector such that $1 = \sum_i \beta_{ij}$. By using $\boldsymbol{\beta}_j$, if we set ℓ sample vectors for standard Nyström method as $\mathbf{w}_j = \mathbf{\Phi}\boldsymbol{\beta}_j$ for $j = 1, ..., \ell$, where $\ell = K$, then the error of the standard Nyström approximation is bounded by*

$$\|\mathbf{K} - \tilde{\mathbf{K}}^{nys}\|_F \le 4n_1\sqrt{C_\mathcal{X}^\kappa K n_1 D} + C_\mathcal{X}^\kappa K n_1 D \|\mathbf{K}_W^\dagger\|_F,$$

where n_i is the number of instances in i-th cluster, $n_1 = \max_i |n_i|$, $c(i) = \text{argmin}_{j=1,2,...,K} \|\mathbf{x}_i - \mathbf{z}_j\|$, D is the normal K-means cost s.t. $D = \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{z}_{c(i)}\|_2^2$, and $C_\mathcal{X}^\kappa$ is the unknown constant for the given kernel function κ s.t. $(\kappa(\mathbf{x}_a, \mathbf{x}_b) - \kappa(\mathbf{x}_c, \mathbf{x}_d))^2 \le C_\mathcal{X}^\kappa \|\mathbf{x}_a - \mathbf{x}_c\|_2^2 + \|\mathbf{x}_b - \mathbf{x}_d\|_2^2$.

Now, we show why kernel K-means sampling for Nyström methods is able to induce a small approximation error. First, we can easily show that the cost of kernel K-means is $(\epsilon_2(\tilde{\mathbf{V}}_K) + \text{tr}(\mathbf{K} - \mathbf{K}_k))$, and provide Lem 8.

Lemma 8. *Let columns of $\tilde{\mathbf{V}}_K \in \mathbb{R}^{n\times K}$ be L_2 normalized membership vectors. Then $\tilde{\mathbf{V}}_K$ satisfies the condition $\tilde{\mathbf{V}}_K^\top \tilde{\mathbf{V}}_K = \mathbf{I}_K$, and the objective of kernel K-means is minimizing the $\epsilon_2(\tilde{\mathbf{V}}_K)$ defined in Lem 4 for $k = K$.*

Proof. It is well known that objective of kernel K-means is

$$\underset{\tilde{\mathbf{V}}_K}{\text{minimize}}\, \text{tr}(\mathbf{K}) - \text{tr}(\tilde{\mathbf{V}}_K \mathbf{\Phi}^\top \mathbf{\Phi} \tilde{\mathbf{V}}_K), \tag{18}$$

where columns of $\tilde{\mathbf{V}}_K \in \mathbb{R}^{n\times K}$ are L_2 normalized membership vectors. Since $\text{tr}(\mathbf{K})$ is a constant for the given kernel matrix $\mathbf{K}$, we can also consider the $\epsilon_2(\tilde{\mathbf{V}}_K)$ as its objective function. □

Thus, if the cost of kernel K-means is low, then kernel K-means sampling will induce small $\epsilon_2(\tilde{\mathbf{V}}_K)$ and $\epsilon_2(\tilde{\mathbf{V}}_k)$. Using Lem 7, Lem 8 and Thm 3, we prove that if we consider L_2 normalized membership vector $\tilde{\mathbf{V}}_K$ of kernel K-means of a low cost and set $\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_k$ for Nyström methods, then the approximation error will be small.

Corollary 3. *Suppose that the k-th eigengap γ_k is nonzero given $\mathbf{K}$. Assume that given the cost of kernel K-means is $(\epsilon_2 + \text{tr}(\mathbf{K} - \mathbf{K}_k))$, and columns of $\tilde{\mathbf{V}}_K$ are L_2 normalized membership vector of given the clustering. If we set $\ell = K$ and $\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_K$, then the standard and one-shot Nyström methods induce the* NRE *and the* MRE *as*

$$\text{NRE}(\tilde{\mathbf{U}}_k) \le \text{NRE}(\mathbf{U}_k) + \sqrt{\frac{2\epsilon_2}{\gamma_k}}$$
$$\|\mathbf{K} - \tilde{\mathbf{K}}_k\|_F \le \|\mathbf{K} - \mathbf{K}_k\|_F + \sqrt{\frac{2\epsilon_2}{\gamma_k}}\,\text{tr}(\mathbf{K}).$$

Remark 3. *Our analysis urges that L_2 normalized membership vectors should be set as $\tilde{\mathbf{V}}_K$ for $\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_K$ instead of L_1 normalized membership vectors. In fact, using L_2 or L_1 normalized membership does not affect the one-shot Nyström approximation, however it brings out the difference in standard Nyström method.*

Remark 4. *By Lem 8 and Cor 3, kernel K-means algorithm is suited for minimizing the ϵ_2 to reduce the error of the Nyström methods. If we consider L_2 membership vectors of the normal K-means, then the normal K-means could be considered as approximate kernel K-means sampling, since its objective function $\text{tr}(\mathbf{K}) - \text{tr}(\tilde{\mathbf{V}}_K \mathbf{X}^\top \mathbf{X} \tilde{\mathbf{V}}_K)$ is similar with Eqn (18), however it does not apply kernel function. Thus, kernel K-means for Nyström methods may induce more accurate approximations rather than normal K-means for Nyström methods, except that both of their clustering results are similar, i.e., kernel function is ineffective.*

Remark 5. *Our analysis is distinct from the work suggested by Zhang & Kwok (2010), First, our proposed error bounds in Cor 3 include the optimal error term and can converge to the optimum for any rank k, but the other can not. Second, there is also a difference between approaches for using the cluster structure. For example, suppose that we have the same clustering result from normal K-means. Then, based on our analysis, we prefer to use*

L_2 membership vectors of clustering for obtaining $\tilde{\mathbf{V}}_K$ and $\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_K$, and construct $\mathbf{C}$ and $\mathbf{K}_W$ by using $\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_K$. However, the other method uses the centroids of clustering to directly compute $\mathbf{C}$ and $\mathbf{K}_W$ by applying kernel function $\kappa(\cdot,\cdot)$. The latter is fast, but the former can be more accurate. To resolve the scalability issue, we can use scalable kernel K-means.

C.6.2 The Leverage Score Sampling

In this section, our goal is showing that leverage score sampling may induce a short subspace distance $\mathrm{PE}(\mathbf{V}_k, \tilde{\mathbf{V}}_k)$, consequently a small approximation error from Nyström methods.

The i-th leverage score of the columns of $\mathbf{K}$ for rank-k is defined as

$$\mathrm{lev}_i = (\mathbf{V}_k\mathbf{V}_k^\top)_{(i,i)},$$

which is the i-th diagonal element of $\mathbf{V}_k\mathbf{V}_k^\top$, where columns of $\mathbf{V}_k \in \mathbb{R}^{n\times k}$ are the true k eigenvectors of $\mathbf{K}$. Since the leverage scores are squared L_2 norm of each row of $\mathbf{V}_k$, if we consider $\tilde{\mathbf{V}}_k$ which consists of k vectors of row indices, then we can provide an alternative form of $\mathrm{PE}(\mathbf{V}_k, \tilde{\mathbf{V}}_k)$ by using leverage scores.

Lemma 9. *Let columns of $\mathbf{V}_k$ be the true k eigenvectors of $\mathbf{K}$ and columns of $\tilde{\mathbf{V}}_k$ be vectors of row indices corresponding to index set I, where $|I| = k$. Then, $\mathrm{PE}(\mathbf{V}_k, \tilde{\mathbf{V}}_k)$ can be characterized by leverage scores*

$$\mathrm{PE}(\mathbf{V}_k, \tilde{\mathbf{V}}_k) = \sqrt{k - \sum_{i\in I}\mathrm{lev}_i}, \tag{19}$$

where lev_i be the i-th leverage score.

Proof. Given $\tilde{\mathbf{V}}_k$, $\mathrm{PE}(\mathbf{V}_k, \tilde{\mathbf{V}}_k) = \|\mathbf{V}_k - \tilde{\mathbf{V}}_k\tilde{\mathbf{V}}_k^\top\mathbf{V}_k\|_F$. Since columns both of $\mathbf{V}_k$ and $\tilde{\mathbf{V}}_k$ are orthonormal,

$$\mathrm{PE}^2(\mathbf{V}_k, \tilde{\mathbf{V}}_k) = \|\mathbf{V}_k - \tilde{\mathbf{V}}_k\tilde{\mathbf{V}}_k^\top\mathbf{V}_k\|_F^2 = \mathrm{tr}(\mathbf{I}_k - \tilde{\mathbf{V}}_k^\top\mathbf{V}_k\mathbf{V}_k^\top\tilde{\mathbf{V}}_k).$$

Therefore, $\mathrm{PE}(\mathbf{V}_k, \tilde{\mathbf{V}}_k) = \sqrt{k - \sum_{i\in I}\mathrm{lev}_i}$. □

Now we provide directly Cor 4, which states that the upper bounds of approximation errors induced by Nyström methods from using any row (or column) index sampling can be characterized by sum of leverage scores.

Corollary 4. *Let $\tilde{\mathbf{V}}_\ell$ be vectors of row indices corresponding to index set J where $|J| = \ell$, and $\tilde{\mathbf{V}}_k$ be a submatrix which consists of k columns of $\tilde{\mathbf{V}}_\ell$ corresponding to index set I of $I \subset J$. If we set input vectors for the standard and one-shot Nyström methods as $\mathbf{W} = \mathbf{\Phi}\tilde{\mathbf{V}}_\ell$, then*

$$\mathrm{NRE}(\tilde{\mathbf{U}}_k) \le \mathrm{NRE}(\mathbf{U}_k) + \sqrt{\frac{2\lambda_1(\mathbf{K})(k - \sum_{i\in I}\mathrm{lev}_i)}{\gamma_k}}$$
$$\|\mathbf{K} - \tilde{\mathbf{K}}_k\|_F \le \|\mathbf{K} - \mathbf{K}_k\|_F + \sqrt{\frac{2\lambda_1(\mathbf{K})(k - \sum_{i\in I}\mathrm{lev}_i)}{\gamma_k}}\,\mathrm{tr}(\mathbf{K}).$$

Remark 6. *To reduce upper bounds both of the* $\mathrm{NRE}(\tilde{\mathbf{U}}_k)$ *and the* $\|\mathbf{K}-\tilde{\mathbf{K}}_k\|_F$ *in Cor 4, the probability of sampling indices has to induce a low expected value of* $\max_{I\subset J}(k-\sum_{i\in I}\mathrm{lev}_i)$, *or equivalently a high expected value of* $\max_{I\subset J}\sum_{i\in I}\mathrm{lev}_i$. *The simple idea which leads to high* $E[\max_{I\subset J}(k-\sum_{i\in I}\mathrm{lev}_i)]$ *is the leverage score sampling which selects indices for rank-k approximation with probability* $p_i=\frac{\mathrm{lev}_i}{k}$. *Thus, the leverage score sampling has an effect which reduces the expectation of* $\min_{\tilde{\mathbf{V}}_k}\mathrm{PE}(\mathbf{V}_k,\tilde{\mathbf{V}}_k)$, *and may induce a small error of Nyström approximation.*