Improving the Modified Nyström Method Using Spectral Shifting

Shusen Wang
College of Computer Science and Technology
Zhejiang University
wss@zju.edu.cn

Chao Zhang
College of Computer Science and Technology
Zhejiang University
zczju@zju.edu.cn

Hui Qian
College of Computer Science and Technology
Zhejiang University
qianhui@zju.edu.cn

Zhihua Zhang
*Department of Computer Science and Engineering
Shanghai Jiao Tong University
zhihua@sjtu.edu.cn

ABSTRACT

The Nyström method is an efficient approach to enabling large-scale kernel methods. The Nyström method generates a fast approximation to any large-scale symmetric positive semidefinete (SPSD) matrix using only a few columns of the SPSD matrix. However, since the Nyström approximation is low-rank, when the spectrum of the SPSD matrix decays slowly, the Nyström approximation is of low accuracy. In this paper, we propose a variant of the Nyström method called the modified Nyström by spectral shifting (SS-Nyström). The SS-Nyström method works well no matter whether the spectrum of SPSD matrix decays fast or slow. We prove that our SS-Nyström has a much stronger error bound than the standard and modified Nyström methods, and that SS-Nyström can be even more accurate than the truncated SVD of the same scale in some cases. We also devise an algorithm such that the SS-Nyström approximation can be computed nearly as efficient as the modified Nyström approximation. Finally, our SS-Nyström method demonstrates significant improvements over the standard and modified Nyström methods on several real-world datasets.

Categories and Subject Descriptors

G.1.0 [**Numerical Analysis**]: General—*Numerical algorithms*; G.1.3 [**Numerical Analysis**]: Numerical Linear Algebra—*Sparse, structured, and very large systems*

Keywords

Kernel approximation; the Nyström method; large-scale machine learning

*Corresponding Author

KDD'14, August 24–27, 2014, New York, NY, USA.

http://dx.doi.org/10.1145/2623330.2623614 .

1. INTRODUCTION

With the advent of the big-data era, how to efficiently learn from big-data has become a major concern and a hot topic of machine learning research. When data are large, many expensive matrix operations, e.g., matrix inverse and eigenvalue decomposition, which cost $\mathcal{O}(m^3)$ time and $\mathcal{O}(m^2)$ space for an $m \times m$ matrix, become computational prohibitive. Such expensive matrix operations are indispensable for many classical machine learning methods like kernel methods [22, 23], so these machine learning methods are infeasible when facing big-data problems. One possible approach to making matrix computation and kernel methods scalable is to use randomized matrix approximations to reduce the time and space costs [18], among which the most famous one is perhaps the Nyström method [5, 10, 14, 15, 20, 26, 30, 31].

The Nyström method approximates an arbitrary symmetric positive semidefinite (SPSD) kernel matrix using a small subset of its columns, and the method reduces the time complexity of many matrix operations from $\mathcal{O}(m^3)$ or $\mathcal{O}(m^2k)$ to $\mathcal{O}(mc^2)$ and space complexity from $\mathcal{O}(m^2)$ to $\mathcal{O}(mc)$, where k is the target rank, c is the number of selected columns, and it holds in general that $k < c \ll m$. In this way, time and space costs are only linearly in m, so many kernel methods can be efficiently solved even when m is large.

Williams & Seeger [30] used the Nyström method to speedup matrix inverse such that the inference of large-scale Gaussian process regression can be efficiently performed. Later on, the Nyström method has been applied to spectral clustering [7, 17], kernel SVMs [33], kernel PCA and manifold learning [26, 32, 33], kernel ridge regression [3], determinantal point processes [1], etc.

However, although the Nyström method is usually effective and efficient, its approximation quality can be very low in some cases. Wang & Zhang [28] showed that the relative-error (with respect to the best rank-k approximation) of the Nyström approximation grows with the matrix size m at least linearly. Thus the approximation can be rather rough when m is large, unless a large number of columns are selected to construct the Nyström approximation, which will violate the intention of using matrix approximations. To improve the approximation quality without sampling a large amount of columns, some other fast matrix approximation models have been proposed. Particularly, Wang & Zhang [28] developed the *modified Nyström method* to generate a low-rank approximation in a similar way to the standard Nyström method. The modified

Nyström method is much more accurate than the standard Nyström method in that it only samples an acceptable amount of columns and its relative error does not grow with matrix size m. In addition, the modified Nyström method only requires the original matrix to be symmetric, which is milder than SPSD required by the standard Nyström method.

The standard/modified Nyström methods generate low-rank approximations to kernel matrices, and their approximation errors cannot be better than the rank c truncated SVD, where c is the number of columns selected by the Nyström methods. When the spectrum of a kernel matrix decays slowly (that is, the $c+1$ to m largest eigenvalues are not small enough), the low-rank approximations constructed by either the truncated SVD or the standard/modified Nyström methods are far from the original kernel matrix. Cortes *et al.* [3] showed that the accuracy of kernel approximations affects the accuracy of learning algorithms. Therefore, when the spectrum of the kernel matrix decays slowly, the standard/modified Nyström methods cannot generate effective approximations to be used in learning algorithms.

To make the approximation still effective even when the spectrum of the original kernel matrix decays slowly, we propose in this paper a new method called *the modified Nyström by spectral shifting (SS-Nyström)*. Unlike the standard/modified Nyström methods which approximate the kernel matrix $\mathbf{K} \in \mathbb{R}^{m\times m}$ by a low-rank factorization $\mathbf{K} \approx \mathbf{C}\mathbf{U}\mathbf{C}^T$, our SS-Nyström approximates $\mathbf{K}$ by $\mathbf{K} \approx \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T + \delta\mathbf{I}_m$, where $\mathbf{C}, \bar{\mathbf{C}} \in \mathbb{R}^{m\times c}$, $\mathbf{U}, \bar{\mathbf{U}} \in \mathbb{R}^{c\times c}$, and $\delta \geq 0$. When the spectrum of $\mathbf{K}$ decays slowly, the term $\delta\mathbf{I}_m$ significantly improves the approximation accuracy. We show that SS-Nyström method has a provably tighter bound than the standard/modified Nyström methods. In sum, this paper offers the following contributions:

- We propose a kernel approximation method called the modified Nyström by Spectral Shifting (SS-Nyström), which is provably superior over the standard/modified Nyström methods (see Theorem 3). SS-Nyström can be even tighter than the truncated SVD in some conditions (See Example 1).

- We devise an efficient algorithm such that the SS-Nyström approximation can be computed nearly as efficient as the modified Nyström approximation. The proposed algorithm is also pass-efficient in that it goes only four passes through the kernel matrix, thus SS-Nyström is still efficient even when data do not fit in RAM.

The kernel approximation models proposed in the very recent work [15, 24] are also variants of the Nyström method and reported to achieve higher approximation accuracy. It is also straightforward to improve the ensemble Nyström method [15] and the memory efficient kernel approximation method [24] using the spectral shifting method proposed in this paper.

The remainder of this paper is organized as follows. In Section 2 we define the notation that will be used in this paper. In Section 3 we formally introduce the standard Nyström method and the modified Nyström method. In Section 4 we formulate our SS-Nyström method and show that our SS-Nyström can speedup several kernel methods in the same way as the standard Nyström method does. In Section 5 we theoretically show the superiority of SS-Nyström over the standard/modified Nyström methods. In Section 6 we devise an efficient algorithm for computing SS-Nyström. In Section 7 we empirically evaluate the SS-Nyström method and the proposed efficient algorithm. The proof of the theorems are deferred to the appendix.

2. NOTATION

For a matrix $\mathbf{A} = [a_{ij}]$, we let $\mathbf{a}_j$ be its j-th column, and $\|\mathbf{A}\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$ be its Frobenius norm. For a squared matrix, the matrix trace $\mathrm{tr}(\cdot)$ is the sum of the diagonal entries. We also let $\mathbf{I}_m$ be the $m \times m$ identity matrix and let $\mathbf{1}_m$ be the size-m all-one vector.

Letting $\rho = \mathrm{rank}(\mathbf{A})$, we write the condensed singular value decomposition (SVD) of $\mathbf{A}$ as $\mathbf{A} = \mathbf{U}_\mathbf{A}\mathbf{\Sigma}_\mathbf{A}\mathbf{V}_\mathbf{A}^T$, where the (i,i)-th entry of $\mathbf{\Sigma}_\mathbf{A} \in \mathbb{R}^{\rho\times\rho}$ is the i-th largest singular value of $\mathbf{A}$, denoted $\sigma_i(\mathbf{A})$. We also let $\mathbf{U}_{\mathbf{A},k}$ and $\mathbf{V}_{\mathbf{A},k}$ be the first k ($< \rho$) columns of $\mathbf{U}_\mathbf{A}$ and $\mathbf{V}_\mathbf{A}$, respectively, and $\mathbf{\Sigma}_{\mathbf{A},k}$ be the $k \times k$ top left block of $\mathbf{\Sigma}_\mathbf{A}$. Then the matrix $\mathbf{A}_k = \mathbf{U}_{\mathbf{A},k}\mathbf{\Sigma}_{\mathbf{A},k}\mathbf{V}_{\mathbf{A},k}^T$ is the "closest" rank-k approximation to $\mathbf{A}$. If $\mathbf{A}$ is a squared matrix, we let $\lambda_i(\mathbf{A})$ be the i-th largest eigenvalue. If $\mathbf{A}$ is SPSD, then the eigenvalue decomposition and SVD are equivalent.

For an $m\times n$ matrix, the full SVD costs time $\mathcal{O}(\min\{m^2n, mn^2\})$, and the rank k truncated SVD costs time $\mathcal{O}(mnk)$. Although multiplying an $m\times n$ matrix by an $n\times p$ matrix runs in $\mathcal{O}(mnp)$ flops, the constant in the big-$\mathcal{O}$ notation is tremendously smaller than that of SVD, and matrix multiplication can be highly efficiently performed in parallel computing facilities. So we instead denote the time complexity of matrix multiplication by $T_{\text{multiply}}(mnp)$, which is far less than $\mathcal{O}(mnp)$ in practice [12, 28].

3. RELATED WORK

Given an $m \times m$ SPSD matrix $\mathbf{K}$, we let $\mathcal{J}$ ($\mathcal{J} \subset [m] \triangleq \{1, 2, \ldots, m\}$ and $|\mathcal{J}| = c$) be an index set computed by some column selection algorithm. Then we let $\mathbf{C} \in \mathbb{R}^{m\times c}$ be the columns of $\mathbf{K}$ indexed by $\mathcal{J}$, and let $\mathbf{W} \in \mathbb{R}^{c\times c}$ be the rows of $\mathbf{C}$ indexed by $\mathcal{J}$. The *standard Nyström method* [30] approximates $\mathbf{K}$ by

$$\tilde{\mathbf{K}}_c^{\text{nys}} = \mathbf{C}\mathbf{U}^{\text{nys}}\mathbf{C}^T = \mathbf{C}\mathbf{W}^\dagger\mathbf{C}^T,$$

where $\mathbf{W}^\dagger$ is the Moore-Penrose inverse of $\mathbf{W}$. The *modified Nyström method* [28, 29] is defined by

$$\tilde{\mathbf{K}}_c^{\text{mod}} = \mathbf{C}\mathbf{U}^{\text{mod}}\mathbf{C}^T = \mathbf{C}\big(\mathbf{C}^\dagger\mathbf{K}(\mathbf{C}^\dagger)^T\big)\mathbf{C}^T.$$

The only difference between the standard and the modified Nyström methods is their intersection matrices: $\mathbf{U}^{\text{nys}} = \mathbf{W}^\dagger$ for the standard Nyström method and $\mathbf{U}^{\text{mod}} = \mathbf{C}^\dagger\mathbf{K}(\mathbf{C}^\dagger)^T$ for the modified Nyström method.

When $\mathrm{rank}(\mathbf{K}) = \mathrm{rank}(\mathbf{U}^{\text{nys}}) = \mathrm{rank}(\mathbf{U}^{\text{mod}})$, that is, when $\mathbf{K}$ is low-rank, both of the standard/modified Nyström approximations are exact [29]. The modified Nyström method is in general more accurate than the standard Nyström method due to $\|\mathbf{K} - \tilde{\mathbf{K}}_c^{\text{mod}}\|_F \leq \|\mathbf{K} - \tilde{\mathbf{K}}_c^{\text{nys}}\|_F$. With the selected columns at hand, it costs time $\mathcal{O}(c^3)$ to compute the standard Nyström approximation and $\mathcal{O}(mc^2) + T_{\text{multiply}}(m^2c)$ to compute the modified Nyström approximation.

The quality of the Nyström approximations is largely determined by whether the selected columns are informative, so a better column selection algorithm makes the Nyström approximations more accurate. In the previous work much attention has been paid to the relative-error column selection algorithms [2, 4, 6, 10, 11], among which the uniform sampling [10] and adaptive sampling [4] are the most widely used ones. The following lemma is the strongest theoretical result for the Nyström methods [28].

LEMMA 1. *Given an $m \times m$ symmetric matrix $\mathbf{K}$ and a target rank k, by selecting $c = \mathcal{O}(k\epsilon^{-2})$ columns of $\mathbf{K}$ to form $\mathbf{C} \in \mathbb{R}^{m\times c}$ using the adaptive sampling based algorithm of [28], the following inequality holds:*

$$\mathbb{E}\big\|\mathbf{K} - \tilde{\mathbf{K}}_c^{mod}\big\|_F \leq (1+\epsilon)\big\|\mathbf{K} - \mathbf{K}_k\big\|_F,$$

where $\mathbf{K}_k$ denotes the top k truncated SVD approximation of $\mathbf{K}$. The algorithm takes time $\mathcal{O}(mc^2 + mk^3\epsilon^{-2/3}) + T_{multiply}(m^2c)$ and space $\mathcal{O}(mc)$ in computing $\mathbf{C}$ and $\mathbf{U}$ of the modified Nyström approximation.

Lemma 1 indicates that by selecting $c = \mathcal{O}(k\epsilon^{-2})$ columns of $\mathbf{K}$, the modified Nyström approximation achieves comparable accuracy as the rank k truncated SVD. When the spectrum of $\mathbf{K}$ decays fast, the approximation generated by the truncated SVD is highly accurate, and so is the modified Nyström approximation. Otherwise, if the bottom $m-k$ singular values (i.e. eigenvalues) of $\mathbf{K}$ are large, then $\|\mathbf{K}-\mathbf{K}_k\|_F$ is large, and so is $\|\mathbf{K}-\tilde{\mathbf{K}}_c^{\text{mod}}\|_F$.

This work is closely related to the matrix ridge approximation (MRA) [34], which improves approximation accuracy by preserving the eigenvalues both large and small. When the bottom eigenvalues of $\mathbf{K}$ are large, MRA is much more accurate than the truncated SVD and the Nyström method [27, 34]. However, MRA is solved by iterative algorithms and is thus not pass-efficient and memory-efficient, it is thus limited to medium-scale data problems. Inspired by MRA, we propose a kernel approximation model which inherits the efficiency of the Nyström method and is effective when the bottom eigenvalues are large.

4. THE SS-NYSTRÖM APPROXIMATION

In the first subsection we formulate and justify our SS-Nyström method. In the second subsection we discuss how to apply SS-Nyström to speedup Gaussian process regression, kernel SVM, and kernel ridge regression problems.

4.1 Problem Formulation

Given a target rank k ($\leq c \ll m$), the SS-Nyström approximation of $\mathbf{K}$ is defined as

$$\tilde{\mathbf{K}}_c^{\text{ss}} \;=\; \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T + \delta\mathbf{I}_m. \tag{1}$$

Here $\delta \geq 0$ is called the spectral shifting term and $\bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T$ is the rank c modified Nyström approximation of $\bar{\mathbf{K}} = \mathbf{K} - \delta\mathbf{I}_m$. Notice that since $\bar{\mathbf{K}}$ is symmetric but possibly not SPSD, the $c \times c$ intersection matrix $\bar{\mathbf{U}}$ is also in general indefinite. Later we will see that the term $\delta\mathbf{I}_m$ has a direct effect on the spectrum of $\mathbf{K}$, that is why we call our method the modified Nyström by *spectrum shifting*.

Now we consider how to choose δ. It follows from the definition directly that the approximation error is $\mathbf{K} - \tilde{\mathbf{K}}_c^{\text{ss}} = \bar{\mathbf{K}} - \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T$; Lemma 1 indicates that by selecting sufficiently many columns of $\bar{\mathbf{K}}$ to construct $\bar{\mathbf{C}}$ and $\bar{\mathbf{U}}$, it holds in expectation that

$$\mathbb{E}\big\|\mathbf{K} - \tilde{\mathbf{K}}_c^{\text{ss}}\big\| \;=\; \mathbb{E}\big\|\bar{\mathbf{K}} - \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T\big\| \;\leq\; (1+\epsilon)\big\|\bar{\mathbf{K}} - \bar{\mathbf{K}}_k\big\|_F.$$

Apparently, for fixed k, the smaller the error $\|\bar{\mathbf{K}} - \bar{\mathbf{K}}_k\|_F$ is, the tighter error bound the SS-Nyström has; if $\|\bar{\mathbf{K}} - \bar{\mathbf{K}}_k\|_F \leq \|\mathbf{K} - \mathbf{K}_k\|_F$, then SS-Nyström has a better error bound than the modified Nyström. Therefore, to make the error bound as strong as possible, we formulate the following optimization problem to compute δ:

$$\min_{\delta \geq 0} \; \big\|\bar{\mathbf{K}} - \bar{\mathbf{K}}_k\big\|_F^2; \qquad \text{s.t. } \bar{\mathbf{K}} = \mathbf{K} - \delta\mathbf{I}_m.$$

However, since $\bar{\mathbf{K}}$ is in general indefinite, it needs all of the eigenvalues of $\mathbf{K}$ to solve the problem exactly. Since computing the full eigenvalue decomposition is expensive, we attempt to relax the problem. Considering that

$$\begin{aligned}\big\|\bar{\mathbf{K}} - \bar{\mathbf{K}}_k\big\|_F^2 \;&=\; \min_{|\mathcal{J}|=m-k} \sum_{j\in\mathcal{J}} \big(\sigma_j(\mathbf{K}) - \delta\big)^2 \\ &\leq\; \sum_{j=k+1}^{m} \big(\sigma_j(\mathbf{K}) - \delta\big)^2, \end{aligned}\tag{2}$$

we seek to minimize the upper bound of $\|\bar{\mathbf{K}} - \bar{\mathbf{K}}_k\|_F^2$ to compute δ, leading to the solution

$$\delta^{\text{opt}} = \frac{1}{m-k}\sum_{j=k+1}^{m}\sigma_j(\mathbf{K}) = \frac{1}{m-k}\Big(\text{tr}(\mathbf{K}) - \sum_{j=1}^{k}\sigma_j(\mathbf{K})\Big). \tag{3}$$

If we choose $\delta = 0$, then SS-Nyström degenerates to the modified Nyström method. The following theorem indicates that the SS-Nyström with any $\delta \in (0, \delta^{\text{opt}}]$ has a stronger relative-error bound than the modified Nyström method.

THEOREM 2. *Give an $m \times m$ SPSD matrix $\mathbf{K}$, we let $\bar{\mathbf{K}} = \mathbf{K} - \delta\mathbf{I}_m$ and δ^{opt} be defined in (3). Then for any $\delta \in (0, \delta^{opt}]$, the following inequality holds:*

$$\big\|\bar{\mathbf{K}} - \bar{\mathbf{K}}_k\big\|_F^2 \;\leq\; \big\|\mathbf{K} - \mathbf{K}_k\big\|_F^2.$$

We give an example in Figure 1 to illustrate why SS-Nyström is useful. From the plot of the eigenvalues we can see that the "tail" of the eigenvalues becomes thinner after the spectral shifting. Specifically, $\|\mathbf{K} - \mathbf{K}_k\|_F^2 = 0.52$ and $\|\bar{\mathbf{K}} - \bar{\mathbf{K}}_k\|_F^2 \leq 0.24$. When the same number of columns are selected to construct the SS-Nyström or the modified Nyström approximations, SS-Nyström has much tighter error bound because $\|\bar{\mathbf{K}} - \bar{\mathbf{K}}_k\|_F^2$ is much smaller than $\|\mathbf{K} - \mathbf{K}_k\|_F^2$.

4.2 Applications to Kernel Methods

We discuss in this section how to speed up matrix inverse and eigenvalue decomposition using the Nyström methods.Many kernel methods will become scalable if the matrix inverse and eigenvalue decomposition can be efficiently solved.

- Gaussian process regression [30], least squares SVM [25], and kernel ridge regression [21] all require computing this kind of linear system:

$$(\mathbf{K} + \alpha\mathbf{I}_m)\mathbf{b} = \mathbf{y}, \tag{4}$$

  which amounts to the matrix inverse problem $\mathbf{b} = (\mathbf{K} + \alpha\mathbf{I}_m)^{-1}\mathbf{y}$. Here α is a constant.

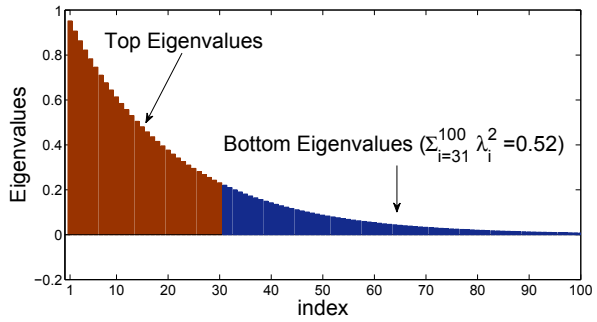
- Spectral clustering [7, 17], kernel PCA [32], and many manifold learning [33, 26] need to perform the truncated eigenvalue decomposition; the sampling algorithm of determinantal point processes [13, 1] performs the full eigenvalue decomposition.

Let $\mathbf{K} \in \mathbb{R}^{m\times m}$ be the kernel matrix, and let the SS-Nyström approximation of $\mathbf{K}$ be defined by

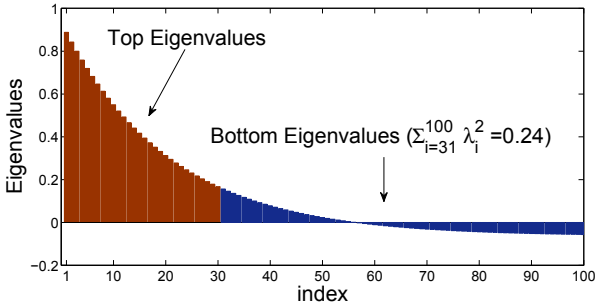
$$\tilde{\mathbf{K}}_c^{\text{ss}} \;=\; \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T + \delta\mathbf{I}_m.$$

We show that when $\mathbf{K}$ is replaced by $\tilde{\mathbf{K}}_c^{\text{ss}}$, the aforementioned linear system and eigenvalue decomposition can be efficiently solved. When using $\tilde{\mathbf{K}}_c^{\text{nys}}$ or $\tilde{\mathbf{K}}_c^{\text{mod}}$ to replace $\mathbf{K}$, one can still use the results by setting $\delta = 0$.

We first show how to approximately compute $\mathbf{b} = (\mathbf{K} + \alpha\mathbf{I}_m)^{-1}\mathbf{y}$. Let $\bar{\mathbf{U}} = \mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^T$ be the condensed eigenvalue decomposition of the intersection matrix of SS-Nyström, where $\mathbf{Z} \in \mathbb{R}^{c\times\rho}$, $\mathbf{\Lambda} \in \mathbb{R}^{\rho\times\rho}$, and $\rho = \text{rank}(\bar{\mathbf{U}}) \leq c$. We expand



(a) Before spectral shifting.



(b) After spectral shifting.

Figure 1: Toy data: 100×100 SPSD matrix whose the t-th eigenvalue is 1.05^{-t}. We set $m = 100$, $k = 30$, and thus $\delta^{\text{opt}} = 0.064$. We plot the eigenvalues of $\mathbf{K}$ in Figure 1(a) and $\bar{\mathbf{K}} = \mathbf{K} - \delta^{\text{opt}}\mathbf{I}_{100}$ in Figure 1(b).

$(\tilde{\mathbf{K}}_c^{\text{ss}} + \alpha\mathbf{I}_m)^{-1}$ by the Sherman-Morrison-Woodbury formula and obtain

$$\begin{aligned}\left(\tilde{\mathbf{K}}_c^{\text{ss}} + \alpha\mathbf{I}_m\right)^{-1} &= \left(\bar{\mathbf{C}}\mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^T\bar{\mathbf{C}}^T + \tau\mathbf{I}_m\right)^{-1} \\ &= \tau^{-1}\mathbf{I}_m - \tau^{-1}\bar{\mathbf{C}}\mathbf{Z}\left(\tau\mathbf{\Lambda}^{-1} + \mathbf{Z}^T\bar{\mathbf{C}}^T\bar{\mathbf{C}}\mathbf{Z}\right)^{-1}\mathbf{Z}^T\bar{\mathbf{C}}^T,\end{aligned}$$

where $\tau = \delta + \alpha$. In this way the linear system (4) can be computed in only $\mathcal{O}(mc^2)$ time and $\mathcal{O}(mc)$ space.

Now we show how to approximately compute the eigenvalue decomposition of $\mathbf{K}$. We let $\bar{\mathbf{C}} = \mathbf{U}_{\bar{\mathbf{C}}}\mathbf{\Sigma}_{\bar{\mathbf{C}}}\mathbf{V}_{\bar{\mathbf{C}}}$ be the condensed SVD of $\bar{\mathbf{C}}$. Suppose $\rho = \text{rank}(\bar{\mathbf{C}})$, we let

$$\mathbf{S} \;=\; \mathbf{\Sigma}_{\bar{\mathbf{C}}}\mathbf{V}_{\bar{\mathbf{C}}}\bar{\mathbf{U}}\mathbf{V}_{\bar{\mathbf{C}}}^T\mathbf{\Sigma}_{\bar{\mathbf{C}}}^T \;\in\; \mathbb{R}^{\rho\times\rho},$$

and we write the eigenvalue decomposition of $\mathbf{S}$ as $\mathbf{S} = \mathbf{U}_{\mathbf{S}}\mathbf{\Lambda}_{\mathbf{S}}\mathbf{U}_{\mathbf{S}}^T$. Now we can write the eigenvalue decomposition of $\tilde{\mathbf{K}}_c^{\text{ss}}$ as

$$\begin{aligned}\tilde{\mathbf{K}}_c^{\text{ss}} \;&=\; (\mathbf{U}_{\mathbf{C}}\mathbf{U}_{\mathbf{S}})\mathbf{\Lambda}_{\mathbf{S}}(\mathbf{U}_{\mathbf{C}}\mathbf{U}_{\mathbf{S}})^T + \delta\mathbf{I}_m \\ &=\; (\mathbf{U}_{\mathbf{C}}\mathbf{U}_{\mathbf{S}})(\mathbf{\Lambda}_{\mathbf{S}} + \delta\mathbf{I}_\rho)(\mathbf{U}_{\mathbf{C}}\mathbf{U}_{\mathbf{S}})^T + \mathbf{U}_\perp(\delta\mathbf{I}_m)\mathbf{U}_\perp^T.\end{aligned}$$

Here $\mathbf{U}_\perp \in \mathbb{R}^{m\times(m-\rho)}$ is a column orthogonal matrix orthogonal to $(\mathbf{U}_{\mathbf{C}}\mathbf{U}_{\mathbf{S}})$.

5. THEORETICAL ANALYSIS

We provide theoretical analysis for the SS-Nyström method in Theorem 3, which shows that SS-Nyström has a much tighter error bound than the modified Nyström method. We also demonstrate in Example 1 that in some cases the SS-Nyström method can be better than any other low-rank matrix approximation methods.

THEOREM 3. *Suppose there is a column selection algorithm $\mathcal{A}_{col}$ such that for any $m \times m$ symmetric matrix $\mathbf{S}$ and target rank k ($\ll m$), by selecting $c \geq C(m,k,\epsilon)$ columns of $\mathbf{S}$ using algorithm $\mathcal{A}_{col}$, the modified Nyström method attains the error bound*

$$\left\|\mathbf{S} - \tilde{\mathbf{S}}_c^{mod}\right\|_F^2 \;\leq\; (1+\epsilon)\left\|\mathbf{S} - \mathbf{S}_k\right\|_F^2.$$

Then for any $m \times m$ SPSD matrix $\mathbf{K}$, we compute δ^{opt} according to (3) and compute $\bar{\mathbf{K}} = \mathbf{K} - \delta^{opt}\mathbf{I}_m$. By using $\mathcal{A}_{col}$ to select $c \geq C(m,k,\epsilon)$ columns of $\bar{\mathbf{K}}$, the SS-Nyström defined in (1) attains the error bound

$$\left\|\mathbf{K} - \tilde{\mathbf{K}}_c^{ss}\right\|_F^2 \;\leq\; (1+\epsilon)\left(\left\|\mathbf{K} - \mathbf{K}_k\right\|_F^2 - \frac{\left[\sum_{i=k+1}^m \lambda_i(\mathbf{K})\right]^2}{m-k}\right).$$

If the columns of $\bar{\mathbf{K}}$ are selected by the adaptive sampling based algorithm of [28], which satisfies the assumption in Theorem 3 and is the best practical algorithm for the modified Nyström method, then the error bound incurred by SS-Nyström is given in the following corollary.

COROLLARY 4. *Suppose we are given an SPSD matrix $\mathbf{K}$. By sampling $c = \mathcal{O}(k\epsilon^{-2})$ columns of $\bar{\mathbf{K}}$ using the adaptive sampling based algorithm of [28], SS-Nyström attains the following error bound:*

$$\mathbb{E}\left\|\mathbf{K} - \tilde{\mathbf{K}}_c^{ss}\right\|_F^2 \leq (1+\epsilon)\left(\left\|\mathbf{K} - \mathbf{K}_k\right\|_F^2 - \frac{\left[\sum_{i=k+1}^m \lambda_i(\mathbf{K})\right]^2}{m-k}\right).$$

Recall from Lemma 1 that the best known error bound of the modified Nyström method is

$$\mathbb{E}\left\|\mathbf{K} - \tilde{\mathbf{K}}_c^{\text{mod}}\right\|_F^2 \;\leq\; (1+\epsilon)\left\|\mathbf{K} - \mathbf{K}_k\right\|_F^2,$$

where $c = \mathcal{O}(k\epsilon^{-2})$ columns are selected from $\mathbf{K}$. When the bottom eigenvalues $\lambda_{k+1}(\mathbf{K}), \cdots, \lambda_m(\mathbf{K})$ are large, we can see from Lemma 1 and Corollary 4 that the error bound of SS-Nyström is much better than that of the modified Nyström method. Here we give an example to demonstrate the superiority of SS-Nyström over the the standard/modified Nyström methods and even the truncated SVD of the same scale.

EXAMPLE 1. *Let $\mathbf{K}$ be an $m \times m$ SPSD matrix such that $\lambda_1(\mathbf{K}) \geq \cdots \geq \lambda_k(\mathbf{K}) > \theta = \lambda_{k+1}(\mathbf{K}) = \cdots = \lambda_m(\mathbf{K}) > 0$. By sampling $c = \mathcal{O}(k)$ columns by the adaptive sampling based algorithm of [28], we have that*

$$\left\|\mathbf{K} - \tilde{\mathbf{K}}_c^{ss}\right\|_F^2 \;=\; 0,$$

and that

$$\begin{aligned}(m-c)\theta^2 \;&=\; \left\|\mathbf{K} - \mathbf{K}_c\right\|_F^2 \\ &\leq\; \left\|\mathbf{K} - \tilde{\mathbf{K}}_c^{mod}\right\|_F^2 \;\leq\; \left\|\mathbf{K} - \tilde{\mathbf{K}}_c^{nys}\right\|_F^2.\end{aligned}$$

In this example SS-Nyström is far better than the other approximation methods if we set θ a large constant.

6. EFFICIENT ALGORITHM

Notice that computing the spectral shifting term δ in (1) according to (3) requires the truncated SVD which costs time $\mathcal{O}(m^2k)$ and space $\mathcal{O}(m^2)$. This can be accelerated by computing the top-k singular values approximately using random projection techniques [2, 12]. We depict the whole algorithm for computing SS-Nyström using random projections in Algorithm 1. The performance of the approximation is analyzed in the following theorem.

Algorithm 1 The Modified Nyström by Spectral Shifting.

1: **Input:** an $m \times m$ SPSD matrix $\mathbf{K}$, a target rank k, the oversampling parameter l.
2: *// compute δ^{opt} approximately*
3: $\mathbf{\Omega} \longleftarrow m \times l$ standard Gaussian matrix;
4: $\mathbf{Q} \longleftarrow$ the l orthonormal basis of $\mathbf{Y} = \mathbf{K\Omega} \in \mathbb{R}^{m\times l}$;
5: $s \longleftarrow$ sum of the top k singular values of $\mathbf{A} = \mathbf{Q}^T\mathbf{K} \in \mathbb{R}^{l\times m}$;
6: $\tilde{\delta} = \frac{1}{m-k}\big(\mathrm{tr}(\mathbf{K}) - s\big) \approx \delta^{\mathrm{opt}}$;
7: *// spectral shifting*
8: $\bar{\mathbf{K}} \leftarrow \mathbf{K} - \tilde{\delta}\mathbf{I}_m \in \mathbb{R}^{m\times m}$;
9: *// compute the modified Nyström approximation for $\bar{\mathbf{K}}$*
10: $\bar{\mathbf{C}} \longleftarrow c$ columns of $\bar{\mathbf{K}}$ selected by some column sampling algorithm;
11: $\bar{\mathbf{U}} \longleftarrow \bar{\mathbf{C}}^{\dagger}\bar{\mathbf{K}}(\bar{\mathbf{C}}^{\dagger})^T \in \mathbb{R}^{c\times c}$;
12: **return** the SS-Nyström approximation $\tilde{\mathbf{K}}_c^{\mathrm{ss}} = \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T + \tilde{\delta}\mathbf{I}_m$.

THEOREM 5. *Let δ^{opt} be defined in (3) and $\tilde{\delta}$, k, l, m be defined in Algorithm 1. The following inequality holds in expectation:*

$$\mathbb{E}\big[\big|\delta^{opt} - \tilde{\delta}\big| \,/\, \delta^{opt}\big] \;\le\; k/\sqrt{l},$$

where the expectation is taken w.r.t. the Gaussian random matrix $\mathbf{\Omega}$ in Algorithm 1. Lines 2–6 in Algorithm 1 compute $\tilde{\delta}$ in time $\mathcal{O}(ml^2) + T_{multiply}(m^2l)$ and space $\mathcal{O}(ml)$.

By using Algorithm 1 to compute δ^{opt} approximately, it costs only $\mathcal{O}(ml^2) + T_{\mathrm{multiply}}(m^2l)$ more time to compute the SS-Nyström approximation than the modified Nyström approximation. Our experiments show that a small l (say, $l = 4k$) is sufficient for obtaining a highly accurate approximation to δ^{opt}, no matter whether the spectrum of $\mathbf{K}$ decays fast or slow. Since it costs $\mathcal{O}(mc^2) + T_{\mathrm{multiply}}(m^2c)$ time to compute the modified Nyström and c can be set as $\mathcal{O}(k\epsilon^{-2})$, if we set $l = 4k$, then the time complexity for computing SS-Nyström is the same as computing the modified Nyström.

7. EXPERIMENTS

We conduct experiments on several real-world datasets to evaluate the method and algorithm proposed in this paper. In Section 7.1 we describe the setup of the experiments. In Section 7.2 we evaluate the fast approximation of δ^{opt} proposed in Section 6. In Section 7.3 we compare SS-Nyström with the standard/modified Nyström methods on several middle-size datasets. In Section 7.4 we compare the methods on a large-scale dataset where the kernel matrix does not fit in RAM.

7.1 The Setup

We perform experiments on several datasets released by UCI [8] and Statlog [19]. We obtain the data collected on the LIBSVM website[1] where the data are scaled to [0,1]. We summarize the datasets in Table 1.

For each dataset, we generate a radial basis function (RBF) kernel $\mathbf{K}^{\alpha}$ defined by

$$k_{ij}^{\alpha} = \exp\Big(-\frac{1}{2\alpha}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2\Big)$$

and a sparse RBF kernel $\mathbf{K}^{\alpha,\nu,C}$ [9] defined by

$$k_{ij}^{\alpha,\nu,C} = \left[\Big(1 - \frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{C}\Big)^{\nu}\right]_+ \cdot \exp\Big(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\alpha}\Big).$$

[1] http://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/

Here $\mathbf{x}_1, \cdots, \mathbf{x}_m$ are the data instances, $\alpha > 0$ is the scaling parameter, $C > 0$ is the cutting-off point, d is the number of attributes, $\nu > (d+1)/2$, and $[z]_+ \triangleq \max\{z, 0\}$. The larger the scaling parameter α is, the faster the spectrum of the kernel decays. For the sparse RBF kernel, following [10], we fix $\nu = \lceil (d+1)/2 \rceil$ and $C = 3\alpha$.

We implement all the algorithms in MATLAB and run the algorithms on a workstation with Intel Xeon 2.40GHz CPUs, 24GB RAM, and 64bit Windows Server 2008 system. To compare the running time, we set MATLAB in single thread mode by the command "maxNumCompThreads(1)".

7.2 Performance of the Approximation to δ^{opt}

We evaluate the accuracy of the approximation to δ^{opt} (Lines 2–6 in Algorithm 1) proposed in Theorem 5. We generate RBF kernel matrices of the listed datasets, where we set the scaling parameter $\alpha = 0.1$ or $\alpha = 1$. We use the error ratio $|\delta^{\mathrm{opt}} - \tilde{\delta}|/\delta^{\mathrm{opt}}$ to evaluate the approximation performance. We repeat the experiments 20 times and plot the average error ratio versus l/k in Figure 2. Here $\tilde{\delta}$, l, and k are defined in Theorem 5. We can see from Figure 2 that the approximation to δ^{opt} is of very high quality: when $l = 4k$, the error ratios are less than 0.03 in all cases. So we set $l = 4k$ in all of the subsequent kernel approximation experiments in order to obtain a low over-sampling rate with a high accuracy at the same time.

7.3 Performance of Kernel Approximation

We evaluate the kernel approximation accuracy of SS-Nyström mainly in comparison with the standard/modified Nyström methods. In this paper our attention is mainly focused on kernel matrices whose spectrum decay slowly, so we set the scaling parameter α a small value. Otherwise if α is large, the bottom eigenvalues will be very small, and consequently the spectral shifting parameter δ^{opt} will be so small that there is no significant difference between SS-Nyström and the modified Nyström. Specifically, we set $k = 50$ and $\alpha = 0.2$ for the dense RBF kernels and $\alpha = 2$ for the sparse RBF kernels. We use Algorithm 1 to compute the SS-Nyström approximation, in which we set $l = 4k$. For each of the standard/modified/SS Nyström methods, we use two algorithms to select columns: the uniform sampling algorithm [10] and the adaptive sampling algorithm [28].

We report the approximation accuracy and running time of each algorithm for each method. The approximation accuracy is evaluated by

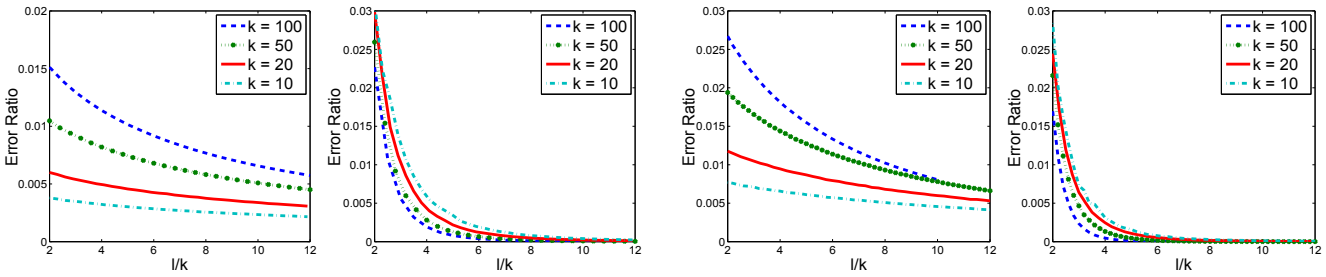
$$\text{Approximation Error} \;=\; \|\mathbf{K} - \tilde{\mathbf{K}}\|_F / \|\mathbf{K}\|_F,$$

where $\tilde{\mathbf{K}}$ is the approximation generated by each method. Every time when we do column sampling, we run each sampling algorithm 10 times and report the minimal approximation error of the 10 repeats since the error bound of each method is actually guaranteed with expectations and we can get a quite accurate approximation within 10 repeats according to [28]. We report the average elapsed time of the 10 repeat rather than the total elapsed time because the 10 repeats can be done in parallel on 10 machines. We depict the approximation errors and the average elapsed time of the dense RBF kernels in Figures 3 and Figure 4 and those of the sparse RBF kernels are in Figure 5 and Figure 6. In the figures, we use $\frac{c}{m}$ as the X-axis because the compared methods have the same RAM cost when c and m are fixed.

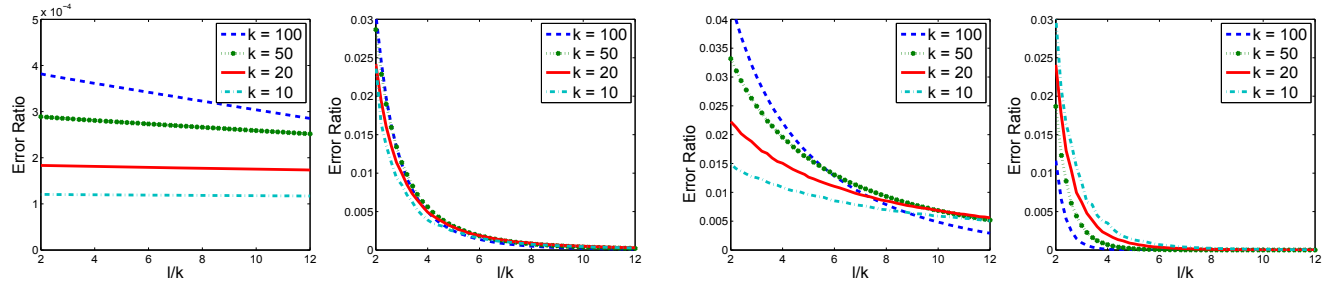
The results clearly show that our SS-Nyström works significantly better than the standard/modified Nyström methods when the spectrum of the kernel matrix decays slowly. As for the running

Table 1: A summary of the datasets for the Nyström approximation.

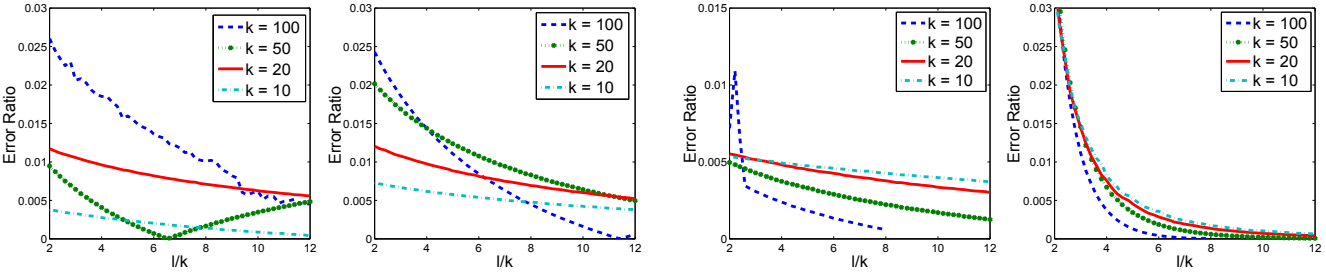
Dataset	MNIST	Letters	Wine Quality	Satimage	Segment	DNA	German	Splice	Breast Cancer
#Instance	60,000	15,000	4,898	4,435	2,310	2,000	1,000	1,000	683
#Attribute	780	16	12	36	19	180	24	60	10



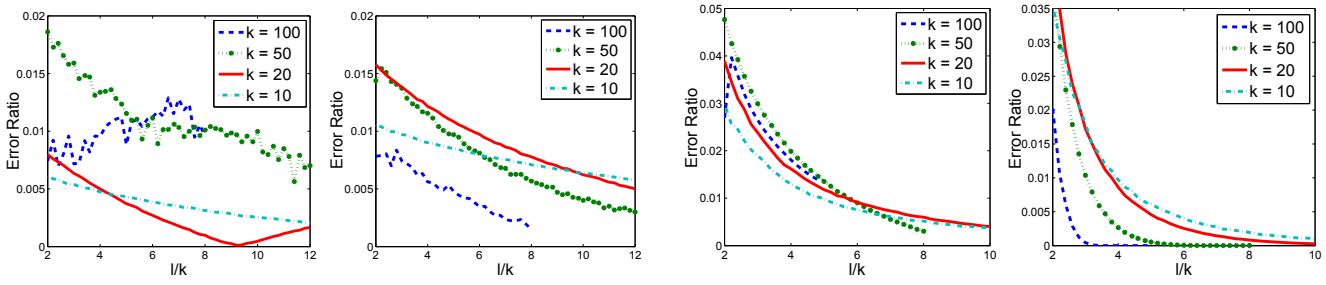
(a) Letters. (b) Wine Quality.



(c) Satimage. (d) Segment.

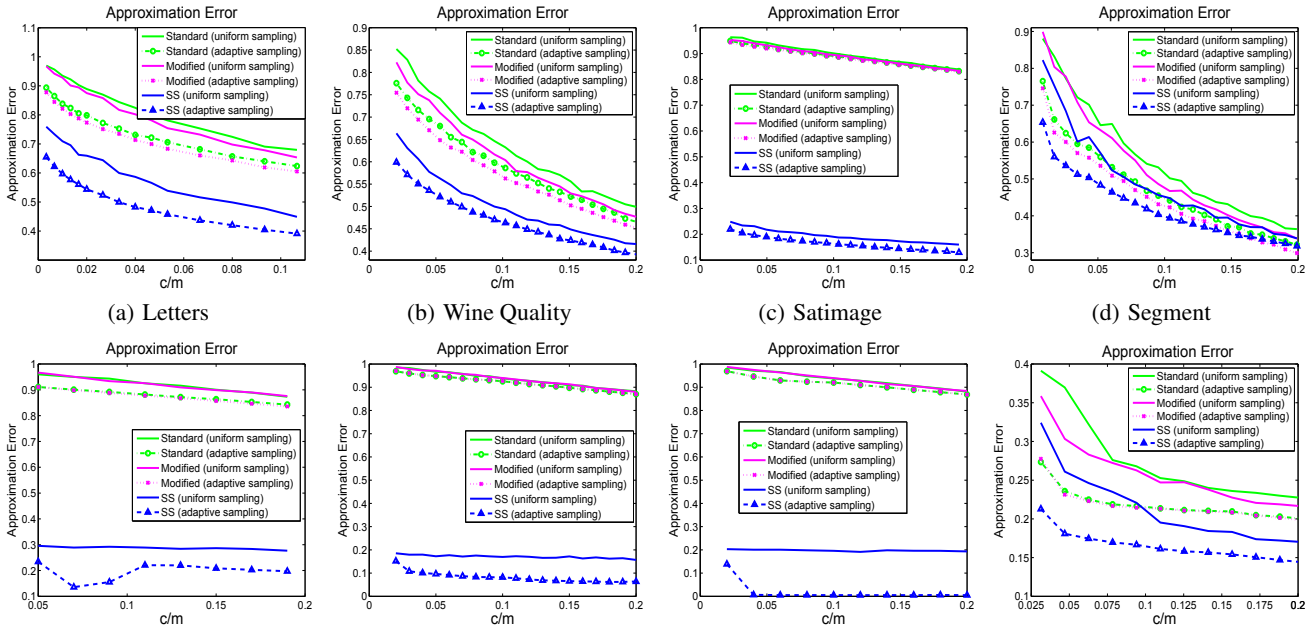


(e) DNA. (f) German.



(g) Splice. (h) Breast Cancer.

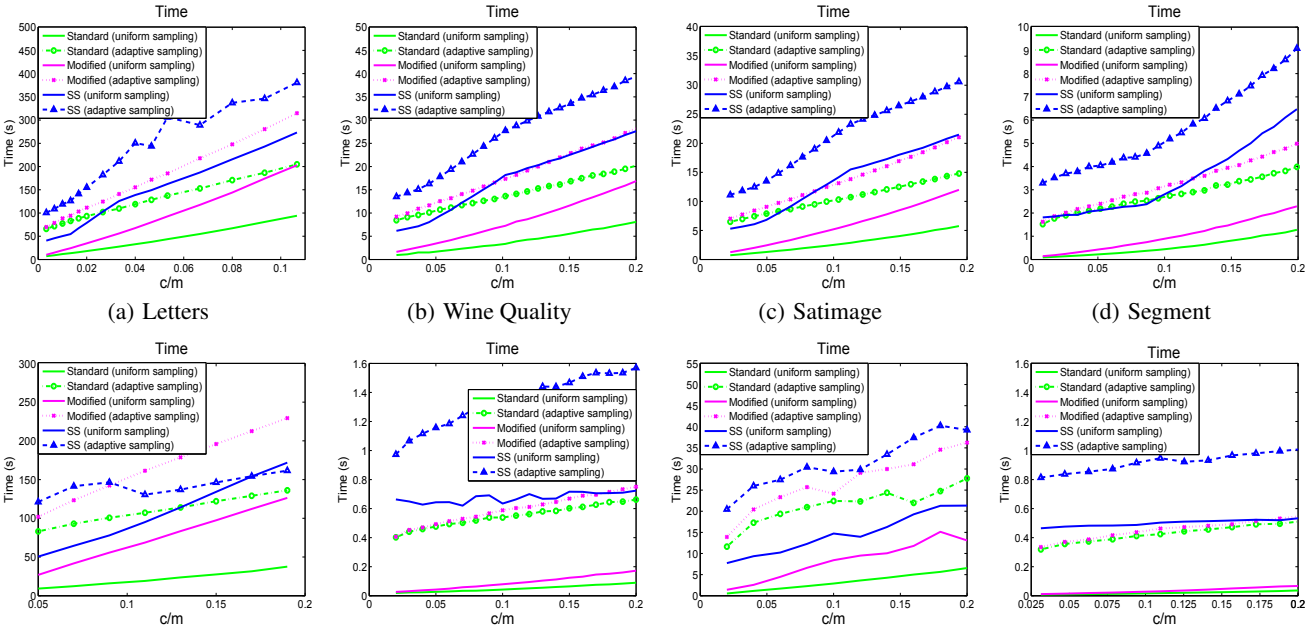
Figure 2: The error ratio $|\delta^{\text{opt}} - \tilde{\delta}|/\delta^{\text{opt}}$ versus l/k. In each subfigure, the left corresponds to the RBF kernel matrix with scaling parameter $\alpha = 0.1$, and the right corresponds to $\alpha = 1$.



(a) Letters (b) Wine Quality (c) Satimage (d) Segment

(e) DNA (f) German (g) Splice (h) Breast Cancer

Figure 3: The kernel approximation error incurred by the standard Nyström, modified Nyström, and SS-Nyström on the dense RBF kernels.



(a) Letters (b) Wine Quality (c) Satimage (d) Segment

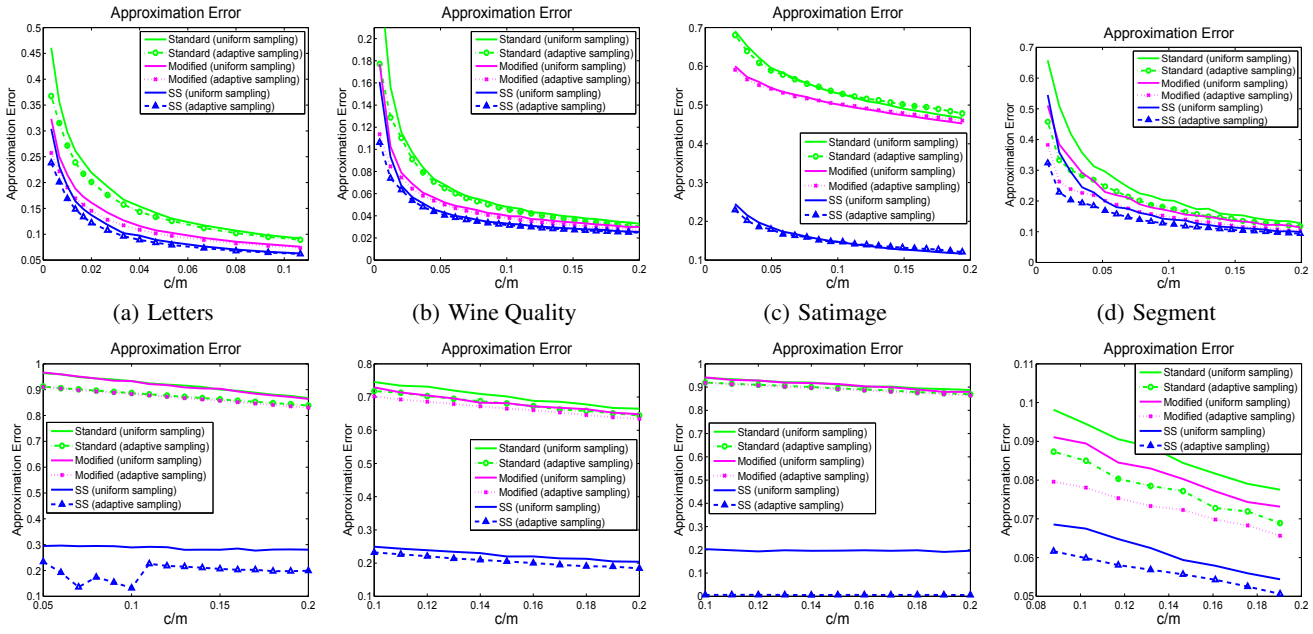
(e) DNA (f) German (g) Splice (h) Breast Cancer

Figure 4: The elapsed time of the standard Nyström, modified Nyström, and SS-Nyström on the dense RBF kernels.

time, our SS-Nyström is a little slower than the modified Nyström because SS-Nyström needs to compute δ^{opt} approximately by randomized SVD, which costs time $\mathcal{O}(mk^2) + T_{\text{multiply}}(m^2k)$ (as we set $l = 4k$). Since it costs time $\mathcal{O}(mc^2) + T_{\text{multiply}}(m^2c)$ to compute the modified Nyström approximation, so our SS-Nyström should be only constant times slower than the modified Nyström; this is verified by experiments.

7.4 Large-Scale Experiment

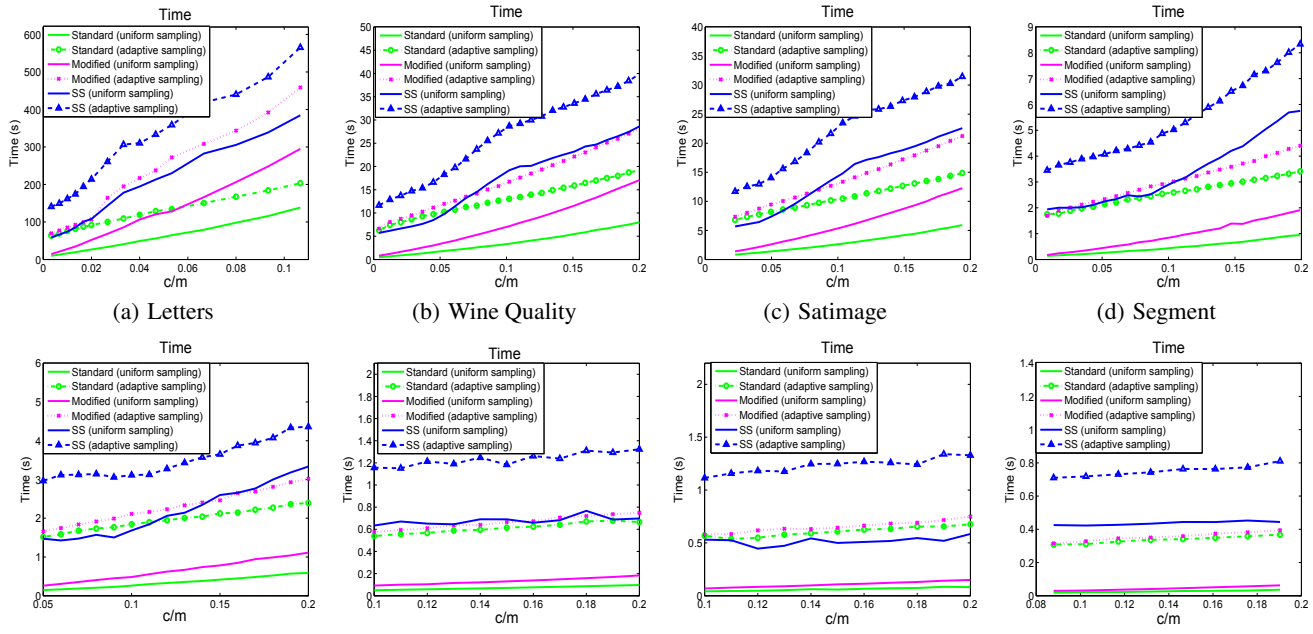
Finally, we compare SS-Nyström with the standard/modified Nyström methods on a large-scale dataset. We use the MNIST



(a) Letters (b) Wine Quality (c) Satimage (d) Segment

(e) DNA (f) German (g) Splice (h) Breast Cancer

Figure 5: The kernel approximation error incurred by the standard Nyström, modified Nyström, and SS-Nyström on the sparse RBF kernels.



(a) Letters (b) Wine Quality (c) Satimage (d) Segment

(e) DNA (f) German (g) Splice (h) Breast Cancer

Figure 6: The elapsed time of the standard Nyström, modified Nyström, and SS-Nyström on the sparse RBF kernels.

[16] dataset which has $60,000$ instance. We generate an RBF kernel with the scaling parameter $\alpha = 5$. The kernel matrix of MNIST has size of $60,000 \times 60,000$ which exceed the RAM of our workstation. We partition the kernel matrix to 30 blocks of size $60,000 \times 2,000$ and store them in the disk; at each time at most one block is loaded into the RAM. We only use uniform sampling to construct the approximations because other sampling methods are much more expensive, and we set $k = c/3$ for SS-Nyström. The standard Nyström method goes one pass through the data, the modified Nyström method goes two passes, and SS-Nyström (Algorithm 1) goes four passes. We report the approximation error (the minimum of 10 repeats) in Figure 7. We can see that the error

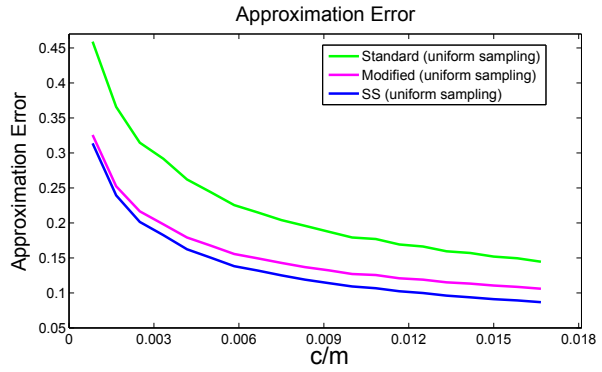


Figure 7: The kernel approximation error incurred by the standard Nyström, modified Nyström, and SS-Nyström on the MNIST dataset.

incurred by SS-Nyström is lower than that of the standard/modified Nyström methods. This set of experiments demonstrates that our proposed SS-Nyström method is still feasible for large-scale data that do not fit in RAM.

8. CONCLUSIONS

The Nyström method is an important kernel approximation method for enabling large-scale machine learning algorithms. In this paper we have proposed the SS-Nyström method which is a variant of the Nyström method and can speedup many kernel methods in the same way as the standard/modified Nyström methods. We have shown that SS-Nyström has a much stronger error bound than the standard/modified Nyström methods. Especially, when the bottom eigenvalues of a kernel matrix are not sufficiently small, the approximation accuracy of the standard/modified Nyström method or even the truncated SVD is unsatisfactory, while our SS-Nyström can still generate approximations of high accuracy. We have also devised an algorithm for computing SS-Nyström efficiently. Finally, the experiments have further demonstrated the effectiveness and efficiency of our SS-Nyström method.

9. ACKNOWLEDGEMENT

Shusen Wang is supported by Microsoft Research Asia Fellowship 2013 and the Scholarship Award for Excellent Doctoral Student granted by Chinese Ministry of Education. Hui Qian is supported by the National Natural Science Foundation of China (No. 61272303) and the National Program on Key Basic Research Project of China (973 Program, No. 2010CB327903). Zhihua Zhang is supported by the National Natural Science Foundation of China (No. 61070239).

APPENDIX

A. PROOF OF THEOREMS

We prove the three theorems of this paper respectively in the following subsections.

A.1 Proof of Theorem 2

PROOF. Since (2) is in convex and δ^{opt} is the minimizer of (2), then for any $\delta \in (0, \delta^{\text{opt}}]$, it holds that

$$\sum_{j=k+1}^{m} \big(\sigma_j(\mathbf{K}) - \delta\big)^2 \;\le\; \sum_{j=k+1}^{m} \big(\sigma_j(\mathbf{K}) - 0\big)^2 \;=\; \big\|\mathbf{K} - \mathbf{K}_k\big\|_F^2.$$

Then the theorem follows by the inequality (2) that any δ in the given interval can result in a smaller error. □

A.2 Proof of Theorem 3

PROOF. The error incurred by SS-Nyström is

$$\begin{aligned}\big\|\mathbf{K} - \tilde{\mathbf{K}}_c^{\text{ss}}\big\|_F^2 \;&=\; \big\|\big(\bar{\mathbf{K}} + \delta^{\text{opt}}\mathbf{I}_m\big) - \big(\bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T + \delta^{\text{opt}}\mathbf{I}_m\big)\big\|_F^2 \\ &=\; \big\|\bar{\mathbf{K}} - \bar{\mathbf{C}}\bar{\mathbf{U}}\bar{\mathbf{C}}^T\big\|_F^2 \;\le\; (1+\epsilon)\big\|\bar{\mathbf{K}} - \bar{\mathbf{K}}_k\big\|_F^2 \\ &=\; (1+\epsilon)\sum_{i=k+1}^{m}\sigma_i^2\big(\bar{\mathbf{K}}\big) \;=\; (1+\epsilon)\sum_{i=k+1}^{m}\lambda_i\big(\bar{\mathbf{K}}^2\big).\end{aligned}$$

Here the inequality follows from the property of the column selection algorithm $\mathcal{A}_{\text{col}}$. The i-th largest eigenvalue of $\bar{\mathbf{K}}$ is $\lambda_i(\mathbf{K}) - \delta^{\text{opt}}$, so the m eigenvalues of $\bar{\mathbf{K}}^2$ are all in the set $\{(\lambda_i(\mathbf{K}) - \delta^{\text{opt}})^2\}_{i=1}^m$. The sum of the least $m-k$ of the m eigenvalues of $\bar{\mathbf{K}}^2$ must be less than or equal to the sum of any $m-k$ of the eigenvalues, thus we have

$$\begin{aligned}\sum_{i=k+1}^{m}\lambda_i\big(\bar{\mathbf{K}}^2\big) \;&\le\; \sum_{i=k+1}^{m}\Big(\lambda_i(\mathbf{K}) - \delta^{\text{opt}}\Big)^2 \\ &=\; \sum_{i=k+1}^{m}\lambda_i^2(\mathbf{K}) - 2\sum_{i=k+1}^{m}\delta^{\text{opt}}\lambda_i(\mathbf{K}) + (m-k)(\delta^{\text{opt}})^2 \\ &=\; \|\mathbf{K} - \mathbf{K}_k\|_F^2 - \frac{1}{m-k}\Big[\sum_{i=k+1}^{m}\lambda_i(\mathbf{K})\Big]^2,\end{aligned}$$

by which the theorem follows. □

A.3 Proof of Theorem 5

PROOF. Let $\tilde{\mathbf{K}} = \mathbf{Q}(\mathbf{Q}^T\mathbf{K})_k$, where $\mathbf{Q}$ is defined in Line 4 in Algorithm 1. It was shown in [2] that

$$\mathbb{E}\|\mathbf{K} - \tilde{\mathbf{K}}\|_F^2 \;\le\; (1 + k/l)\,\|\mathbf{K} - \mathbf{K}_k\|_F^2, \tag{5}$$

where the expectation is taken w.r.t. the random Gaussian matrix $\boldsymbol{\Omega}$.

It follows from Lemma 6 that

$$\|\boldsymbol{\sigma}_{\mathbf{K}} - \boldsymbol{\sigma}_{\tilde{\mathbf{K}}}\|_2^2 \;\le\; \|\mathbf{K} - \tilde{\mathbf{K}}\|_F^2,$$

where $\boldsymbol{\sigma}_{\mathbf{K}}$ and $\boldsymbol{\sigma}_{\tilde{\mathbf{K}}}$ contain the singular values in a descending order. Since $\tilde{\mathbf{K}}$ has a rank at most k, the $k+1$ to n entries of $\boldsymbol{\sigma}_{\tilde{\mathbf{K}}}$ are zero. We split $\boldsymbol{\sigma}_{\mathbf{K}}$ and $\boldsymbol{\sigma}_{\tilde{\mathbf{K}}}$ into vectors of length k and $m-k$:

$$\boldsymbol{\sigma}_{\mathbf{K}} \;=\; \begin{bmatrix}\boldsymbol{\sigma}_{\mathbf{K},k}\\ \boldsymbol{\sigma}_{\mathbf{K},-k}\end{bmatrix} \quad\text{and}\quad \boldsymbol{\sigma}_{\tilde{\mathbf{K}}} \;=\; \begin{bmatrix}\boldsymbol{\sigma}_{\tilde{\mathbf{K}},k}\\ \mathbf{0}\end{bmatrix}$$

and thus

$$\|\boldsymbol{\sigma}_{\mathbf{K},k} - \boldsymbol{\sigma}_{\tilde{\mathbf{K}},k}\|_2^2 + \|\boldsymbol{\sigma}_{\mathbf{K},-k}\|_2^2 \;\le\; \|\mathbf{K} - \tilde{\mathbf{K}}\|_F^2. \tag{6}$$

Since $\|\boldsymbol{\sigma}_{\mathbf{K},-k}\|_2^2 = \|\mathbf{K} - \mathbf{K}_k\|_F^2$, it follows from (5) and (6) that

$$\mathbb{E}\|\boldsymbol{\sigma}_{\mathbf{K},k} - \boldsymbol{\sigma}_{\tilde{\mathbf{K}},k}\|_2^2 \;\le\; \frac{k}{l}\,\|\boldsymbol{\sigma}_{\mathbf{K},-k}\|_2^2.$$

Since $\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \sqrt{k}\|\mathbf{x}\|_2$ for any $\mathbf{x} \in \mathbb{R}^k$, we have that

$$\mathbb{E}\big\|\boldsymbol{\sigma}_{\mathbf{K},k} - \boldsymbol{\sigma}_{\tilde{\mathbf{K}},k}\big\|_1 \;\le\; \frac{k}{\sqrt{l}}\,\big\|\boldsymbol{\sigma}_{\mathbf{K},-k}\big\|_1.$$

Then it follows from (3) and Line 6 in Algorithm 1 that

$$\begin{aligned}\mathbb{E}\left|\delta^{\text{opt}} - \tilde{\delta}\right| &= \mathbb{E}\left[\frac{1}{m-k}\left|\sum_{i=1}^{k}\sigma_i(\mathbf{K}) - \sum_{i=1}^{k}\sigma_i(\tilde{\mathbf{K}})\right|\right] \\ &\leq \frac{1}{m-k}\,\mathbb{E}\left\|\boldsymbol{\sigma}_{\mathbf{K},k} - \boldsymbol{\sigma}_{\tilde{\mathbf{K}},k}\right\|_1 \qquad (7) \\ &\leq \frac{k}{\sqrt{l}}\,\frac{1}{m-k}\left\|\boldsymbol{\sigma}_{\mathbf{K},-k}\right\|_1 = \frac{k}{\sqrt{l}}\,\delta^{\text{opt}}.\end{aligned}$$

□

The following lemma is used to prove the theorem. The lemma is easy to prove, so here we do not show the detailed proof.

LEMMA 6. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be square matrices and* $\boldsymbol{\sigma}_{\mathbf{A}}$ *and* $\boldsymbol{\sigma}_{\mathbf{B}}$ *contain the singular values in a descending order. Then we have that*

$$\|\boldsymbol{\sigma}_{\mathbf{A}} - \boldsymbol{\sigma}_{\mathbf{B}}\|_2^2 \leq \|\mathbf{A} - \mathbf{B}\|_F^2.$$

References

[1] R. H. Affandi, A. Kulesza, E. B. Fox, and B. Taskar. Nyström approximation for large-scale determinantal processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.

[2] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near optimal column-based matrix reconstruction. In *Annual Symposium on Foundations of Computer Science (FOCS)*, 2011.

[3] C. Cortes, M. Mohri, and A. Talwalkar. On the impact of kernel approximation on learning accuracy. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[4] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. *Theory of Computing*, 2(2006):225–247, 2006.

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

[6] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, Sept. 2008.

[7] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.

[8] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[9] M. G. Genton. Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2001.

[10] A. Gittens and M. W. Mahoney. Revisiting the nyström method for improved large-scale machine learning. In *International Conference on Machine Learning (ICML)*, 2013.

[11] V. Guruswami and A. K. Sinop. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1207–1214. SIAM, 2012.

[12] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[13] J. B. Hough, M. Krishnapur, Y. Peres, B. Virág, et al. Determinantal processes and independence. *Probability Surveys*, 3:206–229, 2006.

[14] R. Jin, T. Yang, M. Mahdavi, Y. Li, and Z. Zhou. Improved bounds for the nystrom method with application to kernel classification. *IEEE Transactions on Information Theory*, 59(10):6939–6949, 2013.

[15] S. Kumar, M. Mohri, and A. Talwalkar. Sampling methods for the Nyström method. *Journal of Machine Learning Research*, 13:981–1006, 2012.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[17] M. Li, X.-C. Lian, J. T. Kwok, and B.-L. Lu. Time and space efficient spectral clustering via column sampling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[18] M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[19] D. Michie, D. J. Spiegelhalter, and C. C. Taylor. *Machine Learning, Neural and Statistical Classification*. Prentice Hall, 1994.

[20] E. J. Nyström. Über die praktische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben. *Acta Mathematica*, 54(1):185–204, 1930.

[21] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *International Conference on Machine Learning (ICML)*, 1998.

[22] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2002.

[23] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[24] S. Si, C.-J. Hsieh, and I. Dhillon. Memory efficient kernel approximation. In *International Conference on Machine Learning (ICML)*, pages 701–709, 2014.

[25] J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[26] A. Talwalkar, S. Kumar, M. Mohri, and H. Rowley. Large-scale svd and manifold learning. *Journal of Machine Learning Research*, 14:3129–3152, 2013.

[27] S. Wang, C. Zhang, H. Qian, and Z. Zhang. Using the matrix ridge approximation to speedup determinantal point processes sampling algorithms. In *The 28th AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

[28] S. Wang and Z. Zhang. Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling. *Journal of Machine Learning Research*, 14:2729–2769, 2013.

[29] S. Wang and Z. Zhang. Efficient algorithms and error analysis for the modified Nyström method. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

[30] C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2001.

[31] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[32] K. Zhang and J. T. Kwok. Clustered Nyström method for large scale manifold learning and dimension reduction. *IEEE Transactions on Neural Networks*, 21(10):1576–1587, 2010.

[33] K. Zhang, I. W. Tsang, and J. T. Kwok. Improved Nyström low-rank approximation and error analysis. In *International Conference on Machine Learning (ICML)*, 2008.

[34] Z. Zhang. The matrix ridge approximation: algorithms and applications. *Machine Learning*, pages 1–32, 2014.