A Review of Nyström Methods for Large-Scale Machine Learning

Shiliang Sun*, Jing Zhao, Jiang Zhu

Shanghai Key Laboratory of Multidimensional Information Processing,
Department of Computer Science and Technology, East China Normal University,
500 Dongchuan Road, Shanghai 200241, P. R. China

Abstract

Generating a low-rank matrix approximation is very important in large-scale machine learning applications. The standard Nyström method is one of the state-of-the-art techniques to generate such an approximation. It has got rapid developments since being applied to Gaussian process regression. Several enhanced Nyström methods such as ensemble Nyström, modified Nyström and SS-Nyström have been proposed. In addition, many sampling methods have been developed. In this paper, we review the Nyström methods for large-scale machine learning. First, we introduce various Nyström methods. Second, we review different sampling methods for the Nyström methods and summarize them from the perspectives of both theoretical analysis and practical performance. Then, we list several typical machine learning applications that utilize the Nyström methods. Finally, we make our conclusions after discussing some open machine learning problems related to Nyström methods.

Key words:

Low-rank approximation, Nyström method, sampling method, machine learning

*Corresponding author. Tel.: +86-21-54345183; fax: +86-21-54345119.
Email address: shiliangsun@gmail.com, slsun@cs.ecnu.edu.cn (Shiliang Sun)

1. Introduction

Many large-scale machine learning problems involve generating a low-rank matrix approximation to reduce high time and space complexities. For example, let n be the number of data instances. The Gaussian process regression computes the inverse of an $n \times n$ matrix which takes time $O(n^3)$ and space $O(n^2)$. For the large-scale problems, n can be in the order of tens of thousands to millions, leading to difficulties in operating on, or even storing the matrix.

Various methods [1, 2, 3] have been utilized to generate low-rank matrix approximations. The standard Nyström method is one of the state-of-the-art methods. It selects a subset of columns of the original matrix to build an approximation. In general, the standard Nyström method is used to approximate symmetric positive semidefinite (SPSD) matrices, such as Gram or kernel matrices, or their eigenvalues/eigenvectors. For approximating matrix K, it consists of three steps. 1) Sampling step: it samples a subset of columns of K to form matrix C; 2) Pseudo-inverse step: it performs pseudo-inverse of the matrix W formed by the intersection between those sampled columns and the corresponding rows; 3) Multiplication step: it constructs a matrix by using the formulation $CW^{\dagger}C^{\top}$ to approximate the original matrix. For approximating eigenvalues/eigenvectors, it also consists of three steps. 1) Sampling step: it samples a subset of columns of K to form C; 2) Singular value decomposition (SVD) step: it performs SVD of the matrix W formed by the intersection between those sampled columns and the corresponding rows to get singular values and singular vectors, respectively; 3) Extension step: it uses the Nyström extension to get the approximate eigenvalues/eigenvectors of the original matrix.

The standard Nyström method was first introduced into Gaussian process regression for reducing the computational complexity [4]. By replacing the original matrix with a Nyström approximation and subsequently using the Woodbury formula, the matrix inver-

sion can be easily solved with time complexity $O(\ell^2 n)$, where ℓ columns are sampled. After that, the standard Nyström method has got rapid developments. Several enhanced Nyström methods have been developed to provide more accurate matrix approximation or eigenvector approximation, e.g., density-weighted Nyström (DW-Nyström), ensemble Nyström, modified Nyström and modified Nyström method by spectral shifting (SS-Nyström). In addition, some techniques are developed to improve the inner procedures of the Nyström approximation. For the sampling step, various sampling methods [5, 6, 7] are utilized for the Nyström methods. For the SVD step, recently an approximate SVD [8] that utilizes randomized SVD algorithms [9] was proposed to accelerate the standard Nyström method for some extreme large-scale machine learning applications.

One key aspect of the Nyström methods is the sampling step. It influences the subsequent approximation accuracy and thus the performance of the learning methods [10]. Initially, uniform sampling is adopted when the standard Nyström method was applied [4], and it is also the most widely applied sampling method due to its low time consumption. After that, various sampling methods [6, 11, 12, 13, 14, 15] that focus on selecting the most informative columns are proposed. Thus, we call these sampling methods informative-column sampling. We classify these methods into two classes: 1) fixed sampling; 2) adaptive sampling. For fixed sampling, it means the matrix is sampled with a fixed, non-uniform distribution over the columns. The distribution can be defined by a function on the diagonal entities or the column entities of the original matrix [6, 11]. For adaptive sampling, it means the matrix is sampled with adaptive techniques [12, 13, 16, 17]. Different from fixed sampling, the sampling distribution will be modified at each iteration. Empirical results suggest that a tradeoff between efficiency and accuracy exists for uniform sampling and informative-column sampling as the latter spends more time in finding a concise subset of informative columns but can provide an improved approximation accuracy. The tradeoff also exists for fixed sampling and adaptive sampling. In real world

applications, the tradeoff should be considered carefully before utilizing different sampling techniques.

The standard Nyström method has been applied to many machine learning applications, e.g., manifold learning [18, 19, 20], spectral clustering [21, 22, 23, 24], kernel-based methods such as kernel support vector machine (SVM) [25, 13, 26, 27] and kernel ridge regression [10, 27], signal processing [28, 29] and statistical learning [30, 31]. Talwalkar et al. [18] proved that the standard Nyström method combined with Isomap [32] is an efficient tool to extract the low-dimensional manifold structure given millions of high-dimensional face images, and the approximate Isomap tends to perform better than Laplacian Eigenmaps [33] and is tied to the original Isomap on both clustering and classification with the labeled CMU-PIE [34] data set. In the work of Fowlkes et al. [22], spectral clustering with the standard Nyström method outperforms the traditional Lanczos method [35] where the clustering result is measured by normalized cut [36]. In the work of Williams et al. [26], kernel-based Gaussian processes have been accelerated using the Nyström approximation to the kernel matrix, with the time complexity scaled down from $O(n^3)$ to $O(\ell^2 n)$.

Nyström methods can be seen as a kind of information fusion methods. They use the partial data many times to approximate the values that we are interested, such as the eigenvalues/eigenvectors of a matrix or the inverse of a matrix. When applied to machine learning problems, Nyström methods will bring improvements in efficiency on the premise of not reducing performance much.

The remainder of the paper is organized as follows. Section 2 introduces some preliminary knowledge. In Section 3, we introduce the various Nyström methods. In Section 4, some related low-rank matrix approximation methods are presented to differentiate from the Nyström methods. In Section 5, several sampling methods for the Nyström methods including uniform sampling and informative-column sampling are listed, and we subse-

quently give some comparisons between uniform sampling and informative-column sampling from the perspectives of both theoretical analysis and practical performance. In Section 6, we provide a summary of typical large-scale machine learning applications of the Nyström methods. We make our conclusions in section 8 after discussing several open machine learning problems related to Nyström methods in Section 7.

2. Preliminary Knowledge

2.1. Notations

For an $n \times n$ SPSD matrix $K = [K_{ij}]$, we define $K^{(j)}, j = 1, ..., n$, as the jth column vector of K, $K_{(i)}, i = 1, ..., n$, as the ith row vector of K. For a vector $\mathbf{x} \in R^n$, let $\|\mathbf{x}\|_\xi$, $\xi = 1, 2, \infty$, denote the 1-norm, Euclidean norm, and ∞-norm, respectively. Let Diag(K) denote the vector consisting of the diagonal entries of the matrix K and Σ denote the matrix containing the eigenvalues of K. Then, $\|K\|_2 = \|Diag(\Sigma)\|_\infty$ denotes the spectral norm of K; $\|K\|_F = \|Diag(\Sigma)\|_2$ denotes the Frobenius norm of K; and $\|K\|_\otimes = \|Diag(\Sigma)\|_1$ denotes the trace norm (or nuclear norm) of K. Clearly,

$$\|K\|_2 \leq \|K\|_F \leq \|K\|_\otimes \leq \sqrt{n}\|K\|_F \leq n\|K\|_2. \tag{1}$$

2.2. Best Rank-k Approximation

Given a matrix with rank(K)=p, the SVD of K can be written as $K = U\Sigma V$, where Σ is diagonal and contains the singular values ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$) of K, and U and V have orthogonal columns and contain the left and right singular vectors of K. Let U_k and V_k be the first k ($k < p$) columns of U and V, respectively, and Σ_k be the $k \times k$ top sub-block of Σ. Then, the $n \times n$ matrix $K_k = U_k \Sigma_k V_k$ is the best rank-k approximation to K, when measured with respect to any unitarily-invariant matrix norm, e.g., the spectral, Frobenius, or trace

norm [37]. We have

$$\|K - K_k\|_2 = \lambda_{k+1}, \tag{2}$$

$$\|K - K_k\|_F = (\sum_{i=k+1}^{n} \lambda_i^2)^{1/2}, \tag{3}$$

$$\|K - K_k\|_\otimes = \sum_{i=k+1}^{n} \lambda_i. \tag{4}$$

2.3. Pseudo-Inverse of a Matrix

Another kind of useful Nyström related knowledge is the pseudo-inverse (Moore-Penrose inverse). The pseudo-inverse of an $m \times n$ matrix A can be expressed from the SVD of A as follows. Let the SVD of A be

$$A = U \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} V^\top, \tag{5}$$

where U, V are both orthogonal matrices, and S is a diagonal matrix containing the (non-zero) singular values of A on its diagonal. Then the pseudo-inverse of A is an $n \times m$ matrix defined as

$$A^\dagger = V \begin{pmatrix} S^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^\top. \tag{6}$$

Note that $A^\dagger$ has the same dimension as the transpose of A. If A is square, invertible, then its pseudo-inverse is the true inverse, that is, $A^\dagger = A^{-1}$.

2.4. Orthogonal Projection

In linear algebra and functional analysis, a projection is a linear transformation $\mathcal{P}$ from a vector space to itself such that $\mathcal{P}^2 = \mathcal{P}$. A projection is orthogonal if and only if it is self-adjoint. One way to construct the projection operator on the range space of A is that

$$\mathcal{P}_A = A(A^\top A)^\dagger A^\top = AA^\dagger = HH^\top = UU^\top, \tag{7}$$

where H represents the orthogonal basis on the range space of A and U represents the left singular vectors of A corresponding to the non-zero singular values. Given an $n \times \ell$ matrix C, the projection of K onto the column space of C is defined as $\mathcal{P}_C K = CC^{\dagger} K$.

2.5. Matrix Coherence and Leverage Score

As Nyström related techniques, the leverage score and the matrix coherence that measure the structural nonuniformity should be introduced. The statistical leverage scores of a matrix K are the squared row-norms of any matrix whose columns are obtained by orthogonalizing the columns of the matrix, and the coherence is closely related to the largest leverage score [37]. These quantities play an important role in several machine learning algorithms because they capture the key structural nonuniformity of the matrix. Given matrix K and a rank parameter k, the statistical leverage scores of K relative to the best rank-k approximation to K equal the squared Euclidean norms of the rows of the $n \times k$ matrix U_k

$$\ell_j = \left\| U_{k(j)} \right\|_2^2 .$$

The matrix coherence of K relative to the best rank-k approximation of K is defined by

$$\mu = \frac{n}{k} \max_j \left\| U_{k(j)} \right\|_2^2 .$$

3. Nyström Methods

The standard Nyström method was originally introduced to handle approximation for numerical integration in integral equations. It can be deemed as quadrature methods for integral equation approximation. The standard Nyström method can be used for approximating eigenfunctions. In dealing with matrices, it can be used for approximating SPSD matrices and their eigenvectors/eigenvalues. Besides the standard Nyström method, some

extended methods including DW-Nyström, ensemble Nyström, modified Nyström and SS-Nyström are developed to improve the accuracy of approximation. To illustrate these Nyström methods, we start from the quadrature rule, and then introduce these Nyström methods from the perspectives of eigenvectors/eigenvalues approximation and (or) matrix approximation.

Theoretical analysis and experimental results are two basic ways to evaluate the performance with a specific Nyström method. In most work of the Nyström methods, people bound the spectral reconstruction errors in two forms: 1) relative-error bound; 2) additive-error bound. The relative-error bound is in the form of

$$\left\|K - \widetilde{K}_k\right\|_\xi \leq \alpha \left\|K - K_k\right\|_\xi, \tag{8}$$

while the additive-error bound is in the form of

$$\left\|K - \widetilde{K}_k\right\|_\xi \leq \left\|K - K_k\right\|_\xi + \alpha. \tag{9}$$

Here, $\widetilde{K}_k$ represents the rank-k approximation to K by Nyström methods, K_k represents the best rank-k approximation to K, and α is a constant and ξ represents the spectral norm, Frobenius norm or trace norm.

3.1. Quadrature Method

We consider the quadrature method of solving eigenfunction problem for the operator which is expressed as the convolved integral of a kernel function [38]

$$\int_{\mathcal{D}} \kappa(x, y)\phi_i(x)dx = \lambda_i\phi_i(y), \quad i = 1, ..., N, \tag{10}$$

where λ_i represents each eigenvalue and $\phi_i(x)$ represents the corresponding eigenfunction for the operator. The kernel $\kappa(x, y)$ is defined as the product of two mapping functions (or feature functions) and eigenvalues:

$$\kappa(x, y) = \sum_{i=1}^{N} \lambda_i\phi_i(x)\phi_i(y). \tag{11}$$

The resulting solution of (10) is first found at the set of quadrature node points, and then extended to all points in $\mathcal{D}$ by means of a special interpolation formula. This method requires the use of a quadrature rule. Computing the integral occurring in (10) requires using the quadrature rule

$$\int_{\mathcal{D}} y(x)dx = \sum_{j=1}^{n} w_j y(x_j), \tag{12}$$

where $\{w_j\}_{j=1}^n$ are the weights and $\{x_j\}_{j=1}^n$ are the quadrature points that are determined by the particular quadrature rule. Then we get

$$\int_{\mathcal{D}} \kappa(x,y)\phi_i(x)dx \simeq \sum_{j=1}^{n} w_j \kappa(y,x_j)\phi_i(x_j) = \lambda_i \phi_i(y), i = 1, ..., N. \tag{13}$$

In order to form the symmetric kernel matrix K, the same samples $\{x_k\}_{k=1}^n$ are employed for approximating different values of y. Then the values of function $\phi_i(y)$ with inputs $\{x_k\}_{k=1}^n$ form the vector $[\phi_i(x_1), \phi_i(x_2), ..., \phi_i(x_n)]^\top$. This leads to a system of n algebraic equations

$$\sum_{j=1}^{n} w_j \kappa(x_j,x_k)\phi_i(x_j) = \lambda_i \phi_i(x_k), \quad k = 1, ..., n, \quad i = 1, ..., N. \tag{14}$$

If we write $K = [\kappa(x_j, x_k)]$, $\boldsymbol{\phi}_i = [\phi_i(x_1), \phi_i(x_2), ..., \phi_i(x_n)]^\top$, $U = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_n]$, $W = diag(w_1, w_2, ..., w_n)$ and $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n)$, then the above equations yield the matrix eigenvalue problem

$$KWU = \Lambda U. \tag{15}$$

Since KW is probably asymmetric, (15) can be converted to

$$(W^{\frac{1}{2}} K W^{\frac{1}{2}}) W^{\frac{1}{2}} U = \Lambda W^{\frac{1}{2}} U, \tag{16}$$

which is eigenvalue decomposition of symmetric matrix $(W^{\frac{1}{2}} K W^{\frac{1}{2}})$. If $\widetilde{\Lambda}$ and $\widetilde{U}$ are the eigenvalues and eigenvectors of $(W^{\frac{1}{2}} K W^{\frac{1}{2}})$, respectively, then

$$\Lambda = \widetilde{\Lambda}, \quad U = W^{-\frac{1}{2}} \widetilde{U}. \tag{17}$$

Finally, by substituting the entries of U back into (13), we get the approximate eigenfunction as

$$\phi_i(y) = \frac{1}{\lambda_i} \sum_{j=1}^{n} w_j \kappa(y, x_j) \phi_i(x_j). \tag{18}$$

3.2. Standard Nyström Method

In most machine learning problems, it is more likely that there is a probability density $p(x)$ over the input space which is smoothly varying, rather than being constant within $\mathcal{D}$ and zero outside. In this case the eigenfunction problem is generalized to include the probability function $p(x)$ as

$$\int \kappa(x, y) p(x) \phi_i(x) dx = \lambda_i \phi_i(y). \tag{19}$$

Given iid samples $\{x_j\}_{j=1}^n$ from $p(x)$, this eigenfunction equation is approximated by replacing the integral with an empirical average

$$\int \kappa(x, y) p(x) \phi_i(x) dx \simeq \frac{1}{n} \sum_{j=1}^{n} \kappa(x_j, y) \phi_i(x_j). \tag{20}$$

It is equivalent to the quadrature method in (13) when $w_1 = w_2 =, ..., w_n = \frac{1}{n}$, which yields the matrix eigenvalue problem

$$KWU = \frac{1}{n} KU = \Lambda U \Longleftrightarrow KU = n\Lambda U. \tag{21}$$

This can be regarded as the eigenvalue decomposition of K, with the eigenvalue $\widetilde{\Lambda} = diag(\widetilde{\lambda}_1, \widetilde{\lambda}_2, ..., \widetilde{\lambda}_n)$ and eigenvector $\widetilde{U}$. Combining the fact $K = \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top$ and the kernel definition (11), we get

$$\Lambda = \frac{1}{n}\widetilde{\Lambda}, \quad U = \sqrt{n}\widetilde{U}, \tag{22}$$

Given the results, the interpolation method (18) can be used to compute the eigenfunction. The Nyström approximation to the ith eigenfunction is

$$\phi_i(y) \simeq \frac{1}{n\lambda_i} \sum_{j=1}^{n} \kappa(y, x_j) \phi_i(x_j) = \frac{\sqrt{n}}{\widetilde{\lambda}_i} \sum_{j=1}^{n} \kappa(y, x_j) \widetilde{U}_{(j)}^{(i)}. \tag{23}$$

By using the idea of approximating eigenfunctions, the standard Nyström method can generate an approximation $\widetilde{K}_k^{\text{nys}}$ of an SPSD matrix K based on a sample of ℓ of its columns. Despite sampling methods, we assume that the sample of ℓ columns is given to us. Let C denote the $n \times \ell$ matrix formed by these columns and W be the $\ell \times \ell$ matrix consisting of the intersection of these ℓ columns with the corresponding ℓ rows. Without loss of generality, the columns and rows can be rearranged to [6]

$$K = \begin{bmatrix} W, & K_{21}^\top \\ K_{21}, & K_{22} \end{bmatrix}, \quad C = \begin{bmatrix} W \\ K_{21} \end{bmatrix}. \tag{24}$$

Since K is SPSD, the submatrix W is also SPSD. Let W_k be the best rank-k approximation of W and the eigenvalue decomposition of W_k be $W_k = U_{W,k}\Sigma_{W,k}U_{W,k}^\top$, the standard Nyström method [4] approximates eigenvalues (Σ_k) and eigenvectors (U_k) by using the following extensions

$$\widetilde{\Sigma}_k^{\text{nys}} = (\frac{n}{\ell})\Sigma_{W,k}, \quad \widetilde{U}_k^{\text{nys}} = \sqrt{\frac{\ell}{n}} C U_{W,k}\Sigma_{W,k}^\dagger. \tag{25}$$

The rank-k approximation $\widetilde{K}_k^{\text{nys}}$ of K generated by the standard Nyström method for $k < n$ is computed by

$$\widetilde{K}_k^{\text{nys}} = \widetilde{U}_k^{\text{nys}}\widetilde{\Sigma}_k^{\text{nys}}(\widetilde{U}_k^{\text{nys}})^\top = CW_k^\dagger C^\top. \tag{26}$$

Note that when $k = \ell$, the standard Nyström approximation of K is $CW^\dagger C^\top$. Given C, the time complexity of matrix approximation using the standard Nyström method is $O(k\ell^2) + T_{\text{Multiply}}(n\ell k)$ where $O(k\ell^2)$ is for the eigenvalue decomposition on W and $T_{\text{Multiply}}(n\ell k)$ [39, 40] is for the multiplication in (26). The space complexity is $O(n\ell)$. In this paper, we present the time complexity in the form of $O(*)+T_{\text{Multiply}}(*)$ since the matrix multiplication can be done block-wisely and in parallel.

Regardless of the sampling methods, deterministic error bounds for matrix approximation using the standard Nyström method are given in the following theorem.

Theorem 1 (Deterministic Error Bounds for the Standard Nyström Method [41, 37]). *Let $K \in \mathbb{R}^{n\times n}$ be an arbitrary SPSD matrix K with eigenvalue decomposition $K = U\Sigma U^\top$ where U and Σ are partitioned as*

$$U = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix}. \tag{27}$$

Here U_1 has k-columns and spans the top k-dimensional eigenspace of K, and $\Sigma_1 \in \mathbb{R}^{k\times k}$ is full-rank. We denote the eigenvalues of K with $\lambda_1(K) \geq ... \geq \lambda_n(K)$. Let S denote the sketching matrix which is the product of sampling matrix $R \in \mathbb{R}^{n\times\ell}$ and scaling matrix $D \in \mathbb{R}^{\ell\times\ell}$ and let

$$\Omega_1 = U_1^\top S \quad \text{and} \quad \Omega_2 = U_2^\top S \tag{28}$$

denote the projection of S onto the top and bottom eigenspaces of K, respectively. Then the deterministic error bounds for the standard Nyström method are obtained as

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_2 \leq \|K - K_k\|_2 + \left\|\Sigma_2^{1/2}\Omega_2\Omega_1^\dagger\right\|_2^2, \tag{29}$$

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_F \leq \|K - K_k\|_F + \left\|\Sigma_2^{1/2}\Omega_2\Omega_1^\dagger\right\|_2 \left(\sqrt{2\text{Tr}(\Sigma_2)} + \left\|\Sigma_2^{1/2}\Omega_2\Omega_1^\dagger\right\|_F\right), \tag{30}$$

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_\otimes \leq \|K - K_k\|_\otimes + \left\|\Sigma_2^{1/2}\Omega_2\Omega_1^\dagger\right\|_F^2. \tag{31}$$

Note that the error bounds in Theorem 1 are adapted from [37] by using the facts that $\|\Sigma_2\|_2 = \|K - K_k\|_2$, $\|\Sigma_2\|_F = \|K - K_k\|_F$ and $\|\Sigma_2\|_\otimes = \|K - K_k\|_\otimes$.

Theorem 2 (Lower Error Bounds for the Standard Nyström Method [42]). *Whatever column sampling algorithm is used, there exists an $n \times n$ SPSD matrix K such that the error incurred by the standard Nyström method obeys:*

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_F^2 \geq 1 + \frac{n^2k - \ell^3}{\ell^2(n-k)} \|K - K_k\|_F^2, \tag{32}$$

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_2 \geq \frac{n}{\ell} \|K - K_k\|_2, \tag{33}$$

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_\otimes \geq \frac{n-\ell}{n-k}(1 + \frac{k}{\ell}) \|K - K_k\|_\otimes. \tag{34}$$

Here k is an arbitrary target rank, and ℓ is the number of selected columns.

From the lower error bounds, when the matrix size n is large, the standard Nyström approximation can be very inaccurate unless a large number of columns are selected.

3.3. DW-Nyström Method

The standard Nyström method assigns equal importance to all the chosen samples. In the density-weighted Nyström (DW-Nyström) method [15], a density function $p(x)$ evaluated at the landmark points $Z = \{z_j\}_{j=1}^n$ is explicitly introduced. Then the integral equation can be approximated as

$$\int \kappa(x, y)p(x)\phi_i(x)dx \simeq \frac{1}{c}\sum_{j=1}^{n} p(z_j)\kappa(x_j, y)\phi_i(x_j) = \lambda_i\phi_i(y), \tag{35}$$

where $c = \sum_{j=1}^n p(z_j)$ is the normalization factor. By choosing y at the landmark points and defining $U = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_n]$ with $\boldsymbol{\phi}_i = [\phi_i(z_1), \phi_i(z_2), ..., \phi_i(z_n)]^\top$ and $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ as in the standard Nyström method, we have

$$\widetilde{K}U = c\Lambda U, \tag{36}$$

where $\widetilde{K} \in \mathbb{R}^{n\times n}$ is the density-weighted kernel matrix evaluated at the landmark points,

$$\widetilde{K}_{jk} = p(z_k)\kappa(z_j, z_k), \quad j = 1, ..., n, \quad k = 1, ..., n. \tag{37}$$

After solving eigenvalue decomposition of $\widetilde{K}$, we can get the eigenvectors U and eigenvalues $c\Lambda$. Different from (23), the approximate eigenfunction by the DW-Nyström method can be evaluated at an arbitrary point x as

$$\phi_i(x) \approx \frac{1}{c\lambda_i}\sum_{j=1}^{n} p(z_j)\kappa(x, z_j)\phi_i(z_j). \tag{38}$$

In dealing with matrices, the SPSD matrix K is expressed as in (24). Let $\widetilde{W} = WP$ and $\widetilde{W}_k$ be the best rank-k approximation to $\widetilde{W}$. Then, the approximate eigenvectors of K are given by

$$\widetilde{U}_k^{\text{wny}} = \frac{1}{c} E U_{\widetilde{W},k} \Sigma_{\widetilde{W},k}^{\dagger}, \tag{39}$$

where $E = CP \in \mathbb{R}^{n\times\ell}$ and $P \in \mathbb{R}^{\ell\times\ell}$ is a diagonal matrix such that $P_{kk} = p(z_k)$, $(k = 1, ..., \ell)$. $U_{\widetilde{W},k}$ and $\Sigma_{\widetilde{W},k}$ are the eigenvectors and eigenvalues of $\widetilde{W}_k$. The landmark points $z_k's$ are set by the centers of the data clusters. The size of each cluster is $|S_k|$ which is obtained by the K-means clustering algorithm. In this case, $p(z_k) = \frac{1}{n}|S_k|$ and $c = \sum_{k=1}^{\ell} p(z_k) = 1$. The complexity of computing approximate eigenvectors is $O(\ell^3 + \ell n)$ including K-means $O(\ell n)$. Note that the matrix approximation through spectral reconstruction in the DW-Nyström method is still an open problem as the approximate eigenvalues are not provided [15].

3.4. Ensemble Nyström Method

The ensemble Nyström method [7] was proposed as a meta algorithm which combines the p standard Nyström methods with the mixture weights $w^{(r)}$, $(r = 1, ..., p)$. In particular, it selects a collection of p samples, each sample $C^{(r)}$ containing ℓ' columns of K. Then the ensemble method combines the samples to construct an approximation in the form of

$$\widetilde{K}_k^{\text{ens}} = \sum_{r=1}^{p} w^{(r)} C^{(r)} {W_k^{(r)}}^{\dagger} {C^{(r)}}^{\top} = \sum_{r=1}^{p} w^{(r)} \widetilde{K}_k^{\text{nys(r)}} \tag{40}$$

Typically, the ensemble Nyström method seeks to find out the weights by minimizing $\|K - \widetilde{K}_k^{\text{ens}}\|_F$ or $\|K - \widetilde{K}_k^{\text{ens}}\|_2$. A simple but effective strategy is to set the weights as $w^{(1)} = ... = w^{(p)} = \frac{1}{p}$. For computing, the time complexity of the ensemble Nyström method is $O(pk\ell'^2 + C_w) + T_{\text{Multiply}}(p\ell' kn)$ where C_w is the time for computing w. Kumar et al. [7] indicated that the ensemble sampling naturally fits within distributed computing environments. In this age, solving big data problem with distributed computing techniques

is more and more popular. Not only the sampling procedure could be done in this way, but also the matrix multiplication in the SVD procedure [8] can be performed in parallel.

3.5. Modified Nyström Method

More recently, the modified Nyström method [42] was proposed by borrowing the techniques in CUR matrix decomposition. It was motivated by the fact that the lower error bounds of the standard Nyström method and the ensemble Nyström method are even much worse than the upper bounds of some existed CUR algorithms. The modified Nyström method is formed as

$$\widetilde{K}^{\text{mod}} = CU^{\text{mod}}C^\top = C(C^\dagger K(C^\dagger)^\top)C^\top, \tag{41}$$

which is the projection of K onto the column space of C and the row space of $C^\top$. With the selected columns at hand, the modified Nyström method needs to go only one pass through the data. Although more expensive to compute, the modified Nyström method is a more accurate approximation. Since $U^{\text{mod}} = C^\dagger K(C^\dagger)^\top$ is the minimizer of the optimization problem $\min_U \left\|K - CUC^\top\right\|_F$, the modified Nyström method is in general more accurate than the standard Nyström method in that $\left\|K - CU^{\text{mod}}C^\top\right\|_F \leq \|K - K^{\text{nys}}\|_F$. A lower bound of the modified Nyström Method is established in Theorem 3.

Theorem 3 (The Lower Error Bound for the Modified Nyström Method [40]). *Whatever column sampling algorithm is used, there exists an $n \times n$ SPSD matrix K such that the error incurred by the modified Nyström method obeys*

$$\left\|K - \widetilde{K}^{\text{mod}}\right\|_F^2 \geq \frac{n-\ell}{n-k}\left(1 + \frac{2k}{\ell}\right)\|K - K_k\|_F^2 . \tag{42}$$

From the error bound, one can find that the error doesn't increase as the matrix size increases.

3.6. SS-Nyström Method

When the bottom eigenvalues of a kernel matrix are large, the previous Nyström methods work poorly. Moreover, the previous Nyström methods cannot be directly used for the approximation of K^{-1}. An extension of the modified Nyström method, which is called the modified Nyström method by spectral shifting (SS-Nyström) was proposed by Wang et al. [40] as

$$\widetilde{K}^{\mathrm{ss}} = \bar{C} U^{\mathrm{ss}} \bar{C}^{\top} + \delta^{\mathrm{ss}} I_n. \tag{43}$$

Here $\delta^{\mathrm{ss}} \geq 0$ is called the spectral shifting term. It is inspired by the matrix ridge approximation (MRA) [44]. MRA approximates any SPSD matrix by $AA^{\top} + \delta I_n$ where A is an $n \times \ell$ matrix and $\delta > 0$ is the average of the $n - c$ bottom eigenvalues. The MRA works well no matter whether the bottom eigenvalues are large or small. However, MRA is solved by an iterative algorithm, so it is not efficient. SS-Nyström inherits the efficiency of the Nyström methods and is effective even when the bottom eigenvalues are large.

The SS-Nyström method for matrix approximation is computed in three steps. First, (approximately) compute the initial spectral shifting term

$$\bar{\delta} = \frac{1}{n-k}\left(\mathrm{tr(K)} - \sum_{\mathrm{j}=1}^{\mathrm{k}} \sigma_{\mathrm{j}}(\mathrm{K})\right), \tag{44}$$

and then perform spectral shift $\bar{K} = K - \bar{\delta} I_n$, where $k \leq \ell$ is the target rank. Actually, exactly setting the initial spectral shifting term to be $\bar{\delta}$ is unnecessary because SS-Nyström has a better upper error bound than the modified Nyström method whenever the initial spectral shifting term falls in the interval $(0, \bar{\delta}]$. Second, use some column sampling algorithms to select ℓ columns of $\bar{K}$ to form $\bar{C}$. Finally, with $\bar{C}$ at hand, compute U^{ss} and δ^{ss} by

$$\delta^{\mathrm{ss}} = \frac{1}{n - \mathrm{rank}(\bar{C})}\left(\mathrm{tr}(K) - \mathrm{tr}(\bar{C}^{\dagger} K \bar{C})\right), \tag{45}$$

$$U^{\mathrm{ss}} = \bar{C}^{\dagger} K (\bar{C}^{\dagger})^{\top} - \delta^{\mathrm{ss}} (\bar{C}^{\top} \bar{C})^{\dagger}. \tag{46}$$

Analogously, the SS-Nyström finds the minimizer of the optimization problem $(U^{\text{ss}}, \delta^{\text{ss}}) = \underset{U,\delta}{\operatorname{argmin}} \left\| K - \bar{C}U\bar{C}^\top - \delta I_n \right\|_F$ to obtain the intersection matrix U^{ss} and the spectral shifting term δ^{ss}. However, computing $\bar{\delta}$ according to (44) requires the partial eigenvalue decomposition which costs time $O(n^2k)$ and space $O(n^2)$. This can be accelerated by computing the top k eigenvalues approximately using random projection techniques [40]. This method computes $\bar{\delta}$ in time $O(n\ell^2) + T_{\text{multiply}}(n^2\ell)$ and space $O(n\ell)$.

Given the SS-Nyström approximation of K, the matrix inverse $(K + \alpha I_n)^{-1}$ can be approximately computed. Let $U^{\text{ss}} = Z\Lambda Z^\top$ be the condensed eigenvalue decomposition of the intersection matrix of SS-Nyström, where $Z \in \mathbb{R}^{\ell \times r}$, $\Lambda \in \mathbb{R}^{r \times r}$, and $r = \text{rank}(U^{\text{ss}}) \leq \ell$. $(\widetilde{K}^{\text{ss}} + \alpha I_n)^{-1}$ is expanded by the Woodbury formula

$$(\widetilde{K}^{\text{ss}} + \alpha I_n)^{-1} = \tau^{-1} I_n - \tau^{-1}\bar{C}Z(\tau\Lambda^{-1} + Z^\top \bar{C}^\top \bar{C}Z)^{-1} Z^\top \bar{C}^\top, \tag{47}$$

where $\tau = \delta^{\text{ss}} + \alpha$. Note that when $\alpha = 0$, K^{-1} can be approximately computed.

In addition, the eigenvalue decomposition of K can be computed by using the SVD of $\bar{C}$. Let $\bar{C} = U_{\bar{C}}\Sigma_{\bar{C}}V_{\bar{C}}$ and $S = \Sigma_{\bar{C}}V_{\bar{C}}U^{\text{ss}}V_{\bar{C}}^\top\Sigma_{\bar{C}}^\top$. Then by performing eigenvalue decomposition of S as $S = U_S\Lambda_S U_S^\top$, one can get the eigenvalue decomposition of $\widetilde{K}^{\text{ss}}$ as

$$\widetilde{K}^{\text{ss}} = (U_{\bar{C}}U_S)(\Lambda_S + \delta I_r)(U_{\bar{C}}U_S)^\top + U_\perp(\delta^{\text{ss}} I_n)U_\perp^\top. \tag{48}$$

Here $U_\perp$ is a column orthogonal complementary matrix of $(U_{\bar{C}}U_S)$.

3.7. Discussions on the Nyström methods

We discuss some characteristics of the Nyström methods as follows. The standard Nyström method was proposed earlier than the others and easy to implement. It is most widely applied to the machine learning applications. The DW-Nyström method can only approximate eigenvectors of matrices. The approximation is more accurate than the standard Nyström method empirically [15, 43]. The ensemble Nyström method is more flexible and richer than the others, which can lead better approximation in specific cases [16].

Table 1: Time and space complexities of the Nyström methods on matrix reconstruction.

Methods	Time	Space
Standard Nyström	$O(k\ell^2) + T_{\text{Multiply}}(k\ell n)$	$O(n\ell)$
DW-Nyström	-	-
Ensemble Nyström	$O(pk\ell'^2 + C_w) + T_{\text{Multiply}}(p\ell' kn)$	$O(n\ell')$
Modified Nyström	$O(n\ell^2) + T_{\text{Multiply}}(n^2\ell)$	$O(n^2)$
SS-Nyström	$O(n\ell^2) + T_{\text{Multiply}}(n^2\ell)$	$O(n^2)$

From the lower bound of the matrix approximation, the modified Nyström method is more accurate than the standard Nyström method when the matrix becomes larger. However, the cost is more time and space [40]. The SS-Nyström method is more complex than the previous ones. It can approximate matrices well even when the bottom eigenvalues are large and can directly handle the inverse of matrices [40]. In addition, for clarity, we list the time and space complexities of these Nyström methods on matrix approximation in Table 1.

4. Related Low-Rank Matrix Approximation Methods

Since large-scale machine learning becomes more and more important, a wide range of work on low-rank matrix approximation has been proposed. From the recent literature, it can be summarized as two kinds of methods to approximate matrices, i.e., spectral reconstruction and matrix projection. In particular, given $K_k = U_k \Sigma_k V_k^\top$ where U_k and V_k contain the singular vectors of K corresponding to the top k singular values in Σ_k, $U_k \Sigma_k V_k^\top$ is referred to spectral reconstruction since it uses both the singular values and vectors, and $U_k U_k^\top K$ is referred to matrix projection since it uses only singular vectors to compute the projection of K onto the space spanned by vectors U_k. The standard Nyström method is

a spectral reconstruction based method while the modified Nyström and the SS-Nyström are matrix projection based methods. In addition, some related methods such as approximate SVD [11], column-sampling [45] and CUR [46] are matrix projection based methods while randomized SVD is based on spectral reconstruction. Further, these related methods generate approximation of an arbitrary matrix while the Nyström methods generates approximations only of SPSD matrices. Now, we introduce the characteristics of the related methods.

4.1. Approximate SVD

Let K be an arbitrary matrix and C be its scaled columns, the main idea of approximate SVD is to approximate the left singular vectors $\widetilde{U}$ and singular values $\widetilde{\Sigma}$ of K by the left singular vectors and singular values of the matrix C. The low-rank approximate matrix can be got by matrix projection as $\widetilde{U}\widetilde{U}^\top K$. There are two kinds of algorithms, LinearTimeSVD and ConstantTimeSVD [11], to perform the approximate SVD.

The strategy behind the LinearTimeSVD algorithm is to pick ℓ columns of the matrix K according to probabilities $\{p_i\}_{i=1}^n$, and rescale each $K^{(i_t)}$ by $C^{(t)} = K^{(i_t)}/\sqrt{\ell p_{i_t}}$ to form a matrix $C \in R^{n\times\ell}$ where $i_t \in 1, ..., n$ means the column number of K corresponding to the tth column of C, and then compute the singular values and corresponding left singular vectors of the matrix C. Particularly, the SVD of $C^\top C$ is performed instead of the SVD of C. The detailed procedure to get the left singular vectors of C is as follows. First, SVD decomposition of $C^\top C = V_C \Sigma_C^2 V_C^\top$ is performed to get the right singular vectors of C as V_C and the singular values of C as Σ_C. If given target rank k, $V_{C,k}$ and $\Sigma_{C,k}$ are got from V_C and Σ_C, respectively. Then the left singular vectors of C are obtained by $U_C = CV_{C,k}\Sigma_{C,k}^\dagger$.

The strategy behind the ConstantTimeSVD algorithm is to pick ℓ columns of the matrix K, and rescale each by an appropriate factor to form a matrix $C \in \mathbb{R}^{n\times\ell}$, and then compute approximations to the singular values and left singular vectors of the matrix C using partial

rows of C, which will then be approximations to the singular values and left singular vectors of K.

4.2. Randomized SVD

Randomized SVD [9] was proposed for constructing approximate, low-rank matrix decompositions. In general, it can be used on rectangular matrices. In order to compare with Nyström methods, we give the procedures of randomized SVD on a symmetric matrix K. There are three computational stages in randomized SVD. First, forming an $n \times \ell$ matrix by $Y = K\Omega$, where $\Omega \in \mathbb{R}^{n\times\ell}$ is a Gaussian random matrix. Then, constructing an orthonormal matrix $Q \in \mathbb{R}^{n\times\ell}$ whose columns are orthonormal bases for the range of Y. Finally, performing SVD of $B = Q^\top WQ$ resulting in $B = V\Lambda V^\top$. Here, B can be obtained by solving $B(Q^\top\Omega) = Q^\top Y$. Given above, the SVD of K can be approximated as $W = U\Lambda U^\top$, where $U = QV$. The randomized SVD algorithm can be used on the intersection W of the standard Nyström method to accelerate the approximation for some extreme large-scale machine learning problems.

4.3. Column-Sampling

Column-sampling [45] is a special case of the LinearTimeSVD when it use uniform sampling, i.e., $p_i = \frac{1}{n}$. In this case, C represents the selected columns of the matrix K without scaling. The approximate left singular vectors $\widetilde{U}$ are obtained by the left singular vectors of C and the approximate singular values $\widetilde{\Sigma}$ are obtained by rescaling the singular values of C as $\widetilde{\Sigma} = \sqrt{\frac{n}{\ell}}\Sigma_{C,k}$. Column-sampling has a higher time complexity than the standard Nyström method. Kumar et al. [45] proved that the standard Nyström method is better than column-sampling at spectral reconstruction, while the reconstruction accuracy of eigenvectors of column-sampling is higher.

4.4. CUR

The CUR algorithm [3, 46] seeks to find ℓ columns of K to form a matrix C, r rows to form a matrix R, and an intersection matrix U to approximate an arbitrary matrix K with the matrix multiplication of the above three matrices. Since the CUR algorithm has to select both informative columns and rows while the standard Nyström method needs only informative columns, the sampling step of CUR is harder than that of the standard Nyström method. However, the error bounds of the standard Nyström methods are much weaker than those of the existed CUR algorithms, especially the relative-error bounds. Wang and Zhang [42] borrowed the techniques in the CUR algorithm to improve the standard Nyström method.

4.5. Power Method

Given a unified formulation for the low-rank approximation as $\widetilde{A} = CW^{\dagger}C^{\top}$, one can obtain the optimal rank-k approximation to A by forming an SPSD sketch where the sketching matrix S is an orthonormal basis for the range of A_k, because with such a choice [37],

$$CW^{\dagger}C^{\top} = AS(S^{\top}AS)^{\dagger}S^{\top}A = A(SS^{\top}ASS^{\top})A = A(\mathcal{P}_{A_k}A\mathcal{P}_{A_k})^{\dagger}A = AA_k^{\dagger}A = A_k. \quad (49)$$

The power method can be used to obtain the sketching matrices S_q by taking $S_q = A^qS_0$ where q is a positive integer and $S_0 \in \mathbb{R}^{n\times\ell}$. One can reasonably expect that the sketching matrix S_q produces SPSD sketches of A with a lower additional error. Then the sketching model can be expressed as $C = A^qS$ and $W = S^{\top}A^{2q-1}S$. When $q = 1$, it is the standard Nyström method. When $q = 2$, it is the SPSD sketch proposed by Halko et al. [9] in the form of $A(\mathcal{P}_{AS}A\mathcal{P}_{AS})^{\dagger}A$ which is proved empirically more effective than the form of $\mathcal{P}_{AS}A\mathcal{P}_{AS}$. $\mathcal{P}_{AS}A\mathcal{P}_{AS}$ is the modified Nyström method. However, the power method needs to compute A^qS_0, which costs much more time. Thus, it is applicable when A is such that one can compute the A^qS_0 fast.

5. Sampling Methods for the Nyström Methods

Plenty of sampling methods exist for the Nyström methods. We classify most of these methods to uniform sampling and informative-column sampling. Uniform sampling aims at extracting a subset of columns with the same probability while informative-column sampling aims at extracting a subset of columns that have majority information among all the columns. According to the extracting measure, we classify these sampling methods to fixed sampling and adaptive sampling. Specially, fixed sampling methods include diagonal sampling, column-norm sampling, leverage-score sampling and determinant sampling while adaptive sampling methods include sparse matrix greedy approximation (SMGA) sampling, incomplete Cholesky decomposition (ICL) sampling, K-means sampling, adaptive-full sampling and adaptive-partial sampling. In addition to these sampling methods, there is one kind of deterministic sampling method and two compound sampling methods including near-optimal+adaptive sampling and uniform+adaptive^2 sampling. Here, the symbol ^{2} means using adaptive sampling twice. Considering for approximating SPSD matrix K with $\widetilde{K}_k$, various sampling methods differ in the scaling matrix $D \in \mathbb{R}^{\ell\times\ell}$ and the probabilities $\{p_i\}_{i=1}^n$ to select columns. As in Theorem 1, let $S \in \mathbb{R}^{n\times\ell}$ be the sketching matrix and $R \in \mathbb{R}^{n\times\ell}$ represent the sampling matrix, i.e., $R_{ij} = 1$ if the ith column of K is the jth selected column. C and W used in the Nyström methods are computed by

$$C = KS, \quad W = S^\top KS, \quad S = RD. \tag{50}$$

Now we give concrete descriptions and analysis for these sampling methods which are used to form C.

5.1. Uniform Sampling

Uniform sampling assumes a uniform distribution, so that each column is sampled with the same probability, i.e., $p_i = \frac{1}{n}$. The diagonal elements of the scaling matrix D

are all ones. It is straightforward to show that uniform sampling ignores the structural nonuniformity, which means that there are no special columns or there is no need to find such special columns which probably contain more information. There are two kinds of uniform sampling methods, i.e., sampling with replacement [45] or without replacement [4]. They differ in that columns would be replaced or not after they are selected. We give the error bounds of matrix reconstruction using the standard Nyström method with uniform sampling (with or without replacement) as follows.

Theorem 4 (Error Bounds for the Standard Nyström Method with Uniform Sampling [37]). *Let K be an $n \times n$ SPSD matrix and μ denote the coherence of the top k-dimensional eigenspace of K. Fix a failure probability $\delta \in (0, 1)$ and accuracy factor $\epsilon \in (0, 1)$. If $\ell \geq 2\mu(1-\epsilon)^2 \ln(k/\delta)$, then the corresponding low-rank SPSD approximation satisfies*

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_2 \leq \left(1 + \frac{n}{\epsilon\ell}\right)\|K - K_k\|_2\,, \tag{51}$$

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_F \leq \|K - K_k\|_F + \left(\frac{\sqrt{2}}{\delta\sqrt{\epsilon}} + \frac{1}{\epsilon\delta^2}\right)\|K - K_k\|_\otimes\,, \tag{52}$$

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_\otimes \leq \left(1 + \frac{1}{\delta^2\epsilon}\right)\|K - K_k\|_\otimes\,, \tag{53}$$

with probability at least $1 - 3\delta$.

Besides Theorem 4, error bounds of matrix reconstruction by the standard Nyström method with uniform sampling without replacement are presented in terms of spectral norm and Frobenius norm in Theorem 5.

Theorem 5 (Error Bounds for the Standard Nyström Method with Uniform Sampling Without Replacement [16]). *Let K be an $n \times n$ SPSD matrix. With probability at least $1 - \delta$, the following inequalities hold for any sample of size ℓ:*

$$\left\|K - \widetilde{K}^{\text{nys}}\right\|_2 \leq \|K - K_k\|_2 + \frac{2n}{\sqrt{\ell}}K_{\max}B_{\text{nys}}, \tag{54}$$

$$\left\|K - \widetilde{K}^{\text{nys}}\right\|_F \leq \|K - K_k\|_F + [\frac{64k}{\ell}]^{\frac{1}{4}}nK_{\max}B_{\text{nys}}^{\frac{1}{2}}, \tag{55}$$

where $B_{\text{nys}} = \left[1 + \sqrt{\frac{n-\ell}{n-1/2}\frac{1}{\beta(\ell,n)}\log\frac{1}{\delta}}d_{\max}^{K}/K_{\max}^{\frac{1}{2}}\right]$, $\beta(\ell,n) = 1 - \frac{1}{2\max\{\ell,n-\ell\}}$, $K_{\max} = \max_i K_{ii}$ *and* $d_{\max}^{K} = \max_{ij}\sqrt{K_{ii} + K_{jj} - 2K_{ij}}$.

Uniform sampling is widely used for its convenience and low time consumption. However, there are several drawbacks of uniform sampling. Intuitively, since uniform sampling ignores the matrix information, it would not fit for those data sets with some columns more informative. See the error bounds in Theorem 4, the reconstruction error of the standard Nyström method with uniform sampling is related to the coherence of the original matrix. If some columns are more informative, sampling columns with this method is not accurate enough.

As observed in previous studies [47] together with our analysis, the coherence of matrix μ plays an important role in measuring the approximation performance of the standard Nyström method using uniform sampling. In order to obtain an exact reconstruction of a low-rank matrix via the standard Nyström method, the number of columns needed to sample from K is related to the coherence defined by Talwalkar and Rostamizadeh [47]. In order to exploit the compressive sensing theory [48], [47] defined the coherence of the matrix K, which is adapted from [48], as

$$\mu' = \sqrt{N}\max_{1\le i,j\le n}|U_{k(i)}^{(j)}|, \tag{56}$$

where $U = (u_1, ..., u_n)$ is the eigenvector matrix of K. Intuitively, the coherence measures the degree to which the eigenvectors in U are correlated with the canonical bases. As discussed in the work of Candès and Recht [49], highly coherent matrices are difficult (even impossible) to be randomly recovered via matrix completion algorithms, and this same logic extends to the standard Nyström method with uniform sampling [47].

Theorem 6 (Performance of the Standard Nyström Method with Uniform Sampling [47]). *Let* $K \in \mathbb{R}^{n\times n}$ *be rank-k SPSD matrix and assume* $k \in O(1/\delta)$*, then it suffices to*

sample $\ell \geq O(r\mu'^2 \log(\delta^{-1})$ *columns to have with probability at least* $1-\delta$*,*

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\| = 0. \tag{57}$$

When the uniform sampling without replacement is applied in the ensemble Nyström method, we can get the error bounds as follows.

Theorem 7 (Error Bounds for the Ensemble Nyström Method with Uniform Sampling [16]). *Let* S *be a sample of* $p\ell'$ *columns drawn uniformly at random without replacement from* K*, decomposed into* p *subsamples of size* ℓ'*,* $S_1, ..., S_p$*. For* $r = 1, ..., p$*, let* $\widetilde{K}_k^{\text{nys(r)}}$ *denote the rank-k standard Nyström approximation of* K *based on the sample* S_r*. Then the following inequalities hold for any sample* S *of size* $p\ell'$ *and for any* w *in the unit simplex with probability at least* $1-\delta$*:*

$$\left\|K - \widetilde{K}_k^{\text{ens}}\right\|_2 \leq \|K - K_k\|_2 + \frac{2n}{\sqrt{\ell'}} K_{\max} B_{\text{ens}}, \tag{58}$$

$$\left\|K - \widetilde{K}_k^{\text{ens}}\right\|_F \leq \|K - K_k\|_F + [\frac{64k}{\ell'}]^{\frac{1}{4}} n K_{\max} B_{\text{ens}}^{\frac{1}{2}}, \tag{59}$$

where $B_{\text{ens}} = [1 + w_{\max} p^{\frac{1}{2}} \sqrt{\frac{n-p\ell'}{n-1/2} \frac{1}{\beta(\text{p}\ell', \text{n})} \log \frac{1}{\delta}} d_{\max}^K / K_{\max}^{\frac{1}{2}}]$*,* $\beta(\text{p}\ell', \text{n}) = 1 - \frac{1}{2\max\{\text{p}\ell', \text{n}-\text{p}\ell'\}}$*,* $K_{\max} = \max_i K_{ii}$ *and* $d_{\max}^K = \max_{ij} \sqrt{K_{ii} + K_{jj} - 2K_{ij}}$*.*

5.2. Diagonal Sampling

Diagonal sampling [6] samples ℓ columns of K to form the matrix C with probabilities $\{p_j\}_{j=1}^n$, where

$$p_j = \frac{K_{jj}^2}{\sum_{j=1}^n K_{jj}^2}. \tag{60}$$

The diagonal elements of the scaling matrix depend on the probabilities as $D_{jj} = 1/\sqrt{\ell P_i}$ if $R_{ij} = 1$. Error bounds of matrix reconstruction by the standard Nyström method with diagonal sampling are given below.

Theorem 8 (Error Bounds for the Standard Nyström Method with Diagonal Sampling [6]). *Let* $\eta = 1 + \sqrt{8 * \log(1/\delta)}$ *and fix a failure probability* $\delta \in (0, 1]$ *and approximation factor* $\epsilon \in (0, 1]$. *If* $\ell \geq 64k\eta^2/\epsilon^4$, *with* $1 - \delta$ *probability*

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_F \leq |K - K_k\|_F + \epsilon \sum_{i=1}^{n} K_{ii}^2. \tag{61}$$

If $\ell \geq 4\eta^2/\epsilon^2$, *with* $1 - \delta$ *probability*

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_2 \leq |K - K_k\|_2 + \epsilon \sum_{i=1}^{n} K_{ii}^2. \tag{62}$$

On the theoretical side, this is the first rigorous bound for the standard Nyström method.

5.3. Column-Norm Sampling

Column-norm sampling [11] uses the distribution with the probabilities $\{p_i\}_{i=1}^n$ to choose ℓ columns to form the matrix C, where

$$p_j = \frac{\left|K^{(j)}\right|_2^2}{\|K\|_F^2}. \tag{63}$$

The diagonal elements of the scaling matrix depend on the probabilities as $D_{jj} = 1/\sqrt{\ell P_i}$ if $R_{ij} = 1$. This sampling method has been combined with SVD approximation algorithms to bound the reconstruction error. Let $\eta = 1 + \sqrt{8\log(1/\delta)}$. Then, with probability at least $1 - \delta$, $\left\|KK^\top - CC^\top\right\|_F \leq \frac{\eta}{\sqrt{c}} \|K\|_F^2$. Moreover, it has been used in the CUR decomposition with the reconstruction error bound. As a corollary of a CUR decomposition for a general $m \times n$ matrix A with error bounds of $\|K - CUR\|_\xi \leq \|K - K_k\|_\xi + \epsilon \|K\|_F$, the standard Nyström method with column-norm sampling leads the error bounds as follows.

Theorem 9 (Error Bounds for the Standard Nyström Method with Column-Norm Sampling [6]). *Fix a failure probability* $\delta \in (0, 1]$ *and approximation factor* $\epsilon \in (0, 1]$. *If* $\ell \geq 64k\eta^2/\epsilon^4$, *with* $1 - \delta$ *probability*

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_F \leq |K - K_k\|_F + \epsilon \|K\|_F . \tag{64}$$

If $\ell \geq 4\eta^2/\epsilon^2$*, with* $1-\delta$ *probability*

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_2 \leq |K - K_k\|_2 + \epsilon \|K\|_F \,. \tag{65}$$

5.4. Leverage-Score Sampling

Recall that the leverage scores relative to the best rank-k approximation to K are the squared Euclidean norms of the rows of the $n \times k$ matrix U_k and $\ell_j = \left\|U_{k(j)}\right\|_2^2$. It follows from the orthonormality of U_k that $\sum_j (\ell_j/k) = 1$, and the leverage scores can thus be interpreted as a probability distribution over the columns of K. The scaling matrix D satisfies $D_{jj} = 1/\sqrt{\ell P_i}$ if $R_{ij} = 1$. This kind of sampling was first used in the work of Drineas et al. [46] for general matrices. Since computing the exact leverage scores requires performing SVD of the matrix, the general time complexity of leverage-score sampling is $O(n^2 k)$. Some techniques to compute approximate leverage scores [50, 51] can relieve the burden on computing. The standard Nyström approximation with leverage-score sampling provides improved spectral and Frobenius norm bounds relative to diagonal sampling.

Theorem 10 (Error Bounds for the Standard Nyström Method with Leverage-Score Sampling [37]). *Fix a failure probability* $\delta \in (0, 1]$ *and approximation factor* $\epsilon \in (0, 1]$*. If* $\ell \geq 3200(\epsilon^2)^{-1} k \ln(4k/(\delta))$*, the corresponding standard Nyström approximation with leverage-score sampling satisfies*

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_2 \leq \|K - K_k\|_2 + \left(\epsilon^2 \|K - K_k\|_\otimes\right), \tag{66}$$

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_F \leq \|K - K_k\|_F + \left(\sqrt{2}\epsilon + \epsilon^2\right) \|K - K_k\|_\otimes \,, \tag{67}$$

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_\otimes \leq (1 + \epsilon^2) \|K - K_k\|_\otimes \,, \tag{68}$$

simultaneously with probability at least $1 - 6\delta - 0.6$*.*

5.5. Determinant Sampling

Determinant sampling [23] samples ℓ columns together to form C with the probabilities $\{p(I)\}$ where I represents the multi-index satisfying $|I| = \ell$. In particular, $p(I)$ is defined by

$$p(I) = Z^{-1} \det(K_I), \tag{69}$$

where $Z = \sum_{I,|I|=\ell} \det(K_I)$ is a normalization constant. It is important to point out that the determinant sampling is usually computationally expensive as it requires computing the determinant of the submatrix for the selected columns/rows. Furthermore, computing the probability distribution $p(I) \propto \det(G_I)$ will cost a lot of time since the support of the distribution has cardinality $\binom{n}{\ell}$. The time and space complexities of this sampling method are $O(\binom{n}{\ell}\ell^3)$ and $O(\binom{n}{\ell})$, respectively. For the general case where $\text{rank}(K) \geq k$, we have the following error bound in expectation.

Theorem 11 (The Expectation Error Bound for the Standard Nyström Method with Determinant Sampling [23]). *Let K be a real, $n \times n$, positive quadratic form with eigenvalues $\lambda_1 \geq ... \geq \lambda_n$. Let $\widetilde{K}_k^{\text{nys}}$ be the standard Nyström approximation to K corresponding to multi-index I and $\ell = k$. Then*

$$\mathbb{E}\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_F \leq (k+1) \sum_{i=k+1}^{n} \lambda_i = (k+1) \left\|K - K_k\right\|_{\otimes}. \tag{70}$$

5.6. Sparse Matrix Greedy Approximation (SMGA) Sampling

Sparse Matrix Greedy Approximation (SMGA) [52] provides a low-rank kernel matrix approximation by using a greedy column selection algorithm. It operates by iteratively choosing one sample from a random subset of $m \ll n$ samples. The sampling scheme of SMGA can be used in conjunction with the Nyström methods, which is referred to as SMGA sampling. The time complexity of sampling ℓ columns of K to form C is $O(m\ell^2 n)$. However, there is no error bound for the Nyström approximation with SMGA sampling.

5.7. Incomplete Cholesky Decomposition (ICL) Sampling

Incomplete Cholesky decomposition (ICL) [53] is a variant of Cholesky decomposition which yields a low-rank approximation of K in the form of $\widetilde{K}^{\text{icl}} = \widetilde{X}\widetilde{X}^{\top}$, $\widetilde{X} \in \mathbb{R}^{n\times\ell}$. ICL is performed by a greedy selection process which skips columns below a certain threshold. The strategy of selecting columns in ICL is referred to as ICL sampling. Moreover, the Nyström approximation generated from the ℓ columns of K associated with the columns selected by ICL is identical to $\widetilde{K}^{\text{icl}}$ when $k = \ell$ [16]. The runtime of ICL is $O(\ell^2 n)$.

5.8. K-means Sampling

Zhang et al. [13] proposed a technique to generate informative columns using centroids resulting from K-means clustering. They analyzed that the approximation accuracy of the standard Nyström method is determined by the remarkable points, and subsequently the error bound depends on the distance between the original data and the approximate data. Thus, the adaptive scheme in K-means is feasible for the Nyström methods. Recent advances in speeding up the K-means algorithm [54, 55] also make it particularly suitable for large-scale problems. They proved that the error of the standard Nyström approximation is bounded by

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_F \leq 4T\sqrt{\ell C_x^k eT} + \ell C_x^k T e\left\|W^{-1}\right\|_F, \tag{71}$$

where $T = \max_k |S_k|$, S_k is the kth cluster of the data set, ℓ is the number of clusters, C_x^k is a constant depending on k and the data set, and $e = \sum_{i=1}^{n} \left\|x_i - z_{c(i)}\right\|_2^2$ is the total quantization error of coding each sample x_i with the closest landmark point $z_{c(i)}$. It is necessary to mention that this error bound is neither an additive-error bound nor a relative-error bound.

5.9. Adaptive-Full Sampling

Instead of sampling all ℓ columns of K from a fixed distribution, adaptive-full sampling [5] alternates between selecting a set of columns and updating the distribution over all

the columns. The sampling procedure is without replacement. Starting with an initial distribution over the columns, s columns ($s < l$) are chosen to form a submatrix C'. The probabilities are then updated as a function of previously chosen columns, and s new columns are sampled and incorporated in C'. This process is repeated until ℓ columns have been selected. That is, the iteration number is $\frac{\ell}{s}$. The probability p_i in the distribution is updated according to the reconstruction error for each row of K based on the current C'. The reconstruction is based on matrix projection as $U_{C'}U_{C'}^{\top}K$ where $U_{C'}$ is the eigenvector matrix of K. The bigger the error of the rows is, the higher probability the column should be selected with. Note that the sampling steps require a full pass over K at each step, and hence need $O(n^2)$ time and space.

5.10. Adaptive-Partial Sampling

Kumar et al. [16] extended the method adaptive-full to adaptive-partial, which is also sampling without replacement. At each iterative step, the reconstruction error for each row of C' is measured. The sampling probability of the corresponding column is updated in proportion to this error. The reconstructed $\widetilde{C}'_{k'}$ is obtained by the standard Nyström method with k'-rank. Unlike [5], the reconstruction error for C' is computed as $E = C' - \widetilde{C}'_{k'}$, which is much smaller than K, thus avoiding the $O(n^2)$ computation. Each iteration in this sampling procedure requires $O(n\ell k' + \ell^3)$ time and at most the storage of ℓ columns of K.

5.11. Deterministic Sampling

Deterministic sampling [23] is to choose the columns which contain the largest k diagonal elements of K. The deterministic algorithm requires finding the largest k diagonal elements of K, which can be done in $O(n \log k)$ steps. The worst-case error of the Nyström approximation with deterministic sampling is bounded in Theorem 12.

Theorem 12 (The Error Bound for the Standard Nyström Method with Deterministic Sampling [23]). *Let K be a real positive-definite kernel, I contain the indices of its k largest diagonal elements, and K_k^{nys} be the corresponding standard Nyström approximation. Then the reconstruction error satisfies*

$$\left\|K - \widetilde{K}_k^{\text{nys}}\right\|_F \leq \sum_{i \notin I} K_{ii}. \tag{72}$$

5.12. Near-Optimal+Adaptive Sampling

The near-optimal+adaptive sampling [42] was proposed for the modified Nyström method by combining the near-optimal column sampling [56] and the adaptive sampling [5]. This sampling algorithm consists of three steps: the approximate SVD via random projection [56, 9], the dual-set sparsification algorithm [56], and the adaptive sampling algorithm [5]. The algorithm costs $O(n\ell^2\epsilon^2 + nk^3\epsilon^{-2/3}) + T_{\text{Multiply}}(n^2\ell\epsilon)$ time and $O(n\ell)$ space in computing C. Theorem 13 gives the reconstruction error bound of the modified Nyström method with near-optimal+adaptive sampling.

Theorem 13 (The Error Bound for the Modified Nyström Method with Near-Optimal +Adaptive Sampling [40]). *Given an SPSD matrix $K \in R^{n\times n}$ and a target rank k, the algorithm samples $\ell = O(k\epsilon^{-2})$ columns of K to form C. We run the algorithm $t \geq (2\epsilon^{-1} + 1)\ln(1/\delta)$ times (independently in parallel) and choose the sample that minimizes $\left\|K - \widetilde{K}^{\text{mod}}\right\|_F$, then the inequality*

$$\left\|K - \widetilde{K}^{\text{mod}}\right\|_F \leq (1 + \epsilon)\left\|K - K_k\right\|_F \tag{73}$$

holds with probability at least $1 - \delta$.

The near-optimal+adaptive algorithm can also be used for the SS-Nyström method. If the columns of $\bar{K}$ are selected by this sampling algorithm, the error bound incurred by the SS-Nyström method is given in the following theorem.

Theorem 14 (The Error Bound for the SS-Nyström Method with Near-Optimal + Adaptive Sampling [40]). *Given an SPSD matrix $K \in R^{n\times n}$ and a target rank k, the algorithm samples $\ell = O(k\epsilon^{-2})$ columns of $\bar{K}$ to form $\bar{C}$. We run the algorithm $t \geq (2\epsilon^{-1} + 1)\ln(1/\delta)$ times (independently in parallel) and choose the sample that minimizes $\left\|\bar{K} - \bar{C}(\bar{C}^{\dagger}\bar{K}(\bar{C}^{\dagger})^{\top})\bar{C}^{\top}\right\|_F$, then the inequality*

$$\left\|K - \widetilde{K}_k^{\mathrm{ss}}\right\|_F^2 \leq (1+\epsilon)\left(\|K - K_k\|_F^2 - \frac{[\sum_{i=k+1}^{n}\lambda_i(K)]^2}{n-k}\right) \tag{74}$$

holds with probability at least $1 - \delta$.

5.13. Uniform+Adaptive^2 Sampling

Uniform+adaptive^2 sampling is a compound sampling algorithm which is composed of three steps [40]. In each step, partial columns of $C \in R^{n\times\ell}$ are sampled. Given SPSD matrix K, target rank k, error factor ϵ and matrix coherence μ, the detailed algorithm is described as follows. Firstly, uniformly sample $\ell_1 = 8.7\mu k\log(\sqrt{5}k)$ columns of K without replacement to construct C_1. Then, sample $\ell_2 = 10k\epsilon^{-1}$ columns to construct C_2 using adaptive-full sampling according to the residual $K - \mathcal{P}_{C_1}K$. Finally, sample $\ell_3 = 2\epsilon^{-1}(\ell_1 + \ell_2)$ columns to construct C_2 using adaptive-full sampling according to the residual $K - \mathcal{P}_{[C_1,C_2]}K$. Thus, $C = [C_1, C_2, C_3]$ with $\ell = \ell_1 + \ell_2 + \ell_3$. This algorithm costs $O(n\ell^2\epsilon^2) + T_{\mathrm{Multiply}}(n^2\ell\epsilon)$ time and $O(n\ell)$ space to construct C. Note that the parameter μ is often set to a constant (say 1) since computing the matrix coherence takes lots of time and the exact matrix coherence does not certainly result in the highest accuracy. The error bound for the modified Nyström method with the uniform+adaptive^2 sampling is given by Theorem 15.

Theorem 15 (The Error Bound for the modified Nyström method with Uniform + Adaptive^2 Sampling [40]). *If we sample $\ell = O(k\epsilon^{-2}) + \mu_k\epsilon^{-}1k\log k$ columns of K and run the algorithm $t \geq (20\epsilon^{-1} + 18)\log(1/p)$ times to choose the sample that minimizes*

$\left\|K - \widetilde{K}^{\text{mod}}\right\|_F$, *we can get the error bound of the matrix reconstruction by the modified Nyström method*

$$\left\|K - \widetilde{K}^{\text{mod}}\right\|_F \leq (1 + \epsilon) \|K - K_k\|_F \tag{75}$$

with probability at least $1 - p$.

5.14. Discussions on the Sampling Methods

There exists a tradeoff between efficiency and accuracy for uniform sampling and informative-column sampling as well as for fixed sampling and adaptive sampling. One should consider the tradeoff before applying the Nyström methods to specific applications.

Intuitively, uniform sampling is more efficient since it does not need to retrieve the whole data set before sampling. It is based on the idea that the original data set is uniformly informative. So there is no extra procedure needed to extract the more informative columns. Alternatively, informative-column sampling methods are more accurate since they retrieve the whole data set to measure the information of each column. Adaptive sampling is more accurate and less efficient than fixed sampling, as they regard the information to be changing with each column selected. So they have to retrieve the whole data set for several times. Gittens [41] suggested that uniform sampling should be adopted if the structural nonuniformity of the specific application is low such as for image processing. However, informative-column sampling is also recommended since real data sets of various machine learning applications [3, 57, 58, 59] are with great structural nonuniformity.

From the perspective of reconstruction error bounds, informative-column sampling is more tight. Bounds of various sampling methods have been listed. We can find that the bounds with the uniform sampling are not as tight as the informative-column sampling. This proves theoretically that informative-column sampling is more accurate.

From the perspective of practical performance, different sampling methods were compared in the literature. Several experiments with sampling methods including uniform, diagonal and column-norm sampling were conducted in the work of Kumar et al. [17], which show that uniform sampling is more powerful and unform sampling without replacement outperforms that with replacement. Further, it has been proved that the standard Nyström method with uniform sampling is more suitable when the coherence and the rank are low [47]. Note that low coherence indicates low structural nonuniformity, so no columns are very informative. In addition, some adaptive sampling methods are compared experimentally on the standard Nyström approximation [16]. Among the compared methods (uniform, ICL, SMGA, K-means, adaptive-full and adaptive-partial), K-means sampling performs best. When comparing uniform sampling with leverage-score sampling, it is found that the approximation accuracy depends on the characteristics of the matrix [37]. In particular, uniform sampling does quite well in the cases where matrices are constructed by linear kernels or less dense RBF kernels[1] with a large scale parameter. In these cases, the matrices have relatively low rank and uniform leverage scores. For matrices constructed by dense RBF kernels with smaller scale parameter or sparse RBF kernels, leverage-score sampling tends to perform much better. In addition, the time and space complexities for constructing C with different sampling methods are presented in Table 2. The table is made for providing a clear exhibition for all the introduced sampling methods.

[1]Here, the dense kernel is opposed to the sparse kernel. The degree of dense/sparse is measured by the percentage of nonzero entries in the constructed matrix [37].

Table 2: Time and space complexities for constructing C with different sampling methods.

Category	Sampling Methods	Time	Space
uniform	Uniform	$O(n)$	$O(n)$
fixed	Diagonal	$O(n)$	$O(n)$
fixed	Column-Norm	$O(n^2)$	$O(n)$
fixed	Determinant	$O(\binom{n}{\ell}\ell^3)$	$O(\binom{n}{\ell})$
fixed	Leverage-Score	$O(n^2k)$	$O(n^2)$
adaptive	SMGA	$O(m\ell^2n)$	$O(n^2)$
adaptive	ICL	$O(n\ell^2)$	$O(n^2)$
adaptive	K-means	$O(n\ell L)$	$O(\ell)$
adaptive	Adaptive-Full	$O(n^2)$	$O(n^2)$
adaptive	Adaptive-Partial	$O(\ell^3) + T_{\text{Multiply}}(n\ell^2)$	$O(n\ell)$
deterministic	Deterministic	$O(n \log k)$	$O(n)$
compound	Near-Optimal+Adaptive	$O(n\ell^2\epsilon^2 + nk^3\epsilon^{-2/3}) + T_{\text{Multiply}}(n^2\ell\epsilon)$	$O(n\ell)$
compound	Uniform+Adaptive^2	$O(n\ell^2\epsilon^2) + T_{\text{Multiply}}(n^2\ell\epsilon)$	$O(n\ell)$

6. The Nyström Methods for Large-Scale Machine Learning

We discuss in this section how to speedup 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.

Many methods such as Gaussian process regression, kernel SVM, and kernel ridge regression all require solving this kind of linear system $(K + \alpha I_n)b = y$, which amounts to the matrix inverse problem $b = (K + \alpha I_n)^{-1}y$. Here α is a constant. Given the low-rank approximation $\widetilde{K}_k = LL^\top$, with $L = CW_k^{-\frac{1}{2}}$ (an example in the standard Nyström method),

one can efficiently solve this linear system by utilizing the Woodbury formula

$$(K + \sigma I)^{-1} = \frac{1}{\sigma}(I - L(\sigma I + L^\top L)^{-1}L^\top). \tag{76}$$

Particularly, when $\alpha = 0$, one can use $\widetilde{K}_k^\dagger = (C^\dagger)^\top W_k C^\dagger$ to approximate K^{-1} or use SS-Nyström approximation to get $\widetilde{K}_k^{\text{ss}} = \bar{C}U^{\text{ss}}\bar{C}^\top + \delta^{\text{ss}} I_n$ and then use Woodbury formula as in (47).

There are also some other methods such as spectral clustering, kernel PCA, and manifold learning in need of eigenvalue decomposition. In practice, there are two ways to obtain approximate eigenvalues/eigenvectors of the kernel matrix by Nyström methods. The first is to directly use the standard Nyström extension, which simply computes C and W. It performs the eigenvalue decomposition of W and then extends its eigenvalues/eigenvectors to that of the complete kernel matrix. The SS-Nyström method also provides approximate eigenvectors/eigenvalues as in (48). However, the resultant eigenvectors are not guaranteed to be orthogonal. Thus orthogonalization is needed such as QR decomposition. The second approach first performs a low-rank approximation of $\widetilde{K}_k = LL^\top$, and then applies an additional process [14] to obtain an orthogonal set of approximate eigenvectors. In particular, the top k eigenvectors $U_k = [\mathbf{u}_1, ..., \mathbf{u}_k]$ (first k columns of U) of K can be obtained as $U = LV\Lambda^{-\frac{1}{2}}$, where $V, \Lambda \in \mathbb{R}^{\ell \times \ell}$ are from the eigenvalue decomposition of the $\ell \times \ell$ matrix $L^\top L = V\Lambda V^\top$. Now we introduce some large-scale machine learning applications using Nyström methods in detail.

6.1. Manifold Learning

Manifold learning is a hot research topic in the machine learning area. It aims at extracting low-dimensional structure from high-dimensional data [60, 61]. Instead of assuming global linearity, manifold learning methods make a weaker local-linearity assumption, that is, for nearby points in high-dimensional input space, L_2 distance is assumed to be

a good measure of geodesic along the manifold. However, when data are so large that some calculation is inefficient, some approximations are needed. One typical application is the use of the Nyström approximation to scale up Isomap [18]. Talwalkar et al. [18] showed that the Isomap coupled with Nyström approximation can effectively extract low-dimensional structure from datasets containing millions of images.

Isomap is a representative method in manifold learning. It aims to extract a low-dimensional data representation that best preserves all pairwise distances between input points. It involves three steps. 1) Construct the undirected adjacency graph with k'-nearest-neighbor rule (we use k' to differentiate it from the latter k); 2) Compute approximate geodesic distances and build a similarity matrix K; 3) Calculate the final embedding of the form $Y = (\Sigma_k)^{1/2} U_k^\top$ where Σ_k contains the top k largest eigenvalues of K and $U_k^\top$ are the eigenvectors corresponding to these eigenvalues.

It is easy to identify that Isomap needs eigenvalue decomposition. For those data sets with a large size, directly performing SVD is prohibited. Using the approximate eigenvectors and eigenvalues as in (25), the Nyström low-dimensional embeddings are obtained by

$$\widetilde{Y}^{nys} = \widetilde{\Sigma}_{nys,k}^{1/2} \widetilde{U}_{nys,k}^\top = ((\Sigma_{W,k})^{1/2})^\dagger U_{W,k}^\top C^\top.$$

6.2. Spectral Clustering

Spectral clustering is a clustering method based on the graph theory. It uses the eigenvectors got by eigenvalue decomposition of the similarity matrix to perform clustering. It involves three steps. 1) Calculate the similarity matrix with proper similarity function; 2) Calculate the eigenvalues and eigenvectors of the similarity matrix; 3) Choose a proper k and use the largest k eigenvectors as the input to perform K-means clustering.

Normalized cut [36] is one of the representative methods to implement the spectral clustering. The standard Nyström method is utilized to scale up spectral clustering with

normalized cut in the work of Fowlkes et al. [22], in which the high-resolution image segmentation can be solved efficiently. With approximate eigenvectors produced by the standard Nyström method, the required blocks could be calculated. Experimental results show that the standard Nyström method outperforms the famous Lanczos technique [35].

6.3. Gaussian Process and Other Kernel Methods

Gaussian process [4, 62] is a powerful and popular Bayesian technique. It suffers the cubic scale problem due to the inverse of kernel matrix. With Nyström approximation, Williams and Seeger [4] built spectral reconstruction of the original matrix and subsequently performed Woodbury formula to handle the matrix inversion. This procedure successfully scales Gaussian processes down from $O(n^3)$ to $O(\ell^2 n)$, which allows a very significant speed-up for the handwritten digit recognition without sacrificing accuracy. As the first area in machine learning that involving the Nyström approximation, the significant result inspires other researchers to apply the Nyström methods to other kernel-based methods.

Kernel-based methods [25, 63] are widely spread in machine learning. Kernel SVM, kernel PCA and kernel ridge regression are famous applications [64, 65]. In order to deal with large-scale cases, Fine and Scheinberg [2] utilized the ICL to approximate SVM. However, SVM coupled with the Nyström method could be a feasible direction. Cortes et al. [10] applied the standard Nyström method to several kernel-based methods and analyzed the impact of the approximation on these methods.

7. Open Problems

Now we give some open problems on the Nyström methods for large-scale machine learning applications.

7.1. Multiple-Kernel Learning

In recent years, several methods have been proposed to combine multiple kernels instead of using a single one [66]. These different kernels may correspond to using different notions of similarity or using different features from sources. Consider the simple linear combination $K = \sum_{i=1}^{p} \mu_i K_i$ of the multiple kernels. One way to approximate this matrix is to use the ensemble Nyström method in which the standard Nyström method is applied to each candidate kernel matrix K_i to get matrix $\widetilde{K}_i^{\text{nys}}$ [67]. The parameter μ corresponding to w in the ensemble Nyström method is defined using the approximation error between K_i and $\widetilde{K}_i^{\text{nys}}$. Thus, $\widetilde{K}^{\text{nys}} = \sum_{i=1}^{p} w_i \widetilde{K}_i^{\text{nys}}$. Since the different candidate kernel matrices have different properties such as dense or sparse, linear or nonlinear, sampling methods should be chosen carefully for the multiple Nyström approximations. Moreover, new methods to determine the parameter μ are expected.

7.2. Multi-task Learning

Multi-task learning is the machine learning branch that learns a task together with other related tasks at the same time. It has received a lot of attentions over the past ten years [68, 69, 70]. Consider a simple situation in which the correlations between tasks and between inputs are modeled by $K = K^f \otimes K^x$. Here, K^f specifies the inter-task similarities and K^x is the covariance matrix over inputs. There are different methods to form the matrix K^f. Multi-task Gaussian process proposed by Bonilla et al. [71] uses a free-form covariance matrix K^f over tasks. Some other multi-task methods construct K^f using kernel function $\kappa^f(t, t')$ over the task-descriptor features t. It is an issue to deal with large-scale cases with large numbers of inputs and tasks. Using the standard Nyström method to approximate the large matrix K^x is a natural idea which is adopted in the work of Bonilla et al. [71]. For approximating large matrix K^f, the similarities between tasks can be taken into consideration. K-means sampling is advised to be employed when K^f is constructed

by kernel function. When K^f is in free-form, it is an open problem to approximate K^f. Jointly approximating K^f and K^x is also worth considering.

7.3. Multiple-Output Model

Multiple-output models such as multiple-output Gaussian processes [72, 62] use a multiple-output kernel function to construct the covariance matrix. For large-scale machine learning, the covariance matrix is very large with the size of $ND \times ND$, where N denotes the number of data points and D denotes the dimension number of outputs. A general method to construct the covariance matrix is to use a convolution process. In this case the covariance can not be expressed in a form of Kronecker product of two matrix. It requires to use some methods such as Nyström methods to approximate the large matrix. Considering the above construction method for the matrix, we recommend to divide the matrix into $D \times D$ blocks averagely and use the Nyström method for each block. More precisely, we advise to sample $\ell/(D \times D)$ columns of each submatrix instead of sample ℓ columns of the whole matrix. However, when D is very large, this method is inefficient. More appropriate strategies for approximation should be further developed.

7.4. Cascade Nyström Approximation

In large-scale machine learning, we often encounter the situation to handle operators involving multiple large matrices, where a series of Nyström approximations are needed. For example, we may need to calculate $(K_1 + K_2)^{-1}$, where K_1 and K_2 are both large matrices. Intuitively, it can be solved according to the idea of the ensemble Nyström methods. In order to get higher accuracy, we suggest using the cascade Nyström approximation with adaptive sampling. In particular, we first use $\widetilde{K}_1^{\text{nys}}$ to approximate K_1. Then we approximate K_2 by using the adaptive-full (or adaptive-partial) sampling according to the residual $K_1 - \widetilde{K}_1^{\text{nys}} + K_2 - U_{C'}U_{C'}K_2$. We believe that this strategy has advantages and experimental verification is the future work.

8. Conclusions

In this paper, we review different kinds of Nyström methods for large-scale machine learning. Various sampling methods for the Nyström methods have been listed. We give illustrative comparisons for these methods from the perspectives of both theoretical analysis and practical performance. Then, we show some representative machine learning areas coupled with the Nyström methods, in which manifold learning, spectral clustering and kernel-based methods are highlighted. After that, we propose several open machine learning problems that could be further developed.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Project 61370175, and Shanghai Knowledge Service Platform Project (No. ZF1213).

References

[1] A. Frieze, R. Kannan, S. Vempala, Fast Monte-Carlo algorithms for finding low-rank approximations, in: Proceedings of the 39th Annual IEEE Symposium on Foundations of Computer Science, 1998, pp. 370–378.

[2] S. Fine, K. Scheinberg, Efficient SVM training using low-rank kernel representations, Journal of Machine Learning Research 2 (2001) 243–264.

[3] M. W. Mahoney, P. Drineas, CUR matrix decompositions for improved data analysis, in: Proceedings of the National Academy of Sciences, 2009, pp. 697–702.

[4] C. Williams, M. Seeger, Using the Nyström method to speed up kernel machines, Advances in Neural Information Processing Systems 13 (2001) 682–688.

[5] A. Deshpande, L. Rademacher, S. Vempala, G. Wang, Matrix approximation and projective clustering via volume sampling, in: Proceedings of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms, 2006, pp. 1117–1126.

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

[7] S. Kumar, M. Mohri, A. Talwalkar, Ensemble Nyström method, in: Proceedings of Neural Information Processing Systems, 2009, pp. 1060–1068.

[8] M. Li, J. T. Kwok, B. Lu, Making large-scale Nyström approximation possible, in: Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 631–638.

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

[10] C. Cortes, M. Mohri, A. Talwalkar, On the impact of kernel approximation on learning accuracy, in: Proceedings of 13th International Conference on Artificial Intelligence and Statistics, 2010, pp. 113–120.

[11] P. Drineas, R. Kannan, M. W. Mahoney, Fast Monte Carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix, SIAM Journal of Computing 36 (2006) 158–183.

[12] A. K. Farahat, A. Ghodsi, M. S. Kamel, A novel greedy algorithm for Nyström approximation, in: Proceedings of the 14th International Workshop on Artificial Intelligence and Statistics, 2011, pp. 269–277.

[13] K. Zhang, I. W. Tsang, J. T. Kwok, Improved Nyström low-rank approximation and error analysis, in: Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 1232–1239.

[14] K. Zhang, J. T. Kwok, Clustered Nyström method for large scale manifold learning and dimension reduction, IEEE Transactions on Neural Networks 21 (2009) 121–146.

[15] K. Zhang, J. T. Kwok, Density-weighted Nyström method for computing large kernel eigensystems, Neural Computation 21 (2009) 121–146.

[16] S. Kumar, M. Mohri, A. Talwalkar, Sampling methods for the Nyström method, Journal of Machine Learning Research 13 (2012) 981–1006.

[17] S. Kumar, M. Mohri, A. Talwalkar, Sampling techniques for the Nyström method, in: Proceedings of the 12th International Workshop on Artificial Intelligence and Statistics, 2009, pp. 304–311.

[18] A. Talwalkar, S. Kumar, H. Rowley, Large-scale manifold learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 3129–3152.

[19] A. Talwalkar, S. Kumar, H. Rowley, Large-scale SVD and manifold learning, Journal of Machine Learning Research 14 (2013) 3129–3252.

[20] J. C. Platt, Fast embedding of sparse music similarity graphs, Advances in Neural Information Processing Systems 16 (2004) 511–578.

[21] A. Y. Ng, M. I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, in: Proceedings of Neural Information Processing Systems, 2001, pp. 849–856.

[22] C. Fowlkes, S. Belongie, F. Chung, J. Malik, Spectral grouping using the Nyström method, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2002) 214–225.

[23] M. A. Belabbas, P. J. Wolfe, Spectral methods in machine learning and new strategies for very large datasets, in: Proceedings of the National Academy of Sciences, 2009, pp. 369–374.

[24] M. Li, X. Lian, J. T. Kwok, B. Lu, Time and space efficient spectral clustering via column sampling, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 2297 – 2304.

[25] S. Si, C. Hsieh, I. S. Dhillon, Memory efficient kernel approximation, in: Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 701–709.

[26] C. Williams, C. Rasmussen, V. T. A. Schwaighofer, Observations on the Nyström method for Gaussian process prediction, Technical Report, 2002.

[27] T. Yang, Y. Li, M. Mahdavi, R. Jin, Z. Zhou, Nyström method vs random fourier features: A theoretical and empirical comparison, Advances in Neural Information Processing Systems 24 (2012) 476–484.

[28] P. Parker, P. J. Wolfe, V. Tarok, A signal processing application of randomized low-rank approximation, in: Proceedings of the 13th IEEE Workshop on Statistical Signal Processing, 2005, pp. 345–350.

[29] D. N. Spendley, P. J. Wolfe, Adaptive beamforming using fast low-rank covariance matrix approximations, in: Proceedings of IEEE Radar Conference, 2008, pp. 1–5.

[30] M. A. Belabbas, P. J. Wolfe, On sparse representations of linear operators and the approximation of matrix products, in: Proceedings of the 42nd Annual Conference on Information Sciences and Systems, 2008, pp. 258–263.

[31] M. A. Belabbas, P. J. Wolfe, On landmark selection and sampling in high-dimensional data analysis, Philosophical Transactions of the Royal Society 367 (2009) 4295–4312.

[32] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.

[33] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Advanced Neural Information Processing Systems 13 (2001) 585–591.

[34] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression database, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (2003) 1615 – 1618.

[35] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, H. van der Vorst (Eds.), Templates for the solution of algebraic eigenvalue problems: a practical guide, SIAM, 2000.

[36] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 888–905.

[37] A. Gittens, M. W. Mahoney, Revisiting the Nyström method for improved large-scale machine learning, in: Proceedings of the 26th International Conference on Machine Learning, 2013, pp. 567–575.

[38] C. T. H. Baker, The numerical treatment of integral equations, Oxford:Glarendon Press, 1977.

[39] S. Wang, Z. Zhang, Efficient algorithms and error analysis for the modified Nyström method, in: Proceedings of the 17th International Conference on Artificial Intelligence and Statistics, 2014, pp. 1–11.

[40] S. Wang, L. Luo, Z. Zhang, The Modified Nyström Method: Theories, Algorithms, and Extension, Technical Report, 2014.

[41] A. Gittens, The spectral norm error of the naive Nyström extension, Technical Report, 2011.

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

[43] F. Shang, L. C. Jiao, J. Shi, M. Gong, R. H. Shang, Fast density-weighted low-rank approximation spectral clustering, Data Mining and Knowledge Discovery 23 (2011) 345–378.

[44] Z. Zhang, The matrix ridge approximation: algorithms and applications, Machine Learning 97 (2014) 227–258.

[45] S. Kumar, M. Mohri, A. Talwalkar, On sampling-based approximate spectral decomposition, in: Proceedings of the 26th International Conference on Machine Learning, 2009, pp. 553–560.

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

[47] A. Talwalkar, A. Rostamizadeh, Matrix coherence and the Nyström method, in: Proceedings of the 26th Conference in Uncertainty in Artificial Intelligence, 2010, pp. 572–579.

[48] E. J. Candès, J. Romberg, Sparsity and incoherence in compressive sampling, Inverse Problems 23 (2007) 969–986.

[49] E. J. Candès, B. Recht, Exact matrix completion via convex optimization, Foundations of Computational Mathematics 9 (2009) 717–772.

[50] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, D. P. Woodruff, Fast approximation of matrix coherence and statistical leverage, Journal of Machine Learning Research 13 (2012) 3475–3506.

[51] M. Mohria, A. Talwalkar, Can matrix coherence be efficiently and accurately estimated?, in: Proceedings of the 14th International Workshop on Artificial Intelligence and Statistics, 2011, pp. 534–542.

[52] A. J. Smola, B. Schölkopf, Sparse greedy matrix approximation for machine learning, in: Proceedings of International Conference on Machine Learning, 2000, pp. 911–918.

[53] F. R. Bach, M. I. Jordan, Kernel independent component analysis, Journal of Machine Learning Research 3 (2002) 1–48.

[54] E. Elkan, Using the triangular inequality to accelerate k-means, in: Proceedings of the 21th International Conference on Machine Learning, 2003, pp. 147–153.

[55] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, A. Y. Wu, An efficient k-means clustering algorithm: analysis and implementation, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2001) 881–892.

[56] C. Boutsidis, P. Drineas, M. Magdon-Ismail, Near optimal column-based matrix reconstruction, in: Proceedings of 52nd IEEE Annual Symposium on Foundations of Computer Science, 2011, pp. 305–314.

[57] C. Yip, M. W. Mahoney, A. S. Szalay, I. Csabai, T. Budavari, R. Wyse, L. Dobos, Objective identification of informative wavelength regions in galaxy spectra, The Astronomical Journal 147 (2014) 1–36.

[58] J. C. Platt, Fast embedding of sparse similarity graphs, Advanced Neural Information Processing Systems 15 (2003) 571–578.

[59] P. Paschou, E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, P. Drineas, PCA-correlated SNPs for structure identification in worldwide human populations, PLOS Genetics 3 (2007) 1672–1686.

[60] S. Sun, Tangent space intrinsic manifold regularization for data representation, in: Proceedings of the 1st IEEE China Summit and International Conference on Signal and Information Processing, 2013, pp. 179–183.

[61] Y. Zhou, S. Sun, Semi-supervised tangent space discriminant analysis, Mathematical Problems in Engineering Special Issue on Machine Learning with Applications to Autonomous Systems (2015).

[62] J. Zhao, S. Sun, Variational dependent multi-output Gaussian process dynamical systems, in: Proceedings of the 17th International Conference on Discovery Science, 2014, pp. 350–361.

[63] B. Scholkopf, A. Smola, Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT Press, 2002.

[64] J. Shawe-Taylor, S. Sun, Kernel methods and support vector machines, Book Chapter for E-Reference Signal Processing, Elsevier, 2013.

[65] X. Xie, S. Sun, Multi-view Laplacian twin support vector machines, Applied Intelligence 41 (2014) 1059–1068.

[66] J. Li, S. Sun, Nonlinear combination of multiple kernels for support vector machines, in: Proceedings of the 20th International Conference on Pattern Recognition, 2010, pp. 2889 – 2892.

[67] Z. Wang, W. Jie, D. Gao, A novel multiple Nyström-approximating kernel discriminant analysis, Neurocomputing 119 (2013) 385–398.

[68] S. Sun, Multitask learning for EEG-based biometrics, in: Proceedings of the 19th International Conference on Pattern Recognition, 2008, pp. 1–4.

[69] X. Xie, Multitask centroid twin support vector machines, Neurocomputing 149 (2015) 1085–1091.

[70] J. Zhu, S. Sun, Single-task and multitask sparse Gaussian processes, in: Proceedings of the International Conference on Machine Learning and Cybernetics, 2013, pp. 1033–1038.

[71] E. V. Bonilla, K. Chai, C. Williams, Multi-task Gaussian process prediction, Advances in Neural Information Processing Systems 20 (2008) 153–160.

[72] S. Sun, Infinite mixtures of multivariate Gaussian processes, in: Proceedings of the International Conference on Machine Learning and Cybernetics, 2013, pp. 1011–1016.