
Improved Nyström Low-Rank Approximation and Error Analysis

Kai Zhang
Ivor W. Tsang
James T. Kwok

TWINSSEN@CSE.UST.HK
IVOR@CSE.UST.HK
JAMESK@CSE.UST.HK

Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong

Abstract

Low-rank matrix approximation is an effective tool in alleviating the memory and computational burdens of kernel methods and sampling, as the mainstream of such algorithms, has drawn considerable attention in both theory and practice. This paper presents detailed studies on the Nyström sampling scheme and in particular, an error analysis that directly relates the Nyström approximation quality with the encoding powers of the landmark points in summarizing the data. The resultant error bound suggests a simple and efficient sampling scheme, the k -means clustering algorithm, for Nyström low-rank approximation. We compare it with state-of-the-art approaches that range from greedy schemes to probabilistic sampling. Our algorithm achieves significant performance gains in a number of supervised/unsupervised learning tasks including kernel PCA and least squares SVM.

1. Introduction

Kernel methods play a central role in machine learning and have demonstrated huge success in modelling real-world data with highly complex, nonlinear structures. Examples include the support vector machine, kernel Fisher discriminant analysis and kernel principal component analysis. The key element of kernel methods is to map the data into a kernel-induced Hilbert space $\varphi(\cdot)$ where dot product between points can be computed equivalently through the kernel evaluation $\langle \varphi(x_i), \varphi(x_j) \rangle = K(x_i, x_j)$. Given n sample points, this necessitates the calculation of an $n \times n$ symmetric, positive (semi-)definite kernel matrix. The resultant complexities in terms of both space (quadratic) and time (usually cubic) can be quite demanding for large problems, posing a big challenge on practical applications.

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008. Copyright 2008 by the author(s)/owner(s).

A useful way to alleviate the memory and computational burdens of kernel methods is to utilize the rapid decaying spectra of the kernel matrices (Williams & Seeger, 2000) and perform low-rank approximation in the form of $K = GG'$, where $G \in \mathbb{R}^{n \times m}$ with $m \ll n$. However, the optimal (eigenvalue) decomposition takes $O(n^3)$ time and efficient alternatives have to be sought. In the following, we give a brief review on efficient techniques for low-rank decompositions of symmetric, positive (semi-)definite kernel matrices.

Greedy approaches have been applied in several fast algorithms for approximating the kernel matrix. In (Smola & Schölkopf, 2000), the kernel matrix K is approximated by the subspace spanned by a subset of its columns. The basis vectors are chosen incrementally to minimize an upper bound of the approximation error. The algorithm takes $O(m^2nl)$ time using a probabilistic heuristic, where l is the random subset size. In (Ouibet & Bengio, 2005), a greedy sampling scheme is proposed based on how well a sample point can be represented by a (constrained) linear combination of the current subspace basis in the feature space. Their algorithm scales as $O(m^2n)$. Another well-known greedy approach for low-rank approximation of positive semi-definite matrices is the incomplete Cholesky decomposition (Fine & Scheinberg, 2001; Bach & Jordan, 2005; Bach & Jordan, 2002). It is a variant of the Cholesky decomposition that skip pivots below a certain threshold, and factorizes the kernel matrix K as $K = GG'$ where $G \in \mathbb{R}^{n \times m}$ is a lower triangular matrix.

Another class of low-rank approximation algorithms stem from the Nyström method. The Nyström method was originally designed to solve integral equations (Baker, 1977). Given a kernel matrix K , the Nyström method can be deemed as choosing a subset of m columns (hence rows) $E \in \mathbb{R}^{n \times m}$, and reconstructing the complete kernel matrix by $K \simeq EW^{-1}E'$, where W is the intersection of the selected rows and columns of K . The most popular sampling scheme for Nyström method is random sampling, which leads to fast versions of kernel machines (Williams & Seeger, 2001; Lawrence & Herbrich, 2003) and spectral

clustering (Fowlkes et al., 2004). In (Platt, 2005), several variants of multidimensional scaling are all shown to be related to the Nyström approximation.

There are also a large body of randomized algorithms for low-rank decomposition of arbitrary matrices (Frieze et al., 1998; Achlioptas & McSherry, 2001; Drineas et al., 2003), where the goal is to design column/row sampling probabilities that achieve provable probabilistic bounds. These algorithms are designed for a more general purpose and will not be the focus of this paper. However, we note that one of these randomized algorithms has been recently revised for efficient low-rank approximation of the symmetric Gram matrix (Drineas & Mahoney, 2005). Therefore we will use it as a representative of randomized algorithms in our empirical evaluations. The basic idea of (Drineas & Mahoney, 2005) is to sample the columns of the kernel matrix based on a pre-computed distribution using the norms of the columns. The reconstruction of the kernel matrix is also normalized by the sampling distribution.

In terms of efficiency, greedy approaches usually take $O(m^2n)$ time for sampling, while the random scheme only needs $O(n)$ and is much more efficient. Probabilistic approaches, or randomized algorithms in general, are usually more expensive in that the sampling distributions have to be computed based on the original matrix, which require at least $O(n^2)$. In terms of memory, note that the matrices (E and W) needed in the Nyström method with random sampling can be simply computed on demand. This greatly reduces the memory requirement for very large-scale problems. In contrast, the intermediate matrices for greedy approaches have to be incrementally updated and stored.

Although the Nyström method possesses desirable scaling properties and has been applied with success in various machine learning problems, analysis on its key step of choosing the landmark set is relatively limited. In (Drineas & Mahoney, 2005), a probabilistic error bound is provided on the Nyström low-rank approximation. However, the error bound only applies to the specially designed sampling scheme, which needs to compute the norms of all the rows/columns of the kernel matrix and is hence quite expensive. In (Zhang & Kwok, 2006), a block quantization scheme is proposed for fast spectral embedding. The kernel eigen-system is approximated by first computing a block-wise constant kernel matrix and then extrapolating its eigenvectors through the weighted Nyström extension. However, the error analysis is only on the block-quantization step, and how the Nyström method affects the approximation quality in general remains unclear. Thus, the motivation of this paper is to provide a more concrete analysis on how the sampling scheme (or the choice of the landmark points) in general influences the Nyström low-rank approximation, and to improve the sampling strategy

while still preserving its computational efficiency.

Our key finding is that the Nyström low-rank approximation depends crucially on the quantization error induced by encoding the sample set with the landmark points. This suggests that, instead of applying the greedy or probabilistic sampling, the landmark points can be simply chosen as the k -means cluster centers, which finds a local minimum of the quantization error. To the best of our knowledge, the k -means has not been applied in the Nyström low-rank approximation. The complexity of k -means is only linear in the sample size and dimension and, as our analysis expected, it demonstrates very encouraging performance that is consistently better than all known variants of Nyström. We also compare it with the greedy approach of incomplete Cholesky decomposition and again obtain positive results.

The rest of the paper is organized as follows. In Section 2, we give a brief introduction of the Nyström method. In Section 3, we present an error analysis on how the Nyström low-rank approximation is affected by the chosen landmark points, and propose the k -means algorithm for the sampling step. In Section 4, we compare our approach with a number of state-of-the-art low-rank decomposition techniques (including both greedy and probabilistic sampling approaches). The last section gives concluding remarks.

2. Nyström Method

The Nyström method is originated from the numerical treatment of integral equations of the form

$$\int p(y)k(x, y)\phi_i(y)dy = \lambda_i\phi_i(x), \quad (1)$$

where $p(\cdot)$ is the probability density function, k is a positive definite kernel function, and $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ and ϕ_1, ϕ_2, \dots are the eigenvalues and eigenfunctions of the integral equation, respectively. Given a set of i.i.d. samples $\{x_1, x_2, \dots, x_q\}$ drawn from $p(\cdot)$, the basic idea is to approximate the integral in (1) by the empirical average:

$$\frac{1}{q} \sum_{j=1}^q k(x, x_j)\phi_i(x_j) \simeq \lambda_i\phi_i(x). \quad (2)$$

Choosing x in (2) from $\{x_1, x_2, \dots, x_q\}$ leads to a standard eigenvalue decomposition $K^{(q)}U^{(q)} = U^{(q)}\Lambda^{(q)}$, where $K_{ij}^{(q)} = k(x_i, x_j)$ for $i, j = 1, 2, \dots, q$, $U^{(q)} \in \mathbb{R}^{q \times q}$ has orthonormal columns and $\Lambda^{(q)} \in \mathbb{R}^{q \times q}$ is a diagonal matrix. The eigenfunctions ϕ_i 's and eigenvalues λ_i 's in (1) can be approximated by $U^{(q)}$ and $\Lambda^{(q)}$, as (Williams & Seeger, 2001):

$$\phi_i(x_j) \simeq \sqrt{q}U_{ji}^{(q)}, \quad \lambda_i \simeq \lambda_i^{(q)}/q. \quad (3)$$

This means, the Nyström method using different subset sizes q 's are all approximations to λ_i and ϕ_i in the inte-

gral equation (1). As a result, the Nyström method using a small q can also be deemed as approximating the Nyström method using a large q . Suppose the sample set $\mathcal{X} = \{x_i\}_{i=1}^n$, with the corresponding $n \times n$ kernel matrix K . Then the Nyström method that randomly chooses a subset $\mathcal{Z} = \{z_i\}_{i=1}^m$ of m landmark points will approximate the eigen-system of the full kernel matrix $K\Phi_K = \Phi_K\Lambda_K$ by (Williams & Seeger, 2001)

$$\Phi_K \simeq \sqrt{\frac{m}{n}} E\Phi_Z\Lambda_Z^{-1}, \quad \Lambda_K \simeq \frac{n}{m}\Lambda_Z. \quad (4)$$

Here, $E \in \mathbb{R}^{n \times m}$ with $E_{ij} = k(x_i, z_j)$, and $\Phi_Z, \Lambda_Z \in \mathbb{R}^{m \times m}$ contain the eigenvectors and eigenvalues of $W \in \mathbb{R}^{m \times m}$ where $W_{ij} = k(z_i, z_j)$. Using the approximations in (4), K can be reconstructed as

$$\begin{aligned} K &\simeq \left(\sqrt{\frac{m}{n}} E\Phi_Z\Lambda_Z^{-1} \right) \left(\frac{n}{m}\Lambda_Z \right) \left(\sqrt{\frac{m}{n}} E\Phi_Z\Lambda_Z^{-1} \right)' \\ &= EW^{-1}E'. \end{aligned} \quad (5)$$

Equation (5) is the basis for Nyström low-rank approximation of the kernel matrix (Williams & Seeger, 2001; Fowlkes et al., 2004).

3. Error Analysis of the Nyström Method

In this section we analyze how the Nyström approximation error depends on the choice of landmark points. We first provide an important observation (Section 3.1), and then derive the error bound in more general settings based on a “clustered” data model (Section 3.2-3.4). The error bound gives important insights on the design of efficient sampling schemes for accurate low-rank approximation (Section 3.5).

3.1. Observation

Proposition 1. *Given the data set $\mathcal{X} = \{x_i\}_{i=1}^n$, and the landmark point set $\mathcal{Z} = \{z_j\}_{j=1}^m$. Then the Nyström reconstruction of the kernel entry $K(x_i, x_j)$ will be exact if there exist two landmark points such that $z_p = x_i$, and $z_q = x_j$.*

Proof. Let $K_{x_k, \mathcal{Z}} \in \mathbb{R}^{1 \times m}$ be the similarity between x_k and the landmark points \mathcal{Z} . Then the Nyström reconstruction of the kernel entry will be $K_{x_i, \mathcal{Z}}W^{-1}K'_{x_j, \mathcal{Z}}$, where $W \in \mathbb{R}^{m \times m}$ is the kernel matrix defined on the landmark set \mathcal{Z} . Let $W^{(k)}$ be the k th row of W , then we have $K_{x_i, \mathcal{Z}} = W^{(p)}$ and $K'_{x_j, \mathcal{Z}} = W^{(q)}$ since $x_i = z_p$, and $x_j = z_q$. As a result, the reconstructed entry will be $W^{(p)}W^{-1}(W^{(q)})' = W_{pq} = K(z_p, z_q) = K(x_i, x_j)$. \square

Proposition 1 indicates that the landmark points should be chosen to overlap sufficiently with the original data. However, it is often impossible to use a small landmark set to represent every sample point accurately.

3.2. Approximation Error of Sub-Kernel Matrix

In this section we apply a “clustered” data model to analyze the quality of Nyström low rank approximation. Here, the data clusters can be naturally obtained by assigning each sample to the closest landmark point. As will be seen, this model allows us to derive an explicit error bound for the Nyström approximation.

Again, suppose that the landmark set is $\mathcal{Z} = \{z_i\}_{i=1}^m$, and the whole sample set \mathcal{X} is partitioned into m disjoint clusters \mathcal{S}_k 's. Let $c(i)$ be the function that maps each sample $x_i \in \mathcal{X}$ to the closest landmark point $z_{c(i)} \in \mathcal{Z}$, i.e., $c(i) = \arg \min_{j=1,2,\dots,m} \|x_i - z_j\|$. Our goal is to study the approximation error in (5):

$$\mathcal{E} = \|K - EW^{-1}E'\|_F, \quad (6)$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm.

First, we consider the simpler notion of *partial approximation error* defined as follows.

Definition 1. *Suppose each cluster has T samples¹. Repeat the following sampling process T times: at each time t , pick one sample from each cluster, and denote the set of samples chosen at time t as $\mathcal{X}_{\mathcal{I}_t}$. Then $\mathcal{X} = \{\mathcal{X}_{\mathcal{I}_1} \cup \mathcal{X}_{\mathcal{I}_2} \cup \dots \cup \mathcal{X}_{\mathcal{I}_T}\}$, and the whole kernel matrix will be correspondingly decomposed into T^2 blocks, each of size $m \times m$. Let $K_{\mathcal{I}_i, \mathcal{I}_j}$, and $E_{\mathcal{I}_i, \mathcal{Z}}$ be the $m \times m$ similarity matrices defined on $(\mathcal{X}_{\mathcal{I}_i}, \mathcal{X}_{\mathcal{I}_i})$ and $(\mathcal{X}_{\mathcal{I}_i}, \mathcal{Z})$, respectively, and $W \in \mathbb{R}^{m \times m}$ the kernel matrix defined on \mathcal{Z} . The partial approximation error is the difference between $K_{\mathcal{I}_i, \mathcal{I}_j}$ and its Nyström approximation under the Frobenius norm*

$$\mathcal{E}_{\mathcal{I}_i, \mathcal{I}_j} = \|K_{\mathcal{I}_i, \mathcal{I}_j} - E_{\mathcal{I}_i, \mathcal{Z}}W^{-1}E'_{\mathcal{I}_j, \mathcal{Z}}\|_F. \quad (7)$$

We assume the kernel k satisfies the following property:

$$(k(a, b) - k(c, d))^2 \leq C_{\mathcal{X}}^k (\|a - c\|^2 + \|b - d\|^2), \quad \forall a, b, c, d \quad (8)$$

where $C_{\mathcal{X}}^k$ is a constant depending on k and the sample set \mathcal{X} . The validity of this assumption on a number of commonly used kernels will be proved in Section 3.4.

Proposition 2. *For kernel k satisfying property (8), the partial approximation error $\mathcal{E}_{\mathcal{I}_i, \mathcal{I}_j}$ is bounded by*

$$\begin{aligned} \mathcal{E}_{\mathcal{I}_i, \mathcal{I}_j} &\leq \sqrt{2mC_{\mathcal{X}}^k(e_{\mathcal{I}_i} + e_{\mathcal{I}_j})} + \sqrt{mC_{\mathcal{X}}^k e_{\mathcal{I}_i}} \\ &\quad + \sqrt{mC_{\mathcal{X}}^k e_{\mathcal{I}_j}} + mC_{\mathcal{X}}^k \sqrt{e_{\mathcal{I}_i} e_{\mathcal{I}_j}} \|W^{-1}\|_F. \end{aligned} \quad (9)$$

where $e_{\mathcal{I}_i}$ is the quantization error induced by coding each sample in $\mathcal{X}_{\mathcal{I}_i}$ by the closest landmark point in \mathcal{Z} , i.e.,

$$e_{\mathcal{I}_i} = \sum_{x_i \in \mathcal{X}_{\mathcal{I}_i}} \|x_i - z_{c(i)}\|^2. \quad (10)$$

¹If cluster sizes differ, add “virtual samples” to each cluster such that all the clusters have the same size (which is equal to $T = \max_k |\mathcal{S}_k|$). The virtual samples added to cluster \mathcal{S}_k are chosen as the landmark point z_k for that cluster, so they will not induce extra quantization errors but will loosen the bound.

Proof. We will first define the following matrices

$$\begin{aligned} A_{\mathcal{I}_i, \mathcal{I}_j} &= K_{\mathcal{I}_i, \mathcal{I}_j} - W; B_{\mathcal{I}_i, \mathcal{Z}} = E_{\mathcal{I}_i, \mathcal{Z}} - W; \\ C_{\mathcal{I}_j, \mathcal{Z}} &= E_{\mathcal{I}_j, \mathcal{Z}} - W, \end{aligned} \quad (11)$$

and then show that they have bounded Frobenius norms. Without loss of generality, we specify the indices as follows: $K_{\mathcal{I}_i, \mathcal{I}_j}(p, q) = k(x_{\mathcal{I}_i(p)}, x_{\mathcal{I}_j(q)})$; $E_{\mathcal{I}_i, \mathcal{Z}}(p, q) = k(x_{\mathcal{I}_i(p)}, z_q)$; $E_{\mathcal{I}_j, \mathcal{Z}}(p, q) = k(x_{\mathcal{I}_j(p)}, z_q)$; and $W(p, q) = k(z_p, z_q)$. With property (8), we have

$$\begin{aligned} \|A_{\mathcal{I}_i, \mathcal{I}_j}\|_F^2 &= \sum_{p, q=1}^m (k(x_{\mathcal{I}_i(p)}, x_{\mathcal{I}_j(q)}) - k(z_p, z_q))^2 \\ &\leq C_{\mathcal{X}}^k \sum_{p, q=1}^m (\|x_{\mathcal{I}_i(p)} - z_p\|^2 + \|x_{\mathcal{I}_j(q)} - z_q\|^2) \\ &= mC_{\mathcal{X}}^k \left(\sum_{p=1}^m \|x_{\mathcal{I}_i(p)} - z_p\|^2 + \sum_{q=1}^m \|x_{\mathcal{I}_j(q)} - z_q\|^2 \right) \\ &= 2mC_{\mathcal{X}}^k (e_{\mathcal{I}_i} + e_{\mathcal{I}_j}), \end{aligned}$$

where $e_{\mathcal{I}_i}$ is the same as that in (10) since $c(\mathcal{I}(q)) = q$.

For matrix $B_{\mathcal{I}_i, \mathcal{Z}}$, we have

$$\begin{aligned} \|B_{\mathcal{I}_i, \mathcal{Z}}\|_F^2 &= \sum_{p, q} (k(x_{\mathcal{I}_i(p)}, z_q) - k(z_p, z_q))^2 \\ &\leq mC_{\mathcal{X}}^k \sum_{p=1}^m \|x_{\mathcal{I}_i(p)} - z_p\|^2 = mC_{\mathcal{X}}^k e_{\mathcal{I}_i}, \end{aligned}$$

and similarly for matrix $C_{\mathcal{I}_j, \mathcal{Z}}$ $\|C_{\mathcal{I}_j, \mathcal{Z}}\|_F^2 \leq mC_{\mathcal{X}}^k e_{\mathcal{I}_j}$.

Note that the partial approximation error $\mathcal{E}_{\mathcal{I}_i, \mathcal{I}_j}$ (7) can be re-written as follows using (11).

$$\begin{aligned} \|\mathcal{E}_{\mathcal{I}_i, \mathcal{I}_j}\|_F &= \|W + A_{\mathcal{I}_i, \mathcal{I}_j} - (W + B_{\mathcal{I}_i, \mathcal{Z}})W^{-1}(W + C_{\mathcal{I}_j, \mathcal{Z}})'\|_F \\ &= \|W + A_{\mathcal{I}_i, \mathcal{I}_j} - W' - C_{\mathcal{I}_j, \mathcal{Z}}' - B_{\mathcal{I}_i, \mathcal{Z}} - B_{\mathcal{I}_i, \mathcal{Z}}W^{-1}C_{\mathcal{I}_j, \mathcal{Z}}'\|_F \\ &\leq \|A_{\mathcal{I}_i, \mathcal{I}_j}\|_F + \|B_{\mathcal{I}_i, \mathcal{Z}}\|_F + \|C_{\mathcal{I}_j, \mathcal{Z}}\|_F + \|B_{\mathcal{I}_i, \mathcal{Z}}\|_F \|C_{\mathcal{I}_j, \mathcal{Z}}\|_F \|W^{-1}\|_F \end{aligned}$$

Using the bounds on $\|A_{\mathcal{I}_i, \mathcal{I}_j}\|_F$, $\|B_{\mathcal{I}_i, \mathcal{Z}}\|_F$, $\|C_{\mathcal{I}_j, \mathcal{Z}}\|_F$ together with the definition in (11), we have Proposition 2 \square

3.3. Approximation Error of Complete Kernel Matrix

With the estimated partial approximation error, we can now obtain a bound on the complete error for Nyström approximation (6). The basic idea is to sum up the partial errors $\mathcal{E}_{\mathcal{I}_i, \mathcal{I}_j}$ over all $i, j = 1, 2, \dots, T$.

Proposition 3. *The error of the Nyström approximation (6) is bounded by*

$$\mathcal{E} \leq 4T\sqrt{mC_{\mathcal{X}}^k eT} + mC_{\mathcal{X}}^k Te\|W^{-1}\|_F \quad (12)$$

where $T = \max_k |\mathcal{S}_k|$, and $e = \sum_{i=1}^n \|x_i - z_{c(i)}\|^2$ is the total quantization error of coding each sample $x_i \in \mathcal{X}$ with the closest landmark point $z_j \in \mathcal{Z}$.

Proof. Here we sum up the terms in (9) separately.

$$\begin{aligned} \sum_{i, j=1}^T \sqrt{2mC_{\mathcal{X}}^k (e_{\mathcal{I}_i} + e_{\mathcal{I}_j})} &= \sqrt{2mC_{\mathcal{X}}^k} \sum_{i=1}^T \left(\sum_{j=1}^T \sqrt{e_{\mathcal{I}_i} + e_{\mathcal{I}_j}} \right) \\ &\leq \sqrt{2mC_{\mathcal{X}}^k} \sum_{i=1}^T \left(\sqrt{T} \sqrt{Te_{\mathcal{I}_i} + \sum_{j=1}^T e_{\mathcal{I}_j}} \right) \leq 2T\sqrt{mC_{\mathcal{X}}^k Te} \end{aligned}$$

where $e = \sum_{j=1}^T e_{\mathcal{I}_j} = \sum_{x_i \in \mathcal{X}} \|x_i - z_{c(i)}\|^2$ is the same as defined in proposition 3. Similarly, the second term (and the third term) in (9) can be summarized as

$$\sum_{i, j=1}^T \sqrt{mC_{\mathcal{X}}^k e_{\mathcal{I}_i}} = \sqrt{mC_{\mathcal{X}}^k} \sum_{j=1}^T \left(\sum_{i=1}^T \sqrt{e_{\mathcal{I}_i}} \right) \leq T\sqrt{mC_{\mathcal{X}}^k eT}$$

The last term in (9) can be summarized as

$$\begin{aligned} \sum_{i, j=1}^T mC_{\mathcal{X}}^k \sqrt{e_{\mathcal{I}_i} e_{\mathcal{I}_j}} \|W^{-1}\|_F &= mC_{\mathcal{X}}^k \|W^{-1}\|_F \left(\sum_{i=1}^T \sqrt{e_{\mathcal{I}_i}} \right)^2 \\ &\leq mC_{\mathcal{X}}^k \|W^{-1}\|_F Te \end{aligned}$$

By combining all these terms, we arrive at Proposition 3. \square

3.4. $C_{\mathcal{X}}^k$ Under Different Kernels

In this section, we show that many commonly used kernel functions satisfy the property in (8). Consider the stationary kernel $k(x, y) = \kappa\left(\left\|\frac{x-y}{\sigma}\right\|\right)$, including the Gaussian kernel $\kappa(\alpha) = \exp(-\alpha^2)$, Laplacian kernel $\kappa(\alpha) = \exp(-\alpha)$, and inverse distance kernel $\kappa(\alpha) = (\alpha + \epsilon)^{-1}$. By using the mean value theorem and triangular inequality, we have, for any $a, b, c, d \in \mathbb{R}^d$,

$$\begin{aligned} (k(a, b) - k(c, d))^2 &= (\kappa(\|a-b\|/\sigma) - \kappa(\|c-d\|/\sigma))^2 \\ &= [\kappa'(\xi)/\sigma]^2 (\|a-b\| - \|c-d\|)^2. \end{aligned}$$

Let $v_1 = a - c$ and $v_2 = b - d$. Note that we have $\|c-d\| \leq \|a-b\| + \|v_1\| + \|v_2\|$ and similarly $\|a-b\| \leq \|c-d\| + \|v_1\| + \|v_2\|$. So $\|a-b\| - \|c-d\|$ is always bounded by

$$\begin{aligned} (\|a-b\| - \|c-d\|)^2 &\leq (\|a-c\| + \|b-d\|)^2 \\ &\leq 2(\|a-c\|^2 + \|b-d\|^2). \end{aligned}$$

So $C_{\mathcal{X}}^k$ can be chosen as $\max[2\kappa'(\xi)/\sigma]^2$ which is often bounded ($C_{\mathcal{X}}^k$ is $\frac{1}{2\sigma^2}$ for the Gaussian, $\frac{1}{\sigma^2}$ for the Laplacian, and $\frac{1}{\sigma^2\epsilon^d}$ for the inverse distance). Similarly, for polynomial kernels of the form $k(x, y) = (\langle x, y \rangle + \epsilon)^d$,

$$\begin{aligned} (k(a, b) - k(c, d))^2 &= ((a'b + \epsilon)^d - (c'd + \epsilon)^d)^2 \\ &= (p'(\xi)(a'b - c'd))^2 = (p'(\xi)((a-c)'b + (b-d)'c))^2 \\ &\leq [2p'(\xi)]^2 (\|(a-c)'b\|^2 + \|(b-d)'c\|^2) \\ &\leq [2p'(\xi)R]^2 (\|a-c\|^2 + \|b-d\|^2), \end{aligned}$$

where R is the larger one of the two quantities: the maximum pairwise distance between samples, and maximum distance between samples and the origin point; and $p(z) = z^d$. So $C_{\mathcal{X}}^k$ can be chosen as $\max[\kappa'(\xi)R]^2 = d^2 R^d$.

3.5. Sampling Procedure

The error bound in Proposition 3 provides important insights on how to choose the landmark points in the Nyström method. As can be seen, consistently, that for a number of commonly used kernels, the most important factor that influences the approximation quality is e , the error of quantizing each of the samples in \mathcal{X} with the closest landmark in \mathcal{Z} . If this quantization error is zero, the Nyström low-rank approximation of the kernel matrix will also be exact. This agrees well with the ideal case discussed in Section 3.1.

Motivated by this observation and the fact that k -means clustering can find a local minimum of the quantization error (Gersho & Gray, 1992), we propose to use the centers obtained from the k -means as the landmark points. Here, k is the desired number of landmark points in \mathcal{Z} . The larger the k , the more accurate the approximation though at the cost of higher computations. Despite its simplicity, the k -means procedure can greatly improve the approximation quality compared to other sampling schemes, as will be demonstrated empirically in Section 4. Recent advances in speeding up the k -means algorithms (Elkan, 2003; Kanningo et al., 2001) also make this k -means-based sampling strategy particularly suitable for large-scale problems.

4. Experiments

This section presents empirical evaluations of the various low-rank approximation schemes. First, we discuss how the low rank approximation fits into different applications. One is to solve linear systems of the form $(K + \sigma I)x = a$, where K is the kernel matrix, $\sigma \geq 0$ is a regularization parameter and I is the $n \times n$ identity matrix. Given the low-rank approximation $K \simeq GG'$, the following holds (Williams & Seeger, 2001) by the Woodbury formula

$$(K + \sigma I)^{-1} \simeq \frac{1}{\sigma} (I - G(\sigma I + G'G)^{-1}G'), \quad (13)$$

which only needs $O(m^2n)$ time and $O(mn)$ memory. Therefore, it can be used in speeding up the Gaussian processes (Williams & Seeger, 2001) and least-squares SVM (LS-SVM) (Suykens & Vandewalle, 1999).

The second application is to reconstruct the eigen-system of a matrix approximated by its low-rank decomposition.

Proposition 4. *Given the low-rank approximation $K \approx GG'$, where $G \in \mathbb{R}^{n \times m}$ and $m \ll n$, the top m eigenvectors U of K can be obtained as $U \approx GV\Lambda^{-1/2}$ in $O(m^2n)$ time, where $V, \Lambda \in \mathbb{R}^{m \times m}$ are from the eigenvalue decom-*

position of the $m \times m$ matrix $S = G'G = V\Lambda V'$.

Proof can be found in (Fowlkes et al., 2004). Therefore low-rank approximation is useful for algorithms that rely on eigenvectors of the kernel matrix, such as kernel PCA (Schölkopf et al., 1998), Laplacian eigenmap (Belkin & Niyogi, 2002) and normalized cut.

Note that the Nyström method, when designed originally to solve integral equations, did not provide orthogonal approximations to the kernel eigenfunctions. Thanks to the matrix completion view (5) (Fowlkes et al., 2004; Williams & Seeger, 2001), the Nyström method can be utilized for obtaining orthogonal eigenvectors (Proposition 4), though the time complexity increases from the simple Nyström extension (4) of $O(mn)$ to $O(m^2n)$. In the experiments we focus on the orthogonalized eigenvector approximation.

Table 1. Complexities of basis selection for the different methods.

	Ours	Nyström	Drineas	ICD
time	$O(mn)$	$O(n)$	$O(n^2)$	$O(m^2n)$
space	$O(mn)$	$O(mn)$	$O(mn)$	$O(mn)$

We compare altogether five low-rank approximation algorithms, including: 1. incomplete Cholesky decomposition (ICD)²; 2. Nyström method (with random sampling); 3. the method in (Drineas & Mahoney, 2005); 4. our method (for simplicity, the maximum number of k -means iterations is restricted to 10); 5. SVD. Note that SVD (or eigenvalue decomposition in our context) provides the best low-rank approximation in terms of both the Frobenius norm and spectral norm (Golub & Van Loan, 1996). The complexities of basis selection (i.e., choosing E and W in Nyström, or sampling the columns in (Drineas & Mahoney, 2005) and ICD) in the different algorithms are listed in Table 1. Evaluations are performed in the contexts of kernel matrix approximation (Section 4.1), kernel PCA (Section 4.2), and LS-SVM classification (Section 4.3). We use core(TM)-dual PC with 2.13GHz CPU and the codes are in matlab.

4.1. Approximating the Kernel Matrix

We first examine the performance of the low-rank approximation schemes by measuring their approximation errors (in terms of the Frobenius norm) on the kernel matrix. We choose a number of benchmark data sets from the LIB-SVM archive³, summarized in Table 2. Note that our approximation error bound in Proposition 3 applies to most kernel functions (Section 3.4), and preliminary experimental results with these kernels have shown the superiority of our sampling scheme compared with other low-rank approximation methods. However, due to lack of space, we

²<http://www.di.ens.fr/~fbach/kernel-ica/index.htm>

³<http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>

Table 2. A summary of data sets.

data	german	splice	adult1a	dna
size	1000	1000	1605	2000
dimension	24	60	123	180
data	segment	w1a	svmgd1a	satimage
size	2310	2477	3089	4435
dimension	19	300	4	36

will only report results for the Gaussian kernel $K(x, y) = \exp(-\|x - y\|^2/\gamma)$. Here, γ is chosen as the average squared distance between data points and the mean of each data set. We gradually increase the subset size m from 1% to 10% of the data size. To reduce statistical variability, results of methods 2, 3, and 4 are based on averages over 20 repetitions.

The approximation errors are plotted in Figure 1. As can be seen, our algorithm is only inferior to SVD on most data sets. Moreover, though the method in (Drineas & Mahoney, 2005) involves a more complicated probabilistic sampling scheme, its performance is only comparable or sometimes even worse than the Nyström method with simple random sampling. Similar observations have also been reported in the context of SVD (Drineas et al., 2003). ICD seems to be inferior on several data sets. However, for data sets whose kernel spectra decay rapidly to zero⁴ (such as the `segment`, `svmgd1a` and `satimage`), ICD can also quickly attain performance comparable to others.

We also examine empirically the relationship between \mathcal{E} and e under different sampling schemes. Figure 2 reports the results on the `german` data, where $m = 100$ and each sampling scheme is repeated 100 times. As can be seen, there is a strong, positive correlation between \mathcal{E} and e . This is observed on most data and agrees with our error analysis.

4.2. Kernel PCA

In kernel PCA, the key step is to obtain eigenvectors of the centered kernel matrix HKH , where $H = I - \frac{1}{n}11' \in \mathbb{R}^{n \times n}$. Following Proposition 2 of (Ouibet & Bengio, 2005), with the low-rank decomposition $K \simeq GG'$, the centered kernel matrix can be written as $(HG)(HG)'$ or $(G - \bar{G})(G - \bar{G})'$, where $\bar{G} \in \mathbb{R}^{n \times m}$ and all its rows equal to the mean of rows in G . Hence the top m eigenvectors can be obtained in $O(m^2n)$ time according to Proposition 4.

We evaluate the low rank approximation schemes by the embedding onto the top 3 principal directions. We align the approximate embeddings (\tilde{U}) with the standard KPCA embedding (U) through a linear transform, and report the

⁴Note that the (squared) rank- m approximation error of SVD is $\sum_{i=m+1}^n \sigma_i^2$, where σ_i 's are the singular values of K sorted in descending order (Golub & Van Loan, 1996). Therefore, if SVD's error in Figure 1 drops rapidly, so does the spectrum of K .

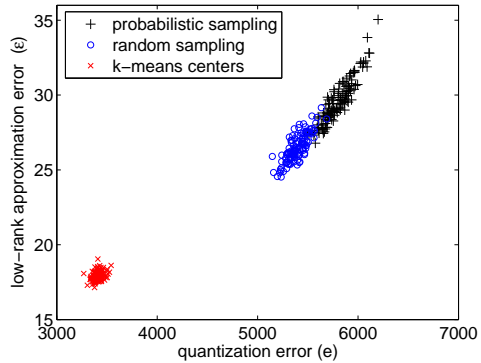


Figure 2. Low-rank approximation error versus quantization error for different sampling schemes.

minimum misalignment error: $\min_{A \in \mathbb{R}^{3 \times 3}} \|U - \tilde{U}A\|_F$. The parameter setting is the same as in Section 4.1, except that we fix $m = 0.05n$ for all the low-rank decomposition algorithms. Again, results of methods 2, 3, 4 in Table 3 are averaged over 20 repetitions. As we can see, our algorithm is the best on most data sets, next comes the standard Nyström and the method by (Drineas & Mahoney, 2005). The time consumptions of all low-rank approximation schemes are significantly lower than SVD.

4.3. Least Squares SVM

Given the kernel matrix K , the training labels $y \in \{\pm 1\}^{n \times 1}$, and the regularization parameter $C > 0$, the LS-SVM classifier $f(x) = \sum_{i=1}^n \alpha_i \phi(x, x_i) + b$ is solved by $b = y'M^{-1}1/y'M^{-1}y$, and $\alpha = M^{-1}(1 - by)$, where 1 is a vector of all ones, and $M = Y(K + I/C)Y$ and $Y = \text{diag}(y)$. Note that $M^{-1} = Y(K + I/C)^{-1}Y$ can be computed efficiently using (13).

We evaluate different low-rank approximation schemes in LS-SVM, using some difficult pairs of the USPS digits⁵. We use Gaussian kernel $\exp(-\|x - y\|^2/\gamma)$ and $C = 0.5$. Table 4 reports the classification performance of the standard LS-SVM, and those with different low-rank approximation schemes, at $m = 0.05n$ and $0.1n$. Again, methods 2, 3, 4 are repeated 20 times. For $m = 0.05n$, our approach is significantly better than methods 1,2,3 with a confidence level that is at least 99.5%. For $m = 0.1n$, ours is also better with a confidence level that is at least 97.5% on the first 7 pairs. For the last 4 pairs, the differences between our approach and methods 1,2,3 are not statistically significant. Note, however, that the testing errors obtained by the various approximation algorithms on these 4 pairs are all close to those of the exact LS-SVM, i.e., all approximation algorithms have reached their possibly best performance.

⁵<http://ftp.kyb.tuebingen.mpg.de/pub/bs/data/>

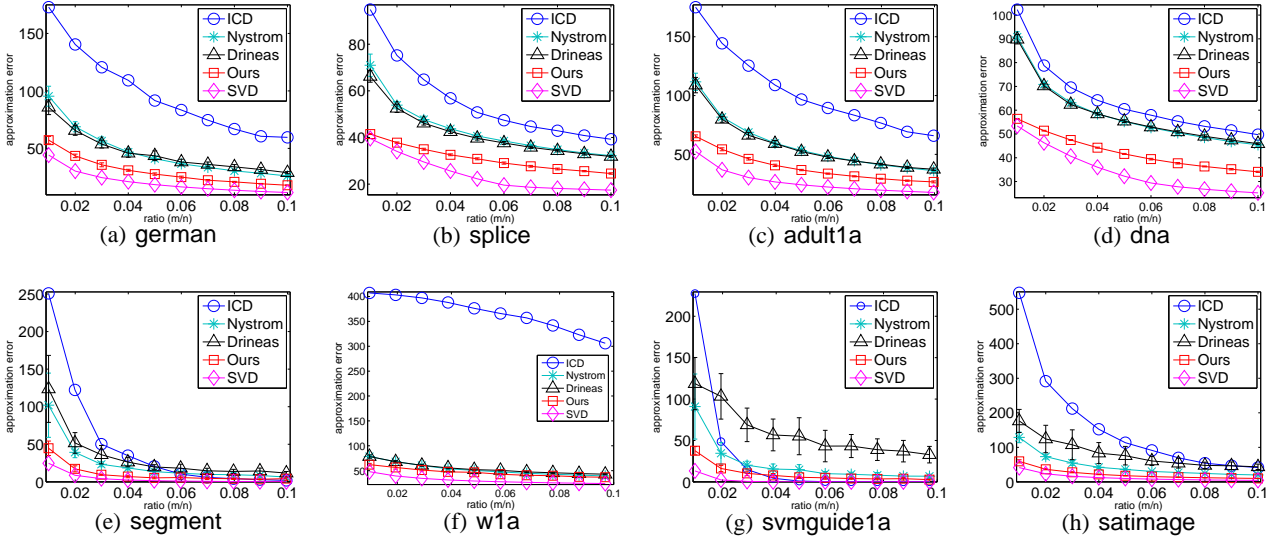


Figure 1. Approximation errors (in terms of the Frobenius norm) on the kernel matrix by different low-rank approximation schemes.

Table 3. Approximation errors and CPU time consumed for the different low-rank approximation schemes in the context of kernel PCA. Due to the lack of space, we do not show the standard deviation of the CPU time.

data	approximation error				CPU time (seconds)				
	Ours	Nyström	Drineas	ICD	SVD	Ours	Nyström	Drineas	ICD
german	$(4.40 \pm 0.58) \times 10^{-2}$	$(2.64 \pm 0.58) \times 10^{-1}$	$(2.71 \pm 0.34) \times 10^{-1}$	5.11×10^{-1}	27.6	0.8	0.03	0.3	0.09
splice	$(3.44 \pm 0.43) \times 10^{-1}$	$(1.06 \pm 0.11) \times 10^0$	$(1.07 \pm 0.11) \times 10^0$	1.27×10^0	24.2	0.9	0.05	0.6	0.1
adult1a	$(4.41 \pm 0.49) \times 10^{-2}$	$(2.86 \pm 0.42) \times 10^{-1}$	$(2.84 \pm 0.66) \times 10^{-1}$	6.19×10^{-1}	134.8	3.0	0.2	4.0	0.7
dna	$(1.88 \pm 0.21) \times 10^{-1}$	$(1.09 \pm 0.08) \times 10^0$	$(1.01 \pm 0.14) \times 10^0$	1.17×10^0	197.0	6.6	0.5	10.6	1.5
segment	$(7.87 \pm 4.43) \times 10^{-4}$	$(8.37 \pm 4.08) \times 10^{-3}$	$(1.84 \pm 0.99) \times 10^{-2}$	2.37×10^{-2}	322.8	4.2	0.3	1.8	1.0
w1a	$(1.55 \pm 0.78) \times 10^{-1}$	$(2.81 \pm 0.62) \times 10^{-1}$	$(6.05 \pm 3.39) \times 10^{-1}$	1.11×10^0	394.0	12.8	1.8	35.3	3.6
svmguide1a	$(5.16 \pm 2.12) \times 10^{-4}$	$(3.71 \pm 2.26) \times 10^{-3}$	$(2.78 \pm 1.60) \times 10^{-2}$	5.07×10^{-4}	650.4	6.7	0.5	2.4	2.3
satimage	$(5.20 \pm 0.97) \times 10^{-4}$	$(6.19 \pm 0.28) \times 10^{-3}$	$(6.80 \pm 1.01) \times 10^{-2}$	2.47×10^{-2}	2762.8	16.1	1.5	15.9	7.5

5. Conclusion

The Nyström method is a useful technique for low-rank approximation. However, analysis on its key step of choosing the landmark points and especially that in terms of approximation quality is still limited. In this paper, we draw an intuitive but important connection between the Nyström approximation quality and the encoding capacities of landmark points. Our analysis suggests the k -means as a natural sampling scheme. Despite its simplicity, the k -means-based sampling gives encouraging performance when empirically compared with state-of-the-art low-rank approximation techniques. One future direction is to utilize label/side information for task-specific decomposition, where one excellent example is (Bach & Jordan, 2005) in the context of incomplete Cholesky decomposition.

Acknowledgments

This research has been partially supported by the Research Grants Council of the Hong Kong Special Administrative

Region under grant 614907.

References

- Achlioptas, D., & McSherry, F. (2001). Fast computation of low rank matrix approximations. *Proceedings of the 23th Annual ACM Symposium on Theory of Computing* (pp. 611 – 618).
- Bach, F., & Jordan, M. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, 3, 1–48.
- Bach, F., & Jordan, M. (2005). Predictive low-rank decomposition for kernel methods. *Proceedings of the 22th International Conference on Machine Learning* (pp. 33 – 40).
- Baker, C. (1977). *The numerical treatment of integral equations*. Oxford: Clarendon Press.
- Belkin, M., & Niyogi, P. (2002). Laplacian eigenmaps and

Table 4. Testing errors (in %) on different pairs of the USPS digits.

data	$m = 0.05n$						$m = 0.1n$				
	LS-SVM	SVD	Ours	Nyström	Drineas	ICD	SVD	Ours	Nyström	Drineas	ICD
1-7	0.97	0.97	0.77±0.09	1.56±0.51	2.69±1.64	4.62	0.97	0.81±0.11	0.94±0.18	1.54±0.35	1.22
2-3	2.47	2.47	2.91±0.55	4.48±1.51	4.54±1.16	4.12	2.47	2.72±0.22	2.77±0.59	2.97±0.59	4.94
2-5	1.67	0.83	1.91±0.62	2.82±1.10	2.98±0.73	5.30	1.67	1.29±0.31	1.57±0.46	1.58±0.51	2.23
3-8	2.71	3.31	3.93±0.84	5.64±1.45	5.52±1.21	6.02	2.71	2.70±0.23	3.78±0.53	3.97±0.55	4.21
5-8	2.76	2.14	3.38±0.65	4.34±1.52	4.28±1.38	4.29	2.14	2.66±0.34	2.82±0.53	3.12±0.58	4.60
6-8	0.59	0.59	1.04±0.35	2.69±1.34	2.24±0.98	5.65	0.89	0.61±0.25	1.03±0.45	1.18±0.44	5.05
8-9	1.45	1.16	1.04±0.50	2.79±2.00	2.52±0.85	2.33	1.45	1.13±0.23	1.22±0.51	1.79±0.37	2.91
2-7	1.44	0.86	0.90±0.10	1.47±0.64	2.45±1.25	3.76	0.86	0.96±0.17	0.87±0.09	1.08±0.39	2.61
3-5	4.91	4.91	6.53±1.18	7.39±1.39	7.03±1.25	7.36	3.98	5.25±0.64	5.22±0.77	5.49±0.83	5.52
4-7	2.88	2.59	2.54±0.40	3.28±0.96	3.30±1.03	8.06	2.59	2.60±0.26	2.41±0.36	2.43±0.45	5.47
5-6	1.21	0.60	1.57±0.63	2.79±1.25	2.52±1.33	3.93	1.21	1.27±0.32	1.25±0.41	1.48±0.29	2.12

spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems 14*.

Drineas, P., Drinea, E., & Huggins, P. (2003). An experimental evaluation of a Monte-Carlo algorithm for singular value decomposition. *Proceedings of 8th Panhellenic Conference on Informatics* (pp. 279–296).

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

Elkan, E. (2003). Using the triangular inequality to accelerate k -means. *Proceedings of the 21th International Conference on Machine Learning* (pp. 147 – 153).

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

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

Frieze, A., Kannan, R., & Vempala, S. (1998). Fast Monte-Carlo algorithms for finding low-rank approximations. *Proceedings of the 39th Annual Symposium on Foundations of Computer Science* (pp. 370 – 378).

Gersho, A., & Gray, R. (1992). *Vector quantization and signal compression*. Boston: Kluwer Academic Press.

Golub, G., & Van Loan, C. (1996). *Matrix computations*. Johns Hopkins University Press. 3rd edition.

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

Lawrence, N. Seeger, M., & Herbrich, R. (2003). Fast sparse Gaussian process methods: The informative vector machine. *Advances in Neural Information Processing Systems*. (pp. 625–632). MIT Press.

Ouimet, M., & Bengio, Y. (2005). Greedy spectral embedding. *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics* (pp. 253–260).

Platt, J. C. (2005). Fastmap, MetricMap, and Landmark MDS are all Nyström algorithms. *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics* (pp. 261–268).

Schölkopf, B., Smola, A., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10, 1299–1319.

Smola, A., & Schölkopf, B. (2000). Sparse greedy matrix approximation for machine learning. *Proceedings of the 17th International Conference on Machine Learning* (pp. 911 – 918).

Suykens, J., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9, 293–300.

Williams, C., & Seeger, M. (2000). The effect of the input density distribution on kernel-based classifiers. *Proceedings of the 17th International Conference on Machine Learning* (pp. 1159–1166).

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

Zhang, K., & Kwok, J. (2006). Block-quantized kernel matrix for fast spectral embedding. *Proceedings of the 23rd international conference on Machine learning* (pp. 1097 – 1104).