Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Nyström Approximation for Sparse Kernel Methods: Theoretical Analysis and Empirical Evaluation

Zenglin Xu
School of Computer Science and Engineering, Big Data Research Center
University of Electronic Science and Technology of China
zenglin@gmail.com

Rong Jin
Dept. of Comp. Sci. & Eng.
Michigan State University

Bin Shen
Dept. of Computer Science
Purdue University

Shenghuo Zhu
Alibaba Group

Abstract

Nyström approximation is an effective approach to accelerate the computation of kernel matrices in many kernel methods. In this paper, we consider the Nyström approximation for sparse kernel methods. Instead of relying on the low-rank assumption of the original kernels, which sometimes does not hold in some applications, we take advantage of the restricted eigenvalue condition, which has been proved to be robust for sparse kernel methods. Based on the restricted eigenvalue condition, we have provided not only the approximation bound for the original kernel matrix but also the recovery bound for the sparse solutions of sparse kernel regression. In addition to the theoretical analysis, we also demonstrate the good performance of the Nyström approximation for sparse kernel regression on real world data sets.

Introduction

Kernel methods (Schölkopf and Smola 2002; Xu et al. 2009) have received a lot of attention in recent studies of machine learning. These methods project data into high-dimensional or even infinite-dimensional spaces via kernel mapping functions. Despite the strong generalization ability induced by kernel methods, they usually suffer from the high computation complexity of calculating the kernel matrix (also called Gram matrix). Although low-rank decomposition techniques(e.g., Cholesky Decomposition (Fine and Scheinberg 2002; Bach and Jordan 2005)), and truncating methods(e.g., Kernel Tapering (Shen, Xu, and Allebach 2014; Furrer, Genton, and Nychka 2006)) can accelerate the calculation of the kernel matrix, they still need to compute the kernel matrix.

An effective approach to avoid the computation cost of computing the entire kernel matrix is to approximate the kernel matrix by the Nyström method (Williams and Seeger 2001), which provides low-rank approximation to the kernel matrix by sampling from its columns. The Nyström method has been proven useful in a number of applications, such as image processing (Fowlkes et al. 2004; Wang et al. 2009), which typically involve computations with large dense matrices. Recent research (Zhang, Tsang, and Kwok 2008; Farahat, Ghodsi, and Kamel 2011; Talwalkar and Rostamizadeh 2010; Kumar, Mohri, and Talwalkar 2012; Mackey, Talwalkar, and Jordan 2011; Gittens and Mahoney 2013) on the Nyström method have shown that the approximation error can be theoretically bounded. Jin et al. (2013) further shows that the approximation error bound can be improved from $O(n/\sqrt{m})$ to $O(n/m^{p-1})$ (where n denotes the number of instances and m denotes the number of dimensions) when the eigenvalues of the kernel matrix satisfy a p-power law distribution.

In this paper, we focus on finding the approximation bound of the Nyström method for sparse kernel methods. Although previous studies have demonstrated the good approximation bounds of the Nyström method for kernel methods, most of which are based on the assumption of the low rank of kernels (Jin et al. 2013). While if kernels are not low rank, Nyström approximations can usually lead to suboptimal performances. To alleviate the strong assumption in the seeking of the approximation bounds, we take a more general assumption that the design matrix $\mathbf{K}$ ensuring the restricted isometric property (Koltchinskii 2011). In particular, the new assumption obeys the restricted eigenvalue condition (Koltchinskii 2011; Bickel, Ritov, and Tsybakov 2009), which has been shown to be more general than several other similar assumptions used in sparsity literature (Candes and Tao 2007; Donoho, Elad, and Temlyakov 2006; Zhang and Huang 2008). Based on the restricted eigenvalue condition, we have provided error bounds for kernel approximation and recovery rate in sparse kernel regression. Thus we can accurately recover the sparse solution even with a modest number of random samples. It is important to note that the expected risk of the learning function will be small by exploiting the generalization error bound for data dependent hypothesis space (Shi, Feng, and Zhou 2011). To further evaluate the performance of the Nyström method for sparse kernel regression, we conduct experiments on both synthetic data and real-world data sets. Experimental results have indicated the huge acceleration of the Nyström method on training time while maintaining the same level of prediction errors.

Nyström Approximation for Sparse Kernel Regression

We consider the regression setting. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ be the collection of training examples, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in [-R, R]$. Let $\kappa(\cdot, \cdot)$ be the kernel function, and $\mathcal{H}_\kappa$ be the Reproduced Kernel Hilbert Space endowed with kernel $\kappa(\cdot, \cdot)$. Without loss of generality, we assume $\kappa(\mathbf{x}, \mathbf{x}) \leq 1$ for any $\mathbf{x} \in \mathbb{R}^d$. The objective of kernel regression is to find a $f \in \mathcal{H}_\kappa$ that best fits the data. It usually results in the following optimization problem:

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2n} \|\mathbf{y} - \mathbf{K}\alpha\|_2^2 + \lambda \|\alpha\|_1 \tag{1}$$

where $\mathbf{K} = [\kappa(\mathbf{x}_i, \mathbf{x}_j)]_{n \times n}$, $\mathbf{y} = (y_1, \ldots, y_n)^\top$, and λ is a regularization parameter that can be usually determined by cross validation. The main computational challenge of kernel regression arises from handling the kernel matrix $\mathbf{K}$, which can be expensive to compute when the number of training examples is large. A common approach to address the computational challenge of kernel regression is to approximate kernel matrix $\mathbf{K}$ by a low rank matrix $\widehat{\mathbf{K}}$. Popular algorithms in this category include random Fourier features (Rahimi and Recht 2007) and Nyström method (Williams and Seeger 2001). Analysis in (Drineas and Mahoney 2005) shows that additional error caused by the low rank approximation of $\mathbf{K}$ can be as high as $O(n/\sqrt{m})$, where m is the number of random samples used by low rank matrix approximation. By making additional assumptions, this error can further reduced to $O(n/m)$ (Jin et al. 2013).

In this paper, we focus on sparse kernel regression, where the sparsity can usually lead to good performance. It assumes there exists a sparse solution α_* such that $f_*(\cdot) = \sum_{i=1}^n \alpha_*^i \kappa(\mathbf{x}_i, \cdot)$ yields a small regression error for all the training examples. To be more precisely, let's denote by $J \in [n]$ the support of α_*. Define $s := |J|$. Since α_* is a sparse solution, we have $s \ll n$. Define ε as the largest regression error made by $f_*(\cdot)$ over the training examples in $\mathcal{D}$, i.e.

$$\varepsilon := \max_{i \in [n]} (f_*(\mathbf{x}_i) - y_i)^2 \tag{2}$$

Since $f_*(\cdot)$ is assumed to be an accurate prediction function, we assume that ε is small. We will show that by assuming a sparse solution for kernel regression, we will be able to accurately recover the solution α_* even with a modest number of random samples for low rank matrix approximation.

The proposed algorithm for sparse kernel regression combines the theory of sparse recovery with the Nyström method. We first random sample m instances from $\mathcal{D}$, denoted by $\widehat{\mathcal{D}} = \{\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_m\}$. Following the Nyström method, we approximate $\mathbf{K}$ by a low rank matrix $\widehat{\mathbf{K}}$ given by

$$\widehat{\mathbf{K}} = \mathbf{K}_a \mathbf{K}_b^{-1} \mathbf{K}_a^\top \tag{3}$$

where $\mathbf{K}_a = [\kappa(\widehat{\mathbf{x}}_i, \mathbf{x}_j)]_{n \times m}$ measures the similarity between every instance in $\mathcal{D}$ and each sampled instance in $\widehat{\mathcal{D}}$, and $\mathbf{K}_b = [\kappa(\widehat{\mathbf{x}}_i, \widehat{\mathbf{x}}_j)]_{m \times m}$. Since $\widehat{\mathbf{K}}$ is a low rank matrix, we can also write $\widehat{\mathbf{K}} = \mathbf{Z}\mathbf{Z}^\top$ as $\mathbf{Z} = \mathbf{K}_a \mathbf{K}_b^{-\frac{1}{2}}$. Given the low rank approximation of $\mathbf{K}$, we will solve the following optimization problem for sparse kernel regression:

$$\min_{\alpha \in \mathbb{R}^n} \quad \frac{1}{2n} \|\mathbf{y} - \widehat{\mathbf{K}}\alpha\|_2^2 + \frac{\gamma}{2} \|\alpha\|_2^2 + \lambda \|\alpha\|_1 \tag{4}$$

where λ and γ are the regularization parameters for the L_2 and L_1 constraints of α, respectively. We note that we introduce an ℓ_2 regularizer in (4) besides the ℓ_1 regularizer. This is closely related to the elastic net regularizer that was originally introduced for Lasso to select groups of correlated features (Zou and Hastie 2005; De Mol, De Vito, and Rosasco 2009). Unlike the elastic-net regularizer, the purpose of introducing ℓ_2 regularizer in (4) is to compensate the error in approximating $\mathbf{K}$ with $\widehat{\mathbf{K}}$.

The optimization problem in (4) is a convex optimization problem, and thus can be solved by a first order method or standard sparse optimization packages, such as (Liu, Ji, and Ye 2009; Schmidt 2010). Alternatively, at each iteration, given the current solution α_t, we can update the solution by solving the following optimization problem

$$\begin{aligned} \alpha_{t+1} \quad = \arg\min_{\alpha \in \mathbb{R}^n} \quad & \theta \|\alpha - \alpha_t\|_2^2 + \frac{1}{n} \alpha^\top \widehat{\mathbf{K}} (\widehat{\mathbf{K}} \alpha_t - \mathbf{y}) \\ & + \gamma \alpha_t^\top \alpha + \lambda \|\alpha\|_1 \end{aligned} \tag{5}$$

Since $\widehat{\mathbf{K}}$ is of low rank, both $\widehat{\mathbf{K}}\mathbf{y}$ and $\widehat{\mathbf{K}}^2\alpha$ can be computed efficiently whose cost are $O(dm)$.

Main Results

We first introduce a few notations that will be used throughout the section. Given a vector $\alpha \in \mathbb{R}^n$ and a subset of indices $J \subseteq [n]$, α_J is a vector of n dimension that includes all the entries of α in J, i.e.

$$[\alpha_J]_i = \alpha_i \mathbf{I}(i \in J)$$

Given a sparse solution $\alpha \in \mathbb{R}^n$, we use $\mathcal{S}(\alpha) \subseteq [n]$ to represent the support set of α, i.e. $\mathcal{S}(\alpha) = \{i \in [n] : \alpha_i \neq 0\}$. We use $\mathbf{K}_{*,j}$ to represent the jth column of kernel matrix $\mathbf{K}$, $\|\mathbf{K}\|_*$ to represent the spectral norm of $\mathbf{K}$, and $\|\mathbf{K}\|_1 = \max_{i \in [n]} \|\mathbf{K}_{*,i}\|_1$ the maximum ℓ_1 norm for all the columns in $\mathbf{K}$.

Similar to most sparse recovery problems, we need to make assumption about the design matrix $\mathbf{K}$ in order to ensure restricted isometric property (Koltchinskii 2011), which was introduced to characterize matrices which are nearly orthogonal. In particular, we assume that the kernel matrix $\mathbf{K}$ obeys the restricted eigenvalue condition (Koltchinskii 2011; Bickel, Ritov, and Tsybakov 2009), which has shown to be more general than several other similar assumptions used in sparsity literature (Candes and Tao 2007; Donoho, Elad, and Temlyakov 2006; Zhang and Huang 2008). More specifically, for any $J \subseteq [n]$, we define a function over a set J, $\beta(J)$ as

$$\begin{aligned} \beta(J) := \min\{\beta \geq 0 : \quad & \beta \|\mathbf{K}\alpha\|_2 \geq \sqrt{n} \|\alpha_J\|_2, \\ & \|\alpha_{J^c}\|_1 \leq 4 \|\alpha_J\|_1\} \end{aligned} \tag{6}$$

and β as

$$\beta := \min\{\beta(J) : J \in [n], |J| \le s\}$$

where parameter 4 is chosen for the convenience of analysis. We assume β, the minimal integer that connects $\|\mathbf{K}\alpha\|_2$ with $\|\alpha_J\|_2$, is not too large for kernel matrix $\mathbf{K}$. To better understand the implication of β for kernel matrix, we have the following theorem for bounding β. Define

$$\tau_+ := \max_{i\in[n]} \frac{1}{\sqrt{n}}\|\mathbf{K}_{*,i}\|_2, \quad \tau_- := \min_{i\in[n]} \frac{1}{\sqrt{n}}\|\mathbf{K}_{*,i}\|_2$$

and

$$\mu := \max_{1\le i<j\le n} \frac{1}{n}\|\mathbf{K}_{*,i}^\top \mathbf{K}_{*,j}\|$$

The following theorem bounds β for kernel matrix $\mathbf{K}$.

Theorem 1 *Suppose $\kappa(\mathbf{x}, \mathbf{x}') \ge 0$ for any $\mathbf{x}$ and $\mathbf{x}'$. If*

$$\frac{8s\mu}{\tau_-} \le \sqrt{\frac{\tau_-}{\tau_+ + s\mu}},$$

then

$$\beta \le \frac{\sqrt{\tau_-}}{\sqrt{\tau_-^{3/2} - 8\mu s\sqrt{\tau_+ + \mu s}}}$$

Furthermore, if

$$16\mu s \le \tau_- \sqrt{\frac{\tau_-}{\tau_+}},$$

we have

$$\beta \le \frac{2}{\tau_-^{1/4}}$$

As indicated by the above theorem, in order to obtain a modest β, the key is to ensure the similarity between any two instances is small enough.

Our analysis is divided into two parts. First, we will show that the sparse solution α_* can be accurately recovered by the optimal solution to (4) if the difference between $\mathbf{K}$ and $\widehat{\mathbf{K}}$ is small. Second, we will bound the spectral norm of $\mathbf{K} - \widehat{\mathbf{K}}$ for the Nyström method, and show it is small for kernel matrix with skewed eigenvalue distribution.

Define

$$\Delta = \mathbf{K} - \widehat{\mathbf{K}}, \quad \delta = \|\Delta\|_*.$$

Theorem 2 *Let $\widehat{\alpha}$ be the optimal solution to (4). If*

$$\begin{aligned} \gamma &\ge \frac{2\delta}{n}\|\mathbf{K}\|_* \\ \lambda &\ge 2\max\left(\frac{\varepsilon\|\mathbf{K}\|_1}{n} + \frac{\varepsilon\delta}{\sqrt{n}} + \frac{\delta\|\alpha_*\|_2}{n}(s+\delta), \gamma\|\alpha_*\|_\infty\right) \end{aligned}$$

we have

$$\|\widehat{\alpha} - \alpha_*\|_1 \le 10\lambda\beta s$$

Next, we will bound δ for the Nyström method. Similar to the previous analysis of Nyström method (e.g. (Gittens 2011)), we define a coherence measure for all the eigenvectors of kernel matrix $\mathbf{K}$

$$\tau := n \max_{1\le i,j\le n} \|\mathbf{U}_{i,j}\|^2$$

where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ include all the eigenvectors of $\mathbf{K}$.

Theorem 3 *Assume $\tau k \le n$. Then, with a probability $1 - 2n^{-3}$, we have*

$$\|\mathbf{K} - \widehat{\mathbf{K}}\|_* \le C'\lambda_{k+1}\sqrt{\frac{n}{m}}\log n,$$

provided that

$$m \ge C\tau k \log n$$

where both C and C' are universal constants.

Remark. Compared to the relative error bound given in (Gittens 2011) where $\|\mathbf{K} - \widehat{\mathbf{K}}\|_* \le O(\lambda_{k+1}n/m)$, our spectral norm bound for the Nyström method is significantly better. The key for the improvement is to explore the orthogonality between $\mathbf{U}_2$ and $\mathbf{U}_1$. To show $\delta = \|\mathbf{K} - \widehat{\mathbf{K}}\|_*$ can be small, consider the case when the eigenvalues of $\mathbf{K}$ follow a power law, i.e. $\lambda_k \le nk^{-p}$, where $p > 1$. By choosing $k = O(m/\log n)$, according to Theorem 3, we have

$$\delta \le O\left(\frac{n^{3/2}}{m^{(2p+1)/2}}[\log n]^{p+1}\right)$$

implying that δ will be small if

$$m \ge n^{3/(2p+1)}\log^2 n$$

Analysis

We first present proof for Theorem 2 and then the proof for Theorem 3.

Proof of Theorem 2

Let $J \subseteq [n]$ be the support set for α_*. The following lemma allows us to relate $\widehat{\alpha}$ to α_*.

Lemma 1 *We have*

$$\begin{aligned} & \frac{1}{n}(\widehat{\alpha} - \alpha_*)^\top \widehat{\mathbf{K}}(\widehat{\mathbf{K}}\widehat{\alpha} - \mathbf{y}) + \gamma\|\widehat{\alpha} - \alpha_*\|_2^2 + \lambda\|\widehat{\alpha}_{J^c}\|_1 \\ \le\ & \lambda\|\widehat{\alpha}_J - \alpha_*\|_1 + \gamma\|\alpha_*\|_\infty\|\widehat{\alpha}_J - \alpha_*\|_1 \end{aligned} \tag{7}$$

The next lemma allows us to bound the first term in the inequality in (7).

Lemma 2 *We have*

$$\begin{aligned} & \frac{1}{n}(\widehat{\alpha} - \alpha_*)^\top \widehat{\mathbf{K}}(\widehat{\mathbf{K}}\widehat{\alpha} - \mathbf{y}) \\ \ge\ & \frac{1}{n}\|\mathbf{K}(\widehat{\alpha} - \alpha_*)\|_2^2 + \frac{1}{n}\|\Delta(\widehat{\alpha} - \alpha_*)\|_2^2 \\ & + \frac{1}{n}\|\widehat{\mathbf{K}}^{\frac{1}{2}}(\widehat{\alpha} - \alpha_*)\|_2^2 - a\|\widehat{\alpha} - \alpha_*\|_1 - b\|\widehat{\alpha} - \alpha_*\|_2^2 \end{aligned}$$

where

$$a := \varepsilon\left(\frac{\|\mathbf{K}\|_1}{n} + \frac{\delta}{\sqrt{n}}\right) + \frac{\delta\|\alpha_*\|_2}{n}(s+\delta), \quad b := \frac{2\delta}{n}\|\mathbf{K}\|_*$$

By choosing

$$\begin{aligned} \gamma &\ge b = \frac{2\delta}{n}\|\mathbf{K}\|_* \\ \lambda &\ge 2\max(a, \gamma\|\alpha_*\|_\infty) \\ &= 2\max\left(\frac{\varepsilon\|\mathbf{K}\|_1}{n} + \frac{\varepsilon\delta}{\sqrt{n}} + \frac{\delta\|\alpha_*\|_2}{n}(s+\delta), \gamma\|\alpha_*\|_\infty\right) \end{aligned}$$

we have

$$\frac{1}{n}\|\mathbf{K}(\widehat{\alpha}-\alpha_*)\|_2^2+\frac{1}{n}\|\Delta(\widehat{\alpha}-\alpha_*)\|_2^2++\frac{\lambda}{2}\|\widehat{\alpha}_{J^c}\|_1\leq 2\lambda\|\widehat{\alpha}_J-\alpha_*\|_1$$

We thus have $\|\widehat{\alpha}_{J^c}\|_1 \leq 4\|\widehat{\alpha}_J-\alpha_*\|_1$, and using the definition of β, we have

$$\begin{aligned}\frac{\|\widehat{\alpha}_J-\alpha_*\|_1^2}{\beta} &\leq \frac{s\|\widehat{\alpha}_J-\alpha_*\|_2^2}{\beta}\\ &\leq \frac{s}{n}\|\mathbf{K}(\widehat{\alpha}-\alpha_*)\|_2^2\leq 2\lambda s\|\widehat{\alpha}_J-\alpha_*\|_1\end{aligned}$$

and therefore

$$\|\widehat{\alpha}_J-\alpha_*\|_1\leq 2\lambda s\beta$$

and

$$\begin{aligned}\|\widehat{\alpha}-\alpha_*\|_1 &= \|\widehat{\alpha}_J-\alpha_*\|_1+\|\widehat{\alpha}_{J^c}\|_1\\ &\leq 5\|\widehat{\alpha}_J-\alpha_*\|_1\leq 10\lambda s\beta\end{aligned}$$

Proof of Theorem 3

Our analysis is based on Theorem 1 (Gittens 2011) given as below.

Theorem 4 *(Theorem 1 from (Gittens 2011)) Let* $\mathbf{A}$ *a PSD matrix of size* n*, and let* $\mathbf{S}$ *be a* $n\times\ell$ *random matrix with each column has exactly one non-zero element with value* 1*. Partition* $\mathbf{A}$ *as*

$$\mathbf{A}=\begin{array}{cc} k & n-k\\ [\mathbf{U}_1 & \mathbf{U}_2]\end{array}\left[\begin{array}{cc}\Sigma_1 & \\ & \Sigma_2\end{array}\right]\left[\begin{array}{c}\mathbf{U}_1^\top\\ \mathbf{U}_2^\top\end{array}\right]$$

and define Ω_1 *and* Ω_2 *as*

$$\Omega_1=\mathbf{U}_1^\top\mathbf{S},\ \Omega_2=\mathbf{U}_2^\top\mathbf{S}$$

Assume Ω_1 *has full row rank. Then, the spectral approximation error of the naive Nyström method extension of* $\mathbf{A}$ *using* $\mathbf{S}$ *as the column sampling matrix satisfies*

$$\begin{aligned}\|\mathbf{A}-\mathbf{C}\mathbf{W}^\dagger\mathbf{C}\|_* &= \|(\mathbf{I}-P_{\mathbf{A}^{1/2}\mathbf{S}})\mathbf{A}^{1/2}\|_2^2\\ &\leq \|\Sigma_2\|_*\left(1+\|\Omega_2\Omega_1^\dagger\|_*^2\right)\end{aligned}\quad(8)$$

where $\mathbf{C}=\mathbf{A}\mathbf{S}$ *and* $\mathbf{W}=\mathbf{S}^\top\mathbf{A}\mathbf{S}$*.*

To map the result in Theorem 4 to the Nyström method, we note $\mathbf{C}$ and $\mathbf{W}$ in Theorem 4 correspond to $\mathbf{K}_a$ and $\mathbf{K}_b$ in (3), respectively. The key of using Theorem 4 is to effectively bound $\|\Omega_2\Omega_1^\dagger\|_*$. To this end, we need the following matrix concentration inequality.

Theorem 5 *(Theorem 1.2 from (Candés and Romberg 2007)) Let* $U\in N\times N$ *include the eigenvectors of a PSD matrix* $\mathbf{A}$ *with coherence* τ*. Let* $\mathbf{U}_1\in R^{N\times k}$ *include the first* k *eigenvectors of* $\mathbf{A}$*. Let* $\mathbf{S}\in\{0,1\}^{n\times\ell}$ *be a random matrix distributed as the first* ℓ *columns of a uniformly random permutation matrix of size* n*. Suppose that* ℓ *obeys* $\ell\geq\tau k\max(\mathbf{C}_a\ln k, C_b\ln(3/\delta))$ *for some positive constants* C_a *and* C_b*. Then*

$$\Pr\left(\left\|\frac{n}{\ell}\mathbf{U}_1^\top\mathbf{S}\mathbf{S}^\top\mathbf{U}_1-\mathbf{I}\right\|_*\geq 1/2\right)\leq\delta.$$

To bound $\|\Omega_2\Omega_1^\dagger\|_*^2$, we write it down explicitly as

$$\|\Omega_2\Omega_1^\dagger\|_*^2=\|\mathbf{U}_2^\top\mathbf{S}\mathbf{S}^\top\mathbf{U}_1(\mathbf{U}_1^\top\mathbf{S}\mathbf{S}^\top\mathbf{U}_1)^{-1}\|_*^2$$

We only consider the event ω_1 where

$$\left\|\frac{n}{m}\mathbf{U}_1^\top\mathbf{S}\mathbf{S}^\top\mathbf{U}_1-\mathbf{I}\right\|_*\leq 1/2$$

According to Theorem 5, we have $\Pr(\omega_1)\geq 1-n^{-3}$ provided that

$$m\geq C\tau k\ln n$$

where C is some universal constant. We then proceed to bound $\|\mathbf{U}_2^\top\mathbf{S}\mathbf{S}^\top\mathbf{U}_1\|_*$ which is given in the following lemma.

Lemma 3 *Assume* $\tau k\leq n$*. Then, with a probability* $1-n^{-3}$*, we have*

$$\|\mathbf{U}_2^\top\mathbf{S}\mathbf{S}^\top\mathbf{U}_1\|_*\leq 8\tau\sqrt{\frac{n}{m}}\log(2n)$$

Combining the results in Theorem 5 and Lemma 3, we have, with a probability $1-2n^{-3}$,

$$\|\Omega_2\Omega_1^\dagger\|_*^2\leq C'\tau\sqrt{\frac{n}{m}}\log n\quad(9)$$

provided that

$$m\geq C\tau k\log n$$

where C and C' are two universal constants. We complete the proof by replacing $\|\Omega_2\Omega_1^\dagger\|_*$ in (8) with the bound in (9).

Empirical Studies

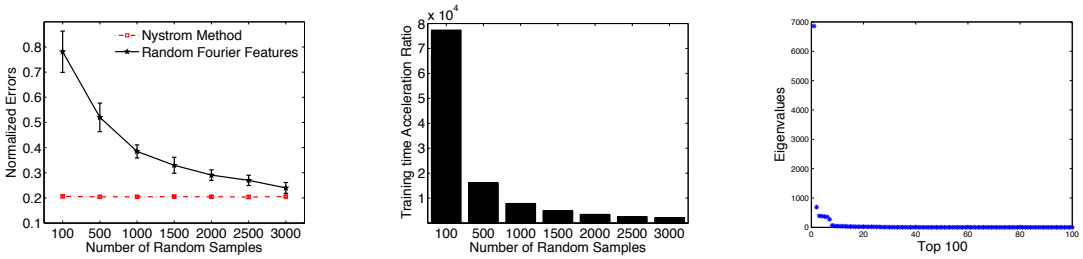
We conduct experiments to verify the efficiency of the proposed approximation algorithm for sparse kernel regression. For comparison, we choose the approximation by random Fourier features, which approximate the shift-invariant kernel matrix with the Fourier transform of a non-negative measure (Rahimi and Recht 2007).

Experiment on Synthetic Data

In order to better show the performance of the Nyström approximation of sparse kernel regression, we design a synthetic data set. We have the following expectations: 1) data containing nonlinearity on the features; and 2) data being embedded with redundant and grouping features. We then generate a 20-dimensional synthetic dataset with 20,000 examples by additive models motivated by an example in (Härdle et al. 2004). And the data are generated by

$$\begin{aligned}y_i &= \sum_{j=1}^{3}f_1(x_{ij})+\sum_{j=4}^{6}f_2(x_{ij})+\sum_{j=7}^{9}f_3(x_{ij})\\ &\quad+\sum_{j=10}^{12}f_4(x_{ij})+\epsilon_i,\end{aligned}\quad(10)$$

where there are four mapping functions: $f_1(x)=-2\sin(2x)+1-\cos(2)$, $f_2(x)=x^2-\frac{1}{3}$, $f_3(x)=x-\frac{1}{2}$, and $f_4(x)=e^{-x}+e^{-1}-1$. We set the rest of mapping functions as $f_j(x)=0$, for $j>9$. x is uniformly distributed



(a) *Normalized error* (b) *Accelerating ratio on training time* (c) *Eigenvalues*

Figure 1: Evaluation on the synthetic data. (a) shows the normalized error in comparison with the method of random Fourier features when the number of samples is changed from 500 to 3000. The lower the better. (b) shows the acceleration ratio of training time for the Nyström approximation over computing the entire kernel matrix. The larger the better. (c) shows the top eigenvalues of the kernel matrix.

in $[0, 1]$, and the noise $\epsilon_i \sim \mathcal{N}(0, 1)$ is a Gaussian noise. The output Y_i is determined by the corresponding features, from 1 to 12, in x_i mapped by f_1, f_2, f_3, and f_4, respectively. The data therefore contain 8 irrelevant features.

We repeat all the algorithms 20 times for each data set. In each run, 50% of the examples are randomly selected as the training data and the remaining data are used for testing. The training data are normalized to have zero mean and unit variance, and the test data are then normalized using the mean and variance of the training data. The regularization parameters λ and γ are selected by 10-fold cross validation. The experiments are conducted on a PC with 2.67GHz CPU and 4GB memory.

We report the following performance measures: the normalized error (i.e. mean square error divided by the norm of training data) on the test data, and the training time. It can be observed that the Nyström approximation achieves much better performance than the method of Random Fourier Features. This is indeed not superposing in that the random Figure 1(a) shows the normalized error.samples in the Nyström method are strongly dependant with the training data. Figure 1(b) shows the acceleration ratio of the training time when varying the number of random samples over the training time when calculating the entire kernel matrix. Especially, when only 500 random samples are used to approximate the kernel matrix, it is over 70,000 times faster than the naive method calculating the entire kernel matrix, while the normalized errors are in the same level. Figure 1(c) partially explains why the acceleration is so big – the eigenvalues of the kernel matrix has the property of fast decade.

Experiment on Real-world Data

We adopted three data sets from other literature and the UCI machine learning repository http://archive.ics.uci.edu/ml/datasets.html. The *CPU* data set also used in (Rahimi and Recht 2007) contains 6,554 examples. The *Bike* data set records the rental behavior and the corresponding weather conditions, precipitation, day of week, season, hour of the day, etc. The *Slice* data set contains 53500 CT images from 74 different patients (43 male, 31 female). Each CT slice is described by two histograms in polar space. The label variable (relative location of an image on the axial axis) was constructed by manually annotating up to 10 different distinct landmarks in each CT Volume with known location. The location of slices in between landmarks was interpolated.

Table 1: Description of the Real-world regression data sets.

Data set	# Training data	# Test data	# Dim
CPU	6554	819	21
Bike	8645	8734	13
Slice	29852	23648	348

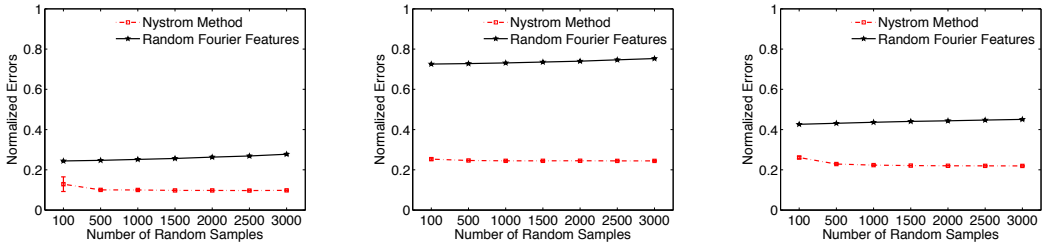
The data information is summarized in Table 1.

We first report the normalized error on the three data sets. As shown in Figure 2, it can be observed that the Nyström approximation achieves much lower normalized errors than the method of Random Fourier Features on all the three data sets. This consistently shows the strong generalization ability of the Nyström method. Especially, even using a small number of selected examples, the Nyström method can achieve very close results with using the entire kernel matrix.

We then report the average training time with standard division for the real-world data sets, as shown in Figure 3. It can be observed that the training procedure after using the Nyström approximation can be very fast even on a large scale data with tens of thousands examples. To further clearly show the advantages over computing the entire kernel matrix, we also report the acceleration ratio of using the Nyström method as shown in Figure 4. It can also be observed that the Nyström method can extremely accelerate the training time. Note that it is very hard to compute the entire kernel matrix for the *Slice* data in a computer with a modest size of memory due to the high storage and computation complexities of large kernel matrices. So we omit the result on the Nyström method.

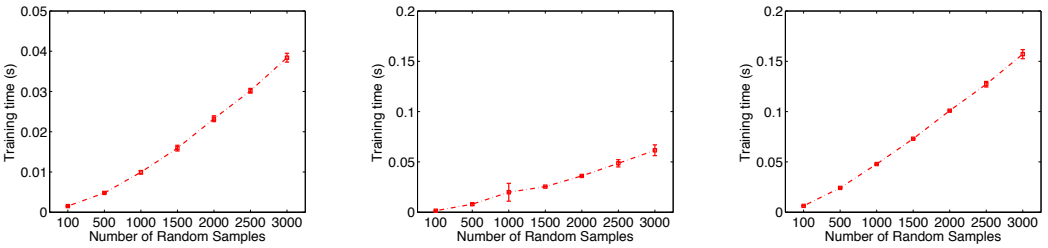
Conclusion

In this paper, we have presented the theoretical analysis and empirical evaluation of the Nyström approximation for sparse kernel regression problems. Based on the restricted



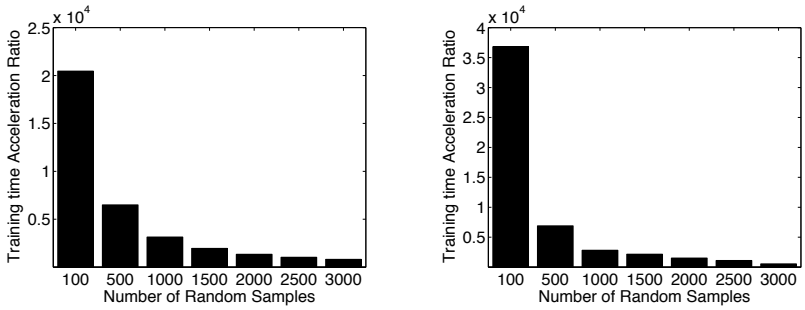
(a) *Normalized error on CPU* (b) *Normalized error on Bike* (c) *Normalized error on Slice*

Figure 2: Test error comparison on three real-world data sets. (a) shows the normalized error on *CPU* in comparison with the method of random Fourier features when the number of samples is changed from 500 to 3000. The lower the better. (b) shows the normalized error on *Bike*. The larger the better. (c) shows the normalized error on *Slice*.



(a) *Training time on CPU* (b) *Training time on Bike* (c) *Training time on Slice*

Figure 3: Training time comparison on three real-world data sets. (a) shows the training time on *CPU* when the number of samples is changed from 500 to 3000. The lower the better. (b) shows the training time on *Bike*. (c) shows the training time on *Slice*.



(a) *Acceleration ratio on CPU* (b) *Acceleration ratio on Bike*

Figure 4: The acceleration ratio of training time for the Nyström approximation over computing the entire kernel matrix. (a)The acceleration ratio on *CPU*. (c) The acceleration ratio on *Bike*. The larger the better.

eigenvalue condition, we not only provide an approximation bound for arbitrary kernels, but also provide a stable recovery rate for sparse kernel methods. In addition to the theoretical analysis, we also demonstrate the good performance of Nyström approximation on real world data sets.

For the future work, we will seek to provide better approximation bounds and recovery rates for sparse kernel classification methods.

Acknowledgement

This work was supported by a Project-985 grant from University of Electronic Science and Technology of China(No. A1098531023601041).

References

Bach, F. R., and Jordan, M. I. 2005. Predictive low-rank decomposition for kernel methods. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, 33–40. New York, NY, USA: ACM.

Bickel, P. J.; Ritov, Y.; and Tsybakov, A. B. 2009. Simultaneous analysis of lasso and dantzig selector. *ANNALS OF STATISTICS* 37(4).

Candés, E., and Romberg, J. 2007. Sparsity and incoherence in compressive sampling. *Inverse Problems* 23(3):969–985.

Candes, E., and Tao, T. 2007. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.* 35(6):2313–2351.

De Mol, C.; De Vito, E.; and Rosasco, L. 2009. Elastic-net regularization in learning theory. *J. Complex.* 25(2):201–230.

Donoho, D. L.; Elad, M.; and Temlyakov, V. N. 2006. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE TRANS. INFORM. THEORY* 52(1):6–18.

Drineas, P., and Mahoney, M. W. 2005. On the Nyström method for approximating a gram matrix for improved kernel-based learning. *JOURNAL OF MACHINE LEARNING RESEARCH* 6:2005.

Farahat, A. K.; Ghodsi, A.; and Kamel, M. S. 2011. In *AISTATS*, 269–277.

Fine, S., and Scheinberg, K. 2002. Efficient svm training using low-rank kernel representations. *J. Mach. Learn. Res.* 2:243–264.

Fowlkes, C.; Belongie, S.; Chung, F.; and Malik, J. 2004. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.

Furrer, R.; Genton, M. G.; and Nychka, D. 2006. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* 15(3).

Gittens, A., and Mahoney, M. W. 2013. Revisiting the Nyström method for improved large-scale machine learning. *CoRR* abs/1303.1849.

Gittens, A. 2011. The spectral norm error of the naive Nyström extension. *CoRR* abs/1110.5305.

Härdle, W.; Müller, M.; Sperlich, S.; and Werwatz, A. 2004. *Nonparametric and Semiparametric Models*. Springer-Verlag Inc.

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

Koltchinskii, V. 2011. *Oracle inequalities in empirical risk minimization and sparse recovery problems. École d'Été de Probabilités de Saint-Flour XXXVIII-2008.* Berlin: Springer.

Kumar, S.; Mohri, M.; and Talwalkar, A. 2012. Sampling methods for the Nyström method. *J. Mach. Learn. Res.* 13:981–1006.

Liu, J.; Ji, S.; and Ye, J. 2009. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University.

Mackey, L. W.; Talwalkar, A.; and Jordan, M. I. 2011. Divide-and-conquer matrix factorization. In *Advances in Neural Information Processing Systems*, 1134–1142.

Rahimi, A., and Recht, B. 2007. Random features for large-scale kernel machines. In *Advances in Neural Infomration Processing Systems*.

Schmidt, M. 2010. *Graphical Model Structure Learning with L1-Regularization*. Ph.D. Dissertation, University of British Columbia.

Schölkopf, B., and Smola, A. 2002. *Learning with Kernels*. Cambridge, MA: MIT Press.

Shen, B.; Xu, Z.; and Allebach, J. P. 2014. Kernel tapering: a simple and effective approach to sparse kernels for image processing. In *International Conference on Image Processing*.

Shi, L.; Feng, Y.-L.; and Zhou, D.-X. 2011. Concentration estimates for learning with L1-regularizer and data dependent hypothesis spaces. *Appl. Comput. Harmon. Anal.* 31.

Talwalkar, A., and Rostamizadeh, A. 2010. Matrix coherence and the nystrom method. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI)*, 572–579.

Wang, J.; Dong, Y.; Tong, X.; Lin, Z.; and Guo, B. 2009. Kernel Nyström method for light transport. *ACM Trans. Graph.* 28(3):29:1–29:10.

Williams, C., and Seeger, M. 2001. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, 682–688. MIT Press.

Xu, Z.; Jin, R.; King, I.; and Lyu, M. 2009. An extended level method for efficient multiple kernel learning. In Koller, D.; Schuurmans, D.; Bengio, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 21 (NIPS)*. 1825–1832.

Zhang, C.-H., and Huang, J. 2008. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Stat.* 36(4):1567–1594.

Zhang, K.; Tsang, I. W.; and Kwok, J. T. 2008. Improved Nyström low-rank approximation and error analysis. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, 1232–1239. New York, NY, USA: ACM.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67:301–320.