Error Analysis of Generalized Nyström Kernel Regression

Hong Chen
Computer Science and Engineering
University of Texas at Arlington
Arlington, TX, 76019
chenh@mail.hzau.edu.cn

Haifeng Xia
Mathematics and Statistics
Huazhong Agricultural University
Wuhan 430070,China
haifeng.xia0910@gmail.com

Weidong Cai
School of Information Technologies
University of Sydney
NSW 2006, Australia
tom.cai@sydney.edu.au

Heng Huang
Computer Science and Engineering
University of Texas at Arlington
Arlington, TX, 76019
heng@uta.edu

Abstract

Nyström method has been successfully used to improve the computational efficiency of kernel ridge regression (KRR). Recently, theoretical analysis of Nyström KRR, including generalization bound and convergence rate, has been established based on reproducing kernel Hilbert space (RKHS) associated with the symmetric positive semi-definite kernel. However, in real world applications, RKHS is not always optimal and kernel function is not necessary to be symmetric or positive semi-definite. In this paper, we consider the generalized Nyström kernel regression (GNKR) with ℓ_2 coefficient regularization, where the kernel just requires the continuity and boundedness. Error analysis is provided to characterize its generalization performance and the column norm sampling strategy is introduced to construct the refined hypothesis space. In particular, the fast learning rate with polynomial decay is reached for the GNKR. Experimental analysis demonstrates the satisfactory performance of GNKR with the column norm sampling.

1 Introduction

The high computational complexity makes kernel methods unfeasible to deal with large-scale data. Recently, the Nyström method and its alternatives (*e.g.*, the random Fourier feature technique [15], the sketching method [25]) have been used to scale up kernel ridge regression (KRR) [4, 23, 27]. The key step of Nyström method is to construct a subsampled matrix, which only contains part columns of the original empirical kernel matrix. Therefore, the sampling criterion on the matrix column affects heavily on the learning performance. The subsampling strategies of Nyström method can be categorized into two types: uniform sampling and non-uniform sampling. The uniform sampling is the simplest strategy, which has shown satisfactory performance on some applications [16, 23, 24]. From different theoretical aspects, several non-uniform sampling approaches have been proposed such as the square ℓ_2 column-norm sampling [3, 4], the leverage score sampling [5, 8, 12], and the adaptive sampling [11]. Besides the sampling strategies, there exist learning bounds for Nyström kernel regression from three measurements: the matrix approximation [4, 5, 11], the coefficient approximation [9, 10], and the excess generalization error [2, 16, 24].

Despite rapid progress on theory and applications, the following critical issues should be further addressed for Nyström kernel regression.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

- *Nyström regression with general kernel.* The previous algorithms are mainly limited to KRR with symmetric and positive semi-definite kernels. For real-world applications, this restriction may be not necessary. Several general kernels have shown the competitive performance in machine learning, *e.g.*, the indefinite kernels for regularized algorithms [14, 20, 26] and PCA [13]. Therefore, it is important to formulate the learning algorithm for *Generalized Nyström Kernel Regression (GNKR).*
- *Generalization analysis and sampling criterion.* Previous theoretical results rely on the symmetric positive semi-definite (SPSD) matrix associated with a Mercer kernel [17]. However, this condition is not satisfied for GNKR, which induces the additional difficulty on error analysis. Can we get the generalization error analysis for GNKR? It is also interesting to explore the sampling strategy for GNKR, *e.g.*, the column-norm sampling in [3, 4].

To address the above issues, we propose the GNKR algorithm and investigate its theoretical properties on generalization bound and learning rate. Inspired from the recent studies for data dependent hypothesis spaces [7, 19], we establish the error analysis for GNKR, which implies that the learning rate with polynomial decay can be reached under proper parameter selection. Meanwhile, we extend the ℓ_2 column norm subsampling in the linear regression [16, 22] to the GNKR setting.

The main contributions of this paper can be summarized as below:

- *GNKR with ℓ_2 regularization.* Due to the lack of Mercer condition associated with general kernel, coefficient regularization becomes a natural choice to replace the kernel norm regularization in KRR. Note that Nyström approximation has the similar role with the ℓ_1 regularization in [7, 18, 20], which addresses the sample sparsity on hypothesis function. Hence, we formulate GNKR by combining the Nyström method and the least squares regression with ℓ_2 regularization in [19, 21].
- *Theoretical and empirical evaluations.* From the view of learning with data dependent hypothesis spaces, theoretical analysis of GNKR is established to illustrate its generalization bound and learning rate. In particular, the fast learning rate arbitrarily close to $O(m^{-1})$ is obtained under mild conditions, where m is the size of subsampled set. The effectiveness of GNKR is also supported by experiments on synthetic and real-world data sets.

2 Related Works

Due to the flexibility and adaptivity, least squares regression algorithms with general kernel have been proposed involving various types of regularization, *e.g.*, the ℓ_1-regularizer [18, 21], the ℓ_2-regularizer [19, 20], and the elastic net regularization [7]. For the Mercer kernel, these algorithms are related closely with the KRR, which has been well understand in learning theory. For the general kernel setting, theoretical foundations of regression with coefficient regularization have been studied recently via the analysis techniques with the operator approximation [20] and the empirical covering numbers [7, 18, 19]. Although rich results on theoretical analysis, the previous works mainly focus on the prediction accuracy without considering the computation complexity for large scale data.

Nyström approximation has been studied extensively for kernel methods recently. Almost all existing studies are relied on the fast approximation of SPSD matrix associated with a Mercer kernel. For the fixed design setting, the expectation of the excess generalization error is bounded for least square regression with the regularizer in RKHS [1, 2]. Recently, the probabilistic error bounds have been estimated for Nyström KRR in [16, 24]. In [24], the fast learning rate with $O(m^{-1})$ is derived for the fixed design regression under the conditions on kernel matrix eigenvalues. In [16], the convergence rate is obtained under the capacity assumption and the regularity assumption. It is worthy notice that the learning bound in [16] is based on the estimates of the sample error, the computation error, and the approximation error. Indeed, the computation error is related with the sampling subset and can be considered as the hypothesis error in [18], which is induced by the variance of hypothesis spaces. Differently from previous works, our theoretical analysis of GNKR is dependent on general continuous kernel and ℓ_2 coefficient regularization.

3 Generalized Nyström Kernel Regression

Let ρ be a probability distribution on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ are viewed as the input space and the output space, respectively. Let $\rho(\cdot|x)$ be the conditional distribution of ρ for given $x \in \mathcal{X}$ and let $\mathcal{F}$ be a measurable function space on $\mathcal{X}$. In statistical learning, the samples $\mathbf{z} := \{z_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n$ are drawn independently and identically from an unknown distribution ρ. The task of least squares regression is to find a prediction function $f : \mathcal{X} \to \mathbb{R}$ such that the expected risk

$$\mathcal{E}(f) = \int_{\mathcal{Z}} (y - f(x))^2 d\rho(x, y)$$

as small as possible. From the viewpoint of approximation theory, this means to search a good approximation of the regression function

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x)$$

based on the empirical risk

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2.$$

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous and bounded kernel function. Without loss of generality, we assume that $\kappa := \sup_{x, x' \in \mathcal{X}} K(x, x') \leq 1$ and for all $|y| \leq 1$ for all $y \in \mathcal{Y}$ throughout this paper.

Besides the given samples $\mathbf{z}$, the hypothesis function space is crucial to reach well learning performance. The following data dependent hypothesis space has been used for kernel regression with coefficient regularization:

$$\mathcal{H}_n = \Big\{ f(x) = \sum_{i=1}^{n} \tilde{\alpha}_i K(x_i, x) : \tilde{\alpha} = (\tilde{\alpha}_1, ..., \tilde{\alpha}_n) \in \mathbb{R}^n, x \in \mathcal{X} \Big\}.$$

Given $\mathbf{z}$, kernel regression with ℓ_2 regularization [19, 20] is formulated as

$$\tilde{f}_{\mathbf{z}} = f_{\tilde{\alpha}_{\mathbf{z}}} = \sum_{i=1}^{n} \tilde{\alpha}_{\mathbf{z},i} K(x_i, \cdot) \tag{1}$$

with

$$\tilde{\alpha}_{\mathbf{z}} = \arg\min_{\tilde{\alpha} \in \mathbb{R}^n} \Big\{ \frac{1}{n} \| K_{nn} \tilde{\alpha} - Y \|_2^2 + \lambda n \tilde{\alpha}^T \tilde{\alpha} \Big\},$$

where $K_{nn} = (K(x_i, x_j))_{i,j=1}^n$, $Y = (y_1, \cdots, y_n)^T$, and $\lambda > 0$ is a regularization parameter.

Even the positive semi-definiteness is not required for the kernel, (3) also can be solved by the following linear system (see Theorem 3.1 in [20])

$$(K_{nn}^T K_{nn} + \lambda n^2 I_n) \tilde{\alpha} = K_{nn}^T Y, \tag{2}$$

where I_n is the n-order unit matrix.

From the viewpoint of learning function in $\mathcal{H}_n$, (1) can be rewritten as

$$\tilde{f}_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}_n} \Big\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda n \|f\|_{\ell_2}^2 \Big\}, \tag{3}$$

where

$$\|f\|_{\ell_2}^2 = \inf \Big\{ \sum_{i=1}^{n} \tilde{\alpha}_i^2 : f = \sum_{i=1}^{n} \tilde{\alpha}_i K(x_i, \cdot) \Big\}.$$

In a standard implementation of (2), the computational complexity is $O(n^3)$. This computation requirement becomes the bottleneck of (3) when facing large data sets. To reduce the computational burden, we consider to find the predictor in a smaller hypothesis space

$$\mathcal{H}_m = \Big\{ f(x) = \sum_{i=1}^{m} \alpha_i K(\bar{x}_i, x) : \alpha = (\alpha_1, ..., \alpha_m) \in \mathbb{R}^m, x \in \mathcal{X}, \{\bar{x}_i\}_{i=1}^m \subset \{x_i\}_{i=1}^n \Big\}.$$

The generalized Nyström kernel regression (GNKR) can be formulated as

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}_m} \Big\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda m \|f\|_{\ell_2}^2 \Big\}. \tag{4}$$

Denote $(K_{nm})_{ij} = K(x_i, \bar{x}_j), (K_{mm})_{jk} = K(\bar{x}_i, \bar{x}_j)$ for $i \in \{1, ..., n\}, j, k \in \{1, ..., m\}$. We can deduce that

$$f_{\mathbf{z}} = \sum_{i=1}^{m} \alpha_{\mathbf{z},i} K(\bar{x}_i, \cdot)$$

with

$$(K_{nm}^T K_{nm} + \lambda mn I_m)\alpha_{\mathbf{z}} = K_{nm}^T Y. \tag{5}$$

The key problem of (4) is how to select the subset $\{\bar{x}_i\}_{i=1}^m$ such that the computational complexity can be decreased efficiently while satisfactory accuracy can be guaranteed. For the KRR, there are several strategies to select the subset with different motivations [5, 11, 12]. In this paper we preliminarily consider the following two strategies with low computational complexity:

- *Uniform Subsampling*. The subset $\{\bar{x}_i\}_{i=1}^m$ is drawn uniformly at random from the input $\{x_i\}_{i=1}^n$.
- *Column-norm Subsampling*. The subset $\{\bar{x}_i\}_{i=1}^m$ is drawn from $\{x_i\}_{i=1}^n$ independently with probabilities $p_i = \frac{\|K_i\|_2}{\sum_{i=1}^n \|K_i\|_2}$, where $K_i = (K(x_1, x_i), ..., K(x_n, x_i))^T \in \mathbb{R}^n$.

Some discussions for the column-norm subsampling will be provided in Section 4.

4 Learning Theory Analysis

In this section, we will introduce our theoretical results on generalization bound and learning rate. The detailed proofs can be found in the supplementary materials.

Inspired from analysis technique in [7, 19], we introduce the intermediate function for error decomposition firstly. Let $\mathcal{F}$ be the square integrable space on $\mathcal{X}$ with norm $\|\cdot\|_{L^2_{\rho_{\mathcal{X}}}}$. For any bounded continuous kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the integral operator $L_K : \mathcal{F} \to \mathcal{F}$ is defined as

$$L_K f(x) = \int_{\mathcal{X}} K(x, t) f(t) d\rho_{\mathcal{X}}(t), \forall x \in \mathcal{X},$$

where $\rho_{\mathcal{X}}$ is the marginal distribution of ρ. Given $\mathcal{F}$ and L_K, introduce the function space

$$\mathcal{H} = \Big\{ g = L_K f, f \in \mathcal{F} \Big\} \text{ with } \|g\|_{\mathcal{H}} = \inf \big\{ \|f\|_{L^2_{\rho_{\mathcal{X}}}} : g = L_K f \big\}.$$

Since $\mathcal{H}$ is sample independent, the intermediate function can be constructed as $g_\lambda = L_K f_\lambda$, where

$$f_\lambda = \arg\min_{f \in \mathcal{F}} \Big\{ \mathcal{E}(L_K f) - \mathcal{E}(f_\rho) + \lambda \|f\|_{L^2_{\rho_{\mathcal{X}}}}^2 \Big\}. \tag{6}$$

In learning theory, usually g_λ is called as the regularized function and

$$D(\lambda) = \inf_{g \in \mathcal{H}} \{ \mathcal{E}(g) - \mathcal{E}(f_\rho) + \lambda \|g\|_{\mathcal{H}}^2 \} = \mathcal{E}(L_K f_\lambda) - \mathcal{E}(f_\rho) + \lambda \|f_\lambda\|_{L^2_{\rho_{\mathcal{X}}}}^2$$

is called the approximation error

To further bridge the gap between g_λ and $f_{\mathbf{z}}$, we construct the stepping stone function

$$\hat{g}_\lambda = \frac{1}{m} \sum_{i=1}^{m} f_\lambda(\bar{x}_i) K(\bar{x}_i, \cdot). \tag{7}$$

The following condition on K is used in this paper, which has been well studied in learning theory literature [18, 19]. Examples include Gaussian kernel, the sigmoid kernel [17], and the fractional power polynomials [13].

Definition 1 *The kernel function K is a C^s kernel with $s > 0$ if there exists some constant $c_s > 0$, such that*

$$|K(t,x) - K(t,x')| \le c_s \|x - x'\|_2^s, \ \ \forall t, x, x' \in \mathcal{X}.$$

The definition of f_ρ tells us $|f_\rho(x)| \le 1$, so it is natural to restrict the predictor to $[-1, 1]$. The projection operator

$$\pi(f)(x) = \min\{1, f(x)\}I\{f(x) \ge 0\} + \max\{-1, f(x)\}I\{f(x) < 0\}$$

has been extensively used in learning theory analysis, *e.g.* [6].

It is a position to present our result on the generalization error bound.

Theorem 1 *Suppose that $\mathcal{X}$ is compact subset of $\mathbb{R}^d$ and $K \in C^s(\mathcal{X} \times \mathcal{X})$ for some $s > 0$. For any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le \tilde{c}_1 \log^2(8/\delta)\Big((1 + m^{-1}\lambda^{-1} + m^{-2}\lambda^{-2} + n^{-\frac{2}{2+p}}\lambda^{-2})D(\lambda) + n^{-\frac{2}{2+p}}\lambda^{-\frac{p}{2+p}}\Big),$$

where constant $\tilde{c}_1$ is independent of m, n, δ, and

$$p = \begin{cases} 2d/(d+2s), & \textit{if } 0 < s \le 1; \\ 2d/(d+2), & \textit{if } 1 < s \le 1 + d/2; \\ d/s, & \textit{if } s > 1 + d/2. \end{cases} \tag{8}$$

Theorem 1 is a general result that applies to Lipschitz continuous kernel. Although the statement appears somewhat complicated at first sight, it yields fast convergence rate on the error when specialized to particular kernels. Before doing so, let us provide a few heuristic arguments for intuition. Theorem 1 guarantees an upper bound of the form

$$\|\pi(f_{\mathbf{z}}) - f_\rho\|^2_{L^2_{\rho_\mathcal{X}}} \le O\Big(c(m,n,\lambda)\inf_f\{\mathcal{E}(L_K f) - \mathcal{E}(f_\rho) + \lambda\|f\|^2_{L^2_{\rho_\mathcal{X}}}\} + n^{-\frac{2}{2+p}}\lambda^{-\frac{p}{2+p}}\Big). \tag{9}$$

Note that a smaller value of λ reduces the approximation error term, but increases the second term associated with the sample error. This inequality demonstrates that the proper λ should be selected to balance the two terms. This quantitative relationship (9) also can be considered as the oracle inequality for GNKR, where the approximation error $D(\lambda)$ only can be obtained by an oracle knowing the distribution.

Theorem 1 tells us that the generalization bound of GNKR depends on the numbers of samples m, n, the continuous degree s, and the approximation error $D(\lambda)$. In essential, the subsampling number m has double impact on generalization error: one is the complexity of data dependent hypothesis space $\mathcal{H}_m$ and the other is the selection of parameter λ.

Now we introduce the characterization of approximation error, which has been studied in [19, 20].

Definition 2 *The target function f_ρ can be approximated with exponent $0 < \beta \le 1$ in $\mathcal{H}$ if there exists a constant $c_\beta \ge 1$ such that $D(\lambda) \le c_\beta \lambda^\beta$ for any $\lambda > 0$.*

If the kernel is not symmetric or positive semi-definite, the approximation condition holds true for $\beta = \frac{2r}{3}$ when $f_\rho \in L^{-r}_{\tilde{K}} \in L^2_{\rho_\mathcal{X}}$, where $L_{\tilde{K}}$ is the integral operator associated with $\tilde{K}(u,v) = \int_\mathcal{X} K(u,x)K(v,x)d\rho_\mathcal{X}, (u,v) \in \mathcal{X}^2$ (see [7]).

Now we state our main results on the convergence rate.

Theorem 2 *Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$. Assume that f_ρ can be approximated with exponent $0 < \beta \le 1$ in $\mathcal{H}$ and $K \in C^s(\mathcal{X} \times \mathcal{X})$ for some $s > 0$. Choose $m \le n^{\frac{1}{2+p}}$ and $\lambda = m^{-\theta}$ for some $\theta > 0$. For any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le \tilde{c}_2 \log^2(8/\delta) m^{-\gamma},$$

where constant $\tilde{c}_2$ is independent of m, δ, and

$$\gamma = \min\Big\{2 - \frac{p\theta}{2+p}, 2 + \beta\theta - 2\theta, \beta\theta, 1 + \beta\theta - \theta\Big\}.$$

Theorem 2 states the polynomial convergence rate of GNKR and indicates its dependence on the subsampling size m as $n \geq m^{2+p}$. Similar observation also can be found in Theorem 2 [16] for Nyström KRR, where the fast learning rate also is relied on the grow of m under fixed hypothesis space complexity. However, even we do not consider the complexity of hypothesis space, the increase of m will add the computation complexity. Hence, a suitable size of m is a trade off between the approximation performance and the computation complexity. When $p \in (0, 2)$, $m = n^{\frac{1}{2+p}}$ means that m can be chosen between $n^{\frac{1}{4}}$ and $\frac{1}{2}$ under the conditions in Theorem 4. In particular, the fast convergence rate $O(m^{-1})$ can be obtained as $K \in C^{\infty}$, $\theta \to 1$, and $\beta \to 1$.

The most related works with Theorems 1 and 2 are presented in [16, 24], where learning bounds are established for Nyström KRR. Compared with the previous results, the features of this paper can be summarized as below.

- *Learning model.* This paper considered Nyström regression with data dependent hypothesis space and coefficient regularization, which can employ general kernel including the indefinite kernel and nonsymmetric kernel. However, the previous analysis just focuses on the positive semi-definite kernel and the regularizer in RKHS. For a fixed design KRR, the fast convergence $O(m^{-1})$ in [24] depends on the eigenvalue condition of kernel matrix. Differently from [24], our result relies on the Lipschitz continuity of kernel and the approximation condition $D(\lambda)$ for the statistical learning setting.
- *Analysis technique.* The previous analysis in [16, 24] utilizes the theoretical techniques for operator approximation and matrix decomposition, which depends heavily on the symmetric positive semi-definite kernel. For GNKR (4), the previous analysis is not valid directly since the kernel is not necessary to satisfy the positive semi-definite or symmetric condition. The flexibility on kernel and the adaptivity on hypothesis space induce the additional difficulty on error analysis. Fortunately, the error analysis is obtained by incorporating the error decomposition ideas in [7] and the concentration estimate techniques in [18, 19]. An interesting future work is to establish the optimal bound of GNKR to extend Theorem 2 in [16] to the general setting.

For the proofs of Theorem 1 and 2, the key idea is using $\hat{g}_\lambda$ as the stepping stone function to bridge $f_{\mathbf{z}}$ and g_λ. Additionally, the connection between $g_\lambda = L_K f_\lambda$ and f_ρ has been well studied in learning theory. Hence, the proofs in Appendix follow from the approximation decomposition.

In remainder of this section, we present a simple analysis for column-norm subsampling.

Given the full samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$ and sampling number m, the key of subsampling is to select a subset of $\mathbf{z}$ with strong inference ability. In other words, we should select the subset with small divergence with the full sample estimator. Following this idea, the optimal subsampling criterion is studied in [28, 22] for the linear regression. Given $\mathbf{z} = \{z_i\}_{i=1}^n$ and K_{nn}, we introduce the objective function $S(p) := S(p_1, ..., p_n) = \sum_{i=1}^n \frac{1-L_{ii}}{p_i} \|K_i\|_2^2$ by extending (16) in [28] to the kernel-based setting. Here $\{p_i\}_{i=1}^n$ are the sampling probabilities with respect to $\{x_i\}_{i=1}^n$ and $L_{ii} = (K_{nn}(K_{nn}^T K_{nn} + \lambda n^2 I_n)^{-1} K_{nn}^T)_{ii}, i \in \{1, ..., n\}$ are basic leverage values obtained from (2). For the fixed design setting, assume that $y_i = K_i^T \alpha_0 + \varepsilon_i, i = 1, ..., n, \alpha_0 \in \mathbb{R}^n$, where $\{\varepsilon_i\}_{i=1}^n$ are drawn identically and independently from $\mathcal{N}(0, \sigma^2)$. Then, for $\lambda = 0$, $\min\limits_p S(p_1, ..., p_n)$ can be transformed as $\min\limits_p Etr((K_{nn})^T (\text{diag}(p))^{-1} K_{nn})$, which is related with the A-optimality or A-criterion for subset selection in [22].

When $L_{ii} \to 0$ for any $i \in \{1, ..., n\}$, we can get the following sampling probabilities.

Theorem 3 *When $h_{ii} = o(1)$ for $1 \leq i \leq n$, the minimizer of $S(p_1, ..., p_n)$ can be approximated by*

$$p_i = \frac{\|K_i\|_2}{\sum_{i=1}^n \|K_i\|_2}, \quad i \in \{1, ..., n\}.$$

Usually, the leverage values are computed by fast approximation algorithms [1, 16] since L_{ii} involves the inverse matrix. Different from the leverage values, the sampling probabilities in Theorem 3 can be computed directly, which just involves the ℓ_2 column-norm of empirical matrix.

Table 1: Average RMSE of GNKR with Gaussian(G)/Epanechnikov(E) kernel under different sampling strategies and sampling size. US:=Uniform subsampling, CS: Column-norm subsampling.

Function	Algorithm	$\sharp 300$	$\sharp 400$	$\sharp 500$	$\sharp 600$	$\sharp 700$	$\sharp 800$	$\sharp 900$	$\sharp 1000$
$f_1(x) = x \sin x$	G-GNKR-US	**0.03412**	0.03145	0.02986	0.02919	0.02897	0.02906	0.02896	0.02908
$x \in [0, 2\pi]$	G-GNKR-CS	0.03420	**0.03086**	**0.02954**	**0.02911**	**0.02890**	**0.02878**	**0.02891**	**0.02889**
	E-GNKR-US	0.10159	0.09653	0.09081	0.08718	0.08515	0.08278	0.08198	0.08024
	E-GNKR-CS	0.09941	0.09414	0.08908	0.08631	0.08450	0.08237	0.08118	0.07898
$f_2(x) = \frac{\sin x}{x}$	G-GNKR-US	**0.03442**	0.03434	0.03418	0.03409	0.03404	0.03400	0.03398	0.03395
$x \in [-2\pi, 2\pi]$	G-GNKR-CS	0.03444	**0.03423**	**0.03419**	**0.03408**	**0.03397**	**0.03397**	0.03396	0.03389
	E-GNKR-US	0.04786	0.04191	0.04073	0.03692	0.03582	0.03493	0.03470	0.03440
	E-GNKR-CS	0.04607	0.03865	0.03709	0.03573	0.03510	0.03441	**0.03316**	**0.03383**
$f_3(x) = sign(x)$	G-GNKR-US	0.29236	0.29102	0.29009	0.28908	0.28867	0.28839	0.28755	0.28742
$x \in [-3, 3]$	G-GNKR-CS	0.29319	0.29071	0.28983	0.28975	0.28903	0.28833	0.28797	0.28768
	E-GNKR-US	0.16170	0.15822	0.15537	**0.15188**	0.15086	0.14889	0.14730	0.14726
	E-GNKR-CS	**0.16500**	**0.15579**	**0.15205**	0.15201	**0.14949**	**0.14698**	**0.14597**	**0.14566**
$f_4(x) = \cos(e^x) + \frac{\sin x}{x}$	G-GNKR-US	0.34916	0.35158	0.35155	0.35148	0.35156	0.35140	0.35136	0.35139
$x \in [-2, 4]$	G-GNKR-CS	0.34909	0.35171	0.35168	0.35133	0.35153	0.35145	0.35141	0.35138
	E-GNKR-US	0.22298	0.21012	0.20265	0.19977	0.19414	0.19126	0.18916	**0.18560**
	E-GNKR-CS	**0.21624**	**0.20783**	**0.20024**	**0.19698**	**0.19260**	**0.18996**	**0.18702**	0.18662

5 Experimental Analysis

Since kernel regression with different types of regularization has been well studied in [7, 20, 21], this section just presents the empirical evaluation of GNKR to illustrate the roles of sampling strategy and kernel function. Gaussian kernel $K_G(x,t) = \exp\big\{-\frac{\|x-t\|_2^2}{2\sigma^2}\big\}$ is used for simulated data and real data. Epanechnikov kernel $K_E(x,t) = \big(1 - \frac{\|x-t\|_2^2}{2\sigma^2}\big)_+$ is used in the simulated experiment. Here, σ denotes the scale parameter selected form $[10^{-5} : 10 : 10^4]$. Following the discussion on parameter selection in [16], we select the regularization parameter of GNKR from $[10^{-15} : 10 : 10^{-3}]$. The best results are reported according to the measure of *Root Mean Squared Error* (RMSE).

5.1 Experiments on synthetic data

Following the empirical studies in [20, 21], we design simulation experiments on $f_1(x) = x \sin x, x \in [0, 2\pi]$, $f_2(x) = \frac{\sin x}{x}, x \in [-2\pi, 2\pi]$, $f_3(x) = sign(x)$, $x \in [-3, 3]$, and $f_4(x) = \cos(e^x) + \frac{\sin x}{x}$, $x \in [-2, 4]$. The function f_i is considered as the truly regression function for $1 \le i \le 4$. Note that f_1, f_2 are smooth, f_3 is not continuous, and f_4 embraces a highly oscillatory part. First, we select 10000 points randomly from the preset interval and generate the dependent variable y according to the corresponding function. Then we divided these data into two parts with equal size. we chose one part as the training samples and the other is regarded as testing samples. For the training samples, the output y is contaminated by Gaussian noise $\mathcal{N}(0, 1)$. For each function and each kernel, we run the experiment 20 times. The average RMSE is shown in Table 1. The results indicate that the column norm subsampling can achieve the satisfactory performance. In particular, GNKR with the indefinite Epanechnikov kernel has better performance than Gaussian kernel for the noncontinuous function f_3 and the non-flat function f_4. This observation is consistent with the empirical result in [21].

5.2 Experiments on real data

In order to better evaluate the empirical performance, four data sets are used in our study including the *Wine Quality, CASP, Year Prediction* datasets (http://archive.ics.uci.edu/ml/) and the *census-house* dataset (http://www.cs.toronto.edu/ delve/data/census-house/desc.html). The detailed information about the data sets are showed in Table 2. Firstly, each data set is standardized by subtracting its mean and dividing its standard deviation. Then, each input vector is unitized. For *CASP* and *Year Prediction*, 20000 samples are drawn randomly from data sets, where half is used for training and the rest is for testing. For other datasets, we random select part samples to training and use the rest part as test set. Table 3 reports the average RMSE over ten trials.

Table 3 shows the performance of two sampling strategies. For *CASP*, and *Year Prediction*, we can see that GNKR with 100 selected samples can achieve the satisfactory performance, which reduce the computation complexity of (2) efficiently. Additionally, the competitive performance of GNKR with Epanechnikov kernel is demonstrated via the experimental results on the four data sets. These empirical examples support the effectiveness of the proposed method.

Table 2: Statistics of data sets

Dataset	**#Features**	**#Instances**	**#Train**	**#Test**	**Dataset**	**#Feature**	**#Instance**	**#Train**	**#Test**
Wine Quality	12	4898	2000	2898	CASP	9	45730	10000	10000
Year Prediction	90	515345	10000	10000	census-house	139	22784	12000	10784

Table 3: Average RMSE ($\times 10^{-3}$) with Gaussian(G)/Epanechnikov(E) kernel under different sampling levels and strategies. US:=Uniform subsampling, CS: Column-norm subsampling.

Function	Algorithm	$\sharp 50$	$\sharp 100$	$\sharp 200$	$\sharp 400$	$\sharp 600$	$\sharp 800$	$\sharp 1000$
Wine Quality	G-GNKR-US	14.567	14.438	14.382	14.292	14.189	14.103	13.936
	G-GNKR-CS	14.563	14.432	14.394	14.225	14.138	14.014	13.936
	E-GNKR-US	13.990	13.928	13.807	13.636	13.473	13.381	**13.217**
	E-GNKR-CS	**13.969**	**13.899**	**13.798**	**13.601**	**13.445**	**13.362**	13.239
CASP	G-GNKR-US	9.275	9.238	9.205	9.222	9.204	9.207	9.205
	G-GNKR-CS	9.220	9.196	9.205	9.193	9.198	9.199	9.198
	E-GNKR-US	4.282	**4.196**	4.213	**4.153**	4.181	4.174	4.180
	E-GNKR-CS	**4.206**	4.249	**4.206**	4.182	**4.172**	**4.165**	**4.118**
Year Prediction	G-GNKR-US	8.806	8.802	8.798	8.795	8.792	8.790	8.782
	G-GNKR-CS	8.806	8.801	8.798	8.793	8.792	8.789	8.781
	E-GNKR-US	7.013	**6.842**	**6.739**	**6.700**	6.676	6.671	**6.637**
	E-GNKR-CS	**7.006**	6.861	6.804	6.705	6.697	**6.663**	6.662
census-house	G-GNKR-US	111.084	111.083	111.082	111.079	111.077	111.074	111.071
	G-GNKR-CS	111.083	111.080	111.080	111.079	111.075	111.071	111.068
	E-GNKR-US	102.731	99.535	99.698	99.718	99.715	**99.714**	99.713
	E-GNKR-CS	**102.703**	**99.528**	**99.697**	**99.716**	**99.714**	99.714	**99.712**

6 Conclusion

This paper focuses on the learning theory analysis of Nyström kernel regression. One key difference with the previous related work is that GNKR uses general continuous kernel function and ℓ_2 coefficient regularization. The stepping-stone functions are constructed to overcome the analysis difficulty induced by the difference. The learning bound with fast convergence is derived under mild conditions and empirical analysis is provided to verify our theoretical analysis.

Acknowledgments

This work was partially supported by U.S. NSF-IIS 1302675, NSF-IIS 1344152, NSF-DBI 1356628, NSF-IIS 1619308, NSF-IIS 1633753, NIH AG049371, and by National Natural Science Foundation of China (NSFC) 11671161. We thank the anonymous NIPS reviewers for insightful comments.

References

[1] A. Alaoui and M. W. Mahoney. Fast randomized kernel methods with statistical guarantees. In *NIPS*, pp. 775–783, 2015.

[2] F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *COLT*, 2013.

[3] P. Drineas, R. Kannan, and M.W. Mahoney. Fast Monte Carlo algorithms for matrices I: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, pp. 158–183, 2006.

[4] P. Drineas and M.W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.*, 6: 2153–2175, 2005.

[5] P. Drineas, M. Magdon-Ismail, M.W. Mahoney, and D.P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, 13: 3475–3506, 2012.

[6] M. Eberts and I. Steinwart. Optimal learning rates for least squares SVMs using Gaussian kernels. In *NIPS*, pp. 1539–1547, 2011.

[7] Y. Feng, S. Lv, H. Huang, and J. Suykens. Kernelized elastic net regularization: generalization bounds and sparse recovery. *Neural Computat.*, 28: 1–38, 2016.

[8] A. Gittens and M.W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. In *ICML*, pp. 567–575, 2013

[9] C.J. Hsieh, S. Si, and I.S. Dhillon. Fast prediction for large scale kernel machines. In *NIPS*, pp. 3689–3697, 2014.

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

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

[12] W. Lim, M. Kim, H. Park, and K. Jung. Double Nyström method: An efficient and accurate Nyström scheme for large-scale data sets. In *ICML*, pp. 1367–1375, 2015.

[13] C. Liu. Gabor-based kernel PCA with fractional power polynomial models for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26: 572–581, 2004.

[14] E. Pekalska and B. Haasdonk. Kernel discriminant analysis with positive definite and indefinite kernels. *IEEE Trans. Pattern. Anal. Mach. Intell.*,31: 1017–1032, 2009.

[15] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, pp. 1177–1184, 2007.

[16] A. Rudi, R. Camoriano, R. Rosasco. Less is more: Nyström computation regularization. In *NIPS*, 1657–1665, 2015.

[17] B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* . MIT Press, 2001.

[18] L. Shi, Y. Feng, and D.X. Zhou. Concentration estimates for learning with ℓ_1-regularizer and data dependent hypothesis spaces. *Appl. Comput. Harmon. Anal.*, 31(2): 286–302, 2011.

[19] L. Shi. Learning theory estimates for coefficient-based regularized regression. *Appl. Comput. Harmon. Anal.*, 34(2): 252–265, 2013.

[20] H. Sun and Q. Wu. Least square regression with indefinite kernels and coefficient regularization. *Appl. Comput. Harmon. Anal.*, 30(1): 96–109, 2011.

[21] H. Sun and Q. Wu. Sparse representation in kernel machines. *IEEE Trans. Neural Netw. Learning Syst.*, 26(10): 2576–2582, 2015.

[22] Y. Wang and A. Singh. Minimax subsampling for estimation and prediction in low-dimensional linear regression. *arXiv*, 2016 (https://arxiv.org/pdf/1601.02068v2.pdf).

[23] C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *NIPS*, pp. 682–688, 2001.

[24] T. Yang, Y.F. Li, M. Mahdavi, R. Jin, and Z.H. Zhou. Nyström method vs random Fourier features: A theoretical and empirical comparison. In *NIPS*, 2012, pp. 485–493.

[25] Y. Yang, M. Pilanci and M. J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *arxiv:1501.06195*, 2015.(http://arxiv.org/abs/1501.06195).

[26] Y. Ying, C. Campbell, and M. Girolami. Analysis of SVM with indefinite kernels. In *NIPS*, pp. 2205–2213, 2009.

[27] K. Zhang, I.W. Tsang, and J.T. Kwok. Improved Nyström low-rank approximation and error analysis. In *ICML*, pp. 1232–1239, 2008.

[28] R. Zhu, P. Ma, M.W. Mahoney, and B. Yu. Optimal subsampling approaches for large sample linear regression. *arXiv:1509.05111*, 2015 (http://arxiv.org/abs/1509.05111).