J. Math. Anal. Appl. 466 (2018) 676–689

Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa

# Distributed regression learning with coefficient regularization

Mengjuan Pang, Hongwei Sun *

*School of Mathematical Science, University of Jinan, Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan 250022, People's Republic of China*

ARTICLE INFO

*Article history:*
Received 18 April 2017
Available online 7 June 2018
Submitted by U. Stadtmueller

*Keywords:*
Distributed learning
Error bound
Learning rates
Integral operator
Coefficient regularization

ABSTRACT

We study distributed regression learning with coefficient regularization scheme in a reproducing kernel Hilbert space (RKHS). The algorithm randomly partitions the sample set $\{z_i\}_{i=1}^N$ into $m$ disjoint sample subsets of equal size, applies the coefficient regularization scheme to each sample subset to produce an output function, and averages the individual output functions to get the final global estimator. We deduce the error bound in expectation in the $L^2$-metric and prove the asymptotic convergence for this distributed coefficient regularization learning. Satisfactory learning rates are then derived under a standard regularity condition on the regression function, which reveals an interesting phenomenon that when $m \le N^s$ and $s$ is small enough, this distributed learning has the same convergence rate compared with the algorithm processing the whole data in one single machine.

## 1. Introduction

In the modern information society, the analysis of big data has become a bottleneck which restricts the development of information technology. Due to the closely relation between the research of regularization algorithm and dimensionality reduction, data mining, image processing, the research and breakthrough of regularization algorithm with big data has great values both on theory and practice.

Kernel methods for single data set have been deeply studied in the past. Various kinds of the kernel based learning algorithms have been discussed thoroughly, include kernel principal component analysis [3,12], regularized kernel network [6,25], online learning based on kernel [22,29,30], gradient learning [17,28], MEE regression [7,11] and so on. But, for big data analysis, these algorithms for the single data set may no longer be applied or available, or at least no longer optimal.

Kernel regression learning algorithm for single data set is to learn all samples $D = \{(x_i, y_i)\}_{i=1}^N$ for one-time. Because that the algorithm is related to the inversion and the characteristic decomposition of kernel matrix $\mathbf{K} = [K(x_i, x_j)]_{i,j=1}^N$, the complexity of this kind algorithms is as high as $\mathcal{O}(N^3)$. Thus

* Corresponding author.
*E-mail address:* ss_sunhw@ujn.edu.cn (H. Sun).

https://doi.org/10.1016/j.jmaa.2018.06.013

researching on low complexity learning algorithms for big data set has become one of the challenges of learning theory. The main methods of reducing the complexity include the low rank approximation of the kernel matrix, such as low rank approximation of the kernel matrix in principal component analysis [1,19], incomplete Cholesky decomposition [8] and so on; the early-stopping method [18,28] in iterative optimization calculation process; and by the idea of greedy algorithms.

In recent years, distributed learning has been widely concerned, [5,13,31] for distributed kernel ridge regression, [16] for distributed conditional entropy models, [4] for distributed local averages, [2,9] for distributed spectral kernel algorithms, [20] for distributed learning with spline and [10] for distributed bias correction regularization network. Its main idea is to divide the sample set into $m$ disjoint subsets, to learn each subset to get prediction function, and then take the weighted average of the $m$ prediction function to get the result of the learning. This method can reduce the complexity of the algorithm to $\mathcal{O}(N^3/m^2)$. In literatures [5,13,31], researchers deduced the error bound and asymptotic convergence rate of the distributed least square regularization algorithms. Especially, Lin–Guo–Zhou [13] further improved the analysis of the distributed kernel ridge regression algorithm in literature [31], and proved that the algorithm can achieve the optimal convergence rate under some conditions. Obviously, the research of distributed regression with kernel method is just beginning, and there are a lot of theoretical and practical issues worth considering.

In view of the importance of distributed regression to big data analysis, this paper will focus on the research of distributed learning with coefficient based regularization. The coefficient regularization was first introduced by Vapnik [26] to design linear programming support vector machines. This regularization network has attracted wide concern, since one can freely choose the regularizer for different purposes, and use indefinite kernels if one has some a priori knowledge and wants to fit the data in certain trend [24]. Moreover, $\ell^1$ regularization can lead to sparse solution [25]. These together with the success of indefinite kernels in some real applications [14,15] motivated the research in the mathematical foundation of coefficient regularization [21,24,27].

The main purpose of this paper is to deduce the error bound and asymptotic convergence rates of this distributed coefficient based regularized learning. Our conclusions show that this distributed learning scheme can not only decrease the complexity of algorithms evidently, but also has good asymptotic convergency, especially it has almost the same learning rates compared with the algorithms for single data set when $m$ is small.

## 2. Assumptions and main results

In regression learning scheme, we usually assume that $X$ is a metric space and $Y = \mathbb{R}$, $Z = X \times Y$ is a probability space with a Borel probability distribution $\rho$. $\rho$ can be decomposed into the condition probability distribution on $Y$ and the marginal probability distribution on $X$, i.e., $\rho = \rho(\cdot|x) \times \rho_X$. Regression function $f_\rho$ reveals some functional relation between input data $x$ and output data $y$, which is defined as

$$f_\rho(x) = \int\limits_Y y\, d\rho(y|x).$$

The goal of regression learning is to learn or approximate $f_\rho$ by samples $D = \{(x_i, y_i)\}_{i=1}^N$ independently drawn according to the distribution $\rho$.

Let $K : X\times X \to \mathbb{R}$ be a continuous, symmetric and positive semi-definite function, called a Mercer kernel. The reproducing kernel Hilbert space $H_K$ associated with a Mercer kernel $K$ is the completion of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the inner product $\langle\cdot,\cdot\rangle_K$ given by $\langle K_x, K_y\rangle_K = K(x, y)$. In the sequel, the Mercer kernel $K$ is supposed to be uniformly bounded, which is equal to

$$\kappa \doteq \sup_{x\in X} \sqrt{K(x,x)} < \infty.$$