Error Analysis for Vector-Valued Regularized Least-Squares Algorithm

Yong Liu and Shizhong Liao
School of Computer Science and Technology
Tianjin University, Tianjin 300072, P. R. China
{yongliu,szliao}@tju.edu.cn

Abstract— *Vector-valued regularized least-squares algorithm (RLS) on vector-valued reproducing kernel Hilbert space (RKHS) has recently received increasing interest in various machine learning problems such as multi-task learning and multi-view learning, but error analysis of the vector-valued RLS is still widely unknown. In this paper, we derive an error bound of the vector-valued RLS, which consists of two parts: sample error bound and approximation error bound. We first present the sample error bound through the concentration inequalities of function-valued random variables. Under a suitable assumption of the approximation error, we propose the total error bound with the derived sample error bound. Furthermore, together with a Ysybakov function, we also present an error bound of the multi-class classification problem in terms of the error bound derived for the vector-valued RLS.*

Keywords: Consistency, vector-valued regularized least-squares algorithm, multi-class classification, multi-view learning

1. Introduction

The regularized least-squares algorithm (RLS) on a reproducing kernel Hilbert space (RKHS) of real-valued functions (i.e., when the output space is equal to $\mathbb{R}$) has been extensively studied in the literature [1]–[6]. In [1], a covering number technique is used to obtain the error bounds expressed in terms of suitable complexity measures of the regression function. In [2], the covering techniques are replaced by estimates of integral operators through concentration inequalities of vector-valued random variables. In [3], entropy methods are used to establish the upper bounds. In [4], [6], the eigenvalues of the integral operator are used as a complexity measure for error analysis.

Following the development of multi-task learning and multi-view learning methods, the vector-valued RLS on a vector-valued RKHS (i.e., when the output space is equal to $\mathbb{R}^d$) has recently attracted considerable attention in the machine learning community. A study of vector-valued learning with kernel methods is started in [7] where the vector-valued RKHS is adopted and the representer theorem for Tikhonov regularization is generalized to the vector-valued setting. In [8], [9], they derive conditions which ensure that the operator-valued kernel is universal (which means that on every compact subset of the input space, every continuous function with values in output space can be uniformly approximated by sections of the kernel). Instead of studying operator-valued kernels and their corresponding RKHS from the perspective of extending Aronszajn's pioneering work [10] to the vector or function valued, Kadri et al. [11] target at advancing the understanding of feature spaces associated with operator-valued kernels. In [6], they study the asymptotic performances of the vector-valued RLS for a suitable class of priors and a assumption that the regression function belong to the RKHS.

Although the vector-valued RLS has recently attracted considerable attention, its error analysis is still widely unknown. In this paper, based on the fact that scalar positive defined kernels can be extended to cope with vector-functions using operator-valued positive kernels, we extend the results of error bounds of the scalar RLS to the vector-valued RLS. We first present finite sample error bounds for the vector-valued RLS both in vector-valued RKHS norm and square integrable norm through the concentration inequalities of function-valued random variables. Then, with the derived sample error bounds, we propose total error bounds under a suitable assumption of approximation error. Furthermore, we consider to use the vector-valued RLS for multi-class classification. Together with a Ysybakov function, we apply the error bounds derived for the vector-valued RLS regression to the multi-class classification problem for error analysis.

The rest of the paper is organized as follows. In Section 2 we consider the vector-valued learning and present the setup of the problem, as well as the basic notions behind the theory of vector-valued RKHS. In Section 3 we present the error bounds for vector-valued RLS regression and discuss their consequences. In Section 4 we generalize the above results to multi-class classification problem. We end in Section 5 with conclusion.

2. Preliminaries and Notations

The problem of supervise learning amounts to inferring an unknown functional relation given a finite training set of examples $\boldsymbol{z} = \{(x_i, y_i)\}_{i=1}^n$. More precisely, the training examples are assumed to be identically and independently distributed according to a fixed, but unknown probability

measure $\rho(x,y)$ on $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$, where usually $\mathcal{Y}\subseteq\mathbb{R}$. Here we are interested in vector-valued learning where $\mathcal{Y}\subseteq\mathbb{R}^d$. A learning algorithm is a map from a training set $\boldsymbol{z}$ to an estimator $f_{\boldsymbol{z}}:\mathcal{X}\to\mathcal{Y}$.

A good estimator should generalize to future examples, this translates into the requirement of having small expected risk

$$\mathcal{E}(f)=\int_{\mathcal{X}\times\mathcal{Y}}\|y-f(x)\|_d^2 d\rho(x,y),$$

where $\|\cdot\|_d$ denotes the euclidean norm in $\mathbb{R}^d$. The minimizer of the expected risk over the space of all the measurable $\mathcal{Y}$-valued functions on $\mathcal{X}$ is the $regression\ function$

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

where $\rho(y|x)$ is the conditional distribution at x induced by ρ. Thus the quality of an estimator $f_{\boldsymbol{z}}$ can be assessed by $\|f_{\boldsymbol{z}}-f_\rho\|_\rho$, where

$$\|f_{\boldsymbol{z}}-f_\rho\|_\rho=\left\{\int_{\mathcal{X}}\|f_{\boldsymbol{z}}(x)-f_\rho(x)\|_d^2 d\rho_{\mathcal{X}}(x)\right\}^{1/2},$$

$\rho_{\mathcal{X}}$ is the marginal distribution of ρ on $\mathcal{X}$.

2.1 Vector-Valued RKHS

In the following we will introduce the vector-valued RKHS. You may refer to [7] for further details and references.

Let $\mathcal{Y}^{\mathcal{X}}$ denote the vector space of all functions $f:\mathcal{X}\to\mathcal{Y}$, $\mathcal{L}(\mathcal{Y})$ the Banach space of bounded linear operators on $\mathcal{Y}$. Note that for $\mathcal{Y}\subseteq\mathbb{R}^d$, the space $\mathcal{L}(\mathcal{Y})$ is the space of $d\times d$ matrices. A function

$$K:\mathcal{X}\times\mathcal{X}\to\mathcal{L}(\mathcal{Y})$$

is said to be an $operator\ valued\ positive\ definite\ kernel$ if for each pair $(x,y)\in\mathcal{X}\times\mathcal{X}$, $K(x,y)\in\mathcal{L}(\mathcal{Y})$ is a self-adjoint operator and

$$\sum_{i,j=1}^n\langle y_i,K(x_i,x_j)y_j\rangle_d\geq 0$$

for every finite set of examples $\{(x_i,y_i)\}_{i=1}^n\subset\mathcal{X}\times\mathcal{Y}$.

For each $x\in\mathcal{X}$ and $y\in\mathcal{Y}$, we form a function $K_xy=K(\cdot,x)y\in\mathcal{Y}^{\mathcal{X}}$ defined by

$$(K_xy)(t)=K(t,x)y\ \text{ for all } t\in\mathcal{X}.$$

Similarly to the scalar case, it can be shown that for any given operator valued kernel K, a unique RKHS $\mathcal{H}_K$ can be defined by considering the completion of the space

$$\operatorname{span}\left\{\sum_{i=1}^n K_{x_i}y_i\Big|\ x_i\in\mathcal{X},y_i\in\mathcal{Y}\right\}$$

with respect to the norm $\|\cdot\|_K$ induced by the inner product

$$\langle f,g\rangle_K=\sum_{i,j=1}^n\langle K(x_j,x_i)\beta_i,w_j\rangle_d,$$

for any

$$f,g\in\operatorname{span}\left\{\sum_{i=1}^n K_{x_i}y_i\Big|\ x_i\in\mathcal{X},y_i\in\mathcal{Y}\right\}$$

with $f=\sum_{i=1}^n K(\cdot,x_i)\beta_i$ and $g=\sum_{i=1}^n K(\cdot,x_i)w_i$. By definition, the kernel K has the following reproducing property, for all $y\in\mathcal{Y}$ and $x\in\mathcal{X}$,

$$\langle f(x),y\rangle_d=\langle f,K_xy\rangle_K\ \text{ for all }\ f\in\mathcal{H}_K. \tag{1}$$

Denote $\kappa=\sqrt{\sup_{x\in\mathcal{X}}\|K(x,x)\|}$. Then (1) implies that

$$\|f\|_\infty:=\sup_{x\in\mathcal{X}}\|f(x)\|_d\leq\kappa\|f\|_K \tag{2}$$

for all $f\in\mathcal{H}_K$.

In this paper, we assume that $\kappa<\infty$ and for some $D\geq 0$, $\|y\|_d\leq D$ almost surely, thus $\|f_\rho\|_\rho\leq D$.

2.2 Vector-Valued RLS Algorithm

In this subsection, we will introduce the vector-valued RLS on the vector-valued RKHS. In this framework the hypothesis space $\mathcal{H}_K$ is a given vector-valued RKHS induced by the operator valued positive definite kernel K, and for any $\lambda>0$, the vector-valued RLS estimator $f_{\boldsymbol{z},\lambda}$ is defined as the solution of the minimizing problem

$$f_{\boldsymbol{z},\lambda}=\operatorname*{arg\,min}_{f\in\mathcal{H}_K}\left\{\frac{1}{n}\sum_{i=1}^n\|f(x_i)-y_i\|_d^2+\lambda\|f\|_K^2\right\}. \tag{3}$$

We know from [7], [12] that the solution $f_{\boldsymbol{z},\lambda}$ uniquely exists, and is given by

$$f_{\boldsymbol{z},\lambda}=\left(\frac{1}{n}S_{\boldsymbol{x}}^*S_{\boldsymbol{x}}+\lambda I\right)^{-1}\frac{1}{n}S_{\boldsymbol{x}}^*\boldsymbol{y}, \tag{4}$$

where the operator $S_{\boldsymbol{x}}^*:\mathcal{Y}^n\to\mathcal{H}_K$ is given by

$$S_{\boldsymbol{x}}^*\boldsymbol{y}=S_{\boldsymbol{x}}^*(y_1,\ldots,y_n)=\sum_{i=1}^n K_{x_i}y_i,$$

and the operator $S_{\boldsymbol{x}}^*S_{\boldsymbol{x}}:\mathcal{H}_K\to\mathcal{H}_K$ is given by

$$S_{\boldsymbol{x}}^*S_{\boldsymbol{x}}f=\sum_{i=1}^n K_{x_i}f(x_i).$$

Our goal is to understand how $f_{\boldsymbol{z},\lambda}$ approximates f_ρ and how the decay of the regularization parameter $\lambda=\lambda(n)$ leads to convergence rates. For the scalar RLS, the rates for this approximation in $L^2_{\rho_X}$ ($\|f_{\boldsymbol{z},\lambda}-f_\rho\|_\rho$) have been considered in [1], [13]–[16], and the approximation in the space $\mathcal{H}_K$ ($\|f_{\boldsymbol{z},\lambda}-f_\rho\|_K$) has been shown in [2], [14]. In this paper, we extend the results of the scalar RLS to the vector-valued RLS. Furthermore, we generalize our results of the vector-valued RLS to multi-class classification problem for error analysis.

3. Error Bounds for Vector-Valued RLS Regression

A data-free limit of (3) is

$$f_\lambda = \arg\min_{f\in\mathcal{H}_K} \left\{ \|f - f_\rho\|_\rho^2 + \lambda \|f\|_K^2 \right\}. \tag{5}$$

By [1], we know that the solution of (5) is

$$f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho, \tag{6}$$

where I is identity operator and $L_K : L^2 \to \mathcal{H}_K$ is an integral operator defined by

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

We will deal with the error $\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K$ by dividing it into two parts $\|f_{\boldsymbol{z},\lambda} - f_\lambda\|_K$ and $\|f_\lambda - f_\rho\|_K$. The first term, $\|f_{\boldsymbol{z},\lambda} - f_\lambda\|_K$, is called the sample error which is made by approximating f_λ through a finite training set $\boldsymbol{z}$. The second term, $\|f_\lambda - f_\rho\|_K$, depends on the choice of $\mathcal{H}_K$ but is independent of sampling, which is called the approximation error.

Theorem 1: Let $\boldsymbol{z} = \{(x_i, y_i)\}_{i=1}^n$ be randomly drawn according to ρ, for all $0 < \delta < 1$, with confidence $1-\delta$,

$$\|f_{\boldsymbol{z},\lambda} - f_\lambda\|_K \le \frac{6\kappa D \log(2/\delta)}{\sqrt{n}\lambda}. \tag{7}$$

Proof: See in Appendix.A. ∎

For the scalar RLS, the error bounds $\|f_{\boldsymbol{z},\lambda} - f_\lambda\|_K$ have been given in [2], [14]. In [14], they show that with confidence $1-\delta$,

$$\|f_{z,\lambda} - f_\lambda\|_K \le \frac{c_1 \log(4/\delta)}{\sqrt{n}\lambda}\left(30 + \frac{c_2 a}{3\sqrt{n\lambda}}\right),$$

in [2], with confidence $1-\delta$,

$$\|f_{z,\lambda} - f_\lambda\|_K \le \frac{c_3 \log(2/\delta)}{\sqrt{n}\lambda},$$

where c_1, c_2 and c_3 are some constants.

To the best of our knowledge, the error bound $\|f_{\boldsymbol{z},\lambda} - f_\lambda\|_K$ for the vector-valued RLS on the vector-valued RKHS had never been given before. Our result fills this gap. By theorem 1, we find that the convergence rate of vector-valued RLS is $O(\frac{1}{\sqrt{n}\lambda})$ as the same as that of the scalar RLS in [2], [14].

Using Theorem 1, we will prove our total error estimate in the $\|\cdot\|_K$ norm.

Theorem 2: Let $\boldsymbol{z} = \{(x_i, y_i)\}_{i=1}^n$ be randomly drawn according to ρ, and assume the approximation error $\|f_\lambda - f_\rho\|_K$ satisfies

$$\|f_\lambda - f_\rho\|_K \le c\lambda^\beta,$$

where $c > 0$ and $\beta > 0$. Then, for any $0 < \delta < 1$, with confidence $1-\delta$,

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K \le 2\log(2/\delta)\left\{\frac{3\kappa D}{\sqrt{n}\lambda} + c\lambda^\beta\right\}. \tag{8}$$

Setting $\lambda = (2\kappa D)^{\frac{1}{\beta+1}} \left(\frac{1}{n}\right)^{\frac{1}{2(\beta+1)}}$, we have

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K \le 4c\log(2/\delta)(2kD)^{\frac{\beta}{\beta+1}} \left(\frac{1}{n}\right)^{\frac{\beta}{2(\beta+1)}}. \tag{9}$$

Proof: Note that

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K \le \|f_{\boldsymbol{z},\lambda} - f_\lambda\|_K + \|f_\lambda - f_\rho\|_K.$$

By Theorem 1 and the assumption $\|f_\lambda - f_\rho\|_K \le c\lambda^\beta$, with confidence $1-\delta$,

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K \le 2\log(2/\delta)\frac{3\kappa D}{\sqrt{n}\lambda} + c\lambda^\beta.$$

Since $0 < \delta < 1$, we have $2\log(2/\delta) > 1$. Therefore,

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K \le 2\log(2/\delta)\left\{\frac{3\kappa D}{\sqrt{n}\lambda} + c\lambda^\beta\right\}.$$

Minimize the $\frac{3\kappa D}{\sqrt{n}\lambda} + c\lambda^\beta$ over $\lambda > 0$, and we obtain

$$\lambda = (2\kappa D)^{\frac{1}{\beta+1}} \left(\frac{1}{n}\right)^{\frac{1}{2(\beta+1)}}.$$

With this choice of λ, we can obtain Theorem 2. ∎

Remark 1: If f_ρ is in the range of L_K^r and $\frac{1}{2} < r \le 1$, the approximation error $\|f_\lambda - f_\rho\|_K \le \lambda^{r-\frac{1}{2}} \|L_K^{-1} f_\rho\|_\rho$, which implies that the assumption $\|f_\lambda - f_\rho\|_K \le c\lambda^\beta$ in Theorem 2 is reasonable.

For the scalar RLS, if f_ρ is in the range of L_K, [14] show that

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K \le c_4 \left(\frac{(\log(4/\delta))^2}{n}\right)^{\frac{1}{6}}.$$

In [2], they improve the above result and obtain that

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K \le c_5 \log(2/\delta)\left(\frac{1}{n}\right)^{\frac{1}{6}}.$$

For the vector-valued RLS, if we also assume that f_ρ is in the range of L_K as the same as that of scalar RLS in [2], [14], then

$$\|f_\lambda - f_\rho\|_K \le \lambda^{\frac{1}{2}} \|L_K^{-1} f_\rho\|_\rho.$$

Therefore, the β in theorem 2 is equal $\frac{1}{2}$. In this case, by Theorem 2, we have

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K \le c_6 \log(2/\delta)\left(\frac{1}{n}\right)^{\frac{1}{6}},$$

the convergence rate of the vector-valued RLS is equal to that of the scalar RLS. When we consider the extreme case, that is, $\beta \to \infty$, the convergence rate is $O(\frac{1}{\sqrt{n}})$.

Using Theorem 2 and $\|f_{\boldsymbol{z},\lambda} - f_\rho\|_\rho \le \kappa \|f_{\boldsymbol{z},\lambda} - f_\rho\|_K$, it is easy to obtain the following corollary.

Corollary 1: Let $\boldsymbol{z} = \{(x_i, y_i)\}_{i=1}^n$ be randomly drawn according to ρ, and assume $\|f_\lambda - f_\rho\|_K \le c\lambda^\beta$, where $c > 0$ and $\beta > 0$. Then, for any $0 < \delta < 1$, with confidence $1-\delta$,

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|_\rho \le 2\kappa\log(2/\delta)\left\{\frac{3\kappa D}{\sqrt{n}\lambda} + c\lambda^\beta\right\}. \tag{10}$$

Setting $\lambda = (2\kappa D)^{\frac{1}{\beta+1}} \left(\frac{1}{n}\right)^{\frac{1}{2(\beta+1)}}$, we have

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|_\rho \leq 4\kappa c \log(2/\delta)(2kD)^{\frac{\beta}{\beta+1}} \left(\frac{1}{n}\right)^{\frac{\beta}{2(\beta+1)}}. \quad (11)$$

In [6], under the assumptions that $\rho \in \mathcal{P}(b,c)$ (see Definition 1 in [6]), $f_\rho \in \mathcal{H}_K$ and the eigenvalues t_n of the integral operator L_K satisfy

$$\alpha \leq n^b t_n \leq \beta,$$

they obtain that

$$\lim_{\tau\to\infty} \limsup_{n\to\infty} \sup_{\rho\in P(b,c)} \mathbb{P}_{\boldsymbol{z}\sim\rho^n} \left[\|f_{\boldsymbol{z},\lambda} - f_\rho\|_\rho > \tau \left(\frac{1}{n}\right)^{\frac{bc}{bc+1}}\right] = 0,$$

where $b < \infty$ and $1 \leq c \leq 2$. The above assumptions may be too strong and therefore may not be satisfied in general cases. In this paper, we only assume that

$$\|f_\lambda - f_\rho\|_K \leq c\lambda^\beta,$$

and if $\beta \geq \frac{2bc}{1-bc}$ and $bc < 1$, by theorem 2, our result yields faster convergence rate. In addition, our proof is much simpler than theirs.

4. Application to Multi-Class Classification

In multi-class classification the examples belong to one of d ($d > 2$) classes. Let $\rho(k|x)$ be the conditional probability for each class, $k = 1, \ldots, d$. A classifier is a function $c : \mathcal{X} \to \{1, 2, \ldots, d\}$, assigning each input point to one of the d classes.

The classification performance can be measured via the misclassification probability

$$R(c) = \mathbb{P}[c(x) \neq y].$$

It is easy to check that the minimizer of the misclassification probability is given by the Bayes rule, defined as

$$b(x) = \arg\max_{k\in\{1,\ldots,d\}} \rho(k|x).$$

In order to use the vector-valued RLS for multi-class classification, we define a coding, that is, a one-to-one map

$$M : \{1, 2, \ldots, d\} \to \mathcal{Y}$$

where $\mathcal{Y} = \{l_1, \ldots, l_d\} \subset \mathbb{R}^d$. In this paper, we define the coding as $l_1 = (1, -1, -1, \ldots, -1)$, $l_2 = (-1, 1, -1, \ldots, -1)$,..., $l_d = (-1, -1, -1, \ldots, 1)$.

We use superscripts to index vector components, so that the squared loss can be written as

$$\|l - f(x)\|_d^2 = \sum_{j=1}^d (l^j - f^j(x))^2.$$

Since the coding is one-to-one, the probability for each coding vector l_k is given by $\rho(k|x)$. The expected risk

$$\mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \|y - f(x)\|_d^2 d\rho(y|x) d\rho_{\mathcal{X}}(x) = \int_{\mathcal{X}} \sum_{k=1}^d \|l_k - f(x)\|_d^2 \rho(k|x) d\rho_{\mathcal{X}}(x),$$

is minimized by the regression function f_ρ, which is expressed as

$$f_\rho(x) = (f_\rho^1(x), \ldots, f_\rho^d(x)) = \int_{\mathcal{Y}} y d\rho(y|x) = \sum_{k=1}^d l_k \rho(k|x).$$

We can write the i-th component of the regression function as

$$f_\rho^i(x) = \sum_{k=1}^d l_k^i \rho(k|x) = \sum_{k=1, k\neq i}^d -\rho(k|x) + \rho(i|x) = \sum_{k=1}^d -\rho(k|x) + \rho(i|x) + \rho(i|x) = 2\rho(i|x) - 1,$$

since $\sum_{k=1}^d \rho(k|x) = 1$. By the definition of the Bayes rule, we have

$$b(x) = \arg\max_{j\in\{1,\ldots,d\}} f_\rho^j(x). \quad (12)$$

The above calculation is simple, but shows us the useful facts: First, the vector-valued RLS algorithm approximating the regression function can be used to learn the Bayes rule for a multi-class problem. Second, once we obtained an estimator for the regression function, Equation (12) shows that the natural way to define a classification rule is to take the argmax of the components of the estimator.

Based on the above idea, the vector-valued RLS estimator $f_{\boldsymbol{z},\lambda}$ for multi-class is defined as the solution of the minimization problem

$$f_{\boldsymbol{z},\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{\frac{1}{n}\sum_{i=1}^n \|f(x_i) - \bar{l}_i\|_d^2 + \lambda\|f\|_{\mathcal{H}_K}^2\right\}. \quad (13)$$

The classifier is given by

$$c(x) = \arg\max_{i\in\{1,\ldots,d\}} f_{\boldsymbol{z},\lambda}^i(x).$$

Instead of estimating the error

$$\mathbb{E}_\rho(I(c(x) \neq b(x))),$$

where $I(c(x) \neq b(x)) = 1$ if $c(x) \neq b(x)$, $I(c(x) \neq b(x)) = 0$ otherwise, in this paper, for applying the previously derived results derived for vector-valued RLS to multi-class classification problems for error analysis, we consider to estimate

$$\|\mathrm{sgn}(f_{\boldsymbol{z},\lambda}) - \mathrm{sgn}(f_\rho)\|_\rho,$$

where $\mathrm{sgn}(f) = (\mathrm{sgn}(f^1), \ldots, \mathrm{sgn}(f^d))$, $\mathrm{sgn}(f^i(x)) = 1$ if $f^i(x) \geq 0$ and $\mathrm{sgn}(f^i(x)) = -1$ otherwise.

Remark 2: If $\mathrm{sgn}(f_{\boldsymbol{z},\lambda})$ approximates $\mathrm{sgn}(f_\rho)$, it implies that $c(x)$ approximates $b(x)$.

4.1 An Error Bound for Multi-Class Classification

In order to estimate the error bound $\|\mathrm{sgn}(f_{\boldsymbol{z},\lambda}) - \mathrm{sgn}(f_\rho)\|_\rho$, we first denote the $misclassification\ set$ of the classifier $\mathrm{sgn}(f_{\boldsymbol{z},\lambda})$ as

$$\mathcal{X}_{f_{\boldsymbol{z},\lambda}} = \Big\{x \in \mathcal{X} \big| \exists i \in \{1, 2, \ldots, d\}, \\ \mathrm{sgn}(f^i_{\boldsymbol{z},\lambda})(x) \neq \mathrm{sgn}(f^i_\rho)(x), \Big\}.$$

Note that

$$\|\mathrm{sgn}(f_{\boldsymbol{z},\lambda}) - \mathrm{sgn}(f_\rho)\|^2_\rho = \\ \int_{\mathcal{X}\setminus\mathcal{X}_{f_{\boldsymbol{z},\lambda}}} \|\mathrm{sgn}(f_{\boldsymbol{z},\lambda})(x) - \mathrm{sgn}(f_\rho)(x)\|^2_d d\rho_{\mathcal{X}}(x) + \\ \int_{\mathcal{X}_{f_{\boldsymbol{z},\lambda}}} \|\mathrm{sgn}(f_{\boldsymbol{z},\lambda})(x) - \mathrm{sgn}(f_\rho)(x)\|^2_d d\rho_{\mathcal{X}}(x) = \\ 0 + \int_{\mathcal{X}_{f_{\boldsymbol{z},\lambda}}} \|\mathrm{sgn}(f_{\boldsymbol{z},\lambda})(x) - \mathrm{sgn}(f_\rho)(x)\|^2_d d\rho_{\mathcal{X}}(x).$$

Note that

$$\int_{\mathcal{X}_{f_{\boldsymbol{z},\lambda}}} \|\mathrm{sgn}(f_{\boldsymbol{z},\lambda})(x) - \mathrm{sgn}(f_\rho)(x)\|^2_d d\rho_{\mathcal{X}}(x) \leq \\ 4d \cdot \rho_{\mathcal{X}}(\mathcal{X}_{f_{\boldsymbol{z},\lambda}}),$$

so we have

$$\|\mathrm{sgn}(f_{\boldsymbol{z},\lambda}) - \mathrm{sgn}(f_\rho)\|^2_\rho \leq 4d \cdot \rho_{\mathcal{X}}(\mathcal{X}_{f_{\boldsymbol{z},\lambda}}), \qquad (14)$$

where $\rho_{\mathcal{X}}$ is the marginal distribution of ρ on $\mathcal{X}$.

In the following, we show that $\mathrm{sgn}(f_{\boldsymbol{z},\lambda})$ approximates $\mathrm{sgn}(f_\rho)$ well if $f_{\boldsymbol{z},\lambda}$ is a good approximation of f_ρ. To this end, we introduce a function motivated by the Tsybakov condition [17] with noise exponent q $(0 < q \leq \infty)$: for some constant $c_q > 0$, $\exists i \in \{1, 2, \ldots, d\}$,

$$\rho_{\mathcal{X}}\left(\left\{x \in \mathcal{X} \big| 0 < |f^i_\rho(x)| \leq c_q t\right\}\right) \leq t^q. \qquad (15)$$

Definition 1: The $Tsybakov\ function$ associated with the probability distribution ρ on $\mathcal{X} \times \mathcal{Y}$ is defined to be the function $S = S_\rho : [0, 1] \to [0, 1]$ given by, $\exists i \in \{1, 2, \ldots, d\}$

$$S(C) = \rho_{\mathcal{X}}\left(\left\{x \in \mathcal{X} \big| f^i_\rho(x) \in [-C, C], \right\}\right), \qquad (16)$$

Let $0 < q < \infty$, we define the q-coefficient as follows (if it is finite)

$$a_q = a_{q,\rho} = \sup_{0<C<1} \frac{S(C)}{C^q}. \qquad (17)$$

By the above definitions, it is easy to verify that for $0 < q < \infty$, the Tsybakov condition (15) holds if and only if $a_q < \infty$ and $S(0) = 0$. We say that ρ has (hard) margin $\tau > 0$ if $S(L) \equiv 0$ when $L \in [0, \tau)$.

Proposition 1: Let $\boldsymbol{z} = \{(x_i, y_i)\}^n_{i=1}$ be randomly drawn according to ρ having q-coefficient $a_q < \infty$ for some $0 < q < \infty$, then

$$\|\mathrm{sgn}(f_{\boldsymbol{z},\lambda}) - \mathrm{sgn}(f_\rho)\|^2_\rho \leq 4d \cdot a_q \kappa^q \|f_{\boldsymbol{z},\lambda} - f_\rho\|^q_K.$$

Proof: By the definition of misclassification set $\mathcal{X}_{f_{\boldsymbol{z},\lambda}}$, $\exists i \in \{1, 2, \ldots, d\}$,

$$\mathcal{X}_{f_{\boldsymbol{z},\lambda}} = \left\{x \in \mathcal{X} \big| \mathrm{sgn}(f^i_{\boldsymbol{z},\lambda})(x) \neq \mathrm{sgn}(f^i_\rho)(x)\right\},$$

we know that for $x \in \mathcal{X}_{f_{\boldsymbol{z},\lambda}}$, $\exists i \in \{1, 2, \ldots, d\}$ such that

$$\mathrm{sgn}(f^i_{\boldsymbol{z},\lambda})(x) \neq \mathrm{sgn}(f^i_\rho)(x).$$

Therefore, $\exists i \in \{1, 2, \ldots, d\}$,

$$|f^i_\rho(x)| \leq |f^i_{\boldsymbol{z},\lambda}(x) - f^i_\rho(x)| \leq \\ \|f_{\boldsymbol{z},\lambda}(x) - f_\rho(x)\|_d \leq \\ \|f_{\boldsymbol{z},\lambda} - f_\rho\|_\infty.$$

This means that the set $\mathcal{X}_{f_{\boldsymbol{z},\lambda}}$ is a subset of (or equal to)

$$\left\{x \in \mathcal{X} \big| f^i_\rho \leq \|f_{\boldsymbol{z},\lambda} - f_\rho\|_\infty, \exists i \in \{1, 2, \ldots, d\}\right\}.$$

By the definition of Tsybakov function, we have

$$\rho(\mathcal{X}_{f_{\boldsymbol{z},\lambda}}) \leq S(\|f_{\boldsymbol{z},\lambda} - f_\rho\|_\infty).$$

By (14) and $\|f\|_\infty \leq \kappa\|f\|_K$, we have

$$\|\mathrm{sgn}(f_{\boldsymbol{z},\lambda}) - \mathrm{sgn}(f_\rho)\|_\rho \leq \\ 4d \cdot S(\|f_{\boldsymbol{z},\lambda} - f_\rho\|_\infty) \leq \\ 4d \cdot S(\kappa\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K).$$

According to the definition of q-coefficient, it is easy to verify that

$$S(\kappa\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K) \leq a_q(\kappa\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K)^q.$$

This verifies the desired bound for $\|\mathrm{sgn}(f_{\boldsymbol{z},\lambda}) - \mathrm{sgn}(f_\rho)\|_\rho$. ■

This proposition shows that $\mathrm{sgn}(f_{\boldsymbol{z},\lambda})$ approximates $\mathrm{sgn}(f_\rho)$ well if $f_{\boldsymbol{z},\lambda}$ is a good approximation of f_ρ in $\|\cdot\|_K$. When ρ has hard margin $\tau > 0$, $S(C) = 0$ for $C < \tau$, it is sufficient to consider the case $\|f_{\boldsymbol{z},\lambda} - f_\rho\|_K \geq \frac{\tau}{\kappa}$ in Proposition 1.

Combining Theorem 2 and Proposition 1 yields the following result.

Corollary 2: Let $\boldsymbol{z} = \{(x_i, y_i)\}^n_{i=1}$ be randomly drawn according to ρ having $a_q < \infty$ for some $0 < q < \infty$ and $\|f_{\boldsymbol{z},\lambda} - f_\rho\| \leq c\lambda^\beta$. Setting

$$\lambda = (2\kappa D)^{\frac{1}{\beta+1}} \left(\frac{1}{n}\right)^{\frac{1}{2(\beta+1)}}.$$

Then with confidence $1 - \delta$,

$$\|\mathrm{sgn}(f) - \mathrm{sgn}(f_\rho)\|^2_\rho \leq 4d \cdot a_q \cdot (Q \log(2/\delta))^q \left(\frac{1}{n}\right)^{\frac{q\beta}{2(\beta+1)}},$$

where $Q = 4c\kappa^q(2kD)^{\frac{\beta}{\beta+1}}$.

Remark 3: The theoretical analysises of the multiclass empirical risk minimization methods in multiclass classification have been given in [18]–[20]. In this paper, we use the vector-valued RLS for multiclass classification, and present the specific convergence rate of the error bound (most of the above work only studied the consistency of multiclass classification, but didn't give the specific convergence rate of error bound).

5. Conclusion

The error analysis of the scalar RLS algorithm has been extensively studied in the literature, but little work has focused on the error analysis of the vector-valued RLS. In this paper, we propose the error bounds of the vector-valued RLS for general operator valued kernels. Furthermore, we consider to use the vector-valued RLS for multi-class classification, and derive an error bound for the multi-class classification problem.

Our analysis extensively uses the special properties of the square loss function, henceforth it would be interesting to extend our approach to other loss functions. We think that our results may be improved by taking into account more information about the structure of the hypothesis space.

Acknowledgments

The work is supported in part by the National Natural Science Foundation of China under grant No. 61170019, the Natural Science Foundation of Tianjin under grant No. 11JCYBJC00700.

Appendix.A

Lemma 1 (De Mol et al. [21] Proposition 6): Let H be a Hilbert space and let ξ be a random variable on $(\mathcal{Z},\rho)$ with values in H. Assume $\|\xi\| \leq C$ almost surely. Denote $\sigma^2(\xi) = \mathbb{E}(\|\xi\|^2)$. Let $\{z_i\}_{i=1}^n$ be independent random drawers of ρ. For any $0 < \delta < 1$, with confidence $1-\delta$,

$$\left\| \frac{1}{n}\sum_{i=1}^{m}[\xi_i - \mathbb{E}(\xi_i)] \right\| \leq \frac{2C\log(2/\delta)}{n} + \sqrt{\frac{2\sigma^2(\xi)\log(2/\delta)}{n}}.$$

Proof: [Proof of Theorem 1] By (4) and (6), we have

$$f_{\boldsymbol{z},\lambda} - f_\lambda = \left(\frac{1}{n}S_{\boldsymbol{x}}^* S_{\boldsymbol{x}} + \lambda I\right)^{-1} \times \left\{\frac{1}{n}S_{\boldsymbol{x}}^*\boldsymbol{y} - \frac{1}{n}S_{\boldsymbol{x}}^* S_{\boldsymbol{x}} f_\lambda - \lambda f_\lambda\right\}.$$

Note that

$$\frac{1}{n}S_{\boldsymbol{x}}^*\boldsymbol{y} - \frac{1}{n}S_{\boldsymbol{x}}^* S_{\boldsymbol{x}} f_\lambda = \frac{1}{n}\sum_{i=1}^{n} K_{x_i}(y_i - f_\lambda(x_i)),$$

and by the definition (6) of f_λ, we have

$$\lambda f_\lambda = L_K(f_\rho - f_\lambda).$$

It follows that, for all $\boldsymbol{z} = \{(x_i,y_i)\}_{i=1}^n$, and $\lambda > 0$,

$$f_{\boldsymbol{z},\lambda} - f_\lambda = \left(\frac{1}{n}S_{\boldsymbol{x}}^* S_{\boldsymbol{x}} + \lambda I\right)^{-1} \Lambda.$$

where

$$\Lambda = \frac{1}{n}\sum_{i=1}^{n} K_{x_i}(y_i - f_\lambda(x_i)) - L_K(f_\rho - f_\lambda).$$

Since $S_{\boldsymbol{x}}^* S_{\boldsymbol{x}}$ is positive semidefinite operator, it is easy to see that

$$\|f_{\boldsymbol{z},\lambda} - f_\lambda\|_K \leq \left\| \left(\frac{1}{n}S_{\boldsymbol{x}}^* S_{\boldsymbol{x}} + \lambda I\right)^{-1} \right\| \|\Lambda\|_K \leq \frac{1}{\lambda}\|\Lambda\|_K.$$

Denote random variable $\xi = K_x(y - f_\lambda(x))$ on $(\mathcal{Z},\rho)$ with values in $\mathcal{H}_K$. According to the reproducing property, we have

$$\|\xi\|_K = \|y - f_\lambda(x)\|_d \sqrt{\|K(x,x)\|} \leq \kappa(D + \|f_\lambda\|_\infty),$$

and

$$\sigma^2(\xi) \leq \kappa^2 \int_{\mathcal{Z}} \|f_\lambda(x) - y\|_d^2 d\rho.$$

Note that the definition of the regression function yields

$$\int_{\mathcal{Z}} \|f(x) - y\|_d^2 d\rho - \int_{\mathcal{Z}} \|f_\rho - y\|_d^2 d\rho = \|f - f_\rho\|_\rho^2. \quad (18)$$

Recall the definition (5) of f_λ. Setting $f = 0$ yields

$$\|f_\lambda - f_\rho\|_\rho^2 + \lambda\|f_\lambda\|_K^2 \leq \|f_\rho\|_\rho^2.$$

Hence

$$\|f_\lambda\|_K \leq \|f_\rho\|_\rho / \sqrt{\lambda}$$

and

$$\|f_\lambda - f_\rho\|_\rho^2 \leq \|f_\rho\|_\rho^2 \leq D^2.$$

Recall the Eq.(2), we have

$$\begin{aligned}\|\xi\|_K \leq & \kappa(D + \|f_\lambda\|_\infty) \leq \\ & \kappa(D + \kappa\|f_\lambda\|_K) \leq \\ & \kappa(D + \kappa\|f_\rho\|_\rho/\sqrt{\lambda}) \leq \\ & \kappa D(1 + \kappa/\sqrt{\lambda}).\end{aligned}$$

By (18), we have

$$\begin{aligned}&\int_{\mathcal{Z}} \|f_\rho(x) - y\|_d^2 d\rho = \\ &\int_{\mathcal{Z}} \|f(x) - y\|_d^2 d\rho - \|f - f_\rho\|_\rho^2 \leq \\ &\int_{\mathcal{Z}} \|f(x) - y\|_d^2.\end{aligned}$$

Setting $f = 0$, then

$$\int_{\mathcal{Z}} \|f_\rho(x) - y\|_d^2 d\rho \leq \int_{\mathcal{Z}} \|0 - y\|_d^2 d\rho \leq D^2,$$

thus

$$\int_{\mathcal{Z}} \|f_\lambda(x) - y\|_d^2 d\rho = \int_{\mathcal{Z}} \|f_\rho - y\|_d^2 d\rho + \|f_\lambda - f_\rho\|_\rho^2 \leq 2D^2.$$

Note that

$$\mathbb{E}(\xi) = \int_{\mathcal{X}} K_x \int_d (y - f_\lambda(x)) d\rho(y|x) d\rho_{\mathcal{X}}(x) = L_K(f_\rho - f_\lambda).$$

This means that

$$\Lambda = \frac{1}{n}\sum_{i=1}^n K_{x_i}(y_i - f_\lambda(x_i)) - L_K(f_\rho - f_\lambda) = \frac{1}{n}\sum_{i=1}^n [\xi(z_i) - \mathbb{E}(\xi)].$$

Using the lemma 1, with confidence $1-\delta$, we have

$$\|\Lambda\| \leq \frac{2\kappa(D + \|f_\lambda\|_\infty)\log(2/\delta)}{n} + \kappa\sqrt{\frac{2\int_{\mathcal{Z}} \|f_\lambda(x) - y\|_d^2 d\rho \log(2/\delta)}{n}} \leq \frac{2\kappa D(1 + \kappa/\sqrt{\lambda})\log(2/\delta)}{n} + 2\kappa D\sqrt{\frac{\log(2/\delta)}{n}}.$$

If $\kappa/\sqrt{n\lambda} \leq 1/(3\log(2/\delta))$, the above estimate can be bounded further as

$$\|\Lambda\| \leq \frac{2\kappa D\log(2/\delta)}{n} + \frac{2\kappa D\log(2/\delta)}{\sqrt{n}}\frac{\kappa}{\sqrt{n\lambda}} + \frac{2\kappa D\log(2/\delta)}{\sqrt{n}}\frac{1}{\sqrt{\log(2/\delta)}} \leq \frac{6\kappa D\log(2/\delta)}{\sqrt{n}}.$$

This yields the bound when $\kappa/\sqrt{n\lambda} \leq 1/(3\log(2/\delta))$.

When $\kappa/\sqrt{n\lambda} > 1/(3\log(2/\delta))$, we have

$$\frac{6\kappa D\log(2/\delta)}{\sqrt{n}\lambda} \geq 2D/\sqrt{\lambda}.$$

In this case, we use

$$\|f_\lambda\|_K \leq \|f_\rho\|_\rho/\sqrt{\lambda} \leq D/\sqrt{\lambda},$$

and the trivial bound

$$\|f_{\boldsymbol{z},\lambda}\| \leq D/\sqrt{\lambda}$$

seen from (4) by setting $f = 0$. Then there holds

$$\|f_{\boldsymbol{z},\lambda} - f_\lambda\|_K \leq 2D/\sqrt{\lambda}$$

with probability 1. So the desired inequality also holds in the second case. This proves Theorem 1. ■

References

[1] F. Cucker and S. Smale, "Best choices for regularization parameters in learning theory: on the bias-variance problem," *Foundations of Computational Mathematics*, vol. 2, no. 4, pp. 413–428, 2002.

[2] S. Smale and D.-X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constructive Approximation*, vol. 26, no. 2, pp. 153–172, 2007.

[3] V. Temlyakov, "Approximation in learning theory," *Constructive Approximation*, vol. 27, no. 1, pp. 33–74, 2008.

[4] I. Steinwart, D. Hush, and C. Scovel, "Optimal rates for regularized least squares regression," in *Proceedings of the 22nd Conference on Learning Theory (COLT 2009)*, 2009, pp. 79–93.

[5] Y. Liu, S. Jiang, and S. Liao, "Eigenvalues perturbation of integral operator for kernel selection," in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013)*, 2013, pp. 2189–2198.

[6] A. Caponnetto and E. D. Vito, "Optimal rates for the regularized least-squares algorithm," *Foundations of Computational Mathematics*, vol. 7, no. 3, pp. 331–368, 2006.

[7] C. A. Micchelli and M. Pontil, "On learning vector-valued functions," *Neural Computation*, vol. 17, no. 1, pp. 177–204, 2005.

[8] A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying, "Universal multi-task kernels," *Journal of Machine Learning Research*, vol. 9, pp. 1615–1646, 2008.

[9] C. Carmele, E. D. Vito, and A. Toigo, "Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem," *Analysis and Applications*, vol. 4, no. 4, pp. 377–408, 2006.

[10] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[11] H. Kadri, A. Rabaoui, P. Preux, E. Duflos, and A. Rakotomamonjy, "Functional regularized least squares classification with operator-valued kernels," in *Proceeding of the 28th International Conference on Machine Learning (ICML 2011)*, 2011, pp. 993–1000.

[12] H. Q. Minh and V. Sindhwani, "Vector-valued manifold regularization," in *Proceeding of the 28th International Conference on Machine Learning (ICML 2011)*, 2011, pp. 57–64.

[13] E. D. Vito, A. Caponnetto, and L. Rosasco, "Model selection for regularized least-squares algorithm in learning theory," *Foundations of Computational Mathematics*, vol. 5, no. 1, pp. 59–85, 2005.

[14] S. Smale and D.-X. Zhou, "Shannon sampling II: Connections to learning theory," *Applied and Computational Harmonic Analysis*, vol. 19, no. 3, pp. 285–302, 2005.

[15] Q. Wu, Y. Ying, and D.-X. Zhou, "Learning rates of least-square regularized regression," *Foundations of Computational Mathematics*, vol. 6, no. 2, pp. 171–192, 2006.

[16] Z. Tong, "Leave-one-out bounds for kernel methods," *Neural Computation*, vol. 15, no. 6, pp. 1397–1437, 2003.

[17] A. B. Tsybakov, "Optimal aggregation of classifiers in statistical learning," *The Annals of Statistics*, vol. 32, pp. 135–166, 2004.

[18] D.-R. Chen and T. Sun, "Consistency of multiclass empirical risk minimization methods based on convex loss," *Journal of Machine Learning Research*, vol. 7, pp. 2435–2447, 2006.

[19] A. Tewari and P. L. Bartlett, "On the consistency of multiclass classification methods," *Journal of Machine Learning Research*, vol. 8, pp. 1007–1025, 2007.

[20] Y. Guermeur, "VC theory of large margin multi-category classifiers," *Journal of Machine Learning Research*, vol. 8, pp. 2551–2594, 2007.

[21] C. D. Mola, E. D. Vito, and L. Rosasco, "Elastic-net regularization in learning theory," *Journal of Complexity*, vol. 25, pp. 201–230, 2009.