LETTER Communicated by Qiang Wu

Generalization Analysis of Fredholm Kernel Regularized Classifiers

Tieliang Gong
adidasgtl@gmail.com
Zongben Xu
zbxu@mail.xjtu.edu.cn
School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China

Hong Chen
chenh@mail.hzav.edu.cn
College of Science, Huazhong Agricultural University, Wuhan 430070, China

Recently, a new framework, Fredholm learning, was proposed for semisupervised learning problems based on solving a regularized Fredholm integral equation. It allows a natural way to incorporate unlabeled data into learning algorithms to improve their prediction performance. Despite rapid progress on implementable algorithms with theoretical guarantees, the generalization ability of Fredholm kernel learning has not been studied. In this letter, we focus on investigating the generalization performance of a family of classification algorithms, referred to as Fredholm kernel regularized classifiers. We prove that the corresponding learning rate can achieve $\mathcal{O}(l^{-1})$ (l is the number of labeled samples) in a limiting case. In addition, a representer theorem is provided for the proposed regularized scheme, which underlies its applications.

1 Introduction

Many scientific problems (e.g., regression and classification) come down to learning a prediction rule from the given finite input-output samples. Kernel tricks and methods based on integral operators provide powerful tools for learning tasks and have become central to machine learning. In order to construct a good predictor, one usually chooses a function from a class of functions (hypothesis space) using regularized learning schemes associated with certain loss functions.

Regularized kernel learning has attracted much attention due to its solid theoretical foundations and successful practical applications. Recently a new kernel learning framework, Fredholm learning, has been proposed by reformulating the learning problem as a regularized Fredholm integral equation (Que, Belkin & Wang, 2014; Que & Belkin, 2013). This framework

Neural Computation **29, 1879–1901** (2017)
doi:10.1162/NECO_a_00967

allows a way to incorporate unlabeled data into learning algorithms and can be interpreted as a special form of kernel method with a data-dependent kernel—the Fredholm kernel. It has been shown that the Fredholm classification algorithm can reduce the variance of kernel function evaluations at data noise and improve the prediction accuracy and robustness of kernel methods (Que et al., 2014).

Despite rapid progress on theoretical and empirical evaluations, the generalization performance of Fredholm kernel learning remains unknown. This letter makes efforts to answer the question. Specifically, we focus on error analysis for a family of classification algorithms, Fredholm kernel regularized classifiers, and establishing the corresponding generalization error bound. We show that the learning rate of Fredholm kernel regularized classifiers can achieve $\mathcal{O}(l^{-1})$ under mild conditions. In addition, we also justify its representer theorem, which makes the solution of Fredholm kernel learning model computation accessible.

The rest of the letter is organized as follows. Section 2 presents basic definitions and some necessary background. Section 3 focuses on establishing the generalization bounds for Fredholm kernel regularized classifiers. We conclude in Section 4.

2 Preliminaries

2.1 Classification in Learning Theory. We begin with a brief review of a binary classification problem. Let (X, d) be a compact metric space and $Y = \{-1, 1\}$ be the output space. A classifier f is a map $f : X \to Y$ that makes a prediction $y = f(\mathbf{x})$ for each $\mathbf{x} \in X$. The mapping relationship between the input and the output can be modeled by a probability ρ on $Z := X \times Y$. Suppose ρ admits a decomposition $\rho(\mathbf{x}, y) = \rho_X(\mathbf{x})\rho(y|\mathbf{x})$ in which $\rho_X(\mathbf{x})$ denotes a marginal probability measure on X and $\rho(y|\mathbf{x})$ denotes a condition probability measure (given $\mathbf{x}$) on Y. Then the prediction ability of a classifier $f : X \to Y$ can be measured by a misclassification error, defined as the probability of incorrect prediction

$$\mathcal{R}(f) := \text{Prob}\{f(X) \neq Y\} = \int_X \text{Prob}_{y \in Y}(y \neq f(\mathbf{x})|\mathbf{x}) d\rho_X. \tag{2.1}$$

Define the regression function

$$f_\rho(\mathbf{x}) = \text{Prob}_Y(y = 1|\mathbf{x}) - \text{Prob}_Y(y = -1|\mathbf{x}) = 2\eta(\mathbf{x}) - 1,$$

where $\eta(\mathbf{x}) = \text{Prob}_Y(y = 1|\mathbf{x})$. The best classifier that minimizes the misclassification error is the Bayes rule, given by

$$f_c(\mathbf{x}) = \operatorname{sgn}(f_\rho)(\mathbf{x}) = \begin{cases} 1, & \text{if } \operatorname{Prob}_Y(y=1|\mathbf{x}) > \operatorname{Prob}_Y(y=-1|\mathbf{x}), \\ -1, & \text{if } \operatorname{Prob}_Y(y=1|\mathbf{x}) \le \operatorname{Prob}_Y(y=-1|\mathbf{x}). \end{cases} \tag{2.2}$$

To formulate the error analysis for Fredholm kernel regularized classifiers, we assume the loss function ϕ satisfies the following condition:

Definition 1. *We say that $\phi : \mathbb{R} \to \mathbb{R}_+$ is a normalized classification loss function if it is convex, differentiable at 0 with $\phi'(0) < 0$, and the smallest zero of ϕ is 1.*

Typical examples of classification loss include hinge loss $\phi_h(t) = (1-t)_+$ for support vector machine (SVM) $\phi_q(t) = (1-t)_+^q$ for SVM q-norm ($q > 1$) soft margin classifier, and least square loss $\phi_{ls}(t) = (1-t)^2$.

Define the expected risk associated with ϕ as

$$\mathcal{E}(f) = \int_{X\times Y} \phi(yf(\mathbf{x}))d\rho(\mathbf{x}, y). \tag{2.3}$$

However, it cannot be computed directly since ρ is usually unknown. Its discretization is therefore often used, which, computable in terms of finite samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^l$, is defined as

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{l}\sum_{i=1}^{l} \phi(y_i f(\mathbf{x}_i)) \tag{2.4}$$

and called empirical risk. Regularized learning schemes, implemented by minimizing a penalized version of empirical risk, aim to find a good approximation of the Bayesian rule. It is expected that minimization can always be taken over a set of functions, known as hypothesis space, which is usually selected as a reproducing kernel Hilbert space (RKHS) associated with a Mercer kernel. Such a kernel $K_{\mathcal{H}}$ is a continuous, symmetric function on $X \times X$ such that the matrix $(K_{\mathcal{H}}(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^m$ is positive semidefinite for any $\{\mathbf{x}_i\}_{i=1}^m \subset X$. It is well known that a specific Mercer kernel corresponds to a unique RKHS (Aronszajn, 1950) $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$. The reproducing property takes the form $\langle K_{\mathcal{H}}(\mathbf{x}, \cdot), f\rangle_{\mathcal{H}} = f(\mathbf{x})$, $\forall \mathbf{x} \in X$, $\forall f \in \mathcal{H}$. The standard regularized classifier is defined by

$$f_{\mathbf{z}}^* = \operatorname*{argmin}_{f\in\mathcal{H}} \left\{ \frac{1}{l}\sum_{i=1}^{l} \phi(y_i f(\mathbf{x}_i)) + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \tag{2.5}$$

where λ is regularization parameter and controls a trade-off between empirical risk and the regularization term.

2.2 Fredholm Learning Framework. Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ be l labeled pairs from distribution $\rho(\mathbf{x}, y)$ and $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$ be the unlabeled data points from marginal distribution $\rho_X(\mathbf{x})$. The goal of semisupervised classification is to construct a reliable classifier by incorporating the information of labeled and unlabeled data together. To this end, we introduce an integral operator $\mathcal{K}_{P_X}$ associated with a kernel function $K(\mathbf{x}, \mathbf{t})$ with $\kappa := \sup_{\mathbf{t},\mathbf{x}} K(\mathbf{x}, \mathbf{t}) < \infty$, defined by

$$(\mathcal{K}_{P_X} f)(\mathbf{x}) = \int K(\mathbf{x}, \mathbf{t}) f(\mathbf{t}) p_X(\mathbf{t}) d\mathbf{t}, \mathcal{K}_{P_X} : L^2_{\rho_X} \to L^2_{\rho_X}, \tag{2.6}$$

where $p_X(\cdot)$ denotes the density function of $\rho_X(\cdot)$ and $L^2_{\rho_X}$ the square integrable function space. Generally, by the law of large numbers, $\mathcal{K}_{P_X}$ can be approximated by unlabeled data from ρ_X as

$$(\mathcal{K}_{\hat{P}_X} f)(\mathbf{x}) = \frac{1}{l+u} \sum_{i=1}^{l+u} K(\mathbf{x}, \mathbf{x}_i) f(\mathbf{x}_i). \tag{2.7}$$

The target of the Fredholm learning framework is then to solve the following optimization problem:

$$\bar{f} = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{l} \sum_{i=1}^{l} \left(y_i - (K_{\hat{P}_X} f)(\mathbf{x}_i)\right)^2 + \lambda \|f\|^2_{\mathcal{H}} \right\}. \tag{2.8}$$

The final classifier is $g(\mathbf{x}) = \operatorname{sgn}\big((K_{\hat{P}_X} \bar{f})\big)(\mathbf{x})$. Notice that equation 2.8 can be considered as an empirical regularized version of the Fredholm integral equation $(\mathcal{K}_{P_X} \bar{f})(\mathbf{x}) = \mathbf{y}$. This is why such a learning scheme is called the Fredholm learning framework. In this framework, the intrinsic hypothesis space $\mathcal{K}_{P_X}\mathcal{H} = \{\mathcal{K}_{P_X} \bar{f} : \bar{f} \in \mathcal{H}\}$ is density dependent.

At first glance, the Fredholm learning framework defined in equation 2.8 looks similar to standard regularized learning, equation 2.5. However, $\mathcal{K}_{\hat{P}_X}$ makes a significant difference. The density-dependent hypothesis space enables us to integrate the information from unlabeled data. In contrast, most kernels used in a traditional kernel learning framework—for example, linear kernel, gaussian kernel, and polynomial kernel—are completely independent of data distribution. In particular, when setting the kernel K to be δ-function, formulation 2.8 is reduced to the standard regularized learning framework.

In addition, the solution of the optimization problem, equation 2.8, is computationally accessible and benefits from the well-known representer theorem in RKHS. That theorem (Que et al., 2014) allows us to transform

equation 2.8 into quadratic optimization in a finite dimensional space. The solution can be represented as

$$\bar{f}(\mathbf{x}) = \frac{1}{l+u}\sum_{j=1}^{l+u} c_j K_{\mathcal{H}}(\mathbf{x}, \mathbf{x}_j), \mathbf{c} = (\mathbf{K}_{l+u}^{\top}\mathbf{K}_{l+u}\mathbf{K}_{\mathcal{H}} + \lambda I)^{-1}\mathbf{K}_{l+u}^{\top}\mathbf{y},$$

where $(\mathbf{K}_{l+u})_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ for $1 \le i \le l, 1 \le j \le l+u$, and $(\mathbf{K}_{\mathcal{H}})_{ij} = K_{\mathcal{H}}(\mathbf{x}_i, \mathbf{x}_j)$ for $1 \le i, j \le l+u$. Let $\mathbf{K}_F = \mathbf{K}_{l+u}\mathbf{K}_{\mathcal{H}}\mathbf{K}_{l+u}^{\top}$ be the $l \times l$ matrix associated with a new kernel, named the Fredholm kernel, defined by

$$K_F(\mathbf{x}, \mathbf{x}') = \frac{1}{(l+u)^2}\sum_{i,j=1}^{l+u} K(\mathbf{x}, \mathbf{x}_i)K_{\mathcal{H}}(\mathbf{x}_i, \mathbf{x}_j)K(\mathbf{x}', \mathbf{x}_j). \tag{2.9}$$

Then the solution of equation 2.8 can be rewritten as

$$(K_{\hat{P}_X}\bar{f})(\mathbf{x}) = \frac{1}{l+u}\sum_{i=1}^{l+u} K(\mathbf{x}, \mathbf{x}_i)\bar{f}(\mathbf{x}_i) = \sum_{t=1}^{l} K_F(\mathbf{x}, \mathbf{x}_t)\gamma_t, \boldsymbol{\gamma} = (\mathbf{K}_F + \lambda I)^{-1}\mathbf{y}.$$

Equation 2.9 involves the "inner" kernel $K_{\mathcal{H}}$ and the "outer" kernel K. It can be proved that the Fredholm kernel defined in equation 2.9 is positive semidefinite if $K_{\mathcal{H}}$ is a positive semidefinite kernel. Note that it is not necessary for the outer kernel to be positive definite or even symmetric, and it can be selected flexibly based on the user's preferences.

Inspired by the Fredholm kernel (Que et al., 2014), we propose a family of Fredholm kernel regularized classifiers given by

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}}\left\{\frac{1}{l}\sum_{i=1}^{l}\phi(y_i \cdot (K_{\hat{P}_X}f)(\mathbf{x}_i)) + \lambda\|f\|_{\mathcal{H}}^2\right\}, \tag{2.10}$$

where $\lambda > 0$ is a regularization parameter.

The main goal of this letter is to investigate the generalization performance of equation 2.10. Specifically, we expect to give an explicit convergence rate for Fredholm kernel regularized classifiers under some mild conditions. The following proposition states the representer theorem for Fredholm kernel regularized classifiers:

Proposition 1. *Assume $\mathcal{K}_{\hat{P}_X}$ is defined in equation 2.7, and ϕ is a classification loss. Then the solution of equation 2.10 is of the form*

$$f_z(\boldsymbol{x}) = \frac{1}{l+u}\sum_{j=1}^{l+u} c_j K_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{x}_j),$$

for some $\boldsymbol{c} = [c_1, c_2, \cdots, c_{l+u}] \in \mathbb{R}^{l+u}$.

The proof of proposition 1 is similar to the analysis of a representer theorem for kernel methods (e.g., Belkin & Niyogi, 2006). We provide its proof in the appendix.

3 Bounds of Generalization Error

The generalization analysis aims at bounding the misclassification error $\mathcal{R}(\text{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c)$. Nevertheless, the algorithm is constructed by minimizing regularized empirical error $\mathcal{E}_{\mathbf{z}}(f)$ associated with the loss function ϕ. Hence it seems necessary to build a bridge between the excess misclassification error and the excess convex risk. Fortunately, researchers have established comparison theorems to solve this problem (Bartlett & Mendelson, 2002; Zhang, 2004; Chen, Wu, Yin & Zhou, 2004; Wu & Zhou, 2005). Here we mention some results that will be used in this letter.

Lemma 1. *(Chen et al., 2004). If an activating loss ϕ satisfies $\phi''(0) > 0$, then there exists a constant $c_\phi > 0$ such that for any measurable function $f : X \to \mathbb{R}$, there holds*

$$\mathcal{R}(sgn(f)) - \mathcal{R}(f_c) \le c_\phi \sqrt{\mathcal{E}(f) - \mathcal{E}(f_\rho)}.$$

Furthermore, we can get tight comparison bounds when the distribution satisfies the Tsybakov noise condition (Tsybakov, 2004).

Definition 2. *We say that ρ has a Tsybakov noise exponent $\alpha \ge 0$ if there exists a constant $c_\alpha > 0$ for every measurable function f:*

$$\rho_X(\{\boldsymbol{x} \in X : sgn(f(\boldsymbol{x})) \ne sgn(f_c(\boldsymbol{x}))\}) \le c_\alpha (\mathcal{R}(f) - \mathcal{R}(f_c))^\alpha. \tag{3.1}$$

Note that all the distributions satisfy equation 3.1 with $\alpha = 0$ and $c_\alpha = 1$. Tsybakov (2004) considered the convergence rate of the risk of a function that minimizes empirical risk over a fixed class and demonstrated that a fast convergence rate $O(l^{-1})$ can be achieved under the Tsybakov noise condition. We assume that ρ satisfies the Tsybakov condition to obtain a fast convergence rate.

Lemma 2 *(Wu, Ying, & Zhou, 2007). Let classification loss ϕ satisfy $\phi''(0) > 0$. If ρ satisfies the Tsybakov noise condition, equation 3.1, for some $\alpha \in [0, 1]$ and $c_\alpha > 0$, then*

$$\mathcal{R}(sgn(f)) - \mathcal{R}(f_c) \leq \left[2c_\phi c_\alpha(\mathcal{E}(f) - \mathcal{E}(f_c))\right]^{\frac{1}{2-\alpha}}, \forall f : X \to \mathbb{R}.$$

Since Fredholm kernel regularized classifiers are obtained by composing the sgn function with a real-value function $f : X \to \mathbb{R}$, we expect to improve the error estimates by projecting the estimator f into $[-1, 1]$.

Definition 3. *The projection operator π is defined on the space of measurable functions $f : X \to \mathbb{R}$ as*

$$\pi(f)(x) = \begin{cases} 1, & \text{if} \quad f(x) > 1, \\ f(x), & \text{if} \quad |f(x)| \leq 1, \\ -1, & \text{if} \quad f(x) < -1. \end{cases}$$

It is easy to check that $\text{sgn}(\pi(f)) = \text{sgn}(f)$. A well-developed approach for conducting a generalization analysis of the regularization algorithm in RKHS is error decomposition, which allows the excess generalization error to be decomposed into sample error and approximation error (Zhou, 2002; Cucker & Smale, 2001). In the Fredholm learning framework, we formulate error decomposition in a similar way by introducing a data-independent regularized function. We first introduce some conditions on the capacity of hypothesis space and the approximation ability of Fredholm learning framework. The covering number (Zhou, 2002, 2003; Shi, Feng, & Zhou, 2011) is used to describe the capacity of a function space.

Definition 4. *Let $\mathcal{M}, d$ be a pseudometric space and $S \subset \mathcal{M}$. For every $\varepsilon > 0$, the covering number $\mathcal{N}(S, \varepsilon, d)$ of S with respect to ε and d is defined as the minimal number of balls of radius ε whose union covers S, that is,*

$$\mathcal{N}(S, \varepsilon, d) = \min\left\{ n \in \mathbb{N} : S \subset \bigcup_{j=1}^{n} B(o_j, \varepsilon) \text{ for some } \{o_j\}_{j=1}^{n} \subset \mathcal{M} \right\},$$

where $B(o_j, \varepsilon) = \{o \in \mathcal{M} : d(o, o_j) \leq \varepsilon\}$ is a ball in $\mathcal{M}$.

Definition 5. *Let $\mathcal{F}$ be a class of functions on sample set $\mathbf{z} := \{z_i\}_{i=1}^m$. The ℓ_2-metric $d_{2,\mathbf{z}}$ is defined on $\mathcal{F}$ by*

$$d_{2,\bar{\mathbf{z}}}(f,g)=\left\{\frac{1}{m}\sum_{i=1}^{m}(f(z_i)-g(z_i))^2\right\}^{1/2}.$$

For every $\varepsilon>0$, the covering number of $\mathcal{F}$ with ℓ_2-metric is defined as

$$\mathcal{N}_2(\mathcal{F},\varepsilon)=\inf\left\{n\in\mathbb{N}:\exists\{f_i\}_{i=1}^n \text{ such that } \mathcal{F}=\bigcup_{i=1}^{n}\{f\in\mathcal{F}:d_{2,\mathbf{z}}(f,f_i)\le\varepsilon\}\right\},$$

and the covering number of $\mathcal{F}$ with ℓ_∞-metric is denoted by $\mathcal{N}_\infty(\mathcal{F},\varepsilon)$. Note that for any function set $\mathcal{F}\subset\mathcal{C}(X)$, there exists $\mathcal{N}_2(\mathcal{F},\varepsilon)\le\mathcal{N}_\infty(\mathcal{F},\varepsilon)$.

For $R>0$, denote

$$\mathcal{B}_R=\{f\in\mathcal{H}:\|f\|_{\mathcal{H}}\le R\}$$

and

$$\bar{\mathcal{B}}_R=\left\{f=\sum_{i=1}^{l+u}\alpha_iK(\cdot,\mathbf{u}_i):\sum_{i=1}^{l+u}|\alpha_i|\le R,\mathbf{u}_i\in X\right\}.$$

Assumption 1 (capacity condition). For the inner kernel $K_{\mathcal{H}}$ and outer kernel K, there exist positive constants s and p such that for any $\varepsilon>0$,

$$\log\mathcal{N}_\infty(\mathcal{B}_1,\varepsilon)\le C_s(1/\varepsilon)^s \quad\text{and}\quad \log\mathcal{N}_\infty(\bar{\mathcal{B}}_1,\varepsilon)\le C_p(1/\varepsilon)^p,$$

where $C_s,C_p>0$ are positive constants independent of ε.

The data-independent regularized function is defined by

$$f_\lambda=\arg\min_{f\in\mathcal{H}}\{\mathcal{E}(\mathcal{K}_{P_X}f)-\mathcal{E}(f_\rho)+\lambda\|f\|_{\mathcal{H}}^2\}. \tag{3.2}$$

Denote

$$(\mathcal{K}_{P_X}f_\lambda)(\mathbf{x})=\int_X f_\lambda(\mathbf{x})K(\mathbf{x},\mathbf{t})d\rho_X(\mathbf{t}),\ \forall\mathbf{x},\mathbf{t}\in X. \tag{3.3}$$

Then the approximation ability of the Fredholm kernel learning scheme can be characterized by

$$\mathcal{D}(\lambda)=\mathcal{E}(\mathcal{K}_{P_X}f_\lambda)-\mathcal{E}(f_\rho)+\lambda\|f_\lambda\|_{\mathcal{H}}^2=\min_{f\in\mathcal{H}}\{\mathcal{E}(\mathcal{K}_{P_X}f)-\mathcal{E}(f_\rho)+\lambda\|f\|_{\mathcal{H}}^2\}.$$

Assumption 2 (approximation condition). There exists a constant $0 < \beta \leq 1$ such that

$$\mathcal{D}(\lambda) \leq c_\beta \lambda^\beta, \quad \forall \lambda > 0,$$

where $c_\beta > 0$ is a constant independent of λ.

Assumption 3. Suppose distribution ρ satisfies Tsybakov noise condition 3.1 and there exists some $\tau \in [0, 1]$ and a constant $C_\tau > 0$ such that

$$\mathbb{E}[(\phi(y(\mathcal{K}_{P_X} f)(\mathbf{x})) - \phi(y f_\rho(\mathbf{x})))^2] \leq C_\tau (\mathcal{E}(\mathcal{K}_{P_X} f) - \mathcal{E}(f_\rho))^\tau, \ \ \forall \|f\|_\infty \leq 1.$$

Following the ideas in Wu and Zhou (2005, 2008), we obtain the following error decomposition:

Proposition 2. *Let $f_{\mathbf{z}}$ be defined as in equation 2.10 with sample set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^l \cup \{x_j\}_{j=l+1}^{l+u}$ and $\lambda > 0$. Then*

$$\mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \leq \mathcal{S}(\mathbf{z}, \lambda) + \mathcal{D}(\lambda) + \mathcal{H}(\mathbf{z}, \lambda), \tag{3.4}$$

where

$$\begin{aligned}
\mathcal{S}(\mathbf{z}, \lambda) &= \mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) + \mathcal{E}_{\mathbf{z}}(\mathcal{K}_{\hat{P}_X} f_\lambda) - \mathcal{E}(\mathcal{K}_{\hat{P}_X} f_\lambda), \\
\mathcal{D}(\lambda) &= \mathcal{E}(\mathcal{K}_{P_X} f_\lambda) - \mathcal{E}(f_\rho) + \lambda \|f_\lambda\|_{\mathcal{H}}^2, \\
\mathcal{H}(\mathbf{z}, \lambda) &= \mathcal{E}(\mathcal{K}_{\hat{P}_X} f_\lambda) - \mathcal{E}(\mathcal{K}_{P_X} f_\lambda),
\end{aligned}$$

are named sample error, approximation error, and hypothesis error, respectively.

Proof. A direct computation shows that

$$\begin{aligned}
&\mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \\
&= \mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) \\
&\quad + \left\{ \mathcal{E}_{\mathbf{z}}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) + \lambda \|f_{\mathbf{z}}\|_{\mathcal{H}}^2 - \mathcal{E}_{\mathbf{z}}(\mathcal{K}_{\hat{P}_X} f_\lambda) - \lambda \|f_\lambda\|_{\mathcal{H}}^2 \right\} \\
&\quad + \mathcal{E}_{\mathbf{z}}(\mathcal{K}_{\hat{P}_X} f_\lambda) - \mathcal{E}(\mathcal{K}_{\hat{P}_X} f_\lambda) + \mathcal{E}(\mathcal{K}_{\hat{P}_X} f_\lambda) - \mathcal{E}(\mathcal{K}_{P_X} f_\lambda) \\
&\quad + \left\{ \mathcal{E}(\mathcal{K}_{P_X} f_\lambda) - \mathcal{E}(f_\rho) + \lambda \|f_\lambda\|_{\mathcal{H}}^2 \right\} - \lambda \|f_{\mathbf{z}}\|_{\mathcal{H}}^2.
\end{aligned}$$

Then the conclusion holds true since both the second term and the last term of the above equality are less than 0. □

Theorem 1. *Let ϕ be a normalized classification loss, and suppose it satisfies increment condition $|\phi(t)| \le c_q|t|^q$ ($q > 0, c_q > 0$). Under assumptions 1, 2, and 3, the following inequality,*

$$\mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le \hat{C} log\,(3/\delta)\Big(\lambda^{-\frac{s}{4-2\tau+\tau s}} l^{-\frac{2}{4-2\tau+\tau s}} + l^{-1} + \lambda^{\beta} + \lambda^{\frac{(\beta-1)q}{2}} l^{-1} + \lambda^{\frac{(\beta-1)p}{2}} l^{-\frac{1}{2-\tau+p}} + \lambda^{\frac{\beta-1}{2}}(l+u)^{-1} + \lambda^{\frac{\beta-1}{2}}(l+u)^{-\frac{1}{2}}\Big),$$

holds with probability at least $1-\delta$, where $\hat{C}$ is a constant independent of l, λ, δ.

Remark 1. It can be observed from theorem 1 that the generalization bound relies on the capacity condition, the approximation condition, and the choice of regularization parameter λ. Specifically, the labeled data play a key role on the generalization bound without the extra assumption on marginal distribution, which is consistent with the theoretical analysis for semisupervised learning (Belkin & Niyogi, 2006; Chen, Zhou, Tang, Li, & Pan, 2013).

Theorem 2. *With the same conditions in theorem 1, taking $\lambda = l^{-\theta}$, we get*

$$\mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le 7\hat{C} log\,(3/\delta) l^{-\theta}, \tag{3.5}$$

which holds with probability at least $1-\delta$, where

$$\theta = min\left\{\frac{2\beta}{(4-2\tau+\tau s)\beta + s}, \quad \frac{2\beta}{1+\beta}, \quad \frac{2\beta}{(2\beta-\beta p+p)(2-\tau+p)}\right\}, \tag{3.6}$$

and $\hat{C}$ is a constant independent to l, λ, and δ.

Remark 2. Let $\tau \to 1$ and $\beta \to 1$. Then it can be observed that the learning rate will be arbitrarily close to $\mathcal{O}(l^{-1})$ for sufficient small s and p, which is regarded as the optimal learning rate in theory. We give two examples to show how the distribution ρ and variance-expectation τ influence the learning rate. When $\phi(t) = \phi_h(t)$, if ρ satisfies the Tsybakov noise condition, equation 3.1, then we know that assumption 3 is valid with $\tau = \frac{\alpha}{1+\alpha}$ and the constant $C_\tau = 8(\frac{1}{2c_\alpha})^{\alpha/(1+\alpha)}$ (Steinwart & Scovel, 2005). When $\phi(t) = \phi_{ls}(t)$, a direct computation implies

$$\begin{aligned}
&\mathbb{E}[(\phi_{ls}(y(\mathcal{K}_{P_X} f)(\mathbf{x})) - \phi_{ls}(y f_\rho(\mathbf{x})))^2] \\
\le\, &16\mathbb{E}[(y - (\mathcal{K}_{P_X} f)(\mathbf{x}))^2 - (y - f_\rho(\mathbf{x}))^2] \\
=\, &16\big(\mathcal{E}(\mathcal{K}_{P_X} f) - \mathcal{E}(f_\rho)\big),
\end{aligned}$$

Assumption 3 holds by taking $C_\tau = 16$ and $\tau = 1$ (Lee, Bartlett, & Williamson, 1996). Theorem 2 illustrates that Fredholm kernel regularized classifiers inherit the convergence characteristic of a standard kernel-based regularization classification algorithm.

Remark 3. In the semisupervised learning literature (Johnson & Zhang, 2007; Chen, Pan, Li, & Tang, 2013), the learning rate is essentially determined by the number of labeled data. Nevertheless, it does not mean that unlabeled data have no effect on the final result. In fact, the estimation of hypothesis error involves the unlabeled data, and some empirical theoretical results illustrate that the unlabeled data are helpful for improving learning performance. However, the effect on the learning rate is limited due to the fact that $u \gg l$.

By theorem 2, a direct corollary can be obtained a:

Corollary 1. *With the same conditions in theorem 2, for any $0 < \delta < 1$, there exists a constant $\hat{C}$ independent of l, λ, δ such that the following inequality,*

$$\mathcal{R}(sgn(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{R}(f_c) \leq \hat{C} \left(\frac{1}{l}\right)^{\theta/2},$$

holds with confidence at least $1 - \delta$, where $\lambda = l^{-\theta}$ and θ is given by equation 3.6. In addition, if ρ satisfies the Tsybakov noise condition with $\alpha \in (0, 1]$, the power $\theta/2$ can be improved to $\frac{\theta}{2-\alpha}$.

We are now in a position to present the proofs of main results based on error decomposition, equation 3.4.

3.1 Estimation of Hypothesis Error. The following lemmas are useful for estimating hypothesis error.

Lemma 3 *(Smale & Zhou, 2007). Let $\mathcal{H}$ be a Hilbert space and ξ be a random variable on Z with values in $\mathcal{H}$. Assume that $\|\xi\| \leq \tilde{M} < \infty$ almost surely. Denote $\sigma^2(\xi) = \mathbb{E}(\xi^2)$. Then for any given independent and identically distributed samples $\{z_i\}_{i=1}^m$ and any $\delta \in (0, 1)$, there holds*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mathbb{E}\xi \right\| \leq \frac{2\tilde{M} log\,(2/\delta)}{m} + \sqrt{\frac{2\mathbb{E}\xi^2 log\,(2/\delta)}{m}}$$

with confidence at least $1 - \delta$.

Lemma 4. *Let ϕ be a normalized classification loss function. Then*

$$\mathcal{E}(\mathcal{K}_{\hat{P}_X} f_\lambda) - \mathcal{E}(\mathcal{K}_{P_X} f_\lambda) \leq \|\phi'\|_\infty \|\mathcal{K}_{\hat{P}_X} f_\lambda - \mathcal{K}_{P_X} f_\lambda\|_\infty.$$

Proof. Define a univariate convex function Q for $\mathbf{x} \in X$ as

$$Q(t) = Q_{\mathbf{x}}(t) := \int_Y \phi(yt) d\rho(y|\mathbf{x}), \quad t \in \mathbb{R}.$$

It is easy to check that the one-side derivatives of $Q(t)$ exist, are nondecreasing, and satisfy $Q'_-(t) \le Q'_+(t)$ for every $t \in \mathbb{R}$. Denote

$$(\mathcal{K}_{P_X} f_\lambda^-)(\mathbf{x}) = \sup\{t \in \mathbb{R}, Q'_-(t) < 0\},$$
$$(\mathcal{K}_{P_X} f_\lambda^+)(\mathbf{x}) = \inf\{t \in \mathbb{R}, Q'_+(t) > 0\}.$$

Then for each $\mathbf{x} \in X$, the univariate function Q is strictly decreasing in $\big(-\infty, (\mathcal{K}_{P_X} f_\lambda^-)(\mathbf{x})\big]$ and strictly increasing in $\big[(\mathcal{K}_{P_X} f_\lambda^+)(\mathbf{x}), +\infty\big)$. For $t \in \big[(\mathcal{K}_{P_X} f_\lambda^-)(\mathbf{x}), (\mathcal{K}_{P_X} f_\lambda^+)(\mathbf{x})\big]$, we have $0 \le Q'_-(t) \le Q'_+(t) \le 0$. Hence, Q is a constant, which is the minimal value on $\mathbb{R}$.

With the function $Q = Q_{\mathbf{x}}$, we can rewrite $\mathcal{E}(\mathcal{K}_{\hat{P}_X} f_\lambda) - \mathcal{E}(\mathcal{K}_{P_X} f_\lambda)$ as

$$\mathcal{E}\big(\mathcal{K}_{\hat{P}_X} f_\lambda\big) - \mathcal{E}\big(\mathcal{K}_{P_X} f_\lambda\big) = \int_X \Big\{ Q\big((\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})\big) - Q\big((\mathcal{K}_{P_X} f_\lambda)(\mathbf{x})\big) \Big\} d\rho_X.$$

Since $\phi'(0) < 0$ and $\phi(t) \ge 0$, we derive $\phi(0) > 0$ and $\phi'_\pm(t) < 0$ for $t < 0$. Let $P(t) = \max\{\phi'_\pm(t), -\phi'_\pm(-t)\}$ for $t > 0$. The only thing we need to prove is that

$$\begin{aligned} &Q\big((\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})\big) - Q\big((\mathcal{K}_{P_X} f_\lambda)(\mathbf{x})\big) \\ &\quad \le P\big(|(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})|\big)\big|(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x}) - (\mathcal{K}_{P_X} f_\lambda)(\mathbf{x})\big| \end{aligned}$$

holds for those $\mathbf{x}$ with $Q\big((\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})\big) - Q\big((\mathcal{K}_{P_X} f_\lambda)(\mathbf{x})\big) > 0$. By the deduction conducted above, such a point $\mathbf{x}$ must satisfy $(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x}) \notin [(\mathcal{K}_{P_X} f_\lambda^-)(\mathbf{x}), (\mathcal{K}_{P_X} f_\lambda^+)(\mathbf{x})]$.

$(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x}) > (\mathcal{K}_{P_X} f_\lambda^+)(\mathbf{x})$, which means that Q is strictly increasing on $[(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x}), +\infty]$; hence, $(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x}) > (\mathcal{K}_{P_X} f_\lambda)(\mathbf{x})$. Then we have

$$\begin{aligned} &Q\big((\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})\big) - Q\big((\mathcal{K}_{P_X} f_\lambda)(\mathbf{x})\big) \\ &\quad \le Q'_-\big((\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})\big)\big((\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x}) - (\mathcal{K}_{P_X} f_\lambda)(\mathbf{x})\big). \end{aligned}$$

Because ϕ is convex, the one-side derivatives ϕ'_- and ϕ'_+ exist, are nondecreasing, and satisfy $\phi'_-(t) \le \phi'_+(t)$ for every $t \in \mathbb{R}$. Consider

$$Q(t) = \eta(\mathbf{x})\phi(t) + (1 - \eta(\mathbf{x}))\phi(-t).$$

Then

$$\begin{aligned} Q'_-(\mathcal{K}_{\hat{P}_X} f_\lambda) &= \eta(\mathbf{x})\phi'_-((\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})) - (1 - \eta(\mathbf{x}))\phi'_+(-(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})) \\ &\le \max\{\phi'_\pm(|(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})|), -\phi'_\pm(-|(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})|)\}. \end{aligned}$$

For $(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x}) < (\mathcal{K}_{P_X} f_\lambda^-)(\mathbf{x})$, Q is strictly decreasing on $(-\infty, (\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})]$. Hence, $(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x}) < (\mathcal{K}_{P_X} f_\lambda)(\mathbf{x})$, and we have

$$\begin{aligned} &Q\big((\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})\big) - Q\big((\mathcal{K}_{P_X} f_\lambda)(\mathbf{x})\big) \\ &\quad \le -Q'_+\big((\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})\big)((\mathcal{K}_{P_X} f_\lambda)(\mathbf{x}) - (\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})). \end{aligned}$$

Because

$$\begin{aligned} -Q'_+\big((\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})\big) &= -\eta(\mathbf{x})\phi'_+((\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})) + (1 - \eta(\mathbf{x}))\phi'_-(-(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})) \\ &\le P(|(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})|), \end{aligned}$$

we find that the previous equation still holds.

Proposition 3. *For any $0 < \delta < 1$, with confidence at least $1 - \delta$, we have*

$$\mathcal{H}(z, \lambda) \le \kappa \|\phi'\|_\infty \sqrt{\frac{\mathcal{D}(\lambda)}{\lambda}} \left(\frac{2\log(1/\delta)}{l+u} + \sqrt{\frac{2\log(1/\delta)}{l+u}} \right). \tag{3.7}$$

Proof. According to lemma 4, we have

$$\mathcal{E}(\mathcal{K}_{\hat{P}_X} f_\lambda) - \mathcal{E}(\mathcal{K}_{P_X} f_\lambda) \le \|\phi'\|_\infty \|\mathcal{K}_{\hat{P}_X} f_\lambda - \mathcal{K}_{P_X} f_\lambda\|.$$

Let $\zeta(\mathbf{x}_i) = f_\lambda(\mathbf{x}_i)K(\cdot, \mathbf{x}_i)$, which is a continuous and bounded variable on X. Then

$$\mathcal{K}_{\hat{P}_X} f_\lambda = \frac{1}{l+u} \sum_{i=1}^{l+u} \zeta(\mathbf{x}_i)$$

and

$$\mathcal{K}_{P_X} f_\lambda = \int K(\cdot, t) f_\lambda(t) d\rho_X(t) = \mathbb{E}\zeta.$$

Since $|\zeta| \le \kappa \|f_\lambda\|_K \le \kappa\sqrt{\mathcal{D}(\lambda)/\lambda}$, $\sigma^2(\zeta) \le \kappa^2 \mathcal{D}(\lambda)/\lambda$ and applying lemma 3 to random variable ζ, we obtain

$$\begin{aligned}\mathcal{H}(\mathbf{z}, \lambda) &\le \|\phi'\|_\infty \left(\frac{1}{l+u} \sum_{i=1}^{l+u} \zeta(\mathbf{x}_i) - \mathbb{E}\zeta \right) \\ &\le \kappa \|\phi'\|_\infty \sqrt{\frac{\mathcal{D}(\lambda)}{\lambda}} \left(\frac{2\log(1/\delta)}{l+u} + \sqrt{\frac{2\log(1/\delta)}{l+u}} \right).\end{aligned}$$

□

3.2 Estimation of Sample Error. In this section, we focus on bounding the sample error. It should be noted that the estimation of sample error $\mathcal{S}(\mathbf{z}, \lambda)$ involves the sample $\mathbf{z}$ and thus runs over a set of functions. Hence, we introduce the following two inequalities (Wu et al., 2007; Zhou & Jetter, 2006) to measure the uniform concentration estimate.

Lemma 5. *Let $\mathcal{F}$ be a set of measurable functions on Z. Assume that there are constants $B, c > 0$ and $\alpha \in [0, 1]$ such that $\|f\|_\infty \le B$ and $\mathbb{E}f^2 \le c(\mathbb{E}f)^\tau$ for every $f \in \mathcal{F}$. If for some $a > 0$ and $s \in (0, 2)$, $log\mathcal{N}_2(\mathcal{F}, \epsilon) \le a\epsilon^{-s}$ for any $\epsilon > 0$, then there exists a constant C_s such that*

$$\begin{aligned}\mathbb{E}(f) - \frac{1}{m}\sum_{i=1}^{m} f(z_i) \le \frac{1}{2}\eta^{1-\tau}(\mathbb{E}f)^\tau + C_s\eta + 2\left(\frac{c\,log(1/\delta)}{m}\right)^{\frac{1}{2-\tau}} \\ + \frac{18B\,log(1/\delta)}{m}\end{aligned}$$

holds with confidence at least $1 - \delta$, where

$$\eta = max\left\{ c^{\frac{2-s}{4-2\tau+\tau s}} \left(\frac{a}{m}\right)^{\frac{2}{4-2\tau+\tau s}}, \quad B^{\frac{2-s}{2+s}} \left(\frac{a}{m}\right)^{\frac{2}{2+s}} \right\}.$$

Lemma 6. *Let $0 \le \tau \le 1$, $c \ge 0$, and $\mathcal{G}$ be a set of functions on Z such that for every $g \in \mathcal{G}$, $\mathbb{E}(g) \ge 0$, $\|g - \mathbb{E}(g)\| \le B$ and $\mathbb{E}(g^2) \le c(\mathbb{E}(g))^\tau$. Then for any $\varepsilon > 0$,*

$$\begin{aligned}&Prob_{z\in Z^m}\left\{ \sup_{g\in\mathcal{G}} \frac{\mathbb{E}(g) - \frac{1}{m}\sum_{i=1}^{m} g(z_i)}{\sqrt{(\mathbb{E}(g))^\tau + \varepsilon^\tau}} \ge 4\varepsilon^{1-\tau/2} \right\} \\ &\le \mathcal{N}(\mathcal{G}, \varepsilon) exp\left\{ -\frac{m\varepsilon^{2-\tau}}{2\left(c + \frac{1}{3}B\varepsilon^{1-\tau}\right)} \right\}.\end{aligned}$$

According to the definition of $f_{\mathbf{z}}$, by taking $f = 0$ in equation 2.10, we can see $\lambda \| f_{\mathbf{z}} \|_{\mathcal{H}}^2 \leq \phi(0)$. Hence, $\| f_{\mathbf{z}} \|_{\mathcal{H}} \leq \sqrt{\frac{\phi(0)}{\lambda}} := \bar{R}$.

Proposition 4. *Suppose normalized classifying loss ϕ satisfies the increment condition with exponent $q > 0$ and a constant $c_q > 0$, $\left|\phi(t)\right| \leq c_q |t|^q$, $\forall t \geq 1$. Under assumptions 1, 2, and 3, for any $0 < \delta < 1$, with confidence at least $1 - \delta$, there holds*

$$\mathcal{S}(z, \lambda) \leq \frac{1}{2}\eta^{1-\tau}(\mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{E}(f_\rho))^\tau + C_s \eta + \frac{1}{2}\mathcal{H}(z, \lambda) + \frac{1}{2}\mathcal{D}(\lambda)$$
$$+2\left(\frac{C_\tau \, log(1/\delta)}{l}\right)^{\frac{1}{2-\tau}} + \frac{18 \, log(1/\delta)}{l} + 20\bar{\varepsilon}, \tag{3.8}$$

where

$$\eta = max\left\{ C_\tau^{\frac{2-s}{4-2\tau+\tau s}} \left(\frac{C_s(\kappa \bar{R} C_0)^s}{l}\right)^{\frac{2}{4-2\tau+\tau s}}, \quad \left(\frac{C_s(\kappa \bar{R} C_0)^s}{l}\right)^{\frac{2}{2+s}} \right\}$$

and

$$\bar{\varepsilon} = \frac{4c_q \kappa^q \, log(1/\delta)}{3l}\left(\frac{\mathcal{D}(\lambda)}{\lambda}\right)^{\frac{q}{2}} + \left(\frac{4C_1 \, log(1/\delta)}{l}\right)^{\frac{1}{2-\tau}}$$
$$+ 4C_p(\kappa \bar{C}\sqrt{\mathcal{D}(\lambda)/\lambda})^p \cdot l^{-\gamma}.$$

Proof. We divide the sample error into two parts,

$$\mathcal{S}(\mathbf{z}, \gamma) = \mathcal{S}_1(\mathbf{z}, \lambda) + \mathcal{S}_2(\mathbf{z}, \lambda)$$
$$:= \left\{ \mathbb{E}(\xi_1) - \frac{1}{l}\sum_{i=1}^{l} \xi_1(\mathbf{z}_i) \right\} + \left\{ \frac{1}{l}\sum_{i=1}^{l} \xi_2(\mathbf{z}_i) - \mathbb{E}(\xi_2) \right\}$$

where $\xi_1 = \phi(y\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})(\mathbf{x})) - \phi(y f_\rho(\mathbf{x}))$, $\xi_2 = \phi(y(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})) - \phi(y f_\rho(\mathbf{x}))$.

We first estimate $\mathcal{S}_2(\mathbf{z}, \lambda)$. Denote $\mathcal{G} = \{g_\lambda : g_\lambda(\mathbf{x}) = \mathcal{K}_{\hat{P}_X} f_\lambda(\mathbf{x}), \mathbf{x} \in X\}$. According to the definition of f_λ, we know that for $g_\lambda \in \mathcal{G}$, $g_\lambda \in \bar{\mathcal{B}}_R$ with $R = \kappa\sqrt{\mathcal{D}(\lambda)/\lambda}$. Let

$$\tilde{\mathcal{H}} := \left\{ h(\mathbf{z}) = \phi(y(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x})) - \phi(y f_\rho(\mathbf{x})), \forall \mathbf{z} \in \{(\mathbf{x}_i, y_i)\}_{i=1}^{l} \cup \{\mathbf{x}_j\}_{j=l+1}^{l+u} \right\},$$

since f_ρ is restricted in $[-1, 1]$. Then we have $|h(\mathbf{z})| \le c_q \kappa^q (\mathcal{D}(\lambda)/\lambda)^{q/2} + 1 := B_\lambda + 1$ and $\|h - \mathbb{E}(h)\| \le 2(B_\lambda + 1)$. Accordingly,

$$\begin{aligned}\|h_1(\mathbf{z}) - h_2(\mathbf{z})\|_\infty &:= \sup_{\mathbf{z}\in Z} \left|\phi(y(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x}_1)) - \phi(y(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x}_2))\right| \\ &\le \|\phi'_-(-1)\| \cdot \|(\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x}_1) - (\mathcal{K}_{\hat{P}_X} f_\lambda)(\mathbf{x}_2)\|_\infty \\ &\le \bar{C} \cdot \kappa \sqrt{\mathcal{D}(\lambda)/\lambda} \cdot \|f_\lambda(\mathbf{x}_1) - f_\lambda(\mathbf{x}_2)\|_\infty.\end{aligned}$$

In connection with assumption 1, this means that

$$\log \mathcal{N}_\infty(\tilde{\mathcal{H}}, \varepsilon) \le \log \mathcal{N}_\infty\left(\bar{\mathcal{B}}_1, \frac{\varepsilon}{\bar{C} \cdot \kappa \sqrt{\mathcal{D}(\lambda)/\lambda}}\right) \le C_p \cdot (\kappa \bar{C} \sqrt{\mathcal{D}(\lambda)/\lambda})^p \varepsilon^{-p}.$$

By lemma 6, we obtain

$$\begin{aligned}&\text{Prob}\left\{\sum_{f\in\bar{\mathcal{B}}_R} \frac{\mathbb{E}(h) - \frac{1}{l}\sum_{i=1}^l h(\mathbf{z}_i)}{\sqrt{(\mathbb{E}(h))^\tau + \varepsilon^\tau}} \ge 4\varepsilon^{1-\tau/2}\right\} \\ &\le \exp\left\{C_p(\kappa\bar{C}\sqrt{\mathcal{D}(\lambda)/\lambda})^p \varepsilon^{-p} - \frac{l\varepsilon^{2-\tau}}{2(C_1 + \frac{1}{3}B_\lambda \varepsilon^{1-\tau})}\right\}.\end{aligned}$$

Assume $\bar{\varepsilon}$ is the positive solution of the following equation:

$$\varphi(\varepsilon) := C_p(\kappa\bar{C}\sqrt{\mathcal{D}(\lambda)/\lambda})^p \varepsilon^{-p} - \frac{l\varepsilon^{2-\tau}}{2(C_1 + \frac{1}{3}B_\lambda \varepsilon^{1-\tau})} = \log(\delta).$$

Then one can check that $\varphi : \mathbb{R}_+ \to \mathbb{R}$ is strictly decreasing. Hence, $\bar{\varepsilon} \le \varepsilon^*$ if $\varphi(\varepsilon^*) \le \log(\delta)$.

Denote $\gamma := \frac{1}{2-\tau+p} \in (0, 1)$. If we take ε^* to be a positive number that satisfies $\varepsilon^* > l^{-\gamma}$ and

$$C_p(\kappa\bar{C}\sqrt{\mathcal{D}(\lambda)/\lambda})^p \varepsilon^{\gamma p} - \frac{l\varepsilon^{2-\tau}}{2(C_1 + \frac{1}{3}B_\lambda \varepsilon^{1-\tau})} \le \log(\delta),$$

we have $\varphi(\varepsilon^*) \le \log(\delta)$.

Note that the inequality satisfied by ε^* can be written as

$$\begin{aligned}&\varepsilon^{2-\tau} - \frac{2/3B_\lambda \log(1/\delta)}{l} \cdot \varepsilon^{1-\tau} - \frac{2C_1 \log(1/\delta) + 2C_p(\kappa\bar{C}\sqrt{\mathcal{D}(\lambda)/\lambda})^p \cdot l^{\gamma p}}{l} \\ &\ge 0.\end{aligned}$$

According to lemma 7 (Cucker & Smale, 2002), we can choose

$$\varepsilon^* = \max\left\{\frac{4B_\lambda \log(1/\delta)}{3l}, \left(\frac{4C_1 \log(1/\delta) + 4C_p(\kappa\bar{C}\sqrt{\mathcal{D}(\lambda)/\lambda})^p \cdot l^{\gamma p}}{l}\right)^{\frac{1}{2-\tau}}\right\} \geq l^{-\gamma},$$

which indicates that

$$\bar{\varepsilon} \leq \frac{4B_\lambda \log(1/\delta)}{3l} + \left(\frac{4C_1 \log(1/\delta)}{l}\right)^{\frac{1}{2-\tau}} + 4C_p(\kappa\bar{C}\sqrt{\mathcal{D}(\lambda)/\lambda})^p \cdot l^{-\gamma}.$$

Applying lemma 6 with $\varepsilon = \bar{\varepsilon}$, we know that for all $f_\lambda \in \bar{\mathcal{B}}_R$,

$$\mathcal{S}_2(\mathbf{z}, \lambda) \leq 4\bar{\varepsilon}^{1-\tau/2}\left(\mathcal{E}(\mathcal{K}_{\hat{P}_X} f_\lambda) - \mathcal{E}(f_\rho)\right)^{\tau/2} + 4\bar{\varepsilon}.$$

Recall the elementary inequality:

$$\frac{1}{a} + \frac{1}{b} = 1 \text{ with } a, b > 1 \Rightarrow c \cdot d \leq \frac{1}{a}c^a + \frac{1}{b}d^b \ \forall c, d \geq 0.$$

Using this for $a = \frac{2}{\tau}$, $b = \frac{2}{2-\tau}$, $\tau \leq 1$, we can find

$$\mathcal{S}_2(\mathbf{z}, \lambda) \leq \frac{1}{2}\left(\mathcal{E}(\mathcal{K}_{\hat{P}_X} f_\lambda) - \mathcal{E}(f_\rho)\right) + 20\bar{\varepsilon}.$$

By the fact that

$$\begin{aligned}\mathcal{E}(\mathcal{K}_{\hat{P}_X} f_\lambda) - \mathcal{E}(f_\rho) &= \mathcal{E}(\mathcal{K}_{\hat{P}_X} f_\lambda) - \mathcal{E}(\mathcal{K}_{P_X} f_\lambda) + \mathcal{E}(\mathcal{K}_{P_X} f_\lambda) - \mathcal{E}(f_\rho) \\ &\leq \mathcal{H}(\mathbf{z}, \lambda) + \mathcal{D}(\lambda),\end{aligned}$$

we get

$$\mathcal{S}_2(\mathbf{z}, \lambda) \leq \frac{1}{2}\mathcal{H}(\mathbf{z}, \lambda) + \frac{1}{2}\mathcal{D}(\lambda) + 20\bar{\varepsilon}. \tag{3.9}$$

We now focus on estimating $\mathcal{S}_1(\mathbf{z}, \lambda)$. Let $\mathcal{F}_{\bar{R}} := \{\phi(y\pi(\mathcal{K}_{\hat{P}_X} f)(\mathbf{x})) - \phi(y f_\rho(\mathbf{x})), f \in \mathcal{B}_{\bar{R}}\}$, and $A = \phi(y\pi(\mathcal{K}_{\hat{P}_X} f)(\mathbf{x})) - \phi(y f_\rho(\mathbf{x}))$. Then

$$\mathbb{E}(A) = \mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f)) - \mathcal{E}(f_\rho), \text{ and } \frac{1}{l}\sum_{i=1}^{l} A(\mathbf{z}_i) = \mathcal{E}_{\mathbf{z}}(\pi(\mathcal{K}_{\hat{P}_X} f)) - \mathcal{E}_{\mathbf{z}}(f_\rho).$$

Since $-1 \le y\pi(\mathcal{K}_{\hat{P}_X} f)(\mathbf{x}) \le 1$, the monotonicity of ϕ tells us that

$$-\phi(-1) \le -\phi(yf_\rho(\mathbf{x})) \le A(\mathbf{z}) \le \phi(y\pi(\mathcal{K}_{\hat{P}_X} f)(\mathbf{x})) \le \phi(-1).$$

Let $C_0 = 2|\phi(-1)|$. Then for any $f_1, f_2 \in \mathcal{B}_{\bar{R}}$, there exists

$$\begin{aligned}
&|\phi(y\pi(\mathcal{K}_{\hat{P}_X} f_1)(\mathbf{x})) - \phi(yf_\rho(\mathbf{x}) - \{\phi(y\pi(\mathcal{K}_{\hat{P}_X} f_2)(\mathbf{x})) - \phi(yf_\rho(\mathbf{x})\}| \\
&\le \frac{C_0}{l+u} |\sum_{i=1}^{l+u} (f_1(\mathbf{x}_i) - f_2(\mathbf{x}_i))K(\mathbf{x}, \mathbf{x}_i)| \\
&\le \kappa C_0 \|f_1 - f_2\|_\infty,
\end{aligned}$$

which implies that

$$\log \mathcal{N}_\infty(\mathcal{F}_{\bar{R}}, \varepsilon) \le \log \mathcal{N}_\infty\left(\mathcal{B}_{\bar{R}}, \frac{\varepsilon}{\kappa \bar{R} C_0}\right) \le C_S(\kappa \bar{R} C_0)^s \varepsilon^{-s}.$$

Assumption 3 tells us that $\mathbb{E}(h^2) \le C_\tau(\mathbb{E}h)^\tau$. Then by lemma 5, the following holds with confidence at least $1 - \delta$,

$$\begin{aligned}
\mathcal{S}_1(\mathbf{z}, \lambda) &= \mathbb{E}(\xi_1) - \frac{1}{l}\sum_{i=1}^{l} \xi_1(\mathbf{z}_i) \\
&\le \frac{1}{2}\eta^{1-\tau}\left(\mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{E}(f_\rho)\right)^\tau + C_s\eta \\
&\quad + 2\left(\frac{C_\tau \log(1/\delta)}{l}\right)^{\frac{1}{2-\tau}} + \frac{18\log(1/\delta)}{l}, \qquad (3.10)
\end{aligned}$$

where

$$\eta = \max\left\{C_\tau^{\frac{2-s}{4-2\tau+\tau s}}\left(\frac{C_s(\kappa\bar{R}C_0)^s}{l}\right)^{\frac{2}{4-2\tau+\tau s}}, \quad \left(\frac{C_s(\kappa\bar{R}C_0)^s}{l}\right)^{\frac{2}{2+s}}\right\}.$$

Combining equations 3.9 and 3.10, we obtain the desired results. □

Proof of Theorem 1. Combining propositions 2, 3, and 4, with confidence at least $1 - 3\delta$, we have

$$
\begin{aligned}
&\mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \\
&\le \frac{1}{2}\eta^{1-\tau}\left(\mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{E}(f_\rho)\right)^{\tau} + C_s\eta + \frac{3}{2}\mathcal{D}(\lambda) \\
&\quad + 2\left(\frac{C_\tau \log(1/\delta)}{l}\right)^{\frac{1}{2-\tau}} + \frac{18\log(1/\delta)}{l} + 20\bar{\varepsilon} \\
&\quad + \frac{3}{2}\kappa\|\phi'\|_\infty\sqrt{\frac{\mathcal{D}(\lambda)}{\lambda}}\left(\frac{2\log(1/\delta)}{l+u} + \sqrt{\frac{2\log(1/\delta)}{l+u}}\right),
\end{aligned} \tag{3.11}
$$

where $\bar{\varepsilon}$ is given in proposition 4. When $\tau = 1$, with confidence at least $1 - 3\delta$, there holds

$$
\begin{aligned}
\mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le 2C_s\eta + 3\mathcal{D}(\lambda) + \frac{(4C_\tau + 36)\log(1/\delta)}{l} + 40\bar{\varepsilon} \\
+ 3\kappa\|\phi'\|_\infty\sqrt{\frac{\mathcal{D}(\lambda)}{\lambda}}\left(\frac{2\log(1/\delta)}{l+u} + \sqrt{\frac{2\log(1/\delta)}{l+u}}\right).
\end{aligned}
$$

When $0 < \tau < 1$, we find equation 3.11 still holds by the elementary inequality: if $a, b > 0$, then

$$
x \le ax^\tau + b,\ x > 0 \Rightarrow x \le \max\{(2a)^{\frac{1}{1-\tau}}, 2b\}.
$$

Considering $f_{\mathbf{z}} \in \mathcal{B}_{\bar{R}}$ with $\bar{R} = \sqrt{\frac{\phi(0)}{\lambda}}$ and setting $C' = 2C_S \max\{C_\tau^{\frac{2-s}{4-2\tau+\tau s}}, 1\}(\kappa C_0)^{\frac{2-s}{4-2\tau+\tau s}}$, and $C'' = 160C_p(\kappa\bar{C})^p c_\beta^{p/2}$, the following inequality holds,

$$
\begin{aligned}
&\mathcal{E}(\pi(\mathcal{K}_{\hat{P}_X} f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \\
&\le C'\lambda^{-\frac{s}{4-2\tau+\tau s}} l^{-\frac{2}{4-2\tau+\tau s}} + 3c_\beta\lambda^\beta + \frac{(4C_\tau + 36)\log(1/\delta)}{3l} \\
&\quad + \frac{160c_q\kappa^q c_\beta^{q/2}\log(1/\delta)}{3}\lambda^{\frac{(\beta-1)q}{2}} l^{-1} + C''\lambda^{\frac{(\beta-1)p}{2}} l^{-\frac{1}{2-\tau+p}} \\
&\quad + 2\kappa\|\phi'\|_\infty \log(1/\delta)\lambda^{\frac{\beta-1}{2}}(l+u)^{-1} + 2\kappa\|\phi'\|_\infty\lambda^{\frac{\beta-1}{2}}(l+u)^{-\frac{1}{2}} \\
&\le \hat{C}\log(3/\delta)\Big(\lambda^{-\frac{s}{4-2\tau+\tau s}} l^{-\frac{2}{4-2\tau+\tau s}} + l^{-1} + \lambda^\beta + \lambda^{\frac{(\beta-1)q}{2}} l^{-1} \\
&\quad + \lambda^{\frac{(\beta-1)p}{2}} l^{-\frac{1}{2-\tau+p}} + \lambda^{\frac{\beta-1}{2}}(l+u)^{-1} + \lambda^{\frac{\beta-1}{2}}(l+u)^{-\frac{1}{2}}\Big)
\end{aligned}
$$

with probability at least $1-\delta$, where $\hat{C}$ is a positive constant independent of l, λ, δ. □

Proof of Theorem 2. Theorem 1 tells us that with confidence $1-\delta$,

$$\begin{aligned}&\mathcal{E}(\pi(\mathcal{K}_{\hat{p}_X} f_{\mathbf{z}})) - \mathcal{E}(f_\rho)\\ \le& \hat{C}\log(3/\delta)\Big(\lambda^{-\frac{s}{4-2\tau+\tau s}} l^{-\frac{2}{4-2\tau+\tau s}} + l^{-1} + \lambda^{\beta} + \lambda^{\frac{(\beta-1)q}{2}} l^{-1}\\ &+ \lambda^{\frac{(\beta-1)p}{2}} l^{-\frac{1}{2-\tau+p}} + \lambda^{\frac{\beta-1}{2}}(l+u)^{-1} + \lambda^{\frac{\beta-1}{2}}(l+u)^{-\frac{1}{2}}\Big).\end{aligned}$$

Since $u \gg l$, without loss of generality, assume that $u = \mathcal{O}(l^2)$ and let

$$\theta = \min\left\{\frac{2\beta}{(4-2\tau+\tau s)\beta+s}, \quad \frac{2\beta}{1+\beta}, \quad \frac{2\beta}{(2\beta-\beta p+p)(2-\tau+p)}\right\}.$$

Taking $\lambda = \lambda(l) = l^{-\theta}$, we can verify that

$$\begin{aligned}&\lambda^{-\frac{s}{4-2\tau+\tau s}} l^{-\frac{2}{4-2\tau+\tau s}} \le \lambda^{\beta}, \quad \lambda^{\frac{(\beta-1)q}{2}} l^{-1} \le \lambda^{\beta},\\ &\lambda^{\frac{\beta-1}{2}}(l+u)^{-1} \le \lambda^{\beta}, \quad \lambda^{\frac{(\beta-1)p}{2}} l^{-\frac{1}{2-\tau+p}} \le \lambda^{\beta}, \quad \lambda^{\frac{\beta-1}{2}}(l+u)^{-\frac{1}{2}} \le \lambda^{\beta}.\end{aligned}$$

Hence, with confidence $1-\delta$, we have

$$\mathcal{E}(\pi(\mathcal{K}_{\hat{p}_X} f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le 7\hat{C}\log(3/\delta) l^{-\theta}.$$

□

4 Conclusion

This letter investigates the generalization performance of Fredholm kernel regularized classifiers. Convergence analysis shows that the fast learning rate with $O(l^{-1})$ can be reached under mild conditions for a family of classification algorithms with a Fredholm kernel. It will be interesting to explore fast optimization and a distributed framework for Fredholm kernel learning with big data.

Appendix: Proof of Representer Theorem

Here we focus on only the case for $\phi = \phi_h$; the proofs for the other cases are similar. Define the empirical loss for the learning problem:

$$L(f_{\mathbf{z}}) = \min_{f\in\mathcal{H}} L(f) = \min_{f\in\mathcal{H}} \left\{\frac{1}{l}\sum_{i=1}^{l} \phi(y_i \cdot (K_{\hat{p}_X} f)(\mathbf{x}_i)) + \lambda \|f\|_{\mathcal{H}}^2\right\}.$$

We first project f onto the subspace of $\mathcal{H}$, $\text{span}\{K_{\mathcal{H}}(\mathbf{x}_i, \cdot) : 1 \le i \le l+u\}$, which is spanned by a kernel function centered at the data points. Then we can obtain the orthogonal decomposition,

$$f_{\mathbf{z}} = f_S + f_\perp,$$

where $f_S \in \mathcal{H}$ is the component along the subspace and $f_\perp$ is the component orthogonal to the subspace. By definition, $f_\perp(\mathbf{x}_i) = \langle f_\perp(\mathbf{x}), K_{\mathcal{H}}(\mathbf{x}, \mathbf{x}_i)\rangle_K = 0$ for $i = 1, 2, \cdots, l+u$. Thus, the empirical loss can be expressed as

$$\begin{aligned}
&\frac{1}{l}\sum_{i=1}^{l} \phi(y_i \cdot (K_{\hat{P}_X} f_{\mathbf{z}})(\mathbf{x}_i)) \\
&= \frac{1}{l}\sum_{i=1}^{l} (1 - y_i \cdot (K_{\hat{P}_X} f_{\mathbf{z}})(\mathbf{x}_i))_+ \\
&= \frac{1}{l}\sum_{i=1}^{l} \left(1 - y_i \cdot \frac{1}{l+u}\sum_{j=1}^{l+u} K_H(\mathbf{x}_i, \mathbf{x}_j)(f_S(\mathbf{x}_j) + f_\perp(\mathbf{x}_j))\right)_+ \\
&= \frac{1}{l}\sum_{i=1}^{l} \left(1 - y_i \cdot \frac{1}{l+u}\sum_{j=1}^{l+u} K_H(\mathbf{x}_i, \mathbf{x}_j) f_S(\mathbf{x}_j)\right)_+ \\
&= \frac{1}{l}\sum_{i=1}^{l} \phi(y_i \cdot (K_{\hat{P}_X} f_S)(\mathbf{x}_i)).
\end{aligned}$$

Hence, the orthogonal component of $f_{\mathbf{z}}$ does not serve any function in empirical risk function. For the regularization term, since

$$\|f_{\mathbf{z}}\|_{\mathcal{H}}^2 = \|f_S\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2,$$

then $\|f_S\|_{\mathcal{H}}^2 \le \|f_{\mathbf{z}}\|_{\mathcal{H}}^2$. By combining the results above, we have

$$\begin{aligned}
L(f_S) &= \frac{1}{l}\sum_{i=1}^{l} \phi(y_i \cdot (K_{\hat{P}_X} f_S)(\mathbf{x}_i)) + \lambda(\|f_S\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2) \\
&\le L(f_{\mathbf{z}}),
\end{aligned}$$

which implies that $L(f)$ is minimized if f lies in the subspace and $f_S = f_{\mathbf{z}}$. The conclusion holds true. □

Acknowledgments

Two anonymous referees carefully read the manuscript for this letter and provided numerous constructive suggestions. As a result, the overall quality of the letter has been noticeably enhanced; we are much indebted to these referees and are grateful for their help. The research was partially supported by National 973 Programming (2013CB329404), the National Natural Science Foundation of China (11671161, 61673015, 11131006).

References

Aronszajn (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, *68*, 337–404.

Bartlett, P., & Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, *3*, 463–482.

Belkin, M., & Niyogi, P. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, *7*, 2399–2434.

Chen, D., Wu, Q., Ying, Y., & Zhou, D. (2004). Support vector machine soft margin classifiers: Error analysis. *J. Mach. Learn. Res.*, *5*, 1143–1175.

Chen, H., Pan, Z., Li, L., & Tang Y., (2013). Learning rates of coefficient-based regularized classifier for density level detection. *Neural Comput.*, *25*, 1107–1121.

Chen, H., Zhou, Y., Tang, Y., Li, L., & Pan, Z. (2013). Convergence rate of semisupervised greedy algorithm. *Neural Networks*, *44*, 44–50.

Cucker, F., & Smale, S. (2001). On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, *39*, 1–49.

Cucker, F., & Smale, S. (2002). Best choices for regularization parameters in learning theory: On the bias-variance problem. *Found. Comput. Math.*, *1*, 413–428.

Johnson, R., & Zhang, T. (2007). On the effectiveness of Laplacian normalization for graph semi-supervised learning. *J. Mach. Learn. Res.*, *8*, 1489–1517.

Lee, W., Bartlett, P., & Williamson, R. (1996). Efficient agnostic learning of neural networks with bound fan-in. *IEEE. Trans. Inf. Theory*, *42*, 2118–2132.

Que, Q., & Belkin, M. (2013). Inverse density as an inverse problem: The Fredholm equation approach. In C. J. C. Burges, L. Bolou, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *26*. Red Hook, NY: Curran.

Que, Q., Belkin, M., & Wang, Y. (2014). Learning with Fredholm kernels. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *27*. Red Hook, NY: Curran.

Shi, L., Feng, Y., & Zhou, D. (2011). Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis space. *Appl. Comput. Harmonic. Anal.*, *31*, 286–302.

Smale, S., & Zhou, D. (2007). Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, *26*, 153–172.

Steinwart, I., & Scovel, C. (2005). Fast rates for support vector machines. In *Proceedings of the 18th Conference on Learning Theory.* New York: Springer.

Tsybakov, A. (2004). Optimal aggression of classifiers in statistical learning. *Ann. Statist.*, *32*, 135–166.

Wu, Q., Ying, Y., & Zhou, D. (2007). Multi-kernel regularized classifiers. *J. Complexity*, *23*, 108–134.

Wu, Q., & Zhou, D. (2005). SVM soft margin classifiers: Linear programming versus quadratic programming. *Neural Comp.*, *17*, 1160–1187.

Wu, Q., & Zhou, D. (2008). Learning with sample dependent hypothesis spaces. *Comput. Math. Appl.*, *56*, 2896–2907.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, *32*, 56–134.

Zhou, D. (2002). The covering number in learning theory. *J. Complexity*, *18*, 739–767.

Zhou, D. (2003). Capacity of reproducing kernel space in learning theory. *IEEE Trans. Inf. Theory*, *49*, 1743–1752.

Zhou, D., & Jetter, K. (2006). Approximation with polynomial kernels and SVM classifiers. *Adv. Comput. Math.*, *25*, 323–344.

Received September 24, 2016; accepted February 6, 2017.