Unified approach to coefficient-based regularized regression✩

Yun-Long Feng∗, Shao-Gao Lv

Joint Advanced Research Center, University of Science and Technology of China, China
City University of Hong Kong, Suzhou, Jiangsu, China

A R T I C L E  I N F O

Article history:
Received 6 January 2011
Received in revised form 28 March 2011
Accepted 16 May 2011

Keywords:
Data dependent hypothesis spaces
Coefficient-based regularization scheme
$l^q$-regularizer
$l^2$-empirical covering number
Learning rates

A B S T R A C T

In this paper, we consider the coefficient-based regularized least-squares regression problem with the $l^q$-regularizer $(1 \leq q \leq 2)$ and data dependent hypothesis spaces. Algorithms in data dependent hypothesis spaces perform well with the property of flexibility. We conduct a unified error analysis by a stepping stone technique. An empirical covering number technique is also employed in our study to improve sample error. Comparing with existing results, we make a few improvements: First, we obtain a significantly sharper learning rate that can be arbitrarily close to $O(m^{-1})$ under reasonable conditions, which is regarded as the best learning rate in learning theory. Second, our results cover the case $q = 1$, which is novel. Finally, our results hold under very general conditions.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

In this paper, we study coefficient-based regularization learning algorithms in a least-squares regression setting. Let $X$ be a compact metric space, $Y = \mathbb{R}$ and $\rho$ be an unknown distribution on $Z := X \times Y$. Regression algorithms aim at producing functions to approximate the regression function $f_\rho$ given by

$$f_\rho(x) = \int_Y y d\rho(y|x),$$

where $\rho(\cdot|x)$ is the conditional distribution of $\rho$ at $x \in X$.

We consider a learning algorithm generated by a coefficient-based regularization scheme in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ [1] associated with a Mercer kernel $K$, where $K : X \times X \rightarrow \mathbb{R}$ is a continuous, symmetric and positive semi-definite function. $\mathcal{H}_K$ is the completion of the linear span of functions $\{K_x = K(x, \cdot) : x \in X\}$, with the inner product for fundamental functions given by $\langle K_x, K_y \rangle_K = K(x, y)$. That is, $\langle \sum_i a_i K_{x_i}, \sum_j \beta_j K_{y_j} \rangle_K = \sum_i a_i \beta_j K(x_i, y_j)$. Moreover, following the reproducing property $f(x) = \langle f, K_x \rangle_K$, we know that for every $f \in \mathcal{H}_K$, $\|f\|_\infty \leq \kappa \|f\|_K$ where $\kappa = \sup_{x \in X} \sqrt{K(x,x)}$.

Given a sample $z := \{(x_i, y_i)\}_{i=1}^m$ drawn independently according to $\rho$, we learn the regression function by

$$f_z = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \eta \Omega(f) \right\}, \quad (1.1)$$

✩ The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 103709].
∗ Corresponding address: University of Science and Technology of China, Joint Advanced Research Center, 215123 Suzhou, Jiangsu, China. Tel.: +86 512 8716 1278.
E-mail addresses: feyl@mail.ustc.edu.cn (Y.-L. Feng), kenan716@mail.ustc.edu.cn (S.-G. Lv).

0898-1221/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2011.05.034
where $\Omega_2(f) = \inf \{ \sum_{i=1}^m |\alpha_i|^q : f = \sum_{i=1}^m a_i K_{x_i} \}, \eta = \eta(m) > 0$ is a regularization parameter, $1 \leq q \leq 2$ and 
\[
\mathcal{H}_{K, \mathcal{X}} = \left\{ \sum_{i=1}^m \alpha_i K_{x_i} : \alpha_i \in \mathbb{R}, i = 1, 2, \ldots, m \right\}.
\]

Throughout the paper, we assume that $\rho(\cdot|x)$ is supported on $[-M, M]$, for some $M > 0$ and each $x \in X$. Under this assumption, it is natural to apply a projection operator, which was introduced into learning algorithms to improve learning rates in [2–4].

**Definition 1.** The projection operator $\pi_M$ is defined on the space of measurable functions $f : X \to \mathbb{R}$ as
\[
\pi_M(f)(x) = \begin{cases} 
M, & \text{if } f(x) > M, \\
\; f(x), & \text{if } -M \leq f(x) \leq M, \\
-M, & \text{if } f(x) < -M.
\end{cases}
\]

The purpose of this paper is to conduct error analysis for the algorithm producing $\pi_M(f_x)$ and derive learning rates when the regularizer takes an $l^p$ norm with a general power index $1 \leq q \leq 2$.

Error analysis for the case $q = 1$ and $q = 2$ without projection has been presented in [5,6]. Moreover, the case $q = 1$ may yield a sparse property, as pointed out in [7]. Error analysis with projection was conducted in [8] when $1 < q \leq 2$ and learning rates of type $\mathcal{O}(m^{-\frac{q}{q-1}})$ were obtained. Their approach could not be extended to the case $q = 1$. To the best of our knowledge, there is no general error analysis that covers the case $1 \leq q \leq 2$. Motivated by this gap, we present our analysis.

In this paper, we elaborate our analysis by employing a stepping-stone approach [9,3] while different regularization parameters are adopted and exploiting an empirical covering number technique. The distance between $f_x$ and $f_p$ in $L^2_{\rho_X}$ space is adopted to measure the efficiency of algorithm (1.1), where $\|f\|_{\rho} = \|f\|_{L^2_{\rho_X}} = \left( \int_X |f(x)|^2 d\rho_X \right)^{\frac{1}{2}}$ and $\rho_X$ is the marginal distribution of $\rho$ on $X$. Let us illustrate our main contribution by a special case of our learning rate described in Theorem 2.

**Theorem 1.** Assume that $X$ is a compact subset of $\mathbb{R}^n$, $K \in C^\infty(\mathbb{R}^n \times \mathbb{R}^n)$, $f_p \in \mathcal{H}_K$. Let $1 \leq q \leq 2$. Then for any $0 < \varepsilon < \frac{1}{2}$ and $0 < \delta < 1$, choose
\[
\eta = \left( \frac{1}{m} \right)^{\frac{2q+2}{2q+1(q+1)}}\left( \frac{1}{m} \right)^{\frac{3q+2}{2q+q+1(q+1)}} \epsilon
\]
with confidence $1 - \delta$ there holds
\[
\|\pi_M(f_x) - f_p\|_{\rho}^2 \leq C \log\left( \frac{2}{\delta} \right) \left( \frac{1}{m} \right)^{\frac{3q+2}{2q+q+1(q+1)}} \epsilon
\]
where $C$ is a positive constant independent of $m$ or $\delta$.

As seen above, under the assumption of Theorem 1, we obtain the learning rate of type
\[
\mathcal{O}\left( m^{-\frac{q}{2q+q+1(q+1)}} \epsilon^{-1} \right).
\]
It can be arbitrary close to $\mathcal{O}(m^{-1})$ by choosing $\epsilon$ to be small enough, which is the best convergence rate in learning theory literature. Comparing with the rate in [8], which is $\mathcal{O}(m^{-\frac{1}{q-1}})$ under the same hypothesis, our learning rate is faster. Moreover, our result holds for the uniform range $q \in [1, 2]$.

The proof of Theorem 1 will be given in Section 3 where the constant $C$ is given explicitly. In Section 2, explanations on the capacity assumption and error analysis will be presented. In Section 3, we derive learning rates of algorithm (1.1) and we will also present our main results of this paper, as stated in Theorem 2. A comparison of learning rates and some discussions can be found in Section 4.

2. Error analysis

As pointed out in [3], algorithm (1.1) can be rewritten as
\[
f_x = f_{a_2} = \sum_{i=1}^m \alpha_{2i} K_{x_i},
\]
with $\alpha_{2} = (\alpha_{2,1}, \alpha_{2,2}, \ldots, \alpha_{2,m})^T$ given by
\[
\alpha_{2} = \arg \min_{\alpha \in \mathbb{R}^m} \left\{ \mathcal{E}_2 \left( \sum_{i=1}^m \alpha_i K_{x_i} \right) + \eta \sum_{i=1}^m |\alpha_i|^q \right\}.
\]
Here $\mathcal{E}_2(\sum_{i=1}^m \alpha_i K_{x_i}) = \frac{1}{m} \sum_{k=1}^m (\sum_{i=1}^m \alpha_i K(x_i, x_k) - y_k)^2$.
We use a stepping stone method as in [3,9] to conduct error analysis for algorithm (1.1). As we can see from (1.1), our learning algorithm works in data dependent spaces which is difficult to deal with, see e.g., [6,3]. We technically convert it into a data independent problem by introducing an empirical target function $f_{\kappa,\lambda}$, which plays a stepping stone role between $f_{\kappa}$ and regularization function $f_{\lambda}$. Here $f_{\kappa,\lambda}$ and $f_{\lambda}$ are given by

$$f_{\kappa,\lambda} = \arg \min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_E(f) + \lambda \|f\|^2_K \right\},$$

$$f_{\lambda} = \arg \min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) + \lambda \|f\|^2_K \right\},$$

where $\mathcal{E}(f) = \int_X (y-f(x))^2 d\rho$ and $\lambda = \lambda(m) > 0$ is another regularization parameter.

**Remark 1.** Note that the regularization parameter $\lambda$ above may be different from the regularization parameter $\eta$ as in algorithm (1.1). In fact, it is an improvement we make which will be proved to be effective. The parameter $\lambda$ will be selected when we derive learning rates in Section 3.

### 2.1. Error decomposition

The following error decomposition scheme was proposed by [3].

**Proposition 1.** Let $\lambda > 0$, we have

$$\mathcal{E}(\pi_M(f_{\kappa})) - \mathcal{E}(f_{\rho}) \leq \mathcal{E}(\pi_M(f_{\kappa})) - \mathcal{E}(f_{\rho}) + \eta \Omega_2(f_{\kappa}) = \delta(z,\lambda) + \mathcal{H}(z,\lambda) + \mathcal{D}(\lambda),$$

where

$$\delta(z,\lambda) = \{ \mathcal{E}(\pi_M(f_{\kappa})) - \mathcal{E}(\pi_M(f_{\lambda})) \} + \{ \mathcal{E}(f_{\lambda}) - \mathcal{E}(f_{\rho}) \},$$

$$\mathcal{H}(z,\lambda) = \{ \mathcal{E}_E(f_{\kappa}) \} + \{ \mathcal{E}_2(f_{\lambda}) + \lambda \|f_{\lambda}\|^2_K \},$$

$$\mathcal{D}(\lambda) = \{ \mathcal{E}(f_{\lambda}) \} - \mathcal{E}(f_{\rho}) + \lambda \|f_{\lambda}\|^2_K.$$

(2.1)

**Remark 2.** $\delta(z,\lambda)$ is called sample error, $\mathcal{H}(z,\lambda)$ is called hypothesis error, caused by the data dependence space $\mathcal{H}_K$, which may not contain the regularizing function $f_{\lambda} \in \mathcal{H}_K$. $\mathcal{D}(\lambda)$ is called approximation error.

**Assumption 1.** $\mathcal{D}(\lambda) \leq c_{\beta} \lambda^{\beta}$ for some $c_{\beta} > 0$ and $0 < \beta \leq 1$.

According to (2.1), $\mathcal{D}(\lambda)$ can be expressed as

$$\mathcal{D}(\lambda) = \inf_{f \in \mathcal{H}_K} \{ \mathcal{E}(f) - \mathcal{E}(f_{\rho}) + \lambda \|f\|^2_K \}.$$

The decay of $\mathcal{D}(\lambda)$ when $\lambda \to 0$ is used to measure the approximation ability of the function space $\mathcal{H}_K$, which is well studied in [10–12]. Denote $L_K$ as the integral operator on $L^2_{\rho_X}$ defined by

$$L_K(f)(x) = \int_X K(x,y)f(y)d\rho_X(y), \quad x \in X, f \in L^2_{\rho_X}.$$

Then it was shown in [12] that Assumption 1 holds when $L_K^{-\beta/2}f_{\rho} \in L^2_{\rho_X}$ with $0 < \beta \leq 1$ and $c_{\beta} = \|L_K^{-\beta/2}f_{\rho}\|^2_{\rho}.$

### 2.2. Estimating sample error

This section is devoted to estimating the sample error $\delta(z,\lambda)$. As shown in Proposition 1

$$\delta(z,\lambda) = \delta_1(z,\lambda) + \delta_2(z,\lambda),$$

where

$$\delta_1(z,\lambda) = \{ \mathcal{E}_2(f_{\lambda}) - \mathcal{E}_2(f_{\rho}) \},$$

$$\delta_2(z,\lambda) = \{ \mathcal{E}(\pi_M(f_{\kappa})) - \mathcal{E}(f_{\rho}) \} - \{ \mathcal{E}_2(\pi_M(f_{\kappa})) - \mathcal{E}_2(f_{\rho}) \}.$$

$\delta_1(z,\lambda)$ can be easily bounded by applying the following one-side Bernstein type probability inequality, see e.g., [13,11].
**Lemma 1.** Let $\xi$ be a random variable on a probability space $Z$ with variance $\sigma^2$ satisfying $|\xi - \mathbb{E}\xi| \leq M$ for some constant $M$. Then for any $0 < \delta < 1$, we have

$$\frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mathbb{E}\xi \leq \frac{2M\log \frac{1}{\delta}}{3m} + \sqrt{\frac{2\sigma^2 \log \frac{1}{\delta}}{m}}.$$

Applying Lemma 1 to the random variable $\xi(z) = (y - f(x))^2 - (y - f_p(x))^2$, where $z = (x, y) \in Z$, we obtain the following upper bound of $\delta_1(z, \lambda)$.

**Proposition 2.** For any $0 < \delta < 1$, with confidence $1 - \delta/2$,

$$\delta_1(z, \lambda) \leq \frac{1}{2} D(\lambda) + \frac{7}{3m} \left( 3M + \kappa \sqrt{D(\lambda)/\lambda} \right)^2 \log \left( \frac{1}{\delta} \right).$$

**Proof.** Following the definition of $D(\lambda)$, we have

$$\|f_i\|_\infty \leq \kappa \|f_i\|_K \leq \kappa D(\lambda)/\lambda.$$

Consider the random variable $\xi(z)$ defined above and recall that $|f_p(x)| \leq M$ almost everywhere, one gets

$$|\xi(z)| \leq (3M + \|f_i\|_\infty)(M + \|f_i\|_\infty) \leq (3M + \kappa D(\lambda)/\lambda)^2 := \tau.$$

Moreover,

$$\mathbb{E}(\xi^2) = \int (2y - f(x) - f_p(x))^2 (f_p(x) - f(x))^2 \, d\rho \leq (3M + \kappa D(\lambda)/\lambda)^2 \mathbb{D}(\lambda).$$

So we have $|\xi - \mathbb{E}\xi| \leq 2\tau$ and $\sigma^2 \leq \mathbb{E}(\xi^2) \leq \tau D(\lambda)$ almost everywhere.

Applying Lemma 1 with $M_2 = 2\tau$ and $\sigma^2 = \tau D(\lambda)$, we get the desired estimation. □

Bounding $\delta_2(z, \lambda)$ is more difficult in the sense that it involves the complexity of the function space $\mathcal{H}_k$. For this purpose, we introduce an empirical covering number technique [14].

**Definition 2.** Let $(\mathcal{M}, d)$ be a pseudo-metric space and $S \subset \mathcal{M}$ a subset. For every $\varepsilon > 0$, the covering number $\mathcal{N}(S, \varepsilon, d)$ of $S$ with respect to $\varepsilon$ and $d$ is defined as the minimal number of balls of radius $\varepsilon$ whose union covers $S$, that is,

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

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

Instead of using the uniform covering number [15], we employ the $\ell^2$-empirical covering number to measure the capacity of $\mathcal{H}_k$. Denote $d_2$ as the normalized metric on the Euclidean space $\mathbb{R}^n$ defined by

$$d_2(a, b) = \left( \frac{1}{n} \sum_{i=1}^{n} |a_i - b_i|^2 \right)^{1/2} \quad \text{for } a = \{a_i\}_{i=1}^{n}, b = \{b_i\}_{i=1}^{n} \in \mathbb{R}^n.$$

**Definition 3.** Let $\mathcal{F}$ be a set of functions on $X, x = (x_i)_{i=1}^{n} \subset X^n$ and $\mathcal{F}|_X = \{(f(x_i))_{i=1}^{n} : f \in \mathcal{F}\} \subset \mathbb{R}^n$. Set $\mathcal{N}_{2, \mathcal{F}}(\mathcal{F}, \varepsilon) = \mathcal{N}(\mathcal{F}|_X, \varepsilon, d_2)$. The $\ell^2$-empirical covering number of $\mathcal{F}$ is defined by

$$\mathcal{N}_{2, \mathcal{F}}(\mathcal{F}, \varepsilon) = \sup_{n \in \mathbb{N}} \mathcal{N}_{2, \mathcal{F}}(\mathcal{F}, \varepsilon), \quad \varepsilon > 0.$$

Denote $B_R$ as the ball of radius $R$ with $R > 0$, where $B_R := \{f \in \mathcal{H}_k : \|f\|_K \leq R\}$. We need the following capacity assumption on $\mathcal{H}_k$, which holds for most RKHSs as explained in Section 2.3.

**Assumption 2.** There exists an exponent $p$, with $0 < p < 2$ and a constant $C_{p,K} > 0$ such that

$$\log \mathcal{N}_{2}(B_1, \varepsilon) \leq C_{p,K} \left( \frac{1}{\varepsilon} \right)^{p}, \quad \forall \varepsilon > 0, \quad \text{(2.2)}$$

where $B_1$ is the unit ball of $\mathcal{H}_k$ defined as above.
Our estimation of $\delta_2(z, \lambda)$ mainly relies on the following concentration inequality which can be found in [14] and the detailed proof therein.

**Lemma 2.** Let $\mathcal{F}$ be a class of measurable functions on $Z$. Assume that there are constants $B, c > 0$ and $\theta \in [0, 1]$ such that $\|f\|_\infty \leq B$ and $\mathbb{E}f^2 \leq c(\mathbb{E}f)^\theta$ for every $f \in \mathcal{F}$. If for some $a > 0$ and $p \in (0, 2)$,

$$\log \mathcal{N}_2(\mathcal{F}, \epsilon) \leq a e^{-p}, \quad \forall \epsilon > 0,$$

then there exists a constant $c_p$ depending only on $p$ such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds

$$\mathbb{E}f - \frac{1}{m} \sum_{i=1}^{m} f(z_i) \leq \frac{1}{2} t^{1-\theta}(\mathbb{E}f)^\theta + c_p t^2 + 2 \left( \frac{c t}{m} \right)^{\frac{1}{\theta}} + \frac{188 t}{m}, \quad \forall f \in \mathcal{F}, \quad (2.3)$$

where

$$t := \max \left\{ \sqrt[2-p]{2^{-2p/(1-p)}} \left( \frac{a M^2}{m} \right), B^{2-p} \left( \frac{a K^2}{m} \right)^{\frac{1}{2-p}} \right\}. $$

Applying Lemma 2 to the function set $F_R$ with $R > 0$ defined by

$$F_R := \{(y - \pi_M(f)(x))^2 - (y - f_p(x))^2 : f \in B_R \}.$$

**Proposition 3.** If $B_1$ satisfies the capacity condition (2.2), then for any $0 < \delta < 1$, with confidence $1 - \delta/2$, $(\mathbb{E}(f) - \mathbb{E}(f_p)) - (\mathbb{E}(f) - \mathbb{E}(f_p))$ can be bounded by

$$\frac{1}{2}(\mathbb{E}(f) - \mathbb{E}(f_p)) + \frac{320M^2}{m} \log \left( \frac{2}{\delta} \right) + C_{p,M,K} \left( \frac{1}{m} \right)^{2/p} R^{2p}, \quad \forall f \in B_R,$$

where $C_{p,M,K} = C_p(4M)^{\frac{4}{2p}} C_{p,K}^{\frac{2}{p}}.$

**Proof.** Let $g \in F_R.$ It follows that

$$\mathbb{E}g = \mathbb{E}(f) - \mathbb{E}(f_p) = \|f - f_p\|_p^2,$$

$$|g(z)| = |(f(x) - f_p(x))|(|f(x) - y| + |f_p(x) - y|) \leq 8M^2,$$

$$\mathbb{E}g^2 = \mathbb{E}(f(x) - f_p(x))^2((f(x) - y) + (f_p(x) - y))^2 \leq 16M^2 \mathbb{E}g.$$

If $g_1, g_2 \in F_R$ following Definition 1, one gets

$$|g_1(z) - g_2(z)| = |(y - f_1(x))^2 - (y - f_2(x))^2| \leq 4M |f_1(x) - f_2(x)|.$$

Therefore

$$\mathcal{N}_{2,x}(F_R, \epsilon) \leq \mathcal{N}_{2,x} \left( B_R, \frac{\epsilon}{4M} \right) \leq \mathcal{N}_{2,x} \left( B_1, \frac{\epsilon}{4MR} \right).$$

This in connection with Definition 3 implies

$$\log \mathcal{N}_2(F_R, \epsilon) \leq C_{p,K}(4M)^{p} R^{p} \left( \frac{1}{\epsilon} \right)^{p}. $$

Applying Lemma 2 with $B = a = 16M^2, \theta = 1$ and $a = C_{p,K}(4M)^{p} R^{p}$, it follows that

$$\mathbb{E}g - \frac{1}{m} \sum_{i=1}^{m} g(z_i) \leq \frac{\mathbb{E}g}{2} + \frac{320M^2}{m} \log \left( \frac{2}{\delta} \right) + C_{p,M,K} R^{2p} \left( \frac{1}{m} \right)^{2/p}, \quad \forall g \in F_R$$

where $C_{p,M,K} = C_p(4M)^{\frac{4}{2p}} C_{p,K}^{\frac{2}{p}}.$

Observe that $\frac{1}{m} \sum_{i=1}^{m} g(z_i) = \mathbb{E}_x(f) - \mathbb{E}_x(f_p)$, our assertion follows. $\square$
2.3. On empirical covering number

This subsection concerns the empirical covering number of hypothesis space $\mathcal{H}_K$. In learning literature, various measurements of evaluating the complexity of hypothesis space have been studied, including VC-dimension, Rademacher complexity, covering number, and entropy number. Among these, covering number is one of the most frequently used measurements, which is well understood [16,17,15,2] and can be found in a large learning literature [6,11,8,9].

Instead of using the uniform covering number, we utilize the empirical covering number in this paper, which is measured by empirical distances. As pointed out in [15,13], this may yield sharper bounds and better generalization performances, which was further proved by [18,14]. To further explain our capacity assumption, let us begin with a general definition of empirical covering number.

**Definition 4.** Let $\mathcal{F}$ be a class of functions on $X$ and $x = \{x_1, \ldots, x_m\} \subset X$. For $1 \leq p \leq \infty$, $f, g \in \mathcal{F}$ define

$$d_{p,x}(f, g) = \left\{ \frac{1}{m} \sum_{i=1}^{m} |f(x_i) - g(x_i)|^p \right\}^{1/p}.$$  

For every $\varepsilon > 0$, the covering number of $\mathcal{F}$ associated with $d_{p,x}$ is

$$\mathcal{N}_{p,x}(\mathcal{F}, \varepsilon) = \min \left\{ n \in \mathbb{N} : \exists f_j \in \mathcal{F} \text{ such that } \mathcal{F} = \bigcup_{j=1}^{n} \{ f \in \mathcal{F} : d_{p,x}(f, f_j) \leq \varepsilon \} \right\}.$$  

The $\ell^p$-empirical covering number of $\mathcal{F}$ is then defined by

$$\mathcal{N}_p(\mathcal{F}, \varepsilon) = \sup_{m \in \mathbb{N}, x \in X^m} \mathcal{N}_{p,x}(\mathcal{F}, \varepsilon).$$

Cases when $p = 1, 2$ and $\infty$ are frequently used in learning theory. Following **Definition 4** and denoting $\mathcal{N}(\mathcal{F}, \varepsilon)$ as the uniform covering number [15] defined with the metric $\| \cdot \|_\infty$, one can easily deduce that

$$\mathcal{N}_1(\mathcal{F}, \varepsilon) \leq \mathcal{N}_2(\mathcal{F}, \varepsilon) \leq \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \leq \mathcal{N}(\mathcal{F}, \varepsilon).$$

Discussions concerning the equivalence relations between the empirical covering number and the uniform covering number can be found in [16]. What is nice about the $\ell^2$-empirical covering number is that it enables one to derive refined probability inequalities by utilizing the entropy integral. Following from the finiteness of Dudley’s entropy integral

$$\int_0^1 \sqrt{\log \mathcal{N}_2(B_1, \eta)} d\eta < \infty,$$

we have

$$\log \mathcal{N}_2(B_1, \varepsilon) \leq c_0 (1/\varepsilon)^s, \quad 0 < s \leq 2,$$

where $c_0$ is a positive constant. Under the circumstances, it is reasonable to claim that **Assumption 2**, with $0 < p < 2$ is satisfied for most RKHSs.

To illustrate, let us consider two classes of examples: Sobolev spaces and RKHSs with compactly supported radial basis function [19], which can also be found in [18].

**Example 1.** Let $X$ be a bounded domain in $\mathbb{R}^n$ and $H^r(X)$ be the Sobolev space of index $r$. When $r > n$, the classical Embedding Theorem tells us that $H^r(X)$ is an RKHS and its unit ball $B_1$ is included in a finite ball of the function space $C^{r-\frac{d}{2}-\frac{\varepsilon}{2}}(X)$ with inclusion bounded where $0 < \varepsilon < r - n$. Then from the classical bounds for covering numbers of the unit ball of $C^{r-\frac{d}{2}-\frac{\varepsilon}{2}}(X)$, we know that

$$\log \mathcal{N}_2(B_1, \eta) \leq c_\varepsilon \eta^{-\frac{n}{r-n/2-\varepsilon}}, \quad \forall \eta > 0.$$  

Hence **Assumption 2** holds with $p = \frac{n}{r-n/2-\varepsilon} < 2$.

**Example 2.** With an index $s \in \mathbb{N}$, define the function $\psi_s(t) = (\max(1 - t^2, 0))^s$ on $\mathbb{R}$ and $\phi_k = \psi_s * \psi_s$ to be the convolution of $\psi_s$ with itself. If an integer $k$ satisfies $0 \leq k < \frac{2s-1}{4}$, then the scaled $k$-th derivatives $\phi_k^{(k)}(2t)$ induce a radial basis function $K(x, y) = \phi_k^{(k)}(2|x - y|)$ with $x, y \in \mathbb{R}^{2k+1}$. It was shown in [19] that $K$ is $C^{2k-2k}$ and is positive definite. Thus when $X$ is a bounded domain in $\mathbb{R}^{2k+1}$, the restriction of $K$ onto $X \times X$ is a Mercer kernel. This in connection with the embedding results from [17,15] tells us that for such a Mercer kernel, **Assumption 2** holds with $p = \frac{2(2k+1)}{2s-2k} < 2$.  


511
2.4. Estimating hypothesis error

The estimation of hypothesis error can be conducted analogously to that in [3]. Denote \( y = (y_1, y_2, \ldots, y_m)^T \), \( K[x] = (K(x_i, x_j))_{i,j=1}^m \) and \( \alpha_x = (\alpha_{x,1}, \ldots, \alpha_{x,m})^T \) as the coefficient of \( f_{x,\lambda} \). Based on a representer theorem and Hölder’s inequality, we come to the following proposition.

Proposition 4. Let \( 1 \leq q \leq 2 \). There holds

\[
\mathcal{H}(z, \lambda) \leq \frac{m\eta M^2}{(m\lambda)^q}. \tag{2.4}
\]

Proof. Recall that the coefficient \( \alpha_x \) of \( f_{x,\lambda} \) satisfies

\( (\lambda m + K[x]) \alpha_x = y. \)

which yields \( \lambda m \alpha_x = y - K[x] \alpha_x. \)

Following a representer theorem, for \( i = 1, 2, \ldots, m \), we get

\[ \alpha_{x,i} = \frac{1}{\lambda m} (y_i - f_{x,\lambda}(x_i)). \]

Using Hölder’s inequality,

\[
\sum_{i=1}^m |\alpha_{x,i}|^q = \frac{1}{(\lambda m)^q} \sum_{i=1}^m |y_i - f_{x,\lambda}(x_i)|^q \leq \frac{m}{(m\lambda)^q} \left( \varepsilon_2(f_{x,\lambda}) \right)^{q/2}.
\]

Hence,

\[
\varepsilon_2(\pi_M(f_x)) + \eta \Omega_2(f_x) \leq \varepsilon_2(f_x) + \eta \Omega_2(f_x) \leq \varepsilon_2(f_{x,\lambda}) + \eta \Omega_2(f_{x,\lambda}) \leq \varepsilon_2(f_{x,\lambda}) + \frac{m\eta}{(m\lambda)^q} \varepsilon_2(f_{x,\lambda})^{q/2} \leq \varepsilon_2(f_{x,\lambda}) + \lambda \|f_{x,\lambda}\|^2_K + \frac{m\eta}{(m\lambda)^q} \varepsilon_2(f_{x,\lambda}) + \lambda \|f_{x,\lambda}\|^2_K^{q/2}.
\]

Recall that \( \varepsilon_2(f_{x,\lambda}) + \lambda \|f_{x,\lambda}\|^2_K \leq \varepsilon_2(0) + \lambda \|0\|^2_K \), we get

\[
\varepsilon_2(\pi_M(f_x)) + \eta \Omega_2(f_x) \leq \varepsilon_2(f_{x,\lambda}) + \lambda \|f_{x,\lambda}\|^2_K + \frac{m\eta M^2}{(m\lambda)^q}.
\]

Notice that

\[
\varepsilon_2(f_{x,\lambda}) + \lambda \|f_{x,\lambda}\|^2_K \leq \varepsilon_2(f_{x}) + \lambda \|f_{x}\|^2_K.
\]

This yields our desired estimation. \( \square \)

3. Deriving learning rates

We are now in a position to derive the learning rate of projected algorithm (1.1). Main results of this paper will be presented in Theorem 2.

Following the error decomposition scheme in Proposition 1 and combining Propositions 2–4, we get the following estimation on the total error.

Proposition 5. Suppose assumptions 1, 2 and 3 hold. Let \( 1 \leq q \leq 2, R > 0 \). If \( 0 < \delta < 1 \), then with confidence \( 1 - \delta \), \( \mathcal{E}(\pi_M(f_x)) - \mathcal{E}(f_x) + \eta \Omega_2(f_x) \) can be bounded by

\[
3\mathcal{D}(\lambda) + 14(3M + \kappa \sqrt{\mathcal{D}(\lambda)/\lambda})^2 + 1920M^2 \log \left( \frac{2}{\delta} \right) + \frac{2m\eta M^2}{(m\lambda)^q} + 2C_{p,M,K} R^{2p} \left( \frac{1}{m} \right)^{\frac{2p}{m}}.
\]

Proposition 5 yields an upper bound of the generalization error. The parameter \( R \) can be generally bounded as follows.

Lemma 3. Let \( 1 \leq q \leq 2 \). Then for any \( z \in \mathbb{Z}^m \), almost surely we have

\[
\|f_x\|_K \leq \kappa m^{1-q} (M^2/\eta)^{1/q}.
\]

Proof. By choosing \( f = 0 \), we deduce that

\[
\eta \sum_{i=1}^m |\alpha_{x,i}|^q = \eta \Omega_2(f_x) \leq \varepsilon_2(f_x) + \eta \Omega_2(f_x) \leq \varepsilon_2(0) + \eta \Omega_2(0) \leq M^2. \tag{3.1}
\]
When $1 < q \leq 2$, applying Hölder’s inequality one gets
\[
\|f_k\|_K = \left\| \sum_{i=1}^{\infty} \alpha_{x,i}K_{ij} \right\|_K \leq K \sum_{i=1}^{\infty} |\alpha_{x,i}| \leq \kappa m^{1-\frac{1}{q}} \left( M^2 \frac{1}{\eta} \right)^{1/q}.
\]

When $q = 1$, the desired result follows from (3.1). Hence our assertion holds. \qed

Replacing $R$ in Proposition 5 with the upper bound presented in Lemma 3, we get our main theorem as follows.

**Theorem 2.** Suppose assumptions 1, 2 and 3 hold. Let $1 \leq q \leq 2$. Choose
\[
\eta = \left( \frac{1}{m} \right)^{\frac{2q(p+q)}{2pq+p+2q+pq}} + 1 - q.
\]

For any $0 < \delta < 1$, with confidence $1 - \delta$, there holds
\[
\mathcal{E}(\pi_M(f_k)) - \mathcal{E}(f_\rho) \leq C \log \left( \frac{2}{\delta} \right) \left( \frac{1}{m} \right)^{\min \left\{ \frac{2q}{2pq+p+2q+pq}, \beta \right\}},
\]
where
\[
C = 2(3c_\beta + 5\kappa^2 + 364M^2 + 2C_{p,M,K}).
\]

**Proof.** Following Proposition 5 and Lemma 3, $\mathcal{E}(\pi_M(f_k)) - \mathcal{E}(f_\rho)$ can be bounded by
\[
C \log \left( \frac{2}{\delta} \right) \left( \lambda^p + \lambda^{\beta-1} \frac{1}{m} + 1 \right)^{\frac{q-1}{2q}} \frac{\eta}{\lambda^q} + \left( \frac{1}{m} \right)^{\frac{2q-2q+2q}{2pq+2q+pq}} \left( \frac{1}{\eta} \right)^{\frac{2q}{2pq+2q+pq}},
\]
where $C$ is given in (3.3).

We choose regularization parameter $\eta$ such that
\[
\left( \frac{1}{m} \right)^{\frac{q-1}{2q}} \frac{\eta}{\lambda^q} = \left( \frac{1}{m} \right)^{\frac{2q-2q+2q}{2pq+2q+pq}} \left( \frac{1}{\eta} \right)^{\frac{2q}{2pq+2q+pq}},
\]
so we have
\[
\eta = \lambda \frac{\frac{q(q+2)}{2q-2q+2q+pq} m^{pq-2-4q+2q+pq}}{\lambda^{pq+2q+pq}}.
\]

Next, we choose regularization parameter $\lambda$ satisfying
\[
\left( \frac{1}{m} \right)^{\frac{q-1}{2q}} \frac{\eta}{\lambda^q} = \lambda^p.
\]

Hence our assertion follows by taking $\lambda = \left( \frac{1}{m} \right)^{\min \left\{ \frac{2q}{2pq+p+2q+pq}, \beta \right\}}$. \qed

At the end of this section we present a proof of Theorem 1 by applying Theorem 2.

**Proof of Theorem 1.** Since $K \in C^\infty(X \times X)$, it follows that condition (2.2) holds for arbitrary small $p > 0$. Moreover, $f_\rho \in \mathcal{H}_K$ implies that $\beta = 1$.

Considering that $0 < p < 2$, we choose $p = \frac{2\epsilon}{1-\epsilon}$ with $0 < \epsilon < \frac{1}{2}$. Applying Theorem 2, it follows that for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds
\[
\|\pi_M(f_k) - f_\rho\|_\rho^2 \leq C \log \left( \frac{2}{\delta} \right) \left( \frac{1}{m} \right)^{1-\frac{3p+2}{2q(p+q)+2q+pq}} \epsilon,
\]
where $C$ is given by (3.3) and we choose $\eta = \left( \frac{1}{m} \right)^{\frac{2q+2-2q+2q+pq}{2pq+2q+pq}} \epsilon$. \qed

4. **Comparison and discussion**

In this section, we present a detailed description and explanation of our contributions by comparing our learning rates with existing results.
Recently, considering the case $1 < q \leq 2$, under the capacity condition on uniform covering numbers
\[
\log N(\eta) \leq c_1 \left( \frac{1}{\eta} \right)^{\frac{1}{q-1}}, \quad \forall \eta > 0,
\]
and the restriction $s < \frac{1}{q-1}$, with confidence at least $1 - \delta$, [8] established the learning rate
\[
\| \pi_m(f_z) - f_z \|_p^2 \leq C_3 \log \left( \frac{2}{\delta} \right) \left( \frac{1}{m} \right) \min_{\beta \geq 0} \left\{ \frac{m^{-\beta}}{\frac{q-1}{q-1} + \beta} \right\}
\]
where $c_1$ and $C_3$ are positive constants independent of $m$ or $\delta$.

Comparing with their results, we make great improvements from three aspects.

Firstly, it is easy to see that our learning rate is much faster. Specifically, a closer look reveals that with sufficient smooth kernels, the learning rate they derived is of type $O \left( m^{-\frac{q-1}{q-1} - \beta} \right)$, which does not perform well when $q$ tends to 1. While as shown in Theorem 2, under the same condition we obtained the learning rate of type $O(1)$ uniformly with $1 \leq q \leq 2$.

Secondly, their results only hold for $1 < q \leq 2$ but do not cover the case $q = 1$. In this sense, our work is essentially different. To the best of our knowledge, our results are novel in learning literature. And last, we remove the constraint on the kernel smooth condition $s < \frac{1}{q-1}$. That is, our results hold for any $0 < p < 2$, but their results no longer hold when $s \in \left[ \frac{1}{q-1}, \infty \right)$ with $1 < q \leq 2$.

Referring to [8], one might wonder what contributes to these improvements, since analysis there was also conducted based on the capacity assumption and the stepping stone technique. In fact, besides making use of advanced concentration inequality (2.3), the slackness of previous restriction on the regularization parameter is another major contributing factor. To clarify this, let us first briefly revisit the stepping stone techniques used there and in this paper.

Based on notations defined above, they studied the following algorithm
\[
f_z = \arg \min_{f \in \mathcal{H}_{K,x}} \{ \varepsilon_{\zeta}(f) + \lambda \Omega_{\zeta}(f) \},
\]
by choosing a stepping stone $f_{z,\lambda}$ given by
\[
f_{z,\lambda} = \arg \min_{f \in \mathcal{H}_{K,x}} \{ \varepsilon_{\zeta}(f) + \lambda \| f \|_K \},
\]
where $\lambda = \lambda(m) > 0$ is a regularization parameter. Turning back to our work, based on the same stepping stone function $f_{z,\lambda}$, we learned the empirical target function $f_z$ via
\[
f_z = \arg \min_{f \in \mathcal{H}_{K,x}} \{ \varepsilon_{\zeta}(f) + \eta \Omega_{\zeta}(f) \},
\]
where $\eta = \eta(\lambda, m) > 0$ is another regularization parameter. Note that the regularization parameter $\eta$ in (4.2) is chosen as a function of $\lambda$ and $m$, instead of choosing $\eta = \lambda$ in (4.1). As illustrated in Proposition 4, this choice of $\eta$ significantly improves the hypothesis error (2.4). To understand this intuitively, let us draw attention to the role that the regularization parameter plays. The regularization parameter $\eta$ specifies the trade-off between the empirical sample error and the smoothness enforced by the penalty term $\Omega_{\zeta}(f)$. By choosing $\eta = \eta(\lambda, m)$, it enables more flexibility when searching for the empirical target function $f_z$ in the hypothesis space $\mathcal{H}_{K,x}$. This enforces the closeness of $f_z$ and $f_{z,\lambda}$ in the sense of (2.1). However, referring to (3.2) one might argue that this choice of $\eta$ may encounter the risk of over-fitting. Fortunately, as revealed in Proposition 3, the risk is eliminated due to the projection operator given in Definition 1.

Finally, our learning rate could be further improved by introducing an iteration technique as done in [11,13,18,20], and we leave it for future study.

Acknowledgments

The authors would like to thank anonymous referees and Dr. Q.W. Xiao for their valuable suggestions.

References