Learning theory estimates for coefficient-based regularized regression

Lei Shi
Shanghai Key Laboratory for Contemporary Applied Mathematics, School of Mathematical Sciences, Fudan University, Shanghai 200433, PR China

ABSTRACT

We consider a coefficient-based regularized regression in a data dependent hypothesis space. For a given set of samples, functions in this hypothesis space are defined to be linear combinations of basis functions generated by a kernel function and sample data. We do not need the kernel to be symmetric or positive semi-definite, which provides flexibility and adaptivity for the learning algorithm. Another advantage of this algorithm is that, it is computationally effective without any optimization processes. In this paper, we apply concentration techniques with $\ell^2$-empirical covering numbers to present an elaborate capacity dependent analysis for the algorithm, which yields shaper estimates in both confidence estimation and convergence rate. When the kernel is $C^\infty$, under a very mild regularity condition on the regression function, the rate can be arbitrarily close to $m^{-1}$.

1. Introduction

In this paper, we consider a learning algorithm for regression, which is generated by a regularization scheme with data dependent hypothesis space and coefficient-based regularizer.

In our setting, functions are defined on a compact metric space $X$ and take values in $Y = \mathbb{R}$. The sampling process is controlled by a Borel probability measure $\rho$ on $Z = X \times Y$. We use the least square loss to define the generalization error for $f : X \rightarrow Y$ as

$$E(f) = \int_Z \left( f(x) - y \right)^2 d\rho.$$  \hspace{2cm} (1.1)

The function which minimizes $E(f)$ is called regression function and is given by

$$f_\rho(x) = \int_Y y d\rho(y|x), \quad x \in X,$$  \hspace{2cm} (1.2)

where $\rho(\cdot|x)$ is the conditional probability measure induced by $\rho$ at $x$. Throughout the paper, we assume $|y| \leq M$ almost surely for some $M > 0$ and the set of samples $Z = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ drawn independently from $\rho$.

The target of a regression algorithm is to use the sample set $Z$ to produce an approximation of $f_\rho$. We denoted this approximation by $f_Z$. For any $f \in L^2_{\rho_X}(X)$, it is easy to check $E(f) - E(f_\rho) = \|f - f_\rho\|_{L^2_{\rho_X}}^2$, where $\rho_X$ is the marginal distribution of $\rho$ on $X$ and $\|f\|_{L^2_{\rho_X}} = (\int_X |f(x)|^2 d\rho_X)^{1/2}$. The learning ability or statistical performance of an algorithm for regression can be measured by the excess generalization error.
\[ \| f_\rho - f_\rho \|^2_{L^2_{\rho X}} = \mathcal{E}(f_\rho) - \mathcal{E}(f_\rho). \] (1.3)

We approximate the generalization error \( \mathcal{E}(f) \) by the empirical error

\[ \mathcal{E}_n(f) = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2. \]

A regularization scheme [6] for regression is composed of the empirical error and a regularization term.

A classical learning algorithm is the regularization scheme in a reproducing kernel Hilbert space (RKHS, see [1] for details and [2] for some generalizations) associated with a Mercer kernel \( K \) on \( \Omega \) on \( \Omega \).

\[ \mathcal{H}_K \text{ is given by} \]

\[ \mathcal{H}_K = \left\{ \sum_{i=1}^{m} \alpha_i K_{x_i} : (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m \right\}, \] (1.5)

where \( K_{x} (\cdot) = K(\cdot, x) \). We do not need the kernel \( K \) to be symmetric or positive semi-definite, which leads to much flexibility. Since every function in this hypothesis space is determined by its coefficients, we investigate the following regularization scheme

\[ f_{\mathcal{L}, \lambda} = \arg \min_{f \in \mathcal{H}_{\mathcal{L}}} \left\{ \mathcal{E}_n(f) + \lambda \| f \|^2_{K} \right\}, \quad \lambda > 0, \] (1.6)

where the regularizer \( \mathcal{L}(f) \), as a functional acting on the functions in such kinds of hypothesis spaces, is defined to be a positive function imposed on the corresponding coefficients of \( f \). This technique is called the coefficient-based regularization which was introduced by Vapnik [15] to linear programming support vector machines.

According to the representer theorem (e.g. see [17,8]), if we take \( K \) to be a Mercer kernel and \( \mathcal{L}(\sum_{i=1}^{m} \alpha_i K_{x_i}) = \alpha^T K(x) \alpha \) with \( \alpha = (\alpha_1, \ldots, \alpha_m)^T \), the learning scheme (1.6) is exactly the same as the classical algorithm (1.4). Another typical and natural choice for \( \mathcal{L}(f) \) is the \( \ell^p \)-norm of the coefficients, which is given by \( \mathcal{L}(\sum_{i=1}^{m} \alpha_i K_{x_i}) = \sum_{i=1}^{m} |\alpha_i|^p \) with \( 1 \leq p < \infty \).

Recently, much attention has been paid to the \( \ell^1 \)-norm regularizer, since the algorithm can produce sparse solutions, and increasing theoretical work has been done for this special case [22,18,9].

In this paper, we define the following coefficient-based regularizer

\[ \mathcal{L}(f) := m \sum_{i=1}^{m} \alpha_i^2 \quad \text{for} \quad f = \sum_{i=1}^{m} \alpha_i K_{x_i}. \] (1.7)

Then the output function of learning algorithm (1.6) is \( f_{\mathcal{L}, \lambda} = \sum_{i=1}^{m} \alpha^2_i K_{x_i} \), where the coefficient vector \( \alpha^2 = (\alpha^2_1, \ldots, \alpha^2_m) \) is given by

\[ \alpha^2 = \arg \min_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{m} \sum_{j=1}^{m} \left( \sum_{i=1}^{m} \alpha_i K(x_j, x_i) - y_j \right)^2 + \lambda m \sum_{i=1}^{m} \alpha_i^2 \right\}. \] (1.8)

Besides the flexibility imposed by removing the symmetry for the kernel, another advantage of this learning algorithm is the effectivity of computations, since \( \alpha^2 \) can be explicitly solved by a linear system of equations. This algorithm has been recently studied in [13]. By using the integral operator technique from [10], Sun and Wu have given a capacity independent estimate for the convergence rates of \( \| f_{\mathcal{L}, \lambda} - f_\rho \|_{L^2_{\rho X}} \). Let \( L_K \) be an integral operator on \( L^2_{\rho X} \) defined by

\[ L_K f(x) = \int_{X} K(x, t) f(t) d\rho_X(t), \quad x \in X. \] (1.9)
Since $X$ is compact and $K$ is continuous, $L_K$ and its adjoint $L_K^*$ are both compact operators. Define $\tilde{K}(u, v) = \int_X K(u, x)K(v, x)d\rho_X(x)$. It is easy to verify that $\tilde{K}$ is a Mercer kernels and $L_K = L_{\tilde{K}}L_K^*$. As a self-adjoint positive operator on $L_{\tilde{K}}^2$, its $r$-th power $L_{\tilde{K}}^r$ is well defined for any $r > 0$. It is proved in [13] that, if $f_\rho$ lies in the range of $L_{\tilde{K}}^r$ for some $r > 0$, then with confidence $1 - \delta$,

$$\|f_{\mathcal{I}, \lambda} - f_\rho\|_{L_{\tilde{K}}^2}^2 = O\left(\frac{1}{\delta}m^{-\min\left(\frac{1}{2r_\rho}, \frac{1}{2}\right)}\right). \tag{1.10}$$

The part involving $\delta$ in the above error bound is $1/\delta$, which is not optimal in learning theory. By using exponential probability inequalities, most error bounds stated in the literature of learning theory (e.g. [10,21]) depend on $\delta$ in the form $\log(1/\delta)$.

The purpose of this paper is to provide a capacity dependent analysis for algorithm (1.8) by applying concentration techniques involving the $\ell^2$-empirical covering numbers (see Definition 3 below). We improve the error bound further in both confidence estimation (the part involving $\delta$) and convergence rates. Our learning rates are given in terms of the measure $\rho$ and the kernel $K$. Firstly, let us state two important results to illustrate our general error analysis described in the next section.

**Theorem 1.** Suppose that $X$ is a compact subset of $\mathbb{R}^n$, $K$ is $C^s$ with some $0 < s \leq 1$ satisfying

$$|k(t, x) - k(t, x')| \leq c_s|x - x'|^s, \quad \forall t, x, x' \in X \tag{1.11}$$

for some constant $c_s > 0$, and $f_\rho$ lies in the range of $L_{\tilde{K}}^r$ for some $r > 0$. Take $\lambda = m^e - \frac{n+2s}{2n+2s}$ with $0 < e < \frac{n+2s}{2n+2s}$. For any $0 < \delta < 1$, with confidence $1 - \delta$, we have

$$\|f_{\mathcal{I}, \lambda} - f_\rho\|_{L_{\tilde{K}}^2}^2 \leq C_\epsilon \left(\frac{20}{\epsilon \delta}\right)^{1/2} m^{-\min(2r_\rho, \frac{n+2s}{2n+2s} - \epsilon)} \tag{1.12}$$

where the constant $C_\epsilon$ is independent of $m$ or $\delta$.

Since $\epsilon$ can be chosen arbitrarily small (but fixed), our error bound is sharper than (1.10) in both confidence estimation and convergence rate. The following result is about a special case when $K$ is $C^\infty$ and $f_\rho \in \mathcal{H}_K$.

**Theorem 2.** Assume $X$ is a compact subset of $\mathbb{R}^n$ and $K \in C^\infty(X \times X)$. If $f_\rho \in \mathcal{H}_K$, let $\lambda = m^e - \frac{n}{2n+2s}$ with $0 < e < 1$ and $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\|f_{\mathcal{I}, \lambda} - f_\rho\|_{L_{\tilde{K}}^2}^2 \leq C_\epsilon \left(\frac{30}{\epsilon \delta}\right)^{1/2} m^{-(1-\epsilon)} \tag{1.13}$$

where the constant $C_\epsilon$ is independent of $m$ or $\delta$.

### 2. Key analysis

Before giving our key analysis, we firstly introduce some known properties about the RKHS $\mathcal{H}_K$ generated by the Mercer kernel $\tilde{K}$. From [3], we know that $\mathcal{H}_K$ is the range of $L_{\tilde{K}}^{3/2}$. For any $f \in \mathcal{H}_K$, there holds

$$\|f\|_\infty \leq \kappa \|f\|_{\tilde{K}} \quad \text{and} \quad \|f\|_\infty = \|L_{\tilde{K}}^{-1/2}f\|_{L_{\tilde{K}}^2} \tag{2.1}$$

where $\kappa = \sup_{x, t \in X} |K(x, t)|$.

The following lemma is about the polar decomposition of compact operators (e.g. [4]), which is very useful for our analysis.

**Lemma 1.** Let $H$ be a separable Hilbert space and $T$ be a compact operator on $H$, then $T$ can be factored as

$$T = \Gamma A \tag{2.2}$$

where $A = (T^*T)^{1/2}$ and $\Gamma$ is a partial isometry on $H$ with $\Gamma^* \Gamma$ being the orthogonal projection onto $\overline{\mathcal{R}(A)}$.

Using above lemma, we immediately get the following proposition.

**Proposition 1.** Consider $\mathcal{H}_K$ as a subspace of $L_{\tilde{K}}^2$, then $L_{\tilde{K}}^* = UL_{\tilde{K}}^{1/2}$ and $L_K = L_{\tilde{K}}^{1/2}U^*$, where $U$ is a partial isometry on $L_{\tilde{K}}^2$ with $U^*U$ being the orthogonal projection onto $\overline{\mathcal{R}(U)}$. 

---

**Note:** The actual content of the document is cut off after the above paragraphs, and the document contains mathematical expressions and equations which are essential for understanding the context of the text. The rest of the content is not visible in the image provided. This indicates that the provided text is a representative slice of a larger document focused on mathematical analysis and learning theory.
2.1. Error decomposition

A useful approach to do error analysis for regularization schemes with sample independent hypothesis spaces is the error decomposition (e.g., [21]), which decomposes the excess generalization error (1.3) into the sum of a sample error and an approximation error. The main difficulty in our setting is the dependence of the hypothesis space \( \mathcal{H}_{K,z} \) on \( z \). In order to overcome this difficulty, we use the operator \( L_K \) to define a function space independent of samples in which we could approximate \( f_\rho \). Note that, by Proposition 1, we have \( L_K = L_K^2 U^* \).

**Definition 1.** Define a Hilbert space

\[
\mathcal{H}_2 = L_K \left( L^2_{\rho x} \right) = \left\{ L_K^{1/2} U^* f : f \in L^2_{\rho x} \right\}
\]

with inner product inherited from \( \mathcal{H}_K \) as

\[
\langle L_K f, L_K g \rangle_{\mathcal{H}_2} = \langle L_K f, L_K g \rangle_{\mathcal{H}_K}.
\]

We also define a class of functions for some given \( R > 0 \) as

\[
B_R = \left\{ f = \sum_{i=1}^d c_i K u_i : d \in \mathbb{N}, c_i \in \mathbb{R}, \{u_i\} \subset X \text{ and } \| f \|_2 \leq R \right\}.
\]

From (2.1), we have

\[
\langle L_K f, L_K g \rangle_{\mathcal{H}_2} = \left\{ L_K^{1/2} U^* f, L_K^{1/2} U^* g \right\}_{\mathcal{H}_K} = \left\{ U^* f, U^* g \right\}_{L^2_{\rho x}},
\]

thus the inner product (2.4) is well-defined.

Since \( X \) is compact, if \( K \in C(K_x \times X) \), both \( \mathcal{H}_2 \) (2.3) and \( B_R \) (2.5) can be regarded as subsets of \( C(X) \).

To formulate the error decomposition for algorithm (1.8), we introduce a regularizing function as

\[
f_\lambda = \arg\min_{f \in \mathcal{H}_2} \left\{ \mathcal{E}(f) + \lambda \| f \|_{\mathcal{H}_2}^2 \right\}.
\]

**Proposition 2.** The function \( f_\lambda \) defined by (2.7) can be expressed as \( f_\lambda = L_K g_\lambda \) where

\[
g_\lambda = U \left( \lambda I + L_K \right)^{-1} L_K^{1/2} f_\rho.
\]

Moreover, \( g_\lambda \) is continuous and

\[
\| f_\lambda \|_{\mathcal{H}_2} = \| g_\lambda \|_{L^2_{\rho x}}.
\]

**Proof.** From (2.3), any function \( f \in \mathcal{H}_2 \) can be expressed as \( L_K^{1/2} U^* g \) for some \( g \in L^2_{\rho x} \). Then from (1.3) and (2.6), we have

\[
\mathcal{E}(f) + \lambda \| f \|_{\mathcal{H}_2}^2 = \left\{ \| f - f_\rho \|_{L^2_{\rho x}}^2 + \lambda \| f \|_{\mathcal{H}_2}^2 \right\} + \mathcal{E}(f_\rho)
\]

\[
= \left\{ \| L_K^{1/2} U^* g - f_\rho \|_{L^2_{\rho x}}^2 + \lambda \| U^* g \|_{L^2_{\rho x}}^2 \right\} + \mathcal{E}(f_\rho)
\]

\[
:= \phi_\lambda (g) + \mathcal{E}(f_\rho).
\]

Computing the functional derivative of \( \phi_\lambda \) in \( L^2_{\rho x} \), we obtain

\[
\nabla \phi_\lambda (g) = 2 \left\{ U L_K U^* g - U L_K^{1/2} f_\rho \right\} + 2 \lambda U U^* g.
\]

From the definition of \( f_\lambda \), we must have \( f_\lambda = L_K g_\lambda \) with \( g_\lambda = \arg\min_{g \in L^2_{\rho x}} \phi_\lambda (g) \), then \( \nabla \mathcal{E}_\lambda (g_\lambda) = 0 \) which implies

\[
U \left( \lambda I + L_K \right) U^* g_\lambda = U L_K^{1/2} f_\rho.
\]

Thus we get the expression (2.8) by solving the above equation for \( g_\lambda \). We rewrite \( g_\lambda \) as

\[
g_\lambda = U \left( \lambda I + L_K \right)^{-1} L_K^{1/2} f_\rho = U L_K^{1/2} (\lambda I + L_K)^{-1} f_\rho = L_K^* (\lambda I + L_K)^{-1} f_\rho.
\]
For any \( x_1, x_2 \in X \), we have
\[
\| g_\lambda(x_1) - g_\lambda(x_2) \| \leq \| K(\cdot, x_1) - K(\cdot, x_2) \|_{L_\rho^2} \lambda^{-1} \| f_\rho \|_{L_\rho^2},
\]
since \( K \) is continuous and \( X \) is compact, \( K \) is uniformly continuous. This implies the continuity of \( g_\lambda \).

Conclusion (2.9) directly follows the following computations
\[
\| g_\lambda \|^2_{L_\rho^2} = (U \lambda I + L_R)^{-1} L_R^2 f_\rho, U (\lambda I + L_R)^{-1} L_R^2 f_\rho)_{L_\rho^2}
= (\lambda I + L_R)^{-1} L_R^2 f_\rho, U^* (\lambda I + L_R)^{-1} L_R^2 f_\rho)_{L_\rho^2}
= (U^* (\lambda I + L_R)^{-1} L_R^2 f_\rho, U^* (\lambda I + L_R)^{-1} L_R^2 f_\rho)_{L_\rho^2}
= \| U^* g_\lambda \|^2_{L_\rho^2} = \| L_K g_\lambda \|^2_{\mathcal{H}_2},
\]
the third equation holds from the fact that \( U^* U \) is the orthogonal projection onto \( \mathcal{H}_R \).

Now we are in the position to give the error decomposition for algorithm (1.8).

**Proposition 3.** Let \( \lambda > 0 \) and \( f_{z,\lambda} = \sum_{i=1}^m \alpha_{z,\lambda}^i K_{i,\lambda} \) with the coefficients given by (1.8). We define \( \hat{f}_{z,\lambda} = \frac{1}{m} \sum_{i=1}^m K_{i,\lambda} g_{z,\lambda} \) with \( g_{z,\lambda} \) expressed as (2.8), then
\[
\mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) + \lambda \mathcal{O}_2(f_{z,\lambda}) \leq S_1(z, \lambda) + S_2(z, \lambda) + \mathcal{H}_1(z, \lambda) + \mathcal{H}_2(z, \lambda) + \mathcal{D}(\lambda),
\]
where
\[
S_1(z, \lambda) = \{ \mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) \} - \{ \mathcal{E}_z(f_{z,\lambda}) - \mathcal{E}_z(f_\rho) \},
S_2(z, \lambda) = \{ \mathcal{E}_z(f_{z,\lambda}) - \mathcal{E}_z(f_\rho) \} - \{ \mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) \},
\mathcal{H}_1(z, \lambda) = \lambda \mathcal{O}_2(\hat{f}_{z,\lambda}) - \lambda \| f_{z,\lambda} \|^2_{\mathcal{H}_2},
\mathcal{H}_2(z, \lambda) = \mathcal{E}(\hat{f}_{z,\lambda}) - \mathcal{E}(f_{z,\lambda}),
\mathcal{D}(\lambda) = \mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) + \lambda \| f_{z,\lambda} \|^2_{\mathcal{H}_2}.
\]

**Proof.** By a direct computation, we have
\[
\mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) + \lambda \mathcal{O}_2(f_{z,\lambda})
= \{ \mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) \} - \{ \mathcal{E}_z(f_{z,\lambda}) - \mathcal{E}_z(f_\rho) \} + \{ \mathcal{E}(f_{z,\lambda}) + \lambda \mathcal{O}_2(f_{z,\lambda}) - \mathcal{E}_z(f_{z,\lambda}) - \lambda \mathcal{O}_2(\hat{f}_{z,\lambda}) \}
+ \{ \mathcal{E}_z(f_{z,\lambda}) - \mathcal{E}(f_\rho) \} - \{ \mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) \} + \{ \mathcal{E}(\hat{f}_{z,\lambda}) - \mathcal{E}(f_{z,\lambda}) \} + \{ \mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_\rho) + \lambda \| f_{z,\lambda} \|^2_{\mathcal{H}_2} \}
= \mathcal{D}(\lambda) = \inf_{g \in L_\rho^2} \{ \| L_K g - f_\rho \|^2_{L_\rho^2} + \lambda \| U^* g \|^2_{L_\rho^2} \}.
\]
The decay of \( \mathcal{D}(\lambda) \) as \( \lambda \to 0 \) measures the approximation ability of the function space \( \mathcal{H}_2 \). We shall assume that for some constants \( q \in (0, 1) \) and \( c_q > 0 \),
\[
\mathcal{D}(\lambda) \leq c_q \lambda^q, \quad \forall \lambda > 0.
\]
The construction of the transitional function $\hat{f}_{2,\lambda}$ is very critical for the error decomposition process. This function plays a stepping stone role [19] between $f_{2,\lambda}$ and regularizing function $\hat{f}_2$. By introducing $\hat{f}_{2,\lambda}$, we could effectively estimate the hypothesis error (see Section 4). This idea can be used to refine the previous error estimates for $\ell^1$-regularized learning schemes in [9] and derive satisfactory learning rates for non-smooth kernels. Actually, motivated by Proposition 3, we could construct similar error decompositions for coefficient-based regularization scheme with $\Omega_k(f) := m^p-1\|\alpha\|_p$ for $1 \leq p \leq 2$.

2.2. Measuring the capacity of hypothesis space by empirical covering numbers

We will use a concentration inequality to estimate sample error, the capacity of the hypothesis space plays an essential role in our analysis. As an important measurement of the capacity of a function set, covering numbers have been well studied in the literature [14,24,25].

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

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

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

The $\ell^2$-empirical covering number of a function set is defined by means of the normalized $\ell^2$-metric $d_2$ on the Euclidean space $\mathbb{R}^k$ given by

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

**Definition 3.** Let $\mathcal{F}$ be a set of functions on $X$, $x = (x_i)_{i=1}^{k} \subset X^k$ and $\mathcal{F}|_x = \{(f(x))_{i=1}^{k}: f \in \mathcal{F}\} \subset \mathbb{R}^k$. Set $\mathcal{N}_2(x, \mathcal{F}, \epsilon) = \mathcal{N}(\mathcal{F}|_x, \epsilon, d_2)$. The $\ell^2$-empirical covering number of $\mathcal{F}$ is defined by

$$
\mathcal{N}_2(\mathcal{F}, \epsilon) = \sup_{k \in \mathbb{N}} \mathcal{N}_2(x, \mathcal{F}, \epsilon), \quad \epsilon > 0.
$$

We shall use the $\ell^2$-empirical covering number to describe the capacity property of the hypothesis space. Recall $B_R$ defined by (2.5), by Cauchy–Schwarz inequality, for any $f = \sum_{i=1}^{d} c_i K_{u_i} \in B_R$, the $\ell^1$-norm of the coefficients satisfies

$$
\sum_{i=1}^{d} |c_i| \leq \sqrt{d \sum_{i=1}^{d} c_i^2} = \|f\|_2 \leq R.
$$

This in connection with Theorem 2 in [9] yields the following bound for the capacity of $B_1$.

**Lemma 2.** If $X$ is a compact subset of $\mathbb{R}^n$ and $K$ satisfies condition (1.11) with some $0 < s \leq 1$ and $\kappa > 0$, then

$$
\log \mathcal{N}_2(B_1, \epsilon) \leq C_{X,s} \left( \frac{1}{\epsilon} \right)^{2n/(n+2s)}, \quad \forall \epsilon > 0,
$$

where $C_{X,s} > 0$ is a constant depending on $s$, $\kappa$, $\kappa$ and $X$.

Note that $\mathcal{N}_2(\mathcal{F}, \epsilon) \leq \mathcal{N}(\mathcal{F}, \epsilon, \| \cdot \|_\infty)$, i.e., the $\ell^2$-empirical covering number is bounded by the uniform covering numbers $\mathcal{N}(\mathcal{F}, \epsilon, \| \cdot \|_\infty)$ defined with respect to the $L^\infty$-metric. Since $B_1$ is a subset of $C^s(X)$ provided $K \in C^s$, a natural bound for the covering numbers of $B_1$ can be derived directly from a classical result about the uniform covering numbers of the function space $C^s(X)$ (see [5]).

**Lemma 3.** If $X$ is a compact subset of $\mathbb{R}^n$ and $K \in C^s(X \times X)$ for some $s > 0$, then there is a constant $\tilde{C}_{X,s} > 0$ depending only on $X$, $s$ and $\|K\|_{C^s}$ such that

$$
\log \mathcal{N}_2(B_1, \epsilon) \leq \tilde{C}_{X,s} \left( \frac{1}{\epsilon} \right)^{n/s}, \quad \forall \epsilon > 0.
$$
2.3. Statement of general analysis

We are in the position to give our general analysis. Assume that \( K \in C^s(X \times X) \) with some \( s > 0 \), the following bound for the \( \ell^2 \)-empirical covering numbers is derived from (2.14) and (2.15)

\[
\log N_2(B_1, \varepsilon) \leq c_{p,K} \left( \frac{1}{\varepsilon} \right)^{p}, \quad \forall \varepsilon > 0, \tag{2.16}
\]

where \( c_{p,K} > 0 \) is a constant independent of \( \varepsilon > 0 \), and \( 0 < p < 2 \) is a power index defined by

\[
p = \begin{cases} 
2n/(n + 2s), & \text{when } 0 < s \leq 1, \\
2n/(n + 2), & \text{when } 1 < s \leq 1 + n/2, \\
n/s, & \text{when } s > 1 + n/2.
\end{cases} \tag{2.17}
\]

**Theorem 3.** Suppose that \( X \) is a compact subset of \( \mathbb{R}^n \) and \( K \in C^s(X \times X) \) for some \( s > 0 \), the approximation error condition (2.13) is valid with some \( q > 0 \) and \( c_q > 0 \). For any \( 0 < \delta < 1 \), take \( \lambda = m^{-\gamma} \) with \( 0 < \gamma < \frac{2}{2 + p} \), where \( p \) is defined by (2.17), then with confidence \( 1 - \delta \), there holds

\[
\| f_x - f_{\rho} \|_{L^2_{\rho X}} \leq \tilde{C}_{K,p,\gamma} (\log(10(\tilde{J} + 1)/\delta))^{\frac{4 - \gamma(2 + p)}{2 - \gamma(2 + p)}} m^{-\gamma' \lambda}, \tag{2.18}
\]

where \( \tilde{J} = \frac{2}{2 - \gamma(2 + p)} \) and \( \tilde{C}_{K,p,\gamma} \) is a constant independent of \( \delta \) or \( m \).

We will prove the main result in Section 6 with \( \tilde{C}_{K,p,\gamma} \) given explicitly.

3. Estimates for the approximation error

We show how to realize the polynomial decay of \( D(\lambda) \) under the assumption that \( f_{\rho} \) is in the range of \( L^r_K \) for some \( r > 0 \).

**Proposition 4.** If \( L^{-r}_K f_\rho \in L^2_{\rho X} \) for some \( r > 0 \), we have

\[
D(\lambda) \leq C_r \| L^{-r}_K f_\rho \|_{L^2_{\rho X}}^{2} \lambda^{\min(2r,1)}, \tag{3.1}
\]

where \( C_r \) is a constant depending on \( r \) only.

**Proof.** We firstly estimate \( \| g_\lambda \|_{L^2_{\rho X}} \) through the expression (2.8) of \( g_\lambda \). If \( L^{-r}_K f_\rho \in L^2_{\rho X} \) with \( 0 < r < \frac{1}{2} \), we have

\[
\| g_\lambda \|_{L^2_{\rho X}} = \left\| (\lambda I + L^r_K)^{-1} L^{-r}_K f_\rho \right\|_{L^2_{\rho X}} \leq \left( \frac{1}{1 + r} \right)^{1/2} \frac{1}{(1 - r)^{1/2}} \left\| L^{-r}_K f_\rho \right\|_{L^2_{\rho X}} \lambda^{-1/2},
\]

the first equality holds since \( U^*U \) is the orthogonal projection onto \( \mathbb{H}^r_K \). If \( r \geq \frac{1}{2} \),

\[
\| g_\lambda \|_{L^2_{\rho X}} = \left\| (\lambda I + L^r_K)^{-1} L^r_K L^{-r}_K f_\rho \right\|_{L^2_{\rho X}} \leq \kappa^{-1/2} \left\| L^{-r}_K f_\rho \right\|_{L^2_{\rho X}},
\]

where the last inequality holds since \( \| L^r_K \| \leq \kappa \). Hence if \( L^{-r}_K f_\rho \in L^2_{\rho X} \) for some \( r > 0 \), we get

\[
\lambda \| g_\lambda \|_{L^2_{\rho X}}^2 \leq c_r^2 \left\| L^{-r}_K f_\rho \right\|_{L^2_{\rho X}}^2 \lambda^{\min(2r,1)}, \tag{3.2}
\]

where

\[
c_r = \begin{cases} 
\frac{1}{(1 + r)^{1/2}} \frac{1}{(1 - r)^{1/2}}, & \text{if } 0 < r < \frac{1}{2}, \\
\kappa^{1/2}, & \text{if } r \geq \frac{1}{2}.
\end{cases}
\]

Since \( f_\lambda = L^r_K g_\lambda = (\lambda I + L^r_K)^{-1} L^r_K f_\rho \), we have

\[
\| f_\lambda - f_\rho \|_{L^2_{\rho X}} = \lambda \left\| (\lambda I + L^r_K)^{-1} f_\rho \right\|_{L^2_{\rho X}}. \tag{3.3}
\]

Then if \( 0 < r < 1 \), we get
Let $H$ be a Hilbert space and $\lambda \geq 1$.

For any $0 \leq r \leq 1$, we get

$$\|f - f_\rho\|_{L^p_H}^2 = \lambda \|\lambda L - (\lambda L)^{-1}L^\ell L^\ell f_\rho\|_{L^p_H}^2 \leq \frac{r^r}{(1-r)^{r-1}} \|L^\ell f_\rho\|_{L^p_H}^r \lambda.$$

Thus if $L^\ell f_\rho \in L^2_{L^p_H}$ for some $r > 0$, we get

$$\|f - f_\rho\|_{L^p_H}^2 \leq \tilde{c}_r^2 \|L^\ell f_\rho\|_{L^p_H}^2 \lambda \min[2, 2]. \quad (3.4)$$

where

$$\tilde{c}_r = \begin{cases} \frac{r^r}{(1-r)^{r-1}}, & \text{if } 0 < r < 1, \\
\kappa^{r-1}, & \text{if } r \geq 1. \end{cases}$$

Finally from (2.12), combining (3.2) and (3.4), we complete our proof by taking $C_r = \tilde{c}^2_r + \tilde{c}^2_r$. $\square$

**4. Estimates for the hypothesis error**

Under the assumption that the sample is independently drawn form $\rho$ and $|y| \leq M$ almost surely, we estimate $\mathcal{H}_1$ and $\mathcal{H}_2$ mainly by the following probability inequality in Hilbert spaces [10].

**Lemma 4.** Let $H$ be a Hilbert space and $\xi$ be a random variable on $(Z, \rho)$ with values in $H$. Assume $\|\xi\| \leq \tilde{M} < \infty$ almost surely. Denote $\sigma^2(\xi) = E(\|\xi\|^2)$. Let $(x_i)_{i=1}^m$ be independent random drawers of $\rho$. For any $0 < \delta < 1$, with confidence $1 - \delta$,

$$\|\frac{1}{m} \sum_{i=1}^m [\xi_i - E(\xi_i)]\| \leq \frac{2M \log (2/\delta)}{m} + \sqrt{\frac{2n \sigma^2(\xi) \log(2/\delta)}{m}}. \quad (4.1)$$

Using this lemma, we can give estimations for $\mathcal{H}_1$ and $\mathcal{H}_2$ as follows.

**Proposition 5.** For any $0 < \delta < 1$, with confidence $1 - \delta$, we have

$$\mathcal{H}_1(z, \lambda) \leq \frac{3\kappa^2 D(\lambda) \log(2/\delta)}{\lambda m} + \frac{1}{2} D(\lambda). \quad (4.2)$$

and

$$\mathcal{H}_2(z, \lambda) \leq 8\kappa^2 (\kappa^2 + 1) \log^2(2/\delta) \left\{ \frac{D(\lambda)}{\lambda^2 m^2} + \frac{D(\lambda)}{\lambda m} \right\} + D(\lambda). \quad (4.3)$$

**Proof.** We first deal with $\mathcal{H}_1$. Since $\hat{f}_{0, \lambda} = \frac{1}{m} \sum_{i=1}^m K_{\xi}(x_i, x_i)$, then $\Omega_2(\hat{f}_{0, \lambda}) = \frac{1}{m} \sum_{i=1}^m g_{\xi}^2(z_i)$. We apply Lemma 4 to the random variable $\xi = g_{\xi}^2(\chi)$ on $\mathcal{X}$ with values in $\mathbb{R}$, then $|\xi| \leq g_{\xi}^2(\chi)$. From (2.10) and (3.3), we get

$$\|g_{\xi}\|_\infty \leq \kappa \|\lambda L - (\lambda L)^{-1} f_\rho \|_{L^p_H} \leq \frac{\kappa}{\lambda} \|f_{\lambda} - f_\rho\|_{L^p_H} \leq \frac{\kappa}{\lambda} \sqrt{D(\lambda)}. \quad (4.4)$$

It is easy to check

$$E(\xi) = \int_{\mathcal{X}} g_{\xi}^2 d\rho = \|g_{\xi}\|_{L^p_{L^2_H}}^2 = \|f_{\lambda}\|_{L^p_{L^2_H}}^2$$

and

$$\sigma^2(\xi) = E(\xi^2) = \int_{\mathcal{X}} g_{\xi}^4 d\rho \leq \|g_{\xi}\|_{L^p_{L^2_H}}^2 \|g_{\xi}\|_{L^p_{L^2_H}}^2 \cdot$$

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

$$\Omega_2(\hat{f}_{0, \lambda}) \leq \|\hat{f}_{0, \lambda}\|_{L^p_{L^2_H}}^2 \leq \frac{2\|g_{\xi}\|_{L^p_{L^2_H}}^2 \log(2/\delta)}{m} + \sqrt{\frac{2\|g_{\xi}\|_{L^p_{L^2_H}}^2 \|g_{\xi}\|_{L^p_{L^2_H}}^2 \log(2/\delta)}{m}}. \quad (4.5)$$
Using $\|g_\lambda\|_{L^2_{pz}}^2 = \|f_\lambda\|_{L^2_{pz}}^2 \leq \frac{D(\lambda)}{\lambda m}$ and (4.4), we get
\[
H_1 = \lambda \left\{ \Omega_2(\hat{f}_{z,\lambda}) - \|f_\lambda\|_{L^2_{pz}}^2 \right\}
\leq 2k^2D(\lambda) \log(2/\delta) + \sqrt{\frac{2k^2D(\lambda) \log(2/\delta)}{\lambda m}}
\leq \frac{3k^2D(\lambda) \log(2/\delta)}{2} + \frac{1}{2} D(\lambda).
\]
As for $H_2$, a simple computation shows that
\[
E(\hat{f}_{z,\lambda}) - E(f_\lambda) = \|\hat{f}_{z,\lambda} - f_\rho\|_{L^2_{pz}}^2 - \|f_\lambda - f_\rho\|_{L^2_{pz}}^2
\leq \|\hat{f}_{z,\lambda} - f_\lambda\|_{L^2_{pz}}^2 \{ \|\hat{f}_{z,\lambda} - f_\lambda\|_{L^2_{pz}}^2 + 2\|f_\lambda - f_\rho\|_{L^2_{pz}}^2 \}
= \|\hat{f}_{z,\lambda} - f_\lambda\|_{L^2_{pz}}^2 + 2\|\hat{f}_{z,\lambda} - f_\lambda\|_{L^2_{pz}}^2 \|f_\lambda - f_\rho\|_{L^2_{pz}}^2
\leq 2\|\hat{f}_{z,\lambda} - f_\lambda\|_{L^2_{pz}}^2 + \|f_\lambda - f_\rho\|_{L^2_{pz}}^2,
\]
(4.6)

We also apply Lemma 4 to the random variables $\xi_i = K_\lambda g_\lambda(x_i)$ on $(X, \rho X)$ with values in the Hilbert space $L^2_{pz}(X)$. It satisfies $E(\xi) = L_2 K g_\lambda = f_\lambda$ and $\|\xi\|_{L^2_{pz}} \leq \kappa \|g_\lambda\|_{\infty}$. It is also easy to check that $\sigma^2(\xi) = E[\xi^2]_{L^2_{pz}} \leq \kappa^2 \|g_\lambda\|_{L^2_{pz}}^2$. Thus we get with confidence $1 - \delta$,
\[
\|\hat{f}_{z,\lambda} - f_\lambda\|_{L^2_{pz}} \leq \frac{2k \|g_\lambda\|_{\infty} \log(2/\delta)}{m} + \sqrt{\frac{2k^2 \|g_\lambda\|_{L^2_{pz}}^2 \log(2/\delta)}{m}}
\leq \frac{2k^2 \sqrt{D(\lambda)} \log(2/\delta)}{\lambda m} + \sqrt{\frac{2k^2D(\lambda) \log(2/\delta)}{\lambda m}}.
\]
Finally, using (4.6), we have
\[
H_2 \leq 2 \left\{ \frac{2k^2 \sqrt{D(\lambda)} \log(2/\delta)}{\lambda m} + \sqrt{\frac{2k^2D(\lambda) \log(2/\delta)}{\lambda m}} \right\}^2 + D(\lambda)
\leq 4 \left\{ \frac{2k^4D(\lambda) \log(2/\delta)}{\lambda^2m^2} + \frac{k^2D(\lambda) \log(2/\delta)}{\lambda m} \right\} + D(\lambda)
\leq 8k^2(k^2 + 1) \log(2/\delta) \left\{ \frac{D(\lambda)}{\lambda^2m^2} + \frac{D(\lambda)}{\lambda m} \right\} + D(\lambda).
\]
This proves the proposition. \(\Box\)

5. Estimates for the sample error

Estimating $S_1$ and $S_2$ are more involved since neither $f_{z,\lambda}$ nor $\hat{f}_{z,\lambda}$ is a single random variable, they both depend on the sample $z$. Therefore, the usual probability inequality such as Lemma 4 does not guarantee the convergence of these two terms. Our concentration estimations for $S_1$ and $S_2$ are based on the following concentration inequality derived in [9].

Lemma 5. Assume $B_1$ satisfy the capacity condition (2.16) with some $0 < p < 2$, then for any $\delta \in (0, 1)$, with confidence $1 - \delta$, we have
\[
\left\{ E(f) - E(f_\rho) \right\} - \left\{ E_2(f) - E_2(f_\rho) \right\}
\leq \frac{1}{2} \left\{ E(f) - E(f_\rho) \right\} + C_{k,p} \log(1/\delta) m^{-p/2} \max\{R^2, M^2\}, \; \forall f \in B_R.
\]
(5.1)

where $C_{k,p}$ is a constant depending only on $k$ and $p$. The same bound also holds true for $\left\{ E_2(f) - E_2(f_\rho) \right\} - \left\{ E(f) - E(f_\rho) \right\}$.

We can use Lemma 5 to estimate $S_1$ and $S_2$ with a properly chosen $R$. We firstly give an estimation for $S_2(z, \lambda)$.
Proposition 6. Assume $\mathcal{B}_1$ satisfies the capacity condition (2.16) with some $0 < p < 2$, then for any $0 < \delta < 1$, with confidence $1 - \delta$,

$$
\mathcal{S}_2(\mathbf{z}, \lambda) \leq D(\lambda) + 4\kappa^2(\kappa^2 + 1)\log^2(5/\delta) \left\{ \frac{D(\lambda)}{\lambda^2 m^2} + \frac{D(\lambda)}{\lambda m} \right\} + C_1 \log^2(5/\delta) \left( \frac{D(\lambda)}{\lambda^2 m} + \frac{D(\lambda)}{\lambda} + 1 \right) m^{-\frac{\lambda}{2 + \lambda^2}},
$$

(5.2)

where $C_1 = C_{k, p}(3\kappa^2 + \frac{3}{2} + M^2)$.

**Proof.** From (4.4) and (4.5), for any $\delta \in (0, 1)$, with confidence $1 - 2\delta/5$, we have

$$
\frac{1}{m} \sum_{i=1}^{m} g_{i, x}^2(x_i) - \|g_{i, x}\|^2_{L^2_x} \leq \frac{2\|g_{i, x}\|^2_{L^\infty_x} \log(5/\delta)}{m} + \frac{1}{2} \left\{ \frac{\|g_{i, x}\|^2_{L^2_x}}{m} \right\},
$$

which implies the existence of a subset $U_1$ of $Z^m$ with measure at most $2\delta/5$ such that

$$
\frac{1}{m} \sum_{i=1}^{m} g_{i, x}^2(x_i) \leq \frac{3\kappa^2 D(\lambda) \log(5/\delta)}{\lambda^2 m} + \frac{3}{2} \|g_{i, x}\|^2_{L^2_x} + \frac{3\kappa D(\lambda)}{2\lambda}, \quad \forall \mathbf{z} \in Z^m \setminus U_1.
$$

This inequality ensures that for every $\mathbf{z} \in Z^m \setminus U_1$, we have $\hat{f}_{x, \lambda} \in \mathcal{B}_{R_\lambda}$ with $R_\lambda = \sqrt{\frac{3\kappa^2 D(\lambda) \log(5/\delta)}{\lambda^2 m} + \frac{3\kappa D(\lambda)}{2\lambda}}$. By Lemma 5, there exists $U_{R_\lambda}$ with measure at most $\delta/5$ such that for every $\mathbf{z} \in Z^m \setminus (U_1 \cup U_{R_\lambda})$, we have $\hat{f}_{x, \lambda} \in \mathcal{B}_{R_\lambda}$ and

$$
\left\{ \mathcal{E}(\hat{f}_{x, \lambda}) - \mathcal{E}(f_\rho) \right\} \leq \left\{ \mathcal{E}(f_{x, \lambda}) - \mathcal{E}(f_\rho) \right\} + C_{k, p} \log(5/\delta) m^{-\frac{\lambda}{2 + \lambda^2}} \max\{R_\lambda^2, M^2\}
$$

$$
\leq \frac{1}{2} \left\{ \mathcal{E}(f_{x, \lambda}) - \mathcal{E}(f_\rho) \right\} + \frac{1}{2} \left\{ \mathcal{E}(f_{x, \lambda}) - \mathcal{E}(f_\rho) \right\} + C_{k, p} \log(5/\delta) m^{-\frac{\lambda}{2 + \lambda^2}} \left\{ R_\lambda^2 + M^2 \right\}
$$

(5.3)

From Proposition 5, we know that there exists a subset $U_2$ of $Z^m$ with measure at most $2\delta/5$ such that for every $\mathbf{z} \in Z^m \setminus U_2$, the total error $\mathcal{E}(f_{x, \lambda}) - \mathcal{E}(f_\rho) \leq 8\kappa^2(\kappa^2 + 1)\log^2(5/\delta) \left\{ \frac{D(\lambda)}{\lambda^2 m^2} + \frac{D(\lambda)}{\lambda m} \right\} + D(\lambda)$.

(5.4)

Finally, let $U = U_1 \cup U_2 \cup U_{R_\lambda}$. The measure of $U$ is at most $\delta$ and for every $\mathbf{z} \in Z^m \setminus U$, both (5.3) and (5.4) hold. Thus we get our conclusion. \qed

As for $\mathcal{S}_1$, since the choice of $R$ is very important in our total error analysis, we directly apply Lemma 5 to estimate $\mathcal{S}_1$ with $R$ to be determined later.

**6. Estimates for the total errors**

**6.1. Deriving the estimator for total error**

For $R > 0$, denote

$$
\mathcal{V}(R) = \{ \mathbf{z} \in Z^m : \|f_{x, \lambda}\|_2 \leq R \}.
$$

(6.1)

We will provide an estimate for the total error as follows.

**Proposition 7.** Under the assumption of Theorem 3, if $0 < \lambda \leq 1$, $0 < \delta < 1$ and $R > 0$, then there is a subset $V_R$ of $Z^m$ with measure at most $\delta$ such that
Lemma 7. Under the same assumptions of Theorem 3, take $\gamma = m^{-\gamma}$ with $\gamma < \frac{2}{2^{1+2p}}$ and $0 < \delta < 1$, then with confidence $1 - \frac{2}{2^{-\gamma}(2\delta + \gamma)}$, there holds
\[
\|f_{z,\lambda}\|_2 \leq \tilde{C}_\gamma \left( \frac{\log 10}{\delta} \right)^{\frac{4-\gamma(2+p)}{2^{1+2p}}} m^{1-\gamma},
\] (6.3)
where
\[
\tilde{C}_\gamma = \left( 2M + \sqrt{7C_\gamma} + \frac{2}{2^{1+2p}} \frac{1}{2^{1+2p}} - 1 + 2\sqrt{C_2(c_q + 1)} \right) (2C_\kappa, p)^{\frac{4-\gamma(2+p)}{2^{1+2p}}}.
\]

Proof. Proposition 5 ensures the existence of $V_1$ and $V_2$ of $Z_m$ with measure at most $\delta/5$ such that
\[
H_1(z, \lambda) \leq \frac{3\kappa^2 D(\lambda) \log(10/\delta)}{\lambda m} + \frac{1}{2} D(\lambda), \quad \forall z \in Z_m \setminus V_1
\]
and
\[
H_2(z, \lambda) \leq 8\kappa^2 (\kappa^2 + 1) \log^2(10/\delta) \left\{ \frac{D(\lambda)}{\lambda^2 m^2} + \frac{D(\lambda)}{\lambda m} \right\} + D(\lambda), \quad \forall z \in Z_m \setminus V_2.
\]
Proposition 6 tells us that there exists a subset $V_3$ of $Z_m$ with measure at most $\delta/2$ such that
\[
S_2(z, \lambda) \leq D(\lambda) + 4\kappa^2 (\kappa^2 + 1) \log^2(10/\delta) \left\{ \frac{D(\lambda)}{\lambda^2 m^2} + \frac{D(\lambda)}{\lambda m} \right\}
+ C_1 \log^2(10/\delta) \left\{ \frac{D(\lambda)}{\lambda^2 m} + \frac{D(\lambda)}{\lambda} + 1 \right\} m^{-\gamma/\tilde{C}_\gamma}, \quad \forall z \in Z_m \setminus V_3.
\]
By Lemma 5, we know that there exists a subset $U_R$ of $Z_m$ with measure at most $\delta/10$ such that for very $z \in W(R) \setminus U_R$,
\[
S_1(z, \lambda) \leq \frac{1}{2} \{E(f_{z,\lambda}) - E(f_{\rho})\} + C_{K, p} \log(10/\delta) m^{-\gamma/\tilde{C}_\gamma} \max\{R^2, M^2\}.
\]
Finally, combining the above four bounds with (2.11), we get our desired result by taking $V_R = V_1 \cup V_1 \cup V_3 \cup U_R$. □

Proposition 7 immediately yields a learning rate when we use a rough bound for $\|f_{z,\lambda}\|_2$.

Lemma 6. For almost every $z \in Z_m$, we have
\[
\|f_{z,\lambda}\|_2 \leq M / \sqrt{\lambda}.
\]
Hence $W(R) = Z_m$ for $0 < \lambda \leq 1$ and $R = M / \sqrt{\lambda}$.

Proof. The definition of $f_{z,\lambda}$ tells us that
\[
\lambda \|f\|_2^2 \leq E(f_{z,\lambda}) + \lambda \Omega(z_{f,\lambda}) \leq E(z) + 0 \leq \frac{1}{m} \sum_{i=1}^{m} (y_i - 0)^2 \leq M^2.
\]
So $\|f_{z,\lambda}\|_2 \leq M / \sqrt{\lambda}$ holds almost surely. □

6.2. Bounding the estimator by iteration

To get better error estimates, we shall apply an iteration technique to improve the rough bound for $\|f_{z,\lambda}\|_2$ given in Lemma 6. We will give a tight bound by using Proposition 7 iteratively. This technique can be found in [12,21].

Lemma 7. Under the same assumptions of Theorem 3, take $\lambda = m^{-\gamma}$ with $\gamma < \frac{2}{2^{1+2p}}$ and $0 < \delta < 1$, then with confidence $1 - \frac{2}{2^{-\gamma}(2\delta + \gamma)}$, there holds
\[
\|f_{z,\lambda}\|_2 \leq \tilde{C}_\gamma \left( \frac{\log 10}{\delta} \right)^{\frac{4-\gamma(2+p)}{2^{1+2p}}} m^{1-\gamma},
\] (6.3)
Thus we get the desired result.

Thus we have

We determine

But

\[ \rho(\text{holde}_f \text{or each} \right) \]

Let us apply the inclusion (6.4) for a sequence for radius \( R \) defined by \( R^{(0)} = M/\sqrt{\lambda} \) and

Taking \( \lambda = m^{-\gamma} \), we have

Since \( 2/\pi - \gamma > 0 \), then \( \lambda m^{-1} - \gamma < 1 \). We also have \( 1/\lambda m^2 \leq 1/\lambda m \leq 1 \), thus \( b_{m,\lambda} \) can be bounded as

Let us apply the inclusion (6.4) for a sequence for radius \( R \) defined by \( R^{(0)} = M/\sqrt{\lambda} \) and

\[ R(j) = a_{m,\lambda} \max\{R(j-1), M\} + b_{m,\lambda}, \quad j \in \mathbb{N}. \]

Lemma 6 gives the identity \( \forall V(0^{(0)}) = Z^m \). Note that \( b_{m,\lambda} \geq C_{m,\lambda} \gamma(2/\pi - \frac{\gamma}{2}) \geq M \). So \( R(j) \geq M \) and we see that

\[ \max\{R(j-1), M\} = R(j-1) \quad \text{and} \quad R(j-1) = a_{m,\lambda} R(j-1) + b_{m,\lambda}. \]

Since (6.4) holds for each \( R(j) \), we have \( \forall V(R(j-1)) \subseteq \forall V(R(j)) \cup V_{R(j-1)} \) with \( \rho(V_{R(j-1)}) \leq \delta \). Apply this inclusion for \( j = 1, 2, \ldots, j \), with \( \delta \) to be determined later. We see that

\[ Z^m = \forall V(R(0)) \subseteq \forall V(R(1)) \cup V_{R(0)} \subseteq \cdots \subseteq \forall V(R(j)) \cup \left( \bigcup_{j=0}^{j-1} V_{R(j)} \right). \]

But \( \rho\left(\bigcup_{j=0}^{j-1} V_{R(j)}\right) \leq J \delta \). So the measure of the set \( \forall V(R(j)) \) is at least \( 1 - J \delta \).

By the definition of the sequence, we have

\[ R(j) = a_{m,\lambda} R^{(0)} + b_{m,\lambda} \sum_{j=0}^{j-1} a_{m,\lambda} \]

Thus we have \( m^{-J(1/\pi - \frac{\gamma}{2}) + \frac{\gamma}{2}} \leq 1 \) and \( (2C_{k,\lambda} \log(10/\delta))^{\frac{1}{2}} \leq (2C_{k,\lambda} \log(10/\delta))^{\frac{2}{2-\gamma(p+2)}} \).

Finally when \( m \geq 2 \), with confidence \( 1 - J \delta \geq 1 - \frac{2}{2-\gamma(p+2)} \delta \), we have

\[ R(j) \leq \left( M + \frac{2^{\frac{1}{\pi} - \frac{\gamma}{2}}}{2^{\frac{1}{\pi} - \frac{\gamma}{2}} - 1} + C_{m,\lambda} + 2\sqrt{2(C_{q} + 1)} \right) (2C_{k,\lambda} \log(10/\delta))^{\frac{2-\gamma(p+2)}{2-\gamma(p+2)}} m^{\frac{1}{2}-\gamma}. \]

Thus we get the desired result. \( \square \)
6.3. Deriving the learning rates

We are in the position to prove our main results.

**Proof of Theorem 3.** For any $0 < \delta < 1$ and $\lambda = m^{-\gamma}$ with $\gamma < \frac{2}{2+p}$, since $\tilde{f} = \frac{2}{z^{-\gamma(2+p)}}$, by Lemma 7, with confidence $1 - \frac{1}{j+1} \delta$, we have

$$\|f_{z,\lambda}\|_z \leq \tilde{C}_{\gamma} \left( \frac{\log(10(\tilde{f} + 1)/\delta)}{\delta} \right)^{\frac{4-\gamma(2+p)}{2(2+p)}} m^{\frac{1-\gamma}{2}}.$$

We take $R = \tilde{C}_{\gamma} \left( \frac{\log(10(\tilde{f} + 1)/\delta)}{\delta} \right)^{\frac{4-\gamma(2+p)}{2(2+p)}} m^{\frac{1-\gamma}{2}}$, then $R \geq M$ and the measure of the subset $\mathcal{W}(R)$ of $Z^m$ is at least $1 - \frac{1}{j+1} \delta$. Applying Proposition 7, we see that there is a subset $V_R$ of $Z^m$ with measure at most $\frac{1}{j+1} \delta$ such that

$$\mathcal{E}(f_{z,\lambda}) - \mathcal{E}(f_{\rho}) \leq 2C_{k,\rho} \log \left( \frac{10(\tilde{f} + 1)/\delta}{\delta} \right) m^{-\frac{2}{2+p}} R^2 + 7c_q m^{-q} \gamma + 4(c_q + 1)C_2 \log^2(10(\tilde{f} + 1)/\delta) m^{-q} \gamma, \quad \forall z \in \mathcal{W}(R) \setminus V_R.$$

But the measure of the set $\mathcal{W}(R) \setminus V_R$ is at least $1 - \delta$. So we know that with confidence at least $1 - \delta$, we have

$$\|f_{z,\lambda} - f_{\rho}\|_{L^{\infty}_X}^2 \leq (2C_{k,\rho} \tilde{C}_{\gamma}^2 + 7c_q + 4(c_q + 1)C_2 \log(10(\tilde{f} + 1)/\delta)) \left( \frac{4-\gamma(2+p)}{2(2+p)} m^{-q} \gamma \right).$$

Thus we complete our proof with $\tilde{C}_{k,\rho,\gamma} = 2C_{k,\rho} \tilde{C}_{\gamma}^2 + 7c_q + 4(c_q + 1)C_2$. \qed

Next, we will give the proofs of Theorem 1 and Theorem 2.

**Proof of Theorem 1.** Since $B_1$ always satisfies the capacity condition (2.16) with $p = \frac{2n}{n+2}$, then

$$\tilde{f} = \frac{2}{2 - \gamma(2+p)} \leq \frac{1}{\epsilon}.$$ 

By Proposition 4, $f_{\rho}$ lies in the range of $L^r_K$ implies

$$\mathcal{D}(\lambda) \leq C_r \|L^{-r}_{-K} f_{\rho}\|_{L^r_{-K}}^2 \lambda^{\min[2r,1]}.$$

Replacing $\tilde{f}$ by $\frac{1}{\epsilon}$ in Theorem 3, when $m \geq 2$, our conclusion holds with $c_q = C_r \|L^{-r}_{-K} f_{\rho}\|_{L^r_{-K}}^2 \lambda^{2r} q = \min[2r,1]$ and $C_e = (4M^2 + 14 + 8C_2 + (\frac{\sqrt{2}q + 2}{\sqrt{2}q - 1}) + (2C_{k,\rho})^{2+\frac{1}{\gamma}})(c_q + 1)$. \qed

**Proof of Theorem 2.** For any $0 < \epsilon < 1$, set $s = \max\{\frac{2-\epsilon}{2\epsilon} - 1, \frac{1}{2}\}$. Since $K$ is $C^\infty$ on $X$, $B_1$ satisfies the capacity condition (2.16) with $p = n/s < 2\epsilon/(2 - \epsilon)$, thus we have $\frac{2}{2+p} \geq 1 - \frac{\epsilon}{2} > 1 - \epsilon = \gamma$, hence we get $\tilde{f} \leq \frac{1}{\epsilon}$. Condition $f_{\rho} \in \mathcal{H}_K$ implies $f_{\rho}$ is in the range of $L^1_K$, thus Proposition 4 ensures the decay condition (2.13) holds with

---

**Table 1**

<table>
<thead>
<tr>
<th>Notations</th>
<th>Meaning</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{E}(f)$</td>
<td>Generalization error for a function $f$</td>
<td>252</td>
</tr>
<tr>
<td>$f_{\rho}$</td>
<td>Regression function</td>
<td>252</td>
</tr>
<tr>
<td>$\mathcal{E}(f) - \mathcal{E}(f_{\rho})$</td>
<td>Excess generalization error for a function $f$</td>
<td>252</td>
</tr>
<tr>
<td>$\mathcal{H}_K$</td>
<td>Reproducing kernel Hilbert space</td>
<td>253</td>
</tr>
<tr>
<td>$\mathcal{H}_{K,L}$</td>
<td>Data dependent hypothesis space</td>
<td>253</td>
</tr>
<tr>
<td>$\mathcal{D}(\lambda)$</td>
<td>Coefficient-based regularization terms for a function $f$</td>
<td>253</td>
</tr>
<tr>
<td>$f_{\lambda}$</td>
<td>Output function of learning algorithm (1.6)</td>
<td>253</td>
</tr>
<tr>
<td>$L^r_K$</td>
<td>Integral operator on $L^r_{-K}$</td>
<td>253</td>
</tr>
<tr>
<td>$\mathcal{D}_0$</td>
<td>Partial isometry on $L^r_{-K}$</td>
<td>254</td>
</tr>
<tr>
<td>$B_R$</td>
<td>Hilbert space defined by $L_K$</td>
<td>255</td>
</tr>
<tr>
<td>$\beta_i$</td>
<td>Class of functions for some given $R &gt; 0$</td>
<td>255</td>
</tr>
<tr>
<td>$\frac{1}{\epsilon}$</td>
<td>Regularizing functions</td>
<td>255</td>
</tr>
<tr>
<td>$\gamma(r,\lambda)$</td>
<td>Transitional function in error decomposition</td>
<td>256</td>
</tr>
<tr>
<td>$\mathcal{S}(\mathcal{E},\epsilon), \mathcal{S}(\mathcal{E},\lambda)$</td>
<td>Sample errors</td>
<td>256</td>
</tr>
<tr>
<td>$\mathcal{N}(\lambda)$</td>
<td>Hypothesis errors</td>
<td>256</td>
</tr>
<tr>
<td>$\mathcal{D}(\lambda)$</td>
<td>Approximation error</td>
<td>256</td>
</tr>
<tr>
<td>$\mathcal{N}(\lambda)$</td>
<td>Covering numbers</td>
<td>257</td>
</tr>
</tbody>
</table>
cq = \frac{5}{4} \| f_\rho \|_k^2 \text{ and } q = 1. \text{ Finally, using Theorem 3, by replacing } \tilde{J} \text{ by } \frac{2}{q}, \text{ when } m \geq 2, \text{ we get our conclusion with }
\tilde{C}_m = \left(4M^2 + 14 + 8C_2 + \frac{2^q+1}{2^q-1} + (2C_{k,p})^{2+\frac{3}{q}}\right)(c_q + 1). \quad \Box

For notations, see Table 1.

References