Learning Rates of Least-Square Regression with ℓ^1-Regularizer

Abstract

1 Introduction

In this paper we consider the least-square regression algorithm with ℓ^1-regularizer. The main result will be the satisfactory learning rates.

In the formal setting of the regression problem, we often have a compact metric space (X, d) as the input space and $Y = \mathbb{R}$ as the output space. Let ρ be a Borel probability measure on $Z = X \times Y$. For each pair $(x, y) \in Z$, the prediction accuracy of a predictor $f : X \to Y$ could be measured by the least-square loss $(f(x) - y)^2$. The generalization error for f with respect to ρ is defined as

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho. \tag{1.1}$$

The function that minimizes the error is called the regression function. It is given by

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

where $\rho(\cdot|x)$ is the conditional probability measure at x induced by ρ. A standard assumption on $\rho(\cdot|x)$ which we will use throughout the paper is that, for some $M \geq 0$, $\rho(\cdot|x)$ is almost everywhere supported on $[-M, M]$, that is, $y \leq M$ almost surely with respect to ρ. It is immediately from the definition of f_ρ that $|f_\rho(x)| \leq M$. In this paper, without loss of generality, we assume $M \geq 1$.

Since ρ is usually unknown, f_ρ cannot be obtain directly. The target of the regression problem is to learn regression function or to find good approximations from a set of samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ which is drawn independently according to the measure

ρ. This is a typical ill-posed problem and regularization technique is needed. Set the empirical error as

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2.$$

It is a discretization of the error $\mathcal{E}(f)$. Given a hypothesis space $\mathcal{H}$ (a set of functions form X to Y) and a regularizer (a penalty functional) $\Omega : \mathcal{H} \to \mathbb{R}_+$, the Tikhonov regularization scheme searching for an approximation of f_ρ is described as

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \big\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\Omega(f)\big\}. \tag{1.3}$$

Here $\lambda = \lambda(m) \geq 0$ usually depended on m is called the regularization parameter. Note that by choosing different hypothesis space $\mathcal{H}$ and regularizer Ω, we will get different learning algorithms. The efficiency of the algorithm is measured by the difference between $f_{\mathbf{z}}$ and the regression function f_ρ. Because of the least-square nature, the measurement is the weighted L^2 metric in $L^2_{\rho_X}$ defined as $\|f\|_{L^2_{\rho_X}} = (\int_X |f(x)|^2 d\rho_X)^{1/2}$, where ρ_X is the marginal distribution of ρ on X. One can easily check that

$$\|f_{\mathbf{z}} - f_\rho\|^2_{L^2_{\rho_X}} = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho). \tag{1.4}$$

So estimating error (1.4) for the algorithm (1.3) is the main goal of theoretic analysis.

One concrete example is the least-square regularization scheme with application of Mercer kernels [2]. Recall that, a Mercer kernel K is a function on $X \times X$ which is continuous and positive semi-definite, $(\mathcal{H}_K, \|\cdot\|_K)$ is the associated reproducing kernel Hilbert space [1], then the scheme is given by

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_K} \big\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|^2_K\big\}. \tag{1.5}$$

It has been well understood due to many literatures [3, 4, 5, 6]. The rates of convergence of $f_{\mathbf{z}}$ to f_ρ in $L^2_{\rho_X}$-space has been estimated by means of properties of ρ and K.

In this paper, we are interested in a least-square learning algorithm for regression with ℓ^1-regularizer. Given a symmetric and continuous kernel function $K : X \times X \to \mathbb{R}$ which is not necessarily positive semi-definite, we consider the following sample dependent hypothesis space defined by

$$\mathcal{H}_{K,\mathbf{z}} = \left\{\sum_{i=1}^{m}\alpha_i K_{x_i} : \alpha_i \in \mathbb{R}\right\},$$

where $K_t(\cdot) = K(\cdot, t)$. The learning algorithm is given by

$$f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_{K,\mathbf{z}}} \big\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\Omega_{\mathbf{z}}(f)\big\}, \tag{1.6}$$

where

$$\Omega_{\mathbf{z}}(f) = \sum_{i=1}^{m} |\alpha_i| \quad for\ f = \sum_{i=1}^{m} \alpha_i K_{x_i}.$$

The motivation we consider this algorithm is that the ℓ^1-regularizer often leads to sparsity of the regression coefficients $\{\alpha_i\}$ with properly chosen regularization parameter λ. This phenomena has been empirically observed in LASSO algorithm [10, 11, 12] and verified in the literature of compressed sensing[13]. Some theoretic work has been done about the least square regression with ℓ^1-regularizer [7, 8, 9]. For example, if K is Lipschitz continuous, it is proved in [8] that

$$\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} = \mathcal{O}(m^{-1/[3(n+1)]}),$$

under the assumption that f_ρ in the range of L_K^2 and ρ_X satisfies condition L_τ with $\tau = n$ (see definition 2). Here, L_K is the integral operator on the space $L^2_{\rho_X}$ defined by $L_K f = \int_X K_u f(u) d\rho_X(u)$. The convergence rate stated is low and depends on the dimension n of the input space X which is often large for learning problems. In [9] under the assumption that $K \in C^s(X \times X)$ with $s \geq 2$, by making fully use of higher order regularity of K and induced approximation, the learning rate can be improved to $\frac{1}{3} - \epsilon$ for any $0 < \epsilon < 1/3$ if K is C^∞ kernels.

Our setting is mainly followed by [9]. We will consider how fast $f_{\mathbf{z},\lambda}$ approximates f_ρ as the sample size m increases. Learning rates will be given in terms of the input space X, and the measure ρ and the kernel K. We assume X is a compact subset of $\mathbb{R}^n$ which satisfies an interior cone condition.

Definition 1. *A subset X of $\mathbb{R}^n$ is said to satisfy an interior cone condition, if there exists an angle $\theta \in (0, \pi/2)$, a radius $r > 0$, and a unit vector $\xi(x)$ for every $x \in X$ such that the cone*

$$C(x, \xi(x), \theta, r) = \left\{x + ty : y \in \mathbb{R}^n, |y| = 1, y^T \xi(x) \geq cos\theta, t \in [0, r]\right\}$$

is contained in X.

The sampling process on X is based on the marginal distribution ρ_X, we assume that ρ_X satisfies condition L_τ.

Definition 2. *A probability measure ρ_X on X is said to be satisfy condition L_τ with $0 < \tau < \infty$ if there exists some $C_\tau > 0$ such that for any ball $B(x, r) = \{u \in X : d(u, x) < r\}$, we have*

$$\rho_X(B(x, r)) \geq C_\tau r^\tau \quad \forall x \in X, 0 < r \leq 1. \tag{1.7}$$

If X satisfies interior cone condition and ρ is the uniform distribution on X, then (1.7) holds with $\tau = n$ and C_τ depends on X.

In this paper, we will apply a refined uniform concentration inequality involving the ℓ^2-empirical covering numbers (see definition 5) to derive learning rates and the iteration method will be used to give a sharper bound for algorithm (1.6). Firstly, let us state a result for C^∞ kernels.

Theorem 1. *Assume that X satisfies an interior cone condition, $K \in C^\infty(X \times X)$, f_ρ lies in the range of L_K^2, and ρ_X satisfies condition L_τ with some $\tau > 0$. Let $0 < \delta < 1$, $0 < \epsilon < 1/2$ and $\lambda = m^{\epsilon/2-1/2}$. If $m > \widetilde{M}_{\delta,\epsilon}$, then with confidence $1-\delta$, we have*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{L^2_{\rho_X}}^2 \le C_{X,\rho,K}\bigg(1 + \log(4(1/\epsilon+1)/\delta)\bigg)m^{-(\frac{1}{2}-\epsilon)},$$

where

$$\widetilde{M}_{\delta,\epsilon} = C_M^{4/\epsilon^2}\bigg\{\log(m+1) + \log(4(1/\epsilon+1)/\delta)\bigg\}^{4\max\{\frac{1}{\epsilon^2},\frac{n}{\tau\epsilon^2}\}},$$

and $C_{X,\rho,K}$, C_M are constant independent of m, δ and ϵ.

Obviously, our result is shaper than [9].

2 Main Result

Except for the regularizers, the main difference between algorithms (1.5) and (1.6) is that, the hypothesis space in the first algorithm is independent on samples. A useful approach for getting learning rate for regularization schemes with sample independent hypothesis spaces is error decomposition [5] which decomposes the total error (1.4) into the sum of a sample error and a regularization error (or approximation error). The main difficulty with algorithm (1.6) is the dependence of the hypothesis space $\mathcal{H}_{K,\mathbf{z}}$ on $\mathbf{z}$. This was pointed out in [7] where a modified error decomposition technique is introduced by means of an extra hypothesis error. In order to do error decomposition for algorithm (1.6), we use ℓ^1 sequences to define a function space.

Definition 3. *Take the Banach Space $\mathcal{H}_0 = \big\{f : f = \sum_{j=1}^\infty \alpha_j K_{x_j}, \{\alpha_j\} \in \ell^1, \{x_j\} \subset X\big\}$ with the norm*

$$\|f\| = \inf\left\{\sum_{j=1}^\infty |\alpha_j| : f = \sum_{j=1}^\infty \alpha_j K_{x_j}\right\}.$$

Since X is compact,when $K \in C^s(X \times X)$ with $s \ge 0$, $\mathcal{H}_0$ can be regarded as a subspace of $C^s(X)$ with the inclusion map $I : \mathcal{H}_0 \to C^s(X)$ bounded as

$$\|f\|_{C^s(X)} \le \|K\|_{C^s(X\times X)}\|f\|, \quad \forall f \in \mathcal{H}_0. \tag{2.1}$$

Denote $\kappa = \|K\|_{C(X\times X)}$. If we set $B_R = \{f \in \mathcal{H}_0 : \|f\| \le R\}$, then for all $R > 0$, the set $I(B_R)$ is compact in $C^s(X)$.

We introduce a regularizing functions as

$$f_\lambda = \arg\min_{f\in\mathcal{H}_0} \{\mathcal{E}(f) + \lambda\|f\|\} \tag{2.2}$$

Using (1.4), the following error decomposition process is used in [8, 9].

Proposition 1. *Let $f_{\mathbf{z},\lambda}$ be defined by (1.6) with $\lambda > 0$. Then*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} \le \mathcal{S}(\mathbf{z},\lambda) + \mathcal{H}(\mathbf{z},\lambda) + \mathcal{D}(\lambda). \tag{2.3}$$

Where

$$\begin{aligned}
\mathcal{S}(\mathbf{z},\lambda) &= \mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\\
\mathcal{H}(\mathbf{z},\lambda) &= \{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})\} - \{\mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\|\}\\
\mathcal{D}(\lambda) &= \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \lambda\|f_\lambda\|
\end{aligned}$$

$\mathcal{S}(\mathbf{z},\lambda)$ is called sample error. Since generally $\mathcal{H}_{K,\mathbf{z}} \subset \mathcal{H}_0$, f_λ may not be in the space $\mathcal{H}_{K,\mathbf{z}}$. The second item $\mathcal{H}(\mathbf{z},\lambda)$ is so called hypothesis error due to the different hypothesis spaces. While the last item $\mathcal{D}(\lambda)$ is called regularization error. Obviously, form (1.4) and definition (2.2)

$$\mathcal{D}(\lambda) = \inf_{f\in\mathcal{H}_0} \left\{\|f - f_\rho\|^2_{L^2_{\rho_X}} + \lambda\|f\|\right\}.$$

Usually we use $\mathcal{D}(\lambda)$ to measure the approximation ability of $\mathcal{H}_0$. For getting rates we shall assume that for some constants $q \in (0,1]$ and $c_q > 0$,

$$\mathcal{D}(\lambda) \le c_q\lambda^q, \quad \forall\lambda > 0. \tag{2.4}$$

A sufficient condition to estimate the regularization error $\mathcal{D}(\lambda)$ is given in [8]. We state it as follows.

Proposition 2. *If $f_\rho = L^s_K g$ for some $0 < s \le 2$ and $g \in L^2_{\rho_X}$, then*

$$\mathcal{D}(\lambda) \le \left(\|g\|^2_{L^2_{\rho_X}} + \kappa\|g\|_{L^2_{\rho_X}}\right)\lambda^{\frac{2s}{s+2}} \qquad \forall\lambda > 0. \tag{2.5}$$

If $K \in C^s(X\times X)$ with $s \ge 2$, applying the symmetry of the kernel and a local polynomial reproduction formula form the literature of multivariate approximation [14, 15], a estimation for hypothesis error $\mathcal{H}(\mathbf{z},\lambda)$ given by [9] is stated as follows.

Proposition 3. *Suppose X satisfies an interior cone condition with radius $R_X > 0$ and angle $\theta \in (0,\pi/2)$, if ρ_X satisfies condition L_τ with some $\tau > 0$, $K \in C^s(X\times X)$ with*

$s \geq 2$, *(2.4) is valid, then for* $0 < \delta < 1$, $0 < \lambda \leq 1$ *and* $m \geq \widetilde{C}_0 \left(\log(2/\delta) + \log(m+1)\right)$, *with confidence* $1 - \frac{\delta}{2}$,

$$\mathcal{H}(\mathbf{z}, \lambda) \leq c_q \lambda^q + \widetilde{C}_1 \lambda^{2(q-1)} \left(\frac{\log(2/\delta) + \log(m+1)}{m} \right)^{\frac{s}{\tau}} \tag{2.6}$$

where $\widetilde{C}_0$ *and* $\widetilde{C}_1$ *depend on* $X, \tau, C_\tau, R_X, \theta, n, s$ *and* $\|K\|_{C^s}$.

In this paper, we focus on estimating the sample error $\mathcal{S}(\mathbf{z}, \lambda)$. Since the quantity $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})$ needs to be estimated by some uniform law of large numbers. To this end, we need the capacity of the hypothesis space, which plays an essential role in sample error estimates. Covering number as an important measurement of the capacity of a function set has been well studied in a lot of literatures [16, 17, 18]. Firstly, we will give a general definition of covering number.

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

$$\mathcal{N}(S, \epsilon, d) = \min \left\{ n \in \mathbb{N} : \exists \{s_j\}_{j=1}^n \subset \mathscr{M} \text{ such that } S \subset \bigcup_{j=1}^n B(s_j, \epsilon) \right\},$$

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

Next, we introduce the covering number of function class. Let d_2 denote the normalized ℓ^2-metric on the Euclidian space $\mathbb{R}^m$ given by

$$d_2(\mathbf{a}, \mathbf{b}) = \left(\frac{1}{m} \sum_{i=1}^m |a_i - b_i|^2 \right)^{1/2}$$

for $\mathbf{a} = (a_i)_{i=1}^m, \mathbf{b} = (b_i)_{i=1}^m \in \mathbb{R}^m$.

Definition 5. *Let* $\mathcal{F}$ *be a set of functions on* X, $\mathbf{x} = (x_i)_{i=1}^m \subset X^m$ *and* $\mathcal{F}|_{\mathbf{x}} = \{(f(x_i))_{i=1}^m : f \in \mathcal{F}\} \subset \mathbb{R}^m$. *The* ℓ^2*-empirical covering number of* $\mathcal{F}$ *associated to* $\mathbf{x}$ *is defined as*

$$\mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \epsilon) = \mathcal{N}(\mathcal{F}|_{\mathbf{x}}, \epsilon, d_2).$$

Moreover, the ℓ^2*-empirical covering number of* $\mathcal{F}$ *is given by*

$$\mathcal{N}_2(\mathcal{F}, \epsilon) = \sup_{m \in \mathbb{N}} \sup_{\mathbf{x} \in X^m} \mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \epsilon).$$

While the uniform covering number of $\mathcal{F}$ *is defined as* $\mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_\infty)$, *the covering number of* $\mathcal{F}$ *with respect to the* L^∞*-metric.*

Note that ℓ^2-empirical covering number is always smaller than the uniform covering number and hence help to yield sharper bounds. Let us state our main result on analysis.

Theorem 2. *Suppose X satisfies an interior cone condition, if ρ_X satisfies condition L_τ with some $\tau > 0$, $K \in C^s(X \times X)$ with $s \geq 2$ and approximation error condition (2.4) is valid. Assume B_1 satisfy the capacity condition with a constant $c_p > 0$ and $p \in (0, 2)$, that is*

$$\log \mathcal{N}_2(B_1, \epsilon) \leq c_p \left(\frac{1}{\epsilon}\right)^p, \ \forall \epsilon > 0. \tag{2.7}$$

Set

$$\gamma_1 = \max\left\{\frac{s}{\tau}, 1\right\}, \quad \gamma_2 = \min\left\{\frac{s}{\tau}, 1\right\}, \tag{2.8}$$

and

$$\gamma_3 = \begin{cases} \min\left\{\frac{1}{2+p}, \frac{\gamma_2}{2(1-q)}\right\} & \text{if } 0 < q < 1 \\ \frac{1}{2+p} & \text{if } q = 1 \end{cases}. \tag{2.9}$$

Take $\lambda = m^{\epsilon-\gamma_3}$ with $0 < \epsilon < \gamma_3$. Define

$$\gamma_5 = \min\left\{\gamma_2 - (3-2q)(\gamma_3 - \epsilon), (q-1)(\gamma_3 - \epsilon)\right\} \text{ and } \beta = \gamma_3/(2\epsilon) + 1/2, \tag{2.10}$$

for any $0 < \delta < 1$ and $m > m_{\delta,\epsilon}$, with confidence $1 - \delta$, we have

$$\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} \leq C_{X,\rho,K} C_0^{2\beta} M^{4\beta} \left(1 + \log(4(\beta+1)/\delta)\right) \left\{\log(m+1) + \log(4(\beta+1)/\delta)\right\}^{2\max\{\beta,\gamma_1\}} m^{-\Theta},$$

where

$$\Theta = \min\left\{\frac{2}{p+2} + 2\gamma_5, \gamma_2 - 2(1-q)(\gamma_3 - \epsilon), q(\gamma_3 - \epsilon)\right\}$$

and

$$m_{\delta,\epsilon} := \max\left\{\widetilde{C}_0 \left(\log(2(\beta+1)/\delta) + \log(m+1)\right), \left\{2C_0 \left(1 + \log(4(\beta+1)/\delta)\right) M^2\right\}^{\frac{1}{2\epsilon}}, \right.$$
$$\left. \left\{4C_1 \left(\log(4(\beta+1)/\delta) + \log(m+1)\right)^{\gamma_1} / M^2\right\}^{\frac{1}{\gamma_4}}, \left\{4C_3/M^2\right\}^{\frac{1}{q(\gamma_3-\epsilon)}} \right\}$$

with γ_4 defined as

$$\gamma_4 = \begin{cases} 2(1-q)\epsilon & \text{if } 0 < q < 1 \\ \gamma_2 & \text{if } q = 1 \end{cases}. \tag{2.11}$$

The constants $C_{X,\rho,K}$, $\widetilde{C}_0$, C_0, C_1 and C_3 is independent of m, δ and ϵ.

Remark 1. *We will show that the capacity condition (2.7) always hold for the space $\mathcal{H}_0$.*

3 Estimates for the Sample Error

Under the assumption that all the samples are independent drawn form ρ and $|y| \leq M$ almost truly, we are in the position to estimate the sample error $\mathcal{S}(\mathbf{z}, \lambda)$ which can be rewritten as

$$\mathcal{S}(\mathbf{z}, \lambda) = \mathcal{S}_1(\mathbf{z}, \lambda) + \mathcal{S}_2(\mathbf{z}, \lambda)$$

where

$$\mathcal{S}_1(\mathbf{z}, \lambda) = \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_{\mathbf{z}}(f_\rho)\} - \{\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho)\}$$

and

$$\mathcal{S}_2(\mathbf{z}, \lambda) = \{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\}.$$

The first term can be estimated by using one-side Bernstein inequality.

Lemma 1. *Let ξ be a random variable on a probability space Z with expectation $\mu = \mathbb{E}\xi$ and variance $\sigma^2(\xi) = \sigma^2$. If $|\xi(z) - \mu| \leq M_\xi$ for almost all $z \in Z$, then*

$$\frac{1}{m}\sum_{i=1}^{m} \xi(z_i) - \mu \leq \frac{2M_\xi \log \frac{1}{\delta}}{3m} + \sqrt{\frac{2\sigma^2 \log \frac{1}{\delta}}{m}} \tag{3.1}$$

with confidence $1 - \delta$.

Proposition 4. *For any $0 < \delta < 1$, with confidence $1 - \delta/4$,*

$$\mathcal{S}_1(\mathbf{z}, \lambda) \leq \frac{7\left(3M + \kappa\mathcal{D}(\lambda)/\lambda\right)^2 \log \frac{4}{\delta}}{3m} + \frac{1}{2}\mathcal{D}(\lambda) \tag{3.2}$$

Proof. Form the definition of $\mathcal{D}(\lambda)$ and (1.4), we know that

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

It follows from (2.1) that

$$\|f_\lambda\|_\infty \leq \kappa\|f_\lambda\| \leq \kappa\mathcal{D}(\lambda)/\lambda.$$

Set $\xi(z) = (y - f_\lambda(x))^2 - (y - f_\rho(x))^2$, since $|f_\rho(x)| \leq M$ almost everywhere, we have

$$|\xi(z)| \leq \left(3M + \|f_\lambda\|_\infty\right)\left(M + \|f_\lambda\|_\infty\right) \leq c := \left(3M + \kappa\mathcal{D}(\lambda)/\lambda\right)^2.$$

Hence $M_\xi = 2c$. Moreover,

$$\mathbb{E}(\xi^2) = \int_Z \left\{f_\lambda(x) + f_\rho(x) - 2y\right\}^2 \left\{f_\lambda(x) - f_\rho(x)\right\}^2 d\rho \leq \left(3M + \|f_\lambda\|_\infty\right)^2 \|f_\lambda - f_\rho\|^2_{L^2_{\rho_X}}$$

which implies that $\sigma^2(\xi^2) \leq \mathbb{E}(\xi^2) \leq c\mathcal{D}(\lambda)$. Now applying lemma 1, with confidence $1 - \delta/4$, we have

$$\mathcal{S}_1(\mathbf{z}, \lambda) = \frac{1}{m}\sum_{i=1}^{m} \xi(z_i) - \mathbb{E}(\xi) \leq \frac{4c \log \frac{4}{\delta}}{3m} + \sqrt{\frac{2c\mathcal{D}(\lambda) \log \frac{4}{\delta}}{m}} \leq \frac{7c \log \frac{4}{\delta}}{3m} + \frac{1}{2}\mathcal{D}(\lambda)$$

the last inequality holds since $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$. □

Before estimating $\mathcal{S}_2(\mathbf{z},\lambda)$, we will state some results about the covering number concerning $\mathcal{H}_0$. Recall B_1 is the unit ball in $\mathcal{H}_0$ and define $\mathrm{Diam}(X) := \max_{x,y\in X}\|x-y\|$.

Proposition 5.

(i) *If K is Lipschitz continuous of order α with $0<\alpha\le 1$, that is for some constant $c_\alpha>0$,*

$$|K(t,x)-K(t,x')|\le c_\alpha|x_1-x_2|^\alpha,\quad \forall t,x,x'\in X. \tag{3.3}$$

Then for all $\epsilon>0$,

$$\log\mathcal{N}_2\left(B_1,\epsilon\right)\le\widetilde{C}_2\left(\frac{1}{\epsilon}\right)^{2n/(n+2\alpha)} \tag{3.4}$$

where $\widetilde{C}_2$ depend on n,α,c_α,κ and $\mathrm{Diam}(X)$.

(ii) *If $K\in C^s(X\times X)$ for some $s>0$, then there is $\widetilde{C}_3>0$ depending on X,s and $\|K\|_{C^s}$ such that for all $\epsilon>0$,*

$$\log\mathcal{N}\left(B_1,\epsilon,\|\cdot\|_\infty\right)\le\widetilde{C}_3\left(\frac{1}{\epsilon}\right)^{n/s}. \tag{3.5}$$

The result (ii) is directly result form [17, 18]. We leave the proof for (i) in the appendix. We will use the following uniform concentration inequality stated in [19].

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

$$\log\mathcal{N}_2\left(\mathcal{F},\epsilon\right)\le a\epsilon^{-p},\quad\forall\epsilon>0, \tag{3.6}$$

then there exist a constant c_p' depending only on p such that for any $t>0$, with probability at least $1-e^{-t}$, there holds

$$\mathbb{E}f-\frac{1}{m}\sum_{i=1}^m f(z_i)\le\frac{1}{2}\eta^{1-\alpha}\left(\mathbb{E}f\right)^\alpha+c_p'\eta+2\left(\frac{ct}{m}\right)^{\frac{1}{2-\alpha}}+\frac{18Bt}{m},\quad\forall f\in\mathcal{F}, \tag{3.7}$$

where

$$\eta:=\max\left\{c^{\frac{2-p}{4-2\alpha+p\alpha}}\left(\frac{a}{m}\right)^{\frac{2}{4-2\alpha+p\alpha}},B^{\frac{2-p}{2+p}}\left(\frac{a}{m}\right)^{\frac{2}{2+p}}\right\}.$$

From (3.4), we know that condition (3.6) in lemma 2 always holds for B_1 if $K\in C^s(X\times X)$ with $s>0$. We apply lemma 2 to a set of function $\mathcal{F}_R$ with $R>0$, where

$$\mathcal{F}_R=\left\{(y-f(x))^2-(y-f_\rho(x))^2|f\in B_R\right\} \tag{3.8}$$

Proposition 6. *Assume B_1 satisfy the capacity condition with a constant $c_p > 0$ and $p \in (0,2)$, that is*

$$\log \mathcal{N}_2(B_1, \epsilon) \leq c_p \left(\frac{1}{\epsilon}\right)^p, \ \forall \epsilon > 0.$$

If $R \geq M$, then for all $f \in B_R$ and $\delta \in (0,1)$, with confidence $1 - \delta/4$, we have

$$\begin{aligned} &\{\mathcal{E}(f) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_\rho)\} \\ &\quad \leq \frac{1}{2}\{\mathcal{E}(f) - \mathcal{E}(f_\rho)\} + C_{\kappa,p}\left(1 + \log(4/\delta)\right)\left(\frac{1}{m}\right)^{\frac{2}{2+p}} R^2 \end{aligned} \tag{3.9}$$

where $C_{\kappa,p} = \max\left\{c_p'(3+\kappa)^{\frac{4-2p}{2+p}}(c_p(2+2\kappa))^{\frac{2}{2+p}}, 20(3+\kappa)^2\right\}$.

Proof. Consider the set $\mathcal{F}_R$. Each function $g \in \mathcal{F}_R$ has the form $g(z) = (y - f(x))^2 - (y - f_\rho(x))^2$ with $f \in B_R$. Hence $\mathbb{E}(g) = \mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}}$, $(1/m)\sum_{i=1}^m g(z_i) = \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_\rho)$ and

$$g(z) = (f(x) - f_\rho(x))\left\{(f(x) - y) + (f_\rho(x) - y)\right\}.$$

Since $\|f\|_\infty \leq \kappa\|f\| \leq \kappa R$ and $|f_\rho(x)| \leq M$, we find that

$$|g(z)| = (\kappa R + M)(\kappa + 3R) \leq (3M + \kappa R)^2.$$

and

$$\mathbb{E}g^2 = \int_Z (2y - f(x) - f_\rho(x))^2 (f(x) - f_\rho(x))^2 \, d\rho \leq (3M + \kappa R)^2 \, \mathbb{E}g.$$

Moreover, since $\forall g_1, g_2 \in \mathcal{F}_R$,

$$|g_1(z) - g_2(z)| = |\,(y - f_1(x))^2 - (y - f_2(x))^2\,| \leq (2M + 2\kappa R)|f_1(x) - f_2(x)|,$$

there holds

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{F}_R, \epsilon) \leq \mathcal{N}_{2,\mathbf{x}}\left(B_R, \frac{\epsilon}{2M + 2\kappa R}\right) \leq \mathcal{N}_{2,\mathbf{x}}\left(B_1, \frac{\epsilon}{R(2M + 2\kappa R)}\right)$$

which implies

$$\log \mathcal{N}_2(\mathcal{F}_2, \epsilon) \leq c_p R^p (2M + 2\kappa R)^p \epsilon^{-p}.$$

Since $R \geq M$, using lemma 2 with $B = c = (3M + \kappa R)^2$, $\alpha = 1$ and $a = c_p R^p (2M + 2\kappa R)^p$, for $\forall g \in \mathcal{F}_R$ and $\delta \in (0,1)$, with confidence $1 - \delta/4$, there holds

$$\begin{aligned} \mathbb{E}g - \frac{1}{m}\sum_{i=1}^m g(z_i) &\leq \frac{1}{2}\mathbb{E}g + c_p'\left\{(3M + \kappa R)^2\right\}^{\frac{2-p}{2+p}}\left(\frac{c_p R^p (2M + 2\kappa R)^p}{m}\right)^{\frac{2}{2+p}} \\ &\quad + 20(3M + \kappa R)^2 \frac{\log(4/\delta)}{m} \\ &\leq \frac{1}{2}\mathbb{E}g + c_p'(3+\kappa)^{\frac{4-2p}{2+p}}(c_p(2+2\kappa))^{\frac{2}{2+p}}\left(\frac{1}{m}\right)^{\frac{2}{2+p}} R^2 + \frac{20(3+k)^2\log(4/\delta)}{m} R^2 \\ &\leq \frac{1}{2}\mathbb{E}g + C_{\kappa,p}\left(1 + \log(4/\delta)\right)\left(\frac{1}{m}\right)^{\frac{2}{2+p}} R^2. \end{aligned}$$

Thus we complete our proof. □

4 Error Bounds in a Weak Form

Now we can derive error bounds. For $R > 0$, denote

$$\mathcal{W}(R) = \{\mathbf{z} \in Z^m : \|f_{\mathbf{z},\lambda}\| \leq R\}. \tag{4.1}$$

Proposition 7. *Suppose X satisfies an interior cone condition, if ρ_X satisfies condition L_τ with some $\tau > 0$, $K \in C^s(X \times X)$ with $s \geq 2$ and approximation error condition (2.4) is valid. Assume B_1 satisfy the capacity condition (2.7). For all $0 < \lambda \leq 1$, $0 < \delta < 1$ and $R > M$, when*

$$m \geq \widetilde{C}_0 \left(\log(2/\delta) + \log(m+1)\right) \tag{4.2}$$

there is a set $V_R \subset Z^m$ with $\rho(V_R) \leq \delta$ such that, for all $\mathbf{z} \in \mathcal{W}(R)\backslash V_R$,

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) &\leq C_0 \left(1 + \log(4/\delta)\right) \left(\frac{1}{m}\right)^{\frac{2}{2+p}} R^2 \\ &+ C_1 \left(\log(4/\delta) + \log(m+1)\right)^{\max\{\frac{s}{\tau},1\}} \lambda^{2(q-1)} \left(\frac{1}{m}\right)^{\min\{\frac{s}{\tau},1\}} + C_2\lambda^q \end{aligned} \tag{4.3}$$

where $C_0 = 2C_{\kappa,p}$, $C_1 = \max\{2\widetilde{C}_1, \frac{14(3M+c_q\kappa)^2}{3}\}$ and $C_2 = 3c_q$.

Proof. When m satisfies (4.2), form proposition 3, there exists a subset $U_1 \in Z^m$ with $\rho(U_1) \leq \delta/2$ such that for every $\mathbf{z} \in Z^m\backslash U_1$,

$$\mathcal{H}(\mathbf{z},\lambda) \leq c_q\lambda^q + \widetilde{C}_1 \left(\log(2/\delta) + \log(m+1)\right)^{\frac{s}{\tau}} \lambda^{2(q-1)} \left(\frac{1}{m}\right)^{\frac{s}{\tau}}.$$

Form proposition 4, there exists a subset $U_2 \in Z^m$ with $\rho(U_2) \leq \delta/4$ such that for every $\mathbf{z} \in Z^m\backslash U_2$,

$$\mathcal{S}_1(\mathbf{z},\lambda) \leq \frac{7\left(3M + \kappa\mathcal{D}(\lambda)/\lambda\right)^2 \log\frac{4}{\delta}}{3m} + \frac{1}{2}\mathcal{D}(\lambda) \leq \frac{7(3M + c_q\kappa)^2 \log(4/\delta)}{3}\lambda^{2(q-1)}\frac{1}{m} + \frac{1}{2}c_q\lambda^q$$

the last inequality holds since $0 < \lambda \leq 1$ and $0 < q \leq 1$. Form proposition 6, for $R \geq M$, there exists a subset U_R with $\rho(U_R) \leq \delta/4$ such that for every $\mathbf{z} \in \mathcal{W}(R)\backslash U_R$,

$$\begin{aligned}\mathcal{S}_2(\mathbf{z},\lambda) &\leq \frac{1}{2}\left\{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)\right\} + C_{\kappa,p}\left(1 + \log(4/\delta)\right)\left(\frac{1}{m}\right)^{\frac{2}{2+p}} R^2 \\ &\leq \frac{1}{2}\left\{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})\right\} + C_{\kappa,p}\left(1 + \log(4/\delta)\right)\left(\frac{1}{m}\right)^{\frac{2}{2+p}} R^2.\end{aligned}$$

Finally, since $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) = \mathcal{H}(\mathbf{z},\lambda) + \mathcal{S}(\mathbf{z},\lambda) + \mathcal{D}(\lambda)$, let $V_R = U_1 \cup U_2 \cup U_R$, we get desire result. □

Learning rate in weak forms can be obtained from proposition 7.

Proposition 8. *Under the assumption of proposition 7. Let* $0 < \delta < 1$ *and* $\lambda = m^{-\vartheta}$, *when* $m \geq \widetilde{C}_0 \left(\log(2/\delta) + \log(m+1)\right)$, *with confidence* $1-\delta$, *we have*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} \leq \widetilde{C}\left(\log\frac{4}{\delta} + \log(m+1)\right)^{\max\{\frac{s}{\tau},1\}} m^{-\Theta}, \tag{4.4}$$

where

$$\Theta = \min\left\{q\vartheta, 1-2(1-q)\vartheta, \frac{s}{\tau} - 2(1-q)\vartheta, \frac{2}{2+p} - 2\vartheta\right\},$$

and $\widetilde{C} = C_0 M^4 + C_1 + C_2$ *is a constant independent of* m *or* δ.

Proof. The definition of $f_{\mathbf{z},\lambda}$ tells us that,

$$\lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \leq \mathcal{E}_{\mathbf{z}}(0) + 0 \leq \frac{1}{m}\sum_{i=1}^{m}(y_i - 0) \leq M^2$$

hold almost surely. Since $\|f_{\mathbf{z},\lambda}\| \leq \Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})$, we have $\|f_{\mathbf{z},\lambda}\| \leq M^2/\lambda$ for all most $\mathbf{z} \in Z^m$. Therefore, $\mathcal{W}(M^2/\lambda) = Z^m$. Take $R := M^2/\lambda \geq M$ due to $M \geq 1$ and $0 < \lambda \leq 1$. Form proposition 7, when $m \geq \widetilde{C}_0 \left(\log(2/\delta) + \log(m+1)\right)$, with confidence $1-\delta$, we have

$$\begin{aligned}\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} &\leq C_0\left(1 + \log(4/\delta)\right) M^4 \lambda^{-2}\left(\frac{1}{m}\right)^{\frac{2}{2+p}} \\ &+ C_1\left(\log(4/\delta) + \log(m+1)\right)^{\max\{\frac{s}{\tau},1\}} \lambda^{2(q-1)}\left(\frac{1}{m}\right)^{\min\{\frac{s}{\tau},1\}} + C_2\lambda^q.\end{aligned}$$

Using this inequality, let $\lambda = m^{-\vartheta}$, we get the desire result. □

Remark 2. *In [9], they give the bound in a similar form as (4.4) with* $\Theta = \min\{q\vartheta, \frac{s}{\tau} - 2(1-q)\vartheta, 1-2(1-q)\vartheta, \frac{1}{1+2n/s} - 2\vartheta\}$. *Our bound is sharper than [9] since* $p \leq n/s$ *always holds form (3.5).*

5 Strong Bound by Iteration

In this section, we will use the iteration technique to obtain strong error estimation. The method in the previous section was rough because we use the bound $\|f_{\mathbf{z},\lambda}\| \leq M^2/\lambda$ which is much worse than the bound for f_λ, namely, $\|f_\lambda\| \leq \mathcal{D}(\lambda)/\lambda$. Since $f_{\mathbf{z},\lambda}$ is a good approximation of f_λ, one would expect $\|f_{\mathbf{z},\lambda}\|$ to have some tighter bound. We shall prove this is the case with high probability by applying proposition 7 iteratively. We will use a similar iteration technology showed in [5]. The strong bound will be proved after two lemmas. Recall the set $\mathcal{W}(R)$ defined by (4.1). The following lemma is a direct result form proposition 7. Here, we mainly use the upper bound for R and the fact that $\lambda\|f_{\mathbf{z},\lambda}\|$ is bounded by $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})$. Recall (2.8), we denote $\gamma_1 = \max\left\{\frac{s}{\tau}, 1\right\}$ and $\gamma_2 = \min\left\{\frac{s}{\tau}, 1\right\}$.

Lemma 3. *Under the assumption of proposition 7. For any $0 < \delta < 1$, $0 < \lambda \leq 1$ and $M \leq R \leq M^2/\lambda$, there is a set $V_R \subset Z^m$ with $\rho(V_R) \leq \delta$ such that*

$$\mathcal{W}(R) \subset \mathcal{W}(a_m M^2 \lambda^{-2} R + b_m \lambda^{-1}) \cup V_R,$$

where

$$a_m = C_0 \left(1 + \log(4/\delta)\right) \left(\frac{1}{m}\right)^{\frac{2}{2+p}},$$
$$b_m = C_1 \left(\log(4/\delta) + \log(m+1)\right)^{\gamma_1} \lambda^{2(q-1)} \left(\frac{1}{m}\right)^{\gamma_2} + C_3 \lambda^q,$$

here $C_3 = \max\{C_2, M\}$.

Next, we will prove with high probability, tighter bound for $\|f_{z,\lambda}\|$ will be obtain by iteratively using lemma 3.

Lemma 4. *Under the assumption of Theorem 2 and γ_1, γ_2, γ_3, γ_4, γ_5, β are defined as the same in Theorem 2. Take $\lambda = m^{\epsilon - \gamma_3}$ with $0 < \epsilon < \gamma_3$. For any $0 < \delta < 1$ and $m \geq m_\delta$, with confidence $1 - \beta\delta$, there holds*

$$\|f_{\mathbf{z},\lambda}\| \leq \left\{ C_0^\beta \left(1 + \log(4/\delta)\right)^\beta M^{2\beta+2} + 2C_1 \left(\log(4/\delta) + \log(m+1)\right)^{\gamma_1} + 2C_3 \right\} m^{-\gamma_5}. \tag{5.1}$$

while

$$m_\delta = \max\left\{ \widetilde{C}_0 \left(\log(2/\delta) + \log(m+1)\right), \left\{2C_0 \left(1 + \log(4/\delta)\right) M^2\right\}^{\frac{1}{2\epsilon}}, \right.$$
$$\left. \left\{4C_1 \left(\log(4/\delta) + \log(m+1)\right)^{\gamma_1} / M^2\right\}^{\frac{1}{\gamma_4}}, \left\{4C_3/M^2\right\}^{\frac{1}{q(\gamma_3-\epsilon)}} \right\}.$$

Proof. We just consider the case $0 < q < 1$, the proof is the same for $q = 1$. For any $0 < \epsilon < \gamma_3$, take $\lambda = m^{-\vartheta}$ with $\vartheta = \gamma_3 - \epsilon$ then

$$\tilde{a}_m := a_m M^2 \lambda^{-2} = C_0 \left(1 + \log\frac{4}{\delta}\right) M^2 \left(\frac{1}{m}\right)^{\frac{2}{2+p} - 2\vartheta} \leq C_0 \left(1 + \log\frac{4}{\delta}\right) M^2 m^{-2\epsilon}. \tag{5.2}$$

We have the trivial bound $b_m \geq C_3 \geq M$ since $0 < q \leq 1$ and $0 < \lambda \leq 1$. Moreover, a simple computation show that

$$\begin{aligned} b_m &= C_1 \left(\log(4/\delta) + \log(m+1)\right)^{\gamma_1} \left(\frac{1}{m}\right)^{\gamma_2 - 2(1-q)\vartheta} + C_3 m^{-q\vartheta} \\ &\leq C_1 \left(\log(4/\delta) + \log(m+1)\right)^{\gamma_1} m^{-2(1-q)\epsilon} + C_3 m^{-q(\gamma_3-\epsilon)}. \end{aligned}$$

If $m \geq m_\delta$, we have $\tilde{a}_m \leq 1/2$ and $b_m \leq M^2/2$. For any $M \leq R \leq \frac{M^2}{\lambda}$, there holds

$$M \leq \tilde{a}_m R + b_m \lambda^{-1} \leq \frac{1}{2} R + M^2/(2\lambda) \leq M^2/\lambda. \tag{5.3}$$

Define a sequence $\{R^{(j)}\}_{j\in\mathbb{N}}$ by $R^{(0)} = M^2/\lambda$ and, for $j \geq 1$,

$$R^{(j)} = \tilde{a}_m R^{(j-1)} + b_m \lambda^{-1}.$$

Then proposition 8 proves that $\mathcal{W}(R^{(0)}) = Z^m$ and (5.3) guarantee lemma 3 holds for each $R^{(j)}$, that is, $\mathcal{W}(R^{(j-1)}) \subseteq \mathcal{W}(R^{(j)}) \cup V_{R^{(j-1)}}$ with $\rho(V_{R^{(j-1)}}) \leq \delta$. Apply this inclusion for $j = 1, 2, \ldots, J$, with J satisfying $\gamma_3/(2\epsilon) - 1/2 \leq J \leq \gamma_3/(2\epsilon) + 1/2$. We see that

$$Z^m = \mathcal{W}(R^{(0)}) \subseteq \mathcal{W}(R^1) \cup V_{R^{(0)}} \subseteq \cdots \subseteq \mathcal{W}(R^{(J)}) \cup \left(\bigcup_{j=0}^{J-1} V_{R^{(j)}} \right).$$

It follows that the measure of the set $\mathcal{W}(R^{(J)})$ is at least $1 - J\delta \geq 1 - \delta(\gamma_3/(2\epsilon) + 1/2)$. By the definition of the sequence, we have

$$R^{(J)} = \tilde{a}_m^J R^{(0)} + b_m \lambda^{-1} \sum_{j=0}^{J-1} \tilde{a}_m^j.$$

Since $\tilde{a}_m \leq \frac{1}{2}$, hence $\sum_{j=0}^{J-1} \tilde{a}_m^j \leq 2$. The bound $\tilde{a}_m \leq C_0 \left(1 + \log \frac{4}{\delta}\right) M^2 m^{-2\epsilon}$ and $R^{(0)} = M^2/\lambda = M^2 m^{\gamma_3 - \epsilon}$ yield

$$\tilde{a}_m^J R^{(0)} \leq C_0^J \left(1 + \log(4/\delta)\right)^J M^{2J+2} m^{\gamma_3 - 2J\epsilon - \epsilon}.$$

But $J \geq \gamma_3/(2\epsilon) - 1/2$ which implies $\gamma_3 - 2J\epsilon - \epsilon \leq 0$. Hence

$$\tilde{a}_m^J R^{(0)} \leq C_0^J \left(1 + \log(4/\delta)\right)^J M^{2J+2}.$$

Moreover,

$$b_m \lambda^{-1} \leq \left(C_1 \left(\log(4/\delta) + \log(m+1)\right)^{\gamma_1} + C_3\right) m^{-\gamma_5}.$$

Note $\gamma_5 \leq 0$, thus we have

$$R^{(J)} \leq \left\{ C_0^J \left(1 + \log(4/\delta)\right)^J M^{2J+2} + 2C_1 \left(\log(4/\delta) + \log(m+1)\right)^{\gamma_1} + 2C_3 \right\} m^{-\gamma_5}.$$

This prove our statement. □

We see that γ_5 given in the lemma is larger than $\epsilon - \gamma_3$ for $0 < q \leq 1$ which implies the bound (5.1) is better than before.

Now we are in the position to prove our main result.

Proof of Theorem 2. For $0 < \delta < 1$, let $\tilde{\delta} := \frac{1}{\beta+1}\delta \in (0,1)$ and let $m_{\delta,\epsilon} = m_{\tilde{\delta}}$ be as in lemma 3. Take

$$R = \left\{ C_0^\beta \left(1 + \log(4/\tilde{\delta})\right)^\beta M^{2\beta+2} + 2C_1 \left(\log(4/\tilde{\delta}) + \log(m+1)\right)^{\gamma_1} + 2C_3 \right\} m^{-\gamma_5}.$$

Let $m \geq m_{\delta,\epsilon}$, lemma 3 tell us that the measure of the set $\mathcal{W}(R)$ is at least $1-\beta\tilde{\delta}$. Applying proposition 7 to the above R, we know that, for each $\mathbf{z} \in \mathcal{W}(R)\backslash V_R$,

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \quad \leq \quad & C_0\left(1+\log(4/\tilde{\delta})\right)\left\{C_0^\beta\left(1+\log(4/\tilde{\delta})\right)^\beta M^{2\beta+2}\right. \\ & \left.+2C_1\left(\log(4/\tilde{\delta})+\log(m+1)\right)^{\gamma_1}+2C_3\right\}^2 m^{-(\frac{2}{2+p}+2\gamma_5)} \\ & +C_1\left(\log(4/\tilde{\delta})+\log(m+1)\right)^{\gamma_1} m^{-(\gamma_2-2(1-q)(\gamma_3-\epsilon))}+C_2 m^{-q(\gamma_3-\epsilon)}.\end{aligned}$$

Since the measure of V_R is at most $\tilde{\delta}$, we know that the above error bounds holds for $\mathbf{z} \in \mathcal{W}(R)\backslash V_R$ which has measure at least $1-\beta\tilde{\delta}-\tilde{\delta} = 1-\delta$. Thus we complete our proof with $C_{X,\rho,K} = C_0(M^2+2C_1+2C_3)^2+C_1+C_2$. □

Next we will give a proof of theorem 1.

Proof of Theorem 1. Take $s=2$ in proposition 2, we will get the estimation of $\mathcal{D}(\lambda)$ which implies condition (2.4) is satisfied with $q=1$. For any $0<\epsilon<1/2$, we take $\lambda = m^{-(1/2-\epsilon/2)}$. As the iteration process in lemma 4, we do it again with this λ. Since s can be choose arbitrarily large and $p \leq n/s$ form proposition 5. If $s \geq \tau$ and $s \geq n/\epsilon$, then $\gamma_1 = \frac{s}{\tau}$, $\gamma_2 = 1$ and

$$\begin{aligned}\tilde{a}_m &= C_0\left(1+\log\frac{4}{\delta}\right)M^2\left(\frac{1}{m}\right)^{\epsilon-\frac{p}{2+p}} \leq C_0\left(1+\log\frac{4}{\delta}\right)M^2\left(\frac{1}{m}\right)^{\epsilon/2}, \\ b_m &= C_1\left(\log(4/\delta)+\log(m+1)\right)^{\frac{s}{\tau}}\left(\frac{1}{m}\right)+C_3\left(\frac{1}{m}\right)^{1/2-\epsilon/2}.\end{aligned}$$

Thus we could take $\beta = \frac{1}{\epsilon}$, $\gamma_5 = 0$ and

$$\begin{aligned}m_{\delta,\epsilon} = \max\Big\{&\widetilde{C}_0\left(\log(2(1/\epsilon+1)/\delta)+\log(m+1)\right), \left\{2C_0\left(1+\log(4(1/\epsilon+1)/\delta)\right)M^2\right\}^{\frac{2}{\epsilon}}, \\ &4C_1\left(\log(4(1/\epsilon+1)/\delta)+\log(m+1)\right)^{\frac{s}{\tau}}/M^2, \left\{4C_3/M^2\right\}^{\frac{2}{1-\epsilon}}\Big\}\end{aligned}$$

If we take $s = \max\{n/\epsilon, \tau/\epsilon\}$, then from theorem 2, if $m > m_{\delta,\epsilon}$, with confidence $1-\delta$, there holds

$$\begin{aligned}&\|f_{\mathbf{z},\lambda}-f_\rho\|^2_{L^2_{\rho_X}} \leq C_{X,\rho,K}C_0^{2/\epsilon}M^{4/\epsilon} \\ &\left(1+\log(4(1/\epsilon+1)/\delta)\right)\left\{\log(m+1)+\log(4(1/\epsilon+1)/\delta)\right\}^{2\max\{1,n/\tau\}\frac{1}{\epsilon}} m^{-(1/2-\epsilon/2)}.\end{aligned}$$

We set $C_M = (\widetilde{C}_0+2C_0+4C_1+4C_3)M^2$, then $\widetilde{M}_{\delta,\epsilon} > M_{\delta,\epsilon}$ and when $m > \widetilde{M}_{\delta,\epsilon}$, we have

$$m^{\epsilon/2} > C_0^{2/\epsilon}M^{4/\epsilon}\left\{\log(m+1)+\log(4(1/\epsilon+1)/\delta)\right\}^{2\max\{1,n/\tau\}\frac{1}{\epsilon}},$$

Thus we complete our proof. □

6 Appendix

In this appendix, we will give the proof for proposition 5 (i). We mainly use the following theorem given in [16].

Theorem 3. *Let Q be a probability measure on a measurable space $(\mathcal{X}, \mathcal{A})$, and let $\mathcal{F}$ be a class of measurable functions with measurable square integrable envelope F such that $QF^2 = \int F^2 dQ < \infty$ and*

$$\mathcal{N}\left(\mathcal{F}, \epsilon \|F\|_{Q,2}, L_2(Q)\right) \leq C\left(\frac{1}{\epsilon}\right)^V, \quad 0 < \epsilon < 1.$$

Then there exists a constant K that depends on C and V only such that

$$\log \mathcal{N}\left(\overline{conv}\mathcal{F}, \epsilon \|F\|_{Q,2}, L_2(Q)\right) \leq K\left(\frac{1}{\epsilon}\right)^{2V/(V+2)}.$$

Here, $\mathcal{N}\left(\mathcal{F}, \epsilon, L_2(Q)\right)$ is the covering number relative to the $L_2(Q)$-norm

$$\|f\|_{Q,2} = \left(\int |f|^2 dQ\right)^{1/2}.$$

An envelope function of a class $\mathcal{F}$ is any function $x \to F(x)$ such that $|f(x)| \leq F(x)$ for every x and $f \in \mathcal{F}$. $conv\mathcal{F}$ is abbreviated to the convex hull of $\mathcal{F}$ which is defined as

$$conv\mathcal{F} = \left\{ \sum_{i=1}^{k} \alpha_i f_i \middle| f_i \in \mathcal{F}, \alpha_i \in \mathbb{R}, \alpha_i \geq 0, \sum_{i=1}^{k} \alpha_i = 1, k \in \mathbb{N} \right\}$$

and $\overline{conv}\mathcal{F}$ denotes its closure with respect to $L_2(Q)$.

Proof of Proposition 5 (i). We set $\mathcal{F}_1 = \{K_x | x \in X\}$ and $\mathcal{F}_2 = \{\mathcal{F}_1\} \cup \{-\mathcal{F}_1\} \cup \{0\}$. Note that B_1 is in L^∞-closure of the set $\mathcal{G} = \left\{ \sum_{i=1}^{\infty} \alpha_i K_{x_i} \middle| \{x_i\} \subset X, \sum_{i=1}^{\infty} |\alpha_i| \leq 1 \right\}$, recall $\|f\| = \inf\left\{ \sum_{j=1}^{\infty} |\alpha_j| : f = \sum_{j=1}^{\infty} \alpha_j K_{x_j} \right\}$, we just verify the claim when $\|f\| = 1$ which implies $\forall 0 < \epsilon < 1$, there exists a sequence $\{\alpha_i^\epsilon\} \in \ell^1$ and points $\{x_i^\epsilon\} \subset X$, such that

$$f = \sum_{i=1}^{\infty} \alpha_i^\epsilon K_{x_i^\epsilon} \quad and \quad 1 - \epsilon \leq \sum_{i=1}^{\infty} |\alpha_i^\epsilon| \leq 1 + \epsilon,$$

set $g = \frac{1}{1+\epsilon} f$ then $g \in \mathcal{G}$ and $\|f - g\|_\infty \leq 2\kappa\epsilon$. Hence $B_1 \subset \overline{conv}\mathcal{F}_2$ because of $\mathcal{G} \subset conv\mathcal{F}_2$. Since

$$\mathcal{N}\left(\mathcal{F}_2, \epsilon, L_2(Q)\right) \leq 2\mathcal{N}\left(\mathcal{F}_1, \epsilon, L_2(Q)\right) + 1$$

and we could choose $F \equiv \kappa$ as the envelope of both $\mathcal{F}_1$ and $\mathcal{F}_2$, thus we turn to estimate $\mathcal{N}(\mathcal{F}_1, \epsilon\kappa, L_2(Q))$. By condition (3.3), we have

$$\mathcal{N}(\mathcal{F}_1, \epsilon\kappa, L_2(Q)) \leq \mathcal{N}(\mathcal{F}_1, \epsilon\kappa, \|\cdot\|_\infty) \leq \mathcal{N}\left(X, (\epsilon\kappa/c_\alpha)^{1/\alpha}\right)$$

where $\mathcal{N}(X,\epsilon)$ denotes the covering number of X with respect to the Euclidean distance. Hence

$$\mathcal{N}(\mathcal{F}_1,\epsilon\kappa,L_2(Q)) \leq \left(\frac{c_\alpha}{\kappa}\right)^{\frac{n}{\alpha}} (3\mathrm{Diam}(X))^n \left(\frac{1}{\epsilon}\right)^{\frac{n}{\alpha}}.$$

Applying theorem 3, we get

$$\log \mathcal{N}\left(B_1,\epsilon,L_2(Q)\right) \leq \log \mathcal{N}\left(\overline{conv}\mathcal{F}_2,\epsilon,L_2(Q)\right) \leq \widetilde{C}\left(\frac{1}{\epsilon}\right)^{2n/(n+2\alpha)}$$

where $\widetilde{C}$ depend on n,α,c_α,κ and $\mathrm{Diam}(X)$. Finally, $\forall m \in \mathbb{N}$, for any samples $\mathbf{x} = \{x_i\}_{i=1}^m \subset X^m$, the above estimates hold true for $Q = \frac{1}{m}\sum_{i=1}^m \delta_{X_i}$. Thus we complete our proof. □

References

[1] N.Aronszajn. Theory of reproducing kernels, Trans. Amer. Math. Soc, volume 68: 337-404, 1950.

[2] T. Evgeniou, M. Pontil, and T. Poggio, Regularization networks and suport vector machine, Adv. comput. Math. **10** (1999), 51–80.

[3] E. De vito, A. caponnetto, and L, Rosasco. Model Selection for Regularized Least-Squares Algorithm in Learning Theory. Foundations of Computational Mathematics, 5(1):59-85, 2005.

[4] S. Smale and D.X.Zhou. Learning Theory Estimates via Intergral Operator and Their Approximation. Constructive Approximation, 26(2):153-172, 2007.

[5] Q.Wu, Y.Ying, and D.X.Zhou. Learning rates of least-square regularized regressioin. Foundations of Computational Mathematics, 6(2): 171-192, 2006.

[6] T.Zhang. Leave-one-out bounds for kernel methods. Neural Computation, 15(6):1397-1437, 2003.

[7] Q.Wu and D.X.Zhou, Learning with sample dependent hypothesis space, Computers and Mathematics with Applications 56(2008), 2896-2907.

[8] Q.W.Xiao and D.X.Zhou, Learning by nonsymmetric kernel with data dependent spaces and ℓ^1-regularizer. Taiwan.J.Math. to appear.

[9] H.Y.Wang, Q.W.Xiao and D.X.Zhou, Learning with ℓ^1-Regularization for Regression. Preprint.

[10] R.Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Stastistical Society. Series B, 58(1):267-288,1996.

[11] T.Zhang. some sharp performance bounds for least square regression with L_1 regularization. Annals of Stastics, 2009.

[12] P.Zhao and B.Yu. On model selection consistency of Lasso. Journal of Machine Learning Research, 7:2541-2563, 2006.

[13] E.Candès and J.Romberg. Sparsity and incoherence in compressive sampling. Inverse Problems, 23:969-985, 2007.

[14] H.Wenland. Local polynomial reproduction and moving least squares approximation. IMA Journal of Mumerical Analysis, 21(1):285-300,2001.

[15] K.Jetter, J.Stoeckler and J.D.Ward. Error estimates for scattered data interpolation on sphere. Advances in Computational Mathematics,13(1):1-50, 2000.

[16] A.W.Van der vaart and J.A.Wellner, Weak Convergence and Emprical Processes, Springer-Verlag, New York, 1996.

[17] D.X.Zhou, The covering number in learning thoery, J.Complexity 18 (2002), 739-767.

[18] D.X.Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory 49(2003), 1743-1752.

[19] Q.Wu, Y.Ying, and D.X.Zhou. Multi-kernel regularized classifiers, Jounal of Complexity 23(2007), 108-134.