Journal of Complexity 27 (2011) 55–67

Contents lists available at ScienceDirect

Journal of Complexity

journal homepage: www.elsevier.com/locate/jco

Optimal learning rates for least squares regularized regression with unbounded sampling[☆]

Cheng Wang[a], Ding-Xuan Zhou[b,*]

[a] College of Mathematical Sciences, Guangxi Normal University, Guilin, Guangxi 541004, PR China

[b] Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China

ARTICLE INFO

Article history:
Received 26 April 2010
Accepted 28 September 2010
Available online 14 October 2010

Keywords:
Learning theory
Least squares regression
Regularization in reproducing kernel Hilbert spaces
Covering number

ABSTRACT

A standard assumption in theoretical study of learning algorithms for regression is uniform boundedness of output sample values. This excludes the common case with Gaussian noise. In this paper we investigate the learning algorithm for regression generated by the least squares regularization scheme in reproducing kernel Hilbert spaces without the assumption of uniform boundedness for sampling. By imposing some incremental conditions on moments of the output variable, we derive learning rates in terms of regularity of the regression function and capacity of the hypothesis space. The novelty of our analysis is a new covering number argument for bounding the sample error.

1. Introduction

Learning algorithms produce approximations of functions from samples. Efficiency of algorithms relies on models relating the approximated functions on a metric space X and samples in $Y = \mathbb{R}$. Here we take a model with a Borel probability measure ρ on $Z := X \times Y$. We assume that a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \subset Z^m$ is drawn independently from ρ and the approximated function is the *regression function* of ρ defined by

$$f_\rho(x) = \int_Y y \mathrm{d}\rho(y|x), \quad x \in X. \tag{1.1}$$

[☆] The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 103508] and the National Science Fund for Distinguished Young Scholars of China [Project No. 10529101].

[*] Corresponding author.
E-mail addresses: wang_cheng20@yahoo.com.cn (C. Wang), mazhou@cityu.edu.hk (D.-X. Zhou).

0885-064X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jco.2010.10.002

Here $\rho(\cdot|x)$ is the conditional distribution of ρ at $x \in X$. One measurement for the efficiency of a learning algorithm is the distance between the approximant produced by the algorithm and f_ρ in the space $L^2_{\rho_X}$ with norm $\|f\|_{L^2_{\rho_X}} = (\int_X |f(x)|^2 \mathrm{d}\rho_X)^{1/2}$ where ρ_X is the marginal distribution of ρ on X.

In this paper we consider a *learning algorithm* generated by a least squares regularization scheme in a *reproducing kernel Hilbert space* (RKHS). Let $K : X \times X \to \mathbb{R}$ be a bounded, symmetric, and positive semi-definite function. The RKHS $\mathcal{H}_K$ associated with the kernel K is the completion of the linear span of functions $\{K_x := K(x, \cdot), x \in X\}$ with the inner product given by $\langle K_x, K_y\rangle_{\mathcal{H}_K} = \langle K_x, K_y\rangle_K = K(x, y)$. Then the learning algorithm for the regression problem is given by the regularization scheme

$$f_{\mathbf{z}} = f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2 + \lambda\|f\|_K^2 \right\}, \tag{1.2}$$

where $\lambda > 0$ is a regularization parameter which may depend on the sample size $\lambda = \lambda(m)$ with $\lim_{m\to\infty} \lambda(m) = 0$.

There has been a large learning theory literature on error analysis for learning algorithm (1.2); see e.g. [4,17,13,6,2,5,11]. Most obtained error bounds are presented under the standard assumption that $|y| \le M$ almost surely for some constant $M > 0$, i.e., $\rho(\cdot|x)$ is supported on $[-M, M]$ for almost every $x \in X$. This standard assumption is abandoned in [2]. There the authors consider a general setting satisfying the condition

$$\int_Y \left(\exp\left\{ -\frac{|y - f_{\mathcal{H}}(x)|^2}{M} \right\} - \frac{|y - f_{\mathcal{H}}(x)|}{M} - 1 \right) \mathrm{d}\rho(y|x) \le \frac{\Sigma^2}{2M^2} \tag{1.3}$$

for ρ_X-almost every $x \in X$ and some constants $M, \Sigma > 0$, where $f_{\mathcal{H}}$ is the orthogonal projection of f_ρ onto the closure of $\mathcal{H}_K$ in $L^2_{\rho_X}$. Bounds for the error $\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}}$ are established under the assumption that $f_{\mathcal{H}}$ actually lies in $\mathcal{H}_K$. How to relax the assumption $f_{\mathcal{H}} \in \mathcal{H}_K$ in the error analysis with $|y| \le M$ is investigated in [11,5].

The main purpose of this paper is to conduct error analysis in another general setting satisfying the following moment hypothesis concerning unbounded outputs.

Moment hypothesis: There exist constants $M > 0$ and $C > 0$ such that

$$\int_Y |y|^\ell \mathrm{d}\rho(y|x) \le C\ell! M^\ell \quad \forall \ell \in \mathbb{N},\ x \in X. \tag{1.4}$$

Remark 1. The moment hypothesis is a natural generalization of condition (1.3) to cases without the restriction $f_{\mathcal{H}} \in \mathcal{H}_K$ or $f_{\mathcal{H}} \in L^\infty(X)$. In fact, they are equivalent (with different constants) in the case $f_{\mathcal{H}} \in L^\infty(X)$. To see this, we notice from the Taylor expansion that the left-hand side of (1.3) equals

$$\sum_{\ell=2}^{\infty} \frac{\int_Y |y - f_{\mathcal{H}}(x)|^\ell \mathrm{d}\rho(y|x)}{\ell! M^\ell}.$$

Then hypothesis (1.4) implies condition (1.3) with M replaced by $3M + 2\|f_{\mathcal{H}}\|_\infty$. Conversely, when (1.3) is valid, we know that for $2 \le \ell \in \mathbb{N}$, $\int_Y |y - f_{\mathcal{H}}(x)|^\ell \mathrm{d}\rho(y|x) \le \Sigma^2 \ell! M^{\ell-2}/2$. Hence (1.4) holds true with $C = \frac{\Sigma^2}{2M^2} + 1$ and M replaced by $\max\{2\|f_{\mathcal{H}}\|_\infty, 2M\}$.

A simple computation verifies (1.4) for Gaussian noise.

Example 1. Let $B > 0$ and $B_0 > 0$. If for each $x \in X$, $|f_\rho(x)| \le B$ and the condition distribution $\rho(\cdot|x)$ is a normal distribution with variance σ_x^2 bounded by B_0, then (1.4) is satisfied with $M = \max\{\sqrt{2}B_0, B\}$ and $C = 4$.

To show some ideas of our error analysis, we first state learning rates of (1.2) in the special case when $f_\rho \in \mathcal{H}_K$ and K is C^∞ on $X \subset \mathbb{R}^n$.

Theorem 1. *Under the moment hypothesis and $f_\rho \in \mathcal{H}_K$, if $X \subset \mathbb{R}^n$ and K is C^∞ on $X \times X$, then for any $0 < \epsilon < 1$ and $0 < \delta < 1$, by taking $\lambda = m^{\epsilon-1}$, with confidence $1-\delta$, we have*

$$\|f_{\mathbf{z}} - f_\rho\|^2_{L^2_{\rho_X}} \leq \tilde{C}_\epsilon m^{\epsilon-1} \left(\log \frac{4}{\delta}\right)^{\frac{4}{\epsilon}+2},$$

where $\tilde{C}_\epsilon$ is a constant independent of m or δ.

Theorem 1 is a corollary of our main result presented in the next section.

2. Main result

Our main result is about learning rates of (1.2) stated under conditions on the approximation ability of $\mathcal{H}_K$ with respect to f_ρ and capacity of $\mathcal{H}_K$.

The approximation ability of the hypothesis space $\mathcal{H}_K$ with respect to f_ρ in the space $L^2_{\rho_X}$ is reflected by approximation error.

Definition 1. The approximation error of the triple $(\mathcal{H}_K, f_\rho, \rho_X)$ is defined as

$$\mathcal{D}(\lambda) = \min_{f \in \mathcal{H}_K} \{\|f - f_\rho\|^2_{L^2_{\rho_X}} + \lambda \|f\|^2_K\}, \quad \lambda > 0. \tag{2.1}$$

We shall assume that for some $0 < \beta \leq 1$ and $C_\beta > 0$,

$$\mathcal{D}(\lambda) \leq C_\beta \lambda^\beta \quad \forall \lambda > 0. \tag{2.2}$$

Remark 2. Our analysis applies when f_ρ is replaced by $f_{\mathcal{H}}$ which would imply $\mathcal{D}(\lambda) \to 0$ as $\lambda \to 0$. So (2.2) is a natural assumption. Note [9] that $\mathcal{D}(\lambda) = o(\lambda)$ would imply $f_\rho \equiv 0$. So $\beta = 1$ in (2.2) is the best we can expect. This case is equivalent to $f_\rho \in \mathcal{H}_K$ when $\mathcal{H}_K$ is dense in $L^2_{\rho_X}$. See [9]. Assumption (2.2) with $0 < \beta < 1$ can be characterized in terms of interpolation spaces [7].

For a general kernel on a general metric space X, we need the capacity of $\mathcal{H}_K$ to quantitatively understand influence of the complexity of the hypothesis space to learning ability of algorithm (1.2). Here we use covering numbers to measure the capacity.

Definition 2. For a subset $\mathcal{F}$ of a metric space and $\eta > 0$, the covering number $\mathcal{N}(\mathcal{F}, \eta)$ is defined to be the minimal integer ℓ such that there exist ℓ disks with radius η covering $\mathcal{F}$.

We shall use this notion for balls $B_R = \{f \in \mathcal{H}_K : \|f\|_K \leq R\}$ as subsets of $L^\infty(X)$.

Definition 3. We say $\mathcal{H}_K$ has polynomial complexity exponent $s > 0$ if for some constant $C_0 > 0$,

$$\log \mathcal{N}(B_1, \eta) \leq C_0 \left(\frac{1}{\eta}\right)^s, \quad \forall \eta > 0. \tag{2.3}$$

Remark 3. When X is a bounded domain in $\mathbb{R}^n$ and $K \in C^\tau(X \times X)$, it is known [18] that (2.3) holds true with $s = \frac{2n}{\tau}$. In particular, if $K \in C^\infty(X \times X)$, condition (2.3) is valid for an arbitrarily small $s > 0$. It would be interesting to extend this covering number bound to unbounded input spaces X. Some ideas from [19] might help.

Now we can state our general result on learning rates for algorithm (1.2).

Theorem 2. *Assume the moment hypothesis (1.4) and condition (2.2) with $0 < \beta \leq 1$. If $\mathcal{H}_K$ has polynomial complexity exponent $s > 0$ and $0 < \epsilon < \frac{\beta}{s+1}$, then by taking $\lambda = m^{\frac{\epsilon}{\beta} - \frac{1}{s+1}}$, for any $0 < \delta < 1$,* *with confidence* $1-\delta$, *we have*

$$\|f_{\mathbf{z}}-f_\rho\|^2_{L^2_{\rho_X}} \leq \tilde{C}_\epsilon m^{\epsilon-\frac{\beta}{s+1}} \left(\log\frac{4}{\delta}\right)^{\frac{\beta(1+\beta)}{(s+1)\epsilon}+2},$$

where $\tilde{C}_\epsilon$ *is a constant depending on* ϵ *but not on* m *or* δ*.*

Theorem 2 establishes learning rates for unbounded sampling processes satisfying the moment hypothesis, which generalizes results in the classical case of uniformly bounded outputs (e.g. [13]). In addition, we do not require the sample size m to be sufficiently large, a restriction imposed as $m \geq m_{\delta,\epsilon}$ in [13].

Theorem 2 provides a confidence-based estimate for the least squares error of the learning algorithm. The dependence of the estimate on the confidence (variance) is in the form of $\log(4/\delta)$ which is mild.

Theorem 2 will be proved in Section 5 and the constant $\tilde{C}_\epsilon$ will be given explicitly. The proof is mainly based on our novel approach to handle unbounded sampling with a new covering number argument which will be presented in Section 4. Note that when $\beta = 1$ and s is small enough, the learning rate stated in Theorem 2 can be arbitrarily close to 1, hence is optimal [2,11,5,3].

To illustrate Theorem 2, we consider the example of $\mathcal{H}_K$ being a Sobolev space $H^\tau(X)$ which consists of functions on $X \subset \mathbb{R}^n$ with all derivatives of order up to $\frac{n}{2} < \tau \in \mathbb{N}$ lying in $L^2(X)$. Take $\mathcal{H}_K = H^\tau(X)$. Then (2.3) holds true with $s = \frac{2n}{\tau}$. If ρ_X is the uniform measure and $f_\rho \in H^r(X)$ for some $0 < r < \tau$, then we know [7] that condition (2.2) is valid with $\beta = \frac{r}{\tau}$. So the conclusion in the following example is a corollary of Theorem 2.

Example 2. Let X be a bounded domain in $\mathbb{R}^n$ and ρ_X be the uniform measure. Assume the moment hypothesis (1.4). If $\mathcal{H}_K = H^\tau(X)$ with $\frac{n}{2} < \tau \in \mathbb{N}, f_\rho \in H^r(X)$ for some $0 < r < \tau$, and $0 < \epsilon < \frac{r}{2n+\tau}$, then by taking $\lambda = m^{\frac{\tau\epsilon}{r}-\frac{\tau}{2n+\tau}}$, for any $0 < \delta < 1$, with confidence $1-\delta$, we have

$$\|f_{\mathbf{z}}-f_\rho\|^2_{L^2_{\rho_X}} \leq \tilde{C}_\epsilon m^{\epsilon-\frac{r}{2n+\tau}} \left(\log\frac{4}{\delta}\right)^{\frac{r(r+\tau)}{\tau(2n+\tau)\epsilon}+2}.$$

Now let us describe two kinds of approaches for error analysis of algorithm (1.2) and compare our learning rates with those in the literature.

The first family of approaches aims at bounding

$$\sup_{f\in\mathcal{F}_\lambda}\left|\int_Z (f(x)-y)^2 d\rho - \frac{1}{m}\sum_{i=1}^m (f(x_i)-y_i)^2\right| \tag{2.4}$$

with a properly chosen function class $\mathcal{F}_\lambda$ and then applying some uniform law of large numbers. Such an approach leads to capacity dependent error bounds for various learning algorithms stated in terms of various quantities measuring capacity of $\mathcal{H}_K$ such as VC-dimension, V_γ-dimension, covering number, and empirical covering number (e.g. [5,10,14]).

A typical optimal learning rate stated in terms of covering numbers can be found in [13]. It asserts under the conditions of Theorem 1 that for $0 < \epsilon < 1$ and $m \geq m_{\delta,\epsilon}$, with confidence $1-\delta$, we have $\|f_{\mathbf{z}}-f_\rho\|^2_{L^2_{\rho_X}} \leq \tilde{C}(\log\frac{2}{\delta})m^{\epsilon-1}$. A minor improvement of our Theorem 1 is to determine the restriction $m \geq m_{\delta,\epsilon}$ specifically in addition to our main contribution of removing the uniform boundedness assumption of $|y| \leq M$.

Another typical optimal learning rate is stated in terms of conditions on eigenvalues $\{\lambda_i\}$ of the integral operator $L_{K,\rho_X} : L^2_{\rho_X} \to L^2_{\rho_X}$ defined by

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

When the eigenvalues satisfy $a_1 i^{-b} \le \lambda_i \le a_2 i^{-b}$ for all $i \in \mathbb{N}$, some $1 < b < \infty$ and $0 < a_1, a_2 < \infty$, under the assumption $f_{\mathcal{H}} \in \mathcal{H}_K$, it was proved in [2] that with confidence $1-\delta$, $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \le \tilde{C}(\log\frac{6}{\delta})^2 m^{-\frac{b}{b+1}}$, where $\mathcal{E}(f)$ is the generalization error defined for $f : X \to \mathbb{R}$ as

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

In the special case of $f_\rho \in \mathcal{H}_K$ and $b \ge \frac{1}{\epsilon} - 1$, the optimal learning rate $O(m^{\epsilon-1})$ is achieved though the eigenvalue condition is difficult to check. Optimal learning rates are also discussed in [5] for algorithm (1.2) with the penalty $\|f\|_K^q$ for some $0 < q < 1$ where the condition $f_{\mathcal{H}} \in \mathcal{H}_K$ is replaced by the uniform boundedness of the eigenvectors of L_{K,ρ_X} or more generally by the norm comparison assumption

$$\|f\|_\infty \le C \|f\|_K^{s'} \|f\|_{L^2_{\rho_X}}^{1-s'} \quad \forall f \in \mathcal{H}_K \tag{2.5}$$

with some constants $C > 0$ and $0 \le s' \le 1$. In [11] the lower bound condition for the eigenvalues in [2] is removed and the restriction $f_{\mathcal{H}} \in \mathcal{H}_K$ is replaced by approximation error condition (2.2) with $0 < \beta \le 1$. When $(\mathcal{H}_K, \rho_X)$ satisfies (2.5) with $s' = \frac{1}{b}$, it was shown in [11] that when $|y| \le M$, with confidence $1-\delta$, $\|\pi_M(f_{\mathbf{z}}) - f_\rho\|^2_{L^2_{\rho_X}} \le \tilde{C}(\log\frac{3}{\delta}) m^{-\frac{b\beta}{b\beta+1}}$, where $\pi_M(f)$ is the projection operator [3,13] defined by

$$\pi_M(f)(x) = \begin{cases} M, & \text{if } f(x) > M, \\ f(x), & \text{if } -M \le f(x) \le M, \\ -M, & \text{if } f(x) < -M. \end{cases}$$

This general result yields optimal learning rate in the special case $\beta = 1$. The upper bound condition for the eigenvalues and the norm comparison assumption (2.5) used in [2,11,5] can be easily verified in some common situations [9], and they have the advantage of applying to bounded input spaces X. It would be interesting to combine advantages of methods from [2,11,5] and our approach. In particular, the following two questions would lead to further study on error analysis:

1. Is it possible to have some criteria for checking the eigenvalue condition and (2.5) for general marginal distributions ρ_X which could be used to prove Theorem 1?
2. Can we extend the covering number approach to unbounded input spaces which can be used to recover results in [11]?

The second family of approaches for error analysis of the least squares algorithm (1.2) is to make full use of the linear nature of the algorithm for bounding the error between $f_{\mathbf{z}}$ and f_λ. In [17], a leave-one-out technique was used to obtain

$$\mathbb{E}(\|f_{\mathbf{z}} - f_\rho\|^2_{L^2_{\rho_X}}) \le \mathcal{D}\left(\frac{\lambda}{2}\right) + \left(\mathcal{E}(f_\rho) + \mathcal{D}\left(\frac{\lambda}{2}\right)\right)\left\{\frac{4\kappa^2}{m\lambda} + \left(\frac{2\kappa^2}{m\lambda}\right)^2\right\},$$

where $\kappa = \sup_{x\in X}\sqrt{K(x,x)}$. When $f_\rho \in \mathcal{H}_K$ and $\mathcal{E}(f_\rho) > 0$, the learning rate would be $\|f_{\mathbf{z}} - f_\rho\|^2_{L^2_{\rho_X}} \le \frac{C}{\delta}(\frac{1}{m})^{\frac{1}{2}}$ with confidence $1-\delta$, corresponding to the choice $\lambda = \frac{1}{\sqrt{m}}$.

In [4], a functional analysis approach was employed to show that $\|f_{\mathbf{z}} - f_\rho\|^2_{L^2_{\rho_X}} \le C(\frac{1}{m}\log\frac{2}{\delta})^{\frac{2}{5}}$ with confidence $1-\delta$ when $f_\rho = \int_X K(\cdot,u)g(u)\mathrm{d}\rho_X(u)$ for some $g \in L^2_{\rho_X}$. An integral operator approach was applied in [6] to prove that under the same condition, with confidence $1-\delta$, there holds $\|f_{\mathbf{z}} - f_\rho\|^2_{L^2_{\rho_X}} \le \widetilde{C}(\log\frac{4}{\delta})^2 m^{-\frac{2}{3}}$.

3. Error decomposition

The error analysis of algorithm (1.2) will be conducted by an error decomposition procedure. The idea of error decomposition has been used in the analysis of regularization schemes [15,3,13,14,8,16]. But the previous approaches for bounding (2.4) cannot be applied here because of the unboundedness of the sampling outputs. We shall adjust the error decomposition technique by bounding the outputs with confidence and then applying a novel covering number argument for a finite set of functions (an η-net) instead of the ball B_R (an infinite set of functions). The detailed procedure is described in Section 4.

Observe that the regression function f_ρ is a minimizer of the generalization error $\mathcal{E}(f)$ and actually we have

$$\|f - f_\rho\|^2_{L^2_{\rho_X}} = \mathcal{E}(f) - \mathcal{E}(f_\rho). \tag{3.1}$$

If we define the empirical error $\mathcal{E}_{\mathbf{z}}(f)$ as

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

then the following error decomposition follows from the relation $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\|f_{\mathbf{z}}\|_K^2 \leq \mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\|_K^2$, as shown in [13].

Lemma 1. *Let f_λ be a minimizer of (2.1). Then*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \lambda\|f_{\mathbf{z}}\|_K^2 \leq \mathcal{S}_1(\mathbf{z}) + \mathcal{S}_2(\mathbf{z}) + \mathcal{D}(\lambda), \tag{3.2}$$

where

$$\mathcal{S}_1(\mathbf{z}) = (\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)) - (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)),$$
$$\mathcal{S}_2(\mathbf{z}) = (\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_{\mathbf{z}}(f_\rho)) - (\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho)).$$

With decomposition (3.2), the error $\|f_{\mathbf{z}} - f_\rho\|^2_{L^2_{\rho_X}} = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)$ can be bounded by estimating the two quantities $\mathcal{S}_1(\mathbf{z})$ and $\mathcal{S}_2(\mathbf{z})$. While the second quantity can be easily analyzed by applying probability inequalities to the random variable $(f_\lambda(x) - y)^2 - (f_\rho(x) - y)^2$ on the space (Z, ρ), the first quantity $\mathcal{S}_1(\mathbf{z})$ is the main task of error analysis for algorithm (1.2): though $\mathcal{S}_1(\mathbf{z})$ can be expressed as $\int_Z \xi_1(z)d\rho - \frac{1}{m}\sum_{i=1}^m \xi_1(z_i)$ with $\xi_1(z) = (f_{\mathbf{z}}(x) - y)^2 - (f_\rho(x) - y)^2$, the major challenge is that ξ_1 is not a single random variable and it depends on the sample $\mathbf{z}$ itself. Our approach to tackle $\mathcal{S}_1(\mathbf{z})$ is a novel covering number argument presented in Section 4.

Our error analysis relies on the following probability inequality for random variables without uniform boundedness [1].

Lemma 2. *Let $X_1, X_2, \ldots, X_m$ be independent random variables with $\mathbb{E}X_i = 0$. If for some constants $M, v > 0$, the bound $\mathbb{E}|X_i|^\ell \leq \frac{1}{2}\ell! M^{\ell-2} v$ holds for every $2 \leq \ell \in \mathbb{N}$, then*

$$\text{Prob}\left\{\sum_{i=1}^m X_i \geq \varepsilon\right\} \leq \exp\left\{-\frac{\varepsilon^2}{2}(mv + M\varepsilon)^{-1}\right\} \quad \forall \varepsilon > 0.$$

In our setting we apply Lemma 2 to random variables $X_i = \mathbb{E}g - g(z_i)$ for a function g on Z where $z_i = (x_i, y_i)$ and $\mathbb{E}g = \int_Z g(z)d\rho$.

Lemma 3. *Denote $\mathbb{E}_{\mathbf{z}}g = \frac{1}{m}\sum_{i=1}^m g(z_i)$ for a measurable function g on Z. If for some $M, v > 0$, the bound $\mathbb{E}|g - \mathbb{E}g|^\ell \leq \frac{1}{2}\ell! M^{\ell-2} v$ holds for $2 \leq \ell \in \mathbb{N}$, then there holds*

$$\text{Prob}_{\mathbf{z}\in Z^m}\{\mathbb{E}g - \mathbb{E}_{\mathbf{z}}g \geq \varepsilon\} \leq \exp\left\{-\frac{m\varepsilon^2}{2(v + M\varepsilon)}\right\} \quad \forall \varepsilon > 0.$$

In particular, the second quantity $\mathcal{S}_2(\mathbf{z})$ in (3.2) can be bounded easily.

Lemma 4. *Under the moment hypothesis, with confidence at least* $1-\frac{\delta}{2}$, *we have*

$$\mathcal{S}_2(\mathbf{z}) \leq \frac{240\{(\kappa+1)^2\mathcal{D}(\lambda)/\lambda + 2(C+1)^2M^2\}}{m}\log\frac{2}{\delta} + 18\mathcal{D}(\lambda).$$

Proof. Consider the function g on the space (Z,ρ) given with $z=(x,y)$ by $g(z)=(f_\rho(x)-y)^2-(f_\lambda(x)-y)^2$. It satisfies for $2\leq\ell\in\mathbb{N}$

$$\begin{aligned}|g(z)|^\ell &= |f_\rho(x)-f_\lambda(x)|^\ell|f_\rho(x)+f_\lambda(x)-2y|^\ell\\ &\leq 3^\ell(|f_\lambda(x)|^\ell+|f_\rho(x)|^\ell+2^\ell|y|^\ell)2^{\ell-2}(|f_\lambda(x)|^{\ell-2}+|f_\rho(x)|^{\ell-2})|f_\lambda(x)-f_\rho(x)|^2.\end{aligned}$$

The moment hypothesis yields for almost every $x\in X$ that

$$\int_Y |y|^\ell \mathrm{d}\rho(y|x) \leq C\ell!M^\ell$$

and

$$|f_\rho(x)| = \left|\int_Y y\mathrm{d}\rho(y|x)\right| \leq CM.$$

The reproducing property of $\mathcal{H}_K$ means

$$\langle f, K_x\rangle_K = f(x), \quad \forall x\in X,\ f\in\mathcal{H}_K. \tag{3.3}$$

It follows that

$$|f(x)| \leq \kappa\|f\|_K \quad \forall f\in\mathcal{H}_K,\ x\in X. \tag{3.4}$$

Note that $\mathcal{D}(\lambda)=\mathcal{E}(f_\lambda)-\mathcal{E}(f_\rho)+\lambda\|f_\lambda\|_K^2$ implies $\|f_\lambda\|_K\leq\sqrt{\mathcal{D}(\lambda)/\lambda}$. So $|f_\lambda(x)|\leq\kappa\sqrt{\mathcal{D}(\lambda)/\lambda}$ for each $x\in X$. Hence

$$\begin{aligned}\mathbb{E}|g|^\ell &= \int_X\int_Y |g(z)|^\ell \mathrm{d}\rho(y|x)\mathrm{d}\rho_X(x) \leq 3^\ell\{\kappa^\ell(\mathcal{D}(\lambda)/\lambda)^{\ell/2}+C^\ell M^\ell+2^\ell C\ell!M^\ell\}\\ &\quad\times 2^{\ell-2}\{\kappa^{\ell-2}(\mathcal{D}(\lambda)/\lambda)^{(\ell-2)/2}+C^{\ell-2}M^{\ell-2}\}\int_X |f_\lambda(x)-f_\rho(x)|^2\mathrm{d}\rho_X(x)\\ &\leq \ell!6^\ell\{(\kappa+1)^2\mathcal{D}(\lambda)/\lambda+2(C+1)^2M^2\}^{\ell-1}\mathcal{D}(\lambda).\end{aligned}$$

It follows that

$$\mathbb{E}|g-\mathbb{E}g|^\ell \leq 2^{\ell+1}\mathbb{E}|g|^\ell \leq 2^{\ell+1}\ell!6^\ell\{(\kappa+1)^2\mathcal{D}(\lambda)/\lambda+2(C+1)^2M^2\}^{\ell-1}\mathcal{D}(\lambda).$$

Denote $M_{1,\lambda}=12\{(\kappa+1)^2\mathcal{D}(\lambda)/\lambda+2(C+1)^2M^2\}$ and $v_{1,\lambda}=24^2M_{1,\lambda}\mathcal{D}(\lambda)$. We find that

$$\mathbb{E}|g-\mathbb{E}g|^\ell \leq \frac{1}{2}\ell!M_{1,\lambda}^{\ell-2}v_{1,\lambda}.$$

Then we apply Lemma 3 and see that for any $\varepsilon>0$,

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\{\mathbb{E}g-\mathbb{E}_{\mathbf{z}}g\geq\varepsilon\} \leq \exp\left\{-\frac{m\varepsilon^2}{2(v_{1,\lambda}+M_{1,\lambda}\varepsilon)}\right\}.$$

Consider the quadratic equation by setting the probability on the right-hand side to be $\frac{\delta}{2}$. The positive solution is

$$\begin{aligned}\varepsilon &= \frac{1}{m}\left\{M_{1,\lambda}\log\frac{2}{\delta}+\sqrt{M_{1,\lambda}^2\log^2\frac{2}{\delta}+2mv_{1,\lambda}\log\frac{2}{\delta}}\right\}\\ &\leq \frac{1}{m}\left\{2M_{1,\lambda}\log\frac{2}{\delta}+24\sqrt{2mM_{1,\lambda}\log\frac{2}{\delta}\mathcal{D}(\lambda)}\right\} \leq \frac{20M_{1,\lambda}}{m}\log\frac{2}{\delta}+18\mathcal{D}(\lambda).\end{aligned}$$

Then with confidence at least $1-\frac{\delta}{2}$, we have

$$\mathbb{E}g-\mathbb{E}_{\mathbf{z}}g \le \frac{20M_{1,\lambda}}{m}\log\frac{2}{\delta}+18\mathcal{D}(\lambda).$$

So our desired bound follows from the identity $\mathbb{E}g-\mathbb{E}_{\mathbf{z}}g=\mathcal{S}_2(\mathbf{z})$ since $\mathbb{E}g=-(\mathcal{E}(f_\lambda)-\mathcal{E}(f_\rho))$ and $\mathbb{E}_{\mathbf{z}}g=-(\mathcal{E}_{\mathbf{z}}(f_\lambda)-\mathcal{E}_{\mathbf{z}}(f_\rho))$. □

4. Novelty dealing with unboundedness

In this section we present our novelty in the error analysis to deal with the error term $\mathcal{S}_1(\mathbf{z})$ in (3.2) for algorithm (1.2) before proving Theorem 2 in the next section. Though $\mathcal{S}_1(\mathbf{z})$ can be written as $\int \xi_1(z)\mathrm{d}\rho-\frac{1}{m}\sum_{i=1}^m \xi_1(z_i)$ with $\xi_1(z)=(f_{\mathbf{z}}(x)-y)^2-(f_\rho(x)-y)^2$, the function ξ_1 is not really a random variable: the function $f_{\mathbf{z}}$ also depends on the sample $\mathbf{z}$. We would follow the covering number approach in [13] to handle this term. However, the lack of uniform boundedness for sample function values causes serious difficulty and the approach in [13] of estimating quantity (2.4) is not applicable directly: to estimate (2.4), we would have to bound $|\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_j)|$, which would lead to bounding $\frac{1}{m}\sum_{i=1}^m |y_i|$ for every $f\in\mathcal{F}_\lambda$. We shall deal with the difficulty in three steps. Our key point is to use a new novel covering number argument.

4.1. Bounding sample values with confidence

The first step in our approach is to bound y with confidence.

Proposition 1. *Under the moment hypothesis there is a subset Z_δ of Z^m with measure at least $1-\frac{\delta}{4}$ such that*

$$\frac{1}{m}\sum_{i=1}^m |y_i| \le M_\delta := CM+4M(1+\sqrt{2C})\frac{\log(4/\delta)}{\sqrt{m}} \quad \forall \mathbf{z}\in Z_\delta.$$

Proof. Let g be the function on Z given by $g(z)=-|y|$. Then for $2\le \ell\in\mathbb{N}$, we have

$$\mathbb{E}|g-\mathbb{E}g|^\ell \le 2^{\ell+1}\mathbb{E}|y|^\ell \le 2^{\ell+1}C\ell!M^\ell \le \frac{1}{2}\ell!(2M)^{\ell-2}C(4M)^2.$$

So we can apply Lemma 3 to obtain for $\varepsilon>0$,

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\{\mathbb{E}g-\mathbb{E}_{\mathbf{z}}g\ge\varepsilon\}\le \exp\left\{-\frac{m\varepsilon^2}{2(C(4M)^2+2M\varepsilon)}\right\}.$$

Setting the right-hand side to be $\frac{\delta}{4}$ and bounding the solution to the corresponding quadratic equation, we know that with confidence at least $1-\frac{\delta}{4}$, there holds

$$\mathbb{E}g-\mathbb{E}_{\mathbf{z}}g\le\frac{4M}{m}\log\frac{4}{\delta}\{1+\sqrt{2Cm}\}.$$

Note that $-\mathbb{E}_{\mathbf{z}}g=\frac{1}{m}\sum_{i=1}^m|y_i|$ and $-\mathbb{E}g=\mathbb{E}|y|\le CM$. Then with confidence $1-\frac{\delta}{4}$, we have

$$\frac{1}{m}\sum_{i=1}^m |y_i| \le CM+4M(1+\sqrt{2C})\frac{\log(4/\delta)}{\sqrt{m}}=M_\delta.$$

This means that there is a subset Z_δ of Z^m with measure at least $1-\frac{\delta}{4}$, such that for every $\mathbf{z}\in Z_\delta$, we have $\frac{1}{m}\sum_{i=1}^m|y_i|\le M_\delta$. This proves Proposition 1. □

4.2. Bounding error difference for a finite function set

The second step in our approach is to bound the error difference $[\mathcal{E}(f) - \mathcal{E}(f_\rho)] - [\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_\rho)]$ for a finite set of functions. For this purpose, we need the following lemma which is a corollary of Lemma 3 by taking ε as $\sqrt{\varepsilon}\sqrt{\varepsilon + |\mathbb{E}g|}$.

Lemma 5. *Let g be a measurable function on Z. If for some $M, c_v > 0$, the bound $\mathbb{E}|g - \mathbb{E}g|^\ell \le \frac{1}{2}\ell! M^{\ell-2} c_v |\mathbb{E}g|$ holds for $2 \le \ell \in \mathbb{N}$, then*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\{\mathbb{E}g - \mathbb{E}_{\mathbf{z}}g \ge \sqrt{\varepsilon}\sqrt{\varepsilon + |\mathbb{E}g|}\} \le \exp\left\{-\frac{m\varepsilon}{2(c_v + M)}\right\} \quad \forall \varepsilon > 0.$$

In the following lemma, $\{f_j\}_{j=1}^{\mathcal{N}}$ is a fixed set of functions which will be chosen as an η-net of the set B_R in our error analysis conducted in Lemma 7.

Lemma 6. *Let $0 < \delta < 1$, $R \ge M$ and $\{f_j\}_{j=1}^{\mathcal{N}}$ be a set of functions in B_R. Then there exists a subset Z'_δ of Z^m with measure at least $1 - \frac{\delta}{4}$ such that*

$$\mathcal{E}(f_j) - \mathcal{E}(f_\rho) - [\mathcal{E}_{\mathbf{z}}(f_j) - \mathcal{E}_{\mathbf{z}}(f_\rho)] \le \varepsilon_{m,\delta,\mathcal{N},R} + \frac{1}{2}\{\mathcal{E}(f_j) - \mathcal{E}(f_\rho)\} \quad \forall j \in \{1, \ldots, \mathcal{N}\},\ \mathbf{z} \in Z'_\delta,$$

where

$$\varepsilon_{m,\delta,\mathcal{N},R} = 520(\kappa + C + 2(C+1))^2 \frac{R^2}{m} \log(4\mathcal{N}/\delta).$$

Proof. Fix $j \in \{1, \ldots, \mathcal{N}\}$. Let g be the function on Z defined by $g(z) = (f_j(x) - y)^2 - (f_\rho(x) - y)^2$. By the moment hypothesis, for $2 \le \ell \in \mathbb{N}$, we have

$$\begin{aligned}\mathbb{E}|g - \mathbb{E}g|^\ell &\le 2^{\ell+1}\mathbb{E}|g|^\ell = 2^{\ell+1}\mathbb{E}\{|f_j(x) - f_\rho(x)|^\ell \cdot |f_j(x) + f_\rho(x) - 2y|^\ell\} \\ &\le 2^{2\ell+1}\mathbb{E}\{|f_j(x) - f_\rho(x)|^2(\kappa R + CM)^{\ell-2}((\kappa R + CM)^\ell + 2^\ell |y|^\ell)\} \\ &\le 2^{2\ell+1}(\kappa R + CM)^{\ell-2}((\kappa R + CM)^\ell + 2^\ell C\ell! M^\ell)\int_X |f_j(x) - f_\rho(x)|^2 d\rho_X.\end{aligned}$$

But $\int_X |f_j(x) - f_\rho(x)|^2 d\rho_X = \mathcal{E}(f_j) - \mathcal{E}(f_\rho) = \mathbb{E}g = |\mathbb{E}g|$. So we know that

$$\mathbb{E}|g - \mathbb{E}g|^\ell \le \frac{1}{2}\ell! M_{3,R}^{\ell-2} v_{3,R}|\mathbb{E}g|$$

with $M_{3,R} = 4(\kappa + C + 2(C+1))^2 R^2$ and $v_{3,R} = 4^3 M_{3,R}$. The constants are independent of j. Thus we can apply Lemma 5 and get

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\mathcal{E}(f_j) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(f_j) - \mathcal{E}_{\mathbf{z}}(f_\rho))}{\sqrt{\varepsilon + \mathcal{E}(f_j) - \mathcal{E}(f_\rho)}} \ge \sqrt{\varepsilon}\right\} \le \exp\left\{-\frac{m\varepsilon}{2(v_{3,R} + M_{3,R})}\right\}.$$

Now we take all these events with $j \in \{1, \ldots, \mathcal{N}\}$ and see that

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\max_{1\le j\le\mathcal{N}}\frac{\mathcal{E}(f_j) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(f_j) - \mathcal{E}_{\mathbf{z}}(f_\rho))}{\sqrt{\varepsilon + \mathcal{E}(f_j) - \mathcal{E}(f_\rho)}} \ge \sqrt{\varepsilon}\right\} \le \mathcal{N}\exp\left\{-\frac{m\varepsilon}{2(v_{3,R} + M_{3,R})}\right\}.$$

Setting the right-hand side to be $\frac{\delta}{4}$, we choose

$$\varepsilon = \frac{2(v_{3,R} + M_{3,R})}{m}\log(4\mathcal{N}/\delta) \le 520(\kappa + C + 2(C+1))^2\frac{R^2}{m}\log(4\mathcal{N}/\delta) = \varepsilon_{m,\delta,\mathcal{N},R}.$$

Then we conclude that with confidence at least $1-\frac{\delta}{4}$, there holds

$$\begin{aligned}\mathcal{E}(f_j)-\mathcal{E}(f_\rho)-[\mathcal{E}_{\mathbf{z}}(f_j)-\mathcal{E}_{\mathbf{z}}(f_\rho)] &\le \sqrt{\varepsilon_{m,\delta,\mathcal{N},R}}\sqrt{\varepsilon_{m,\delta,\mathcal{N},R}+\mathcal{E}(f_j)-\mathcal{E}(f_\rho)}\\ &\le \varepsilon_{m,\delta,\mathcal{N},R}+\frac{1}{2}\{\mathcal{E}(f_j)-\mathcal{E}(f_\rho)\} \quad \forall j\in\{1,\ldots,\mathcal{N}\}.\end{aligned}$$

This proves our statements. □

4.3. A new covering number argument

Now we can describe our new covering number argument. It is based on the observation that $f_{\mathbf{z}}$ is only one function (though it changes with the sample $\mathbf{z}$) and can be very close to one of the functions in the net $\{f_j\}_{j=1}^{\mathcal{N}}$. Hence we can bound $\mathcal{S}_1(\mathbf{z})$ with a single (but varying) function f_j instead of the quantity $\sup_{f\in\mathcal{F}_\lambda}|\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)|$ in (2.4) concerning the whole function set $\mathcal{F}_\lambda$.

Denote

$$W(R)=\{\mathbf{z}\in Z^m:\|f_{\mathbf{z}}\|_K\le R\}.$$

Lemma 7. *Let $0<\delta<1$ and $R\ge M$. Take Z_δ in Proposition 1 and Z'_δ in Lemma 6. Then for every $\mathbf{z}\in Z_\delta\bigcap Z'_\delta\bigcap W(R)$ we have*

$$[\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_\rho)]-[\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_\rho)]\le C_1R^2m^{-\frac{1}{2(s+1)}}\log\frac{4}{\delta}+\frac{1}{2}\{\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_\rho)\},$$

where C_1 is the constant given by

$$C_1=6\kappa+6C+8(1+\sqrt{2C})/M+520(\kappa+C+2(C+1))^2(C_0+1).$$

Proof. Let $\mathbf{z}\in Z_\delta\bigcap Z'_\delta\bigcap W(R_\delta)$. It satisfies $\frac{1}{m}\sum_{i=1}^m|y_i|\le M_\delta$. Moreover,

$$\mathcal{E}(f_j)-\mathcal{E}(f_\rho)-[\mathcal{E}_{\mathbf{z}}(f_j)-\mathcal{E}_{\mathbf{z}}(f_\rho)]\le\varepsilon_{m,\delta,\mathcal{N},R}+\frac{1}{2}\{\mathcal{E}(f_j)-\mathcal{E}(f_\rho)\},\quad\forall j\in\{1,\ldots,\mathcal{N}\}.$$

Let $\eta=Rm^{-\frac{1}{s+1}}$ and $\{f_j\}_{j=1}^{\mathcal{N}}$ with $\mathcal{N}=\mathcal{N}(B_R,\eta)$ be an η-net of the set B_R meaning that for any $f\in B_R$, there exists some $j\in\{1,\ldots,\mathcal{N}\}$ such that $\|f-f_j\|_\infty\le\eta$. Since $\mathbf{z}\in W(R)$, we know that $f_{\mathbf{z}}\in B_R$ and there is some $j_{\mathbf{z}}\in\{1,\ldots,\mathcal{N}\}$ such that $\|f_{\mathbf{z}}-f_{j_{\mathbf{z}}}\|_\infty\le\eta$. Then

$$\begin{aligned}|\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_{j_{\mathbf{z}}})| &= \left|\int_Z[f_{\mathbf{z}}(x)-f_{j_{\mathbf{z}}}(x)][f_{\mathbf{z}}(x)+f_{j_{\mathbf{z}}}(x)-2y]\mathrm{d}\rho\right|\\ &\le \eta\int_Z(|f_{\mathbf{z}}(x)|+|f_{j_{\mathbf{z}}}(x)|+2|y|)\mathrm{d}\rho\le\eta(2\kappa R+2\mathbb{E}|y|)\le2\eta(\kappa R+CM).\end{aligned}$$

From the bound $\frac{1}{m}\sum_{i=1}^m|y_i|\le M_\delta$, we find

$$\begin{aligned}|\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_{j_{\mathbf{z}}})| &= \left|\frac{1}{m}\sum_{i=1}^m[f_{\mathbf{z}}(x_i)-f_{j_{\mathbf{z}}}(x_i)][f_{\mathbf{z}}(x_i)+f_{j_{\mathbf{z}}}(x_i)-2y_i]\right|\\ &\le \eta\frac{1}{m}\sum_{i=1}^m(|f_{\mathbf{z}}(x_i)|+|f_{j_{\mathbf{z}}}(x_i)|+2|y_i|)\le2\eta(\kappa R+M_\delta).\end{aligned}$$

The above two bounds together with Lemma 6 tell us that

$$\begin{aligned}&[\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_\rho)]-[\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_\rho)]\\ &\quad=[\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_{j_{\mathbf{z}}})]+\{[\mathcal{E}(f_{j_{\mathbf{z}}})-\mathcal{E}(f_\rho)]-[\mathcal{E}_{\mathbf{z}}(f_{j_{\mathbf{z}}})-\mathcal{E}_{\mathbf{z}}(f_\rho)]\}+[\mathcal{E}_{\mathbf{z}}(f_{j_{\mathbf{z}}})-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})]\\ &\quad\le2\eta(2\kappa R+CM+M_\delta)+\varepsilon_{m,\delta,\mathcal{N},R}+\frac{1}{2}\{\mathcal{E}(f_{j_{\mathbf{z}}})-\mathcal{E}(f_\rho)\}.\end{aligned}$$

But $|\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{j_{\mathbf{z}}})| \leq 2\eta(\kappa R + CM)$. So the above expression is bounded by

$$2\eta(3\kappa R + 2CM + M_\delta) + \varepsilon_{m,\delta,\mathcal{N},R} + \frac{1}{2}\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)\}.$$

By the covering number condition (2.3), $\varepsilon_{m,\delta,\mathcal{N},R}$ can be bounded as

$$\begin{aligned}\varepsilon_{m,\delta,\mathcal{N},R} &\leq 520(\kappa + C + 2(C+1))^2\frac{R^2}{m}(\log(4/\delta) + C_0(R/\eta)^s)\\ &\leq \frac{520(\kappa + C + 2(C+1))^2C_0R^{2+s}}{m}\eta^{-s} + \frac{520(\kappa + C + 2(C+1))^2}{m}\log\frac{4}{\delta}R^2.\end{aligned}$$

Putting the choice of η and the expression of M_δ, we see that

$$\begin{aligned}[\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)] - [\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)] &\leq \frac{1}{2}\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)\} + \{6\kappa + 6C + 8(1+\sqrt{2C})/M\\ &\quad + 520(\kappa + C + 2(C+1))^2(C_0+1)\}R^2m^{-\frac{1}{s+1}}\log\frac{4}{\delta}.\end{aligned}$$

This proves our statement. □

Combining Lemmas 4 and 7, we get from (3.2) the following bound for $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \lambda\|f_{\mathbf{z}}\|_K^2$.

Proposition 2. *Let $0 < \delta < 1$ and $R \geq M$. There exists a subset V_R of Z^m with measure at most δ such that for every $\mathbf{z} \in W(R) \setminus V_R$,*

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \lambda\|f_{\mathbf{z}}\|_K^2 &\leq 38\mathcal{D}(\lambda) + 2[C_1 + 32^2(C+1)^2]R^2m^{-\frac{1}{s+1}}\log\frac{4}{\delta}\\ &\quad + \frac{480(\kappa+1)^2\mathcal{D}(\lambda)}{\lambda m}\log\frac{2}{\delta}.\end{aligned}$$

5. Deriving error bounds by iteration

We use an iteration technique [12,13,9] to derive our error bounds.

Proof of Theorem 2. From the bound for $\|f_{\mathbf{z}}\|_K$ obtained in Proposition 2, we see that for $R \geq M$ and $\mathbf{z} \in W(R) \setminus V_R$, we have $\|f_{\mathbf{z}}\|_K \leq a_{m,\delta}R + b_{m,\delta}$, where

$$a_{m,\delta} = \sqrt{2[C_1 + 32^2(C+1)^2]}\frac{1}{\sqrt{\lambda}}m^{-\frac{1}{2(s+1)}}\sqrt{\log(4/\delta)}$$

and

$$b_{m,\delta} = \sqrt{38\mathcal{D}(\lambda)/\lambda} + \sqrt{\frac{480(\kappa+1)^2\mathcal{D}(\lambda)}{\lambda^2 m}\log(2/\delta)} + M.$$

It tells us that $W(R) \subseteq W(a_{m,\delta}R + b_{m,\delta}) \bigcup V_R$.

Let us first derive a rough bound for $\|f_{\mathbf{z}}\|_K$. From the definition of $f_{\mathbf{z}}$, we see that

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\|f_{\mathbf{z}}\|_K^2 \leq \mathcal{E}_{\mathbf{z}}(0) + \lambda\|0\|_K^2 = \frac{1}{m}\sum_{i=1}^m y_i^2.$$

It implies that

$$\begin{aligned}\lambda\|f_{\mathbf{z}}\|_K^2 &\leq \frac{1}{m}\sum_{i=1}^m\{y_i^2 - (f_{\mathbf{z}}(x_i) - y_i)^2\} = \frac{1}{m}\sum_{i=1}^m\{f_{\mathbf{z}}(x_i)[2y_i - f_{\mathbf{z}}(x_i)]\}\\ &= \frac{1}{m}\sum_{i=1}^m\{f_{\mathbf{z}}(x_i)2y_i\} - \frac{1}{m}\sum_{i=1}^m\{f_{\mathbf{z}}(x_i)^2\} \leq \frac{2}{m}\sum_{i=1}^m\{f_{\mathbf{z}}(x_i)y_i\} \leq \frac{2}{m}\sum_{i=1}^m y_i\kappa\|f_{\mathbf{z}}\|_K.\end{aligned}$$

It follows that $\|f_{\mathbf{z}}\|_K \leq \frac{2\kappa}{m\lambda}\sum_{i=1}^m y_i$ and we see from Proposition 1 that for $\mathbf{z} \in Z_\delta$,

$$\|f_{\mathbf{z}}\|_K \leq \frac{2\kappa M_\delta}{\lambda}.$$

It means that $Z_\delta \subseteq W(R_\delta^{(0)})$ where $R_\delta^{(0)} = \frac{2\kappa M_\delta}{\lambda} + M$. Define a sequence $\{R_\delta^{(j)}\}_{j=0}^\infty$ by

$$R_\delta^{(j+1)} = a_{m,\delta} R_\delta^{(j)} + b_{m,\delta}.$$

We know that

$$W(R_\delta^{(0)}) \subseteq W(R_\delta^{(1)}) \bigcup V_{R_\delta^{(0)}} \subseteq \cdots \subseteq W(R_\delta^{(J)}) \bigcup \left\{ \bigcup_{j=0}^{J-1} V_{R_\delta^{(j)}} \right\}.$$

Since each set $V_{R_\delta^{(j)}}$ has measure at most δ, the set $W(R_\delta^{(J)})$ has measure at least $1 - (J+1)\delta$.

By the definition of the sequence $\{R_\delta^{(j)}\}$, $R_\delta^{(J)} = a_{m,\delta}^J R_\delta^{(0)} + b_{m,\delta}\sum_{i=0}^{J-1} a_{m,\delta}^i$. Now we take $\lambda = m^{2\epsilon - \frac{1}{s+1}}$ with $0 < \epsilon < \frac{1}{2(s+1)}$. Then

$$a_{m,\delta} \leq C_2 m^{-\epsilon}\sqrt{\log(4/\delta)},$$

where $C_2 = \sqrt{2[C_1 + 32^2(C+1)^2]}$. Putting $\mathcal{D}(\lambda) \leq C_\beta \lambda^\beta$ into $b_{m,\delta}$, we see

$$b_{m,\delta} \leq \{\sqrt{38C_\beta} + (\kappa+1)\sqrt{480C_\beta} + M\} m^{\frac{\beta-1}{2}(2\epsilon - \frac{1}{s+1})}\sqrt{\log(2/\delta)}.$$

Thus with $C_3 = \sqrt{38C_\beta} + (\kappa+1)\sqrt{480C_\beta} + M$, we have

$$\begin{aligned} R_\delta^{(J)} \leq\ & C_2^J \left(\log\frac{4}{\delta}\right)^{\frac{J}{2}+1} M[2\kappa(C + (1+2\sqrt{2C})) + 1] m^{\frac{1}{s+1} - (J+2)\epsilon} \\ & + C_3 m^{(\beta-1)\epsilon + \frac{1-\beta}{2(s+1)}} \sqrt{\log(2/\delta)} J \max\{1, (C_2 m^{-\epsilon}\sqrt{\log(4/\delta)})^{J-1}\}. \end{aligned}$$

Choose J_ϵ to be the smallest integer satisfying $J_\epsilon \geq \frac{1+\beta}{2\epsilon(s+1)} - 2$. Then $J_\epsilon \leq \frac{1+\beta}{2\epsilon(s+1)} - 1 \leq \frac{1}{\epsilon(s+1)} - 1$ and

$$\begin{aligned} R_\delta^{(J_\epsilon)} &\leq \left\{ M[2\kappa(C + (1+2\sqrt{2C})) + 1] C_2^{J_\epsilon} \left(\log\frac{4}{\delta}\right)^{\frac{J_\epsilon}{2}+1} + C_3 J_\epsilon C_2^{J_\epsilon - 1} (\log(4/\delta))^{\frac{J_\epsilon}{2}} \right\} m^{\frac{1-\beta}{2(s+1)}} \\ &\leq C_4 \frac{1}{\epsilon(s+1)} C_2^{\frac{1}{\epsilon(s+1)}} (\log(4/\delta))^{\frac{J_\epsilon}{2}+1} m^{\frac{1-\beta}{2(s+1)}}, \end{aligned}$$

where

$$C_4 = M(2\kappa(C + (1+2\sqrt{2C})) + 1) + C_3.$$

Since the set $W(R_\delta^{(J_\epsilon)})$ has measure at least $1 - (J_\epsilon + 1)\delta$, applying Proposition 2 with $R = R_\delta^{(J_\epsilon)}$, we conclude that with confidence at least $1 - (J_\epsilon + 2)\delta$,

$$\begin{aligned} \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq\ & 38C_\beta m^{2\beta\epsilon - \frac{\beta}{s+1}} + 2(C_1 + 32^2(C+1)^2) C_4^2 \left(\frac{1}{\epsilon(s+1)}\right)^2 \\ & \times C_2^{\frac{2}{\epsilon(s+1)}} m^{-\frac{\beta}{s+1}} (\log(4/\delta))^{J_\epsilon + 3} + 480(\kappa+1)^2 C_\beta m^{-1 + \frac{1-\beta}{s+1} - 2\epsilon(1-\beta)} \log\frac{2}{\delta}. \end{aligned}$$

By setting $\widetilde{\delta} = (J_\epsilon + 2)\delta$ and $\widetilde{\epsilon} = 2\beta\epsilon$, we know that with confidence at least $1 - \widetilde{\delta}$,

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq \frac{C_5}{\widetilde{\epsilon}^2} C_2^{\frac{4\beta}{\widetilde{\epsilon}(s+1)}} m^{\widetilde{\epsilon} - \frac{\beta}{s+1}} \left(\log(4/\widetilde{\delta}) + \log\left(\frac{2}{\widetilde{\epsilon}(s+1)} + 1\right)\right)^{\frac{\beta(1+\beta)}{\widetilde{\epsilon}(s+1)} + 2},$$

where

$$C_5 = 38C_\beta + 2(C_1 + 32^2(C+1)^2)C_4^2\left(\frac{2}{s+1}\right)^2 + 480(\kappa+1)^2C_\beta.$$

This proves the conclusion of Theorem 2 by taking

$$\widetilde{C}_\epsilon = \frac{C_5}{\epsilon^2}C_2^{\frac{4\beta}{\epsilon(s+1)}}\left(1+\log\left(\frac{2}{\epsilon(s+1)}+1\right)\right)^{\frac{\beta(1+\beta)}{\epsilon(s+1)}+2}$$

to be a constant independent of m or δ. □

When K is C^∞ we know from [18,19] that (2.3) holds for any $s > 0$. Take $\beta = 1$ and $s = \frac{\epsilon}{1-\epsilon}$ when $0 < \epsilon < \frac{1}{2}$ in Theorem 2. We know that by taking $\lambda = m^{2\epsilon-1}$, for any $0 < \delta < 1$, with confidence $1-\delta$, we have

$$\|f_{\mathbf{z}} - f_\rho\|^2_{L^2_{\rho_X}} \le \tilde{C}_\epsilon m^{2\epsilon-1}\left(\log\frac{4}{\delta}\right)^{\frac{2}{\epsilon}+2}.$$

This verifies Theorem 1 by scaling 2ϵ to ϵ.

References

[1] G. Bennett, Probability inequalities for the sum of independent random variables, J. Amer. Statist. Assoc. 57 (1962) 33–45.

[2] A. Caponnetto, E. De Vito, Optimal rates for regularized least-squares algorithms, Found. Comput. Math. 7 (2007) 331–368.

[3] D.R. Chen, Q. Wu, Y. Ying, D.X. Zhou, Support vector machine soft margin classifiers: error analysis, J. Mach. Learn. Res. 5 (2004) 1143–1175.

[4] E. De Vito, A. Caponnetto, L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, Found. Comput. Math. 5 (2005) 59–85.

[5] S. Mendelson, J. Neeman, Regularization in kernel learning, Ann. Statist. 38 (2010) 526–565.

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

[7] S. Smale, D.X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. 1 (2003) 17–41.

[8] S. Smale, D.X. Zhou, Online learning with Markov sampling, Anal. Appl. 7 (2009) 87–113.

[9] I. Steinwart, A. Christmann, Support Vector Machines, Springer-Verlag, New York, 2008.

[10] I. Steinwart, D. Hush, C. Scovel, A new concentration result for regularized risk minimizers, E. Giné, V. Kolchinskii, W. Li, J. Zinn (Eds.), High Dimensional Probability IV, Institute of Mathematical Statistics, Beachwood, 2006 pp. 260–275.

[11] I. Steinwart, D. Hush, C. Scovel, Optimal rates for regularized least-squares regression, in: S. Dasgupta, A. Klivans (Eds.), Proceedings of the 22nd Annual Conference on Learning Theory, 2009, pp. 79–93.

[12] I. Steinwart, C. Scovel, Fast rates for support vector machines, Lecture Notes in Comput. Sci. 3559 (2005) 279–294.

[13] Q. Wu, Y. Ying, D.X. Zhou, Learning rates of least-square regularized regression, Found. Comput. Math. 6 (2006) 171–192.

[14] Q. Wu, Y. Ying, D.X. Zhou, Multi-kernel regularized classifiers, J. Complexity 23 (2007) 108–134.

[15] Q. Wu, D.X. Zhou, Analysis of support vector machine classification, J. Comput. Anal. Appl. 8 (2006) 99–119.

[16] G.B. Ye, D.X. Zhou, SVM learning and L^p approximation by Gaussians on Riemannian manifolds, Anal. Appl. 7 (2009) 309–339.

[17] T. Zhang, Leave-one-out bounds for kernel methods, Neural Comput. 15 (2003) 1397–1437.

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

[19] D.X. Zhou, Derivative reproducing properties for kernel methods in learning theory, J. Comput. Appl. Math. 220 (2008) 456–463.