Generalization Performance of Radial Basis Function Networks

Yunwen Lei, Lixin Ding, and Wensheng Zhang

***Abstract*—This paper studies the generalization performance of *Radial Basis Function* (RBF) networks by using local Rademacher complexities. We propose a general result on controlling local Rademacher complexities with the L_1-metric capacity. We then apply this result to estimate RBF networks' complexities, based on which a novel estimation error bound is obtained. An effective approximation error bound is also derived by carefully investigating the Hölder continuity of the ℓ_p loss function's derivative. Furthermore, it is demonstrated that the RBF network minimizing an appropriately constructed structural risk admits a significantly better learning rate when compared to the existing results. An empirical study is also performed to justify the application of our structural risk in model selection.**

***Index Terms*—Structural risk minimization, Learning theory, Local Rademacher complexity, Radial basis function networks**

I. INTRODUCTION

ARTIFICIAL neural networks have proved to be effective modeling strategies in approximating nonlinear relationships between input and output variables [1], [2]. As compared to traditional nonparametric estimation methods, neural networks have an advantage of dimensionality reduction by composition and thus have found great success in various multivariate prediction problems [3]. Among different types of artificial neural networks, *Radial Basis Function* (RBF) networks have received considerable attention since they constitute solutions for regularization problems using certain standard smoothness functionals as stabilizers [1]. Estimating the generalization performance of RBF networks is important to understand the factors influencing models' quality, as well as to suggest possible ways to improve them [4], [5], [6]. This paper investigates the learning ability of RBF networks under the *Structural Risk Minimization* (SRM) principle. Our basic strategy is to consider separately two contradictory factors determining the generalization performance: approximation errors and estimation errors.

Recent years have witnessed a great progress in understanding the approximation power of RBF networks. Park and Sandberg [7] indicated that RBF networks with Gaussian computational nodes admit universal approximation ability. Namely, they are able to approximate with arbitrary accuracy among all square integrable functions on compact subsets of $\mathbb{R}^d$, where d is the input dimension. For band-limited functions with continuous derivatives up to order $r > d/2$, Girosi and Anzellotti [8] used a probability trick to derive an approximation error rate of the form $k^{-1/2}$, where k is the number of neurons. Girosi [9] pioneered the research of applying tools in learning theory to obtain satisfactory approximation error rates for more general kernel classes. Gnecco and Sanguineti [10] refined this result by using the more advanced tool called Rademacher complexity. The tractability issues in RBF network approximation were treated by Kainen et al. [11], [12]. Estimation errors for RBF networks have also been extensively studied in Anthony and Bartlett [13], Niyogi and Girosi [14], Haussler [15], Györfi et al. [16], using standard complexity measures such as pseudo-dimensions and covering numbers.

Some researchers also provided unified viewpoints to simultaneously consider approximation and estimation errors. Barron [17] addressed the combined effect of the approximation and estimation error on the overall accuracy of a network as a prediction rule. However, his approach is based on covering numbers under the supremum norm and therefore the activation functions considered there are required to satisfy the Hölder condition. Niyogi and Girosi [4] removed this restriction by using the more relaxed L_1-metric capacity instead. Unfortunately, their analysis relies on uniform deviation bounds via a Hoeffding type inequality, which ignores the information on variances and could only yield a suboptimal learning rate. Krzyżak and Linder [1] refined these results by applying a ratio-type inequality under the squared loss setting. However, there still exist some weaknesses that could be improved in their discussion for the general ℓ_p loss $\varphi_p(t) := |t|^p, p > 1$:

(1) Under the ℓ_p loss, the discussion of the estimation error in Krzyżak and Linder [1] is based on the uniform (supremum) deviation argument. It may happen that the established model stays far away from achieving this supremum and therefore this deduction could only yield a rather conservative result [18]. On the other hand, most learning algorithms are inclined towards choosing models with small expected errors and thus the uniform deviation over sub-classes with small expected errors is sufficient to control estimation errors [5]. A remarkable concept called local Rademacher complexity has been introduced into the learning theory community to capture this speciality of learning algorithms [5], [19], [20]. Since local Rademacher complexity allows us to concentrate our attention to those sub-classes of primary interest, it always yields considerably improved estimation error bounds

Y. Lei and L. Ding are with State Key Lab of Software Engineering, School of Computer, Wuhan University, Wuhan 430072, China (e-mail: ywlei@whu.edu.cn, lxding@whu.edu.cn).

W. Zhang is with Institute of Automation, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: wensheng.zhang@ia.ac.cn)

This work is supported in part by the National Natural Science Foundation of China (Grant no. 60975050).

when the variance-expectation relation holds [5], [18].

(2) Under the ℓ_p loss, Krzyżak and Linder [1] only exploited the Hölder continuity of φ_p to show that the approximation error $\mathcal{E}(h_k^*) - \mathcal{E}(h^*)$ decays as a linear function of the distance $\|h_k^* - h^*\|^{1/2}$, where h^* is the target function and h_k^* is defined as Eq. (9). However, recent studies indicated that when the derivative $\varphi_p^{'}$ is Hölder continuous, $\mathcal{E}(h_k^*) - \mathcal{E}(h^*)$ can superlinearly decay as a function of $\|h_k^* - h^*\|^{1/2}$ [21]. Consequently, it is worthwhile to investigate the Hölder continuity of $\varphi_p^{'}$ rather than that of φ_p to derive improved approximation error rates.

In this paper we study these issues by providing novel generalization error bounds for RBF networks under the SRM principle. Our main scheme is to apply local Rademacher complexities to refine the existing estimation error bounds and to use the Hölder continuity of $\varphi_p^{'}$ to provide improved approximation error bounds. For this purpose, we first offer a general result on controlling local Rademacher complexities with the L_1-metric capacity. This bound is novel since it is based on the L_1-metric capacity rather than the traditional and larger L_2-metric capacity. Then we apply this general result to control RBF networks' local Rademacher complexities, based on which we derive an effective estimation error bound and construct an appropriate structural risk. The approximation power of RBF networks is investigated by exploiting the Hölder continuity of $\varphi_p^{'}$. It is shown that the RBF network minimizing our structural risk attains a favorable trade-off between approximation and estimation errors, yielding a learning rate significantly better than that in Krzyżak and Linder [1]. We also present an empirical study to support our theoretical deduction.

This paper is organized as follows. In Section II the problem is formulated. The main theorem, as well as its superiority to the results in Krzyżak and Linder [1], is presented in Section III. Section IV addresses local Rademacher complexity bounds. Section V tackles estimation and approximation errors for RBF networks. An empirical study is provided in Section VI. Section VII presents some conclusion remarks.

II. Problem formulation

Before formulating our problem we first introduce some notations that will be used throughout this paper. Given a set $\{Z_1, \dots, Z_n\}$, the associated empirical measure P_n is defined as $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$, where δ_{Z_i} is the Dirac measure supported on the point Z_i [22]. For a measure μ and a measurable function g, we use the notation $\mu g = \int g \mathrm{d}\mu$ to denote the expectation of g. Now, the empirical average of g over $Z_1, \dots, Z_n$ can be abbreviated as $P_n g = \frac{1}{n}\sum_{i=1}^n g(Z_i)$.

For a measure μ and a number $1 \leq q < \infty$, the notation $L_q(\mu)$ means the class of functions with finite norm $\|f\|_{L_q(\mu)} := [\int |f|^q \mathrm{d}\mu]^{1/q}$. The infinity-norm of a function is defined by $\|f\|_\infty := \sup_x |f(x)|$. By the notation $\mathrm{sgn}(x)$ we mean the sign of x, i.e., $\mathrm{sgn}(x) = 1$ if $x \geq 0$ and -1 otherwise. For any $d \in \mathbb{N}^+$, we denote by $\mathbb{S}^{d\times d}$ the class of non-negative definite $d \times d$ matrices. The minimum of two numbers is denoted by $a_1 \wedge a_2 := \min(a_1, a_2)$. By c we denote constants independent of the sample size n, complexity index (number of neurons) k and input dimension d, and their values may change from line to line, or even within the same line.

TABLE I
Notations

$\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$	sample space $\mathcal{Z}$ with input space $\mathcal{X}$ and output space $\mathcal{Y}$
n, k	sample size and number of neurons, respectively
d, b	input dimension and output bound, respectively
$\mathcal{H}_k$	the space of RBF networks with k nodes, Eq. (8)
$\overline{\mathcal{H}}_b$	closure of $\mathcal{H}_b^{'}$ in Eq. (14)
$\mathcal{F}_k$	the k-th loss class, Eq. (16)
$\mathcal{F}_k^*$	the k-th shifted loss class, Eq. (17)
$\mathcal{E}(h)$	generalization error (risk) of h, Eq. (1)
$\mathcal{E}_{\boldsymbol{z}}(h)$	empirical error of h, Eq. (2)
$\widetilde{\mathcal{E}}_{\boldsymbol{z}}(\hat{h}_k)$	the structural risk of $\hat{h}_k$, Eq. (10)
$\hat{h}_k$	ERM model in the class $\mathcal{H}_k$, Eq. (9)
h_n	SRM model, Eq. (4)
h_k^*	best model in the class $\mathcal{H}_k$, Eq. (9)
h^*	target function, $h^* := \mathrm{argmin}_h\, \mathcal{E}(h)$
$\hat{f}_k$	an element in $\mathcal{F}_k^*$ defined by Eq. (26)
P_n	empirical measure
$\mathbb{S}^{d\times d}$	the class of non-negative definite $d \times d$ matrices
$a_1 \wedge a_2$	the minimum between a_1 and a_2
c	a constant independent of n, k and d
$\mathrm{sgn}(x)$	the sign of x
φ_p	ℓ_p loss $\varphi_p(t) = \lvert t\rvert^p$
α_p, β_p	two constants given below Eq. (10)
$L_q(\mu)$	the function class with norm $\|f\|_{L_q(\mu)} = [\int \lvert f\rvert^q \mathrm{d}\mu]^{1/q}$

A. Learning and structural risk minimization

In the machine learning context, we are given an input space $\mathcal{X}$, an output space $\mathcal{Y}$ and a probability measure P defined on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ governing the sampling process [23]. When presented with a sequence of examples $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n)$, the purpose of learning is to construct a prediction rule $h : \mathcal{X} \to \mathcal{Y}$ such that it can perform the prediction as accurately as possible [21], [24], [25]. The local error suffered from using $h(x)$ to predict y is quantified by $\varphi(h(x) - y)$, where φ is a non-negative loss function. Consequently, the quality of a prediction rule h is characterized by its generalization error (also called risk)

$$\mathcal{E}(h) := \int \varphi(h(X) - Y)\mathrm{d}P. \tag{1}$$

The function $h^* := \mathrm{argmin}_h\, \mathcal{E}(h)$ with minimal risk is called the target function, where the minimum is taken over all measurable functions. Since the underlying measure P is often unknown to us, the term $\mathcal{E}(h)$ cannot be directly used to guide the learning process and as an alternative we use the empirical error (empirical risk)

$$\mathcal{E}_{\boldsymbol{z}}(h) := \frac{1}{n}\sum_{i=1}^n \varphi(h(X_i) - Y_i) \tag{2}$$

to approximate $\mathcal{E}(h)$ [26], [27]. Under the famous *Empirical Risk Minimization* (ERM) principle [26], one simply minimizes the empirical risk over a pre-selected hypothesis space $\mathcal{H}$ to obtain the estimator $\hat{h}$, that is, $\hat{h} := \mathrm{argmin}_{h\in\mathcal{H}}\, \mathcal{E}_{\boldsymbol{z}}(h)$.

As the empirical error can be optimistically biased compared to the corresponding generalization error, the direct minimization of $\mathcal{E}_{\boldsymbol{z}}(h)$ may result in overfitting or underfitting [1], [24]. To see this, we identify two factors determining the model's generalization performance by recalling the following bias-variance decomposition [24], [28]:

$$\mathbb{E}\mathcal{E}(\hat{h})-\mathcal{E}(h^*)=\left(\mathbb{E}\mathcal{E}(\hat{h})-\inf_{h\in\mathcal{H}}\mathcal{E}(h)\right)+\left(\inf_{h\in\mathcal{H}}\mathcal{E}(h)-\mathcal{E}(h^*)\right). \tag{3}$$

The first term is often called the estimation error, while the second is the approximation error [24], [28]. The approximation error results from the insufficient representation power of the associated hypothesis space, which can be made arbitrarily small by expanding the searching space [17]. However, this is bound to increase the estimation difficulty and therefore causes a large estimation error [1]. Consequently, the performance of ERM scheme is sensitive to the class $\mathcal{H}$ [23], [29].

An effective strategy to tackle this bias-variance phenomenon is to employ the SRM principle [24], [26]. Unlike ERM, SRM considers a sequence of classes $\mathcal{H}_k, k\in\mathbb{N}^+$ with increasing complexities and then builds a set of candidate models $\hat{h}_k$, one from each class $\mathcal{H}_k, k\in\mathbb{N}^+$. Now, the structural risk $\widetilde{\mathcal{E}}_{\boldsymbol{z}}(\hat{h}_k)$ is established by adding a penalty term reflecting $\mathcal{H}_k$'s complexity into $\mathcal{E}_{\boldsymbol{z}}(\hat{h}_k)$. The ultimate model

$$h_n := \operatorname*{argmin}_{\hat{h}_k, k\in\mathbb{N}^+}\widetilde{\mathcal{E}}_{\boldsymbol{z}}(\hat{h}_k) \tag{4}$$

is derived by minimizing the structural risk over all candidate prediction rules [1], [24]. It is well known that the success of the SRM principle largely depends on the quality of the constructed structural risk, which should balance the empirical accuracy and the complexity of the candidate models [1], [26].

Theorem 1 ([24]). *Assume that for each complexity index $k\in\mathbb{N}^+$, $\hat{h}_k$ minimizes the empirical risk over the k-th hypothesis space $\mathcal{H}_k$. Suppose that for every sample size n, there are positive numbers κ and γ such that for each k an estimate $L_{n,k}$ of $\mathcal{E}(\hat{h}_k)$ is available which satisfies*

$$\Pr\left\{\mathcal{E}(\hat{h}_k)>L_{n,k}+t\right\}\le\kappa e^{-\gamma t} \tag{5}$$

for any $t>0$. Assume that the model h_n is defined by

$$h_n=\operatorname{argmin}_{\hat{h}_k,k\in\mathbb{N}^+}\widetilde{\mathcal{E}}_{\boldsymbol{z}}(\hat{h}_k),\qquad \widetilde{\mathcal{E}}_{\boldsymbol{z}}(\hat{h}_k):=L_{n,k}+\frac{2\log k}{\gamma}.$$

Then the generalization error can be controlled as follows

$$\mathbb{E}\mathcal{E}(h_n)-\mathcal{E}(h^*)\le\min_{k\in\mathbb{N}^+}\Bigg[\mathbb{E}\left(L_{n,k}-\mathcal{E}_{\boldsymbol{z}}(\hat{h}_k)\right)+\left(\inf_{h\in\mathcal{H}_k}\mathcal{E}(h)-\mathcal{E}(h^*)\right)+\frac{2\log k+\log(2e\kappa)}{\gamma}\Bigg]. \tag{6}$$

Theorem 1 justifies the success of the SRM principle by showing that the model minimizing a suitable structural risk can automatically trade-off the approximation and estimation errors. We choose to present it here since it is quite important for the progression of our theoretical discussion. For example, we will use Eq. (5) to guide the construction of our specific structural risk. Furthermore, Eq. (6) allows us to consider separately the approximation and estimation errors when studying the generalization performance of h_n.

B. Radial basis function networks

We consider here RBF networks with one hidden layer, which can be characterized by a kernel $K:\mathbb{R}^+\to\mathbb{R}$. The sample space is of the form $\mathcal{Z}:=\mathcal{X}\times\mathcal{Y}\subset\mathbb{R}^d\times[-b,b]$, where d is the input dimension and b is a positive number. A RBF network with k nodes considered here takes the form [1]

$$h(x)=\sum_{i=1}^k w_iK\left([x-c_i]^TA_i[x-c_i]\right)+w_0, \tag{7}$$

where $w_0,\ldots,w_k$ are real numbers called weights, $c_1,\ldots,c_k\in\mathbb{R}^d$ are centroids and $A_1,\ldots,A_k$ are non-negative definite $d\times d$ matrices determining the receptive field of the kernel function K [1], [2]. Here x^T denotes the transpose of the vector x. Some typical kernels include the Gaussian kernel $K(t)=e^{-t}$, the exponential kernel $K(t)=e^{-\sqrt{t}}$ and the inverse multi-quadratic kernel $K(t)=(1+t)^{-1/2}$ [16]. Neural networks are trained under the SRM principle and the k-th hypothesis space consists of functions that can be expressed as Eq. (7) with k nodes satisfying the weight condition $\sum_{i=0}^k|w_i|\le b$. That is,

$$\mathcal{H}_k=\left\{\sum_{i=1}^k w_iK\left([x-c_i]^TA_i[x-c_i]\right)+w_0:\sum_{i=0}^k|w_i|\le b\right\}. \tag{8}$$

The candidate models $\hat{h}_k, k\in\mathbb{N}^+$ are constructed by minimizing the empirical error in the associated hypothesis spaces under the ℓ_p loss $\varphi_p(t):=|t|^p, p>1$. In order to explicitly indicate the dependence on the class, we use h_k^* and $\hat{h}_k$ to denote the minimizer of the risk and empirical risk over the k-th class, respectively. That is,

$$h_k^*=\operatorname*{argmin}_{h\in\mathcal{H}_k}\mathcal{E}(h)\qquad\text{and}\qquad\hat{h}_k=\operatorname*{argmin}_{h\in\mathcal{H}_k}\mathcal{E}_{\boldsymbol{z}}(h). \tag{9}$$

It should be noted that dependencies of some notations, e.g., $h_k^*,\hat{h}_k,\mathcal{E}(h),\mathcal{E}_{\boldsymbol{z}}(h)$, on the parameter p are hidden for brevity.

III. MAIN RESULTS

The purpose of this paper is to study the generalization performance of RBF networks under the SRM principle (4) with the following specific structural risk:

$$\begin{aligned}\widetilde{\mathcal{E}}_{\boldsymbol{z}}(\hat{h}_k):=\mathcal{E}(h^*)+\beta_p\left[\mathcal{E}_{\boldsymbol{z}}(\hat{h}_k)-\mathcal{E}_{\boldsymbol{z}}(h^*)\right]\\+c\left((kd^2\log n)^{\frac{1}{2-\alpha_p}}+2\log k\right)n^{-\frac{1}{2-\alpha_p}}.\end{aligned} \tag{10}$$

Here the constants are $\alpha_p=2/p\wedge 1$ and $\beta_p=2$ if $1<p\le 2,\beta_p=p/(p-2)$ if $p>2$. Since $\mathcal{E}(h^*)$ and $\mathcal{E}_{\boldsymbol{z}}(h^*)$ remain as constants for all candidate models $\hat{h}_k, k\in\mathbb{N}^+$, the structural risk (10) can also be reformulated as follows:

$$\widetilde{\mathcal{E}}_{\boldsymbol{z}}(\hat{h}_k):=\mathcal{E}_{\boldsymbol{z}}(\hat{h}_k)+\frac{c}{\beta_p}\left((kd^2\log n)^{\frac{1}{2-\alpha_p}}+2\log k\right)n^{-\frac{1}{2-\alpha_p}}. \tag{11}$$

Theorem 2 shows that the risk of the SRM model under the structural risk (10) is indeed within a constant factor of the risk of the best model in the optimal class, i.e., almost as good as if the optimal class has been previously indicated by an "oracle" [30]. Theorem 2 is proved in part D of the appendix. A real-valued function f defined on an interval $[a_1,a_2]$ is said to be of bounded variation if there exists a number V such that $\sum_{i=2}^{m}|f(x_i)-f(x_{i-1})|<V$ for all knots $a_1 \le x_1 < x_2 < \cdots < x_m \le a_2, \forall m \in \mathbb{N}^+$ [13].

Theorem 2 (Main result)**.** *Suppose that the examples $Z_i=(X_i,Y_i), i=1,2,\ldots,n$ are independently drawn according to a probability measure P defined on $\mathcal{Z}:=\mathcal{X}\times[-b,b], b>0$, where $\mathcal{X}\subset\mathbb{R}^d$ is the input space and d is the input dimension. Assume that the loss function is $\varphi_p, p>1$ and the kernel K is of bounded variation satisfying the condition $\sup_t|K(t)|\le 1$. Then for the prediction rule h_n minimizing the structural risk* (10)*, the term $\mathbb{E}\mathcal{E}(h_n)-\mathcal{E}(h^*)$ can be upper bounded by*

$$\min_{k\in\mathbb{N}^+}\left[\beta_p\left(\mathcal{E}(h_k^*)-\mathcal{E}(h^*)\right)+c(kd^2n^{-1}\log n)^{\frac{1}{2-\alpha_p}}\right].$$

Here the definitions of β_p and α_p can be found below Eq. (10)*.*

The key point in proving Theorem 2 is to show that the structural risk (10) is an appropriate upper bound of $\mathcal{E}(\hat{h}_k)$ in the sense of Eq. (5). It is well known that the behavior of $\mathcal{E}(\hat{h}_k)$ heavily relies on the size of the class $\mathcal{H}_k$, which will be studied via the tool called local Rademacher complexity in Section IV. With this complexity bound at hand, Section V-A will apply a Talagrand type inequality to show that the structural risk (10) indeed meets the assumption (5).

Remark 1. For the case $p\neq 2, p>1$, Krzyżak and Linder [1] constructed the structural risk of the form[1]

$$\widetilde{\mathcal{E}}_{\boldsymbol{z}}(\hat{h}_k)=\mathcal{E}_{\boldsymbol{z}}(\hat{h}_k)+c\sqrt{\frac{kd^2\log n}{n}} \tag{12}$$

and indicated that the prediction rule under the associated SRM principle satisfies the bound

$$\mathcal{E}(h_n)-\mathcal{E}(h^*)\le\min_{k\in\mathbb{N}^+}\left[c\sqrt{\frac{kd^2\log n}{n}}+\left(\mathcal{E}(h_k^*)-\mathcal{E}(h^*)\right)\right]. \tag{13}$$

As compared to this result, Theorem 2 provides an exponentially faster learning rate in the sense that the exponent of n is much smaller. Although k appears as a linear term in our bound when $1<p<2$, one should note that this is indeed not a big drawback since the case $k\ll n$ is the one of primary interest. As we will see, Theorem 2 can yield a significantly faster learning rate than that can be derived from Eq. (13) when the target function admits some degree of regularity.

Remark 2. The underlying reason for failing to get these improved rates in Krzyżak and Linder [1] is that their discussion is based on a Hoeffding type inequality, which is bound to control the universal deviation of empirical means from their expectations over the entire class and can only lead to the conservative rate $c(n^{-1/2})$. As a comparison, our improvement is attributed to the following three strategies:

(1) The use of local Rademacher complexity rather than the global counterpart allows us to concentrate our attention to functions that are likely to be picked out by learning algorithms, typically constituting a subset of the original class with small risks.
(2) The variance-expectation relation of the associated shifted loss class, which we will consider in Section V-A, shows that functions in this subset always admit small variances. Consequently, to study the generalization performance of the prediction rule it suffices to consider a sub-class of functions with small variances.
(3) The application of a Talagrand type inequality (Theorem 6) permits us to exploit this information on variances to get refined learning rates.

When the target function h^* satisfies some regularity condition, one can control the approximation error $\mathcal{E}(h_k^*)-\mathcal{E}(h^*)$ by a function of k and thus obtain explicit error bounds for the SRM model h_n. In this paper, the smoothness condition on h^* is formulated by assuming that it belongs to $\overline{\mathcal{H}}_b$. Here $\overline{\mathcal{H}}_b$ is the closure of $\mathcal{H}_b^{'}$ in $L_{2\wedge p}(P_X)$ with

$$\mathcal{H}_b^{'}:=\left\{\sum_{i=1}^{m}t_ib_iK([x-c_i]^TA_i[x-c_i]):t_i>0,\sum_{i=1}^{m}t_i=1,\right.$$
$$\left.|b_i|\le b, c_i\in\mathbb{R}^d, A_i\in\mathbb{S}^{d\times d}, m\in\mathbb{N}^+\right\}. \tag{14}$$

The approximation error will be controlled in Section V-B by using the Hölder continuity of $\varphi_p^{'}$ to relate it to the metric distance $\|h_k^*-h^*\|_{L_{2\wedge p}(P_X)}$, which is more convenient to approach in approximation theory. The proof of Corollary 3 is given in part D of the appendix.

Corollary 3. *Under the same condition of Theorem 2 and if we further assume that $h^*\in\overline{\mathcal{H}}_b$, then*

$$\mathbb{E}\mathcal{E}(h_n)-\mathcal{E}(h^*)\le c\left(\frac{d^2\log n}{n}\right)^{\frac{p}{p+2}\wedge\frac{p}{3p-2}}, \tag{15}$$

where d is the input dimension, n is the sample size and p indicates the loss function.

Remark 3. Under the special case $p=2$, Krzyżak and Linder [1] derived the learning rate $(n^{-1}d^2\log n)^{1/2}$. However, Krzyżak and Linder only offered the learning rate $(n^{-1}d^2\log n)^{1/4}$ for the general loss function $\varphi_p(1<p<2)$. In comparison with these results, it can be clearly seen that the bound presented in Corollary 3 is much improved. Indeed, the exponent $\frac{p}{p+2}\wedge\frac{p}{3p-2}$ in Corollary 3 is always larger than $1/4$ for any $1<p<2$ (for the special case $p=2$, our learning rate recovers the result in Krzyżak and Linder [1]). The reason for this improvement consists in two independent aspects: (1) this corollary is based on a refined estimation error bound established in Theorem 2; (2) using the Hölder continuity of $\varphi_p^{'}$ we derive the following refined inequality on approximation error (see Eq. (44)):

$$\mathcal{E}(h)-\mathcal{E}(h^*)\le 2\|h-h^*\|_{L_p(P_X)}^p,$$

which is much better than the relationship [1]

$$\mathcal{E}(h)-\mathcal{E}(h^*)\le c\|h-h^*\|_{L_p(P_X)}$$

based on the Hölder continuity of φ_p.

[1] Krzyżak and Linder [1] did not consider the effect of d since the input dimension is treated as a constant hidden in the big O notation. However, a closer look of their deduction would recover the exact form of d in Eq. (12), (13).

IV. LOCAL RADEMACHER COMPLEXITY BOUNDS

As a first step to show that the structural risk (10) meets Eq. (5), we need to consider the complexity of the loss class

$$\mathcal{F}_k := \mathcal{F}_{k,p} = \left[|h(X) - Y|^p : h \in \mathcal{H}_k\right], \ k \in \mathbb{N}^+. \tag{16}$$

We use local Rademacher complexity to measure the size of function classes, as it can capture the key property of learning algorithms and can yield a significant improvement on error analysis when the variance-expectation assumption holds. However, as shown in Bartlett et al. [5], the local Rademacher complexity analysis applied to $\mathcal{F}_k$ can only yield an error bound of the form $\mathcal{E}(\hat{h}_k) \leq c\mathcal{E}_{\boldsymbol{z}}(\hat{h}_k) + o(1), c > 1$, which is non-consistent if $\inf_{h\in\mathcal{H}_k} \mathcal{E}(h) > 0$. This problem can be circumvented by applying local Rademacher complexity to, instead of the class $\mathcal{F}_k$, the shifted loss class $\mathcal{F}_k^*$:

$$\mathcal{F}_k^* := \mathcal{F}_{k,p}^* = \left[|h(X)-Y|^p - |h^*(X)-Y|^p : h\in\mathcal{H}_k\right], k\in\mathbb{N}^+. \tag{17}$$

Note that in Eqs. (16), (17), dependencies on p are suppressed for brevity. This section aims to estimate local Rademacher complexity bounds for the shifted loss classes (17). Section V will illustrate how to use these results to obtain satisfactory learning rates. The definition of Rademacher complexity can be traced back to Hans Rademacher and it was first proposed as an effective complexity measure by Koltchinskii [31].

Definition 1 (Rademacher complexities)**.** Let $\mathcal{F}$ be a class of functions on a probability space $(\mathcal{Z}, P)$ and let $Z_1, \ldots, Z_n$ be n points independently drawn from P. Suppose that $\sigma_1, \ldots, \sigma_n$ are n independent Rademacher random variables, i.e., $\Pr\{\sigma_i = 1\} = \Pr\{\sigma_i = -1\} = 1/2$. Introduce the notation

$$R_n\mathcal{F} = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i).$$

The Rademacher complexity $\mathbb{E}R_n\mathcal{F}$ is the expectation of $R_n\mathcal{F}$, and the empirical Rademacher complexity

$$\mathbb{E}_\sigma R_n\mathcal{F} := \mathbb{E}\left[R_n\mathcal{F}\middle|Z_1, \ldots, Z_n\right]$$

is defined as the conditional expectation of $R_n\mathcal{F}$.

Local Rademacher complexities differ from the standard Rademacher complexities in that the supremum is taken over a subset of the original class rather than the whole class. The subsets considered here are defined by the $L_2(P)$ norm or the $L_2(P_n)$ norm. To be precise, we consider here local Rademacher complexities of the form

$$\mathbb{E}_\sigma R_n\{f \in \mathcal{F} : Pf^2 \leq r\} \quad \text{or} \quad \mathbb{E}R_n\{f \in \mathcal{F} : Pf^2 \leq r\}.$$

Local Rademacher complexities can be viewed as functions of r and they allow us to filter out those functions with large variances, which are of little interest since learning algorithms are unlikely to select them [32].

Unfortunately, in the vast majority of cases, the direct computation of (local) Rademacher complexity is extremely difficult if not impossible [22]. The way to bypass this obstacle is to firstly relate it to other complexity measures such as covering numbers and then use these auxiliary concepts to estimate it indirectly.

Definition 2 (Covering numbers)**.** Let $(\mathcal{G}, d)$ be a metric space and let $\mathcal{F}$ be a subset of $\mathcal{G}$. For any $\epsilon > 0$, we say that $\{g_1, \ldots, g_m\} \subset \mathcal{G}$ is an ϵ-cover of $\mathcal{F}$ if

$$\sup_{f\in\mathcal{F}} \min_{1\leq i\leq m} d(f, g_i) \leq \epsilon.$$

The covering number $\mathcal{N}(\epsilon, \mathcal{F}, d)$ is defined as the cardinality of a minimal ϵ-cover of $\mathcal{F}$. When $\mathcal{G}$ is a normed space with norm $\|\cdot\|$, we also denote by $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)$ the covering number of $\mathcal{F}$ with respect to the metric $d(f, g) := \|f - g\|$.

For any probability measure P, we have the following relationship among covering numbers under different metrics [22]:

$$\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_{L_p(P)}) \leq \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_{L_q(P)}), \ \forall 1 \leq p \leq q. \tag{18}$$

In order to remove the dependence on the involved probability measure, we introduce the following L_p-metric capacity (L_p-norm covering numbers) by ranging P_n over all empirical measures supported on n points [4]:

$$\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_p) = \sup_n \sup_{P_n} \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_{L_p(P_n)}).$$

Estimating Rademacher complexities is a classical theme in learning theory. The first breakthrough in this direction is marked by Dudley's entropy integral [33], which captures in an elegant form the relationship between covering numbers and Rademacher complexities. Mendelson [22] extended this classical result to the local Rademacher complexity setting and provided some novel results for classes satisfying general entropy assumptions. These discussions always involve the L_2-metric capacity. In this section we generalize these results by illustrating how to use L_1-norm covering numbers to control local Rademacher complexities. To our best knowledge, this is the first result on estimating local Rademacher complexities via L_1-norm covering numbers.

Theorem 4. *Let $\mathcal{F}$ be a function class with $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq b$, where b is a positive number. Then for any $r > 0$ and sample size n, local Rademacher complexity can be controlled by:*

$$\mathbb{E}R_n\{f \in \mathcal{F} : Pf^2 \leq r\} \leq \inf_{\epsilon>0}\Bigg[2\epsilon + \frac{8b\log\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_1)}{n} + (2\sqrt{2b\epsilon} + \sqrt{r})\sqrt{\frac{2\log\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_1)}{n}}\Bigg].$$

Proof. We first introduce a new random variable

$$Y_r := \sup_{f\in\mathcal{F}:Pf^2\leq r} P_n f^2.$$

The definition of Y_r implies that for any sample, a function $f \in \mathcal{F}$ with $Pf^2 \leq r$ would automatically satisfy the inequality $P_n f^2 \leq Y_r$. Consequently, the following inclusion relationship holds almost surely:

$$\{f \in \mathcal{F} : Pf^2 \leq r\} \subset \{f \in \mathcal{F} : P_n f^2 \leq Y_r\}. \tag{19}$$

Moreover, Y_r meets the following inequality [22, Lemma 3.6]

$$\mathbb{E}Y_r \leq r + 4b\mathbb{E}R_n\{f \in \mathcal{F} : Pf^2 \leq r\}. \tag{20}$$

Putting Eqs. (19), (20) and Lemma 11 together, we have

$$\begin{aligned}
&\mathbb{E}R_n\{f\in\mathcal{F}:Pf^2\le r\}\\
&=\mathbb{E}\mathbb{E}_\sigma R_n\{f\in\mathcal{F}:Pf^2\le r\}\le\mathbb{E}\mathbb{E}_\sigma R_n\{f\in\mathcal{F}:P_nf^2\le Y_r\}\\
&\le\epsilon+\mathbb{E}\left[(\sqrt{2b\epsilon}+\sqrt{Y_r})\sqrt{\frac{2\log\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_{L_1(P_n)})}{n}}\right]\\
&\le\epsilon+(\sqrt{2b\epsilon}+\sqrt{\mathbb{E}Y_r})\sqrt{\frac{2\log\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_1)}{n}}\\
&\le\epsilon+\left(\sqrt{2b\epsilon}+\sqrt{r+4b\mathbb{E}R_n\{f\in\mathcal{F}:Pf^2\le r\}}\right)\\
&\qquad\times\sqrt{\frac{2\log\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_1)}{n}},\qquad\forall\epsilon>0,
\end{aligned}$$

where in the deduction process we have used Jensen's inequality $\mathbb{E}\sqrt{Y_r}\le\sqrt{\mathbb{E}Y_r}$. The above inequality can be viewed as a quadratic inequality of $\sqrt{\mathbb{E}R_n\{f\in\mathcal{F}:Pf^2\le r\}}$ and a direct calculation yields that

$$\begin{aligned}
\mathbb{E}R_n\{f\in\mathcal{F}:Pf^2\le r\}&\le2\epsilon+\frac{8b\log\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_1)}{n}\\
&\quad+(2\sqrt{2b\epsilon}+\sqrt{r})\sqrt{\frac{2\log\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_1)}{n}}.
\end{aligned}$$

The desired inequality follows by taking the infimum over $\epsilon>0$. □

Remark 4. As compared to the existing results, our approach may admit the following superiorities:

(1) It may happen that the estimation of L_1-norm covering numbers is simpler than that of L_2-norm covering numbers. For example, Krzyżak and Linder [1] only discussed L_1-norm covering numbers for RBF networks. Some other examples include the class of uniformly bounded convex functions, for which Guntuboyina and Sen [34] obtained optimal L_1-norm covering number bounds and indicated that the extension of this result to L_2-norm covering numbers requires more involved arguments. Therefore, our result may be more convenient to use.

(2) As shown in Eq. (18), L_1-norm covering numbers are always smaller than L_2-norm covering numbers. Consequently, our result may yield a tighter bound when $\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_1)$ is much smaller than $\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_2)$.

(3) The deduction presented here is simple, while the analysis based on the entropy integral in Mendelson [22] is more involved. To be precise, Mendelson obtained the following entropy integral by resorting to chaining arguments based on the L_2-metric capacity:

$$\mathbb{E}R_n\{f\in\mathcal{F}:Pf^2\le r\}\le c\mathbb{E}\int_0^{\sqrt{Y_r}}\log^{\frac{1}{2}}\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|_2)\mathrm{d}\epsilon. \tag{21}$$

Notice that the random variable $\sqrt{Y_r}$ appears as the upper limit of the integral in Eq. (21) and the basic inequality available to us is a bound on $\mathbb{E}Y_r$ given by Eq. (20). Consequently, one needs some involved strategy to apply Eq. (20) to estimate the integral in Eq. (21). Mendelson's tricky strategy is to bound the integral in Eq. (21) by a function of Y_r in which the variable Y_r appears in a simpler term. Furthermore, this constructed function turns out to be increasing and concave with respect to Y_r, to which Eq. (20) can be readily applied. Notice that the construction of this function is not easy and requires some additional effort. As a comparison, one can clearly see that the variable $\sqrt{Y_r}$ always occurs as a linear term in our deduction and its expectation can be simply bounded by the inequality $\mathbb{E}\sqrt{Y_r}\le\sqrt{\mathbb{E}Y_r}$.

We are now in a position to present local Rademacher complexity bounds for the shifted loss class (17). The proof, which is given in part A of the appendix, relies on the complexity bounds in Theorem 4 and the L_1-norm covering number bounds given by Krzyżak and Linder [1].

Theorem 5. *If K is of bounded variation and satisfies the condition $\sup_t|K(t)|\le1$, then for any input dimension d, sample size n, complexity index k and $r>0$, the local Rademacher complexity of the shifted loss class $\mathcal{F}_k^*$ satisfies:*

$$\mathbb{E}R_n\{f\in\mathcal{F}_k^*:Pf^2\le r\}\le c\left[\frac{kd^2\log n}{n}+\sqrt{\frac{rkd^2\log n}{n}}\right].$$

V. Generalization performance of radial basis function networks

This section discusses the generalization performance of RBF networks by considering separately the estimation and approximation errors. We first apply local Rademacher complexity bounds in Theorem 5 and a Talagrand-type inequality (Theorem 6) to tackle the estimation error bounds, based on which one can show that the structural risk (10) indeed meets the condition (5). Then the approximation power of RBF networks is treated via classical results in approximation theory. The generalization performance of RBF networks is justified by plugging the obtained estimation and approximation error bounds into Eq. (6).

A. Controlling the estimation error

Our discussion on estimation error bounds is based on Theorem 6 due to Bousquet [18] and Blanchard et al. [30], which shows that if the uniform deviation of the empirical processes indexed by sub-classes can be controlled by a sub-root function ϕ, then the uniform deviation over the whole class can also be dominated by the fixed point of ϕ.

Definition 3 ([22])**.** A function $\phi:[0,\infty)\to[0,\infty)$ is called sub-root if it is nondecreasing and if $r\to\phi(r)/\sqrt{r}$ is non-increasing over $r>0$.

It can be checked that any sub-root function ϕ admits a unique positive number r^* satisfying $\phi(r^*)=r^*$. We will refer to such r^* as the fixed point of ϕ in the remainder [5].

Theorem 6 ([30])**.** *Let $\mathcal{F}$ be a class of measurable, square integrable functions such that $Pf-f\le b,\forall f\in\mathcal{F}$. Assume that the convex hull of $\mathcal{F}$ contains the zero function. Let $w(f)$ be a non-negative functional with Var$(f)\le w(f),\forall f\in\mathcal{F}$. Let ϕ be a sub-root function with unique fixed point r^* such that the following inequality holds:*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}:w(f)\le r}(P-P_n)f\right]\le\phi(r),\quad\forall r\ge r^*. \tag{22}$$

Then, for any $t > 0$ and $M > 1/7$, the following inequality holds with probability at least $1 - e^{-t}$:

$$Pf - P_n f \leq \frac{w(f)}{M} + 50Mr^* + \frac{(M+9b)t}{n},\ \forall f \in \mathcal{F}. \quad (23)$$

The sub-classes in Theorem 6 are defined through a non-negative functional $w(f)$, which will be fixed as the specific choice $w(f) := Pf^2$ in this paper. To estimate the term $M^{-1}w(f)$ in Eq. (23), we need an additional assumption called the Bernstein condition.

Definition 4 (Bernstein condition [18])**.** Let $0 < \alpha \leq 1$ and $B > 0$ be two given constants. We say that $\mathcal{F}$ is an (α, B)-Bernstein class with respect to the probability measure P if

$$Pf^2 \leq B(Pf)^\alpha, \qquad \forall f \in \mathcal{F}. \quad (24)$$

Bernstein condition (24) ensures that variances of functions in $\mathcal{F}$ can be controlled through their expectations, which is essential for us to get improved learning rates via the local Rademacher complexity technique. The intuitive example for extracting such condition is the famous Bernstein inequality, where the variance-expectation relation plays a significant role in deriving sharp bounds [22]. Lemma 13 in part B of the appendix guarantees the Bernstein condition for the shifted loss class (17). The estimation error bounds for the prediction rule $\hat{h}_k$ can be controlled by the following theorem, whose proof is given in part B of the appendix.

Theorem 7. *Let P be a probability measure defined on $\mathcal{Z} := \mathcal{X} \times [-b, b] \subset \mathbb{R}^d \times [-b, b], d \in \mathbb{N}^+, b > 0$, from which the examples $Z_i = (X_i, Y_i), i = 1, 2, \ldots, n$ are independently drawn. Assume that the loss function is $\varphi_p, p > 1$, the kernel K is of bounded variation and satisfies that $\sup_t |K(t)| \leq 1$. Then for the hypothesis space defined as Eq. (8) and any $t > 0$, with probability at least $1 - e^{-t}$ there holds*

$$P\hat{f}_k \leq \beta_p P_n \hat{f}_k + c\left[\left(\frac{kd^2 \log n}{n}\right)^{\frac{1}{2-\alpha_p}} + t\left(\frac{1}{n}\right)^{\frac{1}{2-\alpha_p}}\right], \quad (25)$$

where $\alpha_p = 2/p \wedge 1$ and $\hat{f}_k$ is an element in $\mathcal{F}_k^*$ defined by

$$\hat{f}_k(z) := \varphi_p(\hat{h}_k(x) - y) - \varphi_p(h^*(x) - y),\ k \in \mathbb{N}^+. \quad (26)$$

B. Controlling the approximation error

Our approximation error bounds for RBF networks are based on Theorem 8 due to Wu et al. [21], which implies that for a loss function φ with a Hölder continuous derivative, the term $\mathcal{E}(h) - \mathcal{E}(h^*)$ can be approached by studying the distance between h and h^* under the metric $\|\cdot\|_{L_p(P)}$.

Definition 5. Let $I \subset \mathbb{R}$ be an interval with nonempty interior. A function $\varphi : I \to \mathbb{R}$ is called Hölder continuous with exponent α $(0 < \alpha < 1)$ and constant c_0 on I if

$$|\varphi(y) - \varphi(x)| \leq c_0|y - x|^\alpha, \qquad \forall x, y \in I. \quad (27)$$

Theorem 8 ([21])**.** *Assume that $|y - h(x)| \leq M$ and $|y - h^*(x)| \leq M$ almost surely. If the loss function φ is differentiable on $[-M, M]$ and its derivative is Hölder continuous with exponent α and constant c_0, then we have*

$$\mathcal{E}(h) - \mathcal{E}(h^*) \leq \frac{c_0}{1+\alpha}\|h - h^*\|_{L_{1+\alpha}(P)}^{1+\alpha}.$$

Under the assumption $h^* \in \overline{\mathcal{H}}_b$, we have the following approximation error rates. The definition of $\overline{\mathcal{H}}_b$ can be seen from Eq. (14). The proof is given in part C of the appendix.

Theorem 9. *If the target function h^* belongs to $\overline{\mathcal{H}}_b$ and the kernel K is uniformly bounded in the sense that $\sup_t |K(t)| \leq 1$, then for the loss function $\varphi_p, p > 1$ there holds*

$$\mathcal{E}(h_k^*) - \mathcal{E}(h^*) \leq \frac{pc_{p-1}}{2 \wedge p}\left(\frac{b}{\sqrt{k}}\right)^{2\wedge p}, \quad (28)$$

where $c_{p-1} = 2$ if $p \leq 2$ and $c_{p-1} = (p-1)(2b)^{p-2}$ if $p > 2$.

Remark 5. Eq. (28) is an example of dimension-independent bound since the involved convergence rate does not depend on the input dimension d, which, at first glance, may seem inconsistent with the curse of dimensionality: approximation will become harder as the input dimension increases. However, this is indeed not the case since the information on d is hidden in the assumption that $h^* \in \overline{\mathcal{H}}_b$. To clearly see the role of the dimension here, we consider the special case $K(t) = e^{-t}, A = \sigma^{-1}I$ (I is the identify matrix and $\sigma \in \mathbb{R}^+$). For any $r > d/2, q \in [1, \infty)$, the *Bessel-potential* space $(L^{q,r}, \|\cdot\|_{L^{q,r}})$ is defined as the set of functions f that can be expressed as $f = w * \beta_r$, where $*$ stands for the convolution operator, $w \in L_q(\lambda_0)$ (λ_0 is the Lebesgue measure) and β_r is the r-th Bessel potential with $\hat{\beta}_r(s) = (1+\|s\|^2)^{-r/2}$ being its Fourier transform. Expressing β_r in an integral formula as Eq. (12) in Kainen et al. [11] and applying Theorem 2.4 in Kainen et al. [11] to control the variational norm of any $h^* \in L^{1,r}$, one can show that $h^* \in \overline{\mathcal{H}}_b$ if

$$b \geq \|w\|_{L_1(\lambda_0)} 2^{-d/2}\Gamma(r/2 - d/2)/\Gamma(r/2), \quad (29)$$

where $\Gamma(x) := \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$ is the Gamma function. Therefore, the condition $h^* \in \overline{\mathcal{H}}_b$ indeed hides the information on d, which places an appropriate constraint on the target function to allow for a dimension-independent error rate. Furthermore, restating the regularity condition in other ways would automatically reveal the role of the dimension in the approximation process [11], [35]. For example, by assuming $h^* = w * \beta_r \in L^{1,r}$ with w satisfying Eq. (29), one can recover the following dimension-dependent error rate [11], [36]

$$\mathcal{E}(h_k^*) - \mathcal{E}(h^*) \leq \frac{pc_{p-1}}{2 \wedge p}\left(\frac{\|w\|_{L_1(\lambda_0)} 2^{-d/2}\Gamma(r/2 - d/2)}{\sqrt{k}\Gamma(r/2)}\right)^{2\wedge p}.$$

To let the above inequality be nontrivial, we need to impose the constraint $r > d$ (since $\int_0^\infty t^{x-1}e^{-t}\mathrm{d}t = \infty$ if $x \leq 0$). Since the space $L^{1,r}$ will become more and more constrained as r increases, one needs to place a much stronger smoothness assumption on the target function to attain similar approximation error rates for large d, justifying the curse of dimensionality. For a fixed $c_0 \geq 0$ and the degree $r_d := d + c_0$, the factor

$$k(r_d, d) := \left[2^{-d/2}\frac{\Gamma(r_d/2 - d/2)}{\Gamma(r_d/2)}\right]^{2\wedge p}$$

involving d decays exponentially fast as d increases, showing the hyper-tractability behavior for approximation by RBF networks [11, 12].

Remark 6. It is interesting to describe the class of problems that can be addressed by RBF networks with guaranteed approximation error rates, i.e., to illustrate the class of functions belonging to $\overline{\mathcal{H}}_b$. Using Theorem 8.2 in Girosi and Anzellotti [8] one can show that functions g with the integral representation

$$g(x) = \int_{\mathbb{R}^{d^2+d}} K\left([x-c]^t A[x-c]\right) \lambda(\mathrm{d}c\mathrm{d}A)$$

are indeed members of $\overline{\mathcal{H}}_b$. Here λ is a signed measure on $\mathbb{R}^{d^2+d}$ with variation $\|\lambda\| < b$. For the case $K(t) = e^{-t}, t \in \mathbb{R}^+$ and $A = \sigma^{-1}I, \sigma > 0$, other than the Bessel-potential spaces considered in Remark 5, Niyogi and Girosi [4] indicated that $\overline{\mathcal{H}}_b$ contains the Sobolev space $\mathcal{H}^{2m,1}(2m > d)$ consisting of functions whose derivatives up to order $2m$ are integrable. One can also see here that the assumption $h^* \in \overline{\mathcal{H}}_b$ imposes stronger constraints as d increases.

VI. Simulation study

This section aims to justify the effectiveness of the previous theoretical analysis from an empirical perspective. Specifically, we will consider the application of the structural risk (11) in selecting an appropriate complexity index k and compare its behavior with other model selection methods. Instead of the general RBF networks of the form (7), the networks to our attention here take the specific form

$$h(x) = \sum_{i=1}^{k} w_i e^{-\|x-c_i\|^2} + w_0, \tag{30}$$

which allow us to use the standard function *newrb* in the *Matlab Neural Network Toolbox* to train networks. The notation $\|\cdot\|$ in Eq. (30) means the Euclidean norm. We consider here some specific ℓ_p loss functions with $1 < p < 2$, which are more robust, or equivalently less sensitive to "outliers" (bad observations), than the standard squared loss. Values of p close to one are of great importance for robust neural network regression [1]. Concretely, Darken et al. [37] indicated the superiority of $\ell_{1.2}$ to ℓ_2 since $\ell_{1.2}$ is relatively insensitive to "outliers". We do not consider the squared loss here since Krzyżak and Linder [1] obtained a structural risk similar to ours in this case.

As our purpose is to compare different model selection methods rather than the accurate construction of RBF networks, we consider here a two-stage approximation method to train RBF networks under the general ℓ_p loss, $1 < p < 2$. At the first stage, the centroids $c_1, \ldots, c_k$ are approached by the function *newrb* in Matlab, which is exclusively designed for the squared loss. Once the centroids are derived, the calculation of the coefficients w_i is indeed a L_p regression problem and, perhaps more important, a convex optimization problem. We use *CVX* [38], a package for specifying and solving convex programs, to identify the coefficients w_i at the second stage. The constraint on the coefficients as $\sum_{i=0}^{k} |w_i| \leq b$ is ignored here since our main focus is to study the effect of k on the generalization performance. Also, the parameter b relies on the target function's regularity assumption, which is often unknown to us.

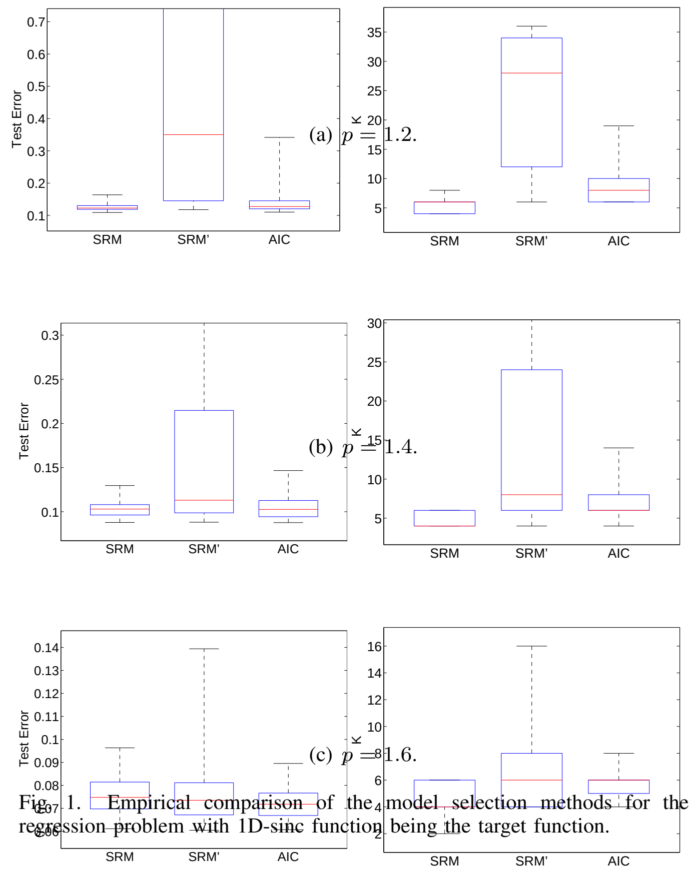


Fig. 1. Empirical comparison of the model selection methods for the regression problem with 1D-sinc function being the target function.

The success of a model selection method heavily relies on a criterion to assess the associated models' quality. Instead of the structural risk (11), we use the following structural risk

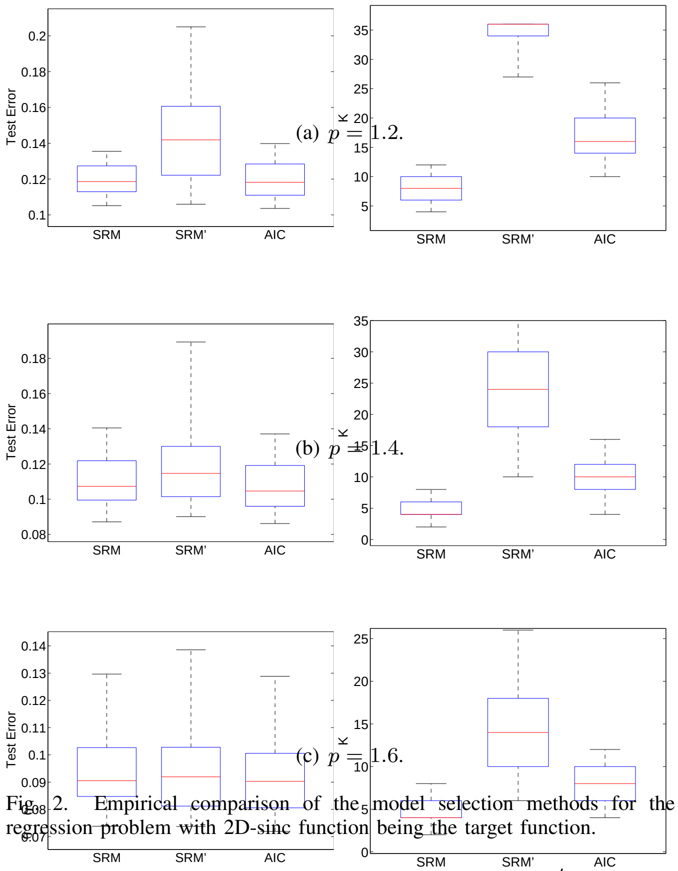
$$\mathrm{SRM}(\hat{h}_k) := \mathcal{E}_{\boldsymbol{z}}(\hat{h}_k) + \lambda \left(\frac{k(d+1)+1}{n} \log n \right)^{\frac{1}{2-\alpha_p}}, \tag{31}$$

where $\lambda > 0$ is a constant. The distinction between Eq. (11) and Eq. (31) consists in two aspects: firstly, the negligible term $2\log k \cdot n^{-\frac{1}{2-\alpha_p}}$ in Eq. (11) is removed here; secondly[2], the term kd^2 in Eq. (11) is replaced by $k(d+1)+1$. Empirical studies imply that $\lambda = \sigma^2$ (σ^2 is the variance of the noise ϵ in Eq. (35)) is an appropriate choice and, in this case, the structural risk (31) reduces to (since $\alpha_p = 1$ if $1 < p < 2$)

$$\mathrm{SRM}(\hat{h}_k) := \mathcal{E}_z(\hat{h}_k) + \frac{k(d+1)+1}{n} \log n \cdot \sigma^2, \tag{32}$$

which coincides with the *Bayesian Information Criterion* (BIC). This fact provides a possible justification of our theoretical discussion as it recovers the well-known BIC proposed from a Bayesian viewpoint. To illustrate the efficiency of the structural risk (32), we perform an empirical comparison between it and two other model selection methods: one based

[2] An intuitive interpretation is that the number of parameters is $k(d+1)+1$ for functions of the form (30), while that for general RBF networks is ckd^2. Indeed, for the specific RBF networks (30), analyzing in a similar way one can show that the term kd^2 in Eq. (11) should be replaced by $k(d+1)+1$.



(a) $p = 1.2$.

(b) $p = 1.4$.

(c) $p = 1.6$.

Fig. 2. Empirical comparison of the model selection methods for the regression problem with 2D-sinc function being the target function.

on the analysis in Krzyżak and Linder [1] ($\mathrm{SRM}^{'}$) and one based on *Akaike Information Criterion* (AIC) [39]:

$$\mathrm{SRM}^{'}(\hat{h}_k) = \mathcal{E}_{\boldsymbol{z}}(\hat{h}_k) + \sqrt{\frac{k(d+1)+1}{n}\log n\sigma^2}, \qquad (33)$$

$$\mathrm{AIC}(\hat{h}_k) = \mathcal{E}_{\boldsymbol{z}}(\hat{h}_k) + \frac{2k(d+1)+2}{n}\sigma^2. \qquad (34)$$

Note that the structural risk (33) is derived from Eq. (12) by replacing kd^2 and c with $k(d+1)+1$ and σ^2, respectively.

The empirical comparison is performed in a controlled manner, for which the data is independently generated from

$$y = f_\rho(x) + \epsilon, \qquad (35)$$

where x follows the uniform distribution over $\mathcal{X}$ and ϵ follows the normal distribution with expectation 0 and variance σ^2. We consider here two specific regression problems, where the corresponding target functions are

$$\text{1D-sinc function:} \quad f_\rho(x) = x^{-1}\sin x \qquad x \in [-10, 10],$$

$$\text{2D-sinc function:} \quad f_\rho(x) = \frac{\sin\sqrt{x_1^2 + x_2^2}}{\sqrt{x_1^2 + x_2^2}} \qquad \boldsymbol{x} \in [-5, 5]^2.$$

We choose the noise variance σ^2 in a way to let the *Signal-Noise-Ratio* (SNR) equal to 4, where SNR is defined as the ratio of the variance of the true output value $f_\rho(x)$ to the variance of the noise ϵ [39]. For simplicity, we assume that σ^2 is accessible to us and thus all the criteria (32), (33), (34) can be directly computed from the data.

For each regression problem, we generate a training set by independently drawing n points from Eq. (35). Then the complexity index k is ranged over the specified set $\{2, 4, 6, \ldots, 36\}$. For each temporarily fixed k, the associated RBF network $\hat{h}_k$ is established by our two-stage approximation method, resulting in a sequence of candidate models. For each model selection method, the quality of the candidate models is assessed by the corresponding criterion (SRM, $\mathrm{SRM}^{'}$, AIC), and the one with the best quality is identified as the ultimate model. The generalization performance of the model chosen by a model selection method is measured through the test error:

$$\mathcal{E}_{\text{test}}(h_n) := \frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}} |h_n(x_i^{'}) - y_i^{'}|^p,$$

where $((x_1^{'}, y_1^{'}), \ldots, (x_{n_{\text{test}}}^{'}, y_{n_{\text{test}}}^{'}))$ is the test sample independently drawn from Eq. (35). Now, the test error and the complexity index k of the chosen model are recorded. We always set $n_{\text{test}} = 500$.

The above experimental procedure is repeated 100 times, with each trial an independent random realization of $n = 50$ training points. The empirical distribution of these test errors, as well as the corresponding complexity indices, are displayed via the standard box plot, with marks at 95-th, 75-th, 50-th, 25-th and 5-th percentile of the empirical distribution.

Fig. 1 exhibits the relative behavior of different model selection methods under the 1D-sinc function and different loss functions ($p = 1.2, 1.4, 1.6$), while Fig. 2 displays their performance for the 2D-sinc function. Both SRM and AIC work well for all the regression problems and all considered ℓ_p loss functions. As a comparison, $\mathrm{SRM}^{'}$ performs relatively poorly in the case $p = 1.2$ and $p = 1.4$. It can also be clearly seen that SRM favors the simplest model, which is mostly consistent with the principle of Occam's razor: among all hypotheses consistent with the facts, choose the simplest.

VII. Conclusions

This paper studies the generalization performance of RBF networks under the SRM principle and general loss functions. We propose a general local Rademacher complexity bound involving the L_1-metric capacity rather than the traditional L_2-metric capacity. We then apply this general result to the RBF network setting to derive substantially improved estimation error bounds. Effective approximation error bounds are also presented by carefully investigating the Hölder continuity of the associated loss function's derivative. It is shown that the RBF network minimizing an appropriate structural risk attains a significantly faster learning rate when compared to the existing results. We also perform an empirical study to justify the application of our structural risk in model selection.

Acknowledgment

The authors would like to thank the editor and the anonymous referees for their insightful and constructive comments, which greatly improved the quality of the paper.

Appendix

A. Proofs on local Rademacher complexity bounds

Lemma 10 ([18])**.** *If $\mathcal{F}$ is a finite class with cardinality N, then for any sample size n and $r > 0$ there holds:*

$$\mathbb{E}_\sigma R_n\{f \in \mathcal{F} : P_n f^2 \le r\} \le \sqrt{\frac{2r \log N}{n}}.$$

Lemma 11. *Let n be the sample size, r and b two positive numbers. For any function class $\mathcal{F}$ with $\sup_{f\in\mathcal{F}} \|f\|_\infty \le b$, we have the following complexity bounds:*

$$\mathbb{E}_\sigma R_n\{f \in \mathcal{F} : P_n f^2 \le r\} \le \inf_{\epsilon>0}\left[\epsilon + (\sqrt{2b\epsilon}+\sqrt{r})\sqrt{\frac{2\log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_{L_1(P_n)})}{n}}\right].$$

Proof. We temporarily fix any parameter $\epsilon > 0$. Let $\mathcal{F}^\triangle$ be a minimal ϵ-cover of the class $\mathcal{F}$ with respect to the norm $\|\cdot\|_{L_1(P_n)}$. Denote by

$$\mathcal{F}_r^\triangle := \{f \in \mathcal{F}^\triangle : \|f\|_{L_2(P_n)} \le \sqrt{2b\epsilon} + \sqrt{r}\}$$

a subset of $\mathcal{F}^\triangle$. For any $f \in \mathcal{F}$, we define $f^\triangle$ as the closest element to f in $\mathcal{F}^\triangle$:

$$f^\triangle := \operatorname*{argmin}_{g\in\mathcal{F}^\triangle} \|f-g\|_{L_1(P_n)}.$$

For any f, g with $\|f\|_\infty \le b, \|g\|_\infty \le b$, we know that

$$\|f-g\|^2_{L_2(P_n)} = \int |f-g|^2 \mathrm{d}P_n \le 2b\|f-g\|_{L_1(P_n)}.$$

Without loss of generality, one can always assume that the set $\mathcal{F}^\triangle$ is also uniformly bounded by b. Now, for any element $f \in \mathcal{F}$ with $P_n f^2 \le r$, it follows from the triangle inequality and the definition of $f^\triangle$ that

$$\begin{aligned}\|f^\triangle\|_{L_2(P_n)} &\le \|f^\triangle - f\|_{L_2(P_n)} + \|f\|_{L_2(P_n)} \\ &\le \sqrt{2b\|f^\triangle - f\|_{L_1(P_n)}} + \|f\|_{L_2(P_n)} \\ &\le \sqrt{2b\epsilon} + \sqrt{r}.\end{aligned} \tag{36}$$

That is, for any $f \in \mathcal{F}$ with $P_n f^2 \le r$ we have $f^\triangle \in \mathcal{F}_r^\triangle$. By the definition of Rademacher complexity, we derive that

$$\begin{aligned}&\mathbb{E}_\sigma R_n\{f \in \mathcal{F} : P_n f^2 \le r\} \\ &= \mathbb{E}_\sigma \sup_{f\in\mathcal{F}:P_nf^2\le r}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i\big(f(X_i) - f^\triangle(X_i)\big) + \frac{1}{n}\sum_{i=1}^n \sigma_i f^\triangle(X_i)\right] \\ &\le \sup_{f\in\mathcal{F}:P_nf^2\le r} \|f - f^\triangle\|_{L_1(P_n)} + \mathbb{E}_\sigma R_n \mathcal{F}_r^\triangle \\ &\le \epsilon + (\sqrt{2b\epsilon}+\sqrt{r})\sqrt{\frac{2\log|\mathcal{F}_r^\triangle|}{n}},\end{aligned}$$

where the last step follows from Lemma 10. The proof is complete since the above inequality holds for any $\epsilon > 0$. □

Proof of Theorem 5. Let V be the total variation of K. For the hypothesis space (8), Krzyżak and Linder [1, Lemma 4] derived the following covering number bounds:

$$\begin{aligned}\mathcal{N}(\epsilon, \mathcal{H}_k, \|\cdot\|_1) &\le \big(e^2(d^2+d+3)\big)^{2(k+1)}\left(\frac{2e(b+\epsilon)}{\epsilon}\right)^{k+1} \\ &\quad\times\left(\frac{Ve(b+\epsilon)}{\epsilon}\right)^{2(k+1)(d^2+d+2)} \\ &\le \left(e^2(d^2+d+3)\sqrt{2e}(Ve)^{d^2+d+2}\right)^{2(k+1)} \\ &\quad\times\left(\frac{2b}{\epsilon}\right)^{(k+1)(2d^2+2d+5)} \qquad \text{if } \epsilon \le b.\end{aligned}$$

Using the above inequality and the structural result [1], [22]

$$\mathcal{N}(\epsilon, \mathcal{F}_k^*, \|\cdot\|_1) = \mathcal{N}(\epsilon, \mathcal{F}_k, \|\cdot\|_1) \le \mathcal{N}(\epsilon/(p(2b)^{p-1}), \mathcal{H}_k, \|\cdot\|_1),$$

one can show that

$$\begin{aligned}\mathcal{N}(\epsilon, \mathcal{F}_k^*, \|\cdot\|_1) &\le \left(e^2(d^2+d+3)\sqrt{2e}(Ve)^{d^2+d+2}\right)^{2(k+1)} \\ &\quad\times\left(\frac{p(2b)^p}{\epsilon}\right)^{(k+1)(2d^2+2d+5)} \\ &= \Bigg(\underbrace{\sqrt{2e^5}(d^2+d+3)(Ve)^{d^2+d+2}(p2^pb^p)^{(2d^2+2d+5)/2}}_{:=A}\Bigg)^{2(k+1)} \\ &\quad\times\epsilon^{-(k+1)(2d^2+2d+5)}.\end{aligned}$$

The above inequality can be rewritten as follows:

$$\begin{aligned}\log\mathcal{N}(\epsilon, \mathcal{F}_k^*, \|\cdot\|_1) &\le 2(k+1)\log A \\ &\quad+ (k+1)(2d^2+2d+5)\log(1/\epsilon).\end{aligned}$$

It can be directly checked that the class $\mathcal{F}_k^*$ is uniformly bounded by $(2b)^p$. Consequently, one can apply Theorem 4 here to derive the following inequality for any $0 < \epsilon < b$

$$\begin{aligned}\mathbb{E}R_n\{f \in \mathcal{F}_k^* : Pf^2 \le r\} &\le 2\epsilon + \left(2\sqrt{2(2b)^p\epsilon}+\sqrt{r}\right) \\ &\quad\times\sqrt{\frac{k+1}{n}}\sqrt{4\log A + 2(2d^2+2d+5)\log(1/\epsilon)} \\ &\quad+\frac{8(2b)^p(k+1)}{n}\big[2\log A + (2d^2+2d+5)\log(1/\epsilon)\big].\end{aligned}$$

Taking the assignment $\epsilon = n^{-1}$ in the above inequality (we assume that $n^{-1} < b$ for brevity), we have

$$\mathbb{E}R_n\{f \in \mathcal{F}_k^* : Pf^2 \le r\} \le c\left[\frac{kd^2\log n}{n} + \sqrt{\frac{rkd^2\log n}{n}}\right].$$

□

B. Proofs on estimation error bounds

Our proof on estimation error bounds heavily relies on the variance-expectation relation for functions in the class (17). For this purpose we first recall the following lemma due to Bartlett et al. [40] and Mendelson [41], which shows that the shifted loss class (17) is indeed an (α, B)-Bernstein class, provided that the involved hypothesis space $\mathcal{H}_k$ is convex.

Lemma 12 ([41, Theorem 6.1]). *Suppose that the hypothesis space $\mathcal{H}$ is convex and the loss function $\varphi_p(t)$ satisfies the condition $|\varphi_p(h(x)-y)| \le M, \forall h \in \mathcal{H}, (x,y) \in \mathcal{Z}$ for some positive constant M. Then the associated shifted loss class*

$$\mathcal{F} = \{\varphi_p(h(x)-y) - \varphi_p(h^*_{\mathcal{H}}(x)-y) : h \in \mathcal{H}\}$$

is an (α_p, B)-Bernstein class, where $\alpha_p = 2/p \wedge 1, B$ is a constant depending on p and M and $h^*_{\mathcal{H}} := \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{E}(h)$.

However, $\mathcal{H}_k$ does not satisfy the convexity assumption in our specific problem and therefore Lemma 12 cannot be directly applied. Fortunately, with a little additional work, we can obtain Lemma 13 at our rescue.

Lemma 13. *Suppose that the response variable Y takes values in the region $[-b, b]$ with probability 1. Then, the shifted loss class $\mathcal{F}^*_k$ is an (α_p, B)-Bernstein class, where $\alpha_p = 1 \wedge 2/p$ and B is a constant depending on p and b.*

Proof. Introduce the auxiliary function class

$$\mathcal{H}_b = \{h \text{ is a measurable function defined on } \mathcal{X} : \|h\|_\infty \le b\}.$$

The convexity of $\mathcal{H}_b$ follows from the above definition. For any function h, it can be verified that $h_b := \max(-b, \min(b, h))$ is a better function for modeling the data in the sense $\mathcal{E}(h_b) \le \mathcal{E}(h)$. Indeed, one can even show that the inequality $|h_b(x) - y| \le |h(x)-y|$ holds for any y with $|y| \le b$. Consequently, the target function h^* lies in $\mathcal{H}_b$ and thus one can apply Lemma 12 to show that

$$\bar{\mathcal{F}}^*_b := \{\varphi_p(h(x)-y) - \varphi_p(h^*(x)-y) : h \in \mathcal{H}_b\}$$

is an (α_p, B)-Bernstein class for some constant B. As a subset of $\bar{\mathcal{F}}^*_b$, $\mathcal{F}^*_k$ is also an (α_p, B)-Bernstein class. ☐

With these preparations, we can now prove Theorem 7 on estimation error bounds. Since the exponent α in Eq. (24) may vary when p takes different values, we consider two cases ($1 < p \le 2$ and $p > 2$) to proceed with our proof.

Proof of Theorem 7. According to Lemma A.5 in [5], the deviation of empirical means from their expectations can be controlled by the associated Rademacher complexity. Therefore, we can obtain from Theorem 5 that

$$\begin{aligned}\mathbb{E}\left[\sup_{f \in \mathcal{F}^*_k : Pf^2 \le r} (P - P_n) f\right] &\le 2\mathbb{E}R_n\{f \in \mathcal{F}^*_k : Pf^2 \le r\} \\ &\le c\left[\frac{kd^2 \log n}{n} + \sqrt{\frac{rkd^2 \log n}{n}}\right].\end{aligned}$$

Introduce the sub-root function

$$\phi(r) := c\left[\frac{kd^2 \log n}{n} + \sqrt{\frac{rkd^2 \log n}{n}}\right].$$

The fixed point r^* of $\phi(r)$ can be calculated by solving a quadratic function, which can be further bounded by

$$r^* \le \frac{ckd^2 \log n}{n}. \tag{37}$$

Furthermore, functions in $\mathcal{F}^*_k$ always satisfy the inequalities

$$Pf - f \le 2(2b)^p \quad \text{and} \quad \operatorname{Var}(f) \le Pf^2, \qquad \forall f \in \mathcal{F}^*_k.$$

Applying Theorem 6 with $w(f) = Pf^2$ and $\mathcal{F} = \mathcal{F}^*_k$, with probability at least $1 - e^{-t}$ there holds:

$$P\hat{f}_k \le P_n \hat{f}_k + M^{-1} P\hat{f}_k^2 + \frac{50Mckd^2 \log n}{n} + \frac{(M + 18(2b)^p)t}{n}, \tag{38}$$

where $\hat{f}_k$ is defined by Eq. (26). Now we can continue our proof by distinguishing two cases according to the value of p:

CASE $1 < p \le 2$. In this case, Lemma 13 guarantees the existence of B satisfying $Pf^2 \le BPf, \forall f \in \mathcal{F}^*_k$ and therefore Eq. (38) reduces to

$$P\hat{f}_k \le P_n \hat{f}_k + \frac{BP\hat{f}_k}{M} + \frac{50Mckd^2 \log n}{n} + \frac{(M + 18(2b)^p)t}{n}.$$

Since the above inequality holds for any $M > 1/7$, one can take the assignment $M = 2B$ to give (we assume $B > 1/14$)

$$P\hat{f}_k \le 2P_n \hat{f}_k + \frac{200Bckd^2 \log n}{n} + \frac{2(2B + 18(2b)^p)t}{n}. \tag{39}$$

CASE $p > 2$. For such p, Lemma 13 implies that the inequality $Pf^2 \le B(Pf)^{2/p}$ holds for some $B > 0$ and any $f \in \mathcal{F}^*_k$. Now, it follows directly from Eq. (38) that

$$\begin{aligned}P\hat{f}_k \le\, & P_n \hat{f}_k + \frac{B}{M}(P\hat{f}_k)^{2/p} \\ & + \frac{50Mckd^2 \log n}{n} + \frac{(M + 18(2b)^p)t}{n} \\ \le\, & P_n \hat{f}_k + \frac{2}{p}\left[(P\hat{f}_k)^{2/p}\right]^{p/2} + \left(1 - \frac{2}{p}\right)\left(\frac{B}{M}\right)^{p/(p-2)} \\ & + \frac{50Mckd^2 \log n}{n} + \frac{(M + 18(2b)^p)t}{n},\end{aligned} \tag{40}$$

where we have used the Hölder inequality [23]

$$p^{-1}a^p + q^{-1}b^q \ge ab, \quad \forall\, p^{-1} + q^{-1} = 1, a, b, p, q > 0.$$

Eq. (40) can be reformulated as follows

$$\begin{aligned}P\hat{f}_k \le\, & \frac{p}{p-2} P_n \hat{f}_k + \left(\frac{B}{M}\right)^{p/(p-2)} \\ & + \frac{50pMckd^2 \log n}{n(p-2)} + \frac{p(M + 18(2b)^p)t}{n(p-2)}.\end{aligned}$$

Plugging $M = (kn^{-1}d^2 \log n)^{(p-2)/(2-2p)}$ into the above inequality, we have

$$\begin{aligned}P\hat{f}_k \le\, & \frac{p}{p-2} P_n \hat{f}_k + \left(B^{\frac{p}{p-2}} + \frac{50pc}{p-2}\right)\left(\frac{kd^2 \log n}{n}\right)^{\frac{p}{2p-2}} \\ & + \frac{18p(2b)^p t}{n(p-2)} + \frac{pt}{p-2}\left(\frac{1}{kd^2 \log n}\right)^{\frac{p-2}{2p-2}}\left(\frac{1}{n}\right)^{\frac{p}{2p-2}}.\end{aligned} \tag{41}$$

Eq. (39) and Eq. (41) can be written in a compact form as Eq. (25), where c is a constant independent of n, k and d. ☐

C. Proofs on the approximation error bounds

To apply Theorem 8 in our context we need to check the Hölder continuity of the signed power function $\psi(x) := \operatorname{sgn}(x)|x|^\alpha$, which is justified by the following lemma.

Lemma 14. *The signed power function* $\psi(x) := \operatorname{sgn}(x)|x|^\alpha, \alpha > 0$ *defined on* $[-M, M]$ *is Hölder continuous with exponent* $1 \wedge \alpha$ *and constant* c_α*, where* $c_\alpha = 2$ *if* $0 < \alpha \le 1$ *and* $c_\alpha = \alpha M^{\alpha-1}$ *if* $\alpha > 1$.

Proof. We consider two cases according to the value of α.

CASE $0 < \alpha \le 1$. For such α, it can be directly checked that the power function $\tilde{\psi}(x) := x^\alpha$ defined on $[0, \infty)$ satisfies the following inequality

$$(x+y)^\alpha \le x^\alpha + y^\alpha \le 2(x+y)^\alpha, \quad \forall x, y \in [0, \infty). \quad (42)$$

Indeed, the first inequality follows from the sub-additive property of $\tilde{\psi}(x)$, whereas the second inequality is due to the non-negativity of x, y.

For numbers x, y with $x \cdot y \ge 0$, Eq. (42) implies that

$$|x|^\alpha = |x - y + y|^\alpha \le |x-y|^\alpha + |y|^\alpha,$$
$$|y|^\alpha = |y - x + x|^\alpha \le |x-y|^\alpha + |x|^\alpha.$$

These two basic inequalities yield that

$$|\operatorname{sgn}(x)|x|^\alpha - \operatorname{sgn}(y)|y|^\alpha| \le |x-y|^\alpha.$$

For numbers x, y with $x \cdot y < 0$, the desired inequality

$$|\operatorname{sgn}(x)|x|^\alpha - \operatorname{sgn}(y)|y|^\alpha| \le 2|x-y|^\alpha$$

is equivalent to $|x|^\alpha + |y|^\alpha \le 2\left(|x| + |y|\right)^\alpha$ (note that $|x-y| = |x| + |y|$), which follows from the right-hand side of Eq. (42).

CASE $\alpha > 1$. In this case, it can be verified that $\psi(x) = \operatorname{sgn}(x)|x|^\alpha$ is differentiable and the derivative is uniformly bounded in that $|\psi^{'}(x)| \le \alpha M^{\alpha-1}, \forall x \in [-M, M]$. Consequently, the Hölder continuity of $\psi(x)$ can be established by

$$|\psi(x) - \psi(y)| = \Big| \int_y^x \psi^{'}(t)\mathrm{d}t \Big| \le \alpha M^{\alpha-1}|x-y|.$$

□

Proof of Theorem 9. For the target function h^* in $\overline{\mathcal{H}}_b$, the monotonicity of the norm $\|\cdot\|_{L_p(P_X)}$ with respect to p and Lemma 1 in Barron [36] guarantee the existence of a function $\tilde{h}_k \in \mathcal{H}_k$ such that

$$\|\tilde{h}_k - h^*\|_{L_{2\wedge p}(P_X)} \le \|\tilde{h}_k - h^*\|_{L_2(P_X)} \le b\sqrt{1/k}. \quad (43)$$

Note that $|y - h(x)| \le 2b, \forall h \in \mathcal{H}_k$ and $|y - h^*(x)| \le 2b$ hold almost surely. Lemma 14 implies that $\varphi_p^{'}(x) = \operatorname{sgn}(x) \cdot p|x|^{p-1}$ is Hölder continuous with exponent $1 \wedge (p-1)$ and constant pc_{p-1}. Consequently, it follows from Theorem 8 that

$$\mathcal{E}(\tilde{h}_k) - \mathcal{E}(h^*) \le \frac{pc_{p-1}}{2 \wedge p} \|\tilde{h}_k - h^*\|^{2\wedge p}_{L_{2\wedge p}(P_X)}, \quad (44)$$

which, coupled with h_k^*'s definition and Eq. (43), yields that

$$\mathcal{E}(h_k^*) - \mathcal{E}(h^*) \le \mathcal{E}(\tilde{h}_k) - \mathcal{E}(h^*) \le \frac{pc_{p-1}}{2 \wedge p} \left(\frac{b}{\sqrt{k}}\right)^{2\wedge p}.$$

□

D. Proofs on the generalization error bounds

Proof of Theorem 2. Theorem 7 implies that the following inequality holds with probability at least $1 - e^{-t}$

$$\mathcal{E}(\hat{h}_k) \le \mathcal{E}(h^*) + \beta_p \left[\mathcal{E}_{\boldsymbol{z}}(\hat{h}_k) - \mathcal{E}_{\boldsymbol{z}}(h^*)\right] + c\left[\left(\frac{kd^2 \log n}{n}\right)^{\frac{1}{2-\alpha_p}} + t\left(\frac{1}{n}\right)^{\frac{1}{2-\alpha_p}}\right].$$

Consequently, for the estimate $L_{n,k}$ defined as

$$L_{n,k} := \mathcal{E}(h^*) + \beta_p \left[\mathcal{E}_{\boldsymbol{z}}(\hat{h}_k) - \mathcal{E}_{\boldsymbol{z}}(h^*)\right] + c(kd^2 n^{-1}\log n)^{\frac{1}{2-\alpha_p}},$$

the inequality (5) holds with $\kappa = 1$ and $\gamma = c^{-1}n^{1/(2-\alpha_p)}$. Now the term $L_{n,k} - \mathcal{E}_{\boldsymbol{z}}(\hat{h}_k)$ can be upper bounded by

$$(\beta_p - 1)\mathcal{E}_{\boldsymbol{z}}(\hat{h}_k) + \mathcal{E}(h^*) - \beta_p\mathcal{E}_{\boldsymbol{z}}(h^*) + c(kd^2n^{-1}\log n)^{\frac{1}{2-\alpha_p}}.$$

Taking the expectation on both sides and using the ERM property $\mathcal{E}_{\boldsymbol{z}}(\hat{h}_k) \le \mathcal{E}_{\boldsymbol{z}}(h_k^*)$, we get

$$\mathbb{E}\left[L_{n,k} - \mathcal{E}_{\boldsymbol{z}}(\hat{h}_k)\right] \le (\beta_p - 1)\left(\mathcal{E}(h_k^*) - \mathcal{E}(h^*)\right) + c(kd^2n^{-1}\log n)^{\frac{1}{2-\alpha_p}}.$$

It can be directly verified that the structural risk defined by Eq. (10) is indeed $L_{n,k} + 2\gamma^{-1}\log k$. Plugging the above inequality into Eq. (6) yields the following result

$$\mathbb{E}\mathcal{E}(h_n) - \mathcal{E}(h^*) \le \min_k \Big[\beta_p\left(\mathcal{E}(h_k^*) - \mathcal{E}(h^*)\right) + c(kd^2n^{-1}\log n)^{\frac{1}{2-\alpha_p}} + (2\log k + \log(2e))cn^{-\frac{1}{2-\alpha_p}}\Big].$$

□

Proof of Corollary 3. For the case $1 < p \le 2$, we can derive from Theorem 9 and Theorem 2 that

$$\mathbb{E}\mathcal{E}(h_n) - \mathcal{E}(h^*) \le \min_k \left[\beta_p ck^{-\frac{p}{2}} + ckd^2n^{-1}\log n\right] \le c\left(\frac{d^2\log n}{n}\right)^{\frac{p}{p+2}},$$

where in the second inequality we simply take the choice $k = (d^2n^{-1}\log n)^{-2/(p+2)}$.

The case $p > 2$ can be analogously addressed by taking $k = (d^2n^{-1}\log n)^{p/(2-3p)}$ in the deduction:

$$\mathbb{E}\mathcal{E}(h_n) - \mathcal{E}(h^*) \le \min_k \left[\beta_p ck^{-1} + c(kd^2n^{-1}\log n)^{\frac{p}{2p-2}}\right] \le c(d^2n^{-1}\log n)^{\frac{p}{3p-2}}.$$

□

REFERENCES

[1] A. Krzyżak and T. Linder, “Radial basis function networks and complexity regularization in function learning,” *IEEE Trans. Neural Netw.*, vol. 9, no. 2, pp. 247–256, 1998.

[2] A. Krzyżak and D. Schafer, “Nonparametric regression estimation by normalized radial basis function networks,” *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 1003–1010, 2005.

[3] A. Barron, “Complexity regularization with application to artificial neural networks,” in *Nonparametric functional estimation and related topics*, G. Roussas, Ed. Boston, MA and Dordrecht: Kluwer Academic Publishers, 1990, pp. 561–576.

[4] P. Niyogi and F. Girosi, "Generalization bounds for function approximation from scattered noisy data," *Adv. Comput. Math.*, vol. 10, no. 1, pp. 51–80, 1999.

[5] P. Bartlett, O. Bousquet, and S. Mendelson, "Local Rademacher complexities," *Ann. Stat.*, vol. 33, no. 4, pp. 1497–1537, 2005.

[6] B. Zou, L.-Q. Li, and Z.-B. Xu, "The generalization performance of ERM algorithm with strongly mixing observations," *Mach. Learn.*, vol. 75, no. 3, pp. 275–295, 2009.

[7] J. Park and I. W. Sandberg, "Universal approximation using radial-basis-function networks," *Neural Comput.*, vol. 3, no. 2, pp. 246–257, 1991.

[8] F. Girosi and G. Anzellotti, "Rates of convergence for radial basis functions and neural networks," in *Artificial Neural Networks for Speech and Vision*, R. Mammone, Ed. London: Chapman and Hall, 1993, pp. 97–113.

[9] F. Girosi, "Approximation error bounds that use VC-bounds," in *Proc. International Conference on Artificial Neural Networks*, F. Fogelman-Soulie and P. Gallinari, Eds., vol. 1, Paris, 1995, pp. 295–302.

[10] G. Gnecco and M. Sanguineti, "Approximation error bounds via rademacher's complexity," *Appl. Math. Sci.*, vol. 2, no. 1-4, pp. 153–176, 2008.

[11] P. C. Kainen, V. Kůrková, and M. Sanguineti, "Complexity of Gaussian-radial-basis networks approximating smooth functions," *J. Complex.*, vol. 25, no. 1, pp. 63–74, 2009.

[12] ——, "Dependence of computational models on input dimension: Tractability of approximation and optimization tasks," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1203–1214, 2012.

[13] M. Anthony and P. Bartlett, *Neural Network Learning: Theoretical Foundations*. New York: Cambridge Univ. Press, 2009.

[14] P. Niyogi and F. Girosi, "On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions," *Neural Comput.*, vol. 8, no. 4, pp. 819–842, 1996.

[15] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Inform. Comput.*, vol. 100, no. 1, pp. 78–150, 1992.

[16] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer-Verlag, 2002.

[17] A. Barron, "Approximation and estimation bounds for artificial neural networks," *Mach. Learn.*, vol. 14, no. 1, pp. 115–133, 1994.

[18] O. Bousquet, "Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms," Ph.D. dissertation, Ecole Polytechnique, 2002.

[19] P. Massart, "Some applications of concentration inequalities to statistics," *Annales de la faculté des sciences de Toulouse*, vol. 9, no. 2, pp. 245–303, 2000.

[20] V. Koltchinskii and D. Panchenko, "Rademacher processes and bounding the risk of function learning," in *Hign Dimensional Probability II*, E. Giné, D. Mason, and J. Wellner, Eds. Boston: Birkhäuser, 2000, pp. 443–458.

[21] Q. Wu, Y.-M. Ying, and D.-X. Zhou, "Learning theory: from regression to classification," in *Topics in Multivariate Approximation and interpolation*, K. Jetter, M. Buhmann, W. Haussmann, R. Schaback, and J. Stoeckler, Eds. Amsterdam: Elsevier, 2006, pp. 257–290.

[22] S. Mendelson, "A few notes on statistical learning theory," in *Advanced Lectures on Machine Learning. Lect. Notes Comput. Sci. 2600*, S. Mendelson and A. Smola, Eds. Berlin: Springer-Verlag, 2003, pp. 1–40.

[23] F. Cucker and D.-X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge: Cambridge Univ. Press, 2007.

[24] P. Bartlett, S. Boucheron, and G. Lugosi, "Model selection and error estimation," *Mach. Learn.*, vol. 48, no. 1, pp. 85–113, 2002.

[25] B. Zou, L.-Q. Li, Z.-B. Xu, T. Luo, and Y.-Y. Tang, "Generalization performance of Fisher linear discriminant based on Markov sampling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 2, pp. 288–300, 2013.

[26] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 2000.

[27] H.-Y. Wang, Q.-W. Xiao, and D.-X. Zhou, "An approximation theory approach to learning with l_1 regularization," *J. Approx. Theory*, vol. 167, pp. 240–258, 2013.

[28] G. Lugosi and A. Nobel, "Adaptive model selection using empirical complexities," *Ann. Stat.*, vol. 27, no. 6, pp. 1830–1864, 1999.

[29] Y.-L. Xu, D.-R. Chen, H.-X. Li, and L. Liu, "Least square regularized regression in sum space," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 4, pp. 635–646, 2013.

[30] G. Blanchard, O. Bousquet, and P. Massart, "Statistical performance of support vector machines," *Ann. Stat.*, vol. 36, no. 2, pp. 489–531, 2008.

[31] V. Koltchinskii, "Rademacher penalties and structural risk minimization," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1902–1914, 2001.

[32] S. Mendelson, "On the performance of kernel classes," *J. Mach. Learn. Res.*, vol. 4, pp. 759–771, 2003.

[33] R. Dudley, "The sizes of compact subsets of Hilbert space and continuity of Gaussian processes," *J. Funct. Anal*, vol. 1, no. 3, pp. 290–330, 1967.

[34] A. Guntuboyina and B. Sen, "l_1 covering numbers for uniformly bounded convex functions," *J. Mach. Learn. Res.:Workshop and Conference Proceedings*, vol. 23, pp. 12.1–12.13, 2012.

[35] H. N. Mhaskar, "On the tractability of multivariate integration and approximation by neural networks," *J. Complex.*, vol. 20, no. 4, pp. 561–590, 2004.

[36] A. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[37] C. Darken, M. Donahue, L. Gurvits, and E. Sontag, "Rate of approximation results motivated by robust neural network learning," in *Proceedings of the 6th Annual Conference on Computational Learning Theory*, ser. COLT '93, L. Pitt, Ed. Santa Cruz, CA: ACM, 1993, pp. 303–309.

[38] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.0 beta," http://cvxr.com/cvx, Sep. 2013.

[39] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.

[40] P. Bartlett, M. Jordan, and J. McAuliffe, "Convexity, classification, and risk bounds," *J. Am. Stat. Assoc.*, vol. 101, no. 473, pp. 138–156, 2006.

[41] S. Mendelson, "Geometric parameters in learning theory," in *Geometric Aspects of Functional Analysis. Lecture Notes in Mathematics. 1850*. Springer-Verlag, 2004, pp. 193–235.

Yunwen Lei received the B.S. degree from the College of Mathematics and Econometrics, Hunan university, Changsha, China, in 2008.

He is currently pursuing his Ph.D. degree at State Key Laboratory of Software Engineering, Wuhan University, Wuhan, China. His main research interests include machine learning, statistical learning theory, basic theory of evolutionary computation and optimization theory.

Lixin Ding received the Ph.D. degree from the State Key Laboratory of Software Engineering (SKLSE), Wuhan University, Wuhan, China, in 1998.

He is currently a Professor at the SKLSE, Wuhan University. He has published more than 60 research articles in domestic and foreign academic journals, such as IEEE Transactions on Communications, IEEE Transactions on Engineering Management, Evolutionary Computation, Neural Computation, Neural Networks, Science China: Information Sciences etc. His research interests include intelligence computation, intelligent information processing and machine learning.

Wensheng Zhang received the Ph.D. degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences (CAS), in 2000.

He joined the Institute of Software, CAS, in 2001. He is a Professor of Machine Learning and Data Mining and the Director of Research and Development Department, Institute of Automation, CAS. He has published over 32 papers in the area of Modeling Complex Systems, Statistical Machine Learning and Data Mining. His research interests include computer vision, pattern recognition and artificial intelligence.