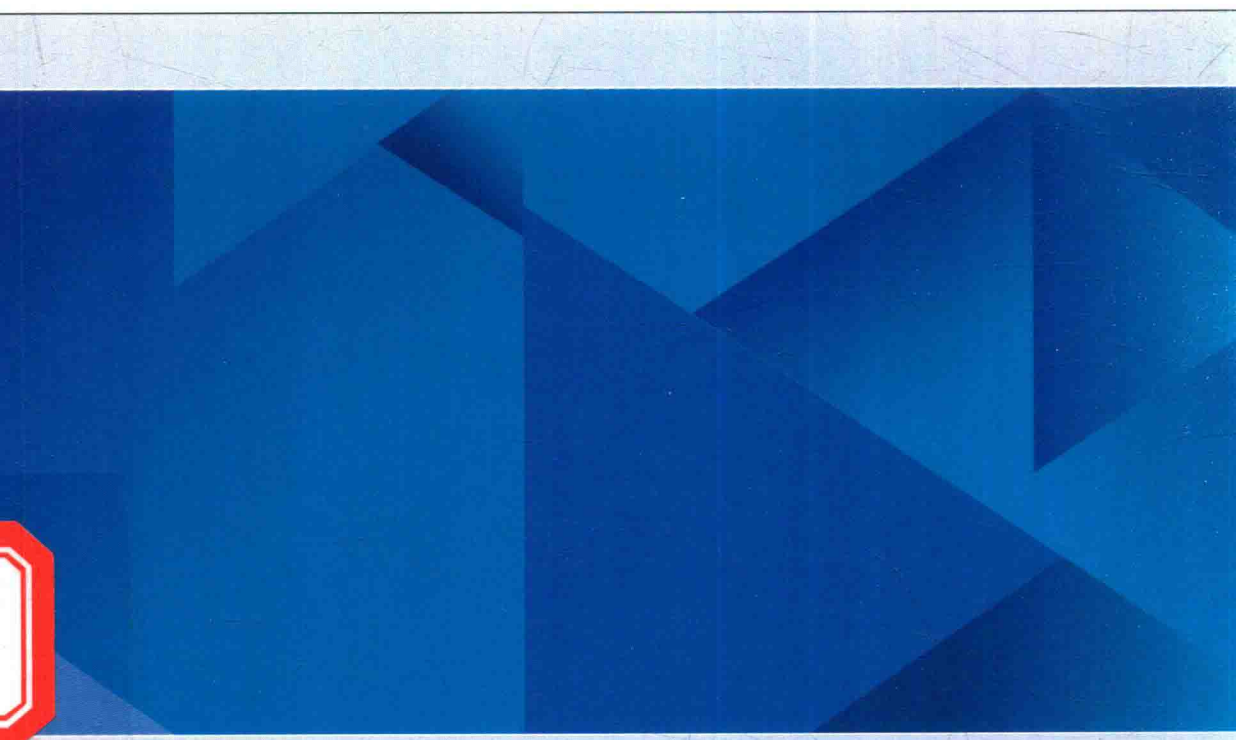
现代数学
·专著版·

Statistical Machine Learning: Error Analysis and Applications

统计机器学习：误差分析与应用

陈洪（Hong Chen） 著



WUHAN UNIVERSITY PRESS
武汉大学出版社

现代数学

Statistical Machine Learning: Error Analysis and Applications

统计机器学习：误差分析与应用

陈洪（Hong Chen） 著

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

统计机器学习:误差分析与应用 = Statistical Machine Learning: Error Analysis and Applications/陈洪著. —武汉: 武汉大学出版社,2017.6
现代数学. 专著版
ISBN 978-7-307-19426-7

Ⅰ.统… Ⅱ.陈… Ⅲ.机器学习—误差分析 Ⅳ.TP181

中国版本图书馆 CIP 数据核字(2017)第 143339 号

责任编辑:顾素萍 责任校对:汪欣怡 版式设计:马 佳

出版发行: **武汉大学出版社** (430072 武昌 珞珈山)
(电子邮件: cbs22@whu.edu.cn 网址: www.wdp.com.cn)
印刷: 虎彩印艺股份有限公司
开本:720×1000 1/16 印张:24.5 字数:412 千字 插页:1
版次:2017 年 6 月第 1 版 2017 年 6 月第 1 次印刷
ISBN 978-7-307-19426-7 定价:65.00 元

Preface

For given data, learning algorithms are usually designed to characterize the intrinsic relationship between the input and its output value. For different learning tasks (e.g., regression, classification, and ranking), various learning methods under different frameworks (e.g., Tikhonov regularization, greedy approximation, and neural networks) have been proposed and show well performance in real-word applications. Besides empirical evaluations, the theoretical foundations of learning algorithms are crucial to better understand its prediction mechanism, and then give us insights to further improve its performance. To attain this goal, learning theory (e.g., consistency, generalization, and robustness) has attracted much attentions in machine learning literatures, which involves the analysis techniques associated with statistics, approximation theory, and optimization. Ideas from all these areas have been brought together in a coherent and natural way to form a subject on machine learning theory.

This book aims to give some theory results for learning to regression, ranking and classification, with a focus on the generalization error analysis of learning algorithms. Special emphasises include regularized kernel methods, stochastic gradient descent method, greedy algorithms, and random weighted networks. To support the theoretical analysis, I also present the empirical evaluations on simulated and real data. As there are various learning problems and models, we try to each chapter focuses on one learning task and each section is self-containment.

I am indebted to Dr Biqin Song, Dr Zhibin Pan, Dr Yi Tang, Dr Jiangtao Peng, Dr Fangchao He, Dr Yulong Wang, and Dr Tieliang Gong for their valuable feedback and suggestion. Much of the material in this

book comes from our joint papers. I am indebted to Suping Gu and the staff of Wuhan University Press for their patience and willingness to help. I have also been supported by the National Natural Science Foundation of China (NSFC) under grant nos. 11671161,11626107, by the Teaching Research Project of Hubei Province (2016167), and by the Fundamental Research Funds for the Central Universities under Grants 2662015PY046, 2014PY025.

Contents

Chapter 1

Learning to Regression: Error Analysis and Applications

In this chapter, we mainly consider the following topics:

- We consider the multi-kernel regularized regression (MKRR) algorithm associated with least square loss over reproducing kernel Hilbert spaces. We provide an error analysis for the MKRR algorithm based on the Rademacher chaos complexity and iteration techniques. The main result is an explicit learning rate for the MKRR algorithm. Two examples are given to illustrate that the learning rates are much improved compared to those in the literature.
- We propose a stochastic gradient descent algorithm for the least square regression with coefficient regularization. An explicit expression of the solution via sampling operator and empirical integral operator is derived. Learning rates are given in terms of the suitable choices of the step size and regularization parameters.
- We propose a new greedy algorithm combining the semi-supervised learning and the sparse representation with the data-dependent hypothesis spaces. The proposed greedy algorithm is able to use a small portion of the labeled and unlabeled data to represent the target function, and to efficiently reduce the computational burden of the semi-supervised learning. We establish the estimation of the generalization error based on the empirical covering numbers. A detail analysis shows that the error has $O(n^{-1})$ decay. Our theoretical result illustrates that the unlabeled data is useful to improve the learning performance under mild conditions.

- The correntropy-induced loss (C-loss) has been employed in learning algorithms to improve their robustness to non-Gaussian noise and outliers recently. Despite its success on robust learning, only little work has been done to study the generalization performance of regularized regression with the C-loss. To enrich this theme, we investigate a kernel-based regression algorithm with the C-loss and ℓ_1-regularizer in data dependent hypothesis spaces. The asymptotic learning rate is established for the proposed algorithm in terms of novel error decomposition and capacity-based analysis technique. The sparsity characterization of the derived predictor is studied theoretically. Empirical evaluations demonstrate its advantages over the related approaches.
- Nyström method has been used successfully to improve the computational efficiency of kernel ridge regression (KRR). Recently, theoretical analysis of Nyström KRR, including generalization bound and convergence rate, has been established based on reproducing kernel Hilbert space (RKHS) associated with the symmetric positive semi-definite kernel. However, in real world applications, RKHS is not always optimal and kernel function is not necessary to be symmetric or positive semi-definite. Here, we consider the generalized Nyström kernel regression (GNKR) with ℓ_2 coefficient regularization, where the kernel just requires the continuity and boundedness. Error analysis is provided to characterize its generalization performance and the column norm sampling is introduced to construct the refined hypothesis space. In particular, the fast learning rate with polynomial decay is reached for the GNKR. Experimental analysis demonstrates the satisfactory performance of GNKR with the column norm sampling.
- Moving least-square method is investigated with examples drawn from unbounded sampling processes. Convergence analysis is established by imposing some incremental conditions on moments of the example output and window width. The derived convergence rates are consistent with the previous work concerning standard

boundedness assumption.

- Extreme learning machine (ELM) has gained increasing attention for its computation feasibility on various applications. However, the previous generalization analysis of ELM relies on the independent and identically distributed (i.i.d.) samples. Now, we go far beyond this restriction by investigating the generalization bound of the ELM classification associated with the uniformly ergodic Markov chains (u.e.M.c.) samples. The upper bound of the misclassification error is estimated for the ELM classification showing that the satisfactory learning rate can be achieved even for the dependent samples. Empirical evaluations on real-word datasets are provided to compare the predictive performance of ELM with independent and Markov sampling.

1.1 Multi-kernel Regularized Regression

Kernel methods such as Support Vector Machines have been extensively used in various learning tasks. The performance of a kernel method largely depends on the data representation via the choice of kernel function. Due to the practical importance of multi-kernel learning, many recent experiments and theoretical studies have been devoted to this subject recently, see, e.g., [71], [81], [82], [131], [141], [144]. The purpose of this section is to improve the estimation of learning rates for the multi-kernel regularized regression (MKRR) algorithm.

Let us recall some basic concepts of statistical learning theory in a regression setting. For details we refer to [119], [26], [27], [30], [130], [144] and references therein.

As usual in the framework of statistical learning theory, we consider a space X of possible inputs (instance space) and a space Y of possible outputs (label set). The product space $Z := X \times Y$ is assumed to be measurable and it is endowed with an unknown probability measure denoted by ρ. Input-output pairs (x, y) are sampled according to ρ. We assume X is a compact subset of $\mathbb{R}^n$ and Y is contained in $[-M, M]$. For

every $x \in X$, let $\rho(y|x)$ be the conditional (w.r.t. x) probability measure on Y and $\rho_X(x)$ be the marginal probability measure on X. Notice that $\rho, \rho(y|x)$ and $\rho_X(x)$ are related via $\rho(x,y) = \rho(y|x)\rho_X(x)$. The error for a measurable function $f : X \to Y$ is the so-called expected risk

$$\mathcal{E}(f) := \int_Z V(y, f(x))\mathrm{d}\rho,$$

where $V(y, f(x))$ is the loss function which measures the cost paid by replacing y with the estimate $f(x)$.

Now, we focus on the least square loss, namely,

$$V(y, f(x)) = (y - f(x))^2.$$

It is known that the function that minimizes the expected risk

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

is the regression function defined by

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

From the assumption $y \in [-M, M]$, we know that $|f_\rho(x)| \leq M$.

Set $\mathbb{N}_m := \{1, 2, ..., m\}$ for any $m \in \mathbb{N}$. A training set of size m is drawn by sampling m independent and identically distributed pairs according to ρ,

$$\boldsymbol{z} := \{z_i,\ i \in \mathbb{N}_m\} = \{(x_i, y_i),\ i \in \mathbb{N}_m\} \in Z^m.$$

We restrict our attention to the uniform convergence of the MKRR with a prescribed set $\mathcal{K}$ of candidate Mercer kernels. We say that $K : X \times X \to \mathbb{R}$ is a Mercer kernel if it is a continuous, symmetric, and positive semi-definite, i.e., for any finite set of distinct points $\{x_1, x_2, ..., x_\ell\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^{\ell}$ is positive semi-definite. The candidate reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated with a Mercer kernel K (see [2]) is defined as the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$, equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ defined by

$$\langle K_x, K_y \rangle_{\mathcal{H}_K} = K(x, y).$$

The reproducing property is given by

$$\langle K_x, f \rangle_{\mathcal{H}_K} = f(x), \quad \forall x \in X, \ f \in \mathcal{H}_K. \tag{1.2}$$

Denote $C(X)$ as the space of continuous functions on X with the supremum norm $\|\cdot\|_\infty$. Because of the continuity of $K \in \mathcal{K}$ and the compactness of X, we have

$$\kappa := \sup_{K \in \mathcal{K}} \sup_{x \in X} \sqrt{K(x, x)} < \infty.$$

So, the reproducing property above tells us

$$\|f\|_\infty \leq \kappa \|f\|_K, \quad \forall f \in \mathcal{H}_K.$$

The regularization scheme of MKRR is defined as a two-layer minimization problem

$$(K_{\mathbf{z},\lambda}, f_{\mathbf{z},\lambda}) := \arg \min_{K \in \mathcal{K}} \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2 + \lambda \|f\|_K^2 \right\}, \tag{1.3}$$

where λ is a positive constant called the regularization parameter. Usually it is chosen to depend on $m : \lambda = \lambda(m)$, and $\lambda(m) \to 0$ as $m \to \infty$.

We set the empirical error with respect to the random samples $\mathbf{z}$ as

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

It is a discretization of the error $\mathcal{E}(f)$. Then, the regularization scheme (1.3) can be rewritten as

$$(K_{\mathbf{z},\lambda}, f_{\mathbf{z},\lambda}) := \arg \min_{K \in \mathcal{K}} \min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|_K^2 \right\}. \tag{1.4}$$

Moreover, we expect that $f_{\mathbf{z},\lambda}$ to be a good approximator of f_ρ as $m \to \infty$.

Our main goal thus is to estimate the excess risk

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \tag{1.5}$$

for the empirical minimizer $f_{\mathbf{z},\lambda}$ provided by the regularization scheme (1.4).

There is a vast literature of statistical performance analysis for multi-kernel learning algorithms, e.g., [71], [82], [131], [144], [141]. These studies are based on different measures of complexity, such as Rademacher averages [71], [82], [144] and empirical covering numbers [131]. Recently, based on the theory of U-processes (see [28]), in [141] Ying and Campbell introduce the Rademacher chaos complexity and establish novel generalization error bounds for the multi-kernel regularized classification. Following the complexity analysis in [141], we consider here the MKRR algorithm (1.4) and investigate its generalization performance. Using the iteration technique in [117], [130], we derive faster convergence rates for the MKRR than previous results in [82], [140], [141], [144]. Combining the Rademacher chaos complexity with the iteration technique to derive the generalization error bound is our theoretical contribution.

Now we introduce some basic definitions and notations. A data free limit of (1.4) is defined as

$$(K_\lambda, f_\lambda) := \arg\min_{K\in\mathcal{K}} \min_{f\in\mathcal{H}_K} \left\{\mathcal{E}(f) + \lambda\|f\|_K^2\right\}. \tag{1.6}$$

While the regularization error of scheme (1.4) is defined as

$$D(\lambda) := \min_{K\in\mathcal{K}} \min_{f\in\mathcal{H}_K} \left\{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\|f\|_K^2\right\}. \tag{1.7}$$

It is easy to verify that f_λ is the minimizer of $D(\lambda)$. The sample error is defined by

$$S_{\mathbf{z},\lambda} = \{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})\} + \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\}. \tag{1.8}$$

Now, we present the following error decomposition which leads to bounds for the difference $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)$ (see [14], [27], [131]).

Proposition 1.1. *Let f_ρ, $f_{\mathbf{z},\lambda}$, and f_λ be defined via the expressions (1.1),(1.4), and (1.6) respectively. Then we have*

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \le \mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) + \lambda\|f_{\mathbf{z},\lambda}\|_{K_{\mathbf{z},\lambda}}^2 \le S_{\mathbf{z},\lambda} + D(\lambda).$$

The bounding technique for sample error (1.8) relies on complexity measures for MKRR. Here, we consider Rademacher chaos complexity as the measurement of hypothesis space complexity, which is defined as below (also, see [105], [141]).

Definition 1.1. *Let $\mathcal{F}$ be a class of functions on $X \times X$ and let $\{\epsilon_i : i \in \mathbb{N}_q\}$ be independent Rademacher random variables. Moreover, let $\mathbf{x} = \{x_i : i \in \mathbb{N}_q\}$ be independent random variables distributed according to a distribution μ on X. The homogeneous Rademacher chaos of order two, with respect to the Rademacher random variables ϵ, is a random variable system defined by*

$$\left\{\hat{\mathcal{U}}_g(\epsilon) = \frac{1}{q} \sum_{i,j \in \mathbb{N}_q,\ i<j} \epsilon_i \epsilon_j g(x_i, x_j) : g \in \mathcal{F}\right\}.$$

We refer to the expectation of its superma

$$\hat{\mathcal{U}}_q(\mathcal{F}) = \mathbb{E}_\epsilon[\sup_{g \in \mathcal{F}} |\hat{\mathcal{U}}_g(\epsilon)|]$$

as the empirical Rademacher chaos complexity.

We refer the reader to Section 3.2 of [28] for a general definition of the Rademacher chaos of an arbitrary order $q \in \mathbb{N}$. It is worth mentioning that the Rademacher process can be regarded as a homogeneous Rademacher chaos process of order one. The nice application of U-processes to the generalization analysis of the ranking and scoring problem is recently developed in [25].

Now we recall the definition of the kernel pseudo-dimension of a class of kernel functions on the product space $X \times X$ (see [1]).

Definition 1.2. *Let $\mathcal{K}$ be a set of reproducing kernel functions mapping from $X \times X$ to $\mathbb{R}$. We say that $S_q = \{(x_i, t_i) \in X \times X : i \in \mathbb{N}_q\}$ is pseudo-shattering by $\mathcal{K}$ if there are real numbers $\{r_i \in \mathbb{R} : i \in \mathbb{N}_q\}$ such that for any $b \in \{-1, 1\}^q$ there is a function $K \in \mathcal{K}$ with the property $\operatorname{sgn}(K(x_i, t_i) - r_i) = b_i$ for any $i \in \mathbb{N}_q$. Then, we define a pseudo-dimension $d_{\mathcal{K}}$ of $\mathcal{K}$ to be the maximum cardinality of S_q that is pseudo-shattered by $\mathcal{K}$.*

The following explicit estimate of the Rademacher chaos complexity is proved in [141].

Lemma 1.1. *Denote the pseudo-dimension of $\mathcal{K}$ by $d_{\mathcal{K}}$. Then, there exists a universal constant c such that, for any random variables $\mathbf{x} = \{x_i : i \in \mathbb{N}_q\}$ distributed according to a distribution μ on X, it holds*

$$\hat{\mathcal{U}}_q(\mathcal{K}) \leq c(1+\kappa)^2 d_{\mathcal{K}} \ln(2eq^2)).$$

It is easy to see that

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \int_X (f(x) - f_\rho(x))^2 \mathrm{d}\rho_X = \|f - f_\rho\|^2_{L^2_{\rho_X}}.$$

Hence the regularization error defined in (1.7) concerns the $L^2_{\rho_X}$-approximation of f_ρ by functions from $\mathcal{H}_{K_\lambda}$; it can be characterized by requiring f_ρ to lie in some interpolation space of the pair $(L^2_{\rho_X}, \mathcal{H}_{K_\lambda})$, as done in [82], [131].

Definition 1.3. *We say the target function f_ρ can be approximated with exponent $0 < \beta \leq 1$ in $\mathcal{H}_{K_\lambda}$ if there exists a constant c_β, such that*

$$D(\lambda) \leq c_\beta \lambda^\beta, \quad \forall \lambda > 0.$$

1.1.1 Error bounds

Now, we present our main results. The proofs of the results will be given in next subsection. Now, we present the learning rate of the MKRR (1.4) by trading off the sample error and the regularization error with iterative technique.

Theorem 1.1. *Assume that f_ρ can be approximated with exponent $0 < \beta \leq 1$ in $\mathcal{H}_{K_\lambda}$. Take $\lambda = (\frac{d_{\mathcal{K}} \ln m}{m})^{\frac{1}{2}}$. Then, for any $0 < \delta < 1$ there exists a constant $\tilde{c}$ independent of m such that*

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \leq \tilde{c}\Big(\frac{d_{\mathcal{K}} \ln m}{m}\Big)^{\frac{\beta}{2}}$$

with confidence $1 - \delta$.

In particular, when $\beta = 1$ and $d_{\mathcal{K}}$ is finite, the convergence rate is the order of $(\frac{\ln m}{m})^{\frac{1}{2}}$. Note that under the same condition for $D(\lambda)$ the

convergence rate of excess error (1.5) given in Corollary 5.4 of [82] is of order $(\frac{\ln m}{\sqrt{m}})^{\frac{\beta}{2(1+\beta)}}$. Thus, by combining the estimate of Rademacher chaos complexity with the iterative technique, we greatly improved the result in [82].

In fact, applying the results in [131] (Theorem 7 and Proposition 3) to investigate the generalization performance of MKRR algorithm, we can derive faster learning rates than Theorem 1.1. However, the conditions on multi-kernels $\mathcal{K}$ in [131] are much stricter than ours. An added capacity condition for the unit ball of the multi-kernel hypothesis space is needed in [131]. Different from the capacity estimate of the hypothesis space based on empirical covering numbers in [131], our estimate depends on the pseudo-dimension of a class of candidate kernel functions.

Now, we give two examples of error rates with the MKRR for learning Gaussian kernels $\mathcal{K}_{sc} := \{\mathrm{e}^{-\sigma\|x-t\|^2} : \sigma \in [0,\infty),\ x,t \in X\}$ and general radial basis kernels

$$\mathcal{K}_{rbf} := \left\{ \int_0^\infty \mathrm{e}^{-\sigma\|x-t\|^2} \mathrm{d}p(\sigma) : p \in \mathcal{M}(\mathbb{R}^+),\ x,t \in X \right\},$$

where $\mathcal{M}(\mathbb{R}^+)$ denotes a class of probabilities on $\mathbb{R}^+$. Also, we denote by $H^s(X)$ the Sobolev space with index $s > 0$ on X (see [106], [140]).

Example 1.1. *Let $X \subseteq \mathbb{R}^n$ be a domain with Lipschitz boundary. Let $f_{\mathbf{z},\lambda}$ be defined by (1.4) with Gaussian candidate kernels $\mathcal{K}_{sc}$ or general radial basis kernels $\mathcal{K}_{rbf}$. Assume $f_\rho \in H^s(X)$ with some $s > 0$. Then the following hold.*

(1) If $n/2 < s \le n/2+2$ then for any $0 < \varepsilon < 2s-n$, with $\lambda = (\frac{\ln m}{m})^{\frac{1}{2}}$, we have

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \le O\Big(\Big(\frac{\ln m}{m}\Big)^{\frac{2s-\varepsilon-n}{2(2s-\varepsilon)}}\Big), \tag{1.9}$$

with probability at least $1-\delta$.

(2) If X is bounded, ρ_X is the Lebesgue measure, and $0 < s \le 2$ then by choosing $\lambda = (\frac{\ln m}{m})^{\frac{1}{2}}$, with probability at least $1-\delta$ we get

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \le O\Big(\Big(\frac{\ln m}{m}\Big)^{\frac{s}{2s+n}}\Big). \tag{1.10}$$

Now we give some comparisons with the learning rates obtained in [141], [144]. By using the comparison inequality for the least square loss in [141], we can easily derive the corresponding convergence rate for regression. If $n/2 < s \leq n/2 + 2$ then for any $0 < \varepsilon < 2s - n$, the convergence rate of Example 1 in [144] yields $O((\ln m)^{\frac{1}{2}} m^{-\frac{2s-n-\varepsilon}{4(4s-n-2\varepsilon)}})$ and the learning rate in [141] yields $O((\ln m)^{\frac{1}{2}} m^{-\frac{2s-n-\varepsilon}{2(4s-n-2\varepsilon)}})$. Since $2s-\varepsilon-n > 0$, our learning rate in (1.9) is faster than the previous results of [141], [144]. Meanwhile, for the case in which ρ_X is the Lebesgue measure and $0 < s \leq 2$, our learning rate $O\big((\frac{\ln m}{m})^{-\frac{s}{2s+n}}\big)$ improves the results of $O\big((\ln m)^{\frac{1}{2}} m^{-\frac{s}{8s+2n}}\big)$ in [144] and of $O\big((\ln m)^{\frac{1}{2}} m^{-\frac{s}{4s+n}}\big)$ in [141].

In the above example we consider the function approximation on a domain of $\mathbb{R}^n$, so the learning rate is poor if the dimension n is large. However, in many situations, the input space X is a low-dimensional manifold embedded in the large dimensional space $\mathbb{R}^n$. In such a situation, the improved learning rates have been investigated in [140] based on the function approximation on Riemannian manifolds. In the sequel, we discuss whether the results presented in [140] can be improved by using the iterative technique.

Example 1.2. *Let X be a connected compact C^∞ submanifold of $\mathbb{R}^n$ which is isometrically embedded and it has dimension d. Let $f_{\mathbf{z},\lambda}$ be defined by (1.4) with Gaussian candidate kernels $\mathcal{K}_{sc}$ or general radial basis kernels $\mathcal{K}_{rbf}$. If $f_\rho \in H^s(X)$ with some $0 < s \leq 1$, then by taking $\lambda = (\frac{\ln m}{m})^{\frac{1}{2}}$, we have*

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) \leq O\Big(\Big(\frac{\ln m}{m}\Big)^{\frac{s}{2(s+d)}}\Big), \tag{1.11}$$

with probability at least $1 - \delta$.

When ignoring the difference of the form to express error rates using expectations and probabilistic inequalities, the learning rate in Example 1.2 improves the result of $O\big((\frac{\ln m}{m})^{\frac{s}{8s+2d}}\big)$ in [140]. Moreover, when $d \ll n$, the order of estimation in (1.11) is much better than the one in (1.10).