Adv Comput Math (2010) 33:97–112
DOI 10.1007/s10444-009-9123-x

Learning from non-identical sampling for classification

Quan-Wu Xiao · Zhi-Wei Pan

Received: 4 August 2008 / Accepted: 22 February 2009 /
Published online: 12 March 2009

Abstract We consider the classification problem by learning from samples drawn from a non-identical sequence of probability measures. The learning algorithm is from Tikhonov regularization schemes associated with convex loss functions and reproducing kernel Hilbert spaces. Our main goal is to provide satisfactory estimates for the excess misclassification error of the produced classifiers.

Keywords Reproducing kernel Hilbert space · Binary classification · Excess misclassification error · Non-identical sampling

Mathematics Subject Classifications (2000) 68T05 · 62H30

1 Introduction

In a binary classification problem, input points are from an input space which is a compact metric space X and outputs from $Y = \{1, -1\}$ representing two classes. A classifier $\mathcal{C}$ is the map $\mathcal{C} : X \to Y$ that makes a prediction $y = \mathcal{C}(x)$ for each $x \in X$.

To model two possibilities of outputs in Y, we assume that each $x \in X$ is assigned a probability measure ρ_x on Y. If a probability measure ρ_X on X

Communicated by Ding-Xuan Zhou.

Q.-W. Xiao (✉) · Z.-W. Pan
Joint Advanced Research Center,
University of Science and Technology of China
and City University of Hong Kong,
Suzhou, Jiangshu 215123, China
e-mail: 50009217@student.cityu.edu.hk

represents the distribution of input points, we can define a probability measure ρ on $Z = X \times Y$ with ρ_X being its marginal distribution on X and ρ_x its conditional distribution at $x \in X$. Then the prediction ability of a classifier is measured by the *misclassification error* defined to be the probability of wrong prediction

$$\mathcal{R}(\mathcal{C}) = \text{Prob}_{(x,y)\in(Z,\rho)}\{y \neq \mathcal{C}(x)\}.$$

The best classifier that minimizes the misclassification error is the *Bayes rule* given by

$$f_c(x) = \begin{cases} 1, & \text{if} \quad \rho_x(y=1) \geq \rho_x(y=-1), \\ -1, & \text{if} \quad \rho_x(y=1) < \rho_x(y=-1). \end{cases}$$

However, f_c is usually unknown since the underlying measures ρ_x are unknown.

The classification problem in learning theory aims at learning f_c from a finite sample $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^m \in Z^m$. Classifiers considered in this paper are induced by real-valued functions $f : X \to \mathbb{R}$ as $\mathcal{C} = \text{sgn}(f)$. For algorithms involving continuous functions, a loss function $\phi : \mathbb{R} \to \mathbb{R}_+$ is used to measure how the output y differs from $sgn(f(x))$ with error $\phi(yf(x))$.

Definition 1 A function $\phi : \mathbb{R} \to \mathbb{R}_+$ is called a *classifying loss function* if it is convex, $\phi'(0) < 0$, and the smallest zero of ϕ is 1.

For instance, the hinge loss $\phi_h(t) = \max\{1-t, 0\}$ for the support vector machine (SVM) [11] and the least-square loss $\phi_{ls}(t) = (1-t)^2$ [10] are classifying loss functions.

Learning algorithms here are kernel methods. A *Mercer kernel* $K : X \times X \to R$ is a continuous and symmetric function such that the matrix $(K(x_i, x_j))_{i,j=1}^l$ is positive semidefinite for any finite set of points $\{x_1, \ldots, x_l\} \subset X$. The Guassian function $K_\sigma(x, x') = \exp\{-|x-x'|^2/(2\sigma^2)\}$ with variance σ is an example of Mercer kernel, and it is C^∞ on $X \times X$. The *reproducing kernel Hilbert space* $\mathcal{H}_K$ associated with the kernel K is defined to be the completion of the linear span of the set of functions $\{K_x = K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ given by $\langle K_x, K_{x'} \rangle_K = K(x, x')$. The reproducing property takes the form

$$\langle K_x, f \rangle_K = f(x) \quad \forall f \in \mathcal{H}_K, \ x \in X. \tag{1.1}$$

Denote $\kappa = \sup_{x\in X} \sqrt{K(x,x)}$, then (1.1) tells us that

$$\|f\|_\infty \leq \kappa \|f\|_K, \quad \forall f \in \mathcal{H}_K. \tag{1.2}$$

The Tikhonov regularization scheme associated with a loss ϕ, a Mercer kernel K and a finite sample $\mathbf{z} \in Z^m$ is given [3] by

$$f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{ \mathcal{E}_{\mathbf{z}}^{\phi}(f) + \lambda \|f\|_K^2 \right\}, \tag{1.3}$$

where $\lambda = \lambda(m) > 0$ is a regularization parameter and $\mathcal{E}_{\mathbf{z}}^{\phi}$ is the *empirical error* given by

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

The minimization over a possibly infinitely dimensional space $\mathcal{H}_K$ indicates good approximation ability of the scheme (1.3), while a representer theorem [12] ensured by the reproducing property (1.1) tells us that a solution to (1.3) takes a form $f_{\mathbf{z},\lambda} = \sum_{i=1}^{m} c_i K_{x_i}$ with $\{c_i\}$ solved by a convex optimization problem over $\mathbb{R}^m$.

There is a large literature on convergence rates of the classification algorithm (1.3), e.g., [9, 13–15, 17]. These results are stated under the assumption that the sample $\mathbf{z}$ is drawn from ρ identically and independently.

A setting of online learning with non-identical sampling was considered for the purpose of regression in [8] and for classification in Hu and Zhou (unpublished manuscript). Following the framework there, we assume a sequence of Borel probability measures $\{\rho^{(i)}\}_{i=1,2,\cdots}$ on Z such that the conditional distribution of each $\rho^{(i)}$ at $x \in X$ is ρ_x, independent of i. Throughout the paper we also assume the independence of the sampling, that is, $\{z_i = (x_i, y_i)\}_{i=1,2,\cdots}$ is a sample drawn from the product probability space $\Pi_{i=1,2,\cdots}(Z, \rho^{(i)})$.

For the error analysis we assume a polynomial convergence of the marginal distributions in the dual $(C^s(X))^*$ of a Hölder space $C^s(X)$ with $0 < s \le 1$. Recall that $C^s(X)$ is the space of all continuous functions with the norm $\|f\|_{C^s(X)} = \|f\|_{C(X)} + |f|_{C^s(X)}$ finite, where $|f|_{C^s(X)} = \sup_{x\neq x'} \frac{|f(x)-f(x')|}{|x-x'|^s}$.

Definition 2 Let $0 < s \le 1$. The sequence $\{\rho_X^{(i)}\}_{i=1,2,\cdots}$ is said to *converge polynomially* to a probability measure ρ_X in $(C^s(X))^*$ if there exist $C_b > 0$ and $b > 0$ such that

$$\|\rho_X^{(i)} - \rho_X\|_{(C^s(X))^*} \le C_b i^{-b}, \quad \forall\, i \in \mathbb{N}. \tag{1.4}$$

Note that the above definition can also be written as

$$\left|\int_X f(x)\, d\rho_X^{(i)} - \int_X f(x) d\rho_X\right| \le C_b i^{-b}\|f\|_{C^s(X)}, \quad \forall\, f \in C^s(X),\ i \in \mathbb{N}. \tag{1.5}$$

Such a sequence of probability measures can be generated by iterations of integral operators associated with stochastic density kernels acting on an initial probability measure [8] .

The main goal of this paper is to show that for the non-identical sampling setting satisfying (1.4), the classifier $\mathrm{sgn}(f_{\mathbf{z},\lambda})$ can learn f_c well. The learning ability, also known as learning rate, is measured by *excess misclassification error* $\mathcal{R}(sgn(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c)$, which is expected to be small when m is large. The probability for $\mathcal{R}$ is taken with respect to the probability measure ρ on Z. Let us demonstrate our error analysis by two examples of special loss functions.

The first example corresponds to the hinge loss ϕ_h.

Theorem 1 *Let $\phi = \phi_h$. Assume (1.4) for the marginal distributions $\{\rho_X^{(i)}\}$, K to be C^∞ with $X \subset \mathbb{R}^n$, and for some $0 < \beta < 1$, $C_\beta > 0$,*

$$\inf_{f \in \mathcal{H}_K} \{\| f - f_c\|_{\mathcal{L}^1_{\rho_X}} + \lambda \| f\|_K^2\} \le C_\beta \lambda^\beta, \qquad \forall \lambda > 0. \tag{1.6}$$

Let $0 < \varepsilon < \frac{1}{2}$. Then for any $0 < \delta < 1$, with confidence $1 - \delta$ we have

$$\mathcal{R}(sgn(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c)$$
$$\le \begin{cases} C_h \log \frac{4}{\delta} \max\left\{m^{-\frac{2\beta}{1+2\beta}}, m^{-\frac{1}{2}+\varepsilon}\right\}, & \text{if } b > 1, \lambda = m^{-\frac{2}{1+2\beta}}, \\ C_h \log \frac{4}{\delta} \max\left\{m^{-\frac{2\beta}{1+2\beta}}(1+\log m), m^{-\frac{1}{2}+\varepsilon}\right\}, & \text{if } b = 1, \lambda = m^{-\frac{2}{1+2\beta}}, \\ C_h \log \frac{4}{\delta} \max\left\{m^{-\frac{2b\beta}{1+2\beta}}, m^{-\frac{1}{2}+\varepsilon}\right\}, & \text{if } 0 < b < 1, \lambda = m^{-\frac{2b}{1+2\beta}}, \end{cases}$$

where C_h is a constant independent of m or δ.

The second example corresponds to the least-square loss. Define the regression function

$$f_\rho(x) = \rho_x(y=1) - \rho_x(y=-1) = \int_Y y d\rho_x, \qquad x \in X,$$

then $f_c = sgn(f_\rho)$.

Theorem 2 *Let $\phi = \phi_{ls}$. Assume (1.4) for the marginal distributions $\{\rho_X^{(i)}\}$, K to be C^∞ with $X \subset \mathbb{R}^n$, and for some $0 < \beta < 1$, $C_\beta > 0$,*

$$\inf_{f \in \mathcal{H}_K} \{\| f - f_\rho\|_{\mathcal{L}^2_{\rho_X}} + \lambda \| f\|_K^2\} \le C_\beta \lambda^\beta, \quad \forall \lambda > 0. \tag{1.7}$$

Denote $\zeta = \max\{2(1-\beta), 1\}$. For any $0 < \varepsilon < \frac{1}{2}$, with confidence $1 - \delta$ we have

$$\mathcal{R}(sgn(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c)$$
$$\le \begin{cases} C_{ls} \log \frac{4}{\delta} \max\left\{m^{-\frac{\beta}{\zeta+2\beta}}, m^{-\frac{1}{2}+\varepsilon}\right\}, & \text{if } b > 1, \lambda = m^{-\frac{2}{\zeta+2\beta}}, \\ C_{ls} \log \frac{4}{\delta} \max\left\{m^{-\frac{\beta}{\zeta+2\beta}}\sqrt{1+\log m}, m^{-\frac{1}{2}+\varepsilon}\right\}, & \text{if } b = 1, \lambda = m^{-\frac{2}{\zeta+2\beta}}, \\ C_{ls} \log \frac{4}{\delta} \max\left\{m^{-\frac{b\beta}{\zeta+2\beta}}, m^{-\frac{1}{2}+\varepsilon}\right\}, & \text{if } 0 < b < 1, \lambda = m^{-\frac{2b}{\zeta+2\beta}}, \end{cases}$$

where C_{ls} is a constant independent of m or δ.

Remark 1 Conditions (1.6) and (1.7) measure how fast f_c and f_ρ are approximated by functions from $\mathcal{H}_K$ in the metric $\mathcal{L}^1_{\rho_X}$ and $\mathcal{L}^2_{\rho_X}$ respectively. They can be stated as some interpolation space conditions for f_c and f_ρ.

Remark 2 It is observed that if $b > 1$, both learning rates for hinge loss and least-square loss are independent of b. Moreover, if $\beta \ge \frac{1}{2}$, the learning rates are arbitrarily close to $m^{-\frac{1}{2}}$, which is the same as that in the i.i.d. case [4, 6, 7, 13, 16] .

Remark 3 In the current non-iid setting, learning rates for an online classification algorithm presented in Hu and Zhou (unpublished manuscript) are of type $E_{z_1,\ldots,z_m}\left(\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c)\right) = O(m^{-\theta})$ in an expectation-based form. They lead to confidence-based estimates of type $O(m^{-\theta}/\delta)$ by the Chebyshev inequality. Our results in Theorems 1 and 2 are in a stronger confidence-based form with $\frac{1}{\delta}$ replaced by $\log\frac{4}{\delta}$.

We will bound the excess misclassification error of $\mathrm{sgn}(f_{\mathbf{z},\lambda})$ for general classifying loss functions, including hinge loss and least-square loss, in Section 2. They are obtained from the analysis in Section 3 and finally proved in Section 4.

2 Framework with a general loss

In this section, we state learning rates for the algorithm (1.3) associated with a general classifying loss ϕ. This is done under the polynomial convergence (1.4) of the marginal distributions and some assumptions for the triple (ϕ, K, ρ).

Denote

$$f_\rho^\phi(x) = \arg\min\left\{\mathcal{E}^\phi(f) : f \quad \text{is a measurable function on } X\right\}$$

where $\mathcal{E}^\phi(f)$ is the *generalization error* defined by

$$\mathcal{E}^\phi(f) = \int_Z \phi(yf(x))\, d\rho.$$

It is shown in [13] that f_ρ^ϕ can be chosen such that $f_\rho^\phi(x) \in [-1, 1]$ for all $x \in X$.

The following lemma proved in [17] and [1] tells us that a classifier $\mathrm{sgn}(f)$ has a small excess misclassification error if f has a small *excess generalization error* $\mathcal{E}^\phi(f) - \mathcal{E}^\phi(f_\rho^\phi)$.

Lemma 1 *If ϕ is a classifying loss such that $\phi''(0)$ exists and is positive, then for any measurable function $f : X \to \mathbb{R}$, it holds that*

$$\mathcal{R}(sgn(f)) - \mathcal{R}(f_c) \le c_\phi\sqrt{\mathcal{E}^\phi(f) - \mathcal{E}^\phi(f_\rho^\phi)} \tag{2.1}$$

for some $c_\phi > 0$. Moreover, if $\phi = \phi_h$, then

$$\mathcal{R}(sgn(f)) - \mathcal{R}(f_c) \le \mathcal{E}^\phi(f) - \mathcal{E}^\phi(f_\rho^\phi). \tag{2.2}$$

Since $\mathcal{E}_{\mathbf{z}}^\phi$ is an empirical version of $\mathcal{E}^\phi$, we would except that $f_{\mathbf{z},\lambda}$ approximates f_ρ^ϕ as $m \to \infty$ and $\lambda \to 0$. This is actually true if f_ρ^ϕ can be approximated by the *regularization function*

$$f_\lambda = \arg\min_{f\in\mathcal{H}_K}\left\{\mathcal{E}^\phi(f) + \lambda\|f\|_K^2\right\}. \tag{2.3}$$

The approximation ability is measured by the decay of the *regularization error*

$$D(\lambda) = \mathcal{E}^{\phi}(f_\lambda) - \mathcal{E}^{\phi}(f_\rho^{\phi}) + \lambda \|f_\lambda\|_K^2 = \min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}^{\phi}(f) - \mathcal{E}^{\phi}(f_\rho^{\phi}) + \lambda \|f\|_K^2 \right\},$$

and the decay rate is described by

$$D(\lambda) \le C_\beta \lambda^\beta, \quad \forall\, 0 < \lambda \le 1, \tag{2.4}$$

for some $0 < \beta < 1$ and $C_\beta > 0$.

The approximation of f_λ by $f_{\mathbf{z},\lambda}$ involves the capacity of the function space $\mathcal{H}_K$. Here the capacity is measured by the covering number of balls $B_R = \{f \in \mathcal{H}_K : \|f\|_K \le R\}$.

Definition 3 Let S be a metric space and $\eta > 0$. We define the *covering number* $\mathcal{N}(S, \eta)$ to be the minimal $l \in \mathbb{N}$ such that there exists l disks in S with radius η covering S.

Denote the covering number of B_1 in $C(X)$ with the metric $\|\cdot\|_\infty$ by $\mathcal{N}(\eta)$.

Definition 4 We say that $\mathcal{H}_K$ has *polynomial complexity exponent* $r > 0$ if for some $C_r > 0$,

$$\log \mathcal{N}(\eta) \le C_r (1/\eta)^r, \quad \forall \eta > 0. \tag{2.5}$$

It was shown in [19] that (2.5) holds if $X \subset \mathbb{R}^n$ and $K \in C^{2n/r}(X \times X)$. In particular, for those $K \in C^\infty(X \times X)$ like Gaussian kernels, (2.5) is valid for arbitrary $r > 0$. See also [18].

Definition 5 We say that the kernel K satisfies the *kernel condition* of order s if for some $\kappa_s > 0$,

$$K \in C^s(X \times X), \quad |K(x,x) - 2K(x,x') + K(x',x')| \le \kappa_s^2 |x - x'|^{2s}, \quad \forall\, x, x' \in X. \tag{2.6}$$

For any $0 < s \le 1$, (2.6) is satisfied if $X \subset \mathbb{R}^n$ and $K \in C^2(X \times X)$ [20]. In particular, (2.6) holds for $K = K_\sigma$ and any $0 < s \le 1$.

The following concepts describe the increment and convexity of ϕ.

Definition 6 We say that ϕ has the *increment exponent* $p \ge 1$ if for some $C_p > 0$,

$$|\phi(t)| \le C_p |t|^p, \quad \forall\, |t| \ge 1. \tag{2.7}$$

Also, ϕ has the *derivative increment exponent* $q \ge 0$ if for some $C_q > 0$ and almost every $t \ge 1$,

$$|\phi'(t)| \le C_q |t|^q. \tag{2.8}$$

Definition 7 A *variance power* τ of $(\phi, \{\rho_X^{(i)}\})$ is a number $0 \le \tau \le 1$ such that for some constant C_τ and any $\|f\|_\infty \le 1$,

$$\int [\phi(yf(x)) - \phi(yf_\rho^\phi(x))]^2 \, d\rho^{(i)} \le C_\tau \left[\int \phi(yf(x)) \, d\rho^{(i)} - \int \phi(yf_\rho^\phi(x)) \, d\rho^{(i)} \right]^\tau \tag{2.9}$$

holds for each i.

Note that (2.9) holds with $\tau = 0$ and $C_\tau = 2C_q^2$ if (2.8) holds for some q. Larger τ are possible when ϕ has high convexity, or some noise conditions [5] are satisfied.

An essential difference between the regression setting and the classification one lies in improvement caused by a projection operator. Define a projection operator π by

$$\pi(f)(x) = \begin{cases} 1, & \text{if } f(x) > 1, \\ f(x), & \text{if } -1 \le f(x) \le 1, \\ -1, & \text{if } f(x) < -1. \end{cases} \tag{2.10}$$

Now we can state our main results on learning rates of the algorithm (1.3) with non-identical sampling for a general classifying loss function.

Theorem 3 *Let $f_\rho \in C^s(X)$ and $\int_Y \phi(yf_\rho^\phi(x))d\rho_x(y) \in C^s(X)$ for some $0 < s \le 1$. Assume that the kernel K satisfies (2.5) with $r > 0$ and (2.6), the loss ϕ satisfies (2.7) with $p \ge 1$ and (2.8) with $q \ge 0$, the triple (ϕ, K, ρ) satisfies (2.4) for some $0 < \beta < 1$. Let $\lambda = m^{-\gamma}$ with $\gamma \le 2/r$. If the sampling sequence $\{\rho^{(i)}\}$ satisfies (1.4) and (2.9), then with confidence $1 - \delta$, we have*

$$\mathcal{E}^\phi(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}^\phi(f_\rho^\phi) \le \widetilde{C} \log \frac{4}{\delta} \max \left\{ \left(\frac{1}{m}\right)^{\min\{\beta\gamma, \frac{1-r\gamma/2}{r+2-\tau}\}}, \omega_b(m) m^{\frac{\gamma\zeta}{2}} \right\},$$

where $\zeta = \max\{(1-\beta)p, (1-\beta)(q+1), 1\}$,

$$\omega_b(m) = \begin{cases} m^{-b}, & \text{if } 0 < b < 1, \\ \frac{1+\log m}{m}, & \text{if } b = 1, \\ \frac{1}{m}, & \text{if } b > 1, \end{cases} \tag{2.11}$$

and $\widetilde{C}$ is a constant independent of m or δ.

As a corollary, when $X \subset \mathbb{R}^n$ and K is C^∞, we can take arbitrarily small r in Theorem 3 and get the following learning rates.

Corollary 1 *Under the assumption of Theorem 3, if $b > 1$, $X \subset \mathbb{R}^n$ and K is C^∞, and if $\beta < \max\{\frac{1}{2(1-\tau)}, \frac{\max\{p,q+1\}}{\max\{p,q+1\}+2(1-\tau)}\}$, then with confidence $1 - \delta$, we have*

$$\mathcal{E}^\phi(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}^\phi(f_\rho^\phi) \le \widetilde{C} \log \frac{4}{\delta} \left(\frac{1}{m}\right)^{\frac{2\beta}{2\beta+\max\{(1-\beta)p,(1-\beta)(q+1),1\}}}.$$

3 Error analysis

This section is devoted to estimates for the sample error.

3.1 Error decomposition

Recall the projection operator π defined by (2.10). While the identity $\mathcal{R}(sgn(\pi(f))) = \mathcal{R}(sgn(f))$ holds true for any measurable function f, the minimal zero 1 and the convexity of ϕ tells us $\mathcal{E}^\phi(\pi(f)) \le \mathcal{E}^\phi(f)$. So we can bound the excess misclassification error by $\mathcal{E}^\phi(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}^\phi(f_\rho^\phi)$ instead of $\mathcal{E}^\phi(f_{\mathbf{z},\lambda}) - \mathcal{E}^\phi(f_\rho^\phi)$.

In the literature of error analysis for regularized classifiers, the excess generalization error is usually decomposed into two parts, sample error and regularization error (e.g., [13] and [14]). In this paper, another type of error caused by non-identical sampling should also be considered.

Denote

$$\mathcal{E}_m^\phi(f) = \frac{1}{m}\sum_{i=1}^m \int_Z \phi(yf(x))\, d\rho^{(i)}. \tag{3.1}$$

Since the conditional distributions for each i are the same, we know that f_ρ^ϕ also minimizes $\mathcal{E}_m^\phi$. Then our error decomposition can be done as follows.

Lemma 2 *Let $f_{\mathbf{z},\lambda}$ be defined by (1.3), and f_λ by (2.3). Then we have*

$$\begin{aligned}\mathcal{E}^\phi(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}^\phi(f_\rho^\phi) \le\ & \left\{[\mathcal{E}^\phi(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^\phi(\pi(f_{\mathbf{z},\lambda}))] + [\mathcal{E}_m^\phi(f_\lambda) - \mathcal{E}^\phi(f_\lambda)]\right\} \\ &+ \left\{[\mathcal{E}_m^\phi(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^\phi(f_\rho^\phi)] - [\mathcal{E}_{\mathbf{z}}^\phi(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}^\phi(f_\rho^\phi)]\right\} \\ &+ \left\{[\mathcal{E}_{\mathbf{z}}^\phi(f_\lambda) - \mathcal{E}_{\mathbf{z}}^\phi(f_\rho^\phi)] - [\mathcal{E}_m^\phi(f_\lambda) - \mathcal{E}_m^\phi(f_\rho^\phi)]\right\} + D(\lambda).\end{aligned} \tag{3.2}$$

Proof By the definition of $f_{\mathbf{z},\lambda}$, we have

$$\mathcal{E}_{\mathbf{z}}^\phi(\pi(f_{\mathbf{z},\lambda})) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2 \le \mathcal{E}_{\mathbf{z}}^\phi(f_{\mathbf{z},\lambda}) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2 \le \mathcal{E}_{\mathbf{z}}^\phi(f_\lambda) + \lambda\|f_\lambda\|_K^2. \tag{3.3}$$

It follows that

$$\begin{aligned}\mathcal{E}^\phi(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}^\phi(f_\rho^\phi) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2 \le\ & \left\{\mathcal{E}^\phi(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}^\phi(\pi(f_{\mathbf{z},\lambda}))\right\} \\ &+ \left\{\mathcal{E}_{\mathbf{z}}^\phi(\pi(f_{\mathbf{z},\lambda})) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2 - \mathcal{E}^\phi(f_\lambda)\right\} \\ &+ \mathcal{E}^\phi(f_\lambda) - \mathcal{E}^\phi(f_\rho^\phi) \\ \le\ & \left\{\mathcal{E}^\phi(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}^\phi(\pi(f_{\mathbf{z},\lambda}))\right\} \\ &+ \left\{\mathcal{E}_{\mathbf{z}}^\phi(f_\lambda) - \mathcal{E}^\phi(f_\lambda)\right\} \\ &+ \left\{\mathcal{E}^\phi(f_\lambda) - \mathcal{E}^\phi(f_\rho^\phi) + \lambda\|f_\lambda\|_K^2\right\}\end{aligned}$$

which can be bounded further by

$$\left\{[\mathcal{E}^{\phi}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}^{\phi}(f_{\rho}^{\phi})] - [\mathcal{E}_{\mathbf{z}}^{\phi}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}^{\phi}(f_{\rho}^{\phi})]\right\}$$
$$+ \left\{[\mathcal{E}_{\mathbf{z}}^{\phi}(f_{\lambda}) - \mathcal{E}_{\mathbf{z}}^{\phi}(f_{\rho}^{\phi})] - [\mathcal{E}^{\phi}(f_{\lambda}) - \mathcal{E}^{\phi}(f_{\rho}^{\phi})]\right\} + D(\lambda).$$

By adding and subtracting the quantities $[\mathcal{E}_m^{\phi}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^{\phi}(f_{\rho}^{\phi})]$ and $[\mathcal{E}_m^{\phi}(f_{\lambda}) - \mathcal{E}_m^{\phi}(f_{\rho}^{\phi})]$, we obtain the desired bound. □

In the bound (3.2), the first term is caused by the drift of non-identical measures $\rho^{(i)}$ from ρ, so it is called the *drift error*. This is an essential error appearing in the non-identical setting and will be treated in Section 3.2. The second and third terms are called the *sample error* which is caused by drawing the sample from each $\rho^{(i)}$ and will be handled in Section 3.3.

3.2 Estimating drift errors

Before discussing the drift error, we estimate C^s norms of some functions.

Lemma 3 *Under the assumption of* (2.4), (2.7) *and* (2.8), *we have*

$$\|\phi(\pi(f_{\mathbf{z},\lambda}))\|_{C^s(x)} \le \kappa_s C_q C_p^{\frac{1}{2}} \lambda^{-\frac{1}{2}} + C_p \tag{3.4}$$

and

$$\|\phi(f_{\lambda})\|_{C^s(x)} \le \left(C_p + \frac{C_q \kappa_s}{\kappa}\right)\left\{\left(\kappa\sqrt{C_\beta}\lambda^{\frac{\beta-1}{2}}\right)^{\max\{p,q+1\}} + 1\right\}. \tag{3.5}$$

Proof Since

$$\frac{|\phi(f(x)) - \phi(f(x'))|}{|f(x) - f(x')|} \le \|\phi'\|_{L^\infty[-\|f\|_\infty, \|f\|_\infty]}$$

for any $f \in C^s(X)$, we have

$$|\phi(f)|_{C^s(X)} = \sup_{x \ne x'} \frac{|\phi(f(x)) - \phi(f(x'))|}{|x - x'|^s} \le \|\phi'\|_{L^\infty[-\|f\|_\infty, \|f\|_\infty]} |f|_{C^s(X)}. \tag{3.6}$$

With the reproducing property (1.1) we know that for any $f \in \mathcal{H}_K$,

$$|f(x) - f(x')| = |\langle f, K_x - K_{x'}\rangle| \le \|f\|_K \sqrt{|K(x,x) - 2K_\sigma(x,x') + K_\sigma(x',x')|}.$$

It follows from the kernel condition (2.6) and $|\pi(f)(x) - \pi(f)(x')| \le |f(x) - f(x')|$ that

$$|\pi(f)|_{C^s(X)} \le |f|_{C^s(X)} = \sup_{x,x' \in X} \frac{|f(x) - f(x')|}{|x - x'|^s} \le \kappa_s \|f\|_K. \tag{3.7}$$

Taking $f = 0$ on the right side of (1.3), we have

$$\|f_{\mathbf{z},\lambda}\|_K \le \sqrt{\phi(0)/\lambda}. \tag{3.8}$$

Note from the convexity and the minimal zero 1 of ϕ that $\phi(0) \le \phi(-1) \le C_p$. It follows from (2.8), (3.6) and (3.7) that

$$|\phi(\pi(f_{\mathbf{z},\lambda}))|_{C^s(X)} \le C_q |\pi(f_{\mathbf{z},\lambda})|_{C^s(X)} \le C_q |f_{\mathbf{z},\lambda}|_{C^s(X)} \le \kappa_s C_q C_p^{\frac{1}{2}} \lambda^{-\frac{1}{2}}.$$

Then (3.4) holds since (2.7) implies $\|\phi(\pi(f_{\mathbf{z},\sigma,\lambda}))\|_{C(X)} \le C_p$.

Under the assumption (2.4), we have

$$\|f_\lambda\|_{C(X)} \le \kappa \|f_\lambda\|_K \le \kappa \sqrt{C_\beta} \lambda^{\frac{\beta-1}{2}}. \tag{3.9}$$

It follows from (2.7) and (2.8) that

$$\|\phi(f_\lambda)\|_{C(X)} \le C_p \max\{\|f_\lambda\|_{C(X)}^p, 1\} \le C_p \max\{\kappa^p C_\beta^{\frac{p}{2}} \lambda^{\frac{p(\beta-1)}{2}}, 1\}.$$

Applying (3.6) and (3.7), we also get

$$|\phi(f_\lambda)|_{C^s(X)} \le C_q \max\{\kappa^q C_\beta^{\frac{q}{2}} \lambda^{\frac{q(\beta-1)}{2}}, 1\} \kappa_s \sqrt{C_\beta} \lambda^{\frac{\beta-1}{2}}.$$

Then we see (3.5) by discussing in two cases of $\kappa\sqrt{C_\beta}\lambda^{\frac{\beta-1}{2}} \le 1$ and $\kappa\sqrt{C_\beta}\lambda^{\frac{\beta-1}{2}} > 1$. □

Now we can estimate the drift error.

Lemma 4 *Assume* (2.4), (2.7), (2.8), *and* (1.4). *Then we have*

$$\left|\mathcal{E}_m^\phi(f_\lambda) - \mathcal{E}^\phi(f_\lambda)\right| \le C_1 \omega_b(m) \left\{\lambda^{\frac{\beta-1}{2}\max\{p,q+1\}} + 1\right\} \tag{3.10}$$

and

$$\left|\mathcal{E}^\phi(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^\phi(\pi(f_{\mathbf{z},\lambda}))\right| \le C_2 \omega_b(m) \left(\lambda^{-\frac{1}{2}} + 1\right) \tag{3.11}$$

where C_1 and C_2 are constants given by

$$C_1 = C_b \widetilde{C}_b \left(4 + \|f_\rho\|_{C^s(X)}\right) \left(C_p + \frac{C_q \kappa_s}{\kappa}\right) \left\{\left(\kappa\sqrt{C_\beta}\right)^{\max\{p,q+1\}} + 1\right\}$$

$$C_2 = C_b \widetilde{C}_b \left(4 + \|f_\rho\|_{C^s(X)}\right) \left(\kappa_s C_q C_p^{\frac{1}{2}} + C_p\right)$$

with

$$\widetilde{C}_b := \begin{cases} \frac{1}{1-b}, & \text{if } 0 < b < 1, \\ 1, & \text{if } b = 1, \\ \frac{b}{b-1}, & \text{if } b > 1. \end{cases}$$

Proof Denote

$$\Delta_i = \int_Z \phi(y f_\lambda(x)) d(\rho^{(i)} - \rho).$$

Since $P(y = 1|x) = (1 + f_\rho(x))/2$ and $P(y = -1|x) = (1 - f_\rho(x))/2$, we have

$$\begin{aligned}\Delta_i &= \int_X \left\{ \frac{1 + f_\rho(x)}{2} \phi(f_\lambda(x)) + \frac{1 - f_\rho(x)}{2} \phi(-f_\lambda(x)) \right\} d\left(\rho_X^{(i)} - \rho_X\right) \\ &\le \frac{1}{2} C_b i^{-b} \left\{ \left\| \left(1 + f_\rho\right) \phi(f_\lambda) \right\|_{C^s(X)} + \left\| \left(1 - f_\rho\right) \phi(-f_\lambda) \right\|_{C^s(X)} \right\}.\end{aligned}$$

Observe that

$$\|fg\|_{C^s(X)} \le \|f\|_{C(X)} \|g\|_{C^s(X)} + \|f\|_{C^s(X)} \|g\|_{C(X)}.$$

Since $\|(1 + f_\rho)\|_{C(X)} \le 2$ and $\|(1 + f_\rho)\|_{C^s(X)} \le 2 + \|f_\rho\|_{C^s(X)}$, we have

$$\left\| \left(1 + f_\rho\right) \phi(f_\lambda) \right\|_{C^s(X)} \le 2 \|\phi(f_\lambda)\|_{C^s(X)} + \left(2 + \|f_\rho\|_{C^s(X)}\right) \|\phi(f_\lambda)\|_{C(X)}.$$

The same bound is valid for $\left\| \left(1 - f_\rho\right) \phi(-f_\lambda) \right\|_{C^s(X)}$. Then with the C^s norms bounded in Lemma 3, we have

$$\Delta_i \le C_b i^{-b} \left(4 + \|f_\rho\|_{C^s(X)}\right) \left(C_p + \frac{C_q \kappa_s}{\kappa} \right) \left\{ \left(\kappa \sqrt{C_\beta} \lambda^{\frac{\beta - 1}{2}} \right)^{\max\{p, q+1\}} + 1 \right\}.$$

Notice that

$$\mathcal{E}_m^\phi(f_\lambda) - \mathcal{E}^\phi(f_\lambda) = \frac{1}{m} \sum_{i=1}^m \Delta_i.$$

Then the bound (3.10) follows by applying the following elementary inequality (which is derived from $\sum_{i=1}^m \frac{1}{i^b} \le 1 + \sum_{i=2}^m \int_{i-1}^i x^{-b} dx \le 1 + \int_1^m x^{-b} dx$) valid for any $b > 0$ and $m \in \mathbb{N}$

$$\frac{1}{1^b} + \frac{1}{2^b} + \ldots + \frac{1}{m^b} \le \begin{cases} \frac{1}{1-b} m^{1-b}, & \text{if } 0 < b < 1, \\ 1 + \log m, & \text{if } b = 1, \\ \frac{b}{b-1}, & \text{if } b > 1. \end{cases} \tag{3.12}$$

The bound (3.11) is proved in the same way. □

3.3 Estimating the sample error with non-identical sampling

With proper probability inequalities, bounding the sample error for the non-identical sampling case has no significant difference from that for the identical sampling case, which has been extensively studied. Therefore, proofs in this subsection are simplified because they are similar to those in Chapter 10 of [2].

The following one-side Bernstein inequality is standard.

Lemma 5 *Let $\{\xi_i\}_{i=1}^m$ be independent random variables on a probability space Z with means $\{\mu_i\}$ and variances $\{\sigma_i^2\}$ satisfying $|\xi_i(z) - \mu_i| \le M$ for each i and almost all $z \in Z$. Let $\Sigma^2 = \sum_{i=1}^m \sigma_i^2$. Then for every $\varepsilon > 0$,*

$$Prob\left\{\frac{1}{m}\sum_{i=1}^{m}[\xi_i - \mu_i] > \varepsilon\right\} \le \exp\left\{-\frac{m^2\varepsilon^2}{2\left(\Sigma^2 + \frac{1}{3}mM\varepsilon\right)}\right\}. \tag{3.13}$$

The sample error involving f_λ can be bounded by Lemma 5 as Proposition 10.18 in [2] by noticing (3.9).

Lemma 6 *Assume (2.4), (2.7), and (2.9). With confidence $1 - \delta/2$, the quantity $[\mathcal{E}_{\mathbf{z}}^{\phi}(f_\lambda) - \mathcal{E}_{\mathbf{z}}^{\phi}(f_\rho^{\phi})] - [\mathcal{E}_m^{\phi}(f_\lambda) - \mathcal{E}_m^{\phi}(f_\rho^{\phi})]$ is bounded by*

$$C_{\beta,p,\tau,\kappa}\log\frac{4}{\delta}\max\left\{\frac{\lambda^{\frac{p(\beta-1)}{2}}}{m}, \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}}\right\} + \mathcal{E}_m^{\phi}(f_\lambda) - \mathcal{E}_m^{\phi}(f_\rho^{\phi}),$$

where $C_{\beta,p,\tau,\kappa} := 2C_p\kappa^p C_\beta^{p/2} + (C_p + 2C_\tau^{1/(2-\tau)})$.

Lemma 5 cannot be used to bound directly the sample error term involving $f_{\mathbf{z},\lambda}$, since the function depends on the sample $\mathbf{z}$. The probability inequality we would use should work for a set of functions, not a single one. The following inequality follows by a standard argument with covering numbers, as Lemma 10.20 in [2].

Lemma 7 *Let $0 \le \tau \le 1$, C_τ, $M \ge 0$, and $\mathcal{G}$ be a set of functions on Z such that for each $g \in \mathcal{G}$ and $i = 1, \ldots, m$, $\mu_i(g) = \int_Z g(z)\, d\rho^{(i)} \ge 0$, $\|g - \mu_i(g)\|_{L^\infty_{\rho^{(i)}}} \le M$, and $\mu_i(g^2) \le C_\tau(\mu_i(g))^\tau$. Then for any $\varepsilon > 0$,*

$$Prob\left\{\sup_{g\in\mathcal{G}}\frac{\frac{1}{m}\sum_{i=1}^m \mu_i(g) - \frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{\frac{1}{m}\sum_{i=1}^m(\mu_i(g))^\tau + \varepsilon^\tau}} > 4\varepsilon^{1-\frac{\tau}{2}}\right\}$$
$$\le \mathcal{N}(\mathcal{G},\varepsilon)\exp\left\{-\frac{m\varepsilon^{2-\tau}}{2\left(C_\tau + \frac{1}{3}M\varepsilon^{1-\tau}\right)}\right\}.$$

Then we can estimate the sample error part involving $f_{\mathbf{z},\lambda}$ as follows.

Lemma 8 *Assume (2.5), (2.7), (2.8), and (2.9). If $\lambda \ge m^{-2/r}$, then for any constant $0 < \delta < 1$, with confidence $1 - \delta/2$, we have*

$$[\mathcal{E}_m^{\phi}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^{\phi}(f_\rho^{\phi})] - [\mathcal{E}_{\mathbf{z}}^{\phi}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}^{\phi}(f_\rho^{\phi})]$$
$$\le C_{p,q,r,\tau}\max\left\{\left(\frac{\log(2/\delta)}{m}\right)^{\frac{1}{2-\tau}}, \left(\frac{1}{m\lambda^{r/2}}\right)^{\frac{1}{r+2-\tau}}\right\} + \frac{1}{2}\left[\mathcal{E}_m^{\phi}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_m^{\phi}(f_\rho^{\phi})\right],$$

where $C_{p,q,r,\tau} := 96\left(\max\{C_p, C_q, C_r, C_\tau, 1\}\right)^2$.

Proof From (3.8) we know that $f_{\mathbf{z},\lambda} \in \left\{f \in \mathcal{H}_K : \|f\|_K \le \sqrt{\phi(0)/\lambda}\right\}$. Apply Lemma 7 to the set

$$\mathcal{G} = \left\{\phi(y\pi(f)(x)) - \phi(yf_\rho^\phi(x)) : \|f\|_K \le \sqrt{\phi(0)/\lambda}\right\}$$

satisfying the assumption with $M = 2\phi(-1)$. Also, the covering number can be bounded as $\mathcal{N}(\mathcal{G}, \varepsilon) \le \mathcal{N}(B_{\sqrt{\phi(0)/\lambda}}, \varepsilon/|\phi'_+(-1)|)$. We know from Lemma 7 and the inequality

$$\sqrt{\left(\mathcal{E}_m^\phi(\pi(f)) - \mathcal{E}_m^\phi(f_\rho^\phi)\right)^\tau + \varepsilon^\tau} \cdot 4\varepsilon^{1-\frac{\tau}{2}} \le \frac{1}{2}\left(\mathcal{E}_m^\phi(\pi(f)) - \mathcal{E}_m^\phi(f_\rho^\phi)\right) + 12\varepsilon$$

that with confidence at least $1 - \delta/2$ there holds for every $f \in B_{\sqrt{\phi(0)/\lambda}}$,

$$[\mathcal{E}_m^\phi(\pi(f)) - \mathcal{E}_m^\phi(f_\rho^\phi)] - [\mathcal{E}_{\mathbf{z}}^\phi(\pi(f)) - \mathcal{E}_{\mathbf{z}}^\phi(f_\rho^\phi)] \le 12\varepsilon^* + \frac{1}{2}\left[\mathcal{E}_m^\phi(\pi(f)) - \mathcal{E}_m^\phi(f_\rho^\phi)\right], \tag{3.14}$$

where ε^* is the smallest ε satisfying

$$C_r\left(\frac{\sqrt{\phi(0)}|\phi'_+(-1)|}{\sqrt{\lambda}\varepsilon}\right)^r - \frac{m\varepsilon^{2-\tau}}{2C_\tau + \frac{4}{3}\phi(-1)\varepsilon^{1-\tau}} \le \log\frac{\delta}{2}.$$

This inequality can be written as

$$\varepsilon^{r+2-\tau} - \log\frac{2}{\delta}\frac{4}{3m}\phi(-1)\varepsilon^{r+1-\tau} - C_r\left(\frac{\sqrt{\phi(0)}|\phi'_+(-1)|}{\sqrt{\lambda}}\right)^r \frac{4}{3m}\phi(-1)\varepsilon^{1-\tau}$$
$$- \frac{2C_\tau}{m}\log\frac{2}{\delta}\varepsilon^r - C_r\left(\frac{\sqrt{\phi(0)}|\phi'_+(-1)|}{\sqrt{\lambda}}\right)^r \frac{2C_\tau}{m} \ge 0.$$

By Lemma 7.2 of [2] we know that

$$\varepsilon^* \le \max\left\{\log\frac{2}{\delta}\frac{16}{3m}\phi(-1), \left(\frac{\sqrt{\phi(0)}|\phi'_+(-1)|}{\sqrt{\lambda}}\right)^{r/(r+1)}\left(\frac{16C_r}{3m}\phi(-1)\right)^{1/(r+1)},\right.$$
$$\left.\left(\frac{8C_\tau}{m}\log\frac{2}{\delta}\right)^{1/(2-\tau)}, \left(\frac{\sqrt{\phi(0)}|\phi'_+(-1)|}{\sqrt{\lambda}}\right)^{r/(r+2-\tau)}\left(\frac{8C_\tau C_r}{m}\right)^{1/(r+2-\tau)}\right\}.$$

Since $\lambda^{r/2}m \ge 1$ we have

$$\varepsilon^* \le 8\left(\max\{C_p, C_q, C_r, C_\tau, 1\}\right)^2 \max\left\{\left(\frac{\log(2/\delta)}{m}\right)^{1/(2-\tau)}, \left(\frac{1}{m\lambda^{r/2}}\right)^{1/(r+2-\tau)}\right\}.$$

Putting this bound for ε^* into (3.14) and taking $f = f_{\mathbf{z},\lambda}$, we see that the desired estimate holds true. □

4 Deriving learning rates

Now we can prove our main result on learning rates.

We should put the error bounds got in Lemmas 4, 6 and 8 into (3.2) to derive the learning rate in Theorem 3.

Proof of Theorem 3 First we apply Lemmas 6 and 8. We know that with confidence $1-\delta$,

$$\begin{aligned}
&\left\{[\mathcal{E}_m^\phi(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}_m^\phi(f_\rho^\phi)]-[\mathcal{E}_{\mathbf{z}}^\phi(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}_{\mathbf{z}}^\phi(f_\rho^\phi)]\right\}\\
&\quad+\left\{[\mathcal{E}_{\mathbf{z}}^\phi(f_\lambda)-\mathcal{E}_{\mathbf{z}}^\phi(f_\rho^\phi)]-[\mathcal{E}_m^\phi(f_\lambda)-\mathcal{E}_m^\phi(f_\rho^\phi)]\right\}\\
&\le (C_{p,q,r,\tau}+C_{\beta,p,\tau,\kappa})\log\frac{4}{\delta}\max\left\{\left(\frac{1}{m}\right)^{\frac{1}{2-\tau}},\left(\frac{1}{m\lambda^{r/2}}\right)^{\frac{1}{r+2-\tau}},\frac{\lambda^{\frac{p(\beta-1)}{2}}}{m}\right\}\\
&\quad+\frac{1}{2}\left[\mathcal{E}_m^\phi(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}_m^\phi(f_\rho^\phi)\right]+\mathcal{E}_m^\phi(f_\lambda)-\mathcal{E}_m^\phi(f_\rho^\phi).
\end{aligned}$$

Next we combine the above bound with Lemma 4 and (3.2) and (2.4) and know that with confidence $1-\delta$, the quantity $\mathcal{E}^\phi(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}^\phi(f_\rho^\phi)$ is bounded by

$$\begin{aligned}
&(C_{p,q,r,\tau}+C_{\beta,p,\tau,\kappa})\log\frac{4}{\delta}\max\left\{\left(\frac{1}{m}\right)^{\frac{1}{2-\tau}},\left(\frac{1}{m\lambda^{r/2}}\right)^{\frac{1}{r+2-\tau}},\frac{\lambda^{\frac{p(\beta-1)}{2}}}{m}\right\}\\
&\quad+2C_1\omega_b(m)\left\{\lambda^{\frac{\beta-1}{2}\max\{p,q+1\}}+1\right\}+\frac{C_2}{2}\omega_b(m)\left(\lambda^{-\frac{1}{2}}+1\right)+C_\beta\lambda^\beta\\
&\quad+\frac{1}{2}\left[\mathcal{E}^\phi(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}_m^\phi(f_\rho^\phi)\right].
\end{aligned}$$

Then we observe that

$$\begin{aligned}
\mathcal{E}^\phi(f_\rho^\phi)-\mathcal{E}_m^\phi(f_\rho^\phi)&=\int_X\left\{\int_Y\phi(yf_\rho^\phi(x))d\rho_x(y)\right\}d\left(\rho_X-\frac{1}{m}\sum_{i=1}^m\rho_X^{(i)}\right)\\
&\le\left\|\int_Y\phi(yf_\rho^\phi(x))d\rho_x(y)\right\|_{C^s(X)}\left\|\rho_X-\frac{1}{m}\sum_{i=1}^m\rho_X^{(i)}\right\|_{(C^s(X))^*}.
\end{aligned}$$

It follows from (3.12) that

$$\mathcal{E}^\phi(f_\rho^\phi)-\mathcal{E}_m^\phi(f_\rho^\phi)\le\left\|\int_Y\phi(yf_\rho^\phi(x))d\rho_x(y)\right\|_{C^s(X)}C_b\widetilde{C}_b\omega_b(m).$$

Thus we see that

$$\mathcal{E}^{\phi}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) \le \widetilde{C} \log \frac{4}{\delta} \max \left\{ \left(\frac{1}{m}\right)^{\frac{1}{2-\tau}}, \lambda^{\beta}, \right.$$
$$\left. \left(\frac{1}{m\lambda^{r/2}}\right)^{\frac{1}{r+2-\tau}}, \omega_b(m) \left(\frac{1}{\sqrt{\lambda}}\right)^{\max\{(1-\beta)p,(1-\beta)(q+1),1\}} \right\},$$

where

$$\widetilde{C} = 2(C_{p,q,r,\tau} + C_{\beta,p,\tau,\kappa}) + 8C_1 + 2C_2 + 2C_{\beta}$$
$$+2 \left\| \int_Y \phi(y f_{\rho}^{\phi}(x)) d\rho_x(y) \right\|_{C^s(X)} C_b \widetilde{C}_b.$$

Finally if we take $\lambda = m^{-\gamma}$ with $\gamma \le 2/r$, we know that with confidence $1-\delta$,

$$\mathcal{E}^{\phi}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) \le \widetilde{C} \log \frac{4}{\delta} \max \left\{ \left(\frac{1}{m}\right)^{\min\{\beta\gamma, \frac{1-r\gamma/2}{r+2-\tau}\}}, \ \omega_b(m) m^{\frac{\gamma\zeta}{2}} \right\}.$$

This proves Theorem 3. □

Theorems 1 and 2 are concluded from Theorem 3 and Lemma 1.

Proof of Theorem 1 For the hinge loss ϕ_h, we can take $p = 1, q = 0, \tau = 0$, and $f_{\rho}^{\phi} = f_c$. Then $\zeta = \max\{p(1-\beta), (q+1)(1-\beta), 1\} = 1$.

Since $K \in C^{\infty}(X \times X)$, r can be arbitrarily small, and $m^{-\frac{2-r\gamma}{2(2+r-\tau)}} \le m^{-\frac{1}{2}+\varepsilon}$ holds.

If $b > 1$, take $\gamma = \frac{2}{1+2\beta}$, then $m^{\frac{\gamma\zeta}{2}} \omega_b(m) = m^{-\gamma\beta} = m^{-\frac{2\beta}{1+2\beta}}$.

If $0 < b < 1$, take $\gamma = \frac{2b}{1+2\beta}$, then $m^{\frac{\gamma\zeta}{2}} \omega_b(m) = m^{-\gamma\beta} = m^{-\frac{2b\beta}{1+2\beta}}$.

If $b = 1$, take $\gamma = \frac{2}{1+2\beta}$, then $m^{-\gamma\beta} < m^{\frac{\gamma\zeta}{2}} \omega_b(m) = m^{-\frac{2\beta}{1+2\beta}}(1 + \log m)$.

With (2.2) and Theorem 3, we can get the result of this theorem with $C_h = \tilde{C}$. □

Proof of Theorem 2 For the least-square loss ϕ_{ls}, we can take $p = 2$, $q = 1$, $\tau = 1$, and $f_{\rho}^{\phi} = f_{\rho}$. Then $\zeta = \max\{2(1-\beta), 1\} = \max\{p(1-\beta), (q+1)(1-\beta), 1\}$.

Since $K \in C^{\infty}(X \times X)$, r can be arbitrarily small, and $m^{-\frac{2-r\gamma}{2(2+r-\tau)}} \le m^{-1+2\varepsilon}$ holds.

If $b > 1$, take $\gamma = \frac{2}{\zeta+2\beta}$, then $m^{-\gamma\beta} = m^{\frac{\gamma\zeta}{2}} \omega_b(m) = m^{-\frac{2\beta}{\zeta+2\beta}}$.

If $0 < b < 1$, take $\gamma = \frac{2b}{\zeta+2\beta}$, then $m^{-\gamma\beta} = m^{\frac{\gamma\zeta}{2}} \omega_b(m) = m^{-\frac{2b\beta}{\zeta+2\beta}}$.

If $b = 1$, take $\gamma = \frac{2}{\zeta+2\beta}$, then $m^{-\gamma\beta} < m^{\frac{\gamma\zeta}{2}} \omega_b(m) = m^{\frac{2\beta}{\zeta+2\beta}}(1 + \log m)$.

Thus, Theorem 2 follows from (2.1) and Theorem 3 with $C_{ls} = c_{\phi}\sqrt{\widetilde{C}}$. □

Acknowledgements The work described in this paper was partially supported by a grant from the Research Grants Council of Hong Kong [Project No. CityU 103206] and National Science Fund for Distinguished Young Scholars of China [Project No. 10529101].

References

1. Chen, D.R., Wu, Q., Ying, Y., Zhou, D.X.: Support vector machine soft margin classifiers: error analysis, J. Mach. Learn. Res. **5**, 1143–1175 (2004)
2. Cucker, F., Zhou, D.X.: Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, Cambridge (2007)
3. Evgeniou, T., Pontil, M., Poggio, T.: Regularization networks and support vector machines. Adv. Comput. Math. **13**, 1–50 (2000)
4. De Vito, E., Caponnetto, A., Rosasco, L.: Model selection for regularized least-squares algorithm in learning theory. Found. Comput. Math. **5**, 59–85 (2005)
5. Mukherjee, S., Wu, Q.: Estimation of gradients and coordinate covariation in classification. J. Mach. Learn. Res. **7**, 2481–2514 (2006)
6. Smale, S., Zhou, D.X.: Shannon sampling and function reconstruction from point values. Bull. Am. Math. Soc. **41**, 279–305 (2004)
7. Smale, S., Zhou, D.X.: Learning theory estimates via integral operators and their approximations. Constr. Approx. **26**, 153–172 (2007)
8. Smale, S., Zhou, D.X.: Online learning with Markov sampling. Anal. Appl. **7**, 87–113 (2009)
9. Steinwart, I., Scovel, C.: Fast rates for support vector machines using Gaussian kernels. Ann. Stat. **35**, 575–607 (2007)
10. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**, 293–300 (1999)
11. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
12. Wahba, G.: Spline Models for Observational Data. SIAM, Philadelphia (1990)
13. Wu, Q., Ying, Y., Zhou, D.X.: Multi-kernel regularized classifiers. J. Complex. **21**, 108–134 (2007)
14. Xiang, D.H., Zhou, D.X.: Classification with Gaussians and convex loss. J. Mach. Learn. Res. (2009, in press)
15. Ying, Y.: Convergence analysis of online algorithms. Adv. Comput. Math. **27**, 273–291 (2007)
16. Yao, Y., Rosasco, L., Caponnetto, A.: On early stopping in gradient descent learning. Constr. Approx. **26**, 298–315 (2007)
17. Zhang, T.: Statistical behavior and consistency of classification methods based on convex risk minimization. Ann. Stat. **32**, 56–85 (2004)
18. Zhou, D.X.: The covering number in learning theory. J. Complex. **18**, 739–767 (2002)
19. Zhou, D.X.: Capacity of reproducing kernel spaces in learning theory. IEEE Trans. Inf. Theory **49**, 1743–1752 (2003)
20. Zhou, D.X.: Derivative reproducing properties for kernel methods in learning theory. J. Comput. Appl. Math. **220**, 456–463 (2008)