Available online at www.sciencedirect.com

ScienceDirect

Journal of Complexity 23 (2007) 108–134

Journal of COMPLEXITY

www.elsevier.com/locate/jco

Multi-kernel regularized classifiers☆

Qiang Wu[1], Yiming Ying[2], Ding-Xuan Zhou*

Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China

Received 13 September 2005; accepted 27 June 2006
Available online 30 August 2006

Abstract

A family of classification algorithms generated from Tikhonov regularization schemes are considered. They involve multi-kernel spaces and general convex loss functions. Our main purpose is to provide satisfactory estimates for the excess misclassification error of these multi-kernel regularized classifiers when the loss functions achieve the zero value. The error analysis consists of two parts: regularization error and sample error. Allowing multi-kernels in the algorithm improves the regularization error and approximation error, which is one advantage of the multi-kernel setting. For a general loss function, we show how to bound the regularization error by the approximation in some weighted L^q spaces. For the sample error, we use a projection operator. The projection in connection with the decay of the regularization error enables us to improve convergence rates in the literature even for the one-kernel schemes and special loss functions: least-square loss and hinge loss for support vector machine soft margin classifiers. Existence of the optimization problem for the regularization scheme associated with multi-kernels is verified when the kernel functions are continuous with respect to the index set. Concrete examples, including Gaussian kernels with flexible variances and probability distributions with some noise conditions, are used to illustrate the general theory.

Keywords: Classification algorithm; Multi-kernel regularization scheme; Convex loss function; Misclassification error; Regularization error and sample error

☆ The work described in this paper was fully supported by a grant from the Research Grants Council of Hong Kong Special Administrative Region, China [Project No. CityU 103303] and a grant from City University of Hong Kong [Project No. 7001816].

* Corresponding author.
E-mail addresses: qiang@stat.duke.edu (Q. Wu), y.ying@cs.ucl.ac.uk (Y. Ying), mazhou@cityu.edu.hk (D.-X. Zhou).

[1] Current address: Institute of Genome Sciences and Policy, Duke University, Durham, NC 27708, USA.

[2] Current address: Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, England, UK.

0885-064X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jco.2006.06.007

1. Introduction

We study binary classification algorithms generated from Tikhonov regularization schemes associated with general convex loss functions and multi-kernel spaces. These algorithms produce *binary classifiers* $\mathcal{C} : X \to \{1, -1\}$, from a compact metric space X (called input space) to the output space $Y = \{1, -1\}$ (representing the two classes). Such a classifier $\mathcal{C}$ yields for each point x the value $\mathcal{C}(x) \in Y$ which is a prediction made for x (when $X \subset \mathbb{R}^n$, x is a vector representing an event with each component corresponding to a specific measurement).

The classifiers considered here have the form $\mathcal{C} = \operatorname{sgn}(f)$, defined as $\operatorname{sgn}(f)(x) = 1$ if $f(x) \geqslant 0$ and $\operatorname{sgn}(f)(x) = -1$ if $f(x) < 0$, induced by real-valued functions. These functions are solutions of some optimization problems associated with a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$, independently drawn according to a (unknown) probability distribution ρ on $Z = X \times Y$. The nature of such an optimization problem (called a Tikhonov regularization scheme) is determined by two objects: a *loss function* and a *hypothesis space*.

Definition 1. A function $\phi : \mathbb{R} \to \mathbb{R}_+$ is called an activating loss (function) for classification if it is convex, $\phi'(0) < 0$, and $\inf_{t\in\mathbb{R}} \phi(t) = 0$.

Typical examples of activating loss include the hinge loss $\phi_h(t) = (1 - t)_+ = \max\{1 - t, 0\}$ for the support vector machine (SVM) classification and the exponential loss $\phi_{\exp}(t) = e^{-t}$ for boosting.

Let ϕ be an activating loss. For a real-valued function f, when $\operatorname{sgn}(f)$ is used for classification or prediction, the local error incurred for the event x and output y will be measured by the value $\phi(yf(x))$. The average of local errors is defined as $\mathcal{E}^\phi(f) = \int_Z \phi(yf(x))\, d\rho$, called the error or *generalization error*.

The convexity of ϕ tells us that the (one-side) derivative ϕ' is non-decreasing. This in connection with the condition $\phi'(0) < 0$ [3] implies that $\phi'(t) \leqslant \phi'(0) < 0$ for $t < 0$. It follows that when $yf(x) < 0$, i.e., when $\operatorname{sgn}(f)(x)$ predicts the class label y incorrectly, the local error is large: $\phi(yf(x)) > \phi(0) > 0$. So local errors are possibly small only if $yf(x) \geqslant 0$. Hence minimizing the generalization error is expected to lead to a function predicting the label satisfactorily. This gives the intuition that ϕ is admissible for classification problems, as verified by many examples in practice.

Since the generalization error involving the unknown distribution ρ is not computable, its discretization is used instead which, computable in terms of the sample $\mathbf{z}$, is defined as

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

and called the *empirical error*. Regularized learning schemes are implemented by minimizing a penalized version of the empirical error over a set of functions, called a *hypothesis space* $\mathcal{H}$, equipped with a functional $\Omega : \mathcal{H} \to \mathbb{R}_+$. The penalty functional Ω reflects constraints imposed on functions from the hypothesis space in various desirable forms.

Definition 2. Given a function $\phi : \mathbb{R} \to \mathbb{R}_+$ and a hypothesis space $\mathcal{H}$ together with a penalty functional Ω, the regularized classifier generated for a sample $\mathbf{z} \in Z^m$ is defined as $\operatorname{sgn}(f_{\mathbf{z}})$, where

$f_{\mathbf{z}}$ is a minimizer of the Tikhonov regularization scheme

$$f_{\mathbf{z}} := \arg\min_{f\in\mathcal{H}} \left\{ \frac{1}{m}\sum_{i=1}^{m} \phi\left(y_i f(x_i)\right) + \lambda\Omega(f) \right\}. \tag{1.1}$$

Here λ is a positive constant called the regularization parameter. It depends on m : $\lambda = \lambda(m)$, and usually $\lambda(m) \to 0$ as m becomes large.

Reproducing kernel Hilbert spaces (RKHSs) are often used as the hypothesis space in (1.1). They play an important role in learning theory because of their reproducing property.

Let $K : X \times X \to \mathbb{R}$ be continuous, symmetric and positive semidefinite, i.e., for any finite set of distinct points $\{x_1, \ldots, x_\ell\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^{\ell}$ is positive semidefinite. Such a function is called a *Mercer kernel*.

The RKHS $\mathcal{H}_K$ associated with the Mercer kernel K is defined (see [1]) 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 $\langle K_x, K_y\rangle_K = K(x, y)$. The reproducing property of $\mathcal{H}_K$ is

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

The classical soft margin classifier [41,12] corresponds to the scheme (1.1) with $\mathcal{H} = \mathcal{H}_K$:

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

In this paper we introduce a multi-kernel setting where $\mathcal{H}$ is the union of a set of RKHSs.

Definition 3. Let $K_\Sigma = \{K_\sigma : \sigma \in \Sigma\}$ be a set of Mercer kernels on X. The multi-kernel space associated with K_Σ is defined to be the union $\mathcal{H}_\Sigma = \bigcup_{\sigma\in\Sigma} \mathcal{H}_{K_\sigma}$. For $f \in \mathcal{H}_\Sigma$, we take

$$\|f\|_\Sigma = \inf\left\{ \|f\|_{K_\sigma} : \ f \in \mathcal{H}_{K_\sigma}, \sigma \in \Sigma \right\}, \tag{1.4}$$

where $\|f\|_{K_\sigma}$ is the RKHS norm of the function f in the RKHS $\mathcal{H}_{K_\sigma}$. Taking $\mathcal{H}_\Sigma$ as the hypothesis space and $\Omega(f) = \|f\|_\Sigma^2$ in (1.1) leads to the following scheme in the multi-kernel space $\mathcal{H}_\Sigma$:

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_\Sigma} \left\{ \frac{1}{m}\sum_{i=1}^{m} \phi\left(y_i f(x_i)\right) + \lambda\|f\|_\Sigma^2 \right\}. \tag{1.5}$$

The corresponding multi-kernel regularized classifier is given by $\operatorname{sgn}(f_{\mathbf{z}})$.

Note that $\mathcal{H}_\Sigma$ may not be a linear space. Denote $\left(\mathcal{H}_{K_\sigma}, \|\cdot\|_{K_\sigma}\right)$ as $(\mathcal{H}_\sigma, \|\cdot\|_\sigma)$ for simplicity. The regularization scheme in the multi-kernel space $\mathcal{H}_\Sigma$ can be rewritten as a two-layer minimization problem:

$$f_{\mathbf{z}} = \arg\min_{\sigma\in\Sigma} \min_{f\in\mathcal{H}_\sigma} \left\{ \frac{1}{m}\sum_{i=1}^{m} \phi\left(y_i f(x_i)\right) + \lambda\|f\|_\sigma^2 \right\}. \tag{1.6}$$

It reduces to (1.3) when Σ contains only one element.

Our study of general multi-kernel schemes is motivated by recent work on learning algorithms with varying kernels. In [10] SVMs with multiple parameters are investigated. In [22,29] mixture

density estimation is considered and Gaussian kernels with variance σ^2 flexible on an interval $[\sigma_1^2, \sigma_2^2]$ with $0 < \sigma_1 < \sigma_2 < +\infty$ are used for deriving bounds. Approximation properties of multi-kernel spaces are studied in [50]. Some algorithms for multi-task learning and learning the kernel function involve kernels from a convex hull of several Mercer kernels and spaces with changing norms, e.g. [18,20,27].

The first natural concern about the optimization problem (1.5) or (1.6) is the existence of a minimizer before efficient algorithms are searched. The existence is assured by the compactness of the index metric set Σ and the continuity of K_σ for $\sigma \in \Sigma$ in the next result following from Proposition 1 given in Section 2.

Theorem 1. *Let ϕ be an activating loss. If the index set Σ is a compact metric space, and for each pair (x, y), the function $K_\sigma(x, y)$ is continuous with respect to $\sigma \in \Sigma$, then a solution $f_{\mathbf{z}}$ to the multi-kernel scheme* (1.6) *exists.*

In particular, $f_{\mathbf{z}}$ exists in the one-kernel setting (1.3). We shall assume the existence of the optimization problem (1.6) throughout the error analysis of multi-kernel regularized classifiers, the main goal of this paper.

Let $(\mathcal{X}, \mathcal{Y})$ be the random variable on $X \times Y$ with the probability distribution ρ. The *misclassification error* for a classifier $\mathcal{C} : X \to Y$ is defined to be the probability of the event $\{\mathcal{C}(\mathcal{X}) \neq \mathcal{Y}\}$,

$$\mathcal{R}(\mathcal{C}) = \text{Prob}\{\mathcal{C}(\mathcal{X}) \neq \mathcal{Y}\} = \int_X P(\mathcal{Y} \neq \mathcal{C}(x)|x)\, d\rho_X. \tag{1.7}$$

Here ρ_X is the marginal distribution on X and $P(\cdot|x)$ is the conditional distribution. Our target of error analysis is to understand how $\text{sgn}(f_{\mathbf{z}})$ approximates the Bayes rule, the best classifier with respect to the misclassification error: $f_c = \arg\inf \mathcal{R}(\mathcal{C})$ with the infimum taken over all classifiers. Denote $\eta(x) = P(\mathcal{Y} = 1|x)$ and recall the regression function

$$f_\rho(x) = \int_Y y\, d\rho(y|x) = P(\mathcal{Y} = 1|x) - P(\mathcal{Y} = -1|x) = 2\eta(x) - 1, \quad x \in X. \tag{1.8}$$

Then the *Bayes rule* is given (e.g. [17]) by the sign of the regression function $f_c = \text{sgn}(f_\rho)$. Estimating the *excess misclassification error*

$$\mathcal{R}(\text{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \tag{1.9}$$

for the multi-kernel regularized classification algorithm (1.6) is our main purpose.

For the one-kernel setting (1.3) and special choices of ϕ, the error analysis has been extensively investigated in the literature, especially when ρ is strictly separable (with a positive margin). Examples of loss functions include

(1) hinge loss ϕ_h for SVM [41,30,35,13,44];
(2) $\phi_q(t) = (1-t)_+^q$ for the SVM q-norm ($q > 1$) soft margin classifier, see [41,23,11];
(3) least-square loss $\phi_{\text{ls}}(t) = (1-t)^2$, see e.g. [14,17,19,28,34,37,47];
(4) exponential loss $\phi_{\exp}(t) = e^{-t}$, see [47,5,24];
(5) logistic regression $\phi(t) = \log(1 + e^{-t})$ or $1/(1 + e^t)$, see [47,5].

For the error bounds, we will focus on activating loss functions achieving zeros, which allows us to provide a powerful analysis.

Definition 4. An activating loss is called a classifying loss for classification if $\phi(t_0) = 0$ for some $t_0 \in \mathbb{R}$. It is called normalized if 1 is the minimal zero of ϕ.

Examples of classifying loss include the hinge loss ϕ_h, the q-norm loss ϕ_q for SVM classification and the least-square loss $\phi_{\text{ls}}(t) = (1-t)^2$. They are all normalized.

Our error analysis will be done in Sections 3–5. It uses an error decomposition procedure for regularization scheme introduced in [11,43], with the aid of an *iteration technique* [36,43] and a *projection operator* hyperlinkbib11[11]. The convergence rates will be stated in terms of the sample size m with proper choices of the regularization parameter $\lambda = \lambda(m) \to 0$. Our analysis yields fast convergence rates which might be improved further in some situations [33]. Let us demonstrate the convergence rates in the SVM case.

Assume $X \subset \mathbb{R}^n$ and for some $s > n$, the multi-kernels K_Σ satisfy

$$\sup_{\sigma\in\Sigma} \|K_\sigma\|_{C^s(X\times X)} < \infty. \tag{1.10}$$

It means that $\{K_\sigma : \sigma \in \Sigma\}$ is a set of C^s Mercer kernels with a uniform bound. Here the C^s norm equals $\|K_\sigma\|_{C^s(X\times X)} := \max_{\alpha_1+\cdots+\alpha_{2n}\leqslant s} \|\frac{\partial^{\alpha_1+\cdots+\alpha_{2n}}}{\partial x_1^{\alpha_1}\cdots\partial x_n^{\alpha_n}\partial y_1^{\alpha_{n+1}}\cdots\partial y_n^{\alpha_{2n}}} K\|_{C(X\times X)}$. The convergence rate for SVM with such multi-kernels which will be proved in Section 5 can be stated as follows.

Theorem 2. *Let $\phi = \phi_h$ and $f_{\mathbf{z}}$ be given by* (1.6). *Assume that for some $0 < \beta \leqslant 1$ and $c_\beta > 0$, we have*

$$\inf_{\sigma\in\Sigma} \inf_{f\in\mathcal{H}_\sigma} \left\{ \|f - f_c\|_{L^1_{\rho_X}} + \lambda \|f\|_\sigma^2 \right\} \leqslant c_\beta \lambda^\beta \quad \forall \lambda > 0. \tag{1.11}$$

If (1.10) *holds for some $s > n$, choose $\lambda(m) = \left(\frac{1}{m}\right)^{\min\left\{\frac{1}{2\beta+(1-\beta)n/s}, \frac{2}{1+\beta}\right\}}$. For any $\varepsilon > 0$ and $0 < \delta < 1$, there exists a constant $\widetilde{c}$ independent of m such that with confidence $1-\delta$,*

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leqslant \widetilde{c} \left(\frac{1}{m}\right)^\theta, \tag{1.12}$$

where $\theta = \min\left\{\frac{\beta}{2\beta+(1-\beta)n/s} - \varepsilon, \frac{2\beta}{1+\beta}\right\}$.

In Theorem 2, ε can be arbitrarily small. Hence the power θ in the learning rate (1.12) is arbitrarily close to $\min\left\{\frac{\beta}{2\beta+(1-\beta)n/s}, \frac{2\beta}{1+\beta}\right\}$. When the kernels are C^∞ and (1.10) holds for any $s > 0$, we see that θ can be arbitrarily close to $\min\left\{\frac{1}{2}, \frac{2\beta}{1+\beta}\right\}$ which equals to $\frac{1}{2}$ when $\beta \geqslant \frac{1}{3}$.

The condition (1.11) measures the approximation power of the multi-kernel space $\mathcal{H}_\Sigma$ in $L^1_{\rho_X}$, acting on the function $f_c = \operatorname{sgn}(f_\rho)$ which involves only the sign of f_ρ. It can be described by some interpolation spaces of the pair $(\mathcal{H}_\Sigma, L^1_{\rho_X})$.

We only assume conditions on the approximation power (1.11) and the smoothness (1.10) in Theorem 2. If further information about the distribution ρ is available, one can expect sharper error estimates. For example, when ρ satisfies a so-called Tsybakov noise condition [39]

$$\rho_X(\{x \in X : 0 < |f_\rho(x)| \leqslant \Delta t\}) \leqslant t^\zeta \quad \forall t > 0, \tag{1.13}$$

with some $\zeta \in [0, \infty]$ and $\Delta > 0$, then the power θ in the error bound (1.12) can be improved to $\theta = \min\left\{\frac{\beta(\zeta+1)}{\beta(\zeta+2)+(\zeta+1-\beta)n/s} - \varepsilon, \frac{2\beta}{1+\beta}\right\}$. This will be shown in Theorem 6 below (in Section 5).

Note that any distribution satisfies (1.13) with $\zeta = 0$. The case $\zeta = \infty$ is the same as $|f_\rho(x)| \geqslant \Delta$ or $f_\rho(x) = 0$, meaning that the two classes are well separated.

Our result is new for the multi-kernel setting. Even for the one-kernel setting $\mathcal{H}_\Sigma = \mathcal{H}_K$, Theorem 2 provides the best convergence rate for the SVM under the same assumption (1.11) of the approximation power of $\mathcal{H}_K$ and the regularity condition of the kernel ($K \in C^s$ with $s > n$): the capacity independent estimates derived by Zhang [47] yield the learning rate (1.12) with $\theta = \beta/(1+\beta)$; under the noise condition (1.13) and some moment conditions on the probability distribution, Steinwart and Scovel [36] obtained the learning rate (1.12) with $\theta = \frac{2\beta(\zeta+1)}{(2+\zeta+\zeta n/s)(1+\beta)} - \varepsilon$. Since $s > n$, our rate is sharper than theirs.

2. Optimization problem for regularization with multi-kernels

We divide the study of the optimization problem (1.6) in two steps.

First, fix $\sigma \in \Sigma$. Denote the optimal solution in the RKHS $\mathcal{H}_\sigma$ as

$$f_{\mathbf{z},\sigma} = \arg\min_{f\in\mathcal{H}_\sigma} \left\{ \frac{1}{m}\sum_{i=1}^{m} \phi(y_i f(x_i)) + \lambda \|f\|_\sigma^2 \right\}.$$

To solve this problem by a dual argument in optimization theory, we define the dual function $\psi : \mathbb{R} \to \mathbb{R}$ of ϕ by

$$\psi(v) = \sup_{u\in\mathbb{R}} \{vu - \phi(u)\}, \quad v \in \mathbb{R}. \tag{2.1}$$

By the reproducing property (1.2), the optimization problem for solving $f_{\mathbf{z},\sigma}$ on $\mathcal{H}_\sigma$ can be reduced into one on $\mathbb{R}^m$. The following relation between the primal problem and its dual is well known (see e.g. [46]):

$$\inf_{f\in\mathcal{H}_\sigma} \left\{ \frac{1}{m}\sum_{i=1}^{m} \phi(y_i f(x_i)) + \lambda \|f\|_\sigma^2 \right\} = \sup_{\alpha\in\mathbb{R}^m} \left\{ \hat{R}(\alpha,\sigma) \right\},$$

where

$$\hat{R}(\alpha,\sigma) := -\frac{1}{m}\sum_{i=1}^{m} \psi(-\alpha_i y_i) - \frac{1}{4m^2\lambda} \sum_{i,j=1}^{m} \alpha_i K_\sigma(x_i, x_j)\alpha_j, \quad \alpha \in \mathbb{R}^m.$$

Moreover, both optimizers exist. If $\hat{\alpha}_\sigma = \arg\max_{\alpha\in\mathbb{R}^m} \hat{R}(\alpha,\sigma)$, then $(\hat{\alpha}_\sigma)_i y_i \geqslant 0$ and

$$f_{\mathbf{z},\sigma}(x) = \frac{1}{2\lambda m}\sum_{i=1}^{m} \left(\hat{\alpha}_\sigma\right)_i K_\sigma(x_i, x).$$

Next, consider the multi-kernel scheme (1.6). A solution $f_{\mathbf{z}}$ can be represented as

$$f_{\mathbf{z}}(x) = \frac{1}{2\lambda m}\sum_{i=1}^{m} \hat{\alpha}_i K_{\hat{\sigma}}(x_i, x)$$

if an optimal point $(\hat{\alpha}, \hat{\sigma})$ of the following "dual problem" exists:

$$(\hat{\alpha}, \hat{\sigma}) = \arg\min_{\sigma\in\Sigma} \max_{\alpha\in\mathbb{R}^m} \left\{\hat{R}(\alpha, \sigma)\right\}. \tag{2.2}$$

We show that under some mild condition, (2.2) can be solved.

Proposition 1. *Under the conditions of Theorem* 1, *an optimal point* $(\hat{\alpha}, \hat{\sigma})$ *of* (2.2) *can be achieved. Hence an optimal solution* $f_{\mathbf{z}}$ *to the multi-kernel regularization scheme* (1.5) *always exists*.

The proof of Proposition 1 will be given in the Appendix.

Example 1. Let $\Sigma = [\sigma_1, \sigma_2]$ with $0 < \sigma_1 \leqslant \sigma_2 < \infty$ and K_σ be the Gaussian kernel $K_\sigma(x, y) = \exp\left\{-\frac{|x-y|^2}{2\sigma^2}\right\}$ on a compact subset X of $\mathbb{R}^n$. Then a solution to the optimization problem (1.6) exists.

It would be interesting to consider the existence when $\Sigma = (0, \infty)$.

3. Error analysis: a general framework

In this section, we give a general framework of our error analysis, consisting of a comparison theorem (reducing (1.9) to an excess generalization error), a projection operator (making random variables uniformly bounded) and an error decomposition procedure (decomposing the excess generalization error into a sum of a regularization error and a sample error). Then the framework provides bounds for the excess misclassification error in terms of a regularization error and a sample error, studied in the next two sections separately.

3.1. Comparison theorems

Similar to the learning rate stated in Theorem 2, the error analysis aims at bounding the excess misclassification error $\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c)$. But the algorithm is designed by minimizing a penalized empirical error $\mathcal{E}_{\mathbf{z}}^{\phi}$ associated with the loss function ϕ. Knowledge of regularization schemes or empirical risk minimization processes would only lead us to expect the convergence of $\mathcal{E}^{\phi}(f_{\mathbf{z}})$ as $m \to \infty$. So relations between misclassification error and generalization error become crucial. Some work on this topic includes [5,47,3]. Here we only mention some comparison theorems which will be used in the paper.

Denote $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. Define

$$f_\rho^{\phi} = \arg\min \mathcal{E}^{\phi}(f)$$

with the minimum taken over all functions $f : X \to \overline{\mathbb{R}}$. Note that f_ρ^{ϕ} always exists since ϕ is convex. It satisfies $\mathrm{sgn}(f_\rho^{\phi}) = f_c$, an admissible condition for the loss function, see [34,3]. Comparison theorems enable us to bound the excess misclassification error (1.9) by estimates for the excess generalization error $\mathcal{E}^{\phi}(f) - \mathcal{E}^{\phi}(f_\rho^{\phi})$.

Proposition 2. *Let* $\phi = \phi_h$ *be the hinge loss. We have* $f_\rho^{\phi_h} = f_c$ *and for every measurable function* $f : X \to \mathbb{R}$,

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c) \leqslant \mathcal{E}^{\phi_h}(f) - \mathcal{E}^{\phi_h}(f_c). \tag{3.1}$$

Proposition 3. *If an activating loss ϕ satisfies $\phi''(0) > 0$, then there exists a constant $c_\phi > 0$ such that for any measurable function $f : X \to \mathbb{R}$, there holds*

$$\mathcal{R}(f) - \mathcal{R}(f_c) \leqslant c_\phi \sqrt{\mathcal{E}^\phi(f) - \mathcal{E}^\phi(f_\rho^\phi)}.$$

The fact $f_c = f_\rho^{\phi_h}$ was proved in [42]. The relation (3.1) in the first comparison theorem, Proposition 2, was proved in [47].

The second comparison theorem, Proposition 3, deals with general activating loss functions. It was explicitly given in [11] following the general results in [3]. Note that if $\phi''(0)$ exists then the convexity of ϕ implies $\phi''(0) \geqslant 0$.

Tighter comparison bounds are possible under some noise conditions. We say that ρ has a Tsybakov *noise exponent* $\alpha \geqslant 0$ if for some $c_\alpha > 0$ and every measurable $f : X \to Y$,

$$\rho_X(\{x \in X : f(x) \neq f_c(x)\}) \leqslant c_\alpha (\mathcal{R}(f) - \mathcal{R}(f_c))^\alpha. \tag{3.2}$$

All distributions satisfy (3.2) with $\alpha = 0$ and $c_\alpha = 1$. The following sharper comparison bound for $\alpha > 0$ follows immediately from [3] which can also be seen from [5, Lemma 6] and Proposition 3.

Corollary 1. *Let ϕ be a classifying loss satisfying $\phi''(0) > 0$. If ρ satisfies the Tsybakov noise condition* (3.2) *for some $\alpha \in [0, 1]$ and $c_\alpha > 0$, then*

$$\mathcal{R}(\operatorname{sgn}(f)) - \mathcal{R}(f_c) \leqslant \left\{2c_\phi c_\alpha \left(\mathcal{E}^\phi(f) - \mathcal{E}^\phi(f_\rho^\phi)\right)\right\}^{1/(2-\alpha)} \quad \forall f : X \to \mathbb{R}.$$

3.2. Projection operator

By comparison theorems, we only need to bound the excess generalization error $\mathcal{E}^\phi(f_{\mathbf{z}}) - \mathcal{E}^\phi(f_\rho^\phi)$ in order to study the performance of the classifier $\operatorname{sgn}(f_{\mathbf{z}})$. But we can do better using the special feature of a classifying loss that it achieves a zero. A key technical tool here is a projection operator.

To simply the notations and statements, we will restrict our discussion only for normalized classifying loss functions. For such a loss function ϕ, we can choose a minimizer f_ρ^ϕ of $\mathcal{E}^\phi(f)$ such that $f_\rho^\phi(x) \in [-1, 1]$ on X. To see this, we set a univariate convex function Q for $x \in X$ as

$$Q(t) = Q_x(t) := \int_Y \phi(yt)\, d\rho(y|x), \quad t \in \mathbb{R}. \tag{3.3}$$

Its one-side derivatives exist, are non-decreasing and satisfy $Q'_-(t) \leqslant Q'_+(t)$ for every $t \in \mathbb{R}$.

Denote

$$f_\rho^-(x) = \sup\left\{t \in \mathbb{R} : Q'_-(t) < 0\right\}, \quad f_\rho^+(x) = \inf\left\{t \in \mathbb{R} : Q'_+(t) > 0\right\}.$$

Theorem 3. *Let ϕ be a normalized classifying loss function. Then*

(a) *for each $x \in X$, the univariate function Q given by* (3.3) *is strictly decreasing on $(-\infty, f_\rho^-(x)]$, strictly increasing on $[f_\rho^+(x), +\infty)$, and is constant on $[f_\rho^-(x), f_\rho^+(x)]$.*

(b) $f_\rho^\phi : X \to \mathbb{R}$ *is a minimizer of the generalization error* $\mathcal{E}^\phi(f)$ *if and only if for almost every* $x \in (X, \rho_X)$, $f_\rho^\phi(x)$ *is a minimizer of* Q, *that is, there holds*

$$f_\rho^-(x) \leqslant f_\rho^\phi(x) \leqslant f_\rho^+(x). \tag{3.4}$$

(c) *We may choose a minimizer* f_ρ^ϕ *of* $\mathcal{E}^\phi$ *satisfying* $f_\rho^\phi(x) \in [-1, 1]$ *for each* $x \in X$.

Proof. Let $x \in X$. Consider the univariate continuous function Q given by (3.3). It is strictly decreasing on the interval $(-\infty, f_\rho^-(x))$, since $Q'_-(t) < 0$ on this interval. In the same way, $Q'_+(t) > 0$ for $t > f_\rho^+(x)$, so Q is strictly increasing on $(f_\rho^+(x), +\infty)$. For $t \in (f_\rho^-(x), f_\rho^+(x))$, we have $0 \leqslant Q'_-(t) \leqslant Q'_+(t) \leqslant 0$, hence Q is constant which is the minimal value of Q on $\mathbb{R}$. This proves (a).

Since $\mathcal{E}^\phi(f) = \int_X Q_x(f(x))\, d\rho_X(x)$, the statement (b) follows directly from (a).

By the assumption, ϕ is convex and has minimal zero 1. This implies that ϕ is strictly decreasing on $(-\infty, 1]$ and non-decreasing on $[1, +\infty)$. So $Q(t) \geqslant Q(1)$ for $t > 1$ and $Q(t) \geqslant Q(-1)$ for $t < -1$. So a minimum of Q can always be achieved on $[-1, 1]$. Hence we may choose f_ρ^ϕ such that $f_\rho^\phi(x) \in [-1, 1]$. This proves the statement (c). □

In what follows we shall always choose f_ρ^ϕ with $|f_\rho^\phi(x)| \leqslant 1$ for normalized classifying loss functions. Then we can make full use of the projection operator introduced in [11].

Definition 5. The projection operator π is defined on the space of measurable functions $f : X \to \mathbb{R}$ as

$$\pi(f)(x) = \begin{cases} 1 & \text{if } f(x) > 1, \\ -1 & \text{if } f(x) < -1, \\ f(x) & \text{if } -1 \leqslant f(x) \leqslant 1. \end{cases} \tag{3.5}$$

It is easy to see that $\pi(f)$ and f induce the same classifier, i.e., $\operatorname{sgn}(\pi(f)) = \operatorname{sgn}(f)$. Apply this fact to comparison theorems. It is sufficient for us to bound the excess generalization error for $\pi(f_{\mathbf{z}})$ instead of $f_{\mathbf{z}}$. This leads to better estimates, as we will see later.

The following property of the projection operator is immediate from the definition of ϕ.

Proposition 4. *If* ϕ *is a normalized classifying loss function, then there holds almost surely*

$$\phi(y\pi(f)(x)) \leqslant \phi(yf(x)). \tag{3.6}$$

Hence for any measurable function f, we have $\mathcal{E}^\phi(\pi(f)) \leqslant \mathcal{E}^\phi(f)$ *and* $\mathcal{E}_{\mathbf{z}}^\phi(\pi(f)) \leqslant \mathcal{E}_{\mathbf{z}}^\phi(f)$.

3.3. Error decomposition

Now we can present the *error decomposition* which leads to bounds of the excess generalization error for $\pi(f_{\mathbf{z}})$. Define

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

Proposition 5. *Let ϕ be a normalized classifying loss and $f_{\mathbf{z}}$ given by* (1.6). *Then*

$$\mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) + \lambda\|f_{\mathbf{z}}\|_{\Sigma}^2 \leqslant \mathcal{D}(\lambda) + \mathcal{S}_{\mathbf{z},\lambda}, \tag{3.7}$$

where $\mathcal{D}(\lambda)$ is the regularization error of the multi-kernel space $\mathcal{H}_{\Sigma}$ defined [32] *as*

$$\mathcal{D}(\lambda) = \inf_{\sigma\in\Sigma}\inf_{f\in\mathcal{H}_{\sigma}}\left\{\mathcal{E}^{\phi}(f) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) + \lambda\|f\|_{\sigma}^2\right\} \tag{3.8}$$

and

$$\mathcal{S}_{\mathbf{z},\lambda} = \left\{\mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}^{\phi}(\pi(f_{\mathbf{z}}))\right\} + \left\{\mathcal{E}_{\mathbf{z}}^{\phi}(f_{\lambda}) - \mathcal{E}^{\phi}(f_{\lambda})\right\}. \tag{3.9}$$

Proof. Write $\mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) + \lambda\|f_{\mathbf{z}}\|_{\Sigma}^2$ as

$$\left\{\mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}^{\phi}(\pi(f_{\mathbf{z}}))\right\} + \left\{\left(\mathcal{E}_{\mathbf{z}}^{\phi}(\pi(f_{\mathbf{z}})) + \lambda\|f_{\mathbf{z}}\|_{\Sigma}^2\right) - \left(\mathcal{E}_{\mathbf{z}}^{\phi}(f_{\lambda}) + \lambda\|f_{\lambda}\|_{\Sigma}^2\right)\right\}$$
$$+ \left\{\mathcal{E}_{\mathbf{z}}^{\phi}(f_{\lambda}) - \mathcal{E}^{\phi}(f_{\lambda})\right\} + \left\{\mathcal{E}^{\phi}(f_{\lambda}) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) + \lambda\|f_{\lambda}\|_{\Sigma}^2\right\}.$$

By Proposition 4, $\mathcal{E}_{\mathbf{z}}^{\phi}(\pi(f_{\mathbf{z}})) \leqslant \mathcal{E}_{\mathbf{z}}^{\phi}(f_{\mathbf{z}})$. This in connection with the definition of $f_{\mathbf{z}}$ tells us that the second term is $\leqslant 0$. Note that $\mathcal{S}_{\mathbf{z},\lambda}$ is just the sum of the first and third terms. By the definition of f_{λ}, the last term equals to $\mathcal{D}(\lambda)$. This proves (3.7). □

The regularization error term $\mathcal{D}(\lambda)$ in the error decomposition (3.7) is independent of the sample and will be discussed in Section 4.

The last term $\mathcal{S}_{\mathbf{z},\lambda}$ in (3.7) is called the *sample error*. Without projection, it is well understood because of the vast literature in learning theory, see [7] and references therein. We are able to improve the sample error estimates, stated in Theorem 5 below, because of the projection operator.

Comparison theorems and the error decomposition help switch the goal of the error analysis to the estimation of the regularization error and the sample error. For instance, to prove Theorem 2, we first apply Proposition 2 to $\pi(f_{\mathbf{z}})$ and then Proposition 5. It tells us that $\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leqslant \mathcal{D}(\lambda) + \mathcal{S}_{\mathbf{z},\lambda}$.

4. Estimating regularization error and approximation error

In this section, we discuss the estimation of the regularization error $\mathcal{D}(\lambda)$ which is non-random and is also called the approximation error. The convexity of ϕ implies that $\phi'_{-}(t) = \phi'_{+}(t) = \phi'(t)$ for almost every $t \in \mathbb{R}$.

Theorem 4. *Let ϕ be a normalized classifying loss. Then*

$$\mathcal{E}^{\phi}(f) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) \leqslant \|\phi'\|_{L^{\infty}[-\|f\|_{\infty},\|f\|_{\infty}]}\|f - f_{\rho}^{\phi}\|_{L^1_{\rho_X}}.$$

If moreover, ϕ is C^1 and ϕ' is absolutely continuous on $\mathbb{R}$, we have

$$\mathcal{E}^{\phi}(f) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) \leqslant \|\phi''\|_{L^{\infty}[-\|f\|_{\infty}-1,\|f\|_{\infty}+1]}\|f - f_{\rho}^{\phi}\|_{L^2_{\rho_X}}^2.$$

Proof. With the function $Q = Q_x$ defined in (3.3), write $\mathcal{E}^{\phi}(f) - \mathcal{E}^{\phi}(f_{\rho}^{\phi})$ as

$$\mathcal{E}^{\phi}(f) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) = \int_X \left\{Q(f(x)) - Q(f_{\rho}^{\phi}(x))\right\} d\rho_X.$$

Since $\phi'(0) < 0$ and $\phi(t) \geqslant 0$, we have $\phi(0) > 0$ and $\phi'_{\pm}(t) < 0$ for $t < 0$. Let $P(t) = \max\left\{\phi'_{\pm}(t), -\phi'_{\pm}(-t)\right\}$ for $t > 0$. We only need to prove

$$Q(f(x)) - Q(f_\rho^\phi(x)) \leqslant P(|f(x)|)|f(x) - f_\rho^\phi(x)| \tag{4.1}$$

for those x with $Q(f(x)) - Q(f_\rho^\phi(x)) > 0$. According to Theorem 3, such a point x satisfies $f(x) \notin [f_\rho^-(x), f_\rho^+(x)]$.

If $f(x) > f_\rho^+(x)$, then Q is strictly increasing on $[f(x), +\infty)$. Hence $f(x) > f_\rho^\phi(x)$. By Theorem 3, we have

$$Q(f(x)) - Q(f_\rho^\phi(x)) \leqslant Q'_-(f(x))\left(f(x) - f_\rho^\phi(x)\right).$$

The convexity of ϕ implies that the one-side derivatives ϕ'_+ and ϕ'_- exist, are non-decreasing, and satisfy $\phi'_-(t) \leqslant \phi'_+(t)$ for any $t \in \mathbb{R}$. Note that $Q(t) = \eta(x)\phi(t) + (1 - \eta(x))\phi(-t)$. Hence

$$\begin{aligned} Q'_-(f(x)) &= \eta(x)\phi'_-(f(x)) - (1 - \eta(x))\phi'_+(-f(x)) \\ &\leqslant \max\left\{\phi'_{\pm}(|f(x)|), -\phi'_{\pm}(-|f(x)|)\right\} \end{aligned}$$

no matter whether $f(x) \geqslant 0$ or not. Thus, (4.1) holds true when $f(x) > f_\rho^+(x)$.

In the same way, if $f(x) < f_\rho^-(x)$, then Q is strictly decreasing on $(-\infty, f(x)]$. Hence $f(x) < f_\rho^\phi(x)$. Theorem 3 yields again

$$Q(f(x)) - Q(f_\rho^\phi(x)) \leqslant -Q'_+(f(x))\left(f_\rho^\phi(x) - f(x)\right).$$

Since $-Q'_+(f(x)) = -\eta(x)\phi'_+(f(x)) + (1 - \eta(x))\phi'_-(-f(x)) \leqslant P(|f(x)|)$, we see that (4.1) also holds when $f(x) < f_\rho^-(x)$. This proves the first statement.

If ϕ is C^1 and ϕ' is absolutely continuous on $\mathbb{R}$, we know from Theorem 3 that $Q'(f_\rho^\phi(x)) = 0$. Hence

$$Q(f(x)) - Q(f_\rho^\phi(x)) = \int_{f_\rho^\phi(x)}^{f(x)} Q'(u) - Q'(f_\rho^\phi(x))\,du \leqslant \frac{\|Q''\|_{L^\infty(I)}}{2}|f(x) - f_\rho^\phi(x)|^2,$$

where I is the interval between $f_\rho^\phi(x)$ and $f(x)$. Then the second statement follows. □

In the above, $L^q_{\rho_X}$ is the L^q space with norm $\|f\|_{L^q_{\rho_X}} = \left\{\int_X |f(x)|^q\,d\rho_X\right\}^{1/q}$. Thus, we can use the rich knowledge from approximation theory to estimate the regularization error. See [11] for details on bounding the regularization error for the SVM q-norm soft margin classifiers by means of K-functionals in $L^q_{\rho_X}$.

One advantage of multi-kernel algorithms is the improvement of regularization errors compared with the one-kernel setting. Let us show this by the example of Gaussian kernels and least-square loss ϕ_{ls}. Here $\phi_{\text{ls}}(yf(x)) = (1 - yf(x))^2 = (y - f(x))^2$ since $y^2 = 1$ for $y \in Y$. So we know [41] that $f_\rho^\phi = f_\rho$ and $\mathcal{E}^{\phi_{\text{ls}}}(f) - \mathcal{E}^{\phi_{\text{ls}}}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}}$.

Example 2. Let $\phi = \phi_{\text{ls}}$ and K_σ be the Gaussian kernel $K_\sigma(x, y) = \exp\left\{-\frac{|x-y|^2}{2\sigma^2}\right\}$ on a compact domain X of $\mathbb{R}^n$ with piecewise smooth boundary. Assume ρ_X is the Lebesgue measure on X.

(1) If $\Sigma = \{\sigma\}$ corresponding to a single Gaussian kernel with variance $\sigma > 0$, then $\mathcal{D}(\lambda) = O(\lambda^\varepsilon)$ for some $\varepsilon > 0$ only if $f_\rho \in C^\infty(X)$.

(2) If $\Sigma = (0, \infty)$ and $f_\rho \in C^1(X)$, then $\mathcal{D}(\lambda) = O(\lambda^{\frac{1}{n+1}})$.

The first statement follows from the analysis in [31] or [16] on the approximation error, since $\|f_\lambda\|_{K_\sigma} \leqslant 1/\sqrt{\lambda}$ and $\mathcal{D}(\lambda) = O(\lambda^\varepsilon)$ implies $\inf_{\|f\|_{K_\sigma} \leqslant R}\{\|f - f_\rho\|^2_{L^2_{\rho_X}}\} = O(R^{-2\varepsilon})$. The error bound in the second statement was achieved [45] by $\sigma = \lambda^{\frac{1}{2n+2}} \in (0, \infty)$. For details on deriving satisfactory learning rates in the case $\Sigma = (0, \infty)$, see [45] where the sample error analysis was done by means of empirical covering numbers. Note that the uniform smoothness condition (1.10) with $s > 0$ does not hold in this case.

More examples and discussion can be found in [50,36,31,45].

5. Sample error estimates and learning rates

We are in a position to estimate the sample error and derive the learning rates. Throughout this section, we assume that the kernels are uniformly bounded in the sense that

$$\kappa := \sup_{\sigma\in\Sigma} \|K\|_{C(X\times X)} < \infty. \tag{5.1}$$

To state our result, we need to further introduce several concepts and notations.

The quantity $\mathcal{E}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi_{\mathbf{z}}(\pi(f_{\mathbf{z}}))$ in the sample error (3.9) needs to be estimated by some uniform law of large numbers. To this end, we need the capacity of the hypothesis space, which plays an essential role in sample error estimates. In this paper, we use the covering numbers measured by empirical distances.

Definition 6. Let $\mathcal{F}$ be a set of functions on Z and $\mathbf{x} = \{z_1, \ldots, z_m\} \subset Z$. The metric $d_{2,\mathbf{x}}$ is defined on $\mathcal{F}$ by

$$d_{2,\mathbf{x}}(f, g) = \left\{\frac{1}{m}\sum_{i=1}^m (f(z_i) - g(z_i))^2\right\}^{1/2}.$$

For every $\varepsilon > 0$, the covering number of $\mathcal{F}$ with respect to $d_{2,\mathbf{x}}$ is defined as

$$\mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \varepsilon) = \inf\left\{\ell \in \mathbb{N} : \exists\{f_i\}_{i=1}^\ell \subset \mathcal{F} \text{ such that } \mathcal{F} = \bigcup_{i=1}^\ell \{f \in \mathcal{F} : d_{2,\mathbf{x}}(f, f_i) \leqslant \varepsilon\}\right\}.$$

The function sets in our situation are balls of the multi-kernel space in the form of $B_R = \{f \in \mathcal{H}_\Sigma : \|f\|_\Sigma \leqslant R\} = \bigcup_{\sigma\in\Sigma}\{f \in \mathcal{H}_\sigma : \|f\|_\sigma \leqslant R\}$. We need the empirical covering number of B_1 defined as

$$\mathcal{N}(\varepsilon) = \sup_{m\in\mathbb{N}} \sup_{\mathbf{x}\in X^m} \mathcal{N}_{2,\mathbf{x}}(B_1, \varepsilon). \tag{5.2}$$

Note that for any function set $\mathcal{F} \subset C(X)$, the empirical covering number $\mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \varepsilon)$ is bounded by $\mathcal{N}(\mathcal{F}, \varepsilon)$, the (uniform) covering number of $\mathcal{F}$ under the metric $\|\cdot\|_\infty$, since $d_{2,\mathbf{x}}(f, g) \leqslant \|f - g\|_\infty$. So in the multi-kernel setting, the behavior of the covering number $\mathcal{N}(\varepsilon)$ can be estimated by the uniform smoothness of kernels in Σ according to [49].

Example 3. If the set Σ of kernels on $X \subset \mathbb{R}^n$ satisfies (1.10) for some $s > 0$, then there is a constant $c_s > 0$ such that $\log \mathcal{N}(\varepsilon) \leqslant c_s\,(1/\varepsilon)^{2n/s}$ for any $\varepsilon > 0$.

For a function $f : Z \to \mathbb{R}$, denote $\mathbb{E}f = \int_Z f(z)\,d\rho$.

Theorem 5. *Let ϕ be a normalized classifying loss. Assume the following conditions with exponents $q > 0$, $\tau \in [0, 1]$ and $p \in (0, 2)$:*

(1) *an increment condition for ϕ with a constant $c_q > 0$,*

$$|\phi(t))| \leqslant c_q |t|^q \quad \forall |t| \geqslant 1, \tag{5.3}$$

(2) *a variance–expectation bound for the pair (ϕ, ρ) with the exponent τ and some $c_\tau > 0$,*

$$\mathbb{E}\left\{\left(\phi(yf(x)) - \phi(yf_\rho^\phi(x))\right)^2\right\} \leqslant c_\tau \left\{\mathcal{E}^\phi(f) - \mathcal{E}^\phi(f_\rho^\phi)\right\}^\tau \quad \forall \|f\|_\infty \leqslant 1, \tag{5.4}$$

(3) *a capacity condition for the function set B_1 with a constant $c_p > 0$*

$$\log \mathcal{N}(\varepsilon) \leqslant c_p \left(\frac{1}{\varepsilon}\right)^p \quad \forall \varepsilon, R > 0, \ \ m \in \mathbb{N}. \tag{5.5}$$

If $\mathcal{D}(\lambda) \leqslant c_\beta \lambda^\beta$ for some $0 < \beta \leqslant 1$ and $c_\beta > 0$, then for any $\varepsilon > 0$ and $0 < \delta < 1$, there exists a constant $\tilde{c}$ independent of m such that, with $\lambda = \lambda(m) = \left(\frac{1}{m}\right)^\gamma$, we have

$$\mathcal{E}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi(f_\rho^\phi) \leqslant \tilde{c} \left(\frac{1}{m}\right)^\theta \tag{5.6}$$

with confidence $1 - \delta$, where

$$\gamma = \min\left\{\frac{2}{\beta(4 - 2\tau + p\tau) + p(1 - \beta)}, \frac{2}{2\beta + q - \beta q}\right\}, \tag{5.7}$$

$$\theta = \min\left\{\frac{2\beta}{\beta(4 - 2\tau + p\tau) + p(1 - \beta)} - \varepsilon, \frac{2\beta}{2\beta + q - \beta q}\right\}. \tag{5.8}$$

The proof of Theorem 5 will be given at the end of this section by using a local Rademacher process.

The increment condition (5.3) is satisfied for many useful loss functions including the hinge loss and least-square loss.

The variance–exponent condition (5.4) for the pair (ϕ, ρ) always holds for $\tau = 0$ with $c_\tau = (\max\{\phi(-1), \phi(1)\})^2$. This can be seen from the fact that $|\phi(yf(x)) - \phi(yf_\rho^\phi(x))| \leqslant \max\{\phi(-1), \phi(1)\}$. Larger exponents τ are possible when ϕ has high convexity (such as ϕ_{ls} in Theorem 7 below) or when the distribution ρ satisfies some conditions (such as the Tsybakov noise condition (1.13) in Theorem 6 below).

Besides Example 3, the capacity condition (5.5) always holds with $p \leqslant 2$ if K_Σ contains only one kernel.

The regularization error $\mathcal{D}(\lambda)$ decays to zero once $\mathcal{H}_\Sigma$ is dense in $C(X)$. By the discussion in Section 4, the decay rate with an exponent β can be estimated if some a priori knowledge on the distribution is available; see [11] for explicit examples.

Let us now show how to apply Theorem 5 to derive learning rates.

Recall Proposition 3 and Corollary 1. A direct corollary of Theorem 5 is as follows.

Corollary 2. *Under the assumption of Theorem* 5, *if* $\phi''(0) > 0$, *then for any* $\varepsilon > 0$ *and* $0 < \delta < 1$, *there is a constant* $\widetilde{c}$ *independent of* m *such that with confidence* $1 - \delta$,

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leqslant \widetilde{c}\left(\frac{1}{m}\right)^{\theta/2}, \tag{5.9}$$

where $\lambda = \left(\frac{1}{m}\right)^{\gamma}$, γ, θ *are given by* (5.7) *and* (5.8), *respectively. If, in addition,* ρ *satisfies the noise condition* (3.2) *with* $0 < \alpha \leqslant 1$, *the power* $\frac{\theta}{2}$ *in* (5.9) *can be improved to* $\frac{1}{2-\alpha}\theta$.

Next we consider two classical classification algorithms: SVM classification and least-square method.

5.1. Learning rates for the SVM classification

For the SVM classification with the hinge loss, we illustrate, as in [4,3], how noise conditions on the distribution ρ raise the variance–expectation exponent τ in (5.4) from 0 (for general distributions) to $\tau = \zeta/(\zeta + 1) > 0$.

Theorem 6. *Let* $\phi = \phi_h$ *and the multi-kernels* $\{K_\sigma : \sigma \in \Sigma\}$ *satisfy* (5.5). *Assume*

$$\inf_{\sigma\in\Sigma} \inf_{f\in\mathcal{H}_\sigma} \left\{\mathcal{E}^{\phi_h}(f) - \mathcal{E}^{\phi_h}(f_c) + \lambda\|f\|_\sigma^2\right\} \leqslant c_\beta \lambda^\beta \quad \forall \lambda > 0 \tag{5.10}$$

with $0 < \beta \leqslant 1$, $c_\beta > 0$, *and that* ρ *satisfies the noise condition* (1.13) *with* $\zeta \in [0, \infty]$ *and* $\Delta > 0$. *Choose* $\lambda = \lambda(m) = \left(\frac{1}{m}\right)^{\min\left\{\frac{2(\zeta+1)}{\beta(\zeta+2)+p(\zeta+1-\beta)}, \frac{2}{\beta+1}\right\}}$. *For any* $\varepsilon > 0$ *and* $0 < \delta < 1$, *there exists a constant* $C_\varepsilon > 0$ *independent of* m *such that with confidence* $1 - \delta$,

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leqslant C_\varepsilon\left(\frac{1}{m}\right)^{\theta}, \quad \theta = \min\left\{\frac{2\beta(\zeta+1)}{2\beta(\zeta+2)+p(\zeta+1-\beta)} - \varepsilon, \frac{2\beta}{1+\beta}\right\}.$$

Proof. Observe that ϕ_h satisfies the increment condition (5.3) with $q = 1$ and $c_q = 2$.

Because of the noise condition (1.13), we know from [36,43] that the condition (5.4) is valid with the exponent $\tau = \frac{\zeta}{\zeta+1}$ and the constant $c_\tau = 8\left(\frac{1}{2\Delta}\right)^{\zeta/(\zeta+1)}$. Then the conclusion follows from Theorem 5 and Proposition 2. □

Theorem 2 stated in the Introduction is a special case of Theorem 6 with multi-kernels having a uniform bound in C^s.

Proof of Theorem 2. By Example 3, (5.5) holds with $p = 2n/s$. Since ϕ_h is Lipschitz, Theorem 4 yields $\mathcal{E}^{\phi_h}(f) - \mathcal{E}^{\phi_h}(f_c) \leqslant \|f - f_c\|_{L^1_{\rho_X}}$. Hence (1.11) implies (5.10). Take $\zeta = 0$ since no assumption on the noise is made. We see Theorem 2 follows from Theorem 6. □

5.2. Learning rates with the least-square loss

Turn to the least-square loss $\phi_{\rm ls}(t) = (1-t)^2$ [37]. The high convexity of $\phi_{\rm ls}$ ensures a large variance–expectation exponent τ in (5.4). In fact, it was proved in [21] (see also [14]) that (5.4) holds true with $\tau = 1$ and $C_\tau = 1$. The increment condition (5.3) for $\phi_{\rm ls}$ is true with $q = 2$. Putting all these into Proposition 3 and Corollary 2, we obtain the following learning rate.

Theorem 7. *Consider* (1.6) *with* $\phi = \phi_{\rm ls}$ *and multi-kernels* $\{K_\sigma : \sigma \in \Sigma\}$ *satisfying* (5.5) *with some* $p \in (0, 2)$. *Assume that for some* $0 < \beta \leqslant 1$ *and* $c_\beta > 0$,

$$\inf_{\sigma\in\Sigma} \inf_{f\in\mathcal{H}_\sigma} \left\{ \|f - f_\rho\|^2_{L^2_{\rho_X}} + \lambda \|f\|^2_\sigma \right\} \leqslant c_\beta \lambda^\beta \quad \forall \lambda > 0. \tag{5.11}$$

Then by choosing $\lambda = \lambda(m) = \left(\frac{1}{m}\right)^{\min\left\{\frac{2}{2\beta+p}, 1\right\}}$, *for any* $\varepsilon > 0$ *and* $0 < \delta < 1$, *there exists a constant* C_ε *independent of* m *such that with confidence* $1 - \delta$,

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leqslant C_\varepsilon \left(\frac{1}{m}\right)^\theta \quad \text{with } \theta = \frac{1}{2} \min\left\{ \frac{2\beta}{2\beta + p} - \varepsilon, \beta \right\}. \tag{5.12}$$

If moreover, ρ *satisfies* (3.2), *then* θ *can be improved to* $\frac{1}{2-\alpha} \min\left\{ \frac{2\beta}{2\beta+p} - \varepsilon, \beta \right\}$. *In particular, when* $\inf_{x\in X} |f_\rho(x)| > 0$, (5.12) *holds with* $\theta = \min\left\{ \frac{2\beta}{2\beta+p} - \varepsilon, \beta \right\}$.

The above learning rate is better than those in the literature, e.g. [15,28,8,47]. When the kernels are C^∞ with (1.10) valid for any $s > 0$, we may take p in Theorem 7 to be arbitrarily small and the power θ in (5.12) becomes $\min\{\frac{1}{2} - \varepsilon, \beta/2\}$.

Example 4. Let $\phi(t) = (1-t)^2$, $\Sigma = [\sigma_1, \sigma_2]$ with $0 < \sigma_1 \leqslant \sigma_2 < \infty$ and K_σ be the Gaussian kernel $K_\sigma(x, y) = \exp\left\{-\frac{|x-y|^2}{2\sigma^2}\right\}$ on $X \subset \mathbb{R}^n$. Assume (5.11). Let $\varepsilon > 0$ and $\lambda = \lambda(m) = \left(\frac{1}{m}\right)^{\min\left\{\frac{1}{\beta}-\varepsilon, 1\right\}}$. Then with confidence $1 - \delta$, we have

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leqslant \widetilde{c} \left(\frac{1}{m}\right)^{\theta/2}, \quad \theta = \min\{1 - \varepsilon, \beta\}.$$

If ρ satisfies the noise condition (3.2) with $0 < \alpha \leqslant 1$, then $\theta/2$ can be improved to $\frac{1}{2-\alpha}\theta = \frac{1}{2-\alpha} \min\{1 - \varepsilon, \beta\}$. When $\inf_{x\in X} |f_\rho(x)| > 0$, we can replace $\theta/2$ by $\min\{1 - \varepsilon, \beta\}$.

5.3. Proof of the main result

To end this section, we prove our main result, Theorem 5. To this end, we shall use the following concentration inequality.

Proposition 6. *Let $\mathcal{F}$ be a set of measurable functions on Z, and $B, c > 0, \tau \in [0,1]$ be constants such that each function $f \in \mathcal{F}$ satisfies $\|f\|_\infty \leqslant B$ and $\mathbb{E}(f^2) \leqslant c(\mathbb{E}f)^\tau$. If for some $a > 0$ and $p \in (0,2)$,*

$$\sup_{m\in\mathbb{N}} \sup_{\mathbf{z}\in Z^m} \log \mathcal{N}_{2,\mathbf{z}}(\mathcal{F},\varepsilon) \leqslant a\varepsilon^{-p} \quad \forall \varepsilon > 0, \tag{5.13}$$

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

$$\mathbb{E}f - \frac{1}{m}\sum_{i=1}^m f(z_i) \leqslant \frac{1}{2}\eta^{1-\tau}(\mathbb{E}f)^\tau + c'_p\eta + 2\left(\frac{ct}{m}\right)^{1/(2-\tau)} + \frac{18Bt}{m} \quad \forall f \in \mathcal{F},$$

where

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

Other concentration inequalities [25] might be used for the error analysis of multi-kernel schemes.

To prove Proposition 6, we need to make some preparation as in [2].

Definition 7. A function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is sub-root if it is non-negative, non-decreasing, and if $\psi(r)/\sqrt{r}$ is non-increasing.

For a sub-root function ψ and any $D > 0$, the equation $\psi(r) = r/D$ has a unique positive solution.

The following proposition is given in [2], see also [4].

Proposition 7. *Let $\mathcal{F}$ be a class of measurable, square integrable functions such that $\mathbb{E}f - f \leqslant b$ for all $f \in \mathcal{F}$. Let ψ be a sub-root function, D be some positive constant and r^* be the unique solution to $\psi(r) = r/D$. Assume that*

$$\mathbb{E}\left[\max\left\{0, \sup_{\substack{f\in\mathcal{F}\\ \mathbb{E}f^2\leqslant r}} \mathbb{E}f - \frac{1}{m}\sum_{i=1}^m f(z_i)\right\}\right] \leqslant \psi(r) \quad \forall r \geqslant r^*.$$

Then for all $t > 0$, and all $K > D/7$, with probability at least $1 - e^{-t}$ there holds

$$\mathbb{E}f - \frac{1}{m}\sum_{i=1}^m f(z_i) \leqslant \frac{\mathbb{E}f^2}{K} + \frac{50K}{D^2}r^* + \frac{(K+9b)t}{m} \quad \forall f \in \mathcal{F}.$$

We need to find the sub-root function ψ in our setting. To this end, introduce the Rademacher variables $\varepsilon_i,\ i = 1,\dots,m$. Then

$$\mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}\\ \mathbb{E}f^2\leqslant r}} \left|\mathbb{E}f - \frac{1}{m}\sum_{i=1}^m f(z_i)\right|\right] \leqslant 2\mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}\\ \mathbb{E}f^2\leqslant r}} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i f(z_i)\right|\right]. \tag{5.14}$$

The right-hand side is called the local Rademacher process. It can be bounded by using empirical covering numbers and the entropy integral. See [40,26].

The following result is a scaled version of Proposition 5.4 in [36] where the case $B = 1$ is given.

Proposition 8. *Let $\mathcal{F}$ be a class of measurable functions from Z to $[-B, B]$. Assume* (5.13) *for some $p \in (0, 2)$ and $a > 0$. Then there exists a constant c_p depending only on p such that*

$$\mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}\\ \mathbb{E}f^2\leqslant r}}\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i f(z_i)\right|\right]\leqslant c_p\max\left\{r^{1/2-p/4}\left(\frac{a}{m}\right)^{1/2}, B^{\frac{2-p}{2+p}}\left(\frac{a}{m}\right)^{2/(2+p)}\right\}.$$

According to Proposition 8 and (5.14), in applying Proposition 7, one should take

$$\psi(r) = 2c_p\max\left\{r^{1/2-p/4}\left(\frac{a}{m}\right)^{1/2}, B^{\frac{2-p}{2+p}}\left(\frac{a}{m}\right)^{2/(2+p)}\right\}. \tag{5.15}$$

Then the solution r^* to the equation $\psi(r) = r/D$ satisfies

$$r^*\leqslant\max\left\{(2c_pD)^{\frac{4}{2+p}}, 2c_pDB^{\frac{2-p}{2+p}}\right\}\left(\frac{a}{m}\right)^{\frac{2}{2+p}}. \tag{5.16}$$

Proof of Proposition 6. Let ψ be defined by (5.15) and r^* be the solution to $\psi(r) = r/D$. Since $\|f\|_\infty\leqslant B$, we have $\mathbb{E}f - f\leqslant b := 2B$ for each $f \in \mathcal{F}$. Choose $K = D/5$. By Proposition 7 and the condition $\mathbb{E}f^2\leqslant c(\mathbb{E}f)^\tau$ we know that with probability at least $1 - e^{-t}$ there holds

$$\mathbb{E}f - \frac{1}{m}\sum_{i=1}^{m}f(z_i)\leqslant\frac{5c}{D}(\mathbb{E}f)^\tau + \frac{10}{D}r^* + \frac{(\frac{D}{5}+18B)t}{m}\quad\forall f\in\mathcal{F}. \tag{5.17}$$

Recall that r^* satisfies (5.16). Take $D = 10c\eta^{\tau-1}$ where η is given in our statement. Then $\frac{5c}{D} = \frac{1}{2}\eta^{1-\tau}$. The expression of η in connection with the bound (5.16) for r^* tells us that $\frac{10}{D}r^*\leqslant\widetilde{c}_p\eta$ where $\widetilde{c}_p$ is a constant depending only on p and c_p, hence only on p. Observe from the choice of D that

$$\frac{Dt}{5m} = \frac{2ct}{m\eta^{1-\tau}}\leqslant 2\max\left\{\eta, \left(\frac{ct}{m}\right)^{1/(2-\tau)}\right\},$$

according to whether $\eta\geqslant\left(\frac{ct}{m}\right)^{1/(2-\tau)}$. Take c'_p to be the constant $\widetilde{c}_p + 2$ depending only on p. Then the desired inequality holds for each $f \in \mathcal{F}$. This proves Proposition 6. □

We now turn to our key analysis and prove Theorem 5. Let us first explain our main ideas.

In the sample error term of (3.7), the quantity $\mathcal{E}^\phi_{\mathbf{z}}(f_\lambda) - \mathcal{E}^\phi(f_\lambda)$ is easy to handle. It can be estimated by the one-side Bernstein inequality for the single random variable $\phi(yf_\lambda(x))$ on Z. This will be done in the first step of the proof with a mild technical modification: consider the random variable $\xi = \phi(yf_\lambda(x)) - \phi(y, f^\phi_\rho(x))$ instead of $\phi(yf_\lambda(x))$.

The quantity $\mathcal{E}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi_{\mathbf{z}}(\pi(f_{\mathbf{z}}))$ is more difficult and we need Proposition 6 to estimate. Here the function set will be $\mathcal{F} = \left\{\phi\left(y\pi(f)(x)\right) - \phi\left(yf_\rho^\phi(x)\right) : f \in B_R\right\}$ with such a radius R that B_R contains $f_{\mathbf{z}}$, i.e., R is a bound of $\|f_{\mathbf{z}}\|_\Sigma$. On the other hand, smaller radius R yields better estimates. Hence good bounds for $\|f_{\mathbf{z}}\|_\Sigma$ play an important role for the sample error estimates.

A rough bound for $\|f_{\mathbf{z}}\|_\Sigma$ immediately follows from the definition of $f_{\mathbf{z}}$. By choosing $f = 0$, we find $\lambda\|f_{\mathbf{z}}\|_\Sigma^2 \leqslant \mathcal{E}^\phi_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\|f_{\mathbf{z}}\|_\Sigma^2 \leqslant \mathcal{E}^\phi_{\mathbf{z}}(0) + \lambda\cdot 0 = \phi(0)$. This proves

Lemma 1. *For every $\lambda > 0$, there holds $\|f_{\mathbf{z}}\|_\Sigma \leqslant \sqrt{\phi(0)/\lambda}$.*

We may use the bound $\sqrt{\phi(0)/\lambda}$ as R in $\mathcal{F}$ and apply Proposition 6 to get some rough estimates for $\mathcal{E}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi_{\mathbf{z}}(\pi(f_{\mathbf{z}}))$. However, the empirical error $\mathcal{E}^\phi_{\mathbf{z}}(f)$ is a good approximation of the generalization error $\mathcal{E}^\phi(f)$. Hence the penalty value $\|f_{\mathbf{z}}\|_\Sigma$ is expected to be close to $\|f_\lambda\|_\Sigma$ which is bounded by $\sqrt{\mathcal{D}(\lambda)/\lambda}$:

$$\lambda\|f_\lambda\|_\Sigma^2 \leqslant \mathcal{E}^\phi(f_\lambda) - \mathcal{E}^\phi(f_\rho^\phi) + \lambda\|f_\lambda\|_\Sigma^2 = \mathcal{D}(\lambda). \tag{5.18}$$

This expectation will be realized by an *iteration technique* used in [36] and [43]. By this technique, we shall show under some assumptions that with high confidence $\|f_{\mathbf{z}}\|_\Sigma$ has a bound arbitrarily close to $\sqrt{\mathcal{D}(\lambda)/\lambda}$ (in the order of λ).

We are in a position to estimate the sample error and prove Theorem 5.

Proof of Theorem 5. Write the sample error as

$$\begin{aligned}\mathcal{S}_{\mathbf{z},\lambda} = &\left\{\left(\mathcal{E}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi(f_\rho^\phi)\right) - \left(\mathcal{E}^\phi_{\mathbf{z}}(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi_{\mathbf{z}}(f_\rho^\phi)\right)\right\}\\ &+ \left\{\left(\mathcal{E}^\phi_{\mathbf{z}}(f_\lambda) - \mathcal{E}^\phi_{\mathbf{z}}(f_\rho^\phi)\right) - \left(\mathcal{E}^\phi(f_\lambda) - \mathcal{E}^\phi(f_\rho^\phi)\right)\right\} := S_1 + S_2.\end{aligned}$$

We divide our estimation into three steps. Take $t \geqslant 1$ which will be determined later. Denote $B = \max\{\phi(-1), \phi(1)\}$.

Step 1: Estimate S_2. Consider the random variable $\xi = \phi(yf_\lambda(x)) - \phi(yf_\rho^\phi(x))$ on Z. Denote

$$\xi = \xi_1 + \xi_2 = \left\{\phi(yf_\lambda(x)) - \phi(y\pi(f_\lambda)(x))\right\} + \left\{\phi(y\pi(f_\lambda)(x)) - \phi(yf_\rho^\phi(x))\right\}.$$

First we bound ξ_1. By (1.2), (5.1) and (5.18), we have $\|f_\lambda\|_\infty \leqslant \kappa\|f_\lambda\|_\Sigma \leqslant \kappa\sqrt{\mathcal{D}(\lambda)/\lambda}$. We may assume the last quantity to be greater than one since otherwise $\xi_1 \equiv 0$. Then the increment condition on ϕ tells us $0 \leqslant \xi_1 \leqslant B_\lambda := c_q\kappa^q\left(\mathcal{D}(\lambda)/\lambda\right)^{q/2}$. Hence $|\xi_1 - \mathbb{E}(\xi_1)| \leqslant B_\lambda$. Applying the one-side Bernstein inequality to ξ_1, we know that for any $\varepsilon > 0$,

$$\text{Prob}\left\{\frac{1}{m}\sum_{i=1}^m \xi_1(z_i) - \mathbb{E}\xi_1 > \varepsilon\right\} \leqslant \exp\left\{-\frac{m\varepsilon^2}{2\left(\sigma^2(\xi_1) + \frac{1}{3}B_\lambda\varepsilon\right)}\right\}.$$

Solving the quadratic equation

$$\frac{m\varepsilon^2}{2\left(\sigma^2(\xi_1) + \frac{1}{3}B_\lambda\varepsilon\right)} = t$$

for ε, we see that there exists a subset U_1 of Z^m with measure at least $1-e^{-t}$ such that for every $\mathbf{z}\in U_1$,

$$\frac{1}{m}\sum_{i=1}^{m}\xi_1(z_i)-\mathbb{E}\xi_1\leqslant\frac{\frac{1}{3}B_\lambda t+\sqrt{\left(\frac{1}{3}B_\lambda t\right)^2+2m\sigma^2(\xi_1)t}}{m}\leqslant\frac{2B_\lambda t}{3m}+\sqrt{\frac{2t}{m}\sigma^2(\xi_1)}.$$

But the fact $0\leqslant\xi_1\leqslant B_\lambda$ implies $\sigma^2(\xi_1)\leqslant B_\lambda\mathbb{E}(\xi_1)$. Therefore, we have

$$\frac{1}{m}\sum_{i=1}^{m}\xi_1(z_i)-\mathbb{E}\xi_1\leqslant\frac{7B_\lambda t}{6m}+\mathbb{E}\xi_1\quad\forall\mathbf{z}\in U_1.$$

Next we consider ξ_2. Since both $y\pi(f_\lambda)(x)$ and $yf_\rho^\phi(x)$ are on $[-1,1]$, ξ_2 is a random variable satisfying $|\xi_2|\leqslant B$. Applying the one-side Bernstein inequality as above, we know that there exists another subset U_2 of Z^m with measure at least $1-e^{-t}$ such that for every $\mathbf{z}\in U_2$,

$$\frac{1}{m}\sum_{i=1}^{m}\xi_2(z_i)-\mathbb{E}\xi_2\leqslant\frac{2Bt}{3m}+\sqrt{\frac{2t\sigma^2(\xi_2)}{m}}.$$

By (5.4), we have $\sigma^2(\xi_2)\leqslant C_\tau(\mathbb{E}\xi_2)^\tau$. Applying the elementary inequality

$$\frac{1}{q}+\frac{1}{q^*}=1\text{ with }q,q^*>1\Longrightarrow a\cdot b\leqslant\frac{1}{q}a^q+\frac{1}{q^*}b^{q^*}\quad\forall a,b\geqslant 0$$

with $q=\frac{2}{2-\tau}$, $q^*=\frac{2}{\tau}$ and $a=\sqrt{\frac{2tC_M}{m}}$, $b=\sqrt{(\mathbb{E}\xi_2)^\tau}$, we see that

$$\sqrt{\frac{2t\sigma^2(\xi_2)}{m}}\leqslant\sqrt{\frac{2tC_\tau}{m}}\cdot\sqrt{(\mathbb{E}\xi_2)^\tau}\leqslant\left(1-\frac{\tau}{2}\right)\left(\frac{2tC_\tau}{m}\right)^{\frac{1}{2-\tau}}+\frac{\tau}{2}\mathbb{E}\xi_2.$$

Hence

$$\frac{1}{m}\sum_{i=1}^{m}\xi_2(z_i)-\mathbb{E}\xi_2\leqslant\frac{2Bt}{3m}+\left(\frac{2tC_\tau}{m}\right)^{\frac{1}{2-\tau}}+\mathbb{E}\xi_2\quad\forall\mathbf{z}\in U_2.$$

Combine the above estimates for ξ_1 and ξ_2 with the fact $\mathbb{E}\xi_1+\mathbb{E}\xi_2=\mathbb{E}\xi\leqslant\mathcal{D}(\lambda)\leqslant c_\beta\lambda^\beta$. We conclude that

$$S_2\leqslant\frac{7B_\lambda t+4Bt}{6m}+\left(\frac{2tC_\tau}{m}\right)^{\frac{1}{2-\tau}}+\mathcal{D}(\lambda)\quad\forall\mathbf{z}\in U_1\cap U_2.\tag{5.19}$$

Step 2: Estimate S_1. By Proposition 5, one has

$$\Delta_{\mathbf{z}}:=\mathcal{E}^\phi(\pi(f_{\mathbf{z}}))-\mathcal{E}^\phi(f_\rho^\phi)+\lambda\|f_{\mathbf{z}}\|_\Sigma^2\leqslant S_1+S_2+\mathcal{D}(\lambda).\tag{5.20}$$

Let $R>0$. Apply Proposition 6 to the function set

$$\mathcal{F}=\left\{\phi\left(y\pi(f)(x)\right)-\phi\left(yf_\rho^\phi(x)\right):f\in B_R\right\}.$$

Since $|\phi(y\pi(f)(x)) - \phi(y\pi(g)(x))| \leqslant |\phi'_-(-1)||\pi(f)(x) - \pi(g)(x)| \leqslant |\phi'_-(-1)||f(x) - g(x)|$, there holds

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{F}, \varepsilon) \leqslant \mathcal{N}_{2,\mathbf{z}}\left(B_R, \frac{\varepsilon}{|\phi'_-(-1)|}\right).$$

Hence (5.5) yields (5.13) with $a = c_p|\phi'_-(-1)|^p R^p$.

Since $\phi(y\pi(f)(x)) \leqslant B$ and $\phi\left(yf_\rho^\phi(x)\right) \leqslant B$, we know that $\|f\|_\infty \leqslant B$ for every $f \in \mathcal{F}$. The assumption (5.4) tells us that $\mathbb{E}f^2 \leqslant c(\mathbb{E}f)^\tau$ with $c = C_\tau$.

Thus all the conditions in Proposition 6 hold, and we know that there is a subset $\mathcal{V}(R)$ of Z^m with measure at least $1 - e^{-t}$ such that for every $\mathbf{z} \in \mathcal{V}(R)$ and every $f \in B_R$,

$$\begin{aligned}&\left(\mathcal{E}^\phi(\pi(f)) - \mathcal{E}^\phi(f_\rho^\phi)\right) - \left(\mathcal{E}_\mathbf{z}^\phi(\pi(f)) - \mathcal{E}_\mathbf{z}^\phi(f_\rho^\phi)\right)\\ &\quad\leqslant \frac{1}{2}\eta_R^{1-\tau}\left(\mathcal{E}^\phi(\pi(f)) - \mathcal{E}^\phi(f_\rho^\phi)\right)^\tau + c'_p\eta_R + 2\left(\frac{C_\tau t}{m}\right)^{\frac{1}{2-\tau}} + \frac{18Bt}{m},\end{aligned} \tag{5.21}$$

where $\eta_R = \eta$ is given in Proposition 6 with $c = C_\tau$ and $a = c_p|\phi'_-(-1)|^p R^p$, i.e.,

$$\eta_R = \max\left\{C_\tau^{\frac{2-p}{4-2\tau+p\tau}}\left(\frac{c_p|\phi'_-(-1)|^p R^p}{m}\right)^{\frac{2}{4-2\tau+p\tau}}, B^{\frac{2-p}{2+p}}\left(\frac{c_p|\phi'_-(-1)|^p R^p}{m}\right)^{\frac{2}{2+p}}\right\}.$$

Let $\mathcal{W}(R)$ be the subset of Z^m defined by

$$\mathcal{W}(R) = \{\mathbf{z} \in U_1 \cap U_2 : f_\mathbf{z} \in B_R\}.$$

Let $\mathbf{z} \in \mathcal{W}(R) \cap \mathcal{V}(R)$. Then (5.21) holds for $f_\mathbf{z}$. Together with the estimate (5.19) for S_2 and (5.20), we know that

$$\begin{aligned}\Delta_\mathbf{z} \leqslant\ & \frac{1}{2}\eta_R^{1-\tau}\left(\mathcal{E}^\phi(\pi(f_\mathbf{z})) - \mathcal{E}^\phi(f_\rho^\phi)\right)^\tau + c'_p\eta_R + 4\left(\frac{C_\tau t}{m}\right)^{1/(2-\tau)}\\ &+ \frac{19Bt + 3B_\lambda t/2}{m} + 2\mathcal{D}(\lambda).\end{aligned}$$

When $\tau = 1$ this yields

$$\Delta_\mathbf{z} \leqslant c''_p\eta_R + 8\left(\frac{C_\tau t}{m}\right)^{1/(2-\tau)} + \frac{38Bt + 3B_\lambda t}{m} + 4\mathcal{D}(\lambda), \tag{5.22}$$

where $c''_p = \max\{2c'_p, 1\}$. Here we have bounded $2c'_p$ by c''_p. When $0 < \tau < 1$, we use the elementary inequality: if $a, b > 0$ and $0 < \tau < 1$, then

$$x \leqslant ax^\tau + b, \quad x > 0 \implies x \leqslant \max\{(2a)^{1/(1-\tau)}, 2b\}.$$

We find that (5.22) still holds.

By the choice of $\lambda = \lambda(m) = (\frac{1}{m})^\gamma$, one easily checks that

$$\eta_R \leqslant c_{p,\tau}\lambda^\beta \max\left\{(R^2\lambda^{1-\beta})^{\frac{p}{4-2\tau+p\tau}}, (R^2\lambda^{1-\beta})^{\frac{p}{2+p}}\right\}$$

for some $c_{p,\tau} > 0$. But $4 - 2\tau + p\tau \geqslant 2 + p$, hence if $R > \lambda^{-(1-\beta)/2}$, then

$$\eta_R \leqslant c_{p,\tau}\lambda^{\beta}(R^2\lambda^{1-\beta})^{\frac{p}{2+p}} = c_{p,\tau}\lambda^{\frac{p+2\beta}{2+p}} R^{\frac{2p}{2+p}}. \tag{5.23}$$

The choice of λ together with the assumption $\mathcal{D}(\lambda) \leqslant c_\beta \lambda^\beta$ and $t > 1$ on the regularization error also implies

$$8\left(\frac{C_\tau t}{m}\right)^{1/(2-\tau)} + \frac{38Bt + 3B_\lambda t}{m} + 4\mathcal{D}(\lambda) \leqslant c_{q,\tau,\beta} t \lambda^\beta \tag{5.24}$$

for some $c_{q,\tau,\beta} > 0$.

Putting the estimates (5.24) and (5.23) into (5.22) we obtain

$$\Delta_{\mathbf{z}} \leqslant c_p'' c_{p.\tau} \lambda^{\frac{p+2\beta}{2+p}} R^{\frac{2p}{2+p}} + c_{q,\tau,\beta} t \lambda^\beta \quad \forall \mathbf{z} \in \mathcal{W}(R) \cap \mathcal{V}(R) \tag{5.25}$$

whenever $R > \lambda^{-(1-\beta)/2}$. This implies that $\|f_{\mathbf{z}}\|_\Sigma \leqslant \sqrt{\Delta_{\mathbf{z}}/\lambda} \leqslant g(R)$, where $g : \mathbb{R}_+ \to \mathbb{R}_+$ is a univariate function defined as

$$g(R) = \sqrt{c_p'' c_{p,\tau}} \lambda^{\frac{\beta-1}{2+p}} R^{\frac{p}{2+p}} + \sqrt{c_{q,\tau,\beta} t}\, \lambda^{(\beta-1)/2}. \tag{5.26}$$

It follows that

$$\mathcal{W}(R) \cap \mathcal{V}(R) \subseteq \mathcal{W}(g(R)) \quad \forall R > \lambda^{-(1-\beta)/2}. \tag{5.27}$$

Step 3: By iteration, find a small ball B_R that, with high confidence, contains $f_{\mathbf{z}}$.

Lemma 1 means that $\mathcal{W}(R_0) = U_1 \cap U_2$ for $R_0 = \sqrt{\phi(0)/\lambda}$.

When $R_0 > \lambda^{-(1-\beta)/2}$, we use our conclusion (5.27) iteratively.

Denote $g^{[0]}(R) = R$, $g^{[1]}(R) = g(R)$ and $g^{[\ell]}(R) = g\left(g^{[\ell-1]}(R)\right)$ for $\ell \geqslant 2$. According to (5.27), if

$$g^{[j]}(R) > \lambda^{-(1-\beta)/2}, \quad j = 0, 1, \ldots, \ell - 1, \tag{5.28}$$

then

$$\mathcal{W}(R) \cap \mathcal{V}(R) \cap \mathcal{V}(g^{[1]}(R)) \cap \cdots \cap \mathcal{V}\left(g^{[\ell-1]}(R)\right) \subseteq \mathcal{W}(g^{[\ell]}(R)). \tag{5.29}$$

Observe that $g(R) = d_0 R^{\frac{p}{2+p}} + d_1$ with $d_0, d_1 > 0$ given in (5.26). Then

$$g^{[2]}(R) = d_0\left(d_0 R^{\frac{p}{2+p}} + d_1\right)^{\frac{p}{2+p}} + d_1 \leqslant d_0^{1+\frac{p}{2+p}} R^{\left(\frac{p}{2+p}\right)^2} + d_1 + d_0 d_1^{\frac{p}{2+p}},$$

and in general, for $\ell \in \mathbb{N}$,

$$\begin{aligned} g^{[\ell]}(R) \leqslant\ & d_0^{1+\frac{p}{2+p}+\cdots+\left(\frac{p}{2+p}\right)^{\ell-1}} R^{\left(\frac{p}{2+p}\right)^\ell} + d_1 + d_0 d_1^{\frac{p}{2+p}} + d_0^{1+\frac{p}{2+p}} d_1^{\left(\frac{p}{2+p}\right)^2} \\ & + \cdots + d_0^{1+\frac{p}{2+p}+\cdots+\left(\frac{p}{2+p}\right)^{\ell-2}} d_1^{\left(\frac{p}{2+p}\right)^{\ell-1}}. \end{aligned}$$

This in connection with the expressions for d_0 and d_1 gives

$$g^{[\ell]}(R) \leqslant d_0^{\frac{2+p}{2}\left\{1-\left(\frac{p}{2+p}\right)^{\ell}\right\}} R^{\left(\frac{p}{2+p}\right)^{\ell}} + \sum_{i=0}^{\ell-1} d_0^{\sum_{j=0}^{i-1}\left(\frac{p}{2+p}\right)^{j}} d_1^{\left(\frac{p}{2+p}\right)^{i}}$$
$$\leqslant c_0^{\frac{2+p}{4}} \lambda^{\frac{(\beta-1)}{2}\left\{1-\left(\frac{p}{2+p}\right)^{\ell}\right\}} R^{\left(\frac{p}{2+p}\right)^{\ell}} + \sum_{i=0}^{\ell-1} c_0^{\frac{2+p}{4}} (c_1 t)^{\left(\frac{p}{2+p}\right)^{i}} \lambda^{\frac{(\beta-1)}{2}},$$

where $c_0 = \max\{1, c_p'' c_{p,\tau}\}$ and $c_1 = \max\{1, \sqrt{c_{q,\tau,\beta}}\}$. In particular, for $R = R_0$, there holds

$$g^{[\ell]}(R_0) \leqslant c_0^{\frac{2+p}{4}} \lambda^{(\beta-1)/2} \left\{ (\phi(0))^{\frac{1}{2}\left(\frac{p}{2+p}\right)^{\ell}} \lambda^{-\frac{\beta}{2}\left(\frac{p}{2+p}\right)^{\ell}} + c_1 t \ell \right\}.$$

For $\varepsilon > 0$, choose $\ell_0 \in \mathbb{N}$ such that $\ell_0 \geqslant \log \frac{1}{2\varepsilon} / \log \frac{2+p}{p}$. Then $\frac{1}{2}\left(\frac{p}{2+p}\right)^{\ell_0} \leqslant \varepsilon$. It follows that

$$g^{[\ell_0]}(R_0) \leqslant c_0^{\frac{2+p}{4}} \lambda^{(\beta-1)/2} \left\{ (\phi(0))^{\frac{1}{2}\left(\frac{p}{2+p}\right)^{\ell_0}} \lambda^{-\beta\varepsilon} + c_1 t \ell_0 \right\}$$

when (5.28) with $\ell = \ell_0$ and $R = R_0$ holds.

When (5.28) with $\ell = \ell_0$ and $R = R_0$ is not valid, we have $g^{[j_0]}(R_0) \leqslant \lambda^{(\beta-1)/2}$ for some $j_0 \in \{0, 1, \ldots, \ell_0 - 1\}$.

Take $\ell_\varepsilon = \ell_0$ when (5.28) with $\ell = \ell_0$ and $R = R_0$ holds and $\ell_\varepsilon = j_0$ otherwise. In both cases, we have

$$g^{[\ell_\varepsilon]}(R_0) \leqslant c_\varepsilon \lambda^{(\beta-1)/2-\beta\varepsilon} =: R_\varepsilon, \tag{5.30}$$

where $c_\varepsilon := c_0^{\frac{2+p}{4}} \left((\phi(0))^{\frac{1}{2}\left(\frac{p}{2+p}\right)^{\ell_0}} + c_1 t \ell_0 \right)$.

Take $\ell = \ell_\varepsilon \leqslant \ell_0$ and $R = R_0$ in (5.29). Since $\mathcal{W}(R_0) = U_1 \cap U_2$, we know that there is a subset $\mathcal{V}_\varepsilon$ of Z^m with measure at most $\ell_0 e^{-t}$ such that

$$U_1 \cap U_2 \subseteq \mathcal{W}(R_\varepsilon) \cup \mathcal{V}_\varepsilon.$$

Then the measure of the set $\mathcal{W}(R_\varepsilon)$ is at least $1 - (\ell_0 + 2)e^{-t}$.

Apply (5.22) with $R = R_\varepsilon$ and notice (5.24). Let $\mathbf{z} \in \mathcal{W}(R_\varepsilon) \cap \mathcal{V}(R_\varepsilon)$. We know that

$$\Delta_{\mathbf{z}} \leqslant c_p'' \eta_{R_\varepsilon} + c_{q,\tau,\beta} t \lambda^{(\beta-1)/2}.$$

It is easy to check that $\eta_{R_\varepsilon} \leqslant c_{p,\tau} c_\varepsilon \left(\frac{1}{m}\right)^{\theta}$. Therefore, with the constant $\widetilde{c} = c_p'' c_{p,\tau} c_\varepsilon + c_{q,\tau,\beta} t$, there holds

$$\mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}^{\phi}(f_\rho^{\phi}) \leqslant \Delta_{\mathbf{z}} \leqslant \widetilde{c} \left(\frac{1}{m}\right)^{\theta}.$$

Taking $t = \log \frac{\ell_0+3}{\delta}$, the measure of the set $\mathcal{W}(R_\varepsilon) \cap \mathcal{V}(R_\varepsilon)$ is at least $1 - \delta$. Then Theorem 5 is proved. □

6. Extensions

A key point of our analysis is to find essential bounds for penalty functional values of regularization schemes. This approach can be extended to regularization schemes with more general loss functions and general penalty functionals.

Let the hypothesis space $\mathcal{H}$ be a function set containing 0. It is assigned a functional $\Omega : \mathcal{H} \to \mathbb{R}_+$ satisfying $\Omega(0) = 0$. Beyond the multi-kernel space $\mathcal{H}_\Sigma$, such a hypothesis space is the linear programming SVM classifier [43] in a one-kernel setting with the penalty functional $\Omega(f)$ defined for $f \in \mathcal{H} = \mathcal{H}_{K,\mathbf{z}} = \left\{\sum_{i=1}^m \alpha_i y_i K_{x_i} : \alpha_i \geqslant 0\right\}$ as $\Omega(f) = \sum_{i=1}^m \alpha_i$.

Let Y be a subset of $\mathbb{R}$, and $V : \mathbb{R}^2 \to \mathbb{R}_+$ be a general loss function.

The general regularization scheme in $\mathcal{H}$ associated with V and the penalty functional Ω is defined for the sample $\mathbf{z}$ as

$$f_{\mathbf{z}}^V = \arg\min_{f\in\mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^m V(y_i, f(x_i)) + \lambda\Omega(f) \right\}. \tag{6.1}$$

All the results we obtained for the multi-kernel regularized classifiers (1.6) can be established for the more general scheme (6.1) under the assumption that the pair (V, ρ) is *M-admissible*: there is a constant $M > 0$ such that $|y| \leqslant M$ almost surely with respect to ρ, and for each $y \in [-M, M]$, $V(y, t)$ is a convex function of the variable $t \in \mathbb{R}$ satisfying

$$\begin{cases} V(y,t) \geqslant V(y,M) & \forall t > M, \\ V(y,t) \geqslant V(y,-M) & \forall t < -M. \end{cases} \tag{6.2}$$

An important family of regularization schemes (6.1) are those for *regression with a general loss function*: take $Y = \mathbb{R}$ and $V(y, f(x)) = \psi(y - f(x))$ where $\psi : \mathbb{R} \to \mathbb{R}_+$ is even, convex and increasing on $[0, +\infty)$ with $\psi(0) = 0$. If $|y| \leqslant M$ almost surely with respect to ρ, then (V, ρ) is M-admissible. Our approach can be used to analyze the convergence of $\int_Z V(y, f_{\mathbf{z}}^V(x))\, d\rho$ to $\inf_{f\in\mathcal{H}} \int_Z V(y, f(x))\, d\rho$, which will be discussed in the future.

Example 5. Let $\varepsilon > 0$. The ε-insensitive norm is the univariate loss function ψ used for regression defined [41] as $\psi(t) = \max\{|t| - \varepsilon, 0\}$. It would be interesting to analyze the convergence of the scheme (6.1) as ε tends to zero.

For the classification algorithm (1.6), some of our error bounds can be extended to non-classifying loss functions (such as the exponential loss), i.e., those activating loss functions whose infimum cannot be achieved. For this purpose, we need a more general projection operator.

Definition 8. For $M > 0$, the projection operator at level M is defined on the space of measurable functions $f : X \to \mathbb{R}$ as

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

Using this projection operator, we can have similar error decompositions by revising the regularization error and introducing level M adapting to the behavior of the loss function (the convergence rate of $\phi(t)$ as $t \to \infty$). Then some learning rates can be obtained, following our approach. Detailed analysis will be done in our future investigation.

Acknowledgment

We thank the referees for their careful reading and constructive suggestions which help us improve the paper.

Appendix A.

Proof of Proposition 1. We first claim that there exists a constant $C(\phi, m)$ depending on ϕ and the sample size m such that

$$\|\hat{\alpha}_\sigma\|_{\ell^\infty(\mathbb{R}^m)} \leqslant C(\phi, m) \quad \forall \sigma \in \Sigma. \tag{A.1}$$

To verify our claim, recall that $\hat{\alpha}_\sigma$ is a maximizer of $\hat{R}(\alpha, \sigma)$. This yields

$$\hat{R}(\hat{\alpha}_\sigma, \sigma) \geqslant \hat{R}(0, \sigma) = -\psi(0) = -\sup_{u\in\mathbb{R}} \{0 - \phi(u)\} = \inf_{u\in\mathbb{R}} \phi(u) = 0.$$

Since K_σ is positive semidefinite, it follows that

$$\sum_{i=1}^m \psi\left(-(\hat{\alpha}_\sigma)_i y_i\right) = -m\hat{R}(\hat{\alpha}_\sigma, \sigma) - \frac{1}{4m\lambda} \sum_{i,j=1}^m (\hat{\alpha}_\sigma)_i K_\sigma(x_i, x_j)(\hat{\alpha}_\sigma)_j \leqslant -m\hat{R}(\hat{\alpha}_\sigma, \sigma) = 0.$$

However, for each $v \in \mathbb{R}$,

$$\psi(-v) = \sup_{u\in\mathbb{R}} \{-uv - \phi(u)\} \geqslant -\phi(0).$$

Therefore, for each $i \in \{1, \ldots, m\}$, we have

$$\psi\left(-(\hat{\alpha}_\sigma)_i y_i\right) \leqslant 0 - \sum_{j\neq i} \psi(-(\hat{\alpha}_\sigma)_j y_j) \leqslant -\sum_{j\neq i} \{-\phi(0)\} = (m-1)\phi(0). \tag{A.2}$$

Now we prove our claim in two cases.

Case 1: $\phi'_+(t) \leqslant 0$ for each $t \in \mathbb{R}$. In this case, ϕ is non-increasing and $\lim_{u\to+\infty} \phi(u) = \inf_{u\in\mathbb{R}} \phi(u) = 0$. This in connection with the definition of the dual function implies

$$\psi(-v) = \sup_{u\in\mathbb{R}} \{-uv - \phi(u)\} \geqslant \lim_{u\to+\infty} \{-uv\} = +\infty \quad \forall v < 0. \tag{A.3}$$

It follows from (A.2) that $(\hat{\alpha}_\sigma)_i y_i \geqslant 0$ for each i.

Definition 1 also tells us that ϕ is strictly decreasing on $(-\infty, 0]$ and $\lim_{t\to-\infty} \phi(t) = +\infty$. Then the inverse function ϕ^{-1} is well defined on $[\phi(0), +\infty)$. Choose $u = \phi^{-1}(\sqrt{v})$ for $v \geqslant (\phi(0))^2$ in the definition of ψ, we see that $\psi(-v) \geqslant -v\phi^{-1}(\sqrt{v}) - \phi\left(\phi^{-1}(\sqrt{v})\right)$. It follows that for any $v \geqslant \max\left\{1, (\phi(-2))^2\right\}$ there holds

$$\psi(-v) \geqslant \sqrt{v}\left\{-\sqrt{v}\phi^{-1}(\sqrt{v}) - 1\right\} \geqslant \sqrt{v}.$$

Hence

$$v \leqslant \max\{1, (\phi(-2))^2, (\psi(-v))^2\} \quad \forall v \in \mathbb{R}.$$

Combining with (A.2), this implies that

$$(\hat{\alpha}_\sigma)_i y_i \leqslant \max\left\{1, (\phi(-2))^2, (m-1)^2 (\phi(0))^2\right\} =: C_1(\phi, m).$$

As $y_i = \pm 1$ and $\operatorname{sgn}((\hat{\alpha}_\sigma)_i) = y_i$, we know that $\left|(\hat{\alpha}_\sigma)_i\right| = \left|(\hat{\alpha}_\sigma)_i y_i\right| = (\hat{\alpha}_\sigma)_i y_i \leqslant C_1(\phi, m)$ for each i. This proves our claim in Case 1: $\|\hat{\alpha}_\sigma\|_{\ell^\infty(\mathbb{R}^m)} \leqslant C_1(\phi, m)$.

Case 2: $\phi'_+(t_0) > 0$ for some $t_0 \in \mathbb{R}$. In this case, $t_0 > 0$ and ϕ is strictly increasing on $[t_0, +\infty)$. Then for $v \leqslant \min\left\{-1, -(\phi(t_0+2))^2\right\}$, there exists some $u_v \geqslant t_0 + 2$ such that $\phi(u_v) = \sqrt{-v}$. Choosing $u = u_v$ in the definition of ψ, we see that $\psi(-v) \geqslant -u_v v - \phi(u_v)$ can be bounded from below as

$$\psi(-v) \geqslant \sqrt{-v}\left\{\sqrt{-v}(t_0+2) - 1\right\} \geqslant \sqrt{-v} \quad \forall v \leqslant \min\left\{-1, -(\phi(t_0+2))^2\right\}. \tag{A.4}$$

On the other hand, since ϕ is strictly decreasing on $(-\infty, 0]$, for $v \geqslant \max\left\{1, (\phi(-2))^2\right\}$ there exists some $u_v \leqslant -2$ such that $\phi(u_v) = \sqrt{v}$. It follows that

$$\psi(-v) \geqslant -u_v v - \phi(u_v) \geqslant \sqrt{v}\left\{-u_v\sqrt{v} - 1\right\} = \sqrt{v} \quad \forall v \geqslant \max\left\{1, (\phi(-2))^2\right\}.$$

This in connection with (A.4) implies that $\psi(-v) > (m-1)\phi(0)$ whenever

$$|v| > \max\left\{(m-1)^2 (\phi(0))^2, (\phi(t_0+2))^2, 1, (\phi(-2))^2\right\} =: C_2(\phi, m).$$

Combining with (A.2), we see again that $\left|(\hat{\alpha}_\sigma)_i\right| = \left|\hat{\alpha}_{\sigma,i} y_i\right| \leqslant C_2(\phi, m)$ for each $i \in \{1, \ldots, m\}$. This proves our claim in Case 2: $\|\hat{\alpha}_\sigma\|_{\ell^\infty(\mathbb{R}^m)} \leqslant C_2(\phi, m)$. Therefore, (A.1) holds with $C(\phi, m) = \max\{C_1(\phi, m), C_2(\phi, m)\}$.

Next, we apply our claim (A.1) to prove the proposition. Denote

$$\hat{G}(\sigma) = \max_{\alpha \in \mathbb{R}^m} \hat{R}(\alpha, \sigma) = \hat{R}(\hat{\alpha}_\sigma, \sigma).$$

To prove the existence of a solution $(\hat{\alpha}, \hat{\sigma}) = (\hat{\alpha}_{\hat{\sigma}}, \hat{\sigma})$ to the problem (2.2), it is sufficient to prove that the function $\hat{G}(\sigma)$ is continuous on the compact metric space (Σ, d_Σ).

Let $\sigma_1, \sigma_0 \in \Sigma$. By the definition of $\hat{G}(\sigma)$ and $\hat{R}(\alpha, \sigma)$, we have

$$\begin{aligned}\hat{G}(\sigma_1) - \hat{G}(\sigma_0) &= \hat{R}(\hat{\alpha}_{\sigma_1}, \sigma_1) - \hat{R}(\hat{\alpha}_{\sigma_0}, \sigma_0) \leqslant \hat{R}(\hat{\alpha}_{\sigma_1}, \sigma_1) - \hat{R}(\hat{\alpha}_{\sigma_1}, \sigma_0) \\ &= \frac{1}{4m^2\lambda} \sum_{i,j=1}^{m} (\hat{\alpha}_{\sigma_1})_i \left(K_{\sigma_0}(x_i, x_j) - K_{\sigma_1}(x_i, x_j)\right) (\hat{\alpha}_{\sigma_1})_j.\end{aligned}$$

By symmetry, there holds

$$\hat{G}(\sigma_0) - \hat{G}(\sigma_1) \leqslant \frac{1}{4m^2\lambda} \sum_{i,j=1}^{m} (\hat{\alpha}_{\sigma_0})_i \left(K_{\sigma_1}(x_i, x_j) - K_{\sigma_0}(x_i, x_j)\right) (\hat{\alpha}_{\sigma_0})_j.$$

By the continuity of $K_\sigma(x_i, x_j)$ at σ_0 for each pair (i, j), we know that for any $\varepsilon > 0$, there exists some $\delta > 0$ such that $\left|K_{\sigma_1}(x_i, x_j) - K_{\sigma_0}(x_i, x_j)\right| \leqslant 4\lambda\varepsilon / (C(\phi, m))^2$ whenever $d_\Sigma(\sigma_1, \sigma_0) <$

δ. It follows from (A.1) and the above two bounds that $\left|\hat{G}(\sigma_1) - \hat{G}(\sigma_0)\right| \leqslant \varepsilon$. This shows the continuity of $\hat{G}$ at σ_0. Since σ_0 is an arbitrary point in Σ, $\hat{G}(\sigma)$ is continuous on Σ. Therefore, a minimizer of $\hat{G}(\sigma)$ in Σ exists: $\hat{\sigma} = \arg\inf_{\sigma\in\Sigma} \hat{G}(\sigma)$. Thus,

$$\inf_{\sigma\in\Sigma} \max_{\alpha} \hat{R}(\alpha, \sigma) = \inf_{\sigma\in\Sigma} \hat{G}(\sigma) = \hat{G}(\hat{\sigma}) = \max_{\alpha} \hat{R}(\alpha, \hat{\sigma}).$$

Moreover the maximizer of $\hat{R}(\alpha, \hat{\sigma})$ always exists. This tells us that the general optimum of $\hat{R}(\alpha, \sigma)$ is achievable. By the relationship between the primal problem and its dual, we obtain the existence of the multi-kernel regularization scheme (1.5). This completes the proof of Proposition 1. □

References

[1] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
[2] P.L. Bartlett, O. Bousquet, S. Mendelson, Local Rademacher complexities, Ann. Statist. 33 (2005) 1497–1537.
[3] P.L. Bartlett, M.I. Jordan, J.D. McAuliffe, Convexity, classification, and risk bounds, J. Amer. Statist. Assoc. 101 (2006) 138–156.
[4] B. Blanchard, O. Bousquet, P. Massart, Statistical performance of support vector machines, preprint, 2003.
[5] B. Blanchard, G. Lugosi, N. Vayatis, On the rate of convergence of regularized boosting classifiers, J. Mach. Learning Res. 4 (2003) 861–894.
[7] S. Boucheron, O. Bousquet, G. Lugosi, Theory of classification: a survey of some recent advances, ESAIM: Probab. Statist. 9 (2005) 323–375.
[8] O. Bousquet, A. Elisseeff, Stability and generalization, J. Mach. Learning Res. 2 (2002) 499–526.
[10] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Choosing multiple parameters for support vector machines, Mach. Learning 46 (2002) 131–159.
[11] D.R. Chen, Q. Wu, Y. Ying, D.X. Zhou, Support vector machine soft margin classifiers: error analysis, J. Mach. Learning Res. 5 (2004) 1143–1175.
[12] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learning 20 (1995) 273–297.
[13] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, MA, 2000.
[14] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. 39 (2001) 1–49.
[15] F. Cucker, S. Smale, Best choices for regularization parameters in learning theory: on the bias–variance problem, Found. Comput. Math. 2 (2002) 413–428.
[16] F. Cucker, D.X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, Cambridge, MA, in press.
[17] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1997.
[18] T. Evgeniou, M. Pontil, Regularized multi-task learning, Proceedings of the 17th SIGKDD Conference on Knowledge Discovery and Data Mining, 2004.
[19] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.
[20] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, M.I. Jordan, Learning the kernel matrix with semidefinite programming, J. Mach. Learning Res. 5 (2004) 27–72.
[21] W.S. Lee, P.L. Bartlett, R.C. Williamson, Efficient agnostic learning of neural networks with bounded fan-in, IEEE Trans. Inform. Theory 42 (1996) 2118–2132.
[22] J. Li, A. Barron, Mixture density estimation, in: S.A. Solla, T.K. Leen, K.R. Muller (Eds.), Advances in Neural Information Processing Systems, vol. 12, Morgan Kaufmann Publishers, San Mateo, 1999.
[23] Y. Lin, Support vector machines and the Bayes rule in classification, Data Mining Knowledge Discovery 6 (2002) 259–275.
[24] G. Lugosi, N. Vayatis, On the Bayes-risk consistency of regularized boosting methods, Ann. Statist. 32 (2004) 30–55.
[25] P. Massart, Some applications of concentration inequalities to statistics, Ann. Fac. Sci. Toulouse Ser. 6 (9) (2000) 245–303.
[26] S. Mendelsen, Improving the sample complexity using global data, IEEE Trans. Inform. Theory 48 (2002) 1977–1991.

[27] C.A. Micchelli, M. Pontil, Learning the kernel function via regularization, J. Mach. Learning Res. 6 (2005) 1099–1125.
[28] S. Mukherjee, R. Rifkin, T. Poggio, Regression and classification with regularization, in: D.D. Denison, M.H. Hansen, C.C. Holmes, B. Mallick, B. Yu (Eds.), Lecture Notes in Statistics: Nonlinear Estimation and Classification, Springer, New York, 2002, pp. 107–124.
[29] A. Rakhlin, D. Panchenko, S. Mukherjee, Risk bounds for mixture density estimation, ESAIM: Probab. Statist. 9 (2005) 220–229.
[30] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, M. Anthony, Structural risk minimization over data-dependent hierarchies, IEEE Trans. Inform. Theory 44 (1998) 1926–1940.
[31] S. Smale, D.X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. 1 (2003) 17–41.
[32] S. Smale, D.X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. 41 (2004) 279–305.
[33] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their applications, Constr. Approx., in press.
[34] I. Steinwart, On the influence of the kernel on the consistency of support vector machines, J. Mach. Learning Res. 2 (2001) 67–93.
[35] I. Steinwart, Support vector machines are universally consistent, J. Complexity 18 (2002) 768–791.
[36] I. Steinwart, C. Scovel, Fast rates for support vector machines, in: Proceedings of the Conference on Learning Theory (COLT-2005), 2005, pp. 279–294.
[37] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (1999) 293–300.
[39] A.B. Tsybakov, Optimal aggregation of classifiers in statistical learning, Ann. Statist. 32 (2004) 135–166.
[40] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer, New York, 1996.
[41] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[42] G. Wahba, Spline Models for Observational Data, SIAM, Philadelphia, PA, 1990.
[43] Q. Wu, D.X. Zhou, SVM soft margin classifiers: linear programming versus quadratic programming, Neural Comput. 17 (2005) 1160–1187.
[44] Q. Wu, D.X. Zhou, Analysis of support vector machine classification, J. Comput. Anal. Appl. 8 (2006) 99–119.
[45] Y. Ying, D.X. Zhou, Learnability of Gaussians with flexible variances, preprint, 2004.
[46] T. Zhang, On the dual formulation of regularized linear systems with convex risks, Mach. Learning 46 (2002) 91–129.
[47] T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, Ann. Statist. 32 (2004) 56–85.
[48] D.X. Zhou, The covering number in learning theory, J. Complexity 18 (2002) 739–767.
[49] D.X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory 49 (2003) 1743–1752.
[50] D.X. Zhou, Density problem and approximation error in learning theory, preprint, 2003.