Journal of Machine Learning Research 5 (2004) 1143–1175 Submitted 9/03; Revised 4/04; Published 9/04

Support Vector Machine Soft Margin Classifiers: Error Analysis

Di-Rong Chen DRCHEN@BUAA.EDU.CN
Department of Applied Mathematics
Beijing University of Aeronautics and Astronautics
Beijing 100083, P. R. CHINA

Qiang Wu WU.QIANG@STUDENT.CITYU.EDU.HK
Yiming Ying YMYING@CITYU.EDU.HK
Ding-Xuan Zhou MAZHOU@MATH.CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Kowloon, Hong Kong, P. R. CHINA

Editor: Peter Bartlett

Abstract

The purpose of this paper is to provide a PAC error analysis for the q-norm soft margin classifier, a support vector machine classification algorithm. It consists of two parts: regularization error and sample error. While many techniques are available for treating the sample error, much less is known for the regularization error and the corresponding approximation error for reproducing kernel Hilbert spaces. We are mainly concerned about the regularization error. It is estimated for general distributions by a K-functional in weighted L^q spaces. For weakly separable distributions (i.e., the margin may be zero) satisfactory convergence rates are provided by means of separating functions. A projection operator is introduced, which leads to better sample error estimates especially for small complexity kernels. The misclassification error is bounded by the V-risk associated with a general class of loss functions V. The difficulty of bounding the offset is overcome. Polynomial kernels and Gaussian kernels are used to demonstrate the main results. The choice of the regularization parameter plays an important role in our analysis.

Keywords: support vector machine classification, misclassification error, q-norm soft margin classifier, regularization error, approximation error

1. Introduction

In this paper we study support vector machine (SVM) classification algorithms and investigate the SVM q-norm soft margin classifier with $1 < q < \infty$. Our purpose is to provide an error analysis for this algorithm in the PAC framework.

Let (X, d) be a compact metric space and $Y = \{1, -1\}$. A binary classifier $f : X \to \{1, -1\}$ is a function from X to Y which divides the input space X into two classes.

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

$$\mathcal{R}(f) := \text{Prob}\{f(\mathcal{X}) \neq \mathcal{Y}\} = \int_X P(\mathcal{Y} \neq f(x)|x) d\rho_X(x). \tag{1}$$

Here ρ_X is the marginal distribution on X and $P(\cdot|x)$ is the conditional probability measure given $\mathcal{X} = x$.

The SVM q-norm soft margin classifier (Cortes and Vapnik, 1995; Vapnik, 1998) is constructed from samples and depends on a reproducing kernel Hilbert space associated with a Mercer kernel.

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

The *Reproducing Kernel Hilbert Space* (RKHS) $\mathcal{H}_K$ associated with the kernel K is defined (Aronszajn, 1950) to be the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K} = \langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_y \rangle_K = K(x, y)$ and

$$\langle K_x, g \rangle_K = g(x), \qquad \forall x \in X, g \in \mathcal{H}_K.$$

Denote $C(X)$ as the space of continuous functions on X with the norm $\|\cdot\|_\infty$. Let $\kappa := \sqrt{\|K\|_\infty}$. Then the above reproducing property tells us that

$$\|g\|_\infty \le \kappa \|g\|_K, \qquad \forall g \in \mathcal{H}_K. \tag{2}$$

Define $\overline{\mathcal{H}}_K := \mathcal{H}_K + \mathbb{R}$. For a function $f = f_1 + b$ with $f_1 \in \mathcal{H}_K$ and $b \in \mathbb{R}$, we denote $f^* = f_1$ and $b_f = b \in \mathbb{R}$. The constant term b is called the *offset*. For a function $f : X \to \mathbb{R}$, the sign function is defined as $\operatorname{sgn}(f)(x) = 1$ if $f(x) \ge 0$ and $\operatorname{sgn}(f)(x) = -1$ if $f(x) < 0$.

Now the *SVM q-norm soft margin classifier* (SVM q-classifier) associated with the Mercer kernel K is defined as $\operatorname{sgn}(f_{\mathbf{z}})$, where $f_{\mathbf{z}}$ is a minimizer of the following optimization problem involving a set of random samples $\mathbf{z} = (x_i, y_i)_{i=1}^m \in Z^m$ independently drawn according to ρ:

$$\begin{aligned} f_{\mathbf{z}} := \arg \min_{f \in \overline{\mathcal{H}}_K} \quad & \frac{1}{2}\|f^*\|_K^2 + \frac{C}{m}\sum_{i=1}^m \xi_i^q, \\ \text{subject to} \quad & y_i f(x_i) \ge 1 - \xi_i, \text{ and } \xi_i \ge 0 \text{ for } i = 1, \ldots, m. \end{aligned} \tag{3}$$

Here C is a constant which depends on m: $C = C(m)$, and often $\lim_{m \to \infty} C(m) = \infty$.

Throughout the paper, we assume $1 < q < \infty$, $m \in \mathbb{N}$, $C > 0$, and $\mathbf{z} = (x_i, y_i)_{i=1}^m$ are random samples independently drawn according to ρ. Our target is to understand how $\operatorname{sgn}(f_{\mathbf{z}})$ converges (with respect to the misclassification error) to the best classifier, the Bayes rule, as m and hence $C(m)$ tend to infinity. Recall the regression function of ρ:

$$f_\rho(x) = \int_Y y d\rho(y|x) = P(\mathcal{Y} = 1|x) - P(\mathcal{Y} = -1|x), \qquad x \in X. \tag{4}$$

Then the *Bayes rule* is given (e.g. Devroye, L. Györfi and G. Lugosi, 1997) by the sign of the regression function $f_c := \operatorname{sgn}(f_\rho)$. Estimating the excess misclassification error

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \tag{5}$$

for the classification algorithm (3) is our goal. In particular, we try to understand how the choice of the regularization parameter C affects the error.

To investigate the error bounds for general kernels, we rewrite (3) as a regularization scheme. Define the loss function $V = V_q$ as

$$V(y, f(x)) := (1 - yf(x))_+^q = |y - f(x)|^q \chi_{\{yf(x) \le 1\}}, \tag{6}$$

where $(t)_+ = \max\{0, t\}$. The corresponding *V-risk* is

$$\mathcal{E}(f) := E(V(y, f(x))) = \int_Z V(y, f(x))d\rho(x,y). \tag{7}$$

If we set the empirical error as

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m}\sum_{i=1}^{m} V(y_i, f(x_i)) = \frac{1}{m}\sum_{i=1}^{m}(1 - y_i f(x_i))_+^q, \tag{8}$$

then the scheme (3) can be written as (see Evgeniou, Pontil and Poggio, 2000)

$$f_{\mathbf{z}} = \arg\min_{f \in \overline{\mathcal{H}_K}} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \frac{1}{2C}\|f^*\|_K^2 \right\}. \tag{9}$$

Notice that when $\overline{\mathcal{H}_K}$ is replaced by $\mathcal{H}_K$, the scheme (9) is exactly the Tikhonov regularization scheme (Tikhonov and Arsenin, 1977) associated with the loss function V. So one may hope that the method for analyzing regularization schemes can be applied.

The definitions of the V-risk (7) and the empirical error (8) tell us that for a function $f = f^* + b \in \overline{\mathcal{H}_K}$, the random variable $\xi = V(y, f(x))$ on Z has the mean $\mathcal{E}(f)$ and $\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) = \mathcal{E}_{\mathbf{z}}(f)$. Thus we may expect by some standard empirical risk minimization (ERM) argument (e.g. Cucker and Smale, 2001; Evgeniou, Pontil and Poggio, 2000; Shawe-Taylor et al., 1998; Vapnik, 1998; Wahba, 1990) to derive bounds for $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_q)$, where f_q is a minimizer of the V-risk (7). It was shown in Lin (2002) that for $q > 1$ such a minimizer is given by

$$f_q(x) = \frac{(1 + f_\rho(x))^{1/(q-1)} - (1 - f_\rho(x))^{1/(q-1)}}{(1 + f_\rho(x))^{1/(q-1)} + (1 - f_\rho(x))^{1/(q-1)}}, \qquad x \in X. \tag{10}$$

For $q = 1$ a minimizer is f_c, see Wahba (1999). Note that $\operatorname{sgn}(f_q) = f_c$.

Recall that for the classification algorithm (3), we are interested in the excess misclassification error (5), not the excess V-risk $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_q)$. But we shall see in Section 3 that $\mathcal{R}(\operatorname{sgn}(f)) - \mathcal{R}(f_c) \le \sqrt{2(\mathcal{E}(f) - \mathcal{E}(f_q))}$. One might apply this to the function $f_{\mathbf{z}}$ and get estimates for (5). However, special features of the loss function (6) enable us to do better: by restricting $f_{\mathbf{z}}$ onto $[-1, 1]$, we can improve the sample error estimates. The idea of the following projection operator was introduced for this purpose in Bartlett (1998).

Definition 1 *The* projection operator π *is defined on the space of measurable functions* $f : X \to \mathbb{R}$ *as*

$$\pi(f)(x) = \begin{cases} 1, & \textit{if } f(x) \ge 1, \\ -1, & \textit{if } f(x) \le -1, \\ f(x), & \textit{if } -1 < f(x) < 1. \end{cases} \tag{11}$$

It is trivial that $\operatorname{sgn}(\pi(f)) = \operatorname{sgn}(f)$. Hence

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \le \sqrt{2(\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q))}. \tag{12}$$

The definition of the loss function (6) also tells us that $V(y, \pi(f)(x)) \le V(y, f(x))$, so

$$\mathcal{E}(\pi(f)) \le \mathcal{E}(f) \qquad \text{and} \qquad \mathcal{E}_{\mathbf{z}}(\pi(f)) \le \mathcal{E}_{\mathbf{z}}(f). \tag{13}$$

According to (12), we need to estimate $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q)$ in order to bound (5). To this end, we introduce a *regularizing function* $f_{K,C} \in \overline{\mathcal{H}_K}$. It is arbitrarily chosen and depends on C.

Proposition 2 *Let $f_{K,C} \in \overline{\mathcal{H}}_K$, and $f_{\mathbf{z}}$ be defined by (9). Then $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q)$ can be bounded by*

$$\left\{\mathcal{E}(f_{K,C}) - \mathcal{E}(f_q) + \frac{1}{2C}\|f^*_{K,C}\|_K^2\right\} + \left\{\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + \mathcal{E}_{\mathbf{z}}(f_{K,C}) - \mathcal{E}(f_{K,C})\right\}. \tag{14}$$

Proof Decompose the difference $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q)$ as

$$\left\{\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}}))\right\} + \left\{\left(\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) + \frac{1}{2C}\|f^*_{\mathbf{z}}\|_K^2\right) - \left(\mathcal{E}_{\mathbf{z}}(f_{K,C}) + \frac{1}{2C}\|f^*_{K,C}\|_K^2\right)\right\}$$
$$+\left\{\mathcal{E}_{\mathbf{z}}(f_{K,C}) - \mathcal{E}(f_{K,C})\right\} + \left\{\mathcal{E}(f_{K,C}) - \mathcal{E}(f_q) + \frac{1}{2C}\|f^*_{K,C}\|_K^2\right\} - \frac{1}{2C}\|f^*_{\mathbf{z}}\|_K^2.$$

By the definition of $f_{\mathbf{z}}$ and (13), the second term is ≤ 0. Then the statement follows. ■

The first term in (14) is called the regularization error (Smale and Zhou, 2004). It can be expressed as a generalized K-functional $\inf_{f \in \overline{\mathcal{H}}_K} \left\{\mathcal{E}(f) - \mathcal{E}(f_q) + \frac{1}{2C}\|f^*\|_K^2\right\}$ when $f_{K,C}$ takes a special choice $\widetilde{f}_{K,C}$ (a standard choice in the literature, e.g. Steinwart 2001) defined as

$$\widetilde{f}_{K,C} := \arg\min_{f \in \overline{\mathcal{H}}_K} \left\{\mathcal{E}(f) + \frac{1}{2C}\|f^*\|_K^2\right\}. \tag{15}$$

Definition 3 *Let V be a general loss function and f^V_ρ be a minimizer of the V-risk (7). The* regularization error *for the regularizing function $f_{K,C} \in \overline{\mathcal{H}}_K$ is defined as*

$$\mathcal{D}(C) := \mathcal{E}(f_{K,C}) - \mathcal{E}(f^V_\rho) + \frac{1}{2C}\|f^*_{K,C}\|_K^2. \tag{16}$$

It is called the regularization error of the scheme *(9) when $f_{K,C} = \widetilde{f}_{K,C}$:*

$$\widetilde{\mathcal{D}}(C) := \inf_{f \in \overline{\mathcal{H}}_K} \left\{\mathcal{E}(f) - \mathcal{E}(f^V_\rho) + \frac{1}{2C}\|f^*\|_K^2\right\}.$$

The main concern of this paper is the regularization error. We shall investigate its asymptotic behavior. This investigation is not only important for bounding the first term in (14), but also crucial for bounding the second term (sample error): it is well known in structural risk minimization (Shawe-Taylor et al., 1998) that the size of the hypothesis space is essential. This is determined by $\mathcal{D}(C)$ in our setting. Once C is fixed, the sample error estimate becomes routine. Therefore, we need to understand the choice of the parameter C from the bound for $\mathcal{D}(C)$.

Proposition 4 *For any $C \geq 1/2$ and any $f_{K,C} \in \overline{\mathcal{H}}_K$, there holds*

$$\mathcal{D}(C) \geq \widetilde{\mathcal{D}}(C) \geq \frac{\widetilde{\kappa}^2}{2C} \tag{17}$$

where

$$\widetilde{\kappa} := \mathcal{E}_0/(1 + \kappa q 4^{q-1}), \qquad \mathcal{E}_0 := \inf_{b \in \mathbb{R}} \{\mathcal{E}(b) - \mathcal{E}(f_q)\}. \tag{18}$$

Moreover, $\widetilde{\kappa} = 0$ if and only if for some $p_0 \in [0,1]$, $P(\mathcal{Y} = 1|x) = p_0$ in probability.

This proposition will be proved in Section 5.

According to Proposition 4, the decay of $\mathcal{D}(C)$ cannot be faster than $O(1/C)$ except for some very special distributions. This special case is caused by the offset in (9), for which $\widetilde{\mathcal{D}}(C) \equiv 0$. Throughout this paper we shall ignore this trivial case and assume $\widetilde{\kappa} > 0$.

When ρ is strictly separable, $\mathcal{D}(C) = O(1/C)$. But this is a very special phenomenon. In general, one should not expect $\mathcal{E}(f) = \mathcal{E}(f_q)$ for some $f \in \overline{\mathcal{H}}_K$. Even for (weakly) separable distributions with zero margin, $\mathcal{D}(C)$ decays as $O(C^{-p})$ for some $0 < p < 1$. To realize such a decay for these separable distributions, the regularizing function will be multiples of a separating function. For details and the concepts of strictly or weakly separable distributions, see Section 2.

For general distributions, we shall choose $f_{K,C} = \widetilde{f}_{K,C}$ in Sections 6 and 7 and estimate the regularization error of the scheme (9) associated with the loss function (6) by means of the approximation in the function space $L^q_{\rho_X}$. In particular, $\widetilde{\mathcal{D}}(C) \leq \mathcal{K}(f_q, \frac{1}{2C})$, where $\mathcal{K}(f_q,t)$ is a K-functional defined as

$$\mathcal{K}(f_q,t) := \begin{cases} \inf_{f \in \overline{\mathcal{H}}_K} \left\{ \|f - f_q\|^q_{L^q_{\rho_X}} + t\|f^*\|^2_K \right\}, & \text{if } 1 < q \leq 2, \\ \inf_{f \in \overline{\mathcal{H}}_K} \left\{ q2^{q-1}(2^{q-1}+1)\|f - f_q\|_{L^q_{\rho_X}} + t\|f^*\|^2_K \right\}, & \text{if } q > 2. \end{cases} \tag{19}$$

In the case $q = 1$, the regularization error (Wu and Zhou, 2004) depends on the approximation in $L^1_{\rho_X}$ of the function f_c which is not continuous in general. For $q > 1$, the regularization error depends on the approximation in $L^q_{\rho_X}$ of the function f_q. When the regression function has good smoothness, f_q has much higher regularity than f_c. Hence the convergence for $q > 1$ may be faster than that for $q = 1$, which improves the regularization error.

The second term in (14) is called the *sample error*. When C is fixed, the sample error $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})$ is well understood (except for the offset term). In (14), the sample error is for the function $\pi(f_{\mathbf{z}})$ instead of $f_{\mathbf{z}}$ while the misclassification error (5) is kept: $\mathcal{R}(\text{sgn}(\pi(f_{\mathbf{z}}))) = \mathcal{R}(\text{sgn}(f_{\mathbf{z}}))$. Since the bound for $V(y, \pi(f)(x))$ is much smaller than that for $V(y, f(x))$, the projection improves the sample error estimate.

Based on estimates for the regularization error and sample error above, our error analysis will provide $\varepsilon(\delta, m, C, \beta) > 0$ for any $0 < \delta < 1$ such that with confidence $1 - \delta$,

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q) \leq (1+\beta)\{\mathcal{D}(C) + \varepsilon(\delta, m, C, \beta)\}. \tag{20}$$

Here $0 < \beta \leq 1$ is an arbitrarily fixed number. Moreover, $\lim_{m \to \infty} \varepsilon(\delta, m, C, \beta) = 0$.

If f_q lies in the $L^q_{\rho_X}$-closure of $\overline{\mathcal{H}}_K$, then $\lim_{t \to 0} \mathcal{K}(f_q, t) = 0$ by (19). Hence $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q) \to 0$ with confidence as m (and hence $C = C(m)$) becomes large. This is the case when K is a universal kernel, i.e., $\mathcal{H}_K$ is dense in $C(X)$, or when a sequence of kernels whose RKHS tends to be dense (e.g. polynomial kernels with increasing degrees) is used.

In summary, estimating the excess misclassification error $\mathcal{R}(\text{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c)$ consists of three parts: the comparison (12), the regularization error $\mathcal{D}(C)$ and the sample error $\varepsilon(\delta, m, C, \beta)$ in (20). As functions of the variable C, $\mathcal{D}(C)$ decreases while $\varepsilon(\delta, m, C, \beta)$ usually increases. Choosing suitable values for the regularization parameter C will give the optimal convergence rate. To this end, we need to consider the tradeoff between these two errors. This can be done by minimizing the right side of (20), as shown in the following form.

Lemma 5 *Let $p,\alpha,\tau > 0$. Denote $c_{p,\alpha,\tau} := (p/\tau)^{\frac{\tau}{\tau+p}} + (\tau/p)^{\frac{p}{\tau+p}}$. Then for any $C > 0$,*

$$C^{-p} + \frac{C^\tau}{m^\alpha} \geq c_{p,\alpha,\tau}\left(\frac{1}{m}\right)^{\frac{\alpha p}{\tau+p}}. \tag{21}$$

The equality holds if and only if $C = (p/\tau)^{\frac{1}{\tau+p}} m^{\frac{\alpha}{\tau+p}}$. This yields the optimal power $\frac{\alpha p}{\tau+p}$.

The goal of the regularization error estimates is to have p in $\mathcal{D}(C) = O(C^{-p})$, as large as possible. But $p \leq 1$ according to Proposition 4. Good methods for sample error estimates provide large α and small τ such that $\varepsilon(\delta, m, C, \beta) = O(\frac{C^\tau}{m^\alpha})$. Notice that, as always, both the approximation properties (represented by the exponent p) and the estimation properties (represented by the exponents τ and α) are important to get good estimates for the learning rates (with the optimal rate $\frac{\alpha p}{\tau+p}$).

2. Demonstrating with Weakly Separable Distributions

With some special cases let us demonstrate how our error analysis yields some guidelines for choosing the regularization parameter C. For weakly separable distributions, we also compare our results with bounds in the literature. To this end, we need the covering number of the unit ball $\mathcal{B} := \{f \in \mathcal{H}_K : \|f\|_K \leq 1\}$ of $\mathcal{H}_K$ (considered as a subset of $C(X)$).

Definition 6 *For a subset $\mathcal{F}$ of a metric space and $\eta > 0$, the* covering number $\mathcal{N}(\mathcal{F}, \eta)$ *is defined to be the minimal integer $\ell \in \mathbb{N}$ such that there exist ℓ disks with radius η covering $\mathcal{F}$.*

Denote the covering number of $\mathcal{B}$ in $C(X)$ by $\mathcal{N}(\mathcal{B}, \eta)$. Recall the constant $\widetilde{\kappa}$ defined in (18). Since the algorithm involves an offset term, we also need its covering number and set

$$\mathcal{N}(\eta) := \left\{(\kappa + \frac{1}{\widetilde{\kappa}})\frac{1}{\eta} + 1\right\}\mathcal{N}(\mathcal{B}, \eta). \tag{22}$$

Definition 7 *We say that the Mercer kernel K has* logarithmic complexity exponent $s \geq 1$ *if for some $c > 0$, there holds*

$$\log \mathcal{N}(\eta) \leq c\,(\log(1/\eta))^s, \qquad \forall \eta > 0. \tag{23}$$

The kernel K has polynomial complexity exponent $s > 0$ *if*

$$\log \mathcal{N}(\eta) \leq c\,(1/\eta)^s, \qquad \forall \eta > 0. \tag{24}$$

The covering number $\mathcal{N}(\mathcal{B}, \eta)$ has been extensively studied (see e.g. Bartlett, 1998; Williamson, Smola and Schölkopf, 2001; Zhou, 2002; Zhou, 2003a). In particular, for convolution type kernels $K(x,y) = k(x-y)$ with $\hat{k} \geq 0$ decaying exponentially fast, (23) holds (Zhou, 2002, Theorem 3). As an example, consider the Gaussian kernel $K(x,y) = \exp\{-|x-y|^2/\sigma^2\}$ with $\sigma > 0$. If $X \subset [0,1]^n$ and $0 < \eta \leq \exp\{90n^2/\sigma^2 - 11n - 3\}$, then (23) is valid (Zhou, 2002) with $s = n+1$. A lower bound (Zhou, 2003a) holds with $s = \frac{n}{2}$, which shows the upper bound is almost sharp. It was also shown in (Zhou, 2003a) that K has polynomial complexity exponent $2n/p$ if K is C^p.

To demonstrate our main results concerning the regularization parameter C, we shall only choose kernels having logarithmic complexity exponents or having polynomial complexity exponents with s small. This means, $\overline{\mathcal{H}_K}$ has small complexity.

The special case we consider here is the deterministic case: $\mathcal{R}(f_c) = 0$. Understanding how to find $\mathcal{D}(C)$ and $\varepsilon(\delta, m, C, \beta)$ in (20) and then how to choose the parameter C is our target. For $M \geq 0$, denote

$$\theta_M = \begin{cases} 0, & \text{if } M = 0, \\ 1, & \text{if } M > 0. \end{cases} \tag{25}$$

Proposition 8 *Suppose $\mathcal{R}(f_c) = 0$. If $f_{K,C} \in \overline{\mathcal{H}}_K$ satisfies $\|V(y, f_{K,C}(x))\|_\infty \leq M$, then for every $0 < \delta < 1$, with confidence at least $1 - \delta$, we have*

$$\mathcal{R}(sgn(f_{\mathbf{z}})) \leq 2 \max\left\{\varepsilon^*, \frac{4M \log(2/\delta)}{m}\right\} + 4\mathcal{D}(C), \tag{26}$$

where with $\mathcal{M} := \sqrt{2C\mathcal{D}(C) + 2C\varepsilon\theta_M}$, $\varepsilon^* > 0$ is the unique solution to the equation

$$\log \mathcal{N}\left(\frac{\varepsilon}{q2^{q+3}\mathcal{M}}\right) - \frac{3m\varepsilon}{2^{q+9}} = \log(\delta/2). \tag{27}$$

(a) *If (23) is valid, then with $\widetilde{c} = 2^{q+9}\{1 + c((q+4)\log 8)^s\}$,*

$$\varepsilon^* \leq \widetilde{c}\left(\frac{(\log m + \log(C\mathcal{D}(C)) + \log(C\theta_M))^s + \log(2/\delta)}{m}\right).$$

(b) *If (24) is valid and $C \geq 1$, then with a constant $\widetilde{c}$ (given explicitly in the proof),*

$$\varepsilon^* \leq \widetilde{c} \log(2/\delta)\left\{\frac{(2C\mathcal{D}(C))^{s/(2s+2)}}{m^{1/(s+1)}} + \frac{(2C)^{s/(s+2)}}{m^{1/(\frac{s}{2}+1)}}\right\}. \tag{28}$$

Note that the above constant $\widetilde{c}$ depends on c, q, and κ, but not on C, m or δ. The proof of Proposition 8 will be given in Section 5.

We can now derive our error bound for weakly separable distributions.

Definition 9 *We say that ρ is* (weakly) separable *by $\overline{\mathcal{H}}_K$ if there is a function $f_{sp} \in \overline{\mathcal{H}}_K$, called a* separating function, *satisfying $\|f^*_{sp}\|_K = 1$ and $y f_{sp}(x) > 0$ almost everywhere. It has* separation exponent $\theta \in (0, +\infty]$ *if there are positive constants γ, c' such that*

$$\rho_X\{x \in X : |f_{sp}(x)| < \gamma t\} \leq c' t^\theta, \qquad \forall\, t > 0. \tag{29}$$

Observe that condition (29) with $\theta = +\infty$ is equivalent to

$$\rho_X\{x \in X : |f_{\text{sp}}(x)| < \gamma t\} = 0, \qquad \forall\, 0 < t < 1.$$

That is, $|f_{\text{sp}}(x)| \geq \gamma$ almost everywhere. Thus, weakly separable distributions with separation exponent $\theta = +\infty$ are exactly strictly separable distributions. Recall (e.g. Vapnik, 1998; Shawe-Taylor et al., 1998) that ρ is said to be *strictly separable* by $\overline{\mathcal{H}}_K$ with margin $\gamma > 0$ if ρ is (weakly) separable together with the requirement $y f_{\text{sp}}(x) \geq \gamma$ almost everywhere.

¿From the first part of Definition 9, we know that for a separable distribution ρ, the set $\{x : f_{\text{sp}}(x) = 0\}$ has ρ_X-measure zero, hence

$$\lim_{t \to 0} \rho_X\{x \in X : |f_{\text{sp}}(x)| < t\} = 0.$$

The separation exponent θ measures the asymptotic behavior of ρ near the boundary of two classes. It gives the convergence with a polynomial decay.

Theorem 10 *If ρ is separable and has separation exponent $\theta \in (0, +\infty]$ with (29) valid, then for every $0 < \delta < 1$, with confidence at least $1-\delta$, we have*

$$\mathcal{R}(sgn(f_{\mathbf{z}})) \le 2\varepsilon^* + \frac{8\log(2/\delta)}{m} + 4(2c')^{\frac{2}{\theta+2}}(C\gamma^2)^{-\frac{\theta}{\theta+2}}, \tag{30}$$

where ε^* is solved by

$$\log \mathcal{N}\left(\frac{\varepsilon}{q2^{q+3}\sqrt{2(2c')^{\frac{2}{\theta+2}}\gamma^{-\frac{2\theta}{\theta+2}}C^{\frac{2}{\theta+2}} + 2C\varepsilon}}\right) - \frac{3m\varepsilon}{2^{q+9}} = \log(\delta/2). \tag{31}$$

(a) *If (23) is satisfied, with a constant $\widetilde{c}$ depending on γ we have*

$$\mathcal{R}(sgn(f_{\mathbf{z}})) \le \widetilde{c}\left(\frac{\log(2/\delta) + (\log m + \log C)^s}{m} + C^{-\frac{\theta}{\theta+2}}\right). \tag{32}$$

(b) *If (24) is satisfied, there holds*

$$\mathcal{R}(sgn(f_{\mathbf{z}})) \le \widetilde{c}\left\{\log\frac{2}{\delta}\left(C^{\frac{s}{(s+1)(\theta+2)}}m^{-\frac{1}{s+1}} + C^{\frac{s}{s+2}}m^{-\frac{1}{s/2+1}}\right) + C^{-\frac{\theta}{\theta+2}}\right\}. \tag{33}$$

Proof Choose $t = (\frac{\gamma^\theta}{2c'C})^{1/(\theta+2)} > 0$ and the function $f_{K,C} = \frac{1}{t}f_{\text{sp}} \in \overline{\mathcal{H}}_K$. Then we have $\frac{1}{2C}\|f^*_{K,C}\|_K^2 = \frac{1}{2Ct^2}$.

Since $yf_{\text{sp}}(x) > 0$ almost everywhere, we know that for almost every $x \in X$, $y = \text{sgn}(f_{\text{sp}})(x)$. This means $f_\rho = \text{sgn}(f_{\text{sp}})$ and then $f_q = \text{sgn}(f_{\text{sp}})$. It follows that $\mathcal{R}(f_c) = 0$ and $V(y, f_{K,C}(x)) = (1 - \frac{|f_{\text{sp}}(x)|}{t})_+^q \in [0,1]$ almost everywhere. Therefore, we may take $M = 1$ in Proposition 8. Moreover,

$$\mathcal{E}(f_{K,C}) = \int_Z \left(1 - \frac{yf_{\text{sp}}(x)}{t}\right)_+^q d\rho = \int_X \left(1 - \frac{|f_{\text{sp}}(x)|}{t}\right)_+^q d\rho_X.$$

This can be bounded by (29) as

$$\int_{\{x:|f_{\text{sp}}(x)|<t\}} \left(1 - \frac{|f_{\text{sp}}(x)|}{t}\right)_+^q d\rho_X \le \rho_X\{x \in X : |f_{\text{sp}}(x)| < t\} \le c'\left(\frac{t}{\gamma}\right)^\theta.$$

It follows from the choice of t that

$$\mathcal{D}(C) \le c'\left(\frac{t}{\gamma}\right)^\theta + \frac{1}{2Ct^2} = (2c')^{\frac{2}{\theta+2}}(C\gamma^2)^{-\frac{\theta}{\theta+2}}. \tag{34}$$

Then our conclusion follows from Proposition 8. ■

In case (a), the bound (32) tells us to choose C such that $(\log C)^s/m \to 0$ and $C \to \infty$ as $m \to \infty$. By (32) a reasonable choice of the regularization parameter C is $C = m^{\frac{\theta+2}{\theta}}$, which yields $\mathcal{R}(\text{sgn}(f_{\mathbf{z}})) = O(\frac{(\log m)^s}{m})$. In case (b),

$$\text{when (24) is valid,} \quad C = m^{\frac{\theta+2}{\theta+s+\theta s}} \Longrightarrow \mathcal{R}(\text{sgn}(f_{\mathbf{z}})) = O(m^{-\frac{\theta}{\theta+s+\theta s}}). \tag{35}$$

The following is a typical example of separable distribution which is not strictly separable. In this example, the separation exponent is $\theta = 1$.

Example 1 *Let $X = [-1/2, 1/2]$ and ρ be the Borel probability measure on Z such that ρ_X is the Lebesgue measure on X and*

$$f_\rho(x) = \begin{cases} 1, & \text{for } -1/2 \le x \le -1/4 \text{ and } 1/4 \le x \le 1/2, \\ -1, & \text{for } -1/4 \le x < 1/4. \end{cases}$$

(a) *If K is the linear polynomial kernel $K(x,y) = x \cdot y$, then $\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) \ge 1/4$.*

(b) *If K is the quadratic polynomial kernel $K(x,y) = (x \cdot y)^2$, then with confidence $1-\delta$,*

$$\mathcal{R}(sgn(f_{\mathbf{z}})) \le \frac{q(q+4)2^{q+11}\left(\log m + \log C + 2\log(2/\delta)\right)}{m} + \frac{16}{C^{1/3}}.$$

Hence one should take C such that $\log C/m \to 0$ and $C \to \infty$ as $m \to \infty$.

Proof The first statement is trivial.

To see (b), we note that $\dim \mathcal{H}_K = 1$ and $\kappa = 1/4$. Also, $\mathcal{E}_0 = \inf_{b \in \mathbb{R}} \mathcal{E}(b) = 1$. Hence $\widetilde{\kappa} = 1/(1+q4^{q-2})$. Since $\mathcal{H}_K = \{ax^2 : a \in \mathbb{R}\}$ and $\|ax^2\|_K = |a|$, $\mathcal{N}(\mathcal{B}, \eta) \le 1/(2\eta)$. Then (23) holds with $s = 1$ and $c = 4q$. By Proposition 8, we see that $\varepsilon^* \le \frac{q(q+4)2^{q+11}(\log(Cm)+\log(2/\delta))}{m}$.

Take the function $f_{\rm sp}(x) = x^2 - 1/16 \in \overline{\mathcal{H}}_K$ with $\|f^*_{\rm sp}\|_K = 1$. We see that $y f_{\rm sp}(x) > 0$ almost everywhere. Moreover,

$$\{x \in X : |f_{\rm sp}(x)| < t\} \subseteq \left[\frac{\sqrt{1-16t}}{4}, \frac{\sqrt{1+16t}}{4}\right] \bigcup \left[-\frac{\sqrt{1+16t}}{4}, -\frac{\sqrt{1-16t}}{4}\right].$$

The measure of this set is bounded by $8\sqrt{2}t$. Hence (29) holds with $\theta = 1$, $\gamma = 1$ and $c' = 8\sqrt{2}$. Then Theorem 10 gives the stated bound for $\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}}))$. ■

In this paper, for the sample error estimates we only use the (uniform) covering numbers in $C(X)$. Within the last a few years, the empirical covering numbers (in ℓ^∞ or ℓ^2), the leave-one-out error or stability analysis (e.g. Vapnik, 1998; Bousquet and Ellisseeff, 2002; Zhang, 2004), and some other advanced empirical process techniques such as the local Rademacher averages (van der Vaart and Wellner, 1996; Bartlett, Bousquet and Mendelson, 2004; and references therein) and the entropy integrals (van der Vaart and Wellner, 1996) have been developed to get better sample error estimates. These techniques can be applied to various learning algorithms. They are powerful to handle general hypothesis spaces even with large capacity.

In Zhang (2004) the leave-one-out technique was applied to improve the sample error estimates given in Bousquet and Ellisseeff (2002): the sample error has a kernel-independent bound $O(\frac{C}{m})$, improving the bound $O(\frac{C}{\sqrt{m}})$ in Bousquet and Ellisseeff (2002); while the regularization error $\widetilde{\mathcal{D}}(C)$ depends on K and ρ. In particular, for $q = 2$ the bound (Zhang, 2004, Corollary 4.2) takes the form:

$$E(\mathcal{E}(f_{\mathbf{z}})) \le \Big(1 + \frac{4\kappa^2 C}{m}\Big)^2 \inf_{f \in \mathcal{H}_K} \Big\{\mathcal{E}(f) + \frac{1}{2C}\|f\|_K^2\Big\} = \Big(1 + \frac{4\kappa^2 C}{m}\Big)^2 \Big\{\widetilde{\mathcal{D}}(C) + \mathcal{E}(f_q)\Big\}. \tag{36}$$

In Bartlett, Jordan and McAuliffe (2003) the empirical process techniques are used to improve the sample error estimates for ERM. In particular, Theorem 12 there states that for a convex set $\mathcal{F}$ of functions on X, the minimizer $\widehat{f}$ of the empirical error over $\mathcal{F}$ satisfies

$$\mathcal{E}(\widehat{f}) \le \inf_{f \in \mathcal{F}} \mathcal{E}(f) + K \max\Big\{\varepsilon^*, \Big(\frac{c_r L^2 \log(1/\delta)}{m}\Big)^{1/(2-\beta)}, \frac{BL\log(1/\delta)}{m}\Big\}. \tag{37}$$

Here K, c_r, β are constants, and ε^* is solved by an inequality. The constant B is a bound for differences, and L is the Lipschitz constant for the loss ϕ with respect to a pseudometric on $\mathbb{R}$. For more details, see Bartlett, Jordan and McAuliffe (2003). The definitions of B and L together tell us that

$$|\phi(y_1 f(x_1)) - \phi(y_2 f(x_2))| \leq LB, \qquad \forall\ (x_1,y_1) \in Z, (x_2,y_2) \in Z, f \in \mathcal{F}. \tag{38}$$

Because of our improvement for the bound of the random variables $V(y, f(x))$ given by the projection operator, for the algorithm (3) the error bounds we derive here are better than existing results when the kernel has small complexity. Let us confirm this for separable distributions with separation exponent $0 < \theta < \infty$ and for kernels with polynomial complexity exponent $s > 0$ satisfying (24). Recall that $\mathcal{D}(C) = O(C^{-\frac{\theta}{\theta+2}})$. Take $q = 2$.

Consider the estimate (36). This together with (34) tells us that the derived bound for $E(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_q))$ is at least of the order $\widetilde{\mathcal{D}}(C) + \frac{C}{m}\widetilde{\mathcal{D}}(C) = O(C^{-\frac{\theta}{\theta+2}} + \frac{C^{\frac{2}{\theta+2}}}{m})$. By Lemma 5, the optimal bound derived from (36) is $O(m^{-\frac{\theta}{\theta+2}})$. Thus, our bound (35) is better than (36) when $0 < s < \frac{2}{\theta+1}$.

Turn to the estimate (37). Take $\mathcal{F}$ to be the ball of radius $\sqrt{2C\widetilde{\mathcal{D}}(C)}$ of $\mathcal{H}_K$ (the expected smallest ball where $f_{\mathbf{z}}^*$ lies), we find that the constant LB in (38) should be at least $\kappa\sqrt{2C\widetilde{\mathcal{D}}(C)}$. This tells us that one term in the bound (37) for the sample error is at least $O\Big(\frac{\sqrt{2C\widetilde{\mathcal{D}}(C)}}{m}\Big) = O(\frac{C^{\frac{1}{\theta+2}}}{m})$, while the other two terms are more involved. Applying Lemma 5 again, we find that the bound derived from (37) is at least $O(m^{-\frac{\theta}{\theta+1}})$. So our bound (35) is better than (37) at least for $0 < s < \frac{1}{\theta+1}$.

Thus, our analysis with the help of the projection operator improves existing error bounds when the kernel has small complexity. Note that in (35), the values of s for which the projection operator gives an improvement are values corresponding to rapidly diminishing regularization ($C = m^\beta$ with $\beta > 1$ being large).

For kernels with large complexity, refined empirical process techniques (e.g. van der Vaart and Wellner, 1996; Mendelson, 2002; Zhang, 2004; Bartlett, Jordan and McAuliffe, 2003) should give better bounds: the capacity of the hypothesis space may have more influence on the sample error than the bound of the random variable $V(y, f(x))$, and the entropy integral is powerful to reduce this influence. To find the range of this large complexity, one needs explicit rate analysis for both the sample error and regularization error. This is out of our scope. It would be interesting to get better error bounds by combining the ideas of empirical process techniques and the projection operator.

Problem 11 *How much improvement can we get for the total error when we apply the projection operator to empirical process techniques?*

Problem 12 *Given $\theta > 0, s > 0, q \geq 1$, and $0 < \delta < 1$, what is the largest number $\alpha > 0$ such that $\mathcal{R}(sgn(f_{\mathbf{z}})) = O(m^{-\alpha})$ with confidence $1-\delta$ whenever the distribution ρ is separable with separation exponent θ and the kernel K satisfies (24)? What is the corresponding optimal choice of the regularization parameter C?*

In particular, for Example 1 we have the following.

Conjecture 13 *In Example 1, with confidence $1-\delta$ there holds $\mathcal{R}(sgn(f_{\mathbf{z}})) = O(\frac{\log(2/\delta)}{m})$ by choosing $C = m^3$.*

Consider the well known setting (Vapnik, 1998; Shawe-Taylor et al., 1998; Steinwart, 2001; Cristianini and Shawe-Taylor, 2000) of strictly separable distributions with margin $\gamma > 0$. In this case, Shawe-Taylor et al. (1998) shows that

$$\mathcal{R}(\operatorname{sgn}(f)) \leq \frac{2}{m}(\log \mathcal{N}(\mathcal{F}, 2m, \gamma/2) + \log(2/\delta)) \tag{39}$$

with confidence $1-\delta$ for $m > 2/\gamma$ whenever $f \in \mathcal{F}$ satisfies $y_i f(x_i) \geq \gamma$. Here $\mathcal{F}$ is a function set, $\mathcal{N}(\mathcal{F}, 2m, \gamma/2) = \sup_{\vec{t} \in X^{2m}} \mathcal{N}(\mathcal{F}, \vec{t}, \gamma/2)$ and $\mathcal{N}(\mathcal{F}, \vec{t}, \gamma/2)$ denotes the covering number of the set $\{(f(t_i))_{i=1}^{2m} : f \in \mathcal{F}\}$ in $\mathbb{R}^{2m}$ with the ℓ^∞ metric. For a comparison of this covering number with that in $C(X)$, see Pontil (2003).

In our analysis, we can take $f_{K,C} = \frac{1}{\gamma} f_{\mathrm{sp}} \in \overline{\mathcal{H}_K}$. Then $M = \|V(y, f_{K,C}(x))\|_\infty = 0$, and $\mathcal{D}(C) = \frac{1}{2C\gamma^2}$. We see from Proposition 8 (a) that when (23) is satisfied, there holds $\mathcal{R}(f_{\mathbf{z}}) \leq \widetilde{c}(\frac{(\log(1/\gamma)+\log m)^s + \log(1/\delta)}{m})$. But the optimal bound $O(\frac{1}{m})$ is also valid for spaces with larger complexity, which can be seen from (39) by the relation of the covering numbers: $\mathcal{N}(\mathcal{F}, 2m, \gamma/2) \leq \mathcal{N}(\mathcal{F}, \gamma/2)$. See also Steinwart (2001). We see the shortcoming of our approach for kernels with large complexity. This convinces the interest of Problem 11 raised above.

3. Comparison of Errors

In this section we consider how to bound the misclassification error by the V-risk. A systematic study of this problem was done for convex loss functions by Zhang (2004), and for more general loss functions by Bartlett, Jordan, and McAuliffe (2003). Using these results, it can be shown that for the loss function V_q, there holds

$$\psi\Big(\mathcal{R}(\operatorname{sgn}(f)) - \mathcal{R}(f_c)\Big) \leq \mathcal{E}(f) - \mathcal{E}(f_q),$$

where $\psi : [0,1] \to \mathbb{R}_+$ is a function defined in Bartlett, Jordan, and McAuliffe (2003) and can be explicitly computed.

In fact, we have

$$\mathcal{R}(\operatorname{sgn}(f)) - \mathcal{R}(f_c) \leq c\sqrt{\mathcal{E}(f) - \mathcal{E}(f_q)}. \tag{40}$$

Such a comparison of errors holds true even for a general convex loss function, see Theorem 34 in Appendix. The derived bound for the constant c in (40) need not be optimal. In the following we shall give an optimal estimate in a simpler form.

The constant we derive depends on q and is given by

$$C_q = \begin{cases} 1, & \text{if } 1 \leq q \leq 2, \\ 2(q-1)/q, & \text{if } q > 2. \end{cases} \tag{41}$$

We can see that $C_q \leq 2$.

Theorem 14 *Let $f : X \to \mathbb{R}$ be measurable. Then*

$$\mathcal{R}(\operatorname{sgn}(f)) - \mathcal{R}(f_c) \leq \sqrt{C_q\,(\mathcal{E}(f) - \mathcal{E}(f_q))} \leq \sqrt{2\,(\mathcal{E}(f) - \mathcal{E}(f_q))}.$$

Proof By the definition, only points with $\mathrm{sgn}(f)(x) \neq f_c(x)$ are involved for the misclassification error. Hence

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c) = \int_X |f_\rho(x)| \chi_{\{\mathrm{sgn}(f)(x) \neq \mathrm{sgn}(f_q)(x)\}} d\rho_X. \tag{42}$$

This in connection with the Schwartz inequality implies that

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c) \leq \left\{ \int_X |f_\rho(x)|^2 \chi_{\{\mathrm{sgn}(f)(x) \neq \mathrm{sgn}(f_q)(x)\}} d\rho_X \right\}^{1/2}.$$

Thus it is sufficient to show for those $x \in X$ with $\mathrm{sgn}(f)(x) \neq \mathrm{sgn}(f_q)(x)$,

$$|f_\rho(x)|^2 \leq C_q \left(\mathcal{E}(f|x) - \mathcal{E}(f_q|x) \right), \tag{43}$$

where for $x \in X$, we have denoted

$$\mathcal{E}(f|x) := \int_Y V_q(y, f(x)) d\rho(y|x). \tag{44}$$

By the definition of the loss function V_q, we have

$$\mathcal{E}(f|x) = (1 - f(x))_+^q \frac{1 + f_\rho(x)}{2} + (1 + f(x))_+^q \frac{1 - f_\rho(x)}{2}. \tag{45}$$

It follows that

$$\mathcal{E}(f|x) - \mathcal{E}(f_q|x) = \int_0^{f(x) - f_q(x)} F(u) du, \tag{46}$$

where $F(u)$ is the function (depending on the parameter $f_\rho(x)$):

$$F(u) = \frac{1 - f_\rho(x)}{2} q(1 + f_q(x) + u)_+^{q-1} - \frac{1 + f_\rho(x)}{2} q(1 - f_q(x) - u)_+^{q-1}, \quad u \in \mathbb{R}.$$

Since $F(u)$ is nondecreasing and $F(0) = 0$, we see from (46) that when $\mathrm{sgn}(f)(x) \neq \mathrm{sgn}(f_q)(x)$, there holds

$$\mathcal{E}(f|x) - \mathcal{E}(f_q|x) \geq \int_0^{-f_q(x)} F(u) du = \mathcal{E}(0|x) - \mathcal{E}(f_q|x) = 1 - \mathcal{E}(f_q|x). \tag{47}$$

But

$$\mathcal{E}(f_q|x) = \frac{2^{q-1} \left(1 - |f_\rho(x)|^2 \right)}{\left\{ (1 + f_\rho(x))^{1/(q-1)} + (1 - f_\rho(x))^{1/(q-1)} \right\}^{q-1}}. \tag{48}$$

Therefore, (43) is valid (with $t = f_\rho(x)$) once the following inequality is verified:

$$\frac{t^2}{C_q} \leq 1 - \frac{2^{q-1} \left(1 - t^2 \right)}{\left\{ (1 + t)^{1/(q-1)} + (1 - t)^{1/(q-1)} \right\}^{q-1}}, \qquad t \in [-1, 1].$$

This is the same as the inequality

$$\left(1 - \frac{t^2}{C_q} \right)^{-1/(q-1)} \leq \frac{(1 + t)^{-1/(q-1)} + (1 - t)^{-1/(q-1)}}{2}, \qquad t \in [-1, 1]. \tag{49}$$

To prove (49), we use the Taylor expansion

$$(1+u)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} u^k = \sum_{k=0}^{\infty} \frac{\prod_{\ell=0}^{k-1}(\alpha-\ell)}{k!} u^k, \qquad u \in (-1,1).$$

With $\alpha = -1/(q-1)$, we have

$$\left(1-\frac{t^2}{C_q}\right)^{-1/(q-1)} = \sum_{k=0}^{\infty} \frac{\prod_{\ell=0}^{k-1}\left(\frac{1}{q-1}+\ell\right)}{k!} \frac{1}{C_q^k} t^{2k}$$

and (all the odd power terms vanish)

$$\frac{(1+t)^{-1/(q-1)}+(1-t)^{-1/(q-1)}}{2} = \sum_{k=0}^{\infty} \frac{\prod_{\ell=0}^{k-1}\left(\frac{1}{q-1}+\ell\right)}{k!} \left(\prod_{\ell=0}^{k-1} \frac{\frac{1}{q-1}+\ell+k}{1+\ell+k}\right) t^{2k}.$$

Note that

$$\prod_{\ell=0}^{k-1} \frac{\frac{1}{q-1}+\ell+k}{1+\ell+k} \geq \frac{1}{C_q^k}.$$

This proves (49) and hence our conclusion. ■

When $\mathcal{R}(f_c)=0$, the estimate in Theorem 14 can be improved (Zhang 2004, Bartlett, Jordan and McAuliffe 2003) to

$$\mathcal{R}(\mathrm{sgn}(f)) \leq \mathcal{E}(f). \tag{50}$$

In fact, $\mathcal{R}(f_c)=0$ implies $|f_\rho(x)|=1$ almost everywhere. This in connection with (48) and (47) gives $\mathcal{E}(f_q|x)=0$ and $\mathcal{E}(f|x)\geq 1$ when $\mathrm{sgn}(f)(x)\neq f_c(x)$. Then (43) can be improved to the form $|f_\rho(x)| \leq \mathcal{E}(f|x)$. Hence (50) follows from (42).

Theorem 14 tells us that the misclassification error can be bounded by the V-risk associated with the loss function V. So our next step is to study the convergence of the q-classifier with respect to the V-risk $\mathcal{E}$.

4. Bounding the Offset

If the offset b is fixed in the scheme (3), the sample error can be bounded by standard argument using some measurements of the capacity of the RKHS. However, the offset is part of the optimization problem and even its boundedness can not be seen from the definition. This makes our setting here essentially different from the standard Tikhonov regularization scheme. The difficulty of bounding the offset b has been realized in the literature (e.g. Steinwart, 2002; Bousquet and Ellisseeff, 2002). In this paper we shall overcome this difficulty by means of special features of the loss function V.

The difficulty raised by the offset can also be seen from the stability analysis (Bousquet and Ellisseeff, 2002). As shown in Bousquet and Ellisseeff (2002), the SVM 1-norm classifier without the offset is uniformly stable, meaning that $\sup_{\mathbf{z}\in Z^m, z_0'\in Z} \|V(y,f_{\mathbf{z}}(x))-V(y,f_{\mathbf{z}'}(x))\|_{L^\infty} \leq \beta_m$ with $\beta_m \to 0$ as $m\to\infty$, and $\mathbf{z}'$ is the same as $\mathbf{z}$ except that one element of $\mathbf{z}$ is replaced by z_0'.

The SVM 1-norm classifier with the offset is not uniformly stable. To see this, we choose $x_0 \in X$ and samples $\mathbf{z} = \{(x_0,y_i)\}_{i=1}^{2n+1}$ with $y_i = 1$ for $i=1,\ldots,n+1$, and $y_i=-1$ for $i=n+2,\ldots,2n+1$.

Take $z_0' = (x_0, -1)$. As x_i are identical, one can see from the definition (9) that $f_{\mathbf{z}}^* = 0$ (since $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) = \mathcal{E}_{\mathbf{z}}(b_{\mathbf{z}} + f_{\mathbf{z}}(x_0))$). It follows that $f_{\mathbf{z}} = 1$ while $f_{\mathbf{z}'} = -1$. Thus, $|f_{\mathbf{z}} - f_{\mathbf{z}'}| = 2$ which does not converge to zero as $m = 2n+1$ tends to infinity. It is unknown whether the q-norm classifier is uniformly stable for $q > 1$.

How to bound the offset is the main goal of this section. In Wu and Zhou (2004) a direct computation is used to realize this point for $q = 1$. Here the index $q > 1$ makes a direct computation very difficult, and we shall use two bounds to overcome this difficulty. By $x \in (X, \rho_X)$ we mean that x lies in the support of the measure ρ_X on X.

Lemma 15 *For any $C > 0, m \in \mathbb{N}$ and $\mathbf{z} \in Z^m$, a minimizer of (9) satisfies*

$$\min_{1 \le i \le m} f_{\mathbf{z}}(x_i) \le 1 \quad \textit{and} \quad \max_{1 \le i \le m} f_{\mathbf{z}}(x_i) \ge -1 \tag{51}$$

and a minimizer of (15) satisfies

$$\inf_{x \in (X, \rho_X)} \widetilde{f}_{K,C}(x) \le 1 \quad \textit{and} \quad \sup_{x \in (X, \rho_X)} \widetilde{f}_{K,C}(x) \ge -1. \tag{52}$$

Proof Suppose a minimizer of (9) $f_{\mathbf{z}}$ satisfies $r := \min_{1 \le i \le m} f_{\mathbf{z}}(x_i) > 1$. Then $f_{\mathbf{z}}(x_i) - (r-1) \ge 1$ for each i. We claim that

$$y_i = 1, \qquad \forall i = 1, \ldots, m.$$

In fact, if the set $I := \{i \in \{1, \ldots, m\} : y_i = -1\}$ is not empty, we have

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}} - (r-1)) = \frac{1}{m} \sum_{i \in I} (1 + f_{\mathbf{z}}(x_i) - (r-1))^q < \frac{1}{m} \sum_{i \in I} (1 + f_{\mathbf{z}}(x_i))^q = \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}),$$

which is a contradiction to the definition of $f_{\mathbf{z}}$. Hence our claim is verified. From the claim we see that $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}} - (r-1)) = 0 = \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})$. This tells us that $\widetilde{f}_{\mathbf{z}} := f_{\mathbf{z}} - (r-1)$ is a minimizer of (9) satisfying the first inequality, hence both inequalities of (51).

In the same way, if a minimizer of (9) $f_{\mathbf{z}}$ satisfies $r := \max_{1 \le i \le m} f_{\mathbf{z}}(x_i) < -1$. Then we can see that $y_i = -1$ for each i. Hence $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}} - r - 1) = \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})$ and $\widetilde{f}_{\mathbf{z}} := f_{\mathbf{z}} - r - 1$ is a minimizer of (9) satisfying the second inequality and hence both inequalities of (51).

Therefore, we can always find a minimizer of (9) satisfying (51).

We prove the second statement in the same way. Suppose $r := \inf_{x \in (X, \rho_X)} \widetilde{f}_{K,C}(x) > 1$ for a minimizer $\widetilde{f}_{K,C}$ of (15). Then $\widetilde{f}_{K,C}(x) - (r-1) \ge 1$ for almost every $x \in (X, \rho_X)$. Hence

$$\mathcal{E}\left(\widetilde{f}_{K,C} - (r-1)\right) = \int_X \left(1 + \widetilde{f}_{K,C}(x) - (r-1)\right)^q P(\mathcal{Y} = -1|x) d\rho_X \le \mathcal{E}(\widetilde{f}_{K,C}).$$

As $f_{K,C}$ is a minimizer of (15), the above equality must hold. It follows that $P(\mathcal{Y} = -1|x) = 0$ for almost every $x \in (X, \rho_X)$. Hence $\widetilde{F}_{K,C} := \widetilde{f}_{K,C} - (r-1)$ is a minimizer of (15) satisfying the first inequality and thereby both inequalities of (52).

Similarly, when $r := \sup_{x \in (X, \rho_X)} \widetilde{f}_{K,C}(x) < -1$ for a minimizer $\widetilde{f}_{K,C}$ of (15). Then $P(\mathcal{Y} = 1|x) = 0$ for almost every $x \in (X, \rho_X)$. Hence $\widetilde{F}_{K,C} := f_{K,C} - r - 1$ is a minimizer of (15) satisfying the second inequality and thereby both inequalities of (52).

Thus, (52) can always be realized by a minimizer of (15). ∎

In what follows we always choose $f_{\mathbf{z}}$ and $\widetilde{f}_{K,C}$ to satisfy (51) and (52), respectively.

Lemma 15 yields bounds for the $\mathcal{H}_K$-norm and offset for $f_{\mathbf{z}}$ and $\widetilde{f}_{K,C}$. Denote $b_{\widetilde{f}_{K,C}}$ as $\widetilde{b}_{K,C}$.

Lemma 16 *For any $C>0, m \in \mathbb{N}, f_{K,C} \in \overline{\mathcal{H}}_K$, and $\mathbf{z} \in Z^m$, there hold*

(a) $\|\widetilde{f}^*_{K,C}\|_K \le \sqrt{2C\widetilde{\mathcal{D}}(C)} \le \sqrt{2C}, \quad |\widetilde{b}_{K,C}| \le 1 + \|\widetilde{f}^*_{K,C}\|_\infty$.

(b) $\|\widetilde{f}_{K,C}\|_\infty \le 1 + 2\kappa\sqrt{2C\widetilde{\mathcal{D}}(C)} \le 1 + 2\kappa\sqrt{2C}, \quad \mathcal{E}(\widetilde{f}_{K,C}) \le \mathcal{E}(f_q) + \widetilde{\mathcal{D}}(C) \le 1$.

(c) $|b_{\mathbf{z}}| \le 1 + \kappa\|f^*_{\mathbf{z}}\|_K$.

Proof By the definition (15), we see from the choice $f = 0+0$ that

$$\widetilde{\mathcal{D}}(C) = \mathcal{E}(\widetilde{f}_{K,C}) - \mathcal{E}(f_q) + \frac{1}{2C}\|\widetilde{f}^*_{K,C}\|_K^2 \le 1 - \mathcal{E}(f_q). \tag{53}$$

Then the first inequality in (a) follows.

Note that (52) gives

$$-1 \le \sup_{x \in (X,\rho_X)} \widetilde{f}_{K,C} \le \widetilde{b}_{K,C} + \|\widetilde{f}^*_{K,C}\|_\infty$$

and

$$\widetilde{b}_{K,C} - \|\widetilde{f}^*_{K,C}\|_\infty \le \inf_{x \in (X,\rho_X)} \widetilde{f}_{K,C} \le 1.$$

Thus, $|\widetilde{b}_{K,C}| \le 1 + \|\widetilde{f}^*_{K,C}\|_\infty$. This proves the second inequality in (a).

Since the first inequality in (a) and (2) lead to $\|\widetilde{f}^*_{K,C}\|_\infty \le \kappa\sqrt{2C\widetilde{\mathcal{D}}(C)}$, we obtain $\|\widetilde{f}_{K,C}\|_\infty \le 1 + 2\|\widetilde{f}^*_{K,C}\|_\infty \le 1 + 2\kappa\sqrt{2C\widetilde{\mathcal{D}}(C)}$. Hence the first inequality in (b) holds.

The second inequality in (b) is an easy consequence of (53).

The inequality in (c) follows from (51) in the same way as the proof of the second inequality in (a). ∎

5. Convergence of the q-Norm Soft Margin Classifier

In this section, we apply the ERM technique to analyze the convergence of the q-classifier for $q > 1$. The situation here is more complicated than that for $q = 1$. We need the following lemma concerning $q > 1$.

Lemma 17 *For $q > 1$, there holds*

$$\left|(x)_+^q - (y)_+^q\right| \le q\,(\max\{x,y\})_+^{q-1}\,|x-y|, \qquad \forall x, y \in \mathbb{R}.$$

If $y \in [-1,1]$, then

$$\left|(1-x)_+^q - (1-y)_+^q\right| \le q4^{q-1}|x-y| + q2^{q-1}|x-y|^q, \qquad \forall x \in \mathbb{R}.$$

Proof We only need to prove the first inequality for $x > y$. This is trivial:

$$(x)_+^q - (y)_+^q = \int_y^x q(u)_+^{q-1} du \leq q(x)_+^{q-1}(x-y).$$

If $y \in [-1,1]$, then $1-y \geq 0$ and the first inequality yields

$$\left|(1-x)_+^q - (1-y)_+^q\right| \leq q\left(\max\{1-x, 1-y\}\right)_+^{q-1} |x-y|. \tag{54}$$

When $x \geq 2y-1$, we have

$$\left|(1-x)_+^q - (1-y)_+^q\right| \leq q2^{q-1}(1-y)^{q-1}|x-y| \leq q4^{q-1}|x-y|.$$

When $x < 2y-1$, we have $1-x < 2(y-x)$ and $x < 1$. This in combination with $\max\{1-x, 1-y\} \leq \max\{1-x, 2\}$ and (54) implies

$$\left|(1-x)_+^q - (1-y)_+^q\right| \leq q\{(1-x)^{q-1} + 2^{q-1}\}|x-y| \leq q2^{q-1}|x-y|^q + q2^{q-1}|x-y|.$$

This proves Lemma 17. ■

The second part of Lemma 17 will be used in Section 6. The first part can be used to verify Proposition 4.

Proof of Proposition 4. It is trivial that $\mathcal{D}(C) \geq \widetilde{\mathcal{D}}(C)$. To show the second inequality, apply the first inequality of Lemma 17 to the two numbers $1 - y\widetilde{f}_{K,C}(x)$ and $1 - y\widetilde{b}_{K,C}$. We see that

$$(1 - y\widetilde{f}_{K,C}(x))_+^q \geq (1 - y\widetilde{b}_{K,C})_+^q - q(1 + |\widetilde{f}_{K,C}(x)| + |\widetilde{b}_{K,C}|)^{q-1}|\widetilde{f}^*_{K,C}(x)|.$$

Notice that $|\widetilde{f}^*_{K,C}(x)| \leq \kappa\|\widetilde{f}^*_{K,C}\|_K$ and by Lemma 16, $|\widetilde{b}_{K,C}| \leq 1 + \kappa\|\widetilde{f}^*_{K,C}\|_K$. Hence

$$\mathcal{E}(\widetilde{f}_{K,C}) \geq \mathcal{E}(\widetilde{b}_{K,C}) - \kappa q 2^{q-1}(1 + \kappa\|\widetilde{f}^*_{K,C}\|_K)^{q-1}\|\widetilde{f}^*_{K,C}\|_K.$$

It follows that when $\|\widetilde{f}^*_{K,C}\|_K \leq \widetilde{\kappa}$, we have

$$\mathcal{E}(\widetilde{f}_{K,C}) - \mathcal{E}(f_q) \geq \mathcal{E}_0 - \kappa q 2^{q-1}(1 + \kappa\widetilde{\kappa})^{q-1}\widetilde{\kappa}.$$

Since $\mathcal{E}_0 \leq 1$, the definition (18) of $\widetilde{\kappa}$ yields $\widetilde{\kappa} \leq 1/(1+\kappa) \leq \min\{1, 1/\kappa\}$. Hence

$$\widetilde{\mathcal{D}}(C) \geq \mathcal{E}(\widetilde{f}_{K,C}) - \mathcal{E}(f_q) \geq \mathcal{E}_0 - \kappa q 4^{q-1}\widetilde{\kappa} = \widetilde{\kappa} \geq \widetilde{\kappa}^2.$$

As $C \geq 1/2$, we conclude (17) in this case.

When $\|\widetilde{f}^*_{K,C}\|_K > \widetilde{\kappa}$, we also have

$$\widetilde{\mathcal{D}}(C) \geq \frac{1}{2C}\|\widetilde{f}^*_{K,C}\|_K^2 \geq \frac{\widetilde{\kappa}^2}{2C}.$$

Thus in both case we have verified (17).

Note that $\widetilde{\kappa} = 0$ if and only if $\mathcal{E}_0 = 0$. This means for some $b_0' \in [-1,1]$, $f_q(x) = b_0'$ in probability. By the definition of f_q, the last assertion of Proposition 4 follows. ■

The loss function V is not Lipschitz, but Lemma 17 enables us to derive a bound for $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}}))$ with confidence, as done for Lipschitz loss functions in Mukherjee, Rifkin and Poggio (2002). Since the function $\pi(f_{\mathbf{z}})$ changes and lies in a set of functions, we shall compare the error with the empirical error for functions from a set

$$\mathcal{F} := \{\pi(f) : f \in B_R + [-B, B]\}. \tag{55}$$

Here $B_R = \{f^* \in \mathcal{H}_K : \|f^*\|_K \le R\}$ and the constant B is a bound for the offset.

The following probability inequality was motivated by sample error estimates for the square loss (Barron 1990, Bartlett 1998, Cucker and Smale 2001, Lee, Bartlett and Williamson 1998) and will be used in our estimates.

Lemma 18 *Suppose a random variable ξ satisfies $0 \le \xi \le M$, and $\mathbf{z} = (z_i)_{i=1}^m$ are independent samples. Let $\mu = E(\xi)$. Then for every $\varepsilon > 0$ and $0 < \alpha \le 1$, there holds*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\mu - \frac{1}{m}\sum_{i=1}^m \xi(z_i)}{\sqrt{\mu+\varepsilon}} \ge \alpha\sqrt{\varepsilon}\right\} \le \exp\left\{-\frac{3\alpha^2 m\varepsilon}{8M}\right\}.$$

Proof As ξ satisfies $|\xi - \mu| \le M$, the one-side Bernstein inequality tells us that

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\mu - \frac{1}{m}\sum_{i=1}^m \xi(z_i)}{\sqrt{\mu+\varepsilon}} \ge \alpha\sqrt{\varepsilon}\right\} \le \exp\left\{-\frac{\alpha^2 m(\mu+\varepsilon)\varepsilon}{2\left(\sigma^2 + \frac{1}{3}M\alpha\sqrt{\mu+\varepsilon}\sqrt{\varepsilon}\right)}\right\}.$$

Here $\sigma^2 \le E(\xi^2) \le ME(\xi) = M\mu$ since $0 \le \xi \le M$. Then we find that

$$\sigma^2 + \frac{1}{3}M\alpha\sqrt{\mu+\varepsilon}\sqrt{\varepsilon} \le M\mu + \frac{1}{3}M(\mu+\varepsilon) \le \frac{4}{3}M(\mu+\varepsilon).$$

This yields the desired inequality. ■

Now we can turn to the error bound involving a function set. Lemma 17 yields the following bounds concerning the loss function V:

$$|\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(g)| \le q \max\left\{(1+\|f\|_\infty)^{q-1}, (1+\|g\|_\infty)^{q-1}\right\}\|f-g\|_\infty \tag{56}$$

and

$$|\mathcal{E}(f) - \mathcal{E}(g)| \le q \max\left\{(1+\|f\|_\infty)^{q-1}, (1+\|g\|_\infty)^{q-1}\right\}\|f-g\|_\infty. \tag{57}$$

Lemma 19 *Let $\mathcal{F}$ be a subset of $C(X)$ such that $\|f\|_\infty \le 1$ for each $f \in \mathcal{F}$. Then for every $\varepsilon > 0$ and $0 < \alpha \le 1$, we have*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{F}} \frac{\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)}{\sqrt{\mathcal{E}(f)+\varepsilon}} \ge 4\alpha\sqrt{\varepsilon}\right\} \le \mathcal{N}\left(\mathcal{F}, \frac{\alpha\varepsilon}{q2^{q-1}}\right)\exp\left\{-\frac{3\alpha^2 m\varepsilon}{2^{q+3}}\right\}.$$

Proof Let $\{f_j\}_{j=1}^J \subset \mathcal{F}$ with $J = \mathcal{N}\left(\mathcal{F}, \frac{\alpha\varepsilon}{q2^{q-1}}\right)$ such that $\mathcal{F}$ is covered by balls centered at f_j with radius $\frac{\alpha\varepsilon}{q2^{q-1}}$. Note that $\xi = V(y, f(x))$ satisfies $0 \le \xi \le (1+\|f\|_\infty)^q \le 2^q$ for $f \in \mathcal{F}$. Then for each j, Lemma 18 tells

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\mathcal{E}(f_j)-\mathcal{E}_{\mathbf{z}}(f_j)}{\sqrt{\mathcal{E}(f_j)+\varepsilon}} \ge \alpha\sqrt{\varepsilon}\right\} \le \exp\left\{-\frac{3\alpha^2 m\varepsilon}{8\cdot 2^q}\right\}.$$

For each $f \in \mathcal{F}$, there is some j such that $\|f-f_j\|_\infty \le \frac{\alpha\varepsilon}{q2^{q-1}}$. This in connection with (56) and (57) tells us that $|\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_j)|$ and $|\mathcal{E}(f)-\mathcal{E}(f_j)|$ are both bounded by $\alpha\varepsilon$. Hence

$$\frac{|\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_j)|}{\sqrt{\mathcal{E}(f)+\varepsilon}} \le \alpha\sqrt{\varepsilon} \qquad \text{and} \qquad \frac{|\mathcal{E}(f)-\mathcal{E}(f_j)|}{\sqrt{\mathcal{E}(f)+\varepsilon}} \le \alpha\sqrt{\varepsilon}.$$

The latter implies that $\sqrt{\mathcal{E}(f_j)+\varepsilon} \le 2\sqrt{\mathcal{E}(f)+\varepsilon}$. Therefore,

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{F}}\frac{\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)}{\sqrt{\mathcal{E}(f)+\varepsilon}} \ge 4\alpha\sqrt{\varepsilon}\right\} \le \sum_{j=1}^J \mathrm{Prob}\left\{\frac{\mathcal{E}(f_j)-\mathcal{E}_{\mathbf{z}}(f_j)}{\sqrt{\mathcal{E}(f_j)+\varepsilon}} \ge \alpha\sqrt{\varepsilon}\right\}$$

which is bounded by $J\cdot\exp\left\{-\frac{3\alpha^2 m\varepsilon}{2^{q+3}}\right\}$. ∎

Take $\mathcal{F}$ to be the set (55). The following covering number estimate will be used.

Lemma 20 *Let $\mathcal{F}$ be given by (55) with $R \ge \widetilde{\kappa}$ and $B = 1+\kappa R$. Its covering number in $C(X)$ can be bounded as follows:*

$$\mathcal{N}(\mathcal{F},\eta) \le \mathcal{N}\left(\frac{\eta}{2R}\right), \qquad \forall \eta > 0.$$

Proof It follows from the fact $\|\pi(f)-\pi(g)\|_\infty \le \|f-g\|_\infty$ that

$$\mathcal{N}(\mathcal{F},\eta) \le \mathcal{N}(B_R + [-B,B],\eta).$$

The latter is bounded by $\{(\kappa+\frac{1}{\kappa})\frac{2R}{\eta}+1\}\mathcal{N}\left(B_R,\frac{\eta}{2}\right)$ since $\frac{2B}{\eta} \le (\kappa+\frac{1}{\kappa})\frac{2R}{\eta}$ and

$$\|(f^*+b_f)-(g^*+b_g)\|_\infty \le \|f^*-g^*\|_\infty + |b_f-b_g|.$$

Note that an $\frac{\eta}{2R}$-covering of $\mathcal{B}$ is the same as an $\frac{\eta}{2}$-covering of B_R. Then our conclusion follows from Definition 6. ∎

We are in a position to state our main result on the error analysis. Recall θ_M in (25).

Theorem 21 *Let $f_{K,C} \in \overline{\mathcal{H}}_K$, $M \ge \|V(y, f_{K,C}(x))\|_\infty$, and $0 < \beta \le 1$. For every $\varepsilon > 0$, we have*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q) \le (1+\beta)\,(\varepsilon + \mathcal{D}(C))$$

with confidence at least $1-F(\varepsilon)$ where $F: \mathbb{R}_+ \to \mathbb{R}_+$ is defined by

$$F(\varepsilon) := \exp\left\{-\frac{m\varepsilon^2}{2M(\Delta+\varepsilon)}\right\} + \mathcal{N}\left(\frac{\beta(\varepsilon+\mathcal{D}(C))^{3/2}}{q2^{q+3}\Sigma}\right)\exp\left\{-\frac{3m\beta^2\,(\mathcal{D}(C)+\varepsilon)^2}{2^{q+9}\,(\Delta+\varepsilon)}\right\} \qquad (58)$$

with $\Delta := \mathcal{D}(C) + \mathcal{E}(f_q)$ and $\Sigma := \sqrt{2C(\Delta+\varepsilon)(\Delta+\theta_M\varepsilon)}$.

Proof Let $\varepsilon > 0$. We prove our conclusion in three steps.

Step 1: Estimate $\mathcal{E}_{\mathbf{z}}(f_{K,C}) - \mathcal{E}(f_{K,C})$.

Consider the random variable $\xi = V(y, f_{K,C}(x))$ with $0 \leq \xi \leq M$. If $M > 0$, since $\sigma^2(\xi) \leq M\mathcal{E}(f_{K,C}) \leq M(\mathcal{D}(C) + \mathcal{E}(f_q))$, by the one-side Bernstein inequality we obtain $\mathcal{E}_{\mathbf{z}}(f_{K,C}) - \mathcal{E}(f_{K,C}) \leq \varepsilon$ with confidence at least

$$1 - \exp\left\{-\frac{m\varepsilon^2}{2(\sigma^2(\xi) + \frac{1}{3}M\varepsilon)}\right\} \geq 1 - \exp\left\{-\frac{m\varepsilon^2}{2M(\Delta+\varepsilon)}\right\}.$$

If $M = 0$, then $\xi = 0$ almost everywhere. Hence $\mathcal{E}_{\mathbf{z}}(f_{K,C}) - \mathcal{E}(f_{K,C}) = 0$ with probability 1. Thus, in both cases, there exists $U_1 \in Z^m$ with measure at least $1 - \exp\left\{-\frac{m\varepsilon^2}{2M(\varepsilon+\Delta)}\right\}$ such that $\mathcal{E}_{\mathbf{z}}(f_{K,C}) - \mathcal{E}(f_{K,C}) \leq \theta_M \varepsilon$ whenever $\mathbf{z} \in U_1$.

Step 2: Estimate $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}}))$.

Let $\mathcal{F}$ be given by (55) with $R = \sqrt{2C(\Delta + \theta_M\varepsilon)}$ and $B = 1 + \kappa R$. By Proposition 4, $R \geq \sqrt{2C\mathcal{D}(C)} \geq \widetilde{\kappa}$. Applying Lemma 19 to $\mathcal{F}$ with $\tilde{\varepsilon} := \mathcal{D}(C) + \varepsilon > 0$ and $\alpha = \frac{\beta}{8}\sqrt{\tilde{\varepsilon}/(\tilde{\varepsilon} + \mathcal{E}(f_q))} \in (0, 1/8]$, we have

$$\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f) \leq 4\alpha\sqrt{\tilde{\varepsilon}}\sqrt{\mathcal{E}(f) - \mathcal{E}(f_q) + \tilde{\varepsilon} + \mathcal{E}(f_q)}, \qquad \forall f \in \mathcal{F} \tag{59}$$

for $\mathbf{z} \in U_2$ where U_2 is a subset of Z^m with measure at least

$$1 - \mathcal{N}\Big(\mathcal{F}, \frac{\alpha\tilde{\varepsilon}}{q2^{q-1}}\Big)\exp\left\{-\frac{3m\alpha^2\tilde{\varepsilon}}{2^{q+3}}\right\} \geq 1 - \mathcal{N}\Big(\frac{\beta(\varepsilon + \mathcal{D}(C))^{3/2}}{q2^{q+3}\Sigma}\Big)\exp\left\{-\frac{3m\beta^2(\mathcal{D}(C)+\varepsilon)^2}{2^{q+9}(\Delta+\varepsilon)}\right\}.$$

In the above inequality we have used Lemma 20 to bound the covering number.

For $\mathbf{z} \in U_1 \cap U_2$, we have

$$\frac{1}{2C}\|f_{\mathbf{z}}^*\|_K^2 \leq \mathcal{E}_{\mathbf{z}}(f_{K,C}) + \frac{1}{2C}\|f_{K,C}^*\|_K^2 \leq \mathcal{E}(f_{K,C}) + \theta_M\varepsilon + \frac{1}{2C}\|f_{K,C}^*\|_K^2$$

which equals to $\mathcal{D}(C) + \theta_M\varepsilon + \mathcal{E}(f_q) = \Delta + \theta_M\varepsilon$. It follows that $\|f_{\mathbf{z}}^*\|_K \leq R$. By Lemma 16, $|b_{\mathbf{z}}| \leq B$. This means, $\pi(f_{\mathbf{z}}) \in \mathcal{F}$ and (59) is valid for $f = \pi(f_{\mathbf{z}})$.

Step 3: Bound $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q)$ using (14).

Let α and $\tilde{\varepsilon}$ be the same as in Step 2. For $\mathbf{z} \in U_1 \bigcap U_2$, both $\mathcal{E}_{\mathbf{z}}(f_{K,C}) - \mathcal{E}(f_{K,C}) \leq \theta_M\varepsilon \leq \varepsilon$ and (59) with $f = \pi(f_{\mathbf{z}})$ hold true. Then (14) tells us that

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q) \leq 4\alpha\sqrt{\tilde{\varepsilon}}\sqrt{(\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q)) + \tilde{\varepsilon} + \mathcal{E}(f_q)} + \tilde{\varepsilon}.$$

Denote $r := \mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q) + \tilde{\varepsilon} + \mathcal{E}(f_q) > 0$, we see that

$$r \leq \tilde{\varepsilon} + \mathcal{E}(f_q) + 4\alpha\sqrt{\tilde{\varepsilon}}\sqrt{r} + \tilde{\varepsilon}.$$

It follows that

$$\sqrt{r} \leq 2\alpha\sqrt{\tilde{\varepsilon}} + \sqrt{4\alpha^2\tilde{\varepsilon} + 2\tilde{\varepsilon} + \mathcal{E}(f_q)}.$$

Hence

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q) = r - (\tilde{\varepsilon} + \mathcal{E}(f_q)) \leq \tilde{\varepsilon} + 8\alpha^2\tilde{\varepsilon} + 4\alpha\tilde{\varepsilon}\sqrt{4\alpha^2 + 2 + \mathcal{E}(f_q)/\tilde{\varepsilon}}.$$

Putting the choice of α into above, we find that

$$\mathcal{E}(\pi(f_{\mathbf{z}}))-\mathcal{E}(f_q)\leq\tilde{\varepsilon}+\frac{\beta^2\tilde{\varepsilon}^2}{8\,(\tilde{\varepsilon}+\mathcal{E}(f_q))}+\frac{\beta\tilde{\varepsilon}}{2}\sqrt{\frac{\tilde{\varepsilon}}{\tilde{\varepsilon}+\mathcal{E}(f_q)}}\sqrt{\frac{\beta^2\tilde{\varepsilon}}{16\,(\tilde{\varepsilon}+\mathcal{E}(f_q))}+2+\frac{\mathcal{E}(f_q)}{\tilde{\varepsilon}}}$$

which is bounded by $(1+\beta)\tilde{\varepsilon}=(1+\beta)(\varepsilon+\mathcal{D}(C))$.

Finally, noting that the measure of $U_1\bigcap U_2$ is at least $1-F(\varepsilon)$, the proof is finished. ■

Observe that the function F is strictly decreasing and $F(0)>1$, $\lim_{\varepsilon\to\infty}F(\varepsilon)=0$. Hence for every $0<\delta<1$ there exists a unique number $\varepsilon>0$ satisfying $F(\varepsilon)=\delta$. Also, for a fixed $\varepsilon>0$, $F(\varepsilon)<\delta$ for sufficiently large m, since $F(\varepsilon)$ can be written as a^m+cb^m with $0<a,b<1$. Therefore, Theorem 21 can be restated in the following form.

Corollary 22 *Let F be given in Theorem 21. For every $0<\delta<1$, define $\varepsilon(\delta,m,C,\beta)>0$ to be the unique solution to the equation $F(\varepsilon)=\delta$. Then with confidence at least $1-\delta$, (20) holds. Moreover,* $\lim_{m\to\infty}\varepsilon(\delta,m,C,\beta)=0$.

Now we can prove the statements we made in Section 2 for the deterministic case.

Proof of Proposition 8. Take $\beta=1$. Since $\mathcal{R}(f_c)=0$, we know that $f_q=f_c$, $\mathcal{E}(f_q)=0$ and $\Delta=\mathcal{D}(C)$. Then $\Sigma\leq\sqrt{2C(\mathcal{D}(C)+\theta_M\varepsilon)}\sqrt{\mathcal{D}(C)+\varepsilon}$ and

$$\mathcal{N}\left(\frac{\beta(\varepsilon+\mathcal{D}(C))^{3/2}}{q2^{q+3}\Sigma}\right)\leq\mathcal{N}\left(\frac{\varepsilon}{q2^{q+3}\mathcal{M}}\right).$$

The function value $F(\varepsilon)$ can be bounded by

$$\exp\left\{-\frac{m\varepsilon^2}{2M(\mathcal{D}(C)+\varepsilon)}\right\}+\mathcal{N}\left(\frac{\varepsilon}{q2^{q+3}\mathcal{M}}\right)\exp\left\{-\frac{3m\varepsilon}{2^{q+9}}\right\}.$$

The first term above is bounded by $\delta/2$ when $\varepsilon\geq\frac{4M\log(2/\delta)}{m}+\mathcal{D}(C)$. The second term is at most $\delta/2$ if $\varepsilon\geq\varepsilon^*$. Therefore

$$\varepsilon(\delta,m,C,\beta)\leq\max\left\{\frac{4M\log(2/\delta)}{m}+\mathcal{D}(C),\varepsilon^*\right\},$$

and (26) follows from Theorem 21.

(a) When (23) holds, noting that the left side of (27) is strictly decreasing, it is easy to check that

$$\varepsilon^*\leq 2^{q+9}\left\{1+c((q+4)\log 8)^s\right\}\frac{(\log m+\log(C\mathcal{D}(C))+\log(C\theta_M))^s+\log(2/\delta)}{m}.$$

This yields the stated bound for ε^* in the case (a).

(b) If (24) is true, then the function defined by

$$\widetilde{F}(\varepsilon):=c\frac{(q2^{q+4})^s\{(2C\mathcal{D}(C))^{s/2}+(2C\varepsilon)^{s/2}\}}{\varepsilon^s}-\frac{3m\varepsilon}{2^{q+9}}$$

satisfies $\widetilde{F}(\varepsilon^*) \geq \log(\delta/2)$. Since $\widetilde{F}$ is a decreasing function, $\varepsilon^* \leq \widetilde{\varepsilon}^*$ whenever $\widetilde{F}(\widetilde{\varepsilon}^*) \leq \log(\delta/2)$.

If $\widetilde{\varepsilon}^* \geq rm^{-1/(s+1)}(2C\mathcal{D}(C))^{s/(2s+2)} \log\frac{2}{\delta}$ with

$$r \geq 2^{q+8}\widetilde{\kappa}^{-s/(s+1)} + q2^{q+10}c^{1/(s+1)}/\log 2, \tag{60}$$

then

$$c(q2^{q+4})^s \frac{(2C\mathcal{D}(C))^{s/2}}{(\widetilde{\varepsilon}^*)^s} - \frac{m\widetilde{\varepsilon}^*}{2^{q+8}} \leq \frac{rm^{s/(s+1)}}{2^{q+8}}(2C\mathcal{D}(C))^{\frac{s}{2s+2}} \log\frac{2}{\delta}\left\{\frac{c(q2^{q+4})^s 2^{q+8}}{(r\log(2/\delta))^{s+1}} - 1\right\}.$$

Since $r \geq q2^{q+10}c^{1/(s+1)}/\log 2$, according to Proposition 4, this can be bounded by

$$\frac{r}{2^{q+8}}m^{s/(s+1)}(2C\mathcal{D}(C))^{s/(2s+2)} \log\frac{2}{\delta}\left\{-\frac{1}{2}\right\} \leq \frac{r}{2^{q+9}}\widetilde{\kappa}^{s/(s+1)} \log\frac{\delta}{2} \leq \frac{1}{2}\log\frac{\delta}{2}.$$

Here we have used the condition $r \geq 2^{q+8}\widetilde{\kappa}^{-s/(s+1)}$.

In the same way, if $\widetilde{\varepsilon}^* \geq rm^{-2/(s+2)}(2C)^{s/(s+2)} \log\frac{2}{\delta}$ with

$$r \geq 2^{q+9} + q^2 4^{q+5} c^{2/(s+2)}/\log 2, \tag{61}$$

we have for $C \geq 1/2$,

$$c(q2^{q+4})^s \left(\frac{2C}{\widetilde{\varepsilon}^*}\right)^{s/2} - \frac{m\widetilde{\varepsilon}^*}{2^{q+9}} \leq \frac{1}{2}\log\frac{\delta}{2}.$$

Combining the above two bounds, we obtain the desired estimate (28) with $\widetilde{c}$ determined by the two conditions (60) and (61). The proof of Proposition 8 is complete. ∎

In the general case, the following bounds hold.

Corollary 23 *For every* $0 < \delta \leq 1$*, with confidence at least* $1-\delta$ *there holds*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_q) \leq 2\max\left\{\frac{2^{q+1}\left(1+\kappa\sqrt{2C\widetilde{\mathcal{D}}(C)}\right)^{q/2}\sqrt{\log(2/\delta)}}{\sqrt{m}}, \varepsilon^*\right\} + 2\widetilde{\mathcal{D}}(C),$$

where ε^* *is the solution to the equation*

$$\log\mathcal{N}\left(\frac{\varepsilon^{3/2}}{q4^{q+2}\sqrt{2C}}\right) - \frac{3m\varepsilon^2}{4^{q+5}} = \log\frac{\delta}{2}. \tag{62}$$

Proof Take $\beta = 1$ and $f_{K,C} = \widetilde{f}_{K,C}$. By Lemma 16, $\|\widetilde{f}_{K,C}\|_\infty \leq 1 + 2\kappa\sqrt{2C\widetilde{\mathcal{D}}(C)}$ and we can take $M = 2^q(1+\kappa\sqrt{2C\widetilde{\mathcal{D}}(C)})^q$. Also, $\Delta = \widetilde{\mathcal{D}}(C) + \mathcal{E}(f_q) \leq 1$. It follows that $\Sigma \leq \sqrt{2C}(\Delta+\varepsilon) \leq \sqrt{2C}(1+\varepsilon)$.

Since $\mathcal{E}(\pi(f_{\mathbf{z}})) \leq 2^q$, we only need to consider the range $\varepsilon \leq 2^q$ to bound $\varepsilon(\delta, m, C, \beta)$. In this range,

$$\mathcal{N}\left(\frac{\beta(\varepsilon+\widetilde{\mathcal{D}}(C))^{3/2}}{q2^{q+3}\Sigma}\right) \leq \mathcal{N}\left(\frac{\varepsilon^{3/2}}{q4^{q+2}\sqrt{2C}}\right).$$

Then $F(\varepsilon)$ can be bounded by

$$\exp\left\{-\frac{m\varepsilon^2}{4^{q+1}(1+\kappa\sqrt{2C\widetilde{\mathcal{D}}(C)})^q}\right\}+\mathcal{N}\left(\frac{\varepsilon^{3/2}}{q4^{q+2}\sqrt{2C}}\right)\exp\left\{-\frac{3m\varepsilon^2}{4^{q+5}}\right\}.$$

Thus $F(\varepsilon)\le\delta$ if

$$\varepsilon\ge\max\left\{\frac{2^{q+1}(1+\kappa\sqrt{2C\widetilde{\mathcal{D}}(C)})^{q/2}\sqrt{\log(2/\delta)}}{\sqrt{m}},\varepsilon^*\right\}$$

where ε^* is the solution to the equation (62). This together with Theorem 21 yields the desired estimate. ■

The bound for the sample error derived in Corollary 23 may be further improved by the well developed empirical process techniques in the literature. We shall discuss this elsewhere.

The total error (14) consists of two parts. We shall not discuss the possibility of further improving the sample error bound here, because it is of the same importance to understand the regularization error. This becomes more important when not much estimate is available for the regularization error. In the previous sections, we could compute $\mathcal{D}(C)$ explicitly only for special cases. Most of the time, ρ is not strictly separable, even not weakly separable. Hence it is desirable to estimate $\mathcal{D}(C)$ explicitly for general distributions. In the following we shall choose $f_{K,C}=\widetilde{f}_{K,C}$ and estimate $\widetilde{\mathcal{D}}(C)$.

6. Error Analysis by Approximation in L^q Spaces

The main result on the convergence analysis given in Section 5 enables us to have some nice observations. These follow from facts on approximation in L^q spaces.

Lemma 24 *If $1<q\le 2$, then*

$$(1+u)_+^q\le 1+|u|^q+qu,\qquad \forall u\in\mathbb{R}.$$

Proof Set the continuous function $f(u):=1+|u|^q+qu-(1+u)_+^q$. Then $f(0)=0$.

Since $0<q-1\le 1$, for $u>0$ we have

$$f'(u)=q\left(1+u^{q-1}-(1+u)^{q-1}\right)\ge 0.$$

Hence $f(u)\ge 0$ for $u\ge 0$.

For $-1<u<0$, we see that

$$f'(u)=q\left(1-(-u)^{q-1}-(1+u)^{q-1}\right)=q\left(1-|u|^{q-1}-(1-|u|)^{q-1}\right)\le 0.$$

Hence $f(u)\ge 0$ for $-1<u<0$.

Finally, when $u<-1$, there holds

$$f'(u)=q\left(1-(-u)^{q-1}\right)\le 0.$$

Therefore, we also have $f(u)\ge f(-1)\ge 0$ for $u\le -1$.

Thus $f(u)\ge 0$ on the whole real line and Lemma 24 is proved. ■

Theorem 25 *Let $f : X \to \mathbb{R}$ be measurable. Then*

$$\mathcal{E}(f) - \mathcal{E}(f_q) \le \begin{cases} \|f - f_q\|^q_{L^q_{\rho_X}}, & \text{if } 1 < q \le 2, \\ q2^{q-1}\|f - f_q\|_{L^q_{\rho_X}} \left(2^{q-1} + \|f - f_q\|^{q-1}_{L^q_{\rho_X}}\right), & \text{if } q > 2. \end{cases}$$

Proof Since $|f_q(x)| \le 1$, by the second inequality of Lemma 17, for each $x \in X$ we have

$$\begin{aligned} \mathcal{E}(f|x) - \mathcal{E}(f_q|x) &= \int_Y (1 - yf(x))^q_+ - (1 - yf_q(x))^q_+ d\rho(y|x) \\ &\le q4^{q-1}|f(x) - f_q(x)| + q2^{q-1}|f(x) - f_q(x)|^q. \end{aligned}$$

It follows that

$$\mathcal{E}(f) - \mathcal{E}(f_q) = \int_X \mathcal{E}(f|x) - \mathcal{E}(f_q|x) d\rho_X \le q4^{q-1}\|f - f_q\|_{L^1_{\rho_X}} + q2^{q-1}\|f - f_q\|^q_{L^q_{\rho_X}}.$$

Then the inequality for the case $q > 2$ follows from the Hölder inequality.

Turn to the case $1 < q \le 2$. It is sufficient to show that for each $x \in X$,

$$\mathcal{E}(f|x) - \mathcal{E}(f_q|x) \le |f(x) - f_q(x)|^q. \tag{63}$$

The definition (10) of f_q tells us that

$$(1 + f_q(x))^{q-1} = \frac{2^{q-1}(1 + f_\rho(x))}{\{(1 + f_\rho(x))^{1/(q-1)} + (1 - f_\rho(x))^{1/(q-1)}\}^{q-1}}$$

and

$$(1 - f_q(x))^{q-1} = \frac{2^{q-1}(1 - f_\rho(x))}{\{(1 + f_\rho(x))^{1/(q-1)} + (1 - f_\rho(x))^{1/(q-1)}\}^{q-1}}.$$

These expressions in connection with (45) imply

$$\mathcal{E}(f|x) = \frac{(1 + f(x))^q_+(1 - f_q(x))^{q-1}}{(1 + f_q(x))^{q-1} + (1 - f_q(x))^{q-1}} + \frac{(1 - f(x))^q_+(1 + f_q(x))^{q-1}}{(1 + f_q(x))^{q-1} + (1 - f_q(x))^{q-1}}$$

and together with (48)

$$\mathcal{E}(f_q|x) = \frac{2(1 - f_q(x))^{q-1}(1 + f_q(x))^{q-1}}{(1 + f_q(x))^{q-1} + (1 - f_q(x))^{q-1}}.$$

Thus (63) follows from the following inequality (by taking $t = f(x)$ and $\theta = f_q(x)$):

$$\begin{aligned} &\frac{(1+t)^q_+(1-\theta)^{q-1}}{(1+\theta)^{q-1} + (1-\theta)^{q-1}} + \frac{(1-t)^q_+(1+\theta)^{q-1}}{(1+\theta)^{q-1} + (1-\theta)^{q-1}} \\ &\qquad -2\frac{(1-\theta)^{q-1}(1+\theta)^{q-1}}{(1+\theta)^{q-1} + (1-\theta)^{q-1}} \le |t - \theta|^q, \qquad \forall t \in \mathbb{R}, \theta \in (-1, 1). \end{aligned} \tag{64}$$

What is left is to verify the inequality (64). Since $-1 < \theta < 1$, we have

$$(1+t)^q_+(1-\theta)^{q-1} = (1-\theta^2)^{q-1}(1+\theta)\left(\frac{1+t}{1+\theta}\right)^q_+ = (1-\theta^2)^{q-1}(1+\theta)\left(1 + \frac{t-\theta}{1+\theta}\right)^q_+.$$

By Lemma 24 with $u = (t-\theta)/(1+\theta)$, we see that

$$(1+t)_+^q(1-\theta)^{q-1} \le (1-\theta^2)^{q-1}(1+\theta)\left(1+\left|\frac{t-\theta}{1+\theta}\right|^q + q\frac{t-\theta}{1+\theta}\right).$$

In the same way, by Lemma 24 with $u = -(t-\theta)/(1-\theta)$, we have

$$(1-t)_+^q(1+\theta)^{q-1} \le (1-\theta^2)^{q-1}(1-\theta)\left(1+\left|\frac{t-\theta}{1-\theta}\right|^q - q\frac{t-\theta}{1-\theta}\right).$$

Combining the above two estimates, we obtain

$$(1+t)_+^q(1-\theta)^{q-1} + (1-t)_+^q(1+\theta)^{q-1} \le 2(1-\theta^2)^{q-1} + |t-\theta|^q\left\{(1-\theta)^{q-1} + (1+\theta)^{q-1}\right\}.$$

This proves our claim (64), thereby Theorem 25. ■

Recall the K-functional given by (19).

Theorem 26 *For each $C > 0$, there holds*

$$\widetilde{\mathcal{D}}(C) \le \mathcal{K}(f_q, \frac{1}{2C}).$$

Proof The case $1 < q \le 2$ is an easy consequence of Theorem 25.

Turn to the case $q > 2$. The special choice $f = 0+0 \in \overline{\mathcal{H}_K}$ and the fact $\|f_q\|_{L^q_{\rho_X}} \le 1$ tell us that for any $t > 0$,

$$\mathcal{K}(f_q,t) = \inf_{\substack{f\in\overline{\mathcal{H}_K} \\ \|f-f_q\|_{L^q_{\rho_X}} \le 1}} \left\{q2^{q-1}(2^{q-1}+1)\|f-f_q\|_{L^q_{\rho_X}} + t\|f^*\|_K^2\right\}.$$

According to Theorem 25, for $f \in \overline{\mathcal{H}_K}$ with $\|f-f_q\|_{L^q_{\rho_X}} \le 1$, we have

$$\mathcal{E}(f) - \mathcal{E}(f_q) \le q2^{q-1}(2^{q-1}+1)\|f-f_q\|_{L^q_{\rho_X}}.$$

Thus,

$$\widetilde{\mathcal{D}}(C) = \inf_{f\in\overline{\mathcal{H}_K}}\left\{\mathcal{E}(f) - \mathcal{E}(f_q) + \frac{1}{2C}\|f^*\|_K^2\right\} \le \mathcal{K}(f_q, \frac{1}{2C})$$

and the proof of Theorem 26 is complete. ■

We can now derive several observations from Theorem 26.

The first observation says that the SVM q-classifier converges when f_q lies in the closure of $\overline{\mathcal{H}_K}$ in $L^q_{\rho_X}$. In particular, for any Borel probability measure ρ, this is always the case if K is a universal kernel since $C(X)$ is dense in $L^q_{\rho_X}$.

Corollary 27 *If f_q lies in the closure of $\overline{\mathcal{H}_K}$ in $L^q_{\rho_X}$, then for every $\varepsilon > 0$ and $0 < \delta < 1$, there exist $C_\varepsilon > 0$, $m_0 \in \mathbb{N}$ and a sequence C_m with $\lim_{m\to\infty} C_m = \infty$ such that*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) \le \varepsilon\right\} \ge 1-\delta, \qquad \forall m \ge m_0,\ C_\varepsilon \le C \le C_m.$$

Proof Since f_q lies in the closure of $\overline{\mathcal{H}}_K$ in $L^q_{\rho_X}$, for every $\varepsilon > 0$ there is some $f_\varepsilon = f^*_\varepsilon + b_\varepsilon \in \overline{\mathcal{H}}_K$ such that $q2^q(2^{q-1}+1)\|f_\varepsilon - f_q\|_{L^q_{\rho_X}} \le \varepsilon^2/16$. Take $C_\varepsilon = 8\|f^*_\varepsilon\|^2_K/\varepsilon^2$. Then for any $C \ge C_\varepsilon$, we have $\frac{1}{2C}\|f^*_\varepsilon\|^2_K \le \varepsilon^2/16$. Theorem 26 tells us that $\widetilde{\mathcal{D}}(C) \le \varepsilon^2/8$.

Take m_0 such that

$$\frac{2^{q+1}(1+\kappa\sqrt{2C_\varepsilon})^{q/2}\sqrt{\log(2/\delta)}}{\sqrt{m_0}} \le \frac{\varepsilon^2}{8}$$

and

$$\log \mathcal{N}\left(\frac{\varepsilon^3}{q4^{q+4}\sqrt{2C_\varepsilon}}\right) - \frac{3m_0\varepsilon^4}{4^{q+8}} \le \log\frac{\delta}{2}.$$

Also, we choose C_m such that

$$\frac{2^{q+1}(1+\kappa\sqrt{2C_m})^{q/2}\sqrt{\log(2/\delta)}}{\sqrt{m}} < \frac{\varepsilon^2}{8}$$

and

$$\log \mathcal{N}\left(\frac{\varepsilon^3}{q4^{q+4}\sqrt{2C_m}}\right) - \frac{3m\varepsilon^4}{4^{q+8}} \le \log\frac{\delta}{2}.$$

Then by Corollary 5.2, $\varepsilon(\delta, m, C, \beta) \le \frac{\varepsilon^2}{8}$ when $m \ge m_0$ and $C_\varepsilon \le C \le C_m$. Together with Theorem 14, our conclusion is proved. ∎

Our second observation from Theorem 26 concerns nonuniversal kernels which nonetheless ensures the convergence of the SVM q-classifier. The point here is that $\mathcal{H}_K$ is not dense in $C(X)$, but after adding the offset the space $\overline{\mathcal{H}}_K$ becomes dense.

Example 2 *Let K be a Mercer kernel on $X = [0,1]$:*

$$K(x,y) = \sum_{j\in J} a_j (x\cdot y)^j,$$

where J is a subset of $\mathbb{N}$, $a_j > 0$ for each $j \in J$, and $\sum_{j\in J} a_j < \infty$. Note that this kernel satisfies $K_0(y) \equiv 0$, hence $f(0) = 0$ for all $f \in \mathcal{H}_K$. Hence the space $\mathcal{H}_K$ is not dense in $C(X)$ and K is not an universal kernel. But if $\sum_{j\in J} \frac{1}{j} = \infty$, then $\overline{\mathcal{H}}_K$ is dense in $C(X)$ (Zhou 2003a) and hence in $L^q_{\rho_X}$. Therefore, the SVM q-classifier associated with the (identical) kernel K converges.

Remark 28 *In Section 4 and the proof of Theorem 21, we have shown how the offset influences the sample error. Proposition 4 and Example 2 tell that it may also influence the approximation error. However, our analysis in the following two sections will not focus on this point and it may be an interesting topic.*

In practical applications, one can use varying kernels for (3).

Definition 29 *Let $\{K_d\}_{d\in\mathbb{N}}$ be a sequence of Mercer kernels on X. We say that the SVM q-classifier associated with the kernels $\{K_d\}$ converges if for a sequence $\{C_m\}_{m\in\mathbb{N}}$ of positive numbers, $f_{\mathbf{z}}$ defined by (3) with $K = K_d$ and $C = C_m$ satisfies the following:*

For every Borel probability measure ρ *on* Z, *and* $0 < \delta < 1$, $\varepsilon > 0$, *for sufficiently large* d *there is some* $m_d \in \mathbb{N}$ *such that*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\{\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \le \varepsilon\} \ge 1-\delta, \qquad \forall m \ge m_d.$$

For a universal kernel K, one may take K_d to be identically K and the convergence holds. But the kernels could change such as the polynomial kernels (Boser, Guyon and Vapnik, 1992). Our third observation from Theorem 26 is to confirm the convergence of the SVM q-classifiers with these kernels.

Proposition 30 *For any* $1 < q < \infty, n \in \mathbb{N}$ *and* $X \subset \mathbb{R}^n$, *the SVM q-classifier associated with* $\{K_d\}_{d=1}^{\infty}$, *the polynomial kernels* $K_d(x,y) = (1 + x\cdot y)^d$, *converges.*

Proposition 30 is a consequence of a quantitative result below.

Let P_d be the space of all polynomials of degree at most d. It is a RKHS $\mathcal{H}_{K_d}$ with the Mercer kernel $K_d(x,y) = (1+x\cdot y)^d$. The rich knowledge from approximation theory tells us that for an arbitrary Borel probability measure on Z, there is a sequence of polynomials $\{p_d \in P_d\}_{d=1}^{\infty}$ such that $\lim_{d\to\infty} \|f_q - p_d\|_{L^q_{\rho_X}} = 0$. The rate of this convergence depends on the regularity of the function f_q (hence the function f_ρ) and the marginal distribution ρ_X. With this in hand, we can now state the result on the convergence of the q-classifier with polynomial kernels K_d.

Corollary 31 *Let* $X \subset \mathbb{R}^n$, *and* ρ *be an arbitrary Borel probability measure on* Z. *Let* $d \in \mathbb{N}$ *and* $K_d(x,y) = (1+x\cdot y)^d$ *be the polynomial kernel. Set* $\|X\| := \sup_{x\in X}|x|$. *Let* $\{p_d \in P_d\}_{d=1}^{\infty}$ *satisfy* $E_d := \|f_q - p_d\|_{L^q_{\rho_X}} \to 0$ *(as* $d \to \infty$*). Set* $N := (n+d)!/(n!d!) + 1$ *and* $0 < \sigma < \frac{2}{q}$. *Then there exists* $m_{q,\sigma} \in \mathbb{N}$ *such that for* $m \ge m_{q,\sigma}$ *and* $C = m^{\sigma}$, *for every* $0 < \delta < 1$, *with confidence* $1-\delta$ *there holds*

$$\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \le \sqrt{q}2^{q+1}\sqrt{E_d} + \frac{\|p_d\|_{K_d}}{m^{\sigma/2}} + \frac{2^{q+5}\big(N\log\big(\frac{2}{\delta}\big)\big)^{1/4}(1+\|X\|^2)^{qd/8}}{m^{\frac{1}{4}-\frac{q\sigma}{8}}}.$$

Proof Take $p_d^* = p_d$ with zero offset in the K-functional. Then by Theorem 26,

$$\widetilde{\mathcal{D}}(C) \le \mathcal{K}(f_q, \frac{1}{2C}) \le q2^{q-1}(2^{q-1}+1)E_d + \frac{1}{2C}\|p_d\|^2_{K_d}.$$

The covering numbers of the finite dimensional space P_d (e.g. Cucker and Zhou, 2004) and (22) give us the estimate:

$$\mathcal{N}(\eta) \le (\kappa + \frac{1}{\tilde{\kappa}})\left(\frac{2}{\eta}+1\right)^N,$$

where $N-1 = (n+d)!/(n!d!)$ is the dimension of the space $P_d(\mathbb{R}^n)$. Also, $\kappa = \sqrt{\|K_d\|_\infty} \le (1+\|X\|^2)^{d/2}$.

Take $C = m^{\sigma}$. By Corollary 23, solving the equation (62) yields

$$\varepsilon(\delta, m, C, \beta) \le \frac{(2^{q+5}\sqrt{N} + \sqrt{\log(\kappa+\frac{1}{\tilde{\kappa}})}(\log m + \log(2/\delta))^{1/2}}{\sqrt{m}} + \frac{4^{q+2}\kappa^{\frac{q}{2}}\sqrt{\log(\frac{2}{\delta})}}{m^{\frac{1}{2}-\frac{q\sigma}{4}}}$$

which is bounded by $4^{5+q}\sqrt{N\log\left(\frac{2}{\delta}\right)}\kappa^{q/2}m^{\frac{q\sigma}{4}-\frac{1}{2}}$ for $m \geq m_{q,\sigma}$. Here $m_{q,\sigma}$ is an integer depending on q,σ (but not on m, d or δ). Then for each $m \geq m_{q,\sigma}$, with confidence $1-\delta$ the desired estimate holds true. ■

Remark 32 *Note that P_d is a finite dimension space. Thus the norms $\|\cdot\|_{K_d}$ and $\|\cdot\|_{L^q_{\rho_X}}$ are equivalent for a fixed d when ρ_X is non-degenerate. It would be interesting to compare the norm $\|p\|_{K_d}$ with $\|p\|_{L^q_{\rho_X}}$ for $p \in P_d$ as d tends to infinity.*

Proof of Proposition 30. For every $\varepsilon > 0$, there exists some $d_0 \in \mathbb{N}$ such that $\sqrt{q}2^{q-1}\sqrt{E_d} \leq \varepsilon/2$ for every $d \geq d_0$. Then by Corollary 31 we can find some $m_{q,\sigma} \leq m_d \in \mathbb{N}$ such that $m_d^{-\sigma/2}\|p_d\|_{K_d} \leq \varepsilon/4$ and $2^{q+5}(N\log(2/\delta))^{1/4}(1+\|X\|^2)^{dq/8}m_d^{\frac{q\sigma}{8}-\frac{1}{4}} \leq \varepsilon/4$. Then for any $m \geq m_d$ we have $\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leq \varepsilon$ with confidence $1-\delta$. This proves Proposition 30. ■

7. Rate of Convergence for the q-Norm Soft Margin Classifier

Corollary 23 and Theorem 26 enable us to get the convergence rate for the SVM q-classifier. The rate depends on the K-functional $\mathcal{K}(f_q,t)$. It can be characterized by the quantity (Smale and Zhou 2003; Zhou 2003b)

$$I_q(g,R) := \inf_{\substack{f\in\overline{\mathcal{H}}_K \\ \|f^*\|_K\leq R}} \left\{\|g-f\|_{L^q_{\rho_X}}\right\}. \tag{65}$$

Define

$$J_q(f_q,R) := \begin{cases} (I_q(f_q,R))^q, & \text{if } 1<q\leq 2, \\ q2^{q-1}(2^{q-1}+1)I_q(f_q,R), & \text{if } q>2. \end{cases}$$

Then the following corollary holds true.

Corollary 33 *For any $t>0$ there holds*

$$\mathcal{K}(f_q,t) \leq \inf_{R>0}\{J_q(f_q,R)+tR^2\}.$$

One may choose appropriate R to estimate the convergence rate of $\mathcal{K}(f_q,t)$, which together with Corollary 23 gives the convergence rate of the V-risk and a strategy of choosing the regularization parameter C. In general, the choice of R depends on the regularity of f_q and the kernel K. Let us demonstrate this by examples.

In what follows let $X \subset \mathbb{R}^n$ have Lipschitz boundary and ρ be a probability measure such that $d\rho_X = dx$ is the Lebesgue measure. Consider $q=2$ and thus $f_q = f_\rho$. We use the approximation error studied in Smale and Zhou (2003) (see also Niyogi and Girosi (1996) for related discussion):

$$I_2^*(g,R) := \inf_{\substack{f^*\in\mathcal{H}_K \\ \|f^*\|_K\leq R}} \{\|g-f^*\|_{L^2}\}$$

to bound the term $I_2(f_\rho,R)$. With the choice $b_f=0$ we obtain

$$I_2(f_\rho,R) \leq I_2^*(f_\rho,R).$$

Note that we disregard the influence of the offset here and thus the rate need not be optimal.

The first example includes spline kernels (Wahba 1990).

Example 3 *Let $X \subset \mathbb{R}^n$ and K be a Mercer kernel such that $\mathcal{H}_K$ is the Sobolev space $H^r(X)$ with $r > n/2$. If f_ρ lies in the Sobolev space $H^s(X)$ with $0 < s < r$, then*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) = O\left(\frac{\sqrt{\log m + C}}{\sqrt{m}} + \frac{C^{n/(4r+3n)}}{m^{2r/(4r+3n)}}\right) + O\left(C^{-s/r}\right).$$

Thus, C should be chosen such that $C \to \infty$, $C/m \to 0$ as $m \to \infty$.

Proof It was shown in Smale and Zhou (2003, Theorem 3.1) that for $0 < \theta < 1$, $I_2^*(f_\rho, R) = O(R^{-\theta/(1-\theta)})$ if and only if f_ρ lies in the interpolation space $(L^2_{\rho_X}, \mathcal{H}_K)_{\theta,\infty}$. It is well known that $H^s(X) \subset (L^2, H^r(X))_{s/r,\infty}$ for $0 < s < r$. Here $\mathcal{H}_K = H^r(X)$ and $d\rho_X = dx$. Therefore, the assumption $f_\rho \in H^s(X)$ tells us that $f_\rho \in (L^2_{\rho_X}, \mathcal{H}_K)_{s/r,\infty}$. Hence there holds

$$I_2(f_\rho, R) \le I_2^*(f_\rho, R) \le C_\rho R^{-s/(r-s)}$$

for some constant C_ρ. Choose $R = C_\rho^{(r-s)/r} C^{(r-s)/2r}$ to obtain

$$\mathcal{K}(f_\rho, \frac{1}{2C}) \le (I_2(f_\rho, R))^2 + \frac{R^2}{2C} \le \frac{3}{2} C_\rho^{2(r-s)/r} C^{-s/r}.$$

Using the well known covering number estimates for Sobolev spaces

$$\log \mathcal{N}(B_R, \eta) \le C_r \left(\frac{1}{\eta}\right)^{n/r}$$

and solving the equation (62), we see that

$$\varepsilon^* \le 2^6 \sqrt{\frac{\log \kappa + \log m + \log C + \log(2/\delta)}{m}} + 2^{6+3n/(4r)} \sqrt{C_r} \frac{C^{n/(4r+3n)}}{m^{2r/(4r+3n)}}.$$

This proves the conclusion. ∎

Example 4 *Let $\sigma > 0, s > 0$ and K be the Gaussian kernel $K(x,y) = \exp\left\{-\frac{|x-y|^2}{\sigma^2}\right\}$.*

(a) *If f_ρ lies in the interpolation space $(L^2_{\rho_X}, \mathcal{H}_K)_{\theta,\infty}$ for some $0 < \theta < 1$, that is, $\mathcal{K}(f_\rho, t) \le C_\theta t^\theta$ for some constant C_θ, then for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) = O\left(\frac{\sqrt{C + (\log m)^{n+1}}}{\sqrt{m}}\right) + O\left(C^{-\theta}\right).$$

This implies the parameter C should be taken to satisfy $C \to \infty$ and $C/m \to 0$ as $m \to \infty$. An asymptotic optimal choice is $C = O(m^{1/(1+2\theta)})$. With this choice, the convergence rate is $O(m^{-\theta/(1+2\theta)})$.

(b) *If $f_\rho \in H^s(X)$ with $s > 0$, then for any $0 < \delta < 1$, with confidence $1 - \delta$, we have*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) = O\left(\frac{\sqrt{C + (\log m)^{n+1}}}{\sqrt{m}}\right) + O\left((\log C)^{-s/2}\right).$$

An asymptotically optimal choice is $C = O\left(\frac{m}{(\log m)^s}\right)$ which gives the convergence rate $O((\log m)^{-s/2})$.

Proof Solving the equation (62) with the covering number estimate (23) yields

$$\varepsilon^* = O\left(\frac{\sqrt{(1+c)(\log C + \log m)^{n+1}}}{\sqrt{m}}\right).$$

Then the statement in (a) follows easily from Corollary 23, Theorem 21 and Corollary 33.

To see the conclusion in (b), we notice that the assumption $f_\rho \in H^s(X)$ provides the approximation error estimates (Smale and Zhou, 2003; Zhou, 2003b)

$$I_2(f_\rho, R) \le I_2^*(f_\rho, R) \le C_s\,(\log R)^{-s/4}$$

for every $R > C_s$, where C_s is a constant depending on s, σ, n and the Sobolev norm of f_ρ. Choose $R = \sqrt{2C}\,(\log C)^{-s/4}$ to obtain

$$\mathcal{K}(f_\rho, \frac{1}{2C}) \le (I_2(f_\rho, R))^2 + \frac{R^2}{2C} \le (2^s C_s^2 + 1)(\log C)^{-s/2}.$$

Then the desired bound follows from Theorem 26 and the above established bound for ε^*. ■

It was shown in Smale and Zhou (2003) that for the Gaussian kernel in Example 4, $I_2^*(g, R) = O(R^{-\varepsilon})$ with $\varepsilon > 0$ only if g is C^∞. Hence logarithmic convergence rate is expected for a Sobolev function f_ρ. However, in practice, one often chooses the different variances σ of the Gaussian kernel according to the different sample size m. With this flexibility, the regularization error can be improved greatly and polynomial convergence rates are possible.

Acknowledgments

The authors would like to thank Professor Zhen Wang for his helpful discussion on the best constants in Theorem 14 and Theorem 25. They are also grateful to the referees for the constructive suggestions and comments.

This work is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 1087/02P] and by City University of Hong Kong [Project No. 7001442]. The corresponding author is Ding-Xuan Zhou. Yiming Ying is on leave from Institute of Mathematics, Chinese Academy of Science, Beijing 100080, P. R. CHINA.

Appendix A. Error Comparison for a General Loss Function

In this appendix for a general convex loss function we bound the excess misclassification error by the excess V-risk. Here the loss function takes the form

$$V(y, f(x)) = \phi(yf(x))$$

for a univariate function $\phi : \mathbb{R} \to \mathbb{R}_+$.

For each $x \in X$, we denote

$$\mathcal{E}(f|x) := \int_Y V(y, f(x))\, d\rho(y|x), \tag{66}$$

then $\mathcal{E}(f|x) = Q(\eta(x), f(x))$. Here $Q : [0,1] \times (\mathbb{R} \cup \{\pm\infty\}) \to \mathbb{R}_+$ is given by

$$Q(\eta, f) = \eta\phi(f) + (1-\eta)\phi(-f),$$

and $\eta : X \to \mathbb{R}$ is defined by $\eta(x) := P(\mathcal{Y} = 1|x)$. Set

$$f_\phi^*(\eta) := \arg \min_{f \in \mathbb{R} \cup \{\pm\infty\}} Q(\eta, f).$$

Then $f_\rho^V(x) = f_\phi^*(\eta(x))$. The main result of Zhang (2004) can be stated as follows.

Theorem A *Let ϕ be convex. Assume $f_\phi^*(\eta) > 0$ when $\eta > 0.5$. Assume there exists $c > 0$ and $s \geq 1$ such that for all $\eta \in [0,1]$,*

$$|0.5 - \eta|^s \leq c^s \left(Q(\eta, 0) - Q(\eta, f_\phi^*(\eta))\right), \tag{67}$$

then for any measurable function $f(x)$:

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c) \leq 2c \left(\mathcal{E}(f) - \mathcal{E}(f_\rho^V)\right)^{1/s}.$$

Further analysis was made by Bartlett, Jordan and McAuliffe (2003). For example, it was proved in Bartlett, Jordan and McAuliffe (2003, Theorem 6) that for a convex function ϕ, $f_\phi^*(\eta) > 0$ for any $\eta > 0.5$ if and only if ϕ is differentiable at 0 and $\phi'(0) < 0$. Borrowing some ideas from Bartlett, Jordan and McAuliffe (2003), we can derive a simple criterion for the condition (67) with $s = 2$. The existence of $\phi''(0)$ means that the function $\phi'(x)$ is well defined in a neighborhood of 0 and is differentiable at 0. Note that the convexity of ϕ implies $\phi''(0) \geq 0$.

Theorem 34 *Let $\phi : \mathbb{R} \to \mathbb{R}_+$ be a convex function such that $\phi''(0)$ exists. If $\phi'(0) < 0$ and $\phi''(0) > 0$, then (67) holds for $s = 2$. Hence for any measurable function f:*

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c) \leq 2c\sqrt{\mathcal{E}(f) - \mathcal{E}(f_\rho^V)}.$$

Proof By the definition of $\phi''(0)$, there exists some $1/2 \geq c_0 > 0$ such that

$$\left|\frac{\phi'(f) - \phi'(0)}{f} - \phi''(0)\right| \leq \frac{\phi''(0)}{2}, \qquad \forall f \in [-c_0, c_0].$$

This implies

$$\phi'(0) + \phi''(0)f - \frac{\phi''(0)}{2}|f| \leq \phi'(f) \leq \phi'(0) + \phi''(0)f + \frac{\phi''(0)}{2}|f|.$$

If $\eta > 1/2$, then for $0 \leq f \leq c_0$,

$$\frac{\partial Q}{\partial f}(\eta, f) = \eta\phi'(f) - (1-\eta)\phi'(-f) \leq (2\eta - 1)\phi'(0) + \phi''(0)f + \frac{\phi''(0)}{2}f.$$

Thus for $0 \leq f \leq \Delta_\eta := \min\{\frac{-\phi'(0)}{\phi''(0)}(\eta - \frac{1}{2}), c_0\}$, we have

$$\frac{\partial Q}{\partial f}(\eta, f) \leq (2\eta - 1)\phi'(0) + \frac{3}{2}\phi''(0)\frac{-\phi'(0)}{\phi''(0)}(\eta - \frac{1}{2}) \leq \frac{\phi'(0)}{2}(\eta - \frac{1}{2}) < 0.$$

Therefore as a function of the variable f, $Q(\eta, f)$ is strictly decreasing on the interval $[0,\Delta_\eta]$. But $f_\phi^*(\eta) > 0$ is its minimal point, hence

$$Q(\eta,0) - Q(\eta, f_\phi^*(\eta)) \geq Q(\eta,0) - Q(\eta,\Delta_\eta) \geq -\frac{\phi'(0)}{2}(\eta - \frac{1}{2})\Delta_\eta.$$

When $\frac{-\phi'(0)}{\phi''(0)}(\eta - \frac{1}{2}) > c_0$, we have $\Delta_\eta = c_0 \geq 2c_0(\eta - \frac{1}{2})$. Hence

$$Q(\eta,0) - Q(\eta, f_\phi^*(\eta)) \geq \frac{-\phi'(0)}{2} \min\left\{2c_0, \frac{-\phi'(0)}{\phi''(0)}\right\} (\eta - \frac{1}{2})^2.$$

That is, (67) holds with $s = 2$ and

$$c = \max\left\{\frac{\sqrt{2\phi''(0)}}{-\phi'(0)}, \sqrt{\frac{1}{-\phi'(0)c_0}}\right\}.$$

The proof for $\eta < 1/2$ is the same by estimating the upper bound of $\frac{\partial Q}{\partial f}(\eta, f)$ for $f < 0$. ■

Turn to the special loss function $V = V_q$ given in (6) by $\phi(t) = (1-t)_+^q$. Applying Theorem 34, we see that the function ϕ satisfies $\phi'(0) = -q < 0$ and $\phi''(0) = q(q-1) > 0$. This verifies (40) and the constant c can be obtained from the proof of Theorem 34.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** (1950), 337–404, 1950.

A. R. Barron. Complexity regularization with applications to artificial neural networks. G. Roussa, editor, in *Nonparametric Functional Estimation*, Kluwer, Dortrecht, pages 561–576, 1990.

P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, **44**: 525–536, 1998.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals Statis.*, 2004, to appear.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. Preprint, 2003.

B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop of Computational Learning Theory*, **5**, Pittsburgh, ACM, pages 144–152, 1992.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, **2**: 499–526, 2002.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, **20**: 273–297, 1995.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, **39**: 1–49, 2001.

F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Monograph manuscript in preparation for Cambridge University Press, 2004.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1997.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Adv. Comput. Math.*, **13**: 1–50, 2000.

W. S. Lee, P. Bartlett, and R. Williamson. The importance of convexity in learning with least square loss. *IEEE Transactions on Information Theory*, **44**: 1974–1980, 1998.

Y. Lin. Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery* **6**: 259–275, 2002.

S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory* **48**: 1977–1991, 2002.

S. Mukherjee, R. Rifkin, and T. Poggio. Regression and classification with regularization. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu, eds., *Lecture Notes in Statistics: Nonlinear Estimation and Classification*, Springer-Verlag, New York, pages 107–124, 2002.

P. Niyogi and F. Girosi. On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. *Neural Comp.*, **8**: 819–842, 1996.

M. Pontil. A note on different covering numbers in learning theory. *J. Complexity* **19**: 665–671, 2003.

F. Rosenblatt. *Principles of Neurodynamics*. Spartan Book, New York, 1962.

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transations on Information Theory* **44**: 1926–1940, 1998.

S. Smale and D. X. Zhou. Estimating the approximation error in learning theory. *Anal. Appl.*, **1**: 17–41, 2003.

S. Smale and D. X. Zhou. Shannon sampling and function reconstruction from point values. *Bull. Amer. Math. Soc.* , **41**: 279–305, 2004.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, **2**: 67–93, 2001.

I. Steinwart. Support vector machines are universally consistent. *J. Complexity*, **18**: 768–791, 2002.

A. Tikhonov and V. Arsenin. *Solutions of Ill-posed Problems*. W. H. Winston, 1977.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

G. Wahba. *Spline models for observational data*. SIAM, 1990.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the Randomized GACV. In Schölkopf, Burges and Smola, eds., *Advances in Kernel Methods - Support Vector Learning*, MIT Press, pages 69–88, 1999.

R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory* **47**: 2516–2532, 2001.

Q. Wu and D. X. Zhou. Analysis of support vector machine classification. Preprint, 2004.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals Statis.*, **32**: 56–85, 2004.

D. X. Zhou. The covering number in learning theory. *J. Complexity*, **18**: 739–767, 2002.

D. X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, **49**: 1743–1752, 2003a.

D. X. Zhou. Density problem and approximation error in learning theory. Preprint, 2003b.