SVM Soft Margin Classifiers: Linear Programming versus Quadratic Programming

Qiang Wu
wu.qiang@student.cityu.edu.hk
Ding-Xuan Zhou
mazhou@cityu.edu.hk
Department of Mathematics, City University of Hong Kong,
Tat Chee Avenue, Kowloon, Hong Kong, China

Support vector machine soft margin classifiers are important learning algorithms for classification problems. They can be stated as convex optimization problems and are suitable for a large data setting. Linear programming SVM classifier is specially efficient for very large size samples. But little is known about its convergence, compared with the well understood quadratic programming SVM classifier. In this paper, we point out the difficulty and provide an error analysis. Our analysis shows that the convergence behavior of the linear programming SVM is almost the same as that of the quadratic programming SVM. This is implemented by setting a stepping stone between the linear programming SVM and the classical 1–norm soft margin classifier. An upper bound for the misclassification error is presented for general probability distributions. Explicit learning rates are derived for deterministic and weakly separable distributions, and for distributions satisfying some Tsybakov noise condition.

1 Introduction

Support vector machines (SVM's) form an important subject in learning theory. They are very efficient for many applications, especially for classification problems.

The classical SVM model, the so-called 1–norm soft margin SVM, was introduced with polynomial kernels by Boser et al. (1992) and with general kernels by Cortes and Vapnik (1995). Since then many different forms of SVM algorithms were introduced for different purposes (e.g. Niyogi and Girosi 1996; Vapnik 1998). Among them the linear programming (LP) SVM (Bradley and Mangasarian 2000; Kecman and Hadzic 2000; Niyogi and Girosi 1996; Pedroso and N. Murata 2001; Vapnik 1998) is an important one because of its linearity and flexibility for large data setting. The term "linear programming" means the algorithm is based on linear programming optimization. Correspondingly, the 1–norm soft margin SVM is also called quadratic programming (QP) SVM since it is based on quadratic programming optimization (Vapnik 1998). Many experiments demonstrate that LP-SVM is efficient and performs even better than QP-SVM for some purposes: capable of solving huge sample size problems (Bradley and Mangasarian 2000), improving the computational speed (Pedroso and N. Murata 2001), and reducing the number of support vectors (Kecman and Hadzic 2000).

While the convergence of QP-SVM has become pretty well understood because of recent works (Steinwart 2002; Zhang 2004; Wu and Zhou 2003; Scovel and Steinwart 2003; Wu et al. 2004), little is known for LP-SVM. The purpose of this paper is to point out the main difficulty and then provide error analysis for LP-SVM.

Consider the binary classification setting. Let (X, d) be a compact metric space and $Y = \{1, -1\}$. A *binary classifier* is a function $f : X \to Y$ which labels every point $x \in X$ with some $y \in Y$.

Both LP-SVM and QP-SVM considered here are kernel based classifiers. A function $K : X \times X \to \mathbb{R}$ is called a *Mercer kernel* if it is 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.

Let $\mathbf{z} = \{(x_1, y_1), \cdots, (x_m, y_m)\} \subset (X \times Y)^m$ be the sample. Motivated by reducing the number of support vectors of the 1–norm soft margin SVM, Vapnik (1998) introduced the LP-SVM algorithm associated to a Mercer Kernel K. It is based on the following *linear programming* optimization problem:

$$\begin{aligned} \min_{\alpha \in \mathbb{R}_+^m,\ b \in \mathbb{R}} \quad & \Big\{ \frac{1}{m} \sum_{i=1}^{m} \xi_i + \frac{1}{C} \sum_{i=1}^{m} \alpha_i \Big\} \\ \text{subject to} \quad & y_i \Big(\sum_{j=1}^{m} \alpha_j y_j K(x_i, x_j) + b \Big) \geq 1 - \xi_i \\ & \xi_i \geq 0,\ i = 1, \cdots, m. \end{aligned} \tag{1.1}$$

Here $\alpha = (\alpha_1, \cdots, \alpha_m)$, ξ_i's are slack variables. The trade-off parameter $C = C(m) > 0$ depends on m and is crucial. If $\big(\alpha_{\mathbf{z}} = (\alpha_{1,\mathbf{z}}, \cdots, \alpha_{m,\mathbf{z}}), b_{\mathbf{z}}\big)$ solves the optimization problem (1.1), the LP-SVM classifier is given by $\operatorname{sgn}(f_{\mathbf{z}})$ with

$$f_{\mathbf{z}}(x) = \sum_{i=1}^{m} \alpha_{i,\mathbf{z}} y_i K(x, x_i) + b_{\mathbf{z}}. \tag{1.2}$$

For a real-valued function $f : X \to \mathbb{R}$, its sign function is defined as $\operatorname{sgn}(f)(x) = 1$ if $f(x) \geq 0$ and $\operatorname{sgn}(f)(x) = -1$ otherwise.

The QP-SVM is based on a *quadratic programming* optimization problem:

$$\begin{aligned} \min_{\alpha \in \mathbb{R}_+^m,\ b \in \mathbb{R}} \quad & \Big\{ \frac{1}{m} \sum_{i=1}^{m} \xi_i + \frac{1}{2\widetilde{C}} \sum_{i,j=1}^{m} \alpha_i y_i K(x_i, x_j) \alpha_j y_j \Big\} \\ \text{subject to} \quad & y_i \Big(\sum_{j=1}^{m} \alpha_j y_j K(x_i, x_j) + b \Big) \geq 1 - \xi_i \\ & \xi_i \geq 0,\ i = 1, \cdots, m. \end{aligned} \tag{1.3}$$

Here $\widetilde{C} = \widetilde{C}(m) > 0$ is also a trade-off parameter depending on the sample size m. If $\big(\widetilde{\alpha}_{\mathbf{z}} = (\tilde{\alpha}_{1,\mathbf{z}}, \cdots, \tilde{\alpha}_{m,\mathbf{z}}), \widetilde{b}_{\mathbf{z}}\big)$ solves the optimization problem (1.3), then the 1–norm soft margin classifier is defined by $\operatorname{sgn}(\widetilde{f}_{\mathbf{z}})$ with

$$\widetilde{f}_{\mathbf{z}}(x) = \sum_{i=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} y_i K(x, x_i) + \widetilde{b}_{\mathbf{z}}. \tag{1.4}$$

Observe that both LP-SVM classifier (1.1) and QP-SVM classifier (1.3) are implemented by *convex* optimization problems. Compared with this, neural network learning algorithms are often performed by nonconvex optimization problems.

The reproducing kernel property of Mercer kernels ensures nice approximation power of SVM classifiers. Recall that the *Reproducing Kernel Hilbert Space* (RKHS) $\mathcal{H}_K$ associated with a Mercer 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 $< \cdot,\cdot >_K$ satisfying $< K_x, K_y >_K = K(x,y)$. The reproducing property is given by

$$< f, K_x >_K = f(x), \qquad \forall x \in X, f \in \mathcal{H}_K. \tag{1.5}$$

The QP-SVM is well understood. It has attractive approximation properties (see (2.2) below) because the learning scheme can be represented as a Tikhonov regularization (Evgeniou et al. 2000) (modified by an offset) associated with the RKHS:

$$\widetilde{f}_{\mathbf{z}} = \underset{f=f^*+b\in\mathcal{H}_K+\mathbb{R}}{\arg\min} \Big\{ \frac{1}{m}\sum_{i=1}^{m} \big(1 - y_i f(x_i)\big)_+ + \frac{1}{2\widetilde{C}} \|f^*\|_K^2 \Big\}, \tag{1.6}$$

where $(t)_+ = \max\{0, t\}$. Set $\overline{\mathcal{H}}_K := \mathcal{H}_K + \mathbb{R}$. For a function $f = f_1 + b_1 \in \overline{\mathcal{H}}_K$, we denote $f^* = f_1$ and $b_f = b_1$. Write $b_{f_{\mathbf{z}}}$ as $b_{\mathbf{z}}$.

It turns out that (1.6) is the same as (1.3) together with (1.4). To see this, we first note that $\widetilde{f}_{\mathbf{z}}^*$ must lies in the span of $\{K_{x_i}\}_{i=1}^m$ according to the representation theorem (Wahba 1990). Next, the dual problem of (1.6) shows (Vapnik 1998) that the coefficient of K_{x_i}, $\alpha_i y_i$, has the same sign as y_i. Finally, the definition of the $\mathcal{H}_K$ norm yields $\|\widetilde{f}_{\mathbf{z}}^*\|_K^2 = \|\sum_{i=1}^m \alpha_i y_i K_{x_i}\|_K^2 = \sum_{i,j=1}^m \alpha_i y_i K(x_i, x_j) \alpha_j y_j$.

The rich knowledge on Tikhonov regularization schemes and the idea of bias-variance trade-off developed in the neural network literature provide a mathematical foundation of the QP-SVM. In particular, the convergence is well understood due to the work done within the last a few years. Here the form (1.6) illustrate some advantages of the QP-SVM: the minimization is

taken over the whole space $\overline{\mathcal{H}}_K$, so we expect the QP-SVM has some good approximation power, similar to the approximation error of the space $\overline{\mathcal{H}}_K$.

Things are totally different for LP-SVM. Set

$$\mathcal{H}_{K,\mathbf{z}} = \left\{ \sum_{i=1}^{m} \alpha_i y_i K(x, x_i) : \alpha = (\alpha_1, \cdots, \alpha_m) \in \mathbb{R}_+^m \right\}.$$

Then the LP-SVM scheme (1.1) can be written as

$$f_{\mathbf{z}} = \arg\min_{f=f^*+b \in \mathcal{H}_{K,\mathbf{z}}+\mathbb{R}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \big(1 - y_i f(x_i)\big)_+ + \frac{1}{C} \Omega(f^*) \right\}. \tag{1.7}$$

Here we have denoted $\Omega(f^*) = \|y\alpha\|_{\ell^1} = \sum_{i=1}^m \alpha_i$ for $f^* = \sum_{i=1}^m \alpha_i y_i K_{x_i}$ with $\alpha_i \geq 0$. It plays the role of a norm of f^* in some sense. This is not a Hilbert space norm, which raises the technical difficulty for the mathematical analysis. More seriously, the hypothesis space $\mathcal{H}_{K,\mathbf{z}}$ depends on the sample $\mathbf{z}$. The "centers" x_i of the basis functions in $\mathcal{H}_{K,\mathbf{z}}$ are determined by the sample $\mathbf{z}$, not free. One might consider regularization schemes in the space of all linear combinations with free centers, but whether the minimization can be reduced into a convex optimization problem of size m, like (1.1), is unknown. Also, it is difficult to relate the corresponding optimum (in a ball with radius C) to $f_{\mathbf{z}}^*$ with respect to the estimation error. Thus separating the error for LP-SVM into two terms of sample error and approximation error is not as immediate as for the QP-SVM or neural network methods (Niyogi and Girosi 1996) where the centers are free. In this paper, we shall overcome this difficulty by setting a stepping stone.

Turn to the error analysis. Let ρ be a Borel probability measure on $Z := X \times Y$ and $(\mathcal{X}, \mathcal{Y})$ be the corresponding random variable. The prediction power of a classifier f is measured by its misclassification error, i.e., the probability of the event $f(\mathcal{X}) \neq \mathcal{Y}$:

$$\mathcal{R}(f) = \text{Prob}\{f(\mathcal{X}) \neq \mathcal{Y}\} = \int_X P(\mathcal{Y} \neq f(x)|x)\, d\rho_X. \tag{1.8}$$

Here ρ_X is the marginal distribution and $\rho(\cdot|x)$ is the conditional distribution of ρ. The classifier minimizing the misclassification error is called the *Bayes*

rule f_c. It takes the form

$$f_c(x) = \begin{cases} 1, & \text{if } P(\mathcal{Y}=1|x) \geq P(\mathcal{Y}=-1|x), \\ -1, & \text{if } P(\mathcal{Y}=1|x) < P(\mathcal{Y}=-1|x). \end{cases}$$

If we define the regression function of ρ as

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

then $f_c = \text{sgn}(f_\rho)$. Note that for a real-valued function f, $\text{sgn}(f)$ gives a classifier and its misclassification error will be denoted by $\mathcal{R}(f)$ for abbreviation.

Though the Bayes rule exists, it can not be found directly since ρ is unknown. Instead, we have in hand a set of samples $\mathbf{z} = \{z_i\}_{i=1}^m = \{(x_1, y_1), \cdots, (x_m, y_m)\}$ $(m \in \mathbb{N})$. Throughout the paper we assume $\{z_1, \cdots, z_m\}$ are independently and identically distributed according to ρ. A classification algorithm constructs a classifier $f_{\mathbf{z}}$ based on $\mathbf{z}$.

Our goal is to understand how to choose the parameter $C = C(m)$ in the algorithm (1.1) so that the LP-SVM classifier $\text{sgn}(f_{\mathbf{z}})$ can approximate the Bayes rule f_c with satisfactory convergence rates (as $m \to \infty$). Our approach provides clues to study learning algorithms with penalty functional different from the RKHS norm (Niyogi and Girosi 1996; Evgeniou et al. 2000). It can be extended to schemes with general loss functions (Rosasco et al. 2004; Lugosi and Vayatis 2004; Wu et al. 2004).

2 Main Results

In this paper we investigate learning rates, the decay of the excess misclassification error $\mathcal{R}(f_{\mathbf{z}}) - \mathcal{R}(f_c)$ as m and $C(m)$ become large.

Consider the QP-SVM classification algorithm $\widetilde{f}_{\mathbf{z}}$ defined by (1.3). Steinwart (2002) showed that $\mathcal{R}(\widetilde{f}_{\mathbf{z}}) - \mathcal{R}(f_c) \to 0$ (as m and $\widetilde{C} = \widetilde{C}(m) \to \infty$), when $\mathcal{H}_K$ is dense in $C(X)$, the space of continuous functions on X with the norm $\|\cdot\|_\infty$. Lugosi and Vayatis (2004) found that for the exponential loss, the excess misclassification error of regularized boosting algorithms

can be estimated by the excess generalization error. An important result on the relation between the misclassification error and generalization error for a convex loss function is due to Zhang (2004). See Bartlett et al. (2003), and Chen et al. (2004) for extensions to general loss functions. Here we consider the hinge loss $V(y, f(x)) = (1 - yf(x))_+$. The generalization error is defined as

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

Note that f_c is a minimizer of $\mathcal{E}(f)$. Then Zhang's results asserts that

$$\mathcal{R}(f) - \mathcal{R}(f_c) \le \mathcal{E}(f) - \mathcal{E}(f_c), \qquad \forall f : X \to \mathbb{R}. \tag{2.1}$$

Thus, the excess misclassification error $\mathcal{R}(\widetilde{f}_{\mathbf{z}}) - \mathcal{R}(f_c)$ can be bounded by the excess generalization error $\mathcal{E}(\widetilde{f}_{\mathbf{z}}) - \mathcal{E}(f_c)$, and the following error decomposition (Wu and Zhou 2003) holds:

$$\mathcal{E}(\widetilde{f}_{\mathbf{z}}) - \mathcal{E}(f_c) \le \Big\{\mathcal{E}\big(\widetilde{f}_{\mathbf{z}}\big) - \mathcal{E}_{\mathbf{z}}\big(\widetilde{f}_{\mathbf{z}}\big) + \mathcal{E}_{\mathbf{z}}\big(\widetilde{f}_{K,\widetilde{C}}\big) - \mathcal{E}\big(\widetilde{f}_{K,\widetilde{C}}\big)\Big\} + \widetilde{\mathcal{D}}(\widetilde{C}). \tag{2.2}$$

Here $\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^m V(y_i, f(x_i))$. The function $\widetilde{f}_{K,\widetilde{C}}$ depends on $\widetilde{C}$ and is defined as

$$\widetilde{f}_{K,C} := \arg\min_{f \in \overline{\mathcal{H}}_K} \Big\{\mathcal{E}(f) + \frac{1}{2C}\|f^*\|_K^2\Big\}, \qquad C > 0. \tag{2.3}$$

The decomposition (2.2) makes the error analysis for QP-SVM easy, similar to that in Niyogi and Girosi (1996). The second term of (2.2) measures the approximation power of $\overline{\mathcal{H}}_K$ for ρ.

Definition 2.1. The *regularization error* of the system (K, ρ) is defined by

$$\widetilde{\mathcal{D}}(C) := \inf_{f \in \overline{\mathcal{H}}_K} \Big\{\mathcal{E}(f) - \mathcal{E}(f_c) + \frac{1}{2C}\|f^*\|_K^2\Big\}. \tag{2.4}$$

The regularization error for a *regularizing function* $f_{K,C} \in \overline{\mathcal{H}}_K$ is defined as

$$\mathcal{D}(C) := \mathcal{E}(f_{K,C}) - \mathcal{E}(f_c) + \frac{1}{2C}\|f_{K,C}^*\|_K^2. \tag{2.5}$$

In Wu and Zhou (2003) we showed that $\mathcal{E}(f) - \mathcal{E}(f_c) \le \|f - f_c\|_{L^1_{\rho_X}}$. Hence the regularization error can be estimated by the approximation in a weighted L^1 space, as done in Smale and Zhou (2003), and Chen et al. (2004).

Definition 2.2. We say that the probability measure ρ can be approximated by $\overline{\mathcal{H}}_K$ with exponent $0 < \beta \le 1$ if there exists a constant c_β such that

$$\widetilde{\mathcal{D}}(C) \le c_\beta C^{-\beta}, \qquad \forall C > 0. \tag{H1}$$

The first term of (2.2) is called the *sample error*. It has been well understood in learning theory by concentration inequalities, e.g. Vapnik (1998), Devroye et al. (1997), Niyogi (1998), Cucker and Smale (2001), Bousquet and Elisseeff (2002).

The approaches developed in Barron (1990), Bartlett (1998), Niyogi and Girosi (1996), and Zhang (2004) separate the regularization error and the sample error concerning $\widetilde{f}_{\mathbf{z}}$. In particular, for the QP-SVM, Zhang (2004) proved that

$$\mathbf{E}_{\mathbf{z}\in Z^m}\big\{\mathcal{E}(\widetilde{f}_{\mathbf{z}})\big\} \le \inf_{f\in\overline{\mathcal{H}}_K}\Big\{\mathcal{E}(f) + \frac{1}{2\widetilde{C}}\|f^*\|_K^2\Big\} + \frac{2\widetilde{C}}{m}. \tag{2.6}$$

It follows that $\mathbf{E}_{\mathbf{z}\in Z^m}\big\{\mathcal{E}(\widetilde{f}_{\mathbf{z}}) - \mathcal{E}(f_c)\big\} \le \widetilde{\mathcal{D}}(\widetilde{C}) + \frac{2\widetilde{C}}{m}$. When (H1) holds, Zhang's bound in connection with (2.1) yields $\mathbf{E}_{\mathbf{z}\in Z^m}\big\{\mathcal{R}(\widetilde{f}_{\mathbf{z}}) - \mathcal{R}(f_c)\big\} = O(\widetilde{C}^{-\beta}) + \frac{2\widetilde{C}}{m}$. This is similar to some well-known bounds for the neural network learning algorithms, see e.g. Theorem 3.1 in Niyogi and Girosi (1996). The best learning rate derived from (2.6) by choosing $\widetilde{C} = m^{1/(\beta+1)}$ is

$$\mathbf{E}_{\mathbf{z}\in Z^m}\big\{\mathcal{R}(\widetilde{f}_{\mathbf{z}}) - \mathcal{R}(f_c)\big\} = O\big(m^{-\alpha}\big), \qquad \alpha = \frac{\beta}{\beta+1}. \tag{2.7}$$

Observe that the sample error bound $\frac{2\widetilde{C}}{m}$ in (2.6) is independent of the kernel K or the distribution ρ. If some information about K or ρ is available, the sample error and hence the excess misclassification error can be improved.

The information we need about K is the capacity measured by covering numbers.

Definition 2.3. Let $\mathcal{F}$ be a subset of a metric space. For any $\varepsilon > 0$, the *covering number* $\mathcal{N}(\mathcal{F}, \varepsilon)$ is defined to be the minimal integer $\ell \in \mathbb{N}$ such that there exist ℓ balls with radius ε covering $\mathcal{F}$.

In this paper we only use the uniform covering number. Covering numbers measured by empirical distances are also used in the literature (van der Vaart and Wellner 1996). For comparisons, see Pontil (2003).

Let $\mathcal{B}_R = \{f \in \mathcal{H}_K : \|f\|_K \le R\}$. It is a subset of $C(X)$ and the covering number is well defined. We denote the covering number of the unit ball $\mathcal{B}_1$ as

$$\mathcal{N}(\varepsilon) := \mathcal{N}\big(\mathcal{B}_1, \varepsilon\big), \qquad \varepsilon > 0. \tag{2.8}$$

Definition 2.4. The RKHS $\mathcal{H}_K$ is said to have *logarithmic complexity exponent* $s \ge 1$ if there exists a constant $c_s > 0$ such that

$$\log \mathcal{N}(\varepsilon) \le c_s \big(\log(1/\varepsilon)\big)^s. \tag{H2}$$

It has *polynomial complexity exponent* $s > 0$ if there is some $c_s > 0$ such that

$$\log \mathcal{N}(\varepsilon) \le c_s \big(1/\varepsilon\big)^s. \tag{H2$'$}$$

The uniform covering number has been extensively studied in learning theory. In particular, we know that for the Gaussian kernel $K(x, y) = \exp\{-|x - y|^2/\sigma^2\}$ with $\sigma > 0$ on a bounded subset X of $\mathbb{R}^n$, (H2) holds with $s = n + 1$, see Zhou (2002); if K is C^r with $r > 0$ (Sobolev smoothness), then (H2$'$) is valid with $s = 2n/r$, see Zhou (2003).

The information we need about ρ is a Tsybakov noise condition (Tsybakov 2004).

Definition 2.5. Let $0 \le q \le \infty$. We say that ρ has *Tsybakov noise exponent* q if there exists a constant $c_q > 0$ such that

$$P_X\big(\{x \in X : |f_\rho(x)| \le c_q t\}\big) \le t^q. \tag{H3}$$

All distributions have at least noise exponent 0. Deterministic distributions (which satisfy $|f_\rho(x)| \equiv 1$) have the noise exponent $q = \infty$ with $c_\infty = 1$.

Using the above conditions about K and ρ, Scovel and Steinwart (2003) showed that when (H1), (H2′) and (H3) hold, for every $\epsilon > 0$ and every $\delta > 0$, with confidence $1 - \delta$,

$$\mathcal{R}(\widetilde{f}_{\mathbf{z}}) - \mathcal{R}(f_c) = O\big(m^{-\alpha}\big), \qquad \alpha = \frac{4\beta(q+1)}{(2q+sq+4)(1+\beta)} - \epsilon. \tag{2.9}$$

When no conditions are assumed for the distribution (i.e., $q = 0$) or $s = 2$ for the kernel (the worse case when empirical covering numbers are used, see van der Vaart and Wellner 1996), the rate is reduced to $\alpha = \frac{\beta}{\beta+1} - \epsilon$, arbitrarily close to Zhang's rate (2.7).

Recently, Wu et al. (2004) improve the rate (2.9) and show that under the same assumptions (H1), (H2′) and (H3), for every $\epsilon, \delta > 0$, with confidence $1 - \delta$,

$$\mathcal{R}(\widetilde{f}_{\mathbf{z}}) - \mathcal{R}(f_c) = O\big(m^{-\alpha}\big),\ \alpha = \min\Big\{\frac{\beta(q+1)}{\beta(q+2) + (q+1-\beta)s/2} - \epsilon, \frac{2\beta}{\beta+1}\Big\}. \tag{2.10}$$

When some condition is assumed for the kernel but not for the distribution, i.e., $s < 2$ but $q = 0$, the rate (2.10) has power $\alpha = \min\big\{\frac{\beta}{2\beta+(1-\beta)s/2} - \epsilon, \frac{2\beta}{\beta+1}\big\}$. This is better than (2.7) or (2.9) (or the rates given in Bartlett et al. 2003; Blanchard et al. 2004, see Chen et al. 2004; Wu et al. 2004 for detailed comparisons) if $\beta < 1$. This improvement is possible due to the projection operator.

Definition 2.6. The *projection operator* π is defined on the space of measurable functions $f : X \to \mathbb{R}$ as

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

The idea of projections appeared in margin-based bound analysis, e.g. Bartlett (1998), Lugosi and Vayatis (2004), Zhang (2002), Anthony and Bartlett (1999). We used the projection operator for the purpose of bounding misclassification and generalization errors in Chen et al. (2004). It helps

us to get sharper bounds of the sample error: probability inequalities are applied to random variables involving functions $\pi(\widetilde{f}_{\mathbf{z}})$ (bounded by 1), not to $\widetilde{f}_{\mathbf{z}}$ (the corresponding bound increases to infinity as C becomes large). In this paper we apply the projection operator to the LP-SVM.

Turn to our main goal, the LP-SVM classification algorithm $f_{\mathbf{z}}$ defined by (1.1). To our knowledge, the convergence of the algorithm has not been verified, even for distributions strictly separable by a universal kernel. What is the main difficulty in the error analysis?

One difficulty lies in the error decomposition: nothing like (2.2) exists for LP-SVM in the literature. Bounds for the regularization or approximation error independent of $\mathbf{z}$ are not available. We do not know whether it can be bounded by a norm in the whole space $\mathcal{H}_K$ or a norm similar to those in Niyogi and Girosi (1996).

In the paper we overcome the difficulty by means of a stepping stone from QP-SVM to LP-SVM. Then we can provide error analysis for general distributions. In particular, explicit learning rates will be presented. To this end, we first make an error decomposition.

Theorem 1. *Let $C > 0$, $0 < \eta \le 1$ and $f_{K,C} \in \overline{\mathcal{H}}_K$. There holds*

$$\mathcal{R}(f_{\mathbf{z}}) - \mathcal{R}(f_c) \le 2\eta\mathcal{R}(f_c) + \mathcal{S}(m, C, \eta) + 2\mathcal{D}(\eta C),$$

where $\mathcal{S}(m, C, \eta)$ is the sample error defined by

$$\mathcal{S}(m, C, \eta) := \Big\{\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}}))\Big\} + (1+\eta)\Big\{\mathcal{E}_{\mathbf{z}}\big(f_{K,C}\big) - \mathcal{E}\big(f_{K,C}\big)\Big\}. \quad (2.11)$$

Theorem 1 will be proved in Section 4. The term $\mathcal{D}(\eta C)$ is the regularization error (Smale and Zhou 2004) defined for a regularizing function $f_{K,C}$ (arbitrarily chosen) by (2.5). In Chen et al. (2004), we showed that

$$\mathcal{D}(C) \ge \widetilde{\mathcal{D}}(C) \ge \frac{\widetilde{\kappa}^2}{2C} \quad (2.12)$$

where

$$\widetilde{\kappa} := \mathcal{E}_0/(1+\kappa), \qquad \kappa = \sup_{x\in X}\sqrt{K(x,x)}, \qquad \mathcal{E}_0 := \inf_{b\in\mathbb{R}}\big\{\mathcal{E}(b) - \mathcal{E}(f_c)\big\}.$$

Also, $\widetilde{\kappa} = 0$ only for very special distributions. Hence the decay of $\mathcal{D}(C)$ cannot be faster than $O(1/C)$ in general. Thus, to have satisfactory convergence rates, C can not be too small, and it usually takes the form of m^τ for some $\tau > 0$. The constant κ is the norm of the inclusion $\mathcal{H}_K \subset C(X)$:

$$\|f\|_\infty \le \kappa \|f\|_K, \qquad \forall f \in \mathcal{H}_K. \tag{2.13}$$

Next we focus on analyzing the learning rates. Since a uniform rate is impossible for all probability distributions as shown in Theorem 7.2 of Devroye et al. (1997), we need to consider subclasses.

The choice of η is important in the upper bound in Theorem 1. If the distribution is deterministic, i.e., $\mathcal{R}(f_c) = 0$, we may choose $\eta = 1$. When $\mathcal{R}(f_c) > 0$, we must choose $\eta = \eta(m) \to 0$ as $m \to \infty$ in order to get the convergence rate. Of course the latter choice may lead to a slightly worse rate. Thus, we will consider these two cases separately.

The following proposition gives the bound for deterministic distributions.

Proposition 2.1. *Suppose $\mathcal{R}(f_c) = 0$. If $f_{K,C}$ is a function in $\overline{\mathcal{H}}_K$ satisfying $\|V(y, f_{K,C}(x))\|_\infty \le M$, then for every $0 < \delta < 1$, with confidence $1 - \delta$ there holds*

$$\mathcal{R}(f_{\mathbf{z}}) \le 32\varepsilon_{m,C} + \frac{20M\log(2/\delta)}{3m} + 8\mathcal{D}(C),$$

where with a constant c_s' depending on c_s, κ and s, $\varepsilon_{m,C}$ is given by

$$\begin{cases} \dfrac{22}{m}\left\{\log\dfrac{2}{\delta} + c_s'\Big[\log\Big(CM\log\dfrac{2}{\delta}\Big) + \log\Big(mC\mathcal{D}(C)\Big)\Big]^s\right\}, & \text{if (H2) holds;} \\ \dfrac{35\log\big(2/\delta\big)}{m}\Big(1 + (c_s')^{\frac{1}{1+s}}(CM)^{\frac{s}{1+s}}\Big) + \dfrac{32c_s'\,(C\mathcal{D}(C))^{\frac{s}{1+s}}}{3m^{1/(1+s)}}, & \text{if (H2}') \text{ holds.} \end{cases}$$

Proposition 2.1 will be proved in Section 6. As corollaries we obtain learning rates for strictly separable distributions and for weakly separable distributions.

Definition 2.7. We say that ρ is *strictly separable* by $\overline{\mathcal{H}}_K$ with margin $\gamma > 0$ if there is some function $f_\gamma \in \overline{\mathcal{H}}_K$ such that $\|f_\gamma^*\|_K = 1$ and $yf_\gamma(x) \ge \gamma$ almost everywhere.

For QP-SVM, the strictly separable case is well understood, see e.g. Vapnik (1998), Cristianini and Shawe-Taylor (2000) and vast references therein. For LP-SVM, we have

Corollary 2.1. *If ρ is strictly separable by $\overline{\mathcal{H}}_K$ with margin $\gamma > 0$ and (H2) holds, then*

$$\mathcal{R}(f_{\mathbf{z}}) \le \frac{704}{m}\left\{\log\frac{2}{\delta} + c_s'\Big(\log m + \log\frac{1}{\gamma^2}\Big)^s\right\} + \frac{4}{C\gamma^2}.$$

In particular, this will yield the learning rate $O\left(\frac{(\log m)^s}{m}\right)$ by taking $C = m/\gamma^2$.

Proof . Take $f_{K,C} = f_\gamma/\gamma$. Then $V(y, f_{K,C}(x)) \equiv 0$ and $\mathcal{D}(C)$ equals $\frac{1}{2C}\|f_\gamma^*/\gamma\|_K^2 = \frac{1}{2C\gamma^2}$. The conclusion follows from Proposition 2.1 by choosing $M = 0$. □

Remark 2.1. For strictly separable distributions, we verify the optimal rate when (H2) holds. Similar rates are true for more general kernels. But we omit details here.

Definition 2.8. We say that ρ is (weakly) separable by $\overline{\mathcal{H}}_K$ if there is some function $f_{\rm sp} \in \overline{\mathcal{H}}_K$, called the *separating function*, such that $\|f_{\rm sp}^*\|_K = 1$ and $yf_{\rm sp}(x) > 0$ almost everywhere. It has *separating exponent* $\theta \in (0, \infty]$ if for some $\gamma_\theta > 0$, there holds

$$\rho_X(0 < |f_{\rm sp}(x)| < \gamma_\theta t) \le t^\theta. \tag{2.14}$$

Corollary 2.2. *Suppose that ρ is separable by $\overline{\mathcal{H}}_K$ with* (2.14) *valid.*

(i) *If (H2) holds, then*

$$\mathcal{R}(f_{\mathbf{z}}) = O\left(\frac{(\log m + \log C)^s}{m} + C^{-\frac{\theta}{\theta+2}}\right).$$

This gives the learning rate $O(\frac{(\log m)^s}{m})$ by taking $C = m^{(\theta+2)/\theta}$.

(ii) *If (H2′) holds, then*

$$\mathcal{R}(f_{\mathbf{z}}) = O\left(\frac{C^{\frac{s}{1+s}}}{m} + \left(\frac{C^{\frac{2s}{\theta+2}}}{m}\right)^{\frac{1}{1+s}} + C^{-\frac{\theta}{\theta+2}}\right).$$

This yields the learning rate $O(m^{-\frac{\theta}{s\theta+2s+\theta}})$ *by taking* $C = m^{\frac{\theta+2}{s\theta+2s+\theta}}$.

Proof . Take $f_{K,C} = C^{\frac{1}{\theta+2}} f_{\mathrm{sp}}/\gamma_\theta$. By the definition of f_{sp}, we have $y f_{K,C}(x) \geq 0$ almost everywhere. Hence $0 \leq V(y, f_{K,C}(x)) \leq 1$. Moreover,

$$\mathcal{E}(f_{K,C}) = \int_X \left(1 - \frac{C^{\frac{1}{\theta+2}}}{\gamma_\theta}|f_{\mathrm{sp}}(x)|\right)_+ d\rho_X = \rho_X\left\{0 < |f_{\mathrm{sp}}(x)| < \gamma_\theta C^{-\frac{1}{\theta+2}}\right\}$$

which is bounded by $C^{-\frac{\theta}{\theta+2}}$. Therefore, $\mathcal{D}(C) \leq (1 + \frac{1}{2\gamma_\theta^2})C^{-\frac{\theta}{\theta+2}}$. Then the conclusion follows from Proposition 2.1 by choosing $M = 1$. □

Example. 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

$$f_\rho(x) = \begin{cases} -1, & \text{if} \quad -1/2 \leq x < 0, \\ 1, & \text{if} \quad 0 < x < 1/2. \end{cases}$$

If we take the linear kernel $K(x, y) = x \cdot y$, then $\theta = 1$, $\gamma_\theta = 1/2$. Since (H2) is satisfied with $s = 1$, the learning rate is $O(\frac{\log m}{m})$ by taking $C = m^3$.

Remark 2.2. The condition (2.14) with $\theta = \infty$ is exactly the definition of strictly separable distribution and γ_θ is the margin.

The choice of $f_{K,C}$ and the regularization error play essential roles to get our error bounds. It influences the strategy of choosing the regularization parameter (model selection) and determines learning rates. For weakly separable distributions we chose $f_{K,C}$ to be multiples of a separating function in Corollary 2.2. For the general case, it can be the choice (2.3).

Let's analyze learning rates for distributions having polynomially decaying regularization error, i.e., (H1) with $\beta \leq 1$. This is reasonable because of (2.12).

Theorem 2. *Suppose that* $\mathcal{R}(f_c) = 0$ *and the hypotheses (H1), (H2′) hold with* $0 < s < \infty$ *and* $0 < \beta \leq 1$, *respectively. Take* $C = m^\zeta$ *with* $\zeta := \min\{\frac{1}{s+\beta}, \frac{2}{1+\beta}\}$. *Then for every* $0 < \delta < 1$ *there exists a constant* $\widetilde{c}$ *depending on* s, β, δ *such that with confidence* $1 - \delta$,

$$\mathcal{R}(f_{\mathbf{z}}) \leq \widetilde{c} m^{-\alpha}, \qquad \alpha = \min\left\{\frac{2\beta}{1+\beta}, \frac{\beta}{s+\beta}\right\}.$$

Next we consider general distributions satisfying Tsybakov condition (Tsybakov 2004).

Theorem 3. *Assume the hypotheses (H1), (H2′) and (H3) with* $0 < s < \infty$, $0 < \beta \leq 1$, *and* $0 \leq q \leq \infty$. *Take* $C = m^\zeta$ *with*

$$\zeta := \min\left\{\frac{2}{\beta+1}, \frac{(q+1)(\beta+1)}{s(q+1)+\beta(q+2+qs+s)}\right\}.$$

For every $\epsilon > 0$ *and every* $0 < \delta < 1$ *there exists a constant* $\widetilde{c}$ *depending on* s, q, β, δ, *and* ϵ *such that with confidence* $1 - \delta$,

$$\mathcal{R}(f_{\mathbf{z}}) - \mathcal{R}(f_c) \leq \widetilde{c} m^{-\alpha}, \ \alpha = \min\left\{\frac{2\beta}{\beta+1}, \frac{\beta(q+1)}{s(q+1)+\beta(q+2+qs+s)} - \epsilon\right\}.$$

Remark 2.3. Since $\mathcal{R}(f_c)$ is usually small for a meaningful classification problem, the upper bound in Theorem 1 tells that the performance of LP-SVM is similar to that of QP-SVM. However, to have convergence rates, we need to choose $\eta = \eta(m) \to 0$ as m becomes large. This makes our rate worse than that of QP-SVM. This is the case when the capacity index s is large. When s is very small, the rate is $O\left(m^{-\alpha}\right)$ with α close to $\min\{\frac{q+1}{q+2}, \frac{2\beta}{\beta+1}\}$, which coincides to the rate (2.10), and is better than the rates (2.7) or (2.9) for QP-SVM. As any C^∞ kernel satisfies (H2′) for an arbitrarily small $s > 0$ (Zhou 2003), this is the case for polynomial or Gaussian kernels, usually used in practice.

Remark 2.4. Here we use a stepping stone from QP-SVM to LP-SVM. So the derived learning rates for the LP-SVM are essentially no worse than those of QP-SVM. It would be interesting to introduce different tools to get

learning rates for the LP-SVM, better than those of QP-SVM. Also, the choice of the trade-off parameter C in Theorem 3 depends on the indices β (approximation), s (capacity), and q (noise condition). This gives a rate which is optimal by our approach. One can take other choices $\zeta > 0$ (for $C = m^\zeta$), independent of β, s, q, and then derive learning rates according to the proof of Theorem 3. But the derived rates are worse than the one stated in Theorem 3. It would be of importance to give some methods for choosing C adaptively.

Remark 2.5. When empirical covering numbers are used, the capacity index can be restricted to $s \in [0, 2]$. Similar learning rates can be derived, as done in Blanchard et al. (2004), Wu et al. (2004).

3 Stepping Stone

Recall that in (1.7), the penalty term $\Omega(f^*)$ is usually not a norm. This makes the scheme difficult to analyze. Since the solution $f_{\mathbf{z}}$ of the LP-SVM has a representation similar to $\widetilde{f}_{\mathbf{z}}$ in QP-SVM, we expect close relations between these schemes. Hence the latter may play roles in the analysis for the former. To this end, we need to estimate $\Omega(\widetilde{f}^*_{\mathbf{z}})$, the l^1–norm of the coefficients of the solution $\widetilde{f}^*_{\mathbf{z}}$ to (1.4).

Lemma 3.1. *For every $\widetilde{C} > 0$, the function $\widetilde{f}_{\mathbf{z}}$ defined by* (1.3) *and* (1.4) *satisfies*

$$\Omega(\widetilde{f}^*_{\mathbf{z}}) = \sum_{i=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} \le \widetilde{C}\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z}}) + \|\widetilde{f}^*_{\mathbf{z}}\|_K^2.$$

Proof. The dual problem of the 1–norm soft margin SVM (Vapnik 1998) tells us that the coefficients $\widetilde{\alpha}_{i,\mathbf{z}}$ in the expression (1.4) of $\widetilde{f}_{\mathbf{z}}$ satisfy

$$0 \le \widetilde{\alpha}_{i,\mathbf{z}} \le \frac{\widetilde{C}}{m} \qquad \text{and} \qquad \sum_{i=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} y_i = 0. \tag{3.1}$$

The definition of the loss function V implies that $1 - y_i\widetilde{f}_{\mathbf{z}}(x_i) \le V\big(y_i, \widetilde{f}_{\mathbf{z}}(x_i)\big)$.

Then

$$\sum_{i=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} - \sum_{i=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} y_i \widetilde{f}_{\mathbf{z}}(x_i) \le \sum_{i=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} V\big(y_i, \widetilde{f}_{\mathbf{z}}(x_i)\big).$$

Applying the upper bound for $\widetilde{\alpha}_{i,\mathbf{z}}$ in (3.1), we can bound the right side above as

$$\sum_{i=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} V\big(y_i, \widetilde{f}_{\mathbf{z}}(x_i)\big) \le \frac{\widetilde{C}}{m} \sum_{i=1}^{m} V\big(y_i, \widetilde{f}_{\mathbf{z}}(x_i)\big) = \widetilde{C}\mathcal{E}_{\mathbf{z}}\big(\widetilde{f}_{\mathbf{z}}\big).$$

Applying the second relation in (3.1) yields

$$\sum_{i=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} y_i \widetilde{b}_{\mathbf{z}} = 0.$$

It follows that

$$\sum_{i=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} y_i \widetilde{f}_{\mathbf{z}}(x_i) = \sum_{i=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} y_i \big(\widetilde{f}^*_{\mathbf{z}}(x_i) + \widetilde{b}_{\mathbf{z}}\big) = \sum_{i=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} y_i \widetilde{f}^*_{\mathbf{z}}(x_i).$$

But $\widetilde{f}^*_{\mathbf{z}}(x_i) = \sum_{j=1}^{m} \widetilde{\alpha}_{j,\mathbf{z}} y_j K(x_i, x_j)$. We have

$$\sum_{i=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} y_i \widetilde{f}_{\mathbf{z}}(x_i) = \sum_{i,j=1}^{m} \widetilde{\alpha}_{i,\mathbf{z}} y_i \widetilde{\alpha}_{j,\mathbf{z}} y_j K(x_i, x_j) = \|\widetilde{f}^*_{\mathbf{z}}\|_K^2.$$

Hence the bound for $\Omega(\widetilde{f}^*_{\mathbf{z}})$ follows. □

Remark 3.1. Dr. Yiming Ying pointed out to us that actually the equality holds in Lemma 3.1. This follows from the KKT conditions. But we only need the inequality here.

4 Error Decomposition

In this section, we estimate $\mathcal{R}(f_{\mathbf{z}}) - \mathcal{R}(f_c)$.

Since $\operatorname{sgn}(\pi(f)) = \operatorname{sgn}(f)$, we have $\mathcal{R}(f) = \mathcal{R}(\pi(f))$. Using (2.1) to $\pi(f)$, we obtain

$$\mathcal{R}(f) - \mathcal{R}(f_c) = \mathcal{R}(\pi(f)) - \mathcal{R}(f_c) \le \mathcal{E}(\pi(f)) - \mathcal{E}(f_c). \tag{4.1}$$

It is easy to see that $V(y, \pi(f)(x)) \le V(y, f(x))$. Hence

$$\mathcal{E}(\pi(f)) \le \mathcal{E}(f) \quad \text{and} \quad \mathcal{E}_{\mathbf{z}}(\pi(f)) \le \mathcal{E}_{\mathbf{z}}(f). \tag{4.2}$$

We are in a position to prove Theorem 1 which, by (4.1), is an easy consequence of the following result.

Proposition 4.1. *Let $C > 0$, $0 < \eta \le 1$ and $f_{K,C} \in \overline{\mathcal{H}}_K$. Then*

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_c) + \frac{1}{C}\Omega(f_{\mathbf{z}}^*) \le 2\eta\mathcal{R}(f_c) + \mathcal{S}(m, C, \eta) + 2\mathcal{D}(\eta C),$$

where $\mathcal{S}(m, C, \eta)$ is defined by (2.11).

Proof. Take $\widetilde{f}_{\mathbf{z}}$ to be the solution of (1.4) with $\widetilde{C} = \eta C$.

We see from the definition of $f_{\mathbf{z}}$ and (4.2) that

$$\Big(\mathcal{E}_{\mathbf{z}}\big(\pi(f_{\mathbf{z}})\big) + \frac{1}{C}\Omega(f_{\mathbf{z}}^*)\Big) - \Big(\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z}}) + \frac{1}{C}\Omega(\widetilde{f}_{\mathbf{z}}^*)\Big) \le 0.$$

This enables us to decompose $\mathcal{E}((\pi(f_{\mathbf{z}})) + \frac{1}{C}\Omega(f_{\mathbf{z}}^*)$ as

$$\mathcal{E}(\pi(f_{\mathbf{z}})) + \frac{1}{C}\Omega(f_{\mathbf{z}}^*) \le \Big\{\mathcal{E}\big(\pi(f_{\mathbf{z}})\big) - \mathcal{E}_{\mathbf{z}}\big(\pi(f_{\mathbf{z}})\big)\Big\} + \Big(\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z}}) + \frac{1}{C}\Omega(\widetilde{f}_{\mathbf{z}}^*)\Big).$$

Lemma 3.1 gives $\Omega(\widetilde{f}_{\mathbf{z}}^*) \le \widetilde{C}\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z}}) + \|\widetilde{f}_{\mathbf{z}}^*\|_K^2$. But $\widetilde{C} = \eta C$. Hence

$$\mathcal{E}(\pi(f_{\mathbf{z}})) + \frac{1}{C}\Omega(f_{\mathbf{z}}^*) \le \Big\{\mathcal{E}\big(\pi(f_{\mathbf{z}})\big) - \mathcal{E}_{\mathbf{z}}\big(\pi(f_{\mathbf{z}})\big)\Big\} + (1+\eta)\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z}}) + \frac{1}{C}\|\widetilde{f}_{\mathbf{z}}^*\|_K^2.$$

Next we use the function $f_{K,C}$ to analyze the second term of the above bound and get

$$\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z}}) + \frac{1}{(1+\eta)C}\|\widetilde{f}_{\mathbf{z}}^*\|_K^2 \le \mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z}}) + \frac{1}{2\widetilde{C}}\|\widetilde{f}_{\mathbf{z}}^*\|_K^2 \le \mathcal{E}_{\mathbf{z}}(f_{K,C}) + \frac{1}{2\widetilde{C}}\|f_{K,C}^*\|_K^2.$$

This bound can be written as $\big\{\mathcal{E}_{\mathbf{z}}(f_{K,C}) - \mathcal{E}(f_{K,C})\big\} + \big\{\mathcal{E}(f_{K,C}) + \frac{1}{2\widetilde{C}}\|f_{K,C}^*\|_K^2\big\}$.

Combining the above two steps, we find that $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_c) + \dfrac{1}{C}\Omega(f_{\mathbf{z}}^*)$ is bounded by

$$\Big\{\mathcal{E}\big(\pi(f_{\mathbf{z}})\big) - \mathcal{E}_{\mathbf{z}}\big(\pi(f_{\mathbf{z}})\big)\Big\} + (1+\eta)\Big\{\mathcal{E}_{\mathbf{z}}\big(f_{K,C}\big) - \mathcal{E}\big(f_{K,C}\big)\Big\}$$
$$+(1+\eta)\Big\{\mathcal{E}\big(f_{K,C}\big) - \mathcal{E}\big(f_c\big) + \tfrac{1}{2\eta C}\|f_{K,C}^*\|_K^2\Big\} + \eta\mathcal{E}(f_c).$$

By the fact $\mathcal{E}(f_c) = 2\mathcal{R}(f_c)$ and the definition of $\mathcal{D}(C)$, we draw our conclusion. □

5 Probability Inequalities

In this section we give some probability inequalities. They modify the Bernstein inequality and extend our previous work in Chen et al. (2004) which was motivated by sample error estimates for the square loss (e.g. Barron 1990; Bartlett 1998; Cucker and Smale 2001, and Mendelson 2002). Recall the Bernstein inequality:

Let ξ be a random variable on Z with mean μ and variance σ^2. If $|\xi-\mu| \le M$, then

$$\text{Prob}\left\{\left|\mu - \frac{1}{m}\sum_{i=1}^m \xi(z_i)\right| > \varepsilon\right\} \le 2\exp\left\{-\frac{m\varepsilon^2}{2(\sigma^2 + \frac{1}{3}M\varepsilon)}\right\}.$$

The one-side Bernstein inequality holds without the leading factor 2.

Proposition 5.1. *Let ξ be a random variable on Z satisfying $\mu \ge 0$, $|\xi - \mu| \le M$ almost everywhere, and $\sigma^2 \le c\mu^\tau$ for some $0 \le \tau \le 2$. Then for every $\varepsilon > 0$ there holds*

$$\text{Prob}\left\{\frac{\mu - \frac{1}{m}\sum_{i=1}^m \xi(z_i)}{(\mu^\tau + \varepsilon^\tau)^{\frac{1}{2}}} > \varepsilon^{1-\frac{\tau}{2}}\right\} \le \exp\left\{-\frac{m\varepsilon^{2-\tau}}{2(c + \frac{1}{3}M\varepsilon^{1-\tau})}\right\}.$$

Proof. The one-side Bernstein inequality tells us that

$$\text{Prob}\left\{\frac{\mu - \frac{1}{m}\sum_{i=1}^m \xi(z_i)}{(\mu^\tau + \varepsilon^\tau)^{\frac{1}{2}}} > \varepsilon^{1-\frac{\tau}{2}}\right\} \le \exp\left\{-\frac{m\varepsilon^{2-\tau}(\mu^\tau + \varepsilon^\tau)}{2\big(\sigma^2 + \frac{1}{3}M\varepsilon^{1-\frac{\tau}{2}}(\mu^\tau + \varepsilon^\tau)^{\frac{1}{2}}\big)}\right\}.$$

Since $\sigma^2 \le c\mu^\tau$, we have

$$\sigma^2 + \frac{M}{3}\varepsilon^{1-\frac{\tau}{2}}(\mu^\tau+\varepsilon^\tau)^{\frac{1}{2}} \le c\mu^\tau + \frac{M}{3}\varepsilon^{1-\tau}(\mu^\tau+\varepsilon^\tau) \le (\mu^\tau+\varepsilon^\tau)\big(c+\frac{1}{3}M\varepsilon^{1-\tau}\big).$$

This yields the desired inequality. □

Note that $f_{\mathbf{z}}$ depends on $\mathbf{z}$ and thus runs over a set of functions as $\mathbf{z}$ changes. We need a probability inequality concerning the uniform convergence. Denote $\mathbf{E}\, g := \int_Z g(z)d\rho$.

Lemma 5.1. *Let $0 \le \tau \le 1$, $M > 0$, $c \ge 0$, and $\mathcal{G}$ be a set of functions on Z such that for every $g \in \mathcal{G}$, $\mathbf{E}\, g \ge 0$, $|g - \mathbf{E}\, g| \le M$ and $\mathbf{E}\, g^2 \le c(\mathbf{E}\, g)^\tau$. Then for $\varepsilon > 0$,*

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

Proof. Let $\{g_j\}_{j=1}^{\mathcal{N}} \subset \mathcal{G}$ with $\mathcal{N} = \mathcal{N}\big(\mathcal{G},\varepsilon\big)$ such that for every $g \in \mathcal{G}$ there is some $j \in \{1,\dots,\mathcal{N}\}$ satisfying $\|g - g_j\|_\infty \le \varepsilon$. Then by Proposition 5.1, a standard procedure (Cucker and Smale 2001; Mukherjee et al 2002; Chen et al. 2004) leads to the conclusion. □

Remark 5.1. Various forms of probability inequalities using empirical covering numbers can be found in the literature. For simplicity we give the current form in Lemma 5.1 which is enough for our purpose.

Let us find the hypothesis space covering $f_{\mathbf{z}}$ when $\mathbf{z}$ runs over all possible samples. This is implemented in the following two lemmas.

By the idea of bounding the offset from Wu and Zhou (2003), and Chen et al. (2004), we can prove the following.

Lemma 5.2. *For any $C > 0, m \in \mathbb{N}$ and $\mathbf{z} \in Z^m$, we can find a solution $f_{\mathbf{z}}$ of (1.7) satisfying $\min\limits_{1\le i\le m} |f_{\mathbf{z}}(x_i)| \le 1$. Hence $|b_{\mathbf{z}}| \le 1 + \|f_{\mathbf{z}}^*\|_\infty$.*

We shall always choose $f_{\mathbf{z}}$ as in Lemma 5.2. In fact, the only restriction we need to make for the minimizer $f_{\mathbf{z}}$ is to choose $\alpha_i = 0$ and $b_{\mathbf{z}} = y^*$, i.e., $f_{\mathbf{z}}(x) = y^*$ whenever $y_i = y^*$ for all $1 \le i \le m$ with some $y^* \in Y$.

Lemma 5.3. *For every $C > 0$, we have $f_{\mathbf{z}}^* \in \mathcal{H}_K$ and $\|f_{\mathbf{z}}^*\|_K \le \kappa\Omega(f_{\mathbf{z}}^*) \le \kappa C$.*

Proof. It is trivial that $f_{\mathbf{z}}^* \in \mathcal{H}_K$. By the reproducing property (1.5),

$$\|f_{\mathbf{z}}^*\|_K = \left(\sum_{i,j=1}^m \alpha_{i,\mathbf{z}}\alpha_{j,\mathbf{z}} y_i y_j K(x_i, x_j)\right)^{1/2} \le \kappa \left(\sum_{i,j=1}^m \alpha_{i,\mathbf{z}}\alpha_{j,\mathbf{z}}\right)^{1/2} = \kappa\Omega(f_{\mathbf{z}}^*).$$

Bounding the solution to (1.7) by the choice $f = 0 + 0$, we have $\mathcal{E}(f_{\mathbf{z}}) + \frac{1}{C}\Omega(f_{\mathbf{z}}^*) \le \mathcal{E}(0) + 0 = 1$. This gives $\Omega(f_{\mathbf{z}}^*) \le C$, and completes the proof. □

By Lemma 5.3 and Lemma 5.2 we know that $\pi(f_{\mathbf{z}})$ lies in

$$\mathcal{F}_R := \big\{\pi(f) : f \in \mathcal{B}_R + \big[-(1+\kappa R), 1+\kappa R\big]\big\} \tag{5.1}$$

with $R = \kappa C$. The following lemma (Chen et al. 2004) gives the covering number estimate for $\mathcal{F}_R$.

Lemma 5.4. *Let $\mathcal{F}_R$ be given by (5.1) with $R > 0$. For any $\varepsilon > 0$ there holds*

$$\mathcal{N}(\mathcal{F}_R, \varepsilon) \le \left(\frac{2(1+\kappa R)}{\varepsilon} + 1\right)\mathcal{N}\left(\frac{\varepsilon}{2R}\right).$$

Using the function set $\mathcal{F}_R$ defined by (5.1), we set for $R > 0$,

$$\mathcal{G}_R = \Big\{V\big(y, f(x)\big) - V(y, f_c(x)) : f \in \mathcal{F}_R\Big\}. \tag{5.2}$$

By Lemma 5.4 and the additive property of the log function, we have

Lemma 5.5. *Let $\mathcal{G}_R$ given by (5.2) with $R > 0$.*

(i) *If (H2) holds, then there exists a constant $c_s' > 0$ such that*

$$\log\mathcal{N}(\mathcal{G}_R, \varepsilon) \le c_s'\Big(\log\frac{R}{\varepsilon}\Big)^s.$$

(ii) *If (H2′) holds, then there exists a constant $c_s' > 0$ such that*

$$\log\mathcal{N}(\mathcal{G}_R, \varepsilon) \le c_s'\Big(\frac{R}{\varepsilon}\Big)^s.$$

The following lemma was proved by Scovel and Steinwart (2003) for general functions $f : X \to \mathbb{R}$. With the projection, here f has range $[-1, 1]$ and a simpler proof is given.

Lemma 5.6. *Assume (H3). For every function $f : X \to [-1, 1]$ there holds*

$$\mathbf{E}\left\{ \Big(V(y, f(x)) - V(y, f_c(x))\Big)^2 \right\} \le 8\big(\frac{1}{2c_q}\big)^{q/(q+1)} \Big(\mathcal{E}(f) - \mathcal{E}(f_c)\Big)^{\frac{q}{q+1}}.$$

Proof. Since $f(x) \in [-1, 1]$, we have $V(y, f(x)) - V(y, f_c(x)) = y(f_c(x) - f(x))$. It follows that

$$\mathcal{E}(f) - \mathcal{E}(f_c) = \int_X (f_c(x) - f(x)) f_\rho(x) d\rho_X = \int_X |f_c(x) - f(x)| \ |f_\rho(x)| d\rho_X$$

and

$$\mathbf{E}\left\{ \Big(V(y, f(x)) - V(y, f_c(x))\Big)^2 \right\} = \int_X |f_c(x) - f(x)|^2 d\rho_X.$$

Let $t > 0$ and separate the domain X into two sets: $X_t^+ := \{x \in X : |f_\rho(x)| > c_q t\}$ and $X_t^- := \{x \in X : |f_\rho(x)| \le c_q t\}$. On X_t^+ we have $|f_c(x) - f(x)|^2 \le 2|f_c(x) - f(x)| \frac{|f_\rho(x)|}{c_q t}$. On X_t^- we have $|f_c(x) - f(x)|^2 \le 4$. It follows from Assumption (H3) that

$$\int_X |f_c(x) - f(x)|^2 d\rho_X \le \frac{2\big(\mathcal{E}(f) - \mathcal{E}(f_c)\big)}{c_q t} + 4\rho_X(X_t^-) \le \frac{2\big(\mathcal{E}(f) - \mathcal{E}(f_c)\big)}{c_q t} + 4t^q.$$

Choosing $t = \big\{(\mathcal{E}(f) - \mathcal{E}(f_c))/(2c_q)\big\}^{1/(q+1)}$ yields the desired bound. □

Take the function set $\mathcal{G}$ in Lemma 5.1 to be $\mathcal{G}_R$. Then a function g in $\mathcal{G}_R$ takes the form $g(x, y) = V\big(y, \pi(f)(x)\big) - V(y, f_c(x))$ with $\pi(f) \in \mathcal{F}_R$. Obviously we have $\|g\|_\infty \le 2$, $\mathbf{E}\, g = \mathcal{E}(\pi(f)) - \mathcal{E}(f_c)$ and $\frac{1}{m}\sum_{i=1}^m g(z_i) = \mathcal{E}_{\mathbf{z}}(\pi(f)) - \mathcal{E}_{\mathbf{z}}(f_c)$. When Assumption (H3) is valid, Lemma 5.6 tells us that $\mathbf{E}\, g^2 \le c\big(\mathbf{E}\, g\big)^\tau$ with $\tau = \frac{q}{q+1}$ and $c = 8\big(\frac{1}{2c_q}\big)^{q/(q+1)}$. Applying Lemma 5.1 and solving the equation

$$\log \mathcal{N}\big(\mathcal{G}_R, \varepsilon\big) - \frac{m\varepsilon^{2-\tau}}{2(c + \frac{1}{3} \cdot 2\varepsilon^{1-\tau})} = \log \delta,$$

we see the following corollary from Lemma 5.5 and Lemma 5.6.

Corollary 5.1. *Let $\mathcal{G}_R$ be defined by (5.2) with $R > 0$ and (H3) hold with $0 \le q \le \infty$. For every $0 < \delta < 1$, with confidence at least $1-\delta$, there holds*

$$\Big\{\mathcal{E}(f) - \mathcal{E}(f_c)\Big\} - \Big\{\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_c)\Big\} \le 4\varepsilon_{m,R} + 4\varepsilon_{m,R}^{\frac{q+2}{2(q+1)}} \Big\{\mathcal{E}(f) - \mathcal{E}(f_c)\Big\}^{\frac{q}{2(q+1)}}$$

for all $f \in \mathcal{F}_R$, where $\varepsilon_{m,R}$ is given by

$$\begin{cases} 5\Big(8\big(\frac{1}{2c_q}\big)^{q/(q+1)} + \frac{1}{3}\Big)\left(\dfrac{\log\frac{1}{\delta} + c_s'(\log R + \log m)^s}{m}\right)^{\frac{q+1}{q+2}}, & \text{if (H2) holds,} \\ 8\Big(8\big(\frac{1}{2c_q}\big)^{q/(q+1)} + \frac{1}{3}\Big)\left(\Big(\dfrac{c_s' R^s}{m}\Big)^{\frac{(q+1)}{q+2+qs+s}} + \Big(\dfrac{\log\frac{1}{\delta}}{m}\Big)^{\frac{q+1}{q+2}}\right), & \text{if (H2}') \text{ holds.} \end{cases}$$

6 Rate Analysis

Let us now prove the main results stated in Section 2. We first prove Proposition 2.1.

Proof of Proposition 2.1. Since $\mathcal{R}(f_c) = 0$, $V(y, f_c(x)) = 0$ almost everywhere and $\mathcal{E}(f_c) = 0$. Take $\eta = 1$ in Proposition 4.1.

We first consider the random variable $\xi = V(y, f_{K,C}(x))$. Since $0 \le \xi \le M$ and $\mathbf{E}\,\xi = \mathcal{E}(f_{K,C}) \le \mathcal{D}(C)$, we have

$$\sigma^2(\xi) \le \mathbf{E}\,\xi^2 \le M\,\mathbf{E}\,\xi \le M\mathcal{D}(C).$$

Applying the one-side Bernstein inequality to ξ, we see by solving the quadratic equation $-\frac{m\varepsilon^2}{2(\sigma^2 + M\varepsilon/3)} = \log\big(\delta/2\big)$ that with probability $1-\delta/2$,

$$\mathcal{E}_{\mathbf{z}}(f_{K,C}) - \mathcal{E}(f_{K,C}) \le \frac{2M\log\big(\frac{2}{\delta}\big)}{3m} + \sqrt{\frac{2\sigma^2(\xi)\log\big(2/\delta\big)}{m}} \le \frac{5M\log\big(\frac{2}{\delta}\big)}{3m} + \mathcal{D}(C). \tag{6.1}$$

Next we estimate $\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}}))$. By the definition of $f_{\mathbf{z}}$, there holds

$$\frac{1}{C}\Omega(f_{\mathbf{z}}^*) \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \frac{1}{C}\Omega(f_{\mathbf{z}}^*) \le \mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z}}) + \frac{1}{C}\Omega(\widetilde{f}_{\mathbf{z}}^*).$$

According to Lemma 3.1, this is bounded by $2\Big(\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z}}) + \frac{1}{2C}\|\widetilde{f}^*_{\mathbf{z}}\|^2_K\Big)$. This in connection with the definition of $\widetilde{f}_{\mathbf{z}}$ yields

$$\frac{1}{C}\Omega(f^*_{\mathbf{z}}) \le 2\Big(\mathcal{E}_{\mathbf{z}}(\widetilde{f}_{\mathbf{z}}) + \frac{1}{2C}\|\widetilde{f}^*_{\mathbf{z}}\|^2_K\Big) \le 2\Big(\mathcal{E}_{\mathbf{z}}(f_{K,C}) + \frac{1}{2C}\|f^*_{K,C}\|^2_K\Big).$$

Since $\mathcal{E}(f_c) = 0$, $\mathcal{D}(C) = \mathcal{E}(f_{K,C}) + \frac{1}{2C}\|f^*_{K,C}\|^2_K$. It follows that

$$\frac{1}{C}\Omega(f^*_{\mathbf{z}}) \le 2\Big(\mathcal{E}_{\mathbf{z}}(f_{K,C}) - \mathcal{E}(f_{K,C}) + \mathcal{D}(C)\Big).$$

Together with Lemma 5.3 and (6.1), this tells us that with probability $1-\delta/2$

$$\|f^*_{\mathbf{z}}\|_K \le \kappa\Omega(f^*_{\mathbf{z}}) \le R := 2\kappa C\Big(\frac{5M\log\big(2/\delta\big)}{3m} + 2\mathcal{D}(C)\Big).$$

As we are considering a deterministic case, (H3) holds with $q = \infty$ and $c_\infty = 1$. Recall the definition of $\mathcal{G}_R$ in (5.2). Corollary 5.1 with $q = \infty$ and R given as above implies that

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) \le 4\varepsilon_{m,C} + 4\sqrt{\varepsilon_{m,C}}\sqrt{\mathcal{E}(\pi(f_{\mathbf{z}}))}$$

with confidence $1-\delta$ where $\varepsilon_{m,C}$ is defined in the statement.

Putting the above two estimates into Proposition 4.1, we have with confidence $1-\delta$,

$$\mathcal{E}(\pi(f_{\mathbf{z}})) \le 4\varepsilon_{m,C} + 4\sqrt{\varepsilon_{m,C}}\sqrt{\mathcal{E}(\pi(f_{\mathbf{z}}))} + \frac{10M\log\big(2/\delta\big)}{3m} + 4\mathcal{D}(C).$$

Solving the quadratic inequality for $\sqrt{\mathcal{E}(\pi(f_{\mathbf{z}}))}$ leads to

$$\mathcal{E}(\pi(f_{\mathbf{z}})) \le 32\varepsilon_{m,C} + \frac{20M\log\big(2/\delta\big)}{3m} + 8\mathcal{D}(C).$$

Then our conclusion follows from (4.1). □

Finally, we turn to the proof of Theorems 2 and 3. To this end, we need a bound for $\|\widetilde{f}^*_{K,C}\|_K$. According to the definition, $\frac{1}{2C}\|\widetilde{f}^*_{K,C}\|^2_K \le \widetilde{\mathcal{D}}(C)$. Then we have

Lemma 6.1. *For every $C > 0$, there hold*

$$\|\widetilde{f}^*_{K,C}\|_K \le \Big(2C\widetilde{\mathcal{D}}(C)\Big)^{1/2} \quad \text{and} \quad \|\widetilde{f}_{K,C}\|_\infty \le 1 + 2\kappa\Big(2C\widetilde{\mathcal{D}}(C)\Big)^{1/2}.$$

Proof of Theorem 2. Take $f_{K,C} = \widetilde{f}_{K,C}$ in Proposition 4.1. Then by Lemma 6.1 we may take $M = 2 + 2\kappa\Big(2C\widetilde{\mathcal{D}}(C)\Big)^{1/2}$. Proposition 2.1 with Assumption (H2′) yields

$$\mathcal{R}(f_{\mathbf{z}}) \le c_{s,\beta,\delta}\Big\{\frac{C^{(1-\beta)s/(s+1)}}{m^{\frac{1}{1+s}}} + \frac{C^{(1-\beta)s/(s+1)}}{m^{\frac{1}{1+s}}}\big(\frac{C^{(1+\beta)/2}}{m}\big)^{\frac{s}{1+s}} + \frac{C^{\frac{1-\beta}{2}}}{m} + C^{-\beta}\Big\}.$$

Take $C = \min\{m^{\frac{1}{s+\beta}}, m^{\frac{2}{1+\beta}}\}$. Then $\frac{C^{(1+\beta)/2}}{m} \le 1$ and the proof is complete. □

Proof of Theorem 3. Denote $\Delta_{\mathbf{z}} = \mathcal{E}\big(\pi(f_{\mathbf{z}})\big) - \mathcal{E}(f_c) + \frac{1}{C}\Omega(f^*_{\mathbf{z}})$. Then we have $\Omega(f^*_{\mathbf{z}}) \le C\Delta_{\mathbf{z}}$. This in connection with Lemma 5.3 yields

$$\|f^*_{\mathbf{z}}\|_K \le \kappa\Omega(f^*_{\mathbf{z}}) \le \kappa C\Delta_{\mathbf{z}}. \tag{6.2}$$

Take $f_{K,C} = \widetilde{f}_{K,\widetilde{C}}$ with $\widetilde{C} = \eta C$ in Proposition 4.1. It tells us that

$$\Delta_{\mathbf{z}} \le 2\eta\mathcal{R}(f_c) + \mathcal{S}(m, C, \eta) + 2\widetilde{\mathcal{D}}(\eta C).$$

Set $\eta = C^{-\beta/(\beta+1)}$. Then $\widetilde{C} = \eta C = C^{1/(\beta+1)}$. By the fact $\mathcal{R}(f_c) \le \frac{1}{2}$ and Assumption (H1),

$$\Delta_{\mathbf{z}} \le \mathcal{S}(m, C, \eta) + (1 + 2c_\beta)C^{-\frac{\beta}{\beta+1}}. \tag{6.3}$$

Recall the expression (2.11) for $\mathcal{S}(m, C, \eta)$. Here $f_{K,C} = \widetilde{f}_{K,\widetilde{C}}$. So we have

$$\begin{aligned}\mathcal{S}(m, C, \eta) = \ & \Big\{\Big(\mathcal{E}\big(\pi(f_{\mathbf{z}})\big) - \mathcal{E}(f_c)\Big) - \Big(\mathcal{E}_{\mathbf{z}}\big(\pi(f_{\mathbf{z}})\big) - \mathcal{E}_{\mathbf{z}}(f_c)\Big)\Big\} \\ & +(1+\eta)\Big\{\Big(\mathcal{E}_{\mathbf{z}}\big(\widetilde{f}_{K,\widetilde{C}}\big) - \mathcal{E}_{\mathbf{z}}(f_c)\Big) - \Big(\mathcal{E}\big(\widetilde{f}_{K,\widetilde{C}}\big) - \mathcal{E}(f_c)\Big)\Big\} \\ & +\eta\Big\{\mathcal{E}_{\mathbf{z}}(f_c) - \mathcal{E}(f_c)\Big\} =: \mathcal{S}_1 + (1+\eta)\mathcal{S}_2 + \eta\mathcal{S}_3.\end{aligned}$$

Take $t \ge 1, C \ge 1$ to be determined later. For $R \ge 1$, denote

$$\mathcal{W}(R) := \{\mathbf{z} \in Z^m : \|f^*_{\mathbf{z}}\|_K \le R\}. \tag{6.4}$$

For $\mathcal{S}_1$, we apply Corollary 5.1 with $\delta = e^{-t} \le 1/e$. We know that there is a set $V_R^{(1)} \subset Z^m$ of measure at most $\delta = e^{-t}$ such that

$$\mathcal{S}_1 \le c_{s,q} t \left\{ \Big(\frac{R^s}{m}\Big)^{\frac{q+1}{q+2+qs+s}} + \Big(\frac{R^s}{m}\Big)^{\frac{q+1}{q+2+qs+s}\cdot\frac{q+2}{2(q+1)}} \Delta_{\mathbf{z}}^{\frac{q}{2(q+1)}} \right\}, \quad \forall \mathbf{z} \in \mathcal{W}(R) \setminus V_R^{(1)}.$$

Here $c_{s,q} := 32\big(8\big(\frac{1}{2c_q}\big)^{q/(q+1)} + \frac{1}{3}\big)(c'_s + 1) \ge 1$ is a constant depending only on q and s.

To estimate $\mathcal{S}_2$, consider $\xi = V\big(y, \widetilde{f}_{K,\widetilde{C}}(x)\big) - V(y, f_c(x))$ on (Z, ρ). By Lemma 6.1, we have

$$\|\widetilde{f}_{K,\widetilde{C}}\|_\infty \le 1 + 2\kappa\sqrt{2\widetilde{C}\widetilde{\mathcal{D}}(\widetilde{C})} \le 1 + 2\kappa\sqrt{2c_\beta} C^{\frac{1-\beta}{2(\beta+1)}}.$$

Write $\xi = \xi_1 + \xi_2$ where

$$\xi_1 := V\big(y, \widetilde{f}_{K,\widetilde{C}}(x)\big) - V\big(y, \pi(\widetilde{f}_{K,\widetilde{C}})(x)\big),\ \xi_2 := V\big(y, \pi(\widetilde{f}_{K,\widetilde{C}})(x)\big) - V(y, f_c(x)).$$

It is easy to check that $0 \le \xi_1 \le 2\kappa\sqrt{2c_\beta} C^{\frac{1-\beta}{2(\beta+1)}}$. Hence $\sigma^2(\xi_1)$ is bounded by $2\kappa\sqrt{2c_\beta} C^{\frac{1-\beta}{2(\beta+1)}} \mathbf{E}\,\xi_1$. Then the one-side Bernstein inequality with $\delta = e^{-t}$ tells us that there is a set $V^{(2)} \subset Z^m$ of measure at most $\delta = e^{-t}$ such that for every $\mathbf{z} \in Z^m \setminus V^{(2)}$, there holds

$$\frac{1}{m}\sum_{i=1}^m \xi_1(z_i) - \mathbf{E}\,\xi_1 \le \frac{4\kappa\sqrt{2c_\beta} C^{\frac{1-\beta}{2(\beta+1)}} t}{3m} + \sqrt{\frac{2\sigma^2(\xi_1)t}{m}} \le \frac{10\kappa\sqrt{2c_\beta} C^{\frac{1-\beta}{2(\beta+1)}} t}{3m} + \mathbf{E}\,\xi_1.$$

For ξ_2, by Lemma 5.6,

$$\sigma^2(\xi_2) \le 8\big(\frac{1}{2c_q}\big)^{q/(q+1)} (\mathbf{E}\,\xi_2)^{\frac{q}{q+1}}.$$

But $|\xi_2| \le 2$. So the one-side Bernstein inequality tells us again that there is a set $V^{(3)} \subset Z^m$ of measure at most $\delta = e^{-t}$ such that for every $\mathbf{z} \in Z^m \setminus V^{(3)}$, there holds

$$\frac{1}{m}\sum_{i=1}^m \xi_1(z_i) - \mathbf{E}\,\xi_1 \le \frac{4t}{3m} + \sqrt{\frac{4\sigma^2(\xi_1)t}{m}} \le \frac{4t}{3m} + 32\big(\frac{1}{2c_q}\big)^{\frac{q}{q+2}} \Big(\frac{t}{m}\Big)^{\frac{q+1}{q+2}} + \mathbf{E}\,\xi_2.$$

Here we have used the following elementary inequality with $b := (\mathbf{E}\,\xi_2)^{\frac{q}{2q+2}}$ and $a := \big(32(\frac{1}{2c_q})^{q/(q+1)}t/m\big)^{1/2}$:

$$a \cdot b \le \frac{q+2}{2q+2}a^{(2q+2)/(q+2)} + \frac{q}{2q+2}b^{(2q+2)/q}, \qquad \forall a, b > 0.$$

Combing the two estimates for ξ_1, ξ_2 with the fact that $\mathbf{E}\,\xi = \mathbf{E}\,\xi_1 + \mathbf{E}\,\xi_2 = \mathcal{E}\big(\widetilde{f}_{K,\widetilde{C}}\big) - \mathcal{E}(f_c) \le \widetilde{\mathcal{D}}(\widetilde{C}) \le c_\beta C^{-\beta/(\beta+1)}$ we see that

$$\mathcal{S}_2 \le c_{q,\beta}t\left(\frac{C^{\frac{1-\beta}{2(\beta+1)}}}{m} + \Big(\frac{1}{m}\Big)^{\frac{q+1}{q+2}} + C^{\frac{-\beta}{\beta+1}}\right), \qquad \forall \mathbf{z} \in Z^m \setminus V_R^{(2)} \setminus V_R^{(3)},$$

where $c_{q,\beta} := 10\kappa\sqrt{2c_\beta}/3 + \frac{4}{3} + 32(\frac{1}{2c_q})^{q/(q+1)} + c_\beta$ is a constant depending on q, β.

The last term is $\mathcal{S}_3 \le 1$.

Putting the above three estimates for $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ to (6.3), we find that for every $\mathbf{z} \in \mathcal{W}(R) \setminus V_R^{(1)} \setminus V^{(2)} \setminus V^{(3)}$ there holds

$$\Delta_{\mathbf{z}} \le 2c_{s,q}t\Big(\frac{R^s}{m}\Big)^{\frac{(q+1)}{q+2+qs+s}} + 8c_{q,\beta}t\left\{\Big(\frac{1}{m}\Big)^{\frac{q+1}{q+2}} + C^{-\frac{\beta}{\beta+1}}\Big(\frac{C^{1/2}}{m} + 1\Big)\right\}. \quad (6.5)$$

Here we have used another elementary inequality for $\alpha = q/(2q+2) \in (0,1)$ and $x = \Delta_{\mathbf{z}}$:

$$x \le ax^\alpha + b, \quad a, b, x > 0 \Longrightarrow x \le \max\{(2a)^{1/(1-\alpha)}, 2b\}.$$

Now we can choose C to be

$$C := \min\left\{m^2,\, m^{\frac{(q+1)(\beta+1)}{s(q+1)+\beta(q+2+qs+s)}}\right\}. \quad (6.6)$$

It ensures that $\big(\frac{1}{m}\big)^{\frac{q+1}{q+2}} \le C^{-\frac{\beta}{\beta+1}}$ and $\big(\frac{1}{m}\big)^{\frac{(q+1)}{q+2+qs+s}} \le C^{-\frac{s(q+1)+\beta(q+2+qs+s)}{(\beta+1)(q+2+qs+s)}}$. With this choice of C, (6.5) implies that with a set $V_R := V_R^{(1)} \cup V_R^{(2)} \cup V_R^{(3)}$ of measure at most $3e^{-t}$,

$$\Delta_{\mathbf{z}} \le C^{-\frac{\beta}{\beta+1}}\left\{2c_{s,q}t\Big(C^{-\frac{1}{\beta+1}}R\Big)^{\frac{s(q+1)}{q+2+qs+s}} + 24c_{q,\beta}t\right\}, \quad \forall \mathbf{z} \in \mathcal{W}(R) \setminus V_R. \quad (6.7)$$

We shall finish our proof by using (6.2) and (6.7) iteratively.

Start with the bound $R = R^{(0)} := \kappa C$. Lemma 5.3 verifies $\mathcal{W}(R^{(0)}) = Z^m$. At this first step, by (6.7) and (6.2) we have $Z^m = \mathcal{W}(R^{(0)}) \subseteq \mathcal{W}(R^{(1)}) \cup V_{R^{(0)}}$, where

$$R^{(1)} := \kappa C^{\frac{1}{\beta+1}} \Big\{ \left(2c_{s,q}t(\kappa+1)\right) C^{\frac{\beta}{\beta+1}\cdot\frac{s(q+1)}{q+2+qs+s}} + 24c_{q,\beta}t \Big\}.$$

Now we iterate. For $n = 2, 3, \cdots$, we derive from (6.7) and (6.2) that

$$Z^m = \mathcal{W}(R^{(0)}) \subseteq \mathcal{W}(R^{(1)}) \cup V_{R^{(0)}} \subseteq \cdots \subseteq \mathcal{W}(R^{(n)}) \cup \big(\cup_{j=0}^{n-1} V_{R^{(j)}}\big),$$

where each set $V_{R^{(j)}}$ has measure at most $3e^{-t}$ and the number $R^{(n)}$ is given by

$$R^{(n)} = \kappa C^{\frac{1}{\beta+1}} \Big\{ \left(2c_{s,q}t(\kappa+1)\right)^n C^{\frac{\beta}{\beta+1}\cdot\left(\frac{s(q+1)}{q+2+qs+s}\right)^n} + 24c_{q,\beta}t(\kappa+1)n \Big\}.$$

Note that $\epsilon > 0$ is fixed. We choose $n_0 \in \mathbb{N}$ to be large enough such that

$$\left(\frac{s(q+1)}{q+2+qs+s}\right)^{(n_0+1)} \leq \epsilon \left(s + \frac{2s}{\beta} + \frac{q+2}{q+1}\right).$$

In the n_0-th step of our iteration we have shown that for $\mathbf{z} \in \mathcal{W}(R^{(n_0)})$,

$$\|f_{\mathbf{z}}^*\|_K \leq \kappa C^{\frac{1}{\beta+1}} \Big\{ \left(2c_{s,q}t(\kappa+1)\right)^{n_0} C^{\frac{\beta}{\beta+1}\cdot\left(\frac{s(q+1)}{q+2+qs+s}\right)^{n_0}} + 24c_{q,\beta}t(\kappa+1)n_0 \Big\}.$$

This together with (6.5) gives

$$\Delta_{\mathbf{z}} \leq c(s,q,\beta,\epsilon) t^{n_0} \max\left\{ m^{-\frac{2\beta}{\beta+1}},\, m^{-\frac{\beta(q+1)}{s(q+1)+\beta(s+2+qs+s)}+\epsilon} \right\}.$$

This is true for $\mathbf{z} \in \mathcal{W}(R^{(n_0)}) \setminus V_{R^{(n_0)}}$. Since the set $\cup_{j=0}^{n_0} V_{R^{(j)}}$ has measure at most $3(n_0+1)e^{-t}$, we know that the set $\mathcal{W}(R^{(n_0)}) \setminus V_{R^{(n_0)}}$ has measure at least $1 - 3(n_0+1)e^{-t}$. Note that $\mathcal{E}\big(\pi(f_{\mathbf{z}})\big) - \mathcal{E}(f_c) \leq \Delta_{\mathbf{z}}$. Take $t = \log(\frac{3(n_0+1)}{\delta})$. Then the proof is finished by (4.1). □

Acknowledgments

This work is partially supported by the Research Grants Council of Hong Kong [Project No. CityU 103704] and by City University of Hong Kong [Project No. 7001442]. The corresponding author is Ding-Xuan Zhou.

References

Anthony, M., and Bartlett, P. L. (1999). Neural Network Learning: Theoretical Foundations. Cambridge University Press.

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

Barron, A. R. (1990). Complexity regularization with applications to artificial neural networks. In *Nonparametric Functional Estimation* (G. Roussa, ed.), 561–576. Dortrecht: Kluwer.

Bartlett, P. L. (1998). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. IEEE Trans. Inform. Theory, 44, 525–536.

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

Blanchard, B., Bousquet, O., and Massart, P. (2004). Statistical performance of support vector machines. Preprint.

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

Bousquet, O., and Elisseeff, A. (2002). Stability and generalization. J. Machine Learning Research, 2, 499–526.

Bradley, P. S., and Mangasarian, O. L. (2000). Massive data discrimination via linear support vector machines. Optimization Methods and Software, 13, 1–10.

Chen, D. R., Wu, Q., Ying, Y., and Zhou, D. X. (2004). Support vector machine soft margin classifiers: error analysis. J. Machine Learning Research, 5, 1143–1175.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learning, 20, 273–297.

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

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

Devroye, L., Györfi, L., and Lugosi, G. (1997). A probabilistic Theory of Pattern Recognition. New York: Springer-Verlag.

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

Kecman, V., and Hadzic, I. (2000). Support vector selection by linear programming. Proc. of IJCNN, 5, 193–198.

Lugosi, G., and Vayatis, N. (2004). On the Bayes-risk consistency of regularized bossting methods. Ann. Statis., 32, 30–55.

Mendelson, S. (2002). Improving the sample complexity using global data. IEEE Trans. Inform. Theory, 48, 1977–1991.

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

Niyogi, P. (1998). The Informational Complexity of Learning. Kluwer.

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

Pedroso, J. P., and Murata, N. (2001). Support vector machines with different norms: motivation, formulations and results. Pattern recognition Letters, 22, 1263–1272.

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

Rosasco, L., De Vito, E., Caponnetto, A., Piana, M., and Verri, A. (2004). Are loss functions all the same? Neural Comp., 16, 1063–1076.

Scovel, C., and Steinwart, I. (2003). Fast rates for support vector machines. Preprint.

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

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

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

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. Ann. Statis., 32, 135–166.

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

Vapnik, V. (1998). Statistical Learning Theory. John Wiley & Sons.

Wahba, G. (1990). Spline Models for Observational Data. SIAM.

Wu, Q., Ying, Y., and Zhou, D. X. (2004). Multi-kernel regularized classifiers. Preprint.

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

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. Ann. Statis., 32, 56–85.

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. J. Machine Learning Research, 2, 527–550.

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

Zhou, D. X. (2003). Capacity of reproducing kernel spaces in learning theory. IEEE Trans. Inform. Theory, 49, 1743–1752.