Journal of Machine Learning Research 10 (2009) 1447-1468 Submitted 7/07; Revised 3/09; Published 7/09

Classification with Gaussians and Convex Loss

Dao-Hong Xiang DAOHONGXIANG@GMAIL.COM
Ding-Xuan Zhou MAZHOU@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, China

Editor: John Shawe-Taylor

Abstract

This paper considers binary classification algorithms generated from Tikhonov regularization schemes associated with general convex loss functions and varying Gaussian kernels. Our main goal is to provide fast convergence rates for the excess misclassification error. Allowing varying Gaussian kernels in the algorithms improves learning rates measured by regularization error and sample error. Special structures of Gaussian kernels enable us to construct, by a nice approximation scheme with a Fourier analysis technique, uniformly bounded regularizing functions achieving polynomial decays of the regularization error under a Sobolev smoothness condition. The sample error is estimated by using a projection operator and a tight bound for the covering numbers of reproducing kernel Hilbert spaces generated by Gaussian kernels. The convexity of the general loss function plays a very important role in our analysis.

Keywords: reproducing kernel Hilbert space, binary classification, general convex loss, varying Gaussian kernels, covering number, approximation

1. Introduction

In this paper we study binary classification algorithms generated from Tikhonov regularization schemes associated with general convex loss functions and varying Gaussian kernels.

Let X be a compact subset of $\mathbb{R}^n$ (input space) and $Y = \{1,-1\}$ (representing the two classes). Classification algorithms produce *binary classifiers* $C : X \to Y$. The misclassification error is used to measure the prediction power of a classifier C. If ρ is a probability distribution on $Z := X \times Y$, then the *misclassification error* of C is defined by

$$\mathcal{R}(C) = \text{Prob}\{C(x) \neq y\} = \int_X P(y \neq C(x)|x) d\rho_X.$$

Here ρ_X is the marginal distribution of ρ on X and $P(y|x)$ is the conditional distribution at $x \in X$. The classifier minimizing the misclassification error is called the Bayes rule f_c and is given by

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

The performance of a classifier C can be measured by the *excess misclassification error* $\mathcal{R}(C) - \mathcal{R}(f_c)$.

The classifiers considered here are induced by real-valued functions $f : X \to \mathbb{R}$ as $C_f = \text{sgn}(f)$ which is defined by $\text{sgn}(f)(x) = 1$ if $f(x) \geq 0$ and $\text{sgn}(f)(x) = -1$ otherwise. The real-valued

functions are generated from Tikhonov regularization schemes associated with general convex loss functions and varying Gaussian kernels.

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

Examples of classifying loss functions include the least-square loss $\phi_{ls}(t) = (1-t)^2$, the hinge loss $\phi_h(t) = (1-t)_+ = \max\{1-t, 0\}$ for support vector machine (SVM) algorithms, and the r-norm SVM loss with $1 \le r < \infty$ defined by $\phi_r(t) = (\phi_h(t))^r$.

The *Gaussian kernel* with variance $\sigma > 0$ is the function on $X \times X$ given by

$$K^\sigma(x,x') = \exp\Big\{-\frac{|x-x'|^2}{2\sigma^2}\Big\}. \tag{1}$$

It defines (Aronszajn, 1950) a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\sigma$.

With the loss ϕ and Gaussian kernel K^σ, the *Tikhonov regularization scheme* is defined (Wahba, 1990; Evgeniou et al., 2000; Cristianini and Shawe-Taylor, 2000) with a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ as the solution $f_{\mathbf{z}} = f^\phi_{\mathbf{z},\sigma,\lambda}$ to the following minimization problem

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

Here λ is a positive constant called the *regularization parameter*. Throughout the paper we assume that the sample $\mathbf{z}$ is drawn independently according to the distribution ρ.

The purpose of this paper is to estimate the excess misclassification error $\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c)$ as $m \to \infty$. Convergence rates will be derived under the choice of the parameters

$$\lambda = \lambda(m) = m^{-\gamma}, \quad \sigma = \sigma(m) = \lambda^\zeta = m^{-\gamma\zeta} \tag{3}$$

for some $\gamma, \zeta > 0$ and conditions on the distribution ρ and the loss ϕ. This has been done for the SVM in Steinwart and Scovel (2007) with the loss ϕ_h. Here we consider the error analysis with a general loss function ϕ (De Vito et al., 2004).

Let us demonstrate our main results by stating learning rates for the least-square loss $\phi = \phi_{ls}$. The rates will be proved in Section 4. They are given by means of a Tsybakov noise condition (Tsybakov, 2004) and a function smoothness condition stated in terms of Sobolev spaces. Since $\phi_{ls}(yf(x)) = (1-yf(x))^2 = (y-f(x))^2$ for $y \in Y$, a minimizer of $\int_Z \phi_{ls}(yf(x))d\rho$ is the *regression function* defined by

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

Definition 2 *Let $0 \le q \le \infty$. We say that ρ satisfies the Tsybakov noise condition with exponent q if there exists a constant $C_q > 0$ such that*

$$\rho_X(\{x \in X : |f_\rho(x)| \le C_q t\}) \le t^q, \qquad \forall t > 0. \tag{5}$$

Note that (5) always holds for $q = 0$ with $C_q = 1$. So setting the index $q = 0$ in (5) is the same as removing the Tsybakov noise condition. The case $q = \infty$ means $|f_\rho(x)| \geq C_q$ for almost every $x \in (X, \rho_X)$.

Recall the Sobolev space $H^s(\mathbb{R}^n)$ with index $s > 0$ consisting of all functions in $L^2(\mathbb{R}^n)$ with the semi-norm $|f|_{H^s(\mathbb{R}^n)} = \left\{(2\pi)^{-n} \int_{\mathbb{R}^n} |\xi|^{2s} |\hat{f}(\xi)|^2 d\xi\right\}^{\frac{1}{2}}$ finite where $\hat{f}$ is the Fourier transform of f defined for $f \in L^1(\mathbb{R}^n)$ as $\hat{f}(\xi) = \int_{\mathbb{R}^n} f(x) e^{-ix\cdot\xi}\, dx$.

Theorem 1 *Let $\phi = \phi_{ls}$. Assume (5) for some $q \in [0, \infty]$ and $\frac{d\rho_X}{dx} \in L^2(X)$. If for some $s > 0$, f_ρ equals the restriction onto X of some function in $H^s(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$, then by taking $\sigma = \lambda^\zeta$ with $0 < \zeta \leq \frac{1}{n+s}$ and $\lambda = m^{-\frac{1}{\zeta(s+2n+2)}}$, for any $0 < \delta < 1$, with confidence $1 - \delta$, we have*

$$\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leq \widetilde{C}_{\rho,s,q,n} m^{-\theta_{ls}} \log \frac{2}{\delta} \qquad \text{with} \quad \theta_{ls} = \frac{(q+1)s}{(q+2)(s+2n+2)}, \tag{6}$$

where $\widetilde{C}_{\rho,s,q,n}$ is a constant independent of m or δ.

When Tsybakov noise condition (5) is not assumed, we can still use Theorem 1 by setting $q = 0$ and obtain learning rate (6) with $\theta_{ls} = \frac{s}{2s+4n+4}$.

When q tends to infinity, the power index θ_{ls} in (6) has the limit $\frac{s}{s+2n+2}$ which can be very close to 1 for large s. So the learning rate can be $O(m^{\varepsilon-1})$ for arbitrarily small $\varepsilon > 0$ when q and s are large enough. To be more specific, for $0 < \varepsilon < 1$, when $q > \frac{1}{\varepsilon} - 2$ and $s \geq (2n+2)(q+2)\frac{1-\varepsilon}{\varepsilon(q+2)-1}$, we have $\theta_{ls} \geq 1 - \varepsilon$.

Remark 1 *We show that the power index θ_{ls} for learning rate (6) can be $1 - \varepsilon$ for arbitrarily small $\varepsilon > 0$. This result is new for scheme (2) associated with $\phi = \phi_{ls}$ and a single Gaussian kernel with changing variance $\sigma = \sigma(m)$. The same learning rates are achieved in the literature in two different settings: one is for the same least square regularization scheme associated with a single fixed Gaussian kernel, but under the much stronger condition that f_ρ lies in the range of powers of an integral operator associated with a fixed Gaussian kernel, requiring $f_\rho \in C^\infty$ (Zhang, 2004; De Vito et al., 2005; Smale and Zhou, 2007). The other setting is to allow flexible variances of Gaussians in (2), see Ying and Zhou (2007) and Wu et al. (2007).*

When the decision boundary $\{x \in X : f_\rho(x) = 0\}$ has measure zero and $\frac{d\rho_X}{dx} \in L^2(X)$, the smoothness condition for an extension of f_ρ implies (5) for some $q > 0$. In general, noise condition (5) does not require smoothness of f_ρ in domains away from the decision boundary.

Note that as $t \to -\infty$, the hinge loss ϕ_h for the SVM studied in Steinwart and Scovel (2007) increases slowly: $\phi_h(t) = O(|t|)$, while the least-square loss ϕ_{ls} in Theorem 1 increases moderately with $\phi_{ls}(t) = O(|t|^2)$. Difficulty arises for the error analysis with a general loss ϕ when $\phi(t)$ increases fast such as $\phi = \phi_r$ with very large r or the *exponential-hinge loss* we introduce in this paper as

$$\phi_{eh}(t) = \max\{e^{1-t} - 1, 0\} = \begin{cases} e^{1-t} - 1, & \text{if } t \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

The reason is random variables of form $\xi = \phi(yf(x))$ with $(x, y) = z \in (Z, \rho)$ are involved and large norms $\|f\|_{L^\infty(X)}$ would lead to large bounds for ξ. We shall use special properties of Gaussian

kernels and construct functions $f_{\sigma,\lambda}$ which are uniformly bounded and have powerful approximation ability (see (9) and (10) below). With this construction, we can do the analysis well for the general loss ϕ by dealing with uniformly bounded random variables in an error decomposition approach (see (13) below). In particular, explicit learning rates will be given in Section 4 for the r-norm SVM loss ϕ_r (Theorem 4) and the exponential-hinge loss ϕ_{eh} (Theorem 5). Comparing with Theorem 1, we shall provide at the end of Section 4 an approximation theory viewpoint to the effect of various loss functions for learning algorithm (2): the exponential-hinge loss has some advantages over ϕ_{ls} and ϕ_r, the r-norm SVM loss ϕ_r may have worse performance when $r > 2$.

We list key notations used in the paper in a table given in Appendix B.

2. Two Special Properties of Gaussians and Key Bounds

The novelty in our approach for general ϕ and kernels K^σ arises from two special properties of the Gaussian kernels with changing variance $\sigma > 0$: nice approximation scheme and low capacity of the RKHS, described in Sections 2.1 and 2.3.

2.1 Regularizing Functions Generated by Gaussians

A data-free limit of (2) is a function $\widetilde{f}_{\sigma,\lambda}$ defined in terms of the *generalization error* $\mathcal{E}^\phi$ as

$$\widetilde{f}_{\sigma,\lambda} := \arg\min_{f\in\mathcal{H}_\sigma}\{\mathcal{E}^\phi(f)+\lambda\|f\|^2_{\mathcal{H}_\sigma}\},\ \text{where } \mathcal{E}^\phi(f)=\int_Z \phi(yf(x))\,d\rho. \tag{7}$$

This is the regularizing function used in the literature (De Vito et al., 2005; Yao, 2008; Zhang, 2004). It works well for the error analysis when the loss ϕ increases slowly or moderately (as $t\to-\infty$) such as $\phi=\phi_h$ or ϕ_{ls}.

In this paper we consider a general loss ϕ. When $\phi(t)$ increases fast (as $t\to-\infty$), applying the regularizing function $\widetilde{f}_{\sigma,\lambda}$ in the error analysis (described in Section 2.2) may lead to a random variable $\phi(y\widetilde{f}_{\sigma,\lambda}(x))$ of large bound.

The first novelty of this paper is to construct a function $f_{\sigma,\lambda}$ (which plays the role of a regularizing function in an error decomposition approach discussed in subsection 2.2) by special approximation ability of Gaussian kernels. The constructed function has two advantages. On one hand, it is uniformly bounded (with respect to both λ and σ) so that the random variable $\phi(yf_{\sigma,\lambda}(x))$ involved in the error analysis is bounded. On the other hand, it plays the same role as $\widetilde{f}_{\sigma,\lambda}$ in achieving nice bounds for the approximation error. The construction of the explicit approximation scheme for $f_{\sigma,\lambda}$ is done under a Sobolev smoothness condition of a measurable function f_ρ^ϕ minimizing $\mathcal{E}^\phi$, that is, for a. e. $x\in X$,

$$f_\rho^\phi(x)=\arg\min_{t\in\mathbb{R}}\int_Y\phi(yt)\,d\rho(y|x)=\arg\min_{t\in\mathbb{R}}\{\phi(t)P(y=1|x)+\phi(-t)P(y=-1|x)\}.$$

Theorem 2 *Assume that for some $s>0$,*

$$f_\rho^\phi=\tilde{f}_\rho^\phi|_X \text{ for some } \tilde{f}_\rho^\phi\in H^s(\mathbb{R}^n)\cap L^\infty(\mathbb{R}^n) \text{ and } \frac{d\rho_X}{dx}\in L^2(X). \tag{8}$$

Then we can find functions $\{f_{\sigma,\lambda}\in\mathcal{H}_\sigma : 0<\sigma\le 1,\lambda>0\}$ such that

$$\|f_{\sigma,\lambda}\|_{L^\infty(X)} \le \widetilde{B}, \tag{9}$$

$$\mathcal{D}(\sigma,\lambda):=\mathcal{E}^\phi(f_{\sigma,\lambda})-\mathcal{E}^\phi(f_\rho^\phi)+\lambda\|f_{\sigma,\lambda}\|^2_{\mathcal{H}_\sigma} \le \widetilde{B}(\sigma^s+\lambda\sigma^{-n}) \tag{10}$$

for $0 < \sigma \le 1, \lambda > 0$, *where* $\widetilde{B} \ge 1$ *is a constant independent of* σ *or* λ.

Theorem 2 will be proved in Appendix A in a more general form as Theorem 6 where the constant $\widetilde{B}$ is given explicitly.

Remark 2 *A usual assumption in the literature (Zhang, 2004) for deriving learning rates is that for some* $0 < \beta \le 1$ *and* $C_\beta > 0$,

$$\widetilde{\mathcal{D}}(\lambda) = \min_{f \in \mathcal{H}_\sigma} \{\mathcal{E}^\phi(f) - \mathcal{E}^\phi(f_\rho^\phi) + \lambda \|f\|^2_{\mathcal{H}_\sigma}\} \le C_\beta \lambda^\beta \qquad \forall \lambda > 0. \tag{11}$$

This is hardly satisfied for a single fixed K^σ *due to the analyticity of the Gaussian kernel (Smale and Zhou, 2003; Cucker and Zhou, 2007). When we choose a changing Gaussian kernel with* $\sigma = \lambda^\zeta$ *for some* $\zeta > 0$, *decay (11) of the approximation error is valid in many cases (as shown in Theorem 2). Under assumption (11), one has the bound*

$$\|\widetilde{f}_{\sigma,\lambda}\|_{L^\infty(X)} \le \|\widetilde{f}_{\sigma,\lambda}\|_{\mathcal{H}_\sigma} \le \sqrt{\widetilde{\mathcal{D}}(\lambda)/\lambda} \le \sqrt{C_\beta}\lambda^{\frac{\beta-1}{2}}.$$

Hence a natural bound for the random variable $\phi_{eh}(y\widetilde{f}_{\sigma,\lambda}(x))$ *would be* $\exp\{\sqrt{C_\beta}\lambda^{\frac{\beta-1}{2}}\}$ *which increases exponentially fast as* $\lambda \to 0$ *(polynomially fast with degree* $\frac{r(1-\beta)}{2}$ *for* $\phi = \phi_r$ *when* r *is very large). This shows difficulty in choosing* $\widetilde{f}_{\sigma,\lambda}$ *and demonstrates novelty in choosing the function* $f_{\sigma,\lambda}$ *from Theorem 2 for the error analysis with a general loss* ϕ.

When $\sigma = \lambda^\zeta$ for some $0 < \zeta < \frac{1}{n}$, (10) of Theorem 2 tells us that the function $f_{\sigma,\lambda}$ yields an approximation order similar to (11) while (9) ensures the uniform boundedness of $\phi(yf_{\sigma,\lambda}(x))$, better than the function $\widetilde{f}_{\sigma,\lambda}$ for the error decomposition described below.

2.2 Error Decomposition and Projection Operator

The excess misclassification error $\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c)$ for the classifier $\mathrm{sgn}(f)$ can be bounded by means of the *excess generalization error* $\mathcal{E}^\phi(f)) - \mathcal{E}^\phi(f_\rho^\phi)$ according to some comparison theorems (Zhang, 2004; Chen et al., 2004; Bartlett et al., 2006). For example, it was proved in Zhang (2004) that for $\phi = \phi_h$ and any measurable function $f : X \to \mathbb{R}$, we have

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

For a classifying loss ϕ with $\phi''(0) > 0$, it was proved in Chen et al. (2004) and Bartlett et al. (2006) that for some $c_\phi > 0$,

$$\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c) \le c_\phi \sqrt{\mathcal{E}^\phi(f) - \mathcal{E}^\phi(f_\rho^\phi)}. \tag{12}$$

For the least square loss and ρ satisfying the Tsybakov noise condition, a comparison theorem improving (12) will be given in Section 4 and will be used to prove Theorem 1.

Classifiers in this paper are obtained by taking signs of real-valued functions. Since the smallest zero of ϕ is 1, we can take $f_\rho^\phi(x) \in [-1,1]$ for each $x \in X$, which we shall assume throughout the paper. We may improve the error estimates (Chen et al., 2004) by replacing values of f by projections onto $[-1,1]$.

Definition 3 *The projection operator π on the space of functions on X is defined by*

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

Trivially $\operatorname{sgn}(\pi(f)) = \operatorname{sgn}(f)$. Then we can use (12) with $f = \pi(f_{\mathbf{z}})$ to bound the excess misclassification error $\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c)$ by means of the excess generalization error $\mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}^{\phi}(f_{\rho}^{\phi})$ which in turn can be estimated by an error decomposition technique (Wu and Zhou, 2006). Define the *empirical error* associated with the loss ϕ as

$$\mathcal{E}_{\mathbf{z}}^{\phi}(f) = \frac{1}{m}\sum_{i=1}^{m} \phi(y_i f(x_i)) \qquad \text{for } f : X \to \mathbb{R}.$$

Then we have the following error decomposition which will be proved in Section 3.

Lemma 1 *Let ϕ be a classifying loss, $f_{\mathbf{z}}$ be defined by (2) and $f_{\sigma,\lambda} \in \mathcal{H}_{\sigma}$. Then*

$$\mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) \le \mathcal{D}(\sigma,\lambda) + \mathcal{S}_{\mathbf{z}}(f_{\sigma,\lambda}) - \mathcal{S}_{\mathbf{z}}(\pi(f_{\mathbf{z}})), \tag{13}$$

where the quantity $\mathcal{S}_{\mathbf{z}}(f)$ is defined for $f \in C(X)$ by

$$\mathcal{S}_{\mathbf{z}}(f) = [\mathcal{E}_{\mathbf{z}}^{\phi}(f) - \mathcal{E}_{\mathbf{z}}^{\phi}(f_{\rho}^{\phi})] - [\mathcal{E}^{\phi}(f) - \mathcal{E}^{\phi}(f_{\rho}^{\phi})].$$

When we use the regularizing function $f_{\sigma,\lambda}$ given in Theorem 2, the bound (10) deals with $\mathcal{D}(\sigma,\lambda)$, the first term of (13). The uniform bound (9) for $\|f_{\sigma,\lambda}\|_{L^{\infty}(X)}$ ensures that the second term $\mathcal{S}_{\mathbf{z}}(f_{\sigma,\lambda})$ of (13), which can be expressed as $\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mathbf{E}(\xi)$ with the random variable $\xi(z) = \phi(y f_{\sigma,\lambda}(x)) - \phi(y f_{\rho}^{\phi}(x))$, can be easily handled. The crucial remaining term $\mathcal{S}_{\mathbf{z}}(\pi(f_{\mathbf{z}}))$ of (13) involves the set of functions $\{f_{\mathbf{z}}\}_{\mathbf{z}\in Z^m}$ and can be treated by various empirical process techniques such as Rademacher average and entropy integral. Here we use the specialty of the Gaussians that the RKHS has low capacity, hence the last term of (13) can be estimated efficiently and simply by means of covering numbers.

2.3 Applying Tight Bounds for Covering Numbers

The second novelty of this paper is to make full use of the special low capacity property of the Gaussian kernels that a tight bound for covering numbers of the unit ball of the RKHS $\mathcal{H}_{\sigma}$ leads to nice estimates for the last term $\mathcal{S}_{\mathbf{z}}(\pi(f_{\mathbf{z}}))$ of (13) for the error analysis.

Definition 4 *For a subset S of $C(X)$ and $\eta > 0$, the covering number $\mathcal{N}(S,\eta)$ is the minimal integer $l \in \mathbb{N}$ such that there exist l disks with radius η covering S.*

The covering numbers of unit balls of classical function spaces have been well studied in the literature (Edmunds and Triebel, 1996). As an example, take $X = [0,1]^n$ and $s > 0$. The covering numbers of the unit ball $B_1(C^s(X))$ of the space $C^s(X)$ has the asymptotic behavior

$$c_s'(\frac{1}{\eta})^{n/s} \le \log \mathcal{N}(B_1(C^s(X)),\eta) \le c_s''(\frac{1}{\eta})^{n/s}, \tag{14}$$

where the positive constants c'_s, and c''_s are independent of $0<\eta<1$. In particular, since a Gaussian kernel K^σ is C^∞, an embedding result from Zhou (2003) tells us that $\log \mathcal{N}(B_1,\eta) \le C''_s(\frac{1}{\eta})^{n/s}(\frac{1}{\sigma})^{2n}$ where $s>0$ can be arbitrarily large but the constant C''_s depends on s. Here $B_1 = B_{1,\sigma} = \{f \in \mathcal{H}_\sigma : \|f\|_{\mathcal{H}_\sigma} \le 1\}$ is the unit ball of $\mathcal{H}_\sigma$ and is regarded as a compact subset of $C(X)$. A crucial improved bound for the covering number of B_1 was given in Zhou (2002) with $(\frac{1}{\eta})^{n/s}$ replaced by $(\log \frac{1}{\eta})^{n+1}$ as follows.

Proposition 1 *There exists a constant $C_0>0$ depending only on X and n such that*

$$\log \mathcal{N}(B_1,\eta) \le C_0\Big((\log\frac{1}{\eta})^{n+1} + \frac{1}{\sigma^{2(n+1)}}\Big) \qquad \forall\, 0<\eta<1, 0<\sigma\le 1. \tag{15}$$

The constant C_0 can be taken as $(124n)^{n+2}$ when $X=[0,1]^n$. Bound (15) is almost sharp in the sense that for some $C'_0>0$ given in Zhou (2003),

$$\log \mathcal{N}(B_1,\eta) \ge C'_0\Big((\log\frac{1}{\eta})^{n/2} + \frac{1}{\sigma^{n}}\Big).$$

The logarithmic term $(\log\frac{1}{\eta})^{n+1}$ appearing in the tight bound (15) is better than the polynomial term $(\frac{1}{\eta})^{n/s}$ in (14). This enables us to derive efficient error bounds for the algorithm (2) involving Gaussian kernels, by a simple covering number argument without other empirical process techniques or iteration techniques used in Steinwart and Scovel (2007) and Wu et al. (2007). To demonstrate explicitly why tight bound (15) helps, we state the following result which is needed for estimating confidence and will be proved in Appendix B.

Lemma 2 *Let $0\le\tau\le 1$ and $C_1>0$. Let $0<\delta<1$ and λ,σ take form (3) with some $\gamma>0$ and $0<\zeta<\frac{1}{2\gamma(n+1)}$. Denote $\varepsilon^*(m,\lambda,\sigma,\delta/2)$ as the smallest positive number ε satisfying*

$$1-\mathcal{N}\Big(B_1, \frac{\lambda\varepsilon}{\sqrt{\phi(0)}|\phi'_+(-1)|}\Big)\exp\left\{-\frac{m\varepsilon^{2-\tau}}{2C_1+\frac{2}{3}\phi(-1)\varepsilon^{1-\tau}}\right\} \ge 1-\frac{\delta}{2}. \tag{16}$$

Then we have

$$\varepsilon^*(m,\lambda,\sigma,\delta/2) \le C_2 m^{-\frac{1-2\gamma\zeta(n+1)}{2-\tau}}\log\frac{2}{\delta}, \tag{17}$$

where C_2 is the constant independent of m,λ,σ or δ.

2.4 Key Bounds

We are in a position to present our key bounds for the excess generalization error $\mathcal{E}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi(f_\rho^\phi)$ which will be used to get rates for the excess misclassification error $\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c)$. To achieve tight bounds, we need the following definition.

Definition 5 *A variancing power $\tau=\tau_{\phi,\rho}$ of the pair (ϕ,ρ) is a number τ in $[0,1]$ such that for any $\widetilde{B}\ge 1$, there exists some constant $C_1=C_1(\widetilde{B})>0$ satisfying*

$$\mathbf{E}\Big\{\Big[\phi(yf(x))-\phi(yf_\rho^\phi(x))\Big]^2\Big\} \le C_1\Big[\mathcal{E}^\phi(f)-\mathcal{E}^\phi(f_\rho^\phi)\Big]^\tau \qquad \forall\, f: X\to[-\widetilde{B},\widetilde{B}]. \tag{18}$$

Remark 3 *For $\phi = \phi_{ls}$, we can take $\tau = 1$, see Evgeniou et al. (2000) and Cucker and Zhou (2007). For $\phi = \phi_h$, we can take $\tau = 0$, and an improved power $\tau = \frac{q}{q+1}$ if the Tsybakov noise condition (5) is satisfied (Steinwart and Scovel, 2007; Wu and Zhou, 2005). In general, $\tau_{\phi,\rho}$ depends on the strong convexity of ϕ and noise conditions for ρ.*

Theorem 3 *Let $\sigma = \lambda^{\zeta}$ and $\lambda = m^{-\gamma}$ for some $0 < \zeta < \frac{1}{n}$ and $0 < \gamma < \frac{1}{2\zeta(n+1)}$. If (8) is valid for some $s > 0$, then for any $0 < \delta < 1$, with confidence $1 - \delta$ we have*

$$\mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) \le \widetilde{C} m^{-\theta} \log \frac{2}{\delta} \tag{19}$$

where

$$\theta = \min\Big\{ s\zeta\gamma, \gamma(1 - n\zeta), \frac{1 - 2\gamma\zeta(n+1)}{2 - \tau} \Big\}, \tag{20}$$

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

Theorem 3 will be proved in the next section and the constant $\widetilde{C}$ will be given explicitly.

3. Error Analysis

In this section we derive the key error bounds stated in Theorem 3 by estimating the right-hand side of (13) in Lemma 1 (which is proved here).

3.1 Proof of Lemma 1

Write the regularized excess generalization error as

$$\begin{aligned}
&\mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) + \lambda \|f_{\mathbf{z}}\|_{\mathcal{H}_\sigma}^2 = \Big\{ \mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}^{\phi}(\pi(f_{\mathbf{z}})) \Big\} \\
&+ \Big\{ \Big[\mathcal{E}_{\mathbf{z}}^{\phi}(\pi(f_{\mathbf{z}})) + \lambda \|f_{\mathbf{z}}\|_{\mathcal{H}_\sigma}^2 \Big] - \Big[\mathcal{E}_{\mathbf{z}}^{\phi}(f_{\sigma,\lambda}) + \lambda \|f_{\sigma,\lambda}\|_{\mathcal{H}_\sigma}^2 \Big] \Big\} \\
&+ \Big\{ \mathcal{E}_{\mathbf{z}}^{\phi}(f_{\sigma,\lambda}) - \mathcal{E}^{\phi}(f_{\sigma,\lambda}) \Big\} + \Big\{ \mathcal{E}^{\phi}(f_{\sigma,\lambda}) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}) + \lambda \|f_{\sigma,\lambda}\|_{\mathcal{H}_\sigma}^2 \Big\}.
\end{aligned}$$

Since ϕ is convex and its smallest zero is 1, we find a special property of the projection operator that $\phi(y\pi(f)(x)) \le \phi(yf(x))$ for any function f on X. Hence $\mathcal{E}_{\mathbf{z}}^{\phi}(\pi(f)) \le \mathcal{E}_{\mathbf{z}}^{\phi}(f)$. This in connection with the definition of $f_{\mathbf{z}}$ tells us that the second term on the right-hand side above is at most zero. By subtracting and adding $\mathcal{E}^{\phi}(f_{\rho}^{\phi})$ in the first and third terms we see $\mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}^{\phi}(f_{\rho}^{\phi})$ is bounded as in (13). This proves Lemma 1. ■

Let us turn to estimate $\mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}^{\phi}(f_{\rho}^{\phi})$ by (13). We first bound $\mathcal{S}_{\mathbf{z}}(f_{\sigma,\lambda})$, the term involving $f_{\sigma,\lambda}$. It can be written as $\frac{1}{m}\sum_{i=1}^{m} \xi(z_i) - \mathbf{E}(\xi)$ with ξ the random variable on (Z,ρ) given by $\xi(z) = \phi(yf_{\sigma,\lambda}(x)) - \phi(yf_{\rho}^{\phi}(x))$.

Lemma 3 *Let $\tau = \tau_{\phi,\rho}$ and $f_{\sigma,\lambda} \in \mathcal{H}_\sigma$ satisfy (9). For any $0 < \delta < 1$, with confidence $1 - \frac{\delta}{2}$, the term $\mathcal{S}_{\mathbf{z}}(f_{\sigma,\lambda})$ of (13) can be bounded as*

$$\mathcal{S}_{\mathbf{z}}(f_{\sigma,\lambda}) \le 2(\|\phi\|_{C[-\widetilde{B},\widetilde{B}]} + C_1) \log \frac{2}{\delta} m^{-\frac{1}{2-\tau}} + \mathcal{E}^{\phi}(f_{\sigma,\lambda}) - \mathcal{E}^{\phi}(f_{\rho}^{\phi}).$$

Proof Consider the random variable $\xi(z) = \phi(yf_{\sigma,\lambda}(x)) - \phi(yf_\rho^\phi(x))$ on (Z,ρ). It satisfies $-\phi(-1) \le \xi \le \|\phi\|_{C[-\widetilde{B},\widetilde{B}]}$. Hence $|\xi - \mathbf{E}(\xi)| \le 2\|\phi\|_{C[-\widetilde{B},\widetilde{B}]}$. We apply the one side Bernstein inequality and know that

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\Big\{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbf{E}(\xi) > \varepsilon\Big\} \le \exp\Big\{-\frac{m\varepsilon^2}{2(\sigma^2(\xi) + \frac{2}{3}\|\phi\|_{C[-\widetilde{B},\widetilde{B}]}\varepsilon)}\Big\} \qquad \forall \varepsilon > 0.$$

Here $\sigma^2(\xi)$ is the variance of ξ. Solving the quadratic equation for ε by setting the above probability bound to be $\delta/2$, we see that with confidence at least $1-\delta/2$,

$$\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbf{E}(\xi) \le \frac{4\|\phi\|_{C[-\widetilde{B},\widetilde{B}]}\log\frac{2}{\delta}}{3m} + \frac{\sqrt{2m\sigma^2(\xi)\log\frac{2}{\delta}}}{m}.$$

Using (18) involving the variancing power $\tau = \tau_{\phi,\rho}$ in Definition 5, we have $\sigma^2(\xi) \le \mathbf{E}(\xi^2) \le C_1(\mathbf{E}(\xi))^\tau$. This in connection with Young's inequality implies

$$\frac{\sqrt{2m\sigma^2(\xi)\log\frac{2}{\delta}}}{m} \le \sqrt{\frac{2\log\frac{2}{\delta}C_1(\mathbf{E}(\xi))^\tau}{m}} \le (1-\frac{\tau}{2})\Big(\frac{2\log\frac{2}{\delta}C_1}{m}\Big)^{\frac{1}{2-\tau}} + \frac{\tau}{2}\mathbf{E}(\xi).$$

Therefore, with confidence at least $1-\delta/2$,

$$\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbf{E}(\xi) \le \frac{4\|\phi\|_{C[-\widetilde{B},\widetilde{B}]}\log\frac{2}{\delta}}{3m} + \Big(\frac{2\log\frac{2}{\delta}C_1}{m}\Big)^{\frac{1}{2-\tau}} + \mathbf{E}(\xi).$$

Since $\mathbf{E}(\xi) = \mathcal{E}^\phi(f_{\sigma,\lambda}) - \mathcal{E}^\phi(f_\rho^\phi)$, our conclusion follows. ■

The sample error term $-\mathcal{S}_{\mathbf{z}}(\pi(f_{\mathbf{z}}))$ in (13) can be expressed as $\int \xi_{\mathbf{z}} d\rho - \frac{1}{m}\sum_{i=1}^m \xi_{\mathbf{z}}(z_i)$ with $\xi_{\mathbf{z}}(z) = \phi(yf_{\mathbf{z}}(x)) - \phi(yf_\rho^\phi(x))$. However, $\xi_{\mathbf{z}}$ is not a single random variable since $\mathbf{z}$ is a random sample itself. This is the essential difficulty. Here we use the specialty of low capacity of the RKHS $\mathcal{H}_\sigma$ and overcome the difficulty by a simple covering number argument over a ball of $\mathcal{H}_\sigma$ where $f_{\mathbf{z}}$ lies.

Lemma 4 *For any $\lambda > 0$ and $\mathbf{z} \in Z^m$, there holds*

$$\|f_{\mathbf{z}}\|_{\mathcal{H}_\sigma} \le \sqrt{\phi(0)/\lambda}.$$

The proof follows easily by taking $f = 0$ in the definition of $f_{\mathbf{z}}$ as in De Vito et al. (2005), De Vito et al. (2004) and Hardin et al. (2004).

Let ξ be a random variable on Z with mean $\mu \ge 0$ and variance $\sigma^2 \le c\mu^\tau$ for some $0 \le \tau \le 2$ and $c \ge 0$. If $|\xi - \mu| \le B$ almost surely for some $B \ge 0$, then the one-side Bernstein inequality implies

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\Big\{\frac{\mu - \frac{1}{m}\sum_{i=1}^m \xi(z_i)}{\sqrt{\mu^\tau + \varepsilon^\tau}} > \varepsilon^{1-\frac{\tau}{2}}\Big\} \le \exp\Big\{-\frac{m\varepsilon^{2-\tau}}{2(c + \frac{1}{3}B\varepsilon^{1-\tau})}\Big\} \qquad \forall\, \varepsilon > 0.$$

Applying this probability inequality to random variables of type $\xi(z) = \phi(y(\pi f)(x)) - \phi(yf_\rho^\phi(x))$ and using a standard argument (Wu et al., 2007; Yao, 2008; Ying, 2007) with covering numbers for the ball $\{f \in \mathcal{H}_\sigma : \|f\|_{\mathcal{H}_\sigma} \le \sqrt{\phi(0)/\lambda}\}$ of the RKHS $\mathcal{H}_\sigma$, we find the following bound.

Lemma 5 *Let $\tau = \tau_{\phi,\rho}$ satisfy (18) with $\widetilde{B}$ being 1. For any $\varepsilon > 0$, we have*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\Big\{ \sup_{\|f\|_{\mathcal{H}_\sigma}\le\sqrt{\phi(0)/\lambda}} \frac{[\mathcal{E}^\phi(\pi(f)) - \mathcal{E}^\phi(f_\rho^\phi)] - [\mathcal{E}_{\mathbf{z}}^\phi(\pi(f)) - \mathcal{E}_{\mathbf{z}}^\phi(f_\rho^\phi)]}{\sqrt{(\mathcal{E}^\phi(\pi(f)) - \mathcal{E}^\phi(f_\rho^\phi))^\tau + \varepsilon^\tau}} \le 4\varepsilon^{1-\frac{\tau}{2}} \Big\}$$
$$\ge 1 - \mathcal{N}\Big(B_1, \frac{\sqrt{\lambda}\varepsilon}{\sqrt{\phi(0)}|\phi'_+(-1)|}\Big) \exp\Big\{ - \frac{m\varepsilon^{2-\tau}}{2C_1 + \frac{2}{3}\phi(-1)\varepsilon^{1-\tau}} \Big\}.$$

Recall the definition of $\varepsilon^*(m,\lambda,\sigma,\delta/2)$ in Lemma 2. It satisfies (16) which means that the probability in Lemma 5 is bounded by $1-\frac{\delta}{2}$ from below when $\varepsilon = \varepsilon^*(m,\lambda,\sigma,\delta/2)$.

Proposition 2 *Let $f_{\sigma,\lambda} \in \mathcal{H}_\sigma$ satisfy (9). For any $0 < \delta < 1$, with confidence at least $1-\delta$, we have*

$$\mathcal{E}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi(f_\rho^\phi) \le 4\mathcal{D}(\sigma,\lambda) + 40\varepsilon^*(m,\lambda,\sigma,\delta/2) + 4(\|\phi\|_{C[-\widetilde{B},\widetilde{B}]} + C_1)\log\frac{2}{\delta} m^{-\frac{1}{2-\tau}}.$$

Proof Applying Lemma 3, we know that there is a subset V_1 of Z^m with measure at least $1-\frac{\delta}{2}$ such that for $\mathbf{z} \in V_1$,

$$\mathcal{S}_{\mathbf{z}}(f_{\sigma,\lambda}) \le 2(\|\phi\|_{C[-\widetilde{B},\widetilde{B}]} + C_1)\log\frac{2}{\delta} m^{-\frac{1}{2-\tau}} + \mathcal{D}(\sigma,\lambda).$$

By Lemma 5 and Lemma 4, taking $\varepsilon = \varepsilon^*(m,\lambda,\sigma,\delta/2)$, we see that there exists another subset V_2 of Z^m with measure at least $1-\frac{\delta}{2}$ such that for $\mathbf{z} \in V_2$,

$$\begin{aligned} -\mathcal{S}_{\mathbf{z}}(\pi(f_{\mathbf{z}})) &= [\mathcal{E}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi(f_\rho^\phi)] - [\mathcal{E}_{\mathbf{z}}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}^\phi(f_\rho^\phi)] \\ &\le 4[\varepsilon^*(m,\lambda,\sigma,\delta/2)]^{1-\frac{\tau}{2}} \sqrt{\Big[\mathcal{E}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi(f_\rho^\phi)\Big]^\tau + [\varepsilon^*(m,\lambda,\sigma,\delta/2)]^\tau} \\ &\le \Big(1-\frac{\tau}{2}\Big) 4^{\frac{2}{2-\tau}} \varepsilon^*(m,\lambda,\sigma,\delta/2) + \frac{\tau}{2}[\mathcal{E}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi(f_\rho^\phi)] + 4\varepsilon^*(m,\lambda,\sigma,\delta/2). \end{aligned}$$

Here we have used the elementary inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and Young's inequality.

Adding the above two bounds and observing that $0 \le \tau \le 1$ implies $\frac{1}{1-\tau/2} \le 2$ we know from Lemma 1 that for $\mathbf{z} \in V_1 \cap V_2$,

$$\mathcal{E}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi(f_\rho^\phi) \le 4\mathcal{D}(\sigma,\lambda) + 40\varepsilon^*(m,\lambda,\sigma,\delta/2) + 4(\|\phi\|_{C[-\widetilde{B},\widetilde{B}]} + C_1)\log\frac{2}{\delta} m^{-\frac{1}{2-\tau}}.$$

Since the measure of $V_1 \cap V_2$ is at least $1-\delta$, our conclusion holds true. ■

Now we are in a position to prove Theorem 3.

3.2 Proof of Theorem 3

From condition (8) and the parameter form $\sigma = \lambda^\zeta$ with $0 < \zeta < \frac{1}{n}$, we know by Theorem 2 that

$$\mathcal{D}(\sigma,\lambda) \le \widetilde{B}\lambda^{\min\{s\zeta, 1-n\zeta\}}.$$

Putting bound (17) for $\varepsilon^*(m,\lambda,\sigma,\delta/2)$ from Lemma 2 into Proposition 2, we see from the parameter form $\lambda = m^{-\gamma}$ that with confidence at least $1-\delta$,

$$\begin{aligned}\mathcal{E}^{\phi}(\pi(f_{\mathbf{z}})) - \mathcal{E}^{\phi}(f_\rho^{\phi}) &\leq 4\widetilde{B}\lambda^{\min\{s\zeta,1-n\zeta\}} + 40C_2 m^{-\frac{1-2\gamma\zeta(n+1)}{2-\tau}} \log\frac{2}{\delta} \\ &\quad + 4(\|\phi\|_{C[-\widetilde{B},\widetilde{B}]} + C_1)\log\frac{2}{\delta} m^{-\frac{1}{2-\tau}} \leq \widetilde{C} m^{-\theta}\log\frac{2}{\delta}.\end{aligned}$$

Here θ is given by (20) and $\widetilde{C}$ is the constant independent of m and δ given by

$$\widetilde{C} = 4\widetilde{B} + 40C_2 + 4(\|\phi\|_{C[-\widetilde{B},\widetilde{B}]} + C_1).$$

This proves (19) and hence Theorem 3. ■

4. Deriving Learning Rates

In this section we apply Theorem 3 to derive learning rates with various loss functions. For the least square loss, to prove Theorem 1 we need the following comparison theorem improving (12).

Proposition 3 *If $\phi = \phi_{ls}$ and ρ satisfies noise condition (5) for some $q \in [0,\infty]$, then for every measurable function $f : X \to \mathbb{R}$, we have*

$$\mathcal{R}(sgn(f)) - \mathcal{R}(f_c) \leq 2C_q^{-\frac{q}{q+2}} \left\{\mathcal{E}^{\phi_{ls}}(f) - \mathcal{E}^{\phi_{ls}}(f_\rho)\right\}^{\frac{q+1}{q+2}}.$$

Proof Denote $X_f = \{x \in X : \mathrm{sgn}(f)(x) \neq f_c(x)\}$. It is known that $\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c) = \int_{X_f} |f_\rho(x)| d\rho_X$. See, for example, Equation (9.14) of Cucker and Zhou (2007).

When $q < \infty$, take $t = \left(\|f - f_\rho\|_{L^2_{\rho_X}}/C_q\right)^{\frac{2}{q+2}} > 0$. We separate the set X_f into two parts, one with $|f_\rho(x)| \leq C_q t$ and the other with $|f_\rho(x)| > C_q t$ where $|f_\rho(x)| \leq |f_\rho(x)|^2/(C_q t) \leq |f(x) - f_\rho(x)|^2/(C_q t)$. We find from (5) that

$$\begin{aligned}\int_{X_f} |f_\rho(x)| d\rho_X &\leq \int_{\{x\in X_f:\, |f_\rho(x)|\leq C_q t\}} C_q t d\rho_X + \int_{\{x\in X_f:\, |f_\rho(x)|> C_q t\}} |f(x) - f_\rho(x)|^2/(C_q t) d\rho_X \\ &\leq C_q t \rho_X(\{x \in X : |f_\rho(x)| \leq C_q t\}) + \|f - f_\rho\|^2_{L^2_{\rho_X}}/(C_q t) \\ &\leq C_q t^{q+1} + \|f - f_\rho\|^2_{L^2_{\rho_X}}/(C_q t) = 2C_q \left(\|f - f_\rho\|_{L^2_{\rho_X}}/C_q\right)^{\frac{2q+2}{q+2}}.\end{aligned}$$

This gives the desired bound for the case $q < \infty$ since $\|f - f_\rho\|^2_{L^2_{\rho_X}} = \mathcal{E}^{\phi_{ls}}(f) - \mathcal{E}^{\phi_{ls}}(f_\rho)$.

When $q = \infty$, noise condition (5) means $|f_\rho(x)| \geq C_q$ and hence $|f_\rho(x)| \leq |f_\rho(x)|^2/C_q$ for almost every $x \in (X, \rho_X)$. So $\int_{X_f} |f_\rho(x)| d\rho_X \leq \int_{X_f} |f(x) - f_\rho(x)|^2/C_q d\rho_X = \|f - f_\rho\|^2_{L^2_{\rho_X}}/C_q$ which is what we want. ■

Now we can derive learning rates with the least square loss.

4.1 Proof of Theorem 1

The assumptions on $\frac{d\rho_X}{dx}$ and f_ρ verify condition (8). Then by Theorem 2 with $\phi = \phi_{ls}$, we find functions $f_{\sigma,\lambda}$ satisfying (9) and (10) for $0 < \sigma \le 1, \lambda > 0$.

The choice $\sigma = \lambda^\zeta$ with $0 < \zeta \le \frac{1}{n+s} < \frac{1}{n}$ and $\lambda = m^{-\gamma}$ with $\gamma = \frac{1}{\zeta(s+2n+2)}$ tell us that $0 < \gamma < \frac{1}{2\zeta(n+1)}$. Therefore all conditions of Theorem 3 are valid. Moreover, a specialty of the least square loss is $\tau = 1$ in (18). So by Theorem 3, for any $0 < \delta < 1$, with confidence $1-\delta$, (19) holds with

$$\theta = \min\Big\{\frac{s}{s+2n+2}, \frac{1-n\zeta}{\zeta(s+2n+2)}, 1-\frac{2(n+1)}{s+2n+2}\Big\} = \frac{s}{s+2n+2}.$$

This bound for the excess generalization error $\mathcal{E}^\phi(\pi(f_{\mathbf{z}})) - \mathcal{E}^\phi(f_\rho^\phi)$ together with Proposition 3 yields the desired bound (6) for the excess misclassification error $\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c)$ with the constant $\widetilde{C}_{\rho,s,q,n} = 2C_q^{-\frac{q}{q+2}}\widetilde{C}^{\frac{q+1}{q+2}}$. The proof of Theorem 1 is complete. ■

Let us derive learning rates with the r-norm SVM loss $\phi = \phi_r$ $(1 < r < \infty)$ for which we have (Chen et al., 2004)

$$f_\rho^\phi(x) = f_\rho^{\phi_r}(x) = \frac{(1+f_\rho(x))^{1/(r-1)} - (1-f_\rho(x))^{1/(r-1)}}{(1+f_\rho(x))^{1/(r-1)} + (1-f_\rho(x))^{1/(r-1)}}, \qquad x \in X. \tag{21}$$

Theorem 4 *Let $\phi = \phi_r$ with $1 < r < \infty$. Assume (8) for some $s > 0$. Take $\sigma = \lambda^\zeta$ with $0 < \zeta \le \frac{1}{n+s}$ and $\lambda = m^{-\gamma}$ with $\gamma = \frac{1}{\zeta(s+2n+2)}$ for $1 < r \le 2$ and $\gamma = \frac{1}{\zeta(2s(1-1/r)+2n+2)}$ for $2 < r < \infty$. Then for any $0 < \delta < 1$, with confidence $1-\delta$, we have*

$$\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) \le \widetilde{C}_{\rho,r} m^{-\theta_r} \log\frac{2}{\delta} \quad \text{with } \theta_r = \begin{cases} \frac{s}{2(s+2n+2)}, & \text{if } 1 < r \le 2, \\ \frac{s}{4(s(1-1/r)+n+1)}, & \text{if } 2 < r < \infty. \end{cases} \tag{22}$$

Proof The convexity of ϕ_r gives the variancing power (Bartlett et al., 2006) as

$$\tau = \tau_{\phi_r,\rho} = \begin{cases} 1, & \text{if } 1 < r \le 2, \\ \frac{2}{r}, & \text{if } 2 < r < \infty. \end{cases}$$

Take $\sigma = \lambda^\zeta$ with $0 < \zeta \le \frac{1}{n+s} < \frac{1}{n}$ and choose $\lambda = m^{-\gamma}$ with $\gamma = \frac{1}{\zeta((2-\tau)s+2n+2)}$. We see that $0 < \gamma < \frac{1}{2\zeta(n+1)}$. Hence all conditions of Theorem 3 are valid and we conclude that for any $0 < \delta < 1$, with confidence $1-\delta$, (19) holds with $\theta = \frac{s}{(2-\tau)s+2n+2}$. This bound for the excess generalization error together with comparison relation (12) caused by $\phi_r''(0) = r(r-1) > 0$ yields the desired bound (22) for the excess misclassification error with the constant $\widetilde{C}_{\rho,r} = c_{\phi_r}\sqrt{\widetilde{C}}$. The proof of Theorem 4 is complete. ■

When $\phi = \phi_{eh}$, a simple computation shows that the function f_ρ^ϕ is given by

$$f_\rho^{\phi_{eh}}(x) = \begin{cases} \frac{1}{2}\log\frac{1+f_\rho(x)}{1-f_\rho(x)}, & \text{if } -(e^2-1)/(e^2+1) \le f_\rho(x) \le (e^2-1)/(e^2+1), \\ 1, & \text{if } f_\rho(x) > (e^2-1)/(e^2+1), \\ -1, & \text{if } f_\rho(x) < -(e^2-1)/(e^2+1). \end{cases} \tag{23}$$

Theorem 5 *Let $\phi = \phi_{eh}$. Assume (8) for some $s > 0$. Take $\sigma = \lambda^{\zeta}$ with $0 < \zeta \le \frac{1}{n+s}$ and $\lambda = m^{-\frac{1}{\zeta(2s+2n+2)}}$. Then for any $0 < \delta < 1$, with confidence $1-\delta$, we have*

$$\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) \le \widetilde{C}_{\rho,eh} m^{-\theta_{eh}} \log \frac{2}{\delta} \qquad \text{with} \quad \theta_{eh} = \frac{s}{4s+4n+4}. \tag{24}$$

Proof Take $\tau = 0$ and $\sigma = \lambda^{\zeta}$ with $0 < \zeta \le \frac{1}{n+s} < \frac{1}{n}$. Choose $\lambda = m^{-\gamma}$ with $\gamma = \frac{1}{\zeta(2s+2n+2)}$ in Theorem 3. We see that for any $0 < \delta < 1$, with confidence $1-\delta$, (19) holds with $\theta = \frac{s}{2s+2n+2}$. This bound together with comparison relation (12) again (as $\phi''_{eh}(0) = e > 0$) yields the desired bound (24) with $\widetilde{C}_{\rho,eh} = c_{\phi_{eh}} \sqrt{\widetilde{C}}$. This proves Theorem 5. ■

Remark 4 *When $s \le \frac{1}{r-1}$, the extension condition of f_ρ stated in Theorem 1 implies assumption (8) of $f_\rho^{\phi_r}$ required in Theorem 4. In fact, the extension of the function $f_\rho^{\phi_r}$ onto $\mathbb{R}^n$ can be defined by taking values of the extended function of f_ρ in (21). After composing with the function $t \to t^{1/(r-1)}$ on $\mathbb{R}$, smoothness of functions in the Sobolev space H^s is kept for $s \le \frac{1}{r-1}$. When $s \le 1$, the same condition for f_ρ implies assumption (8) of $f_\rho^{\phi_{eh}}$ needed for Theorem 5, as seen from expression (23).*

It is possible to refine learning rates (22) and (24) by improving comparison relation (12) when Tsybakov noise condition (5) is satisfied. We omit the discussion here.

Error analysis with $\phi = \phi_r$ was done in Chen et al. (2004) under assumption (11). Our learning rates in Theorem 4 are new since our assumption on Sobolev smoothness is weaker. The learning rates for $\phi = \phi_{eh}$ in Theorem 5 are also new.

We are in a position to get from Theorems 1, 4 and 5 some theoretical clues on the effect of various loss functions for learning algorithm (2). We know from Smale and Zhou (2003) that when $\phi = \phi_{ls}$ the approximation error and hence learning rates can essentially be characterized by regularities of the function f_ρ. So here we give some comparisons under the same regularity assumption (8) for the function f_ρ^ϕ with some $s > 0$. Under this assumption (removing the Tsybakov noise condition by taking $q = 0$ in Theorem 1), the learning rates derived in Theorems 1, 4 and 5 for $\phi = \phi_{ls}, \phi_r, \phi_{eh}$ take the same form $\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}})) - \mathcal{R}(f_c) = O(m^{-\theta} \log \frac{2}{\delta})$ with the power index θ very close, all lying in the range $[\frac{s}{4s+4n+4}, \frac{s}{2s+4n+4}]$. However, the index s in regularity assumption (8) for the function f_ρ^ϕ might vary dramatically, leading to varying power index θ for the learning rates.

Note that the function f_ρ^ϕ with $\phi = \phi_r, \phi_{eh}$ depends explicitly on the regression function f_ρ corresponding the least-square loss. The dependence of the function $f_\rho^{\phi_{eh}}$ on f_ρ has an advantage of ignoring any irregularity appearing in the domain where $|f_\rho(x)| > (e^2-1)/(e^2+1)$. This can be seen from the following example where f_ρ has a singularity at 0 while $f_\rho^{\phi_{eh}} \equiv 1$ is C^∞.

Example 1 *Let $X = [-1,1]$, $0 < \alpha < \frac{1}{14}$ and ρ be the distribution given by $d\rho_X = \frac{1}{2}dx$ and $f_\rho(x) = 1 - \frac{1}{5}|x|^\alpha$ which means $P(y = 1|x) = 1 - \frac{1}{10}|x|^\alpha$. It is well known that the function $|x|^\alpha$ lies in the Sobolev space $H^s(X)$ if and only if $s < \alpha + \frac{1}{2}$. So regularity assumption (8) is satisfied for ϕ_{ls} if and only if $s < \alpha + \frac{1}{2}$. Then from Theorem 1, we see the learning rate $\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) = O(m^{-\theta_{ls}} \log \frac{2}{\delta})$ with $\theta_{ls} = \frac{s}{s+2+2}$ arbitrarily close to $\frac{1+2\alpha}{9+2\alpha} < \frac{1}{8}$. However, for the exponential-hinge loss ϕ_{eh}, we have $f_\rho^{\phi_{eh}} \equiv 1$ which follows from expression (23) and the definition $f_\rho(x) = 1 - \frac{1}{5}|x|^\alpha \ge$*

$1-\frac{1}{5} > (e^2-1)/(e^2+1)$ *on* X. *Therefore, regularity assumption (8) is satisfied for an arbitrarily large s and Theorem 5 yields the learning rate* $\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) = O(m^{-\theta_{eh}} \log \frac{2}{\delta})$ *with* θ_{eh} *arbitrarily close to* $\frac{1}{4}$. *Thus for learning algorithm (2), the exponential-hinge loss has some advantages over* ϕ_{ls} *(and* ϕ_r *as shown in the next example).*

The dependence of the function $f_\rho^{\phi_r}$ on f_ρ involves a power function $u \to u^{1/(r-1)}$ which might cause irregularity. This is demonstrated by the following example where the singularity of the function f_ρ at 0 is worsened for the function $f_\rho^{\phi_r}$ when r is large.

Example 2 *Let* X *and* ρ *be as in Example 1. When* $\phi = \phi_r$ *with* $r > 2$, *the function* $f_\rho^{\phi_r}$ *in (21) equals*

$$f_\rho^{\phi_r}(x) = \frac{(2-\frac{1}{5}|x|^\alpha)^{1/(r-1)} - (\frac{1}{5}|x|^\alpha)^{1/(r-1)}}{(2-\frac{1}{5}|x|^\alpha)^{1/(r-1)} + (\frac{1}{5}|x|^\alpha)^{1/(r-1)}}.$$

Regularity assumption (8) is satisfied for ϕ_r *if and only if* $s < \frac{\alpha}{r-1} + \frac{1}{2}$. *Then Theorem 4 yields the learning rate* $\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) = O(m^{-\theta_r} \log \frac{2}{\delta})$ *with* $\theta_r = \frac{s}{4(s(1-1/r)+1+1)}$ *arbitrarily close to* $\frac{r(2\alpha+r-1)}{4(5r+2\alpha-1)(r-1)}$. *This power index is always less than that of* ϕ_{ls} *or* ϕ_{eh}. *It shows that the loss* ϕ_r *has worse performance than* ϕ_{ls} *and* ϕ_{eh}, *at least for some distributions.*

5. Further Discussion

Let us discuss further generalizations and connections briefly here. More details will be provided in our future study.

The first extension is to a manifold setting. If X is a connected compact C^∞ submanifold of $\mathbb{R}^n$ without boundary and its dimension is $d \le n$, then the covering number estimate (15) holds with n replaced by the manifold dimension d. Proposition 2 and Lemma 2 are still valid with n replaced by d. Learning rates in Theorems 1, 4 and 5 can be improved with n replaced by d if approximation error estimates similar to Theorem 2 can be established in the manifold setting. One can use ideas for convolution type approximation schemes on $\mathbb{R}^n$ (Pan, et al., 2008) to define higher order operators on manifolds and then get estimates for the regularization error.

The second connection is to multi-kernel regularization schemes (Wu et al., 2007; Argyriou et al., 2006; Chapelle et al., 2002) defined as

$$f_{\mathbf{z},\lambda}^{\phi} = \arg \min_{0<\sigma<\infty} \min_{f \in \mathcal{H}_\sigma} \Big\{ \frac{1}{m} \sum_{i=1}^{m} \phi(y_i f(x_i)) + \lambda \| f \|_{\mathcal{H}_\sigma}^2 \Big\}.$$

In this scheme the variance parameter σ is chosen automatically while the learning rate derived in Ying and Zhou (2007) is at most $O(m^{-1/6})$. It would be interesting to investigate how to choose the parameter σ in (2).

The last questions is about more general loss functions. In our analysis we assume that the convex loss ϕ has a zero which excludes the logistic loss $\phi(t) = \log(1+e^{-t})$. One might generalize our analysis to get some error bounds for the scheme with loss functions without zero by using a general projection operator π_M with level $M > 0$ given by

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

Acknowledgments

We would like to thank the referees for their constructive suggestions and comments. The work described in this paper was partially supported by a grant from the Research Grants Council of Hong Kong [Project No. CityU 104007] and National Science Fund for Distinguished Young Scholars of China [Project No. 10529101]. The corresponding author is Ding-Xuan Zhou.

Appendix A. Approximation Scheme by Gaussians

This appendix provides a proof of Theorem 2 which is a corollary of the following more general theorem. The approximation error is estimated by means of a convolution type scheme constructed by Gaussians with a Fourier analysis technique (Schaback and Werner, 2006; Steinwart and Scovel, 2007; Steinwart et al. , 2006).

Theorem 6 *Assume that for some $s > 0$, f_ρ^ϕ is the restriction of some $\tilde{f}_\rho^\phi \in H^s(\mathbb{R}^n)$ onto X, and the density function $g = \frac{d\rho_X}{dx}$ exists and lies in $L^2(X)$.*

(1) If $\tilde{f}_\rho^\phi \in L^\infty(\mathbb{R}^n)$, then we can find a set of functions $\{f_{\sigma,\lambda} \in \mathcal{H}_\sigma\}_{0\le\sigma\le1,\lambda>0}$ such that

$$\|f_{\sigma,\lambda}\|_{L^\infty(X)} \le \widetilde{B}, \tag{25}$$

$$\mathcal{D}(\sigma,\lambda) \le \widetilde{B}(\sigma^s + \lambda\sigma^{-n}), \quad \forall\, 0<\sigma\le 1, \lambda>0, \tag{26}$$

where $\widetilde{B}$ is a constant independent of σ or λ.

(2) If for some $r \ge 1$ and $C_\phi > 0$,

$$|\phi'_+(t)| \le C_\phi |t|^{r-1} \qquad \forall |t| \ge 1. \tag{27}$$

then we can find $\{f_{\sigma,\lambda} \in \mathcal{H}_\sigma\}$ such that

$$\|f_{\sigma,\lambda}\|_{L^\infty(X)} \le \widetilde{B}'\sigma^{-\frac{n}{2}}, \tag{28}$$

$$\mathcal{D}(\sigma,\lambda) \le \widetilde{B}'(\sigma^{s-\frac{n(r-1)}{2}} + \lambda\sigma^{-n}), \quad \forall\, 0<\sigma\le 1, \lambda>0, \tag{29}$$

where $\widetilde{B}'$ is a constant independent of σ or λ.

Proof Take some trigonometric polynomial $\tilde{a}(\xi) = \sum_{j\in J} a_j e^{-ij\cdot\xi}$ on $\mathbb{R}^n$ with a finite subset J of $\mathbb{Z}^n$ such that for some $C_s > 0$ depending only on s and n, we have

$$|e^{-\frac{|\xi|^2}{2}}\tilde{a}(\xi) - 1| \le C_s|\xi|^s \quad \forall\, \xi \in \mathbb{R}^n.$$

This can be done by choosing the coefficients $(a_j)_{j\in J}$ of $\tilde{a}$ satisfying the linear system

$$\tilde{a}(0) = 1 \quad \text{and} \quad D^\alpha(e^{-\frac{|\xi|^2}{2}}\tilde{a}(\xi))(0) = 0, \qquad \alpha \in \mathbb{Z}^n, 0 < |\alpha| < s.$$

So J and $(a_j)_{j\in J}$ depend only on s and n.

Define

$$\tilde{f}_\sigma(x) = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n \int_{\mathbb{R}^n} K^\sigma(x,y) \sum_{j\in J} a_j \tilde{f}_\rho^\phi(y - \sigma j)\, dy, \quad x \in \mathbb{R}^n.$$

We first estimate $\|\tilde{f}_\sigma - \tilde{f}_\rho^\phi\|_{L^2(\mathbb{R}^n)}$.

Define a function $\tilde{k}^\sigma$ on $\mathbb{R}^n$ by $\tilde{k}^\sigma(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{|x|^2}{2\sigma^2}}$, we know that $\hat{\tilde{k}^\sigma}(\xi) = e^{-\frac{|\sigma\xi|^2}{2}}$ and $\tilde{f}_\sigma(x) = \tilde{k}^\sigma * (\sum_{j\in J} a_j \tilde{f}_\rho^\phi(\cdot - \sigma j))$. This in connection with the fact that the Fourier transform of $\tilde{f}_\rho^\phi(\cdot - \sigma j)$ equals $e^{-i\sigma j\cdot\xi}\hat{\tilde{f}}_\rho^\phi(\xi)$ implies

$$\hat{\tilde{f}}_\sigma(\xi) = \hat{\tilde{k}^\sigma}(\xi)\sum_{j\in J} a_j e^{-i\sigma j\cdot\xi}\hat{\tilde{f}}_\rho^\phi(\xi) = e^{-\frac{|\sigma\xi|^2}{2}}\tilde{a}(\sigma\xi)\hat{\tilde{f}}_\rho^\phi(\xi).$$

It follows that

$$\begin{aligned}\|\tilde{f}_\sigma - \tilde{f}_\rho^\phi\|^2_{L^2(\mathbb{R}^n)} &= (2\pi)^{-n}\|\hat{\tilde{f}}_\sigma - \hat{\tilde{f}}_\rho^\phi\|^2_{L^2(\mathbb{R}^n)} = (2\pi)^{-n}\int_{\mathbb{R}^n}|e^{-\frac{|\sigma\xi|^2}{2}}\tilde{a}(\sigma\xi) - 1|^2|\hat{\tilde{f}}_\rho^\phi(\xi)|^2\,d\xi \\ &\le (2\pi)^{-n}C_s^2\int_{\mathbb{R}^n}|\sigma\xi|^{2s}|\hat{\tilde{f}}_\rho^\phi(\xi)|^2\,d\xi \le C_s^2\sigma^{2s}(2\pi)^{-n}\int_{\mathbb{R}^n}|\xi|^{2s}|\hat{\tilde{f}}_\rho^\phi(\xi)|^2\,d\xi.\end{aligned}$$

That is,

$$\|\tilde{f}_\sigma - \tilde{f}_\rho^\phi\|_{L^2(\mathbb{R}^n)} \le C_s\|\tilde{f}_\rho^\phi\|_{H^s(\mathbb{R}^n)}\sigma^s. \tag{30}$$

Then we bound $\|\tilde{f}_\sigma\|_{\mathcal{H}_\sigma(\mathbb{R}^n)}$. Here $\mathcal{H}_\sigma(\mathbb{R}^n)$ is the RKHS generated by the Mercer kernel $K^\sigma(x,y)$ on $\mathbb{R}^n$. By the inner product in $\mathcal{H}_\sigma(\mathbb{R}^n)$, we know that $\langle K^\sigma(\cdot,y), K^\sigma(\cdot,z)\rangle_{\mathcal{H}_\sigma(\mathbb{R}^n)} = K^\sigma(y,z)$. So we have

$$\|\tilde{f}_\sigma\|^2_{\mathcal{H}_\sigma(\mathbb{R}^n)} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{2n}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}K^\sigma(y,z)\sum_{j\in J}a_j\tilde{f}_\rho^\phi(y-\sigma j)\,dy\sum_{l\in J}a_l\tilde{f}_\rho^\phi(z-\sigma l)\,dz.$$

By the elementary inequality $|uv| \le \frac{u^2+v^2}{2}$ and $\int_{\mathbb{R}^n}K^\sigma(y,z)\,dz = (\sqrt{2\pi}\sigma)^n$, we see that

$$\begin{aligned}\|\tilde{f}_\sigma\|^2_{\mathcal{H}_\sigma(\mathbb{R}^n)} &\le \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{2n}\sum_{j\in J}\sum_{l\in J}|a_j||a_l|\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}K^\sigma(y,z)\frac{|\tilde{f}_\rho^\phi(y-\sigma j)|^2 + |\tilde{f}_\rho^\phi(z-\sigma l)|^2}{2}\,dydz \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{2n}\sum_{j,l\in J}|a_j||a_l|(\sqrt{2\pi}\sigma)^n\|\tilde{f}_\rho^\phi\|^2_{L^2(\mathbb{R}^n)}.\end{aligned}$$

That is,

$$\|\tilde{f}_\sigma\|_{\mathcal{H}_\sigma(\mathbb{R}^n)} \le C_s'\|\tilde{f}_\rho^\phi\|_{L^2(\mathbb{R}^n)}(\sqrt{2\pi}\sigma)^{-\frac{n}{2}}$$

where $C_s' := \sum_{j\in J}|a_j|$ is a constant depending only on s and n.

Take $f_{\sigma,\lambda} = \tilde{f}_\sigma|_X$, the restriction of $\tilde{f}_\sigma$ onto X. By basic facts about RKHS (Aronszajn, 1950), we know that $f_{\sigma,\lambda} = \tilde{f}_\sigma|_X \in \mathcal{H}_\sigma$ and

$$\|f_{\sigma,\lambda}\|_{\mathcal{H}_\sigma} \le \|\tilde{f}_\sigma\|_{\mathcal{H}_\sigma(\mathbb{R}^n)} \le C_s'\|\tilde{f}_\rho^\phi\|_{L^2(\mathbb{R}^n)}(\sqrt{2\pi}\sigma)^{-\frac{n}{2}}. \tag{31}$$

Now we can derive the desired bounds.

(1) When $\tilde{f}_\rho^\phi \in L^\infty(\mathbb{R}^n)$, for any $x\in\mathbb{R}^n$, we have

$$|\tilde{f}_\sigma(x)| \le \sum_{j\in J}|a_j|\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\int_{\mathbb{R}^n}e^{-\frac{|x-y|^2}{2\sigma^2}}\,dy\|\tilde{f}_\rho^\phi\|_{L^\infty(\mathbb{R}^n)} = C_s'\|\tilde{f}_\rho^\phi\|_{L^\infty(\mathbb{R}^n)}.$$

It follows that (25) holds true and

$$\begin{aligned}\mathcal{E}^{\phi}(f_{\sigma,\lambda})-\mathcal{E}^{\phi}(f_{\rho}^{\phi}) &\leq \int_X\int_Y \sup\Big\{|\phi'_+(\xi)|:|\xi|\leq \max\{C'_s\|\tilde{f}_{\rho}^{\phi}\|_{L^\infty(\mathbb{R}^n)},\|f_{\rho}^{\phi}\|_{L^\infty(X)}\}\Big\}\\ &\quad |f_{\sigma,\lambda}(x)-f_{\rho}^{\phi}(x)|d\rho(y|x)g(x)dx.\end{aligned}$$

By the Schwarz inequality we see that

$$\mathcal{E}^{\phi}(f_{\sigma,\lambda})-\mathcal{E}^{\phi}(f_{\rho}^{\phi})\leq \sup\{|\phi'_+(\xi)|:|\xi|\leq (C'_s+1)\|\tilde{f}_{\rho}^{\phi}\|_{L^\infty(\mathbb{R}^n)}\}\|f_{\sigma,\lambda}-f_{\rho}^{\phi}\|_{L^2(X)}\|g\|_{L^2(X)}.$$

This bound in connection with (30) and (31) implies (26) with the constant $\widetilde{B}$ given by

$$\begin{aligned}\widetilde{B} &= \max\Big\{C'_s\|\tilde{f}_{\rho}^{\phi}\|_{L^\infty(\mathbb{R}^n)},(C'_s)^2\|\tilde{f}_{\rho}^{\phi}\|^2_{L^2(\mathbb{R}^n)}(2\pi)^{-\frac{n}{2}},\\ &\qquad \sup\{|\phi'_+(\xi)|:|\xi|\leq (C'_s+1)\|\tilde{f}_{\rho}^{\phi}\|_{L^\infty(\mathbb{R}^n)}\}C_s\|\tilde{f}_{\rho}^{\phi}\|_{H^s(\mathbb{R}^n)}\|g\|_{L^2(X)},\Big\}.\end{aligned}$$

(2) Without the condition $\tilde{f}_{\rho}^{\phi}\in L^\infty(\mathbb{R}^n)$, we bound $\|f_{\sigma,\lambda}\|_{L^\infty(X)}$ directly from the expression of $\tilde{f}_\sigma$. For $x\in\mathbb{R}^n$, we have

$$\begin{aligned}|\tilde{f}_\sigma(x)| &\leq \sum_{j\in J}|a_j|\Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n\Big\{\int_{\mathbb{R}^n}(K^\sigma(x,y))^2\,dy\Big\}^{\frac12}\Big\{\int_{\mathbb{R}^n}|\tilde{f}_{\rho}^{\phi}(y-\sigma j)|^2\,dy\Big\}^{\frac12}\\ &= \sum_{j\in J}|a_j|(2\sqrt{\pi}\sigma)^{-\frac{n}{2}}\|\tilde{f}_{\rho}^{\phi}\|_{L^2(\mathbb{R}^n)}.\end{aligned}$$

It follows that $\|f_{\sigma,\lambda}\|_{L^\infty(X)}\leq C'_s(2\sqrt{\pi})^{-\frac{n}{2}}\|\tilde{f}_{\rho}^{\phi}\|_{L^2(\mathbb{R}^n)}\sigma^{-\frac{n}{2}}$.

To derive the bound for the excess generalization error, we notice from (27) that

$$\begin{aligned}&\mathcal{E}^{\phi}(f_{\sigma,\lambda})-\mathcal{E}^{\phi}(f_{\rho}^{\phi})=\int_X\int_Y(\phi(yf_{\sigma,\lambda}(x))-\phi(yf_{\rho}^{\phi}(x)))\,d\rho(y|x)d\rho_X(x)\\ &\leq \int_X\int_Y \sup\Big\{|\phi'_+(\xi)|:|\xi|\leq \max\{C'_s(2\sqrt{\pi}\sigma)^{-\frac{n}{2}}\|\tilde{f}_{\rho}^{\phi}\|_{L^2(\mathbb{R}^n)},\|f_{\rho}^{\phi}\|_{L^\infty(X)}\}\Big\}\\ &\quad |f_{\sigma,\lambda}(x)-f_{\rho}^{\phi}(x)|d\rho(y|x)g(x)dx.\end{aligned}$$

If we denote $C''_s=\max\{C_\phi,C_\phi\|f_{\rho}^{\phi}\|^{r-1}_{L^\infty(X)},C_\phi[C'_s(2\sqrt{\pi})^{-\frac{n}{2}}\|\tilde{f}_{\rho}^{\phi}\|_{L^2(\mathbb{R}^n)}]^{r-1}\}$, then

$$\mathcal{E}^{\phi}(f_{\sigma,\lambda})-\mathcal{E}^{\phi}(f_{\rho}^{\phi})\leq C''_s\sigma^{-\frac{n(r-1)}{2}}\|f_{\sigma,\lambda}-f_{\rho}^{\phi}\|_{L^2(X)}\|g\|_{L^2(X)}\leq C''_s\|g\|_{L^2(X)}C_s\|\tilde{f}_{\rho}^{\phi}\|_{H^s(\mathbb{R}^n)}\sigma^{s-\frac{n(r-1)}{2}}.$$

Therefore,

$$\mathcal{D}(\sigma,\lambda)\leq C''_s\|g\|_{L^2(X)}C_s\|\tilde{f}_{\rho}^{\phi}\|_{H^s(\mathbb{R}^n)}\sigma^{s-\frac{n(r-1)}{2}}+\lambda(C'_s)^2\|\tilde{f}_{\rho}^{\phi}\|^2_{L^2(\mathbb{R}^n)}(\sqrt{2\pi}\sigma)^{-n}.$$

This verifies (28) and (29) by taking

$$\widetilde{B'}=\max\Big\{C'_s(2\sqrt{\pi})^{-\frac{n}{2}}\|\tilde{f}_{\rho}^{\phi}\|_{L^2(\mathbb{R}^n)},C''_s\|g\|_{L^2(X)}C_s\|\tilde{f}_{\rho}^{\phi}\|_{H^s(\mathbb{R}^n)},(C'_s)^2\|\tilde{f}_{\rho}^{\phi}\|^2_{L^2(\mathbb{R}^n)}(2\pi)^{-n/2}\Big\}.$$

The proof of Theorem 6 is complete. ■

In our main results, only part (1) of Theorem 6, that is, Theorem 2 is used. A main assumption, condition (8), gives the restriction $\widetilde{f}_\rho^\phi \in H^s(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$. When we do not know whether f_ρ^ϕ can be extended to a uniformly bounded function on $\mathbb{R}^n$, we can use part (2) of Theorem 6. This might be the case when $\phi = \phi_h$, as mentioned in the following.

A geometric noise condition was introduced in Steinwart and Scovel (2007). This condition with exponent $\alpha > 0$ means

$$\int_X |f_\rho(x)| \exp\left\{-\frac{\tau_x^2}{t}\right\} d\rho_X(x) = O\left(t^{\frac{\alpha n}{2}}\right) \tag{32}$$

where

$$\tau_x = \begin{cases} \inf_{f_\rho(u)\geq 0} |x-u|, & \text{if } f_\rho(x) < 0, \\ \inf_{f_\rho(u)\leq 0} |x-u|, & \text{if } f_\rho(x) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

It does not assume smoothness of functions. An interesting result in Steinwart and Scovel (2007) asserts that when $\phi = \phi_h$, geometric noise condition (32) with exponent $0 < \alpha < \infty$ leads to $\mathcal{D}(\sigma,\lambda) \leq \widetilde{B}'''(\sigma^{\alpha n} + \lambda\sigma^{-n})$. With this estimate, under Tsybakov noise condition (5), learning rates are obtained in Steinwart and Scovel (2007). For example, when $\alpha > \frac{q+2}{2q}$, for an arbitrarily small $\varepsilon > 0$, with confidence $1-\delta$,

$$\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leq \widetilde{C}_\varepsilon \left(\log\frac{4}{\delta}\right)^2 \left(\frac{1}{m}\right)^{\frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4}-\varepsilon}. \tag{33}$$

Since no Sobolev smoothness is assumed for $f_\rho^{\phi_h} = f_c$ (Wahba, 1990), we need to use the regularizing function $\widetilde{f}_{\sigma,\lambda}$ defined by (7) and derive by some detailed computations that with confidence $1-\delta$,

$$\mathcal{R}(sgn(f_{\mathbf{z}})) - \mathcal{R}(f_c) \leq \widetilde{C}_{\rho,h} \log\frac{4}{\delta} \left(\frac{1}{m}\right)^{\frac{(q+1)\alpha n}{(q+2)\alpha n+2(q+1)(n+1)}}.$$

This rate is slightly worse than (33), though the estimate for the confidence is slightly better. It raises the question of improving Theorem 6 under various noise conditions.

Appendix B. Role of Tight Bounds for Covering Numbers

In this appendix we prove Lemma 2 which shows a special role of the tight bound (15) for covering numbers concerning Gaussian kernels. In fact, we have the following more general result.

Proposition 4 *Let $\Delta \geq 1$ be arbitrary. Then $\varepsilon^*(m,\lambda,\sigma,\delta/2)$ defined by (16) satisfies*

$$\varepsilon^*(m,\lambda,\sigma,\delta/2) \leq \widetilde{C}_2 \left(\frac{\log\frac{2}{\delta} + \sigma^{-2(n+1)/(2-\tau)} + (\log m)^{(n+1)/(2-\tau)}}{m^{\frac{1}{2-\tau}}} + \frac{\sigma^{-2(n+1)}}{m} + \frac{\sqrt{\phi(0)}}{\sqrt{\lambda} m^{\Delta}}\right),$$

where $\widetilde{C}_2$ is the constant depending on $C_0, C_1, \tau, \phi'_+(-1), \phi(-1), \Delta$ and is given by

$$\widetilde{C}_2 = \max\left\{|\phi'_+(-1)|, (4C_1)^{\frac{1}{2-\tau}}, 2(4C_0C_1)^{\frac{1}{2-\tau}}\Delta^{n+1}, 2\phi(-1)(1+C_0+C_0\Delta^{n+1})\right\}.$$

Proof Observe from (15) that as a function on $(0,+\infty)$, the logarithm of the middle term of (16) is bounded by

$$h(\varepsilon) := C_0\Big(\Big(\log\frac{\sqrt{\phi(0)}|\phi'_+(-1)|}{\sqrt{\lambda}\varepsilon}\Big)^{n+1} + \frac{1}{\sigma^{2(n+1)}}\Big) - g(\varepsilon),$$

where g is the strictly increasing function on $(0,\infty)$ defined by

$$g(\varepsilon) = \frac{m\varepsilon^{2-\tau}}{2C_1 + \frac{2}{3}\phi(-1)\varepsilon^{1-\tau}}.$$

Set

$$\begin{aligned}\mathcal{B} &= \frac{\sqrt{\phi(0)}|\phi'_+(-1)|}{\sqrt{\lambda}m^{\Delta}} + \left(\frac{4C_1(\log\frac{2}{\delta} + \frac{C_0}{\sigma^{2(n+1)}}) + 4C_0C_1(\Delta\log m)^{n+1}}{m}\right)^{\frac{1}{2-\tau}} \\ &\quad + \frac{4\phi(-1)}{3m}\left(\log\frac{2}{\delta} + \frac{C_0}{\sigma^{2(n+1)}} + C_0(\Delta\log m)^{n+1}\right).\end{aligned}$$

If $\frac{2}{3}\phi(-1)\mathcal{B}^{1-\tau} \le 2C_1$, then

$$g(\mathcal{B}) \ge \frac{m\mathcal{B}^{2-\tau}}{4C_1} \ge \log\frac{2}{\delta} + \frac{C_0}{\sigma^{2(n+1)}} + C_0(\Delta\log m)^{n+1}.$$

If $\frac{2}{3}\phi(-1)\mathcal{B}^{1-\tau} > 2C_1$, then

$$g(\mathcal{B}) \ge \frac{m\mathcal{B}^{2-\tau}}{\frac{4}{3}\phi(-1)\mathcal{B}^{1-\tau}} = \frac{m\mathcal{B}}{\frac{4}{3}\phi(-1)} \ge \log\frac{2}{\delta} + \frac{C_0}{\sigma^{2(n+1)}} + C_0(\Delta\log m)^{n+1}.$$

Thus in either case we have

$$g(\mathcal{B}) \ge \log\frac{2}{\delta} + \frac{C_0}{\sigma^{2(n+1)}} + C_0(\Delta\log m)^{n+1}.$$

On the other hand, since $\mathcal{B} \ge \frac{\sqrt{\phi(0)}|\phi'_+(-1)|}{\sqrt{\lambda}m^{\Delta}}$, we also see that $\log\frac{\sqrt{\phi(0)}|\phi'_+(-1)|}{\mathcal{B}\sqrt{\lambda}} \le \Delta\log m$. It follows that

$$h(\mathcal{B}) \le C_0(\Delta\log m)^{n+1} - \log\frac{2}{\delta} - C_0(\Delta\log m)^{n+1} = \log\frac{\delta}{2}.$$

But the function h is strictly decreasing. So $\varepsilon^*(m,\lambda,\sigma,\delta/2) \le \mathcal{B}$. The the desired bound for $\varepsilon^*(m,\lambda,\sigma,\delta/2)$ follows with the constant $\widetilde{C}_2$. The proof of Proposition 4 is complete. ■

Now we can prove Lemma 2 by the special form of λ,σ.

B.1 Proof of Lemma 2

Take $\Delta = \frac{\gamma}{2} + 1$ in Proposition 4. Then we know from the special form (3) of λ and σ that

$$\varepsilon^*(m,\lambda,\sigma,\delta/2) \le \widetilde{C}_2\left\{\frac{\log\frac{2}{\delta}}{m^{\frac{1}{2-\tau}}} + \Big(\frac{1}{m}\Big)^{\frac{1-2\gamma_2^*(n+1)}{2-\tau}} + \Big(\frac{(\log m)^{n+1}}{m}\Big)^{\frac{1}{2-\tau}} + \frac{\sqrt{\phi(0)}}{m}\right\}.$$

notation	meaning	pages
$\mathcal{R}(\mathcal{C})$	misclassification error for a classifier $\mathcal{C}$	1447
f_c	Bayes rule which minimizes $\mathcal{R}$	1447
$\mathcal{R}(\mathcal{C})-\mathcal{R}(f_c)$	excess misclassification error	1447
ϕ	loss function for classification	1448
σ	variance parameter for the Gaussian kernel	1448, 1449, 1458
λ	regularization parameter	1448, 1449, 1458
$f_{\mathbf{z}}$	learning scheme (2)	1448
$\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z}}))-\mathcal{R}(f_c)$	excess misclassification error for classifier $\mathrm{sgn}(f_{\mathbf{z}})$	1448, 1449, 1458
f_ρ	regression function	1448, 1449, 1457
$\mathcal{E}^\phi(f)$	generalization error for a function f	1450
f_ρ^ϕ	minimizer of $\mathcal{E}^\phi$	1450
$\mathcal{E}^\phi(f)-\mathcal{E}^\phi(f_\rho^\phi)$	excess generalization error for a function f	1451, 1452, 1454, 1457
$f_{\sigma,\lambda}$	regularizing function constructed in Theorem 2	1450, 1452, 1454, 1461
$\mathcal{D}(\sigma,\lambda)$	regularization error or approximation error	1450, 1452, 1456, 1461
$\tau=\tau_{\phi,\rho}$	variancing power defined in Definition 5	1453, 1454, 1456, 1458

Table 1: NOTATIONS

Observe the elementary inequality (Yao, 2008; Ye and Zhou, 2007)

$$\exp\{-cx\}\le(\frac{a}{ec})^a x^{-a}\quad\forall x,c,a>0.$$

Taking $x=\log m$, $a=n+1$ and $c=2\gamma\zeta(n+1)$, we have

$$(\log m)^{n+1}\le(\frac{1}{2e\gamma\zeta})^{n+1}m^{2\gamma\zeta(n+1)}.$$

Hence (17) holds true with the constant $C_2=\widetilde{C}_2(4+2\sqrt{\phi(0)})$. The proof of Lemma 2 is complete. ■

References

A. Argyriou, R. Hauser, C. A. Micchelli, and M. Pontil. A DC-programming algorithm for kernel selection. *Proceedings of the Twenty-Third International Conference on Machine Learning*, 2006.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

O. Chapelle, V. N. Vapnik, O. Bousquet and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46: 131–159, 2002.

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

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

F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5:59–85, 2006.

E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5:1363–1390, 2004.

D. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, 1996.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

D. Hardin, I. Tsamardinos, and C. F. Aliferis. A theoretical characterization of linear SVM-based feature selection. *Proc. of the 21st Int. Conf. on Machine Learning*, Banff, Canada, 2004.

S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for empirical risk minimization. *Advances in Computational Mathematics*, 25:161–193, 2006.

Z. W. Pan, D. H. Xiang, Q. W. Xiao, and D. X. Zhou. Parzen windows for multi-class classification. *Journal of Complexity*, 24:606–618, 2008.

R. Schaback and J. Werner. Linearly constrained reconstruction of functions by kernels, with applications to machine learning. *Advances in Computational Mathematics*, 25:237–258, 2006.

S. Smale and D. X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007.

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

S. Smale and D. X. Zhou. Shannon sampling and function reconstruction from point values. *Bulletin of the American Mathematical Society*, 41:279–305, 2004.

I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Annals of Statistics*, 35:575–607, 2007.

I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52:4635–4643, 2006.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Q. Wu, Y. Ying, and D. X. Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23:108–134, 2007.

Q. Wu and D. X. Zhou. Analysis of support vector machine classification. *Journal of Computational Analysis and Applications*, 8:99–119, 2006.

Q. Wu and D. X. Zhou. SVM soft margin classifiers: linear programming versus quadratic programming. *Neural Computation*, 17:1160–1187, 2005.

Y. Yao. On complexity issue of online learning algorithms. *IEEE Transactions on Information Theory*, to appear.

G. B. Ye and D. X. Zhou. Fully online classification by regularization. *Applied and Computational Harmonic Analysis*, 23:198–214, 2007.

G. B. Ye and D. X. Zhou. Learning and approximation by Gaussians on Riemannian manifolds. *Advances in Computational Mathematics*, 29:291–310, 2008.

Y. Ying. Convergence analysis of online algorithms. *Advances in Computational Mathematics*, 27:273–291, 2007.

Y. Ying and D. X. Zhou. Learnability of Gaussians with flexible variances. *Journal of Machine Learning Research*, 8:249–276, 2007.

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

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

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

D. X. Zhou and K. Jetter. Approximation with polynomial kernels and SVM classifiers. *Advances in Computational Mathematics*, 25:323–344, 2006.