Topics in Multivariate Approximation and Interpolation
K. Jetter et al., Editors
© 2006 Elsevier B.V. All rights reserved

Learning Theory: From Regression to Classification

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

Department of Mathematics, City University of Hong Kong,
83 Tat Chee Avenue, Kowloon, Hong Kong, P. R. China

Abstract

We give a brief survey of regularization schemes in learning theory for the purposes of regression and classification, from an approximation theory point of view. First, the classical method of empirical risk minimization is reviewed for regression with a general convex loss function. Next, we explain ideas and methods for the error analysis of regression algorithms generated by Tikhonov regularization schemes associated with reproducing kernel Hilbert spaces. Then binary classification algorithms given by regularization schemes are described with emphasis on support vector machines and noise conditions for distributions. Finally, we mention further topics and some open problems in learning theory.

Key words: Learning theory, regression, classification, regularization scheme, reproducing kernel Hilbert space, error analysis

2000 MSC: 68T05, 62J02

Email addresses: 50004455@student.cityu.edu.hk (Qiang Wu),
ymying@cityu.edu.hk (Yiming Ying),
mazhou@math.cityu.edu.hk (Ding-Xuan Zhou).

[1] Corresponding author

1. Introduction

Learning theory started with the classical work on statistical learning theory and support vector machines [75]. It studies learning of function relations from samples and provides powerful learning algorithms for the purpose of regression, classification, feature selection, and many other applications [2,26,76]. These learning algorithms are mostly reduced to well studied convex optimization procedures and are very effective for practical problems of large sample size and huge dimensions [31,51,59]. One example is support vector machine algorithms which are implemented by convex quadratic programming optimizations and can handle large size samples [75,20]. Compared with the vast applications of learning algorithms due to fast developed computing tools, the theoretical analysis of machine learning procedures is still in its starting stage. The purpose of this paper is to give a brief survey on recent developments of learning theory including regression and classification, from an approximation theory point of view.

2. Empirical Risk Minimization for Regression

The main idea of the classical statistical learning theory is Empirical Risk Minimization (ERM). We shall describe this method for regression [62,31,21].

The goal of the regression problem is to find a desired function $f^* : X \to Y$ or its approximation from samples. Here X is a complete metric space (e.g. a subset of $\mathbb{R}^n$) and $Y = \mathbb{R}$. The sample set $\mathbf{z} = \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$ is produced by the desired (unknown) function f^*. To allow noise, we would not expect the sample function value y_i to be exactly $f^*(x_i)$, but its approximation $y_i \approx f^*(x_i)$. The most adopted model in learning theory is to assume that the samples are drawn according to a probability distribution ρ on the product space $Z := X \times Y$, and that the desired function f^* is its regression function f_ρ. If we denote ρ_X as the marginal distribution of ρ on X and $\rho(\cdot|x)$ its conditional distribution at $x \in X$, then the *regression function* is defined as

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X.$$

That is, for a fixed point $x \in X$, the conditional probability distribution $\rho(\cdot|x)$ on $Y = \mathbb{R}$ reflects the noise while its mean is exactly the value of the desired function at x: $f^*(x) = f_\rho(x)$.

To learn the function f_ρ, we draw the set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ of m random samples according to ρ. Various learning algorithms can be applied to the samples and produce approximations of the regression function. ERM forms a rich family of learning algorithms.

The idea of ERM learning algorithms is to minimize empirical errors with respect to a loss function ψ. The loss function measures how the sample value y approximates the function value $f(x)$ by evaluating $\psi(y - f(x))$.

Definition 1. We say that $\psi : \mathbb{R} \to \mathbb{R}_+$ is a *regressing loss function* if it is even, convex, and continuous with $\psi(0) = 0$.

For $(x, y) \in Z$, the value $\psi(y - f(x))$ is the local error suffered from the use of f as a model for the process producing y at x. The condition $\psi(0) = 0$ ensures a zero error when $y = f(x)$. Examples of loss functions for regression include the least-square loss and Vapnik's ϵ-insensitive norm.

Example 2. The least-square loss corresponds to the loss function $\psi(t) = t^2$. For $\epsilon > 0$, the *ϵ-insensitive norm* is the loss function defined by

$$\psi(t) = \psi_\epsilon(t) = \begin{cases} |t| - \epsilon, & \text{if } |t| > \epsilon, \\ 0, & \text{otherwise.} \end{cases}$$

The *error* (generalization error) associated with the loss function ψ is defined by

$$\mathcal{E}(f) = \int_Z \psi(y - f(x)) d\rho. \tag{1}$$

2.1. Example: Regression with the least-square loss

To show the basic idea of ERM, consider the least-square method. Here the error or risk $\mathcal{E}(f)$ for a function $f : X \to Y$ takes the form

$$\mathcal{E}(f) = \int_Z \psi(y - f(x)) d\rho = \int_Z (y - f(x))^2 d\rho. \tag{2}$$

Separating ρ into the marginal distribution and conditional distributions, we see from $\int_Y \{y - f_\rho(x)\} d\rho(y|x) = 0$ that

$$\begin{aligned} \mathcal{E}(f) &= \int_X \left\{ \int_Y (y - f_\rho(x))^2 d\rho(y|x) + (f_\rho(x) - f(x))^2 \right\} d\rho_X(x). \\ &= \mathcal{E}(f_\rho) + \int_X (f(x) - f_\rho(x))^2 d\rho_X(x) \geq \mathcal{E}(f_\rho). \end{aligned}$$

If we denote the weighted L^2 norm as $\|f\|_{L^2_{\rho_X}} = (\int_X |f(x)|^2 d\rho_X)^{1/2}$, then

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}} \geq 0. \tag{3}$$

It means that the regression function f_ρ minimizes the error.

Since the samples are drawn according to ρ, the law of large numbers tells us that for a fixed function $f : X \to Y$, the random variable $\xi(z) = (y - f(x))^2$ on (Z, ρ) satisfies $\frac{1}{m} \sum_{i=1}^m \xi(z_i) \to E(\xi)$ in probability. But $E(\xi) = \int_Z \xi(z) d\rho$ is exactly $\mathcal{E}(f)$, the error of f, while $\frac{1}{m} \sum_{i=1}^m \xi(z_i) = \frac{1}{m} \sum_{i=1}^m (y_i - f(x_i))^2$. We denote this as $\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{i=1}^m (y_i - f(x_i))^2$ and call it the empirical error of f at $\mathbf{z}$. Since $\mathcal{E}_{\mathbf{z}}(f) \to \mathcal{E}(f)$, we expect that a minimizer of the empirical error $\mathcal{E}_{\mathbf{z}}(f)$ is a good

approximation of the minimizer f_ρ of the error $\mathcal{E}(f)$. This is the idea of ERM. If one allows the minimizer to be taken over the set of all measurable functions, he can choose a function with $f(x_i) = y_i$ making $\mathcal{E}_{\mathbf{z}}(f) = 0$, leading to an overfitting. To avoid this, we choose a set $\mathcal{H}$ of functions on X and take the minimization of the empirical error only over $\mathcal{H}$ to get a minimizer

$$f_{\mathbf{z}} := \arg\min_{f\in\mathcal{H}} \mathcal{E}_{\mathbf{z}}(f) = \arg\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2. \tag{4}$$

It turns out that $f_{\mathbf{z}}$ is a good approximation of f_ρ when $\mathcal{H}$ is suitably chosen.

2.2. ERM regression with a general loss function

Now we can define the ERM learning algorithm for the regression problem associated with a general regressing loss function ψ.

Definition 3. Let $\mathcal{H}$ be a set of uniformly bounded measurable functions on X and $\psi : \mathbb{R} \to \mathbb{R}_+$ be a regressing loss function. The *empirical target function* $f_{\mathbf{z}}$ associated with the sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ is defined as

$$f_{\mathbf{z}} := \arg\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\psi(y_i - f(x_i)). \tag{5}$$

We call $\mathcal{H}$ the *hypothesis space* for the regression problem.

If we define the *empirical error* or *risk* associated with the loss function ψ as

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

then $f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}_{\mathbf{z}}(f)$.

The theory of ERM studies how $f_{\mathbf{z}}$ approximates (with respect to the error) the *target function* defined as

$$f_{\mathcal{H}} := \arg\min_{f\in\mathcal{H}} \mathcal{E}(f) = \arg\min_{f\in\mathcal{H}} \int_Z \psi(y - f(x))d\rho. \tag{6}$$

To describe this approximation, we decompose the *sample error*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) + \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}).$$

Since $f_{\mathbf{z}}$ minimizes the empirical error over $\mathcal{H}$, there holds $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) \le 0$. Therefore, we have an *ERM error decomposition*:

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \le \{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})\} + \{\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}})\}. \tag{7}$$

Set a random variable $\xi(z) = \psi(y - f_{\mathcal{H}}(x))$ on (Z, ρ) corresponding to the target function $f_{\mathcal{H}}$, we find that in probability $\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) = \frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi) \to 0$ (as $m \to \infty$). Thus, the second term in (7) converges to zero and its convergence rate can be estimated by some well-known probability inequalities such as Bernstein and Hoeffding inequalities.

Lemma 4. *Let ξ be a random variable on a probability space Z with mean μ and variance σ^2 satisfying $|\xi - \mu| \le M$ almost surely. Then, for every $\epsilon > 0$, we have*

$$\textit{[Bernstein]} \quad Prob_{\mathbf{z}\in Z^m}\left\{\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu > \epsilon\right\} \le \exp\left\{-\frac{m\epsilon^2}{2(\sigma^2 + \frac{1}{3}M\epsilon)}\right\},$$

$$\textit{[Hoeffding]} \quad Prob_{\mathbf{z}\in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| > \epsilon\right\} \le 2\exp\left\{-\frac{m\epsilon^2}{2M^2}\right\}.$$

The Bernstein inequality in the above form is called a one-side inequality.

The first term in (7) is more involved. If one writes $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})$ as $\int_Z \xi(z)d\rho - \frac{1}{m}\sum_{i=1}^{m}\xi(z_i)$ with $\xi(z) = \psi(y - f_{\mathbf{z}}(x))$, then ξ is not a single random variable, it depends on the sample $\mathbf{z}$. Thus the usual law of large numbers does not guarantee the convergence of the first term of (7) to zero. The main part of the classical statistical learning theory deals with this error term. It is called the *theory of uniform convergence*, playing the role of some uniform law of large numbers. To see this, we consider the quantity

$$\sup_{f\in\mathcal{H}} |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)| \tag{8}$$

which bounds the first term of (7), hence estimates the error $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}})$. The theory of uniform convergence studies the convergence of this quantity to zero.

2.3. Sample error bounds using covering numbers

To get convergence rates for the ERM scheme, quantitative behaviors of capacity of the hypothesis space $\mathcal{H}$ are needed. There is a rich literature concerning the error analysis (7) by means of the capacity of the hypothesis space $\mathcal{H}$. In particular, when $\mathcal{H}$ is a compact subset of $C(X)$, the error can be estimated by covering numbers [39,21,9,58].

Definition 5. For a subset $\mathcal{F}$ of a metric space and $\eta > 0$, the *covering number* $\mathcal{N}(\mathcal{F}, \eta)$ is defined to be the minimal integer $\ell \in \mathbb{N}$ such that there exist ℓ disks with radius η covering $\mathcal{F}$. For a set $\mathcal{H}$ of continuous functions on X, we denote $\mathcal{N}(\mathcal{H}, \eta)$ the covering number of $\mathcal{H}$, as a subset of $C(X)$.

For the estimates of covering numbers, see [89,90]. The following result for the least-square loss was provided in [21].

Proposition 6. *Let $\psi(t) = t^2$ and $\mathcal{H}$ be a compact subset of $C(X)$. Assume that, for all $f \in \mathcal{H}$, $|f(x) - y| \le M$ almost everywhere. Let $\sigma^2 = \sup_{f\in\mathcal{H}}\sigma^2((y - f(x))^2)$ where $\sigma^2((y - f(x))^2)$ is the variance of the random variable $\xi(z) = (y - f(x))^2$ on (Z, ρ). Then, for all $\epsilon > 0$,*

$$Prob_{\mathbf{z}\in Z^m}\{\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_{\mathcal{H}})\le\epsilon\}\ge 1-\mathcal{N}\left(\mathcal{H},\frac{\epsilon}{16M}\right)2\exp\left\{-\frac{m\epsilon^2}{8(4\sigma^2+\frac{1}{3}M^2\epsilon)}\right\}.$$

To show the idea, we give a similar result with a general loss function. We say that ψ is *Lipschitz* s $(0<s\le 1)$ on $[-M,M]$ if there is some constant $C\ge 0$ such that

$$|\psi(t)-\psi(t')|\le C|t-t'|^s,\qquad \forall t,t'\in[-M,M]. \tag{9}$$

Theorem 7. *Let $\mathcal{H}$ be a subset of $C(X)$ such that for every $f\in\mathcal{H}$, there holds $|f(x)-y|\le M$ almost surely. If $\psi:\mathbb{R}\to\mathbb{R}_+$ is a regressing loss function satisfying (9), then for every $\epsilon>0$ there holds*

$$Prob_{\mathbf{z}\in Z^m}\{\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_{\mathcal{H}})\le\epsilon\}\ge 1-\mathcal{N}\left(\mathcal{H},\left(\frac{\epsilon}{8C}\right)^{1/s}\right)2\exp\left\{-\frac{m\epsilon^2}{128C^2M^{2s}}\right\}.$$

Proof. Let $\{f_j\}_{j=1}^N\subset\mathcal{H}$ with $N=\mathcal{N}(\mathcal{H},\eta)$ and $\eta:=\left(\epsilon/(4C)\right)^{1/s}$ such that for every $f\in\mathcal{H}$ there is some $j\in\{1,\dots,N\}$ satisfying $\|f-f_j\|_\infty\le\eta$. Then according to the Lipschitz condition given in our assumption, we see from $|y-f(x)|\le M$ that

$$|\mathcal{E}(f)-\mathcal{E}(f_j)|\le\int_Z|\psi(y-f(x))-\psi(y-f_j(x))|d\rho\le\int_Z C|f(x)-f_j(x)|^s d\rho\le C\eta^s$$

and that almost surely

$$|\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_j)|\le\frac{1}{m}\sum_{i=1}^m|\psi(y_i-f(x_i))-\psi(y_i-f_j(x_i))|\le C\eta^s.$$

Thus

$$|\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)|\le|\mathcal{E}(f_j)-\mathcal{E}_{\mathbf{z}}(f_j)|+2C\eta^s.$$

By the choice of η, we find $2C\eta^s=\epsilon/2$. Therefore, the event $\{\mathbf{z}\in Z^m:|\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)|>\epsilon \text{ for some } f\in\mathcal{H}\}$ is contained in the event $\cup_{j=1}^N\{\mathbf{z}\in Z^m:|\mathcal{E}(f_j)-\mathcal{E}_{\mathbf{z}}(f_j)|>\epsilon/2\}$. It follows that

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)|>\epsilon\right\}\le\sum_{j=1}^N\mathrm{Prob}_{\mathbf{z}\in Z^m}\{|\mathcal{E}(f_j)-\mathcal{E}_{\mathbf{z}}(f_j)|>\epsilon/2\}.$$

For each j, we apply Lemma 4 to the random variable $\xi(z)=\psi(y-f_j(x))$ on (Z,ρ). It satisfies $\mu=\mathcal{E}(f_j)$ and $\frac{1}{m}\sum_{i=1}^m\xi(z_i)=\mathcal{E}_{\mathbf{z}}(f_j)$. Moreover, $|\xi(z)-\mu|$ can be expressed as

$$|\psi(y-f_j(x))-\int\psi(y'-f_j(x'))d\rho(z')|=|\int\psi(y-f_j(x))-\psi(y'-f_j(x'))d\rho(z')|$$

which is bounded by $\int C|y-f_j(x)-(y'-f_j(x'))|^s d\rho(z')\le 2CM^s$. Then we have from the Hoeffding inequality in Lemma 4 that

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\{|\mathcal{E}(f_j)-\mathcal{E}_{\mathbf{z}}(f_j)|>\epsilon/2\}\le 2\exp\left\{-\frac{m(\epsilon/2)^2}{2(2CM^s)^2}\right\}.$$

It follows that

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)|>\epsilon\right\}\leq\mathcal{N}(\mathcal{H},\eta)2\exp\left\{-\frac{m\epsilon^2}{32C^2M^{2s}}\right\}.$$

Since $f_{\mathbf{z}}$ and $f_{\mathcal{H}}$ are both in $\mathcal{H}$, by (7) we have $\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_{\mathcal{H}})\leq 2\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)|\leq 2\epsilon$ with the above confidence. Then our conclusion follows after replacing ϵ by $\epsilon/2$. □

By setting the confidence to be $1-\delta$, we can get convergence rates from Theorem 7. Recall the elementary fact [22] that the positive solution ϵ^* to the equation $\epsilon^p-c_1\epsilon^q-c_2=0$ with $c_1,c_2>0$ and $p>q>0$ satisfies $\epsilon^*\leq\max\{(2c_1)^{1/(p-q)},(2c_2)^{1/p}\}$.

Corollary 8. *If the assumption of Theorem 7 holds, and with some $r,C_0>0$ the covering number satisfies $\log\mathcal{N}(\mathcal{H},\eta)\leq C_0\eta^{-r}$, then for every $0<\delta<1$, we have with confidence $1-\delta$,*

$$\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_{\mathcal{H}})\leq\widetilde{C}\max\left\{\sqrt{\frac{\log(2/\delta)}{m}},m^{\frac{1}{2+r/s}}\right\},$$

where $\widetilde{C}:=\max\{16CM^s,(256C^2M^{2s}C_0(8C)^{r/s})^{1/(2+r/s)}\}$.

The condition for the covering number is satisfied when the hypothesis space is a ball of Sobolev spaces embedded in $C(X)$. When r is small enough (the smoothness is high enough for the case of Sobolev spaces), the bound in Corollary 8 is almost of the form $O(1/\sqrt{m})$. This can be improved to a bound of the form $O(1/m^{1-\epsilon})$ when the hypothesis space is convex and the loss function satisfies some high convexity condition, see [39,4,21] and the discussion in Section 3.3 by means of ratio probability inequalities.

2.4. Theory of uniform convergence

The theory of uniform convergence also characterizes those function sets $\mathcal{H}$ such that the quantity (8) tends to zero in probability as $m\to\infty$. This is required for all Borel probability distributions μ, which leads to the following definition.

Definition 9. We say that a set $\mathcal{H}$ of real-valued functions on a metric space X is *uniform Glivenko-Cantelli* (UGC) if for every $\epsilon>0$,

$$\lim_{\ell\to+\infty}\sup_{\mu}\mathrm{Prob}\left\{\sup_{m\geq\ell}\sup_{f\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m}f(x_i)-\int_X f(x)d\mu\right|\geq\epsilon\right\}=0$$

where the supremum is taken with respect to all Borel probability distributions μ on X, and Prob denotes the probability with respect to the samples $x_1,x_2,\ldots$ independently drawn according to such a distribution μ.

When $\mathcal{H}$ consists of indicator functions (characteristic functions), the UGC property can be characterized by VC-dimensions [75]. For general sets of real-valued functions, it can be characterized by V_γ-dimensions [1].

Definition 10. Let $\mathcal{H}$ be a set of functions from X to $[0,1]$. We say that $A \subset X$ is V_γ shattered by $\mathcal{H}$ if there is a number $\alpha \in \mathbb{R}$ with the following property: for every subset E of A there exists some function $f_E \in \mathcal{H}$ such that $f_E(x) \leq \alpha - \gamma$ for every $x \in A \setminus E$, and $f_E(x) \geq \alpha + \gamma$ for every $x \in E$. The V_γ dimension of $\mathcal{H}$, $V_\gamma(\mathcal{H})$, is the maximal cardinality of a set $A \subset X$ which is V_γ shattered by $\mathcal{H}$.

The concept of V_γ-dimension is related to many others about capacity of function sets in approximation theory or functional analysis: covering number, entropy number, packing number, metric entropy and more. See [1,26,52,49,55].

The following characterization of the UGC property was given in [1].

Theorem 11. *Let $\mathcal{H}$ be a set of functions from X to $[0,1]$. Then it is UGC if and only if the V_γ dimension of $\mathcal{H}$ is finite for every $\gamma > 0$.*

Theorem 11 may be used to verify the convergence of ERM schemes when the hypothesis space $\mathcal{H}$ is a UGC set, but is not compact with respect to the norm $\|\cdot\|_\infty$ in $C(X)$. In [83] we consider the union of unit balls of reproducing kernel Hilbert spaces $(\mathcal{H}_K, \|\cdot\|_K)$ (to be defined below) associated with a set of Mercer kernels K, and find conditions for such a hypothesis space to be UGC. In particular, for the Gaussian kernels, we verify the UGC property.

Example 12. Let X be a compact subset of $\mathbb{R}^n$ and $\{K_\sigma\}_{0<\sigma<\infty}$ be the set of Gaussian kernels with $K_\sigma(x,y) = \exp(-\frac{|x-y|^2}{\sigma^2})$. Then the set

$$\mathcal{H} = \cup_{0<\sigma<\infty} \{f \in \mathcal{H}_{K_\sigma} : \|f\|_{K_\sigma} \leq 1\}$$

is UGC, but its closure in $C(X)$ is not compact.

Quantitative behavior of the convergence of the ERM scheme with the hypothesis space in Example 12 is not fully understood, though we have some partial estimates [83,48] by means of local Rademacher averages.

There are still many open fundamental problems about the UGC property. As an example, let us consider the empirical covering numbers.

Definition 13. For $\mathbf{x} = (x_i)_{i=1}^m \in X^m$, $1 \leq p \leq \infty$, and $\mathcal{H} \subset C(X)$, the *ℓ^p-empirical covering number* $\mathcal{N}_p(\mathcal{H}, \mathbf{x}, \eta)$ is the covering number of $\mathcal{H}|_{\mathbf{x}} = \{(f(x_i))_{i=1}^m : f \in \mathcal{H}\}$, as a subset of $\mathbb{R}^m$ under the metric defined for $f, g \in C(X)$ as $d_{p,\mathbf{x}}(f,g) = \left\{\frac{1}{m}\sum_{i=1}^m |f(x_i) - g(x_i)|^p\right\}^{1/p}$. The *metric entropy* of $\mathcal{H}$ is

$$H_m(\mathcal{H}, \eta) = \sup_{\mathbf{x} \in X^m} \log \mathcal{N}_\infty(\mathcal{H}, \mathbf{x}, \eta), \qquad m \in \mathbb{N}, \eta > 0.$$

It is known [28] that a function set $\mathcal{H}$ from X to $[0,1]$ is UGC if and only if $\lim_{m\to\infty} H_m(\mathcal{H},\eta)/m = 0$ for every $\eta > 0$. In this case, there holds $H_m(\mathcal{H},\eta) = O(\log^2 m)$ for every $\eta > 0$. It is conjectured in [1] that $H_m(\mathcal{H},\eta) = O(\log m)$ is true for every $\eta > 0$. The following problem is a weak form.

Problem 14. Is it true that for some $\alpha \in [1,2)$, every UGC set $\mathcal{H}$ satisfies

$$H_m(\mathcal{H},\eta) = O(\log^\alpha m), \qquad \forall \eta > 0?$$

2.5. Approximation error and the ψ-regression function

The target function $f_{\mathcal{H}}$ defined by (6) is the one minimizing the error over $\mathcal{H}$. If we allow the set of all measurable functions, the minimizer is the function we desire with respect to the regressing loss function ψ.

Definition 15. Given the regressing loss function ψ and a probability distribution ρ on Z, the *ψ-regression function* is the one minimizing the error, given by

$$f_\rho^\psi(x) = \arg\min_{t\in\mathbb{R}} \int_Y \psi(y-t)d\rho(y|x), \qquad x \in X.$$

The *approximation error* associated with the hypothesis space $\mathcal{H}$ is defined as

$$\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho^\psi) = \inf_{f\in\mathcal{H}} \mathcal{E}(f) - \mathcal{E}(f_\rho^\psi).$$

The convexity of ψ tells us some information about f_ρ^ψ. For $x \in X$, denote

$$\Psi(t) = \Psi_x(t) = \int_Y \psi(y-t)d\rho(y|x) = \int_Y \psi(t-y)d\rho(y|x), \qquad t \in \mathbb{R}. \tag{10}$$

Then Ψ is also a convex function on $\mathbb{R}$. We know that $\Psi''(t) \geq 0$ almost everywhere. The left and right one-side derivatives of Ψ, $\Psi'_-(t)$ and $\Psi'_+(t)$, exist, are nondecreasing, and satisfy $\Psi'_-(t) \leq \Psi'_+(t)$ for every $t \in \mathbb{R}$. Hence the following fact holds.

Theorem 16. *Let ψ be a regressing loss function and $x \in X$. Then:*

(a) *The univariate function Ψ given by (10) is strictly decreasing on $(-\infty, f_\rho^-(x)]$, strictly increasing on $[f_\rho^+(x), +\infty)$, and is constant on $[f_\rho^-(x), f_\rho^+(x)]$, where*

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

(b) *$f_\rho^\psi(x)$ is a minimizer of Ψ, that is, $f_\rho^-(x) \leq f_\rho^\psi(x) \leq f_\rho^+(x)$.*

(c) *If $\epsilon \geq 0$ is the maximal zero of ψ and $|y| \leq M$ almost surely, then $f_\rho^\psi(x) \in [-M-\epsilon, M+\epsilon]$.*

Now we turn to the approximation error estimates. Denote a weighted L^p norm as $\|f\|_{L^p_{\rho_X}} = (\int_X |f(x)|^p d\rho_X)^{1/p}$ for $p \geq 1$.

Theorem 17. *Assume $|y - f(x)| \le M$ and $|y - f_\rho^\psi(x)| \le M$ almost surely.*

(a) *If ψ is a Lipschitz s regressing loss function satisfying (9), then*

$$\mathcal{E}(f) - \mathcal{E}(f_\rho^\psi) \le C \int_X |f(x) - f_\rho^\psi(x)|^s d\rho_X \le C\|f - f_\rho^\psi\|_{L^1_{\rho_X}}^s \le C\|f - f_\rho^\psi\|_{L^2_{\rho_X}}^{s/2}.$$

(b) *If ψ is C^1 on $[-M, M]$, and its derivative is Lipschitz s satisfying (9), then*

$$\mathcal{E}(f) - \mathcal{E}(f_\rho^\psi) \le C\|f - f_\rho^\psi\|_{L^{1+s}_{\rho_X}}^{1+s} \le C\|f - f_\rho^\psi\|_{L^2_{\rho_X}}^{(1+s)/2}.$$

(c) *If $\psi''(u) \ge c > 0$ for every $u \in [-M, M]$, then*

$$\mathcal{E}(f) - \mathcal{E}(f_\rho^\psi) \ge \frac{c}{2}\|f - f_\rho^\psi\|_{L^2_{\rho_X}}^2.$$

Proof. By the definition of the error, we have

$$\mathcal{E}(f) - \mathcal{E}(f_\rho^\psi) = \int_X \left\{ \int_Y \psi(y - f(x)) - \psi(y - f_\rho^\psi(x)) d\rho(y|x) \right\} d\rho_X.$$

Then the statement in (a) follows from the Lipschitz condition (9) with $t = y - f(x), t' = y - f_\rho^\psi(x) \in [-M, M]$.

If ψ is C^1 on $[-M, M]$, and ψ' satisfies (9), then for each $x \in X$, the function $\Psi = \Psi_x$ is also C^1 and $\Psi'(f_\rho^\psi(x)) = 0$ since $f_\rho^\psi(x)$ is a minimum. Moreover, $\Psi'(t) = \int_Y \psi'(t-y) d\rho(y|x)$. It implies from the Lipschitz condition (9) for ψ' that for t with $|t - y| \le M$,

$$|\Psi'(t) - \Psi'(f_\rho^\psi(x))| = \left| \int_Y \psi'(t - y) - \psi'(f_\rho^\psi(x) - y) d\rho(y|x) \right| \le C|t - f_\rho^\psi(x)|^s.$$

Then

$$\Psi(f(x)) - \Psi(f_\rho^\psi(x)) = \int_{f_\rho^\psi(x)}^{f(x)} \Psi'(t) - \Psi'(f_\rho^\psi(x)) dt \le C|f(x) - f_\rho^\psi(x)|^{1+s}.$$

Hence

$$\mathcal{E}(f) - \mathcal{E}(f_\rho^\psi) = \int_X \Psi(f(x)) - \Psi(f_\rho^\psi(x)) d\rho_X \le C \int_X |f(x) - f_\rho^\psi(x)|^{1+s} d\rho_X.$$

This proves the statement in (b).

Now assume $\psi''(u) \ge c > 0$ for every $u \in [-M, M]$. If $f(x) > f_\rho^\psi(x)$, for any $t \in [f_\rho^\psi(x), f(x)]$ we have $\Psi''(t) = \int_Y \psi''(t-y) d\rho(y|x) \ge c$. Hence $\Psi'(t) - \Psi'(f_\rho^\psi(x)) = \int_{f_\rho^\psi(x)}^t \Psi''(u) du \ge c(t - f_\rho^\psi(x))$. Then

$$\Psi(f(x)) - \Psi(f_\rho^\psi(x)) \ge \int_{f_\rho^\psi(x)}^{f(x)} c(t - f_\rho^\psi(x)) dt = \frac{c}{2}(f(x) - f_\rho^\psi(x))^2.$$

In the same way, if $f(x) < f_\rho^\psi(x)$, we have $\Psi'(t) - \Psi'(f_\rho^\psi(x)) \le c(t - f_\rho^\psi(x))$ for any $t \in [f(x), f_\rho^\psi(x)]$. Hence $\Psi(f(x)) - \Psi(f_\rho^\psi(x)) \ge \frac{c}{2}(f(x) - f_\rho^\psi(x))^2$. In both cases, we get

$$\mathcal{E}(f) - \mathcal{E}(f_\rho^\psi) = \int_X \Psi(f(x)) - \Psi(f_\rho^\psi(x)) d\rho_X \ge \frac{c}{2} \int_X |f(x) - f_\rho^\psi(x)|^2 d\rho_X.$$

This verifies the result. □

To see the approximation of f_ρ^ψ by $f_{\mathbf{z}}$, we need to get the error between $f_{\mathbf{z}}$ and f_ρ^ψ in some function spaces, from the excess generalization error $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho^\psi)$. Part (c) of Theorem 17 ensures this in the case when ψ'' is bounded from below.

Problem 18. It would be interesting for a general loss function ψ to derive lower bounds of the following form with $c, p > 0$:

$$\mathcal{E}(f) - \mathcal{E}(f_\rho^\psi) \geq c \int_X |f(x) - f_\rho^\psi(x)|^p d\rho_X.$$

When ψ is the ϵ-insensitive norm, it is Lipschitz 1 with $C = 1$ and $M = \infty$. It follows from (a) of Theorem 17 that $\mathcal{E}(f) - \mathcal{E}(f_\rho^\psi) \leq \|f - f_\rho^\psi\|_{L^1_{\rho_X}}$.

When ψ is the least-square loss, it is C^1 and its derivative satisfies (9) with $s = 1$, $C = 2$ and $M = \infty$. Hence $\mathcal{E}(f) - \mathcal{E}(f_\rho^\psi) \leq \|f - f_\rho^\psi\|^2_{L^2_{\rho_X}}$ (in fact, equality holds).

Theorem 17 (parts (a) and (b)) gives a way to estimate the approximation error by means of results from approximation theory [91].

For many ERM schemes, the hypothesis is a ball of a dense subspace $(H, \|\cdot\|_H)$ of a Banach space $(B, \|\cdot\|_B)$ of functions on X. For example, B may be some weighted L^p space $L^p_{\rho_X}$ and H is a Sobolev space or a reproducing kernel Hilbert space. For such a setting, the approximation error can be characterized by interpolation spaces. Recall that the $\mathcal{K}$-functional of the couple (B, H) is defined for $f \in B$ as

$$\mathcal{K}(f, t) = \inf_{g \in B} \{\|f - g\|_B + t\|g\|_H\}, \qquad t > 0.$$

For $0 < \theta < 1$, the interpolation space $(B, H)_{\theta,\infty}$ consists of those elements $f \in B$ such that the following norm is finite

$$\|f\|_{\theta,\infty} = \sup_{t>0} \left\{\mathcal{K}(f, t)/t^\theta\right\}.$$

The next result can be found in [64].

Theorem 19. *Let $(B, \|\cdot\|_B)$ be a Banach space and $(H, \|\cdot\|_H)$ be a dense subspace with $\|g\|_B \leq \|g\|_H$ for $g \in H$. Let $0 < \theta < 1$. If $f \in (B, H)_{\theta,\infty}$, then*

$$I(f, R) := \inf \{\|f - g\|_B : g \in H, \|g\|_H \leq R\} \leq \|f\|^{\frac{1}{1-\theta}}_{\theta,\infty} \left(\frac{1}{R}\right)^{\frac{\theta}{1-\theta}}.$$

Conversely, if $I(f, R) \leq C \left(\frac{1}{R}\right)^{\frac{\theta}{1-\theta}}$ for all $R > 0$, then $f \in (B, H)_{\theta,\infty}$ and $\|f\|_{\theta,\infty} \leq 2C^{1-\theta}$.

The most prominent function spaces used in learning theory are reproducing kernel Hilbert spaces. The importance of these function spaces can be seen from regularization schemes.

3. Regularization Schemes for Regression

3.1. Regularization schemes and convex optimization

Another approach in learning theory is to replace the ERM by the minimization of a penalized empirical error over a function space. Reproducing kernel Hilbert spaces generated by Mercer kernels [3] are often used for machine learning.

Definition 20. We say that $K : X \times X \to \mathbb{R}$ is a *Mercer kernel* if it is continuous, symmetric and positive semidefinite, that is, 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.

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

The reproducing property takes the form

$$\langle K_x, f \rangle_K = f(x), \qquad \forall\, x \in X,\ f \in \mathcal{H}_K. \tag{11}$$

Denote $\kappa = \sup_{x \in X} \sqrt{K(x, x)}$. Then the above reproducing property tells us that

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

The *regularization scheme* for the regression problem associated with the RKHS $\mathcal{H}_K$ takes the form

$$f_{\mathbf{z},\lambda} := \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} \psi(y_i - f(x_i)) + \lambda \|f\|_K^2 \right\}. \tag{12}$$

Here $\lambda > 0$ is a constant called the *regularization parameter*. Usually it is chosen to depend on m, $\lambda = \lambda(m)$, and $\lim_{m \to \infty} \lambda(m) = 0$. It plays the role of R in ERM.

The Hilbert space structure of $\mathcal{H}_K$ and the convexity of ψ reduce the scheme (12) to a convex optimization problem over $\mathbb{R}^m$. First, if we denote the projection in $\mathcal{H}_K$ onto the finite dimensional space $\mathcal{H}_{K,\mathbf{x}} = \text{span}\{K_{x_i}\}_{i=1}^m$ as P, then for any $f \in \mathcal{H}_K$ and $i = 1, \ldots, m$, there holds $\langle Pf, K_{x_i} \rangle_K = \langle f, K_{x_i} \rangle_K$, meaning that $(Pf)(x_i) = f(x_i)$ and hence $\mathcal{E}_{\mathbf{z}}(Pf) = \mathcal{E}_{\mathbf{z}}(f)$. This tells us that $f_{\mathbf{z},\lambda}$ must lie in $\mathcal{H}_{K,\mathbf{x}}$. Second, if we write $f_{\mathbf{z},\lambda} = \sum_{i=1}^m c_{i,\mathbf{z}} K_{x_i}$, then $c_{\mathbf{z}} := (c_{i,\mathbf{z}})_{i=1}^m$ is a minimizer of a convex function over $\mathbb{R}^m$:

$$c_{\mathbf{z}} = \arg \min_{c \in \mathbb{R}^m} \frac{1}{m} \sum_{i=1}^{m} \psi\left(y_i - \sum_{j=1}^{m} c_j K(x_i, x_j) \right) + \lambda \sum_{i,j=1}^{m} c_i K(x_i, x_j) c_j.$$

This is a convex optimization problem in $\mathbb{R}^m$, which can be solved by many efficient algorithms from optimization theory. If ψ is a convex quadratic spline such as the least-square loss or the ϵ-insensitive loss, then the convex optimization problem is a quadratic programming, and simpler algorithms are available.

3.2. Error decomposition and regularization error

The convergence of the regularization scheme (12) can be studied by the *excess generalization error* $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho^\psi)$.

Let $\widetilde{f}_\lambda$ be an arbitrary function from $\mathcal{H}_K$. The penalized excess generalization error $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho^\psi) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2$ can be decomposed as

$$\{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})\} + \left\{\mathcal{E}_{\mathbf{z}}(\widetilde{f}_\lambda) - \mathcal{E}(\widetilde{f}_\lambda)\right\} + \left\{\mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho^\psi) + \lambda\|\widetilde{f}_\lambda\|_K^2\right\}$$
$$+ \left\{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2 - \left(\mathcal{E}_{\mathbf{z}}(\widetilde{f}_\lambda) + \lambda\|\widetilde{f}_\lambda\|_K^2\right)\right\}.$$

The last term is at most 0, according to the definition of $f_{\mathbf{z},\lambda}$. Thus, $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho^\psi) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2$ can be bounded by

$$\{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})\} + \left\{\mathcal{E}_{\mathbf{z}}(\widetilde{f}_\lambda) - \mathcal{E}(\widetilde{f}_\lambda)\right\} + \left\{\mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho^\psi) + \lambda\|\widetilde{f}_\lambda\|_K^2\right\}.$$

The last term above is called the regularization error [65]. It does not depend on the sample.

Definition 21. The *regularization error* for a *regularizing function* $\widetilde{f}_\lambda \in \mathcal{H}_K$ is defined as

$$\widetilde{\mathcal{D}}(\lambda) = \mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho^\psi) + \lambda\|\widetilde{f}_\lambda\|_K^2.$$

It is called the *regularization error of the scheme* (12), denoted as $\mathcal{D}(\lambda)$, when $\widetilde{f}_\lambda = f_\lambda$:

$$f_\lambda = \arg\inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f) - \mathcal{E}(f_\rho^\psi) + \lambda\|f\|_K^2\right\}. \tag{13}$$

The regularization error can be estimated by Theorem 17 and the rich knowledge of approximation theory in function spaces. When ψ is the least-square loss, $f_\rho^\psi = f_\rho$ and $\mathcal{E}(f) - \mathcal{E}(f_\rho^\psi) = \|f - f_\rho\|_{L^2_{\rho_X}}^2$. In this case, the regularization error can be easily bounded by the integral operator $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ defined as

$$L_K(f)(x) = \int_X K(x,y)f(y)d\rho_X(y), \qquad x \in X, f \in L^2_{\rho_X}.$$

Note that the range of L_K is in $\mathcal{H}_K$, hence L_K can also be regarded [71] as an operator from $L^2_{\rho_X}$ to $\mathcal{H}_K$. The following result can be found in [66]. The condition $L_K^{-r} f_\rho \in L^2_{\rho_X}$ means $f_\rho = L_K^r g$ for some $g \in L^2_{\rho_X}$ with $\|g\|_{L^2_{\rho_X}} = \|L_K^{-r} f_\rho\|_{L^2_{\rho_X}}$. When $0 < r < 1/2$, this means that $f_\rho \in (L^2_{\rho_X}, \mathcal{H}_K)_{r,\infty}$.

Theorem 22. *Define f_λ by (13). If $L_K^{-r} f_\rho \in L^2_{\rho_X}$, then*

$$\mathcal{D}(\lambda) = \|f_\lambda - f_\rho\|_{L^2_{\rho_X}}^2 + \lambda\|f_\lambda\|_K^2 \le \lambda^{2r}\|L_K^{-r} f_\rho\|_{L^2_{\rho_X}}^2, \qquad \text{if } 0 < r \le \frac{1}{2}$$

and

$$\|f_\lambda - f_\rho\|_{L^2_{\rho_X}} \le \lambda^r\|L_K^{-r} f_\rho\|_{L^2_{\rho_X}}, \qquad \text{if } 0 < r \le 1.$$

When $\frac{1}{2} < r \leq 1$, *we have*

$$\|f_\lambda - f_\rho\|_K \leq \lambda^{r-\frac{1}{2}} \|L_K^{-r} f_\rho\|_{L^2_{\rho_X}}.$$

Notice that for $\frac{1}{2} < r \leq 1$, Theorem 22 yields $\|f_\lambda - f_\rho\|^2_{L^2_{\rho_X}} = O(\lambda^{2r})$ while $\mathcal{D}(\lambda)$ is at most $O(\lambda)$.

Turn back to the error decomposition. One can write the sample error as (7) and then derive convergence rates of some weak form (of type $O(1/\sqrt{m})$) by Hoeffding's inequality, as in Corollary 8. But we can do better with the Bernstein inequality by considering the excess random variable $\psi(y-f(x))-\psi(y-f_\rho^\psi(x))$, not $\psi(y-f(x))$. Then we have a *regularization scheme error decomposition*:

$$\begin{aligned}&\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho^\psi) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2 \\ &\quad \leq \{E(\xi_1) - E_{\mathbf{z}}(\xi_1)\} + \{E_{\mathbf{z}}(\xi_2) - E(\xi_2)\} + \widetilde{\mathcal{D}}(\lambda). \end{aligned} \tag{14}$$

Here

$$\xi_1 := \psi(y - f_{\mathbf{z},\lambda}(x)) - \psi(y - f_\rho^\psi(x)), \qquad \xi_2 := \psi(y - \widetilde{f}_\lambda(x)) - \psi(y - f_\rho^\psi(x))$$

and for $\xi = \xi(x,y)$, we denote $E(\xi) = \int_Z \xi(x,y)d\rho$ and $E_{\mathbf{z}}(\xi) = \frac{1}{m}\sum_{i=1}^m \xi(x_i,y_i)$. We remark that $E(\xi_1)$ is not the expected value of a single random variable since it depends on $\mathbf{z}$.

To show how (14) improves the analysis of the error decomposition of the form (7), we first look at the middle term $E_{\mathbf{z}}(\xi_2) - E(\xi_2)$.

Lemma 23. *Assume* $|\psi(y - \widetilde{f}_\lambda(x)) - \psi(y - f_\rho^\psi(x))| \leq M$ *almost surely. If for some* $\tau \in [0,1]$ *and* $C_\psi > 0$ *there holds*

$$\int \left(\psi(y - \widetilde{f}_\lambda(x)) - \psi(y - f_\rho^\psi(x))\right)^2 d\rho \leq C_\psi \left\{\mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho^\psi)\right\}^\tau,$$

then for any $0 < \delta < 1$, *with confidence at least* $1-\delta$, *we have*

$$E_{\mathbf{z}}(\xi_2) - E(\xi_2) \leq \frac{2M\log(1/\delta)}{3m} + \left(\frac{2C_\psi \log(1/\delta)}{m}\right)^{\frac{1}{2-\tau}} + \frac{1}{2}\widetilde{\mathcal{D}}(\lambda).$$

Proof. Consider the random variable ξ_2 on (Z,ρ). It satisfies $|\xi_2| \leq M$ and

$$\sigma^2 \leq E(\xi_2^2) \leq C_\psi \left\{\mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho^\psi)\right\}^\tau \leq C_\psi \left(\widetilde{\mathcal{D}}(\lambda)\right)^\tau.$$

Applying the one-side Bernstein inequality from lemma 4 to ξ_2, we know that for any $\epsilon > 0$, with confidence at least

$$1 - \exp\left\{-\frac{m\epsilon^2}{2(\sigma^2 + \frac{1}{3}M\epsilon)}\right\}$$

there holds $E_{\mathbf{z}}(\xi_2) - E(\xi_2) \leq \epsilon$. Take $\epsilon = \epsilon^*$ to be the positive solution of the quadratic equation

$$-\frac{m\epsilon^2}{2(\sigma^2 + \frac{1}{3}M\epsilon)} = \log\delta.$$

That is, $m\epsilon^2 - \frac{2M}{3}\log(1/\delta)\epsilon - 2\sigma^2\log(1/\delta) = 0$. Then

$$\epsilon^* = \frac{\frac{M}{3}\log(1/\delta) + \sqrt{(\frac{M}{3}\log(1/\delta))^2 + 2m\sigma^2\log(1/\delta)}}{m}$$
$$\leq \frac{2M}{3m}\log(1/\delta) + \sqrt{\frac{2\log(1/\delta)}{m}C_\psi\left(\widetilde{\mathcal{D}}(\lambda)\right)^\tau}.$$

Recall the elementary inequality

$$\frac{1}{q} + \frac{1}{q^*} = 1 \text{ with } q, q^* > 1 \Longrightarrow a \cdot b \leq \frac{1}{q}a^q + \frac{1}{q^*}b^{q^*}, \qquad \forall a, b \geq 0.$$

Using this with $q = \frac{2}{2-\tau}, q^* = \frac{2}{\tau}$ and $a = \sqrt{\frac{2C_\psi\log(1/\delta)}{m}}, b = \left(\widetilde{\mathcal{D}}(\lambda)\right)^{\tau/2}$, we see that with confidence $1-\delta$, $E_{\mathbf{z}}(\xi_2) - E(\xi_2) \leq \epsilon^*$ can be bounded as stated. □

Definition 24. We say that ψ has *variancing exponent* $\tau \in [0,1]$ if for any $M > 0$ there is a constant $C_{\psi,M}$ such that

$$\int \left(\psi(y - f(x)) - \psi(y - f_\rho^\psi(x))\right)^2 d\rho \leq C_{\psi,M}\left\{\mathcal{E}(f) - \mathcal{E}(f_\rho^\psi)\right\}^\tau$$

for any measurable function f with $|y - f(x)|, |y - f_\rho^\psi(x)| \leq M$ almost surely.

The least-square loss has variancing exponent $\tau = 1$ with $C_{\psi,M} = 4M^2$.

3.3. Error analysis by ratio probability inequalities

Now we show the error analysis for the regularization scheme (12). The error rate has the form $O((\frac{1}{m})^{1-\epsilon})$ with ϵ depending on the regularization error and capacity of $\mathcal{H}_K$. In particular, when K is C^∞ and the regularization error is $O(\lambda)$, ϵ can be arbitrarily small.

Ratio probability inequalities such as Talagrand's inequalities using empirical covering numbers [10,11,5,6] play an essential role in deriving good sample error estimates. Here, to illustrate the idea, we give the following simple form [80].

Proposition 25. *Let ξ be a random variable on Z satisfying $\mu = E(\xi) \geq 0$, $|\xi - \mu| \leq M$ almost surely, and $\sigma^2 \leq C\mu^\tau$ for some $0 \leq \tau \leq 2$. Then for every $\varepsilon > 0$*

$$Prob_{\mathbf{z}\in Z^m}\left\{\frac{\mu - \frac{1}{m}\sum_{i=1}^m \xi(z_i)}{(\mu^\tau + \varepsilon^\tau)^{\frac{1}{2}}} > \varepsilon^{1-\frac{\tau}{2}}\right\} \leq \exp\left\{-\frac{m\varepsilon^{2-\tau}}{2(C + \frac{1}{3}M\varepsilon^{1-\tau})}\right\}.$$

This proposition is an easy consequence of the one-side Bernstein inequality applied to the event $\mu - \frac{1}{m}\sum_{i=1}^m \xi(z_i) > \varepsilon^{1-\frac{\tau}{2}}(\mu^\tau + \varepsilon^\tau)^{\frac{1}{2}}$.

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

Proposition 26. *Let $0 \le \tau \le 1$, $M > 0$, $C \ge 0$, and $\mathcal{F}$ be a set of functions on Z such that for every $f \in \mathcal{F}$, $E(f) \ge 0$, $|f - E(f)| \le M$ and $E(f^2) \le C(E(f))^\tau$. Then for $\varepsilon > 0$, there holds*

$$Prob_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{F}} \frac{E(f) - \frac{1}{m}\sum_{i=1}^m f(z_i)}{((E(f))^\tau + \varepsilon^\tau)^{\frac{1}{2}}} > 4\varepsilon^{1-\frac{\tau}{2}}\right\} \le \mathcal{N}(\mathcal{F},\varepsilon)\exp\left\{\frac{-m\varepsilon^{2-\tau}}{2(C + \frac{1}{3}M\varepsilon^{1-\tau})}\right\}.$$

The proof of this proposition is the same as the procedure in Theorem 7, by means of covering numbers.

The error estimates for (12) follow from Proposition 26 and the regularization error. As an example, consider the scheme associated with the least-square loss

$$f_{\mathbf{z},\lambda} := \arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^m (y_i - f(x_i))^2 + \lambda\|f\|_K^2\right\}. \tag{15}$$

We have the following result presented in [79].

Theorem 27. *Let $f_{\mathbf{z},\lambda}$ be defined by (15). Assume $|y| \le M$ almost surely and that f_ρ is not identically zero. If for some $s, C_0 > 0$, $\log\mathcal{N}(\eta) \le C_0\,(1/\eta)^s$, then for any $0 < \zeta < \frac{1}{1+s}$, $0 < \delta < 1$ and $m \ge m_{\delta,\zeta}$, with confidence $1 - \delta$ there holds*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{L^2_{\rho_X}}^2 \le \frac{\widetilde{C}}{\|f_\rho\|_{L^2_{\rho_X}}^2}\log\left(\frac{2}{\delta}\right)\mathcal{D}\left(m^{-\zeta}\right) \qquad \text{by taking } \lambda = m^{-\zeta}$$

where $m_{\delta,\zeta}$ and $\widetilde{C}$ are constants depending on C_0, s, ζ, κ, M, and $m_{\delta,\zeta}$ also on δ.

In particular, if K is C^∞ on $X \subset \mathbb{R}^n$ and $f_\rho \in \mathcal{H}_K$, then by taking $\lambda = \lambda(m) = m^{2\epsilon-1}$ with an arbitrary $\epsilon > 0$, for any $0 < \delta < 1$, with confidence $1 - \delta$ there holds

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{L^2_{\rho_X}}^2 \le \widetilde{C}\log(\frac{2}{\delta})\left(\frac{1}{m}\right)^{1-\epsilon}$$

for $m \ge m_{\delta,\epsilon}$, where the constants $m_{\delta,\epsilon}$ and $\widetilde{C}$ are independent of m.

Let us briefly mention the idea of deriving the error bounds stated in Theorem 27. For details, see [79]. Take $\widetilde{f}_\lambda = f_\lambda$. Denote $B_R = \{f \in \mathcal{H}_K : \|f\|_K \le R\}$. The main point is to apply Proposition 26 with $\tau = 1$ to the function set

$$\mathcal{F}_R = \left\{(f(x) - y)^2 - (f_\rho(x) - y)^2 : f \in B_R\right\} \tag{16}$$

for various $R > 0$. Proposition 26 says that for all $\varepsilon > 0$ and $R \geq M$,

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in B_R}\frac{\mathcal{E}(f)-\mathcal{E}(f_\rho)-\big(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_\rho)\big)}{\sqrt{\mathcal{E}(f)-\mathcal{E}(f_\rho)+\varepsilon}}\leq\sqrt{\varepsilon}\right\}$$
$$\geq 1-\mathcal{N}\left(B_1,\frac{3\varepsilon}{4(\kappa+3)^2R^2}\right)\exp\left\{-\frac{3m\varepsilon}{160(\kappa+3)^2R^2}\right\}.$$

Denote $v^*(m,\delta)$ as the unique positive solution to the equation

$$\log\mathcal{N}(B_1,\eta)-\frac{m\eta}{40}=\log\delta.$$

Then the confidence becomes $1-\delta/2$ if $\varepsilon=\frac{4(\kappa+3)^2R^2}{3}v^*(m,\delta/2)$. It follows that there is a set $V_R'\subseteq Z^m$ of measure at most $\delta/2$ such that for all $f\in B_R$ and $\mathbf{z}\in Z^m\setminus V_R'$,

$$\mathcal{E}(f)-\mathcal{E}(f_\rho)-\big(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_\rho)\big)\leq\frac{1}{2}\big(\mathcal{E}(f)-\mathcal{E}(f_\rho)\big)+\frac{4}{3}(\kappa+3)^2R^2v^*(m,\delta/2).$$

Thus, if we denote

$$\mathcal{W}(R)=\{\mathbf{z}\in Z^m:\|f_{\mathbf{z},\lambda}\|_K\leq R\},\tag{17}$$

then for $\mathbf{z}\in\mathcal{W}(R)\setminus V_R'$, $f_{\mathbf{z},\lambda}\in B_R$ and $\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho)-\big(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})-\mathcal{E}_{\mathbf{z}}(f_\rho)\big)$ becomes

$$E(\xi_1)-\frac{1}{m}\sum_{i=1}^m\xi_1(z_i)\leq\frac{1}{2}\big(\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho)\big)+\frac{4}{3}(\kappa+3)^2R^2v^*(m,\delta/2).$$

On the other hand, since ξ_2 satisfies $|\xi_2-E(\xi_2)|\leq 2\big(\kappa\sqrt{\mathcal{D}(\lambda)/\lambda}+3M\big)^2$, Lemma 23 tells us that there is another subset $V_R''\subset Z^m$ of measure at most $\delta/2$ such that for all $\mathbf{z}\in Z^m\setminus V_R''$,

$$\frac{1}{m}\sum_{i=1}^m\xi_2(z_i)-E(\xi_2)\leq\mathcal{D}(\lambda)\left(1+\frac{4\kappa^2\log(2/\delta)}{m\lambda}\right)+\frac{36M^2\log(2/\delta)}{m}.$$

Thus for all $\mathbf{z}\in\mathcal{W}(R)\setminus(V_R'\cup V_R'')$,

$$\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho)+\lambda\|f_{\mathbf{z},\lambda}\|_K^2\leq 3(\kappa+3)^2R^2v^*(m,\delta/2)$$
$$+\mathcal{D}(\lambda)\left(4+\frac{8\kappa^2\log(2/\delta)}{m\lambda}\right)+\frac{72M^2\log(2/\delta)}{m}.\tag{18}$$

The key point to derive satisfactory error bounds is to use (18) iteratively. First, by taking $f=0$ in the the definition (15) of the scheme, we see that $\|f_{\mathbf{z},\lambda}\|_K\leq R^{[0]}:=M/\sqrt{\lambda}$. So $\mathcal{W}(R^{[0]})=Z^m$. Then we apply (18) to sharpen the norm and get $Z^m=\mathcal{W}(R^{[0]})\subseteq\mathcal{W}(R^{[1]})\cup V^{[1]}$ where $V^{[1]}=V_{R^{[0]}}'\cup V_{R^{[0]}}''$ and

$$R^{[1]}=(\kappa+3)R\sqrt{3v^*(m,\delta/2)/\lambda}+\sqrt{\frac{\mathcal{D}(\lambda)}{\lambda}}\left(4+\frac{8\kappa^2\log(2/\delta)}{m\lambda}\right)^{1/2}+M\sqrt{\frac{72\log(2/\delta)}{m\lambda}}.$$

By the choice of λ, $R^{[1]}<R^{[0]}$ (for sufficiently large m). That is, the bound $R^{[0]}$ for $\|f_{\mathbf{z},\lambda}\|_K$ has been sharpened to $R^{[1]}$ (with confidence $1-\delta$). Iterating this process

J times with $\frac{1}{1-\zeta(1+s)} - 1 \leq J \leq \frac{1}{1-\zeta(1+s)}$, we can show that with confidence $1 - \frac{\delta}{1-\zeta(1+s)}$ there holds

$$\|f_{\mathbf{z},\lambda}\|_K \leq R^{[J]} = \widetilde{c}\sqrt{\log(2/\delta)}\left(\sqrt{\mathcal{D}(\lambda)/\lambda} + 1\right).$$

Then (18) with $R = R^{[J]}$ yields the desired error bound for $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)$ stated in Theorem 27.

3.4. Leave-one-out capacity independent error bounds

Another interesting approach for the error analysis of learning schemes is the so-called leave-one-out techniques [75]. They derive error bounds from the behavior of local error changes when only one sample point is removed or replaced. Let us state some results for the regularization scheme (15) associated with $\psi(t) = t^2$.

In [14,86], leave-one-out error and stability analysis were used to derive the expected value of general learning schemes. For the scheme (15), the result in [86] can be expressed as

$$E_{\mathbf{z}\in Z^m}\big(\mathcal{E}(f_{\mathbf{z},\lambda})\big) \leq \left(1 + \frac{2\kappa^2}{m\lambda}\right)^2 \inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f) + \lambda\|f\|_K^2\right\}.$$

In terms of the regularization error, it can be restated as

$$E_{\mathbf{z}\in Z^m}\big(\|f_{\mathbf{z},\lambda} - f_\rho\|_{L^2_{\rho_X}}^2\big) \leq \mathcal{D}(\lambda) + \Big(\mathcal{E}(f_\rho) + \mathcal{D}(\lambda)\Big)\left\{\frac{2\kappa^2}{m\lambda} + \left(\frac{\kappa^2}{m\lambda}\right)^2\right\}.$$

Choosing $\lambda = 1/\sqrt{m}$, the derived learning rate is $\big(\frac{1}{m}\big)^{\frac{1}{2}}$ in expectation when $f_\rho \in \mathcal{H}_K$ and $\mathcal{E}(f_\rho) > 0$. By the Markov inequality, $\|f_{\mathbf{z},\lambda} - f_\rho\|_{L^2_{\rho_X}}^2 \leq \frac{C}{\delta}\big(\frac{1}{m}\big)^{\frac{1}{2}}$ with confidence $1 - \delta$.

In [24], a functional analysis approach was employed for the error analysis of the scheme (15). The main result asserts that for any $0 < \delta < 1$, with confidence $1 - \delta$,

$$\big|\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\lambda)\big| \leq \frac{M\kappa^2}{\sqrt{m}}\left(1 + \frac{\kappa}{\sqrt{\lambda}}\right)\left(1 + \sqrt{2\log(2/\delta)}\right).$$

The corresponding learning rate in [24] is the following: when f_ρ lies in the range of L_K, that is, $L_K^{-1} f_\rho \in L^2_{\rho_X}$, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_{L^2_{\rho_X}}^2 \leq C\big(\frac{\log(2/\delta)}{m}\big)^{\frac{2}{5}}, \qquad \text{if} \qquad \lambda = \big(\frac{\log(2/\delta)}{m}\big)^{\frac{1}{5}}.$$

Thus the confidence is improved from $1/\delta$ to $\log(2/\delta)$, while the rate is weakened to $\big(\frac{1}{m}\big)^{\frac{2}{5}}$.

In [66], a modified McDiarmid inequality [44,82] was used to improve the capacity independent error bounds. If f_ρ is in the range of L_K, then for any $0 < \delta < 1$, with confidence $1 - \delta$ there hold

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_K^2 \leq \widetilde{C}\left(\frac{\big(\log(4/\delta)\big)^2}{m}\right)^{\frac{1}{3}} \qquad \text{by taking} \qquad \lambda = \left(\frac{\big(\log(4/\delta)\big)^2}{m}\right)^{\frac{1}{3}}$$

and

$$\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} \le \widetilde{C}\frac{\log(4/\delta)}{\sqrt{m}} \qquad \text{by taking} \qquad \lambda = \left(\frac{\left(\log(4/\delta)\right)^2}{m}\right)^{\frac{1}{4}}.$$

Thus the confidence for the learning rate $m^{-\frac{1}{2}}$ is improved. Moreover, the error in the $\mathcal{H}_K$-norm can be estimated.

In [67], a Bennett inequality for vector-valued random variables with values in Hilbert spaces is applied, which yields better error bounds: If f_ρ is in the range of L_K, then we have

$$\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} \le \widetilde{C}\left(\log(4/\delta)\right)^2\left(\frac{1}{m}\right)^{2/3} \qquad \text{by taking} \qquad \lambda = \log(4/\delta)\left(\frac{1}{m}\right)^{1/3}.$$

All the above results are capacity independent error bounds. When $f_\rho \in \mathcal{H}_K$ and $s < 2$, the learning rate given by Theorem 27 with $r = 1/2$ is better than existing results.

4. Classification Algorithms via Regularization

Classification algorithms provide ways to separate data into different classes. Consider the case of two different classes, labeled by 1 or -1. A binary classifier $f : X \to Y = \{1, -1\}$ is a function from X to Y which divides the input space X into two classes. It makes a decision for each case (an input $x \in X$) measured in many quantitative ways. If X is a subset of $\mathbb{R}^n$, each point $x \in X$ is a vector with n components corresponding to values of n measurements. These measurements are used to make the decision or to predict some event. For weather prediction, we observe the current weather in n different ways and get a vector $x \in X \subset \mathbb{R}^n$, then a classifier (a method for weather prediction) gives an output value $y = f(x) \in Y$. It predicts whether it will rain (if $f(x) = 1$) or not (if $f(x) = -1$). For the diagnosis of a special disease, a patient is tested by n different instruments which produce a vector $x \in X \subset \mathbb{R}^n$. Then a doctor makes a decision whether the patient gets the disease based on the vector x, according to the values $f(x) \in Y$. A good doctor corresponds to a good classifier. A good classifier has small error, measured by misclassification error.

Definition 28. Let ρ be a probability distribution on $Z = X \times Y$. The *misclassification error* $\mathcal{R}(f)$ of a classifier $f : X \to Y$ is defined to be the probability of a wrong prediction, that is, of the event $\{f(x) \neq y\}$,

$$\mathcal{R}(f) = \text{Prob}_{\mathbf{z}\in Z}\{f(x) \neq y\} = \int_X P(y \neq f(x)|x)d\rho_X.$$

By discussing for each $x \in X$, we can easily see that the best classifier minimizing the misclassification error, called the *Bayes rule*, is given by $f_c(x) = \text{sgn}(f_\rho)$, where

$$\operatorname{sgn}(f)(x) = \begin{cases} 1, & \text{if } f(x) \geq 0, \\ -1, & \text{if } f(x) < 0. \end{cases}$$

The purpose of classification algorithms is to find good approximations $f_{\mathbf{z}}$ of the Bayes rule from the random sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ drawn according to the probability distribution ρ. We hope that the approximating classifier $f_{\mathbf{z}}$ will approach the Bayes rule as long as the number of samples is large enough, in the sense that $\mathcal{R}(f_{\mathbf{z}}) \to \mathcal{R}(f_c)$ as $m \to \infty$.

Here we discuss classification algorithms generated from regularization schemes associated with reproducing kernel Hilbert spaces. The classifier is $\operatorname{sgn}(f_{\mathbf{z},\lambda})$ where

$$f_{\mathbf{z},\lambda} := \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m \phi(y_i f(x_i)) + \lambda \|f\|_K^2 \right\}. \tag{19}$$

Here $\phi : \mathbb{R} \to \mathbb{R}_+$ is a loss function for classification.

4.1. Support vector machine classification

Before we study the general classification scheme (19), we describe a classical learning algorithm, the support vector machine (SVM) classifier [13].

Let us start with the hard margin SVM by hyperplanes in the separable case. Suppose $X \subset \mathbb{R}^n$. The sample $\mathbf{z}$ consists of two classes: $I = \{i : y_i = 1\}$ and $II = \{j : y_j = -1\}$. We say that these two classes are separable by a hyperplane $w \cdot x = b$ with $w \in \mathbb{R}^n$, $|w| = 1$, and $b \in \mathbb{R}$ if

$$\begin{cases} w \cdot x_i > b, & \text{if } i \in I, \\ w \cdot x_j < b, & \text{if } j \in II. \end{cases}$$

That is, points from the different classes lie on different sides of the hyperplane. We always assume that I and II are both nonempty.

Note that the (signed) distance from a point $x^* \in \mathbb{R}^n$ to the plane $w \cdot x = 0$ is $w \cdot x^*$, where a negative distance means x^* is in the other side of w. For $b \in \mathbb{R}$, the hyperplane $w \cdot x = b$ is parallel to $w \cdot x = 0$ with distance b. Thus, the distance from $x^* \in \mathbb{R}^n$ to the hyperplane $w \cdot x = b$ is $w \cdot x^* - b$. Positive and negative signs of the distance stand for different sides of the hyperplane $w \cdot x = b$.

Now if the two classes of the sample $\mathbf{z}$ are separable by the hyperplane $w \cdot x = b$, we know that $b_1(w) := \min_{i \in I} \{w \cdot x_i - b\} > 0$ and $b_2(w) := \max_{j \in II} \{w \cdot x_j - b\} < 0$. To balance, we shift the hyperplane to $w \cdot x = c(w)$, where $c(w) = \frac{1}{2}\{b_1(w) + b_2(w)\}$. Then we see that with $\Delta(w) = \frac{1}{2}\{b_1(w) - b_2(w)\} > 0$, there holds

$$\begin{cases} w \cdot x_i - c(w) \geq \min\limits_{\ell \in I} w \cdot x_\ell - c(w) = \Delta(w), & \text{if } i \in I, \\ w \cdot x_j - c(w) \leq \max\limits_{\ell \in II} w \cdot x_\ell - c(w) = -\Delta(w), & \text{if } j \in I. \end{cases}$$

We call the hyperplanes $w \cdot x = c(w) \pm \Delta(w)$ the separating hyperplanes. We know that each separating hyperplane contains some points from that class. The distance

between these two separating hyperplanes is $2\Delta(w)$. So the number $\Delta(w)$ is called the margin associated with the hyperplane $w \cdot x = c(w)$.

The SVM seeks the best hyperplane by requiring $\Delta(w)$ to be the largest. Given the data $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$, we solve the optimization problem

$$\Delta(w^*) = \max_{|w|=1} \frac{1}{2} \left\{ \min_{y_i=1} w \cdot x_i - \max_{y_j=-1} w \cdot x_j \right\}. \tag{20}$$

If w^* is a solution, then $\mathbf{z}$ is separable by a hyperplane if and only if $\Delta(w^*) > 0$. In this case, the hyperplane, called the *optimal hyperplane*, is given by $w \cdot x = c(w^*)$, and the number $\Delta(w^*)$ is called the *maximal margin* of the data. The vectors x_i on the separating hyperplanes $w^* \cdot x = c(w^*) \pm \Delta(w^*)$ are called *support vectors*, because these are the vectors determining the separating hyperplanes. The classifier is then given by $f(x) = \operatorname{sgn}(w^* \cdot x - c(w^*))$.

The optimization problem (20) has a unique solution for separable data, which is stated in [75] as follows.

Theorem 29. *If $\mathbf{z}$ is separable by a hyperplane, then the optimization problem (20) has a unique solution w^*. Moreover $w^* = \frac{\widetilde{w}}{|\widetilde{w}|}$ where $\widetilde{w}$ solves the optimization problem*

$$\widetilde{w} := \arg \min_{w \in \mathbb{R}^n} \min_{b \in \mathbb{R}} |w|^2 \tag{21}$$
$$\text{subject to } y_i(w \cdot x_i - b) \geq 1, \quad i = 1, \ldots, m$$

Also, $\Delta(w^*) = \frac{1}{|\widetilde{w}|}$ is the maximal margin.

4.2. SVM soft margin classifier

The optimization problem (21) is called the SVM hard margin classifier. It works well for separable data. But most data are not separable by hyperplanes.

In the nonseparable case, there is no $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ such that $\mathbf{z}$ is separated by the hyperplane $w \cdot x = b$. Then we use the *SVM soft margin classifier* by introducing slack variables $\xi = (\xi_i)_{i=1}^m$

$$(\widetilde{w}, \widetilde{b}) := \arg \min_{w \in \mathbb{R}^n, b \in \mathbb{R}} \min_{\xi \in \mathbb{R}^n} \left\{ |w|^2 + \frac{2C}{m} \sum_{i=1}^m \xi_i \right\}, \tag{22}$$
$$\text{subject to } y_i(w \cdot x_i - b) \geq 1 - \xi_i, \xi_i \geq 0, \quad i = 1, \ldots, m$$

Here $C > 0$ is a trade-off parameter. The SVM soft margin classifier is then defined as $\operatorname{sgn}(\widetilde{w} \cdot x - \widetilde{b})$.

The soft margin classifier (22) can be reduced to the hard margin classifier (21) in the separable case by taking $C = +\infty$ and $\xi_i = 0$.

Define the *hinge loss* ϕ by $\phi(t) = \max\{1 - t, 0\}$, we see that the optimal ξ in (22) is given by $\xi_i = \max\{1 - y_i(w \cdot x_i - b), 0\} = \phi(y_i(w \cdot x_i - b))$. Hence (22) can be expressed by means of the hinge loss ϕ as a regularization scheme

$$\min_{w\in\mathbb{R}^n, b\in\mathbb{R}} \left\{ \frac{1}{m}\sum_{i=1}^{m} \phi(y_i(w\cdot x_i - b)) + \frac{1}{2C}|w|^2 \right\}.$$

If we choose a Mercer kernel $K(x,y) = x\cdot y$, then $|w|^2 = \|w\cdot x\|_K^2$ and the above scheme can be written [19] as

$$\min_{f\in\mathcal{H}_K, b\in\mathbb{R}} \left\{ \frac{1}{m}\sum_{i=1}^{m} \phi(y_i(f(x_i) - b)) + \frac{1}{2C}\|f\|_K^2 \right\}. \tag{23}$$

This is the regularization scheme (19) associated with the hinge loss ϕ, except the constant term b, called the *offset*. We shall not discuss the offset. For an approach, see [17].

4.3. Classifying loss functions

The goal of the regularization scheme (19) is to approximate the Bayes rule f_c by $\operatorname{sgn}(f_{\mathbf{z},\lambda})$. Note that f_c is a minimizer of the misclassification error $\mathcal{R}(f) = \int_Z \chi_{\{y\neq f(x)\}} d\rho = \int_Z \chi_{\{yf(x)<0\}} d\rho$. It would then be natural to discretize $\mathcal{R}(f)$ into $\frac{1}{m}\sum_{i=1}^{m} \chi_{\{y_i f(x_i)<0\}}$ and consider the regularization scheme (19) with the misclassification loss

$$\phi_0(t) = \chi_{\{t<0\}}.$$

However, the corresponding optimization problem is not convex when we write $f_{\mathbf{z},\lambda} = \sum_{i=1}^{m} c_i K_{x_i}$. Thus, we choose the loss function ϕ in (19) according to two criteria: (1) ϕ should be convex; (2) ϕ should be close to ϕ_0.

Definition 30. 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 $\inf_{t\in\mathbb{R}} \phi(t) = 0$.

For $(x,y) \in Z$, the value $\phi(yf(x))$ is the local error suffered from the use of f as a model for the process reproducing y at x. The condition $\phi'(0) < 0$ tells us that the error has a major drop as the sign of $f(x)$ changes from $-y$ to y (turning from the other class to the class labeled by y). The last requirement means that for some pairs (x,y) and function f, the error $\phi(yf(x))$ is small. This verifies the ability of "classifying".

Example 31.

(a) For $1 \le q < \infty$, the SVM q-norm soft margin classifier corresponds to the loss function $\phi(t) = (\max\{1-t, 0\})^q$.
(b) The least-square loss corresponds to $\phi(t) = (1-t)^2$ since $y^2 = 1$ for $y \in Y$ implies $\phi(yf(x)) = (1 - yf(x))^2 = (y - f(x))^2$.
(c) The exponential loss is $\phi(t) = e^{-t}$, see [87,43].
(d) The logistic regression is $\phi(t) = \log(1 + e^{-t})$, see [87,43].

In particular, the hinge loss $\phi(t) = \max\{1-t, 0\}$ is a good approximation of ϕ_0.

Define the *ϕ-error* or ϕ-risk associated with the loss ϕ as

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

and the *empirical ϕ-error* as $\mathcal{E}_{\mathbf{z}}^{\phi}(f) = \frac{1}{m}\sum_{i=1}^{m}\phi(y_i f(x_i))$. Then the classification algorithm (19) can be studied as the regression scheme (12). In particular, we can see that $\mathcal{E}^{\phi}(f_{\mathbf{z},\lambda}) \to \inf_{f\in\mathcal{H}_K}\mathcal{E}^{\phi}(f)$ with confidence (as $m \to \infty$ and $\lambda(m) \to 0$). However, for the classification algorithm, we are interested in the misclassification error, not the ϕ-error. In [81], we presented a counterexample, and showed that in general one cannot have $\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z},\lambda})) \to \inf_{f\in\mathcal{H}_K}\mathcal{R}(\mathrm{sgn}(f))$ with confidence (as $m \to \infty$) for any $\lambda = \lambda(m)$.

Problem 32. If $\mathcal{H}_K$ is not dense in $C(X)$, for which distributions ρ and choice $\lambda = \lambda(m)$ can we have $\lim_{m\to\infty}\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z},\lambda})) = \inf_{f\in\mathcal{H}_K}\mathcal{R}(\mathrm{sgn}(f))$ with confidence?

Fortunately, the excess misclassification error $\mathcal{R}(\mathrm{sgn}(f)) - \mathcal{R}(f_c)$ can be estimated by the excess ϕ-error $\mathcal{E}^{\phi}(f) - \mathcal{E}^{\phi}(f_\rho^{\phi})$ when $\mathcal{H}_K$ is dense in $C(X)$ [68]. Here f_ρ^{ϕ} is a minimizer of the ϕ-error:

$$f_\rho^{\phi}(x) = \arg\min_{t\in\mathbb{R}}\int_Y \phi(yt)d\rho(y|x) = \arg\min_{t\in\mathbb{R}}\Phi_x(t), \qquad x \in X, \tag{24}$$

and $\Phi_x(t) = \phi(t)P(y=1|x) + \phi(-t)P(y=-1|x)$. For the hinge loss, $f_\rho^{\phi} = f_c$. The properties of f_ρ^{ϕ} can be investigated as for f_ρ^{ψ} in Theorem 16, see [41,78].

Zhang [87] derived the following result for the hinge loss. See [11,43] for related works on boosting methods.

Theorem 33. *Let $\phi(t) = \max\{1-t, 0\}$. For any measurable function $f : X \to \mathbb{R}$,*

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

For general loss functions, the problem was studied in [7,17]. We get the following comparison theorem in [17].

Theorem 34. *If ϕ is a classifying loss function satisfying $\phi''(0) > 0$, then there is a constant c_ϕ such that for any measurable function $f : X \to \mathbb{R}$, there holds*

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

Thus, we can estimate the excess misclassification error by means of the excess ϕ-error. But the binary classification problem has a special feature. The classifiers take values in $\{1, -1\}$ and are often signs of real-valued functions. Therefore, it is natural for us [17] to introduce the concept of projection.

Definition 35. The *projection operator* π_M is defined for $M > 0$ on the space of measurable functions $f : X \to \mathbb{R}$ as

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

By the projection and Theorems 33 and 34, for any $M > 0$ we only need to estimate $\mathcal{E}^\phi(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}^\phi(f_\rho^\phi)$ for the classification algorithm (19).

4.4. Error analysis for regularized classifiers

There has been an extensive investigation on the error analysis of kernel related classification algorithms, especially for SVM with the hinge loss. ERM methods were analyzed in [75,62,5,20,41]. A model selection process was introduced in [10,43,11,69] for error bounds. Leave-one-out techniques including stability analysis were used in [14,86,7]. Modified ERM methods were employed in [70]. A regularization approach was used in [81,17].

Based on previous methods, the error analysis for the general classification algorithm (19) associated with a general classifying loss function ϕ was given in [78].

We need the ℓ^2-empirical covering number $\mathcal{N}_2(\mathcal{H}, \mathbf{x}, \eta)$. The advantage of using this empirical covering number for the RKHS $\mathcal{H}_K$ is that the following is always true for some $p \in (0, 2]$ and $c_p > 0$:

$$\sup_{\mathbf{x}\in X^m, m\in\mathbb{N}} \log \mathcal{N}_2(B_1, \mathbf{x}, \eta) \le c_p \, (1/\eta)^p, \qquad \forall 0 < \eta \le 1. \tag{25}$$

In fact, we have a sharper bound: $\int_0^1 \left\{\sup_{\mathbf{x}\in X^m, m\in\mathbb{N}} \log \mathcal{N}_2(B_1, \mathbf{x}, \eta)\right\}^{1/2} d\eta < \infty$. For more detailed discussions on empirical covering numbers, see [74,6,10].

As for regressing loss functions, we need an exponent concerning the dominance of the variance by expected values.

Definition 36. We say that the classifying loss ϕ has *variancing exponent* $\tau \in [0, 1]$ if for each $M > 0$ there is a constant C_M such that for any measurable function $f : X \to [-M, M]$, there holds

$$\int \left(\phi(yf(x)) - \phi(yf_\rho^\phi(x))\right)^2 d\rho \le C_M \left\{\mathcal{E}^\phi(f) - \mathcal{E}^\phi(f_\rho^\phi)\right\}^\tau. \tag{26}$$

The variancing exponent τ depends on the strong convexity [45,6,7] of ϕ. For example, for the q-norm SVM loss $\phi(t) = (\max\{1 - t, 0\})^q$, we have $\tau = 1$ if $1 < q \le 2$, and $\tau = 2/q$ if $q > 2$, $\tau = 0$ if $q = 1$.

Theorem 37. *Assume that K satisfies (25), the distribution ρ satisfies (26) and*

$$\mathcal{D}(\lambda) = \inf_{f\in\mathcal{H}_K} \left\{\mathcal{E}^\phi(f) - \mathcal{E}^\phi(f_\rho^\phi) + \lambda \|f\|_K^2\right\} \le c_\beta |\lambda|^\beta, \qquad \forall \lambda > 0 \tag{27}$$

for some $\beta \in (0,1]$ *and* $c_\beta > 0$. *If* ϕ *has a minimal zero* M *and with some* $q > 0$ *it satisfies* $\phi(t) \le c_q|t|^q$ *for any* $|t| \ge M$, *then for any* $\epsilon > 0$ *and* $0 < \delta < 1$, *there is a constant* $\widetilde{c}$ *independent on* m *such that with* $\lambda = \lambda(m) = m^{-\gamma}$ *there holds*

$$\mathcal{E}^\phi(\pi_M(f_{\mathbf{z},\lambda})) - \mathcal{E}^\phi(f_\rho^\phi) \le \widetilde{c}m^{-\theta}$$

with confidence $1-\delta$, *where* $\gamma := \min\left\{\frac{2}{\beta(4-2\tau+p\tau)+p(1-\beta)}, \frac{2}{2\beta+q-\beta q}\right\}$ *and*

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

In Theorem 37, we assume that the loss ϕ has zeros on $\mathbb{R}$ and ϕ has at most polynomial increment. This excludes the exponential loss $\phi(t) = e^{-t}$or the logistic loss $\phi(t) = \log(1+e^{-t})$. Though the convergence with these loss functions were studied in [11,43] and some error analysis was done in [78], the convergence rates are in general unknown.

Problem 38. For the exponential loss $\phi(t) = e^{-t}$ or the logistic loss $\phi(t) = \log(1+e^{-t})$, what is the convergence rate in terms of the variancing exponent τ, the order β for the regularization error, and capacity index p?

For the hinge loss, $\tau = 0$ and $q = 1$. So Theorem 37 in connection with Theorem 33 yields the convergence rate $\mathcal{R}(\mathrm{sgn}(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c) = O(m^{-\theta})$ with $\theta = \min\left\{\frac{2\beta}{4\beta+p(1-\beta)} - \epsilon, \frac{2\beta}{\beta+1}\right\}$. When K is C^∞, p can be arbitrarily small, and learning rate θ can be arbitrarily close to $\frac{1}{2}$ if $\beta \ge 1/3$.

In the same way, we can get learning rates for the least-square loss, which can be arbitrarily close to $\frac{1}{2}$ when enough approximation order is assumed.

When the distribution ρ satisfies some noise conditions, the learning rates can be further improved.

4.5. Noise conditions for classification algorithms

In this subsection we discuss some noise conditions. To demonstrate the main idea, we restrict our discussion to the hinge loss $\phi(t) = \max\{1-t, 0\}$.

The classical noise condition is for strictly separable data.

Definition 39. We say that the probability distribution ρ is ***strictly separable*** by $\mathcal{H}_K$ with margin $\Delta > 0$ if there is some $f_{\mathrm{sp}} \in \mathcal{H}_K$ such that $\|f_{\mathrm{sp}}\|_K = 1$ and $yf_{\mathrm{sp}}(x) \ge \Delta$ almost surely.

The strictly separable condition tells us that $\mathcal{R}(f_c) = 0$ and $\mathcal{E}^\phi(f_c) = 0$. By taking the function f_{sp}/Δ in the definition (19), we see that $\mathcal{E}_{\mathbf{z}}^\phi(f_{\mathbf{z},\lambda}) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2 \le \frac{\lambda}{\Delta^2}$, since $\phi(yf_{\mathrm{sp}}(x)/\Delta) = 0$ for almost every $(x,y) \in Z$. It follows that $\|f_{\mathbf{z},\lambda}\|_K \le \frac{1}{\Delta}$

and $\mathcal{E}_{\mathbf{z}}^{\phi}(f_{\mathbf{z},\lambda}) \leq \frac{\lambda}{\Delta^2}$. Thus we can apply the ratio probability inequality, Proposition 26, with $\tau = 1$ to the function set $B_{1/\Delta}$ to estimate the first term of the following error decomposition

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z},\lambda})) \leq \mathcal{E}^{\phi}(f_{\mathbf{z},\lambda}) \leq \mathcal{E}^{\phi}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}^{\phi}(f_{\mathbf{z},\lambda}) + \frac{\lambda}{\Delta^2}.$$

For any $0 < \delta < 1$, with confidence $1 - \delta$ we have

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z},\lambda})) \leq 2\varepsilon^*(m,\delta) + \frac{2\lambda}{\Delta^2},$$

where $\varepsilon^*(m,\delta)$ is the unique positive solution to the equation

$$\log \mathcal{N}(B_1, \frac{\Delta\varepsilon}{4}) - \frac{3m\varepsilon}{128(1+\kappa/\Delta)} = \log\delta.$$

In particular, if $\log \mathcal{N}(B_1,\eta) = O((\log(1/\eta))^p)$ for some $p > 0$, then

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z},\lambda})) \leq \widetilde{c}\left\{\frac{(\log m)^p}{m}\log(2/\delta)\big(1+\frac{1}{\Delta}\big) + \frac{\lambda}{\Delta^2}\right\}.$$

A weaker separable condition was introduced in [17] as follows.

Definition 40. We say that the probability distribution ρ is *(weakly) separable* by $\mathcal{H}_K$ if there is some $f_{\mathrm{sp}} \in \mathcal{H}_K$ such that $\|f_{\mathrm{sp}}\|_K = 1$ and $yf_{\mathrm{sp}}(x) > 0$ almost surely. It has *separation exponent* $\theta \in (0,\infty]$ if there are positive constants Δ, c_θ such that

$$\rho_X\left\{x \in X : |f_{\mathrm{sp}}(x)| < \Delta t\right\} \leq c_\theta t^\theta, \qquad \forall t > 0.$$

When $\theta = \infty$, this concept is reduced to the strictly separable condition.

Take a regularizing function $f_\lambda = (c_\theta/\lambda)^{\frac{1}{2+\theta}}\Delta^{-\frac{\theta}{2+\theta}} f_{\mathrm{sp}}$ and consider the error decomposition

$$\mathcal{E}^{\phi}(f_{\mathbf{z},\lambda}) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2 \leq \mathcal{E}^{\phi}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}^{\phi}(f_{\mathbf{z},\lambda}) + \mathcal{E}_{\mathbf{z}}^{\phi}(f_\lambda) - \mathcal{E}^{\phi}(f_\lambda) + \mathcal{E}^{\phi}(f_\lambda) + \lambda\|f_\lambda\|_K^2.$$

We can apply the ratio probability inequality, Proposition 26, with $\tau = 1$ again to the function set B_R with $R := 2(c_\theta/\lambda)^{\frac{1}{2+\theta}}\Delta^{-\frac{\theta}{2+\theta}} + \sqrt{\frac{7\log(2/\delta)}{6m\lambda}}$. Then we know that if $\log \mathcal{N}(B_1,\eta) = O((\log(1/\eta))^p)$ for some $p > 0$, then with confidence $1-\delta$,

$$\mathcal{R}(\operatorname{sgn}(f_{\mathbf{z},\lambda})) \leq \widetilde{c}\left\{m^{-\frac{\theta}{1+\theta}}(\log m)^p + \frac{\log(2/\delta)}{m} + \lambda^{\frac{\theta}{2+\theta}}\right\}.$$

The previous two noise conditions are for deterministic distributions since $\mathcal{R}(f_c) = 0$. Another noise condition, defined in [73], is also for non-deterministic distributions.

Definition 41. We say that the probability distribution ρ has *Tsybakov's noise exponent* $\zeta \in [0,\infty]$ if for some constant $c_\zeta > 0$,

$$\rho_X\left\{x \in X : |f_\rho(x)| \leq c_\zeta t\right\} \leq t^\zeta, \qquad \forall t > 0.$$

For some detailed discussions, see [67]. The Tsybakov's noise condition helps improve the variancing exponent τ in (26). The following result can be found in [70]. We give a simple proof in [80].

Theorem 42. *If ρ has Tsybakov's noise exponent $\zeta \in [0,\infty]$, then for every function $f : X \to [-1,1]$, (26) holds with $\tau = \frac{\zeta}{\zeta+1}$, $M = 1$ and $C_1 = 8\big(1/(2c_\zeta)\big)^{\zeta/(\zeta+1)}$.*

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

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

and

$$\int \Big(\phi(yf(x)) - \phi(yf_c(x))\Big)^2 d\rho = \int_X |f_c(x) - f(x)|^2 d\rho_X.$$

Let $t > 0$ and separate the domain X into two sets: $X_t^+ = \{x \in X : |f_\rho(x)| > c_\zeta t\}$ and $X_t^- = \{x \in X : |f_\rho(x)| \le c_\zeta t\}$. On X_t^+ we have $|f_c(x) - f(x)|^2 \le 2|f_c(x) - f(x)| \frac{|f_\rho(x)|}{c_\zeta t}$. On X_t^- we have $|f_c(x) - f(x)|^2 \le 4$. It follows from the Tsybakov's noise condition that

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

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

Theorem 42 together with Theorems 37 and 33 yields the following convergence rate.

Corollary 43. *Let $\phi(t) = \max\{1-t, 0\}$. Assume (25) and (27) for some $\beta \in (0,1]$. If ρ has Tsybakov's noise exponent $\zeta \in [0,\infty]$, then for any $\epsilon > 0$ and $0 < \delta < 1$, there is a constant $\widetilde{c}$ independent on m such that with confidence $1-\delta$*

$$\mathcal{R}(sgn(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c) \le \widetilde{c} m^{-\theta},$$

where $\theta := \min\left\{\frac{(\zeta+1)\beta}{\beta(\zeta+2)+(\zeta+1-\beta)p/2} - \epsilon, \frac{2\beta}{\beta+1}\right\}$.

The rate derived in [87] is $\theta = \frac{\beta}{1+\beta}$, at most $1/2$ when $\beta = 1$. The rate in [70] is $\theta = \frac{4\beta(1+\zeta)}{(1+\beta)(4+2\zeta+p\zeta)} - \epsilon$. Corollary 43 improves these results because of the projection operator. Another noise condition can be found in [70,61].

5. Further Topics

Learning theory is a rapidly growing field. Many people are working on it, theoretically or practically, from different points of view. But there are still a lot of open

areas. For some topics, even fundamental questions remain unsolved. In this survey, we only discuss the regularization approach. It works well with reproducing kernel Hilbert spaces, for regression and classification. Let us briefly mention some further research directions involving regularization schemes.

(1) *Feature selection.* One purpose is to understand structures of high dimensional data. Topics include manifold learning (e.g. [8,53,27,18]) and dimensionality reduction (see [32] and references therein). Another purpose is to determine important features (variables) of huge dimensional functions. Two approaches are filter method and wrapper method [37]. There are some attempts to use regularization schemes like support vector machines for feature selection, e.g. [33,34]. But not so much is known about feature selection in general. See the special issue [32]. Recently, a least-square type algorithm for learning covariances via gradients was introduced in [50]. It involves solving linear systems only and provides promising results for variable selection in some applications of gene expression analysis.

(2) *Multi-kernel regularization schemes.* Let $K_\Sigma = \{K_\sigma : \sigma \in \Sigma\}$ be a set of Mercer kernels on X such as Gaussian kernels with variances σ^2 running over $(0, \infty)$. The *multi-kernel regularization scheme* associated with K_Σ is defined as

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

Here $V : \mathbb{R}^2 \to \mathbb{R}_+$ is a general loss function. This multi-kernel scheme is motivated by recent work on learning algorithms with varying kernels. In [16] support vector machines with multiple parameters are investigated. In [40,57] mixture density estimation is considered and Gaussian kernels with variance σ^2 changing on an interval $[\sigma_1^2, \sigma_2^2]$ with $0 < \sigma_1 < \sigma_2 < +\infty$ are used for deriving bounds. Multi-task learning algorithms involve kernels from a convex hull of several Mercer kernels and spaces with changing norms, e.g. [30,35]. Learning of kernel functions is studied in [38,47]. The existence of minimizers of the multi-kernel scheme has been verified for convex hulls of several kernels in [38,47], and for the case when the set of kernels is compact with respect to a metric in Σ in [78], based on the dual formulation in [88]. The error analysis of multi-kernel regularization scheme for classification with has been done in [78].

Problem 44. It would be desirable to have good algorithms for the multi-kernel regularization scheme.

Another related class of multi-kernel regularization schemes are those generated by polynomial kernels $\{K_d(x, y) = (1 + x \cdot y)^d\}$ with $d \in \mathbb{N}$, or in general dot product kernels [54,42]. In [92] we provide some convergence rates in the univariate case $n = 1$ for the multi-kernel regularized classifiers generated by polynomial kernels. The method is Durrmeyer operators from approximation theory. But the situation in the multivariate case is not so clear due to the difficulty of bounding the RKHS norm of the multivariate Durrmeyer operators.

(3) *Regularization with penalty functionals other than* $\|f\|_K^2$. One interesting class deals with the KL-divergence or the entropy such as the maximum entropy discrimination [36]. The convergence and error analysis is not so clear due to the complicated penalty functional (which is not symmetric). Another class is the ν-SVM algorithms [60]. In [80] we investigate the linear programming SVM. For this algorithm, the hypothesis space depends on the sample $\mathbf{z}$, which makes the analysis difficult [77]. We get some error bounds and show that the convergence of the linear programming SVM is not slower than the 1-norm soft margin SVM [80]. But it is unknown whether the convergence of the linear programming SVM is actually faster. Other regularization schemes include one-class SVM [72] and learning of vector-valued functions [46].

(4) *Online learning algorithms.* These algorithms improve the efficiency of learning methods. The convergence is investigated in [75,15], while the error with respect to the step size has been analyzed for the least-square regularized regression in [63] and for regularized classification with a general classifying loss in [84]. But the analysis for online algorithms and related game theory problems needs further study.

(5) *Connections to sampling theory and signal processing.* In many fields like sampling theory, signal processing, and inverse problems [29], the nodes $\overline{x} = \{x_i\}_{i=1}^m$ are deterministic and one needs to recover the function relation from values at these nodes. So the regularization scheme has the same form, but only $\{y_i\}_{i=1}^m$ are (random) samples. Some error analysis is done in [65]. In particular, we consider the following scheme

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

Here $\bar{t}$ is a discrete subset of X and $\mathcal{H}_{K,\bar{t}}$ is the closed subspace of $\mathcal{H}_K$ generated by $\{K_t\}_{t\in\bar{t}}$. The convergence of the above scheme for the case when $\bar{t} \subset \overline{x}$ is part of the random sample, by means of $\#\bar{t}/m \to 0$ has not been investigated so far.

Acknowledgments

This work is supported partially by the Research Grants Council of Hong Kong, Project No. CityU 1144/01P.

References

1. Alon, N., Ben-David, S., Cesa-Bianchi, N. and Haussler, D., Scale-sensitive dimensions, uniform convergence and learnability, *J. ACM* **44** (1997), 615–631.
2. Anthony, M. and Bartlett, P. L., *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, Cambridge, 1999.

3. Aronszajn, N., Theory of reproducing kernels, *Trans. Amer. Math. Soc.* **68** (1950), 337–404.
4. Barron, A. R., Complexity regularization with applications to artificial neural networks, in: *Nonparametric Functional Estimation* (G. Roussa, Ed.), Kluwer, Dordrecht, 1990, pp. 561–576.
5. Bartlett, P. L., 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. Information Theory* **44** (1998), 525–536.
6. Bartlett, P. L., Bousquet, O. and Mendelson, S., Local Rademacher complexities, *Ann. Stat.*, to appear.
7. Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D., Convexity, classification, and risk bounds, preprint, 2003.
8. Belkin, M. and Niyogi, P., Semisupervised learning on Riemannian manifolds, *Machine Learning* **56** (2004), 209–239.
9. Binev, P., Cohen, A., Dahmen, W., DeVore, R. and Temlyakov, V., Universal algorithms for learning theory, Part I: piecewise constant functions, preprint, 2004.
10. Blanchard, B., Bousquet, O. and Massart, P., Statistical performance of support vector machines, preprint, 2003.
11. Blanchard, B., Lugosi, G. and Vayatis, N., On the rate of convergence of regularized boosting classifiers, *J. Mach. Learning Res.* **4** (2003), 861–894.
12. Breiman, L., Arcing classifiers, *Ann. Stat.* **26** (1998), 801–824.
13. Boser, B. E., Guyon, I. and Vapnik, V., A training algorithm for optimal margin classifiers, in: *Proc. Fifth Annual Workshop of Computational Learning Theory* (D. Haussler, Ed.), ACM, Pittsburgh, 1992, pp. 144–152.
14. Bousquet, O. and Elisseeff, A., Stability and generalization, *J. Mach. Learning Res.* **2** (2002), 499–526.
15. Cesa-Bianchi, N., Long, P. M. and Warmuth, M. K., Worst-case quadratic loss bounds for prediction using linear functions and gradient descent, *IEEE Trans. Neural Networks* **7** (1996), 604–619.
16. Chapelle, O., Vapnik, V., Bousquet, O. and Mukherjee, S., Choosing multiple parameters for support vector machines, *Machine Learning* **46** (2002), 131–159.
17. Chen, D. R., Wu, Q., Ying, Y. and Zhou, D. X., Support vector machine soft margin classifiers: error analysis, *J. Machine Learning Res.* **5** (2004), 1143–1175.
18. Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F. and Zucker, S. W., Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps, *Proc. National Academy of Sciences of USA* **102** (2005), 7426–7431.
19. Cortes, C. and Vapnik, V., Support-vector networks, *Machine Learning* **20** (1995), 273–297.
20. Cristianini, C. and Shawe-Taylor, J., *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, 2000.
21. Cucker, F. and Smale, S., On the mathematical foundations of learning, *Bull. Amer. Math. Soc.* **39** (2001), 1–49.

22. Cucker, F. and Smale, S., Best choices for regularization parameters in learning theory: On the bias-variance problem, *Foundations Comput. Math.* **2** (2002), 413–428.
23. Cucker, F. and Zhou, D. X., *Learning Theory: an Approximation Theory Viewpoint*, Cambridge University Press, in preparation.
24. De Vito, E., Caponnetto, A. and Rosasco, L., Model selection for regularized least-squares algorithm in learning theory, *Foundations Comput. Math.* **5** (2005), 59–85.
25. De Vito, E., Rosasco, L., Caponnetto, A., Piana, M. and Verri, A., Some properties of regularized kernel methods, *J. Machine Learning Res.* **5** (2004), 1363–1390.
26. Devroye, L., Györfi, L. and Lugosi, G., *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1997.
27. Donoho, D. L., For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution, preprint, 2004.
28. Dudley, R. M., Giné, E. and Zinn, J., Uniform and universal Glivenko-Cantelli classes, *J. Theor. Prob.* **4** (1991), 485–510.
29. Engl, H. W., Hanke, M. and Neubauer, A., *Regularization of Inverse Problems*, Mathematics and Its Applications, Vol. 375, Kluwer, Dordrecht, 1996.
30. Evgeniou T. and Pontil, M., Regularized multi-task learning, in: *Proc. 17th SIGKDD Conf. on Knowledge Discovery and Data Mining* (W. Kim, R. Kohavi, J. Gehrke and W. DuMouchel, Eds.), ACM, Seattle, 2004.
31. Evgeniou, T., Pontil, M. and Poggio, T., Regularization networks and support vector machines, *Adv. Comput. Math.* **13** (2000), 1–50.
32. Guyon, I. and Elisseeff, A., An introduction to variable and feature selection, *J. Machine Learning Res.* **3** (2003), 1157–1182.
33. Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., Gene selection for cancer classification using support vector machines, *Machine Learning* **46** (2002), 389–422.
34. Hardin, D., Tsamardinos, I. and Aliferis, C. F., A theoretical characterization of linear SVM-based feature selection, in: *21st Int. Conf. on Machine Learning*, Proc. Banff 2004 (C. E. Brodley, Ed.), ACM, Seattle, 2004.
35. Herbster, M., Relative loss bounds and polynomial-time predictions for the K-LMS-NET algorithm, in: *Proc. 15th Int. Conf. on Algorithmic Learning Theory*, (J. Case and A. Maruoka, Eds.), Lecture Notes in Computer Science, Vol. 3244, Springer, Berlin, 2004.
36. Jaakkola, T., Meila, M. and Jebara, T., Maximum entropy discrimination, MIT AI-Lab Technical Report, Boston, 1999.
37. Kohavi, R. and John, G., Wrappers for feature selection, *Artificial Intelligence* **97** (1997), 273–324.
38. Lanckriet, G. R. G., Cristianini, N., Bartlett, P., El Ghaoui, L. and Jordan, M. I., Learning the kernel matrix with semidefinite programming, *J. Machine Learning Res.* **5** (2004), 27–72.
39. Lee, W. S., Bartlett, P. and Williamson, R., The importance of convexity in learning with least square loss, *IEEE Trans. Information Theory* **44** (1998),

1974–1980.

40. Li, J. and Barron, A., Mixture density estimation, in: *Advances in Neural Information Processing Systems, Vol. 12* (S. A. Solla, T. K. Leen and K. R. Muller, Eds.), Morgan Kaufman Publ., San Mateo, 1999, pp. 279–285.
41. Lin, Y., Support vector machines and the Bayes rule in classification, *Data Mining and Knowledge Discovery* **6** (2002), 259–275.
42. Lu, F. and Sun, H., Positive definite dot product kernels in learning theory, *Adv. Comput. Math.* **22** (2005), 181–198.
43. Lugosi, G. and Vayatis, N., On the Bayes-risk consistency of regularized boosting methods, *Ann. Stat.* **32** (2004), 30–55.
44. McDiarmid, C., Concentration, in: *Probabilistic Methods for Algorithmic Discrete Mathematics* (M. Habib, C. McDiarmid, J. Ramirez-Alfonsin and B. Reed, Eds.), Springer, Berlin, 1998, pp. 195–248.
45. Mendelson, S., Improving the sample complexity using global data., *IEEE Trans. Information Theory* **48** (2002), 1977–1991.
46. Micchelli, C. A. and Pontil, M., On learning vector-valued functions, *Neural Comp.* **17** (2005), 177–204.
47. Micchelli, C. A. and Pontil, M., Learning the kernel function via regularization, technical report, Dept. of Computer Science, University College London, 2004.
48. Micchelli, C. A., Pontil, M., Wu, Q. and Zhou, D. X., Error bounds for learning the kernel, preprint, 2005.
49. Mukherjee, S., Niyogi, P., Poggio, T. and Rifkin, R., Learning theory: stability is sufficient for generalization and necessary and sufficient for empirical risk minimization, *Adv. Comput. Math.*, to appear.
50. Mukherjee, S. and Zhou, D. X., Learning covariances via gradients, preprint, 2005.
51. Niyogi, P., *The Informational Complexity of Learning*, Kluwer, Dordrecht, 1998.
52. Niyogi, P. and Girosi, F., On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions, *Neural Comp.* **8** (1996), 819–842.
53. Niyogi, P., Smale, S. and Weinberger, S., Finding the homology of submanifolds with high confidence from random samples, preprint, 2004.
54. Pinkus, P., Strictly positive definite kernels on a real inner product space, *Adv. Comput. Math.* **20** (2004), 263–271.
55. Poggio, T., Rifkin, R., Mukherjee, S. and Niyogi, P., General conditions for predictivity in learning theory, *Nature* **428** (2004), 419–422.
56. Poggio, T. and Smale, S., The mathematics of learning: dealing with data, *Notices Amer. Math. Soc.* **50** (2003), 537–544.
57. Rakhlin, A., Panchenko, D. and Mukherjee, S., Risk bounds for mixture density estimation, *ESAIM: Probability and Statistics* **9** (2005), 220-229.
58. Schaback, R. and Werner, J., Linearly constrained reconstruction of functions by kernels, with applications to machine learning, *Adv. Comput. Math.*, to appear.

59. Schölkopf, B. and Smola, A. J., *Learning with Kernels*, MIT Press, Cambridge, 2002.
60. Schölkopf, B., Smola, A. J., Williamson, R. C. and Bartlett, P. L., New support vector algorithms, *Neural Comp.* **12** (2000), 1207–1245.
61. Scovel, C., Hush, D. and Steinwart, I., Learning rates for density level detection, *Anal. Appl.*, to appear.
62. Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C. and Anthony, M., Structural risk minimization over data-dependent hierarchies, *IEEE Trans. Information Theory* **44** (1998), 1926–1940.
63. Smale, S. and Yao, Y., Online learning algorithms, *Found. Comput. Math.*, to appear.
64. Smale, S. and Zhou, D. X., Estimating the approximation error in learning theory, *Anal. Appl.* **1** (2003), 17–41.
65. Smale, S. and Zhou, D. X., Shannon sampling and function reconstruction from point values, *Bull. Amer. Math. Soc.* **41** (2004), 279–305.
66. Smale, S. and Zhou, D. X., Shannon sampling II. Connections to learning theory, *Appl. Comput. Harmonic Analysis*, to appear.
67. Smale, S. and Zhou, D. X., Learning theory estimates via integral operators and their approximations, *Constr. Approximation*, special issue on Learning Theory, submitted.
68. Steinwart, I., Support vector machines are universally consistent, *J. Complexity* **18** (2002), 768–791.
69. Steinwart, I., On the influence of the kernel on the consistency of support vector machines, *J. Machine Learning Res.* **2** (2001), 67–93.
70. Steinwart, I. and Scovel, C., Fast rates for support vector machines, in: *Proceedings of the Conference on Learning Theory, COLT-2005*, to appear.
71. Sun, H., Mercer theorem for RKHS on noncompact sets, *J. Complexity*, to appear.
72. Tax, D. M. J. and Duin, R. P. W., Support vector domain description, *Pattern Recognition Letters* **20** (1999), 1191–1199.
73. Tsybakov, A. B., Optimal aggregation of classifiers in statistical learning, *Ann. Stat.* **32** (2004), 135–166.
74. van der Vaart, A. W. and Wellner, J. A., *Weak Convergence and Empirical Processes*, Springer, New York, 1996.
75. Vapnik, V., *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.
76. Wahba, G., *Spline Models for Observational Data*, CBMS-NSF Reg. Conf. Ser. in Applied Mathematics, Vol. 59, SIAM, Philadelphia, 1990.
77. Wu, Q., *Classification and Regularization in Learning Theory*, PhD thesis, City University of Hong Kong, 2005.
78. Wu, Q., Ying, Y. and Zhou, D. X., Multi-kernel regularized classifiers, submitted, 2004.
79. Wu, Q., Ying, Y. and Zhou, D. X., Learning rates of least-square regularized regression, *Found. Comput. Math.*, to appear.
80. Wu, Q. and Zhou, D. X., SVM soft margin classifiers: linear programming versus quadratic programming, *Neural Comp.* **17** (2005), 1160–1187.

81. Wu, Q. and Zhou, D. X., Analysis of support vector machine classification, submitted, 2004.
82. Ying, Y., McDiarmid inequalities of Bernstein and Bennett forms, Technical Report, City University of Hong Kong, 2004.
83. Ying, Y. and Zhou, D. X., Learnability of Gaussians with flexible variances, submitted, 2004.
84. Ying, Y. and Zhou, D. X., Online regularized classification algorithms, preprint, 2005.
85. Yurinsky, Y., *Sums and Gaussian Vectors*, Lecture Notes in Mathematics, Vol. 1617, Springer, Berlin, 1995.
86. Zhang, T., Leave-one-out bounds for kernel methods, *Neural Comp.* **15** (2003), 1397–1437.
87. Zhang, T., Statistical behavior and consistency of classification methods based on convex risk minimization, *Ann. Stat.* **32** (2004), 56–85.
88. Zhang, T., On the dual formulation of regularized linear systems with convex risks, *Machine Learning* **46** (2002), 91–129.
89. Zhou, D. X., The covering number in learning theory, *J. Complexity* **18** (2002), 739–767.
90. Zhou, D. X., Capacity of reproducing kernel spaces in learning theory, *IEEE Trans. Information Theory* **49** (2003), 1743–1752.
91. Zhou, D. X., Density problem and approximation error in learning theory, preprint, 2003.
92. Zhou, D. X. and Jetter, K., Approximation with polynomial kernels and SVM classifiers, *Adv. Comput. Math.*, to appear.