International Journal of Wavelets, Multiresolution
and Information Processing
Vol. 9, No. 4 (2011) 549–570

DOI: 10.1142/S0219691311004213

GENERALIZATION BOUNDS OF REGULARIZATION ALGORITHMS DERIVED SIMULTANEOUSLY THROUGH HYPOTHESIS SPACE COMPLEXITY, ALGORITHMIC STABILITY AND DATA QUALITY

XIANGYU CHANG* and ZONGBEN XU†

Institute for Information and System Sciences
Xi'an Jiaotong University, Xi'an 710049, P. R. China
**xiangyuchang@gmail.com*
†*zbxu@mail.xjtu.edu.cn*

BIN ZOU

Department of Mathematics, Hubei University
Wuhan 430062, P. R. China
zoubin0502@hubu.edu.cn

HAI ZHANG

Department of Mathematics, Northwest University
Xi'an 710069, P. R. China
zhanghainwu@tom.com

Received 3 December 2009
Revised 31 March 2010

A main issue in machine learning research is to analyze the generalization performance of a learning machine. Most classical results on the generalization performance of regularization algorithms are derived merely with the complexity of hypothesis space or the stability property of a learning algorithm. However, in practical applications, the performance of a learning algorithm is not actually affected only by an unitary factor just like the complexity of hypothesis space, stability of the algorithm and data quality. Therefore, in this paper, we develop a framework of evaluating the generalization performance of regularization algorithms combinatively in terms of hypothesis space complexity, algorithmic stability and data quality. We establish new bounds on the learning rate of regularization algorithms based on the measure of uniform stability and empirical covering number for general type of loss functions. As applications of the generic results, we evaluate the learning rates of support vector machines and regularization networks, and propose a new strategy for regularization parameter setting.

Keywords: Learning rate; regularization algorithm; algorithmic stability; hypothesis space; sample error; regularization error.

AMS Subject Classification: 62B10, 62H30, 62G05

1. Introduction

Recently there has been a great increase in the interest for theoretical issues in the machine learning community, mainly due to the fact that statistical learning theory has demonstrated its usefulness by providing the ground of developing successful and well-founded learning algorithms such as support vector machines (SVMs).[23] This renewed interest for theory naturally boosted the development of performance bounds for learning machines.[2,3,7,8,17,20,24] Until recently, three main approaches have been proposed to study the generalization performance of a learning machine.

The first approach is based on the theory of uniform convergence of empirical risks to their expected risks.[2,3,23] The scholars utilize the measure of space complexity, for instance, the VC-dimension,[23] covering number,[7,8,18,24,27,28] V_γ-dimension and P_γ-dimension,[12] Rademarcher average[4] to estimate the upper bound of the difference between empirical risks and their expect risks. For example, Vapnik[23] first established the bounds of the rate of uniform convergence and on the relative uniform convergence for a set of loss functions based on VC-dimension, and then obtained the generalization bounds of ERM algorithms. Cucker and Smale[8] considered the least squares error through decomposing the error into the sample error and the approximation error, then obtained the generalization bound in terms of covering number of hypothesis space. Bousquet[3] applied Rademarcher average to establish the generalization bounds of ERM algorithms. However, all of these results are merely in terms of the complexity of hypothesis space, in other words, the obtained bounds for a learning algorithm are the same, even when different training samples are used.

The second approach is based on sensitivity analysis. The basic point of view is that for a good learning algorithm, the outputs of the algorithm should not have significant disturbance when the training set has a little change. According to this viewpoint, Devroye,[11] Bousquet,[5] Kutin and Niyogi[15] introduced various definitions of algorithmic stability, and then they obtained the generalization bounds of learning algorithms in terms of the measures of algorithmic stability, such as, uniform stability, error stability, and hypothesis stability. All of these results are independent of the complexity of hypothesis space, that is, how the hypothesis space influences the learning ability of a learning algorithm is totally ignored.

The third approach is based on the information of data, that is, the information of training samples. The basic point is that the performance of an algorithm is affected by randomness of input samples. To estimate the algorithmic performance, Bousquet,[3] Koltchinskii and Panchenko[16] used Rademarcher average (contains the samples information) and empirical covering number[23,19] (contains the samples information and the hypothesis space information) to obtain the generalization bound of learning algorithms.

It is our point of view that, in real application of machine learning, the performance of a learning algorithm is affected not only by the complexity of hypothesis space, stability of learning algorithm and data information, but by some other

factors like sampling mechanism and sample quality as well. More importantly, how those factors determine the performance of a learning algorithm is by no means in an independent and separate manner. It should be a consequence of synthesized and simultaneous action of all the involved factors. From this point of view, a more reasonable evaluation for performance of a learning algorithm should be consequence of such synthesized influence of all the factors. Therefore, in this paper we derive generalization bounds of regularization algorithms through combinatively using the measures of hypothesis space complexity, algorithmic stability and data quality. We will show that the obtained new generalization bounds generalize the previously known results[5,8] derived respectively from the space complexity and the algorithmic stability, and sharpen them in certain situations.

The paper is organized as follows: In Sec. 2, we will introduce the necessary notion and notations, and present several useful inequality tools. In Sec. 3, the bounds of the sample error will be developed in terms of uniform stability of the algorithm and empirical covering number. In Secs. 4 and 5, we will use the tool of K-functional theorems, and drive the bounds on the learning performance of regularization algorithms, particularly, the regularization networks and support vector machines. Finally, we conclude the paper with some useful remarks in Sec. 6.

2. Preliminaries

In this section we introduce the definitions and notations used throughout the paper.

2.1. *Notion and notations*

Let $(\mathcal{X}, d)$ be a compact metric space and $\mathcal{Y}$ is a subset of $\mathbb{R}$. Suppose that ρ is a fixed but unknown probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We consider a training sample set

$$\mathbf{z} = \{z_1 = (x_1, y_1), z_2 = (x_2, y_2), \ldots, z_m = (x_m, y_m)\}$$

of size m in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ drawn independent and identically distributed (i.i.d.) from the unknown distribution ρ. For the training sample set $\mathbf{z}$, we build, for all $i = 1, 2, \ldots, m$, a series of modified (change-one) sample sets as follows: replace the ith element of the samples set $\mathbf{z}$ by

$$\mathbf{z}^i = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m\},$$

where the sample z_i' is assumed to be drawn from $\mathcal{Z}$ according to the distribution ρ and independent from $\mathbf{z}$.

The goal of machine learning from the samples set $\mathbf{z}$ is to find a function $f : \mathcal{X} \to \mathcal{Y}$ such that when new unlabeled samples are given, the function f can forecast them reasonably. Let

$$\mathcal{E}(f) = \mathrm{E}_z[\ell(f, z)] = \int_{\mathcal{Z}} \ell(f, z) d\rho$$

be the expected risk (or error) of function f, where $\ell(f,z)$ is a non-negative loss function. In the past research,[21] the margin-based loss functions such as the hinge loss, the AdaBoost loss, the logistic loss and the least square loss are widely used in classification applications, and the distance-based loss functions such as the least squares loss, Huber's insensitive loss, the logistic loss, and the ε-insensitive loss are frequently adopted in regression applications. Our aim in the present paper is to discuss the general learning problems, so we will consider general forms of the loss functions $\ell(f,z)$ below.

The learning problem is thus to find a function from a hypothesis space based on the training set $\mathbf{z}$ so as to minimize the expected risk $\mathcal{E}(f)$. Since the distribution ρ is unknown and we only know the samples set $\mathbf{z}$, the minimizer of the expected risk cannot be directly computed. The Empirical Risk Minimization (ERM) principle[23] then advocate that instead of minimizing the expected risk, an approximate solution is found through minimizing the so-called empirical risk (or empirical error) defined by

$$\mathcal{E}_m(f) = \frac{1}{m}\sum_{i=1}^{m} \ell(f, z_i),$$

which is directly computable due to its free from the unknown distribution ρ for any given function f. To find the desired function by the ERM principle based on the training set $\mathbf{z}$, the most natural way is to restrict it to a hypothesis space $\mathcal{H}$. In general, the learning problem can be formulated as finding the minimizer of the expected risk over a hypothesis space $\mathcal{H}$. The problem is ill-posed in general, so a natural way for dealing with the problem is to use the regularization technique.[6,22] The classical regularization theory, as we will consider here, formulates the problem as a variational problem of finding the function f that minimizes the functional

$$\mathcal{R}_m(f) := \mathcal{E}_m(f) + \lambda \|f\|_K^2,$$

where λ is a real positive number, called the regularization parameter, and $\|\cdot\|_K$ is a norm in a Reproducing Kernel Hilbert Space[1] (RKHS) $\mathcal{H}$ which is defined by a positive semidefinite function K.

Assume $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous, symmetric and positive semidefinite function, that is, for any finite set of distinct points $\{x_1, x_2, \ldots, x_l\} \subset \mathcal{X}$, the matrix $(K(x_i, x_j))_{i,j=1}^{l}$ is positive semidefinite. Such a function is called a Mercer kernel. The RKHS $\mathcal{H}$ associated with the kernel K is defined as the linear span of the set of functions $\{K_x = K(x,\cdot) : \forall\, x \in \mathcal{X}\}$ with the inner product $\langle , \rangle_{\mathcal{H}} = \langle , \rangle_K$ satisfying $\langle K_x, K_y \rangle_K = K(x,y)$, that is, $\langle \sum_i \alpha_i K_{x_i}, \sum_j \beta_j K_{y_j} \rangle_K = \sum_{i,j} \alpha_i \beta_j K(x_i, y_j)$. The reproducing property takes the form

$$\langle K_x, f \rangle = f(x), \quad \forall\, x \in \mathcal{X}, \quad f \in \mathcal{H}$$

which then implies that for any $f \in \mathcal{H}$, $\|f\|_\infty \le \kappa \|f\|_K$, where

$$\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x,x)}. \tag{2.1}$$

Let f_ρ be a function minimizing the risk $\mathcal{E}(f)$ over all measurable functions, i.e.

$$f_\rho = \arg\min_f \mathcal{E}(f) = \arg\min_f \int_{\mathcal{Z}} \ell(f,z)d\rho. \tag{2.2}$$

We denote by $f_{\mathbf{z}}$ the function minimizing the regularization empirical risk $\mathcal{R}_m(f)$ over the RKHS $\mathcal{H}$, i.e.

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \mathcal{R}_m(f) = \arg\min_{f\in\mathcal{H}} \{\mathcal{E}_m(f) + \lambda\|f\|_K^2\}. \tag{2.3}$$

Then we consider the function $f_{\mathbf{z}}$ as an approximation of the target function f_ρ. The crucial problem is how small the difference $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)$. We will study this problem in the present paper.

To estimate the difference $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)$, we need some basic assumptions on the hypothesis space $\mathcal{H}$ and the loss function $\ell(f,z)$:

(i) We suppose that $\mathcal{H}$ is contained in a ball $B_R = \{f \in \mathcal{H} : \|f\|_K \le R, R > 0\}$ of a RKHS with a C^∞ Mercer kernel on a compact subset of an Euclidean space $\mathbb{R}^d$. The interested reader can consult Ref. 29 for various concrete examples of the hypothesis space $\mathcal{H}$. This assumption implies that there exists a constant C_h independent of $h > d, \eta > 0$ such that (Ref. 29)

$$\mathcal{N}(\mathcal{H},\eta) \le \exp\left\{\frac{RC_h}{\eta}\right\}^{\frac{2d}{h}}. \tag{2.4}$$

(ii) We denote $\mathcal{H}' = \mathcal{H} \cup \{f_\rho\}$, and define

$$B = \sup_{f\in\mathcal{H}'} \max_{z\in\mathcal{Z}} \ell(f,z), \quad L = \sup_{g_1,g_2\in\mathcal{H}', g_1\neq g_2} \max_{z\in\mathcal{Z}} \frac{|\ell(g_1,z) - \ell(g_2,z)|}{|g_1 - g_2|}.$$

We assume that B and L are both finite in this paper.

2.2. *Main tools*

In this subsection we introduce the notion of uniform stability and some useful inequality tools. Let $f_{\mathbf{z}^i}$ be a function minimizing the (change-one) empirical risk defined by the change-one sample set $\mathbf{z}^i, 1 \le i \le m$ over $\mathcal{H}$, i.e.

$$f_{\mathbf{z}^i} = \arg\min_{f\in\mathcal{H}} \left\{ \frac{1}{m}\sum_{j=1}^m \ell(f,z_j) + \lambda\|f\|_K^2 \right\}, \quad z_j \in \mathbf{z}^i. \tag{2.5}$$

Such $f_{\mathbf{z}^i}$ (also $f_{\mathcal{H}}$, $f_{\mathbf{z}}$) exists since $\mathcal{H}$ is compact by the assumption (i).

Definition 1.[5] The regularization algorithm (2.3) is said to be uniform stable with respect to the loss function $\ell(f,z)$, if there is a non-negative constant β_m (where m is the size of sample set $\mathbf{z}$) such that

$$\forall\, z \in \mathcal{Z}, \quad \forall\, i \in \{1,2,\ldots,m\}, \quad \|\ell(f_{\mathbf{z}},z) - \ell(f_{\mathbf{z}^i},z)\|_\infty \le \beta_m. \tag{2.6}$$

In this case, we say that the algorithm is β_m-uniform stable.

Remark 1. The quantity β_m measures the uniform stability extent of the learning algorithm, and it is normally assumed to be a function of $1/m$ with the property that $\beta_m \to 0$ as $m \to \infty$.[5]

Note that the minimization (2.3) is taken over the discrete quantity $\mathcal{E}_m(f)$, so, intuitively, we have to regulate the capacity of the function set $\mathcal{H}$. Here the capacity will be measured by the covering number and the empirical covering number in this paper.

Definition 2.[10] For a compact set $\mathcal{H}$ in a metric space and $\eta > 0$, the covering number $\mathcal{N}(\mathcal{H}, \eta)$ of the function class $\mathcal{H}$ is the minimal integer $k \in \mathbb{N}$ such that there exist k balls in $\mathcal{H}$ with radius η covering $\mathcal{H}$.

Definition 3.[19,23] Let $\mathcal{H}$ be a class of bounded functions defined on $\mathcal{X}$, and let

$$\mathbf{x} = (x_i)_{i=1}^m \in \mathcal{X}^m, \quad \mathcal{H}|_x = \{(f(x_i))_{i=1}^m : f \in \mathcal{H}\} \subset \mathbb{R}^m.$$

For $1 \le p \le \infty$, we define the p-norm empirical covering number of $\mathcal{H}$ by

$$\begin{aligned} \mathcal{N}_{p,x}(\mathcal{H}, \varepsilon) &= \mathcal{N}(\mathcal{H}|_x, \varepsilon, d_p), \\ \mathcal{N}_p(\mathcal{H}, \varepsilon, m) &= \sup_{x \in \mathcal{X}^m} \mathcal{N}_{p,x}(\mathcal{H}, \varepsilon), \\ \mathcal{N}_p(\mathcal{H}, \varepsilon) &= \sup_{m \in \mathbb{N}^+} \mathcal{N}_{p,x}(\mathcal{H}, \varepsilon), \end{aligned}$$

where $d_p(\mathbf{r_1}, \mathbf{r_2}) = (\frac{1}{m}\sum_{i=1}^m |r_{1i} - r_{2i}|^p)^{\frac{1}{p}}$ is the ℓ^p-metric on the Euclidean space $\mathbb{R}^m$ for all $\mathbf{r_1} = (r_{1i})_{i=1}^m, \mathbf{r_2} = (r_{2i})_{i=1}^m \in \mathbb{R}^m$.

Remark 2. The empirical covering number is a very important quantity which simultaneously contains the information of hypothesis space and training samples. The interested reader can consult Ref. 19 for the details.

To estimate the bound on learning performance of the regularization algorithm, we will use the following three useful inequalities. The first one is the classical Bernstein's inequality,[8] the second one is the inequality due to Cucker and Smale,[9] and the third one is the inequality developed by Wu, Ying and Zhou[25] with the square loss function.

Lemma 1.[8] *Let ξ be a random variable on a space $\mathcal{Z}$ with expectation $\mu = \mathrm{E}(\xi)$. If $|\xi(z) - \mu| \le M_1$ for almost all $z \in \mathcal{Z}$, and the variance $\sigma^2(\xi) = \sigma^2$ is known, then for all $\varepsilon > 0$,*

$$\mathrm{Prob}_{z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mu\right| \ge \varepsilon\right\} \le 2\exp\left\{\frac{-m\varepsilon^2}{2(\sigma^2 + M_1\varepsilon/3)}\right\}.$$

Lemma 2.[9] *Let $c_1, c_2 > 0$, and $s > q > 0$. Then the equation*

$$x^s - c_1 x^q - c_2 = 0$$

has a unique positive zero x^*. *In addition*

$$x^* \le \max\{(2c_1)^{1/(s-q)}, (2c_2)^{(1/s)}\}.$$

Lemma 3.[25] *Let* $f_{\mathbf{z}}$ *defined by* (2.3). *Then for any* $f \in \mathcal{H}$, *there holds*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le \{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\|f\|_K^2\} + \{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_m(f_{\mathbf{z}}) + \mathcal{E}_m(f) - \mathcal{E}(f)\}. \tag{2.7}$$

The first term in the right-hand side of (2.7) is called the regularization error, and the second term is called the sample error. Based on this estimation, to bound the difference $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)$, we have to first bound the regularization error and the sample error respectively. These will be done respectively in Secs. 3 and 4.

3. Bounds of Sample Error

In this section, we present an upper bound estimation on the sample error of regularization algorithms based on uniform stability, complexity of hypothesis and data quality. For this purpose, we first establish two new concentration inequalities.

Theorem 1. *Suppose that the regularization algorithm* (2.3) *is* β_m*-uniform stable with respect to the loss function* $\ell(f, z)$, *then for any* $\varepsilon > 0$,

$$\mathrm{Prob}_{z^m}\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_m(f_{\mathbf{z}}) \ge 2\varepsilon + 2\beta_m\} \le 2\mathcal{N}_1\left(\frac{\varepsilon}{4}, \mathcal{H}, m\right)\exp\left\{\frac{-m\varepsilon^2}{8B^2}\right\}.$$

Proof. Let $L_{\mathbf{z}}(f) = \mathcal{E}(f) - \mathcal{E}_m(f)$. By (2.5) and (2.6), we have that for all $i \in \{1, 2, \ldots, m\}$

$$\begin{aligned}
|L_{\mathbf{z}}(f_{\mathbf{z}}) - L_{\mathbf{z}}(f_{\mathbf{z}^i})| &= |\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_m(f_{\mathbf{z}}) - [\mathcal{E}(f_{\mathbf{z}^i}) - \mathcal{E}_m(f_{\mathbf{z}^i})]| \\
&\le |\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathbf{z}^i})| + |\mathcal{E}_m(f_{\mathbf{z}^i}) - \mathcal{E}_m(f_{\mathbf{z}})| \\
&\le \int_{\mathcal{Z}} |\ell(f_{\mathbf{z}}, z) - \ell(f_{\mathbf{z}^i}, z)| d\rho + \frac{1}{m}\sum_{j=1}^{m} |\ell(f_{\mathbf{z}}, z_j) - \ell(f_{\mathbf{z}^i}, z_j)| \\
&\le 2\beta_m.
\end{aligned}$$

So we have $|L_{\mathbf{z}}(f_{\mathbf{z}})| \le 2\beta_m + |L_{\mathbf{z}}(f_{\mathbf{z}^i})|$.

Assume that for all $i \in \{1, 2, \ldots, m\}, \sup_{1\le i\le m} |L_{\mathbf{z}}(f_{\mathbf{z}^i})| \le 2\varepsilon$. Then we get $|L_{\mathbf{z}}(f_{\mathbf{z}})| \le 2\beta_m + 2\varepsilon$. It follows that

$$\mathrm{Prob}_{z^m}\left\{\sup_{1\le i\le m} |L_{\mathbf{z}}(f_{\mathbf{z}^i})| \le 2\varepsilon\right\} \le \mathrm{Prob}_{z^m}\{|L_{\mathbf{z}}(f_{\mathbf{z}})| \le 2\beta_m + 2\varepsilon\}.$$

Thus

$$\begin{aligned}
\mathrm{Prob}_{z^m}\{|L_{\mathbf{z}}(f_{\mathbf{z}})| \ge 2\beta_m + 2\varepsilon\} &\le \mathrm{Prob}_{z^m}\left\{\sup_{1\le i\le m} |L_{\mathbf{z}}(f_{\mathbf{z}^i})| \ge 2\varepsilon\right\} \\
&\le \mathrm{Prob}_{z^m}\left\{\sup_{f\in\mathcal{H}} |L_{\mathbf{z}}(f)| \ge 2\varepsilon\right\}.
\end{aligned} \tag{3.1}$$

In addition, by Theorem 17.7 in Ref. 19, we have that for the above ε,

$$\mathrm{Prob}_{z^m}\left\{\sup_{f\in\mathcal{H}}|L_{\mathbf{z}}(f)|\le 2\varepsilon\right\}\ge 1-4\mathcal{N}_1\left(\frac{\varepsilon}{4},\mathcal{H},m\right)\exp\left\{\frac{-m\varepsilon^2}{8B^2}\right\}. \tag{3.2}$$

By inequalities (3.1) and (3.2), we get

$$\mathrm{Prob}_{z^m}\{|\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}_m(f_{\mathbf{z}})|\le 2\varepsilon+2\beta_m\}\ge 1-4\mathcal{N}_1\left(\frac{\varepsilon}{4},\mathcal{H},m\right)\exp\left\{\frac{-m\varepsilon^2}{8B^2}\right\}.$$

This implies Theorem 1. □

Remark 3. In the classical statistical learning literature,[23] the quantity $\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}_m(f_{\mathbf{z}})$ is called the generalization error of the function $f_{\mathbf{z}}$. Vapnik,[23] Bousquet,[5] Cucker and Smale[8] have made some upper bound estimations on the generalization error respectively in terms of VC-dimension, algorithmic stability and covering number. They have, however, only considered the effect of each respective measure. Differently from those in Refs. 5, 8 and 23, Theorem 1 here gives the bound of generalization error simultaneously through algorithmic stability and the empirical covering number which contains the information of space complexity and data quality.

However, it should be noted that to estimate the empirical covering number is very difficult in general, so we present another estimation based on using the covering number with uniform metric below.

Theorem 2. *Assume that the regularization learning algorithm (2.3) is β_m-uniform stable with respect to the loss function $\ell(f,z)$, and the variance $D[\ell(f,z)]\le\sigma^2$ for any $f\in\mathcal{H}$ and $z\in\mathcal{Z}$. Then for any $\delta\in(0,1]$, the inequality*

$$\mathcal{E}(f_{\mathbf{z}})\le\mathcal{E}_m(f_{\mathbf{z}})+\varepsilon(m,\delta)+2\beta_m$$

holds with confidence at least $1-\delta$ provided that

$$m\ge\max\left\{\frac{4\ln(1/\delta)(\sigma^2+B^2/3)}{B^2},\frac{4(\sigma^2+B^2/3)(2LRC_h)^{\frac{2d}{h}}}{B^{\frac{2h+2d}{h}}}\right\},$$

where

$$\varepsilon(m,\delta)\le 2\max\left\{\left[\frac{4\ln(1/\delta)(\sigma^2+B^2/3)}{m}\right]^{\frac{1}{2}},\left[\frac{4(\sigma^2+B^2/3)(2LRC_h)^{\frac{2d}{h}}}{m}\right]^{\frac{h}{2d+2h}}\right\}.$$

Proof. We split the proof into three steps.

Step 1. By inequality (3.1), we get that for any $\varepsilon>0$,

$$\mathrm{Prob}_{z^m}\{|L_{\mathbf{z}}(f_{\mathbf{z}})|\ge 2\beta_m+2\varepsilon\}\le\mathrm{Prob}_{z^m}\left\{\sup_{f\in\mathcal{H}}|L_{\mathbf{z}}(f)|\ge 2\varepsilon\right\}. \tag{3.3}$$

Now we bound the term of the right-hand side of inequality (3.3). Indeed,

$$\mathrm{Prob}_{z^m}\left\{\sup_{f\in\mathcal{H}}|L_{\mathbf{z}}(f)|\geq 2\varepsilon\right\}=\mathrm{Prob}_{z^m}\left\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\right\}.$$

Let $\mathcal{H}=\mathcal{H}_1\cup\mathcal{H}_2\cup\cdots\cup\mathcal{H}_n$. A same argument with that conducted in Ref. 8, shows that for any $\varepsilon>0$, whenever

$$\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon,$$

there exists $k, 1\leq k\leq n$, such that

$$\sup_{f\in\mathcal{H}_k}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon.$$

This implies the equivalence

$$\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\Leftrightarrow\exists\, k,\quad 1\leq k\leq n,\quad \text{s.t.}\ \sup_{f\in\mathcal{H}_k}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon. \tag{3.4}$$

By equivalence (3.4), and by the fact that the probability of a union of events is bounded by the sum of the probabilities of these events, we obtain

$$\mathrm{Prob}_{z^m}\left\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\right\}\leq\sum_{k=1}^{n}\mathrm{Prob}_{z^m}\left\{\sup_{f\in\mathcal{H}_k}|\mathcal{E}(f)-\mathcal{E}_m(f)|\geq 2\varepsilon\right\}. \tag{3.5}$$

Step 2. We estimate the term of the right-hand side of inequality (3.5). Let the balls $D_k, k\in\{1,2,\ldots,n\}$, be a cover of $\mathcal{H}$ with center at f_k and radius $\frac{\varepsilon}{2L}$. Then, for all $\mathbf{z}\in\mathcal{Z}^m$ and all $f\in D_k$,

$$\begin{aligned}
|L_{\mathbf{z}}(f)-L_{\mathbf{z}}(f_k)|&\leq|\mathcal{E}(f)-\mathcal{E}(f_k)|+|\mathcal{E}_m(f)-\mathcal{E}_m(f_k)|\\
&\leq|\mathrm{E}_z[\ell(f,z)]-\mathrm{E}_z[\ell(f_k,z)]|+\left|\frac{1}{m}\sum_{i=1}^{m}\ell(f,z_i)-\frac{1}{m}\sum_{i=1}^{m}\ell(f_k,z_i)\right|\\
&\leq 2\|\ell(f,z)-\ell(f_k,z)\|_\infty\\
&\leq 2L\cdot\|f-f_k\|_\infty\leq 2L\cdot\frac{\varepsilon}{2L}=\varepsilon.
\end{aligned}$$

This then implies that for any $\mathbf{z}\in\mathcal{Z}^m$ and all $f\in D_k$

$$\sup_{f\in D_k}|L_{\mathbf{z}}(f)|\geq 2\varepsilon\Rightarrow|L_{\mathbf{z}}(f_k)|\geq\varepsilon.$$

We thus conclude that for any $k\in\{1,2,\ldots,n\}$,

$$\mathrm{Prob}_{z^m}\left\{\sup_{f\in D_k}|L_{\mathbf{z}}(f)|\geq 2\varepsilon\right\}\leq\mathrm{Prob}_{z^m}\left\{|L_{\mathbf{z}}(f_k)|\geq\varepsilon\right\}.$$

By Lemma 1, we have that for any $\varepsilon > 0$,

$$\mathrm{Prob}_{z^m}\left\{\sup_{f\in D_k}|L_{\mathbf{z}}(f)| \geq 2\varepsilon\right\} \leq 2\exp\left\{\frac{-m\varepsilon^2}{2(\sigma^2 + B\varepsilon/3)}\right\}. \tag{3.6}$$

From inequalities (3.5) and (3.6), we thus obtain

$$\mathrm{Prob}_{z^m}\left\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f) - \mathcal{E}_m(f)| \geq 2\varepsilon\right\} \leq 2\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{2L}\right)\exp\left\{\frac{-m\varepsilon^2}{2(\sigma^2 + B\varepsilon/3)}\right\}. \tag{3.7}$$

Combining inequalities (3.1) and (3.7) shows that for all $\varepsilon > 0$,

$$\mathrm{Prob}_{z^m}\{|\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_m(f_{\mathbf{z}})| \geq 2\beta_m + 2\varepsilon\} \leq 2\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{2L}\right)\exp\left\{\frac{-m\varepsilon^2}{2(\sigma^2 + B\varepsilon/3)}\right\}. \tag{3.8}$$

Step 3. We suppose $0 < \varepsilon < B$. Then the exponential part in the right-hand side of inequality (3.8) becomes

$$\frac{-m\varepsilon^2}{2(\sigma^2 + B\varepsilon/3)} \leq \frac{-m\varepsilon^2}{2(\sigma^2 + B^2/3)}.$$

By assumption (2.4), we have that for any $\varepsilon > 0$

$$\begin{aligned}&\mathrm{Prob}_{z^m}\{|\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_m(f_{\mathbf{z}})| \geq 2\beta_m + 2\varepsilon\}\\ &\leq 2\exp\left\{\left(\frac{\varepsilon}{2LRC_h}\right)^{\frac{-2d}{h}} - \frac{m\varepsilon^2}{2(\sigma^2 + B^2/3)}\right\}.\end{aligned} \tag{3.9}$$

Let us rewrite inequality (3.9) in an equivalent form. Write

$$\delta = \exp\left\{\left(\frac{\varepsilon}{2LRC_h}\right)^{\frac{-2d}{h}} - \frac{m\varepsilon^2}{2(\sigma^2 + B^2/3)}\right\}.$$

Then $0 < \delta \leq 1$ and it follows that

$$\varepsilon^{(2+\frac{2d}{h})} - \varepsilon^{\frac{2d}{h}} \cdot \frac{2\ln(1/\delta)(\sigma^2 + B^2/3)}{m} - \frac{2(\sigma^2 + B^2/3)(2LRC_h)^{\frac{2d}{h}}}{m} = 0.$$

By Lemma 2, we can find that the solution of above equation with respect to ε is given by

$$\varepsilon \doteq \varepsilon(m,\delta) \leq \max\left\{\left[\frac{4\ln(1/\delta)(\sigma^2 + B^2/3)}{m}\right]^{\frac{1}{2}}, \left[\frac{4(\sigma^2 + B^2/3)(2LRC_h)^{\frac{2d}{h}}}{m}\right]^{\frac{h}{2d+2h}}\right\}.$$

In addition, if

$$m \geq \max\left\{\frac{4\ln(1/\delta)(\sigma^2 + B^2/3)}{B^2}, \frac{4(\sigma^2 + B^2/3)(2LRC_h)^{\frac{2d}{h}}}{B^{\frac{2h+2d}{h}}}\right\}$$

we have $\varepsilon < B$. Thus by inequality (3.9), the proof of Theorem 2 is completed. □

Remark 4. The generalization bound provided in Theorem 2 is simultaneously through algorithmic stability and covering number with uniform metric. Different

from Theorem 1, the data information is not involved in the bound, however, as an outcome, the derived bound is computationally tractable.

Theorem 2 immediately implies the following bound on the sample error:

Proposition 1. *For any* $\delta \in (0, 1/2]$, *the inequality*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_m(f_{\mathbf{z}}) + \mathcal{E}_m(f) - \mathcal{E}(f) \le \varepsilon(m, \delta) + 2\beta_m + \left[\frac{2\ln(1/\delta)(\sigma^2 + B^2/3)}{m}\right]^{\frac{1}{2}}$$

holds with probability at least $1 - 2\delta$ *provided that*

$$m \ge \max\left\{\frac{4\ln(1/\delta)(\sigma^2 + B^2/3)}{B^2}, \frac{4(\sigma^2 + B^2/3)(2LRC_h)^{\frac{2d}{h}}}{B^{\frac{2h+2d}{h}}}\right\}.$$

Here

$$\varepsilon(m, \delta) \le \max\left\{\left[\frac{4\ln(1/\delta)(\sigma^2 + B^2/3)}{m}\right]^{\frac{1}{2}}, \left[\frac{4(\sigma^2 + B^2/3)(2LRC_h)^{\frac{2d}{h}}}{m}\right]^{\frac{h}{2d+2h}}\right\}.$$

Proof. By Lemma 1, we have that for any $\varepsilon > 0$,

$$\mathrm{Prob}_{z^m}\{|L_{\mathbf{z}}(f)| \ge \varepsilon\} \le 2\exp\left\{\frac{-m\varepsilon^2}{2(\sigma^2 + B\varepsilon/3)}\right\}.$$

So,

$$\mathrm{Prob}_{z^m}\{\mathcal{E}(f) - \mathcal{E}_m(f) \ge \varepsilon\} \le \exp\left\{\frac{-m\varepsilon^2}{2(\sigma^2 + B\varepsilon/3)}\right\}.$$

Let us suppose $\varepsilon < B$. Then for any $\delta \in (0, 1/2]$, let $\delta = \exp\{\frac{-m\varepsilon^2}{2(\sigma^2 + B^2/3)}\}$, we have

$$\varepsilon'(m, \delta) = \left[\frac{2\ln(1/\delta)(\sigma^2 + B^2/3)}{m}\right]^{\frac{1}{2}}.$$

Therefore, we conclude that for any $\delta \in (0, 1/2]$, with confidence at least $1 - \delta$, the inequality

$$\mathcal{E}(f) - \mathcal{E}_m(f) \le \varepsilon'(m, \delta) \tag{3.10}$$

holds as long as $m \ge \frac{2\ln(1/\delta)(\sigma^2 + B^2/3)}{B^2}$. In addition, by Theorem 2, we also have that for the same δ, the inequality

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_m(f_{\mathbf{z}}) \le \varepsilon(m, \delta) + 2\beta_m \tag{3.11}$$

holds with probability at least $1 - \delta$ provided

$$m \ge \max\left\{\frac{4\ln(1/\delta)(\sigma^2 + B^2/3)}{B^2}, \frac{4(\sigma^2 + B^2/3)(2LRC_h)^{\frac{2d}{h}}}{B^{\frac{2h+2d}{h}}}\right\}.$$

Combining inequalities (3.10) and (3.11), we then complete the proof of Proposition (3). □

In application (particularly, when it is specialized to concrete regularization algorithms), the parameters L, B appeared in Theorem 1, Theorem 2 and Proposition 1 have to be specified. While if the problem is given, parameter R usually is constant, that is hypothesis space is given (C_h is constant). The specification of those parameters depends tightly upon the form of loose function and the hypothesis space. As a preparation of application, we below give some remarks on specification of those parameters. The parameter L, as introduced in the basic assumption (ii) of Sec. 2, is a Lipschitz constant of loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$. To have a reasonable estimation on L, we first notice that, viewed as an unitary function, the loss function ℓ is convex whenever the basic assumption (ii) is met. This implies that

$$L = \sup_{g_1, g_2 \in \mathcal{H}, g_1 \neq g_2} \max_{z \in \mathcal{Z}} \frac{|\ell(g_1, z) - \ell(g_2, z)|}{|g_1 - g_2|}$$

is well defined and finite. Assume $\mathcal{Y} = [0, b]$ is the range of the learning machine. Normally it is assumed that for any $y \in \mathcal{Y}$, if $f(x) = 0$ then $\ell(f, y) \leq C_0$, we also note $\ell(0, y) \leq C_0$.

In what follows, we will mainly consider the square loss function $\ell(f, y) = (f(x)-y)^2$, the absolute value loss function $\ell(f, y) = |f(x)-y|$, and the ϵ-insensitive loss function $\ell(f, y) = |f(x) - y|_\epsilon$ for regression application. And we consider the square loss function $\ell(f, y) = (1 - yf(x))^2$, the hinge loss function $\ell(f, y) = \max\{1-yf(x), 0\} = |1-yf(x)|_+$, and the logistic loss function $\ell(f, y) = \frac{\ln(1+e^{-yf(x)})}{\ln 2}$ for classification. In those cases, the assumption $\ell(0, y) \leq C_0 (\forall\, y \in [0, b])$ immediately implies

$$|\ell(f, y) - \ell(0, y)| \leq L|f(x)|, \quad \forall\, f \in \mathcal{H}, \quad \forall\, (x, y) \in \mathcal{Z}$$

and $|\ell(f, y)| \leq Lb + |\ell(0, y)| \leq Lb + C_0 := B$. But this bound can be further improved in regression cases. We can calculate all the parameters, L, C_0, and B, as shown in Table 1.

Remark 5. In order to show significances and values of Theorems 1 and 2, we compare the results in Theorems 1 and 2 with the previously known fundamental results in Refs. 8 and 5. For regression problem, Cucker and Smale[8] established the bound (see Theorem B in Ref. 8) on the rate of the uniform empirical risks for least square loss function by Bernstein's inequality. They established the following basic theorem.

Table 1. The parameters specification in Theorems 1 and 2.

Problem	Loss	L	C_0	B
Regression	$(f(x) - y)^2$	$2b$	b^2	b^2
Regression	$\lvert f(x) - y\rvert$	1	b	b
Regression	$\lvert f(x) - y\rvert_\epsilon$	1	b	b
Classification	$(1 - yf(x))^2$	$2b + 2$	1	$2b^2 + 2b + 1$
Classification	$\lvert 1 - yf(x)\rvert_+$	1	1	$b + 1$
Classification	$\frac{\ln(1+e^{-yf(x)})}{\ln 2}$	$\frac{e^b}{(e^b+1)\ln 2}$	1	$\frac{be^b}{(e^b+1)\ln 2} + 1$

Theorem B.[8] *Let $\mathcal{H}$ be a compact subset of $C(\mathcal{X})$. Assume that, for all $f \in \mathcal{H}, |f(x) - y| \le M$ almost everywhere. Then, for all $\varepsilon > 0$,*

$$\mathrm{Prob}_{z^m}\left\{\sup_{f\in\mathcal{H}}|\mathcal{E}(f)-\mathcal{E}_m(f)|\ge\varepsilon\right\}\le 2\mathcal{N}\Big(\mathcal{H},\frac{\varepsilon}{8M}\Big)\exp\left\{\frac{-m\varepsilon^2}{4(2\sigma^2+M^2\varepsilon/3)}\right\},$$

where $\sigma^2 = \sigma^2(\mathcal{H}) = \sup_{f\in\mathcal{H}}\sigma^2(f_Y^2)$, $\mathcal{E}_m(f) = \frac{1}{m}\sum_{i=1}^m (f(x_i)-y_i)^2$.

Obviously, from Table 1 in the case $M = b$, so for $f_{\mathbf{z}} \in \mathcal{H}$, Theorem B can be rewritten as

$$\mathrm{Prob}_{z^m}\{\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}_m(f_{\mathbf{z}})\ge\varepsilon\}\le\mathcal{N}\Big(\mathcal{H},\frac{\varepsilon}{8b}\Big)\exp\left(-\frac{m\varepsilon^2}{4(2\sigma^2+1/3b^2\varepsilon)}\right).$$

On the other hand, we notice that inequality (3.8) in the proof of Theorem 2 can be written as

$$\mathrm{Prob}_{z^m}\{\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}_m(f_{\mathbf{z}})\ge\gamma\}\le\mathcal{N}\left(\mathcal{H},\frac{\gamma-2\beta_m}{8b}\right)\exp\left(-\frac{m(\gamma-2\beta_m)^2}{4(2\sigma^2+1/3b^2(\gamma-2\beta_m))}\right),$$

where $\gamma = 2\varepsilon + 2\beta_m$. Comparing these two estimations, we can find that Theorem B is just the special case of inequality (3.8) corresponding to $\beta_m = 0$. That is, the generalization bound derived in Ref. 8 is validated only for those learning algorithms that are absolutely stability, in the sense that for any two different input samples sets, the output of learning algorithms is always the same. This is clearly only suitable for the ideal algorithm. In real applications, the size m of training samples is finite, or small, any change of training samples inevitably lead to change of the output of a leaning algorithm. So an absolutely stable learning algorithm hardly exists. Instead, we can reasonably ask that a good learning algorithm should have some kinds of uniform stability, just as done in our Theorems 1 and 2.

Based on also the uniform stability framework, Bousquet and Elisseeff[5] established the following generalization bound of a learning algorithm:

$$\mathrm{Prob}\{|R-R_{emp}|\ge 2\varepsilon+2\beta_m\}\le 2\exp\left\{\frac{-8m\varepsilon^2}{(4m\beta_m+B)^2}\right\}, \tag{3.12}$$

where $R = \mathrm{E}_z[\ell(\mathcal{A},z)]$, $R_{emp} = \frac{1}{m}\sum_{i=1}^m \ell(\mathcal{A},z)$. Note that $\mathcal{A}$ here is corresponding to $f_{\mathbf{z}}$ defined in (2.3). Comparing inequality (3.12) with the inequality in Theorem 1, we can find that the left-hand side of inequalities (3.12) and the inequality in Theorem 1 have the same form. Thus with the same parameters ε, B, and m, when the empirical covering number $\mathcal{N}_1(\frac{\varepsilon}{4},\mathcal{H},m)$ satisfies

$$\ln 2\mathcal{N}_1\left(\frac{\varepsilon}{4},\mathcal{H},m\right)+\frac{8m\varepsilon^2}{(4m\beta_m+B)^2}=\frac{m\varepsilon^2}{8B^2}, \tag{3.13}$$

inequality (3.12) degenerates to the inequality in Theorem 1. This shows that the obtained generalization bounds in Theorem 1 also contain the previously known results in Ref. 5 as a special case. In addition, in Eq. (3.13), we have established an interesting connection between stability and the empirical covering number of hypothesis space $\mathcal{H}$, namely, we have proved that an algorithm having a search

space of finite empirical covering number must be uniform stable, and the uniform stability constant β_m, is bounded by the empirical covering number of the hypothesis space.

4. Bounds of Regularization Error

In this section, we estimate the regularization error and provide an upper bound estimation on the regularization error by using the K-functional tool. We first recall some results on the K-functional.

Let $L_K: \mathcal{L}^2_{\rho_x}(\mathcal{X}) \to \mathcal{L}^2_{\rho_x}(\mathcal{X})$ denote the following integral transform

$$(L_K f)(x) = \int K(x,t) f(t) d\rho_x(t), \quad \forall\, x \in \mathcal{X}, \tag{4.1}$$

where $\rho(y|x)$ is the conditional probability measure on x induced by ρ. Such transformation defines a Fredholm operator. It is known that when the function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a Mercer kernel and the Fredholm operator L_k is positive and self-adjoint, there exists an orthogonal basis $\{\phi_1, \phi_2, \ldots\}$ of $\mathcal{L}^2_{\rho_x}(\mathcal{X})$ consisting of eigenfunctions of L_K. If λ_k is the kth eigenvalue corresponding to the eigenfunction ϕ_k, then the set $\{\lambda_k\}$ is either finite or $\lambda_k \to 0$ when $k \to \infty$. It is known that the set $\{\sqrt{\lambda_k}\phi_k : \lambda_k > 0\}$ is an orthonormal system in RKHS $\mathcal{H}$.

Lemma 4.[10] *Suppose that the RKHS $\mathcal{H}$ is independent of the measure ρ_x. Then if ρ_x is non-degenerate and* $\dim(\mathcal{H}) = \infty$, *$L_K$ has infinitely many positive eigenvalues $\lambda_k, k \geq 1$, and*

$$\mathcal{H} = \left\{ f = \sum_{k=1}^{\infty} a_k \sqrt{\lambda_k}\phi_k : \{a_k\} \subset \ell^2 \right\}.$$

If $\dim(\mathcal{H}) < \infty, L_K$ has only m positive eigenvalues $\lambda_k, k \geq 1$, and

$$\mathcal{H} = \left\{ f = \sum_{k=1}^{p} a_k \sqrt{\lambda_k}\phi_k : \{a_k\} \subset \mathbb{R}^p \right\}.$$

By Lemma 4, we can define the mapping $L_K^{\frac{1}{2}} : \mathcal{L}^2_{\rho_x}(\mathcal{X}) \to \mathcal{H}$ through

$$L_K^{\frac{1}{2}}\Big(\sum a_k \phi_k\Big) = \sum a_k \sqrt{\lambda_k}\phi_k.$$

It is an isomorphism between the closed span of $\{\phi_k : \lambda_k > 0\}$ in $\mathcal{L}^2_{\rho_x}(\mathcal{X})$. Thus for any function $f \in \mathcal{H}$, there exists some $g \in \mathcal{L}^2_{\rho_x}(\mathcal{X})$ such that $f = L_K^{\frac{1}{2}} g$. It is easy to know that $\|f\|_K = \|g\|_{\mathcal{L}^2_{\rho_x}(\mathcal{X})}$.

Lemma 5.[8,10] *Let $\mathcal{F}$ be a Hilbert space and A a self-adjoint, strictly positive compact operator on $\mathcal{F}$. Let $s, r \in \mathbb{R}$ such that $s > r > 0$. Then for any $R > 0$ and $a \in \mathcal{F}$*

$$\min_{b, \text{s.t.} \|A^{-s}b\| \leq R} \|b - a\| \leq R^{\frac{r}{r-s}} \|A^{-r}a\|^{\frac{s}{s-r}}.$$

By Lemmas 4 and 5, we can establish the following bound on the regularization error of regularization algorithms.

Theorem 3. *Suppose $\mathcal{H}$ is a RKHS associated with a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let L_K be defined by (4.1), and $B_R = \{f \in \mathcal{H} : \|f\|_K \leq R\}$ with a fixed positive real number R. If $f_\rho \in \mathcal{L}^2_{\rho_x}(\mathcal{X})$, then there exists $\widetilde{f} \in B_R$ such that the regularization error of regularization algorithms satisfies*

$$\mathcal{E}(\widetilde{f}) - \mathcal{E}(f_\rho) + \lambda \|\widetilde{f}\|_K^2 \leq \lambda R^2 + LC_r R^{\frac{r}{2(r-1)}},$$

where $r \in (0, 1)$ is a constant and C_r is a constant dependent of r.

Proof. By Hölder inequality, for all $f \in \mathcal{H}$, we have

$$\begin{aligned}\{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda \|f\|_K^2\} &\leq \lambda R^2 + \int |\ell(f(x), y) - \ell(f_\rho(x), y)| d\rho \\ &\leq \lambda R^2 + L \int |f - f_\rho| d\rho_x \\ &\leq \lambda R^2 + L \|f - f_\rho\|^{\frac{1}{2}}_{\mathcal{L}^2_{\rho_x}(\mathcal{X})}.\end{aligned}$$

Applying Lemmas 4 and 5 with $\mathcal{F} = \mathcal{L}^2_{\rho_x}(\mathcal{X}), s = 1, A = L_K^{\frac{1}{2}}$ and $a = f_\rho$, and by using the fact that for any $f \in \mathcal{H}$

$$\|L_K^{-\frac{1}{2}} f\|_{\mathcal{L}^2_{\rho_x}(\mathcal{X})} = \|A^{-1} f\|_{\mathcal{L}^2_{\rho_x}(\mathcal{X})} = \|f\|_K,$$

we get that there exists $\widetilde{f} \in B_R$ such that

$$\widetilde{f} = \arg \min_{\|f\|_K \leq R} \|f - f_\rho\|_{\mathcal{L}^2_{\rho_x}(\mathcal{X})} = \arg \min_{\|L_K^{-\frac{1}{2}} f\|_{\mathcal{L}^2_{\rho_x}(\mathcal{X})} \leq R} \|f - f_\rho\|_{\mathcal{L}^2_{\rho_x}(\mathcal{X})},$$

and

$$\|\widetilde{f} - f_\rho\|_{\mathcal{L}^2_{\rho_x}(\mathcal{X})} \leq C_r R^{\frac{r}{(r-1)}}.$$

So take $f = \widetilde{f}$ and we can find

$$\{\mathcal{E}(\widetilde{f}) - \mathcal{E}(f_\rho) + \lambda \|\widetilde{f}\|_K^2\} \leq \lambda R^2 + LC_r R^{\frac{r}{2(r-1)}},$$

where r is a constant and satisfies $0 < r < 1$. This arrives to Theorem 3. □

Remark 6. In Theorem 3, we have supposed that $f_\rho \in \mathcal{L}^2_{\rho_x}(\mathcal{X})$, which is a normal assumption. If we further assume that f_ρ has more desirable property, say, differentiable, sharper bound of the regularization error can be derived.

5. Specifications to Regularization Algorithms

Combining the bound of the samples error (Proposition 3) derived in Sec. 4 and the bound of the regularization error (Theorem 3) established in Sec. 4, we immediately can obtain the following bound on the generalization performance of regularization algorithms.

Theorem 4. *Assume that the regularization algorithm (2.3) is β_m-uniform stable with respect to the loss function $\ell(f, z)$, the variance $D[\ell(f, z)] \leq \sigma^2$ for any $f \in \mathcal{H}$*

and $z \in \mathcal{Z}$, and the hypothesis space is a ball of a RKHS $\mathcal{H}$ associated with the Mercer kernel K, whose radius is R, that is, $B_R = \{f \in \mathcal{H} : \|f\|_K \le R\}$. Then for any $\delta \in (0, 1/2)$ and $r \in (0, 1)$, the estimation

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le \varepsilon(m, \delta) + \left\{ \frac{2\ln(1/\delta)(\sigma^2 + B^2/3)}{m} \right\}^{\frac{1}{2}} + \lambda R^2 + C_r R^{\frac{r}{2(r-1)}} + 2\beta_m$$

holds with confidence at least $1 - 2\delta$ provided that

$$m \ge \max\left\{ \frac{4\ln(1/\delta)(\sigma^2 + B^2/3)}{B^2}, \frac{4(\sigma^2 + B^2/3)(2LRC_h)^{\frac{2d}{h}}}{B^{\frac{2h+2d}{h}}} \right\},$$

where

$$\varepsilon(m, \delta) \le \max\left\{ \left[\frac{8\ln(1/\delta)(\sigma^2 + B^2/3)}{m} \right]^{\frac{1}{2}}, \left[\frac{8(\sigma^2 + B^2/3)(2LRC_h)^{\frac{2d}{h}}}{m} \right]^{\frac{h}{2d+2h}} \right\}.$$

Remark 7. Note that $\lambda(m) \to 0$ as $m \to \infty$. So, from Theorem 4, we can conclude

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \to 0, \quad \text{as } m \to \infty, \quad \text{and} \quad R \to \infty.$$

This shows that the regularization algorithm (2.3) is consistent. Additionally, the bound provided in Theorem 4 is dependent not only on the complexity of hypothesis space but also on the algorithmic stability. To our knowledge, such a bound estimation simultaneously through complexity and stability measures is the first time to be established.

Theorem 4 can be applied to derive generalization bounds or learning rate estimations of various concrete regularization algorithms. As example, in the following, we apply Theorem 4 to derive the learning rate of regularization networks and support vector machines (SVMs), and, by the way, to formulate a new strategy to choose the regularization parameter λ.

Example 1 (Regularization Networks). A regularization network[13,14] is an algorithm that is used to train a feedforward neural network. The algorithm has the following form

$$f_{\mathbf{z}} = \arg\min_{f \in B_R} \left\{ \frac{1}{m} \sum_{j=1}^{m} (f(x_i) - y_i)^2 + \lambda \|f\|_K^2 \right\}. \tag{5.1}$$

From Table 1, in this case, if suppose $\mathcal{Y} = [0, b]$ and $f(x) \in [0, b]$ for any $x \in \mathcal{X}$ and any $f \in B_R$, then we have $L = 2b$ and $B = b^2$, and furthermore, the uniform stability parameter β_m can be bounded by Ref. 5

$$\beta_m \le \frac{2\kappa^2 b^2}{\lambda m},$$

where κ is defined as in (2.1). Since the loss function of regularization networks generally take the form of is the square loss, we have

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x),$$

which is the minimizer of the expected risk over all measurable functions. Thus by Theorem 4 we easily obtain the following proposition.

Proposition 2. *For regularization networks, if we suppose* $D[f(x) - y] \le \sigma^2$ *and* $f(x) \in [0, b]$ *for any* $x \in \mathcal{X}$ *and* $f \in B_R$, *then for any* $\delta \in (0, 1/2]$, *the estimation*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le \varepsilon(m, \delta) + \left\{ \frac{2\ln(1/\delta)(\sigma^2 + b^4/3)}{m} \right\}^{\frac{1}{2}} + \lambda R^2 + C_r R^{\frac{r}{2(r-1)}} + 2\beta_m \tag{5.2}$$

holds with confidence at least $1 - 2\delta$. *In addition, from bound* (5.2), *we can suggest the strategy of setting the regularization parameter* λ *of regularization networks as* $\lambda := \lambda^* = \frac{2\kappa b}{R\sqrt{m}}$. *In this case, if, furthermore, take* $R = m^\xi$ *with* $0 < \xi < \min\{\frac{h}{2d}, \frac{1}{2}\}$, *then the learning rate satisfies*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le O(m^{-\eta}),$$

where $\eta = \min\{\frac{h - 2d\xi}{2d + 2h}, \frac{1}{2} - \xi, \frac{\xi r}{2(1-r)}\}$.

Proof. By Theorem 4 we have that for any $\delta \in (0, 1/2]$, the bound on learning rate of regularization networks obey to

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le \varepsilon(m, \delta) + \left\{ \frac{2\ln(1/\delta)(\sigma^2 + b^4/3)}{m} \right\}^{\frac{1}{2}} + \lambda R^2 + C_r R^{\frac{r}{2(r-1)}} + 2\beta_m \tag{5.3}$$

which holds with confidence at least $1 - 2\delta$ as long as

$$m \ge \max\left\{ \frac{4\ln(1/\delta)(\sigma^2 + b^4/3)}{b^4}, \frac{4(\sigma^2 + b^4/3)(4bRC_h)^{\frac{2d}{h}}}{(b)^{\frac{4(h+d)}{h}}} \right\}.$$

Here

$$\varepsilon(m, \delta) \le \max\left\{ \left[\frac{8\ln(1/\delta)(\sigma^2 + b^4/3)}{m} \right]^{\frac{1}{2}}, \left[\frac{8(\sigma^2 + b^4/3)(4bRC_h)^{\frac{2d}{h}}}{m} \right]^{\frac{h}{2d+2h}} \right\}.$$

For simplicity, let us denote the right-hand side of inequality (5.3) by $\Delta_1(m, R)$. Therefore, for any regularization parameter λ in (5.1), whenever $\beta_m = \frac{2\kappa^2 b^2}{\lambda m}$, we have

$$\Delta_1(m, R) = \varepsilon(m, \delta) + \left\{ \frac{2\ln(1/\delta)(\sigma^2 + b^4/3)}{m} \right\}^{\frac{1}{2}} + \frac{4\kappa^2 b^2}{\lambda m} + \lambda R^2 + C_r R^{\frac{r}{2(r-1)}}.$$

Because

$$\frac{4\kappa^2 b^2}{\lambda m} + \lambda R^2 \geq 2\sqrt{\frac{4\kappa^2 b^2 R^2}{m}} = \frac{4\kappa b R}{\sqrt{m}}, \tag{5.4}$$

we have

$$\Delta_1^*(m,R) := \varepsilon(m,\delta) + \left\{\frac{2\ln(1/\delta)(\sigma^2 + b^4/3)}{m}\right\}^{\frac{1}{2}} + \frac{4\kappa b}{\sqrt{m}} + C_r R^{\frac{r}{2(r-1)}},$$

where the equality (5.4) holds if and only if $\lambda = \lambda^* = \frac{2\kappa b}{R\sqrt{m}}$. Then we can find that

$$\lim_{R\to\infty}\lim_{m\to\infty}\Delta_1^*(m,R) = 0.$$

In addition, if we take $R = m^{\xi}, 0 < \xi < \min\{\frac{h}{2d}, \frac{1}{2}\}$, then

$$\Delta_1^*(m,R) = O(m^{-\eta}),$$

where $\eta = \min\{\frac{h-2d\xi}{2d+2h}, \frac{1}{2} - \xi, \frac{\xi r}{2(1-r)}\}$. This implies Proposition 5. □

Example 2 (SVMs). According to Ref. 23, SVMs are the regularization algorithms defined by

$$f_{\mathbf{z}} = \arg\min_{f\in B_R}\left\{\frac{1}{m}\sum_{j=1}^{m}|1 - y_i f(x_i)|_+ + \lambda\|f\|_K^2\right\},$$

where $|1 - yf(x)|_+ = \max\{1 - yf(x), 0\}$. By Remark 4, if we suppose $D[|f(x) - y|_+] \leq \sigma^2, f(x) \in [0,b]$ for any $x \in \mathcal{X}$ and any $f \in B_R$, and take $\mathcal{Y} = [0,b]$, and then $L = 1$ and $B = b + 1$. So the uniform stability parameter β_m can be bounded by Ref. 5

$$\beta_m \leq \frac{\kappa^2}{2\lambda m}.$$

Since the loss function of SVMs is the hinge loss, the function that minimizes the expected risk is given by Refs. 11 and 26

$$f_\rho(x) = \operatorname{sgn}\left(\int_{\mathcal{Y}} y d\rho(y|x)\right) = \operatorname{sgn}(\operatorname{Prob}(y = 1|x) - \operatorname{Prob}(y = -1|x)),$$

where $\rho(y|x)$ is the conditional probability of x induced by ρ, so when applied to SVMs, Theorem 4 immediately implies the following proposition.

Proposition 3. *For SVMs, if we suppose $D[|1 - yf(x)|_+] \leq \sigma^2$ and $f(x) \in [0,b]$ for all $x \in \mathcal{X}$ and $f \in B_R$, then, for any $\delta \in (0, 1/2]$, the estimation*

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq \varepsilon(m,\delta) + \left\{\frac{2\ln(1/\delta)(\sigma^2 + (1+b)^4/3)}{m}\right\}^{\frac{1}{2}} + \lambda R^2$$
$$+ C_r R^{\frac{r}{2(r-1)}} + 2\beta_m \tag{5.5}$$

holds with confidence at least $1 - 2\delta$. In addition, by the bound (5.5), we can suggest the strategy of setting the regularization parameter of SVMs as $\lambda := \lambda^* = \frac{\kappa}{R\sqrt{m}}$. In this case, if, furthermore, take $R = m^{\xi}$ with where $0 < \xi < \min\{\frac{h}{2d}, \frac{1}{2}\}$, then the

learning rate satisfies

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq O(m^{-\eta}),$$

where $\eta = \min\{\frac{h-2d\xi}{2d+2h}, \frac{1}{2} - \xi, \frac{\xi r}{2(1-r)}\}$.

Remark 8. As it is known, the regularization algorithms can be formulated either as

$$\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i) + \lambda\|f\|_K^2, \tag{5.6}$$

or as

$$\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i) \quad \text{s.t. } \|f\|_K^2 \leq R, \tag{5.7}$$

where the expression (5.6) is the so-called Lagrange form of the expression (5.7). The equivalence of these two formulations comes from the fact that for any R, there exists a regularization parameter λ such that the solution of the two problems are the same. In SVMs, Vapnik applied the SRM principle[23] to tackle the problem of how to choose the regularization parameter λ for any R. The SRM principle consists in solving the second problem for a series of values of R and then choosing the value that minimizes a generalization bound that depends on the VC-dimension of the set $\{f : \|f\|_K \leq R\}$. The VC-dimension is, however, not easy to compute and so only loose upper bounds can be found. In the present work, from another point of view, we have estimated the upper bounds of the sample error and regularization error, and then suggested the strategies for controlling the tradeoff between the sample error and regularization error. In fact, when the parameter $R \to +\infty$, the expression (5.7) becomes the unconstrained optimization problem, so the regularization parameter λ must tend to 0. Interestingly, our strategies suggested by Propositions 2 and 3 (namely $\lambda = \frac{2\kappa b}{R\sqrt{m}}$ in Proposition 2 and $\lambda = \frac{\kappa}{R\sqrt{m}}$ in Proposition 3) exactly possess such property. This shows the rationality of the suggested parameter setting strategies, and, furthermore, the significance of Propositions 2 and 3 is highlighted.

In Propositions 2 and 3, we have derived the generalization bound of SVMs and regularization networks simultaneously through the space complexity and algorithmic stability. Comparing Proposition 4.1 in Ref. 25 with Proposition 2 obtained in this paper, we can find that we all studied the learning performance of the regularization networks and obtained the leaning rate of the regularization networks, but the difference of research approaches are obvious, above all strategies of the regularization parameter setting are different. In the Propositions 4.1 of Ref. 25, in order to obtain the generalization bounds of regularization networks, authors manually took the regularization parameter $\lambda = m^{-1/(1+\beta)(1+s)}$, and then used

this proposition to explain the learning rate of the regularization networks is less than $1/2$, that is

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq O(m^{-\eta'}),$$

where $0 < \eta' \leq 1/2$. Obviously, by (5.6) and (5.7) we know our approach that choose the parameters by minimizing the obtained generalization bound is more valuable. Because it conforms to reality better than the manually taking the regularization parameter (regularization parameter is not related to the parameter R does not conform to reality).

Finally, we would like to remark that the analysis method adopted in this paper has an additional advantage that it can be conveniently used to yield practical strategies of setting the regularization parameters, just as demonstrated in Propositions 2 and 3. The previous approaches obviously have no such advantage.

6. Conclusion

The existing approaches for estimating the generalization performance of a learning algorithm are in terms of unitary measure on hypothesis complexity or algorithmic stability. It is our point of view that the performance of a learning algorithm is affected by no means with an unitary factor like hypothesis space complexity, algorithmic stability, data quality and sampling mechanism, but actually with the combinative effect of all the unitary factors. Based on this viewpoint, we have proposed to bound the generalization performance of regularization learning algorithms simultaneously in terms of the hypothesis space complexity, algorithmic stability and data quality. The obtained generalization bound estimations sharpen or generalize those derived from the traditional approaches based on using unitary measure. We have shown that the new approach has an additional advantage that it can be conveniently used to yield practical strategies of setting the regularization parameters (say choose the parameters by minimizing the obtained generalization bound). The obtained generic results have been specialized to two typical regularization algorithms: the regularization networks and support vector machines, showing the significance and usefulness of the new approach.

There are many problems that deserve further research along the line of the present work. For example, to find an efficient way of estimating the empirical covering number involved in Theorem 1, to answer what type of algorithmic stability that is essential for bounding the generalization performance of a learning algorithm, to systematically compare the performance of the regularization networks and support vector machines with the suggested regularization parameter setting strategies and the other known strategies. All those problems are under our current research.

Acknowledgments

The authors thank the referee for the useful comments towards the improvement of this paper. The research is supported by the National 973 Project (2007CB311002) and Natural Science Foundations of China (60975036, 11001227, 61070225).

References

1. N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.* **68** (1950) 337–404.
2. P. L. Bartlett, The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network, *IEEE Trans. Inform. Theory* **44** (1998) 525–536.
3. O. Bousquet, New approaches to statistical learning theory, *Ann. Inst. Statist. Math.* **55** (2003) 371–389.
4. P. L. Bartlett, Fast rates for estimation error and oracle inequalities for model selection, *Econometric Theory* **24** (2008) 545–552.
5. O. Bousquet and A. Elisseeff, Stability and generalization, *J. Mach. Learn. Res.* **2** (2002) 499–526.
6. M. Bertero, Regularization methods for linear inverse problems, in *Inverse Problems*, ed. C. G. Talenti (Springer-Verlag, Berlin, 1986), pp. 52–112.
7. D. R. Chen, Q. Wu, Y. M. Ying and D. X. Zhou, Support vector machine soft margin classifiers: Error analysis, *J. Mach. Learn. Res.* **5** (2004) 1143–1175.
8. F. Cucker and S. Smale, On the mathematical foundations of learning, *Bull. Amer. Math. Soc.* **39** (2002) 1–49.
9. F. Cucker and S. Smale, Best choices for regularization parameters in learning theory: On the bias-variance problem, *Found. Comput. Math.* **2** (2002) 413–428.
10. F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint* (Cambridge University Press, Cambridge, 2007).
11. L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition* (Springer-Verlag, New York, 1996).
12. T. Evgeniou and M. Pontil, On the V-gamma dimension for regression in reproducing kernel Hilbert spaces, in *Proc. of Algorithmic Learning Theory*, Lecture Notes in Computer Science, Vol. 1720 (Springer, 1999), pp. 106–117.
13. T. Ebgeniou, M. Pontil and T. Poggio, Regularization networks and support vector machines, *Adv. Comput. Math.* **13** (2000) 1–50.
14. F. Girosi, M. Jones and T. Poggio, Regularization theory and neural networks architectures, *Neural Comput.* **7** (1995) 219–269.
15. S. Kutin and P. Niyogi, Almost-everywhere algorithmic stability and generalization error, Technical Report TR-2002-03, Department of Computer Science, The University of Chicago (2002).
16. V. Koltchinskii and D. Panchenko, Rademacher processes and bounding the risk of function learning, *High Dimensional Probability* **47** (2000) 1902–1914.
17. H. Li, N. Chen and Y. Y. Tang, Local learning estimates by integral operators, *Int. J. Wavelets Multiresolut. Inf. Process.* **5** (2010) 695–712.
18. L. Q. Li, Regularized least square regression with spherical polynomial kernels, *Int. J. Wavelets Multiresolut. Inf. Process.* **6** (2009) 781–801.
19. A. Martinand and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations* (Cambridge University Press, Cambridge, UK, 1999).
20. S. Smale and D. X. Zhou, Shannon sampling and function reconstruction from point values, *Bull. Amer. Math. Soc.* **41** (2004) 279–305.
21. I. Steinwart, D. Hush and C. Scovel, Learning from dependent observations, *J. Multivariate Anal.* **100** (2009) 175–194.
22. A. N. Tiknonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems* (W. H. Winston, Washington, DC, 1977).
23. V. Vapnik, *Statistical Learning Theory* (John Wiley, New York, 1998).
24. Q. Wu and D. X. Zhou, SVM soft margin classifiers: Linear programming versus quadratic programming, *Neural Comput.* **17** (2005) 1160–1187.

25. Q. Wu, Y. Ying and D. X. Zhou, Learning rates of least-square regularized regression, *Found. Comput. Math.* **6** (2006) 171–192.
26. G. Wahba, Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV, in *Advances in Kernel Methods — Support Vector Learning*, eds. B. Schölkopf, C. Burges and A. Smola (MIT Press, 1999), pp. 69–88.
27. Y. L. Xu and D. R. Chen, Learning rates of regularized regression for functional data, *Int. J. Wavelets Multiresolut. Inf. Process.* **6** (2009) 839–850.
28. B. Zou, L. Q. Li and Z. B. Xu, The generalization performance of ERM algorithm with strongly mixing observations, *Mach. Learn.* **25**(2) (2009) 188–200.
29. D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, *IEEE Trans. Inform. Theory* **49** (2003) 1743–1752.