Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Learning with sample dependent hypothesis spaces

Qiang Wu[a], Ding-Xuan Zhou[b,*]

[a] Department of Statistical Science, Institute of Genome Sciences and Policy, Duke University, Durham, NC 27708, USA

[b] Department of Mathematics, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 4 July 2007
Received in revised form 27 May 2008
Accepted 10 September 2008

Keywords:
Learning theory
Regularization scheme
Error analysis
Sample dependent hypothesis spaces
Approximation error

ABSTRACT

Many learning algorithms use hypothesis spaces which are trained from samples, but little theoretical work has been devoted to the study of these algorithms. In this paper we show that mathematical analysis for these algorithms is essentially different from that for algorithms with hypothesis spaces independent of the sample or depending only on the sample size. The difficulty lies in the lack of a proper characterization of approximation error. To overcome this difficulty, we propose an idea of using a larger function class (not necessarily linear space) containing the union of all possible hypothesis spaces (varying with the sample) to measure the approximation ability of the algorithm. We show how this idea provides error analysis for two particular classes of learning algorithms in kernel methods: learning the kernel via regularization and coefficient based regularization. We demonstrate the power of this approach by its wide applicability.

1. Introduction

In learning theory, a widely used approach is regularization, or a penalized version of the classical structural or empirical risk minimization [1–3]. This approach, given an input metric space X, an output space $Y \subset \mathbb{R}$ and a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \in (X \times Y)^m$ independently and identically drawn according to an underlying distribution ρ on $Z = X \times Y$, searches over a set $\mathcal{H}$ of functions from X to Y, called the *hypothesis space*, for a function

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda \Omega(f) \right\}. \tag{1.1}$$

Here $\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{i=1}^m \ell(y, f(x))$ is the empirical risk with $\ell : \mathbb{R}^2 \to \mathbb{R}_+$ a loss function measuring the prediction error as $\ell(y, f(x))$ if $f(x)$ is used to predict the real output y, λ is a nonnegative regularization parameter, and $\Omega : \mathcal{H} \to \mathbb{R}_+$ is a penalty functional which usually satisfies $\Omega(0) = 0$ for $0 \in \mathcal{H}$. When $\lambda = 0$, (1.1) becomes the classical empirical risk minimization (ERM) scheme [1].

Many learning algorithms fall into the setting of (1.1) with specific loss function, hypothesis space and penalty functional, including regularization networks [3] and support vector machines (SVM) [4]. The choice of the loss function ℓ usually depends on learning problems, for example, the least square loss $(y - f(x))^2$ for regression and the hinge loss $(1 - yf(x))_+ = \max\{1 - yf(x), 0\}$ for classification with support vector machines.

We are interested in the generalization ability of the scheme (1.1). Our purpose is to bound the (excess) *generalization error* $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\ell^*)$, where

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

[*] Corresponding author.
E-mail addresses: qiang@stat.duke.edu (Q. Wu), mazhou@cityu.edu.hk (D.-X. Zhou).

0898-1221/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2008.09.014

is the expected risk and

$$f_\ell^* = \arg\min \mathcal{E}(f)$$

with the minimum taken over all measurable functions is the target function. This has received much attention in the literature and there have been lot of works on this topic, for cases where the loss function, hypothesis space and penalty functional are specified, see e.g. [5–10] and the references therein. Among them the most useful case is the regularization in a reproducing kernel Hilbert space (RKHS) associated with a Mercer kernel.

A Mercer kernel K on X is a continuous symmetric function $K : X \times X \to \mathbb{R}$ which is positive semi-definite in the sense that the matrix $\left(K(x_i, x_j)\right)_{i,j=1}^m$ is positive semi-definite for any set $\{x_i\}_{i=1}^m \subseteq X$. The RKHS associated with the Mercer kernel K is defined [11] as the completion of the linear span of the set of functions $\{K(x, \cdot) : x \in X\}$ with the inner product satisfying $\langle K(x, \cdot), K(x', \cdot)\rangle_K = K(x, x')$.

Example 1. Let K be a Mercer kernel on X and $\mathcal{H} = \mathcal{H}_K$, $\Omega(f) = \|f\|_K^2$. Then (1.1) becomes the regularization algorithm in RKHS

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_K} \left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2\right\}. \tag{1.2}$$

A main advantage of such a setting is that this algorithm reduces to a finite dimensional optimization problem due to the representer theorem [12,13], which holds because of the reproducing property of the RKHS.

The choice of the hypothesis space $\mathcal{H}$ and the penalty functional Ω are crucial for the performance of the algorithm (1.1).

One classical approach is to choose the hypothesis space $\mathcal{H}$ according to some a priori knowledge on the sampling distribution ρ. In this case the hypothesis space is independent of the sample. A typical example is the set of linear functions in linear regression and linear discrimination analysis.

Another approach in the literature concerns more the fit of the hypothesis space to the data. A large class of learning algorithms with this approach choose the hypothesis space and the penalty functional according to the sample size only: $\mathcal{H} = \mathcal{H}_m$ and $\Omega = \Omega_m$. That is,

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_m} \left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\Omega_m(f)\right\}. \tag{1.3}$$

A typical example is tuning the kernel parameter according to the sample size [14–16] where with a Gaussian kernel $K_\sigma(x, y) = \exp\left\{-\frac{|x-y|^2}{2\sigma^2}\right\}$ of variance $\sigma = \sigma_m$ depending on m, the hypothesis space is $\mathcal{H} = \mathcal{H}_{K_{\sigma_m}}$ and $\Omega(f) = \|f\|_{K_{\sigma_m}}^2$.

In the literature, almost all the results on the error analysis of the scheme (1.1) focus on (1.3), for algorithms with hypothesis spaces independent of the sample or only depending on the sample size, see [1,6,4,3,17,7,18,14,15,9] and the references therein. Usually the error bounds are obtained by balancing the sample (estimation) error and the approximation error, which leads to suitable choices of parameters according to the sample size m. Even when the hypothesis space depends on the sample size, the error analysis can still be done by first fixing the parameterized hypothesis space and then optimizing the regularization parameter. So throughout the paper we regard the hypothesis spaces depending only on the sample size as “sample independent”.

The main purpose of this paper is to study learning algorithms of the form (1.1) with the hypothesis space depending on the sample $\mathbf{z}$, not only on the sample size m. That is, $\mathcal{H} = \mathcal{H}_{\mathbf{z}}$ with a penalty functional also depending on the sample: $\Omega = \Omega_{\mathbf{z}}$. So the scheme (1.1) now takes the following sample dependent form

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_{\mathbf{z}}} \left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\Omega_{\mathbf{z}}(f)\right\}. \tag{1.4}$$

Learning algorithms of the type (1.4) include model selection via resampling [19], adaptive model selection [20], universal algorithms [21,22], tuning kernel parameter via sample values [23], learning the kernel [24,25] and coefficient based regularization [26,13,27] in kernel methods. Let us list two special examples here. For more details and mathematical analysis, see Sections 3–5.

Example 2. Let $X \subset \mathbb{R}^n$ and $K_\sigma(x, y) = \exp\left\{-\frac{|x-y|^2}{2\sigma^2}\right\}$ be Gaussian kernels with $0 < \sigma < \infty$. Define

$$\sigma_{\mathbf{z}} = \arg\min_{0<\sigma<\infty} \min_{f\in\mathcal{H}_{K_\sigma}} \left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_{K_\sigma}^2\right\}.$$

Taking $\mathcal{H} = \mathcal{H}_{K_{\sigma_{\mathbf{z}}}}$ and $\Omega(f) = \|f\|_{K_{\sigma_{\mathbf{z}}}}^2$ in (1.1) yields a special example of the sample dependent scheme (1.4) for learning the kernel:

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_{\sigma_{\mathbf{z}}}} \left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_{K_{\sigma_{\mathbf{z}}}}^2\right\}.$$

Example 3. Let $1 \le p < \infty$ and $K \in C(X \times X)$ (not necessarily positive semi-definite). Choose

$$\mathcal{H}_{\mathbf{z}} = \left\{ \sum_{i=1}^{m} \alpha_i K(x_i, \cdot) : (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m \right\} \quad \text{and} \quad \Omega_{\mathbf{z}} \left(\sum_{i=1}^{m} \alpha_i K(x_i, \cdot) \right) = \sum_{i=1}^{m} |\alpha_i|^p.$$

Then (1.1) becomes a special example of the sample dependent scheme (1.4) where $f_{\mathbf{z}} = \sum_{i=1}^{m} \alpha_{\mathbf{z},i} K(x_i, \cdot)$ with $\alpha_{\mathbf{z}} = (\alpha_{\mathbf{z},1}, \ldots, \alpha_{\mathbf{z},m})^T$ given by

$$\alpha_{\mathbf{z}} = \arg\min_{\alpha \in \mathbb{R}^m} \left\{ \mathcal{E}_{\mathbf{z}} \left(\sum_{i=1}^{m} \alpha_i K(x_i, \cdot) \right) + \lambda \sum_{i=1}^{m} |\alpha_i|^p \right\}.$$

In the literature, no much attention has been paid to the error analysis of the sample dependent scheme (1.4) which is of the same importance. To the best of our knowledge, only two special cases have been considered. One is for those algorithms that can be written as two-stage minimization problems with the first layer minimization associated with a sample independent hypothesis space. The error analysis can be done by making full use of this special feature of these algorithms, for example, in [20–22] for the adaptive model selection and universal estimators. The other is the linear programming SVM for which the error analysis is done in [28] by relating it to the well-known classical SVM.

An immediate question is whether methods for analyzing the scheme (1.3) still work for (1.4). Unfortunately, this is not the case. We shall point out in Section 2 that the error analysis for the scheme (1.4) with sample dependent hypothesis space is essentially different from and more difficult than that for (1.3). Then we will focus on the error analysis for the scheme (1.4). Our main contribution is to provide a general approach, by which we study two classes of algorithms in kernel methods: learning the kernel (like Example 2) and coefficient based regularization (like Example 3). The main advantage of our approach is its generality and wide applicability. Of course, for particular algorithms, better error bounds might be obtained by means of special features of specific algorithms.

2. Error decomposition in mathematical analysis

When the hypothesis space $\mathcal{H}$ is independent of the sample or depends only on the sample size, a widely used and powerful method for the error analysis of the scheme (1.1) based on ERM and regularization is the *error decomposition* method which bounds the generalization error by the sum of sample error and approximation (or regularization) error. The sample error is usually estimated by concentration inequalities [29,2,3,30] and the approximation error by rich knowledge from approximation theory [14,8,31,32,18]. This method has been well understood. Let us explain this briefly for $f_{\mathbf{z}}$ given by (1.1).

2.1. Error decomposition with sample independent hypothesis

The main idea of error decomposition is the law of large numbers by which we have $\mathcal{E}_{\mathbf{z}}(f) \to \mathcal{E}(f)$ in probability (as $m \to \infty$) for a fixed function f. So we expect that $f_{\mathbf{z}}$ defined by (1.1) is a good approximation of its sample independent or noise-free limit defined by

$$f_{\lambda,\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \left\{ \mathcal{E}(f) + \lambda \Omega(f) \right\}. \tag{2.1}$$

Write

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\ell^*) &= \{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})\} + \left\{(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}})) - \left(\mathcal{E}_{\mathbf{z}}(f_{\lambda,\mathcal{H}}) + \lambda\Omega(f_{\lambda,\mathcal{H}})\right)\right\} \\ &\quad + \left\{\mathcal{E}_{\mathbf{z}}(f_{\lambda,\mathcal{H}}) - \mathcal{E}(f_{\lambda,\mathcal{H}})\right\} + \left\{\mathcal{E}(f_{\lambda,\mathcal{H}}) - \mathcal{E}(f_\ell^*) + \lambda\Omega(f_{\lambda,\mathcal{H}})\right\} - \lambda\Omega(f_{\mathbf{z}}).\end{aligned}$$

Since $f_{\mathbf{z}}$ is a minimizer of the penalized empirical risk, the second term is ≤ 0. Also, $-\lambda\Omega(f_{\mathbf{z}}) \le 0$. So we have the following error decomposition

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\ell^*) \le \left\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \mathcal{E}_{\mathbf{z}}(f_{\lambda,\mathcal{H}}) - \mathcal{E}(f_{\lambda,\mathcal{H}})\right\} + \mathcal{D}(\lambda) \tag{2.2}$$

where

$$\mathcal{D}(\lambda) := \left\{\mathcal{E}(f_{\lambda,\mathcal{H}}) - \mathcal{E}(f_\ell^*) + \lambda\Omega(f_{\lambda,\mathcal{H}})\right\} = \inf_{f \in \mathcal{H}} \left\{\mathcal{E}(f) - \mathcal{E}(f_\ell^*) + \lambda\Omega(f)\right\}.$$

The first term in the error decomposition (2.2) is called the *sample error*. Here the quantity $\mathcal{E}_{\mathbf{z}}(f_{\lambda,\mathcal{H}}) - \mathcal{E}(f_{\lambda,\mathcal{H}})$ can be estimated by applying standard probability inequalities to the random variable $\ell(y, f_{\lambda,\mathcal{H}}(x))$ in (Z, ρ). The other quantity $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})$ involves a set of random variables $\ell(y, f(x))$ with f running over a subset of $\mathcal{H}$, which leads to the theory of uniform convergence [1] studying those function sets $\mathcal{F}$ satisfying $\sup_{f \in \mathcal{F}} |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)| \to 0$ with confidence (as $m \to \infty$). Thus bounds for the sample error for a specific hypothesis space $\mathcal{H}$ can be derived by concentration inequalities [2] and depend on the capacity of the hypothesis space [33,34].

The last term $\mathcal{D}(\lambda)$ in the error decomposition (2.2) is independent of the sample. It characterizes the approximation ability of the hypothesis space $\mathcal{H}$ (with respect to the target function f_ℓ^*) and is called the *approximation error* or regularization error [35,32].

The above error decomposition procedure is now standard in learning theory and also works for the scheme (1.3). Its key feature provided by the sample independent scheme (1.3) is that the approximation error $\mathcal{D}(\lambda)$ does not depend on the sample $\mathbf{z}$.

2.2. Error decomposition with sample dependent hypothesis

Things become completely different for the scheme (1.4) where the hypothesis space depends on the sample. At first glance, it seems that the error decomposition procedure (2.2) is still valid. However, this needs us to formally define a function $f_{\lambda,\mathcal{H}_\mathbf{z}}$ as a minimizer of the penalized expected risk over $\mathcal{H}_\mathbf{z}$ and then the approximation error as $\mathcal{E}(f_{\lambda,\mathcal{H}_\mathbf{z}}) - \mathcal{E}(f_\ell^*) + \lambda\Omega_\mathbf{z}(f_{\lambda,\mathcal{H}_\mathbf{z}})$. This quantity depends on the sample $\mathbf{z}$. So the approximation error defined in such a way involves not only the approximation ability of a fixed hypothesis space with respect to the target function but also the variance of the sample. This is the essential difficulty of the problem. As we know there is no general approach in the literature to overcome this difficulty. Error decomposition methods like (2.2) cannot be applied directly to analyze the generalization performance of algorithms with sample dependent hypothesis spaces.

In this paper we propose an error decomposition method which works for sample dependent hypothesis spaces. The key idea lies in a universal hypothesis and proper definition of the approximation error.

Definition 1. We say that a class $\mathcal{H}_0$ of functions on X is a *universal hypothesis* associated with the scheme (1.4) if $\bigcup_{m\in\mathbb{N}}\bigcup_{\mathbf{z}\in Z^m}\mathcal{H}_\mathbf{z}\subseteq\mathcal{H}_0$. The *approximation error* associated with $\mathcal{H}_0$ and a penalty functional $\Omega_0:\mathcal{H}_0\to\mathbb{R}_+$ is defined as

$$\mathcal{D}_0(\lambda) = \inf_{f\in\mathcal{H}_0}\left\{\mathcal{E}(f) - \mathcal{E}(f_\ell^*) + \lambda\Omega_0(f)\right\}, \quad \lambda > 0. \tag{2.3}$$

Note that the universal hypothesis is not necessarily a linear space. The approximation error (2.3) is independent of the sample $\mathbf{z}$. A minimizer

$$f_{\lambda,\mathcal{H}_0} = \arg\min_{f\in\mathcal{H}_0}\{\mathcal{E}(f) + \lambda\Omega_0(f)\} \tag{2.4}$$

of the approximation error will help to realize the error decomposition: write

$$\begin{aligned}\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\ell^*) = {} & \{\mathcal{E}(f_\mathbf{z}) - \mathcal{E}_\mathbf{z}(f_\mathbf{z})\} + \left\{(\mathcal{E}_\mathbf{z}(f_\mathbf{z}) + \lambda\Omega_\mathbf{z}(f_\mathbf{z})) - \left(\mathcal{E}_\mathbf{z}(f_{\lambda,\mathcal{H}_0}) + \lambda\Omega_0(f_{\lambda,\mathcal{H}_0})\right)\right\} \\ & + \left\{\mathcal{E}_\mathbf{z}(f_{\lambda,\mathcal{H}_0}) - \mathcal{E}(f_{\lambda,\mathcal{H}_0})\right\} + \left\{\mathcal{E}(f_{\lambda,\mathcal{H}_0}) - \mathcal{E}(f_\ell^*) + \lambda\Omega_0(f_{\lambda,\mathcal{H}_0})\right\} - \lambda\Omega_\mathbf{z}(f_\mathbf{z}).\end{aligned}$$

Define the *sample error* as

$$\mathcal{S}(\mathbf{z},\lambda) = \{\mathcal{E}(f_\mathbf{z}) - \mathcal{E}_\mathbf{z}(f_\mathbf{z})\} + \left\{\mathcal{E}_\mathbf{z}(f_{\lambda,\mathcal{H}_0}) - \mathcal{E}(f_{\lambda,\mathcal{H}_0})\right\} \tag{2.5}$$

which may be estimated in terms of the capacity of $\mathcal{H}_0$.

If we define further the *hypothesis error* $\mathcal{P}(\mathbf{z},\lambda)$ as

$$\mathcal{P}(\mathbf{z},\lambda) = (\mathcal{E}_\mathbf{z}(f_\mathbf{z}) + \lambda\Omega_\mathbf{z}(f_\mathbf{z})) - \left(\mathcal{E}_\mathbf{z}(f_{\lambda,\mathcal{H}_0}) + \lambda\Omega_0(f_{\lambda,\mathcal{H}_0})\right), \tag{2.6}$$

then we obtain an *error decomposition* for the algorithm (1.4) with the sample dependent hypothesis space as

$$\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\ell^*) \le \mathcal{S}(\mathbf{z},\lambda) + \mathcal{P}(\mathbf{z},\lambda) + \mathcal{D}_0(\lambda). \tag{2.7}$$

This procedure may be regarded as a generalization of the error decomposition technique in (2.2): if the hypothesis space is sample independent, one can take $\mathcal{H}_0 = \mathcal{H}$ and $\Omega_0 = \Omega$. Then the hypothesis error is at most zero and (2.7) reduces to (2.2).

2.3. Choosing the universal hypothesis

For the sample dependent case, in general the hypothesis error $\mathcal{P}(\mathbf{z},\lambda)$ does not satisfy $\mathcal{P}(\mathbf{z},\lambda) \le 0$: the functional Ω_0 is different from the functional $\Omega_\mathbf{z}$ in the definition of $f_\mathbf{z}$. Hence the sum of the sample error and the approximation error need not bound the generalization error $\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\ell^*)$. This forms the essential difference and difficulty.

The error decomposition (2.7) still yields satisfactory error analysis if we can choose the universal hypothesis $\mathcal{H}_0$ properly. A proper choice of $(\mathcal{H}_0, \Omega_0)$ is important not only for estimating the hypothesis error $\mathcal{P}(\mathbf{z},\lambda)$ (mainly caused by the difference between the penalty functionals Ω_0 and $\Omega_\mathbf{z}$), but also for bounding the sample error $\mathcal{S}(\mathbf{z},\lambda)$ (depending mainly on the capacity of $\mathcal{H}_0$) as well as the approximation error $\mathcal{D}_0(\lambda)$ (measuring the approximation ability of $(\mathcal{H}_0, \Omega_0)$). Fortunately this is possible for many algorithms. Here we propose two approaches for choosing $(\mathcal{H}_0, \Omega_0)$.

The first approach to choose the universal hypothesis $\mathcal{H}_0$ works for learning algorithms which can be formulated as two-stage optimization problems (see Example 2). Let $\{\mathcal{H}_\sigma : \sigma\in\Sigma\}$ be a set of function spaces together with penalty functionals $\{\Omega_\sigma\}$. Take

$$\sigma_\mathbf{z} = \arg\min_{\sigma\in\Sigma}\min_{f\in\mathcal{H}_\sigma}\{\mathcal{E}_\mathbf{z}(f) + \lambda\Omega_\sigma(f)\}$$

and define $f_{\mathbf{z}}$ by (1.4) with $\mathcal{H}_{\mathbf{z}} = \mathcal{H}_{\sigma_{\mathbf{z}}}$ and $\Omega_{\mathbf{z}} = \Omega_{\sigma_{\mathbf{z}}}$. For such learning algorithms, we can take $\mathcal{H}_0 = \cup_{\sigma\in\Sigma}\mathcal{H}_\sigma$ and $\Omega_0(f) = \inf\{\Omega_\sigma(f) : f \in \mathcal{H}_\sigma, \sigma \in \Sigma\}$. Then the error decomposition (2.7) can be applied. This approach will be demonstrated in Section 3 for the algorithm of learning the kernel via regularization.

Our second approach works when all the sample dependent hypothesis spaces $\mathcal{H}_{\mathbf{z}}$ are subspaces of a Banach space or even a Hilbert space (see Example 3). We naturally choose this Banach space as the universal hypothesis $\mathcal{H}_0$. If we can find a penalty functional Ω_0 on $\mathcal{H}_0$ such that the quantity

$$(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z}})) - \left(\mathcal{E}_{\mathbf{z}}(f_\mu) + \mu\Omega_0(f_\mu)\right) \tag{2.8}$$

can be bounded efficiently where

$$f_\mu := \arg\min_{f\in\mathcal{H}_0}\{\mathcal{E}(f) + \mu\Omega_0(f)\}$$

for some $\mu = \mu(\lambda) > 0$, then the error decomposition (2.7) can be used to provide satisfactory error bounds. Here to bound (2.8) the penalty functional Ω_0 on $\mathcal{H}_0$ should be closely related to the penalty functional $\Omega_{\mathbf{z}}$ on $\mathcal{H}_{\mathbf{z}}$. This approach will be demonstrated for the algorithm of coefficient based regularization, with a positive semi-definite kernel in Section 4 and a general kernel in Section 5.

3. Learning the kernel via regularization

If the Mercer kernel K is fixed, the algorithm (1.2) is well understood, see e.g. [7,8,14,35,31] and the references therein. A crucial problem for this algorithm is the choice of the kernel. It essentially determines the performance of the algorithm, as has been proved both practically and theoretically. This has recently motivated the research topic of learning the kernel, see e.g. [24,23,36,37,25]. These learning algorithms construct a kernel $K_{\mathbf{z}}$ which depends on the sample $\mathbf{z}$. With such a kernel, (1.2) becomes an algorithm with sample dependent hypothesis space and the classical analysis for regularization schemes with a fixed kernel does not work. Here we show how our idea can be applied to this setting.

Let $\mathcal{K}$ be a set of Mercer kernels on X. For every $K \in \mathcal{K}$, let

$$\mathcal{Q}_\lambda(K) = \min_{f\in\mathcal{H}_K}\left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2\right\}.$$

The approach of learning the kernel via regularization finds a kernel $K_{\mathbf{z}} \in \mathcal{K}$ by

$$K_{\mathbf{z}} = \arg\min_{K\in\mathcal{K}}\mathcal{Q}_\lambda(K). \tag{3.1}$$

Note that Example 2 is a special case where $\mathcal{K}$ is a set of Gaussians. Under very mild conditions, this approach is shown to be solvable [25,31] and prevent overfitting [38].

Consider the function $f_{\mathbf{z}}$ defined by (1.2) with the kernel being $K_{\mathbf{z}}$ learned by (3.1). The hypothesis space here is the RKHS $\mathcal{H}_{K_{\mathbf{z}}}$ and the penalty functional is $\Omega_{\mathbf{z}}(f) = \lambda\|f\|_{K_{\mathbf{z}}}^2$. We use our idea of error decomposition in this setting. Define $\mathcal{H}_0 = \bigcup_{K\in\mathcal{K}}\mathcal{H}_K$. It is easy to check that $\bigcup_{\mathbf{z}}\mathcal{H}_{K_{\mathbf{z}}} \subseteq \mathcal{H}_0$ for every $K_{\mathbf{z}} \in \mathcal{K}$. Define, for $f \in \mathcal{H}_0$, $\Omega_0(f) = \inf\left\{\|f\|_K^2, f \in \mathcal{H}_K, K \in \mathcal{K}\right\}$ and let

$$f_{\lambda,\mathcal{H}_0} = \arg\min_{f\in\mathcal{H}_0}\{\mathcal{E}(f) + \lambda\Omega_0(f)\}.$$

Proposition 2. *Let $f_{\mathbf{z}}$ be the solution of (1.2) with the kernel $K_{\mathbf{z}}$ given by (3.1). With $\mathcal{H}_0$, Ω_0 and $f_{\lambda,\mathcal{H}_0}$ defined as above, there holds*

$$\mathcal{P}(\mathbf{z},\lambda) = \left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\|f_{\mathbf{z}}\|_{K_{\mathbf{z}}}^2\right) - \left(\mathcal{E}_{\mathbf{z}}(f_{\lambda,\mathcal{H}_0}) + \lambda\Omega_0(f_{\lambda,\mathcal{H}_0})\right) \le 0.$$

Proof. Since $f_{\lambda,\mathcal{H}_0} \in \mathcal{H}_0$, there exists a subset $\mathcal{K}_\lambda$ of $\mathcal{K}$ such that $f_{\lambda,\mathcal{H}_0} \in \mathcal{H}_K$ for every $K \in \mathcal{K}_\lambda$ and $\Omega_0(f_{\lambda,\mathcal{H}_0}) = \inf_{K\in\mathcal{K}_\lambda}\|f_{\lambda,\mathcal{H}_0}\|_K^2$. By the definition of $K_{\mathbf{z}}$ and $f_{\mathbf{z}}$, we have

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\|f_{\mathbf{z}}\|_{K_{\mathbf{z}}}^2 &= \mathcal{Q}_\lambda(K_{\mathbf{z}}) = \min_{K\in\mathcal{K}}\mathcal{Q}_\lambda(K) \le \min_{K\in\mathcal{K}_\lambda}\mathcal{Q}_\lambda(K)\\ &\le \min_{K\in\mathcal{K}_\lambda}\left\{\mathcal{E}_{\mathbf{z}}(f_{\lambda,\mathcal{H}_0}) + \lambda\|f_{\lambda,\mathcal{H}_0}\|_K^2\right\}\\ &= \mathcal{E}_{\mathbf{z}}(f_{\lambda,\mathcal{H}_0}) + \lambda\Omega_0(f_{\lambda,\mathcal{H}_0}).\end{aligned}$$

This proves the conclusion. ■

By Proposition 2, the hypothesis error $\mathcal{P}(\mathbf{z},\lambda)$ vanishes and the error decomposition (2.7) reduces to (2.2). Then one can estimate the error bounds and find the learning rates, as done in [38,39].

The learning algorithm (1.2) with $K_{\mathbf{z}}$ given by (3.1) is also known as learning in the multi-kernel spaces $\mathcal{H}_0$ in [31,38], since it can be formulated as

$$f_{\mathbf{z}} = \arg\min_{K\in\mathcal{K}}\min_{f\in\mathcal{H}_K}\left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2\right\}.$$

In this formulation, it is a regularization scheme with sample independent hypothesis space and the conclusion in Proposition 2 is an easy consequence of the definition of $f_{\mathbf{z}}$ and $f_{\lambda,\mathcal{H}_0}$. Here we give a different point of view. Note that not all algorithms for learning the kernel can be written as double layer minimization problems. Though our idea presented here does not provide new results, we hope it may shed light on the study of other approaches for learning the kernel.

4. Coefficient based regularization with positive semi-definite kernels

The research on coefficient based regularization dates back to the study of ridge regression in the 1970's [26]. It is usually used in statistics and learning theory when one needs to fit the data in certain trend, for instance, a linear function in ridge regression and a linear combination of simple classifiers in boosting. It searches for a function over the linear span of a set of base functions. To be precise, denote by I an index set and $\{h_i\}_{i\in I}$ a set of functions on X. Then the hypothesis space is $\mathcal{H} = \operatorname{span}\{h_i\}_{i\in I}$ and the penalty functional is

$$\Omega(f) = \sum_{i\in I} S(\alpha_i) \quad \text{for } f = \sum_{i\in I} \alpha_i h_i \in \mathcal{H},$$

where $S : \mathbb{R} \to \mathbb{R}_+$ is even and nondecreasing on $[0, +\infty)$. Typical choices are $S(t) = |t|^p$, $1 \le p \le +\infty$, in the ℓ^p regularization. This method has attracted much attention recently because it leads to sparse solutions and may be useful in signal processing and feature subset selection [13,27].

Whether the hypothesis space of the coefficient based regularization scheme is sample dependent or sample independent is determined by the choice of base functions h_i. In kernel methods, the base functions have the form $h_i(x) = K(x_i', x)$ with $K : X \times X \to \mathbb{R}$ a given kernel. When x_i' coincides with the sample pattern x_i, it leads to the sample dependent hypothesis space

$$\mathcal{H}_{K,\mathbf{z}} = \left\{ f_{\boldsymbol{\alpha}} : f_{\boldsymbol{\alpha}}(x) = \sum_{i=1}^{m} \alpha_i K(x_i, x), \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m \right\}. \tag{4.1}$$

The coefficient based regularization algorithm with the hypothesis space $\mathcal{H}_{K,\mathbf{z}}$ is

$$f_{\mathbf{z}} = f_{\boldsymbol{\alpha}_{\mathbf{z}}} = \arg\min_{f_{\boldsymbol{\alpha}} \in \mathcal{H}_{K,\mathbf{z}}} \{\mathcal{E}_{\mathbf{z}}(f_\alpha) + \lambda \Omega(f_{\boldsymbol{\alpha}})\}. \tag{4.2}$$

We see that Example 3 belongs to this model.

Another advantage of this setting is that the kernel K is not necessarily positive semi-definite. This may be useful in some cases and will be studied in detail in the next section.

In this section we consider the case when the kernel is positive semi-definite. We will show that (4.2) is closely related to (1.2) in this case. Firstly, the reproducing kernel Hilbert space is well studied and hence the approximation error can be well estimated. Secondly, the hypothesis error can be well estimated by bounding the quantity (2.8). Lastly, rich knowledge about the algorithm (1.2) enhances our understanding of (4.2).

Now assume that the kernel K involved in the algorithm (4.2) is positive semi-definite. Choose $\mathcal{H}_0 = \mathcal{H}_K$ and $\Omega_0(f) = \eta\|f\|_K^2$ for $f \in \mathcal{H}_K$ with $\eta > 0$ a parameter to be determined later. Obviously $\bigcup_{m\in\mathbb{N}} \bigcup_{\mathbf{z}\in Z^m} \mathcal{H}_{K,\mathbf{z}} \subseteq \mathcal{H}_0$, so $\mathcal{H}_0$ is a universal hypothesis.

To estimate the hypothesis error $\mathcal{P}(\mathbf{z}, \lambda)$ and hence activate the error decomposition (2.7), we relate (4.2) to (1.2).

Let $\mu = \eta\lambda$. To avoid confusion, here we denote by $f_{\mathbf{z},\mu}^+$ the solution to (1.2) with regularization parameter μ, i.e.,

$$f_{\mathbf{z},\mu}^+ = \arg\min_{f\in\mathcal{H}_K} \left\{\mathcal{E}_{\mathbf{z}}(f) + \mu\|f\|_K^2\right\}. \tag{4.3}$$

It will play the role of a stepping stone between $f_{\mathbf{z}}$ and $f_{\lambda,\mathcal{H}_0} = f_\mu$ defined by

$$f_\mu = \arg\min_{f\in\mathcal{H}_K} \left\{\mathcal{E}(f) + \mu\|f\|_K^2\right\}.$$

The representer theorem (see e.g. [12,13]) asserts that $f_{\mathbf{z},\mu}^+ = f_{\boldsymbol{\alpha}_\mu} \in \mathcal{H}_{K,\mathbf{z}}$ for some $\boldsymbol{\alpha}_\mu = (\alpha_{\mu,1}, \ldots, \alpha_{\mu,m}) \in \mathbb{R}^m$. Hence a comparison between $f_{\mathbf{z}}$ and $f_{\mathbf{z},\mu}^+$ is possible. On the other hand, f_μ is a data-free limit of $f_{\mathbf{z},\mu}^+$. The following result bounds the hypothesis error with satisfactory rates.

Proposition 3. *Let $\mathbf{z} \in Z^m$. Assume $\ell(y, 0) \le \widetilde{M}$ almost surely for some $\widetilde{M} > 0$. If there exist two nonnegative constants $C_1 = C_1(\mu, m)$ and $C_2 = C_2(\mu, m)$ such that*

$$\Omega(f_{\mathbf{z},\mu}^+) = \sum_{i=1}^{m} S(\alpha_{\mu,i}) \le C_1 \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mu}^+) + C_2 \|f_{\mathbf{z},\mu}^+\|_K^2 \tag{4.4}$$

and η is chosen to satisfy $\eta \ge \frac{C_2}{1+\lambda C_1}$ (that is, $\mu \ge \frac{\lambda C_2}{1+\lambda C_1}$), then

$$(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}})) - \left(\mathcal{E}_{\mathbf{z}}(f_\mu) + \mu\|f_\mu\|_K^2\right) \le \lambda C_1 \widetilde{M}.$$

Proof. By the fact $f^+_{\mathbf{z},\mu} \in \mathcal{H}_{K,\mathbf{z}}$, we know that

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}}) \le \mathcal{E}_{\mathbf{z}}(f^+_{\mathbf{z},\mu}) + \lambda\Omega(f^+_{\mathbf{z},\mu}).$$

The relation (4.4) and the assumption on η further bound the right-hand side by

$$\begin{aligned}(1+\lambda C_1)\mathcal{E}_{\mathbf{z}}(f^+_{\mathbf{z},\mu}) + \lambda C_2\|f^+_{\mathbf{z},\mu}\|_K^2 &= (1+\lambda C_1)\left(\mathcal{E}_{\mathbf{z}}(f^+_{\mathbf{z},\mu}) + \frac{\lambda C_2}{1+\lambda C_1}\|f^+_{\mathbf{z},\mu}\|_K^2\right)\\ &\le (1+\lambda C_1)\left(\mathcal{E}_{\mathbf{z}}(f^+_{\mathbf{z},\mu}) + \mu\|f^+_{\mathbf{z},\mu}\|_K^2\right).\end{aligned}$$

But

$$\mathcal{E}_{\mathbf{z}}(f^+_{\mathbf{z},\mu}) + \mu\|f^+_{\mathbf{z},\mu}\|_K^2 \le \mathcal{E}_{\mathbf{z}}(0) + \mu\cdot 0 \le \widetilde{M}$$

and

$$\mathcal{E}_{\mathbf{z}}(f^+_{\mathbf{z},\mu}) + \mu\|f^+_{\mathbf{z},\mu}\|_K^2 \le \mathcal{E}_{\mathbf{z}}(f_\mu) + \mu\|f_\mu\|_K^2.$$

Our conclusion follows. ■

From Proposition 3, $\mathcal{P}(\mathbf{z},\lambda) \le \lambda C_1\widetilde{M}$ when (4.4) holds. Thus the key for an efficient error decomposition of scheme (4.2) is an inequality of form (4.4) for the scheme (1.2). Moreover, for the bound $\lambda C_1\widetilde{M}$ of the hypothesis error to be effective, we would require $\lambda C_1 \to 0$ as $m \to \infty$ with proper choice of λ to guarantee the consistency. Fortunately, this is true in most cases. Next we illustrate this for two particular learning algorithms: the kernel regression and linear programming SVM.

4.1. Kernel regression

Ridge regression tries to fit the data by a linear model. If we fit the data by linear combinations of kernel functions evaluated at the sampling points, this is just the scheme (4.2) with least square loss and penalty functional $\Omega(f_{\boldsymbol{\alpha}}) = \sum_{i=1}^m \alpha_i^2$:

$$f_{\mathbf{z}} = f_{\boldsymbol{\alpha}_{\mathbf{z}}} \quad \text{where } \boldsymbol{\alpha}_{\mathbf{z}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^m}\left\{\frac{1}{m}\sum_{i=1}^m (f_{\boldsymbol{\alpha}}(x_i) - y_i)^2 + \lambda\|\boldsymbol{\alpha}\|_{\ell^2}^2\right\}. \tag{4.5}$$

By computing the partial derivatives, we see that the solution of (4.5) is given by $f_{\mathbf{z}} = f_{\boldsymbol{\alpha}_{\mathbf{z}}}$ with $\boldsymbol{\alpha}_{\mathbf{z}}$ satisfying the linear system

$$(\lambda m I_m + (K[\mathbf{x}])^2)\boldsymbol{\alpha} = K[\mathbf{x}]\mathbf{y},$$

where $K[\mathbf{x}] = (K(x_i,x_j))_{i,j=1}^m$ and $\mathbf{y} = (y_1,\ldots,y_m)^T$.

For this algorithm, we have the following conclusion.

Theorem 4. *With the least square loss, the solution $f^+_{\mathbf{z},\mu} = f_{\boldsymbol{\alpha}_\mu}$ to (4.3) satisfies*

$$\sum_{i=1}^m \alpha_{\mu,i}^2 = \frac{1}{m\mu^2}\mathcal{E}_{\mathbf{z}}(f^+_{\mathbf{z},\mu}).$$

Hence with the choice $\eta = 1$ (so that $\mu = \lambda$) we have

$$\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\sum_{i=1}^m \alpha_{\mathbf{z},i}^2\right) - \left(\mathcal{E}_{\mathbf{z}}(f_\mu) + \mu\|f_\mu\|_K^2\right) \le \frac{M^2}{m\lambda}$$

if $|y| \le M$ almost surely.

Proof. Recall that the coefficient $\boldsymbol{\alpha}_\mu$ for $f^+_{\mathbf{z},\mu}$ satisfies

$$(\mu m I_m + K[\mathbf{x}])\boldsymbol{\alpha}_\mu = \mathbf{y}.$$

This gives $\mu m\boldsymbol{\alpha}_\mu = \mathbf{y} - K[\mathbf{x}]\boldsymbol{\alpha}_\mu$. But $f^+_{\mathbf{z},\mu}(x_i) = \sum_{j=1}^m \alpha_{\mu,j}K(x_j,x_i)$ is just the ith component of the vector $K[\mathbf{x}]\boldsymbol{\alpha}_\mu$. We have

$$\mu^2 m^2\sum_{i=1}^m \alpha_{\mu,i}^2 = \sum_{i=1}^m (y_i - f^+_{\mathbf{z},\mu}(x_i))^2 = m\mathcal{E}_{\mathbf{z}}(f^+_{\mathbf{z},\mu}).$$

This proves the first claim. It shows that (4.4) holds with $C_1 = \frac{1}{m\mu^2}$ and $C_2 = 0$. The choice $\eta = 1$ implies $\mu = \lambda > 0 = \frac{\lambda C_2}{1+\lambda C_1}$. Hence our conclusion follows from Proposition 3 and the bound $\ell(y,0) = y^2 \le M^2$. ■

By Theorem 4, in the error decomposition (2.7) of scheme (4.5), the hypothesis error is well bounded: $\mathcal{P}(\mathbf{z},\lambda) \le \frac{M^2}{m\lambda}$. It decays of order $O(\frac{1}{m\lambda})$. This is rather fast. We will not go into the details of estimating error bounds and learning rates because they are standard in learning theory and out of the scope of this paper. But we refer the reader to [9] for some techniques.

4.2. Linear programming SVM classification

Support vector machine classification algorithms use the hinge loss $\ell(y, f(x)) = (1 - yf(x))_+$ with $Y = \{\pm 1\}$ containing only two labels. The classical kernel SVM searches for a classifier $\mathrm{sgn}(f_{\mathbf{z}})$ by taking signs of a real-valued function $f_{\mathbf{z}} \in \mathcal{H}_K$ obtained by a regularization scheme of form (1.2). It is implemented by convex quadratic programming optimization [1]. The linear programming SVM was motivated by the idea of having a solution with sparser representation [1]. It searches for a classifier $\mathrm{sgn}(f_{\mathbf{z}})$ generated by a real-valued function $f_{\mathbf{z}}$ produced in $\mathcal{H}_{K,\mathbf{z}}$ by the algorithm

$$f_{\mathbf{z}} = f_{\boldsymbol{\alpha}_{\mathbf{z}}} \quad \text{where } \boldsymbol{\alpha}_{\mathbf{z}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^m} \left\{ \frac{1}{m}\sum_{i=1}^m (1 - y_i f_{\boldsymbol{\alpha}}(x_i))_+ + \lambda\|\boldsymbol{\alpha}\|_{\ell^1} \right\}.$$

It can be implemented by convex linear programming optimization and have the ability of handling huge data. For this algorithm, we have the following result.

Theorem 5. *With the hinge loss, the solution $f^+_{\mathbf{z},\mu} = f_{\alpha_\mu}$ to (4.3) satisfies*

$$\sum_{i=1}^m |\alpha_{\mu,i}| = \frac{1}{2\mu}\mathcal{E}_{\mathbf{z}}(f^+_{\mathbf{z},\mu}) + \|f^+_{\mathbf{z},\mu}\|_K^2.$$

Hence with a choice $\eta \geq 1/2$ we have

$$\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\sum_{i=1}^m |\alpha_{\mathbf{z},i}|\right) - \left(\mathcal{E}_{\mathbf{z}}(f_\mu) + \mu\|f_\mu\|_K^2\right) \leq \frac{1}{2\eta}.$$

The first part follows from the KKT conditions of the optimization problem for (4.3) while the second follows from Proposition 3. This result holds true even when an offset term is involved (see [28]). In order to show the consistency or find learning rates, one needs to choose $\eta = \eta(m, \lambda) \to \infty$ as $m \to \infty$ and at the same time $\mu = \eta\lambda \to 0$ as $\lambda \to 0$. This, in general, will lead to learning rates slightly worse than that of the classical quadratic programming SVM. But when the kernel space $\mathcal{H}_K$ has low capacity, they may have the same rates. For details we refer to [28].

4.3. Sharper bound for noise-free distributions

We have given a general clue to activate the error decomposition (2.7) for the coefficient based regularization scheme (4.2) by bounding the hypothesis error in Proposition 3. This bound is independent of the underlying distribution from which the sample is drawn. It might be improved if the distribution is noise free in the sense that $\mathcal{E}(f^*_\ell) = 0$. This noise-free condition means to reconstruct a function from exact data in regression, or the underlying distribution is deterministic in classification problems.

Proposition 6. *Under the assumptions of Proposition 3, if in addition $\mathcal{E}(f^*_\ell) = 0$, then there holds*

$$(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}})) - \left(\mathcal{E}_{\mathbf{z}}(f_\mu) + \mu\|f_\mu\|_K^2\right) \leq \lambda C_1\left\{\left(\mathcal{E}_{\mathbf{z}}(f_\mu) - \mathcal{E}(f_\mu)\right) + \mathcal{D}(\mu)\right\} \tag{4.6}$$

where

$$\mathcal{D}(\mu) = \inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f) + \mu\|f\|_K^2\right\}$$

is just the approximation error.

Proof. By the proof of Proposition 3 we know that

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}}) \leq (1 + \lambda C_1)\left(\mathcal{E}_{\mathbf{z}}(f^+_{\mathbf{z},\mu}) + \mu\|f^+_{\mathbf{z},\mu}\|_K^2\right) \leq (1 + \lambda C_1)\left(\mathcal{E}_{\mathbf{z}}(f_\mu) + \mu\|f_\mu\|_K^2\right).$$

The conclusion follows by writing

$$\mathcal{E}_{\mathbf{z}}(f_\mu) + \mu\|f_\mu\|_K^2 = \left(\mathcal{E}_{\mathbf{z}}(f_\mu) - \mathcal{E}(f_\mu)\right) + \left(\mathcal{E}(f_\mu) + \mu\|f_\mu\|_K^2\right)$$

and using the fact that the second term on the right-hand side equals $\mathcal{D}(\mu)$ by the definition of f_μ. Since $\mathcal{E}(f^*_\ell) = 0$, it is just the approximation error. ■

In many situations, there holds $\mathcal{D}(\mu) \to 0$ as $\mu \to 0$ (otherwise, the algorithms would not lead to a good approximation of the target function f^*_ℓ). Then the term in the brace on the right-hand side of (4.6) decays to 0 as $m \to \infty$ and $\mu \to 0$. Hence the bound in (4.6) is sharper than that given in Proposition 3. By Proposition 6, to guarantee the consistency for the noise-free setting it suffices to choose the parameter η so that λC_1 is bounded. This also leads to faster learning rates. For an example, see [28].

We remark that in order to establish results in this section we provided some relations to the classical regularization schemes in the corresponding RKHS, which is of independent interest.

5. Coefficient based regularization with general kernels

A main advantage of coefficient based regularization schemes lies in that one may use very general kernels instead of positive semi-definite ones. This may be useful when a priori knowledge is available or one hopes to fit the data for some trends.

When a general kernel is used, a Hilbert space approach may be insufficient for characterizing the approximation error. Instead, we will use a Banach space depending on the kernel K.

Assume that $K : X \times X \to \mathbb{R}$ is continuous. We consider the Banach space $\mathcal{F}_K$ of all functions of the form

$$f(x) = \sum_{i=1}^{\infty} \alpha_i K(x_i', x), \quad x_i' \in X$$

with the norm

$$\|f\| = \inf \left\{ \sum_{i=1}^{\infty} |\alpha_i| : f(x) = \sum_{i=1}^{\infty} \alpha_i K(x_i', x), x_i' \in X \right\}.$$

It is easy to see that $\mathcal{F}_K$ can be embedded into $L^\infty(X)$ and

$$\|f\|_\infty \le \tilde{\kappa} \|f\| \quad \forall f \in \mathcal{F}_K$$

with $\tilde{\kappa} = \|K\|_\infty$. Moreover, for every $\mathbf{z} \in Z^m$, there holds $\mathcal{H}_{K,\mathbf{z}} \subseteq \mathcal{F}_K$. Hence $\mathcal{F}_K$ may play the role of $\mathcal{H}_0$ for the error decomposition.

Note that one cannot use $\mathcal{F}_K$ as the hypothesis space and $\|f\|$ the regularizer in the regularization scheme. It may cause serious computational difficulty since no representer theorem guarantees the solution to have a simple form. But $\mathcal{F}_K$ may be used as a universal hypothesis to characterize the approximation error and hence be useful for the mathematical analysis. To realize the error decomposition via $\mathcal{F}_K$, define $\mathcal{H}_0 = \mathcal{F}_K$, $\Omega_0(f) = \|f\|$ and

$$f_\lambda = \arg \inf_{f \in \mathcal{F}_K} \{\mathcal{E}(f) + \lambda \|f\|\}.$$

The existence of f_λ is not known. But this is not essential in our analysis because a sequence of approximating functions plays the same role.

Now we can proceed the error decomposition procedure. To bound the hypothesis error, we need some basic assumptions on the algorithm and some elementary concepts.

Let us illustrate the idea by studying the following algorithm:

$$f_{\mathbf{z}} = \arg \min_{f_{\boldsymbol{\alpha}} \in \mathcal{H}_{K,\mathbf{z}}} \left\{ \mathcal{E}_{\mathbf{z}}(f_{\boldsymbol{\alpha}}) + \lambda \sum_{i=1}^{m} |\alpha_i| \right\}. \tag{5.1}$$

Firstly, we assume that the loss function $\ell(y, f(x))$ is locally Lipschitz in the sense that for every $B > 0$ there exists $L(B)$ such that

$$|\ell(y, f(x)) - \ell(y, g(x))| \le L(B)|f(x) - g(x)|$$

for every $x \in X$ and any functions f, g with $\|f\|_\infty, \|g\|_\infty \le B$. Obviously $L(B)$ is increasing. Since we do not need the bound to be sharp, we assume below that $L(B)$ is continuous from the right.

We need the following notation concerning the *uniform continuity* of the kernel K on the metric space (X, d):

$$\omega_K(\delta) = \sup_{t \in X} \sup_{d(x,x') \le \delta} |K(x, t) - K(x', t)|.$$

It is easy to see that $\lim_{\delta \to 0} \omega_K(\delta) = 0$ due to the compactness of X.

Definition 7. A point set $\{x_1, \ldots, x_m\} \subseteq X$ is said to be *Δ-dense* if for every $x \in X$ there exists some $1 \le i \le m$ such that $d(x, x_i) \le \Delta$.

If $\Delta \ge \sup_{x \in X} \min_{1 \le i \le m} d(x, x_i)$, then $\{x_1, \ldots, x_m\}$ is Δ-dense.

The following bound on f_λ is an easy consequence of the fact

$$\lambda \|f_\lambda\| \le \mathcal{E}(f_\lambda) + \lambda \|f_\lambda\| \le \mathcal{E}(0) + \lambda \cdot 0 \le \widetilde{M}.$$

Lemma 8. *If $\ell(y, 0) \le \widetilde{M}$ almost surely, there holds*

$$\|f_\lambda\|_\infty \le \tilde{\kappa} \|f_\lambda\| \le \tilde{\kappa} \frac{\widetilde{M}}{\lambda}.$$

Now we can bound the quantity (2.8) and hence the hypothesis error in the error decomposition (2.7).

Theorem 9. *If $\{x_1, \ldots, x_m\}$ is Δ-dense in X, then the solution $f_{\mathbf{z}}$ to (5.1) with $\Omega(f_{\boldsymbol{\alpha}}) = \|\boldsymbol{\alpha}\|_{\ell^1}$ satisfies*

$$(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}})) - (\mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\|) \le L\left(\frac{\tilde{\kappa}\widetilde{M}}{\lambda}\right)\frac{\widetilde{M}}{\lambda}\omega_K(\Delta).$$

Proof. For any $0 < \tau < 1$, there exists an expression

$$f_\lambda(x) = \sum_{j=1}^{\infty} \beta_j K(t_j, x)$$

such that $\{t_j\} \subset X$ and

$$\|f_\lambda\| \le \sum_{j=1}^{\infty} |\beta_j| \le \|f_\lambda\| + \tau \le \frac{\widetilde{M}}{\lambda} + \tau$$

the last by Lemma 8. Choose $N_0 \in \mathbb{Z}_+$ such that $\sum_{j=N_0+1}^{\infty} |\beta_j| \le \tau$. Then

$$\left\| f_\lambda - \sum_{j=1}^{N_0} \beta_j K(t_j, \cdot)\right\|_\infty \le \tilde{\kappa}\left\| \sum_{j=N_0+1}^{\infty} \beta_j K(t_j, \cdot)\right\| \le \tilde{\kappa}\tau.$$

Since $\{x_1, \ldots, x_m\}$ is Δ-dense in X, for every t_j, there exists some $x(t_j) \in \{x_1, \ldots, x_m\}$ such that $d(x(t_j), t_j) \le \Delta$. So we have

$$\left\| \sum_{j=1}^{N_0} \beta_j K(x(t_j), \cdot) - \sum_{j=1}^{N_0} \beta_j K(t_j, \cdot)\right\|_\infty \le \sum_{j=1}^{N_0} |\beta_j|\omega_K(\Delta) \le \left(\frac{\widetilde{M}}{\lambda} + \tau\right)\omega_K(\Delta).$$

It follows that

$$\left\| \sum_{j=1}^{N_0} \beta_j K(x(t_j), \cdot) - f_\lambda\right\|_\infty \le \left(\frac{\widetilde{M}}{\lambda} + \tau\right)\omega_K(\Delta) + \tilde{\kappa}\tau.$$

It is easy to see that both $\sum_{j=1}^{N_0} \beta_j K(x(t_j), \cdot)$ and f_λ are bounded in $L^\infty(X)$ by $\tilde{\kappa}\sum_{j=1}^{\infty} |\beta_j| \le \tilde{\kappa}(\frac{\widetilde{M}}{\lambda} + \tau)$. Since $\ell(y, f(x))$ is locally Lipschitz, we have

$$\left|\mathcal{E}_{\mathbf{z}}\left(\sum_{j=1}^{N_0} \beta_j K(x(t_j), \cdot)\right) - \mathcal{E}_{\mathbf{z}}(f_\lambda)\right| \le L\left(\tilde{\kappa}\left(\frac{\widetilde{M}}{\lambda} + \tau\right)\right)\left(\left(\frac{\widetilde{M}}{\lambda} + \tau\right)\omega_K(\Delta) + \tilde{\kappa}\tau\right).$$

Notice the fact $\sum_{j=1}^{N_0} \beta_j K(x(t_j), \cdot) \in \mathcal{H}_{K,\mathbf{z}}$. There holds

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\sum_{i=1}^{m} |\alpha_{i,\mathbf{z}}| &\le \mathcal{E}_{\mathbf{z}}\left(\sum_{j=1}^{N_0} \beta_j K(x(t_j), \cdot)\right) + \lambda\sum_{j=1}^{N_0} |\beta_j| \\ &\le \mathcal{E}_{\mathbf{z}}(f_\lambda) + L\left(\tilde{\kappa}\left(\frac{\widetilde{M}}{\lambda} + \tau\right)\right)\left(\left(\frac{\widetilde{M}}{\lambda} + \tau\right)\omega_K(\Delta) + \tilde{\kappa}\tau\right) + \lambda\left(\|f_\lambda\| + \tau\right).\end{aligned}$$

Let $\tau \to 0$. We obtain

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\sum_{i=1}^{m} |\alpha_{i,\mathbf{z}}| \le \mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\| + L(\tilde{\kappa}\frac{\widetilde{M}}{\lambda})\frac{\widetilde{M}}{\lambda}\omega_K(\Delta).$$

This completes the proof. ■

Remark 1. One may also consider the algorithms with regularizer $\Omega(f_{\boldsymbol{\alpha}}) = \sum_{i=1}^{m} |\alpha_i|^p$ for $p > 1$. In that case, one need to use $\Omega_0(f) = \eta\|f\|$ with a suitable choice of the parameter η.

Remark 2. Though we have activated the error decomposition (2.7) for the algorithm with general kernels, it is still hard to estimate the error bounds and learning rates. The reason is that $\mathcal{F}_K$ is only a Banach space. Its capacity is not easy to control. The approximation error associated with $\mathcal{F}_K$ is also difficult. We will not go into the details of these problems but leave them for future research. But the hypothesis error is not so hard in this case. Theorem 9 tells that it depends on the quantity $\omega_K(\Delta)$.

Let us present some discussions on this term in what follows.

Definition 10. A probability P is *nondegenerate* on X if its support is X, i.e., every open subset of X has positive probability measure.

Proposition 11. *Suppose P is nondegenerate and $\{x_i\}_{i=1}^m$ are random samples drawn according to P. Then for every $\varepsilon > 0$, there exists some $\mu_\varepsilon > 0$ such that $\{x_i\}_{i=1}^m$ are ε-dense with probability at least $1 - \mathscr{N}(X, \frac{\varepsilon}{2}) \exp(-m\mu_\varepsilon)$, where $\mathscr{N}(X, \frac{\varepsilon}{2})$ is the covering number of X by balls of radius $\varepsilon/2$.*

Proof. Let B_j, $j = 1, \ldots, \mathscr{N} = \mathscr{N}(X, \frac{\varepsilon}{2})$ be the balls with radius $\frac{\varepsilon}{2}$ covering X. Then $\mu_\varepsilon = \min_j P(B_j) > 0$ since P is nondegenerate. Let

$$E_j = \left\{ \{x_i\}_{i=1}^m \subseteq X : \{x_i\}_{i=1}^m \bigcap B_j \neq \emptyset \right\}$$

and $E = \bigcap_j E_j$. It is easy to check that every element in E forms an ε-dense set. Thus, the conclusion follows from the fact that the measure of the set E is at least $1 - \mathscr{N}(1-\mu_\varepsilon)^m \geq 1 - \mathscr{N} \exp(-m\mu_\varepsilon)$. ■

Proposition 11 shows that Δ usually converges fast if the marginal distribution ρ_X is nondegenerate. This means for large data setting, the term $\omega_K(\Delta)$ will be very small for a smooth kernel.

Theorem 9 tells that the performance of coefficient based regularization schemes depends on the distribution of the sampling points $\{x_i\}_{i=1}^m$ in X. The quantity Δ and hence $\omega_K(\Delta)$ becomes smaller as the sample size increases. This fact verifies the idea of improving the performance by semi-supervised learning [40,41], i.e., adding some unlabelled data into the sample. Intuitively, this is somewhat equivalent to an idea of using a relatively larger hypothesis space.

6. Conclusions and discussions

We observed that many algorithms in the literature use hypothesis spaces trained from samples. However, little theoretical work has been devoted to the study of these algorithms. We showed some essential differences between these algorithms and those with sample independent hypothesis spaces. We also point out the difficulty to deal with these algorithms: the lack of a proper characterization of the approximation error. To overcome this difficulty, we propose the idea of using a universal hypothesis containing the union of all possible hypothesis spaces (varying with the sample) to measure the approximation ability of the algorithm. When this is used in the error decomposition procedure, an additional nontrivial term called hypothesis error appears, comparing with the algorithms with sample independent hypothesis spaces. We show that the hypothesis error can be estimated satisfactorily in many cases by bounding the quantity (2.8). This is illustrated for two particular classes of learning algorithms in kernel methods: learning the kernel via regularization and coefficient based regularization. It shows that our approach is widely applicable.

Note that our approach involves a universal hypothesis and is a rather general clue. Such a generality is sometimes also a drawback. We provided two approaches for choosing the universal hypothesis $\mathscr{H}_0$ where the penalty Ω_0 should be closely related to the data dependent penalty so that the hypothesis error can be controlled. However, there is no general theory yet and further study for choosing $(\mathscr{H}_0, \Omega_0)$ is needed.

Recall the error decomposition (2.7). Except the estimation of the hypothesis error, another concern is how much is lost in the estimation of the sample error by taking the relatively large function class $\mathscr{H}_0$. This may depend on the underlying distribution, the algorithm, $\mathscr{H}_0$ and the penalty functional Ω_0. See for example the results in [28,38].

As a very general approach we do not expect the derived error estimate to be tight in most cases. One may find some other approaches to study the performance of algorithms with sample dependent hypothesis spaces, especially when the algorithm has some special structures and special efficient methods are available. For example, when the algorithm has a two-layer minimization form, one may use the methods in [20,31,38,21,22]. As for the kernel regression via coefficient based regularization, one can bound the difference between $f_{\mathbf{z}}$ and $f_{\mathbf{z},\mu}^+$ directly because of their explicit forms. But for the coefficient based regularization with loss functions other than the square loss, no evidence shows this approach works. Our approach shows its power by a uniform solution (Proposition 3) to these problems.

There are many other algorithms falling into the setting of learning with sample dependent hypothesis spaces. We cannot work on them completely in this paper. However, we believe our idea will shed light on the study of these algorithms.

Two topics in kernel method may be interesting for the future work. One is on the learning of kernel functions. There are many methods to construct a kernel through samples, see e.g. [23,42,37]. The analysis should not be so easy as we have done in Section 3 for the regularization method. Another is the regularization scheme with the sample dependent choice of the regularization parameter. The regularization parameter restricts the space that $f_{\mathbf{z}}$ lies in. When it is chosen in such a way that depends on the sample, the regularization scheme becomes one with sample dependent hypothesis space. As far as we know, this problem is far from being well understood, even for the classical choice via cross validation [12].

Acknowledgements

The second author's work is partially supported by the Research Grants Council of Hong Kong [Project No. CityU 103405] and National Science Fund for Distinguished Young Scholars of China [Project No. 10529101].

References

[1] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, New York, 1998.
[2] M. Anthony, P. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, 1999.
[3] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.
[4] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, 2000.
[5] G. Blanchard, G. Lugosi, N. Vayatis, On the rate of convergence of regularized boosting classifiers, J. Machine Learning Res. 4 (2003) 861–894.
[6] O. Bousquet, A. Elisseeff, Stability and generalization, J. Machine Learning Res. 2 (2002) 499–526.
[7] T. Zhang, Leave-one-out bounds for kernel methods, Neural Comp. 15 (2003) 1397–1437.
[8] D. Chen, Q. Wu, Y.M. Ying, D.X. Zhou, Support vector machine soft margin classifiers: Error analysis, J. Machine Learning Res. 5 (2004) 1143–1175.
[9] Q. Wu, Y.M. Ying, D.X. Zhou, Learning rates of least-square regularized regression, Found. Comput. Math. 6 (2006) 171–192.
[10] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. 26 (2007) 153–172.
[11] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
[12] G. Wahba, Spline Models for Observational Data, SIAM, 1990.
[13] B. Schölkopf, A.J. Smola, Learning with Kernels, MIT Press, 2002.
[14] C. Scovel, I. Steinwart, Fast rates for support vector machines, in: Proc. of the 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, pp. 279–294.
[15] R. Vert, J. Vert, Consistency and convergence rates of one-class SVMs and related algorithms, J. Machine Learning Res. 7 (2006) 817–854.
[16] D.H. Xiang, D.X. Zhou, Classification with Gaussians and convex loss, 2008, preprint.
[17] F. Cucker, S. Smale, Best choices for regularization parameters in learning theory, Found. Comput. Math. 2 (2002) 413–428.
[18] Q. Wu, D.X. Zhou, Analysis of support vector machine classification, J. Comp. Anal. Appl. 8 (2006) 99–119.
[19] V. Cherkassky, F. Mulier, Learning from Data: Concepts, Theory, and Methods, Wiley, New York, 1998.
[20] G. Lugosi, A. Nobel, Adaptive model selection using empirical complexities, Ann. Stat. 27 (1999) 1830–1864.
[21] P. Binev, Al. Cohen, W. Dahmen, R. DeVore, V. Temlyakov, Universal algorithms for learning theory Part I: Piecewise constant functions, J. Machine Learning Res. 6 (2005) 1297–1321.
[22] V. Temlyakov, On universal estimators in learning theory, 2005, preprint.
[23] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Choosing multiple parameters for support vector machines, Machine Learning 46 (2002) 131–159.
[24] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, M.I. Jordan, Learning the kernel matrix with semidefinite programming, J. Machine Learning Res. 5 (2004) 27–72.
[25] C.A. Micchelli, M. Pontil, Learning the kernel function via regularization, J. Machine Learning Res. 6 (2005) 1099–1125.
[26] A.E. Hoerl, R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67.
[27] D. Donoho, For most large underdetermined systems of equations, the minimal ℓ^1-norm solution is also the sparsest near-solution, 2004, preprint.
[28] Q. Wu, D.X. Zhou, SVM soft margin classifiers: Linear programming virsus quadratic programming, Neural Comp. 17 (2005) 1160–1187.
[29] V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer-Verlag, New York, 1982.
[30] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. 39 (2001) 1–49.
[31] Q. Wu, Y. Ying, D.X. Zhou, Multi-kernel regularized classifiers, J. Complexity 23 (2007) 108–134.
[32] S. Smale, D.X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. 41 (2004) 279–305.
[33] D.X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory 49 (2003) 1743–1752.
[34] D.X. Zhou, The covering number in learning theory, J. Complexity 18 (2002) 739–767.
[35] S. Smale, D.X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. 1 (2003) 17–41.
[36] Y. Lin, H.H. Zhang, Component selection and smoothing in smoothing spline analysis of variance models — COSSO, Institute of Statistics Mimeo Series 2556, NCSU, 2003.
[37] M. Herbster, Relative loss bounds and polynomial-time predictions for the K-LMS-NET algorithm, in: Proc. 15-th Int. Conf. Algorithmic Learning Theory, 2004.
[38] C.A. Micchelli, M. Pontil, Q. Wu, D.X. Zhou, Error bounds for Learning the kernel, 2005, preprint.
[39] Y. Ying, D.X. Zhou, Learnability of Gaussians with flexible variances, J. Machine Learning Res. 8 (2007) 249–276.
[40] M. Belkin, P. Niyogi, Semi-supervised learning on Riemannian manifolds, Machine Learning 56 (2004) 209–239.
[41] K.P. Bennett, A. Demiriz, Semi-supervised support vector machines, in: D. Cohn, M. Kearns, S. Solla (Eds.), Advances in Neural Information Processing System, MIT Press, 1999, pp. 368–374.
[42] S. Mukherjee, V. Vapnik, Support vector method for multivariate density estimation, CBCL/AI Memo, 1999.