SCIENCE CHINA
Mathematics

• ARTICLES •

September 2012 Vol. 55 No. 9: 1859–1868
doi: 10.1007/s11425-012-4438-3

Generalization errors of Laplacian regularized least squares regression

CAO Ying* & CHEN DiRong

Department of Mathematics, LMIB, Beihang University, Beijing 100191*, China*
Email: caoying@ss.buaa.edu.cn, drchen@buaa.edu.cn

Received November 25, 2011; accepted January 26, 2012; published online July 25, 2012

Abstract Semi-supervised learning is an emerging computational paradigm for machine learning, that aims to make better use of large amounts of inexpensive unlabeled data to improve the learning performance. While various methods have been proposed based on different intuitions, the crucial issue of generalization performance is still poorly understood. In this paper, we investigate the convergence property of the Laplacian regularized least squares regression, a semi-supervised learning algorithm based on manifold regularization. Moreover, the improvement of error bounds in terms of the number of labeled and unlabeled data is presented for the first time as far as we know. The convergence rate depends on the approximation property and the capacity of the reproducing kernel Hilbert space measured by covering numbers. Some new techniques are exploited for the analysis since an extra regularizer is introduced.

Keywords semi-supervised learning, graph Laplacian, covering number, learning rate

MSC(2010) 62H30, 68T05

Citation: Cao Y, Chen D R. Generalization errors of Laplacian regularized least squares regression. Sci China Math, 2012, 55(9): 1859–1868, doi: 10.1007/s11425-012-4438-3

1 Introduction

Semi-supervised learning is an approach for learning from limited supervision by utilizing large amounts of inexpensive, unlabeled observations. Not only does this approach make an appeal as a model for natural learning, but it is also of potentially great practical significance in most applications of machine learning. From an engineering standpoint, data without label would be available cheaply and automatically in large quantity, but manual labeling for the purposes of training learning algorithms is often very time consuming, expensive, and error-prone. As a result, semi-supervised learning has been of growing interest over the past few years.

Since the early 1990's, a considerable amount of work has been done in the problem of learning from labeled and unlabeled data, including semi-supervised and transductive learning [2, 5, 6, 20, 21, 23, 26]. In particular, some regularization based algorithms have been proposed as well [4]. Recently, researchers have tried to develop some theoretical understanding of generalization performance of these methods. However, the existent studies are almost about the transductive learning, for example, the margin method [21] and the graph-based methods [2, 10, 11]. Such transductive learning approaches do not naturally extend to the semi-supervised case where novel test examples need to be classified (predicted). We note that the generalization errors of a semi-supervised classification method are estimated in [13] under a so-called

*Corresponding author

strong cluster assumption. It is based on density level sets estimation, and the convergence rate is achieved when one has consistent estimators of the clusters.

The purpose of this paper is to derive the generalization error bounds of the Laplacian regularized least squares algorithm (hereafter, LapRLS). It is a semi-supervised algorithm for regression problem, which arises out of the framework proposed in [4]. Over the last two decades, graph Laplacian has been applied to a wide range of clustering and semi-supervised learning. The LapRLS algorithm uses the graph Laplacian to present an additional regularization term, and exploits the geometry of the probability distribution that generates the data. In contrast to the variety of purely graph-based approaches in a transductive setting, LapRLS results in a natural out-of-sample extension from the data set (labeled and unlabeled) to novel examples. While plentiful experiments were performed with LapRLS and comparisons were made with the standard regularized least squares (hereafter, RLS) in [4], the crucial issue of generalization performance is still poorly understood. In this paper, we present generalization error bounds depending on the number of labeled and unlabeled examples and illustrate how unlabeled data improve the error bounds. Our work brings together three distinct concepts that have received some independent attention recently in machine learning: data-dependent regularization in reproducing kernel Hilbert spaces (RKHS) [14–16], the convergence of graph Laplacian [3, 7, 12, 19], and error analysis in RKHS [9, 22].

Our approach is mainly an elaborate analysis of the excess generalization error, which is often decomposed into the sum of a sample error and an approximation error. The main difficulty of analysis is the extra regularization term formulated by labeled and unlabeled samples via a graph Laplacian. Hence, a new error decomposition technique is introduced by means of an additional manifold error. Due to the Laplacian-based regularizer, the target function f_z is pushed towards a small region of hypothesis space, where functions are smooth with respect to both the ambient space and the intrinsic geometry of the probability distribution. That is to say, the extra regularizer is expected to limit the domain of the target optimization and thus to decrease sample error. Sindhwani et al. [15, 16] and Rosenberg [14] utilized this regularizer to present a modified data-dependant reproducing kernel and proved that it could improve the sample error estimation in multi-view learning measured by Rademacher complexity. However, the approximation error was not considered there and the convergence rate depending on the sample size was not proposed. In this paper, we derive a refined bound of f_z to reduce the sample error, and moreover bound the approximation and manifold errors, by using the properties of graph Laplacian and its limit version. Finally, the learning rate is established and the improvement of error bounds in terms of the number of samples is presented.

2 Laplacian regularized least squares algorithm

In this paper, we assume that the input space X is a compact metric space and the output space $Y = \mathbb{R}$. Let $\rho = \rho_{X,Y}$ be a probability distribution on $Z := X \times Y$ according to which examples are generated for function learning. The sample set z can be divided into two subsets z_1 and z_2, where $z_1 = \{(x_i, y_i)\}_{i=1}^{l}$ is a collection of labeled data drawn independently from ρ and $z_2 = \{x_j\}_{j=l+1}^{l+u}$ is a typically much larger collection of unlabeled data generated according to the marginal distribution ρ_X of ρ.

In regression problem, we will learn a predictive function over a set of functions and the set is generally an appropriately chosen reproducing kernel Hilbert space (RKHS). Recall that there is a one-to-one correspondence between RKHSs and Mercer kernels. Let $K : X \times X \to \mathbb{R}$ be continuous, symmetric, and positive semidefinite, i.e., given an arbitrary finite set $\{x_1, \ldots, x_n\} \subset X$ of points, the matrix $K = (K(x_i, x_j))_{i,j=1}^{n}$ is positive semi-definite. Such a function is called a Mercer kernel. The RKHS $\mathcal{H}_K$ associated with the kernel K is the completion of $\operatorname{span}\{K_x = K(x, \cdot) : x \in X\}$, with respect to the inner product given by $\langle K_x, K_y \rangle_K = K(x, y)$. See [1] and [9, Chapter 4] for details. Let $\kappa = \sup_{x \in X} \sqrt{K(x, x)} = \sup_{x,y \in X} \sqrt{|K(x, y)|}$. Then by $f(x) = \langle f, K_x \rangle_K, f \in \mathcal{H}_K$, we have

$$|f(x)| \leqslant \kappa \|f\|_{\mathcal{H}_K}, \qquad \forall f \in \mathcal{H}_K, x \in X. \tag{2.1}$$

Hereinafter, $\|\cdot\|_K = \|\cdot\|_{\mathcal{H}_K}$.

The framework for semi-supervised learning in [4] is based on the assumption that the high-dimensional input data truly resides on a low-dimensional manifold. The idea is reasonable in many real-world problems. For example, in vision, the images we get when viewing an object from different positions form a three-dimensional manifold in the image space. Therefore, the predictive function is supposed to be smooth with respect to the manifold, the intrinsic geometry of ρ_X. In other words, we assume that, if points close together on the manifold, then their predictions are similar. Hence, the graph Laplacian is introduced to give the intrinsic regularizer.

Given the sample set z, we construct a weighted undirected graph $G = (V, E)$ with vertex set $V = z$. Let W_{ij} be the edge weights in the data adjacency graph. In this paper, the weights W_{ij} is given by a similarity function $W(x, x') = \exp\{-\|x - x'\|^2/2\sigma^2\}$. The unnormalized graph Laplacian of G is defined as $L = D - W$, where D is a diagonal matrix with diagonal entries $D_{ii} = \sum_{j=1}^{l+u} W_{ij}$.

The LapRLS algorithm [4] solves the optimization problem

$$f_z = \arg\min_{f\in\mathcal{H}_K} \frac{1}{l}\sum_{i=1}^{l}(f(x_i) - y_i)^2 + \frac{\lambda_1}{(l+u)^2}\hat{f}^{\mathrm{T}}L\hat{f} + \lambda_2\|f\|_K^2, \tag{2.2}$$

where $\hat{f} = (f(x_1), \ldots, f(x_{l+u}))^{\mathrm{T}}$ is the sample of f on z. The regularizer $\lambda\|f\|_K^2$ is added to avoid overfitting. The second term is a smoothness penalty intuitively. Note that

$$\hat{f}^{\mathrm{T}}L\hat{f} = \frac{1}{2}\sum_{i,j=1}^{l+u}(f(x_i) - f(x_j))^2 W_{ij}. \tag{2.3}$$

Therefore, the regularizer ensures f_z satisfies the manifold smoothness assumption. The *regularization parameters* λ_1 and λ_2 are constants that control the complexity of function in both the ambient space and the intrinsic geometry of ρ_X. Note that if $\lambda_1 = 0$ the LapRLS algorithm turns into the RLS algorithm, a fully supervised method.

As discussed in [4], the target function f_z admits the representation of the form

$$f_z = \sum_{i=1}^{l+u}\alpha_i^z K(x_i, \cdot).$$

The coefficient vector $\alpha^z = (\alpha_1^z, \ldots, \alpha_{l+u}^z)^{\mathrm{T}} \in \mathbb{R}^{l+u}$ is determined by the optimization problem

$$\alpha^z = \arg\min_{\alpha\in\mathbb{R}^{l+u}} \frac{1}{l}(Y - JK\alpha)^{\mathrm{T}}(Y - JK\alpha) + \frac{\lambda_1}{(l+u)^2}\alpha^{\mathrm{T}}KLK\alpha + \lambda_2\alpha^{\mathrm{T}}K\alpha,$$

where K is the $(l+u)\times(l+u)$ Gram matrix; $Y = (y_1, \ldots, y_l, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^{l+u}$ and J is an $(l+u)\times(l+u)$ diagonal matrix with the first l diagonal entries as 1 and the rest 0.

Figures 1 and 2 show the experimental results in [4] for LapRLS and RLS algorithms applied to binary classification of handwritten digits. The training set is formed by the first 400 images for each digit in the USPS training set. The remaining images formed the test set. In Figure 1, the error rates of these two algorithms are compared at the break-even points in the precision-recall curves for 45 binary classification problems. The labeled examples are 2 images randomly chosen from each class ($l = 2$) and the unlabeled examples are the rest ($u = 398$). The error rate for each classification problem is averaged over 10 random choices of labeled examples. Figure 2 presents the performance in terms of precision-recall break-even points of RLS and LapRLS as a function of the number of labeled examples, on the test set and the unlabeled set. These experiments demonstrate that LapRLS results in significant improvements over inductive classification, i.e., RLS.

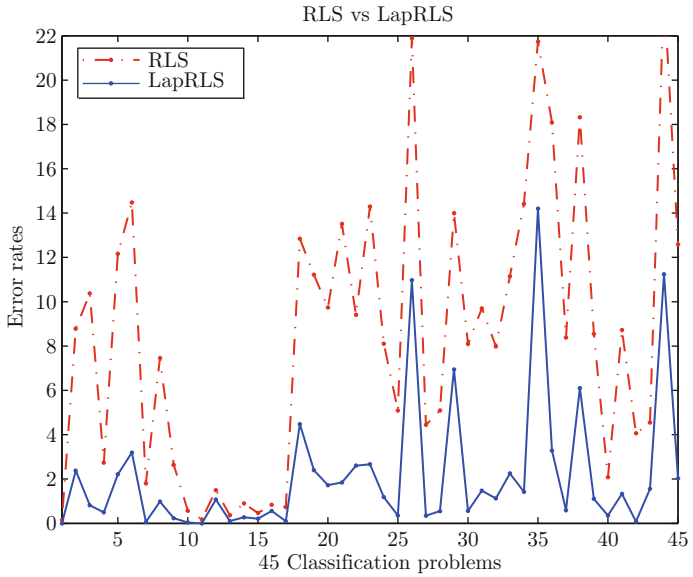

Figure 1 Error rates at precision-recall break-even points for 45 binary classification problems

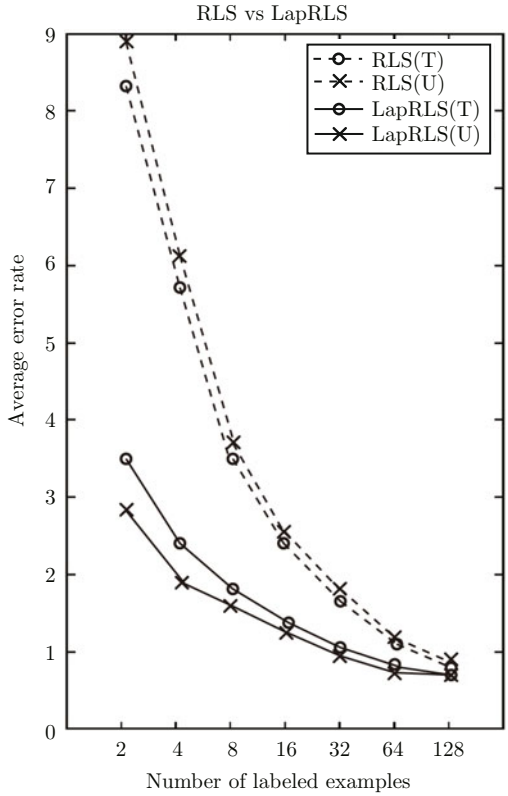

Figure 2 Mean error rate as a function of number of the labeled points

3 Problem setting

Recall that in the least square regression problem, the error for a function $f : X \to Y$ is defined as

$$\varepsilon(f) = \int_Z (f(x) - y)^2 d\rho. \tag{3.1}$$

The function that minimizes the error is called the *regression function*. It is given by

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X. \tag{3.2}$$

Here $\rho(y|x)$ is the conditional probability measure at x induced by ρ. Throughout this paper, we assume that for some $M \geqslant 0$, $|y| \leqslant M$ almost surely (with respect to ρ). It follows from the definition (3.2) that $|f_\rho(x)| \leqslant M$.

The target of the regression problem is to find good approximations of the regression function from random samples. The goodness of the approximation of f_ρ by f_z is usually measured by $\|f_z - f_\rho\|_{L^2_{\rho_X}}$. It is clear that, for any measurable function $f : X \to \mathbb{R}$ (see, e.g., [8]),

$$\|f - f_\rho\|^2_{L^2_{\rho_X}} = \varepsilon(f) - \varepsilon(f_\rho). \tag{3.3}$$

Therefore, we will be concerned with the estimate of the *excess generalization error* $\varepsilon(f_z) - \varepsilon(f_\rho)$. Hereinafter, $\|\cdot\|_2$ is used instead of $\|\cdot\|_{L^2_{\rho_X}}$.

With the *empirical error* defined as

$$\hat{\varepsilon}(f) = \frac{1}{l}\sum_{i=1}^{l}(f(x_i) - y_i)^2,$$

the scheme (2.2) can be written as

$$f_z = \arg\min_{f\in\mathcal{H}_K} \hat{\varepsilon}(f) + \frac{\lambda_1}{(l+u)^2}\hat{f}^{\mathrm{T}} L\hat{f} + \lambda_2\|f\|^2_K. \tag{3.4}$$

A usual approach for getting the learning rates for regularization schemes is error decomposition. Here, a new error decomposition technique is introduced by means of a modified *regularized error* and an extra *manifold error*.

Define an operators L_w on $\mathcal{L}^2_{\rho_x}(X)$ as

$$L_w f(x) = f(x)p(x) - \int_X f(x')W(x,x')d\rho_x(x'), \tag{3.5}$$

with $p(x) = \int W(x,x')d\rho_x(x')$. This is the limit version of graph Laplacian L (see [7]). Applying [9, Proposition 4.5], it is clear that $\|L_w\| \leqslant 2\omega^2$ with $\omega = \sup_{x,x'\in X} W(x,x')$. Moreover, (3.5) tells us that

$$\langle f_\rho, L_w f_\rho\rangle_2 = \frac{1}{2}\iint_X (f_\rho(x) - f_\rho(x'))^2 W(x,x')d\rho_X(x)d\rho_X(x') \geqslant 0. \tag{3.6}$$

Now we introduce a *regularizing function* denoted by f_λ,

$$f_\lambda = f_{\lambda_1,\lambda_2} := \arg\min_{f\in\mathcal{H}_K} \|f - f_\rho\|^2_{L^2_{\rho_X}} + \lambda_1\langle f, L_w f\rangle_2 + \lambda_2\|f\|^2_K. \tag{3.7}$$

It is easy to see $\varepsilon(f_z) - \varepsilon(f_\rho) \leqslant \varepsilon(f_z) - \varepsilon(f_\rho) + \frac{\lambda_1}{(l+u)^2}\hat{f}_z{}^{\mathrm{T}} L\hat{f}_z + \lambda_2\|f_z\|^2_K$, which can be bounded by

$$\mathcal{D}(\lambda) + \mathcal{S}(z,\lambda) + \mathcal{M}(z,\lambda),$$

where

$$\begin{aligned}
&\mathcal{D}(\lambda) = \varepsilon(f_\lambda) - \varepsilon(f_\rho) + \lambda_1\langle f, L_w f\rangle_2 + \lambda_2\|f\|^2_K,\\
&\mathcal{S}(z,\lambda) = \varepsilon(f_z) - \hat{\varepsilon}(f_z) + \hat{\varepsilon}(f_\lambda) - \varepsilon(f_\lambda),\\
&\mathcal{M}(z,\lambda) = \lambda_1\left(\frac{1}{(l+u)^2}\hat{f}_\lambda{}^{\mathrm{T}} L\hat{f}_\lambda - \langle f, L_w f\rangle_2\right).
\end{aligned}$$

The regularization error, denoted by $\mathcal{D}(\lambda)$, is supposed to tend to 0 as λ_1 and λ_2 become small. It is expected that the decay of $\mathcal{D}(\lambda)$ is not slower than that of the regularized error in the RLS setting. The quantity $\mathcal{S}(z,\lambda)$ is called *sample error*. Since $\hat{\varepsilon}(f)$ is a district quantity, estimating the sample error involves the size of the hypothesis space. In the supervised regularization scheme, as λ_2 increases, f_z is pulled towards a region Ω of $\mathcal{H}_K$ with small $\|f\|_K$. Now, due to an extra regularizer, f_z is restricted to a subset of Ω. Hence, this reduces the sample error and also, of course, produces the manifold error, denoted by $\mathcal{M}(z,\lambda)$. However, if our manifold smoothness assumption is true, this quantity will have little effect on the generalization error bounds.

In this paper, our task is to establish the generalization error bounds of LapRLS by an elaborate analysis of the excess generalization error $\varepsilon(f_z)-\varepsilon(f_\rho)$. The regularized and manifold errors are estimated in Section 4. In Section 5, the improvement of error bounds in terms of the number of labeled and unlabeled data is presented.

4 Estimate of the regularized and manifold errors

As mentioned above, the extra regularization term $\frac{\lambda_1}{(l+u)^2}\hat{f}_z{}^{\mathrm{T}}L\hat{f}_z$ is expected to reduce the sample error. Meanwhile, the decay of the regularized error should not be slow, and the effect of $\mathcal{M}(z,\lambda)$ on the generalization error bounds is negligible. In this section, we will estimate the regularized and manifold errors to illustrate the ideas.

The rate of the regularization error is not only important for measuring the ability of approximating f_ρ by functions from $\mathcal{H}_K$, but also crucial for bounding the sample error. The regularized error of least-square supervised learning has been well understood [17, 18].

Now, we introduce some notations. For a kernel $K(x,y)$, the integral operator $S_K:\mathcal{L}^2_\rho\to\mathcal{H}_K$ is defined by

$$S_Kf(x)=\int_X K(x,y)f(y)d\rho(y),\quad x\in X. \tag{4.1}$$

Clearly, as K is a Mercer kernel, S_K is a self-adjoint, positive semi-definite and compact operator. Therefore, S_K^α is well defined for any $\alpha>0$. It is well known that $\mathcal{H}_K=S_K^{1/2}(\mathcal{L}^2_\rho)$. See [9, Chapter 4] for details.

It is easy to see from (3.6) that, for any $f\in\mathcal{H}_K$,

$$\lambda_1\langle f,L_wf\rangle_2+\lambda_2\|f\|^2_K\leqslant(2\omega^2\kappa^2\lambda_1+\lambda_2)\|f\|^2_K. \tag{4.2}$$

Then we present the estimation of $\mathcal{D}(\lambda)$.

Proposition 4.1. *Suppose that K is a Mercer kernel such that $S_K^{-\alpha/2}f_\rho\in\mathcal{L}^2_\rho$ for some $0<\alpha\leqslant 1$. Then we have*

$$\mathcal{D}(\lambda)\leqslant(2\omega^2\kappa^2\lambda_1+\lambda_2)^\alpha\|S_K^{-\alpha/2}f_\rho\|^2_2.$$

Proof. It follows from (3.3) and (3.7) that

$$\mathcal{D}(\lambda)=\inf_{f\in\mathcal{H}_K}\{\|f-f_\rho\|^2_{L^2_{\rho_X}}+\lambda_1\langle f,L_wf\rangle_2+\lambda_2\|f\|^2_K\}.$$

Due to (4.2), we find

$$\mathcal{D}(\lambda)\leqslant\tilde{\mathcal{D}}(2\omega^2\kappa^2\lambda_1+\lambda_2):=\inf_{f\in\mathcal{H}_K}\|f-f_\rho\|^2_{L^2_{\rho_X}}+(\lambda_2+2\omega^2\kappa^2\lambda_1)\|f\|^2_K. \tag{4.3}$$

Under the above assumptions, applying [9, Propositon 8.5] yields

$$\tilde{\mathcal{D}}(2\omega^2\kappa^2\lambda_1+\lambda_2)\leqslant(2\omega^2\kappa^2\lambda_1+\lambda_2)^\alpha\|S_K^{-\alpha/2}f_\rho\|^2_2. \tag{4.4}$$

Obviously, our statement follows from (4.3) and (4.4). □

The above result tells us that the regularized error $\mathcal{D}(\lambda)$ decays at the same rate as that of RLS, as long as the parameters λ_1 and λ_2 are chosen appropriately.

Next, the bounds of manifold error will be derived taking advantage of the results about the graph Laplacian L and its limit version L_w.

Recall that in this paper $W(x, x')$ is chosen as the Gaussian kernel. The bound of the manifold error $\mathcal{M}(z, \lambda)$ follows from the analysis of the approximation of L_w by L in [7].

Proposition 4.2. *For all $0 < \delta < 1$, with confidence at least $1 - 3\delta$, there holds*

$$\mathcal{M}(z, \lambda) \leqslant \frac{16\omega^2\lambda_1\mathcal{D}(\lambda)\log(2/\delta_1)}{\lambda_2\sqrt{u+l}}.$$

Proof. Recall that

$$\mathcal{M}(z, \lambda) = \lambda_1\left(\frac{1}{(l+u)^2}\hat{f}_\lambda{}^{\mathrm{T}}\mathrm{L}\hat{f}_\lambda - \int_X f_\lambda L_w f_\lambda d\rho_X\right).$$

Applying [7, Proposition 4.7] yields, with confidence at least $1 - 3\delta_1$,

$$\mathcal{M}(z, \lambda) \leqslant \frac{16\omega^2\lambda_1\|f_\lambda\|_\infty^2\log(2/\delta_1)}{\sqrt{u+l}}.$$

Besides, it is straightforward that

$$\|f_\lambda\|_K \leqslant \frac{\sqrt{\mathcal{D}(\lambda)}}{\sqrt{\lambda_2}}.$$

Then the proof is completed. □

In practice, the size u of the unlabeled samples can be much larger than l, even $u/l \to \infty$. Therefore, the extra error $\mathcal{M}(z, \lambda)$ can be diminished to an arbitrarily small degree by sufficiently large u.

Note that the selection of the parameters λ_1 and λ_2 is crucial to the rates of regularized and manifold errors. We defer this discussion to later as the decay of these parameters also determines the sample error estimate.

5 Sample error and the bound of f_z

In this section, the sample error is estimated and then the generalization error bounds of LapRLS follows. Moreover, we would show the improvement of the error bounds by adding an extra regularizer with unlabeled samples.

5.1 Sample error

We are now in a position to estimate sample error $\mathcal{S}(z, \lambda)$. Write it as

$$\mathcal{S}(z, \lambda) = \left\{E(\xi_1) - \frac{1}{l}\sum_{i=1}^{l}\xi_1(z_i)\right\} + \left\{E(\xi_2) - \frac{1}{l}\sum_{i=1}^{l}\xi_2(z_i)\right\}, \tag{5.1}$$

where

$$\xi_1 = (f_z(x) - y)^2 - (f_\rho(x) - y)^2, \qquad \xi_2 = (f_\lambda(x) - y)^2 - (f_\rho(x) - y)^2.$$

Obviously, ξ_2 is a fixed random variable with mean $E(\xi_2) = \varepsilon(f_\lambda) - \varepsilon(f_\rho)$. Hence, the last term of (5.1) is a typical quantity that can be estimated by probability inequalities. However, ξ_1 should not be considered as a fixed random variable, since f_z changes with the sample z runs over a set of functions. Here, we just abuse the notion $E(\xi_2) = \varepsilon(f_\lambda) - \varepsilon(f_\rho)$. The bound of the first term of (5.1) involves the capacity of the function space $\mathcal{H}_K$, which is measured by the covering number of the balls

$$B_R := \{f \in \mathcal{H}_K; \|f\|_K \leqslant R\}.$$

Definition 5.1. For a subset $\mathcal{S}$ of a metric space and $\eta > 0$, the covering number $\mathcal{N}(\mathcal{S}, \eta)$ is defined to be the minimal $l \in \mathbb{N}$ such that there exist l disks with radius η covering $\mathcal{S}$.

When $\mathcal{S}$ is compact, this number is finite. For simplicity, we denote the covering number of B_1 in $C(X)$ with the metric $\|\cdot\|_\infty$ by $\mathcal{N}(\eta)$.

Definition 5.2. The RKHS associated with the Mercer kernel K has polynomial complexity exponent s, if we have

$$\log \mathcal{N}(\eta) \leqslant C_0(1/\eta)^s, \quad \forall\, \eta > 0, \tag{5.2}$$

for some $s > 0$.

It was known that (5.2) holds if $K \in C^{2d/s}(X)$ with $X \subset \mathbb{R}^d$. In particular, if $K \in C^\infty(X)$, (5.2) is valid for any $s > 0$. The Gaussian kernel is such an example. Refer to [24, 25] for more details about the covering number.

For supervised learning, there are some studies on the sample error, see, e.g., [9, 22]. By similar methods, it is easy to derive a preliminary sample error bound.

Lemma 5.3. *If the kernel K satisfies* (5.2)*, for some $R > 0$ and $0 < \delta < 1$, with confidence at least $1 - \delta - \mathrm{Prob}_z\{f_z \notin B_R\}$, the sample error $\mathcal{S}(z, \lambda)$ is bounded by*

$$\frac{1}{2}(\varepsilon(f_z) - \varepsilon(f_\rho)) + \frac{4}{3}(\kappa + 3)^2 R^2 \mu(l, \delta) + \mathcal{D}(\lambda)\left(1 + \frac{4\kappa^2 \log(2/\delta)}{l\lambda_2}\right) + \frac{36M^2 \log(2/\delta)}{l},$$

where

$$\mu(l, \delta) = \max\left\{\frac{80 \log(2/\delta)}{l}, \left(80C_0/l\right)^{1/(s+1)}\right\}.$$

What is left is to give a suitable $R > 0$ such that $\mathrm{Prob}_z\{f_z \notin B_R\}$ is small enough. Analysis similar to that of RLS [9, 22] shows that

$$\lambda_2 \|f_z\|_K^2 \leqslant \hat{\varepsilon}(0) \leqslant M^2.$$

Clearly, taking $R = M/\sqrt{\lambda_2} \geqslant M$ when $0 < \lambda_2 \leqslant 1$ yields $\mathrm{Prob}_z\{f_z \notin B_R\} = 0$. With this rough bound, a weak error estimation follows from Propositions 4.1, 4.2 and Lemma 5.3.

Corollary 5.4. *Suppose that the kernel K satisfies* (5.2) *and $S_K^{-\alpha/2} f_\rho \in \mathcal{L}_\rho^2$ for some $0 < \alpha \leqslant 1$. For any $0 < \delta < 1$, with confidence at least $1 - 4\delta$,*

$$\|f_z - f_\rho\|_2^2 \leqslant \zeta_{l,u,\lambda,\delta},$$

where

$$\begin{aligned}\zeta_{l,u,\lambda,\delta} = {} & c_1(\lambda_2 + 2\omega^2\kappa^2\lambda_1)^\alpha\left(4 + \frac{8\kappa^2 \log(2/\delta)}{l\lambda_2} + \frac{32\omega^2\lambda_1 \log(2/\delta_1)}{\lambda_2\sqrt{u+l}}\right) \\ & + \frac{3(\kappa+3)^2 M^2 \mu(l, \delta)}{\lambda_2} + \frac{72M^2 \log(2/\delta)}{l},\end{aligned} \tag{5.3}$$

with $c_1 = \|S_K^{-\alpha/2} f_\rho\|_2^2$.

The result is rough because we use the bound $\|f_z\|_K \leqslant M/\sqrt{\lambda_2}$, which is obtained by the same method as in the supervised case. However, in semi-supervised learning, unlabeled data is expected to reduce the size of the function class in which f_z lies, due to an additional regularizer.

5.2 The modified error bounds

Our next concern is to show how unlabeled data contributes to the choice of R. It easily follows from (3.4) that

$$\frac{\lambda_1}{(u+l)^2} \hat{f}_z{}^{\mathrm{T}} L \hat{f}_z + \lambda_2 \|f_z\|_K^2 \leqslant \hat{\varepsilon}(0) \leqslant M^2. \tag{5.4}$$

That is to say, f_z belongs to a set Ω given by

$$\Omega := \left\{ f \in \mathcal{H}_K; \lambda_2 \|f\|_K^2 \leqslant M^2 - \frac{\lambda_1}{(u+l)^2} \hat{f}^{\mathrm{T}} L \hat{f} \right\}.$$

By (2.3), one finds $\hat{f}^{\mathrm{T}} L \hat{f} \geqslant 0$. Hence, the bound of $\|f_z\|_K$ would be improved by estimating the term $\frac{\lambda_1}{(u+l)^2} \hat{f_z}^{\mathrm{T}} L \hat{f_z}$.

In what follows, we make the assumption: f_ρ is not a constant function.

This assumption is reasonable because the constant function is not of much practical interest for regression. It is clear from (3.6) that $\langle f_\rho, L_w f_\rho \rangle_2 = 0$ holds if and only if f_ρ is a constant almost everywhere, since $W(x, x') > 0$. Consequently, the above assumption verifies $\langle f_\rho, L_w f_\rho \rangle_2 \geqslant b_0$, for some $b_0 > 0$.

Under the above assumption, a modified bound of $\|f_z\|_K$ is proposed.

Proposition 5.5. *If f_ρ is not a constant function, for $0 < \delta < 1$, $l \geqslant l_{\delta,\lambda}$ and $u \geqslant u_{\delta,\lambda}$, there holds with confidence at least $1 - 8\delta$,*

$$\|f_z\|_K \leqslant \sqrt{\frac{M^2 - b_0 \omega^2 \lambda_1/2}{\lambda_2}}.$$

Here, $l_{\delta,\lambda}$ and $u_{\delta,\lambda}$ are minimal positive integers such that $1024 \|f_\rho\|_2^2 \zeta_{l,u,\lambda,\delta} \leqslant b_0^2$ and

$$u_{\delta,\lambda} \geqslant \max\{4\omega^2 \big(\log(2/\delta)\big)^{2+2/s}, 4C_0^2 \omega^{4+s} (192M^2)^{2+s} / \lambda_2^{2+s}\}, \tag{5.5}$$

where $\zeta_{l,u,\lambda,\delta}$ is given as (5.3).

Proof. On the one hand, the analysis in the proof of [7, Theorem 4.11] shows

$$\frac{1}{(l+u)^2} \hat{f_z}^{\mathrm{T}} L \hat{f_z} \geqslant \int f_z(x) L_w f_z(x) d\rho_x - \frac{48\omega^3 M^2}{\lambda_2} \left(\frac{2\omega C_0}{\sqrt{u+l}} \right)^{1/(s+1)},$$

with confidence $1 - 5\delta$, when $u \geqslant 4\omega^2 (\log(2/\delta))^{2+2/s}$.

On the other hand,

$$\int f_z(x) L_w f_z(x) d\rho_x \geqslant \int f_\rho(x) L_w f_\rho(x) d\rho_x - 2\omega^2 \|f_\rho - f_z\|_2 \big(2\|f_\rho\|_2 + \|f_\rho - f_z\|_2\big).$$

Consequently, one finds by Corollary 5.4,

$$\frac{1}{(l+u)^2} \hat{f_z}^{\mathrm{T}} \mathrm{L} \hat{f_z} \geqslant \frac{b_0 \omega^2}{2},$$

provided that $l \geqslant l_{\delta,\lambda}$ and $\geqslant u_{\delta,\lambda}$. Then our statement follows from (5.4). □

Now, with the modified bound $R = \sqrt{(M^2 - b_0 \omega^2 \lambda_1/2)/\lambda_2}$, we will establish the learning rate incorporating the regularized error estimation and the effect of the extra error $\mathcal{M}(z, \lambda)$.

Theorem 5.6. *Suppose that f_ρ is not a constant function, the kernel K satisfies* (5.2) *and $S_K^{-\alpha/2} f_\rho \in \mathcal{L}_\rho^2$ for some $0 < \alpha \leqslant 1$. Take $\lambda_1 = u^\theta$, $\lambda_2 = l^\theta$, with $\theta = 1/[(1+\alpha)(1+s)]$, For any $0 < \delta < 1$, $l \leqslant l_\delta$, and $u \geqslant u_{\delta,l}$ with confidence at least $1 - 8\delta$, there holds*

$$\|f_z - f_\rho\|_2 \leqslant \tilde{C}_1 \log(2/\delta) l^{-\alpha\theta} - \tilde{C}_2 (u/l)^{-\theta} l^{-\frac{1}{1+s}}, \tag{5.6}$$

where

$$\begin{aligned}
&u_{\delta,l} = \max\{u_{\delta,\lambda_2}, l, \tilde{C}_1^2 \big(\log(2/\delta)\big)^2 l^{2\theta} / \tilde{C}_2^2\},\\
&l_\delta = \max\{80 C_0^{-1/s} \big(\log(2/\delta)\big)^{(1+1/s)}, C_3^{(1+1/\alpha)(1+s)}\},\\
&\tilde{C}_1 = 4c_\alpha (1 + 2\omega^2\kappa^2)^\alpha (1 + 2\kappa^2) + 3M^2 (24 + (\kappa+3)^2 (80C_0)^{1/(1+s)}),\\
&\tilde{C}_2 = 3/4 b_0 \omega^2 (\kappa+3)^2 (80C_0)^{1/(1+s)}.
\end{aligned}$$

Although the learning rate is not improved, the error is reduced by $\tilde{C}_2(u/l)^{-\theta}l^{-\frac{1}{1+s}}$.

In this paper, we present a result that shows how the unlabeled data reduces the error bounds. A refined bound R of the target function is proposed, and yet there is still a lot of room for improvement. For instance, the iteration technique [22] to enhance the learning rate and a more elaborate measure of the complexity of Ω. These will be the issues in our future work.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 11171014 and 11101024) and National Basic Research Program of China (973 Project) (Grant No. 2010CB731900).

References

1 Aronszajn N. Theory of reproducing kernels. Trans Amer Math Soc, 1950, 68: 337–404
2 Belkin M, Niyogi P. Semi-supervised learning on riemannian manifolds. Machine Learning, 2004, 56: 209–239
3 Belkin M, Niyogi P. Towards a theoretical foundation for Laplacian-based manifold methods. J Comput System Sci, 2008, 74: 1289–1308
4 Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res, 2006, 7: 2399–2434
5 Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts. In: Proceedings of the 18th International Conference on Machine Learning. Waltham: Morgan Kaufmann Publishers Inc., 2001, 19–26
6 Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory. New York: ACM, 1998, 92–100
7 Cao Y, Chen D-R. Consistency of regularized spectral clustering. Appl Comput Harmon Anal, 2010, 30: 319–336
8 Cucker F, Smale S. On the mathematical foundations of learning. Bull Amer Math Soc, 2002, 39: 1–50
9 Cucker F, Zhou D-X. Learning Theory: An Approximation Theory Viewpoint. Cambridge: Cambridge University Press, 2007
10 Johnson R, Zhang T. On the effectiveness of Laplacian normalization for graph semi-supervised learning. J Mach Learn Res, 2007, 8: 1489–1517
11 Johnson R, Zhang T. Graph-based semi-supervised learning and spectral kernel design. IEEE Trans Inform Theory, 2008, 54: 275–288
12 Luxburg Von U, Belkin M, Bousquet O. Consistency of spectral clustering. Ann Statist, 2008, 36: 555–586
13 Rigollet P. Generalization error bounds in semi-supervised classification under the cluster assumption. J Mach Learn Res, 2007, 8: 1369–1392
14 Rosenberg D. Semi-supervised learning with multiple views. Ph.D. Thesis, California: University of California, 2008
15 Rosenberg D, Sindhwani V, Bartlett P, et al. Multiview point cloud kernels for semisupervised learning [lecture notes]. IEEE Signal Processing Magazine, 2009, 26: 145–150
16 Sindhwani V, Niyogi P, Belkin M. Beyond the point cloud: from transductive to semi-supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. New York: ACM, 2005, 824–831
17 Smale S, Zhou D-X. Estimating the approximation error in learning theory. Anal Appl, 2003, 1: 17–41
18 Smale S, Zhou D-X. Shannon sampling ii: Connections to learning theory. Appl Comput Harmon Anal, 2005, 19: 285–302
19 Smale S, Zhou D-X. Geometry on probability spaces. Constr Approx, 2009, 30: 311–323
20 Szummer M, Jaakkola T. Partially labeled classification with markov random walks. Ad Neural Inform Proc Syst, 2002, 2: 945–952
21 Vapnik V. Statistical Learning Theory. New York: Wiley-Interscience, 1998
22 Wu Q, Ying Y, Zhou D-X. Learning rates of least-square regularized regression. Found Comput Math, 2006, 6: 171–192
23 Zhou D, Bousquet O, Lal T, et al. Learning with local and global consistency. Adv Neural Inform Processing Syst, 2004, 16: 321–328
24 Zhou D-X. The covering number in learning theory. J Complexity, 2002, 18: 739–767
25 Zhou D-X. Capacity of reproducing kernel spaces in learning theory. IEEE Trans Inform Theory, 2003, 49: 1743–1752
26 Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th International Conference on Machine Learning. California: AAAI Press, 2003