International Journal of Wavelets, Multiresolution
and Information Processing
Vol. 8, No. 5 (2010) 695–712

DOI: 10.1142/S0219691310003729

World Scientific
www.worldscientific.com

LOCAL LEARNING ESTIMATES BY INTEGRAL OPERATORS

HONG LI* and NA CHEN†

School of Mathematics and Statistics
Huazhong University of Science and Technology
Wuhan 430074, P. R. China
**hongli@mail.hust.edu.cn*
†*chenna0407@yahoo.com.cn*

YUAN Y. TANG

Center for Pattern Recognition and Information Intelligence
Chongqing University, Chongqing 400044, P. R. China
yytang@comp.hkbu.edu.hk

Received 26 March 2009
Revised 22 March 2010

In this paper, we consider the problem of local risk minimization on the basis of empirical data, which is a generalization of the problem of global risk minimization. A new local risk regularization scheme is proposed. The error estimate for the proposed algorithm is obtained by using probabilistic estimates for integral operators. Experiments are presented to illustrate the general theory. Simulation results on several artificial real datasets show that the local risk regularization algorithm has better performance.

Keywords: Local risk regularization; integral operator; sample error; reproducing kernel Hilbert space.

AMS Subject Classification: 22E46, 53C35, 57S20

1. Introduction

Learning problem can be viewed as the problem of choosing the desired dependence on the basis of empirical data. The goal of learning is to recover the unknown function from the knowledge of the given dataset. It is well known that such a problem is ill-posed as there exists an infinity of functions that pass perfectly through the data. One way to transform this problem into a well-posed one is to assume that the function presents some smoothness properties. In recent years there has been a growing interest around regularization theory, (see Refs. 1, 4 and 15), in which the ill-conditioned estimation from data problem is transformed into a well-conditioned problem by means of a stabilizer, which is a functional with specifical properties.

Regularization theory plays an important role in learning theory. Probabilistic upper bounds on the excess risk of the empirical estimators by regularization algorithms are known, fast learning rate can be obtained by iteration methods,[21] and optimal rates are established by assuming some smoothness condition on the regression function.[7,9] Covering number technique has been used to obtain explicit bounds expressed in terms of suitable complexity measures of the regression function.[6,10] These works are all capacity-based approaches with capacity described by VC-dimension, covering numbers,[5] Rademacher complexities, etc. Capacity analysis is rather general and can lead to fast learning rates. However, the drawback is that the capacity is difficult to estimate. For example, the covering numbers of RKHS are only well estimated for certain very smooth kernels such as polynomial and Gaussians and the underlying input space X is finite-dimensional compact set. Little studies are available when X is of infinite dimension to our best knowledge. The covering techniques have been replaced by estimates of integral operators through concentration inequalities of vector valued random variables.[3,12–14,16] In this paper, we will focus on the capacity independent error bounds and learning rates by using the integral operator technique which can be used to estimate both the estimation error and the approximation error.

It is clear that the problem of global estimation of a function have heavy demands on the estimation functions to obtain an appropriate level of approximation. In fact, a better solution can be obtained by estimating the desired function on subspaces of the input space. That is to say, estimate the function in a vicinity of the point of interest. In other words, better accuracy can be achieved by approximating the function locally. To consider such a problem the model of the local risk minimization is introduced in Refs. 2, 17 and 18.

We consider local risk minimization schemes associated with the least square loss in reproducing kernel Hilbert spaces (RKHS).[11] The goal of local risk minimization is the estimation of function in a vicinity of the point of interest. This problem has already been studied in the statistical learning literature.[18] The well-known classical methods such as the method of K-nearest neighbors in pattern recognition, the Watson–Nadaraya method (method of moving average) in the regression estimation problem and the efficient statistical modeling of wavelet coefficients for image denoising in Ref. 8 are the simplest local methods for solving the problem of minimizing the local risk functional, which uses the empirical risk minimizing principle.

In this paper, we propose a new local risk minimization scheme which modified regularization scheme utilizing the local information of a vicinity of point x_0. It preserves local and global information together.[19,22] It is effective for preserving the local information among the given local point x_0 and hence can lead to good classification results for the datasets. Compared with the operators defined in Ref. 12, the local sampling operator and the local integral operator are defined. We generalize the integral theory into this proposed local method, and apply the operators for function reconstruction to the proposed learning scheme. In particular, we show

that a regression function can be approximated by a local risk regularization scheme f_{z,x_0} in $\mathcal{H}_K$. Then the capacity independent error bounds and learning rates are derived by estimates of local integral operators through concentration inequalities of vector valued random variables. Experiments show that the proposed algorithm can work well for the real life examples since it preserves the local structure of data.

The rest of this paper is organized as follows. Section 2 reviews the basic theory of local risk minimization briefly. A new learning scheme of local risk regularization (LRR) is proposed. The relationship between the proposed scheme and regularized least squares (RLS) algorithm is given in Sec. 3, and a generalization error bound is proved in Sec. 4. Section 5 illustrates our algorithm by simulation experiments. Some conclusions are given in Sec. 6.

2. The Proposed Algorithm and Function Reconstruction

In the following, we assume the samples are drawn from a probability measure ρ on $Z = X \times Y$ with a compact metric space X and $Y = R$. Our primary objective is the regression function of ρ defined as

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

Here $\rho(y|x)$ is the conditional distribution at x induced by ρ.

Let $K : X \times X \to R$ be continuous, symmetric and positive semi-definite, i.e. for any finite set of distinct points $\mathbf{x} = \{x_1, x_2, \ldots, x_m\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^m$ is positive semi-definite. Such a kernel is called a Mercer kernel. The reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated with the kernel K is defined to be the closure of the linear span of the set of functions $\{K_x = K(x, \cdot) : x \in X\}$ with the inner product denoted as $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_y \rangle_K = K(x, y)$. The reproducing property takes the form

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

Denote $\kappa = \sqrt{\sup_{x \in X} K(x, x)}$. Then reproducing property implies that $\mathcal{H}_K \subset C(X)$ and

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

We first give a description of the problem of local estimation of a function, that is the estimation of function in a vicinity of the point of interest. It has been clarified that why local function estimation can be better than global by Vapnik.[18] If we are interested in the estimation of function around x_0, we hope that points whose distance to x_0 is small have great influence on the solution, points away from x_0 have little influence on the solution. In all the previous considerations we use a loss function defined by some variable (x, y). In local risk minimization model we introduce the specific structure of loss functions which consider a nonnegative

function $k(x, x_0; \beta)$ that embodies the concept of vicinity. This function depends on a point x_0 and a "locality" parameter $\beta \in (0, \infty)$ and satisfies conditions:

$$0 \leq k(x, x_0; \beta) \leq 1 \quad \text{and} \quad k(x_0, x_0; \beta) = 1.$$

For example, both the "hard threshold" vicinity function

$$k_1(x, x_0; \beta) = \begin{cases} 1, & \text{if } \|x - x_0\| < \dfrac{\beta}{2}; \\ 0, & \text{otherwise.} \end{cases}$$

and the "soft threshold" vicinity function

$$k_2(x, x_0; \beta) = \exp\left\{-\frac{(x - x_0)^2}{\beta^2}\right\}$$

meet these conditions. Define the value

$$\mathcal{K}(x_0, \beta) = \int_X k(x, x_0; \beta) d\rho_X.$$

We consider the following loss function:

$$Q(y, f(x, \alpha)) = (y - f(x, \alpha))^2,$$

the local risk functional

$$\mathcal{E}(f, x_0; \beta) = \int_Z (y - f(x))^2 \frac{k(x, x_0; \beta)}{\mathcal{K}(x_0, \beta)} d\rho, \tag{2.1}$$

and the local empirical functional

$$\mathcal{E}_z(f, x_0; \beta) = \frac{1}{m} \sum_{i=1}^{m} (y - f(x_i))^2 \frac{k(x_i, x_0; \beta)}{\mathcal{K}(x_0, \beta)}. \tag{2.2}$$

The goal is to minimize the local risk functional. Note that the problem of local risk minimization on the basis of empirical data is a generalization of the problem of global risk minimization. (In the last problem we have to minimize (2.1), where $k(x, x_0, \beta) = 1$.) It is well known that such a problem is ill-posed as there exists an infinity of functions that pass perfectly through the data. One way to transform this problem into a well-posed one is to assume that the function presents some smoothness properties. So we propose the following learning scheme which considers both local and global conditions for function estimation.

Learning Scheme. The regularized version of the local learning problem takes the form

$$f_{z,x_0} = \arg \min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_z(f, x_0; \beta) + \|f\|_K^2 \right\}, \tag{2.3}$$

where $\mathcal{E}_z(f, x_0; \beta)$ is defined in (2.2). We do not use parameter λ as in RLS to balance the trade-off between fitness of f to the data and the smoothness of f since the parameter is hidden in the local value $\mathcal{K}(x_0, \beta)$.

Note that f_ρ is a minimizer of (2.1), we can estimate the function in the vicinity of a given point x_0 in a reproducing kernel Hilbert space, as well as investigate how f_{z,x_0} approximates f_ρ, and how the choices of the local point x_0 and the local parameter β lead to optimal convergence rates.

To understand (2.3), we define the local sampling operator $\hat{S}_{\mathbf{x}} : \mathcal{H}_K \to \ell^2(\mathbf{x})$ by

$$\hat{S}_{\mathbf{x}}(f) = (f(x_i) \cdot k(x_i, x_0; \beta))_{i=1}^m.$$

It is an intuitional generalization of sampling operator $S_{\mathbf{x}} = (f(x_i))_{i=1}^m$, (see Ref. 12), by considering a nonnegative function $k(x, x_0; \beta)$ that embodies the concept of vicinity. We shall also assume that $\hat{S}_{\mathbf{x}}$ is bounded.

Denote $\hat{S}_{\mathbf{x}}^T$ as the adjoint of $\hat{S}_{\mathbf{x}}$. Then for each $c \in \ell^2(\mathbf{x})$, there holds

$$\begin{aligned}
\langle f, \hat{S}_{\mathbf{x}}^T c\rangle_K &= \langle \hat{S}_{\mathbf{x}} f, c\rangle_{\ell^2(\mathbf{x})} \\
&= \sum_{i=1}^m c_i f(x_i) k(x_i, x_0; \beta) \\
&= \sum_{i=1}^m c_i \langle f, K_{x_i}\rangle_K k(x_i, x_0; \beta) \\
&= \left\langle f, \sum_{i=1}^m c_i K_{x_i} k(x_i, x_0, \beta) \right\rangle_K,
\end{aligned}$$

for all $f \in \mathcal{H}_K$. It follows that

$$\hat{S}_{\mathbf{x}}^T c = \sum_{i=1}^m c_i K_{x_i} k(x_i, x_0; \beta), \quad c \in \ell^2(\mathbf{x}).$$

Theorem 2.1. *If* $\frac{1}{m}\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}} + \mathcal{K}(x_0, \beta)$ *is invertible, then* f_{z,x_0} *exists, is unique and*

$$f_{z,x_0} = \left(\frac{1}{m}\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}} + \mathcal{K}(x_0, \beta)\right)^{-1} \frac{1}{m}\hat{S}_{\mathbf{x}}^T y. \tag{2.4}$$

Proof. From definition (2.2), we know that

$$\begin{aligned}
&\mathcal{E}_z(f, x_0; \beta) + \|f\|_K^2 \\
&\quad = \frac{1}{m \cdot \mathcal{K}(x_0, \beta)} \sum_{i=1}^m (y_i - f(x_i))^2 k(x_i, x_0; \beta) + \|f\|_K^2 \\
&\quad = \frac{1}{m \cdot \mathcal{K}(x_0, \beta)} \left(\sum_{i=1}^m y_i^2 - 2\sum_{i=1}^m y_i f(x_i) + \sum_{i=1}^m f^2(x_i)\right) k(x_i, x_0; \beta) + \|f\|_K^2 \\
&\quad = \frac{1}{m \cdot \mathcal{K}(x_0, \beta)} \left(\sum_{i=1}^m y_i^2 k(x_i, x_0; \beta) - 2\langle f, \hat{S}_{\mathbf{x}}^T y\rangle_K + \langle f, \hat{S}_{\mathbf{x}}^T S_{\mathbf{x}}\rangle_K\right) + \langle f, f\rangle_K.
\end{aligned}$$

Taking the functional derivative for $f \in \mathcal{H}_K$, we see that any minimizer f_{z,x_0} of (2.3) satisfies

$$\left(\frac{1}{m}\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}} + \mathcal{K}(x_0, \beta)\right) f = \frac{1}{m}\hat{S}_{\mathbf{x}}^T y.$$

This proves Theorem 2.1. □

Proposition 2.1. *The operator $S_{\mathbf{x}}$ satisfies*

$$\left\| \left(\frac{1}{m}\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}} + \mathcal{K}(x_0, \beta\right)^{-1} \right\| \leq \frac{1}{\frac{1}{m}\lambda_{\mathbf{x}}^2 + \mathcal{K}(x_0, \beta)},$$

where $\lambda_{\mathbf{x}}^2 := \inf_{f \in \mathcal{H}_K} \frac{\langle \hat{S}_{\mathbf{x}} f, S_{\mathbf{x}} f\rangle_{\ell^2(\mathbf{x})}}{\|f\|_K^2}$.

Proof. Let $v \in \mathcal{H}_K$ and $u = \left(\frac{1}{m}\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}} + \mathcal{K}(x_0, \beta)\right)^{-1} v$. Then

$$\left(\frac{1}{m}\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}} + \mathcal{K}(x_0, \beta)\right) u = v.$$

Taking inner products on both sides with u, we have

$$\frac{1}{m}\langle \hat{S}_{\mathbf{x}} u, S_{\mathbf{x}} u\rangle_{\ell^2(\mathbf{x})} + \mathcal{K}(x_0, \beta)\|u\|_K^2 = \langle v, u\rangle_K \leq \|v\|_K \|u\|_K.$$

The definition of $\lambda_{\mathbf{x}}^2$ tells us that

$$\langle \hat{S}_{\mathbf{x}} u, S_{\mathbf{x}} u\rangle_{\ell^2(\mathbf{x})} \geq \lambda_{\mathbf{x}}^2 \|u\|_K^2.$$

It follows that

$$\left(\frac{1}{m}\lambda_{\mathbf{x}}^2 + \mathcal{K}(x_0, \beta)\right) \|u\|_K^2 \leq \|v\|_K \|u\|_K.$$

Hence $\|u\|_K \leq \left(\frac{1}{m}\lambda_{\mathbf{x}}^2 + \mathcal{K}(x_0, \beta)\right)^{-1} \|v\|_K$. This is true for every $v \in \mathcal{H}_K$. Then the conclusion follows. □

Observe that $\hat{S}_{\mathbf{x}}^T : \ell^2(\mathbf{x}) \to \mathcal{H}_K$ is given by $\hat{S}_{\mathbf{x}}^T c = \sum_{i=1}^m c_i K_{x_i} k(x_i, x_0; \beta)$. Then $\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}}$ satisfies

$$\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}} f = \sum_{i=1}^m f(x_i) K_{x_i} k(x_i, x_0; \beta), \quad \forall\, f \in \mathcal{H}_K.$$

So $\frac{1}{m}\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}}$ is a good approximation of the local integral operator $\hat{L}_K : L_{\rho_X}^2 \to \mathcal{H}_K$ defined by

$$\hat{L}_K(f)(x) = \int_X K(x, y) f(y) k(y, x_0; \beta) d\rho_X(y), \quad x \in X.$$

Denote f_{x_0} is a minimizer of the following optimization problem:

$$f_{x_0} := \arg\min_{f\in\mathcal{H}_K} \left\{\mathcal{E}(f, x_0; \beta) - \mathcal{E}(f_\rho, x_0; \beta) + \|f\|_K^2\right\}. \tag{2.5}$$

Since

$$\mathcal{E}(f, x_0; \beta) - \mathcal{E}(f_\rho, x_0; \beta) = \int_X \frac{(f_\rho(x) - f(x))^2 k(x, x_0; \beta)}{\mathcal{K}(x_0, \beta)} d\rho_X, \tag{2.6}$$

we can rewrite (2.5) as

$$\min_{f\in\mathcal{H}_K} \int_X (f_\rho(x) - f(x))^2 k(x, x_0; \beta) d\rho_X + \mathcal{K}(x_0, \beta)\|f\|_K^2. \tag{2.7}$$

Since $\mathcal{K}(x_0, \beta) > 0$, a solution of (2.7) exists, is unique and given by

$$f_{x_0} = (\hat{L}_K + \mathcal{K}(x_0, \beta))^{-1} \hat{L}_K f_\rho. \tag{2.8}$$

Our goal is to understand how f_{z,x_0} approximates f_ρ and how the decay of the parameters x_0 and β lead to convergence rates.

3. The Relationship between LRR and RLS

A straightforward connection between the proposed LRR scheme and the regularized least square algorithm is discussed. We bring the local risk regularization into the framework of the regularization approach in which the sensitivity measure corresponds to a local point x_0 and parameter $\beta \in (0, \infty)$ is introduced.

As it is often done in the LRR algorithm we control the solution by choosing a local point x_0 of interest and a local parameter β to get the estimation function in a vicinity of the point x_0 of interest.

To describe the regularized least square algorithm it is convenient to remember that from the learning scheme in Ref. 14,

$$f_{z,\lambda} := \arg\min_{f\in\mathcal{H}_K} \left\{\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2 + \lambda\|f\|_K^2\right\}. \tag{3.1}$$

We know that a solution $f_{z,\lambda}$ of (3.1) exists, is unique and given by

$$f_{z,\lambda} = \left(\frac{1}{m}S_{\mathbf{x}}^T S_{\mathbf{x}} + \lambda I\right)^{-1} \frac{1}{m} S_{\mathbf{x}}^T y, \tag{3.2}$$

where $S_{\mathbf{x}}$ is the sampling operator.

Equations (2.4) and (3.2) show that solutions of the local risk regularization and RLS algorithms have the same form. The problem of local risk regularization on the basis of empirical data is a generalization of the problem of regularized least square algorithm, because we can transform it into the framework of the regularization approach by choosing $k(x, x_0, \beta) = 1$.

In the LRR scheme, we introduce the specific structure of loss functions which consider a nonnegative function $k(x, x_0; \beta)$ that embodies the concept of vicinity. The function $k(x, x_0; \beta)$ controls the local information around x_0 by achieving the

effect that points close to x_0 have great influence on the solution, points away from x_0 have little influence on the solution. That is to say, RLS considers global information only while LRR considers both local and global information. Thus, LRR can perform better than RLS since it preserves the locality properties. We choose the regularization parameter λ for a fast rate of convergence to the regression function in RLS algorithm, but the convergence rate depends on the local point x_0 and parameter β in local risk regularization algorithm. For different point of interest, we can get a better solution by choosing local parameter β.

Recall that the goodness of the approximation f_{z,x_0} is measured by $\|f_{z,x_0}-f_\rho\|_K$. As usual, we write

$$\|f_{z,x_0} - f_\rho\|_K \le \|f_{z,x_0} - f_{x_0}\|_K + \|f_{x_0} - f_\rho\|_K,$$

where f_{x_0} is defined by Eq. (2.8).

So the error is split into two parts: the first term on the right-hand side called the sample error (or estimation error) and the second term called the approximation error. In this paper, we only provide a simpler approach with the sample error.

4. Error Analysis

The aim of this section is to give a probabilistic upper bound on the expect risk of the solution given by the local risk regularization algorithm. In fact we show that the risk bounds obtained in Ref. 14 can be straightforwardly rephrased in this setting.

We assume that for some $M \ge 0, |y| \le M$ almost surely, that is, $\rho(y|x)$ is supported on $[-M, M]$ for almost every $x \in X$. Then $\|f_\rho\|_\rho \le \|f_\rho\|_\infty \le M$.

Theorem 4.1. *Let* $\mathbf{z}$ *be randomly drawn according to* ρ *satisfying* $|y| \le M$ *almost surely. Then for any* $0 < \delta < 1$, *with confidence* $1 - \delta$ *there holds*

$$\|f_{z,x_0} - f_{x_0}\|_K \le \frac{6\kappa M \log\left(\frac{2}{\delta}\right)}{\sqrt{m}\mathcal{K}(x_0,\beta)}.$$

Proof. Recall the function defined by (2.4), we know that

$$\begin{aligned} f_{z,x_0} - f_{x_0} &= \left(\frac{1}{m}\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}} + \mathcal{K}(x_0,\beta)\right)^{-1} \frac{1}{m}\hat{S}_{\mathbf{x}}^T y - f_{x_0} \\ &= \left(\frac{1}{m}\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}} + \mathcal{K}(x_0,\beta)\right)^{-1} \left(\frac{1}{m}\hat{S}_{\mathbf{x}}^T y - \frac{1}{m}\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}} f_{x_0} - \mathcal{K}(x_0,\beta) f_{x_0}\right). \end{aligned}$$

Observe that

$$\frac{1}{m}\hat{S}_{\mathbf{x}}^T y - \frac{1}{m}\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}} f_{x_0} = \frac{1}{m}\sum_{i=1}^m (y_i - f_{x_0}(x_i)) K_{x_i} k(x_i, x_0; \beta)$$

and by the definition of f_{x_0},

$$\mathcal{K}(x_0,\beta)f_{x_0} = \hat{L}_K(f_\rho - f_{x_0}).$$

It follows that for all $\{(x_i,y_i)\}_{i=1}^m$, $x_0 \in \mathbf{x}$ and $\beta > 0$,

$$f_{z,x_0} - f_{x_0} = \left(\frac{1}{m}\hat{S}_{\mathbf{x}}^T S_{\mathbf{x}} + \mathcal{K}(x_0,\beta)\right)^{-1} \times \left(\frac{1}{m}\sum_{i=1}^m (y_i - f_{x_0}(x_i))k(x_i,x_0;\beta)K_{x_i} - \hat{L}_K(f_\rho - f_{x_0})\right).$$

Considering Proposition 2.1, we have

$$\|f_{z,x_0} - f_{x_0}\|_K \le \frac{1}{\mathcal{K}(x_0,\beta)}\triangle,$$

where

$$\triangle = \left\|\frac{1}{m}\sum_{i=1}^m (y_i - f_{x_0}(x_i))k(x_i,x_0;\beta)K_{x_i} - \hat{L}_K(f_\rho - f_{x_0})\right\|_K.$$

To estimate $\triangle$, we apply the Bennett inequality (see Ref. 14) to the random variable $\xi = (y - f_{x_0}(x))K_x k(x,x_0;\beta)$ on (Z,ρ) with values in the Hilbert space $\mathcal{H}_K$. It satisfies

$$\begin{aligned} E(\xi) &= \int_Z (y - f_{x_0}(x))k(x,x_0;\beta)K_x d\rho \\ &= \int_X (f_\rho(x) - f_{x_0}(x))k(x,x_0;\beta)K_x d\rho_X \\ &= \hat{L}_K(f_\rho - f_{x_0}) \end{aligned}$$

and

$$\|\xi\|_K = |y - f_{x_0}(x)||k(x,x_0;\beta)|\|K_x\|_K \le \kappa(M + \|f_{x_0}\|_\infty) =: \widetilde{M},$$

$$\sigma^2(\xi) = E(\|\xi\|^2) \le \kappa^2 \int_Z (y - f_{x_0}(x))^2 d\rho.$$

It follows from the Bennett inequality that with confidence $1-\delta$ there holds

$$\triangle \le \frac{2\kappa(M + \|f_{x_0}\|_\infty)\log(2/\delta)}{m} + \kappa\sqrt{\frac{2\log(2/\delta)}{m}\int_Z (y - f_{x_0}(x))^2 k(x,x_0;\beta)d\rho}.$$

From (2.6) that

$$\begin{aligned} &\int_Z (y - f(x))^2 k(x,x_0;\beta)d\rho - \int_Z (y - f_\rho(x))^2 k(x,x_0;\beta)d\rho \\ &\quad = \int_X (f_\rho(x) - f(x))^2 k(x,x_0;\beta)d\rho_X. \end{aligned} \tag{4.1}$$

Recall the definition of f_{x_0} in (2.7). Taking $f = 0$ yields

$$\int_X (f_\rho(x) - f_{x_0}(x))^2 k(x, x_0; \beta) d\rho_X + \mathcal{K}(x_0, \beta) \|f_{X_0}\|_K^2$$

$$\leq \int_X f_\rho^2(x) k(x, x_0; \beta) d\rho_X \leq \|f_\rho\|_\rho^2.$$

Hence

$$\|f_{x_0}\|_K \leq \frac{\|f_\rho\|_\rho}{\sqrt{\mathcal{K}(x_0, \beta)}} \leq \frac{M}{\sqrt{\mathcal{K}(x_0, \beta)}} \tag{4.2}$$

and

$$\int_X (f_\rho(x) - f_{x_0}(x))^2 k(x, x_0; \beta) d\rho_X \leq \|f_\rho\|_\rho^2 \leq M^2.$$

It follows from (4.1) with $f = f_{x_0}$ that

$$\int_Z (y - f_{x_0}(x))^2 k(x, x_0; \beta) d\rho$$

$$= \int_Z (y - f_\rho(x))^2 k(x, x_0; \beta) d\rho + \int_Z (f_\rho(x) - f_{x_0}(x))^2 k(x, x_0; \beta) d\rho$$

$$\leq 2M^2.$$

Therefore, with confidence $1 - \delta$ we have

$$\triangle \leq \frac{2\kappa \left(M + \dfrac{\kappa M}{\sqrt{\mathcal{K}(x_0, \beta)}} \right) \log(2/\delta)}{m} + \kappa \sqrt{\frac{4M^2 \log(2/\delta)}{m}}$$

$$\leq \frac{2\kappa M \log(2/\delta)}{\sqrt{m}} \left(\frac{1}{\sqrt{m}} + \frac{\kappa}{\sqrt{m \mathcal{K}(x_0, \beta)}} + \frac{1}{\sqrt{\log(2/\delta)}} \right).$$

In fact, $\frac{1}{\sqrt{m}} \leq 1$ and $\frac{1}{\sqrt{\log(2/\delta)}} \leq 1$ since $0 < \delta < 1$ and $m \geq 1$.

If $\frac{\kappa}{\sqrt{m \mathcal{K}(x_0, \beta)}} \leq \frac{1}{3 \log(2/\delta)} < 1$, it is obviously that the desired bound holds.

When $\frac{\kappa}{\sqrt{m \mathcal{K}(x_0, \beta)}} > \frac{1}{3 \log(2/\delta)}$, we have $\frac{6\kappa M \log(2/\delta)}{\sqrt{m} \mathcal{K}(x_0, \beta)} \geq \frac{2M}{\sqrt{\mathcal{K}(x_0, \beta)}}$. In this case, we use (4.2) and the trivial bound $\|f_{z,x_0}\|_K \leq M/\sqrt{\mathcal{K}(x_0, \beta)}$ seen from (2.3) by taking $f = 0$. Then there holds

$$\|f_{z,x_0} - f_{x_0}\|_K \leq \|f_{z,x_0}\|_K + \|f_{x_0}\|_K \leq \frac{2M}{\sqrt{\mathcal{K}(x_0, \beta)}} \leq \frac{6\kappa M \log\left(\dfrac{2}{\delta}\right)}{\sqrt{m} \mathcal{K}(x_0, \beta)}$$

with probability 1. The desired inequality also holds in the second case. This proves Theorem 2. □

The error bound depends on the number of examples m, the local parameter β, the given point of interest and some prior information on the probability

distribution ρ. Notice the similarity between the bound $\frac{6\kappa M \log(2/\delta)}{\sqrt{m\lambda}}$ of RLS algorithm in Ref. 14 and the error estimate of the above theorem, we know that LRR algorithm can achieve the same convergence rate of $O(m^{-1/2})$ for $\|f_{z,x_0} - f_{x_0}\|_K$ as RLS algorithm does. Thus, future work will focus on how the decay of the parameters x_0 and β lead to convergence rates.

5. Experiments

What is left is to show the effectiveness of LRR. In this section, artificial problems are studied to evaluate the performance of the proposed algorithm firstly, then a series numerical experiments are conducted for classification on a broad range of datasets, including the partial UCI database (the UCI Machine Learning Repository) and the partial Benchmark database.[20] In the experiments, we compare the local risk minimization model with the RLS algorithm associated with Mercer kernels.

5.1. *Behavior of the LRR estimates on artificial problems*

In this section, we compare RLS and LRR on an artificial database of binary class problems in normal distribution.

Normal distribution is the most common distribution of the samples in real world problems. Many learning theory and algorithms are derived on the premise that the patterns follow the normal distribution because of the convenient properties. We present an artificial database in normal distribution. The database contains four datasets (see Fig. 1) in which each class contains 150 samples and the samples are generated randomly from the bivariate normal distribution. The means in the two classes are $[0, 0]$ and $[2, 2]$, respectively, and the variance is uniform diag$[0.5, 0.5]$. In these datasets, we stochastically select 30 samples in respective classes to combine the training set and take the remaining 240 samples as the testing set, as can be seen from Figs. 1 and 2.

The goal of our algorithm is to estimate the function in a vicinity of the point of interest while preserving most of the local information among the given local point x_0 and hence can lead to good classification results for the datasets. To get valid comparison with classical RLS algorithm, on each classification task, we use the same parameters of heat kernels as those used for LRR. The RLS algorithm determined the parameter $\lambda = 0.5$, and LRR determined $\beta = 6$, $\lambda = 0.5$. While for vicinity function, both "hard threshold" and "soft threshold" vicinity function are used in these experiments. The local point x_0 is chosen randomly in two classes.

In Table 1, LRR^1 represents the classifier constructed by LRR when x_0 is chosen in the first class, and LRR^2 represents that x_0 is in the second class. From this table we can observe that LRR^1 can classify the testing samples in the first class with higher accuracy than other classifiers, and LRR^2 can classify the testing samples in the second class with higher accuracy than other classifiers, which is correspondent

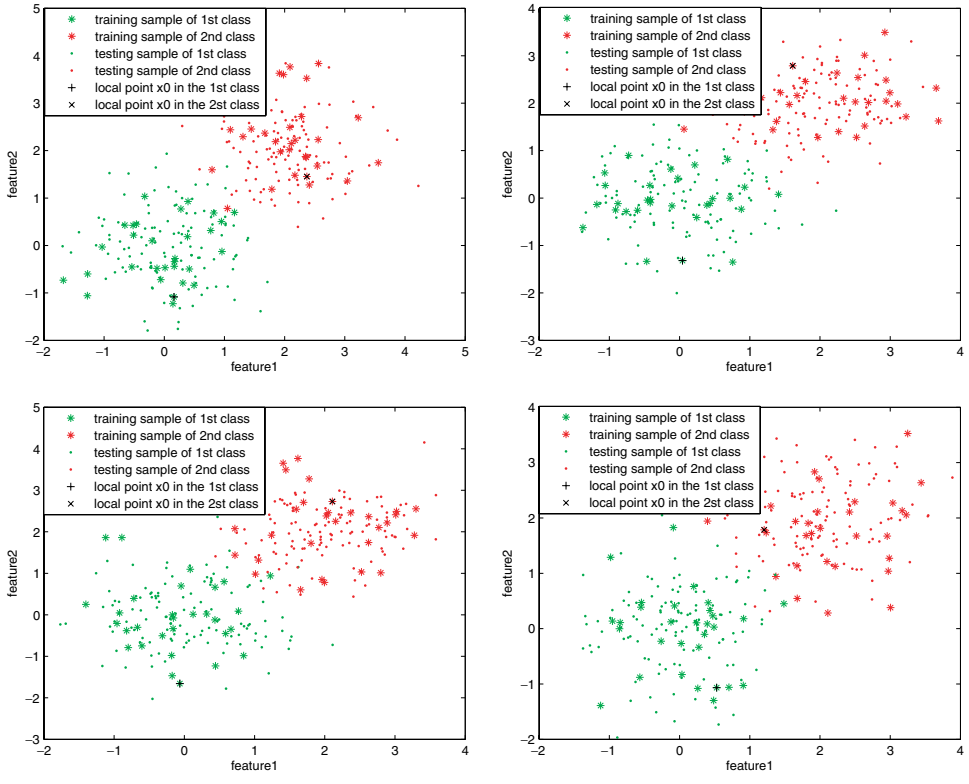

Fig. 1. The training and testing samples in the artificial datasets in normal distribution, the local point x_0 is chosen at random in both classes.

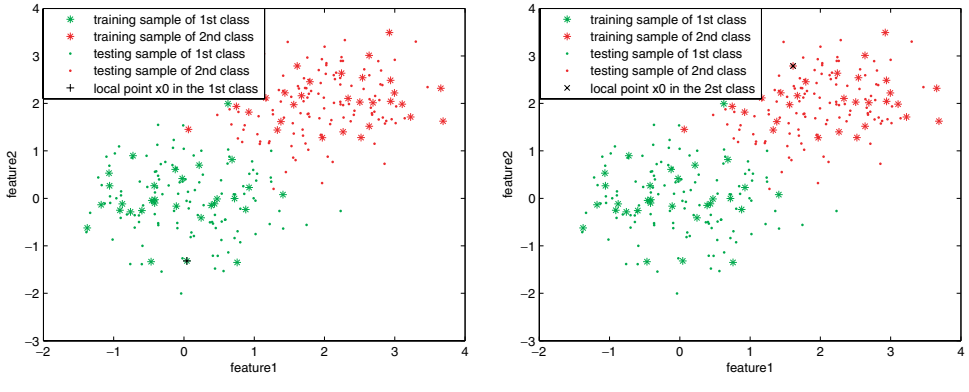

Fig. 2. The training and testing samples in the second artificial dataset in normal distribution, the local point x_0 is chosen at random in the first (left) and the second class (right). Both "hard threshold" and "soft threshold" vicinity function are used in this dataset.

Table 1. Testing accuracy of two classes in four datasets in normal distribution.

Class	LRR^1	LRR^2	RLS
1	**0.88333**	0.86667	0.84167
2	0.75833	**0.76670**	0.72500
1	**0.91667**	0.87500	0.89167
2	0.77500	**0.78333**	0.75833
1	**0.96667**	0.95833	0.91667
2	0.76667	**0.76667**	0.75000
1	**0.95833**	0.93333	0.91667
2	0.79167	**0.80000**	0.77500

Note: "1" ("2") denotes that x_0 is chosen in the 1st (2nd) class.

Table 2. Classification accuracy for the testing sets in four datasets in normal distribution.

Dataset	LRR^1	LRR^2	RLS	LRR
1	0.82083	0.81667	0.78333	**0.82500**
2	0.84583	0.82917	0.82500	**0.85000**
3	0.86667	0.86250	0.83333	**0.86667**
4	0.87500	0.86667	0.84583	**0.87917**

Note: "1" ("2") denotes that x_0 is chosen in the 1st (2nd) class.

with our local theory mentioned above. Consequently, in local learning scheme, we can construct two classifiers for two classes problems, one (LRR^1) with x_0 in the first class, the other (LRR^2) with x_0 in the second class. Then LRR^1 is used for classifying the testing samples in the first class, and LRR^2 for the second class (denoted by LRR in the table). The results in Table 2 indicate that the average accuracies of LRR^1, LRR^2 and LRR are all higher than that of RLS, and it is worth mentioning that the accuracies of LRR is significantly higher than that of RLS.

Table 3 shows the testing accuracies of the classical RLS algorithm and the proposed algorithm with "hard threshold" vicinity function (LRR^h) and "soft threshold" vicinity function (LRR^s). As we see from the table, the LRR estimate performs better than RLS no matter which vicinity function is used in LRR. Further comparison illustrates that the classification accuracies of LRR

Table 3. Classification accuracy for the testing set in the second dataset with "hard threshold" and "soft threshold" vicinity function.

	Testing accuracy		
x_0	LRR^h	LRR^s	RLS
Class 1	0.86250	**0.86667**	0.83333
Class 2	0.84167	**0.86250**	0.83333

with "soft threshold" vicinity function are higher than that of LRR with "hard threshold" vicinity function in all cases. This fact validates that "soft threshold" vicinity function can preserve local and global information perfectly, while "hard threshold" vicinity function eliminates some important global information, since $k_1(x, x_0; \beta) = 0$ for samples far away from the given point x_0. Therefore, the rest of the experiments are based on "soft threshold" vicinity function, and we consider the case in which only one point x_0 is given for simplicity.

5.2. *Experiments on partial Benchmark database*

For a second experiment, the Benchmark database is used in this test. These datasets all contain two classes. We use the training and testing sets offered by the database. Table 4 presents a brief description of these datasets, which are typical sets that the training and testing samples are sampled unevenly within classes. In this experiment, we compare LRR with RLS in different cases.

Experimental results in Table 5 indicate that LRR outperforms RLS consistently in almost all the datasets. Although LRR is superior to RLS in almost all the datasets, the gaps between the accuracies are relatively small in these datasets

Table 4. The dimension, training and testing set size of the nine datasets in the Benchmark database.

Dataset	Dimension	Training set size	Testing set size
B-cancer	9	200	77
Diabetes	8	468	300
F-Solar	9	666	400
German	20	700	300
Heart	13	170	100
Thyroid	5	140	75
Image	18	1,300	1,010
Titanic	3	150	2,051
Splice	60	1,000	2,175

Table 5. The average, training and testing accuracy of the nine datasets in the Benchmark database.

	Testing accuracy		Training accuracy		Average accuracy	
Dataset	LRR	RLS	LRR	RLS	LRR	RLS
B-cancer	**0.76623**	0.74026	0.98000	0.98000	**0.92058**	0.91336
Diabetes	**0.74000**	0.73000	0.95359	0.95359	**0.89453**	0.89063
F-Solar	**0.66500**	0.65500	0.70120	0.70270	**0.68762**	0.68480
German	**0.76333**	0.74333	1	1	**0.92900**	0.92300
Heart	**0.81000**	0.79000	1	1	**0.92963**	0.92222
Thyroid	0.96000	0.96000	0.99286	0.99286	0.98140	0.98140
Image	0.96337	**0.96436**	0.97846	0.97692	**0.97186**	0.97143
Titanic	0.77084	0.77084	0.80667	0.80667	0.77328	0.77328
Splice	**0.72046**	0.71724	1	1	0.80630	**0.80850**

except for B-cancer, owing to the relatively uneven sampling of the training and testing samples within classes. It is impressive that LRR nearly always achieve the higher accuracy no matter around which local point in the dataset. Moreover, we can compare the results in Ref. 20 with our results. It is clear that LRR achieves an ideal solution.

If we are interested in the estimation of function around x_0, we hope that points close to x_0 have great influence on the solution, points away from x_0 have little influence on the solution. To see how the given point x_0 of interest affects the accuracy, we compare the classification accuracies on the dataset with different points, as shown in Fig. 3. We can see that an "optimal" local point exists for a given task. It seems that although LRR can preserve locality properties of a vicinity of the local point x_0, it is little affected by the local point x_0 and is stable for a wide range of values. This is an advantage of LRR since in most cases, working in subspaces of input space is much easier than in the whole space.

5.3. *Experiments on partial UCI database*

To demonstrate the practical applicability of the proposed algorithm, we also choose four datasets, Iris, Ionosphere, Wdbc and Pid, in the UCI database as examples, where for multi-class problems, two classes are selected randomly for classification. As shown in Table 6, in the four classes, Iris is a small-scale one with the amounts of samples between classes evenly, Ionosphere and Wdbc are a bit large-scale and more uneven, and Pid is more large-scale sets and the samples in each class distribute more unevenly.

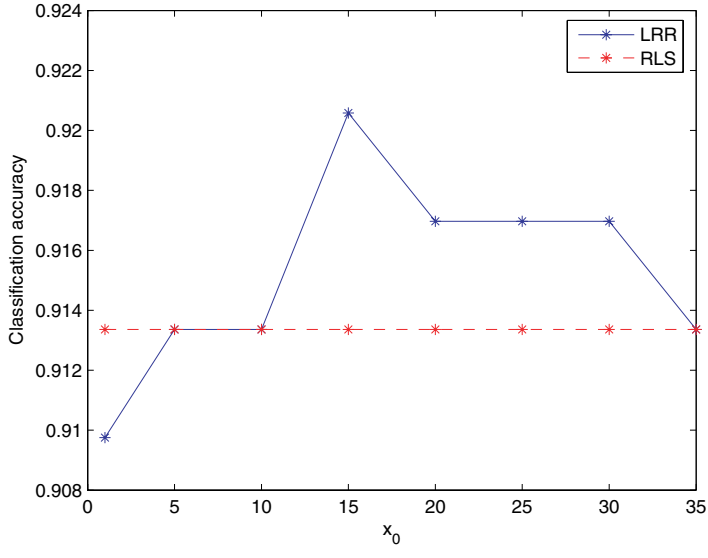

Fig. 3. Shows classification accuracies with different point of interest on the B-cancer dataset with x_0 from the first point to the 35th one.

Table 6. The dimension and the respective class size of the four datasets in the UCI database.

Datasets	Dimension	Class I size	Class II size
Iris	4	50	50
Ionosphere	34	225	126
Wdbc	30	212	357
Pid	8	500	268

Table 7. The training and testing size of the four datasets in the UCI database

Dataset	Training set size	Testing set size
Iris	45	55
Ionosphere	185	166
Wdbc	336	233
Pid	430	338

As can be seen in Table 7, we mix all the samples of two classes and stochastically select almost half of samples as the training set. The remaining samples are taken as the testing set.

To the best of our knowledge, the classification accuracy of LRR is much more sensitive to the value of β. It is vital to choose a suitable value of β, which can be shown by the "soft threshold" vicinity function

$$k_2(x, x_0; \beta) = \exp\left\{-\frac{(x - x_0)^2}{\beta^2}\right\}$$

we choose in our experiments. To emphasize the practical nature of this result, a series of numerical classification experiments are run based on the four datasets in the UCI database.

In the experiments, we do with the following heuristic argument: simplifying to the case when the value of β is chosen to be far greater than the distance between x_0 and other training points, note that $k_2(x, x_0; \beta)$ will be approximately equal to 1. Hence, the proposed algorithm performs very similarly to RLS and the accuracy of the two algorithms are almost the same, as can be seen from the solutions of the two algorithms. On the other hand, when the value of β is far smaller than the distance between x_0 and other training points, $k_2(x, x_0; \beta)$ will be approximately equal to 0. The first term in (2.3) vanishes, LRR cannot reach the ideal performance therefore.

The first observation is that the proposed algorithm performs extremely well in these experiments. In all of our experiments, we consider the case where the training and testing samples are selected randomly, and β is chosen randomly, requiring only not too large and not too small. Experimental results in Table 8 indicate that the accuracy of LRR is always higher than that of RLS. Furthermore, the difference between LRR and RLS is significant. By analyzing and contrasting the

Table 8. Classification accuracy for the testing set in the four datasets.

	Testing accuracy	
Dataset	LRR	RLS
Iris	**0.78182**	0.62338
Ionosphere	**0.84940**	0.74699
Wdbc	**0.86667**	0.60944
Pid	**0.70118**	0.67456

efficiency and effect of the two algorithms, we note that our estimation procedures are remarkably accurate as long as the value of the parameter β is chosen suitably.

6. Conclusions

In this paper, we first improve the local risk minimization scheme with a simplified development, and define the local sampling operator and the local integral operator. Different from traditional regularization methods, LRR can preserve the locality properties since it considers a nonnegative function $k(x, x_0; \beta)$ that embodies the concept of vicinity. Then the defined operators are applied to the proposed learning scheme for function reconstruction, and the integral theory is generalized into this local method. In particular, we show that a regression function can be approximated by a local risk minimization scheme f_{z,x_0} in $\mathcal{H}_K$, and the error estimate for the proposed algorithm has been proven by using probabilistic estimates for integral operators. The experimental results demonstrate that LRR performs better than RLS. We provide a simpler approach with the sample error in this paper. Future work will focus on how the decay of the parameters x_0 and β lead to convergence rates, and since the previous works have focused on independent samples, while independence is a restrictive assumption and may be violated in many real data analysis. In the future, we can study the learning performance of the proposed learning scheme (2.3) for dependent samples.

Acknowledgments

The research is supported in part by NSFC under grant No. 10771053 and by Natural Science Foundation of Hubei Province under grant No. 2009CDB387.

References

1. F. Bauer, S. Pereverzev and L. Rosasco, On regularization algorithms in learning theory, *J. Complexity* **23** (2007) 52–72.
2. L. Bottou and V. Vapnik, Local learning algorithms, *Neural Comput.* **4** (1992) 888–900.
3. A. Caponnetto and E. De Vito, Optimal rates for regularization least-square algorithm, *Found. Comput. Math.* **7** (2007) 331–368.

4. F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint* (Cambridge University Press, 2007).
5. F. Cucker and S. Smale, On the mathematical foundation of learning, *Bull. Amer. Soc.* **39** (2001) 1–49.
6. F. Cucker and S. Smale, Best choices for regularization parameters in learning theory: On the bias-variance problem, *Found. Comput. Math.* **2**(4) (2002) 413–428.
7. H. Chen, L. Q. Li and Y. Y. Tang, Analysis of classification with a reject option, *Int. J. Wavelets Multiresolut. Inf. Process.* **7**(3) (2009) 375–385.
8. Y. Chen, Z. C. Ji and C. J. Hua, Efficient statistical modeling of wavelet coefficients for image denoising, *Int. J. Wavelets Multiresolut. Inf. Process.* **7**(5) (2009) 629–641.
9. L. Györfi, M. Kohler, A. Krzyzak and H. Walk, *A Distribution-Free Theory of Nonparametric Regression* (Springer-Verlag, 2002).
10. L. Q. Li, Regularized least square regression with spherical polynomial kernels, *Int. J. Wavelets Multiresolut. Inf. Process.* **7**(6) (2009) 781–801.
11. A. Rakotomamonjy and S. Canu, Frames, reproducing kernel, regularization and learning, *J. Mach. Learn. Res.* **6** (2005) 1485–1515.
12. S. Smale and D. X. Zhou, Shannon sampling and function reconstruction from point values, *Bull. Amer. Math. Soc.* **41** (2004) 279–305.
13. S. Smale and D. X. Zhou, Shannon sampling II. Connections to learning theory, *Appl. Comput. Harmon. Anal.* **19** (2005) 285–302.
14. S. Smale and D. X. Zhou, Learning theory estimates via integral operators and approximations, *Constr. Approx.* **26** (2007) 153–172.
15. A. Tikhonov and V. Arsenin, *Solutions of Ill-Posed Problems* (W. H. Winston, 1977).
16. E. De Vito, A. Caponnetto and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, *Found. Comput. Math.* **5**(1) (2005) 59–85.
17. V. Vapnik, Principles of risk minimization for learning theory, in *Advances in Neural Information Processing Systems*, Vol. 4 (Morgan Kaufmann, 1992), pp. 831–838.
18. V. Vapnik, *Statistical Learning Theory* (John Wiley and Sons, 1998).
19. Y. T. Wei, H. Li and L. Q. Li, Tensor locality sensitive discriminant analysis and its complexity, *Int. J. Wavelets Multiresolut. Inf. Process.* **7**(6) (2009) 865–880.
20. H. Xue, S. C. Chen and X. Q. Zeng, Classifier learning with a new locality regularization method, *Pattern Recogn.* **41** (2008) 1479–1490.
21. Y. L. Xu and D. R. Chen, Learning rates of regularized regression for functional data, *Int. J. Wavelets Multiresolut. Inf. Process.* **7**(6) (2009) 839–850.
22. D. X. Zhou, O. Bousquet, T. N. Lal, J. Weston and B. Scholkopf, Learning with local and global consistency, in *Advances in Neural Information Processing Systems 16*, eds. S. Thrun, L. Saul and B. Scholkopf (MIT Press, 2004), pp. 321–328.