TAIWANESE JOURNAL OF MATHEMATICS
Vol. 14, No. 5, pp. 1821-1836, October 2010
This paper is available online at http://www.tjm.nsysu.edu.tw/

LEARNING BY NONSYMMETRIC KERNELS WITH DATA DEPENDENT SPACES AND ℓ^1-REGULARIZER

Quan-Wu Xiao and Ding-Xuan Zhou

Abstract. We study a learning algorithm for regression. The algorithm is a regularization scheme with ℓ^1 regularizer stated in a hypothesis space trained from data or samples by a nonsymmetric kernel. The data dependent nature of the algorithm leads to an extra error term called hypothesis error, which is essentially different from regularization schemes with data independent hypothesis spaces. By dealing with regularization error, sample error and hypothesis error, we estimate the total error in terms of properties of the kernel, the input space, the marginal distribution, and the regression function of the regression problem. Learning rates are derived by choosing suitable values of the regularization parameter. An improved error decomposition approach is used in our data dependent setting.

1. INTRODUCTION

In a regression problem, we work with an input metric space (X, d) and an output space $Y = \mathbb{R}$. A function $f : X \to Y$ makes a prediction of the output $y \in Y$ at $x \in X$ by $f(x)$. The prediction accuracy may be measured by the least-square loss $(f(x) - y)^2$. Let ρ be a probability measure on $Z := X \times Y$. The prediction ability of f is quantitatively measured by the *generalization error*

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

Received September 19, 2008, accepted February 17, 2009.
Communicated by Sen-Yen Shaw.
2000 *Mathematics Subject Classification*: 68T05, 62J02.
Key words and phrases: Learning theory, Nonsymmetric kernel, Data dependent hypothesis spaces, Regularization scheme, Error analysis.
The work described in this paper is supported partially by the Research Grants Council of Hong Kong (Project No. CityU 103206), National Science Fund for Distinguished Young Scholars of China (Project No. 10529101), and National Basic Research Program of China (Project No. 973-2006CB303102).

Decompose ρ into the marginal distribution ρ_X on X and the conditional distributions $\rho(y|x)$ at $x \in X$. The function minimizing $\mathcal{E}(f)$ is called the *regression function* given by

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

Since ρ is usually unknown, f_ρ cannot be obtained directly. We can learn f_ρ from samples. Throughout the paper we assume that a sample $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^m$ of size m is drawn independently according to the measure ρ.

Kernel method is an important tool in learning theory. A well studied kernel-based algorithm for the regression problem is the least-square regularization scheme. If $K : X \times X \to \mathbb{R}$ is a continuous positive semi-definite kernel and $(\mathcal{H}_K, \|\cdot\|_K)$ is the associated reproducing kernel Hilbert space [1], then the scheme is given by

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|_K^2 \right\}, \tag{1.1}$$

where $\mathcal{E}_{\mathbf{z}}(f)$ is the *empirical error*

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

and $\lambda > 0$ is a regularization parameter. The hypothesis space $\mathcal{H}_K$ is data independent. Mathematical analysis of learning algorithm (1.1) has been well understood [4, 13, 7, 8].

In this paper we abandon the symmetry (and of course positive semi-definiteness) of the kernel and consider a regularization scheme with ℓ^1-regularizer which is a learning algorithm associated with data dependent hypothesis spaces [9]. Here a kernel function $K : X \times X \to \mathbb{R}$ is a continuous function. The hypothesis space depends on the sample $\mathbf{z}$ and is defined by

$$\mathcal{F}_{\mathbf{z}} = \left\{ \sum_{i=1}^m \alpha_i K_{x_i} : \alpha_i \in \mathbb{R} \right\}, \tag{1.2}$$

where $K_t(\cdot) = K(\cdot, t)$. The learning algorithm is given by

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{F}_{\mathbf{z}}} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda \Omega_{\mathbf{z}}(f) \right\}, \tag{1.3}$$

where

$$\Omega_{\mathbf{z}}(f) = \sum_{i=1}^m |\alpha_i| \qquad \text{for} \quad f = \sum_{i=1}^m \alpha_i K_{x_i}.$$

Example 1. Let φ and $\widetilde{\varphi}$ be two continuous functions on $\mathbb{R}^n$ bounded by $C_0(1+|x|)^{-(n+1)/2}$ with some constant C_0. For $s > n/2$, the kernel

$$\Phi(x,t) = \sum_{j=0}^{\infty} 2^{j(n-2s)} \sum_{k\in\mathbb{Z}^n} \varphi(2^j x - k)\widetilde{\varphi}(2^j t - k), \qquad x,t \in \mathbb{R}^n$$

is applicable to algorithm (1.3) for any $X \subset \mathbb{R}^n$. This nonsymmetric kernel appears naturally in the study of dual wavelets or frames in wavelet analysis [6, 11]. It has the flexibility of having good representation for $f_{\mathbf{z},\lambda}$ while keeping strong approximation ability.

The ℓ^1-regularizer often leads to some sparse properties, as shown in [5], which will be discussed for algorithm (1.3) somewhere else.

In this article, we mainly consider how fast $f_{\mathbf{z},\lambda}$ approximates f_ρ as m increases. Learning rates will be given in terms of properties of the input space X, the measure ρ, and the kernel K.

Definition 1. The *covering number* $\mathcal{N}(X,r)$ of the metric space X is the minimal $l \in \mathbb{N}$ such that there exist l open balls in X with radius r covering X.

Covering numbers are used to describe the complexity of X. We shall assume

$$\mathcal{N}(X,r) \le C_\eta \left(\frac{1}{r}\right)^\eta \qquad \forall 0 < r \le 1 \tag{1.4}$$

for some $\eta > 0$ and $C_\eta > 0$.

Definition 2. A probability measure ρ_X on X is said to satisfy condition L_τ with $0 < \tau < \infty$ if there exists some $C_\tau > 0$ such that for any ball $B(x,r) = \{u \in X : d(u,x) < r\}$, we have

$$\rho_X\left(B(x,r)\right) \ge C_\tau r^\tau \qquad \forall x \in X, 0 < r \le 1. \tag{1.5}$$

Remark 1. When $X \subset \mathbb{R}^n$, condition (1.4) is valid with $\eta = n$. If moreover, X satisfies a cone condition given in [2] and ρ is the uniform distribution on X, then (1.5) holds with $\tau = n$, and C_τ depends on X.

Definition 3. We say that the kernel K satisfies a Lipschitz condition of order (α,β) with $0 < \alpha, \beta \le 1$ if for some $C_\alpha, C_\beta > 0$, we have

$$|K(x,t) - K(x,t')| \le C_\alpha (d(t,t'))^\alpha, \qquad \forall x,t,t' \in X \tag{1.6}$$

and

$$|K(x,t) - K(x',t)| \le C_\beta (d(x,x'))^\beta, \qquad \forall t,x,x' \in X. \tag{1.7}$$

The kernel K defines an integral operator $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ by

$$L_K f(x) = \int_X K(x,t) f(t) d\rho_X(t), \quad x \in X.$$

Since X is compact and K is continuous, L_K and its dual L_K^T are compact operators, and $L_K L_K^T : L^2_{\rho_X} \to L^2_{\rho_X}$ is a self-adjoint positive operator with decreasing eigenvalues $\{\lambda_k^2\}_{k=1}^\infty$ (with $\lambda_k \geq 0$) and eigenfunctions $\{\phi_k\}_{k=1}^\infty$ forming an orthonormal basis of $L^2_{\rho_X}$.

Define $|L_K|^s = (L_K L_K^T)^{\frac{s}{2}}$ to be the operator on $L^2_{\rho_X}$ given by

$$|L_K|^s (\sum_{k=1}^\infty c_k \phi_k) = \sum_{k=1}^\infty c_k \lambda_k^s \phi_k, \quad \{c_k\}_k \in \ell^2.$$

We shall assume a regularity condition that f_ρ lies in the range of $|L_K|^s$ for some $s > 0$.

Now we can state our main result. Throughout the paper we assume $|y| \leq M$ almost surely.

Theorem 1. *Suppose X satisfies (1.4) with $\eta > 0$, the kernel K satisfies a Lipschitz condition of order (α, β) with $0 < \alpha, \beta \leq 1$, ρ_X satisfies condition L_τ with $\tau > 0$, and f_ρ lies in the range of $|L_K|^s$ for some $0 < s \leq 2$. Let $\lambda = m^{-\theta}$ with $\theta > 0$. Denote*

$$\Theta = \min\left\{\frac{1}{1+\eta/\beta} - 2\theta, 1 - \frac{2\theta(2-s)}{s+2}, \frac{1}{2} - \frac{2\theta(1-s)}{s+2}, \frac{\alpha}{\tau} - \frac{2\theta(2-s)}{s+2}, \frac{2\theta s}{s+2}\right\}.$$

Then for any $0 < \delta < 1$, with confidence $1 - \delta$ it holds that

$$\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} \leq C \left(\log\frac{4}{\delta} + \log(m+1)\right)^{\max\left\{1, \frac{\alpha}{\tau}\right\}} m^{-\Theta}.$$

where C is a constant independent of m or δ.

The proof of Theorem 1 will be given in Section 6 where the constant C is given explicitly. Note that $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}}$ for any measurable function f.

A special case of Theorem 1 is the following learning rate when K is Lipschitz on $X \times X$ ($\alpha = \beta = 1$) and f_ρ lies in the range of $L_K L_K^T$ ($s = 2$).

Corollary 1. *Assume K is Lipschitz and f_ρ lies in the range of $L_K L_K^T$. Suppose X satisfies (1.4) with $\eta > 0$ and ρ_X satisfies condition L_τ with $\tau \geq 1$. Then with confidence $1 - \delta$,*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} \leq C \left(\log\frac{4}{\delta} + \log(m+1)\right) m^{-\Theta},$$

where

$$\Theta = \begin{cases} \dfrac{1}{3(1+\eta)}, & \textit{if } \tau \leq 3(1+\eta), \lambda = m^{-\frac{1}{3(1+\eta)}}, \\ \dfrac{1}{\tau}, & \textit{if } \tau > 3(1+\eta), \lambda = m^{-\theta} \textit{ with } \dfrac{1}{\tau} \leq \theta \leq \dfrac{\tau-(1+\eta)}{2\tau(1+\eta)}. \end{cases}$$

Remark 2. By restricting to the support of ρ_X, condition (1.4) with $\tau < \infty$ is a reasonable assumption. The index τ measures the degree of uniformality of the distribution ρ_X on X. When $\eta = n$ and $\tau = n$, we see that the learning rate in Corollary 1 is $O(m^{-\frac{1}{3(1+n)}})$ which is very low. This is mainly due to an error term called hypothesis error below, caused by the data dependent nature of algorithm (1.3). The estimate for this error term we obtain in Section 4 is based on the Lipschitz-α regularity of the kernel K, which might be improved when higher order regularities of K are imposed (as for bounding covering numbers in [14] and estimating local approximation error for scattered data interpolation [10]). This is an interesting topic for further study.

2. Error Decomposition

A useful approach for getting learning rates for regularization schemes with sample independent hypothesis spaces is error decomposition [8] which decomposes the total error $\|f_{\mathbf{z},\lambda} - f_\rho\|_{L^2_{\rho_X}}$ into the sum of a sample error and a regularization error (or approximation error). The main difficulty with algorithm (1.3) is the dependence of the hypothesis space $\mathcal{F}_{\mathbf{z}}$ on the data $\mathbf{z}$. This was pointed out in [9] where a modified error decomposition technique is introduced by means of an extra hypothesis error. Our setting here is more general than that in [9] because the kernel K here is not necessarily symmetric. Our purpose is to complete the error analysis of algorithm (1.3) in this more general setting. Estimates for the regularization error and sample error are new while key ideas for bounding the hypothesis error are from [9].

We consider the Banach space $\mathcal{F}_0$ consisting of all functions of this form

$$f = \sum_{j=1}^{\infty} \alpha_j K_{x_j}, \qquad \{\alpha_j\} \in \ell^1, \ \{x_j\} \subset X$$

with the norm

$$\|f\| = \inf \left\{ \sum_{j=1}^{\infty} |\alpha_j| : f = \sum_{j=1}^{\infty} \alpha_j K_{x_j} \right\}.$$

Since X is compact, $\mathcal{F}_0$ can be regarded as a subset of $C(X)$ with the inclusion map $I : \mathcal{F}_0 \to C(X)$ bounded as

$$\|f\|_\infty \le \kappa \|f\| \qquad \forall f \in \mathcal{F}_0 \tag{2.1}$$

with $\kappa = \|K\|_{C(X\times X)}$. Note that $\mathcal{F}_{\mathbf{z}} \subset \mathcal{F}_0$ for any $\mathbf{z} \in Z^m$.

To formulate the error decomposition for algorithm (1.3), we introduce a regularizing function as

$$f_\lambda = \arg\min_{f\in\mathcal{F}_0} \left\{\mathcal{E}(f) + \lambda\|f\|\right\}. \tag{2.2}$$

We can always replace f_λ by a sequence of approximating functions in our analysis if a minimizer of (2.2) does not exist.

Definition 4. The *sample error* for algorithm (1.3) is defined as

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

The *hypothesis error* takes the form

$$\mathcal{P}(\mathbf{z},\lambda) = \left\{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})\right\} - \left\{\mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\|f_\lambda\|\right\},$$

while the *regularization error* is given by

$$\mathcal{D}(\lambda) = \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \lambda\|f_\lambda\| = \inf_{f\in\mathcal{F}_0}\left\{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\|f\|\right\}.$$

Then we have the following error decomposition.

Lemma 1. *Let $f_{\mathbf{z},\lambda}$ be defined by (1.3) with $\lambda > 0$. Then*

$$\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} \le \mathcal{S}(\mathbf{z},\lambda) + \mathcal{P}(\mathbf{z},\lambda) + \mathcal{D}(\lambda). \tag{2.3}$$

Proof. A simple computation shows that

$$\mathcal{S}(\mathbf{z},\lambda) + \mathcal{P}(\mathbf{z},\lambda) + \mathcal{D}(\lambda) = \mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}).$$

But $\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \ge 0$. So the desired bound (2.3) follows from the identity $\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho) = \|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}}$. ∎

3. Estimating the Regularization Error

Since K is not assumed to be symmetric, the regularization error needs to be bounded in a way different from that for positive definite kernels [7, 8].

Lemma 2. *Let λ_k^2 be the positive eigenvalues of $L_K L_K^T$, and ϕ_k be the corresponding normalized eigenfunctions in $L^2_{\rho_X}$. Then,*

$$\sum_k \lambda_k^2 \leq \kappa^2 \quad \textit{and} \quad \|\phi_k\| \leq \frac{1}{\lambda_k}.$$

Proof. Define a kernel $\tilde{K} : X \times X \to \mathbb{R}$ by

$$\tilde{K}(u,v) = \int_X K(u,x)K(v,x)d\rho_X(x).$$

It is easy to verify that $\tilde{K}$ is a Mercer kernel with $\|\tilde{K}\|_{C(X\times X)} \leq \kappa^2$, and $L_{\tilde{K}} = L_K L_K^T$.

By Mercer's Theorem (e.g. [3]) we know that

$$\sum_k \lambda_k^2 = \sum_k \lambda_k^2 \int_X \phi_k(x)^2 d\rho_X(x) = \int_X \tilde{K}(x,x)d\rho_X(x) \leq \kappa^2.$$

Observe that

$$\phi_k = \frac{1}{\lambda_k^2} L_K L_K^T \phi_k = \frac{1}{\lambda_k^2} L_{\tilde{K}} \phi_k = \frac{1}{\lambda_k^2} \int_X \int_X K(\cdot,x)K(v,x)\phi_k(v)d\rho_X(v)d\rho_X(x).$$

Then ϕ_k can be written as

$$\phi_k = \int_X \left\{ \frac{1}{\lambda_k^2} \int_X K(v,x)\phi_k(v)d\rho_X(v) \right\} K_x d\rho_X(x),$$

a linear combination of the functions $K_x (x \in X)$ with coefficients $\int_X K(v,x)\phi_k(v)$ $d\rho_X(v)$. So by the definition of the norm $\|\cdot\|$ we have

$$\|\phi_k\| \leq \int_X \left| \frac{1}{\lambda_k^2} \int_X K(v,x)\phi_k(v)d\rho_X(v) \right| d\rho_X(x) = \frac{1}{\lambda_k^2} \int_X \left| L_K^T \phi_k(x) \right| d\rho_X(x).$$

By the Schwarz inequality,

$$\|\phi_k\| \leq \frac{1}{\lambda_k^2} \|L_K^T \phi_k\|_{L^2_{\rho_X}} = \frac{1}{\lambda_k^2} \sqrt{\langle L_K L_K^T \phi_k, \phi_k \rangle_{L^2_{\rho_X}}} = \frac{1}{\lambda_k}.$$

This proves the desired bounds. ∎

The first inequality above can be easily seen from the trace of the integral operator $L_{\tilde{K}}$ associated with the symmetric kernel $\tilde{K}$, while the second inequality cannot since the norm $\|\cdot\|$ is different from $\|\cdot\|_{\tilde{K}}$.

The regularization error $\mathcal{D}(\lambda)$ can now be bounded as follows.

Proposition 1. *If* $f_\rho = |L_K|^s g$ *for some* $0 < s \leq 2$ *and* $g \in L^2_{\rho_X}$, *then*

$$\mathcal{D}(\lambda) \leq C_1 \lambda^{\frac{2s}{s+2}} \qquad \forall \lambda > 0, \tag{3.1}$$

where $C_1 = \|g\|^2_{L^2_{\rho_X}} + \kappa \|g\|_{L^2_{\rho_X}}$.

Proof. By annihilating eigenfunctions with zero eigenvalues, we may write $g = \sum_{\lambda_k > 0} a_k \phi_k$. Then $\|g\|^2_{L^2_{\rho_X}} = \sum_{\lambda_k>0} a_k^2 < \infty$ and $f_\rho = \sum_{\lambda_k>0} a_k \lambda_k^s \phi_k$.

If $0 < \lambda \leq \lambda_1^{s+2}$, then there exists some $N \in \mathbb{N}$ such that $\lambda_{N+1} < \lambda^{\frac{1}{s+2}} \leq \lambda_N$. Choose $f = \sum_{k=1}^N a_k \lambda_k^s \phi_k$. For $1 \leq k \leq N$, we have $\lambda_k \geq \lambda_N \geq \lambda^{\frac{1}{s+2}}$. So by Lemma 2 and the Schwarz inequality we obtain

$$\begin{aligned} \|f\| &\leq \sum_{k=1}^N |a_k| \lambda_k^s \|\phi_k\| \leq \sum_{k=1}^N |a_k| \lambda_k^{s-1} \\ &= \sum_{k=1}^N |a_k| \lambda_k^{s-2} \lambda_k \leq \lambda^{\frac{s-2}{s+2}} \sum_{k=1}^N |a_k| \lambda_k \leq \kappa \|g\|_{L^2_{\rho_X}} \lambda^{\frac{s-2}{s+2}}. \end{aligned}$$

On the other hand,

$$\|f - f_\rho\|^2_{L^2_{\rho_X}} = \left\| \sum_{k>N} a_k \lambda_k^s \phi_k \right\|^2_{L^2_{\rho_X}} = \sum_{k>N} a_k^2 \lambda_k^{2s} \leq \lambda^{\frac{2s}{s+2}} \|g\|^2_{L^2_{\rho_X}}.$$

Then

$$\mathcal{D}(\lambda) \leq \|f - f_\rho\|^2_{L^2_{\rho_X}} + \lambda \|f\| \leq \left(\|g\|^2_{L^2_{\rho_X}} + \kappa \|g\|_{L^2_{\rho_X}} \right) \lambda^{\frac{2s}{s+2}}. \tag{3.2}$$

If $\lambda > \lambda_1^{s+2}$, by taking $f = 0 \in \mathcal{F}_0$ we still have

$$\mathcal{D}(\lambda) \leq \|f_\rho\|^2_{L^2_{\rho_X}} = \sum_{\lambda_k>0} a_k^2 \lambda_k^{2s} \leq \sum_{\lambda_k>0} a_k^2 \lambda_1^{2s} \leq \|g\|^2_{L^2_{\rho_X}} \lambda^{\frac{2s}{s+2}}.$$

This in connection with (3.2) tells us that (3.1) holds true. ∎

Notice from Proposition 1 that if $f_\rho = |L_K|^s g$ for some $0 < s \leq 2$ and $g \in L^2_{\rho_X}$, then

$$\|f_\lambda\| \leq C_1 \lambda^{\frac{s-2}{s+2}} \qquad \forall \lambda > 0. \tag{3.3}$$

4. Estimating the Hypothesis Error

In this section we bound the hypothesis error $\mathcal{P}(\mathbf{z}, \lambda)$ by using ideas of Proposition 11 and Theorem 9 in [9].

Definition 5. A point set $\{x_1, \ldots, x_m\} \subset X$ is said to be Δ-dense if for every $x \in X$ there exists some $1 \le i \le m$ such that $d(x, x_i) \le \Delta$.

Lemma 3. *If ρ_X satisfies condition L_τ with $\tau > 0$, and $\{x_i\}_{i=1}^m$ is a sample independently drawn from ρ_X, then for any $0 < \delta < 1$, with confidence $1 - \frac{\delta}{2}$, $\{x_i\}_{i=1}^m$ is Δ-dense provided that $\Delta > 0$ satisfies*

$$\log \mathcal{N}(X, \frac{\Delta}{2}) - \frac{mC_\tau}{2^\tau}\Delta^\tau \le \log \frac{\delta}{2}. \tag{4.1}$$

Proof. Let $\{B_j, j = 1, \ldots, \mathcal{N} = \mathcal{N}(X, \frac{\Delta}{2})\}$ be balls with radius $\frac{\Delta}{2}$ covering X. By the definition of condition L_τ, $\rho_X(B_j) \ge C_\tau \left(\frac{\Delta}{2}\right)^\tau$ holds for each j. Hence the probability for the event $\{x_i\}_{i=1}^m \bigcap B_j = \emptyset$ is at most $\left(1 - C_\tau \left(\frac{\Delta}{2}\right)^\tau\right)^m$. So the probability for $\{x_i\}_{i=1}^m \bigcap B_j = \emptyset$ to be true for at least one $j \in \{1, \ldots, m\}$ is at most

$$\mathcal{N}\left(1 - C_\tau \left(\frac{\Delta}{2}\right)^\tau\right)^m \le \mathcal{N} \exp\left\{-mC_\tau \left(\frac{\Delta}{2}\right)^\tau\right\}.$$

It follows that with confidence at least $1 - \mathcal{N} \exp\left\{-mC_\tau \left(\frac{\Delta}{2}\right)^\tau\right\}$, none of the events $\{x_i\}_{i=1}^m \bigcap B_j = \emptyset$ with $j = 1, \ldots, \mathcal{N}$ happens. That means, each ball B_j contains at least one sample point, which implies that $\{x_i\}_{i=1}^m$ is Δ-dense in X. This proves our conclusion. ∎

Lemma 4. *If $\{x_i\}_{i=1}^m$ is Δ-dense in X, f_ρ lies in the range of $|L_K|^s$ for some $0 < s \le 2$, and the kernel K satisfies (1.6), then*

$$\mathcal{P}(\mathbf{z}, \lambda) \le 2C_\alpha \left(C_1^2 \kappa \lambda^{\frac{2(s-2)}{s+2}} + C_1 M \lambda^{\frac{s-2}{s+2}}\right)\Delta^\alpha.$$

Proof. Since $f_\lambda \in \mathcal{F}_0$ satisfies $\|f_\lambda\| \le C_1 \lambda^{\frac{s-2}{s+2}}$ by (3.3), for any $\iota > 0$, it can be written as $f_\lambda = \sum_{j=1}^\infty \beta_j K_{t_j}$ with $t_j \in X$ and

$$\|f_\lambda\| \le \sum_{j=1}^\infty |\beta_j| \le \|f_\lambda\| + \iota \le C_1 \lambda^{\frac{s-2}{s+2}} + \iota. \tag{4.2}$$

Then there exists some $N_0 \in \mathbb{N}$ such that $\sum_{j=N_0+1}^{\infty} |\beta_j| \leq \iota$ and

$$\left\| \sum_{j=1}^{N_0} \beta_j K_{t_j} - f_\lambda \right\|_\infty \leq \kappa \left\| \sum_{j=N_0+1}^{\infty} \beta_j K_{t_j} \right\| \leq \kappa\iota. \tag{4.3}$$

Since $\{x_i\}_{i=1}^m$ is Δ-dense in X, for every t_j, there exists some $x(t_j) \in \{x_i\}_{i=1}^m$ such that $d(x(t_j), t_j) \leq \Delta$. Then from (1.6) and (4.2) we have

$$\left\| \sum_{j=1}^{N_0} \beta_j K_{x(t_j)} - \sum_{j=1}^{N_0} \beta_j K_{t_j} \right\|_\infty \leq C_\alpha \sum_{j=1}^{N_0} |\beta_j| \Delta^\alpha \leq C_\alpha \left(C_1 \lambda^{\frac{s-2}{s+2}} + \iota \right) \Delta^\alpha.$$

Combining with (4.3), we have

$$\left\| \sum_{j=1}^{N_0} \beta_j K_{x(t_j)} - f_\lambda \right\|_\infty \leq \kappa\iota + C_\alpha \left(C_1 \lambda^{\frac{s-2}{s+2}} + \iota \right) \Delta^\alpha.$$

For any $f_1, f_2 \in L^\infty(X)$ and $(x, y) \in Z$, it holds almost surely

$$|(f_1(x) - y)^2 - (f_2(x) - y)^2| \leq (\|f_1\|_\infty + \|f_2\|_\infty + 2M)\|f_1 - f_2\|_\infty.$$

Since both $L^\infty(X)$ norms of $\sum_{j=1}^{N_0} \beta_j K_{x(t_j)}$ and f_λ are bounded by $\kappa \left(C_1 \lambda^{\frac{s-2}{s+2}} + \iota \right)$, we have

$$\begin{aligned} &\left| \mathcal{E}_{\mathbf{z}} \left(\sum_{j=1}^{N_0} \beta_j K_{x(t_j)} \right) - \mathcal{E}_{\mathbf{z}} (f_\lambda) \right| \\ &\leq 2 \left(\kappa C_1 \lambda^{\frac{s-2}{s+2}} + \kappa\iota + M \right) \left(\kappa\iota + C_\alpha \left(C_1 \lambda^{\frac{s-2}{s+2}} + \iota \right) \Delta^\alpha \right). \end{aligned}$$

Notice that $\sum_{j=1}^{N_0} \beta_j K_{x(t_j)} \in \mathcal{F}_{\mathbf{z}}$. By the definition of $f_{\mathbf{z},\lambda}$, we see that $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \leq \mathcal{E}_{\mathbf{z}} \left(\sum_{j=1}^{N_0} \beta_j K_{x(t_j)} \right) + \lambda \sum_{i=j}^{N_0} |\beta_j|$ can be bounded by

$$\mathcal{E}_{\mathbf{z}}(f_\lambda) + 2 \left(\kappa C_1 \lambda^{\frac{s-2}{s+2}} + \kappa\iota + M \right) \left(\kappa\iota + C_\alpha \left(C_1 \lambda^{\frac{s-2}{s+2}} + \iota \right) \Delta^\alpha \right) + \lambda(\|f_\lambda\| + \iota).$$

Letting $\iota \to 0$, we have

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \leq \mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\Omega_0(f_\lambda) + 2C_\alpha \left(C_1^2 \kappa \lambda^{\frac{2(s-2)}{s+2}} + C_1 M \lambda^{\frac{s-2}{s+2}} \right) \Delta^\alpha.$$

This completes the proof of Lemma 4. ∎

The final confidence-based estimation for the hypothesis error is now obtained.

Proposition 2. *If X satisfies (1.4), ρ_X satisfies condition L_τ with $\tau > 0$, f_ρ lies in the range of $|L_K|^s$ for some $0 < s \leq 2$, and K satisfies (1.6), then for any $0 < \delta < 1$, with confidence $1 - \frac{\delta}{2}$ it holds*

$$\mathcal{P}(\mathbf{z}, \lambda) \leq C_2(\lambda^{\frac{2(s-2)}{s+2}} + \lambda^{\frac{s-2}{s+2}}) \left(\frac{\log(2/\delta) + \log(m+1)}{m}\right)^{\frac{\alpha}{\tau}}, \tag{4.4}$$

where C_2 is a constant independent of λ, m or δ.

Proof. We need to find a solution Δ to (4.1) in order to bound the hypothesis error by Lemma 4. To this end, we consider the strictly decreasing function h on $(0, \infty)$ defined by

$$h(t) = \log \mathcal{N}\left(X, \frac{t}{2}\right) - \frac{mC_\tau}{2^\tau} t^\tau.$$

Take

$$\Delta = \tilde{A}\left(\frac{\log(2/\delta) + \log(m+1)}{mC_\tau}\right)^{\frac{1}{\tau}}$$

where

$$\tilde{A} = 2\left(1 + \left(\frac{\eta}{\tau}\right)^{\frac{1}{\tau}} + C_\eta^{\frac{1}{\eta}} C_\tau^{\frac{1}{\tau}}\right).$$

Then we apply bound (1.4) for the covering number and see that

$$\begin{aligned} h(\Delta) \leq \log &\left(C_\eta \left(\frac{2}{\tilde{A}}\right)^\eta\right) \\ &+ \frac{\eta}{\tau} \log\left(\frac{mC_\tau}{\log(2/\delta) + \log(m+1)}\right) - \frac{\tilde{A}^\tau}{2^\tau}\left(\log(2/\delta) + \log(m+1)\right). \end{aligned}$$

From the definition of $\tilde{A}$, we see that $\tilde{A} \geq 2$, $\frac{\eta}{\tau} \leq (\frac{\tilde{A}}{2})^\tau$, and $\tilde{A}^\eta \geq C_\eta 2^\eta C_\tau^{\frac{\eta}{\tau}}$. It follows that

$$\begin{aligned} h(\Delta) &\leq \log \frac{C_\eta 2^\eta C_\tau^{\frac{\eta}{\tau}}}{\tilde{A}^\eta} \\ &+ \frac{\eta}{\tau}\log m - \frac{\eta}{\tau}\log\log\left[\frac{2}{\delta}(m+1)\right] - \log\frac{2}{\delta} - \frac{\tilde{A}^\tau}{2^\tau}\log(m+1) \leq \log\frac{\delta}{2}. \end{aligned}$$

That is, Δ satisfies inequality (4.1). By Lemma 3, with confidence at least $1 - \frac{\delta}{2}$, $\{x_i\}_{i=1}^m$ is Δ-dense. Then desired bound (4.4) follows from Lemma 4 with the constant C_2 given by

$$C_2 = 2C_\alpha(C_1^2\kappa + C_1 M)\tilde{A}^\alpha C_\tau^{-\frac{\alpha}{\tau}}.$$

The proof of Proposition 2 is complete. ∎

5. Estimating the Sample Error

Let $\mathcal{S}_1(\mathbf{z},\lambda) = \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_{\mathbf{z}}(f_\rho)\} - \{\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho)\}$ and $\mathcal{S}_2(\mathbf{z},\lambda) = \{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\}$, then $\mathcal{S}(\mathbf{z},\lambda) = \mathcal{S}_1(\mathbf{z},\lambda) + \mathcal{S}_2(\mathbf{z},\lambda)$. We bound these two parts of the sample error below.

Let $\xi(z) = \xi(x,y) = (f_\lambda(x) - y)^2 - (f_\rho(x) - y)^2$ be a random variable on Z. Then $\mathcal{S}_1(\mathbf{z},\lambda) = \frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbb{E}\xi$. Bounding $\|f_\lambda\|_\infty$ by (3.3) and (2.1), and bounding the variance of ξ by (3.1), a direct application of the one-side Bernstein inequality as in [8, 12] yields the following estimation.

Lemma 5. *Let $0 < \lambda \leq 1$. For any $0 < \delta < 1$, with confidence $1 - \frac{\delta}{4}$, it holds that*

$$\mathcal{S}_1(\mathbf{z},\lambda) \leq C_3 \left\{ \frac{\lambda^{\frac{2(s-2)}{s+2}}}{m} + \frac{\lambda^{\frac{2(s-1)}{s+2}}}{\sqrt{m}} \right\} \log\frac{4}{\delta} \tag{5.1}$$

where $C_3 = 2\left(\kappa^2 C_1^2 + 4M^2 + \sqrt{\kappa C_1}(\kappa C_1 + 2M)\right)$.

It is more difficult to bound $\mathcal{S}_2(\mathbf{z},\lambda)$ because it involves the sample $\mathbf{z}$ through $f_{\mathbf{z},\lambda}$. We use a probability inequality that handles a class of functions in $\mathcal{F}_0$. Such an inequality uses covering numbers in $\mathcal{F}_0$ to describe the complexity of $\mathcal{F}_0$. We bound the covering numbers in $\mathcal{F}_0$ firstly, and the following lemma plays an important role.

Lemma 6. *Suppose the kernel K satisfies* (1.7). *For any $f \in \mathcal{F}_0$ and $\Delta > 0$, we have*

$$|f(x) - f(x')| \leq C_\beta \|f\| (d(x,x'))^\beta \qquad \forall x, x' \in X.$$

Proof. Let $\iota > 0$. The function f can be written as $f = \sum_{j=1}^\infty \alpha_j K_{t_j}$ such that $t_j \in X$ and

$$\|f\| \leq \sum_{j=1}^\infty |\alpha_j| \leq \|f\| + \iota.$$

Then for $x, x' \in X$, we have

$$|f(x) - f(x')| = \left| \sum_{j=1}^\infty \alpha_j K(x,t_j) - \sum_{j=1}^\infty \alpha_j K(x,t_j) \right| \leq \sup_{t\in X} |K(x,t) - K(x',t)| \sum_{j=1}^\infty |\alpha_j|.$$

This in connection with (1.7) implies

$$|f(x) - f(x')| \leq C_\beta (d(x,x'))^\beta (\|f\| + \iota).$$

Letting $\iota \to 0$, we get what the lemma states. ■

Denote $B_R = \{f \in \mathcal{F}_0 : \|f\| \leq R\}$. Recall that $I(B_1)$ is a subset of $C(X)$. We are interested in its covering numbers $\mathcal{N}(I(B_1), r)$.

Lemma 7. *Let K satisfy* (1.7) *and X satisfy* (1.4). *Then for any $0 < r \leq 1$,*

$$\log \mathcal{N}(I(B_1), r) \leq C_\eta \left(\frac{4C_\beta}{r}\right)^{\frac{\eta}{\beta}} \log\left(2 + \frac{4\kappa}{r}\right).$$

Proof. Let $\Delta = (r/4C_\beta)^{\frac{1}{\beta}}$. Take $\mathbf{x} = \{x_i\}_{i=1}^N$ with $N = \mathcal{N}(X, \Delta)$ such that $\mathbf{x}$ is Δ-dense in X.

Any function $f \in B_1$ is continuous and

$$\|f\|_{C(X)} \leq \kappa \|f\| \leq \kappa.$$

So $-\kappa \leq f(x_i) \leq \kappa$ for each i. Hence, $(v_i - 1)r/2 \leq f(x_i) \leq v_i r/2$ for some $v_i \in J = \{-n+1, \ldots, n\}$ where $n = [2\kappa/r]$ is the smallest integer larger than $2\kappa/r$.

For $v = (v_1, \ldots, v_N) \in J^N$, define

$$V_v = \{f \in B_1 \mid (v_i - 1)r/2 \leq f(x_i) \leq v_i r/2, \ \forall\, i = 1, \ldots, N\}.$$

Then $I(B_1) = \bigcup_{v \in J^N} I(V_v)$. If $f, g \in V_v$, then by Lemma 6, for each $i \in \{1, \ldots, N\}$,

$$\begin{aligned}\max_{d(x,x_i)\leq\Delta} |f(x) - g(x)| &\leq |f(x_i) - g(x_i)| + \max_{d(x,x_i)\leq\Delta} |f(x) - f(x_i)| \\ &\quad + \max_{d(x,x_i)\leq\Delta} |g(x) - g(x_i)| \\ &\leq r/2 + 2C_\beta \Delta^\beta = r.\end{aligned}$$

But

$$\|f - g\|_{C(X)} = \max_{1\leq i\leq N} \max_{d(x,x_i)\leq\Delta} |f(x) - g(x)|.$$

Therefore, $I(V_v)$ has radius at most r as a subset of $C(X)$. That is, $\{I(V_v)\}_{v\in J^N}$ is an r-covering of $I(B_1)$. Therefore $\mathcal{N}(I(B_1), r)$ is bounded by the number of sets of type V_v with $v \in J^N$. Hence,

$$\log \mathcal{N}(I(B_1), r) \leq N \log(2n) \leq \mathcal{N}(X, \Delta) \log\left(2 + \frac{4\kappa}{r}\right),$$

and the desired estimate holds true. ■

For every $\varepsilon > 0$ and $R \geq M$, the following inequality as a uniform law of large numbers for a class of functions can be easily seen as Proposition 8.15 in [3]

$$
\begin{aligned}
&\text{Prob}\left\{\sup_{f\in B_R}\frac{\mathcal{E}(f)-\mathcal{E}(f_\rho)-(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}(f)-\mathcal{E}(f_\rho)+\varepsilon}}\leq\sqrt{\varepsilon}\right\} \\
&\geq 1-\mathcal{N}\left(I(B_1),\frac{\varepsilon}{(\kappa+3)^2R^2}\right)\exp\left\{-\frac{m\varepsilon}{54(\kappa+3)^2R^2}\right\}.
\end{aligned}
\tag{5.2}
$$

With this inequality, we have the following bound for $\mathcal{S}_2(\mathbf{z},\lambda)$.

Lemma 8. *Let K satisfy (1.7) and X satisfy (1.4). If $0 < \lambda \leq 1$, then with confidence $1-\frac{\delta}{4}$, it holds that*

$$
\mathcal{S}_2(\mathbf{z},\lambda)\leq\frac{1}{2}\left(\mathcal{E}(f_{\mathbf{z},\lambda})-\mathcal{E}(f_\rho)\right)+\frac{C_4(\log(4/\delta)+\log(m+1))}{\lambda^2}m^{-\frac{1}{1+\eta/\beta}},
\tag{5.3}
$$

where C_4 is independent of m, λ or δ.

Proof. Let $g:\mathbb{R}_+\to\mathbb{R}$ be the function given by

$$
g(r)=\log\mathcal{N}(I(B_1),r)-\frac{mr}{54}.
$$

Then g is strictly decreasing and for each $0<\delta\leq 1$ there is a unique minimum $r=\varepsilon^*(m,\delta/4)$ satisfying $g(r)\leq\log(\delta/4)$.

Take

$$
\tilde{r}=\max\left\{\frac{108\log(4/\delta)}{m},\tilde{B}m^{-\frac{1}{1+\eta/\beta}}\log(m+1)\right\}
$$

where

$$
\tilde{B}=\left\{108C_\eta(4C_\beta)^{\eta/\beta}[\log(2+4\kappa)+1]\right\}^{\frac{1}{1+\eta/\beta}}+2.
$$

Then $\frac{m\tilde{r}}{108}\geq\log\frac{4}{\delta}$ and by Lemma 7,

$$
\begin{aligned}
g(\tilde{r}) &\leq C_\eta\left(\frac{4C_\beta}{\tilde{r}}\right)^{\frac{\eta}{\beta}}\log\left(2+\frac{4\kappa}{\tilde{r}}\right)-\frac{m\tilde{r}}{108}-\log\frac{4}{\delta} \\
&\leq C_\eta\left(\frac{4C_\beta}{\tilde{r}}\right)^{\frac{\eta}{\beta}}\left\{\log\left(2+\frac{4\kappa}{\tilde{r}}\right)-\frac{m\tilde{r}^{1+\frac{\eta}{\beta}}}{108C_\eta(4C_\beta)^{\frac{\eta}{\beta}}}\right\}-\log\frac{4}{\delta}.
\end{aligned}
$$

The definition of $\tilde{r}$ tells us that $\log\frac{1}{\tilde{r}}\leq\log\left[\frac{1}{\tilde{B}\log(m+1)}m^{\frac{1}{1+\eta/\beta}}\right]\leq\frac{1}{1+\eta/\beta}\log m$. Then

$$
\begin{aligned}
g(\tilde{r})\leq C_\eta\left(\frac{4C_\beta}{\tilde{r}}\right)^{\frac{\eta}{\beta}}&\left\{\log(2+4\kappa)+\frac{1}{1+\eta/\beta}\log m\right. \\
&\left.-\frac{\tilde{B}^{1+\frac{\eta}{\beta}}}{108C_\eta(4C_\beta)^{\eta/\beta}}(\log(m+1))^{1+\frac{\eta}{\beta}}\right\}+\log\frac{\delta}{4}\leq\log\frac{\delta}{4}.
\end{aligned}
$$

Therefore $\varepsilon^*(m, \delta/4) \leq \tilde{r}$.

By taking $f = 0$ in the definition (1.3) of $f_{\mathbf{z},\lambda}$, we see that

$$\lambda\|f_{\mathbf{z},\lambda}\| \leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \leq \mathcal{E}_{\mathbf{z}}(0) \leq M^2.$$

So $f_{\mathbf{z},\lambda} \in B_R$ with $R = M^2/\lambda$. Take $\varepsilon = (\kappa+3)^2 R^2 \varepsilon^*(m, \delta/4)$ in (5.2). With confidence $1 - \frac{\delta}{4}$, we have

$$\begin{aligned}\mathcal{S}_2(\mathbf{z}, \lambda) &\leq \frac{1}{2}\left(\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)\right) + (\kappa+3)^2 R^2 \varepsilon^*(m, \delta/4) \\ &\leq \frac{1}{2}\left(\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho)\right) + \frac{(\kappa+3)^2 M^4}{\lambda^2}\tilde{r}.\end{aligned}$$

Thus the desired bound holds true with $C_4 := (\kappa+3)^2 M^4 \max\{108, \tilde{B}\}$. ∎

6. Deriving the Learning Rate

We can now derive the learning rate by combining the results obtained in Proposition 1, Proposition 2, Lemma 5 and Lemma 8.

Proof. [Proof of Theorem 1]. Let $\lambda = m^{-\theta}$ with $\theta > 0$. We have $0 < \lambda \leq 1$. From (3.1) of Proposition 1, we know that $\mathcal{D}(\lambda) \leq C_1 m^{-\frac{2\theta s}{s+2}}$.

By Proposition 2, with confidence $1 - \frac{\delta}{2}$,

$$\mathcal{P}(\mathbf{z}, \lambda) \leq 2C_2 \left(\log\frac{2}{\delta} + \log(m+1)\right)^{\frac{\alpha}{\tau}} m^{\frac{2\theta(2-s)}{s+2} - \frac{\alpha}{\tau}}.$$

By Lemma 5, with confidence $1 - \frac{\delta}{4}$,

$$\mathcal{S}_1(\mathbf{z}, \lambda) \leq C_3 \log\frac{4}{\delta} m^{-\min\left\{1 - \frac{2\theta(2-s)}{s+2},\ \frac{1}{2} - \frac{2\theta(1-s)}{s+2}\right\}}.$$

Combining the above estimates with Lemma 8 and Lemma 1, we see that with confidence $1 - \delta$,

$$\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} \leq \frac{1}{2}\|f_{\mathbf{z},\lambda} - f_\rho\|^2_{L^2_{\rho_X}} + C\left(\log\frac{4}{\delta} + \log(m+1)\right)^{\max\left\{1, \frac{\alpha}{\tau}\right\}} m^{-\Theta},$$

where $C = C_3 + C_4 + 2C_2 + C_1$ is a constant independent of m or δ. The proof of Theorem 1 is complete. ∎

References

1. N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.*, **68** (1950), 337-404.

2. R. A. Adams and J. F. Fournier, *Sobolev Spaces*, second ed., Elsevier, 2003.
3. F. Cucker and D. X. Zhou, *Learning theory: An approximation theory viewpoint*, Cambridge University Press, 2007.
4. E. De Vito, A. Caponnetto and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, *Found. Comput. Math.*, **5** (2005), 59-85.
5. D. Donoho, For most large undetermined systems of linear equations the minimal ℓ^1-norm solution is the sparsest solution, *Comm. Pure Appl. Math.*, **59** (2006), 797-829.
6. R. Opfer, Multiscale kernels, *Adv. Comput. Math.*, **25** (2006), 357-380.
7. S. Smale and D. X. Zhou, Learning theory estimates via integral operators and their approximations, *Constructive Approximation*, **26** (2007), 153-172.
8. Q. Wu, Y. Ying and D. X. Zhou, Learning rates of least-square regularized regression, *Foundations of Computational Mathematics*, **6** (2006), 171-192.
9. Q. Wu and D. X. Zhou, Learning with sample dependent hypothesis spaces, *Computers and Mathematics with Applications*, **56** (2008), 2896-2907.
10. Z. M. Wu and R. Schaback, Local error estimates for radial basis function interpolation of scattered data, *IMA J. Numer. Anal.*, **13** (1993), 13-27.
11. Y. S. Xu and H. Z. Zhang, Refinable kernels, *J. Mach. Learn. Res.*, **8** (2007), 2083-2120.
12. Y. Ying, Convergence analysis of online algorithms, *Adv. Comput. Math.*, **27** (2007), 273-291.
13. T. Zhang, Leave-one-out bounds for kernel methods, *Neural Comp.*, **15** (2003), 1397-1437.
14. D. X. Zhou, The covering number in learning theory, *J. Complexity*, **18** (2002), 739-767.

Quan-Wu Xiao
Joint Advanced Research Center in Suzhou,
University of Science and Technology of China
and City University of Hong Kong,
Suzhou, Jiangshu 215123,
P. R. China
E-mail: qwxiao@mail.ustc.edu.cn

Ding-Xuan Zhou
Department of Mathematics,
City University of Hong Kong,
Kowloon, Hong Kong,
P. R. China
E-mail: mazhou@cityu.edu.hk