Statistica Sinica **16**(2006), 425-439

CONVERGENCE RATES OF COMPACTLY SUPPORTED RADIAL BASIS FUNCTION REGULARIZATION

Yi Lin and Ming Yuan

University of Wisconsin-Madison and Georgia Institute of Technology

Abstract: Regularization with radial basis functions is an effective method in many machine learning applications. In recent years classes of radial basis functions with compact support have been proposed in the approximation theory literature and have become more and more popular due to their computational advantages. In this paper we study the statistical properties of the method of regularization with compactly supported basis functions. We consider three popular classes of compactly supported radial basis functions. In the setting of estimating a periodic function in a white noise problem, we show that regularization with (periodized) compactly supported radial basis functions is rate optimal and adapts to unknown smoothness up to an order related to the radial basis function used. Due to results on equivalence of the white noise model with many important models including regression and density estimation, our results are expected to give insight on the performance of such methods in more general settings than the white noise model.

Key words and phrases: method of regularization, nonparametric estimation, radial basis functions, rate of convergence, reproducing kernel.

1. Introduction

Radial basis functions (RBF's) are popular tools in function approximation and have been used in many machine learning applications. Examples include RBF regularization networks and, more recently, support vector machines. See, for example, Girosi, Jones and Poggio (1993), Smola, Schölkopf and Müller (1998), Wahba (1999) and Evgeniou, Pontil and Poggio (2000). Traditional radial basis functions have global support. In recent years RBF's with compact support have been proposed in the approximation theory literature (Wu (1995); Wendland (1995); Buhmann (1998)), and have become more and more popular in function approximation and machine learning applications due to their computational advantages. In this paper, we first present a brief overview of the commonly used compactly supported radial basis functions, and then give some theoretical results on the asymptotic properties of the method of regularization with compactly supported RBF's.

The radial basis functions have the form $\Phi(x) = \phi(\|x\|)$ for vector $x \in R^d$. Here ϕ is a univariate function defined on $[0, \infty)$ and $\|\cdot\|$ is the ordinary Euclidean norm on R^d. To be applicable in the method of regularization, a radial basis function must satisfy that $K(x, y) \equiv \Phi(x-y)$ is positive definite or conditionally positive definite. A function $K(x, y)$: $R^d \times R^d \to R$ is said to be positive definite if for any positive integer n, and any distinct $x_1, \ldots, x_n \in R^d$, we have

$$\sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k K(x_j, x_k) > 0 \tag{1.1}$$

for any nonzero vector $(a_1, \ldots, a_n)$. It is said to be conditionally positive definite of an order m if (1.1) holds for any nonzero vector $(a_1, \ldots, a_n)$ satisfying $\sum_{i=1}^{n} a_i p(x_i) = 0$ for all polynomials p of degree less than m. Thus positive definiteness is the same as conditionally positive definiteness of zero order. Any positive definite or conditional positive definite function $K(\cdot, \cdot)$ is associated with a reproducing kernel Hilbert space of functions on R^d. For an introduction to the (conditional) positive definite functions and reproducing kernel Hilbert spaces, the readers are referred to Wahba (1990). Schaback (1997) provided a survey of reproducing kernel Hilbert spaces corresponding to radial basis functions. We say a radial basis function Φ (or the corresponding ϕ) is (conditionally) positive definite if $K(x, y) = \Phi(x-y) = \phi(\|x-y\|)$ is (conditionally) positive definite.

For a positive definite radial basis function Φ, the squared norm in the associated reproducing kernel Hilbert space H_Φ can be written as $J(f) = (2\pi)^{-d/2} \int_{R^d} |\tilde{f}(\omega)|^2 / \tilde{\Phi}(\omega) d\omega$ for any function $f \in H_\Phi$. Here $\tilde{f}$ is the Fourier transform of f,

$$\tilde{f}(\omega) = (2\pi)^{-\frac{d}{2}} \int_{R^d} f(x) e^{-ix^T\omega} dx,$$

and, similarly, $\tilde{\Phi}$ is the the Fourier transform of Φ. For a conditional radial basis function, $J(f)$ is a squared semi-norm.

Let f be a function of interest in a nonparametric function estimation problem. The method of regularization with a radial basis function takes the form

$$\min_{f \in H_\Phi} [L(f, \text{data}) + \lambda J(f)], \tag{1.2}$$

where L is the empirical loss, often taken to be the negative log-likelihood. The smoothing parameter λ controls the trade-off between minimizing the empirical loss and obtaining a smooth solution. For example, in nonparametric regression

$$y_j = f(x_j) + \delta_j, \quad j = 1, \ldots, n, \tag{1.3}$$

where $x_j \in R^d$, $j = 1, \ldots, n$, are the regression inputs, y_j's are the responses, and δ_j's are independent Gaussian noises. In this case we may take $L(f, \text{data}) = \sum_{j=1}^{n} (y_j - f(x_j))^2$ in (1.2).

In many situations including regression and generalized regression, the representer theorem (Kimeldorf and Wahba (1971)) for regularization over reproducing kernel Hilbert spaces guarantees that the solution to (1.2) over H_Φ falls in the finite dimensional space spanned by $\{\Phi(\cdot - x_j), j = 1, \ldots, n\}$. For example, in the regression problem, when Φ is positive definite the solution to (1.2) has the form $\hat{f} = \sum_{j=1}^n c_j \Phi(x - x_j)$, and the coefficient vector $c = (c_1, \ldots, c_n)'$ can be solved by minimizing

$$\|y - Kc\|^2 + \lambda c' Kc. \tag{1.4}$$

Here, with a little abuse of notation, K stands for the n by n matrix $(\Phi(x_i - x_j))$, and y stands for the vector $(y_1, \ldots, y_n)'$. It is clear that the solution to (1.4) can be obtained by solving the linear system

$$(K + \lambda I)c = y. \tag{1.5}$$

Traditional radial basis functions are globally supported. Common examples are given by $\phi(r) = r^{2m} \log(r)$ (thin plate spline RBF), $\phi(r) = e^{-\varrho r^2/2}$ (Gaussian RBF), $\phi(r) = (c^2 + r^2)^{1/2}$ (the multiquadrics RBF), and $\phi(r) = (c^2 + r^2)^{-1/2}$ (the inverse multiquadrics RBF). In recent years radial basis functions with compact support have been constructed. These RBF's are computationally appealing and have become more and more popular in practical applications. In Section 2 we give a brief introduction to them that focuses on aspects that are most relevant to statistical applications.

In Sections 3−4 we present some results on the asymptotic properties of the method of regularization with (periodized) compactly supported radial basis functions. The asymptotic properties will be studied in the problem of estimating periodic functions in the nonparametric white noise model:

$$Y_n(t) = \int_{-\pi}^{t} f(u)du + n^{-\frac{1}{2}} B(t), \quad t \in [-\pi, \pi], \tag{1.6}$$

where $B(t)$ is a standard Brownian motion on $[-\pi, \pi]$ and we observe $Y_n = (Y_n(t), -\pi \le t \le \pi)$. We consider the situation where the function f belongs to a certain Sobolev ellipsoid of periodic functions on $[-\pi, \pi]$, $H^m(Q)$, with unknown m:

$$H^m(Q) = \{f \in L^2(-\pi, \pi) : f \text{ is } 2\pi\text{-periodic}, \int_{-\pi}^{\pi} [f(t)]^2 + [f^{(m)}(t)]^2 dt \le Q\}. \tag{1.7}$$

It is well known that spaces of periodic functions induce periodic reproducing kernels, see Schaback (1997) and Smola, Schölkopf and Müller (1998), for example. Therefore, for the estimation of periodic functions, it is appropriate to consider the periodic version of the radial basis functions. For any function

Φ, the corresponding periodized function with period 2π is given by: $\Phi_0(s) = \sum_{k\in Z^d}\Phi(s-2\pi k)$, where Z is the set of integers.

Consideration of the white noise model is motivated by the results on its equivalence to nonparametric regression (Brown and Low (1996)), density estimation (Nussbaum (1996)), spectral density estimation (Golubev and Nussbaum (1998)), and nonparametric generalized regression (Grama and Nussbaum (1997)). We choose to work with periodic function estimation because it allows a detailed and unified asymptotic analysis. We believe the results obtained in this paper also give insights on the statistical properties of radial basis function regularization in more general situations. For simplicity, we work in the case of univariate white noise model, but similar results can be obtained in higher dimensional situations.

Our results in Section 3 show that for (1.6) with $f \in H^m(Q)$, the method of regularization with periodized compactly supported radial basis functions achieves the optimal rate of convergence, and that this is true whenever m does not exceed a certain number known for the specific compactly supported radial basis function used, and the tuning parameter λ is appropriately chosen. In Section 4 we show that the smoothing parameter λ can be chosen adaptively in our situation without any loss in the rate of convergence. Section 5 contains some discussion. The technical proofs are given in the Appendix.

Throughout this paper, $a_n \sim b_n$ means that there exists $0 < c_1 < c_2 < \infty$ such that $c_1 < a_n/b_n < c_2$, for all $n = 1, 2, \ldots$.

2. Radial Basis Functions with Compact Support

In this section we briefly review those aspects of commonly used, compactly supported radial basis functions that are most relevant to machine learning and statistical applications. The review is mostly based on material in the papers by Wendland (1995), Schaback and Wendland (2000), and the book by Buhmann (2003). Interested readers are referred to them for further details.

There are several popular classes of radial basis functions with compact support. These classes of radial basis functions are constructed by finding compactly supported radial symmetric functions with positive (or nonnegative and not identically zero) Fourier transforms. Let $\Phi(x) \equiv \phi(\|x\|)$ be a bounded integrable radially symmetric function from R^d to R, with Fourier transform $\tilde{\Phi}$. Then for any $n \in Z$, distinct $x_1, \ldots, x_n \in R^d$, and nonzero vector $(a_1, \ldots, a_n)$, it is easy to show that,

$$\sum_{j,k=1}^{n} a_j a_k \phi(\|x_j - x_k\|) = (2\pi)^{-\frac{d}{2}} \int_{R^d} |\sum_{j=1}^{n} a_j e^{ix_j^T\omega}|^2 \tilde{\Phi}(\omega) d\omega. \qquad (2.1)$$

Since the exponentials are linear independent on every open subset of R^d, we see from (2.1) that if $\tilde{\Phi}$ is positive (or nonnegative and not identically zero), then Φ is a positive definite radial basis function. In fact, this is a special case of the famous Bochner's theorem that completely characterizes translation invariant positive definite functions through Fourier transforms.

2.1. Wendland functions

A particularly interesting part of the work on radial basis functions with compact support is due to Wendland, initiated in part by Wu (1995). The Wendland functions are of the form

$$\phi(r) = \begin{cases} p(r) & \text{if } 0 \le r \le 1, \\ 0 & \text{if } r > 1, \end{cases} \tag{2.2}$$

with a univariate polynomial p. They are supported in the unit ball, but a different support can easily be achieved by scaling.

Define an operator I by $Ig(r) \equiv \int_r^\infty g(t)tdt$, and write $\phi_l(r) = (1-r)_+^l$. Then the Wendland functions are

$$\phi_{d,k}(r) = I^k \phi_{\lfloor \frac{d}{2} \rfloor + k + 1}(r),$$

where I^k stands for the I-operator applied k times, and $\lfloor s \rfloor$ stands for the maximum integer less than or equal to s. It is easy to see that $\phi_{d,k}$ is compactly supported, and is a polynomial within its support. It has been shown (Wendland (1995)) that $\Phi_{d,k}(x) \equiv \phi_{d,k}(\|x\|)$ has strictly positive Fourier transform except when $d=1$ and $k=0$, in which case the Fourier transform is nonnegative and not identical zero. Thus in all cases $\phi_{d,k}(r)$ gives rise to a positive definite radial basis function. It was also shown in Wendland (1995) that $\phi_{d,k}(r)$ is $2k$ times continuously differentiable, and is of minimal polynomial degree among all positive definite radial basis functions of the form (2.2) that is $2k$ times continuously differentiable.

The explicit forms of Wendland functions can be computed directly from the definition or, alternatively, from the formula (Wendland (1995)):

$$I^k (1-r)_+^\ell = \sum_{i=0}^{k} \beta_{i,k} r^i (1-r)_+^{\ell+2k-i},$$

with coefficients $\beta_{0,0} = 1$, and

$$\beta_{j,k+1} = \sum_{i=j-1}^{k} \frac{\beta_{i,k}[i+1]_{i-j+1}}{(\ell+2k-i+1)_{i-j+2}}, \quad 0 \le j \le k+1,$$

if the term for $i = -1$ and $j = 0$ is ignored. Here $(\cdot)_k$ and $[\cdot]_k$ are defined by $[q]_{-1} = (1+q)^{-1}$, $(q)_0 = [q]_0 = 1$, $(q)_k = q(q+1)\cdots(q+k-1)$ and $[q]_k = q(q-1)\cdots(q-k+1)$.

By direct calculation we obtain the following examples of Wendland functions: if $\ell = \lfloor d/2 \rfloor + k + 1$, we have

$$\phi_{d,k}(r) = \begin{cases} (1-r)_+^{\ell} & \text{when } k = 0, \\ (1-r)_+^{\ell+1}\{(\ell+1)r+1\}/(\ell+1)_2 & \text{when } k = 1, \\ (1-r)_+^{\ell+2}\{(\ell+1)(\ell+3)r^2+3(\ell+2)r+3\}/(\ell+1)_4 & \text{when } k = 2. \end{cases} \tag{2.3}$$

2.2. Buhmann functions

A different technique for generating compactly supported radial basis functions is due to Buhmann (1998, 2000). Let $g(\beta) = (1-\beta^{\mu})_+^{\nu}$,with $0 < \mu \le 1/2$ and $\nu \ge 1$, and define

$$\phi_{d,\gamma,\alpha}(r) = \int_{r^2}^{\infty} (1 - \frac{r^2}{\beta})^{\gamma} \beta^{\alpha} g(\beta) d\beta,$$

and $\Phi_{d,\gamma,\alpha}(x) \equiv \phi_{d,\gamma,\alpha}(\|x\|)$, where d is the dimension of x. It is easy to see that $\phi_{d,\gamma,\alpha}(r)$ is compactly supported. It is supported in the unit ball, but a different support can easily be achieved by scaling. It has been shown (Buhmann (2003)) that

1. For $d = 1$, the Fourier transform of $\Phi_{d,\gamma,\alpha}(x)$ is everywhere positive if (i) $\gamma \ge 1$, $-1 < \alpha \le \gamma/2$, or (ii) $\gamma \ge 1/2$, $-1 < \alpha \le \min(1/2, \gamma - 1/2)$;
2. For $d > 1$, the Fourier transform of $\Phi_{d,\gamma,\alpha}(x)$ is everywhere positive if $\gamma \ge (d-1)/2$ and $-1 < \alpha \le [\gamma - (d-1)/2]/2$.

Thus $\phi_{d,\gamma,\alpha}$ induces positive definite radial basis functions and $\Phi_{d,\gamma,\alpha}(x) \in C^{1+\lceil 2\alpha \rceil}(R^d)$, where $\lceil s \rceil$ stands for the minimum integer greater than or equal to s.

For example, if $d = 1$ and we take $\mu = 1/2$, $\nu = 1$, $\alpha = 1/2$, and $\gamma = 1$, we get

$$\phi_{1,1,1/2}(r) = \begin{cases} -\frac{1}{2}r^4 + \frac{4}{3}r^3 - r^2 + \frac{1}{6} & \text{if } 0 \le r \le 1, \\ 0 & \text{if } r > 1, \end{cases}$$

This function is twice continuously differentiable. If $d = 2$ and we take $\mu = 1/2$, $\nu = 1$, $\alpha = 0$, and $\gamma = 2$, we get

$$\phi_{2,2,0}(r) = \begin{cases} 4r^2 \log r + r^4 - \frac{16}{3}r^3 + 4r^2 + \frac{1}{3} & \text{if } 0 \le r \le 1, \\ 0 & \text{if } r > 1. \end{cases}$$

This function is once continuously differentiable.

2.3. Product functions

A well-known property of positive definite functions is that the product of positive definite functions is positive definite. See, for example, Proposition 3.12 of Cristianini and Shawe-Taylor (2000). This suggests a strategy for constructing new compactly supported radial basis function based on known ones. If $\phi_1(r)$ is a positive definite radial basis function and $\phi_2(r)$ is a positive definite function radial basis function with compact support, $\phi(r) = \phi_1(r)\phi_2(r)$ is a positive definite radial basis function with compact support. This approach was proposed in Gaspari and Cohn (1999). As an example, we take ϕ_1 to be the Gaussian kernel $e^{-\varrho r^2/2}$ and ϕ_2 to be a Wendland or Buhmann function.

3. Rates of Convergence in Sobolev Spaces

To study the convergence rates of the method of regularization with periodized radial basis function in the white noise problem (1.6), we first transform the problem into a form that is particularly amenable to analysis. Let $\{\xi_0(t) = (2\pi)^{-1/2}, \xi_{2\ell-1}(t) = \pi^{-1/2}\sin(\ell t), \text{ and } \xi_{2\ell}(t) = \pi^{-1/2}\cos(\ell t)\}$ be the classical trigonometric basis in $L_2(-\pi,\pi)$, and let $\theta_\ell = (f, \xi_\ell)$ be the corresponding Fourier coefficients of f, where $(f,\phi) = \int_{-\pi}^{\pi} f(t)\xi(t)dt$ denotes the usual inner product in $L_2(-\pi,\pi)$. By converting the functions into the corresponding sequences of Fourier coefficients, we can see that the white noise problem (1.6) is equivalent to the following Gaussian sequence model:

$$y_\ell = \theta_\ell + \epsilon_\ell, \quad \ell = 0, 1, \dots, \tag{3.1}$$

where ϵ_ℓ's are independent $N(0, 1/n)$ variables. The condition that f belongs to $H^m(Q)$, the Sobolev ellipsoid of periodic functions defined in (1.7) with unknown m and Q, is equivalent to the condition that

$$\sum_{\ell=0}^{\infty} \rho_\ell \theta_\ell^2 \le Q, \tag{3.2}$$

where $\rho_0 = 1$, and $\rho_{2\ell-1} = \rho_{2\ell} = \ell^{2m} + 1$.

To study the method of regularization (1.2) with the periodized radial basis function $\Phi_0(s) = \sum_{k\in Z}\Phi(s - 2\pi k)$, we need the norm or seminorm of the reproducing kernel Hilbert space corresponding to Φ_0. The classes of compactly supported radial basis functions we consider in this paper are positive definite, therefore we seek the norm in the reproducing kernel Hilbert space.

Proposition 1. *Let Φ be a positive definite radial basis function with compact support, and Φ_0 be the periodization of Φ: $\Phi_0(s) = \sum_{k\in Z}\Phi(s - 2\pi k)$. Then Φ_0*

is positive definite and the norm in the reproducing kernel Hilbert space corresponding to Φ_0 *is given by*

$$\|f\|^2 = \sum_{\ell=0}^{\infty} \beta_\ell \theta_\ell^2, \tag{3.3}$$

where the θ*'s are the Fourier coefficients of* f, *and* $\beta_0 = (2\pi)^{-1/2}[\tilde{\Phi}(0)]^{-1}$, $\beta_{2\ell-1} = \beta_{2\ell} = (2\pi)^{-1/2}[\tilde{\Phi}(\ell)]^{-1}$, $\ell = 1, 2, \ldots$.

Therefore the method of regularization with Φ_0 corresponds to

$$\min \sum_{\ell=0}^{\infty} (y_\ell - \theta_\ell)^2 + \lambda \sum_{\ell=0}^{\infty} \beta_\ell \theta_\ell^2 \tag{3.4}$$

with $\beta_{2\ell-1} = \beta_{2\ell} = (2\pi)^{-1/2}[\tilde{\Phi}(\ell)]^{-1}$.

Theorem 1. *Assume* $f \in H^m(Q)$ *with* $m \geq 1$ *in the white noise model* (1.6). *Consider the method of regularization estimator* $\hat{\theta}$ *at* (3.4) *with* $\beta_{2\ell-1} = \beta_{2\ell} = (2\pi)^{-1/2}[\tilde{\Phi}(\ell)]^{-1}$.

1. *If* $\phi(r) = \sigma^{-1}\phi_{d,k}(r/\sigma)$ *is a scaled Wendland radial basis function of compact support with scale* σ, *and* $k \geq 1$, $m \leq 2k+2$, *then when* $\lambda_{(n)} \sim n^{-(2k+2)/(2m+1)}$ *we have*
$$\sup_{\theta \in H^m(Q)} \sum_\ell E(\hat{\theta}_\ell - \theta_\ell)^2 \sim n^{-\frac{2m}{2m+1}}.$$
That is, the optimal rate of convergence in the Sobolev space is achieved if the smoothing parameter is appropriately chosen.
2. *If* $\phi(r) = \sigma^{-1}\phi_{d,\gamma,\alpha}(r/\sigma)$ *is a scaled Buhmann radial basis function of compact support with scale* σ, *and* $m \leq 3 + 2\alpha$, *then when* $\lambda_{(n)} \sim n^{-(3+2\alpha)/(2m+1)}$ *we have*
$$\sup_{\theta \in H^m(Q)} \sum_\ell E(\hat{\theta}_\ell - \theta_\ell)^2 \sim n^{-\frac{2m}{2m+1}}.$$
That is, the optimal rate of convergence in the Sobolev space is achieved if the smoothing parameter is appropriately chosen.
3. *If* $\phi(r) = G(r)\phi_1(r)$ *where* $G(r) = e^{-\varrho r^2/2}$ *is the Gaussian radial basis function and* ϕ_1 *is a scaled Wendland function* (*or scaled Buhmann function*) *given in part* 1 (*or* 2), *then*
$$\sup_{\theta \in H^m(Q)} \sum_\ell E(\hat{\theta}_\ell - \theta_\ell)^2 \sim n^{-\frac{2m}{2m+1}}$$
when the conditions of part 1 (*or* 2) *hold.*

4. Adaptive Tuning

The proper choice of the tuning parameters depends on the unknown smoothness order m of the estimand. In this section, we consider choosing the tuning parameters λ and σ adaptively with Mallows' C_p. We show that the tuning parameters chosen by the unbiased estimator of risk give rise to an estimator that has the same asymptotic risk as the estimator with the optimal (theoretical) tuning parameter. Thus no asymptotic efficiency is lost due to not knowing m.

Formally, we take a finite number of σ's, $\sigma_1, \ldots, \sigma_J$, and tune λ and σ jointly over λ and $\sigma_j \in \{\sigma_1, \ldots, \sigma_J\}$. In our asymptotic considerations, a range of $[0,1]$ for λ suffices, since λ should go to zero. In practice we may use a slightly larger range. The tuning of σ in addition to λ is motivated by the common practice of tuning both λ and σ in the implementation of radial basis function regularization. All our asymptotic results go through if we fix one σ and only tune λ (this corresponds to $J = 1$ in our setting). However, with a finite sample size it is desirable to tune both λ and σ. In fact, it is clear that in the finite sample situation, fixing σ at a value that is too small compared to the distances between sample points leads to poor estimation performance.

The tuning is based on the C_p criterion. Writing $\tau_\ell = (1 + \lambda\beta_\ell)^{-1}$, our estimator is $\hat{\theta}_\ell = \tau_\ell y_\ell$. We can express the risk of our estimator as

$$\sum E(\hat{\theta}_\ell - \theta_\ell)^2 = \frac{1}{n}\sum_{\ell=0}^{\infty}\tau_\ell^2 + \sum_{\ell=0}^{\infty}(1-\tau_\ell)^2\theta_\ell^2.$$

Now an unbiased estimator for θ_ℓ^2 is $y_\ell^2 - (1/n)$. Plugging in, we get that

$$\sum_{\ell=0}^{\infty}[(\tau_\ell^2 - 2\tau_\ell)(y_\ell^2 - \frac{1}{n}) + \frac{1}{n}\tau_\ell^2] = \sum_{\ell=0}^{\infty}[(\tau_\ell^2 - 2\tau_\ell)y_\ell^2 + \frac{2}{n}\tau_\ell] \tag{4.1}$$

is an unbiased estimator of $\sum E(\hat{\theta}_\ell - \theta_\ell)^2 - \sum \theta_\ell^2$. We choose λ^* and σ^* that minimize (4.1), and use the corresponding estimator $\hat{\theta}^*$. Li (1986, 1987) established the asymptotic optimality of C_p in many nonparametric function estimation methods, including the method of regularization. Kneip (1994) studied the adaptive choice among ordered linear smoothers with the unbiased risk estimator. A family of ordered linear smoothers satisfy the condition that for any member $\hat{\theta}_\ell = \tau_\ell y_\ell$, $\ell = 0, 1, \ldots$, of the family, we have $\tau_\ell \in [0,1]$, $\forall \ell$; and for any two members of the family $\tau_\ell y_\ell$ and $\tau'_\ell y_\ell$, $\ell = 0, 1, \ldots,$, we have either $\tau_\ell \geq \tau'_\ell$, $\forall \ell$, or $\tau'_\ell \geq \tau_\ell$, $\forall \ell$.

Theorem 2. *Consider* (3.1) *and the method of regularization* (3.4) *with* $\phi(r)$ *defined as in Theorem* 1(1), 1(2) *or* 1(3). *Suppose* λ^* *and* σ^* *minimize* (4.1) *over* $\lambda \in [0,1]$ *and* $\sigma \in \{\sigma_1, \ldots, \sigma_J\}$, *and* $\hat{\theta}^*$ *is the corresponding method of regularization estimator. Then*

$$\sup_{\theta \in H^m(Q)} \sum E(\hat{\theta}^*_\ell - \theta_\ell)^2 \sim n^{-\frac{2m}{2m+1}}.$$

Therefore the adaptive method of regularization with compactly supported radial basis function estimator $\hat{\theta}^*$ achieves the optimal rate in $H^m(Q)$. In the case of product radial basis functions, one may want to tune $\varrho_k \in \{\varrho_1, \ldots, \varrho_K\}$ as well as $\lambda \in [0,1]$ and $\sigma \in \{\sigma_1, \ldots, \sigma_J\}$. The estimator obtained still achieves the optimal rate in $H^m(Q)$, as can be proved in similar fashion.

5. Discussion

Compactly supported radial basis functions have also been used in spatial statistics to model and approximate spatial covariance. Gneiting (2002) is a good source of compactly supported covariances. Furrer, Genton and Nychka (2005) proposed a method to taper the spatial covariance function to zero beyond a certain range, using an appropriate compactly supported radial radial basis function. This gives an approximation to the standard linear spatial predictor that is both accurate and computationally efficient. They show that their approximate taper-based method makes it possible to analyze and fit very large spatial data sets, and gives a linear predictor that is nearly the same as the exact solution.

Radial basis function regularization has been shown to give excellent performance in practical applications. Direct implementation with traditional globally supported radial basis functions require a computation of order $O(n^3)$. Compactly supported radial basis functions have computational advantages. The results in this paper suggest that regularization with compactly supported radial basis functions enjoy good statistical properties. Further study is needed to investigate these properties more generally. Another interesting research topic is how to efficiently make use of their computational advantages. This is especially important when analyzing large datasets. When the compactly supported radial basis function is used in the method of regularization, the kernel matrix K in (1.5), for example, is sparse. Sparse matrix computation can be used directly in the non-iterative evaluation of the estimator in (1.5). This is the approach taken in Zhang, Genton and Liu (2004). Another approach to improving computation speed is to solve (1.5) with an iterative method, such as the conjugate gradient method. Due to the sparsity and positive definiteness of the linear system (1.5), the conjugate gradient method is expected to be very effective. There have been a number of effective proposals in the approximation theory literature for fast computation in function interpolation with compactly supported radial basis functions. See, for example, Schaback and Wendland (2000) and Floater and Iske (1996). It is promising that adapting these ideas for function approximation in noise free situations to function estimation in noisy situations can lead to fast and efficient machine learning and statistical methods based on compactly supported radial basis functions.

Acknowledgement

Lin's research was supported in part by NSF Grant DMS-0134987.

Appendix

Proof of Proposition 1. Applying Lemma 14 of Williamson, Smola and Schölkopf (2001) (see also Lin and Brown (2004)), we get

$$\Phi_0(s-t) = (2\pi)^{-\frac{1}{2}}\tilde{\Phi}(0) + \sum_{\ell=1}^{\infty} 2(2\pi)^{-\frac{1}{2}}\tilde{\Phi}(\ell)\cos[\ell(s-t)],$$

which can be rewritten as

$$\Phi_0(s-t) = (2\pi)^{\frac{1}{2}}\tilde{\Phi}(0)\xi_0(s)\xi_0(t) + \sum_{\ell=1}^{\infty}(2\pi)^{\frac{1}{2}}\tilde{\Phi}(\ell)\xi_{2\ell-1}(s)\xi_{2\ell-1}(t) + \sum_{\ell=1}^{\infty}(2\pi)^{\frac{1}{2}}\tilde{\Phi}(\ell)\xi_{2\ell}(s)\xi_{2\ell}(t),$$

where the ξ_j form the classical trigonometric basis in $L_2(-\pi,\pi)$. Then ξ_ℓ, $\ell = 0, 1, \ldots$, are the eigenvectors of the reproducing kernel $\Phi_0(s-t)$, and the corresponding eigenvalues are β_ℓ^{-1}, where $\beta_0^{-1} = (2\pi)^{1/2}\tilde{\Phi}(0)$, $\beta_{2\ell-1}^{-1} = \beta_{2\ell}^{-1} = (2\pi)^{1/2}\tilde{\Phi}(\ell)$, $\ell = 1, 2, \ldots$. Now applying Lemma 1.1.1 of Wahba (1990), we get that for any function f in the reproducing kernel Hilbert space induced by Φ_0, $\|f\|^2 = \sum_{\ell=0}^{\infty}\beta_\ell(f,\xi_\ell)^2$, and the proposition is proved.

Proof of Theorem 1(1). Since $k \geq 1$, we have $\tilde{\Phi}_{d,k}(\omega) \sim (1+\|\omega\|^2)^{-(1+2k+d)/2}$ (Theorem 2.1 of Wendland (1998)). Therefore, for our scaled Wendland function Φ and $d = 1$, we have $\tilde{\Phi}(\omega) = \tilde{\Phi}_{d,k}(\sigma\omega) \sim (1+\sigma^2\omega^2)^{-(k+1)}$, and then $\beta_{2\ell-1} = \beta_{2\ell} \sim (1+\sigma^2\ell^2)^{k+1}$.

Solving the minimization problem (3.4), we get the method of regularization estimator $\hat{\theta}_\ell = (1+\lambda\beta_\ell)^{-1}y_\ell$. We consider the variance and bias of $\hat{\theta}$ separately.

First, since $\lambda \sim n^{-(2k+2)/(2m+1)}$, we have,

$$\begin{aligned}\sum_{\ell} var\hat{\theta}_\ell = \frac{1}{n}\sum(1+\lambda\beta_\ell)^{-2} &\sim \frac{2}{n}\sum_{\ell=0}^{\infty}[1+\lambda(\sigma\ell)^{2k+2}]^{-2}\\ &\sim \frac{2}{n}\int_0^\infty [1+\lambda(\sigma x)^{2k+2}]^{-2}dx\\ &= \frac{2}{n}\lambda^{-\frac{1}{2k+2}}\sigma^{-1}\int_0^\infty (1+y^{2k+2})^{-2}dy\\ &\sim n^{-1}\lambda^{-\frac{1}{2k+2}}\sigma^{-1} \sim n^{-\frac{2m}{2m+1}}.\end{aligned}$$

For the bias part, we have

$$\begin{aligned}\sup_{\theta\in H^m(Q)}\sum_\ell(E\hat{\theta}_\ell-\theta_\ell)^2 &= \sup_{\theta\in H^m(Q)}\sum_{\ell=0}^{\infty}(1+\lambda^{-1}\beta_\ell^{-1})^{-2}\theta_\ell^2\\ &= \sup_{\theta\in H^m(Q)}\sum_{\ell=0}^{\infty}(1+\lambda^{-1}\beta_\ell^{-1})^{-2}\rho_\ell^{-1}(\rho_\ell\theta_\ell^2)\\ &= Q\max_\ell[(1+\lambda^{-1}\beta_\ell^{-1})^{-2}\rho_\ell^{-1}]\\ &\sim \max_{0\le\ell<\infty}[\{1+\lambda^{-1}(1+\sigma^2\ell^2)^{-(k+1)}\}^{-2}(1+\ell^{2m})^{-1}].\end{aligned}$$

Here $\rho_{2\ell-1}=\rho_{2\ell}=1+\ell^{2m}$ are the coefficients in the condition (3.2) of the Sobolev ellipsoid $H^m(Q)$. Define $B_\lambda(\ell)=\{1+\lambda^{-1}(1+\sigma^2\ell^2)^{-(k+1)}\}^2(1+\ell^{2m})$. Then

$$\sup_{\theta\in H^m(Q)}\sum_\ell(E\hat{\theta}_\ell-\theta_\ell)^2 \sim [\min_{\ell\ge0}B_\lambda(\ell)]^{-1}. \tag{A.1}$$

Now $B_\lambda(0)=(1+\lambda^{-1})^2$ and

$$\begin{aligned}\min_{\ell\ge1}B_\lambda(\ell) &\sim \min_{\ell\ge1}\{1+\lambda^{-1}(\sigma^2\ell^2)^{-(k+1)}\}^2\ell^{2m}\\ &\sim \min_{x\ge1}\{1+\lambda^{-1}(\sigma^2x^2)^{-(k+1)}\}^2x^{2m}\\ &= \min_{x\ge1}\{x^m+\lambda^{-1}\sigma^{-(2k+2)}x^{m-(2k+2)}\}^2.\end{aligned}$$

If $m=2k+2$, then obviously the last expression is $\{1+\lambda^{-1}\sigma^{-(2k+2)}\}^2 \sim \lambda^{-2} = \lambda^{-2m/(2k+2)}$. When $m<2k+2$, direct calculation shows that the minimum in the last expression is achieved at $x=(2k+2-m)^{1/(2k+2)}m^{-1/(2k+2)}\lambda^{-1/(2k+2)}\sigma^{-1}$, and the minimum is of order $\lambda^{-2m/(2k+2)}$. Since $m\le 2k+2$, we have $\min_{\ell\ge0}B_\lambda(\ell) \sim \lambda^{-2m/(2k+2)}$.

From (A.1) and $\lambda\sim n^{-(2k+2)/(2m+1)}$, we get

$$\sup_{\theta\in H^m(Q)}E\sum_\ell(\hat{\theta}_\ell-\theta_\ell)^2=\sum_\ell var\hat{\theta}_\ell+\sup_{\theta\in H^m(Q)}\sum_\ell(E\hat{\theta}_\ell-\theta_\ell)^2 \sim n^{-\frac{2m}{2m+1}}.$$

Proof of Theorem 1(2). We note that $\tilde{\Phi}_{d,\gamma,\alpha}(\omega) \sim (1 + \|\omega\|^2)^{-(1+\alpha+d/2)}$ (Buhmann (2003, p.158)). The rest of the proof is similar to that of Theorem 1(1).

Proof of Theorem 1(3). In the following, $a = k+1$ if ϕ_1 is Wendland function and $a = \alpha + 3/2$ if ϕ_1 is Buhmann function. We have

$$\begin{aligned}\tilde{\Phi}(\omega) &= \int_{-\infty}^{\infty} \tilde{\Phi}_1(s)\tilde{G}(\omega - s)ds \\ &\sim \int_{-\infty}^{\infty} (1 + \sigma^2 s^2)^{-a} e^{-\frac{(s-\omega)^2}{2\varrho}} ds \\ &\sim E(1 + \sigma^2 Z_{\omega,\varrho}^2)^{-a},\end{aligned}$$

where $Z_{\omega,\varrho}$ is $N(\omega, \varrho)$. Since $E(1 + \sigma^2 Z_{\omega,\varrho}^2)^{-a}$ is a positive and continuous function in ω, when $|\omega| \to \infty$ we have $E(1 + \sigma^2 Z_{\omega,\varrho}^2)^{-a} = \omega^{-2a} E[1/\omega^2 + (\sigma + \sigma Z_{0,\varrho/\omega^2})^2]^{-a} \sim \omega^{-2a}$. We get that $\tilde{\Phi}(\omega) \sim E(1 + \sigma^2 Z_{\omega,\varrho}^2)^{-a} \sim (1 + \omega^2)^{-a}$. The rest of the proof is similar to that of Theorem 1(1).

Proof of Theorem 2. It is easy to check that, for any fixed $\sigma \in \{\sigma_1, \ldots, \sigma_J\}$, the method of regularization estimators with varying λ form a family of ordered linear smoothers. Applying the result in Kneip (1994) (see also expression (15) in Cavalier, Golubev, Picard and Tsybakov (2002)) shows that there exist positive constants C_1 and C_2 such that, for any θ with $\sum \theta_l^2 < +\infty$ and any positive constant B, we have

$$\sum E(\hat{\theta}_\ell^* - \theta_\ell)^2 \le (1 + C_1 B^{-1}) \min_{\lambda,\sigma_j} \{\sum E(\hat{\theta}_\ell - \theta_\ell)^2\} + n^{-1} C_2 B. \tag{A.2}$$

Therefore we have

$$\begin{aligned}&\sup_{\theta \in H^m(Q)} \sum E(\hat{\theta}_\ell^* - \theta_\ell)^2 \\ &\le (1 + C_1 B^{-1}) \sup_{\theta \in H^m(Q)} \min_{\lambda,\sigma_j} \{\sum E(\hat{\theta}_\ell - \theta_\ell)^2\} + n^{-1} C_2 B \\ &\le (1 + C_1 B^{-1})) \min_{\lambda,\sigma_j} \sup_{\theta \in H^m(Q)} \{\sum E(\hat{\theta}_\ell - \theta_\ell)^2\} + n^{-1} C_2 B.\end{aligned}$$

Now take $B = (\log n)^{1/3}$ and the conclusion of the theorem then follows from Theorem 1.

References

Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24**, 2384-2398.

Buhmann, M. D. (1998). Radial functions on compact support. *Proc. Edinb. Math. Soc.* (2) **41**, 33-46.

Buhmann, M. D. (2000). A new class of radial basis functions with compact support. *Math. Comput.* **70**, 307-318.

Buhmann, M. D. (2003). *Radial Basis Functions: Theory and Implementations.* Cambridge University Press, Cambridge.

Cavalier, L., Golubev, G. K., Picard, D. and Tsybakov, A. B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.* **30**, 843-874.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press, Cambridge.

Evgeniou, T., Pontil, M. and Poggio, T. (2000). Regularization networks and support vector machines. *Adv. Comput. Math.* **13**, 1-50.

Floater, M. S. and Iske, A. (1996). Multistep scattered data interpolation using compactly supported radial basis functions. *J. Comput. Appl. Math.* **73**, 65-78.

Furrer, R., Genton, M. G. and Nychka, D. (2005). Covariance tapering for interpolation of large spatial datasets. Available at www.cgd.ucar.edu/stats/pub/furrer/manuscripts/.

Gaspari, G. and Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.* **125**, 723-757.

Girosi, F., Jones, M. and Poggio, T. (1993). Priors, stabilizers and basis functions: from regularization to radial, tensor and additive splines. M.I.T. Artificial Intelligence Laboratory Memo 1430, C.B.C.I. Paper 75.

Gneiting, T. (2002). Compactly supported correlation functions. *J. Multivariate Anal.* **83**, 493-508.

Golubev, G. and Nussbaum, M. (1998). Asymptotic equivalence of spectral density and regression estimation. Technical report, Weierstrass Institute for Applied Analysis and Stochastics, Berlin.

Grama, I. and Nussbaum, M. (1997). Asymptotic equivalence for nonparametric generalized linear models. Technical report, Weierstrass Institute for Applied Analysis and Stochastics, Berlin.

Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22**, 835-866.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33**, 82-95.

Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14**, 1101-1112.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15**, 958-975.

Lin, Y. and Brown, L. D. (2004). Statistical properties of the method of regularization with periodic Gaussian reproducing kernel. *Ann. Statist.* **32**, 1723-1743.

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24**, 2399-2430.

Schaback, R. (1997). Native Hilbert spaces for radial basis functions I. In *Proceedings of Chamonix* 1996 (Edited by A. Le Méhauté, C. Rabut and L. L. Schumaker), 1-18.

Schaback, R. and Wendland, H. (2001). Characterization and construction of radial basis functions. In *Multivariate Approximation and Applications* (Edited by N. Dyn, D. Leviatan, D. Levin and A. Pinkus), 1-24. Cambridge University Press, Cambridge.

Smola, A. J., Schölkopf, B. and Müller, K. (1998). The connection between regularization operators and support vector kernels. *Neural Networks* **11**, 637-649.

Wahba, G. (1990). *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics. Philadelphia, Pennsylvania.

Wahba, G. (1999). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In *Advances in Kernel Methods - Support Vector Learning* (Edited by Schölkopf, Burges and Smola), 69-88. MIT Press.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.* **4**, 389-396.

Wendland, H. (1998). Error estimates for interpolation by compactly supported radial basis functions of minimal degree. *J. Approx. Theory* **93**, 258-272.

Williamson, R. C., Smola, A. J. and Schölkopf, B. (2001). Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Trans. Inform. Theory* **47**, 2516-2532.

Wu, Z. (1995) Multivariate compactly supported positive definite radial functions. *Adv. Comput. Math.* **4**, 283-292.

Zhang, H. H., Genton, M. G. and Liu, P. (2004). Compactly supported radial basis function kernels. Available at www4.stat.ncsu.edu/~hzhang/research.html.

Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706.

E-mail: yilin@stat.wisc.edu

School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive, Atlanta, GA 30332.

E-mail: myuan@isye.gatech.edu

(Received April 2005; accepted August 2005)