

Department of Economics

Boosting Estimation of RBF Neural Networks for Dependent Data

George Kapetanios and Andrew P. Blake

Working Paper No. 588

March 2007

ISSN 1473-0278



Queen Mary
University of London

Boosting estimation of RBF neural networks for dependent data

George Kapetanios*
Queen Mary, University of London

Andrew P. Blake†
Bank of England

February 19, 2007

Abstract

This paper develops theoretical results for the estimation of radial basis function neural network specifications, for dependent data, that do not require iterative estimation techniques. Use of the properties of regression based boosting algorithms is made. Both consistency and rate results are derived. An application to nonparametric specification testing illustrates the usefulness of the results.

Keywords: Neural Networks, Boosting.

AMS 2000 Classification: 82C32.

1 Introduction

The consideration of flexible nonlinear specifications has played a significant part in the development of nonparametric modeling in statistics and econometrics. Such nonlinear specifications form part of a toolkit that has been used to provide approximations to unknown functions in diverse areas such as, e.g., nonparametric specification testing, time series model building and specification of diffusion process models.

A major issue in the use of flexible nonlinear specifications is the need for robust and efficient estimation algorithms. Unfortunately, the problem of estimating such nonlinear models has meant that traditionally focus has been restricted to series

*Department of Economics, Queen Mary, University of London, Mile End Road, London E1 4NS, UK. Email: G.Kapetanios@qmul.ac.uk.

†Monetary Analysis, Bank of England, Threadneedle Street, London EC2R 8AH, UK. Email: Andrew.Blake@bankofengland.co.uk.

expansions, i.e. specifications that involve linear combinations of basis functions, such as trigonometric functions or polynomials. Such basis functions do not involve unknown parameters and therefore, estimation boils down to linear least squares estimation of the linear combination coefficients.

Such restrictions, however, have considerable costs in the sense that many classes of powerful flexible nonlinear specifications are excluded. One such class is neural networks. Neural networks are similar to other classes of approximators in that basis functions are linearly combined to provide an approximation¹. However, these basis functions typically involve unknown parameters. Since these parameters need to be estimated and the number of nodes, formed from the basis function, may be quite large, estimation of neural network specifications is not trivial. The estimation problem has been addressed in ways specific to the application being considered. For example in the case of neglected nonlinearity testing in regression models, work by Lee, White, and Granger (1993) has rested on the use of randomly generated nodes which bypasses the need for estimation. More importantly, in this case, the attendant problem of lack of identification under the null hypothesis of no nonlinearity is thus solved.

In a series of papers, Blake and Kapetanios (2007, 2000, 2003a,b) have introduced a new class of neural networks in the context of a diverse set of testing problems in econometrics. These neural network specifications based on radial basis functions (RBF), provide a novel way for alleviating the aforementioned estimation (and in some cases identification) problem although these specifications are used in the context of testing rather than estimation in the above papers. Although radial basis functions involve the use of nodes that contain unknown parameters, these parameters can

¹For a good review see Bishop (1995). Cybenko (1989), Hornik, Stinchcombe, and White (1989) and Park and Sandberg (1991) provide basic approximation results for large classes of neural network specifications.

be selected in such a way that removes the need for nonlinear estimation. Then, linear least squares completes the estimation of the neural network specification. Furthermore, the selection of the basis function parameters has an extremely useful by-product. It provides a ranking for the nodes which is not readily available for neural networks unlike trigonometric or polynomial approximations where the ranking of the nodes is natural. This estimation approach proved extremely effective. In most testing contexts where the RBF neural networks (RBFNN) were used, they were either clear favourites in terms of test power or very close to the favourites. The main advantage of this line of work compared to standard series expansions such as trigonometric expansions is the fact that through the use of parameters the actual set of nodes used for approximation adapts in a data driven way to the problems at hand. This extends the idea of adaptation which, in this context, is usually taken to mean that the number of nodes is chosen in a data dependent way. So, in the case of RBFNN the adaptation is dual since the number of nodes can also be adaptively selected.

The above line of work built on a literature that focused mainly on practical applicability and relevance. It did not stress theoretical rigour but small sample performance. Of course, another vast strand of the statistical and econometric literature focused on the theoretical properties of nonparametric methods based on series expansions. That work provides a theoretical account of the properties of these methods when applied to problems such as estimation or specification testing. Examples of such work include Bierens (1984), Aerts, Claeskens, and Hart (1999), Newey (1997) and, more recently, Guerre and Lavergne (2005) and Guay and Guerre (2006). Reviews may be found in, e.g., Hart (1997) and Pagan and Ullah (2000). Clearly, there is a gap on whether the existing theoretical analysis relates to neural network specifications and especially RBFNNs. This is clearly of interest since the good small

sample performance of RBFNNs suggests that they merit further focus.

This is the aim of the current paper. We provide a theoretical analysis of RBFNNs. In particular, the method of selecting the parameters of the nodes is analysed using the fact that this method bears very close similarity to a form of boosting. Boosting refers to a set of algorithms which have become very popular in disciplines such as machine learning and, more recently, statistics, in the context of classification and prediction (see, e.g., Freund and Schapire (1996), Friedman, Hastie, and Tibshirani (2000), Schapire (2002), Friedman (2001) and Buhlmann (2006)). The link between boosting and neural networks is not new. For example, one of the early references on boosting in machine learning, uses neural networks (see Drucker, Schapire, and Simard (1993)). Further, greedy algorithms which are closely related to boosting have been considered in the context of neural network training by, e.g., Jones (1992). However, our treatment has a number of distinctive features. In particular, the formal statistical link and results developed between boosting and RBFNNs is to our knowledge novel. Another major distinctive feature is the attention paid to problems arising out of the consideration of dependent data which is of great importance in developing forecasting models. The focus of the paper is solely theoretical. We feel that existing small sample evidence in terms of specification testing, is more than compelling in favour of RBFNNs.

The structure of the paper is as follows: Section 2 presents the preliminary setting of the paper. Section 3 presents the main theoretical result. Section 4 presents an application to nonparametric specification testing. Finally, Section 5 concludes. Proofs are relegated to the Appendix.

2 Setup

Consider a regression model of the form

$$y_t = \mu(x_t) + \epsilon_t \quad (1)$$

The aim is to estimate the unknown regression function by an RBFNN series expansion of the form

$$\hat{\mu}(x_t) = \sum_{i=1}^m c_i \psi(x_t, t_i, \sigma_T) \quad (2)$$

where the RBF nodes, $\psi(x_t, t_i, \sigma_T)$, are radially symmetrical, integrable, bounded functions and t_i are referred to as the centres of the RBFs. Examples include the Gaussian function of the form $\exp\left(-\left(\frac{\|x-t_i\|}{\sigma_T}\right)^2\right)$, or the multiquadratic function $\left(1 + \left(\frac{\|x-t_i\|}{\sigma_T}\right)^2\right)^{-1}$, $\sigma_T > 0$, where $\|\cdot\|$ denotes Euclidean distance. Obviously, estimation of (2) is challenging since unlike standard series expansions, there are two problems that need attention: the first is that $\psi(x, t_i, \sigma_T)$ contain unknown parameters, in particular the centres, and the second is that the nodes are not ranked so that the choice of the nodes in the series expansion is not obvious. Once the order of the nodes and the centres are determined the series expansion can be estimated by least squares.

A popular algorithm for solving the above problem has been suggested by Orr (1995). In a series of papers, Blake and Kapetanios (2007, 2000, 2003a,b) have modified that algorithm for specifically econometric applications with some success. In this paper we modify it further to bring it more in line with the regression based boosting algorithm of Buhlmann (2006). We define this new algorithm as Algorithm 1 below, and label it as the *(RBF) Boosting Algorithm*.

Algorithm 1 *(RBF) Boosting algorithm*

1. Let σ_T be some sequence such that $\sigma_T = o(1)$. We construct the initial set of T RBF nodes given by: $\Psi^{(1, \dots, T)} = \{\psi(x, x_1, \sigma_T), \psi(x, x_2, \sigma_T), \dots, \psi(x, x_T, \sigma_T)\}$.

2. These are ranked according to their ability to reduce the residual variance, when each $\psi(x_t, x_i, \sigma_T)$, $i = 1, \dots, T$, is entered individually in (2).
3. The node that minimises the residual variance becomes the first node in the ranking of the nodes. Denote this node by $\psi(x, x_{\mathcal{S}_1}, \sigma_T)$. Denote the residual from the regression of y_t on $\psi(x_t, x_{\mathcal{S}_1}, \sigma_T)$, by $y_t^{(1)}$. Let $\tilde{\mathcal{S}}_1 = \{\mathcal{S}_1\}$. Let $\Psi^{(1, \dots, T)/\tilde{\mathcal{S}}_1}$ be the set of nodes in $\Psi^{(1, \dots, T)}$ apart from the nodes indexed by the elements of $\tilde{\mathcal{S}}_1$.
4. Set $i = 1$.
5. The nodes in $\Psi^{(1, \dots, T)/\tilde{\mathcal{S}}_1}$ are ranked according to their ability to reduce the residual variance of $y_t^{(i)}$, when $y_t^{(i)}$ is regressed on each $\psi(x_t, x_i, \sigma_T)$, $i \in \tilde{\mathcal{S}}_1$.
6. The node that minimises the residual variance becomes the $i + 1$ -th node in the ranking of the nodes. Denote this node by $\psi(x, x_{\mathcal{S}_{i+1}}, \sigma_T)$. Denote the residual from the regression of $y_t^{(i)}$ on $\psi(x_t, x_{\mathcal{S}_{i+1}}, \sigma_T)$, by $y_t^{(i+1)}$. Let $\tilde{\mathcal{S}}_{i+1} = \tilde{\mathcal{S}}_{i+1} \cup \{\mathcal{S}_{i+1}\}$. Let $\Psi^{(1, \dots, T)/\tilde{\mathcal{S}}_{i+1}}$ be the set of nodes in $\Psi^{(1, \dots, T)}$ apart from the nodes indexed by the elements of $\tilde{\mathcal{S}}_{i+1}$.
7. If $i = m$ for some $m = m_T \rightarrow \infty$ stop, else set $i = i + 1$ and go to Step 5.

Some remarks are in order for this algorithm.

Remark 1 The choice for m is not discussed in Algorithm 1 apart from noting that $m \rightarrow \infty$. Theorem 1 suggests that the maximum possible rate is logarithmic in T .

Remark 2 The sequence σ_T is left unspecified in Algorithm 1. The proof of Theorem 1 suggests that the choice $\sigma_T = O((\ln \ln T)^{-1})$ is acceptable. Given the very slow rate involved, it is reasonable to consider ad hoc data-based values following the practice established by Orr (1995). Accordingly, in practice this tuning parameter is set such that $\sigma_T = \sigma$ where $\sigma = 2 \max_t |x_t - x_{t-1}|$.

Remark 3 *The choice of the initial set of RBF nodes given by:*

$$\Psi^{(1,\dots,T)} = \{\psi(x, x_1, \sigma_T), \psi(x, x_2, \sigma_T), \dots, \psi(x, x_T, \sigma_T)\}$$

may be straightforwardly generalised to $\Psi^{(1,\dots,p_T)}$ where p_T is chosen to reflect a subset of the observations or possibly be of a larger order than T . Theorem 1 allows under appropriate conditions both cases. Therefore, in the ensuing theoretical analysis p_T is left unspecified as long as $p_T \rightarrow \infty$.

Remark 4 *Algorithm 1 is more computationally demanding than that used in Blake and Kapetanios (2007, 2000, 2003a,b). There the nodes are ranked only once according to their ability to reduce the residual variance, when entered individually in (2). Clearly, Algorithm 1 is likely to provide a better fit than the approach of Blake and Kapetanios (2007, 2000, 2003a,b), although the two algorithms are very similar. The cost is a potential increase in computational effort of the order of $T(T+1)/2$. In practice this is likely to be substantially less as the stopping rule, m , will limit the number of nodes added and halt the computational task.*

Remark 5 *Although the discussion in this paper is couched in terms of RBFNNs it is worth noting that extensions to other neural network specifications such as neural networks based on logistic function nodes are possible once a grid of possible parameter values is constructed. One such specification is considered in White (2006) where an algorithm is constructed but no formal theoretical justification for it is given. The advantage of RBFNNs, in the context of Algorithm 1, is the fact that the construction of the grid is obtained by using the actual sample observations thus ensuring an appropriate coverage of the relevant state space for the processes under consideration.*

3 Theoretical Results

In this section we present our main theoretical result. The following assumptions will be needed.

Assumption 1 $E|\epsilon_t^s| < \infty$ for some $s > \max(2/\xi, 4)$ where ξ is defined in Lemmas 1 and 2 in the appendix.

Assumption 2 $\mu(\cdot)$ is L_2 -bounded.

Assumption 3 Either of the following assumptions hold: (i) Let \mathcal{F}_t be the Borel field generated by $(x_1, \epsilon_0), \dots, (x_t, \epsilon_{t-1})$. The sequence $\{\epsilon_t\}_{t=-\infty}^{\infty}$ is a martingale difference sequence with $E(\epsilon_t|\mathcal{F}_t) = 0$, $E(\epsilon_t^2|\mathcal{F}_t) = \sigma^2(x_{t-1})$ where $\sigma(\cdot)$ is continuous and bounded away from zero. (ii) $\{\epsilon_t\}_{t=-\infty}^{\infty}$ is a zero mean sequence with finite variance σ^2 . $\{x_t\}_{t=-\infty}^{\infty}$ and $\{\epsilon_t\}_{t=-\infty}^{\infty}$ are independent sequences.

Assumption 4 x_t is a stationary vector L_2 -NED (near epoque dependent) process of size -3 on some α mixing process η_{1t} of size $-C$, $C > 1$. ϵ_t is a stationary L_2 -NED process of size -3 on some α mixing process η_{2t} of size $-C$, $C > 1$. $p_T = o(T^{1/4})$.

Assumption 5 x_t and ϵ_t are a stationary vector and stationary scalar α -mixing processes with α -mixing coefficients given by $\alpha(k) = C_1 C_2^k$, $C_1 > 0$, $0 < C_2 < 1$. $p_T = O(T^{C_3})$ for some $C_3 > 0$.

Remark 6 Assumptions 4 and 5 provide alternative dependence structures for x_t and ϵ_t . Note the dependence of the rate for p_T on these dependence assumptions. Assumptions 4 is much weaker: firstly because it does not assume mixing and second because the mixing process on which x_t and ϵ_t depend, have α -mixing coefficients which decline at a polynomial rate rather than the exponential rate of assumption 5. The stronger dependence assumption 5 however allows for a much faster rate of increase in p_T .

Then, the following theorem proved in the appendix holds:

Theorem 1 *Let assumptions 1-3 and assumption 4 or assumption 5 hold. The estimate of the regression function $\mu(x_t)$, obtained using the iterative boosting algorithm 1 and denoted $\hat{\mu}(x_t)$, satisfies $\hat{\mu}(x_t) - \mu(x_t) = o_p(m^{-1/C_1})$, for all $C_1 > 6$ and some sequence $\sigma_T = o(1)$, if $m < \log_a T$, for all a that satisfy $\log_a e < \frac{\ln(5/2)}{4}$ if the conditions of Lemma 1 are satisfied. If further, the conditions of Lemma 2 are satisfied then $m < \log_a T$, for all a that satisfy $\log_a e < \frac{\ln(5/2)}{2}$. As a by-product of this estimation, an ordering of the radial basis function neural network nodes is obtained.*

To the best of our knowledge, this theorem provides the first consistency and rate result for a boosting algorithm in the context of neural networks for dynamic models.

Remark 7 *The rate of convergence to the true unknown regression function μ , given in Theorem 1, is rather sharp. Not all logarithmic rates are accommodated. The nature of the logarithmic rates allowed depends crucially on the dependence assumption made about x_t and ϵ_t as well as the tail behaviour of ϵ_t as we can see from the conditions of Lemma 2.*

4 Application to Nonparametric Specification Testing

In this section we provide an application of the result of Theorem 1 in the context of nonparametric specification testing following Guay and Guerre (2006). Let the true model be given by (1). Then, a hypothesis of interest is that $\mu(\cdot)$ belongs to some parametric family $\{m(\cdot; \theta), \theta \in \Theta \subset \mathbb{R}^w\}$. The null hypothesis then becomes

$$H_0 : \mu(\cdot) = m(\cdot; \theta) \text{ for some } \theta \in \Theta$$

Assuming the existence of some estimator $\hat{\theta}_T$ for θ obtained by some estimation method such as, e.g., nonlinear least squares, a set of residuals, \hat{u}_t is obtained. Then,

the null hypothesis may be tested by testing for the presence of some function of x_t , say $\mu_1(x_t)$ in a regression model of the residuals. Guay and Guerre (2006) use a trigonometric based series expansion for this purpose. We suggest use of an RBFNN expansion along the lines of the previous section. The rest of the testing framework of Guay and Guerre (2006) is retained. Once an ordered set of m RBFNN nodes is available via algorithm 1, this set is used in place of the set of trigonometric functions. Guay and Guerre (2006) suggest the use of a data dependent method to determine the final number of nodes to enter in the testing regression. This method depends on a penalty term of order $(\ln \ln T)^{1/2}$ to counterbalance the increase in fit from the use of more nodes in the testing regression. This is similar to the method adopted in Blake and Kapetanios (2007, 2000, 2003a,b) to construct various specification tests. The penalty terms used in Blake and Kapetanios (2003b) are the ones associated with either the Akaike or the Bayesian information criteria. These penalties are not optimal in the sense of Guay and Guerre (2006) since the Akaike penalty term results in a test which does not have an asymptotic χ^2 approximation whereas the Bayesian criterion, with a penalty term of order $\ln T$, is too parsimonious. In the context of the information criterion-based work of Blake and Kapetanios (2003b) the Hannan-Quinn criterion with a penalty term of order $\ln \ln T$ seems a more appropriate choice. Note that Guay and Guerre (2006) allow for the minimum number of nodes to be of order $(\ln T)^C$ $C > 0$ which for $0 < C < 1$ is acceptable according to Theorem 1, whereas they allow for a polynomial order for the maximum number of nodes which is not available for the RBFNN approximation since the number of nodes that can be ordered via algorithm 1 is of logarithmic order of magnitude. Below we provide a formal justification for using the RBFNN approximation in the framework of Guay and Guerre (2006).

For this section the assumptions of Section 3 are augmented and superseded where

appropriate by the following assumptions.

Assumption 6 x_t and ϵ_t are a stationary vector and stationary scalar α -mixing processes with α -mixing coefficients given by $\alpha(k) = C_1 k^{-C_2}$, $C_1 > 0$, $C_2 > 1$. $p_T = o(T^{1/4})$.

Assumption 7 $E(\epsilon_t^8) < \infty$.

Assumption 8 x_t has a density $f(\cdot)$ which is bounded away from zero and infinity.

Assumption 9 The parameter set Θ is a subset of \mathbb{R}^p and the following conditions hold. (i) The regression function $m(x; \theta)$ is twice continuously differentiable with bounded first and second derivatives. (ii) For any L_2 -bounded function $\mu(\cdot)$, there exists a parameter sequence, θ_T in Θ such that $T^{1/2}(\hat{\theta}_T - \theta_T) = O_p(1)$, with $\theta_T = \theta$ if $\mu(\cdot) = m(\cdot, \theta)$ for some θ in Θ .

Assumption 10 For the regression model $y_t = \mu(x_t) + \epsilon_t$, $\sup |\hat{\sigma}(x) - \sigma(x)| = O_p(v_T)$ and all $d/2$ derivatives of $\hat{\sigma}(x)$ are bounded from above by v_T , where $v_T = o(T^{1/C})$, for some $C > 0$.

Remark 8 Assumption 6 is considerably weaker than assumption 5 but considerably stronger than assumption 4. Assumption 7 strengthens assumption 1. Assumptions 9 and 10 are taken almost verbatim from Guay and Guerre (2006) and are technical ones needed to prove Theorems 1-3 of that paper.

Then, the following theorem, proved in the appendix, holds.

Theorem 2 Under assumptions 1-3 (i) and assumptions 6-10, the results of Theorems 1 and 3 of Guay and Guerre (2006) hold with the rate of polynomial approximation of the unknown regression function via the series expansion changed from

$-s/d$ to C , $C < 1/6$ where s is the Holder smoothness order of the unknown regression function and the maximum allowable rate of growth for the maximum allowable number of nodes, K_{\max} , changed from a polynomial rate in T to a logarithmic rate as described in Theorem 1.

Remark 9 *It is also straightforward to see that a version of Theorem 2 of Guay and Guerre (2006) holds. In particular, since only a logarithmic rate is allowed for the number of nodes in the testing regression, the RBFNN based nonparametric specification test cannot detect polynomially small local alternatives but only logarithmically small ones, unlike the trigonometric based nonparametric specification test of Guay and Guerre (2006).*

5 Conclusions

The use of series expansions as flexible nonlinear specifications for a variety of estimation and testing problems in statistics and econometrics is widespread. Limits to their use arise because many series expansions consist of basis functions that contain parameters. These parameters need to be somehow estimated. This necessitates the use of iterative techniques with the attendant computational and robustness costs.

On the other hand such series expansions give rise to methods that have excellent small sample properties as the work of Blake and Kapetanios (2007, 2000, 2003a,b) suggests. This paper formalises the methodology adopted in these papers and shows that it can provide a consistent estimate of an unknown regression function. A result on the rate of the approximation is also obtained. Use of theory on boosting algorithms is used in the process of deriving these results. The paper concludes with an application of the theoretical result to nonparametric specification testing.

A Proofs

A.1 Lemmas

The following two lemmas are needed for the main results.

Lemma 1 *Under Assumptions 1-3 and assumption 4 or assumption 5, with $0 < \xi < 1/2$:*

$$\sup_{1 \leq j, k \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) g_k(x_t) - E(g_j(x_t) g_k(x_t)) \right| = O_p(T^{-\xi/2}), \quad (3)$$

$$\sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \epsilon_t \right| = O_p(T^{-\xi/2}), \quad (4)$$

$$\sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T f(x_t) g_j(x_t) - E(f(x_t) g_j(x_t)) \right| = O_p(T^{-\xi/2}) \quad (5)$$

and

$$\sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) y_t - E(g_j(x_t) y_t) \right| = O_p(T^{-\xi/2}), \quad (6)$$

where g_j , $j = 1, \dots, p_T$ are bounded continuous functions, $f(\cdot) = \sum_{j=1}^{p_T} \beta_j g_j(\cdot)$ and $\sum_{j=1}^{p_T} |\beta_j| = o(T^{1/s})$ for all $s > 0$.

Proof of Lemma 1. For (3) we need to consider the different implications of assumptions 4 and 5 for the bounded quantity $g_j(x_t)$, $j = 1, \dots, p_T$. We start with assumption 4 which allows a greater extent of temporal dependence in x_t at the expense of a slower rate of increase in p_T . Let the autocovariance function of $g_j(x_t) g_k(x_t)$ be denoted by $c_{jk, \tau}$. By Markov's inequality it easily follows that

$$\Pr \left(\left| \sum_{i=1}^T g_j(x_t) g_k(x_t) - E(g_j(x_t) g_k(x_t)) \right| > \epsilon \right) \leq \frac{2T}{\epsilon^2} \sum_{\tau=0}^T c_{jk, \tau}. \quad (7)$$

We examine the behaviour of the RHS of (7). By assumption 4 x_t is an $L_2 - NED$ process of size -3 . By the fact that $g_j(\cdot)$ is a bounded function for all j , $g_j(\cdot)$ satisfies the following uniform Lipschitz condition for all finite constant vectors a and b and some finite constant scalar C

$$|g_j(a) - g_j(b)| \leq C \rho(a, b),$$

where $\rho(a, b) = \sum_{i=1}^d |a_i - b_i|$. Then, by Theorem 17.12 of Davidson (1994) it follows that $g_j(x_t)$ is an $L_2 - NED$ process of size -3 . Note that $\|g_j(x_t)\|_r \leq \infty$ for all finite r . Then, by example 17.17 of Davidson (1994), $g_j(x_t)g_k(x_t)$ is an $L_2 - NED$ process of size $-3(r-2)/2(r-1)$ for all finite r which implies that it is an $L_2 - NED$ process of size $-3/2$. Then, by Theorem 17.7 of Davidson (1994), and the mixing assumption in assumption 4, $\sum_{\tau=0}^T c_{jk,\tau} < \infty$ since the Theorem requires a NED size of -1 . Setting $\epsilon = T\epsilon$ in (7) gives a rate of convergence to zero for the RHS of (7) of T^{-1} . Now, (3) holds if

$$\Pr \left(T^{\xi/2} \sup_{1 \leq j, k \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t)g_k(x_t) - E(g_j(x_t)g_k(x_t)) \right| > \epsilon \right) = o(1), \quad (8)$$

for all $0 < \xi < 1/2$. But

$$\Pr \left(T^{\xi/2} \sup_{1 \leq j, k \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t)g_k(x_t) - E(g_j(x_t)g_k(x_t)) \right| > \epsilon \right) \leq \quad (9)$$

$$p_T^2 \Pr \left(T^{\xi/2} \left| T^{-1} \sum_{i=1}^T g_j(x_t)g_k(x_t) - E(g_j(x_t)g_k(x_t)) \right| > \epsilon \right).$$

Using (7), the RHS of (9) is majorised by $Cp_T^2 T^{\xi-1}$. But, by assumption 4, $p_T = o(T^{1/4})$. Hence, $Cp_T^2 T^{\xi-1} = o(1)$ and (8) follows. We now prove (3) under assumption 5. Now, p_T is allowed to grow at a faster polynomial rate than $1/4$ but this implies that the rate obtained in (7) is too slow. We therefore make use of Bernstein's inequality. Theorem 3.3 of White and Wooldridge (1991) gives a Bernstein inequality allowing for α -mixing stationary x_t . Using Theorem 3.49 of White (1999), we note that if x_t is α -mixing of a given size then $g_j(x_t)g_k(x_t)$ is also α -mixing of the same size. Then, noting that $g_j(x_t)g_k(x_t)$ has a finite upper bound, we get from Theorem 3.3 of White and Wooldridge (1991) that, for some finite constants C_1 and C_2 and for all $0 < \beta < 1$

$$\Pr \left(\left| \sum_{i=1}^T g_j(x_t)g_k(x_t) - E(g_j(x_t)g_k(x_t)) \right| > \epsilon \right) \leq C_1 \exp \left(-C_2 \epsilon T^{-\frac{1}{2}} \right). \quad (10)$$

Setting $\epsilon = T^{1-\xi/2}\epsilon$ we get

$$p_T^2 \Pr \left(T^{\xi/2} \left| T^{-1} \sum_{i=1}^T g_j(x_t) g_k(x_t) - E(g_j(x_t) g_k(x_t)) \right| > \epsilon \right) \leq C_1 p_T^2 \exp(-C_2 \epsilon T^{1/2-\xi/2}).$$

Thus,

$$C_1 p_T^2 \exp(-C_2 \epsilon T^{1/2-\xi/2}) = o(1),$$

for all $\epsilon > 0$, as long as $p_T = O(T^q)$ for all $q > 0$.

We now consider (4). Once again we consider the different implications of assumptions 4 and 5. Starting with assumption 4 we note that by the Markov inequality we get

$$\Pr \left(\left| \sum_{i=1}^T g_j(x_t) \epsilon_t \right| > \epsilon \right) \leq \frac{2T}{\epsilon^2} \sum_{\tau=0}^T c_{j,\tau}, \quad (11)$$

where $c_{j,\tau}$ denotes the autocovariance function of $g_j(x_t)\epsilon_t$. If assumption 3 (i) holds then all autocovariances are trivially zero and so $\sum_{\tau=0}^T c_{j,\tau} < \infty$. We now examine the situation under assumption 3 (ii). A difference between the treatment of $g_j(x_t)g_k(x_t)$ and $g_j(x_t)\epsilon_t$ arises since $g_j(x_t)\epsilon_t$ is not bounded. By assumption 4, $\|\epsilon_t\|_r \leq \infty$ for some $r > 4$. Since both $g_j(x_t)$ and ϵ_t are $L_2 - NED$ processes of size -3 it follows by example 17.17 of Davidson (1994), that $g_j(x_t)\epsilon_t$ is an $L_2 - NED$ process of size $-3(r-2)/2(r-1)$ for $r > 4$. Hence, $g_j(x_t)\epsilon_t$ is an $L_2 - NED$ process of, at most, size -1 . Thus, by Theorem 17.7 of Davidson (1994), and mixing assumption part of assumption 4, $\sum_{\tau=0}^T c_{j,\tau} < \infty$. Thus

$$\Pr \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \epsilon_t \right| > \epsilon \right) \leq p_T \Pr \left(T^{\xi/2} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \epsilon_t \right| > \epsilon \right). \quad (12)$$

Then, the RHS of (12) is majorised by $C p_T T^{\xi-1}$ which, by assumption 4, that $p_T = o(T^{1/4})$, is $o(1)$. Next, we establish (4) under assumption 5. Direct use of Bernstein's inequality for dependent processes is not possible as it applies to bounded random variables. So we use a truncation argument to get the inequality we need. Let

$$\tilde{\epsilon}_t = \begin{cases} \epsilon_t, & \text{if } |\epsilon_t| \leq C_T \\ \text{sign}(C_T)C_T, & \text{if } |\epsilon_t| > C_T \end{cases}$$

where C_T is a sequence to be defined below. We have that

$$\begin{aligned} & \Pr \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \epsilon_t \right| > \epsilon \right) \leq \\ & \Pr \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \tilde{\epsilon}_t - E(g_j(x_t) \tilde{\epsilon}_t) \right| > \epsilon/3 \right) + \\ & \Pr \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) (\epsilon_t - \tilde{\epsilon}_t) \right| > \epsilon/3 \right) + \\ & I \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T E(g_j(x_t) (\epsilon_t - \tilde{\epsilon}_t)) \right| > \epsilon/3 \right), \end{aligned} \quad (13)$$

where $I(\cdot)$ denotes the indicator function. We look, in turn, at the three terms on the RHS of (13). For the first term we can use the Bernstein inequality of (10). However, it takes the slightly different form below since $\tilde{\epsilon}_t$ is not bounded by a constant but by C_T .

$$\Pr \left(\left| \sum_{i=1}^T g_j(x_t) \tilde{\epsilon}_t - E(g_j(x_t) \tilde{\epsilon}_t) \right| > \epsilon \right) \leq C_1 \exp \left(\frac{-C_2 \epsilon T^{-\frac{1}{2}}}{C_T} \right).$$

Then, setting $\epsilon = T^{1-\xi/2} \epsilon$ and $C_T = T^{\xi/2}$, we have that

$$\begin{aligned} & \Pr \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \tilde{\epsilon}_t - E(g_j(x_t) \tilde{\epsilon}_t) \right| > \epsilon/3 \right) \leq \\ & C_1 p_T \exp(-C_2 \epsilon T^{1/2-\xi}) = o(1). \end{aligned}$$

We next look at the second term of the RHS of (13). We have that

$$\begin{aligned} & \Pr \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) (\epsilon_t - \tilde{\epsilon}_t) \right| > \epsilon/3 \right) \leq \Pr(\exists t \text{ such that } |\epsilon_t| > C_T) \leq \\ & T \Pr(|\epsilon_t| > C_T) \leq T \frac{E|\epsilon_t|^s}{C_T^s}. \end{aligned}$$

But

$$T \frac{E|\epsilon_t|^s}{C_T^s} = O(T^{1-s\xi/2}) = o(1)$$

since, by assumption 1, $s > 2/\xi$. Finally, we consider the third term of the RHS of (13). By the uncorrelatedness of $g(x_t)$ and ϵ_t and the boundedness of $g_j(x_t)$ the term can be bounded by

$$T^{\xi/2} E(g_j(x_t) (\epsilon_t - \tilde{\epsilon}_t)) \leq T^{\xi/2} E g_j(x_t) E(\epsilon_t - \tilde{\epsilon}_t) \leq C |E(\epsilon_t - \tilde{\epsilon}_t)|.$$

This can be bounded by

$$\begin{aligned}
|E(\epsilon_t - \tilde{\epsilon}_t)| &\leq \int I(|x| > C_T) (C_T + |x|) dP_\epsilon(x) = \\
&C_T \Pr(|\epsilon_t| > C_T) + \int |x| I(|x| > C_T) dP_\epsilon(x) \leq \\
C_T^{1-s} E|\epsilon_t|^s + (E|\epsilon_t|^2)^{1/2} \Pr(|\epsilon_t| > C_T)^{1/2} &= o(C_T^{-2}) = o(T^{-\xi}).
\end{aligned}$$

This proves (4) under both assumptions 4 and 5. We next consider (5),

$$\begin{aligned}
\sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T f(x_t) g_j(x_t) - E(f(x_t) g_j(x_t)) \right| &\leq \\
\sum_{j=1}^{p_T} |\beta_j| \sup_{1 \leq j, k \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) g_k(x_t) - E(g_j(x_t) g_k(x_t)) \right| &\leq \\
\left(\sum_{j=1}^{p_T} |\beta_j| \right) O_p(T^{-\xi/2}). &
\end{aligned}$$

But $\sum_{j=1}^{p_T} |\beta_j| = o(T^{1/s})$ for all $s > 0$ hence giving the result. Finally, (6) easily follows from (4) and (5). ■

Lemma 2 *Let Assumptions 1-3 and assumption 5 hold and assume that $0 < \xi < b$ for any $1/2 < b < 1$. Further, assume that*

$$\Pr(|\epsilon_t| > a) \leq C_1 \exp(-C_2 a^p)$$

where $C_1 > 0$, $C_2 > 0$ and $p > 1$. Then, Lemma 1 holds.

Proof of Lemma 2. The result of the Lemma will be established if we show that (4) holds under the conditions of the Lemma since it is easy to see from the proof of Lemma 1 that (3) and (5) follow under the conditions of the current Lemma. We revisit the analysis of the Bernstein inequality for unbounded random variables used in Lemma 1. A different truncation argument is then used. Let

$$\tilde{\epsilon}_t = \begin{cases} \epsilon_t, & \text{if } |\epsilon_t| \leq C_T \\ 0, & \text{if } |\epsilon_t| > C_T \end{cases} \quad \text{and} \quad \bar{\epsilon}_t = \begin{cases} 0, & \text{if } |\epsilon_t| \leq C_T \\ \epsilon_t, & \text{if } |\epsilon_t| > C_T \end{cases}$$

where C_T is a sequence to be defined below. Then,

$$\Pr \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \epsilon_t \right| > \epsilon \right) \leq \quad (14)$$

$$\begin{aligned} & \Pr \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \tilde{\epsilon}_t - E(g_j(x_t) \tilde{\epsilon}_t) + T^{-1} \sum_{i=1}^T g_j(x_t) \bar{\epsilon}_t - E(g_j(x_t) \bar{\epsilon}_t) \right| > \epsilon \right) \leq \\ & \Pr \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \tilde{\epsilon}_t - E(g_j(x_t) \tilde{\epsilon}_t) \right| > \epsilon/2 \right) + \\ & \Pr \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \bar{\epsilon}_t - E(g_j(x_t) \bar{\epsilon}_t) \right| > \epsilon/2 \right). \end{aligned}$$

For the first term of the RHS of (14), we can use Theorem 3.3 of White and Wooldridge (1991) to get

$$\Pr \left(\left| T^{-1} \sum_{i=1}^T g_j(x_t) \tilde{\epsilon}_t - E(g_j(x_t) \tilde{\epsilon}_t) \right| > \epsilon/2 \right) \leq C_1 p_T \exp \left(-C_2 \epsilon \frac{T^{1/2}}{C_T} \right).$$

Then,

$$\begin{aligned} & \Pr \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \tilde{\epsilon}_t - E(g_j(x_t) \tilde{\epsilon}_t) \right| > \epsilon/2 \right) \leq \\ & C_1 p_T \exp \left(-C_2 \epsilon \frac{T^{1/2-\xi/2}}{C_T} \right) \leq C_1 p_T \exp \left(-C_2 \epsilon \frac{T^{(1-b)/2}}{C_T} \right). \end{aligned}$$

We let $C_T = T^q$. It is clear that we need $(1-b)/2 > q$. Then,

$$C_1 p_T \exp \left(-C_2 \epsilon \frac{T^{(1-b)/2}}{C_T} \right) = o(1)$$

for all polynomial rates of growth for p_T . We next examine the second term of the RHS of (14). Using

$$\begin{aligned} & \Pr \left(T^{\xi/2} \sup_{1 \leq j \leq p_T} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \bar{\epsilon}_t - E(g_j(x_t) \bar{\epsilon}_t) \right| > \epsilon/2 \right) \quad (15) \\ & \leq p_T \Pr \left(T^{\xi/2} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \bar{\epsilon}_t - E(g_j(x_t) \bar{\epsilon}_t) \right| > \epsilon/2 \right) \end{aligned}$$

we focus on the RHS of (15). Then we have, using Markov's inequality,

$$\Pr \left(T^{\xi/2} \left| T^{-1} \sum_{i=1}^T g_j(x_t) \bar{\epsilon}_t - E(g_j(x_t) \bar{\epsilon}_t) \right| > \epsilon/2 \right) \leq$$

$$\Pr \left(T^{\xi/2-1} \sum_{i=1}^T |g_j(x_t)\bar{\epsilon}_t - E(g_j(x_t)\bar{\epsilon}_t)| > \epsilon/2 \right) \leq CT^{\xi/2-1} \sum_{i=1}^T E |g_j(x_t)\bar{\epsilon}_t|.$$

By the boundedness of $g_j(\cdot)$ and Holder's inequality,

$$\begin{aligned} E |g_j(x_t)\bar{\epsilon}_t| &= E |g_j(x_t)\epsilon_t I(|\epsilon_t| > C_T)| \leq \\ &C (E(|\epsilon_t|^p))^{1/p} (E(I(|\epsilon_t| > C_T))^u)^{1/u} = \\ &C (E(|\epsilon_t|^p))^{1/p} \Pr(|\epsilon_t| > C_T)^{1/u} \end{aligned}$$

where $p^{-1} + u^{-1} = 1$. Since $(E(|\epsilon_t|^p))^{1/p} < \infty$ and

$$\Pr(|\epsilon_t| > C_T) \leq C_1 \exp(-C_2 C_T^p),$$

it follows that

$$\begin{aligned} \Pr \left(T^{\xi/2-1} \sum_{i=1}^T |g_j(x_t)\bar{\epsilon}_t - E(g_j(x_t)\bar{\epsilon}_t)| > \epsilon/2 \right) &\leq CT^{\xi/2-1} \sum_{i=1}^T \exp(-C_2 C_T^p)^{1/u} \leq \\ &CT^{\xi/2-1} \sum_{i=1}^T \exp(-C_2 C_T^p) \leq CT^{\xi/2} \exp(-C_2 C_T^p) = \\ &CT^{\xi/2} \exp(-C_2 T^{qp}) = o(1), \end{aligned}$$

for all $p > 0$ and $q > 0$. Hence, the result follows. ■

A.2 Proofs of Theorems

Proof of Theorem 1. We split the problem in two parts: the approximation part and the estimation part. The approximation part relates to approximating $\mu(x)$ by a approximating function of the form

$$\psi(x; p_T) = \sum_{i=1}^{p_T} c_i \psi(x, t_i, \sigma_T), \quad (16)$$

where $\psi(x, t_i, \sigma_T)$ is a radial basis function node with centre t_i and radius σ_T . The first part of the approximation proof relates to the ability of sums of the form (16) to

approximate L_2 -bounded functions and the conditions required for such an approximation. For that we consider the work of Park and Sandberg (1991). Let

$$\tilde{\psi}(x; p_T) = \sum_{i=1}^{p_T} \bar{\psi}(x, t_i, \sigma_T) \mu^c(x) \left(\frac{2\tau_T}{p_T \sigma_T} \right)^d, \quad (17)$$

where

$$\bar{\psi}(x, t_i, \sigma_T) = \frac{\psi(x, t_i, \sigma_T)}{\int_{\mathbb{R}^r} \bar{\psi}(x, t_i, \sigma_T)}, \quad (18)$$

t_i is a partitioning of $[-\tau_T, \tau_T]^d$ such that all partition intervals are $o(p_T^{-C})$ for some $0 < C < 1$ and $\mu^c(x)$ is some continuous function that approximates arbitrarily well $\mu(x)$. This latter fact is possible since the space of continuous functions is dense in the space of L_2 bounded functions. Then, Park and Sandberg (1991) show that

$$\tilde{\psi}(x; p_T) - \mu(x) = o(1),$$

for all x not belonging to some set on \mathbb{R} of measure zero, as $p_T \rightarrow \infty$, $\sigma_T \rightarrow 0$ and $\tau_T \rightarrow \infty$. The latter two limits can have arbitrarily slow rates with respect to T . It is clear that (17) is of the form (16) with

$$c_i = \frac{\mu^c(x)}{\int_{\mathbb{R}^r} \bar{\psi}(x, t_i, \sigma_T)} \left(\frac{2\tau_T}{p_T \sigma_T} \right)^d.$$

So

$$\sup_{T \in \mathbb{R}} \sum_{i=1}^{p_T} |c_i| = o\left((\log_a T)^C\right), \quad (19)$$

for all $C > 0$, if σ_T and τ_T converge to zero and ∞ at slow enough rates; e.g. they behave as $(\ln \ln T)^{-1}$ and $\ln \ln T$ respectively. Finally, Girosi and Anzellotti (1993) show that the approximation has a rate of $p_T^{1/2}$. This concludes the first part of the proof.

The second part relates to the estimation part. Given the above approximation argument we now assume the existence of a representation of the form (16) for the regression function $\mu(x_t)$. We wish to estimate a representation of the form

$$\psi(x; m) = \sum_{i=1}^m c_i \psi(x, t_i, \sigma_T), \quad (20)$$

where t_i , $i = 1, \dots, m$ are centres that are obtained by some partition of $[-\tau_T, \tau_T]^d$, τ_T and σ_T are defined above, $m \rightarrow \infty$, $m = o(p_T)$ and, more importantly, order the centres, t_i via the boosting algorithm. To do that we use the framework of Buhlmann (2006). That framework is not directly applicable to our setting because it deals with independent observations. We therefore extend a number of results there to accommodate our needs. Let

$$\hat{\psi}(x; m) = \sum_{i=1}^m \hat{c}_{\mathcal{S}_i} \psi(x, t_{\mathcal{S}_i}, \sigma_T)$$

denote the estimated regression function after m iterations of the boosting algorithm where $(\mathcal{S}_1, \mathcal{S}_2, \dots)$ denotes the re-ordering of the centres $(1, 2, \dots)$ obtained by the boosting algorithm. Then, Theorem 1 of Buhlmann (2006) states that $\hat{\psi}(x; m)$ converges to $\psi(x; p_T)$ as $m \rightarrow \infty$ at a slow enough rate, i.e. $m = o(\ln T)$ and $p_T = O(e^{T^{1-\xi}})$. In order to use this result in our framework we need to (i) accommodate dependence in the data, (ii) allow for $\sup_{T \in \mathbb{R}} \sum_{i=1}^{p_T} |c_i| \rightarrow \infty$ and (iii) determine a rate at which $\hat{\psi}(x; m)$ converges to $\psi(x; p_T)$. We deal with each issue in turn. First we substitute Lemma 1 of Buhlmann (2006) with our Lemmas 1 and 2 which deal with dependent data. Secondly, we need to deal with the unboundedness of $\sup_{T \in \mathbb{R}} \sum_{i=1}^{p_T} |c_i|$. Note that it is sufficient for our results to only allow for a rate of growth of $\sup_{T \in \mathbb{R}} \sum_{i=1}^{p_T} |c_i|$ that is arbitrarily slow with respect to T . Accommodating this unboundedness can be done by examining Theorem 5.1 of Temlyakov (2000) which is used in (6.5) of Buhlmann (2006). Let the remainder function at the i -th step of the boosting algorithm for some original regression function f , be denoted by $R^i f$. Further, let b be defined by

$$\left| \langle R^{i-1} \psi(x; p_T), \psi(x, t_{\mathcal{S}_i}, \sigma_T) \rangle \right| \geq b \sup_{1 \leq j \leq p_T} \left| \langle R^{i-1} \psi(x; p_T), \psi(x, t_j, \sigma_T) \rangle \right|,$$

where $t_{\mathcal{S}_i}$ is the centre selected at the i -th step of the boosting algorithm. Then, by

Theorem 5.1 of Temlyakov (2000),

$$\|R^i\psi(x; p_T)\| \leq \left(\sup_{T \in \mathbb{R}} \sum_{i=1}^{p_T} |c_i| \right) (1+m)^{\frac{-b}{2(2+b)}}.$$

Then, if

$$(1+m)^{\frac{-b}{2(2+b)}} = O\left((\log_a T)^{\frac{-b}{2(2+b)}}\right),$$

it follows that by letting $\sup_{T \in \mathbb{R}} \sum_{i=1}^{p_T} |c_i|$ grow slowly enough, as in, e.g., (19),

$$\left(\sup_{T \in \mathbb{R}} \sum_{i=1}^{p_T} |c_i| \right) (1+m)^{\frac{-b}{2(2+b)}} = O\left((\log_a T)^{\frac{-b}{2(2+b)}}\right),$$

as well. Finally, we need to determine a rate at which $\hat{\psi}(x; m)$ converges to $\psi(x; p_T)$.

But

$$\left| \hat{\psi}(x; m) - \psi(x; p_T) \right| \leq \|R^m \mu\|. \quad (21)$$

Then, using (6.17), (6.19) of Buhlmann (2006) and considering sharper bounds in the analysis preceding (6.19) of Buhlmann (2006),

$$\|R^m \psi(x; p_T)\| \leq C_1 (1+m)^{\frac{-b}{2(2+b)}} + C_2 m (5/2)^m T^{-\xi} + C_3 C_4^m T^{-\xi}, \quad (22)$$

for all $C_4 > 5/2$, on the set A_T , where A_T denotes the set of events where (3)-(6) simultaneously occur. The sharper bounds referred to above relate to the following.

The third term of the RHS of (22) arises out of bounding $\left\| R^m \psi(x; p_T) - \tilde{R}^m \psi(x; p_T) \right\|$ where $\tilde{R}^m \psi(x; p_T)$ is a ‘semi’-population version of the remainder function R that uses population covariances rather than sample covariances. Buhlmann (2006) shows that

$$\left\| R^m \psi(x; p_T) - \tilde{R}^m \psi(x; p_T) \right\| \leq \left\| R^{m-1} \psi(x; p_T) - \tilde{R}^{m-1} \psi(x; p_T) \right\| + C_3 (5/2)^{m-1} T^{-\xi}.$$

Then

$$\left\| R^m \psi(x; p_T) - \tilde{R}^m \psi(x; p_T) \right\| \leq C_6 T^{-\xi} \sum_{i=1}^m (5/2)^{m-1} \leq C_6 T^{-\xi} (5/2)^m,$$

rather than

$$\left\| R^m \psi(x; p_T) - \tilde{R}^m \psi(x; p_T) \right\| \leq C_6 T^{-\xi} 3^m,$$

given in (6.19) of Buhlmann (2006). We note that in algorithm 1, b can be arbitrarily close to 1. Since by Lemmas 1 and 2 $\Pr(A_T) = 1 - O(T^{-\xi/8})$, it follows that

$$\|R^m \psi(x; p_T)\| \leq C_1 m^{-C_5} + C_3 C_4^m T^{-\xi}, \quad (23)$$

for all $C_5 > 6$ and all $C_4 > 5/2$. If Lemma 1 holds then if $m < \log_a T$, for all a that satisfies $\log_a e < \frac{\ln(5/2)}{4}$, it follows that there exists $C_4 > 5/2$ such that $C_4^m T^{-\xi/2} < T^{\log_a C_4 - \xi/2}$ and since $\xi < 1/2$, $\log_a C_4 - \xi/2 < 0$. If Lemma 2 holds then $\xi < 1$ and so a only needs to satisfy $\log_a e < \frac{\ln(5/2)}{2}$. Under these conditions, the second term of the RHS of (23) declines polynomially in T , whereas the first term declines at a slower logarithmic rate, in T , which therefore dominates. Overall

$$\|\hat{\mu}(x_t) - \mu(x_t)\| \leq \|R^m \psi(x; p_T)\| + \|\mu - \psi(x; p_T)\| = o_p(m^{-1/C_1}), \quad (24)$$

for all x_t and for all $C_1 > 6$, proving the theorem. Note that the above proof does not explicitly consider the possible heteroscedasticity of ϵ_t . However, the extension to this case follows easily upon noting Corrolary 1 of Buhlmann (2006) and the martingale difference assumption in assumption 3. ■

Proof of Theorem 2. The proof consists of showing that all conditions used in Theorems 1 and 3 of Guay and Guerre (2006) and therefore by extension, in the relevant parts of Propositions 1, 2 and Lemmas 1, A.1-A3 of the same paper, for the trigonometric series expansion, hold for the neural network expansion apart from the different polynomial approximation rate. These conditions, and the location of their use in the context of Guay and Guerre (2006), in parentheses, are (A1) uniform boundedness and orthonormality of the basis functions used to construct the approximation to the unknown regression function, (Lemmas A.1-A.3); (A2) The cardinality of the set of the possible number of nodes for the approximation should

be $\ln T$, (Lemma A.2); (A3) The series expansion approximates the unknown regression function at a polynomial rate (Lemma 1). (A2) and (A3) follow immediately from Theorem 1 and algorithm 1. We investigate (A1). The set of radial basis functions is uniformly bounded by definition for any radial basis function. However, the ordered set of functions arising out of the boosting algorithm is not orthonormal. Nevertheless, it can be made orthonormal using a number of possible orthonormalisation algorithms. We consider the Gram-Schmidt orthonormalisation algorithm. Let $\Psi_m = \{\psi(x, t_1, \sigma_T), \dots, \psi(x, t_m, \sigma_T)\}$ denote a set of radial basis functions used, in a regression, to approximate μ_1 . Let the transformed set of functions be denoted $\check{\Psi}_m = \{\check{\psi}(x, t_1, \sigma_T), \dots, \check{\psi}(x, t_m, \sigma_T)\}$ where $\check{\Psi}_m$ has been obtained from Ψ_m by Gram-Schmidt orthonormalisation as follows:

$$\check{\psi}(x, t_1, \sigma_T) = \frac{\psi(x, t_1, \sigma_T)}{\|\psi(x, t_1, \sigma_T)\|} \quad (25)$$

$$\check{\psi}(x, t_2, \sigma_T) = \frac{\psi(x, t_2, \sigma_T) - \left\langle \psi(x, t_2, \sigma_T), \check{\psi}(x, t_1, \sigma_T) \right\rangle \check{\psi}(x, t_1, \sigma_T)}{\left\| \psi(x, t_2, \sigma_T) - \left\langle \psi(x, t_2, \sigma_T), \check{\psi}(x, t_1, \sigma_T) \right\rangle \check{\psi}(x, t_1, \sigma_T) \right\|} \quad (26)$$

...

$$\check{\psi}(x, t_m, \sigma_T) = \frac{\psi(x, t_m, \sigma_T) - \sum_{i=1}^{m-1} \left\langle \psi(x, t_m, \sigma_T), \check{\psi}(x, t_i, \sigma_T) \right\rangle \check{\psi}(x, t_i, \sigma_T)}{\left\| \psi(x, t_m, \sigma_T) - \sum_{i=1}^{m-1} \left\langle \psi(x, t_m, \sigma_T), \check{\psi}(x, t_i, \sigma_T) \right\rangle \check{\psi}(x, t_i, \sigma_T) \right\|} \quad (27)$$

In order to prove the equivalence of using either Ψ_m or $\check{\Psi}_m$ in a regression to approximate μ_1 we simply note that for all i

$$\check{\psi}(x, t_i, \sigma_T) = \sum_{j=1}^i \check{c}_{ji} \psi(x, t_j, \sigma_T)$$

where the \check{c}_{ji} 's are determined in the recursions (25)-(27). Therefore,

$$\psi(x; m) = \sum_{i=1}^m \check{c}_i \check{\psi}(x, t_i, \sigma_T) = \sum_{i=1}^m \check{c}_i \left(\sum_{j=1}^i \check{c}_{ji} \psi(x, t_j, \sigma_T) \right) =$$

$$\sum_{i=1}^m \sum_{j=1}^i \check{c}_i \check{c}_{ji} \psi(x, t_j, \sigma_T) = \sum_{i=1}^m c_i \psi(x, t_i, \sigma_T)$$

where by grouping appropriate terms

$$c_i = \sum_{\ell=i}^m \check{c}_\ell \check{c}_{i\ell}$$

This completes the proof. ■

References

- AERTS, M., G. CLAESKENS, AND J. D. HART (1999): “Testing the Fit of a Parametric Function,” *Journal of the American Statistical Association*, 94, 869–879.
- BIERENS, H. J. (1984): “Model Specification Testing of Time Series Regression,” *Journal of Econometrics*, 26, 323–353.
- BISHOP, C. M. (1995): *Neural Networks for Pattern Recognition*. Oxford University Press.
- BLAKE, A. P., AND G. KAPETANIOS (2000): “A Radial Basis Function Artificial Neural Network Test for ARCH,” *Economics Letters*, 69, 15–23.
- (2003a): “Pure Significance Tests of the Unit Root Hypothesis Against Nonlinear Alternatives,” *Journal of Time Series Analysis*, 24(3), 253–267.
- (2003b): “A Radial Basis Function Artificial Neural Network Test for Neglected Nonlinearity,” *The Econometrics Journal*, 6(2), 357–373.
- (2007): “Testing for ARCH in the presence of nonlinearity of unknown form in the conditional mean,” *Forthcoming in the Journal of Econometrics*.
- BUHLMANN, P. (2006): “Boosting for high-dimensional linear models,” *Annals of Statistics*, 34, 559–583.

- CYBENKO, G. (1989): “Approximation by Superpositions of a Sigmoidal Function,” *Mathematics of Control, Signals and Systems*, 2, 304–314.
- DAVIDSON, J. (1994): *Stochastic Limit Theory*. Oxford University Press.
- DRUCKER, H., R. E. SCHAPIRE, AND P. Y. SIMARD (1993): “Boosting Performance in Neural Networks,” *International Journal of Pattern Recognition and Artificial Intelligence*, 7, 705–719.
- FREUND, Y., AND R. SCHAPIRE (1996): “Experiments with a New Boosting Algorithm,” *Machine Learning: Proc. 13th Intern. Conf. Morgan Kaufman*.
- FRIEDMAN, J. (2001): “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics*, 29, 1189–1232.
- FRIEDMAN, J., T. HASTIE, AND R. TIBSHIRANI (2000): “Additive Logistic Regression: A Statistical View of Boosting,” *Annals of Statistics*, 28, 337–374.
- GIROSI, F., AND G. ANZELLOTI (1993): “Rates of Convergence for Radial Basis Functions and Neural Networks,” in *Artificial Neural Networks for Speech and Vision*, ed. by R. J. Mammone. Chapman and Hall.
- GUAY, A., AND E. GUERRE (2006): “A Data-Driven Nonparametric Specification Test for Dynamic Regression Models,” *Econometric Theory*, 22, 543–586.
- GUERRE, E., AND P. LAVERGNE (2005): “Data-Driven Rate-Optimal Specification Testing in Regression Models,” *Annals of Statistics*, 33, 840–870.
- HART, J. D. (1997): *Nonparametric Smoothing and Lack-Of-Fit Tests*. Springer, New York.
- HORNIK, K., M. STINCHCOMBE, AND H. WHITE (1989): “Multi-Layer Feedforward Networks and Universal Approximators,” *Neural Network*, 2, 359–366.

- JONES, L. K. (1992): “A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training,” *Annals of Statistics*, 20, 608–613.
- LEE, T. H., H. WHITE, AND C. W. J. GRANGER (1993): “Testing for Neglected Nonlinearity in Time Series Models: A Comparison of Neural Network Methods and Alternative Tests,” *Journal of Econometrics*, 56, 269–290.
- NEWBY, W. K. (1997): “Convergence Rates and Asymptotic Normality for Series Estimators,” *Journal of Econometrics*, 79, 147–168.
- ORR, M. J. (1995): “Regularisation in the Selection of Radial Basis Function Centers,” *Neural Computation*, 7(3), 606–623.
- PAGAN, A., AND A. ULLAH (2000): *Nonparametric Econometrics*. Cambridge University Press.
- PARK, J., AND I. W. SANDBERG (1991): “Universal Approximation using Radial-Basis-Function Networks,” *Neural Computation*, 3(4), 246–257.
- SCHAPIRE, R. (2002): “The Boosting Approach to Machine Learning: An Overview,” in *MSRI Workshop on Nonlinear Estimation and Classification*, ed. by D. Denison, M. Hansen, C. Holmes, B. Mallick, and B. Yu. Springer.
- TEMLYAKOV, V. N. (2000): “Weak Greedy Algorithms,” *Advances in Computational Mathematics*, 12, 213–227.
- WHITE, H. (1999): *Asymptotic Theory for Econometricians*. Academic Press.
- (2006): “Approximate Nonlinear Forecasting Methods,” in *Handbook of Economics Forecasting*, ed. by G. Elliott, C. W. J. Granger, and A. Timmermann. Elsevier.

WHITE, H., AND J. WOOLDRIDGE (1991): "Some Results on Sieve Estimation With Dependent Observations," in *Nonparametric And Semiparametric Methods in Econometrics and Statistics*, ed. by W. Barnett, J. Powell, and G. Tauchen. Cambridge University Press.

**This working paper has been produced by
the Department of Economics at
Queen Mary, University of London**

**Copyright © 2007 George Kapetanios and Andrew P. Blake
All rights reserved**

**Department of Economics
Queen Mary, University of London
Mile End Road
London E1 4NS
Tel: +44 (0)20 7882 5096
Fax: +44 (0)20 8983 3580
Web: www.econ.qmul.ac.uk/papers/wp.htm**