VARYING DEGREE POLYNOMIAL REGRESSION

I. RATES OF CONVERGENCE*

(Abbreviated Title: VARYING DEGREE POLYNOMIAL REGRESSION)

by

Russell D. Wolfinger
SAS Institute Inc.

and

A. Ronald Gallant
North Carolina State University

November 1989**

AMS 1980 subject classifications. Primary 62J05; secondary 62E20, 62G05, 62F12.

Key words and phrases. Regression, nonparametric regression, polynomial regression, rates of convergence, Sobolev norm, metric entropy, M-estimator.

**This research was supported by National Science Foundation Grant SES-8808015, North Carolina Agricultural Experiment Station Projects NCO-5593, NCO-3879, and the NCSU PAMS Foundation.

SUMMARY

for

VARYING DEGREE POLYNOMIAL REGRESSION

I. RATES OF CONVERGENCE

by

Russell D. Wolfinger and A. Ronald Gallant

For a regression model $y_t = g^0(x_t) + e_t$, the unknown function g^0 is estimated consistently using varying degree polynomial regression. Consistency is with respect to a class of norms similar to weighted Sobolev norms, and the method of proof involves metric entropy calculations that enable the development of rates of convergence in these norms. The estimator is viewed as the solution to some member of a wide class of optimization problems, including those associated with least squares, maximum likelihood, and M-estimators; an M-estimator is used to illustrate the main results. Asymptotic normality is shown in a subsequent paper.

1. Introduction. Suppose we observe univariate data $\{y_t\}_{t=1}^n$ generated according to

$$y_t = g^o(x_t) + e_t \qquad 1 \le t \le n$$

where g^o is an unknown regression function possessing m derivatives, the x_t's are observed iid realizations from the beta(a,b) distribution (a and b are known), and the e_t's are unobserved realizations from some distribution $\mathcal{P}(e)$ and are independent of the x_t's. Our objective is to consistently estimate g^o in a norm including derivatives up to order $\ell < m$. Our approach is follow Example 1 of Cox (1988) and fit the parametric model

$$y_t = \sum_{j=1}^{p_n} \theta_j^o \varphi_j(x) + e_t \qquad 1 \le t \le n$$

where $\varphi_j(\cdot)$ is a multiple of the j^{th} Jacobi polynomial, θ_j^o is the j^{th} parameter, and p_n is some increasing function of n satisfying $p_n \le n$. Cox calls this procedure varying degree polynomial regression; it is also known as semi-nonparametric regression (Gallant, 1985; Wolfinger, 1989) and truncated series regression (Andrews, 1988). It is a member of the class of sieves (Grenander, 1980).

In the above context, varying degree polynomial regression is a competitor with nonparametric procedures such as kernel and spline estimators. Though in many cases these procedures may be preferable, polynomial regression does have advantages. First, as Cox mentions, p_n will often be much smaller than n, so the polynomial estimates can be viewed as being "simpler" than kernels or splines. Polynomial regression estimates also allow one to take advantage of knowledge about the design distribution (the beta distribution in our case).

If nothing else, this often results in a lack of boundary conditions which are often prevalent in the nonparametric procedures. Another advantage of varying degree polynomial regression over kernels and splines is that it appears to be more amenable to the incorporation of relevent scientific theory, as is the case with the series estimate. For example, from econometrics, it seems quite difficult to impose the conditional moment restrictions generated by the I-CAPM model (see Gallant and Tauchen, 1989) using kernel estimators. As for spline estimators, they do appear to be theoretically more suited to such an imposition, but their natural form is not in terms of deviations from the leading special case of the relevant theory. (For the above example, the leading special case is a VAR law of motion and a constant relative risk utility function.) Also, splines can be difficult to compute under complex constraints. The above comments are not meant to minimize the importance of kernels and splines, but simply to illustrate that they may not be as useful in certain instances as would a parametric technique possessing nonparametric properties.

Abusing notation slightly, let θ_n^o be the p_n-vector with j^{th} element θ_j^o. Cox estimates θ_n^o by least squares, and we extend his results by considering estimators that can be viewed as the solution to an optimization problem of the form

$$\underset{\theta \in \Theta_n}{\text{minimize:}} \quad (1/n) \sum_{t=1}^{n} s(y_t, x_t, \varphi_t'\theta)$$

where Θ_n is some subset of $\mathbb{R}^{p_n}$, φ_t is the p_n-vector with j^{th} element $\varphi_j(x_t)$, and $s(\cdot,\cdot,\cdot)$ is a suitable objective function. We call such an estimator a least mean distance estimator, with possible examples being least squares, maximum likelihood, and M-estimators. The theory for least mean distance estimators

with p_n bounded can be found in Chapter 3 of Gallant (1987); we thus extend these results to the case where p_n is increasing with n. Our goal is to find constraints on p_n that yield consistency for this entire class, and as such one would not expect to achieve any near optimal rates for any one member of the class. Nonetheless, we do make a comparison with the rates given in Yohai and Maronna (1979) and Portnoy (1984) for M-estimators. We also compare our rates with those that Cox obtains for least squares estimators.

We now describe the general framework and preliminary assumptions. We assume the true regression function, g^0, is m-times differentiable with $D^m g(x)$ absolutely continuous, where D^m is the m^{th} differentiation operator. We also assume that it has the expansion

$$g^0(x) = \sum_{j=1}^{\infty} \theta_j^0 \varphi_j(x)$$

where θ_j^0 is the generalized Fourier coefficients corresponding to the j^{th} basis function $\varphi_j(\cdot)$, defined by

$$\varphi_j(x) = c_{ab}(j) P_{j-1}^{(a-1,b-1)}(1-2x)$$

where $P_{j-1}^{(\alpha,\beta)}$, $\alpha,\beta > -1$, $j = 0,1,2, \ldots$, denote the Jacobi polynomials defined in Abramowitz and Stegun (1964) and Szego (1975). The normalizing constants c_{ab} are

$$c_{ab}^2(j) = \frac{(2j + a + b -3)\ \Gamma(j+ a + b - 2)\ \Gamma(j)\ \Gamma(a)\ \Gamma(b)}{\Gamma(j + a -1)\ \Gamma(j + b - 1)\ \Gamma(a + b)},$$

and can be shown to satisfy

$$c_{ab}(j) \approx j^{1/2}$$

where the symbol $\approx$ means that the r.h.s. can be bounded above and below by constant multiples of the l.h.s. Define a norm on g^0 by

$$\|g^0\|_{2m}^2 = \sum_{j=1}^{\infty} j^{2m} (\theta_j^0)^2$$

which we assume is finite. Note that this is precisely the $\|\cdot\|_{2m}$ norm defined by Cox; it is one member of his scale of norms that vary with powers of j in the above expression. Under the differentiability assumption above, he shows for all nonnegative even integers $\rho \leq 2m$

$$\|g^0\|_{\rho}^2 \approx \int_0^1 \{[g^0(x)]^2 + [D^{\rho/2} g^0(x)]^2 [x(1-x)]^{\rho/2}\} \beta(x)\, dx$$

where $\beta(x)$ is the beta(a,b) density. This norm is thus close to being a weighted Sobolev norm, and we prove our consistency results in $\|\cdot\|_{2\ell}$. Cox shows that the $\|\cdot\|_{2h+\epsilon}$ norm is stronger than the supremum norm, where

$$h = \max(a,b,1/2)$$

and ϵ is any positive real number. For this and other reasons, we assume that $\ell \geq h + 1/2$, and thus our results imply consistency in sup norm. We also use the usual Euclidean norm $\|\cdot\|_0$, which we write as $\|\cdot\|$.

REMARK 1.1. The above identification of $\|\cdot\|_{\rho}$ results from our choice of the beta(a,b) design distribution and the associated Jacobi polynomials multiplied by the normalizing constants. This a simplification of Cox's use of an "asymptotic" design distribution, but it appears that our results can be extended to this scenario. The extension to other design distributions such as the Gaussian (requiring Hermite polynomials) and the gamma (requiring Laguerre polynomials) also seems possible; but, as Cox mentions, a general theory is

desirable.

Define

$$\hat{\theta}_n = \underset{\theta \in \Theta_n}{\operatorname{argmin}} \, (1/n) \sum_{t=1}^{n} s(y_t, x_t, \varphi_t'\theta)$$

which we assume exists and is unique, and where

$$\Theta_n = \{\theta \in \mathbb{R}^{p_n} \mid \|\theta\|_{2\ell} \leq B\}$$

$$\|\theta\|_{2\ell} = \sum_{j=1}^{p_n} j^{2\ell} \theta_j^2$$

and $B = \|g^0\|_{2\ell} + 1$. B is thus unknown, but in practice this will be irrelevant because the choice of p_n will guarantee that $\hat{\theta}_n$ satisfies $\|\hat{\theta}_n\|_{2\ell} < B$ a.s. for n sufficiently large. As our estimate of $g^0(x)$, define

$$\hat{g}_n(x) = \varphi(x)'\hat{\theta}_n$$

where $\varphi(x)$ is the p_n-vector with j^{th} element $\varphi_j(x)$.

Next, define the matrix Φ as being the $n \times p_n$ matrix with rows $\varphi_t' = \varphi(x_t)'$. Note that $\Phi'\Phi$ is the usual $X'X$ regression matrix and that it is nonsingular wpl. We assume that a finite constant Λ exists such that

$$\lambda_{max}(\Phi'\Phi/n)^{-1} \leq \Lambda$$

for all n, where $\lambda_{max}(\cdot)$ denotes the maximum eigenvalue.

Finally, we assume that the objective function s is real valued and has the form $s[Y(e,x),x,g(x)]$, where $Y(e,x) = g^0(x) + e$ and g is some function of x. To be reasonable, s should be some measure of distance between $g_0(x)$ and $g(x)$, and to avoid measurability problems we assume that it is continuous in all

three of its arguments.

We now illustrate our main results with an example.

EXAMPLE (M-estimator). Use the objective function

$$s[Y(e,x),x,\varphi(x)'\theta] = \rho[g^0(x) + e - \varphi(x)'\theta]$$

where

$$\rho(u) = \log \cosh(u/2).$$

Define also

$$\Psi(u) = (d/du)\,\rho(u) = 1/2 \tanh(u/2).$$

REMARK 1.2. In future work we plan to use an iteratively rescaled M-estimator. This will entail the development of theory permitting the estimate of a nuisance parameter in the objective function. As this tends to clutter analysis, we omit it. For more detail concerning this robust estimator with bounded influence, see Huber (1964) and Gallant (1987).

Assume that the errors possess finite r^{th} moments, where $r > 2 + 1/\ell$, and that $\mathcal{E}\,\rho(e) = \mathcal{E}\,\Psi(e) = 0$. This would be satisfied if the error density is symmetric about zero.

Our results impose a trade off between the constraint on p_n and the rate of convergence in the $\|\cdot\|_{2\ell}$ norm; we analyze the two extremes. The first is to force p_n to increase as fast as possible and let the rate in the $\|\cdot\|_{2\ell}$ norm become very slow. If we set

$$\alpha = \frac{\ell - (1+\gamma)(1+2\ell)/r}{h + 2\ell(1+2\ell)} - \delta$$

for a sufficiently small $\gamma > 0$ and $\delta > 0$, and let $p_n \approx n^{\alpha}$, then

$$\|\hat{g}_n - g^o\|_{2\ell} = o(1) \quad \text{a.s.}$$

and provided $m \geq (1/\alpha + 2h + 1)/2$

$$\sum_{t=1}^{n} [\hat{g}_n(x_t) - g^o(x_t)]^2 = O_p(p_n)$$

conditionally on $\{x_t\}$. The latter rate of convergence is the same as that obtained by Yohai and Maronna (1979) and Portnoy (1984), but we must increase p_n much slower than the approximate $p_n = o(\sqrt{n})$ they require. Indeed, the largest possible α is 2/13 (corresponding to $\ell = 1$, $h = 1/2$, and r large). This is the price that we pay to obtain convergence in the strong $\|\cdot\|_{2\ell}$ norm.

The second extreme achieves the best rate of convergence in the $\|\cdot\|_{2\ell}$ norm. Unfortunately this entails slowing p_n to the extent that we lose the above $O_p(p_n)$ result. Nonetheless, assuming m is finite, we set

$$\alpha = \frac{\ell - (1+\gamma)(1+2\ell)/r}{h + 2m(1+2\ell)} - \delta$$

for a sufficiently small $\gamma > 0$ and $\delta > 0$, and let $p_n \approx n^{\alpha}$. The resulting rate of convergence is

$$\|\hat{g}_n - g^o\|_{2\ell} = o[n^{-\alpha(m-\ell) + \epsilon}] \quad \text{a.s.}$$

for a sufficiently small $\epsilon > 0$. For comparison, Cox (1988) obtains

$$p_n \approx n^{1/(2m+1)}$$

and

$$\mathcal{E}\, \|\hat{g}_n - g^o\|_{2\ell} = O[n^{-(m-\ell)/(2m+1)}]$$

when the rates on his variance and bias terms are balanced.

As a specific example, let $m = 3$, $\ell = 2$, $h = 1/2$, and $r = 8$. Then our results yield $\alpha = 11/244 < 1/22$, while Cox requires $p_n \approx n^{1/7}$ with a $O(n^{-1/7})$ rate of convergence. So in this case his rates are roughly 3 times better than ours, but keep in mind that his is a rate of mean square convergence, while ours is almost sure convergence. Also, his results are only for least squares estimators, while ours are for more general optimizers. Both Cox's and our rate constraints on p_n need only be enforced up to a constant multiplier, so in practice they may not differ considerably.

The remainder of the paper is organized as follows. In Section 2 we list our primary assumptions, and in Section 3 we state our main results. These are then applied to the example in Section 4, proving the claims made above. The proofs of the main theorems are given in Section 5. The Appendix gives the proof of one of the lemmas stated in Section 3.

2. The assumptions. Our main assumptions are grouped into two categories: the objective function (Assumptions S1-S6), and rate constraints (Assumptions P1-P5).

ASSUMPTION S1. A real number $r > 2 + 1/\ell$ and a constant R exist such that

$$\mathcal{E} \sup_{\theta \in \Theta_n} s^r[Y(e,x),x,\varphi(x)'\theta] \leq R < \infty \quad \text{for all } n.$$

ASSUMPTION S2. A sequence of positive Borel-measurable functions $\{m_n(\cdot,\cdot)\}$ exists such that

$$|s[Y(e,x),x,\varphi(x)'\theta_1] - s[Y(e,x),x,\varphi(x)'\theta_2]| \leq m_n(e,x)\, \|\theta_1 - \theta_2\|$$

for all e, x, and for all $\theta_1, \theta_2 \in \Theta_n$. A real number $q \geq r$ and a sequence of constants $\{M_n\}$ exist such that

$$\mathcal{E}\, m_n^q(e,x) \leq M_n^q < \infty \quad \text{for all } n.$$

For the next assumption, we define some new notation. First, let

$$g_n^o(x) = \sum_{j=1}^{p_n} \theta_j^o\, \varphi_j(x)$$

be the truncated counterpart of g^o. Next, we write $s(\theta)$ as an abbreviation for $s[Y(e,x),x,\varphi(x)'\theta]$. Also, define

$$\theta_n^* = \operatorname*{argmin}_{\theta \in \Theta_n} \mathcal{E}\, s(\theta)$$

which we assume exists and is unique.

ASSUMPTION S3. A real-valued function $L(\ell,p_n)$ exists such that

$$\|\hat{g}_n - g_n^o\|_{2\ell}^2 = O[L(\ell,p_n)]\ \mathcal{E}\ \{[s(\hat{\theta}_n) - s(\theta_n^*)] + [g^o - g_n^o]^2\} \quad \text{a.s.}$$

where by $\mathcal{E}\, s(\hat{\theta}_n)$ we mean $\mathcal{E}\, s(\theta)$ evaluated at $\theta = \hat{\theta}_n$, i.e., the expectation operator is not applied to $\hat{\theta}_n$.

ASSUMPTION S4. $(\partial/\partial g)s[g^0(x) + e,x,g]$ exists and is nondecreasing in g for all e and x.

ASSUMPTION S5. For every n, define the following function:

$$X_n(e,x) = (\partial/\partial g)s[g^0(x) + e,x,g]\Big|_{g = g_n^0(x)}.$$

Then constants v and c_L exist such that

$$\sup_{0<n<\infty} \sup_{x\in\mathcal{X}} \mathcal{E}_{\mathcal{P}}[X_n(e,x)]^2 \leq v < \infty$$

$$\sup_{x\in\mathcal{X}} [\mathcal{E}_{\mathcal{P}} X_n(e,x)]^2 \leq c_L \sup_{x\in\mathcal{X}} [g^0(x) - g_n^0(x)]^2$$

where $\mathcal{X} = (0,1)$ by $\mathcal{E}_{\mathcal{P}}$ we mean expectation with respect to $\mathcal{P}(e)$ only.

ASSUMPTION S6. Define the following function:

$$w(e,x,\theta,z) = (\partial^2/\partial g^2)s[g^0(x) + e,x,g]\Big|_{g = \varphi(x)'\theta - z}$$

which we assume exists. Then positive numbers a,b,c, and q (different from any used previously) exist such that if we define the set

$$A = \{e \in \mathcal{E},\ x \in \mathcal{X},\ \theta \in \Theta_n,\ z \in \mathbb{R} : |e| < a,\ |z| < b\}$$

then

$$\inf_A w(e,x,\theta,z) \geq c$$

and the constant a satisfies $\mathcal{P}(a) - \mathcal{P}(-a) = q$, where we are using $\mathcal{P}(\cdot)$ to denote the distribution function of the errors.

REMARK 2.1 Assumptions S1 and S2 are similar to Conditions (W) and (L) of Severini and Wong (1987). Assumption S3 provides a straightforward way of showing consistency in the $\|\cdot\|_{2\ell}$ norm. Note that consistency can be shown in any norm satisfying a similar assumption, provided the metric entropy calculation (given below) and the resulting rate constraints are appropriately changed. Assumptions S4-S6 borrow from Yohai and Maronna (1979).

ASSUMPTION P1. A real number α exists such that $0 < \alpha < 1$ and the truncation point p_n satisfies $p_n \approx n^{\alpha}$ and $p_n \leq n$ for every n.

For the next assumption, we need the following definition. For an arbitrary metric space Ξ with metric d, and for a positive real number ϵ, we define the metric entropy (or ϵ-entropy) $H(\epsilon,\Xi,d)$ as the natural logarithm of the minimum number of balls of d-radius ϵ needed to cover Ξ, i.e.

$$H(\epsilon,\Xi,d) = \log m_0$$

where m_0 is the smallest m such that there exist $(\xi_1,\ldots,\xi_m)$ in Ξ satisfying

$$\sup_{\xi\in\Xi} \min_{1\leq j\leq m} d(\xi,\xi_j) \leq \epsilon.$$

In this case the $(\xi_1,\ldots,\xi_{m_0})$ would be the centers of the covering balls.

ASSUMPTION P2. A real number β exists satisfying $0 < \beta < 1/2$ and a sequence of positive constants $\{\epsilon_n\}$ exists such that $\epsilon_n \approx n^{-\beta}$ and

$$H(2\epsilon_n/M_n,\Theta_n,\|\cdot\|) - n\epsilon_n^2 + (1+\epsilon)\log(n)$$

is bounded as $n \to \infty$ for some $\epsilon > 0$, where M_n is from Assumption S1.

ASSUMPTION P3. Using $L(\ell,p_n)$ from Assumption S3, p_n, ϵ_n, and m satisfy the following two constraints for some $\gamma > 0$:

$$\text{(i)} \quad \lim_{n\to\infty} L(\ell,p_n)\, \epsilon_n\, n^{(1+\gamma)/r} \quad \text{exists and is bounded.}$$

$$\text{(ii)} \quad \lim_{n\to\infty} p_n^{2m}\, \epsilon_n\, n^{(1+\gamma)/r} = \infty.$$

ASSUMPTION P4. The real number α from Assumption P1 satisfies

$$\alpha \geq 1/(2m - 2h - 1),$$

and m is large enough so that this lower bound on the growth rate of p_n does not conflict with any upper bound given in other assumptions.

ASSUMPTION P5. Define

$$B(p) = \sum_{j=1}^{p} \sup_{x\in\mathcal{X}} \varphi_j^2(x),$$

Then

$$\lim_{n\to\infty} p_n\, B(p_n)/n = 0.$$

REMARK 2.2. Assumptions P1-P3 represent the rate constraints we use to show consistency in the $\|\cdot\|_{2\ell}$ norm. The metric entropy used in Assumption P2 is computed in Lemma 3.2 below. Assumption P3 implies that $L(\ell,p_n)$ and m must satsify $L(\ell,p_n) = o(p_n^{2m})$. Assumption P4 provides a lower bound on the growth rate of p_n; it is needed in showing the $O_p(p_n)$ result. Concerning Assumption P5, a result from Cox shows that $B(p) = O(p^{2h})$ as $p \to \infty$. So the assumption is implied by $\lim_{n\to\infty} p_n^{2h+1}/n = 0$.

3. Statement of the main results.

LEMMA 3.1.

$$\sup_{0<n<\infty} \sup_{\theta\in\theta_n} \sup_{x\in\mathcal{X}} |g^0(x) - \varphi(x)'\theta| < \infty.$$

PROOF. Use the fact from Cox (1988) that

$$\sup_{x\in\mathcal{X}} |\varphi_j(x)| = O(j^{h-1/2})$$

and the fact that for all n the coefficients of any θ in θ_n must satisfy

$$\theta_j = o(j^{-\ell})$$

for all j. Then note that

$$\sup_{\theta\in\theta_n} \sup_{x\in\mathcal{X}} |\varphi(x)'\theta| = \sum_{j=1}^{p_n} o(j^{-\ell})\, O(j^{h-1/2})$$

$$= o(p_n^{h-\ell+1/2})$$

which will converge to zero as $n \to \infty$ as long as $\ell \geq h + 1/2$. $g^0(x)$ must also be bounded because $m > \ell$. ∎

LEMMA 3.2. *Assume $\ell \geq 2$, and let ϵ be a positive real number. Then a positive constant c_H exists such that*

$$H(\epsilon,\theta_n,\|\cdot\|) \leq c_H\,(2B/\epsilon)^{1/\ell}$$

for all n.

PROOF. See the appendix. ∎

THEOREM 3.3. *Under Assumptions S1-S2 and P1-P2,*

$$\sup_{\theta\in\theta_n} |\mathcal{E}_n s(\theta) - \mathcal{E}\, s(\theta)| = o(\epsilon_n n^{(1+\gamma)/r}) \quad a.s.$$

where $\mathcal{E}_n$ denotes integration by the empirical measure of $\{e_t,x_t\}_{t=1}^n$, and recall we are using $s(\theta)$ as an abbreviation for $s[Y(e,x),x,\varphi(x)'\theta]$.

COROLLARY. *Under the hypotheses of Theorem 3.3,*

$$|\mathcal{E}\, s(\hat{\theta}_n) - \mathcal{E}\, s(\theta_n^*)| \;=\; o(\epsilon_n n^{(1+\gamma)/r}) \quad a.s.$$

where by $\mathcal{E}\, s(\hat{\theta}_n)$ we mean $\mathcal{E}\, s(\theta)$ evaluated at $\theta = \hat{\theta}_n$, i.e., the expectation operator is not applied to $\hat{\theta}_n$.

PROOF. Using the fact that $\hat{\theta}_n$ and θ_n^* are the optimizers of $\mathcal{E}_n s(\theta)$ and $\mathcal{E}\, s(\theta)$ respectively,

$$\begin{aligned} |\mathcal{E}\, s(\hat{\theta}_n) - \mathcal{E}\, s(\theta_n^*)| &= [\mathcal{E}\, s(\hat{\theta}_n) - \mathcal{E}_n s(\hat{\theta}_n)] + [\mathcal{E}_n s(\hat{\theta}_n) - \mathcal{E}\, s(\theta_n^*)] \\ &\le o(\epsilon_n n^{(1+\gamma)/r}) + [\mathcal{E}_n s(\theta_n^*) - \mathcal{E}\, s(\theta_n^*)] \quad a.s. \\ &= o(\epsilon_n n^{(1+\gamma)/r}) \quad a.s. \quad \blacksquare \end{aligned}$$

THEOREM 3.4. *Under Assumptions S1-S3 and P1-P3,*

$$\|\hat{g}_n - g^o\|_{2\ell} \;=\; o[L(\ell,p_n)\, \epsilon_n\, n^{(1+\gamma)/r}]^{1/2} + O[p_n^{\ell-m}] \quad a.s.$$

LEMMA 3.5.

$$\sup_{x\in\mathcal{X}} |g^o(x) - g_n^o(x)| \;=\; o(p_n^{h-m+1/2}).$$

PROOF. Argue as in Lemma 3.1, using the fact that

$$\theta_j^o \;=\; o(j^{-m}). \quad \blacksquare$$

Define

$$h_t \;=\; (\Phi'\Phi)^{-1/2}\varphi_t$$

where $(\Phi'\Phi)^{1/2}$ is the Cholesky factor of $\Phi'\Phi$.

LEMMA 3.6. *Under Assumptions S5 and P4,*

$$P(\| \sum_{t=1}^{n} X_{nt}\, h_t \| \geq k) \;=\; 1/k^2\; O(p_n)$$

for any $k > 0$, where $X_{nt} = X_n(e_t, x_t)$ from Assumption S5.

PROOF. Define $M_{nt} = \mathcal{E}_{\mathcal{P}} X_{nt}$ and $\xi_{nt} = X_{nt} - M_{nt}$. Then

$$\mathcal{E}_{\mathcal{P}}[\| \sum_{t=1}^{n} X_{nt}\, h_t \|]^2 \;\leq\; 2\, \mathcal{E}_{\mathcal{P}}[\| \sum_{t=1}^{n} \xi_{nt}\, h_t \|]^2 \;+\; 2\, \| \sum_{t=1}^{n} M_{nt}\, h_t \|^2.$$

By independence and the first part of Assumption S5, the first term on the right hand side is bounded above by $v \sum_{t=1}^{n} \|h_t\|^2$ which equals vp_n by definition of $\{h_t\}$. Define M to be the n-vector $\{M_{nt}\}$ and H to be the $n \times p_n$ matrix with rows h_t', and let $\lambda_{max}(A)$ denote the maximum eigenvalue of a real symmetric matrix A. Then the second term on the right hand side equals

$$\begin{aligned} 2\, M'HH'M \;&\leq\; 2\, M'M\, \lambda_{max}(H'H) \\ &\leq\; 2\, n\, c_L \sup_{x \in \mathcal{I}} [g^o(x) - g_n^o(x)]^2 \end{aligned}$$

by the second part of Assumption S5 and the fact that H'H is the identity matrix. By Lemma 4 the final term is $o(np_n^{2h-2m+1})$, which by Assumption P4 is $o(1)$. Applying Markov's inequality yields the desired result. ∎

THEOREM 3.7. *Under Assumptions S1-S6 and P1-P5,*

$$\sum_{t=1}^{n} [\hat{g}_n(x_t) - g_n^o(x_t)]^2 \;=\; O_p(p_n).$$

4. Application to the example. We first verify Assumptions S1-S6 for the M-estimator.

Assumption S1 follows from the fact that $\rho(u) \leq 1/2|u|$, and then applying Lemma 3.1 and the moment assumption on the errors.

For Assumption S2, use Taylor's theorem and the Cauchy-Schwartz inequality:

$$\begin{aligned} |\rho(y - \varphi'\theta_1) - \rho(y - \varphi'\theta_2)| &\leq |-\Psi(y - \varphi'\bar{\theta})|\ |\varphi'(\theta_1 - \theta_2)| \\ &\leq |\Psi(y - \varphi'\bar{\theta})|\ \|\varphi\|\ \|\theta_1 - \theta_2\| \end{aligned}$$

where we are writing φ for $\varphi(x)$ and $\bar{\theta}$ is on the line segment joining θ_1 and θ_2. We can thus set

$$m_n(e,x) = \|\varphi(x)\|$$

since $|\Psi(\cdot)| \leq 1$. By a result from Cox, $\mathcal{E}\ \|\varphi(x)\|^2 = p_n$, and so

$$\mathcal{E}\ m_n^q(e,x) \leq p_n\ [B(p_n)]^{q/2-1}$$

where $B(\cdot)$ is defined in Assumption P5. We can thus set

$$M_n = p_n^{1/q}\ [B(p_n)]^{1/2-1/q}$$

and q can be arbitrarily large.

For Assumption S3, again use Taylor's theorem:

$$\begin{aligned} \mathcal{E}\ s(\hat{\theta}_n) &= \mathcal{E}\ \rho[g^0(x) + e - \varphi'\hat{\theta}_n] \\ &= \mathcal{E}\ \rho(e) + \mathcal{E}\ \Psi(e)\ [g^0(x) - \varphi'\hat{\theta}_n] \\ &\quad + \mathcal{E}\ \mathrm{sech}^2[g^0(x) + \lambda e - \varphi'\hat{\theta}_n]\ [g^0(x) - \varphi'\hat{\theta}_n]^2 \end{aligned}$$

for some λ in $[0,1]$. The first two terms in the final sum are zero by assumption, and the final term is bounded below by

$$c\ \mathcal{E}\ [g^0(x) - \varphi'\hat{\theta}_n]^2$$

for some positive constant c depending only on the error distribution. This is true because the argument of $\text{sech}^2(\cdot)$ is bounded above by $|e| + C$, where C is the supremum bound of $g^0(x) - \varphi'\theta$ given in Lemma 3.1. We thus have

$$\begin{aligned}\mathcal{E}\ s(\hat{\theta}_n) &\geq c\ \mathcal{E}\ [g^0(x) - \varphi'\hat{\theta}_n]^2 \\ &= c\ [\|\hat{\theta}_n - \theta_n^0\|^2 + \mathcal{E}\ [g_n^0(x) - g^0(x)]^2]\end{aligned}$$

Using the same Taylor series expansion and the fact that $\text{sech}^2(\cdot)$ is bounded above by 1, we also have

$$\begin{aligned}\mathcal{E}\ s(\theta_n^*) &\leq \mathcal{E}\ s(\theta_n^0) \\ &\leq \mathcal{E}\ [g_n^0(x) - g^0(x)]^2.\end{aligned}$$

Combining these results we obtain

$$\begin{aligned}|\mathcal{E}\ s(\hat{\theta}_n) - \mathcal{E}\ s(\theta_n^*)| &= \mathcal{E}\ s(\hat{\theta}_n) - \mathcal{E}\ s(\theta_n^*) \\ &\geq c\ \|\hat{\theta}_n - \theta_n^0\|^2 + (c - 1)\ \mathcal{E}\ [g_n^0(x) - g^0(x)]^2 \\ &\geq (c/p_n^{2\ell})\ \|\hat{g}_n - g_n^0\|_{2\ell}^2 + (c - 1)\ \mathcal{E}\ [g_n^0(x) - g^0(x)]^2.\end{aligned}$$

where we have used the crude inequality $\|\theta\|_{2\ell}^2 \leq p_n^{2\ell}\ \|\theta\|^2$. We can thus choose $L(\ell, p_n) = p_n^{2\ell}$.

Assumption S4 follows from the fact that $\Psi(\cdot)$ is nondecreasing.

For Assumption S5,

$$X_n(e,x) = -\Psi[g^0(x) - g_n^0(x) + e].$$

Now

$$\mathcal{E}_{\mathcal{P}}[X_n(e,x)]^2 \leq 1$$

for all n and x because $|\tanh(\cdot)|$ is bounded above by 1. Using the assumption that $\mathcal{E}_{\mathcal{P}}\Psi(e) = 0$, we have by Taylor's theorem

$$\begin{aligned} |\mathcal{E}_{\mathcal{P}}X_n(e,x)| &= |[g^0(x) - g_n^0(x)]\ \mathcal{E}_{\mathcal{P}}\operatorname{sech}^2[g^0(x) - g_n^0(x) + \tilde{e}]| \\ &\leq |g^0(x) - g_n^0(x)| \end{aligned}$$

because $\operatorname{sech}^2(\cdot)$ is bounded above by 1.

Assumption S6 follows from an argument similar to the one used in verifying Assumption S3. Here

$$\begin{aligned} w(e,x,\theta,z) &= (\partial^2/\partial g^2)s[g^0(x) + e,x,g]\Big|_{g = \varphi'\theta - z} \\ &= 1/4\ \operatorname{sech}^2\{[g^0(x) + e - \varphi'\theta + z]/2\}. \end{aligned}$$

Using the above mentioned argument, the quantitiy in the brackets can be bounded above by $|e| + |z| + C$, where C is some constant independent of x. So as long as we have positive constants a and q satisfying

$$\mathcal{P}(a) - \mathcal{P}(-a) = q$$

and b is an arbitrary positive number, then we can choose

$$c = 1/4\ \operatorname{sech}^2\{[C + a + b]/2\}$$

which is positive.

REMARK 4.1. Assumptions S1-S6 are easier to verify for the least squares objective function. The same results thus hold for least squares estimators, provided of course the rate constraints given below are satisfied.

We now verify Assumptions P1-P5 for our first choice of α. Assumption P1 is satisfied trivially. For Assumption P2, applying Lemma 3.2 and plugging in our choice of M_n necessitates

$$\{p_n^{1/q} [B(p_n)]^{1/2-1/q} / \epsilon_n\}^{1/\ell} - n\epsilon_n^2 + (1+\epsilon)\log(n)$$

be bounded as $n \to \infty$ for some $\epsilon > 0$. In terms of α and β, this is implied by

$$1 - 2\beta > \alpha[1 + 2h(q/2 - 1)]/q\ell + \beta/\ell$$

or

$$\ell > h\alpha - \alpha(2h - 1)/q + (1+2\ell)\beta.$$

For the first extreme, we want the rate of convergence in Assumption P3 (i) to just be bounded. For our $L(\ell,p_n)$, and in terms of α and β, this is achieved by setting

$$\beta = 2\ell\alpha + (1+\gamma)/r$$

which immediately guarantees Assumption P3 (ii). Plugging this into the final inequality above yields

$$\alpha < \frac{\ell - (1+\gamma)(1+2\ell)/r}{h - (2h - 1)/q + 2\ell(1+2\ell)}$$

and our choice of α satisfies this constraint; the fact that q can be made arbitrarily large ensures that it is tight. Note that Assumption P4 was used

as a hypothesis, and Assumption P5 follows from our choice of α and Remark 2.2.

The second extreme follows from making Assumption P3 (ii) the binding constraint, i.e. letting β be as large as possible. This entails setting

$$\beta = 2m\alpha + (1+\gamma)/r - \epsilon$$

for a small $\epsilon > 0$. The selection of α follows exactly as above. Unfortunately, Assumption P4 cannot be satisfied in this case.

5. Proof of main theorems. In this section are given the proofs of the theorems stated in Section 3. To summarize, we use the techniques of Pollard (1984) and Severini and Wong (1987) to prove Theorem 3.3, which can be viewed as a uniform strong law of large numbers with a rate. This theorem essentially changes Lemma 1 of Severini and Wong from convergence in probability to almost sure convergence. Theorem 3.4 simply makes use of Theorem 3.3 via Assumption S3 to obtain convergence in the $\|\cdot\|_{2\ell}$ norm. Theorem 3.7 borrows heavily from Yohai and Maronna (1979).

PROOF OF THEOREM 3.3. Define

$$\mathcal{F}_n = \{f \mid f(e,x;\theta) = s[Y(e,x),x,\varphi(x)'\theta]/n^{(1+\gamma)/r},\ \theta \in \theta_n\}$$

and note that

$$P\left(\sup_{\theta\in\theta_n} |\mathcal{E}_n s(\theta) - \mathcal{E}s(\theta)| > \epsilon_n n^{(1+\gamma)/r}\right) \leq P\left(\sup_{f\in\mathcal{F}_n} |\mathcal{E}_n f - \mathcal{E}f| > \epsilon_n\right).$$

Let $\mathcal{E}_n^0 f = (1/n)\sum_{t=1}^n \sigma_t f(e_t,x_t)$ where σ_t takes on the values ± 1 with equal probability, independently of $\{e_t,x_t\}_{t=1}^n$. From Pollard (1984, p. 31)

$$P\left(\sup_{f\in\mathcal{F}_n} |\mathcal{E}_n f - \mathcal{E}f| > 8\epsilon_n\right) \leq 4\, P\left(\sup_{f\in\mathcal{F}_n} |\mathcal{E}_n^0 f| > 2\epsilon_n\right)$$

provided $\mathrm{Var}(\mathcal{E}_n f)/(4\epsilon_n)^2 \leq 1/2$ for each $f \in \mathcal{F}_n$. Note that by Assumption S1

$$\mathrm{Var}(\mathcal{E}_n f) \leq (1/n^{1+2(1+\gamma)/r})\, R^{1/2}$$

for $f \in \mathcal{F}_n$, so ϵ_n must satisfy

$$(1/\epsilon_n^2 n^{1+2(1+\gamma)/r}) \leq 8/R^{1/2}.$$

In terms of β this is implied by $1 + 2(1+\gamma)/r - 2\beta > 0$, and this follows from Assumption P2. Now define the following three sequences of sets:

$$A_n = \left\{ \sup_{f \in \mathcal{F}_n} |\mathcal{E}_n^0 f| > 2\epsilon_n \right\}$$

$$B_n = \left\{ \sup_{f \in \mathcal{F}_n} \mathcal{E}_n f^2 > 1 \right\}$$

$$C_n = \left\{ N_1(\epsilon_n, \mathcal{F}_n) > \exp[H(2\epsilon_n/M_n, \Theta_n, \|\cdot\|)] \right\}$$

where $N_1(\epsilon_n, \mathcal{F}_n)$ is the smallest m such that there exist $(\theta_1, \ldots, \theta_m)$ in Θ_n satisfying

$$\sup_{\theta \in \Theta_n} \min_{1 \le j \le m} \mathcal{E} |f(e,x;\theta) - f(e,x;\theta_j)| < 2\epsilon.$$

N_1 is the covering number used in Chapter II of Pollard (1984). Now,

$$P(A_n) \le P(A_n B_n^c C_n^c) + P(B_n) + P(C_n).$$

Our strategy is to show that the right hand side of this expression is summable, and then apply the Borel-Cantelli lemma. For the first term, we condition on $\{e_t, x_t\}_{t=1}^n$, with respect to which both B_n and C_n are measurable. Using the indicator function $1(\cdot)$ and the Hoeffding inequality from Pollard (1984, p. 31)

$$P\left(A_n B_n^c C_n^c \mid \{e_t, x_t\}\right) = 1(B_n^c C_n^c)\, \mathcal{E}\left[1(A_n) \mid \{e_t, x_t\}\right]$$

$$\le 1(B_n^c C_n^c)\, 2\, N_1(\epsilon_n, \mathcal{F}_n) \exp\left[-n\epsilon_n^2/(2 \max_j \mathcal{E}_n g_j^2)\right]$$

where the maximum runs over all $N_1(\epsilon_n, \mathcal{F}_n)$ functions $\{g_j\}$ in the approximating class. This maximum can be replaced by the supremum of $\mathcal{E}_n f^2$ over $f \in \mathcal{F}_n$, and by definition of B_n and C_n the right hand side is bounded above by $2 \exp[H(2\epsilon_n/M_n, \Theta_n, \|\cdot\|) - n\epsilon_n^2/2]$. This bound no longer depends upon the

conditioning random variables, and thus it bounds the unconditional probability as well. It is summable by Assumption P2. For the second term, using Markov's inequality ($r/2^{th}$ power) and Assumption S1,

$$P(B_n) = P\left(\mathcal{E}_n \sup_{\theta\in\Theta_n} s^2[Y(e,x),x,\varphi(x)'\theta] > n^{2(1+\gamma)/r}\right)$$
$$\leq R/n^{1+\gamma}$$

which is summable. Finally, as in the proof of Lemma 1 of Severini and Wong (1987)

$$\sup_{f\in\mathcal{F}_n} \min_j \mathcal{E}_n|f - g_j| = \sup_{\theta\in\Theta_n} \min_j \mathcal{E}_n|s(\theta) - s(\theta_j)|/n^{(1+\gamma)/r}$$
$$\leq \sup_{\theta\in\Theta_n} \min_j \mathcal{E}_n m_n(e,x)\, \|\theta - \theta_j\|/n^{(1+\gamma)/r}$$
$$\leq \left(\sup_{\theta\in\Theta_n} \min_j \|\theta - \theta_j\|\right) \mathcal{E}_n m_n(e,x)/n^{(1+\gamma)/r}.$$

So by Markov's inequality (q^{th} power) and Assumption S2

$$P(C_n) \leq P(\mathcal{E}_n m_n(e,x)/n^{(1+\gamma)/r} > M_n)$$
$$\leq 1/n^{q(1+\gamma)/r}$$

which is summable because $q \geq r$. The final result follows from the fact that for any $\epsilon > 0$, ϵ_n can be replaced by $\epsilon\epsilon_n$ in the above argument. ∎

PROOF OF THEOREM 3.4. Apply the triangle inequality:

$$\|\hat{g}_n - g^o\|_{2\ell} \leq \|\hat{g}_n - g_n^o\|_{2\ell} + \|g_n^o - g_n^o\|_{2\ell}$$

The second term on the r.h.s represents a pure approximation error and is $O(p_n^{\ell-m})$ by the statement given after Theorem 2.2 of Cox. Considering the first term on the r.h.s, first note that

$$\mathcal{E}\,[g_n^o(x) - g^o(x)]^2 = O(p_n^{-2m})$$

by the same statement from Cox. So Assumption S3 and the Corollary to Theorem 1 allow us to say that

$$\|\hat{g}_n - g_n^o\|_{2\ell}^2 = o[L(\ell,p_n)\,\epsilon_n\,n^{(1+\gamma)/r}] \quad \text{a.s.}$$

provided

$$p_n^{-2m} = o(\epsilon_n\,n^{(1+\gamma)/r}),$$

which is Assumption P3(ii). Assumption P3(i) ensures that we have consistency in the $\|\cdot\|_{2\ell}$ norm. ∎

REMARK 5.1. Our proof strategy seems quite different from the functional analytic approach of Cox. We are unsure of the extent to which his results depend upon the projection nature of least squares estimators, and whether or not they can be generalized to our setting. Our approach also differs from compactness arguments a la Jennrich (1969). They appear to yield consistency only in a norm one order less than the one defining the estimation subspace (see Elbadawi, Gallant, and Souza, 1983).

PROOF OF THEOREM 3.7. First, note that

$$\sum_{t=1}^{n} [\hat{g}_n(x_t) - g_n^o(x_t)]^2 = \|\Phi(\hat{\theta}_n - \theta_n^o)\|.$$

We make the following definitions:

$$X(e,x,\beta,z) = (\partial/\partial g)s[g^0(x) + e,x,g]\Big|_{g = h(x)'\beta - z}$$

where $h(x) = (\Phi'\Phi)^{-1/2}\varphi(x)$,

$$\beta_n^0 = (\Phi'\Phi)^{1/2}\theta_n^0$$

$$\hat{\beta}_n = (\Phi'\Phi)^{1/2}\hat{\theta}_n$$

and

$$U(\xi,L) = \sum_{t=1}^{n} X(e_t,x_t,\beta_n^0,Lh_t'\xi)(h_t'\xi).$$

Then we have

$$P(p_n^{-1/2}\ \|\Phi(\hat{\theta}_n - \theta_n^0)\| \geq L) \leq P(\sup_{\|\xi\|=1} p_n^{-1/2}\ U(\xi,p_n^{1/2}L) \geq 0).$$

To see why this holds, note that

$$U[\frac{\beta_n^0 - \hat{\beta}_n}{\|\beta_n^0 - \hat{\beta}_n\theta\|},\|\beta_n^0 - \hat{\beta}_n\|] = \sum_{t=1}^{n} X(e_t,x_t,\hat{\beta}_n,0)\ h_t'[\frac{\beta_n^0 - \hat{\beta}_n}{\|\beta_n^0 - \hat{\beta}_n\|}]$$

$$= 0$$

by the first order conditions of the optimization problem, which hold a.s. for n sufficiently large by Theorem 3.4. So the event

$$\{p_n^{-1/2}\ \|\Phi(\hat{\theta}_n - \theta_n^0)\| \geq L\} = \{\|\beta_n^0 - \hat{\beta}_n\| \geq p_n^{1/2}L\}$$

$$\subseteq \{U[\frac{\beta_n^0 - \hat{\beta}_n}{\|\beta_n^0 - \hat{\beta}_n\|},p_n^{1/2}L] \geq 0\}$$

by Assumption S4, because regardless of whether $h_t'(\beta_n^o - \hat{\beta}_n)$ is positive or negative, the effect of decreasing $\|\beta_n^o - \hat{\beta}_n\|$ to $p_n^{1/2}L$ (in the second argument of $U[\cdot,\cdot]$) is to increase U. The inequality follows by taking the supremum over $\|\xi\|=1$.

Note that

$$X(e_t,x_t,\beta_n^o,0) = X_n(e_t,x_t) = X_{nt},$$

so by Taylor's theorem

$$p_n^{-1/2}\, U(\xi,p_n^{1/2}L) = p_n^{-1/2} \sum_{t=1}^{n} X_{nt}\ (h_t'\xi)$$

$$- L \sum_{t=1}^{n} w(e_t,x_t,\beta_n^o,\lambda p_n^{1/2}Lh_t'\xi)\ (h_t'\xi)^2$$

where λ is between 0 and 1. Looking at the two terms on the r.h.s. separately, put

$$A_n = \sup_{\|\xi\|=1} |p_n^{-1/2} \sum_{t=1}^{n} X_{nt}\ (h_t'\xi)|.$$

Then

$$P(A_n \geq Lcq/2) = 1/L^2\ O(1)$$

by Lemma 3.6. Therefore, by choosing L large enough, we can make

$$P(A_n \geq Lcq/2) \leq \delta/2$$

for any $\delta > 0$.

Now define

$$B_n = \inf_{\|\xi\|=1} \sum_{t=1}^{n} w(e_t, x_t, \beta_n^0, \lambda p_n^{1/2} L h_t' \xi) \; (h_t' \xi)^2 .$$

Note that for $\|\xi\|=1$

$$\begin{aligned} |\lambda p_n^{1/2} L h_t' \xi| &\leq p_n^{1/2} L \; \|h_t\| \; \|\xi\| \\ &\leq (p_n \gamma_n)^{1/2} L \end{aligned}$$

where $\gamma_n = \max_{1 \leq t \leq n} \|h_t\|^2$. Note $\gamma_n \leq \lambda_{max}(\Phi'\Phi/n)^{-1} \; B(p_n)/n$, so this final bound can be made less than b for n sufficiently large by Assumption P5. This allows us to invoke Assumption S6, and conclude that for n sufficiently large

$$P(B_n < cq/2) \leq (4r^2/c^2q^2) \; \gamma_n p_n$$

by Lemma 2 of Yohai and Maronna (1979). The right hand side of this expression converges to zero again by Assumption P5. So for n and L sufficiently large

$$\begin{aligned} P(p_n^{-1/2} \; \|\Phi(\hat{\theta}_n - \theta_n^0)\| \geq L) &\leq P(\sup_{\|\xi\|=1} p_n^{-1/2} \; U(\xi, p_n^{1/2} L) \geq 0) \\ &\leq P(A_n - LB_n \geq 0) \\ &\leq P(A_n \geq Lcq/2) + P(B_n < cq/2) \\ &\leq \delta \end{aligned}$$

for any $\delta > 0$. ∎

APPENDIX

PROOF OF LEMMA 3.2. We use an argument of Smolyak (1960) which was originally given by Kolmogorov and Tihomirov (1959). Recall that

$$\theta_n = \{\theta \in \mathbb{R}^{p_n} \mid \|\theta\|_{2\ell} \le B\}$$

Let M be the maximum number of non-intersecting balls of $\|\cdot\|$-radius $\epsilon/2$ with centers in θ_n and let θ be a point in θ_n. Notice that the $\epsilon/2$-neighborhood of θ in $\|\cdot\|$ must intersect one of the M balls, and hence θ is within $\|\cdot\|$-distance ϵ of one of the centers of these balls. Therefore

$$\exp[H(\epsilon,\theta_n,\|\cdot\|)] \le M.$$

We now find an upper bound for M. Note that all of the M balls lie within the p_n-dimensional ellipsoid with vertices $\pm (B/j^{\ell} + \epsilon/2)$, $j = 1,\ldots,p_n$. If we call this ellipsoid $\theta_{n,\epsilon}$, then $M \le \mathcal{V}/v$, where $\mathcal{V}$ is the volume of $\theta_{n,\epsilon}$ and v is the volume of a p_n-dimensional ball of $\|\cdot\|$-radius $\epsilon/2$. By integration, the volume of a p-dimensional ellipsoid with vertices $\pm a_j$, $j = 1,\ldots,p$, is

$$\frac{\pi^{p/2}}{\Gamma(p/2 + 1)} \prod_{j=1}^{p} |a_j|.$$

Therefore

$$M \le \frac{\pi^{p/2}}{\Gamma(p/2 + 1)} \prod_{j=1}^{p} (B/j^{\ell} + \epsilon/2) \times \frac{\Gamma(p/2 + 1)}{\pi^{p/2}} (2/\epsilon)^p$$

$$= \prod_{j=1}^{p} (2B/\epsilon j^{\ell} + 1)$$

where $p = p_n$. Taking the (natural) logarithm and using an integral approximation we have

$$\log M = \sum_{j=1}^{p} \log(2B/\epsilon j^{\ell} + 1)$$

$$\leq \int_0^p \log(2B/\epsilon y^{\ell} + 1)\, dy$$

$$= (2B/\epsilon)^{1/\ell} \int_q^{\infty} (1/x^2)\log(x^{\ell} + 1)\, dx$$

where $q = (1/p)(2B/\epsilon)^{1/\ell}$ and we have used the transformation $x = (1/y)(2B/\epsilon)^{1/\ell}$. Now for $\ell \geq 2$, the integral can be bounded above by

$$c_H \int_0^{\infty} (1/x^2)\log(x^2 + 1)\, dx = c\pi$$

where c_H is some positive constant and we have used formula 4.295.2, p. 560 from Gradshteyn and Ryzhik (1980). ∎

REMARK A.1. For the case $\ell = 1$, the bound from Lemma 2 still holds as long as q defined above is bounded away from zero. For this case, we would thus need to assume that

$$\lim_{n\to\infty} (2BM_n/\epsilon_n)^{1/\ell} / p_n = \infty.$$

REFERENCES

Abramowitz, Milton, and Irene A. Stegun (1965) *Handbook of Mathematical Functions*. New York: Dover Publications.

Andrews, Donald W. K. (1988) "Asymptotic Normality of Series Estimators for Various Nonparametric and Semiparametric Models," Cowles Foundation Discussion Paper No. 874, Yale University.

Cox, Dennis D. (1988) "Approximation of Least Squares Regression on Nested Subspaces," *The Annals of Statistics* 16, 713-732.

Elbadawi, Ibrahim, A. Ronald Gallant, and Geraldo Souza (1983) "An Elasticity Can be Estimated Consistently Without A Priori Knowledge of Functional Form," *Econometrica* 51, 1731-1752.

Gallant, A. Ronald (1985) "Identification and Consistency in Seminonparametric Regression", Invited paper, World Congress of the Econometric Society. In Truman Bewley, ed. *Advances in Econometrics*, New York, Cambridge University Press, 1987.

Gallant, A. Ronald (1987) *Nonlinear Statistical Models*. New York: John Wiley and Sons.

Gallant, A. Ronald, and George Tauchen (1989) "Seminonparametric Estimation of Conditionally Constrained Heterogeneous Processes: Asset Pricing Applications," *Econometrica* 57, 1091-1120.

Gradshteyn, I.S., and I.M. Ryzhik (1980) *Table of Integrals, Series, and Products*. New York: Academic Press.

Grenander, Ulf (1980) *Abstract Inference*. New York: John Wiley and Sons.

Huber, Peter J. (1964) "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics* 35, 73-101.

Jennrich, R.I. (1969) "Asymptotic Properties of Nonlinear Least Squares Estimation," *The Annals of Mathematical Statistics* 40, 633-643.

Kolmogorov, A.N. and V.M. Tihomirov (1959) "ϵ-entropy and ϵ-capacity of Sets in Functional Spaces," *Uspehi Mat. Nauk* 14, 3-86; English translation (1961) *American Mathematical Society Translations* Ser. 2, Vol. 17, 277-364.

Pollard, David (1984) *Convergence of Stochastic Processes.* New York: Springer Verlag.

Portnoy, Stephen (1984) "Asymptotic Behavior of M-Estimators of p Regression Parameters when p^2/n is Large. I. Consistency," *The Annals of Statistics* 12, 1298-1309.

Severini, T.A. and W.H. Wong (1987) "Convergence Rates of Maximum Likelihood and Related Estimates in General Parameter Spaces," unpublished manuscript, Department of Statistics, University of Chicago.

Smolyak, S.A. (1960) "The ϵ-entropy of classes $\mathcal{E}_s^{\alpha,k}(B)$ and $W_s^{\alpha}(B)$ in the $\mathcal{L}^2$-metric," *Soviet Math. Dokl.* 1, 192-195.

Wolfinger, Russell D. (1989) *Rates of Convergence and Asymptotic Normality in Semi-Nonparametric Regression.* North Carolina State University, Ph.D. dissertation.

Yohai, Victor J., and Ricardo A. Maronna (1979) "Asymptotic Behavior of M-Estimators for the Linear Model," *The Annals of Statistics* 7, 258-268.