The Annals of Statistics
1992, Vol. 20, No. 1, 608–613

A SIMPLE LEMMA ON GREEDY APPROXIMATION IN HILBERT SPACE AND CONVERGENCE RATES FOR PROJECTION PURSUIT REGRESSION AND NEURAL NETWORK TRAINING[1]

BY LEE K. JONES

University of Massachusetts, Lowell

A general convergence criterion for certain iterative sequences in Hilbert space is presented. For an important subclass of these sequences, estimates of the rate of convergence are given. Under very mild assumptions these results establish an $O(1/\sqrt{n})$ nonsampling convergence rate for projection pursuit regression and neural network training; where n represents the number of ridge functions, neurons or coefficients in a greedy basis expansion.

1. Introduction. We consider an iterative sequence f_n in a real Hilbert space H, approximating some $\bar{f}$ where the iterations involve computations with restrictive subsets of H.

EXAMPLE. Let $\bar{f} = E(Y|X)$ be a standard regression function with Y a one-dimensional random variable, let P be the probability measure for the d-dimensional random variable X and let $\| \ \|_P$ be the norm in $H = L_2(P)$. Assuming finite $\|\bar{f}\|_P$ and sufficient regularity for minima to exist, Friedman–Stuetzle projection pursuit regression (PPR) starts with $f_0 = 0$ and at stage $n + 1$ chooses a unit vector a_{n+1} and a function g_{n+1} such that

$$f_{n+1} = f_n + g_{n+1}(a_{n+1}^t x) \quad \text{and} \quad \| f_{n+1} - \bar{f} \|_P \text{ is minimum.} \tag{1}$$

It is shown in [2] that, in fact, $g_{n+1}(z) = E(Y - f_n | a_{n+1}^t X = z)$. In the practical sampling form, one has noisy values of $\bar{f}$ (values of Y) on a finite set. At stage $n + 1$ (using a given statistical routine for each a and a numerical minimization algorithm over a) one finds the best fit of the form of a ridge function, $g_{n+1}(a^t x)$, to the noisy values of $\bar{f} - f_n$ at these points. In [3]this is done with a sophisticated nonlinear smoothing routine. For this and other examples see [3] and [5]. We discuss only the nonsampling theory in which $\bar{f}$ is known and convenient approximations are sought.

Norm convergence of the Friedman–Stuetzle procedure was shown in [6]. But examination of the proof indicates that the convergence may be very slow. We show that this convergence can be accelerated by changing f_{n+1} in (1) to be an optimal convex combination of f_n and a ridge function. (Since this only increases by 1 the dimension of the optimization space, all algorithms treated

Received March 1990; revised June 1991.
[1]Research supported in part by ONR-N6014-85-C2503.
AMS 1980 *subject classification*. Primary 62H99.
Key words and phrases. Projection pursuit, greedy expansion, neural network.

henceforth will optimize over such convex combinations.) Also, one may impose restrictions on the choice of ridge functions available at each iteration and discuss how these affect convergence.

For instance, only ridge functions with g_n of the form $g_n(s) = c_n h(s - t_n)$ may be used as in the case of neural networks with a single hidden layer and fixed activation function h. Or, at each stage, the approximation increments may be restricted to a given set of basis functions. Our results will then provide convergence rates in terms of the number of coefficients in an iterative basis expansion.

The emphasis throughout is on improving the approximation in a single step using members of a restricted class in a greedy fashion. This is important in high-dimensional applications where it is computationally prohibitive to optimize simultaneously over several ridge directions. For convergence rates we also make regularity assumptions on $\bar{f}$ (which are reasonable for applications). There are more extensive results on rates in [2] for $d = 2$, $P =$ standard Gaussian and simultaneous optimization over equispaced ridge directions.

To examine these issues, we state and prove an abstract lemma about the convergence of f_n (to $\bar{f}$) when the degree of approximation of f_{n+1} (to $\bar{f}$) is related to the degree of approximation of f_n in terms of a set of elements $\mathscr{P}_n$ which plays the abstract role of the ridge functions. For a variety of cases including projection pursuit regression under very mild assumptions, results on the rate of convergence will also be derived.

2. Preliminaries. Let H be a real Hilbert space with norm $\| \|$. Let $f_n, \bar{f} \in H$. We are further given a sequence of subsets of H, $\mathscr{P}_n$, which we call the projectors at stage n. Set $e_n = \| f_n - \bar{f} \|$ and finally let

$$r_n = \inf_{\substack{0 \le \alpha \le 1 \\ \phi \in \mathscr{P}_n}} \| (1 - \alpha) f_n + \alpha \phi - \bar{f} \|. \tag{2}$$

DEFINITION. f_n is called *relaxed* with respect to $\bar{f}$ if $e_{n+1} \le r_n$. f_n is called *asymptotically relaxed* with respect to $\bar{f}$ if $\limsup(e_{n+1} - r_n) \le 0$.

Clearly any convergent sequence is asymptotically relaxed. A relaxed variant of PPR (an algorithm leading to a relaxed sequence) is to minimize (assuming sufficient regularity conditions) $\| (1 - \alpha) f_n + \alpha g(a^t x) - \bar{f} \|_P$ over $0 \le \alpha \le 1$ and a, g as before. The $(n + 1)$st estimate is then

$$f_{n+1} = (1 - \alpha_n) f_n + \alpha_n g_{n+1}(a_{n+1}^t x). \tag{3}$$

Here $\mathscr{P}_n$ is the set of all ridge functions and the sequence f_n is clearly relaxed. Since the ridge functions are closed under scalar multiplication, in the practical algorithm, one finds the best fit of form $g(a^t x)$ to noisy values of

$\bar{f} - (1 - \alpha) f_n$ for given a, α and then optimizes over a, α. Note again that this procedure increases the dimension of the optimization domain by only 1 but will have the convergence rates given below. Note also that the sequence will be relaxed if we use a full linear version, that is, f_{n+1} is an optimal linear combination of $f_1, \ldots, f_n$ and a ridge function. Hence our convergence results will apply to this situation. Finally, for an arbitrary sequence $\mathscr{P}_n$, the analogous version of (3) will be called a relaxed algorithm.

3. Convergence results.

LEMMA. *Suppose $\bar{f}$ is in the closure of $\bigcup_{n=1}^{\infty}$ convex hull $(\bigcap_{m \geq n} \mathscr{P}_m)$. Then f_n is asymptotically relaxed with respect to $\bar{f}$ if and only if $f_n \to \bar{f}$.*

Before proving the lemma we note that, in the relaxed PPR case, $\bar{f}$ may be approximated to within ε by a finite linear combination of elements which are in every $\mathscr{P}_n$ (using Fourier analysis). Since the $\mathscr{P}_n$ are closed under scalar multiplication, this combination may be assumed to be convex. Hence the hypothesis of the lemma is satisfied in this case.

PROOF OF LEMMA. It is enough to show convergence under the assumption of asymptotic relaxation: Suppose $f_n \nrightarrow \bar{f}$; then $\limsup \| f_n - \bar{f} \| = \rho > 0$. Also $\rho < \infty$ since r_n, e_n and hence $\| f_n \|$ are bounded. Choose a subsequence n_i such that $\| f_{n_i} - \bar{f} \| \to \rho$ and consider the subsequence $m_i = n_i - 1$. Now

$$\limsup_i \| f_{m_i} - \bar{f} \| \leq \limsup_n \| f_n - \bar{f} \| = \rho.$$

Also

$$\limsup_i \left(\| f_{n_i} - \bar{f} \| - \| f_{m_i} - \bar{f} \| \right) = \limsup_i \left(\| f_{m_i+1} - \bar{f} \| - \| f_{m_i} - \bar{f} \| \right)$$

$$\leq \limsup_i (e_{m_i+1} - r_{m_i}) \leq 0$$

by asymptotic relaxation. This implies $\liminf_i \| f_{m_i} - \bar{f} \| \geq \rho$. Hence, $\| f_{m_i} - \bar{f} \| \to \rho$. The rest of the proof follows by the following argument: There is an N such that we can choose $\phi_1, \ldots, \phi_L \in \bigcap_{n \geq N} \mathscr{P}_n$ with $\| \sum_1^L \beta_j (\phi_j - \bar{f}) \| < \rho/2$ for some convex weights $\beta_1, \ldots, \beta_L$. Let $M = \max_j \| \phi_j - \bar{f} \|$. Clearly $M \geq \rho$, for otherwise if $M < \rho$, we have $\limsup r_n \leq M$ and $\limsup e_{n+1} \leq M$ contradicting $e_{n_i} \to \rho$. Now, given i, there exists a ϕ_{j_i} such that $(\phi_{j_i} - \bar{f}, u_i) < \rho/2$, where u_i is a unit vector in the direction of $f_{m_i} - \bar{f}$; for otherwise $\| \sum_1^L \beta_j (\phi_j - \bar{f}) \| \geq \sum_1^L \beta_j (\phi_j - \bar{f}, u_i) \geq \rho/2$.

Using asymptotic relaxation and convex weights α, $1-\alpha$ we have:

$$\begin{aligned}
\limsup_i \| f_{n_i} - \bar{f} \| &= \limsup_i \| f_{m_i+1} - \bar{f} \| \\
&\le \limsup_i \| (1-\alpha)(f_{m_i} - \bar{f}) + \alpha(\phi_{j_i} - \bar{f}) \| \\
&= \limsup_i \Big((1-\alpha)^2 \| f_{m_i} - \bar{f} \|^2 + \alpha^2 \| \phi_{j_i} - \bar{f} \|^2 \\
&\qquad + 2\alpha(1-\alpha)(\phi_{j_i} - \bar{f}, f_{m_i} - \bar{f}) \Big)^{1/2} \\
&\le \limsup_i \Big((1-\alpha)^2 \| f_{m_i} - \bar{f} \|^2 + \alpha^2 M^2 \\
&\qquad + \alpha(1-\alpha)\rho \| f_{m_i} - \bar{f} \| \Big)^{1/2} \\
&= \big((1-\alpha)\rho^2 + \alpha^2 M^2 \big)^{1/2}.
\end{aligned}$$

By minimizing over $0 \le \alpha \le 1$, this bound can be made (set $\alpha = \rho^2/2M^2$) equal to $\rho(1 - (\rho^2/4M^2))^{1/2}$ which is less than ρ, a clear contradiction. □

Using the idea of the proof, we may get bounds on the rate of convergence in the relaxed case under the additional assumption that $\bar{f}$ lies in the closure of the convex hull of some collection $\mathcal{G}$ of elements of H, where $\|g\| \le M'$ for all $g \in \mathcal{G}$ and $\mathcal{G} \subset \mathcal{P}_n$ for every n. As we show in Section 4, this is the case for relaxed PPR under mild assumptions on $\bar{f}$ and P. We now derive the bounds:

THEOREM. *Under the assumptions of the preceding paragraph, the approximation error is* $O(1/\sqrt{n})$.

PROOF. First for any $\delta > 0$, since we may approximate $\bar{f}$ arbitrarily closely by a convex combinations of elements of $\mathcal{G}$, we may find $\phi_1, \dots, \phi_s \in \mathcal{G}$ such that

$$\left(f_{n-1} - \bar{f}, \sum_1^s \alpha_i \phi_i - \bar{f} \right) < \delta$$

with $\alpha_i \ge 0, \sum_1^s \alpha_i = 1$. This we rewrite as

$$\sum_1^s \alpha_i (f_{n-1} - \bar{f}, \phi_i - \bar{f}) < \delta.$$

Clearly at least one of the inner products in this sum is less than δ. Therefore, for any $\delta > 0$, we have shown the existence of $g \in \mathcal{G}$ such that $(f_{n-1} - \bar{f},$

$g - \bar{f}) < \delta$. Hence

$$e_n^2 \le \inf_{\substack{0\le\alpha\le 1\\ g\in\mathscr{G}}} \left\|(1-\alpha)\left(f_{n-1}-\bar{f}\right)+\alpha\left(g-\bar{f}\right)\right\|^2$$

$$\le \inf_{0\le\alpha\le 1}\left((1-\alpha)^2 e_{n-1}^2+\alpha^2\left(M'+\|\bar{f}\|\right)^2\right).$$

Setting

$$\overline{M} = M' + \|\bar{f}\| \quad \text{and} \quad \alpha = \frac{e_{n-1}^2}{e_{n-1}^2+\overline{M}^2},$$

we get after some calculation

$$e_n^2 \le \overline{M}^2\left(1+\frac{\overline{M}^2}{e_{n-1}^2}\right)^{-1}, \tag{4}$$

which yields

$$\frac{1}{e_n^2} \ge \frac{1}{\overline{M}^2}+\frac{1}{e_{n-1}^2} \ge \frac{2}{\overline{M}^2}+\frac{1}{e_{n-2}^2} \ge \cdots \ge \frac{n}{\overline{M}^2}+\frac{1}{e_0^2}$$

so that

$$e_n^2 \le \overline{M}^2\left(n+\left(\frac{\overline{M}^2}{e_0^2}\right)\right)^{-1} \tag{5}$$

which demonstrates that the approximation error after n iterations is $O(1/\sqrt{n})$. □

Note that in the asymptotically relaxed case, the $O(1/\sqrt{n})$ rate will hold if $e_{n+1} - r_n$ converges to 0 sufficiently fast. Also it is immediate that, in the relaxed PPR case, this rate holds when $\bar{f}$ is itself a finite ridge expansion.

4. Applications to regression and neural networks. Suppose that in the relaxed PPR example, both $\bar{f}$ and its d-dimensional Fourier transform $\hat{\bar{f}}$, have finite $L_1(R^d)$ norm, $\|\ \|_1$. Then, if P is absolutely continuous with respect to Lebesgue measure, $\bar{f}$ can be represented (via the Fourier inversion theorem) in $L_2(P)$ as

$$\bar{f} = \int_{R^d} \frac{\left|\hat{\bar{f}}(\omega)\right|}{\|\hat{\bar{f}}\|_1} \mathscr{R}e\left\{\frac{\hat{\bar{f}}(\omega)}{(2\pi)^d\left|\hat{\bar{f}}(\omega)\right|}\|\hat{\bar{f}}\|_1 e^{i\omega^t x}\right\} d\omega.$$

Hence we may apply (5) where $\mathscr{G}$ consists of those ridge functions which are the real part expressions in the above integral. We see immediately that $\overline{M} = (\|\hat{\bar{f}}\|_1/(2\pi)^d) + \|\bar{f}\|_P$ and $e_0 = \|\bar{f}\|_P$, both of which could be estimated from data in practical applications. Alternatively, making $\mathscr{P}_n \equiv \mathscr{G}$ yields a greedy trigonometric expansion with n coefficients for which (5) holds.

A similar argument holds if $\bar{f}$ is periodic with period 2π in each variable and has absolutely summable discrete Fourier transform. In particular, consider the family of $\bar{f}_\beta = \sum_{m=2}^{\infty}(1/m(\ln m)^2)\cos(\beta a_m^t x)$; $\beta = 1, 2, \ldots$ for which

the a_m have positive integral components and are nonparallel. Let P be the uniform measure on a cube with sides of length 2π. Now if we let $\beta \to \infty$, then the first ridge direction of PPR approaches a_2. Iterating this argument we may, by making β sufficiently large, force the approximation error after n stages of relaxed PPR to be arbitrarily close to

$$\sqrt{\sum_{n+1}^{\infty} \frac{\pi^d}{m^2(\ln m)^4}} \approx O\left(\frac{1}{\sqrt{n}\,(\ln n)^2}\right).$$

Since the $\bar{M}$ and e_0 in (5) are the same for all $\bar{f}_\beta$, (5) is a best possible bound based on the number n of ridge functions in relaxed PPR (to within a logarithmic factor).

Finally we give some applications to neural networks: A neural network with a single hidden layer with n neurons provides approximation to $\bar{f}$ of the form $(*)\sum_{i=1}^{n} c_i h(a_i^t x - t_i)$ as output when x is the network input. The activation function h is fixed and the network is trained by selecting the parameters c_i, a_i, t_i, n. Setting $\mathscr{P}_n = \{ch(a^t x - t); c, a, t\}$, we may train with the relaxed algorithm. If the forms $(*)$ are dense in $L_2(P)$ (which has been demonstrated for squashing h in [4] and continuous sigmoidal h in [1]), then by the lemma the network outputs converge to $\bar{f}$. Also, if $\bar{f}$ is of the form $(*)$ (with n possibly infinite and h bounded) and the $|c_i|$ are summable, then the theorem implies that this convergence is $O(1/\sqrt{n})$. (This was noted by A. R. Barron.)

Acknowledgment. The author is greatly indebted to Andrew R. Barron for a critique of an earlier manuscript.

REFERENCES

[1] CYBENKO, C. (1989). Approximations by superposition of a sigmoidal function. *Math. Control Signals Systems* **2** 303–314.

[2] DONOHO, D. L. and JOHNSTONE, I. M. (1989). Projection-based approximation and a duality with kernel methods. *Ann. Statist.* **17** 58–106.

[3] FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

[4] HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* **2** 359–366.

[5] HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

[6] JONES, L. K. (1987). On a conjecture of Huber concerning the convergence of projection pursuit regression. *Ann. Statist.* **15** 880–882.

INSTITUTE FOR VISUALIZATION
AND PERCEPTION RESEARCH
AND DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MASSACHUSETTS AT LOWELL
ONE UNIVERSITY AVENUE
LOWELL, MASSACHUSETTS 01854