The Annals of Statistics
2000, Vol. 28, No. 5, 1379–1389

LOCAL GREEDY APPROXIMATION FOR NONLINEAR REGRESSION AND NEURAL NETWORK TRAINING

BY L. K. JONES[1]

University of Massachusetts, Lowell

A criterion for local estimation and approximation in nonlinear regression and neural network training is introduced and motivated. Nth-order greedy approximation for the regression (or target) function based on the criterion is shown to converge at rate $O(1/N^{1/2})$ in the nonsampling case.

1. Introduction. For many applications in regression, classification, or neural network training an estimate of expected response (class, output), $f(x)$, is desired at or close to (in an appropriate sense) one fixed predictor (observation, input) vector. This estimate should depend heavily on predictor vectors in the sample which are close to the given fixed predictor vector. For example, in [2], a local technique based on a set of closest vectors in Euclidean distance is shown to be a significant improvement over previous methods for optical character recognition. Restriction to a local region in the feature vector space allows estimation to be fine tuned to letters with similar structure. For such classification problems the images may be easily normalized in size and correctly oriented and the noise level is low. Unfortunately, in many cases where there is low signal-to-noise ratio and/or a continuum of random predictor orientations described by many parameters, Euclidean distance is usually not a possible measure of closeness due to the curse of dimensionality which is exhibited in the following special case: suppose the predictor distribution is uniform on a ball of radius one in d-dimensional space $\mathbf{R}^d$. Assume we are interested in estimating the expected response at or close to the origin 0. It might be reasonable to use only sample vectors inside a ball of radius $\rho < 1$, assuming Euclidean distance as a measure of closeness. But, since the probability of a sample vector lying in the smaller ball is ρ^d, it is necessary that sample sizes are exponential in d to get enough close vectors for accurate estimation. Clearly, the curse persists for many general predictor distributions and distances which are topologically equivalent to Euclidean distance.

To avoid the curse, one might assume $f(x)$ has a ridge approximation,

$$f(x) \cong \sum_{n=1}^{N} c_n g_n(a_n^t x) \tag{1}$$

with $a_n \in \mathbf{A}$, $g_n \in \mathbf{G}$, where $\mathbf{A}$ is a given class of vectors in $\mathbf{R}^d$ and $\mathbf{G}$ is a given class of real valued functions on $\mathbf{R}$. Hence we assume f is nearly a linear

Received November 1998; revised November 1999.
[1]Supported in part by NSF Grant DMS-95-05199.
AMS 1991 *subject classifications*. 62H99
Key words and phrases. Greedy approximation, local training.

combination of functions of projections onto $\mathbf{R}$. (More generally the a's could be $d\,x\,m$ matrices and the g's could be real functions on $\mathbf{R}^m$: all results mentioned or proved here extend straightforwardly to this case of ridge approximation with projections onto $\mathbf{R}^m$). Greedy algorithms for finding the expansion might then be used, that is obtain the "best" approximation of $f(x)$ of the form $g(a^t x)$ ($a \in \mathbf{A}$, $g \in \mathbf{G}$), subtract this from $f(x)$ and repeat the process with the residual, etc. In the implementation with samples, the procedure involves solving simple regression (one-dimensional predictor) problems. If $\mathbf{A}$ is the set of d unit vectors in the coordinate directions, the predictor measure is uniform on the unit cube and $\mathbf{G}$ is all functions square integrable on the unit interval, then any $f(x)$ of the form (1) (w.l.o.g. $N \leq d$, g_n orthogonal to 1 for $n > 1$) is recovered in N greedy steps where "best" means best in L_2 of the unit cube. As we shall see below, $f(x)$ can also be efficiently greedily recovered in more general settings. But this greedy method seems insensitive to the need to favor local structure. Indeed, in the previous example the g_i's would be recovered in the order given by the size of the integrals of their squares. The last to be recovered might be the only one which is nonzero near the fixed predictor of interest. So in this paper we derive a greedy algorithm which is sensitive to the need for a good local approximation and at the same time shares the global efficiency of other greedy algorithms. In the next section we examine some of the results in more general settings before deriving our modification. Readers familiar with recent results on greedy approximation and estimation and complexity of neural net training may want to go immediately to Section 3.

2. Previous work on nonlocal greedy algorithms and computational feasibility.

EXAMPLE 1. If $f \in L_2(P)$ with P the predictor measure, if $\mathbf{A}$ is all unit vectors in $\mathbf{R}^d$ and if $\mathbf{G}$ is the set of all measurable functions, we might approximate in a (pure) greedy fashion: first, find a_0, g_0 such that $f_1(x) = g_0(a_0^t x)$ best approximates $f(x)$ in $L_2(P)$; then determine a_1, g_1 such that $g_1(a_1^t x)$ best approximates $f(x) - f_1(x)$ and set $f_2(x) = f_1(x) + g_1(a_1^t x); \cdots$; find a_n, g_n such that $g_n(a_n^t x)$ best approximates $f(x) - f_n(x)$ and set $f_{n+1} = f_n + g_n(a_n^t x), \ldots$. This is the projection pursuit regression algorithm (PPR) (approximation form) of [8]. For any a in the nth step the best g is just $E(Y - f_n(X) \mid a^t X)$. For practical implementation with a sample $\{(x_i, y_i)\}$, in the nth step for each a the best g is found using a variable bandwidth smoother applied to the simple regression problem with predictors $a^t x_i$ and (residual) responses $r_i = y_i - f_n(x_i)$. The quality of the fit, the average squared error of the smooth, is minimized over a. One such simple smoother is described as follows: for any x select the closest 5% of $\{a^t x_i\}$ to x and let $g(x)$ be the least squares linear fit at x based on this 5%. We now employ a global optimization algorithm to solve (approximately)

$$\min_a \sum (r_i - g(a^t x_i))^2. \tag{S}$$

Hence the objective function is at each step the same except that the residuals are from the previous step.

There are many refinements in the sampling case above: **A** and **G** may be appropriately restricted, a more general loss function may be used. See [7]. We will treat only squared error loss and will motivate a weighted variation of (S). The form of the variation is motivated in the same way for more general loss functions.

EXAMPLE 2. Assume that **A** is the set of unit vectors in $\mathbf{R}^d$, the c_i are real, and **G** consists of all translated dilations of a fixed bounded function, $\sigma(t)$, which is sigmoidal [i.e., $\sigma(-\infty) = 0$ and $\sigma(+\infty) = 1$]. The approximation is the output of the single hidden layer neural network. A modified projection pursuit algorithm could be applied: a would vary over unit vectors as in Example 1 but $g(a^t x)$ would be replaced by the best approximation of the form $c\sigma(ra^t x - t)$. For the sampling case a, c, r and t are determined in step n for which this form best fits the data $\{(x_i, y_i - f_n(x_i))\}$ $(= \{(x_i, r_i)\})$ in the least squares sense.

The advantage of the greedy procedures (if they converge) is that we are solving a sequence of N d-dimensional optimization problems as opposed to one Nd-dimensional problem. The convergence of the procedure and its rate of convergence are the focus of recent investigations. In [9] weak convergence in $L_2(P)$ of PPR under mild assumptions is established and in [10] strong convergence is proved. In [15] an extension of [10] is proved which has strong convergence for Example 2 as a corollary.

A relaxed extension of the above greedy algorithms has been investigated. Let $f(x)$ lie in the closure of the convex hull of a subset F_1 of $L_2(P)$ whose elements h have norms, $\|h\|$, bounded by a fixed constant B. Assume $f_0(x) = 0$ and at step n pick $g_n \in F_2(F_1 \subset F_2)$ and $\alpha_n \in [0, 1]$ minimizing $\|f(x) - \alpha g(x) - (1 - \alpha)f_n(x)\|$ over α, g. Set $f_{n+1}(x) = (1 - \alpha_n)f_n(x) + \alpha_n g_n(x)$. In [11] $\|f(x) - f_n(x)\|$ was shown to be $O(1/n^{1/2})$. [In [5] it was shown to be $O(1/n^{(q-1)/q})$ if $\|h\|$ is (more robustly) norm in $L_q(P)$, $q \le 2$).] This can now be applied to the two examples:

EXAMPLE 1. (Cont.) (1) Take $F_2 = \{g(a^t x)\colon g \in \mathbf{G},\ a \in \mathbf{A}\}$. In [11] it was further shown that, if the Fourier transform of $f(x)$ has bounded $L_1(\mathbf{R}^d)$ norm, then $f(x)$ lies in the closure of the convex hull of multiples of sinusoidal functions bounded in absolute value by some fixed B. Taking these multiples to be F_1, one applies the relaxed greedy procedure and establishes that PPR (relaxed approximation form) converges at rate $O(1/n^{1/2})$. For the sampling case one may use a variable bandwidth smooth of the data $D(a, \alpha) = \{(a^t x_i, y_i - (1 - \alpha)f_n(x_i))\}$, call it h_α, for the term αg at step n and then determine a_n and α_n which solve

$$\text{(RS)} \qquad \min_{\alpha, a} \sum((1 - \alpha)r_i + \alpha y_i - h_\alpha(a^t x_i))^2.$$

In [6] a modification which is nonlinear in f_n was proposed which accelerated convergence for several practical problems.

EXAMPLE 2. (Cont.) (2) if $f(x)$ is in the closure of the convex hull of $F_1 = \{c\sigma(\mathrm{ra}^t x - t)\colon |c| \leq B; a, r, t \text{ arbitrary}\}$ take $F_2 = \{c\sigma(ra^t x - t)\colon c, a, r, t \text{ arbitrary}\}$ and apply the relaxed algorithm. Then the output error in $L_2(P)$ norm is $O(1/n^{1/2})$. In [1] it was shown that such a B exists if P has bounded support and the Fourier transform of $f(x)$ has a finite first moment in $L_1(\mathbf{R}^d)$. In the sampling form, at step n, one finds the term $\alpha g = c\sigma(ra^t x - t)$ which fits $\{(x_i, y_i - (1-\alpha)f_n(x_i))\}$ so that the average squared error is minimized as a function of a, α, r, c, t. If the a's in $\mathbf{A}$ are restricted to have a limited number of non-zero components and $f(x)$ is in the closure of the convex hull of bounded multiples of the associated sigmoids then it follows from [14] that for each $\varepsilon > 0$ we may efficiently construct a network whose output is within ε of $f(x)$ with probability at least $1 - \varepsilon$ from a sample whose size is bounded by a polynomial in d and $1/\varepsilon$. In cases where $\mathbf{A}$ is constrained due to higher order smoothness assumptions on $f(x)$, some practical bounds on expected mean squared error are computed in [12] explicitly as a function of sample size for the ridge estimation problem.

Consider this pure version of the relaxed extension: at step n choose $g \in F_1$ and real c such that $\|f - f_n - cg\|$ is minimum and set $f_{n+1}(x) = f_n(x) + cg(x)$. In [4] it is shown that the approximation error is $O(1/n^{1/6})$. Finally, we remark that certain regularity assumptions were necessary in the cited references for the minima to exist. In several of these investigations, variants of the results were given where at step n one needs only to be within $\tau(n)$ of the infimum for some (tolerance) sequence $\tau(n)$. We will formulate our local approximation results in terms of such sequences.

Another topic of current interest is the computational tractability (capability of being solved with a number of steps which is polynomial in the description length of the data set) of nearly determining the best neural network parameters based on the sample $\{x_i, y_i\}$. When a linear combination of two (or more) sigmoidal functions is sought which (nearly) best fits a sample, some negative results have been obtained: in [3] it is shown that the problem of determining whether two classes of vectors can be separated by thresholding a linear combination of two Heaviside sigmoidal functions [$\sigma(x)$ is 1 for $x \geq 0$, 0 for $x < 0$] is intractable. In [13] it is shown that the problem of finding a linear combination of two Lipschitzian sigmoidal functions for which the sum of squared errors (under suitable normalization of the response data) is within 1/10 of the infimum possible is intractable. Similar results were obtained in [13] for achieving a sum of squared errors within $1/(4n^5)$ of the infimum possible by convex combinations of n sigmoidal functions. In [16] the results of [13] were improved by demonstrating that sum of squared errors could be replaced by average squared error (but with somewhat smaller dimension dependent distances to the infimum). In particular, the techniques of [16], based on recent results on approximation algorithms for NP-Hard problems, yielded a proof of the intractability of fitting the sample with a multiple of one sigmoid to

within a constant of the infimum in the least squares sense. Hence greedily training one node at a time (to within the previous constant of the infimum squared error) is intractable. However these negative results are all worst case analyses and are asymptotic in nature. "Average" complexities may be polynomial. More research is needed in this direction. Also, with the massively parallel architectures of future machines, it may be feasible to perform 100 greedy steps for dimensions about 20 (while training a 100-node network with its 2200 parameters might still be too time-consuming). Hence further investigation of greedy approximation schemes is warranted.

3. A local approximation criterion and its motivation. Unfortunately, as we pointed out earlier, the above methods yield global expansions which have the following drawbacks for the local estimation problem:

1. The ridge approximation is designed to fit the data in the sense of an average and not necessarily to fit well close to any particular point.
2. An expansion is assumed to exist which is approximately valid at every x; simpler expansions may be valid locally and these may be more easily estimated.
3. A local expansion may be easier to interpret.

We now motivate the form of our proposed method. Henceforth **A** consists of a subset of unit vectors in $\mathbf{R}^d$ and **G** is a subset of real valued functions on **R**. Call the ridge functions formed from **A** and **G** admissible. By sphering the available data and shifting so that $f(x)$ is desired at or close to the origin 0, we search for ridge expansions which approximately hold in a region containing a ball of radius ε about 0. Assume a relaxed greedy approximation in the nth step: set $f_0(x) = 0$ and $f_{n+1}(x) = (1-\alpha)f_n(x) + \alpha g(a^t x)$ with $\alpha g(a^t x)$ best fitting $f(x) - (1-\alpha)f_n(x)$ in this region. The right norm for this approximation one might argue, is $|||h||| = \|h(x)I(x^t x < \varepsilon^2)\|$ where $I(\cdot)$ = indicator function for $\cdot$. However, in the practical sampling situation there is little chance of any sample vectors lying in the ε ball (hence little chance of estimating this norm) unless our sample size is exponential in d.

So we need a norm which measures closeness to 0 in a weaker sense. If closeness is characterized by nearness in a one-dimensional projection, then appropriate norms are of the form

$$\|h\|_{a1} = \|h(x)I(|a^t x| < \varepsilon)\|$$

For uniform P these norms give equal weight at P-supported points in $|a^t x| < \varepsilon$. If a term $g(a^t x)$ were (nearly) linear in this region then this fact would be consistent with the equal weighting of squared deviations at each sample predictor in the simple linear prediction problem. Before discussing how to define the best approximation in the nth step, we make a minor modification. Note that a stable reconstruction from projections should require some information from the region $|a^t x| \geq \varepsilon$. Hence we further define

$$\|h\|_{a2} = \|h(x)I(|a^t x| \geq \varepsilon)\|$$

and propose that the appropriate norms for measuring closeness are given by

$$\|h\|_a^2 = \|h\|_{a1}^2 + w\|h\|_{a2}^2, \tag{2}$$

where w is a regularization parameter, $0 < w < 1$. In practical implementations w may be varied and optimally determined using cross validation. Here we discuss only rates of approximation using these norms. An interesting open question in the case $w = 0$ is posed in Section 3.

Although we only want an expansion near 0 we will produce an expansion on each subset in a partition of $\mathbf{R}^d$. We now motivate the criterion for the best approximation in the nth step. Since different a's will be used in different steps and the norms in (2) are difficult to compare, we argue that the appropriate criterion to be minimized as a function of a is

$$\|f - f_{n+1}\|_a / \|f - f_n\|_a.$$

Thus we should search for the direction vector in $\mathbf{A}$ which optimizes the percentage decrease in approximation error when measured relative to this direction. It seems quite plausible that, say for $\varepsilon = w = 1/4$, considerably more local information will be obtained by such a procedure than by the global algorithm. Indeed for the simple preliminary example (with just the d coordinate directions) the g_i with smallest $\|\ \|$ would be chosen first if its support were in a band of width 1/2 and its $\|\ \|$ exceeds 0.5 of that of any other g_i, provided the other g_i's had supports outside the 1/4-ball. Thus practical use is anticipated where neither ε nor w is very small. If we now allow α to take two different values in the nth step (depending on whether ε exceeds $a^t x$), then the approximation error is shown in Section 2 to be $O(1/n^{1/2})$ in the original norm, $\|\ \|$.

4. The local greedy approximation algorithm and its convergence. The precise description of the algorithm proceeds as follows: Let $f_0 = 0$ and $e_n = f - f_n$. Given $n > 0$, let

$$\mathscr{R}_{n+1} = \operatorname*{infimum}_{\substack{0\le\alpha\le 1,\ 0\le\beta\le 1\\ a\in\mathbf{A},\ g\in\mathbf{G}}} \frac{\|f - (1-\alpha)f_n - \alpha g(a^t x)\|_{a1}^2 + w\|f - (1-\beta)f_n - \beta g(a^t x)\|_{a2}^2}{\|f - f_n\|_{a1}^2 + w\|f - f_n\|_{a2}^2}$$

Let $\tau(n) = O(1/n)$ be a decreasing (tolerance) sequence of positive real numbers. Now find any α, β, a and g subject to the constraints of the preceding infimum such that

$$R_{n+1} = \frac{\|f - f_{n+1}\|_a^2}{\|f - f_n\|_a^2} = \frac{\|f - f_{n+1}\|_{a1}^2 + w\|f - f_{n+1}\|_{a2}^2}{\|f - f_n\|_{a1}^2 + w\|f - f_n\|_{a2}^2} < \mathscr{R}_{n+1} + \tau(n),$$

where

$$f_{n+1}(x) = (1-\alpha)f_n(x) + \alpha g(a^t x) \quad \text{if } |a^t x| < \varepsilon \tag{3}$$

$$= (1-\beta)f_n(x) + \beta g(a^t x) \quad \text{if } |a^t x| \ge \varepsilon. \tag{4}$$

(If the ratio R_{n+1} can be made smaller by changing α and/or β to 0, then we do so.)

By the condition in parentheses $R_{n+1} \le 1$; f_{n+1} is a fixed convex combination of $n+1$ ridge functions provided $x^t x < \varepsilon^2$. Outside of this ball, f_{n+1} is a different convex combination of the same ridge functions. Also in any neighborhood of the form $\mathcal{N}_k = \{x\colon |a_i^t x| < \varepsilon \text{ for } i = 1, 2, \ldots, k\}$, where a_i is the ith unit vector of the algorithm, f_{k+1} is a fixed convex combination of $k+1$ ridge functions and f_{n+1} is a convex combination of f_{k+1} and ridge functions from steps beyond k.

Consider the application to the two examples (see Section 2). In Example 1, if a, α, β are chosen at step n, the infimum R_{n+1} is obtained using $E(Y-(1-\alpha)f_n(X) \mid a^t X)$ for αg in (3) and $E(Y-(1-\beta)f_n(X) \mid a^t X)$ for βg in (4). Hence in the sampling case at step n use a smooth h_α of $D(a, \alpha)$ for αg in (3) and a smooth h_β of $D(a, \beta)$ for βg in (4) where a, α, β are chosen to solve

$$\text{(RLS)} \qquad \min_{\alpha,\beta,a} \frac{\sum_{i:\ |a^t x_i|<\varepsilon}((1-\alpha)r_i+\alpha y_i-h_\alpha(a^t x_i))^2+w\sum_{i:\ |a^t x_i|\ge\varepsilon}((1-\beta)r_i+\beta y_i-h_\beta(a^t x_i))^2}{\sum_{i:\ |a^t x_i|<\varepsilon} r_i^2+w\sum_{i:\ |a^t x_i|\ge\varepsilon} r_i^2}$$

A pure greedy local sampling algorithm is obtained by setting $\alpha = \beta = 0$ in the above. In this case the objective function is a ratio of weighted versions of the original PPR objective function of (S). The optimization is still in $d-1$ dimensions. Relaxation increases it only by 2. For Example 2 at step n set $g = c\sigma(ra^t x - t)$ in (3) and (4) where $a, \alpha, \beta, c, r, t$ are chosen to minimize the objective function of (RLS) one would obtain after replacing the h_α term by $\alpha c\sigma(ra^t x_i - t)$ and the h_β term by $\beta c\sigma(ra^t x_i - t)$.

The rest of this article is devoted to showing the algorithm convergence in $\|\ \|$ at the $n^{-1/2}$ rate. In the remaining analysis the following simple fact will be repeatedly used: $(x+c)/(x+d)$ is increasing for positive x if $c \le d$ and decreasing for positive x if $c \ge d$. Recall that, for any a, $\|\cdot\|^2 = \|\cdot\|_{a1}^2 + \|\cdot\|_{a2}^2$.

LEMMA 1. *Let $M = (1-w)/w$. Then for any n, $\lambda(0 \le \lambda \le 1)$, $h \in \mathbf{G}$, $u \in \mathbf{A}$,*

$$\mathscr{R}_{n+1} \le \frac{M + \|f-(1-\lambda)f_n - \lambda h(u^t x)\|^2/\|f-f_n\|^2}{M+1}.$$

PROOF. If $\|f-(1-\lambda)f_n - \lambda h(u^t x)\| \ge \|f-f_n\|$, then the result is obvious; otherwise for any $\alpha(0 \le \alpha \le 1)$,

$$\mathscr{R}_{n+1} \le \frac{[(1-w)\|f-(1-\alpha)f_n-\alpha h(u^t x)\|_{u1}^2+w(\|f-(1-\alpha)f_n-\alpha h(u^t x)\|_{u1}^2+\|f-(1-\lambda)f_n-\lambda h(u^t x)\|_{u2}^2)]}{(1-w)\|f-f_n\|_{u1}^2+w\|f-f_n\|^2}$$

Now we can ensure that

$$\begin{aligned}&\|f-(1-\alpha)f_n-\alpha h(u^t x)\|_{u1}^2\\ &\quad\le \min\{\|f-f_n\|_{u1}^2, \|f-(1-\lambda)f_n-\lambda h(u^t x)\|_{u1}^2\}\end{aligned}$$

by taking $\alpha = 0$ or λ. Therefore,

$$\mathscr{R}_{n+1} \le \frac{(1-w)\|f-f_n\|_{u1}^2+w\|f-(1-\lambda)f_n-\lambda h(u^t x)\|^2}{(1-w)\|f-f_n\|_{u1}^2+w\|f-f_n\|^2}$$

$$= \frac{((1-w)\|f-f_n\|_{u1}^2)/(w\|f-f_n\|^2)+\|f-(1-\lambda)f_n-\lambda h(u^t x)\|^2/\|f-f_n\|^2}{((1-w)\|f-f_n\|_{u1}^2)/(w\|f-f_n\|^2)+1}$$
$$\leq \frac{M+\|f-(1-\lambda)f_n-\lambda h(u^t x)\|^2/\|f-f_n\|^2}{M+1}.$$

LEMMA 2.

$$\frac{\|e_{n+1}\|^2}{\|e_n\|^2} \leq \frac{M+R_{n+1}}{M+1}$$

where $M=(1-w)/w$.

PROOF. For the α, β, a and g defining f_{n+1},

$$\frac{\|e_{n+1}\|^2}{\|e_n\|^2} = \frac{\|e_{n+1}\|_a^2+(1-w)\|e_{n+1}\|_{a2}^2}{\|e_n\|_a^2+(1-w)\|e_n\|_{a2}^2}$$
$$= \frac{R_{n+1}+((1-w)\|f-(1-\beta)f_n-\beta g(a^t x)\|_{a2}^2)/(\|f-f_n\|_{a1}^2+w\|f-f_n\|_{a2}^2)}{1+((1-w)\|f-f_n\|_{a2}^2)/(\|f-f_n\|_{a1}^2+w\|f-f_n\|_{a2}^2)}.$$

Since $\|f-(1-\beta)f_n-\beta g(a^t x)\|_{a2}^2 \leq \|f-f_n\|_{a2}^2$ (otherwise R_{n+1} could be reduced by changing β to 0), the above is

$$\leq \frac{R_{n+1}+((1-w)\|f-f_n\|_{a2}^2)/(\|f-f_n\|_{a1}^2+w\|f-f_n\|_{a2}^2)}{1+((1-w)\|f-f_n\|_{a2}^2)/(\|f-f_n\|_{a1}^2+w\|f-f_n\|_{a2}^2)}$$
$$\leq \frac{R_{n+1}+(1-w)/w}{1+(1-w)/w} = \frac{M+R_{n+1}}{M+1}.$$

LEMMA 3. *Suppose f lies in the closure of the convex hull of a class $\mathbf{Q}$ of admissible ridge functions with each element of $\mathbf{Q}$ having $\|\ \| \leq B$. Then*

$$E_{n+1} = \operatorname*{infimum}_{\substack{0\leq\lambda\leq 1\\ h\text{admissible}}} \frac{\|f-(1-\lambda)f_n-\lambda h\|^2}{\|f-f_n\|^2} \leq \frac{4B^2}{4B^2+\|e_n\|^2}.$$

PROOF (From [11]). 0 lies in the closure of the convex hull of $f-\mathbf{Q}$. Hence,

$$\operatorname*{infimum}_{h\in\mathbf{Q}}(f-f_n, f-h) \leq 0,$$

where $(,)$ is the inner product in $L_2(P)$. Therefore, writing $f-(1-\lambda)f_n-\lambda h$ as $(1-\lambda)(f-f_n)+\lambda(f-h)$ and expanding the infimum expression for E_{n+1} in terms of $(,)$, we get

$$E_{n+1} \leq \operatorname*{infimum}_{\substack{0\leq\lambda\leq 1\\ h\in\mathbf{Q}}} \frac{(1-\lambda)^2\|e_n\|^2+\lambda^2\|f-h\|^2+2\lambda(1-\lambda)(f-f_n, f-h)}{\|e_n\|^2}$$
$$\leq \operatorname*{infimum}_{0\leq\lambda\leq 1} \frac{(1-\lambda)^2\|e_n\|^2+4\lambda^2B^2}{\|e_n\|^2},$$

which is [take $\lambda = \|e_n\|^2(\|e_n\|^2+4B^2)^{-1}$] $\leq 4B^2/(4B^2+\|e_n\|^2)$.

THEOREM. *Assume f lies in the closure of the convex hull of $\mathbf{Q}$, a subset of admissible ridge functions with $L_2(P)$ norms bounded by B. Then $\|e_n\|$ is $O(1/n^{1/2})$. In particular, the local greedy approximation algorithm converges at the rate $O(1/n^{1/2})$ for the PPR and neural network training examples. Also f_n converges at the rate $O(1/n^{1/2})$ in the neighborhood $\mathcal{N}_k$, where it is a convex combination of ridge functions with those from the first k steps having the same relative weights.*

PROOF. By Lemma 2,

$$\frac{\|e_{n+1}\|^2}{\|e_n\|^2} \le \frac{M + R_{n+1}}{M+1} \le \frac{M + \mathscr{R}_{n+1} + \tau(n)}{M+1}$$

which is

$$\le \frac{M + ((M + E_{n+1})/(M+1)) + \tau(n)}{M+1}$$

by Lemma 1 and the definition of E_{n+1}. Finally, by Lemma 3, we get

$$\begin{aligned}
\frac{\|e_{n+1}\|^2}{\|e_n\|^2} &\le \frac{M + (M + (4B^2/(4B^2 + \|e_n\|^2)))/(M+1) + \tau(n)}{M+1} \\
&= \frac{H + (4B^2/(4B^2 + \|e_n\|^2))}{H+1} + \frac{\tau(n)}{M+1} \quad (\text{setting } H = M^2 + 2M) \\
&= \frac{4B^2H + H\|e_n\|^2 + 4B^2}{4B^2H + H\|e_n\|^2 + \|e_n\|^2 + 4B^2} + \frac{\tau(n)}{M+1} \quad \text{which is} \\
&\le \frac{4B^2H + H\|e_0\|^2 + 4B^2}{4B^2H + H\|e_0\|^2 + \|e_n\|^2 + 4B^2} + \frac{\tau(n)}{M+1} \\
&\qquad (\text{since } \|e_n\| \text{ is decreasing by Lemma 2}) \\
&= \frac{T + \tau'(n)}{T + \|e_n\|^2} \quad [\text{setting } T = 4B^2H + H\|e_0\|^2 + 4B^2 \text{ and} \\
&\qquad\qquad \tau'(n) = \tau(n)(T + \|e_n\|^2)(M+1)^{-1}].
\end{aligned}$$

Note that $\tau'(n)$ is decreasing and is $O(1/n)$. Now, rewriting the inequality and iterating, we get

$$\begin{aligned}
\frac{1}{\|e_{n+1}\|^2} &\ge \frac{1}{\|e_n\|^2} \frac{T + \tau'(n) + \|e_n\|^2 - \tau'(n)}{T + \tau'(n)} = \frac{1}{\|e_n\|^2} + \frac{1 - (\tau'(n)/\|e_n\|^2)}{T + \tau'(n)} \\
&\ge \frac{1}{\|e_l\|^2} + \sum_{k=l}^{n} \frac{1 - (\tau'(k)/\|e_k\|^2)}{T + \tau'(k)}.
\end{aligned}$$

Now $\tau'(n) \le Sn^{-1}$ for some S and all $n = 1, 2, 3, \ldots$. Let us first examine even n's. Suppose $\|e_n\|^2 > 4Sn^{-1}$. Then the numerators of the above summands are all $\ge 1/2$ for $k \ge n/2$. Take $l = n/2$. It follows that

$$\frac{1}{\|e_{n+1}\|^2} \ge \frac{n}{4(T + \tau'(0))} \quad \text{or} \quad \|e_{n+1}\|^2 \le 4(T + \tau'(0))n^{-1}.$$

If, on the other hand, $\|e_n\|^2 \le 4Sn^{-1}$ for the even n in question then $\|e_{n+1}\|^2 \le 4Sn^{-1}$ by the monotonicity of $\|e_n\|$. Hence $\|e_n\|^2$ is $O(1/n)$ for n odd. Using this monotonicity one more time, we see that $\|e_n\|^2$ is $O(1/n)$ for n even and the proof is complete. □

5. Conclusions. We have derived the form of a local greedy approximation based on experiences with the curse of dimensionality. Convergence at the $1/n^{1/2}$ rate in the $L_2(P)$ norm has been established. An interesting open question is under what conditions does $1/n^{1/2}$ convergence hold in the ε ball for the case $w = 0$?

As has been noted, this same form [see (3)(4)] of relaxed greedy approximation may be motivated when $\|f - f_n\|^2$, $\|f - f_n\|_{a1}^2$, etc. are replaced by expected values of a more general loss function $\Psi(y, \delta)$, that is, $E(\Psi(Y, f_n(X)))$, $E(\Psi(Y, f_n(X))I(|a^t X| < \varepsilon))$, etc. At step n in the sampling case α, β, a, g are chosen to minimize

$$\frac{\sum_{i:|a^t x_i|<\varepsilon} \Psi(y_i, (1-\alpha)f_n(x_i) + \alpha g(a^t x_i)) + w \sum_{i:|a^t x_i|\ge\varepsilon} \Psi(y_i, (1-\beta)f_n(x_i) + \beta g(a^t x_i))}{\sum_{i:|a^t x_i|<\varepsilon} \Psi(y_i, f_n(x_i)) + w \sum_{i:|a^t x_i|\ge\varepsilon} \Psi(y_i, f_n(x_i))}$$

w may be interpreted as a local–global trade-off parameter and should be an important choice in implementation with data. One open question is at what rate is the approximation form convergent for general loss functions?

Acknowledgments. The author thanks A. R. Barron and Y. Makovoz for valuable suggestions in preparing the manuscript.

REFERENCES

[1] BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* **40** 930–945.

[2] BOTTOU, L. and VAPNIK, V. (1992). Local learning algorithms. *Neural Computation* **4** 888–900.

[3] BLUM, A. L. and RIVEST, R. L. (1992). Training a 3-Node Neural Network is NP-Complete. *Neural Networks* **5** 117–127.

[4] DEVORE, R. A. and TEMLYAKOV, V. N. (1996). Some remarks on greedy algorithms. *Adv. Comput. Math* **5** 173–187.

[5] DONOHUE, M. J., GURVITZ, L., DARKEN, C. and SONTAG, E. (1994). Rates of convex approximation in non-Hilbert spaces. *Constr. Approx.* **13** 187–220.

[6] FLICK, T. E., JONES, L. K., PRIEST, R. and HERMAN, C.(1990). Projection pursuit classification. *Pattern Recognition* **23** 1367–1376.

[7] FRIEDMAN, J. H. (1999). Greedy function approximation: a gradient boosting machine. Technical report, Stanford Univ.

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

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

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

[11] JONES, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist.* **20** 608–613.

[12] JONES, L. K. (1994). Good weights and hyperbolic kernels for neural networks, projection pursuit, and pattern classification: Fourier strategies for extracting information from high-dimensional data. *IEEE Trans. Inform. Theory* **40** 439–454.
[13] JONES, L. K. (1997). The Computational intractability of training sigmoidal neural networks. *IEEE Trans. Inform. Theory* **43** 167–173.
[14] LEE, W. S. and BARTLETT, P. L., and WILLIAMSON, R. C. (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inform. Theory* **42** 2118–2132.
[15] REJTO, L. and WALTER, G. G. (1992). Remarks on projection pursuit regression and density estimation. *Stochastic Anal. Appl.* **10** 213–222.
[16] VU, V. H. (1998). On the infeasibility of training neural networks with small mean-squared error. *IEEE Trans. Inform. Theory* **44** 2892–2900.

DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF MASSACHUSETTS LOWELL
ONE UNIVERSITY AVENUE
LOWELL, MASSACHUSETTS 01854