REGRESSION EFFICACY AND THE CURSE OF DIMENSIONALITY[‡]

MAXWELL B. STINCHCOMBE[*] AND DAVID M. DRUKKER[†]

ABSTRACT. This paper gives a geometric representation of a class of non-parametric regression estimators that includes series expansions (Fourier, wavelet, Tchebyshev and others), kernels and other locally weighted regressions, splines, and artificial neural networks. For any estimator having this geometric representation, there is no curse of dimensionality — asymptotically, the error goes to 0 at the parametric rate. Regression efficacy measures the amount of variation in the conditional mean of the dependent variable, Y, that can be achieved by moving the explanatory variables across their whole range. The dismally slow, dimension-dependent rates of convergence are calculated using a class of target functions in which efficacy is infinite, and the analysis allows for the possibility that the dependent variable, Y, may be an ever-receding target.

I. INTRODUCTION

The starting point is a probability space $(\Omega, \mathcal{F}, P)$ and an independent and identically distributed (iid) sequence $(Y_i, (X_{1,i}, X_{2,i}, \ldots))_{i=1}^n$ in $L^p(\Omega, \mathcal{F}, P)$, $p \in [1, \infty)$. Interest centers on estimating the target functions,

$$f_d(x_1, \ldots, x_d) := E\left(Y \middle| (X_1, \ldots, X_d) = (x_1, \ldots, x_d)\right) \tag{1}$$

from the realization $(Y_i(\omega), (X_{1,i}(\omega), \ldots, X_{d,i}(\omega)))_{i=1}^n$. This paper provides an asymptotic analysis of the question, "How large must n be to nonparametrically estimate $f_d(\cdot)$ to any given degree of precision?" Of particular interest is the relation between d and n.

I.A. **Different Answers.** There are, in the literature, two very different answers, the usual one, due to Stone (1982), is applicable to all nonparametric regression techniques, the second due to Barron (1993), is applicable to the nonparametric regression technique known as single-layer feedforward (slff) artificial neural networks (ann's) with sigmoidal activation functions.

1. The usual asymptotic analysis yields the following answer: if f_d belongs to a particular dense class, $\mathbb{V}_d^{Lip}$, then for a desired degree of precision, ϵ, there is a constant C, independent of ϵ, such that n must satisfy $n^{-\frac{1}{2+d}} < C\epsilon$. The C may depend on the distribution of the data and the nonparametric technique. Further, all nonparametric techniques have this property.
2. The slff ann analysis yields the following answer: if f_d belongs to a different dense class, $\mathbb{V}_d^{ann}$, then for a desired degree of precision, ϵ, there is a constant C, independent of ϵ but dependent on d, such that n must satisfy $n^{-\frac{1}{2}} < C\epsilon$. As above, C may also depend

Date: February 15, 2012.

[‡]We owe many thanks to Xiaohong Chen, Graham Elliot, Jinyong Hahn, James Hamilton, Qi Li, Dan Slesnick, Hal White, Paul Wilson, and an anonymous referee for numerous insights, questions, references, conversations and corrections.

on the distribution governing the data and on the non-parametric regression technique through the specific choice of sigmoidal activation function.

If $C\epsilon = 1/100$ for both approaches, and d is a largish positive integer, say 7, the usual analysis suggests that one needs 10^{18} independent observations, not a practical data requirement, while the ann analysis suggests that one needs but 10^4 independent observations, a large but not impractical data requirement. This impracticality is known as "the curse of dimensionality."[1] It should be noted that the dependence of C on d in the ann analysis might, in principle, lead to the re-emergence of the curse. This paper shows that the ann type of analysis can be done with a constant that does not depend on d for a very wide collection of nonparametric techniques.

I.B. **The Source of the Difference.** The difference in the two types of analysis arises from different assumptions about the classes, $\mathbb{V}_d^{Lip}, \mathbb{V}_d^{ann} \subset L^p(\Omega, \mathcal{F}, P)$, containing the target functions $f_d(\cdot)$. In both cases, the assumed classes are dense, and being dense, they are impossible to reject on the basis of data with smooth measurement error. The **regression efficacy** of explanatory variables is the amount by which the conditional mean of Y varies as the values of $(X_1, \ldots, X_d)$ move across their range.[2] The boundedness/unboundedness of regression efficacy and the possibility/impossibility of ever-receding targets are two of the ways in which the classes differ.

1. In the dense class $\mathbb{V}_d^{Lip}$ used in the curse analysis, efficacy is unbounded in d, the number of explanatory variables. This means that there may infinitely many groups of explanatory variables, each of them having the same ability to vary the conditional mean of Y. By contrast, the dense classes used in the ann analysis has a bound on efficacy that is independent of d. This argument should not be taken as being a final statement of affairs, although unbounded regression efficacy is counter-intuitive, we give an example below of a sequence $(Y, (X_1, X_2, \ldots, X_d, X_{d+1}, \ldots))$ with efficacy that is unbounded in d.
2. The target functions in (1) above work with a fixed Y, as does the ann analysis. In particular this means that there is a fixed joint distribution governing the data. Implicit in the curse analysis is the possibility that we are varying Y as we vary d — instead of calculating the errors in our attempts to estimate $E\left(Y\middle|(X_1, \ldots, X_d)\right)$, we may be calculating the errors in an attempt to estimate $E\left(Y_d\middle|(X_1, \ldots, X_d)\right)$ where the sequence Y_d may be divergent, i.e. ever-receding.

I.C. **Outline.** The next section begins with notation, two norms, and the basic implications that come from breaking up total errors into an approximation errors and estimation errors. It then explains how the main result of the paper, Theorem A yields the result that the total error is bounded by the estimation error in nonparametric regression. The two norms are

[1]The immense literature following on Stone's analysis has expanded the curse results far beyond his use of the sup norm to include the L^p-norms, the Sobolev norms, examined the partial role that smoothness of the target can play in overcoming the curse, and extended the analysis well beyond regression problems. Barron worked with single-layer feedforward ann's, as did Mhaskar and Michelli (1995) in a slightly different context. Yukich, Stinchcombe and White (1995) improved Barron's result in several directions, Chen and White (1999) improved it even further, Chen (2007) is a survey.

[2]From the *Oxford English Dictionary*, efficacy is the "Power or capacity to produce effects." While we think of regression efficacy as causal efficacy, the referee has quite correctly pointed out that there need not be a causal or even a structural component to efficacy as discussed here.

the Lipschitz norm and the efficacy norm, which is a variant of Arzelá's multidimensional variation norm.[3] The approximation error part of the curse analysis uses sets of targets functions having uniformly bounded Lipschitz norm, the approximation error part of the ann analysis uses sets of targets having uniformly bounded efficacy norm.

The following section gives two kinds of intuition about efficacy. The first has to do with the change in the amount of 'information' contained in $(X_1, \ldots, X_d)$ and the amount contained in $(X_1, \ldots, X_{d+1})$. The second compares the implications of Lipschitz bounds and of efficacy bounds in the special case that the conditional mean functions are affine and the regressors, $(X_1, \ldots, X_d)$ are independent. In this particular case, one can directly see how bounded/unbounded efficacy works, and how ever-receding targets can arise.

The penultimate section gives the dimension independent geometric representation of nonparametric regression estimators, and demonstrates that several of the well-known estimators have this structure. The last section gives possible extensions and conclusions.

II. Norms, Density, and Rates

We begin with notation, then turn to the contrasting norms and their basic denseness property. After this, we turn to the source of the curse results and the contrast with the ann results.

II.A. **Notation.** $L^0 = L^0(\Omega, \mathcal{F}, P)$ denotes the set of $\mathbb{R}$-valued random variables, $L^p = L^p(\Omega, \mathcal{F}, P) \subset L^0$ the set of random variables with finite p'th norm, $p \in [1, \infty)$. For any sub-σ-field $\mathcal{G} \subset \mathcal{F}$, $L^0(\mathcal{G}) \subset L^0$ is the set of $\mathcal{G}$-measurable random variables and $L^p(\mathcal{G}) := L^p \cap L^0(\mathcal{G})$.

$\mathbb{X} = \{X_a : a \in \mathbb{N}\} \subset L^2$ denotes the set of possible explanatory variables, $\mathscr{X}_d$ denotes $\sigma(X_1, \ldots, X_d)$, the smallest σ-field making $X_1, \ldots, X_d$ measurable, and $\mathscr{X}$ denotes $\sigma(\mathbb{X})$, the smallest σ-field making every X_a in $\mathbb{X}$ measurable. We assume that $Y \in L^p$ for some $p \in [1, \infty)$ so that the set of all conceivable target functions is $L^p(\mathscr{X})$. The set of all possible targets based on some finite set of regressors is $\bigcup_d L^p(\mathscr{X}_d)$, and this set is dense in $L^p(\mathscr{X})$.

II.B. **A Tale of Two Norms.** By Doob's Theorem (e.g. Dellacherie and Meyer (1978, Theorem I.18, p. 12-13)), $L^p(\mathscr{X}_d)$ is the set of functions of the form $\omega \mapsto g(X_1(\omega), \ldots, X_d(\omega))$ having finite p'th moment, g a measurable function from $\mathbb{R}^d$ to $\mathbb{R}$.

For each $d \in \mathbb{N}$, $C(\mathbb{R}^d)$ denotes the set of continuous functions on $\mathbb{R}^d$, and the obvious extension/restriction identifies $C(\mathbb{R}^d)$ with $C_d \subset C(\mathbb{R}^{\mathbb{N}})$, the elements of $C(\mathbb{R}^{\mathbb{N}})$ that depend on only the first d components, $(x_1, \ldots, x_d)$ of the infinite length vectors $(x_1, x_2, \ldots, x_d, x_{d+1}, \ldots)$. For $x, y \in \mathbb{R}^d$, $e_d(x, y) := \sqrt{(x-y)\cdot(x-y)}$ denotes the Euclidean distance between the d-dimensional vectors x and y.

Definition 1. *The* ***Lipschitz norm*** *of an $f_d \in C(\mathbb{R}^d)$ is*

$$\|f_d\|_{Lip} = \sup_{x \in \mathbb{R}^d} |f_d(x)| + \sup_{x \neq y} \frac{|f_d(x) - f_d(y)|}{e_d(x,y)}$$

whenever this is finite. The ***Lipschitz constant of*** *f_d is $\sup_{x \neq y} \frac{|f_d(x) - f_d(y)|}{e_d(x,y)}$. $C_d^{Lip}(B) \subset C(\mathbb{R}^d)$ denotes the set of f_d with Lipschitz norm B or less. A sequence of functions f_d in $C(\mathbb{R}^{\mathbb{N}})$ with $f_d \in C_d$ is* ***uniformly Lipschitz*** *if for some B, each f_d belongs to $C_d^{Lip}(B)$.*

[3] See Adams and Clarkson (1933) for an extensive comparison of the many non-equivalent definitions of bounded variation for functions of two or more variables.

We will be interested in the maximal total variability in the conditional mean of Y as the explanatory variables move monotonically across their range. Recall that the **total variation** of a function $f : \mathbb{R} \to \mathbb{R}$ is $TV(f) = \sup \sum_i |f(x_{i+1}) - f(x_i)|$ where the supremum is taken over all finite subsets $x_1 < x_2 < \cdots < x_I$ of $\mathbb{R}$. A **wide-sense monotonic path in** $\mathbb{R}^d$ is a function $t \mapsto x(t)$ from $\mathbb{R}$ to $\mathbb{R}^d$ such that for each $i \in \{1, \ldots, d\}$, the function $x_i(t)$ is either non-decreasing or non-increasing.

Definition 2. *The* ***monotonic total variation*** *of a function $f_d \in C(\mathbb{R}^d)$ is $MTV(f) = \sup_x TV(f_d \circ x)$ where the supremum is taken over wide-sense monotonic paths in $\mathbb{R}^d$.*[4] *The* ***monotonic total variation norm*** *or* ***efficacy norm*** *is*

$$\|f_d\|_{MTV} = |f_d(0)| + MTV(f_d)$$

whenever this is finite. $C_d^{MTV}(B) \subset C(\mathbb{R}^d)$ denotes the set of f_d with monotonic total variation norm B or less. A sequence of functions f_d in $C(\mathbb{R}^{\mathbb{N}})$ with $f_d \in C_d$ is ***uniformly efficacy bounded*** *if for some B, each f_d belongs to $C_d^{MTV}(B)$.*

The range of functions with a Lipschitz constant B grows with d, but not so quickly as their monotonic total variation.

Example 1. *If $f_d : [-1, +1]^d \to \mathbb{R}$ belongs to $C_d^{Lip}(B)$, then*

$$\left[\max_{x \in [-1,+1]^d} f_d(x) - \min_{y \in [-1,+1]^d} f_d(y)\right] \leq 2B\sqrt{d} \tag{2}$$

because $\max_{x,y \in [-1,+1]^d} e(x, y) = 2\sqrt{d}$. By contrast, for f_d having Lipschitz constant B, $MTV(f_d) \leq 2Bd$ because the longest monotonic paths in $[-1, +1]^d$ are of length $2d$.[5]

An implication of the previous example is that the ratio $\|f\|_{MTV}/\|f\|_{Lip}$ is unbounded on those parts of $C(\mathbb{R}^{\mathbb{N}})$ for which both norms are finite. The next example demonstrates that $\|f\|_{Lip}/\|f\|_{MTV}$ is also unbounded.

Example 2. *For $f_d(x) := \max\{0, 1 - e_d(x, 0)\}$ and $f_{d,n}(x) := \frac{1}{n} f_d(n^2 x)$, $\|f_{d,n}\|_{Lip} \uparrow \infty$ and $\|f_{d,n}\|_{MTV} \downarrow 0$.*

Lusin's theorem and standard approximation results deliver the following.

Lemma 1. *If $Y \in L^p(\Omega, \mathcal{F}, P)$, $f(x_1, x_2, \ldots) = E\left(Y \middle| (X_1, X_2, \ldots) = (x_1, x_2, \ldots)\right)$, $p \in [1, \infty)$ and $\epsilon > 0$, then there exists $g \in C(\mathbb{R}^{\mathbb{N}})$ such that $\|f - g\|_p < \epsilon$ and the sequence $g_d := E\left(Y \middle| (X_1, \ldots, X_d) = (x_1, \ldots, x_d)\right)$ is both uniformly Lipschitz and uniformly efficacy bounded.*

II.C. **Estimation and Approximation Errors.** Reiterating, interest centers on estimating $f_d(x_1, \ldots, x_d) := E\left(Y \middle| (X_1, \ldots, X_d) = (x_1, \ldots, x_d)\right)$ from iid data $(Y_i, (X_{1,i}, \ldots, X_{d,i}))_{i=1}^n$ assuming that each f_d belongs to some vector subspace, $\mathbb{V}'_d$, of $\mathbb{V}_d := L^p(\mathcal{X}_d)$. Let μ be the true joint distribution of the data and $\widehat{\mu}_n(\omega)$ the empirical joint distribution of the data. A sequence of non-parametric estimators, $\widehat{f}_n$, is typically of the form

$$\widehat{f}_n = \text{argmin}_{g \in \Theta_{\kappa(n)}} \left[\int (y - g(x))^2 \, d\widehat{\mu}_n(y, x)\right]^{1/2} \tag{3}$$

[4] Taking the supremum over the subset of monotonic *increasing* paths delivers the Arzelá norm.

[5] One could reconcile these by replacing $e(x, y)$ by the distance $d_1(x, y) = \sum_{i \leq d} |x_i - y_i|$ in the definition of Lipschitz functions, but this seems to be contrary to common usage.

where $(\Theta_\kappa)_{\kappa=1}^\infty$ is a sequence of subsets of $\mathbb{V}'_d$, $\kappa(n) \uparrow \infty$ and $(\Theta_\kappa)_{\kappa=1}^\infty$ are chosen so that $f \in \operatorname{climinf} \Theta_{\kappa(n)}$ with probability 1 (where $\operatorname{climinf} A_n = \{g \in \mathbb{V} : \forall \epsilon > 0,\ \|g - A_n\| < \epsilon \text{ for all large } n\}$ is the closed liminf of a sequence of sets A_n.).

A useful contrast with (3) arises if μ is perfectly known. Define $f^*_{\kappa(n)}$ as

$$f^*_{\kappa(n)} = \operatorname{argmin}_{g \in \Theta_{\kappa(n)}} \left[\int (y - g(x))^2 \, d\mu(y, x)\right]^{1/2}. \tag{4}$$

The total error, $\|\widehat{f}_n - f\|$, can be bounded by the sum of an estimation error, ϵ_n, and an approximation error, a_n,

$$\epsilon_n + a_n := \underbrace{\|\widehat{f}_n - f^*_{\kappa(n)}\|}_{\text{estimation error}} + \underbrace{\|f^*_{\kappa(n)} - f\|}_{\text{approx. error}} \geq \|\widehat{f}_n - f\|. \tag{5}$$

The larger is $\Theta_{\kappa(n)}$, the smaller is a_n. The tradeoff is that a larger $\Theta_{\kappa(n)}$ leads to overfitting, which shows up as a larger ϵ_n. Most analyses of $\|\widehat{f}_n - f\|$ begin with a dense set, $\mathbb{V}'_d \subset \mathbb{V}_d$, of targets. The set $\mathbb{V}'_d$ is chosen so that one can calculate $\epsilon_n(\kappa)$ and $a_n(\kappa)$ as functions of κ. With this in place, one then chooses $\kappa(n)$ to minimize $\epsilon_n(\kappa) + a_n(\kappa)$.

Stone (1982) showed that the "optimal" rate of convergence is $r_n = n^{-1/(2+d)}$. By optimal, Stone meant that if the sequence f_d is uniformly Lipschitz, then for any nonparametric regression technique, any sequence of estimators, $\widehat{f}_n$, satisfies

$$\|\widehat{f}_n - f\| \geq \mathcal{O}_P(n^{-1/(2+d)}), \tag{6}$$

and that some sequence satisfies (6) with equality.

By denseness, no data with smooth measurement error can reject the hypothesis that $f_d \in C_d^{Lip}$. It seems that this should make the Lipschitz assumption unobjectionable, but it is where dimensionality enters. An extremely clear example of how this works in $L^2(\mathcal{X})$ is Newey (1997). He shows that, if μ satisfies some easy-to-verify and quite general conditions and the target, f, satisfies the uniform approximation condition $\sup_x \inf_{g \in \Theta_\kappa} |f(x) - g(x)| = \mathcal{O}\left(\frac{1}{\kappa^\alpha}\right)$, then

$$\|f - \widehat{f}_n\|^2 = \mathcal{O}_P\left(\tfrac{\kappa}{n} + \tfrac{1}{\kappa^{2\alpha}}\right). \tag{7}$$

Ignoring some of the finer detail, the κ/n term in Newey's result corresponds to the square of the estimation error, and the $\kappa^{-2\alpha}$ to the square of the approximation error.[6] To balance the tradeoffs, one picks $\kappa = \kappa(n)$ to minimize $\frac{\kappa}{n} + \frac{1}{\kappa^{2\alpha}}$.

If $f : [-1, +1]^d \to \mathbb{R}$ has Lipschitz constant B, we must evaluate f at roughly $\left(\frac{2B}{\epsilon}\right)^d$ (carefully chosen) points to pin down f to within ϵ at all points in its domain. For many classes Θ_κ this yields, for every $f \in C_d^{Lip}$, $\sup_x \inf_{g \in \Theta_\kappa} |f(x) - g(x)| = \mathcal{O}\left(\frac{1}{\kappa^\alpha}\right)$ with $\alpha = \frac{1}{d}$. Minimizing $\frac{\kappa}{n} + \frac{1}{\kappa^{2/d}}$ yields $\kappa = n^{\frac{d}{2+d}}$, evaluating the minimand at the solution gives the cursed rate from (6),

$$\|f - \widehat{f}_n\|^2 = \mathcal{O}_P\left(\tfrac{\kappa}{n} + \tfrac{1}{\kappa^{2/d}}\right) = \mathcal{O}_P\left(n^{-\frac{2}{2+d}}\right), \text{ or } \|f - \widehat{f}_n\| = \mathcal{O}_P\left(n^{-\frac{1}{2+d}}\right). \tag{8}$$

Artificial neural networks can accurately fit sparse high dimensional data. A theoretical basis for this empirical observations was given in Barron (1993). He showed that for every d, there is a dense set of functions, $\mathbb{V}_d^{ann}$, depending on the architecture of the networks, such that for all $f \in \mathbb{V}_d^{ann}$, the following variant of the uniform approximation condition

[6] See his equation (A.3), p. 163, for the omitted detail.

(7), $\sup_x \inf_{g\in\Theta_\kappa} |f(x)-g(x)| \leq C(d)\left(\frac{1}{\kappa^\alpha}\right)$, is satisfied with $\alpha = \frac{1}{2}$. Ignoring the dependence on d in the constant $C(d)$, we have $\|f-\widehat{f}_n\|^2 = \mathcal{O}_P\left(\frac{\kappa}{n}+\frac{1}{\kappa^{2\alpha}}\right) = \mathcal{O}_P\left(\frac{\kappa}{n}+\frac{1}{\kappa}\right)$. Minimizing yields $\kappa(n)=\sqrt{n}$ so that $\|f-\widehat{f}_n\|^2 = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

It is in principle possible that $d \mapsto C(d)$ grows explosively enough to vitiate this analysis and return us to the curse world. As we will see, this need not happen, either for ann estimators, nor for any of the other main classes of nonparametric estimators.

Theorem A (below) shows that for any estimator having a particular geometric representation, for any r_n converging to 0, no matter how quickly, and any $\kappa(n)$ increasing to ∞, no matter how slowly, there exists a dense $\mathbb{V}' \subset \mathbb{V}$ for which the approximation error satisfies $a_n = \mathcal{O}(r_n)$. The geometric representation covers a class of non-parametric regression estimators that includes, but is not limited to, series expansions (Fourier, wavelet, Tchebyshev and others), kernels and other locally weighted regressions, splines, wavelets, and artificial neural networks.

Through the following steps, we have dimension independent rates of convergence: First, pick $\kappa(n) \uparrow \infty$ in such a fashion that consistency is guaranteed, typically this requires $\kappa(n)/n \downarrow 0$; Second, calculate $e_n = \|\widehat{f}_n - f^*_{\kappa(n)}\|$; Third, invoke Theorem A to guarantee the existence of a dense class of targets, $\mathbb{V}'$, such that for all $f \in \mathbb{V}'$, $a_n = \|f^*_{\kappa(n)} - f\| = \mathcal{O}(e_n)$. Fourth, observe that $\|\widehat{f}_n - f\| \leq e_n + a_n = \mathcal{O}(e_n)$.

III. Intuitions About Efficacy

To gain intuition about bounded/unbounded efficacy, we will first discuss the possible decrease in the amount of 'information' gained about Y in the move from the set of regressors $(X_1,\ldots,X_d)$ to the set of $(X_1,\ldots,X_{d+1})$. After this we will specialize to the case of affine conditional expectations and bounded range, non-degenerate, independent explanatory vectors $(X_1,\ldots,X_d)$. Here one can directly see how ever-receding targets may arise, and also count the differences in the numbers of regressors that matter.

III.A. **The Information Contained in a Set of Regressors.** In general, one does not expect that the X_k should be mutually independent. This might arise if the X_k are random draws from some larger set of possible explanatory variables. We first discuss the intuitions from the case that they arise from iid draws from $L^2(\Omega,\mathcal{F},P)$, then from the case that they arise from processes that are approximately recurrent.

Let $\Delta(L^2(\Omega,\mathcal{F},P))$ denote the set of probability distributions on $L^2(\Omega,\mathcal{F},P)$, here viewed as the set of possible explanatory variables. Suppose first that the X_k are iid draws from some $\nu \in \Delta(L^2)$, that is, suppose that there is some probability law generating the regressors from amongst all possible regressors. By the generalized Glivenko-Cantelli theorem, the empirical distribution, ν_d, of $(X_1,\ldots,X_d)$ converges to ν. This means that the additional explanatory power to be gained by projecting Y onto the span of $(X_1,\ldots,X_d)$ must be going to 0. Another way to see how this is operating is to note that the support of any ν must be approximately flat.

Another sort of intuition would come into play if the process generating the X_k had the property that one long set of regressors, say $X_k,\ldots,X_{k+n_1}$, contained much the same information about Y as could be found in $X_{k'},\ldots,X_{k'+n_2}$, where $k' > k+n_1$. Such a situation would arise if e.g. the X_k were drawn according to a smooth Markov process with a unique ergodic distribution ν. When the random variables X_k and $X_{k'}$ are close

to each other in $L^2(\Omega, \mathcal{F}, P)$, then continuity would tend to make the distribution of the next excursion from the neighborhood containing the two of them have close to the same distribution. This would mean that we would expect that the information in $X_{k'}, \ldots, X_{k'+n_2}$ that is above and beyond what can be found in $X_k, \ldots, X_{k+n_1}$ should be small.[7]

III.B. **Affine Conditional Expectations and Iid Regressors.** In the case that the regressors or iid and all of the conditional expectations of Y given $X_1, \ldots, X_d$ are affine, it is particularly easy to see how the difference between bounded and unbounded efficacy works, and how ever-receding targets can arise. For the rest of this section, and only for the rest of this section, we assume that:

(1) the X_k, are mutually independent, take values in $[-1, +1]$, have mean 0, and are not degenerate in the limit, i.e. $\liminf_k \operatorname{Var}(X_k) = \underline{\sigma} > 0$, and
(2) the condition expectation of Y is an affine function for all d, i.e. $f_d(x_1, \ldots, x_d) = E(Y|(X_1, \ldots, X_d) = (x_1, \ldots, x_d))$ is of the form $\beta_0 + \sum_{a \leq d} \beta_a x_a$ for some sequence $(\beta_a)_{a \in \mathbb{N}}$.

In this context, we examine how efficacy interacts with properties of Y; how the uniform Lipschitz assumption allows ever-receding targets; and how bounds on the numbers of important regressors work.

Several of the arguments depend on the *three series theorem* — if R_a is a sequence of independent random variables, then the convergence of the three series, $\sum_a P(|R_a| > c)$, $\sum_a E\left(R_a \cdot 1_{|R_a| \leq c}\right)$, and $\sum_a \operatorname{Var}\left(R_a \cdot 1_{|R_a| \leq c}\right)$ for some $c > 0$ implies that $\sum_a R_a$ converges a.e., and if $\sum_a R_a$ converges a.e., then the three series converge for all $c > 0$ (see e.g. Billingsley (2008, p. 290).)

We begin with an elementary result.

Lemma 2. *If $\beta_a \in \mathbb{R}$, $a \in \{0, 1, \ldots\}$ is a sequence in $\mathbb{R}$, then the sequence of affine function $f_d = \beta_0 + \sum_{a \leq d} \beta_a x_a$ on $[-1, +1]^d$ has uniform Lipschitz bound B if and only if $\sup_{A \subset \mathbb{N}} \sum_{a \in A} |\beta_a| \leq B\sqrt{\#A}$, and has uniform efficacy bound $2B$ if and only if $\sum_a |\beta_a| \leq B$.*

Proof. For any non-empty $A \subset \mathbb{N}$, if f is affine and $|\beta_a| \neq 0$ only for $a \in A$, then $\max_{x \neq y} |f(x) - f(y)|/e(x, y)$ is $\sum_{a \in A} |\beta_a|/\sqrt{\#A}$, yielding the first part of the Lemma. For the second part, note that the monotonic total variation of an affine f on $[-1, +1]^d$ is $2\sum_{a \leq d} |\beta_a|$. □

The condition $\sum_a |\beta_a| \leq B$ is the crucial part of Tibshirani's (1996) *least absolute shrinkage and selection operator* (lasso) models, and we will examine the connection in more detail in §III.D. Somewhat counterintuitively, one can have integrable Y, affine conditional expectations, and unbounded efficacy, i.e. $\sum_a |\beta_a| = \infty$.

Example 3. *Suppose that the X_a are iid and that $\beta_a = \mathcal{O}\left(\frac{1}{a}\right)$. For any $c > 0$, for all large a, $P(|R_a| > c) = 0$. This implies that for large a, $E\left(R_a \cdot 1_{|R_a| \leq c}\right) = 0$ and $Var\left(R_a \cdot 1_{|R_a| \leq c}\right) = \mathcal{O}\left(\frac{1}{a^2}\right)$. The requisite three series converge, so $Y_d := \beta_0 + \sum_{a \leq d} \beta_a X_a$ converges a.e. to some random variable Y. Since the variance of the Y_d is uniformly bounded, the Y_d are uniformly integrable, hence Y is integrable. Thus, conditional expectations can be affine while $\sum_a |\beta_a| = \infty$.*

[7] I am grateful to Graham Elliot and Jim Hamilton for these points.

III.C. **Receding Targets.** The affine structure plus the minimal assumptions on Y necessary for the existence of a target function lead to further restrictions on the β_a's.

Lemma 3. *If* $Y \in L^1(\Omega, \mathcal{F}, P)$ *and* $E(Y|(X_1, \ldots, X_d) = (x_1, \ldots, x_d)) = \beta_0 + \sum_{a \le d} \beta_a x_a$, *then* $\sum_a |\beta_a|^2 < \infty$.

Since $\mathrm{Var}(Y) = E(\mathrm{Var}(Y|X_1, \ldots, X_d)) + \mathrm{Var}(E(Y|X_1, \ldots, X_d))$, we know that the variance of the $f_d(X_1, \ldots, X_d)$ is bounded in d when $Y \in L^2(\Omega, \mathcal{F}, P)$. A slightly more involved argument yields the same conclusion more generally.

Proof. Martingale convergence implies that $Y_d := E(Y|(X_1, \ldots, X_d)) \to Y_{\mathscr{X}} := E(Y|\mathscr{X})$ a.e. If $\sum_a |\beta_a|^2$ diverges, then there exists an increasing sequence $1 = D_1 < D_2 < \cdots < D_k < \cdots$ such that $\sum_{a=D_k}^{D_{k+1}-1} |\beta_a|^2 > 2$. For every ω for which $Y_d(\omega)$ converges, the random variables $R_k(\omega) := \sum_{a=D_k}^{D_{k+1}-1} \beta_a X_a(\omega)$ must go to 0. However, for all large k, the variance of R_k is at least $3\underline{\sigma}$, contradicting the three series theorem. □

If $Y \notin L^1(\Omega, \mathcal{F}, P)$, then $E(Y|X)$ does not exist for any random vector X. The following example gives a uniformly Lipschitz class of affine $f_d(\cdot)$'s for which no $Y \in L^1(\Omega, \mathcal{F}, P)$ can satisfy $E(Y|(X_1, \ldots, X_d)) = f_d(X_1, \ldots, X_d)$.

Example 4. *If* $|\beta_a| = \frac{1}{\sqrt{a}}$, *then* $\sum_{a \in A} |\beta_a| = \mathcal{O}(\sqrt{\#A})$ *so that the sequence* $f_d = \beta_0 + \sum_{a \le d} \beta_a x_a$ *is uniformly Lipschitz by Lemma* 2. *Since* $\sum_a \beta_a^2$ *diverges, Lemma* 3 *implies that there is no* $Y \in L^1(\Omega, \mathcal{F}, P)$ *having affine conditional expectations* $f_d(x_1, \ldots, x_d) = \beta_0 + \sum_{a \le d} \beta_a x_a$.

If we define $Y_d = \beta_0 + \sum_{a \le d} \beta_a X_a$ in Example 4, then, by the three series theorem, the sequence Y_d diverges. The implication is that the Lipschitz worst case analyses may be based on ever-receding targets, so that, instead of calculating the errors in our attempts to estimate $E(Y|(X_1, \ldots, X_d))$, we may be calculating the errors in an attempt to estimate $E(Y_d|(X_1, \ldots, X_d))$ for an ever receding sequence Y_d.

III.D. **Number of Regressors Intuitions.** The condition $\sum_a |\beta_a| \le B$ for uniformly bounded efficacy (Lemma 2) implies that as the number of regressors grows, the amount by which any further regressors can affect the conditional mean of Y goes to 0. Another model which suggests this involves random parameters, and is also related to Tibshirani's (1996) lasso models. We will suppose that the β_a's are independent random variables with $E|\beta_a| = 1$, scale them as a function of d so that the functions $\beta_0 + \sum_{a \le d} \beta_a x_a$ satisfy Lipschitz or efficacy bounds, and ask the question, "How many of the d regressors can be ignored while still making an error of less than ϵ?"

Satisfying the Lipschitz constraint on average and being requires multiplying the β_a's by something on the order of $1/\sqrt{d}$. By contrast, if we bound the causal efficacy of the explanatory variables, we must multiply the β_a's by something on the order of $1/d$. Let $|\beta|_{(a)}$ be the a'th order statistic of the $|\beta_a|$'s. For given d and $\epsilon > 0$, let $N = N(d, \epsilon)$ be the largest integer satisfying $\frac{1}{\sqrt{d}} \sum_{a \le N} E|\beta|_{(a)} < \epsilon$ and $M = M(d, \epsilon)$ the largest satisfying $\frac{1}{d} \sum_{a \le M} E|\beta|_{(a)} < \epsilon$.

Example 5. *If the* $|\beta_a|$ *are independent exponentials with mean* 1, *then the difference between the order statistics,* $|\beta|_{(a+1)} - |\beta|_{(a)}$, *are independent exponentials with means* $1/(d-a)$ (*e.g. Feller* (1971, I.6, pp. 19-20). *From this,* $N(20, 0.05) = 4$ *while* $M(20, 0.05) = 13$. *On*

average, 4 *of the* 20 *regressors can be ignored if* f *has a Lipschitz constant of* 1*, while* 13 *of* 20 *can be ignored if the monotonic total norm of* f *is* 1*.*

Since the β_a are multiplied by something going to 0 as d increases, it is their tail behavior that determines $N(d,\epsilon)$ and $M(d,\epsilon)$ when d is larger. If the tails of the $|\beta_a|$ are thinner than the exponential tails, e.g. they have Gaussian tails, then even fewer of the regressors matter, both N and M are smaller. For some tail behaviors, the ratios N/d and M/d go to 0 at different rates as $d \uparrow \infty$.

The dimension dependent growth of total efficacy is behind the slower rates of convergence in higher dimensions. Here, varying the distributional assumptions about the regression coefficients shows that this may not be the relevant approximation. One suspects that in many empirical situations, the total efficacy is often small relative to d because relatively few regressors turn out to matter very much. This is behind the success of Tibshirani's (1996) lasso models, and, as part of an extended comparison of parametric and nonparametric methods, Breiman (2001) discusses several general classes of high-dimensional situations in which this kind of ratio result holds.

IV. The Geometry of Dimension Independent Rates

The previous section strongly suggests that the rate of convergence analyses of non-parametric regression should focus on efficacy bounded classes of functions rather than the efficacy unbounded class of Lipschitz functions. The result in this section goes further, and gives a unified, dimension-independent, geometric representation of a class of non-parametric regression estimators that includes, but is not limited to, series expansions (Fourier, wavelet, Tchebyshev and others), kernels and other locally weighted regressions, splines, wavelets, and artificial neural networks. The geometric representation allows one to identify, for each of these regression techniques, classes that function as Barron's efficacy bounded class, $\mathbb{V}_d^{ann}$.

IV.A. **Spaces of Targets.** Let μ denote the distribution of $(Y,(X_1,\ldots,X_d))$ in $\mathbb{R}^{1+d}$ and μ_X the (marginal) distribution of the explanatory variables, $(X_1,\ldots,X_d)$. The target function is $x \mapsto f(x) := E\left(Y\middle|X=x\right)$ from the support of μ_X to $\mathbb{R}$. Throughout, the target is assumed to belong to a space of functions $\mathbb{V} \subset L^p(\mathbb{R}^d,\mu_X)$ for some $p \in [1,\infty)$ endowed with a norm that makes it a separable, infinite dimensional Banach space such as the following.

(1) $\mathbb{V} = L^2(\mathbb{R}^d,\mu_X)$, typically used in Fourier series analysis, wavelets, and other orthogonal series expansions.
(2) $\mathbb{V} = L^p(\mathbb{R}^d,\mu_X)$ spaces, $p \in [1,\infty)$, typically used when higher (or lower) moment assumptions are appropriate.
(3) $\mathbb{V} = C(D)$, the continuous functions on a compact domain $D \subset \mathbb{R}^d$ satisfying $\mu_X(D) = 1$, with norm $\|f\|_\infty := \max_{x\in D}|f(x)|$.
(4) $\mathbb{V} = C^m(D)$, the space of m-times continuously differentiable functions, $m \in \mathbb{N}$, on a compact domain D having a smooth boundary and satisfying $\mu_X(D) = 1$, with norm $\sup_{x\in D}\sum_{|\alpha|\le m}|D^\alpha f(x)|$, typically used when smoothness of the target is an appropriate assumption.[8]

[8]Here, α is a multi-index, $\alpha = (\alpha_1,\ldots,\alpha_d)$, $\alpha_i \in \{0,1,\ldots\}$, and $|\alpha| := \sum_i \alpha_i$.

(5) $\mathbb{V} = S^{m,p}(\mathbb{R}^d, \mu_X)$, $p \in [1, \infty)$, the Sobolev spaces, defined as the completion of the set $C^{m,p}(\mathbb{R}^d, \mu_X)$, the m-times continuously differentiable functions on $\mathbb{R}^d$, with norm $\|f\| = \sum_{|\alpha| \leq m} \left[\int |D^\alpha f(x)|^p \, d\mu_X(x) \right]^{\frac{1}{p}} < \infty$, are typically used when approximation of a function and its derivatives rather than uniform approximation is appropriate.

The sets C_d^{Lip}, C_d^{MTV}, and $C_d^{Lip} \cap C_d^{MTV}$ are dense in all of these spaces. They are also negligible in a sense to be made clear below.

IV.B. **Compactly Generated Two-Way Cones.** An estimator of an $f \in \mathbb{V}$ is a sequence of functions $\widehat{f_n} \in \mathbb{V}$ where each $\widehat{f_n}$ depends on the data $(Y_i(\omega), (X_{1,i}(\omega), \ldots, X_{d,i}(\omega)))_{i=1}^n$. For the nonparametric techniques studied here, the $\widehat{f_n}$ are of the form $\widehat{f_n}(x) = \sum_k \beta_k c_k(x)$ where $\beta_k \in \mathbb{R}$ and $c_k \in \mathbb{V}$. What varies among the estimators are the functions c_k, the number of terms in the summation, and the dependence of both on ω. The geometry that is common to nonparametric regression estimators is that there is a sequence, $C_{\kappa(n)} = C_{\kappa(n)}(\omega) \subset \mathbb{V}$ of **compactly generated two-way cones** with the property that $\widehat{f_n} \in C_{\kappa(n)}$.

$U = \{f \in \mathbb{V} : \|f\| < 1\}$ denotes the unit ball in $\mathbb{V}$, its closure is $\overline{U}$, and $\partial U = \{f \in \mathbb{V} : \|f\| = 1\}$ is its boundary. For $E \subset \mathbb{V}$, $\mathbf{sp}\, E$ is the span of E, that is the set of all *finite* linear combinations of elements of E, and $\overline{\mathbf{sp}}\, E$ is the closure of the span of E.

For $S \subset \mathbb{R}$, $S \cdot E := \{s \cdot f : f \in E,\ s \in S\}$ is the set of scalar multiples of elements of E with scalars belonging to S. It is worth noting that in the following definition, a cone need not be convex, e.g. the non-negative axes in $\mathbb{R}^d$ are a cone, and that a two-way cone may contain linear subspaces.

Definition 3. *A set $F \subset \mathbb{V}$ is a **cone** if $F = \mathbb{R}_+ \cdot F$, that is, if F is closed under multiplication by non-negative scalars. A set $C \subset \mathbb{V}$ is a* **two-way cone** *if $C = \mathbb{R} \cdot C$. A two-way cone is* **compactly generated** *if there exists a compact $E \subset \overline{U}$, $0 \notin E$, such that $C = \mathbb{R} \cdot E$.*

IV.C. **Examples.** We turn to examples of commonly used nonparametric estimators that belong to sequences of compactly generated two-way cones. Series estimators (Fourier series, wavelets, splines, and the various polynomial schemes), as well as broad classes of artificial neural network estimators belong to *nested* sequences of compactly generated two-way cones. Kernel estimators and other locally weighted regression schemes on compact domains belong to a *non-nested* sequence of compactly generated two-way cones. Throughout, it is important to note that the sequence of cones will often depend not only on n, the number of data points, but on ω through the data, $(Y_i(\omega), (X_{1,i}(\omega), \ldots, X_{d,i}(\omega)))_{i=1}^n$.

IV.C.1. *Series estimators.* Fourier series, wavelets, splines, and the various polynomial schemes specify a countable set $E = \{e_k : k \in \mathbb{N}\} \subset \partial U$ with the property that $\overline{\mathbf{sp}}\, E = \mathbb{V}$. Descriptions of the specific e_k for Fourier series, for the various polynomial schemes, and for wavelets are widely available. The estimator based on n data points, $\widehat{f_n}$, is a function of the form

$$\widehat{f_n}(x) = \sum_{k \leq \kappa(n)} \widehat{\beta}_k e_k(x). \tag{9}$$

The dependence on ω arises because the function is chosen to best fit the data. The estimators $\widehat{f_n}$ belong to $C_{\kappa(n)} := \mathbf{sp}\,\{e_1, \ldots, e_{\kappa(n)}\}$. Being a finite dimension subspace of $\mathbb{V}$, each $C_{\kappa(n)}$ is a compactly generated two-way cone, e.g. generated by $\mathbf{sp}\,\{e_1, \ldots, e_{\kappa(n)}\} \cap \partial U$.

Since $\overline{\mathbf{sp}}\, E = \mathbb{V}$, having $\lim_n \kappa(n) = \infty$ guarantees that the $\widehat{f_n}$ can approximate any function. To avoid overfitting and its implied biases, not letting $\kappa(n)$ go to infinity too

quickly, e.g. $\kappa(n)/n \to 0$ guarantees consistency. If $\kappa(n) \to \infty$ is regarded a sequence of parameters to be estimated e.g. by cross-validation, then $\kappa(n)$ also depends on ω, which yields the random sequence $\omega \mapsto C_{\kappa(n)}(\omega)$.

IV.C.2. *Kernel and locally weighted regression estimators.* Kernel estimators for functions on a compact domain typically begin with a function $K : \mathbb{R} \to \mathbb{R}$, supported (i.e. non-zero) only on $[-1, +1]$, having its maximum at 0 and satisfying three integral conditions: $\int_{-1}^{+1} K(u)\, du = 1$, $\int_{-1}^{+1} uK(u)\, du = 0$, and $\int_{-1}^{+1} u^2 K(u)\, du \neq 0$. Univariate kernel regression functions are (often) of the form

$$\widehat{f}_n(x) = \textstyle\sum_{i=1}^n \widehat{\beta}_i g(x\big|X_i, h_n) = \sum_{i=1}^n \widehat{\beta}_i K(\frac{1}{h_n}(x - X_i)). \tag{10}$$

Here $\kappa(n) = n$ and $C_{\kappa(n)}(\omega) = \mathbf{sp}\,\{K(\frac{1}{h_n}(x - X_i(\omega)) : i = 1, \ldots, n\}$.

When the kernel function, $K(\cdot)$, is smooth and its derivatives satisfy $\lim_{|u|\to 1} K^{(\alpha)}(u) = 0$, and the X_i belong to a compact domain, D, the estimator $\widehat{f}_n$ belongs to $C^m(D)$ for any m, and the $C^m(D)$-norm or one of the S^p_m-norms might be used. If the kernel function, $K(\cdot)$, function is continuous but not smooth, the $\widehat{f}_n$ belong to $C_b(\mathbb{R})$, hence to $L^p(\mathbb{R}, \mu_X)$. For any compact $D \subset \mathbb{R}$, the restrictions of the $\widehat{f}_n$ to D belong to $C(D)$.

In all of these cases, the n-data points, X_i, $i = 1, \ldots, n$, and the window-size parameter h_n, define n non-zero functions, $g(\cdot\big|\theta_{i,n})$, $\theta_{i,n} = (X_i, h_n)$. The estimator, $\widehat{f}_n$, belongs to the span of these n functions. As established above, the span of a finite set of non-zero functions is a compactly generated two-way cone.

The considerations for choosing the window-sizes, h_n, parallel those for choosing the $\kappa(n)$ in the series expansions. They can be chosen, either deterministically or by cross-validation, so that $h_n \to 0$, to guarantee that the kernel estimators can approximate any function, but not too quickly, so as to avoid overfitting.

The considerations for multivariate kernel regression functions are almost entirely analogous. These estimators are often of the form

$$\widehat{f}_n(x) = \textstyle\sum_{i=1}^n \widehat{\beta}_i g(x\big|X_i, h_n) = \sum_{i=1}^n \widehat{\beta}_i K(\frac{1}{h_n}\|x - X_i\|) \tag{11}$$

where $h_n \downarrow 0$ and the X_i are points in the compact domain $D \subset \mathbb{R}^d$.

Locally weighted linear/polynomial regressions have different $g_i(\cdot\big|\theta_{i,n})$, see e.g. Stone (1982). In all of these cases, when the domain is compact, so are the sets of possible parameters for the functions g_i, and the mapping from parameters to functions is continuous. This again implies that the $\widehat{f}_n$ belong to the span of a finite (hence compact) set not containing 0.

IV.C.3. *Artificial neural networks.* Single hidden layer feedforward (slff) estimators with activation function $g : \mathbb{R} \to \mathbb{R}$ often take $E \subset \mathbb{V}$ as $E = \{x \mapsto g(\gamma' \tilde{x}) : \gamma \in \Gamma\}$. Here $x \in \mathbb{R}^d$, $\tilde{x}' := (1, x')' \in \mathbb{R}^{d+1}$, and Γ is a compact subset of $\mathbb{R}^{d+1}$ with non-empty interior. The slff estimators are functions of the form

$$\widehat{f}_n(x) = \textstyle\sum_{k \leq \kappa(n)} \widehat{\beta}_k g(\widehat{\gamma}_k' \tilde{x}), \tag{12}$$

where the $\widehat{\gamma}_k$ belongs to Γ. Specifically, $C_{\kappa(n)} = \{\sum_{k \leq \kappa(n)} \beta_k c_k : c_k \in E\}$ is the compactly generated two-way cone of slff estimators.

If $\kappa(n) \to \infty$, $\kappa(n)/n \to 0$, and $\overline{\mathbf{sp}}\, E = \mathbb{V}$, then the total error goes to 0. Various sufficient conditions on g that guarantee $\overline{\mathbf{sp}}\, E = \mathbb{V}$ with compact Γ in the Banach spaces

listed above are given in Hornik *et. al.* (1989, 1990), Stinchcombe and White (1990, 1998), Hornik (1993), Stinchcombe (1999). Also as above, $\kappa(n)$ may be regarded as a parameter, estimated by cross-validation.

When g is continuous and Γ is compact, then E is a compact subset of $C(D)$ for any compact $D \subset \mathbb{R}^d$. When g is bounded, as is essentially always assumed, E is a compact subset of $L^p(\mathbb{R}^d, \mu_X)$ for any $p \in [1, \infty)$. When g is bounded and measurable, as in the case of the frequently used 'hard limiter,' $g(x) = 1_{[0,\infty)}(x)$, and μ_X has a density with respect to Lebesgue measure, E is a compact subset of $L^p(\mathbb{R}^d, \mu_X)$, $p \in [1, \infty)$. When g is smooth, e.g. the ubiquitous logistic case of $g(x) = e^x/(1 + e^x)$, and Γ compact, then E is a compact subset of $C^m(D)$, and of $S^p_m(\mathbb{R}^d, \mu_X)$ for any m and any $p \in [1, \infty)$.

Aside from notational complexity, essentially the same analysis shows that multiple hidden layer feedforward networks output functions are also expressible as the elements of the span of a compact set E.[9]

Radial basis network estimators most often take E_n to be a set of the form $E_n = \{x \mapsto g(\frac{1}{\lambda_n}(x - \gamma)'\Sigma(x - \gamma)) : \gamma \in \Gamma, \lambda_n \geq \underline{\lambda}_n\}$, Γ a compact subset of $\mathbb{R}^d$ containing the domain, Σ a fixed positive definite matrix, $\underline{\lambda}_n \downarrow 0$ but not too quickly, g a continuous function. The estimators are functions of the form

$$\widehat{f}_n(x) = \textstyle\sum_{k \leq \kappa(n)} \widehat{\beta}_k g(\widehat{\gamma}_k' \tilde{x}), \tag{13}$$

The continuity of g implies that the E_n have compact closure. For the common choices of g in the literature, $g(0) \neq 0$ so that $0 \notin E_n$.

IV.D. **Rates and Consistency.** In the examples just given, the sequence of compactly generated two-way cones become dense, and may be either deterministic or random. The cones becoming dense is **consistency**.

Definition 4. *A random sequence of compactly generated two-way cones, $\omega \mapsto C_{\kappa(n)}(\omega)$, is **consistent** if for all $g \in \mathbb{V}$, $P(\cup_N \cap_{n \geq N} [d(g, C_{\kappa(n)}(\cdot)) < \epsilon]) = 1$.*

IV.E. **Results.** For any sequence of sets, B_n, $[B_n \text{ i.o.}] := \bigcap_m \bigcup_{n \geq m} B_n$ is read as "B_n infinitely often," while $[B_n \text{ a.a.}] := \bigcup_m \bigcap_{n \geq m} B_n$ is read as "B_n almost always." For a compactly generated two-way cone, C, of estimators, and $r > 0$, the set $C + r \cdot U$ is the set of all targets that are within r of set of estimators contained in C. Consistency can be rewritten as "for all $\epsilon > 0$, $P([C_{\kappa(n)} + \epsilon \cdot U \text{ a.a.}] = \mathbb{V}) = 1$." Of particular interest will be sets of the form $[C_{\kappa(n)} + r_n \cdot U \text{ a.a.}]$ where $r_n \to 0$ and $C_{\kappa(n)}$ is a sequence of compactly generated two-way cones.

This section proves Lemmas 4 and 5, which yield the following.

Theorem A. *For any consistent nonparametric regression technique with estimators belonging to a sequence, $C_{\kappa(n)}$, of compactly generated two-way cones, and for any $r_n \to 0$, a dense, shy set of targets can be approximated at the rate $\mathcal{O}(r_n)$.*

"Shyness" is defined below, and provides useful information about the sets of targets. Within the Banach spaces of functions listed above (and many others), C_d^{Lip} and C_d^{MTV} form dense, shy sets of functions, as does their intersection.

[9]Consistency issues for multiple layer feedforward networks are addressed in the approximation theorems of Hornik, Stinchcombe, and White (1989, 1990), and Hornik (1993).

For $M \in \mathbb{N}$, define $A_n^M := C_{\kappa(n)} + Mr_n \cdot U$. Fix a sequence of sets of estimators $C_{\kappa(n)}$. For $g \in \mathbb{V}$, there exists a subsequence, n', such that $d(g, C_{n'}) = \mathcal{O}(r_{n'})$ if and only if $g \in [A_n^M \text{ i.o.}]$ for some $M \in \mathbb{N}$. If we do not allow subsequences, we have $d(g, C_n) = \mathcal{O}(r_n)$ if and only if $g \in [A_n^M \text{ a.a.}]$ for some M.

Definition 5. *The set of* $\mathcal{O}(r_n)$***-accumulatable targets*** *is* $\cup_M [A_n^M \text{ i.o.}]$*, and the set of* $\mathcal{O}(r_n)$***-approximable targets*** *is* $\mathcal{T}(r_n) := \cup_M [A_n^M \text{ a.a.}]$.

Lemma 4. $P(\mathcal{T}(r_n) \text{ is dense}) = 1$ *if and only if the* $C_{\kappa(n)}$ *are consistent.*

Proof. Suppose that $C_{\kappa(n)}$ is consistent. Let $\mathcal{G} = \{g_j : j \in \mathbb{N}\}$ be a dense subset of $\mathbb{V}$. Define $B_j^m = \cup_N \cap_{n \geq N} [(g_j + \frac{1}{m} \cdot U) \cap (C_{\kappa(n)}(\omega) + r_n \cdot U) \neq \emptyset]$. Since the $C_{\kappa(n)}$ are consistent, $P(B_j^m) = 1$. Therefore, $P(\cap_{m,j} B_j^m) = 1$. Finally, the event that $d(g_j, \mathcal{T}(r_n)) < 1/m$ for every m contains $\cap_{m,j} B_j^m$.

Suppose now that $C_{\kappa(n)}$ is not consistent, i.e. there exists $g \in \mathbb{V}$ and $\epsilon > 0$ such that $P([d(g, C_{\kappa(n)}) < \epsilon \text{ a.a.}]) < 1$, equivalently, $P([d(g, C_{\kappa(n)}) \geq \epsilon \text{ i.o.}]) > 0$. For all M, $Mr_n < \epsilon$ for all but finitely many n. Therefore, $P(\mathcal{T}(r_n) \cap (g + \epsilon \cdot U) = \emptyset) > 0$. That is, the probability that $\mathcal{T}(r_n)$ is dense is less than 1. □

The Lipschitz functions and the functions with bounded efficacy satisfy the following notion of a negligible subset of an infinite dimensional space.[10]

Definition 6. *A subset* S *of a universally measurable* $S' \subset \mathbb{V}$ *is* ***shy*** *or* ***Haar null*** *if there exists a compactly supported probability* η *such that* $\eta(S' + g) = 0$ *for all* $g \in \mathbb{V}$.

Lemma 5. *If* r_n *goes to* 0 *more slowly than* r'_n*, then* $\mathcal{T}(r_n) \setminus \mathcal{T}(r'_n)$ *is shy.*

For ease of later reference, we separately record the following easy observation.

Lemma 6. *If* C *is a compactly generated two-way cone, then it is closed, has empty interior, and* $C \cap F$ *is compact for every closed, norm bounded* F.

Proof of Lemma 5: It is sufficient to show that the set of $\mathcal{O}(r_n)$-accumulatable targets is shy because $\mathcal{T}(r_n) = \cup_M [A_n^M \text{ a.a.}] \subset \cup_M [A_n^M \text{ i.o.}]$, and any subset of a shy set is shy.

A set $F \subset \mathbb{V}$ is approximately flat if for every $\epsilon > 0$, there is a finite dimensional subspace W of $\mathbb{V}$ such that $F \subset W + \epsilon \cdot U$. Every compact set is approximately flat — let F_ϵ be a finite ϵ-net and take $W = \mathbf{sp}\, F_\epsilon$. From Stinchcombe (2001, Lemma 1), for any sequence F_n of approximately flat sets, $[(F_n + r_n \cdot U) \text{ i.o.}]$ is shy. Since the countable union of shy sets is shy, $\cup_M [(F_n + Mr_n \cdot U) \text{ i.o.}]$ is shy.

Fix arbitrary $R > 0$. It is sufficient to prove that $(R \cdot U) \cap [A_n^M \text{ i.o.}]$ is shy. Fix arbitrary $\eta > 0$. $R \cdot U$ is a subset of the closed, norm bounded set $R \cdot (1 + \eta)\overline{U}$. By Lemma 6, the set $F_n = C_n \cap (R \cdot (1 + \eta)\overline{U})$ is compact. Since compact sets are approximately flat, $S = [(F_n + Mr_n \cdot U) \text{ i.o.}]$ is shy. Since $r_n \to 0$ and $\eta > 0$, $[(R \cdot U) \cap [A_n^M \text{ i.o.}]] \subset S$. □

Proof of Theorem A: Lemma 4 shows that consistency of the non-parametric regression technique with estimators given by a sequence $C_{\kappa(n)}$ of compactly generated two-way cones and denseness of $\mathcal{T}(r_n)$ are equivalent. Lemma 5 shows that $\mathcal{T}(r_n)$ is shy. □

[10] There are several related notions of negligible sets in infinite dimensional spaces, detailed in Benyamini and Lindenstrauss (2000, Ch. 6). Anderson and Zame (2001) cover some of the uses of shy (Haar null) sets in economic theory, and greatly extend the applicability of the notion.

V. Conclusions and Complements

Most of the analyses of the rates of convergence for nonparametric regression arrive at dismal results with even a moderate number of regressors. The key assumption driving these results is that the target function, $f(x_1, \ldots, x_d) = E\left(Y|(X_1, \ldots, X_d) = (x_1, \ldots, x_d)\right)$ is uniformly Lipschitz. This assumption can never be rejected by data. Replacing the Lipschitz functions by sets of functions sharing this unrejectability shows that the order of the rate of convergence is given by the order of the estimation error, that dimension-dependent approximation error need play no role.

Examples suggest that dimension dependence of the complexity of a regression function is more tightly tied to its monotonic total variation than to any measure of its smoothness. These examples also demonstrate that how the variation depends on the dimensionality may vary from one set of problems or distribution over problems to another. Experience suggests that the variation, both in linear and non-linear regression, is often small.

Together, the results and examples suggest that rates of convergence calculated using Lipschitz functions are mis-leading, that what matters is some measure of variability. This puts correspondingly more weight on the criteria of interpretability and generalization for the judging competing nonparametric approaches.

There are a number of subsidiary points to be made.

V.A. **Comparison and Estimation of Dense Sets.** As well as comparing $\mathcal{T}(r_n)$ and $\mathcal{T}(r'_n)$ for the same nonparametric regression technique, one can also compare these sets across regression techniques. For example, Barron (1993) fixes a pair of rates, r_n and r'_n with $r'_n = \boldsymbol{o}(r_n)$, and shows that for the ann techniques that he considers, $\mathcal{T}_{ann}(r'_n)$ cannot be approximated by any series expansion at a rate r_n. Reversing his example in L^2 requires a permutation of the basis elements, and gives rise to a set $\mathcal{T}_{series}(r'_n)$ that cannot be approximated by any variant of his ann technique at a rate r_n.

This seems to be part of a more general pattern. Pick a pair of sequences r_n, r'_n with $r'_n = \boldsymbol{o}(r_n)$. From Lemma 4, $\mathcal{T}(r_n) \setminus \mathcal{T}(r'_n)$ and $\mathcal{T}(r'_n)$ are disjoint, dense sets of nonparametric targets. We conjecture that for generic pairs of sequences, $C_{1,\kappa(n)}$ and $C_{2,\kappa'(n)}$, of compactly generated two-way cones, $\mathcal{T}_1(r'_n) \setminus \mathcal{T}_2(r_n) \neq \emptyset$ and $\mathcal{T}_2(r'_n) \setminus \mathcal{T}_1(r_n) \neq \emptyset$.

As has been noted, with smooth classical measurement error (or with errors in variables), it is not possible to reject (say) $H_{Eff} : f_d$ is uniformly efficacy bounded in favor of the larger alternative hypothesis, $H_{Lip} : f_d$ is uniformly Lipschitz bounded. If one could estimate Lipschitz or efficacy norms, then in principle one could test the alternative hypotheses against each other, but this estimation problem seems extraordinarily difficult.

V.B. **Comparisons Across Rates.** If r_n and r'_n both go to 0 but r_n goes more slowly, then the dense class $\mathcal{T}(r_n)$ is larger than the dense class $\mathcal{T}(r'_n)$. Lemma 5 shows that the difference between the sets, $\mathcal{T}(r_n) \setminus \mathcal{T}(r'_n)$, is shy. Shy subsets are an infinite dimensional extension of the finite dimensional Lebesgue null set notion non-genericity. This gives partial information about the size of the difference between the two sets. It is only partial information because the proof simply shows that the larger of the two sets is Haar null, and any subset of a null set is a null set. Two points:

(1) Much to be desired is an improvement on this partial result. Something that would, despite the impossibility of data ever distinguishing between the dense sets, allow one

to distinguish, at least theoretically, more finely between sets of targets $\mathcal{T}(r_n)$ and $\mathcal{T}(r'_n)$. However, Lemma 5 shows that trying to resurrect the curse of dimensionality in rates of convergence requires one to say that one non-generic dense set of functions is clearly preferable to another non-generic dense set of functions, and that it's preferable because it yields worse results.

(2) For finite dimensional parametric estimation, superefficiency can happen on Lebesgue null sets (e.g. Lehmann and Casella (1983, Ch. 6.2)). For infinite dimensional non-parametric estimation, Brown, Low, and Zhao (1997) show that it can happen "everywhere," that is, at all points in the dense sets of targets $\mathcal{T}(r_n)$ that are typically used. It seems that behind this result is the same approximately-flat-but-not-flat infinite dimensional geometry that yields the denseness of the $\mathcal{T}(r_n)$ classes.[11]

V.C. **Smoothness.** Another aspect of the work on the curse rates of approximation is that smoother targets lead to faster approximation. For example, if the target f is assumed to have s continuous derivatives, and these derivatives are Lipschitz, then Stone's rate of approximation is increased to $\mathcal{O}_P(n^{-1/(2+[d/s])})$. The dense classes, $\mathbb{V}_{ann}$, in the dimension independent ann rate of approximation work are defined by an integrability condition on various transforms of the gradient of the target. Niyogi and Girosi (1999) note that this suggests that $s = s(d)$ in such a fashion that $[d/s]$ stays small for the $\mathbb{V}_{ann}$ and d increases.

One might guess that something similar is at work in the classes $\mathcal{T}(r_n)$ that are analyzed here. However, this kind of smoothness argument is problematic for three separate kinds of reasons. First, for many classes of ann's, the dense set of targets are not only infinitely smooth, they are analytic. It is hard to see how smoothness could vary with dimension in this context. Second, for many other classes of ann's, the dense set of targets contain discontinuous functions, and smoothness cannot enter. Finally, the work here provides a plethora of dense classes for which the dimensionality of the regressors plays no role, and it seems unlikely that there is some special smoothness structure common to the different dense sets that work for the different techniques.

V.D. **More on Negligible Sets.** By definition, S is shy if and only if $\eta(S+g)$ is shy for all g and some compactly supported probability η. If $\mathbb{V} = \mathbb{R}^k$, the finite dimensional case here ruled out by assumption, one can take η to be the uniform distribution on $[0,1]^k$ and show that S is shy if and only if it is a Lebesgue null set if and only if for every non-degenerate Gaussian distribution ν, $\nu(S) = 0$. Stinchcombe (2001) showed that there is no similar comfortable Bayesian interpretation of shy sets in the infinite dimensional contexts studied here.

Other relevant properties of this class of shy sets are:

(1) shy sets have no interior so that prevalent sets are dense;
(2) the countable union of shy sets is shy, equivalently the countable intersection of prevalent sets is prevalent; and
(3) if $\mathbb{V}$ is infinite dimensional if and only if compact sets are shy.

Lemmas 5 and Theorem A used shy sets. These results would not hold if we replaced shy sets with the original, more restrictive, class of infinite dimensional null sets due to Aronszajn (1976). These are now called **Gauss null** sets because Aronszajn's definition is

[11] I am grateful to Xiaohong Chen and Jinyong Hahn for these last two points.

now known to be equivalent to the following: S is Gauss null if and only if for every non-degenerate Gaussian distribution, ν, on $\mathbb{V}$, $\nu(S) = 0$ (see Benyamini and Lindenstrauss, 2000, Ch. 6). Every Gauss null set is shy, but the reverse is not true. It can be shown that the sets $\cup_n C_{\kappa(n)}$ of estimators are Gauss null, but not that $[C_{\kappa(n)} + r_n \cdot U \text{ a.a.}]$ is not.

V.E. **Possible Extensions and Generalizations.** There are several additional points to be made.

1. One can think of the analysis of affine conditional means with independent regressors of §III as a very special class of parametrized models. Suppose, more generally, that C_κ is smoothly parametrized by a κ-dimensional vector with κ fixed. Standard results imply that $\|\widehat{f}_n - f^*_\kappa\| = \epsilon_{\kappa,n} = \mathcal{O}(n^{-1/2})$. If instead of being fixed, we let κ depend on d and on n. If $\kappa(d,n) \uparrow \infty$, as required for consistency, but $\kappa(d,n)$ grows very slowly, then the $n^{-1/2}$ rate of approximation slows as little as one desires.
2. If the data is not iid but has some time series structure, one expects that the estimation error in (5) will not be $\mathcal{O}(n^{-1/2})$ for fixed κ, but something slower. Again, since Lemmas 4 and 5 concern approximation error, total error for the nonparametric regressions covered here would also go to 0 at this ineluctably slower rate if we were outside of the iid case.
3. It is hard to imagine nonparametric techniques with estimators that do not belong to a sequence of compactly generated two-way cones. For example, in the above discussion of the locally weighted regression schemes and the artificial neural network estimators, we made use of compact domain assumptions to ease the exposition, and this led to the compactly generated conclusion. However, since the distribution of the data is tight, one can replace the compact domains with a sequence of compact domains having the property that with probability 1, the estimators belong to the associated sequence of compactly generated two-way cones.
4. The proof of Lemma 4 can be easily adapted to show that consistency is equivalent to $\mathcal{T}(r_n)$ containing a dense linear subspace of $\mathbb{V}$ with probability 1. Cohen *et. al.* (2001) characterize some of these dense linear subspaces for wavelet expansions.
5. All of the above has been phrased as regression analysis of conditional means. Since Lemmas 4 and 5 concern the approximation error, one could also, with essentially no changes, consider, e.g., conditional quantile regression and/or loss functions other than mean squared loss. At whatever rate the estimation error goes to 0, there is a dense class of nonparametric targets with the approximation error going to 0 at the same rate.
6. The use of Banach spaces for the set of targets is not crucial. The compactly generated assumption must be slightly modified in locally convex, complete, separable, metric vector spaces, but the main result driving the shyness proofs is Stinchcombe (2001, Lemma 1), which applies in such spaces. For example, one could take $\mathbb{V} = C(\mathbb{R}^d)$ with the topology of uniform convergence on compact sets, or any other of the other Frechet spaces that appear in non-parametric regression analyses.
7. It is a reasonable conjecture that the same results hold for density estimation as hold for regression analysis. Following Davidson and McKinnon (1987), the target densities can be modeled as points in a convex subset of the positive orthant in a Hilbert space. Lemma 4 should go through fairly easily, but Lemma 5 may be more difficult. The shyness argument requires extending Stinchcombe (2001, Lemma 1) to Anderson and Zame's (2001) relatively shy sets.

8. It can be shown that if C is a compactly generated two-way cone, then the open set $C+U$ is not dense in $\mathbb{V}$. The role of the compact set E not containing 0 in the definition of compactly generated cones can be seen in the following, which should be compared to Lemma 6.

Example 6. *If x_n is a countable dense subset of ∂U and E is the closure of $\{x_n/n : n \in \mathbb{N}\}$, then E is a compact subset of the closed, norm bounded set $\overline{U}$. However, the two-way cone $\mathbb{R} \cdot E$ is not compactly generated, not closed, and is dense, so that $\mathbb{R} \cdot E + \epsilon \cdot U = \mathbb{V}$ for any $\epsilon > 0$.*

V.F. **Studying the Difficulty of Nonparametric Problems.** Finally, some rather preliminary simulation data suggest that it is possible to characterize the difficulty of nonparametric problems by studying when the root-n consistency "kicks in." More specifically, let d be the number of regressors and let $n(d)$ be the number of data points beyond which the root-n asymptotics provide a reasonable guide. The higher is the function $d \mapsto n(d)$, the more difficult the problem. Being uniformly higher for many different nonparametric techniques constitutes a strong indication that the problem is considerably more difficult. We leave this for future research.

References

Adams, C. and J. A. Clarkson (1933). On Definitions of Bounded Variation for Functions of Two Variables. *Transactions of the American Mathematical Society* **35**(4), 824-854.

Anderson, R. and W. Zame (2001). Genericity with Infinitely Many Parameters. *Advances in Theoretical Economics*: Vol. 1: No. 1, Article 1.

Aronszajn, N. (1976). Differentiability of Lipschitzian Mappings Between Banach Spaces. *Studia Mathematica* **LVII**, 147-190.

Barron, A. (1993). Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Transactions on Information Theory* **39**(3), 930-945.

Benyamini, Y. and J. Lindenstrauss (2000). *Geometric Nonlinear Functional Analysis.* Providence, R.I.: American Mathematical Society, Colloquium publications (American Mathematical Society) v. 48.

Billingsley, P. (2008). *Probability and Measure.* John Wiley and Sons, New York.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science* **16**(3), 199-231.

Brown, L. D., M. G. Low, and L. H. Zhao (1997). Superefficiency in Nonparametric Function Estimation. *Annals of Statistics* **25**(6), 2607-2625.

Chen, X. (2007). Large Sample Sieve Estimation of Nonparametric Models. *Handbook of Econometrics* **6**, 5549-5632.

Chen, X. and H. While (1999). Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators. *IEEE Tran. Information Theory* **45**, 682-691.

Cohen, A., R. DeVore, and G. Kerkyacharian (2001). Maximal Spaces with Given Rate of Convergence for Thresholding Algorithms. *Applied and Computational Harmonic Analysis* **11**, 167-191.

Davidson, R. and J. G. McKinnon (1987). Implicit Alternatives and the Local Power of Test Statistics. *Econometrica* **55**(6), 1305-1329.

Dellacherie, C. and P.-A. Meyer (1978). *Probabilities and Potential*, vol. 29 of North-Holland Mathematics Studies. North-Holland Publishing Co., Amsterdam.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications, v. II*. John Wiley and Sons, New York.

Hornik, K. (1993). Some New Results on Neural Network Approximation. *Neural Networks* **6**(8), 1069-1072.

Hornik, K, M. Stinchcombe and H. White (1989). Multi-layer Feedforward Networks are Universal Approximators. *Neural Networks* **2**, 359-366.

Hornik, K, M. Stinchcombe and H. White (1990). Universal Approximation of an Unknown Mapping and its Derivatives using Multilayer Feedforward Networks. *Neural Networks* **3**, 551-560.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation*, 2nd ed. Springer-Verlag, New York.

Mhaskar, H. N. and C. A. Michelli (1995). Degree of Approximation by Neural and Translation Networks with a Single Hidden Layer. *Advances in Applied Mathematics* **16**, 151-183.

Newey, W. (1996). Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics* **79**, 147-168.

Niyogi, P. and F. Girosi (1999). Generalization bounds for function approximation from scattered noisy data. *Advances in Computational Mathematics* **10**, 51-80.

Stinchcombe, M. (1999). Neural Network Approximation of Continuous Functionals and Continuous Functions on Compactifications. *Neural Networks* **12**, 467-477.

Stinchcombe, M. (2001). The Gap Between Probability and Prevalence: Loneliness in Vector Spaces. *Proceedings of the American Mathematical Society* **129**, 451-457.

Stinchcombe, M. and H. White (1990). Approximating and Learning Unknown Mappings Using Multilayer Feedforward Networks with Bounded Weights. *Proceedings of the International Joint Conference on Neural Networks*, Washington, D. C., **III**, 7-16. San Diego, CA.: SOS Printing.

Stinchcombe, M. and H. White (1998). Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative. *Econometric Theory* **14**, 295-325.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* **10**, 1040-1053.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**(1), 267-288.

Yukich, J., M. Stinchcombe, and H. White (1995). Sup-Norm Approximation Bounds for Networks through Probabilistic Methods. *IEEE Transactions on Information Theory* **41**(4), 1021-1027.

*Dept. of Economics, U.T. Austin, 1 University Station - C3100, Austin, TX 78712, ph: 1-512-475-8515, e-mail max.stinchcombe@gmail.com. †StataCorp, 4905 Lakeway Drive, College Station, TX 77845, ph: 1-979-696-4600, e-mail ddrukker@stata.com.