

CAUSAL EFFICACY AND THE CURSE OF DIMENSIONALITY[‡]

MAXWELL B. STINCHCOMBE

ABSTRACT. There is a unified, dimension-independent, 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, wavelets, and artificial neural networks. For these estimators, the rate at which the distance between the estimator and the target goes to 0 is given by the estimation error and is independent of dimension. This dimension-independent result does not contradict the class of results that go under the name of the “curse of dimensionality.” Those results are based on an implicit assumption that the target function is an ever receding target, that the amount of variability of the target increases without bound as the number of regressors grows.

1. INTRODUCTION

Interest centers on estimating the target function, $f(x) := E(Y|X = x)$ where $Y \in \mathbb{R}^1$, $X \in \mathbb{R}^d$, and (Y, X) is a random vector with distribution μ . Regression analyses form an estimate, $\hat{f}_n(\cdot)$ based on n data points $(Y_i, X_i)_{i=1}^n$. One assumes that as n grows, the empirical distribution of the data, $\hat{\mu}_n$, converges to μ .

A non-parametric estimator, \hat{f}_n , solves

$$\min_{g \in \Theta_{\kappa(n)}} \left[\int (y - g(x))^2 d\hat{\mu}_n(y, x) \right]^{1/2}. \quad (1.1)$$

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

Date: February 28, 2007.

[‡]I owe many thanks to Xiaohong Chen, Jinyong Hahn, Qi Li, Dan Slesnick, Hal White, and Paul Wilson for numerous helpful conversations and corrections. Hopefully, there aren't too many errors left in the paper to blame them for.

By contrast with (1.1), with μ perfectly known, one solves

$$\min_{g \in \Theta_{\kappa(n)}} \left[\int (y - g(x))^2 d\mu(y, x) \right]^{1/2}. \quad (1.2)$$

Denote the solution to (1.2) by $f_{\kappa(n)}^*$. 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 ,

$$\epsilon_n + a_n := \underbrace{\|\widehat{f}_n - f_{\kappa(n)}^*\|}_{\text{estimation error}} + \underbrace{\|f_{\kappa(n)}^* - f\|}_{\text{approx. error}} \geq \|\widehat{f}_n - f\|. \quad (1.3)$$

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, $\mathbb{V}' \subset \mathbb{V}$, of targets. The set \mathbb{V}' 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 $\epsilon_n(\kappa) + a_n(\kappa)$.

This paper 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. Theorem A shows that for any estimator having this 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 $\mathbb{V}' = \mathbb{V}'(r_n, \kappa(n)) \subset \mathbb{V}$ for which $a_n = \mathcal{O}(r_n)$.

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 (often 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 $\|\widehat{f}_n - f\| \leq e_n + a_n = \mathcal{O}(e_n)$.

This dimension-independent result does not contradict the class of results that go under the name of the ‘‘curse of dimensionality.’’ They are often taken to imply that the rate at which $\|\widehat{f}_n - f\|$ goes to 0 is very slow when d is even moderately large, say, 3 or more, and to imply that the data requirements are wildly impractical when d is, say, 7 or more. The difference is that the curse results invoke a different dense set \mathbb{V}' , the Lipschitz functions, in their calculation of the rates of approximation.

One can see the outlines of the resolution to the apparent contradiction in the construction and use of dense sets, \mathbb{V}_{ann} , of targets in the proofs of the results giving fast, dimension-independent rates of convergence for a number of artificial neural network regression techniques.¹ Essentially, the geometric approach used in this paper gives a general construction of dense sets for most of the non-parametric regression techniques presently in use.

The next section shows how the curse results depend on the implicit assumption that the target function, f , is an ever receding target. More specifically, the curse results assume that as d increases, there is an unbounded increase in the causal efficacy of the explanatory variables.² The following section gives and proves the dimension independent result. The geometric constructs may be unfamiliar, and §4 shows that the various regression techniques mentioned above are special cases. The last two sections contain complements and conclusions.

2. EVER RECEDING TARGETS

Probably the most influential papers on the curse of dimensionality for non-linear regression are due to Stone (1980, 1982). He used Lipschitz functions in his calculations of “optimal rates.” Causal efficacy is the amount that changes in the values of explanatory variables, (X_1, \dots, X_d) , can shift the conditional expectation of Y . Stone’s use of Lipschitz functions allows causal efficacy to be unbounded in d . The implications and incidence of bounded and unbounded causal efficacy are most clearly seen in the special case linear regression.

2.a. The Lipschitz Assumption and Dimension Dependence. Stone (1980, 1982) defined an “optimal” rate of convergence, and showed that it is $r_n = n^{-1/(2+d)}$ when the data $(Y_i, X_i)_{i=1}^n$ is iid. By optimal, Stone meant that if the target function, f , is assumed only to have a Lipschitz constant, then for any regression technique, any sequence of estimators, \hat{f}_n , based on n iid data points, satisfies

$$\|\hat{f}_n - f\| \geq \mathcal{O}_P(n^{-1/(2+d)}), \quad (2.1)$$

¹Barron (1993) found this for single-layer feedforward artificial neural networks, 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 (2006) is a survey.

²From the *Oxford English Dictionary*, efficacy is the “Power or capacity to produce effects.”

and that some sequence satisfies (2.1) with equality.

The Lipschitz assumption seems unobjectionable, but it is where dimensionality enters the analysis in an exponential fashion. It seems unobjectionable because the Lipschitz functions, \mathbb{V}_{Lip} are dense in the set of targets all possible targets, \mathbb{V} . Denseness means that data can never reject $H_0: f \in \mathbb{V}_{Lip}$ in favor of $H_A: f \notin \mathbb{V}_{Lip}$, and this ought to mean that there is no harm in the assumption. However, if one assumes that the target has a Lipschitz constant that holds for all d , then the approximation error depends exponentially on d .

An extremely clear example of how this works can be taken from Newey (1997). He shows that, if μ satisfies some easy-to-verify and quite general conditions, the data is iid, and the target, f , satisfies the uniform approximation condition

$$\sup_x \inf_{g \in \Theta_\kappa} |f(x) - g(x)| = \mathcal{O}\left(\frac{1}{\kappa^\alpha}\right), \quad (2.2)$$

then

$$\|f - \hat{f}_n\|^2 = \mathcal{O}_P\left(\frac{\kappa}{n} + \frac{1}{\kappa^{2\alpha}}\right). \quad (2.3)$$

Ignoring some of the finer detail, the κ/n term in Newey's proof corresponds to the square of the estimation error, and the $\kappa^{-2\alpha}$ to the square of the approximation error.³ To balance the tradeoffs, one picks $\kappa = \kappa(n)$ to minimize $\frac{\kappa}{n} + \frac{1}{\kappa^{2\alpha}}$.

Suppose that $f: [-1, +1]^d \rightarrow \mathbb{R}$ belongs to \mathbb{V}_{Lip} , the set of Lipschitz functionse. In terms of the variability of f , the worst case has $|f(x) - f(x')| = B \cdot e(x, x')$ for most pairs $x, x' \in [-1, +1]^d$ (where $e(\cdot, \cdot)$ is the Euclidean metric and B is the Lipschitz constant of the target f). In this worst case, if we evaluate f at roughly $\left(\frac{2B}{\epsilon}\right)^d$ (carefully chosen) points, we know f to within ϵ at all points in its domain. For many classes Θ_κ this yields, for every $f \in \mathbb{V}_{Lip}$, $\sup_x \inf_{g \in \Theta_\kappa} |f(x) - g(x)| = \mathcal{O}\left(\frac{1}{\kappa^\alpha}\right)$ with $\alpha = \frac{1}{d}$. Substituting this in (2.2) means that finding the best tradeoff between estimation and approximation error reduces to minimizing $\frac{\kappa}{n} + \frac{1}{\kappa^{2/d}}$. Solving yields $\kappa = n^{\frac{d}{2+d}}$, evaluating the minimand at the solution gives

$$\|f - \hat{f}_n\|^2 = \mathcal{O}_P\left(\frac{\kappa}{n} + \frac{1}{\kappa^{2/d}}\right) = \mathcal{O}_P\left(n^{-\frac{2}{2+d}}\right), \text{ or } \|f - \hat{f}_n\| = \mathcal{O}_P\left(n^{-\frac{1}{2+d}}\right). \quad (2.4)$$

It is at this point that one observes the ‘‘curse of dimensionality.’’ If one wishes $n^{-\frac{1}{2+d}} < 0.10$ with (say) seven regressors, one sets $n^{1/(2+7)} = 10$ and find that one needs on the order of $n = 10^9$ independent observations, a large number.

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

2.b. Lipschitz Worst Cases Are Impossibly Bad. Notice that if we are not in the worst case, that is, if function f is less variable over much of its domain, then we need to evaluate it at fewer points in order to pin it down. This in turn means that fewer observations are needed to achieve any given degree of accuracy. When Y , the left-hand side in a regression, is a random variable, it sometimes turns out that the variability of the conditional expectation function has a bound that is much sharper than the Lipschitz bound.

For expository simplicity, suppose that the possible explanatory variables, (X_1, X_2, \dots) take values in $[-1, +1]$, have mean 0, and that they are not degenerate in the limit, that is, $\liminf_d \text{Var}(X_d) = \underline{\sigma} > 0$.

If $g : [-1, +1]^d \rightarrow \mathbb{R}$ has Lipschitz constant B , then $\max_{x, x' \in [-1, +1]^d} |g(x) - g(x')| \leq 2B\sqrt{d}$ because $2\sqrt{d} = \max_{x, x' \in [-1, +1]^d} e(x, x')$. If the conditional expectation functions, $f_d(x_1, \dots, x_d) := E(Y | (X_1, \dots, X_d) = (x_1, \dots, x_d))$, have Lipschitz constant B , then the conditional expectation of Y can vary by $2B\sqrt{d}$ over the range of the explanatory variables.

Since $\text{Var}(Y) = E(\text{Var}(Y | X_1, \dots, X_d)) + \text{Var}(E(Y | X_1, \dots, X_d))$, we know that the variance of the $f_d(X_1, \dots, X_d)$ is bounded when Y has finite variance. Combined with functional form assumptions on the f_d , this may provide extra limits on the variability of the regression function. Indeed, functional form assumptions may provide extra limits even without finite variances.

If $f_d(x_1, \dots, x_d) = \beta_0 + \sum_{a \leq d} \beta_a x_a$ is affine, the Lipschitz bound corresponds to $\sum_{a \leq d} |\beta_a| \leq 2B \cdot \sqrt{d}$. By contrast, a bound on total efficacy requires that $\sum_a |\beta_a|$ be bounded by a number independent of d .

Example 1. If $\beta_a = \mathcal{O}\left(\frac{1}{\sqrt{a}}\right)$, then $\sum_{a \leq d} |\beta_a| = \mathcal{O}(\sqrt{d})$. This satisfies the Lipschitz bound, but goes to ∞ , violating the efficacy bound.

If $\beta_a = 1$ for $a = k^2$, $k \in \mathbb{N}$, and $\beta_a = 0$ otherwise, then $\sum_{a \leq d} |\beta_a| = \mathcal{O}(\sqrt{d})$. Again, this satisfies the Lipschitz bound, but total efficacy is unbounded.

Lemma 1. If the (X_1, X_2, \dots) are independent, take values in $[-1, +1]$, have mean 0, $\liminf_d \text{Var}(X_d) = \underline{\sigma} > 0$, and $E(Y | (X_1, \dots, X_d) = (x_1, \dots, x_d)) = \beta_0 + \sum_{a \leq d} \beta_a x_a$, then $\sum_a |\beta_a|^2 < \infty$.

Proof: If Y has finite variance, then for all d , $\text{Var}(Y) \geq \text{Var}(\sum_{a \leq d} \beta_a X_a) = \sum_{a \leq d} |\beta_a|^2 \text{Var}(X_a)$. This cannot happen if $\sum_a |\beta_a|^2$ diverges because $\underline{\sigma} > 0$.

Suppose now that Y has first but not second moments. Martingale Convergence implies that $Y_d := E(Y|(X_1, \dots, X_d)) \rightarrow Y_{\mathcal{X}} := E(Y|\mathcal{X})$ a.e. where $\mathcal{X} = \sigma(\{X_a : a \in \mathbb{N}\})$. Suppose that $\sum_a |\beta_a|^2$ diverges. This implies that there exists an increasing sequence $1 = D_1 < D_2 < \dots < D_k < \dots$ 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 3σ . ■

Thus, if Y is a random variable with finite expectation and the conditional expectations are affine, the last result showed that the patterns of β_a 's given Example 1 cannot happen. However, the arguments in Lemma 1 do not imply that the causal efficacy of the explanatory variables is bounded (more's the pity). The argument depends on the Three-Series Theorem: Suppose that R_a is a sequence of independent random variables. The convergence of the three series

$$\sum_a P(|R_a| > c), \quad \sum_a E(R_a \cdot 1_{|R_a| \leq c}), \quad \sum_a \text{Var}(R_a \cdot 1_{|R_a| \leq c})$$

for some c 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 .⁴

Example 2. *Suppose that the $X_a \in [-1, +1]$ are iid, have mean 0, and set $R_a = \beta_a X_a$, $\beta_a = \mathcal{O}(\frac{1}{a})$. For any $c > 0$, for all large a , $P(|R_a| > c) = 0$. This implies that for large a , $E(R_a \cdot 1_{|R_a| \leq c}) = 0$ and $\text{Var}(R_a \cdot 1_{|R_a| \leq c}) = \mathcal{O}(\frac{1}{a^2})$. Thus, all three series converge, so that $Y_d := \beta_0 + \sum_{a \leq d} \beta_a X_a$ converges a.e. to some random variable Y . Since the variance of the Y_d is uniformly bounded, the sequence Y_d is uniformly integrable, and Y is integrable. Thus, $Y_d = E(Y|(X_1, \dots, X_d))$ can be affine while $\sum_a |\beta_a| = \infty$. That is, causal efficacy can be unbounded even in the presence of affine conditional expectations.*

2.c. Number of Regressors Intuitions. From the previous, we see that modeling Y as a random variable suggests the set of affine functions with a Lipschitz bound is too large. Another model which suggests this involves random parameters.

⁴For a proof, see e.g. Theorem 22.8, Billingsley (1995).

Suppose that the β_a 's are independent random variables with $E|\beta_a| = 1$. Satisfying the Lipschitz constraint on average 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 them 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 $\epsilon > 0$, one can ask how many of the d regressors can be ignored and still make an error less than ϵ . That is, let $N = N(d, \epsilon)$ be the largest integer satisfying $\sum_{(a) \leq N} E \frac{1}{\sqrt{d}} |\beta|_{(a)} < \epsilon$ and $M = M(d, \epsilon)$ the largest satisfying $\sum_{(a) \leq M} E \frac{1}{d} |\beta|_{(a)} < \epsilon$.

Example 3. *If the $|\beta_a|$ are independent exponentials, 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)). If $d = 20$ and $\epsilon = 0.05$, this yields $N = 4$ and $M = 13$. This means that, 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 total variability 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 large. 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 behavior, 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.⁵

2.d. The Non-Comparability of Efficacy and Lipschitz Bounds. A potential worry is that bounds on causal efficacy may vitiate consistency arguments for non-parametric regression. This worry is unfounded. Though the Lipschitz norm and the Efficacy norm are non-comparable, the set of possible targets simultaneously satisfying both kinds of bounds are dense. This means that the

⁵As part of an extended comparison of parametric and nonparametric methods, Breiman (2001) discusses several general classes of high-dimensional situations in which this is true.

ever receding targets are used for the dismally bad “optimal” rate of convergence calculations, but are not needed for consistency.

2.d.1. *Notation and Preliminaries.* $L^0 = L^0(\Omega, \mathcal{F}, P)$ denotes the set of \mathbb{R} -valued random variables, and $L^2 = L^2(\Omega, \mathcal{F}, P) \subset L^0$ the set of square integrable random variables. 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 $L^2(\mathcal{G}) = L^2 \cap L^0(\mathcal{G})$ the set of \mathcal{G} -measurable, square integrable random variables.

$\mathbb{X} = \{X_a : a \in \mathbb{N}\} \subset L^0$ denotes the set of possible explanatory variables, \mathcal{X}_d denotes $\sigma(X_1, \dots, X_d)$, the smallest σ -field making X_1, \dots, X_d measurable, and \mathcal{X} denotes $\sigma(\mathbb{X})$, the smallest σ -field making every X_a in \mathbb{X} measurable.

Assume that $Y \in L^2$. The set of all possible target functions is $L^2(\mathcal{X})$. The set of all possible estimators targets when X_1, \dots, X_d are the regressors is $L^2(\mathcal{X}_d)$. The set of all possible targets based on some finite set of regressors is $\bigcup_d L^2(\mathcal{X}_d)$.

By Martingale Convergence, $\bigcup_d L^2(\mathcal{X}_d)$ is dense in $L^2(\mathcal{X})$, because for all $Y \in L^2$, $Y_d := E(Y|\mathcal{X}_d) \rightarrow Y_{\mathcal{X}} := E(Y|\mathcal{X})$ a.e., and $\|Y_d - Y_{\mathcal{X}}\| \downarrow 0$.

2.d.2. *Non-Comparable Norms.* By Doob’s Theorem (e.g. Dellacherie and Meyer (1978, Theorem I.18, p. 12-13)), $L^2(\mathcal{X}_d)$ is the set of functions of the form $\omega \mapsto g((X_1, \dots, X_d)(\omega))$ that are square integrable where g is a measurable function from \mathbb{R}^d to \mathbb{R} . $C_d \subset L^2(\mathcal{X}_d)$ denotes the subset of $L^2(\mathcal{X}_d)$ with g continuous. By Lusin’s Theorem, C_d is dense in $L^2(\mathcal{X}_d)$, which in turn implies that $\bigcup_d C_d$ is dense in $\bigcup_d L^2(\mathcal{X}_d)$, hence in $L^2(\mathcal{X})$.

Definition 1. *The **Lipschitz norm** of a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is*

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

*whenever this is finite. The **Lipschitz constant** of f is $\sup_{x \neq y} \frac{|f(x) - f(y)|}{e(x,y)}$.*

$C_d^{Lip}(B) \subset C_d \subset L^2(\mathcal{X}_d)$ denotes the subset of possible targets for which g has Lipschitz norm B or less. $C_d^{Lip} := \bigcup_B C_d^{Lip}(B)$ is the set of all targets that are functions of X_1, \dots, X_d and have Lipschitz norm B or less. $C^{Lip} := \bigcup_{d,B} C_d^{Lip}(B)$ is the set of all Lipschitz functions of finitely many regressors.

A **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 , the function $x_i(t)$ is either non-decreasing or non-increasing. Note

that the different components of $t \mapsto x(t)$ can move in different directions, e.g. $x(t) = (-t, +t)$, and need not be unbounded, e.g. $x(t) = (e^{-t}, 1/(1 + e^{-t}))$.

Definition 2. The *total variation* of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ is $TV(f) = \sup \sum_i |f(x_{i+1}) - f(x_i)|$ where the supremum is taken over all finite subsets $x_1 < x_2 < \dots < x_I$ of \mathbb{R} . The *monotonic total variation* of a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is $MTV(f) = \sup_x TV(f \circ x)$ where the supremum is taken over monotonic paths. Finally, the *monotonic total variation norm* is

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

$C_d^{MTV}(B) \subset C_d \subset L^2(\mathcal{X}_d)$ denotes the subset of possible targets for which g has monotonic total variation norm B or less. $C_d^{MTV} := \bigcup_B C_d^{MTV}(B)$ is the set of all targets that are functions of X_1, \dots, X_d and have monotonic total variation norm B or less. $C^{MTV} := \bigcup_{d,B} C_d^{MTV}(B)$ is the set of all functions of finitely many regressors that have bounded monotonic total variation.

For any f , $\|f\|_\infty \leq \min\{\|f\|_{Lip}, \|f\|_{MTV}\}$. However, the two norms are non-comparable, that is, there are functions for which the ratio of the two norms is as large or as small as one pleases.

Example 4. Let $f(x) = \max\{0, 1 - |x|\}$ and $g_m(x) = \max\{0, 1 - d(x, \{1, \dots, m\})\}$ so that $f = g_1$. For $f_n(x) := \frac{1}{n}f(n^2x)$, $\|f_n\|_{Lip} = \frac{1}{n} + n \uparrow \infty$ and $\|f_n\|_{MTV} = 0 + \frac{1}{n} \downarrow 0$. For all m , $\|\frac{1}{n}g_m(x)\|_{Lip} = \frac{1}{n} + \frac{1}{n} \downarrow 0$ and $\|\frac{1}{n}g_m(x)\|_{MTV} = 0 + \frac{2m}{n} \uparrow \infty$ if $m = n^2$.

Lemma 2. For any $\{X_1, \dots, X_d\} \subset L^0$, $C_d^{Lip} \cap C_d^{MTV}$ is dense in $L^2(\mathcal{X}_d)$.

Proof: Let $C_c(\mathbb{R}^d)$ denote the of continuous function with compact support. By Lusin's Theorem, $C_c(X_1, \dots, X_d) := \{f(X_1, \dots, X_d) : f \in C_c(\mathbb{R}^d)\}$ is dense in $L^2(\mathcal{X}_d)$. Convoluting with C^∞ -mollifiers with bounded support shows that the same is true for $C_c^\infty(X_1, \dots, X_d)$, the smooth elements of $C_c(X_1, \dots, X_d)$. Any $f \in C_c^\infty$ is Lipschitz because both $|f|$ and $\|Df\|$ are continuous functions, hence bounded over any compact set. Any $f \in C_c^\infty$ has bounded monotonic total variation because $\|Df\|$ is bounded, which implies that its integral is uniformly bounded over monotonic paths in the compact support of f . Thus, $[C^{Lip} \cap C^{MTV}]$ contains the dense set $C_c^\infty(X_1, \dots, X_d)$, and is therefore dense. ■

3. DIMENSION INDEPENDENT RATES

The ability of artificial neural networks to fit high dimensional data with relatively few parameters suggests that total efficacy is often small relative to d in the context of non-linear regression as well. A theoretical basis for these empirical observations was given in Barron (1993). He showed that for every d , there is a dense set of functions, \mathbb{V}_{ann}^d , depending on the architecture of the networks, such that for all $f \in \mathbb{V}_{ann}^d$, the uniform approximation condition (2.2), $\sup_x \inf_{g \in \Theta_\kappa} |f(x) - g(x)| = \mathcal{O}\left(\frac{1}{\kappa^\alpha}\right)$, is satisfied with $\alpha = \frac{1}{2}$. Returning to (2.3), one then has $\|f - \widehat{f}_n\|^2 = \mathcal{O}_P\left(\frac{\kappa}{n} + \frac{1}{\kappa^{2\alpha}}\right) = \mathcal{O}_P\left(\frac{\kappa}{n} + \frac{1}{\kappa}\right)$. Minimizing yields $\kappa(n) = \sqrt{n}$ so that $\|f - \widehat{f}_n\|^2 = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

This section 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 classes like \mathbb{V}_{ann}^d for each of the consistent version of these regression techniques.

Fix any sequence r_n going to 0 and sequence $\kappa(n)$ going to ∞ . Lemma 3 shows that for any of the consistent versions of the regression techniques in this class, the class of targets, $\mathcal{T}(r_n) = \{f \in \mathbb{V} : \|f - \Theta_{\kappa(n)}\| = \mathcal{O}(r_n)\}$, is dense. The direct implication is that for a dense class of targets, the estimation error, ϵ_n , determines the rate at which the total error goes 0. To see why, fix a sequence $\kappa(n) \rightarrow \infty$ and set $r_n = \epsilon_n$ where ϵ_n is the associated estimation error in (1.3). For every target f in the dense class $\mathcal{T}(r_n)$, $\|\widehat{f}_n - f\| \leq \|\widehat{f}_n - f_{\kappa(n)}^*\| + \|f_{\kappa(n)}^* - f\| = \mathcal{O}_P(r_n) + \mathcal{O}(r_n) = \mathcal{O}_P(r_n)$.

In wavelet and spline estimation, Cohen *et. al.* call $\mathcal{T}(r_n)$ a *maximal space*. If r_n is one of Stone's rates, then $\mathcal{T}(r_n)$ includes the Lipschitz functions. If $r_n = n^{-1/4}$, then $\mathcal{T}(r_n)$ includes the ann target functions identified in Barron (1993).

3.a. Targets. The target is a function $x \mapsto f(x)$ from the support of μ_X to \mathbb{R} that is to be estimated. Typically $f(x)$ is (a version of) the conditional expectation $f(x) = E(Y|X = x)$.⁶

⁶If $f(x) = E(h(Y)|X = x)$ for $h : \mathbb{R} \rightarrow \mathbb{R}$, we have other forms of regression. For example, if $h(y) = \gamma y 1_{(-\infty, 0]}(y) - (1 - \gamma) 1_{(0, \infty)}(y)$, $\gamma \in (0, 1)$, we have quantile regression. Etc.

The target, f , in nonparametric regression belongs to a space of functions \mathbb{V} , assumed throughout to be a separable, infinite dimensional Banach space, e.g.

- (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 $D \subset \mathbb{R}^d$, 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 with a smooth boundary and norm $\sup_{x \in D} \sum_{|\alpha| \leq m} |D^\alpha f(x)|$, typically used when smoothness of the target is an appropriate assumption,⁷
- (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 $\|f\| = \sum_{|\alpha| \leq m} [\int |D^\alpha f(x)|^p d\mu_X(x)]^{\frac{1}{p}} < \infty$, are typically used when probabilistic 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.

3.b. Estimators. An estimator of a target $f \in \mathbb{V}$ is a sequence of functions $\hat{f}_n \in \mathbb{V}$ where each \hat{f}_n depends on the data $((Y_1, X_1), \dots, (Y_n, X_n))$. For the nonparametric techniques studied here, the \hat{f}_n are of the form $\hat{f}_n(x) = \sum_k \beta_k c_k(x)$ 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, $C_{\kappa(n)} = C_{\kappa(n)}(\omega, f) \subset \mathbb{V}$ of compactly generated two-way cones with the property that $\hat{f}_n \in C_{\kappa(n)}$.

Definition 3. *The sequence $(\omega, f) \mapsto C_{\kappa(n)}(\omega, f)$ is **consistent** if for all $g \in \mathbb{V}$ and all $\epsilon > 0$, $P(\cup_N \cap_{n \geq N} [d(g, C_{\kappa(n)}(\cdot, f)) < \epsilon]) = 1$.*

If the $C_{\kappa(n)}$ are nested, then the sequence is consistent if and only if $\cup_n C_{\kappa(n)}$ is, with probability 1, dense in \mathbb{V} .

⁷Here, α is a multi-index, $\alpha = (\alpha_1, \dots, \alpha_d)$, $\alpha_i \in \{0, 1, \dots\}$, and $|\alpha| := \sum_i \alpha_i$.

3.c. **Geometry.** $U = \{f \in \mathbb{V} : \|f\| < 1\}$ is the unit ball in \mathbb{V} , its closure is \bar{U} , and $\partial U = \{f \in \mathfrak{X} : \|f\| = 1\}$ is its boundary. For $E \subset \mathfrak{X}$, $\mathbf{sp} E$ 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 $S \subset \mathbb{R}$, $S \cdot E$ is the set $\{s \cdot f : f \in E, s \in S\}$ of scalar multiples of elements of E with scalars belonging to S . 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. Allowing for multiplication by both negative and positive scalars gives **two-way cones**.

Definition 4. *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 $E \subset \bar{U}$, $0 \notin E$, such that $C = \mathbb{R} \cdot E$.*

Example 5 (Series estimators). *If $C = \mathbf{sp} \{e_1, \dots, e_\kappa\}$ with $e_k \neq 0$, then C is a κ -dimensional subspace of \mathbb{V} , and is a compactly generated two-way cone.*

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 “ $C_{\kappa(n)}$ is **consistent** if 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 $r_n \rightarrow 0$ and $C_{\kappa(n)}$ is a sequence of compactly generated two-way cones.

3.d. **Results.** This section proves Lemmas 3 and 4, 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 $r_n \rightarrow 0$, 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. Both C_d^{Lip} and C_d^{MTV} are dense, shy sets of functions, as is their intersection.

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 3. $P(\mathcal{T}(r_n) \text{ is dense}) = 1$ 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 $B_j^m = \cup_N \cap_{n \geq N} [(g_j + \frac{1}{m} \cdot U) \cap (C_{\kappa(n)}(\omega, f) + r_n \cdot U) \neq \emptyset]$. 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.⁸

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 $\eta(S' + f) = 0$ for all $f \in \mathbb{V}$.

Lemma 4. 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 5. 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 4: 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.

⁸There 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.

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 5, 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 $r_n \rightarrow 0$ and $\eta > 0$, $[(R \cdot U) \cap [A_n^M \text{ i.o.}]] \subset S$. ■

Proof of Theorem A: Lemma 3 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 4 shows that $\mathcal{T}(r_n)$ is shy. ■

4. EXAMPLES

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.

4.a. **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 Fourier series the various polynomial schemes as linear subspaces are widely available in textbooks on functional analysis. For wavelets, see e.g. Debnath (2002), for splines see e.g. Eubank (1999). The estimator based on n data points, \hat{f}_n , is a function of the form

$$\hat{f}_n(x) = \sum_{k \leq \kappa(n)} \hat{\beta}_k e_k(x). \quad (4.1)$$

The estimators \hat{f}_n belong to $C_{\kappa(n)} := \mathbf{sp} \{e_1, \dots, e_{\kappa(n)}\}$. Being a finite dimension subspace of \mathbb{V} , each $C_{\kappa(n)}$ is a compactly generated two-way cone.

Since $\overline{\mathbf{sp}} E = \mathbb{V}$, having $\lim_n \kappa(n) = \infty$ guarantees that the \hat{f}_n can approximate any function. To avoid overfitting and its implied biases, not letting $\kappa(n)$ go to infinity too quickly, e.g. $\kappa(n)/n \rightarrow 0$ guarantees consistency. If $\kappa(n) \rightarrow \infty$ is regarded a sequence of parameters to be estimated e.g. by cross-validation, then $\kappa(n)$ depends on both ω and f , which yields $C_{\kappa(n)} = C_{\kappa(n)}(\omega, f)$.

4.b. **Kernel and locally weighted regression estimators.** Kernel estimators for functions on a compact domain typically begin with a function $K : \mathbb{R} \rightarrow$

\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

$$\widehat{f}_n(x) = \sum_{i=1}^n \widehat{\beta}_i g(x|X_i, h_n) = \sum_{i=1}^n \widehat{\beta}_i K\left(\frac{1}{h_n}(x - X_i)\right). \quad (4.2)$$

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

When the kernel function, $K(\cdot)$, is smooth and all of its derivatives, $K^{(\alpha)}$, satisfy $\lim_{|u| \rightarrow 1} K^{(\alpha)}(u) = 0$, 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_m^p -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 , $i = 1, \dots, n$, and the window-size parameter h_n , define n non-zero functions, $g(\cdot|\theta_{i,n})$, $\theta_{i,n} = (X_i, h_n)$. 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 $h_n \rightarrow 0$, 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

$$\widehat{f}_n(x) = \sum_{i=1}^n \widehat{\beta}_i g(x|X_i, h_n) = \sum_{i=1}^n \widehat{\beta}_i K\left(\frac{1}{h_n}\|x - X_i\|\right) \quad (4.3)$$

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 $g_i(\cdot|\theta_{i,n})$, 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.

4.c. Artificial neural networks. Single hidden layer feedforward (slff) estimators with activation function $g : \mathbb{R} \rightarrow \mathbb{R}$ 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}^{n+1}

with non-empty interior. The slff estimators are functions of the form

$$\widehat{f}_n(x) = \sum_{k \leq \kappa(n)} \widehat{\beta}_k g(\widehat{\gamma}'_k \tilde{x}), \quad (4.4)$$

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 $\kappa(n) \rightarrow \infty$, $\kappa(n)/n \rightarrow 0$, and $\overline{\text{sp}} E = \mathbb{V}$, then the total error goes to 0. Various sufficient conditions on g that guarantee $\overline{\text{sp}} E = \mathbb{V}$ in the contexts to be described just below are given in Stinchcombe and White (1992, 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 $S_m^p(\mathbb{R}^d, \mu_X)$ 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 . Consistency issues for multiple layer feedforward networks are addressed in Hornik, Stinchcombe, and White (1989, 1990)

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

$$\widehat{f}_n(x) = \sum_{k \leq \kappa(n)} \widehat{\beta}_k g(\widehat{\gamma}'_k \tilde{x}), \quad (4.5)$$

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 $0 \notin E_n$. For the consistency properties of these neural networks, see Park and Sandberg (1991, 1993a, b).

5. COMPLEMENTS

There are a number of subsidiary points, grouped here into two kinds of comparisons, the lack of an obvious role for smoothness in the intuitions, some additional information on negligible sets, and possible extensions and generalizations.

5.a. Comparisons Across Techniques. 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 $r'_n = o(r_n)$, 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 only 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 .

5.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 4 shows that the difference between the sets, $\mathcal{T}(r_n) \setminus \mathcal{T}(r'_n)$, is Haar null. Haar null 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 4 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 nonparametric 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.⁹

5.c. **Smoothness.** Another aspect of the work on optimal 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. 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.

5.d. **More on Negligible Sets.** From the definition, S is shy iff $S + f$ is shy for all f . 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 iff it is a Lebesgue null set. Other relevant properties of the 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,
- (3) if \mathbb{V} is infinite dimensional, then compact sets are shy.

Examining naive random search in function spaces gave rise to the key result used in the shyness proofs here, Lemma 1 in Stinchcombe (2001). This result also shows that there is no comfortable Bayesian interpretation of shy sets.

⁹I am grateful to Xiaohong Chen and Jinyong Hahn for these last two points.

5.e. **Possible Extensions and Generalizations.** There are several additional points to be made.

- (1) If C_κ is smoothly parametrized and κ is fixed, then standard finite dimensional results imply that $\|\widehat{f}_n - f_\kappa^*\| = \epsilon_{\kappa,n} = \mathcal{O}(n^{-1/2})$. If $\kappa(n) \uparrow \infty$, as required for consistency, but $\kappa(n)$ grows very slowly, the $n^{-1/2}$ rate slows so little as one desires. This observation provides little practical guidance.
- (2) The proof of Lemma 3 can be easily adapted to show that for nested $C_{\kappa(n)}$, 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.
- (3) Pick a pair of sequences r_n, r'_n with $r'_n = \mathbf{o}(r_n)$. From Lemma 3, $\mathcal{T}(r_n) \setminus \mathcal{T}(r'_n)$ and $\mathcal{T}(r'_n)$ are disjoint, dense sets of nonparametric targets. I 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 $\mathcal{T}_2(r'_n) \setminus \mathcal{T}_1(r_n) \neq \emptyset$.
- (4) All of the above has been phrased as regression analysis of conditional means. Since Lemmas 3 and 4 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.
- (5) The use of Banach spaces for the set of targets is not crucial. The main result driving the proofs is Stinchcombe (2001, Lemma 1), which applies in locally convex, complete, separable, metric vector 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 Frechet spaces that appear in nonparametric regression analyses.
- (6) 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 as points in a convex subset of the positive orthant in a Hilbert space. Lemma 3 should go through fairly easily, but Lemma 4 may be more difficult. The argument requires extending Stinchcombe (2001, Lemma 1) to what are called relatively shy sets in Anderson and Zame (2001).
- (7) If the data is not iid but has some time series structure, one expects that the estimation error in (1.3) will not be $\mathcal{O}(n^{-1/2})$ for fixed κ , but something slower. Again, since Lemmas 3 and 4 concern approximation error, total error for the nonparametric regressions covered here would also go to 0 at this slower rate.

- (8) In the above discussion of the locally weighted regression schemes and the artificial neural network estimators, I made use of compact domain assumptions to ease the exposition. Since the distribution of the data, \mathbb{Z} , in $\times_{i \in \mathbb{N}} \mathbb{R}^{1+d}$ 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.
- (9) Lemmas 4 and A used Haar null sets. They would not hold with the original and more restrictive class of infinite dimensional null sets due to Aronszajn (1976), now called **Gauss null** sets (Benyamini and Lindenstrauss, 2000, Ch. 6). S is Gauss null iff for every non-degenerate Gaussian distribution, η , on \mathbb{V} , $\eta(S) = 0$. Every Gauss null set is Haar null, 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.
- (10) 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 5.

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 \bar{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 $\epsilon > 0$.*

6. CONCLUSIONS

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, $f(x) = E(Y|X = x)$, belongs to the set of Lipschitz functions. 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 various approaches.

7. REFERENCES

- 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.
- 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. (2006). Large sample sieve estimation of nonparametric models. Forthcoming *Handbook of Econometrics*.
- 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.

- Feller, W. (1971). *An Introduction to Probability Theory and its Applications, v. II*. New York, N.Y.: John Wiley and Sons.
- Hornik, K. (1993). Some New Results on Neural Network Approximation. *Neural Networks* **6**(8), 1069-1072.
- Horowitz, J. L. and S. Lee (2005). Nonparametric Estimation of an Additive Quantile Regression Model. *Journal of the American Statistical Association* **100**(472), 1238-1249.
- Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation*, 2nd ed. New York, N.Y.: Springer-Verlag.
- 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.
- Pollard, D. (1984). *Convergence of Stochastic Processes*. New York, N.Y.: Springer-Verlag.
- 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.
- Yatchew, A. (1998). Nonparametric Regression Techniques in Economics. *Journal of Economic Literature* **36**, 669-721.

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.

DEPARTMENTS OF ECONOMICS, UNIVERSITY OF TEXAS AT AUSTIN AND CALIFORNIA INSTITUTE OF TECHNOLOGY, e-mail: maxwell@eco.utexas.edu