Gel'fand n-Widths and the Method of Least Squares

by

David L. Donoho
Department of Statistics
University of California, Berkeley

Technical Report No. 282
December 1990

Department of Statistics
University of California
Berkeley, California 94720

Gel'fand n-widths and the Method of Least Squares

David L. Donoho
Department of Statistics
University of California, Berkeley

November 9, 1990

Abstract

Consider the problem of estimating a function f known to lie in a convex, compact subset $\mathcal{F}$ of $L_2[0,1]$, when one observes data on f containing white Gaussian noise. We establish an upper bound on the integrated mean-squared error of least-squares estimates which uses the asymptotic properties of the Gel'fand n-widths of $\mathcal{F}$. The bound shows that if the Gel'fand n-widths tend to zero at a faster rate than the Kolmogorov linear n-widths, least squares must outperform every orthogonal series estimator, at the level of minimax rates of convergence. We rely heavily on Carl's Theorem, a recent development in the Geometry of Banach Spaces. As an application, we resolve a question about the performance of least-squares estimates in estimating decreasing functions from noisy sampled data.

Acknowledgements. The author would like to thank Lucien Birgé for helpful conversations at the Ecole d'Ete de Probabilités in Saint Flour, and to Prof. P.L. Hennequin for arranging my visit there. Conversations with Henryk Woźniakowski about Gel'fand n-widths were also helpful. This work was supported by NSF DMS-88-10192, and by NASA NCA2-0488.

Key Words and Phrases. White Noise Model. Minimax Risk. n-widths of Gel'fand and Kolmogorov. Estimating a decreasing function.

1 Introduction

Suppose we observe data $Y(t), t \in [0,1]$ characterized by

$$Y(dt) = f(t)dt + \sigma W(dt) \quad t \in [0,1] \tag{1}$$

with W a Wiener Process and σ a formal noise level parameter which we think of as small. We know a priori that $f \in \mathcal{F}$, a convex, compact class of functions, and we wish to estimate f using the information Y.

Section 2 below defines a notion of least-squares estimate $\hat{f}$ for this infinite-dimensional setting. Define $R^*_{LS}(\sigma) = \sup_{\mathcal{F}} E\|\hat{f} - f\|_2^2$; the worst risk in $\mathcal{F}$ of a least-squares estimate. Define the *Gel'fand n-widths*

$$d^n(\mathcal{F}, L^2) = \inf_{L_i \text{ linear}} \sup_{\mathcal{F}} \{\|f\|_2 : L_i f = 0,\ i = 1, ..., n\} \tag{2}$$

These n-widths measure geometric properties of $\mathcal{F}$; compare Pinkus (1985). We show here that they may also be used to bound the risk of least squares estimates. Section 3 and 4 below develop inequalities which imply

Theorem 1 *Let $\mathcal{F}$ be convex, compact, and centrosymmetric. Let $d^n \leq C_0 n^{-\alpha}$; there is a constant C_1, so that*

$$R^*_{LS}(\sigma) \leq C_1 (\sigma^2)^{\frac{2\alpha}{2\alpha+1}} \tag{3}$$

This result has interesting implications which stem from the fact that Gel'fand n-widths, rather than Kolmogorov n-widths, are being employed. Let $\mathcal{P}_n$ denote orthogonal projection onto an n-dimensional subspace of $L_2[0,1]$. The *Kolmogorov linear n-width* is defined as

$$d_n(\mathcal{F}, L^2) = \inf_{\mathcal{P}_n} \sup_{\mathcal{F}} \|f - \mathcal{P}_n f\|_{L_2[0,1]}$$

We note that $d^n \leq d_n$. In section 5 below we discuss orthogonal series estimates $\hat{f}_n = \sum_{i=1}^n y_i \phi_i$, $y_i = \int_0^1 \phi_i Y(dt)$, with $(\phi_i)_{i=1}^\infty$ a complete orthonormal system in $L_2[0,1]$. With $R^*_{OS}(\sigma; \mathcal{F}) = \inf_n \inf_{(\phi_i)} \sup_{\mathcal{F}} E\|\hat{f}_n - f\|^2$, the minimax risk among orthogonal series estimates, we have by an easy calculation that $d_n \asymp n^{-a}$ if and only if $R^*_{OS}(\sigma) \asymp (\sigma^2)^{\frac{2a}{2a+1}}$. Comparison with Theorem 1 yields:

Corollary 1 *If the Kolmogorov linear n-widths satisfy* $d_n \asymp n^{-a}$ *and the Gel'fand n-widths satisfy* $d^n \leq cn^{-\alpha}$, *then* $\alpha > a$ *implies that* $R^*_{OS}(\sigma) >> R^*_{LS}(\sigma) \quad \sigma \to 0$.

In view of $d^n \leq d_n$, we have that least squares either performs about equally as well as orthogonal series, or else it beats orthogonal series estimates convincingly. Under certain geometric symmetry conditions on $\mathcal{F}$ (Donoho, Liu, and MacGibbon, 1990) minimax orthogonal series estimates are within a factor 4 of minimax among all linear estimates; so in such cases the condition $d_n >> d^n$ implies that least squares beats *all* linear estimates in a minimax sense.

The n-widths are, of course, fundamental quantities outside of statistical estimation, in approximation theory (Pinkus, 1986) and computational complexity (Traub, Wasilkowski, Woźniakowski, 1988). The relationship $d^n(\mathcal{F}, L^2) << d_n(\mathcal{F}, L^2)$ is also of importance in those fields. Matei (1990), for example, points out the following implications for the *approximation problem.* Suppose we wish to choose n linear functionals $L_1, ..., L_n$, for the purpose of approximating a function f. The values $(L_i f)_{i=1}^n$ will be used to construct an approximation to f, via some rule $\hat{f} = T(L_1 f, ..., L_n f)$. [There is no noise in this problem!] Then if all nonlinear rules T are allowed, and $\mathcal{F}$ is centrosymmetric:

$$\inf_{T, L_1, ..., L_n} \sup_{\mathcal{F}} \|\hat{f} - f\| = d^n$$

On the other hand, if we restrict ourselves to coefficient functionals $L_i f = \int f \phi_i$ with respect to an orthonormal set (ϕ_i), and to the reconstruction rules $T(L_1 f, ..., L_n f) = \sum_{i=1}^n (L_i f)\phi_i$, then

$$\inf_{(\phi_i)_{i=1}^n} \sup_{\mathcal{F}} \|\hat{f} - f\| = d_n$$

Hence the relationship $d^n << d_n$ implies that general approximation schemes to f do better than orthogonal linear ones.

Hence *a single geometric condition on* $\mathcal{F}$ *has parallel implications in statistical estimation and in optimal recovery.* For other parallels between the two subjects, see Donoho (1989).

Does the condition ever hold? In fact it is well-known (Pinkus, 1985) that $d^n(\mathcal{F}, L^2) << d_n(\mathcal{F}, L^2)$ whenever $\mathcal{F} = W^{m,p}$ a standard L_p Sobolev ball, and $p < 2$. See section 5 for more details.

Does the infinite-dimensional white noise estimation problem lead to results for recognizable estimation problems with finite data? Yes. In section 6 we apply these results to determining performance of least squares estimates in the traditional sampled data model $y_i = f(t_i) + sz_i$, with $z_i \overset{iid}{\sim} N(0,1)$. We determine the rates of convergence of least squares estimates of a decreasing function and thereby show that certain $\log n$-type terms appearing in earlier work of Nemirovskii et al. (1983) and van de Geer (1988) may be omitted.

The main tool of this paper is the theorem of Carl (1981) relating Gelfand n-widths of sets to their Entropy numbers. Part of our paper amounts to the interpretation of his inequality as a statement that the worst-case performance of least-squares over a function class is always within a universal constant of the minimax performance of orthogonal series rules over the same class. The other part of our paper is to interpret differences between types of n-width in statistical terms.

2 Least Squares in the White Noise Model

We begin with a few remarks on the definition of least squares estimates for the white noise model (1). Let $(\phi_i)_{i=1}^{\infty}$ be a complete orthonormal system for $L_2[0,1]$. Define

$$\begin{aligned} \theta_i(f) &= \int_0^1 \phi_i(t) f(t) dt \quad i = 1, 2, \ldots \\ y_i &= \int_0^1 \phi_i(t) Y(dt) \quad i = 1, 2, \ldots \end{aligned}$$

The $\theta_i(f)$ are the Fourier-Bessel coefficients of f with respect to the system $(\phi_i)_{i=1}^{\infty}$, and the y_i obey $y_i = \theta_i(f) + \sigma z_i$, with $z_i \overset{iid}{\sim} N(0,1)$; we call (y_i) the empirical Fourier-Bessel coefficients.

At first glance $\|y - \theta(f)\|^2 = \sum_{i=1}^{\infty} (y_i - \theta_i(f))^2$ is the "squared distance" between f and y, and our impulse, based on experience with finite dimensional least-squares, is to compare trial models $\tilde{f}$ by the value of $\|y - \theta(\tilde{f})\|^2$, seeking a model for which this squared distance is smallest. Unfortunately, in this infinite dimensional situation, we have almost surely that $\|y - \theta(\tilde{f})\|^2 = +\infty$ for all $\tilde{f} \in \mathcal{F}$. To address this difficulty, we set $\|v\|_n = \sqrt{\sum_{i=1}^{n} v_i^2}$ and write, formally,

$$Q(\tilde{f}) = \lim_{n \to \infty} \{ \|y - \theta(\tilde{f})\|_n^2 - \|y\|_n^2 \} \tag{4}$$

the subtracted term is intended to make $Q(\tilde{f})$ finite. The least-squares estimator is then, formally, the minimizer of $Q(\tilde{f})$:

$$\hat{f} = \arg\min Q(\tilde{f}) : \ \tilde{f} \in \mathcal{F} \tag{5}$$

To justify these formal definitions, we introduce the *isonormal process*, a Gaussian stochastic process $Z(f)$ indexed by $f \in \mathcal{F}$. It is defined by

$$Z(f) = \sum_{i=1}^{\infty} \theta_i(f) z_i \tag{6}$$

where $z_i = \int_0^1 \phi_i(t) W(dt)$ are iid $N(0,1)$; equivalently $Z(f) = \int_0^1 f(t) W(dt)$. For each fixed $f \in L_2$, $Z(f) \sim N(0, \int f^2)$.

Lemma 1 *Suppose that $\mathcal{F}$ is compact for the L_2 topology, and that the process $\{Z(f) : f \in \mathcal{F}\}$ has uniformly continuous sample paths on $\mathcal{F}$ on an event Ω of probability one. On the event Ω, the limit in (4) exists, and $Q(\tilde{f})$ is a bounded, uniformly continuous function of $\tilde{f} \in \mathcal{F}$, where continuity is with respect to L_2 convergence.*

It follows that, with probability one, (5) makes sense and an $\hat{f}$ satisfying (5) exists.

Proof. Now

$$\begin{aligned} & \|y - \theta(\tilde{f})\|_n^2 - \|y\|_n^2 \\ &= -2\sigma \sum_{i=1}^{n} z_i \theta_i(\tilde{f}) + \sum_{i=1}^{n} (\theta_i(f) - \theta_i(\tilde{f}))^2 - \sum_{i=1}^{n} \theta_i^2(f) \end{aligned}$$

Letting $\mathcal{P}_n f = \sum_{i=1}^{n} \theta_i(f) \phi_i$ denote the orthogonal projection of f onto the span of $(\phi_1, ..., \phi_n)$, we may rewrite the formal expression (4) as

$$Q(\tilde{f}) = \lim_{n \to \infty} \{-2\sigma Z(\mathcal{P}_n \tilde{f}) + \|\mathcal{P}_n(\tilde{f} - f)\|^2 - \|\mathcal{P}_n f\|^2\}; \tag{7}$$

we still need to justify that the limit exists and has the indicated properties. As $\mathcal{F}$ is compact, $\sup_{\tilde{f} \in \mathcal{F}} \|\tilde{f} - \mathcal{P}_n \tilde{f}\| \to 0$ as $n \to \infty$, and so

$$\begin{aligned} \mathcal{P}_n(f - \tilde{f}) &\to f - \tilde{f} \\ \mathcal{P}_n f &\to f \end{aligned}$$

strongly in L_2, uniformly in $\mathcal{F}$. Hence uniformly in $\mathcal{F}$

$$\begin{aligned} \|\mathcal{P}_n(f-\tilde{f})\| &\to \|f-\tilde{f}\| \\ \|\mathcal{P}_n f\| &\to \|f\| \end{aligned}$$

Also, on the event Ω, $Z(\mathcal{P}_n f) \to Z(f)$ and $Z(\mathcal{P}_n \tilde{f}) \to Z(\tilde{f})$ uniformly in $\tilde{f} \in \mathcal{F}$. It follows that on Ω, the right hand side of (7) has a limit, which is

$$-2\sigma Z(\tilde{f}) + \|f-\tilde{f}\|^2 - \|f\|^2 \tag{8}$$

Also on Ω, this is a uniformly continuous, bounded function on $\mathcal{F}$. The proof of Lemma 1 is complete.

Now note that as $\hat{f}$ is a least squares estimate, we must have

$$Q(\hat{f}) \leq Q(f)$$

which implies, on Ω,

$$\|\hat{f}-f\|^2 \leq 2\sigma(Z(\hat{f}) - Z(f)) \tag{9}$$

This fundamental identity is the source of all our estimates of $E\|\hat{f}-f\|^2$. The idea to bound the error of least squares estimates by the increments of stochastic processes was first developed, in a finite dimensional setting, by van de Geer (1988).

We use (9) as follows. Let $\mathcal{W}$ denote the modulus of continuity of the stochastic process Z over $\mathcal{F}$:

$$\mathcal{W}(\delta) = \sup\{Z(f) - Z(g) : \|f-g\| \leq \delta,\ f,g \in \mathcal{F}\} \tag{10}$$

Then from (7) we see that if Ω holds, then $\|\hat{f}-f\| \geq \delta$ only if

$$\delta^2 \leq 2\sigma\mathcal{W}(\delta) \tag{11}$$

Let $\Delta = \Delta(\sigma; \mathcal{F}, Z)$ be the random variable which is the largest δ for which (11) still holds. Then we have

$$\|\hat{f}-f\| \leq \Delta \quad \text{on } \Omega \tag{12}$$

and so $E\Delta^2(\sigma) \geq R^*_{LS}(\sigma)$.

In short, we get a upper bound on the error of least squares estimates using the modulus of continuity of Z.

A heuristic analysis of the situation may help the reader at this point. Let $\omega(\delta) = A\delta^r$ for some $r \in (0,1)$ and consider the behavior of the deterministic quantity

$$\bar{\Delta}(\sigma) = \sup\{\delta : \ 2\sigma\omega(\delta) - \delta^2 \geq 0\}$$

A little algebra gives

$$\bar{\Delta}(\sigma) = (2A\sigma)^{\frac{1}{2-r}}$$

Hence, if the random quantity $\mathcal{W}(\delta)$ "behaves like" $A\delta^r$, we may guess that $\Delta(\sigma)$ "behaves like" $\sigma^{\frac{1}{2-r}}$ and hence that $\|\hat{f} - f\|_2 = O_p(\sigma^{\frac{1}{2-r}})$.

3 Modulus of Continuity and Entropy Numbers

The "modern theory" of Gaussian processes, see Dudley (1973), shows how to bound the modulus of continuity $\mathcal{W}(\delta)$ in terms of the *covering numbers* of $\mathcal{F}$:

$$N(\varepsilon) = \inf\{card(\mathcal{H}) : \ \inf_{h \in \mathcal{H}} \sup_{f \in \mathcal{F}} \|h - f\|_2 \leq \varepsilon\}$$

We use as our reference the book of Adler (1990) Chapters 3, 4 and 5. Combining the inequalities (4.8), (4.9) and (4.48) in that reference, we get

Proposition 1 *If* $\int_0^\infty (\log N(\varepsilon))^{1/2} d\varepsilon < \infty$ *there exists a constant* K *so that*

$$E\mathcal{W}(\delta) \leq K(\delta |\log \delta| + \int_0^\delta (\log N(\varepsilon))^{1/2} d\varepsilon)$$

Compare Dudley (1973, Corollary 2.4, Page 75). From this, we have immediately

Corollary 2 *If* $Diam(\mathcal{F}) = \sup\{\|f_1 - f_2\| : \ f_i \in \mathcal{F}\} \leq 1$ *and if* $N(\varepsilon) \leq C_1 \exp\{C_2 \varepsilon^{-1/2\alpha}\}$ *there exists a constant* A *such that*

$$E\mathcal{W}(\delta) \leq A\delta^{1-1/2\alpha}$$

Hence, the deterministic function $\omega(\delta) = E\mathcal{W}(\delta)$ has at worst the behavior $A\delta^r$ with $r = 1 - 1/2\alpha$; the heuristic mentioned earlier suggests that $\frac{1}{2-r} = \frac{1}{1+1/2\alpha} = \frac{2\alpha}{2\alpha+1}$ is a bound on the rate of convergence $\|\hat{f} - f\|_2$ to zero.

To make this rigorous, we need to show that the sample path properties of $\mathcal{W}(\delta)$ are not significantly different than the properties of $\omega(\delta)$. For Gaussian processes, the remarkable inequality of Borrell assures us that this is in fact true:

Borrell's Inequality (Adler, 1990, Theorem 2.1, page 40)
Let $X(h)$ be a zero mean Gaussian process with a.s. continuous sample paths on $\mathcal{H}$. Put $\|X\| = \sup_{h\in\mathcal{H}} X(h)$, and $\sigma^2_{\mathcal{H}} = \sup\{Var(X(h)),\ h \in \mathcal{H}\}$. Then

$$\Pr\{|\|X\| - E\|X\|| > \lambda\} \leq 2\exp\{-\frac{\lambda^2}{2\sigma^2_{\mathcal{H}}}\}.$$

Borrell's inequality gives, with $\mathcal{H} = \{g = f_1 - f_2 :\ f_i \in \mathcal{F}, \|g\| \leq \delta\}$, and $\sigma^2_{\mathcal{H}} = \delta^2$ that for any $\lambda > E\mathcal{W}(\delta)$

$$\Pr\{\mathcal{W}(\delta) > \lambda\} \leq 2\exp\{-\frac{(\lambda - E\mathcal{W}(\delta))^2}{2\delta^2}\} \tag{13}$$

We use (13) as follows. Set $\lambda_B = B\delta^{1-1/2\alpha}, B > A$ so that $\lambda_B > E\mathcal{W}(\delta)$. Then $(\lambda_B - E\mathcal{W}(\delta))^2 \geq (\lambda_B - \lambda_A)^2$; hence

$$\Pr\{\mathcal{W}(\delta) > \lambda_A\} \leq 2\exp\{-\frac{(\lambda_B - \lambda_A)^2}{2\delta^2}\}$$

On the other hand, set $\lambda_{\sigma,\delta} = \delta^2/2\sigma$. By (11)

$$\Pr\{\mathcal{W}(\delta) > \lambda_{\sigma,\delta}\} \geq \Pr\{\Delta(\sigma) > \delta\}$$

If we choose δ so that $\lambda_{\sigma,\delta} = \lambda_B$, we have

$$\Pr\{\Delta(\sigma) > \delta\} \leq 2\exp\{-\frac{(\lambda_B - \lambda_A)^2}{2\delta^2}\}$$

With a little algebra, we get that with $j \geq 1$,

$$\Pr\{\Delta(\sigma) > 2^j\lambda_A\} \leq AC(\sigma)^j$$

with $C(\sigma) \to 0$ as $\sigma \to 0$. A slight refinement of the argument gives bounds on $E\Delta(\sigma)^2$.

Proposition 2 *Let* $Diam(\mathcal{F}) \leq 1$ *and* $N(\varepsilon) \leq C_1 \exp\{C_2 \varepsilon^{-1/\alpha}\}$. *Then for a constant* C_3,

$$E\Delta(\sigma)^2 \leq C_3(\sigma^2)^{\frac{2\alpha}{2\alpha+1}} \tag{14}$$

This validates the rate derived heuristically from properties of $E\mathcal{W}(\delta)$.

4 Entropy Numbers and Gel'fand n-widths

G. Pisier, in his recent book *The Volume of Convex Bodies and Banach Space Geometry* (1989), surveys various relations between Gaussian processes indexed by compact convex sets and the geometry of those sets. We use results developed there to relate the rate bounds just given to the Gel'fand n-widths. Define the *entropy numbers*

$$e_n = \inf\{\varepsilon : \ N(\varepsilon) \leq 2^n\} \tag{15}$$

These numbers are very analogous to n-widths d_n and d^n in several ways, as Pisier mentions in Chapter 6. For us, we need the result that e_n is in some sense *smaller* than d_n and d^n.

Proposition 3 (Carl (1983)) *Let $\mathcal{F}$ be a convex, compact, and centrosymmetric ($f \in \mathcal{F}$ implies $-f \in \mathcal{F}$) subset of L_2. There exists a constant $\rho_\alpha > 0$ so that*

$$\sup_n n^\alpha e_n \leq \rho_\alpha \sup_n n^\alpha d^n$$

We combine this with the observation that

Corollary 3 *If the Gel'fand n-widths satisfy $d^n \leq C_0 n^{-\alpha}$, then*

$$N(\varepsilon) \leq 2\exp\{C_1 \varepsilon^{-1/\alpha}\}$$

with $C_1 = \log(2)(\rho_\alpha C_0)^{1/\alpha}$.

Proof: Pick $C > C_0$. As $N(e_n+) \leq 2^n$, $N(Cn^{-\alpha}) \leq 2^n$. Then with $\varepsilon_n = Cn^{-\alpha}$, and $N(\varepsilon_n) \leq \exp\{\log(2)(\varepsilon_n/C)^{-1/\alpha}\}$. For given $\varepsilon > 0$, let $m(\varepsilon) = \inf\{\varepsilon_n : \varepsilon_n > \varepsilon\}$; and so $\varepsilon_m > \varepsilon$ and

$$\begin{aligned} N(\varepsilon) \leq N(\varepsilon_m) &\leq \exp\{\log(2)(\varepsilon_{m(\varepsilon)}/C)^{-1/\alpha}\} \\ &\leq \exp\{\log(2)[(\varepsilon/C)^{-1/\alpha} + 1]\} \\ &= 2\exp\{\log(2)(\varepsilon/C)^{-1/\alpha}\} \quad \square \end{aligned}$$

Combining now Propositions 1, 2, and 3 gives Theorem 1 of the introduction.

5 Linear Estimation and Kolmogorov n-widths

Let $\mathcal{P}_n$ denote an orthogonal projection on an n-dimensional linear subspace of L_2. The Kolmogorov linear n-width is defined as

$$d_n = \inf_{\mathcal{P}_n} \sup_{f \in \mathcal{F}} \|\mathcal{P}_n f - f\|_2 \tag{16}$$

In general $d_n \geq d^n$.

Consider an orthogonal series estimator in the white noise problem

$$\tilde{f}_n = \sum_{i=1}^{n} y_i \phi_i \tag{17}$$

where $y_i = \int_0^1 \phi_i Y(dt)$ and (ϕ_i) is a complete orthonormal system for L_2. With $\mathcal{P}_n f = \sum_{i=1}^n \theta_i(f)\phi_i$, we have $Bias^2(\tilde{f}_n) = \|\mathcal{P}_n f - f\|^2$ and $\int_0^1 Var(\tilde{f}_n(t))dt = \sum_{i=1}^n Var(y_i) = n\sigma^2$. Consequently, $MSE(\tilde{f}_n, f) = \|\mathcal{P}_n f - f\|^2 + n\sigma^2$. For the minimax orthogonal series risk we have

$$\begin{aligned} R^*_{OS}(\sigma) &= \inf_n \inf_{(\phi_i)_{i=1}^\infty} \sup_{\mathcal{F}} \|\mathcal{P}_n f - f\|^2 + n\sigma^2 \\ &= \inf_n d_n^2 + n\sigma^2 \end{aligned} \tag{18}$$

It follows from an obvious calculation that if $d_n \asymp n^{-a}$, $a > 0$, then

$$R^*_{OS}(\sigma) \asymp (\sigma^2)^{\frac{2a}{2a+1}} \tag{19}$$

Comparing (19) with Theorem 1 gives Corollary 1 of the introduction.

Pinkus (1985, Chapter VII, page 232) presents results on the asymptotics of n-widths of Sobolev balls $W^{m,p}([0,1]) = \{f : ||f^{(m)}||_{L^p[0,1]} \leq 1\}$.

Proposition 4 *Let $m \geq 2$. Then*

$$\begin{aligned} d_n(W^{m,p}, L_2) &\asymp \begin{cases} n^{-m} & 2 \leq p \leq \infty \\ n^{-m+1/p-1/2} & 1 \leq p \leq 2 \end{cases} \\ d^n(W^{m,p}, L_2) &\asymp n^{-m} \; 1 \leq p \leq \infty \end{aligned} \tag{20}$$

Consequently, if $p < 2$, least-squares significantly outperforms orthogonal series estimates. If $p \geq 2$, it follows from an additional argument that least-squares and minimax orthogonal series estimates are rate-equivalent.

As an example, let $m = 1, p = 1$. Then $d_n \asymp n^{-1/2}$, but $d^n \asymp n^{-1}$. It follows that

$$R^*_{LS}(\sigma) \asymp (\sigma^2)^{2/3}$$

while

$$R^*_{OS}(\sigma) \asymp (\sigma^2)^{1/2}$$

[The rate $(\sigma^2)^{2/3}$ is in fact optimal among all measurable procedures; the rate $(\sigma^2)^{1/2}$ is optimal among all linear procedures. Compare Donoho and Johnstone (1990)].

6 Application: Estimating a Decreasing Function

Consider now a sampled-data estimation problem: we observe

$$y_i = f(t_i) + sz_i \tag{21}$$

with $i = 1, ..., n$, $z_i \overset{iid}{\sim} N(0,1)$, $t_i = i/n$. We know a priori that $f \in \mathcal{D} = \{f : f\text{decreasing on}[0,1], \|f\|_\infty \leq 1\}$. We wish to apply the discrete least-squares estimator

$$\tilde{f}_n = \arg\min \sum_{i=1}^{n} (y_i - f(t_i))^2 \quad f \in \mathcal{D} \tag{22}$$

where we make the convention that $\tilde{f}_n$ is piecewise constant on $(t_{i-1}, t_i]$.

Define

$$\tilde{R}(n) = \sup_{\mathcal{D}} E\|\tilde{f}_n - f\|^2 \tag{23}$$

this quantity has been studied by Nemirovskii, Tsybakov, and Polyak (1985) and by van de Geer (1988, 1990); compare also forthcoming work of Birgé and Massart.

By the work of these authors, we know that for large n,

$$cn^{-2/3} \leq \tilde{R}(n) \leq Cn^{-2/3}\log(n)^\beta \tag{24}$$

for some constants c, C and some power β. Closest to our point of view are the papers of van de Geer and of Birgé and Massart; these authors attempt

to bound covering numbers for $\mathcal{D}$ in L_2 distance but they are unable to avoid certain logarithmic factors in these calculations.

We would like to point out here that essentially as a direct consequence of Theorem 1 we can improve the upper bound to (24) and get

Theorem 2

$$\tilde{R}(n) \asymp n^{-2/3} \tag{25}$$

While the improvement of (25) over (24) is not dramatic, it is definitive. Morover, it is obtained in obtained in a simple and natural way which has many other applications.

The driving idea is that of *White Noise Approximation*, namely that the data (21) are essentially equivalent to the white noise data (1) with noise level $\sigma = s/\sqrt{n}$.

Set $T_i = ((i-1)/n, i/n]$, except $T_1 = [0, 1/n]$. Let S_n denote the linear operator defined by $S_n f = \sum_{i=1}^n f(t_i) 1_{T_i}$; this delivers a step function approximation to f, based on samples at the t_i.

Note that

$$S_n \mathcal{D} \subset \mathcal{D} \tag{26}$$

and we record that $\mathcal{D}$, as a subset of $\mathcal{F}$, enjoys

$$\sup_{f \in \mathcal{F}} \|f - S_n f\|_2 \leq \frac{1}{\sqrt{n}} \tag{27}$$

This inequality tells us that with $f_n = S_n f$, $\|f - f_n\|^2 = O(1/n)$; this is negligible compared with worst case $E\|\tilde{f}_n - f\|^2$ and so estimating f_n is materially the same as estimating f.

Define now the stochastic process

$$Z_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g(t_i) z_i \tag{28}$$

where z_i are the iid noise values in (21). As in (22) we constrain our estimates $\tilde{f}_n$ to satisfy $\tilde{f}_n = S_n f_n$. Arguments along the lines of section 2 lead to the inequality

$$\|\tilde{f}_n - f_n\|^2 \leq 2\sigma (Z_n(\tilde{f}_n) - Z_n(f_n)) \tag{29}$$

We use this inequality as in section 2. Define

$$\mathcal{W}_n(\delta) = \sup\{Z_n(f_1) - Z_n(f_2) : \|f_1 - f_2\| \leq \delta,\ f_i \in S_n\mathcal{D}\}$$

and

$$\Delta_n = \sup\{\delta : \frac{2s}{\sqrt{n}}\mathcal{W}_n(\delta) - \delta^2 \geq 0\}$$

So that

$$\Delta_n \geq \|\tilde{f}_n - f_n\|. \tag{30}$$

Now we reduce the problem to the study of Z and $\mathcal{W}$. On an appropriate probability space, the r.v. z_i in (21) and r.v. W in (1) are related by

$$z_i = |T_i|^{-1/2} \int_{T_i} W(dt) \tag{31}$$

With this convention, we have the crucial identity

$$Z_n(g) = Z(S_n g) \tag{32}$$

It follows that if $f_i \in S_n\mathcal{D}$

$$Z_n(f_i) = Z(f_i)$$

and so

$$\begin{aligned} \mathcal{W}_n(\delta) &= \sup\{Z(f_i) - Z(f_2) : \|f_1 - f_2\| \leq \delta,\ f_i \in S_n\mathcal{D}\} \\ &\leq \sup\{Z(f_i) - Z(f_2) : \|f_1 - f_2\| \leq \delta,\ f_i \in \mathcal{D}\} \\ &= \mathcal{W}(\delta; \mathcal{D}) \end{aligned}$$

Consequently on an appropriate probability space,

$$\Delta_n \leq \Delta(\frac{s}{\sqrt{n}}; \mathcal{D})$$

and

$$\begin{aligned} \|\tilde{f}_n - f\| &\leq \Delta(\frac{s}{\sqrt{n}}) + \sup_{\mathcal{D}} \|f - f_n\| \\ &\leq \Delta(\frac{s}{\sqrt{n}}) + \frac{1}{\sqrt{n}} \end{aligned} \tag{33}$$

To complete the analysis, we note that $\mathcal{D} \subset \mathcal{F}$, and that $\mathcal{F}$ is the closure, in L_2-norm, of the set $W^{1,1}$, which has Gel'fand n-widths $d^n \leq C/n$ for some constant C. In an obvious notation

$$\mathcal{W}(\delta; \mathcal{D}) \leq \mathcal{W}(\delta; \mathcal{F})$$

and also

$$\Delta(\sigma; \mathcal{D}) \leq \Delta(\sigma; \mathcal{F})$$

By Proposition 1

$$E\Delta^2(\sigma; \mathcal{F}) \leq C(\sigma^2)^{2/3}$$

Combining this with (33) gives

$$\tilde{R}(n) \leq \text{Const } s^{4/3} n^{-2/3} \quad n > n_0$$

and proves Theorem 2.

7 Efficiency of Least-squares

Theorem 2 together with other knowsn results (e.g. Nemirovskii, Tsybakov, and Polyak, 1985) may be interpreted as saying that least-squares is *efficient*, in the sampled-data problem, as regards minimax rates of convergence: no other method can estimate functions over $\mathcal{D}$ with a faster rate of convergence to zero of $\sup_{\mathcal{D}} E||\hat{f} - f||^2$.

This raises the natural question whether a parallel result holds in the problem with white noise data (1). It does; here is a simple criterion of general usefulness.

Let $b_{n,2}(\mathcal{F})$ denote the largest radius of an n-dimensional ball which can be inscribed inside the class $\mathcal{F}$. We know that $b_{n,2} \leq d^n$; see Pinkus (1985), Page 14. When the two quantities are rate-equivalent, least-squares is efficient.

Lemma 2 *If $b_{n,2} \asymp d^n \asymp n^{-\alpha}$ then for the minimax risk among all estimates $R_N^*(\sigma, \mathcal{F})$ we have*

$$R_N^*(\sigma; \mathcal{F}) \asymp R_{LS}^*(\sigma; \mathcal{F}) \asymp (\sigma^2)^{\frac{2\alpha}{2\alpha+1}} \quad as \quad \sigma \to 0.$$

Proof. We use the notation of Donoho, Liu and MacGibbon (1990). Let $b_{n,\infty}$ denote the largest radius of an n-dimensional hypercube that fits inside $\mathcal{F}$, i.e. the largest d such that for some orthonormal set $(\varphi_1, \ldots, \varphi_n)$

$$d \cdot \sum_{i=1}^{n} \pm_i \varphi_i \in \mathcal{F}$$

for all choices of signs $(\pm_i)$. Let $n_0 = \sup\{n : b_{n,\infty} \geq \sigma\}$. Then, by Donoho, Liu and MacGibbon (1990),

$$R_N^*(\sigma; \mathcal{F}) \geq \sigma^2 \frac{1}{2.22} n_0(\sigma).$$

Suppose that $b_{n,2} \asymp n^{-\alpha}$. Then from

$$b_{n,\infty} \leq b_{n,2} \leq \sqrt{n}\, b_{n,\infty}$$

we have $n_0(\sigma) \geq c_1(\sigma^2)^{\frac{-1}{2\alpha+1}}$. Hence

$$\begin{aligned} R_N^*(\sigma) &\geq \sigma^2 c_2(\sigma^2)^{\frac{-1}{2\alpha+1}} \\ &= c_2(\sigma^2)^{\frac{2\alpha}{2\alpha+1}}. \end{aligned}$$

Thus no estimator can have a worst case risk with a better rate of convergence to zero than that which we already know is achieved by the least-squares estimate.

The lemma is immediately applicable whenever the Gel'fand n-widths are established to be of order $n^{-\alpha}$ by a two part argument, where part of the argument invloves using the lower bound $b_{n,2} \leq d^n$, and the other part of the argument involves developing an upper bound on d^n of the same order as $b_{n,2}$. An example of such an argument is in Pinkus (1985) Chapter VII, pp. 234 et seq. where it is used to calculate the Gel'fand n-widths of the Sobolev ball $W^{m,p}$.

Lemma 3 *For* $1 \leq p \leq 2$,

$$b_{n,2}(W^{m,p}) \asymp d^n(W^{m,p}, L_2) \;\; n \to \infty$$

Proof. Evidently, it is sufficient to show that $b_{n,2} \geq cn^{-m}$. Let ϕ be a C^∞ function of compact support in $[0,1]$. Let $\mathcal{F}_n$ denote the set $\{f = \sum_{k=0}^{n-1} \alpha_k \phi_{n,k}(t)\}$, where $\phi_{n,k}(t) = \phi(nt-k)$, $k = 0, \ldots, n-1$. Evidently the $\phi_{n,k}$ are orthogonal; let $\psi_{n,k} = \phi_{n,k}/||\phi_{n,k}||$ be the corresponding orthonormal set. Define the n-dimensional sphere $S_n(d) = \{f : f \in \mathcal{F}_n, ||f||_2 \leq d\}$. Then

$$\begin{aligned} b_{n,2} &= \sup\{d : S_n(d) \subset \mathcal{F}\} \\ &= \sup\{d : \sum \alpha_k^2 \leq d^2 \text{ implies } ||\sum \alpha_k \psi_{n,k}^{(m)}||_p \leq 1\} \end{aligned}$$

On the other hand,

$$||\sum \alpha_k \psi_{n,k}||_p = ||\psi_{n,0}^{(m)}||_p \cdot ||\alpha||_{l_{p,n}}$$

and

$$||\alpha||_{l_{p,n}} \leq n^{1/p-1/2}||\alpha||_{l_{2,n}}$$

Hence

$$b_{n,2} \geq \sup\{d : dn^{1/p-1/2}||\psi_{n,0}^{(m)}||_p\}$$

A simple change of variables gives

$$||\psi_{n,0}^{(m)}||_p = n^{m+1/2-1/p}||\psi^{(m)}||_p$$

and we get $b_{n,2} \geq cn^{-m}$ as desired.

8 Discussion

8.1 Sharpness of n-width bounds

Is the upper bound based on n-widths *sharp* – does it describe the true rate of convergence of the least-squares estimate? This question is intimately connected with the extent to which the Gel'fand numbers provide lower bounds on the exponent in the modulus of continuity for Gaussian processes. A recent result in the Geometry of Banach Spaces, due to Pajor and Tomczak-Jagermann (1986), indicates that Gel'fand numbers may be used to *lower-bound* the expectation of the supremum of a Gaussian process (compare also Pisier, 1989). Sufficiently strong and general results of this kind, adapted for study of the modulus of continuity rather than the supremum, could perhaps be used to show that the bounds from Gel'fand numbers are in some sense sharp. In any case, we know of no example where they fail to be sharp.

8.2 Calculation of Gel'fand n-widths

Gel'fand numbers are not always easy to calculate! The L_2 n-widths of $W^{m,p}$ for $p < 2$ remained for many years an open problem, which turned out differently than many people anticipated. The eventual solution rests on work of Kashin (1977), which introduced the notion of almost-orthogonal decompositions for $l_{1,n}$, a notion that spawned a great deal of work in the Geometry of Banach Spaces. Even this breakthrough did not solve the problem directly, but only through duality relations between Gel'fand numbers and Kolmogorov numbers. For example, the functionals L_i optimal for the Gel'fand n-widths of $W^{m,p}$, $p < 2$, are unknown; their properties have only been inferred probabilistically. The actual calculation of Gel'fand n-widths may therefore lead to deep and difficult, but perhaps interesting problems.

8.3 n-widths and Statistical Estimation

We now know of three kinds of n-widths with important role for statistical estimation: *Bernstein n-widths:* In Donoho, Liu, and MacGibbon (1990), it is shown that the Bernstein n-widths $b_{n,\infty}$ determine, to within logarithmic factors, the optimal rate of convergence, among all measurable estimates, for a certain class of spaces possessing unconditional bases. As Donoho and

Johnstone (1990) show, the n-widths $b_{n,\infty}$ determine the optimal rate of convergence in Besov-spaces.

Kolmogorov n-widths: In Donoho, Liu, and MacGibbon (1990) it is also shown that the Kolmogorov n-widths d_n determine the optimal rate of convergence among all linear estimates over certain "orthosymmetric" function spaces.

Gel'fand n-widths: These bound the rate of convergence of least squares, and they may even determine the optimal rate of convergence among all estimates, for a certain class of Sobolev spaces.

At the very least, the Bernstein n-widths, which refer to all estimates, are more fundamental for orthosymmetric spaces than the Kolmogorov n-widths, which only refer to linear estimates.

In the recent A.N. Kolmogorov Memorial Issue of the *Annals of Statistics* (September 1990), a significant theme of papers by Centsov (1990) and Has'minskii and Ibragimov (1990) is the importance of Kolmogorov n-widths for determining rates of convergence of statistical estimators of orthogonal series type. Although the problems considered are problems of density estimation rather than the white noise model considered here, the basic reasoning is parallel to section 5 of this paper.

The developments of this paper should make clear that orthogonal series estimates, while interesting and useful, are not universally optimal, and so in this sense Kolmogorov n-widths are less fundamental than other notions of "massiveness" of functional classes.

In fact, Kolmogorov himself was not satisfied with the n-widths that bear his name and worked energetically on the notions of ε-entropy and ε-capacity of functional classes in order to get at notions of optimality among all methods of approximation (Kolmogorov, 1956) (Kolmogorov and Tikhomirov, 1959).

These notions, in turn, have borne significant fruit in a statistical setting. The ε-entropy and ε-capacity are intimately related with the covering number notions of section 3 above; and so as we have seen, may be used to obtain continuity properties of Gaussian processes; at another level, these properties may be used to obtain properties of maximum-likelihood estimates. This idea is used in Birgé, Le Cam, van de Geer, and many other papers.

Here we complete the circle by showing that a certain notion of n-width, the Gel'fand n-width, (in some sense dual to the Kolmogorov linear n-width) may often be used to the same end as the covering numbers and lead to an

understanding of optimal nonlinear procedures.

References

[1] Adler, R. (1989) An introduction to Continuity, Extrema, and Related topics for General Gaussian Processes. Manuscript.

[2] Birgé, L. (1983) Approximation dans les espaces métriques et théorie de l'estimation. *Zeit. fur Wahr. und verw. Geb.* **65** 181-237.

[3] Birgé, L. and Massart, P. (to appear)

[4] Carl, B. (1981) Entropy numbers, s-numbers, and eigenvalue problems. *J. Functional Anal.* **41** 290-306.

[5] Centsov, N.N. (1990) The unfathomable influence of Kolmogorov. *Ann. Statist.* **18**.

[6] Donoho, D.L. (1989) Statistical Estimation and Optimal Recovery. Technical Report 213, Department of Statistics, University of California, Berkeley.

[7] Donoho, D.L. (1989) Function estimation and the white noise model. to appear in *Ecole d'Ete de Probabilités 1990 (Saint Flour)*. Springer.

[8] Donoho, D.L., Liu, R.C., and MacGibbon, K.B. (1990) Minimax risk over hyperrectangles, and implications. *Annals of Statistics* **18**, 1416-1437.

[9] Donoho, D.L. and Johnstone, I.M. (1990) Wavelets and optimal nonlinear function estimates. Technical Report. Department of Statistics, University of California, Berkeley.

[10] Dudley, R. (1973) Sample functions of the Gaussian Process. *Ann. Probab.* **1** 66-103.

[11] Has'minskii, R.Z. and Ibragimov, I.A. (1990) On density estimation in view of Kolmogorov's ideas in approximation theory. *Ann. Statist.* **18**.

[12] Kashin, B. (1977) Sections of some finite-dimensional sets and classes of smooth functions. *Izv. Akad. Nauk SSSR* **41** 334-351. (Russian)

[13] Kolmogorov, A.N. (1956) Asymptotic characteristics of some completely bounded metric spaces. *Dokl. Akad. Nauk. SSSR* **108** 585-589.

[14] Kolmogorov, A.N. and Tikhonov, V.M. (1959) ε-entropy and ε-capacity of sets in functional spaces. *Uspekhi Mat. Nauk* **14**, 3-86. (in Russian). *American Math. Soc. Translations* **17** 277-364, (1961). (Engl. Translation).

[15] Le Cam. L. (1974) Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38-53.

[16] Matei (1990) *Journal of Complexity*

[17] Nemirovskii, A.S., Polyak, B.T. and Tsybakov, B.T. (1985) Rates of convergence of nonparametric estimators of maximum likelihood type. *Problems of Information Transmission.* **21** 258-272.

[18] Pajor, A. and Tomczak-Jaegermann, N. (1986) Subspaces of small codimension of finite-dimensional Banach Spaces. *Proc. Am. Math. Soc.* **97** 637-642.

[19] Pinkus, A. (1985) *n-widths in Approximation Theory.* Springer, New York.

[20] Pinkus, A. (1986) n-widths and Optimal Recovery in *Approximation Theory*, Proceedings of Symposia in Applied Mathematics, vol 36, Carl de Boor, Editor. American Math. Soc., Providence, R.I.

[21] Pisier, G. (1989) *The Volume of Convex Bodies and Banach Space Geometry.* Cambridge University Press.

[22] Traub, J.F., Wasilkowski, G.W., and Woźniakowski, H. (1988) *Information-based complexity.* Addison-Wesley, Reading, MA.

[23] Van de Geer, S. (1987) A new approach to least-squares estimation, with applications. *Ann. Statist.* **15** 587-602.