The Annals of Statistics
1989, Vol. 17, No. 1, 58–106

PROJECTION-BASED APPROXIMATION AND A DUALITY WITH KERNEL METHODS

BY DAVID L. DONOHO[1] AND IAIN M. JOHNSTONE[2]

University of California, Berkeley and Stanford University

Projection pursuit regression and kernel regression are methods for estimating a smooth function of several variables from noisy data obtained at scattered sites. Methods based on local averaging can perform poorly in high dimensions (curse of dimensionality). Intuition and examples have suggested that projection based approaches can provide better fits. For what sorts of regression functions is this true? When and by how much do projection methods reduce the curse of dimensionality?

We make a start by focusing on the two-dimensional problem and study the L^2 approximation error (bias) of the two procedures with respect to Gaussian measure. Let RA stand for a certain PPR-type approximation and KA for a particular kernel-type approximation. Building on a simple but striking duality for polynomials, we show that RA behaves significantly better than the minimax rate of approximation for radial functions, while KA performs significantly better than the minimax rate for harmonic functions. In fact, the rate improvements carry over to large classes, RA behaving very well for functions with enough angular smoothness (oscillating slowly with angle), while KA behaves very well for functions with enough Laplacian smoothness, (oscillations averaging out locally). The rate improvements matter: They are equivalent to lowering the dimensionality of the problem. For example, for functions with nice tail behavior, RA behaves as if the dimensionality of the problem were 1.5 rather than its nominal value 2. Also, RA and KA are complementary: For a given function, if one method offers a dimensionality reduction, the other does not.

1. Introduction. The recent introduction of projection pursuit regression (PPR) by Friedman and Stuetzle (1981) [see also Friedman (1985)] expands the growing list of procedures for estimating a smooth function of several variables from scattered, noisy data. Such a list includes kernel regression (KR) [e.g., Stone (1977) and Collomb (1981)], nearest neighbor regression, partitioned or tree structured regression [Breiman, Friedman, Olshen and Stone (1984)], multivariate smoothing splines [e.g., Wahba and Wendelberger (1980)] and fitting of multivariate polynomials and orthogonal series.

PPR algorithms produce a fitted multivariate regression function $\hat{f}$ of the form

$$\hat{f}_n(x) = \sum_{i=1}^{n} \hat{g}_i(\hat{\alpha}_i^t x), \tag{1.1}$$

Received June 1985; revised April 1988.

[1]Supported in part by an NSF Postdoctoral Research Fellowship and in part by the Mathematical Sciences Research Institute, Berkeley.

[2]Supported in part by NSF Grants MCS-80-24649 and DMS-84-51750 and in part by the Mathematical Sciences Research Institute, Berkeley.

AMS 1980 *subject classifications*. Primary 62J02; secondary 62H99, 41A10, 41A25, 42C10.

Key words and phrases. Projection pursuit regression, kernel regression, curse of dimensionality, rates of convergence, Hermite polynomials, harmonic functions, radial functions, angular smoothness, Laplacian smoothness.

where x is a vector, $\{\alpha_i\}$ are unit vectors and each $\alpha^t x$ may be thought of as a projection of x. The ith term $\hat{g}_i(\cdot)$ in the sum is constant along $\alpha_i^t x = c$ and so is often called a ridge function: The estimate at a given point can be thought of as based on averages over certain (adaptively chosen) narrow strips $\{x: |\alpha_i^t x - t_i| \leq \varepsilon\}$. This contrasts with kernel and other local averaging procedures, in which the smoothing is done over small balls of the form $\{x: |x - x_0| \leq \varepsilon'\}$. One of the aims of this paper is to show that, in a certain setting, projection-based and local-averaging based function estimates have complementary properties.

A second notable feature in Friedman and Stuetzle's original discussion of PPR is their suggestion that PPR may be immune, in a certain sense, to the curse of dimensionality. The curse refers to the tendency for nonparametric regression procedures to perform very badly in high dimensions when the sample data are limited, due to the fact that high dimensional space is mostly empty.

Important questions are raised by the wide variety of available procedures. For what sorts of regression functions might we prefer smoothing procedures based on (nonlocal) projections to local averaging methods? Can one identify when and by how much projection methods reduce the curse of dimensionality? These general issues motivated the research reported here, even though the formal results can be presented only in a regrettably restricted setting.

We begin with a concrete example of the contrasting properties of these two averaging schemes. This can be simply stated in terms of a noiseless, approximation theory analog of the estimation problem. Translation of the results to an estimation setting is briefly discussed in Section 9.

Let $f(x, y)$ be a polynomial of degree m in two variables. Let RA (ridge approximation) stand for the scheme of approximating f by a sum of n ridge functions,

$$\hat{f}_n(x, y) = \sum_{i=1}^{n} g_i(x \cos \theta_i + y \sin \theta_i). \tag{1.2}$$

Let KA (kernel approximation) stand for the scheme of approximating f by the convolution

$$\hat{f} = f * K_\sigma$$

where K_σ denotes a kernel function of bandwidth σ (more details are in Section 7).

Measure the approximation error incurred by these two schemes with respect to the $L^2(\Phi_2)$ norm

$$\| f - \hat{f} \|^2 = \int (f - \hat{f})^2 \, d\Phi_2,$$

where Φ_2 is the standard Gaussian measure on $\mathbb{R}^2$, $\Phi_2(dx, dy) = (2\pi)^{-1} \exp\{-(x^2 + y^2)/2\} \, dx \, dy$. With respect to this measure, harmonic polynomials ($\Delta f = (\partial^2/\partial x^2 + \partial^2/\partial y^2) f = 0$) are more difficult to approximate by RA then radial polynomials ($f(x, y) = h(x^2 + y^2)$) of the same degree. (Figure 1 displays radial and harmonic polynomials of degree 8.) Table 1 gives a comparison of the best attainable error by a ridge approximation to harmonic and radial polynomials of degree 16. Evidently, many fewer terms are needed to get a good approximation to the radial than to the harmonic polynomial.

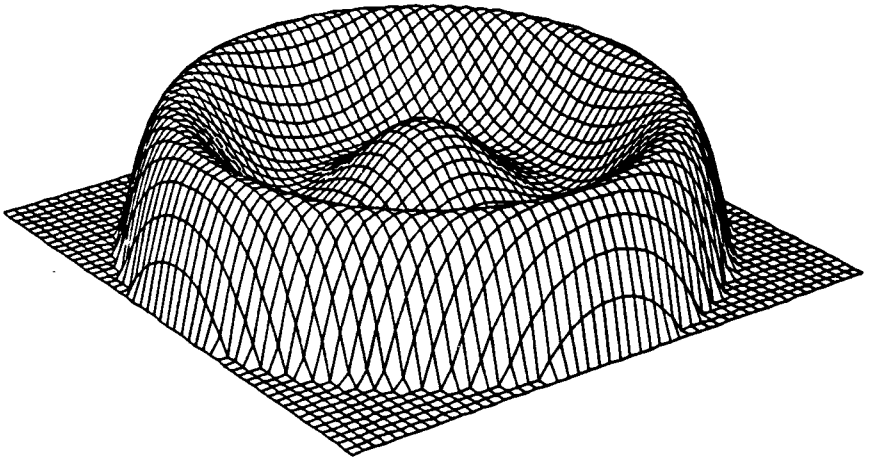

FIG. 1a. *A radial polynomial of degree* 8 (J_{44}) *computed for* $r = (x^2 + y^2)^{1/2} \leq 3.2$. *Reference plane is* $z = J_{44}$ (3.2). (*The remaining* 0*'s of* $\partial/\partial r J_{44}$ *occur at* $r > 3.2$.)

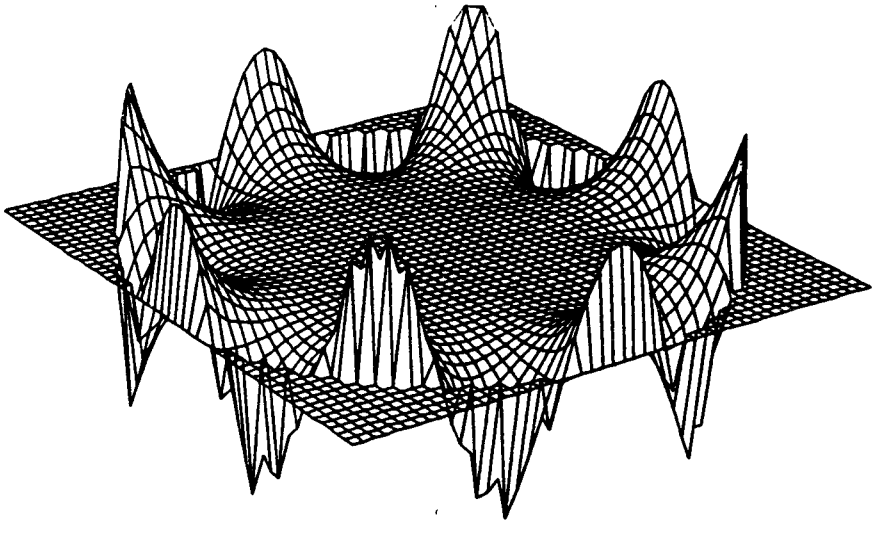

FIG. 1b. *A harmonic polynomial of degree* 8 ($Re J_{8,0}$) *computed for* $r = (x^2 + y^2)^{1/2} \leq 1$. *Reference plane is* $z = 0$.

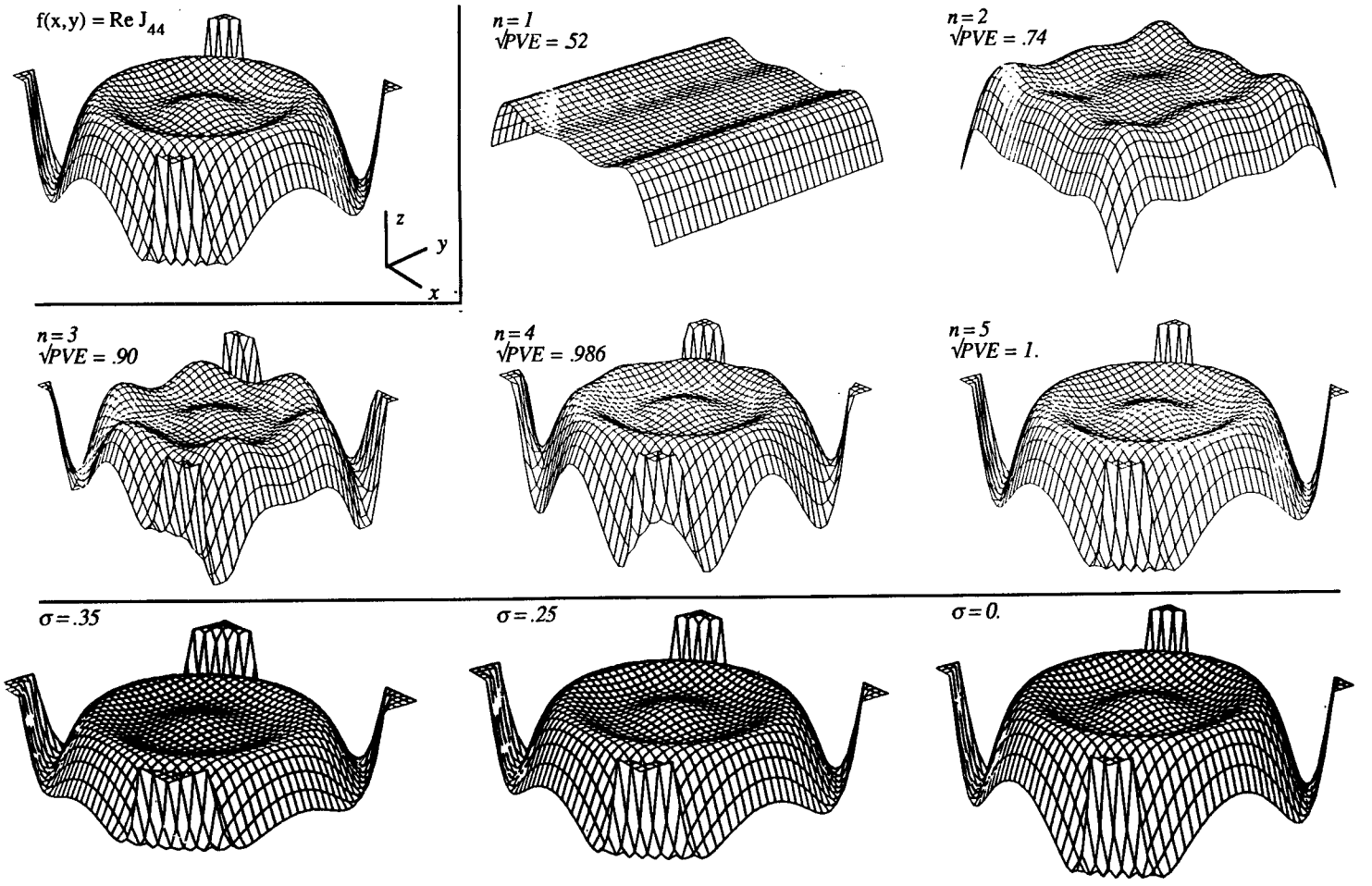

FIG. 1c. *Projection approximation: Radial case.*

Notes: Functions evaluated for $(x, y) \in [-3.4, 34]^2$ *(large values are clipped).* PA *(top): best approximation to 8th degree radial using* $n = 1, 2, 3, 4, 5$

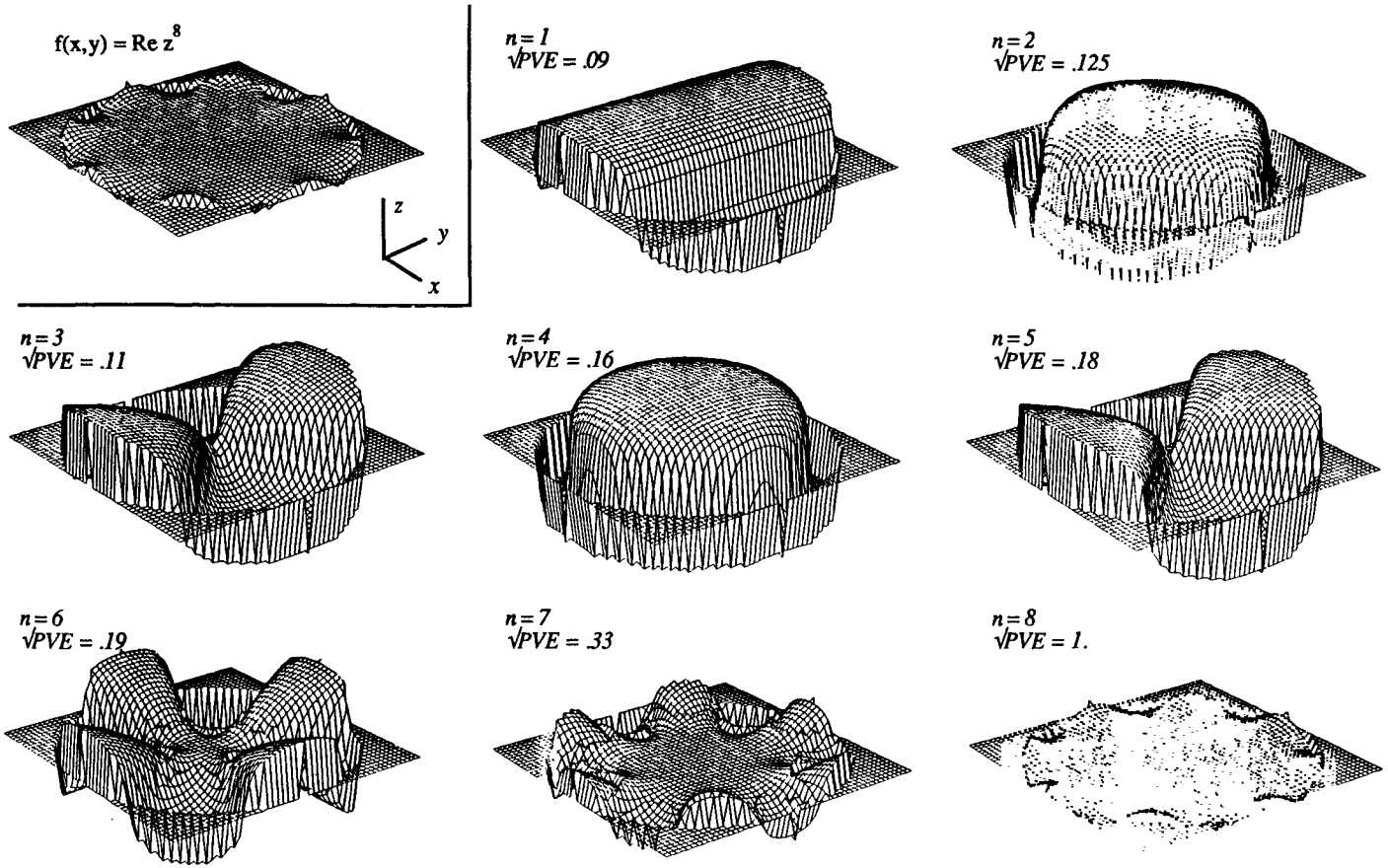

FIG. 1d. *Projection approximation: Harmonic case.*

Notes: *Best approximation to 8th degree harmonic polynomial using* $n = 1, 2, \ldots, 8$ *equally-spaced directions in* $[0, \pi]$. *Functions evaluated for* $(x, y) \in [-1, 1]^2$. *Note that the z coordinate is* $(1 + \log^+(|f|))\ \mathrm{sgn}(f)$ *and is set to* 0 *for* $x^2 + y^2 > 1$.

TABLE 1

Relative error $\|f - \hat{f}_n\|/\|f\|$ *for best ridge function approximation* $\hat{f}_n$ *using* n *equally-spaced directions* $\theta_i = i\pi/n$, $i = 1, \ldots, n$, *to harmonic and radial polynomials of degree* 16 (Re $J_{16,0}$, $J_{8,8}$, *respectively*). [*Notation defined below* (1.2) *and in Section* 4, *computed from* (1.3) *and* (1.4).]

Relative error in RA approximation		
Directions	Harmonic	radial
1	1.000	0.8036
2	0.9999	0.6072
3	0.9999	0.4109
4	0.9999	0.2206
5	0.9998	0.0801
6	0.9998	0.0183
7	0.9998	0.0025
8	0.9998	0.0002
9	0.9998	0.0
10	0.9998	0.0
11	0.9995	0.0
12	0.9989	0.0
13	0.9964	0.0
14	0.9835	0.0
15	0.8824	0.0
16	0.0	0.0

The situation is reversed for KA. Table 2 gives numbers corresponding to those of Table 1 for smoothing with the Gaussian kernel Φ_σ of bandwidth σ. Evidently, harmonics suffer no bias under this form of smoothing, whereas for radials the error is substantial for large σ.

The story is graphed in Figures 1a and 1b for the degree eight polynomials shown in figures 1a and 1b. The radial polynomial can be exactly represented as a sum of five ridge functions, but the approximation in terms of root mean-square error (or root percent variance explained) is quite good already for two and three directions. By contrast, for the harmonic polynomial of Figure 1c, approximation is dismal even for seven directions, while exact representation must occur for this degree 8 polynomial with eight directions. The situation is reversed for kernel approximation: The harmonic is reproduced at all bandwidths because of the mean value approximation and so is not shown. The radial polynomial is flattened by smoothing, leading to relatively large root mean-square errors. Note however that its qualitative shape is preserved: This point is discussed further in Section 9.

This contrasting behavior of the two approximation methods holds for general functions. To explain this, we use a complete system of orthogonal polynomials $\{J_{kl}\}$ for the bivariate functions in $L^2(\Phi)$. For $l \geq k$, $J_{kl}(r, \theta) = \gamma_k e^{i(l-k)\theta} r^{l-k} L_k^{l-k}(r^2/2)$, where L_k^α is a degree k Laguerre polynomial [e.g., Szegö (1939)] and $\gamma_k = (-2)^k k!$ The basis element J_{kl} has radial oscillation $k \wedge l$ and angular oscillation $k - l$, for a total oscillation equal to its degree $k + l$. Here

TABLE 2

Relative error $\|f - \hat{f}_\sigma\|/\|f\|$ *for kernel approximation* $\hat{f}_\sigma = f * \Phi_\sigma$, *using a Gaussian measure with bandwidth* σ *for harmonic and radial polynomials of degree* 16 ($\mathrm{Re}\, J_{16,0}$, $J_{8,8}$, *respectively*). *Computed from formula* (7.0).

Relative error in kernel approximation		
Bandwidth	Harmonic	Radial
0.05	0	0.0004
0.10	0	0.0064
0.15	0	0.0326
0.20	0	0.1044
0.25	0	0.2622
0.30	0	0.5715
0.35	0	1.1478
0.40	0	2.2070
0.45	0	4.1732
0.50	0	7.9059

radial oscillation refers to the number of 0's of J_{kl} along a ray from the origin and angular oscillation to (half) the number of 0's encountered in circling about the origin at fixed radius. As a result, the harmonic (respectively, radial) polynomials correspond to basis elements J_{kl} with no radial (angular) oscillation, $k \wedge l = 0$ ($k - l = 0$).

Figures 2a and 2b portray the relative error in approximating a basis element J_{kl} by RA and KA, respectively. A glance shows that (at least for the tuning constants used to produce the figure) RA works well for those J_{kl} with $k - l$ and $k + l$ small and KA works well for those J_{kl} where kl is small. In short, if f is a function whose J_{kl} expansion

$$f \sim \sum c_{k,l} J_{k,l}$$

is concentrated at low angular oscillation, then RA would appear to be well suited to approximating f; whereas, if f 's expansion concentrates at low radial oscillation, one expects KA to do well.

This paper develops mathematical results to document this complementarity of RA and KA. For simplicity, we start with polynomials. The results are then stated for more general functions in terms of rates of convergence and smoothness classes. Abstractly the situation is as in Figure 3, which depicts the space $\mathscr{F}_p$ of functions p-times differentiable in $L^2(\Phi)$ quadratic mean. Let θ be the angle parameter in polar coordinates. The subspace $\mathscr{A}_{pq}$ consists of angularly smooth functions (those $f \in F_p$ which are $q > p/2$ times $\partial/\partial\theta$-differentiable in quadratic mean). Its members in an appropriate sense have a J_{kl} expansion concentrated near basis elements with $k - l = 0$. The subspace $\mathscr{L}_{pr}$ consists of functions with Laplacian smoothness (those $f \in \mathscr{F}_p$ which are $r > p/2$ times Δ-differentiable in quadratic mean). Its members in an appropriate sense have a J_{kl} expansion concentrated near $k \wedge l = 0$. As forecasted by the heuristics above, PA can be

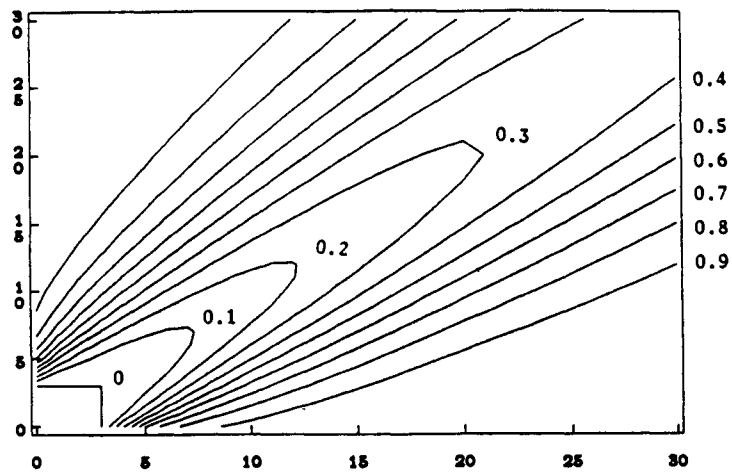

FIG. 2a. RA *approximation error. Contour plot of relative approximation error* $\|\hat{J}_{kl} - J_{kl}\|/\|J_{kl}\|$, *where* $\hat{J}_{kl}$ *is best approximation to* J_{kl} *using ridge functions in* $n = 4$ *equally-spaced directions,* $\theta_i = \pi i/4$. *(Computed from Lemma* 6.2.) *Approximation is best for polynomials that are close to radial* ($|k - l|$ *small*).

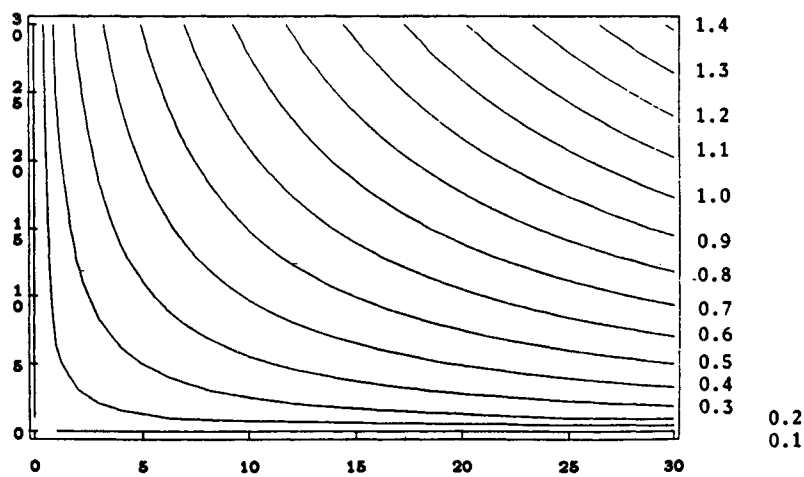

FIG. 2b. KA *approximation error. Contour plot of relative approximation error* $\|\hat{J}_{kl} - J_{kl}\|/\|J_{kl}\|$, *where* $\hat{J}_{kl} = J_{kl} * \Phi_\sigma$ *where* Φ *denotes Gaussian measure on* R^2 *and the bandwidth* $\sigma = 0.2$. [*Computed from* (7.0).] *Approximation is best for polynomials that are close to harmonic* ($k \wedge l$ *close to* 0).

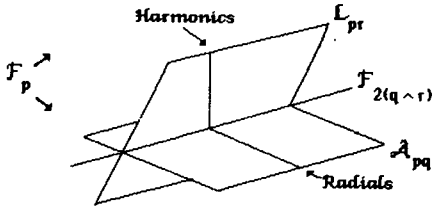

FIG. 3. *The relation between function classes.*

tuned to achieve a better rate of approximation uniformly over (spheres in) $\mathcal{A}_{pq}$ than over $\mathcal{F}_p$, while KA can be tuned to achieve a better rate over (spheres in) $\mathcal{L}_{pr}$ than over all $\mathcal{F}_p$.

The rate improvements are significant: Over $\mathcal{A}_{pq}$, RA behaves in a certain sense as if the dimensionality of the problem were 1.5 instead of 2; over $\mathcal{L}_{pq}$, KA behaves as if the dimensionality were 1. Furthermore, these rate improvements are essentially best possible, in an appropriate sense.

The notion of complementarity can also be expressed as follows. The typical function in $\mathcal{A}_{pq}$ does not belong to $\mathcal{L}_{pr}$ and vice versa: One cannot expect to get a rate improvement over the minimax rate simultaneously for KA and RA.

Contents. The paper has 10 sections. Sections 2 and 3 develop tools for discussing approximation to harmonic and radial functions; Sections 4–7 develop the J_{kl} basis, notions of smoothness classes, rate improvements, and complementarity and Sections 8–10 interpret the results, issuing various warnings and suggestions for further work. The remainder of the introduction surveys the contents of the technical sections, Sections 2–7: For an overview, the reader might then proceed directly to Section 8.

The paper is based on L^2 approximation theory with respect to Gaussian measure. The price of this trade of generality for tractability is paid in the discussions of Sections 9 and 10.

Section 2 considers computations of best $L^2(\Phi)$ approximations by sums of ridge functions. For a fixed set of ridge directions, there is an explicit procedure for computing such best approximations. By expanding each of the ridge functions occurring in the best approximation in terms of Hermite polynomials, one obtains a sequence of finite-dimensional least-squares problems for the best approximation. When the set of directions is equally spaced, the eigenstructure of these problems is especially simple, arising from circulant matrices.

Section 3 applies these results to compute the best approximation to both radial and harmonic functions using equally spaced ridge directions. Discrete Fourier analysis enters, and the formulas for $L^2(\Phi)$ approximation error in these cases involve certain binomial probabilities. In the following, $a \equiv b\ [m]$ stands for "$a = b$ modulo m."

PROPOSITION 1.1. *Let f be a polynomial of degree m orthogonal to all polynomials of lower degree. Let $\hat{f}_n$ denote the best $L^2(\Phi)$ approximation to f using sums of ridge functions in the n ($\leq m$) directions $\theta_j = \theta_0 + j\pi/n$, $j =$*

$0, \ldots, n-1$, *as* θ_0 *varies in* $[0, \pi/n)$. *Let* S_m *denote the simple symmetric* (± 1) *random walk on the integers. If* f *is radial, then*

$$\|\hat{f}_n\|^2/\|f\|^2 = P(S_m = 0)/P(S_m \equiv 0\,[2n]). \tag{1.3}$$

If f *is harmonic, then*

$$\|\hat{f}_n\|^2/\|f\|^2 = P(S_m = \pm m)/P(S_m \equiv \pm m[2n]). \tag{1.4}$$

Examination of these probabilities shows, as in Table 1, that while harmonics are difficult to approximate using sums of ridge functions, radials are relatively easy to approximate. Because of the link to binomial probabilities, simple asymptotic analysis using the Gaussian limit and Hoeffding's inequality yields rates of approximation.

PROPOSITION 1.2. *Let* $R_{n,m}f$ *denote the best approximation to* f *by a sum of* n *ridge polynomials (in equally spaced ridge directions) each of degree less than or equal to* m. *Let* $N = nm + 1$ *denote the total complexity of* $R_{n,m}f$. *If* f *is radial and has* p *Cartesian derivatives in* $L^2(\Phi)$ *quadratic mean, then* n *and* $m(n)$ *can be chosen so that*

$$\|R_{n,m}f - f\|^2 = O(N^{-p/1.5})$$

as $n \to \infty$. *If* f *is harmonic and has* p *Cartesian derivatives in* $L^2(\Phi)$ *quadratic mean, then* n *and* m *can be chosen so that*

$$\|R_{n,m}f - f\|^2 = O(N^{-p/2})$$

and there exist such f *for which*

$$\liminf_{n\to\infty} N^{p/2}\|R_{n,m}f - f\|^2 > 0,$$

regardless of the choice of n, m.

Since $N^{p/2}$ is the minimax approximation rate for p-smooth functions, the result says that RA can improve on the minimax rate for smooth functions which are *radial*, but not for those which are *harmonic*. Propositions 1.1 and 1.2 are proved in Section 3.

To prepare for extension of these results to more general classes of functions, Section 4 develops the properties of the J_{kl} basis for $L^2(\Phi)$. It gives the irreducible representations of the orthogonal group in $L^2(\Phi)$. In this basis, harmonic and radial functions play distinguished roles as extreme elements in various senses. Differentiation operators such as ordinary partial derivatives, the Laplacian and angular differentiation all have convenient multiplier representations in this basis, and so various Sobolev-type measures of smoothness are expressible as moments of the coefficients of a function in this basis. In addition, the projections of a function are simple to compute in terms of its expansion in this basis. In this way, the basis helps solve a number of basic technical problems arising in the study of RA.

This machinery allows one to define quantitative notions of smoothness. Section 5 formalizes the Cartesian, angular and Laplacian smoothness classes $\mathcal{F}_p$,

$\mathscr{A}_{pq}$ and $\mathscr{L}_{pr}$ mentioned earlier and establishes their complementarity properties.

The smoothness class $\mathscr{A}_{pq}$ is the focus in Section 6, which contains our main theorem for RA. In particular, if $f \in \mathscr{A}_{pq}$, n and $m(n)$ can be chosen so that

$$\|R_{n,m} f - f\| = O\left((N/\log^2 N)^{-p/1.5}\right).$$

In general, however, the rate RA can attain over $\mathscr{F}_p$ is no better than $N^{-p/2}$. Thus RA (when well tuned) converges at a faster rate on $\mathscr{A}_{pq}$ than on $\mathscr{F}_p$.

Section 7 gives some KA counterparts to these results. Letting $N = (\sigma^{-2})^d$ denote the complexity of $f * \Phi_\sigma$, we have, if f is harmonic, from the mean value property of harmonic functions:

$$\|\Phi_\sigma * f - f\| = 0,$$

while if f is radial and has two derivatives in mean square

$$\|\Phi_\sigma f - f\| = O(N^{-2/2}).$$

Again, these facts generalize: If $f \in \mathscr{L}_{pp}$, there is a kernel $K^{(p)}$ so that

$$\|K_\sigma^{(p)} * f - f\| = O(N^{-p/1}),$$

while on $\mathscr{F}_p$, in general, one can expect no better than

$$\|K_\sigma^{(p)} * f - f\| = O(N^{-p/2}).$$

Together Sections 6 and 7 display the complementary roles of the subspaces $\mathscr{L}_{pr}$ and $\mathscr{A}_{pq}$ for KA and RA.

Section 8 discusses interpretations of this work. From the viewpoint of minimax rates of convergence, one expects in the present setting to have a best squared approximation error for p-smooth functions on the order $N^{-p/d}$, where d is the dimension of the independent variables. Therefore, when the rate is actually $r^* > p/d$ for a function f which is p-smooth but not more than p-smooth, one can speak of a rate improvement over the minimax rate and an effective dimensionality $d^* < d$. Denoting $d^*(\mathscr{M}_p, P)$ the denominator in the worst rate of convergence p/d^* of the approximation error by method P over the set $\mathscr{M}_p$, we have

$$d^*(\mathscr{A}_{pp}, \mathrm{RA}) = 1.5,$$
$$d^*(\mathscr{F}_p, \mathrm{RA}) = d^*(\mathscr{L}_{pr}, \mathrm{RA}) = 2,$$

while

$$d^*(\mathscr{L}_{pp}, \mathrm{KA}) = 1,$$
$$d^*(\mathscr{F}_p, \mathrm{KA}) = d^*(\mathscr{A}_{pq}, \mathrm{KA}) = 2.$$

Thinking of d^* as a performance index, RA appears to be designed (i.e., performs well) for functions in $\mathscr{A}_{pq}$, while KA appears to be designed for functions in $\mathscr{L}_{pr}$. This suggests some broad analogies with Bayes and minimax problems in simultaneous estimation.

Section 9 discusses difficulties with overly broad interpretations of Section 8 and its predecessors. Chief amongst these is sensitivity of $L^2(\Phi)$ measures of

smoothness to boundary behavior. This is brought on by noncompactness of the support of Φ. Furthermore, the approximation theory results have to be translated into results on estimation in the case of noisy data obtained at irregular sites. If the measure on X is far from Gaussian (e.g., uniform on the unit disc), it seems that interesting and different behavior occurs, requiring further study. Section 10 closes the paper briefly by discussing the relation of these results to tomography and other work and areas for further research.

An informal summary without proofs of the results in this paper appears in Donoho and Johnstone (1986).

Finally, we collect some notation for later use. Let $\mathbf{t}, \boldsymbol{\alpha}, \mathbf{x}, (\mathbf{k})$, etc., denote vectors in $\mathbb{R}^d$ or $\mathbb{Z}^d$; we use multiindex notation $\mathbf{t}^{\mathbf{k}} = t_1^{k_1} t_2^{k_2} \cdots t_d^{k_d}$, $\mathbf{k}! = k_1! k_2! \cdots k_d!$, $|\mathbf{k}| = k_1 + \cdots + k_d$, $\binom{m}{\mathbf{k}} = m!/(k_1! k_2! \cdots k_d!)$ and so on. The Kronecker delta sets: $\delta_l^k = 1$ if $k = l$, 0, otherwise, and $\delta_{\mathbf{l}}^{\mathbf{k}} = 1$ if $k_i = l_i$ for all i, 0, otherwise. $a \pm b[n]$ means a is congruent to b modulo n. For partial derivatives of a bivariate function $D_x^i f = \partial^i f(x, y)/\partial x^i$, $D_y^i f = \partial^j f(x, y)/\partial y^j$, and for d-variate functions $D^{\mathbf{k}} f = D_{x_1}^{k_1} D_{x_2}^{k_2} \cdots D_{x_d}^{k_d} f$. The Laplacian of f is $\Delta f = \sum_i^0 D_{x_j}^2 f = \sum_i \partial^2 f/\partial x_i^2$. Finally, $C(\alpha, \beta, \ldots, f, \ldots)$ denotes a constant whose value depends only on the arguments $\alpha, \beta, \ldots, f, \ldots$.

2. Best approximation by sums of ridge functions. The best approximation error obtainable by the projection pursuit scheme is given by the function f_n of the form

$$f_n(X) = \sum_{i=1}^{n} g_i(\alpha_i^t X), \tag{2.0}$$

minimizing $\| f - f_n \|^2$, where, as remarked in the introduction, $\| \cdot \|$ represents the norm of $L^2(\Phi_2)$. In general, this is a difficult problem because it depends in a highly nonlinear way on the directions θ_i. However, when the directions are fixed, the problem reduces to a least-squares problem. Because we are working in $L^2(\Phi_2)$, the solution has a convenient representation in terms of Hermite polynomials: The Hermite expansion of each of the g_i can be found by solving one least-squares problem of rank d for each degree d in the expansion of g_i. In the special case where the directions are equally spaced, each of these least-squares problems has a particularly convenient structure arising from the circulant nature of its design matrix.

This section has three parts. The first reviews basic facts about Hermite expansions and projections. The second describes the solution to the approximation problem just stated. The third specializes to the case where the directions are equally spaced.

2.1. *Hermite polynomial expansions.* The Hermite polynomials of a single Gaussian variable form a complete orthogonal set in $L^2(\mathbb{R}^1, \Phi_1)$ and we use the definition

$$H_m = \frac{1}{\phi_1}(-D)^m \phi_1, \qquad m \geq 0, \tag{2.1}$$

where D denotes differentiation and $\phi_1(x) = (2\pi)^{-1/2}e^{-x^2/2}$. They satisfy the orthogonality relations

$$\langle H_k, H_l\rangle = E\big[H_k(X)H_l(X)\big] = k!\delta_l^k, \qquad k, l \geq 0, \tag{2.2}$$

where $\delta_l^k = 1$, if $k = l$, and 0, otherwise.

Suppose now that Φ_d is standard Gaussian measure on $\mathbb{R}^d$ with identity covariance matrix and mean 0. Since the components of $\mathbf{X} = (X_1, \ldots, X_d) \sim \Phi_d$ are independent, a natural choice for a complete orthogonal basis for $L^2(\Phi_d)$ is

$$H_{\mathbf{k}}(\mathbf{x}) = \prod_{i=1}^{d} H_{k_i}(x_i), \qquad \mathbf{k} = (k_1, \ldots, k_d) \in Z_+^d.$$

The generating function for $\{H_{\mathbf{k}}\}$ is

$$\exp\big\{-1/2\big(|\mathbf{t}|^2 - 2\mathbf{t}\cdot\mathbf{x}\big)\big\} = \sum_{\mathbf{k}} \mathbf{t}^{\mathbf{k}} H_{\mathbf{k}}(\mathbf{x})/\mathbf{k}!, \qquad \mathbf{t} \in \mathbb{C}^d, \mathbf{x} \in \mathbb{R}^d,$$

and, because of independence, the obvious analog of (2.2) holds:

$$EH_{\mathbf{k}}(\mathbf{X})H_{\mathbf{l}}(\mathbf{X}) = \delta_{\mathbf{l}}^{\mathbf{k}}\mathbf{k}! = \delta_{\mathbf{l}}^{\mathbf{k}} \prod_{i=1}^{d} k_i!. \tag{2.3}$$

The space $\mathscr{P}_m$ of polynomials in d variables of degree at most m is spanned by the basis elements $H_{\mathbf{k}}(\mathbf{x})$ of degree $|\mathbf{k}| = \sum_{i=1}^d k_i \leq m$. For $m \geq 1$, define $\mathbb{P}_m^0$ to be the orthogonal complement (with respect to the inner product $\langle f, g\rangle = \int f\bar{g}\, d\Phi_d$) of $\mathscr{P}_{m-1}$ in $\mathscr{P}_m$ and let $\mathscr{P}_0^0$ be the space of constant functions. The orthogonality relations (2.3) show that $\mathscr{P}_m^0$ consists of polynomials $\sum_{\mathbf{k}} a_{\mathbf{k}} H_{\mathbf{k}}$ for which $a_{\mathbf{k}} \neq 0$ only if $|\mathbf{k}| = m$. Hence P_m^0 is identical to the space of pure or homogeneous Hermite polynomials of degree m (sometimes called Wick polynomials in the statistical mechanics literature). An arbitrary function $f \in L^2(\Phi_d)$ has an L^2-convergent Hermite expansion

$$f = \sum c_{\mathbf{k}} H_{\mathbf{k}}, \qquad c_{\mathbf{k}} = Ef H_{\mathbf{k}}/\mathbf{k}!.$$

By grouping all terms of pure Hermite degree m into a degree m polynomial $f_{(m)} \in \mathscr{P}_m^0$, we may also write

$$f = \sum_m f_{(m)}.$$

Hence $L^2(\Phi_d)$ splits into the direct sum of orthogonal subspaces $\mathscr{P}_m^0$ and so also

$$\|f\|^2 = \sum_m \|f_{(m)}\|^2. \tag{2.3'}$$

A well-known identity expresses a ridge polynomial of pure Hermite degree m in terms of the basis elements $H_{\mathbf{k}}(\mathbf{x})$ of pure degree m. We have

$$H_m(\boldsymbol{\alpha}\cdot\mathbf{x}) = \sum_{|\mathbf{k}|=m} \binom{m}{\mathbf{k}} \boldsymbol{\alpha}^{\mathbf{k}} H_{\mathbf{k}}(\mathbf{x}) \tag{2.4}$$

and its special cases

$$H_m(x\cos\theta + y\sin\theta) = \sum_{k=0}^{m} \binom{m}{k} \cos^k\theta \sin^{m-k}\theta H_k(x)H_{m-k}(y). \tag{2.5}$$

The proofs can exploit either the generating functions or the definition (2.1) [e.g., McKean (1973), page 200].

A basic projection formula. If a single direction $\boldsymbol{\alpha}$ is to be used in (2.0), then the best ridge function approximation to $f(\mathbf{x}) \in L^2(\Phi)$ in direction $\boldsymbol{\alpha}$ is $E[\, f(\mathbf{X}) | \boldsymbol{\alpha} \cdot \mathbf{X} = \boldsymbol{\alpha} \cdot \mathbf{x}]$. This connection between ridge approximation and conditional expectation indicates the usefulness of the following formula that is quite special to the Gaussian measure, but see also Davison and Grunbaum [(1981), Section 4].

LEMMA 2.1. $E[H_{\mathbf{k}}(\mathbf{X}) | \boldsymbol{\alpha}\mathbf{X}] = \boldsymbol{\alpha}^{\mathbf{k}} H_{|\mathbf{k}|}(\boldsymbol{\alpha}\mathbf{X})$, $|\boldsymbol{\alpha}| = 1$, $\mathbf{k} \in Z_+^d$.

Since $H_{\mathbf{k}}(\boldsymbol{\alpha} \cdot \mathbf{X})$ is a polynomial in $\mathscr{P}_m^0$, the result says that conditional expectation carries $\mathscr{P}_m^0$ into $\mathscr{P}_m^0$. It allows us to analyze the actions of projections on each subspace separately. For the proof simply evaluate the conditional expectation

$$E\left[\exp\left\{-\tfrac{1}{2}(|\mathbf{t}|^2 - 2\mathbf{t} \cdot \mathbf{X})\right\} | \boldsymbol{\alpha} \cdot \mathbf{X}\right] \tag{2.6}$$

of the generating function of $\{H_{\mathbf{k}}\}$ and equate terms in the resulting series expansion.

COROLLARY 2.2. *For* $d = 2$,

$$E\left[H_k(X)H_l(Y) | X\cos\theta + Y\sin\theta\right] = \cos^k\theta \sin^l\theta H_{k+l}(X\cos\theta + Y\sin\theta).$$

2.2. *Best approximation by sums of ridge functions.* Let $f(\mathbf{x}) \in L^2(\Phi_d)$ and consider the best ridge function approximation $g(\mathbf{x})$ having the form (2.0) using n directions $\{\theta_0, \ldots, \theta_{n-1}\}$. Quality of approximation is measured by mean-square error with respect to standard Gaussian measure on $\mathbb{R}^2$. Both $f = \Sigma f_{(k)}$ and $g = \Sigma g_{(k)}$ have decompositions into sums of polynomials of pure Hermite degree k [and each $g_{(k)}$ has the form $g_{(k)}(\mathbf{x}) = \sum_{i=0}^{n-1} c_{i,k} H_k(\theta_i \cdot \mathbf{x})$, with the constants $\{c_{i,k}\}$ to be determined]. From (2.3′) we find

$$\|f - g\|^2 = \sum_{k=0}^{\infty} \|f_{(k)} - g_{(k)}\|^2,$$

so that the least-squares problem decomposes into a collection of simpler least-squares problems, one for each degree k.

We may, therefore, focus on a fixed degree, m say, and will now also restrict to $d = 2$ dimensions. We abuse notation a little by setting $\theta_i = (\cos\theta_i, \sin\theta_i)$ and $\mathbf{x} = (x, y)$. The problem is now to fit to $f \in \mathscr{P}_m^0$ a function g of the form

$$g(x, y) = \sum_{i=0}^{n-1} c_i H_m(x\cos\theta_i + y\sin\theta_i), \tag{2.7}$$

where the constants c_i are to be determined by least-squares. Write $h_i = h_i(x, y)$ for $H_m(x\cos\theta_i + y\sin\theta_i)$: Then least-squares solutions $\hat{c}_i$ are obtained from the normal equations

$$\sum_j \langle h_i, h_j \rangle \hat{c}_j = \langle h_i, f \rangle, \tag{2.8}$$

where $\langle f, g\rangle = \int f\bar{g}\, d\Phi_2$. Let us compute $\langle h_i, h_j\rangle = EH_m(\theta_i \mathbf{X})H_m(\theta_j \mathbf{X})$. Since $\mathscr{L}(\mathbf{X})$ is orthogonally invariant, we may equally well replace θ_i and θ_j by $\mathbf{e}_1 = (1,0)$ and $\theta_{j-i} = (\cos(\theta_j - \theta_i), \sin(\theta_j - \theta_i))$, respectively. Thus primed, we can apply the projection lemma via Corollary 2.2 to find

$$\begin{aligned}\langle h_i, h_j\rangle &= E\{H_m(\theta_{j-i}\cdot \mathbf{X})EH_m(X_1|\theta_{j-i}\cdot \mathbf{X})\}\\ &= \cos^m(\theta_j - \theta_i)EH_m^2(\theta_{j-i}\cdot \mathbf{X})\\ &= m!\cos^m(\theta_j - \theta_i).\end{aligned}$$

Thus an important role is played by the $n \times n$ information matrix $\Gamma = (\Gamma_{ij}^{(m)}) = (\cos^m(\theta_i - \theta_j))$.

2.3. *Equally spaced directions: The spectrum of* $\Gamma^{(m)}$. When the directions are equally spaced, we may take $\theta_i = i\pi/n$, for $i = 0, 1, \ldots, n-1$. [Of course, θ' and $\pi + \theta'$ correspond to the same direction, so it suffices to pick $\theta_i \in [0, \pi)$.] Thus $\Gamma_{ij} = \cos^m((i-j)\pi/n)$.

Hence $\Gamma_{ij} = z(i-j)$ with $z(j) = \cos^m(j\pi/n)$ and $z(j+n) = (-1)^m z(j)$. So Γ is a circulant matrix when m is even. To unify the treatment of the cases m odd and m even, introduce the $2n \times 2n$ matrix,

$$\tilde{\Gamma} = (\Gamma_{ij}) = \cos^m\left((i-j)\frac{\pi}{n}\right), \qquad 0 \le i, j \le 2n-1.$$

Note that

$$\tilde{\Gamma} = \begin{pmatrix} \Gamma & (-1)^m\Gamma \\ (-1)^m\Gamma & \Gamma \end{pmatrix}$$

and that $\tilde{\Gamma}$ is a circulant for *all* values of m. Now observe that if $(\gamma^t, (-1)^m\gamma^t)$ is an eigenvector of $\tilde{\Gamma}$ with eigenvalue λ, then γ is an eigenvector of Γ with eigenvalue $\lambda/2$.

Let $\lambda_k = 2\pi k/2n$ denote the kth Fourier frequency associated with $k \in \mathbf{Z}_{2n} = \{0, 1, \ldots, 2n-1\}$. Being circulant, $\tilde{\Gamma}$ has eigenvalues given by the discrete Fourier transform on $\mathbf{Z}_{2n}$ of $\{z(j)\}$,

$$\hat{z}(\lambda_k) = \sum_{j=0}^{2n-1} e^{-i\lambda_k j}\cos^m(j\pi/n),$$

and has corresponding eigenvectors

$$\tilde{\mathbf{v}}_k = (v_k(j))_{j=0}^{2n-1} = (e^{-i\lambda_k j}); \tag{2.9}$$

see, e.g., Brillinger [(1981), page 73]. Let us apply this to the eigenstructure of Γ. Since $\tilde{v}_k(j+n) = (-1)^k\tilde{v}_k(j)$, we may write $\tilde{\mathbf{v}}_k = (\mathbf{w}_k, (-1)^k\mathbf{w}_k)$. Therefore, $\mathbf{w}_k$ is an eigenvector of Γ for those $k \in \{0, 1, \ldots, 2n-1\}$ for which $(-1)^k = (-1)^m$. For such k,

$$\mathbf{w}_k = (w_k(j))_{j=0}^{n-1} = (e^{-i\lambda_k j})$$

are eigenvectors for Γ corresponding to eigenvalues

$$\zeta_k^{(m)} = \hat{z}(\lambda_k)/2 = \sum_{j=0}^{n-1} e^{-i\lambda_k j}\cos^m(j\pi/n).$$

It is helpful to think of $\zeta_k^{(m)}$ in terms of a simple symmetric random walk on $\mathbf{Z}_{2n}$, in particular the copy Λ of $\mathbf{Z}_{2n}$ having elements $\{\lambda_0, \lambda_1, \ldots, \lambda_{2n-1}\}$. Let $\mathbf{p}$ denote the probability distribution assigning mass $\frac{1}{2}$ to $\lambda_{\pm 1}$ (of course, all addition is now modulo $2n$) and write $\mathbf{p}^{*m}$ for m fold convolution of $\mathbf{p}$. Write $w^\vee(j) = (2n)^{-1}\sum_{k=0}^{2n-1} e^{i\lambda_k j} w(\lambda_k)$ for the (inverse) Fourier transform of a function defined on Λ: The point is that $\cos j\pi/n = 2n\mathbf{p}^\vee(j)$ and hence $z(j) = \cos^m(j\pi/n) = 2n(\mathbf{p}^{*m})^\vee(j)$. By the inversion theorem, therefore,

$$\tilde{z}(\lambda_k) = 2n\mathbf{p}^{*m}(\lambda_k).$$

In terms of independent, fair, ± 1 Bernoulli variables $U_1, \ldots, U_m$ with sum S_m, we have

$$\tilde{z}(\lambda_k) = 2nP(S_m \equiv k[2n]).$$

Note that for m odd (even) $\mathbf{p}^{*m}$ (and S_m) is supported on odd (even) values of $\mathbf{Z}_{2n}$, so that $\tilde{z}(\lambda_k)$ vanishes for all even (odd) values of k,

Return now to the spectrum of Γ: $\{\zeta_k^{(m)}\}$ are proportional to the probabilities of those points $k \in \mathbf{Z}_{2n}$ with the same (odd/even) parity as m. Hence

$$\operatorname{rank} \Gamma = |\operatorname{supp} S_m| = (m+1) \wedge n.$$

The proposition below summarizes the discussion.

PROPOSITION 2.3. *The $n \times n$ matrix $\Gamma^{(m)} = (\cos^m((i-j)\pi/n))$ has eigenvalues $\zeta_k^{(m)} = nP(S_m \equiv k\ [2n])$, where k runs over the n integers in $\{0, 1, \ldots, 2n-1\}$ for which $k \equiv m \bmod 2$. Clearly $\zeta_k^{(m)} = \zeta_{-k}^{(m)}$ and the corresponding eigenspace is spanned by the vectors $\mathbf{u}_k = (u_k(j))_{j=0}^{n-1} = (\cos \lambda_k j)$ and $\mathbf{v}_k = (v_k(j))_{j=0}^{n-1} = (\sin \lambda_k j)$. These spaces are two-dimensional unless $k = 0$ or n, in which case $\mathbf{v}_k$ degenerates to 0.*

3. Best ridge function approximation of radial and harmonic functions. This section applies the tools just developed to the analysis of radial and harmonic functions. These two classes of functions, as suggested in the introduction, represent extreme cases for approximation by sums of ridge functions. The section has two parts. The first describes best approximation to radial and harmonic polynomials, while the second develops asymptotic tools to discuss approximation to general radial and harmonic functions. Then follows the result discussed in the introduction, namely that the effective dimensionality of radials is only 1.5 for RA, while for harmonics it is the same as the nominal value 2.

Although the section title refers to best approximations, we actually compute only best *equispaced* approximations. In fact, we believe these to be best approximations for radials and harmonics in $\mathscr{P}_{2^k}^0$, but we do not prove this here. (However Davison and Grunbaum [(1981), Section 9] contains a counterexample in a related but different setting.) The motivation for studying harmonic and radial functions only becomes clear in Section 4. We will make a few forward references to results there.

3.1. *Approximation of polynomials.* Consider first radial functions $f(x) = g(|x|)$ and use n directions equally spaced about the circle: Without essential

loss, we shall take $\theta_i = i\pi/n$, $i = 0, 1, \ldots, n-1$. These directions remain fixed through the rest of this section. In view of the discussion of Section 2.2, we will first study the case where $f(x, y) \in \mathscr{P}_m^0$. Note that for f to be a radial polynomial, m must be even. We need to fit to $f(x, y)$ a ridge sum of the form

$$\hat{f}_n = \sum_{i=0}^{n-1} \hat{c}_i h_i, \tag{3.1}$$

where $h_i(x, y) = H_m(x \cos\theta_i + y \sin\theta_i)$ and $\{\hat{c}_i\}$ are best in a least-squares sense. The normal equations (2.8) become (in the notation of Section 2)

$$\Gamma\hat{\mathbf{c}} = c_m \mathbf{1},$$

where $c_m = \langle f, h_i \rangle / m!$ does not depend on i *because* f *is radial.* But $\mathbf{1}$ is an eigenvector of Γ corresponding to the eigenvalue $\zeta_0 \equiv nP(S_m \equiv 0\ [2n])$ (Proposition 2.3), so that we may take $\hat{c}_i \equiv c_m/\zeta_0$. (The solution is not unique if $m < n - 1$.)

The total variance explained by the best n term approximation is, therefore,

$$\begin{aligned} \|\hat{f}_n\|^2 &= c_m^2 \left\| \sum_{i=0}^{n-1} h_i \right\|^2 / \zeta_0^2 \\ &= m! c_m^2 \mathbf{1}^t \Gamma^{(m)} \mathbf{1} / \zeta_0^2 = m! c_m^2 / P(S_m \equiv 0\ [2n]), \end{aligned}$$

again because $\Gamma^{(m)}\mathbf{1} = \zeta_0 \mathbf{1}$.

What is the percentage of $\|f\|^2$ explained by $\hat{f}_n$? Any polynomial in $\mathscr{P}_m^0$ can be represented using $m + 1$ directions: $\mathscr{P}_m^0$ is an $(m + 1)$-dimensional space, and (2.9) shows that Γ is nonsingular, confirming that the family $\{h_j\}_{j=0}^m$ is in this case linearly independent. In particular our radial function $f = \hat{f}_{m+1}$, and we have the first part of Proposition 1.1:

$$\begin{aligned} \frac{\|\hat{f}_n\|^2}{\|f\|^2} &= \frac{P(S_m = 0)}{P(S_m = 0\ [2n])} \\ &= \binom{m}{m/2} \Big/ \left[\binom{m}{m/2} + 2\binom{m}{m/2 + 2n} + 2\binom{m}{m/2 + 4n} + \cdots \right], \end{aligned} \tag{3.2}$$

where we have used the binomial formula for $P(S_m = j)$.

Conversely, fixing the number of directions, one can obtain information from (3.2) about how much the fit deteriorates as the degree m increases. This was done numerically for various values of n and m in Tables 1, 2 and Figure 1b in the Introduction. Asymptotics appear later in the section.

Now let us turn to harmonic polynomials and consider their approximation by ridge functions in equally spaced directions $\theta_j = \theta_0 + j\pi/n$, $j = 0, 1, \ldots, n-1$. Again, we first consider an arbitrary harmonic f belonging to $\mathscr{P}_m^0$: Write it as $f = \mathrm{Re}(\alpha z^m)$, where $\alpha = s_0 e^{im\varphi_0} \in \mathbb{C}$. The quality of approximation of f may depend on θ_0, but we will set $\theta_0 = 0$ as this is equivalent to replacing φ_0 by $\varphi_0 - \theta_0$.

A best approximation $\hat{f}_n = \hat{c} \cdot \mathbf{h} = \Sigma \hat{c}_j h_j$ is again given by a solution of the normal equations (2.8). To compute $\langle h_i, f \rangle$, we borrow from Corollary 4.2 the

identity $E[Z^m|\theta_j \cdot \mathbf{X}] = e^{im\theta_j}H_m(\theta_j \cdot \mathbf{X})$ and calculate

$$E\left[Z^m h_j\right] = e^{im\theta_j}EH_m^2(\theta_j \cdot \mathbf{X}) = m!e^{im\theta_j} = m!w(j),$$

where $\mathbf{w} = (w(j))_{j=0}^{n-1}$ is an eigenvector of $\Gamma^{(m)}$ associated with the eigenvalue $\zeta = \zeta^{(-m)} = nP(S_m \equiv -m\ [2n])$ in Proposition 2.3 [cf. also (2.9)]. Thus a solution of the normal equations is given by

$$\hat{c} = \frac{1}{\zeta}\mathrm{Re}(\alpha\mathbf{w}).$$

The total variance explained by the best fit using n directions is

$$\begin{aligned} \|\hat{f}_n\|^2 &= \|\hat{c} \cdot \mathbf{h}\|^2 = m!\hat{c}\Gamma\hat{c} \\ &= \frac{m!}{\zeta}|\mathrm{Re}(\alpha\mathbf{w})|^2 \\ &= \frac{m!}{P(S_m \equiv -m[2n])} \times \begin{cases} |\alpha|^2/2, & m \not\equiv 0\ [n], \\ (\mathrm{Re}\,\alpha)^2, & m \equiv 0\ [n], \end{cases} \end{aligned} \tag{3.2'}$$

since $\mathbf{w} = \mathbf{u} + i\mathbf{v}$, $\mathbf{u} \perp \mathbf{v}$, $|\mathbf{u}|^2 + |\mathbf{v}|^2 = n$ and $|\mathbf{u}|^2 - |\mathbf{v}|^2 = nI\{m \equiv 0\ [n]\}$. It can be seen that the orientation θ_0 of the equally spaced directions affects the quality of approximation of $\hat{f}_n$ only when m is an integer multiple of n—in these cases best approximation occurs if $\theta_0 = \varphi_0$. Henceforth, we shall assume that $\theta_0 = \varphi_0$, so that $\mathrm{Re}\,\alpha$ in (3.2′) can be replaced by $|\alpha|^2$. Formula (3.2′) can then be rewritten in an unified way by noting that $\mathscr{P}(S_m \equiv -m\ [2n]) = \mathscr{P}(S_m \equiv m\ [2n])$ because of symmetry, and that $-m \equiv +m\ [2n]$ only when $m \equiv 0\ [n]$. We obtain

$$\|\hat{f}_n\|^2 = m!|\alpha|^2/\mathscr{P}(S_m \equiv \pm m\ [2n]).$$

As in the radial case, f can always be represented using $n = m + 1$ directions, so that $f = \hat{f}_{m+1}$ and

$$\frac{\|\hat{f}_n\|^2}{\|f\|^2} = \frac{\mathscr{P}(S_m \equiv \pm m\ [2m+2])}{\mathscr{P}(S_m \equiv \pm m\ [2n])}, \tag{3.3}$$

which is equivalent to (1.4) and completes the proof of Proposition 1.1. This expression contrasts markedly with (3.2), involving (equivalence classes of) the extremes of the support of $\mathscr{L}(S_m)$ as opposed to (equivalence classes of) the mode of $\mathscr{L}(S_m)$. Again, we can explicitly study behavior as a function of degree m and number of directions n.

3.2. *Some asymptotics.* It is already evident from (3.2) and Figure 1b that a relatively small number of directions suffice to approximate well a relatively high degree radial polynomial. To see how high the degree can get, it is helpful to derive some central limit type bounds for the binomial probabilities involved.

LEMMA 3.1. *If* $|\nu| < m/2$, *then*

$$\varepsilon(n, m, \nu) := P(S_m \neq \nu | S_m \equiv \nu[2n]) \leq C_1\sqrt{m}\,e^{-2n(n-|\nu|)/m}. \tag{3.4}$$

PROOF. Assume without loss of generality that $\nu \geq 0$. Then clearly

$$\begin{aligned}\varepsilon(n, m, \nu) &\leq 2P(S_m \leq \nu - 2n)/P(S_m = \nu) \\ &\leq 2\exp\{-2(2n - \nu)^2/4m\}/P(S_m = \nu)\end{aligned}$$

by Hoeffding's inequality [Hoeffding (1963), Theorem 2, and Pollard (1984), page 191]. Since for $\nu < m/2$, $P(S_m = \nu) \geq c_0 m^{-1/2} \exp\{-\nu^2/2m\}$, as may be checked from Stirling's formula; this gives the result directly. □

From (3.2), if f is a radial polynomial of pure Hermite degree m, then by substituting $\nu = 0$ in (3.4), we find that the percentage of unexplained variance using n directions is

$$\|f - \hat{f}_n\|^2/\|f\|^2 \leq C_1\sqrt{m}\, e^{-2n^2/m}, \tag{3.5}$$

which is vanishingly small for large n if $m(n) \leq (2 - \varepsilon)n^2/\log n$. Thus for a homogeneous radial polynomial f of degree m, an excellent approximation is possible using $N \sim \sqrt{m \log m}$ coefficients as opposed to the $m + 1$ coefficients that occur if f is simply written out in terms of the basis $H_k(x)H_{m-k}(y)$.

To extend these ideas to general radial functions in $L^2(\Phi_2)$, consider the expansion

$$f(r) \sim \sum_{\substack{s \text{ even} \\ s \geq 0}} \gamma_s R_s(r), \tag{3.6}$$

where $R_s(r)$ are orthogonal radial polynomials of degree s and norm 1. [In the notation of Section 4, $R_s(r) = T_{s/2, s/2}(r)/\|T_{s/2, s/2}\|$]. As is well known in classical Fourier series, conditions on the rate of decay of the coefficients in (3.6) correspond to smoothness conditions on f. Here is a version of these results adapted for our later use (and proved in Section 5); P_m denotes projection onto polynomials of degree at most m.

LEMMA 3.2. *If $m \geq p$, then*

$$m^p\|f - P_m f\|^2 \leq B_p \sum_{i=1}^{p} \binom{p}{i} \|D_x^i D_y^{p-i} f\|^2.$$

Let $\mathscr{F}_p$ denote the class of functions f having p derivatives in $L^2(\Phi_2)$. An approximation $P_m f$ using a bivariate polynomial of degree m involves $N = (m + 1)(m + 2)/2 = O(m^2)$ coefficients and leaves an error of order $m^{-p} = O(N^{-p/2})$. In the language of the Introduction, we have a rate of approximation and effective dimensionality

$$r^*(\mathscr{F}_p, \mathrm{PA}) = p/2, \qquad d^*(\mathscr{F}_p, \mathrm{PA}) = 2, \tag{3.7}$$

where PA stands for polynomial approximation.

Suppose now that f is radial. Let us compute by contrast to (3.7) the approximation error of a degree m ridge function approximation to f using n (equally spaced) directions: We denote this by $R_{n,m} f$. This involves $N = mn + 1$

coefficients. Denote by $\hat{R}_{s,n}$ the best n (equispaced) direction approximation to R_s: We have from (3.5)

$$\begin{aligned}\|f - R_{n,m}f\|^2 &= \sum_{\substack{s \text{ even} \\ s \le m}} \gamma_s^2 \|R_s - \hat{R}_{s,n}\|^2 + \sum_{s>m} \gamma_s^2 \|R_s\|^2 \\ &\le C_1 \sum_{s \le m} \gamma_s^2 \sqrt{s}\, e^{-2n^2/s} + C_2 m^{-p} \\ &\le C_1 \|f\|^2 \sqrt{m}\, e^{-2n^2/m} + C_2 m^{-p}.\end{aligned}$$

If we now set $m = m(n) = n^2/\delta \log n$, we find for sufficiently large δ and N,

$$\begin{aligned}\|f - R_{n,m}f\|^2 &\le C\left(\frac{1}{n^{2\delta-1}\sqrt{\log n}} + \frac{\log^p n}{n^{2p}}\right) \\ &\le C(N/\log N)^{-p/1.5},\end{aligned}$$

where C depends on f and δ.

Ignoring $\log N$ terms, for the subspace $\mathscr{R}_p$ of radial functions within $\mathscr{F}_p$, this leads to the improved rate and dimensionality results claimed in the Introduction,

$$r^*(\mathscr{R}_p, \mathrm{RA}) = p/1.5, \qquad d^*(\mathscr{R}_p, \mathrm{RA}) = 1.5.$$

The asymptotic woes of equispaced ridge approximation for harmonic functions are readily derived from (3.3). Suppose that $m > n$: Then clearly there exist integers $\pm j$ with $|j| \le m - 2$ for which $j \equiv m$ or $-m[2n]$. Therefore, we must have for f harmonic in $\mathscr{P}_m^0$,

$$P(S_m \equiv \pm m[2n]) \ge P(|S_m| = m \text{ or } m-2),$$

and hence,

$$\|\hat{f}_n\|^2/\|f\|^2 \le \frac{P(|S_m| = m)}{P(|S_m| = m \text{ or } m-2)} = \frac{1}{m+1}. \tag{3.8}$$

Thus even if only one less direction is used than is needed to perfectly represent f as a ridge sum, the percentage of variation explained is still negligible.

Consider now a general harmonic function f, decomposed into polynomials f_k of pure Hermite degree k: $f = \Sigma_k f_k$. The best degree m equispaced ridge approximation to f using n directions $\theta_0 + j\pi/n$, $j = 0, \ldots, n-1$, is $R_{n,m}f = \sum_{k=1}^{m} \hat{f}_{k,n}$. For any choice of θ_0, $\hat{f}_{k,n} = f_k$ for $k < n$, and by adjusting θ_0 as in the discussion before (3.3), we can ensure that $\hat{f}_{n,n} = f_n$. Therefore, on using (3.8) for $n < k \le m$, we find

$$\begin{aligned}\|f - R_{n,m}f\|^2 &= \sum_{k=n+1}^{m} \|f_k - \hat{f}_{k,n}\|^2 + \|f - P_m f\|^2 \\ &\ge \frac{n}{n+1}\|P_m f - P_n f\|^2 + \|f - P_m f\|^2.\end{aligned} \tag{3.9}$$

Thus, for example, if $N = m_0^2/2$ coefficients are used to approximate f using a polynomial of degree less than or equal to m_0, then the approximation error is $\|f - P_{m_0}f\|^2$, whereas if $n < m_0$ equally spaced ridge directions are taken, and the best ridge approximation of degree $m^2/2n$ is used, the approximation error

is, by (3.9), at least

$$\frac{n}{n+1}\|P_{m^2/2n}f - P_n f\|^2 + \|f - P_{m^2/2n}f\|^2 \geq \frac{n}{n+1}\|f - P_{m_0}f\|^2.$$

So, in a very strong sense, equispaced ridge approximation for harmonic functions does no better than the use of bivariate polynomials, whose rate of approximation is $p/2$ and whose effective dimensionality is 2 in general [cf. (3.7)]. This completes the proof of Proposition 1.2.

4. A basis for $L^2(\Phi_2)$. The results of the last section raise several questions. Are radials and harmonics extreme cases for ridge approximation? Does the good performance of RA carry over to broader classes of functions than just radials?

This section describes an orthogonal basis for $L^2(\Phi_2)$ which is orthogonally equivariant, just as is the problem of approximating by sums of ridge functions. It exhibits harmonics and radials as special elements, behaves conveniently with respect to projection, and it gives a convenient representation for various differentiation operators. These properties are the building blocks for our main results in later sections.

We will write the two coordinates (x, y) in complex form $z = x + iy = re^{i\theta}$ and use the commuting differentiation operators D_z, $D_{\bar{z}}$ defined by

$$D_z = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right), \qquad D_{\bar{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right).$$

Note that

$$D_z z = D_{\bar{z}}\bar{z} = 1, \qquad D_z\bar{z} = D_{\bar{z}}z = 0.$$

The basis consists of (complex) polynomials defined for integers, $k, l \geq 0$ by

$$J_{k,l}(z, \bar{z}) = e^{z\bar{z}/2}(-2D_z)^k(-2D_{\bar{z}})^l e^{-z\bar{z}/2}, \tag{4.1}$$

which have generating function

$$G(s, t) := \sum_{k,l}\frac{s^k}{k!}\frac{t^l}{l!}J_{k,l}(z, \bar{z}) = \exp\{s\bar{z} + tz - 2st\}. \tag{4.2}$$

The second equality may be seen by the following formal calculation, which relies on the commutativity of the operators (to apply the product law for exponentials) and the properties of the univariate Hermite polynomial generating function,

$$\begin{aligned}
G(s, t) &= e^{z\bar{z}/2}\sum_k \frac{(-2sD_z)^k}{k!}\sum_l \frac{(-2tD_{\bar{z}})^l}{l!}e^{-z\bar{z}/2} \\
&= e^{z\bar{z}/2}e^{-2sD_z - 2tD_{\bar{z}}}e^{-z\bar{z}/2} \\
&= e^{(x^2+y^2)/2}e^{-uD_x - ivD_y}e^{-(x^2+y^2)/2} \qquad (u = s + t,\ v = t - s) \\
&= e^{-u^2/2 + ux}e^{-(iv)^2/2 + ivy} \\
&= e^{-(2st - s\bar{z} - tz)}.
\end{aligned}$$

The usual argument [based on computing the product generating function $G(s, t)G(s', t')$ in two ways, using the independence and Gaussianity and equating coefficients] yields the orthogonality relations

$$E J_{kl}(z, \bar{z})\overline{J_{mn}(z, \bar{z})} = \delta_m^k \delta_n^l k! l! 2^{k+l}. \tag{4.3}$$

Generating function methods further yield the identities

$$\begin{aligned} \overline{J_{kl}(z, \bar{z})} &= J_{lk}(z, \bar{z}), \\ D_z J_{kl} &= l J_{k, l-1}, \\ D_{\bar{z}} J_{kl} &= k J_{k-1, l}. \end{aligned} \tag{4.4}$$

The J_{kl}, $k \le l$, are orthogonal polynomials in z and $\bar{z}$ of degree $k + l$. Since

$$(-2D_z)e^{-z\bar{z}/2} = \bar{z}e^{-z\bar{z}/2}, \qquad (-2D_{\bar{z}})e^{-z\bar{z}/2} = ze^{-z\bar{z}/2}, \tag{4.5}$$

it is clear that

$$J_{0k} = z^k = r^k e^{ik\theta}, \qquad J_{k0} = \bar{z}^k = r^k e^{-ik\theta}.$$

Thus J_{0k} and J_{k0} are holomorphic and antiholomorphic, respectively. In particular, the real and imaginary parts are harmonic. The usefulness of the $\{J_{kl}\}$ in Gaussian projection computations comes partly from their representation as a product of a circular harmonic and a radial polynomial,

$$\begin{aligned} J_{kl}(z, \bar{z}) &= S_{l-k}(\theta)T_{k, l}(r) \\ &= e^{i(l-k)\theta} \sum_{j=0}^{k \wedge l} \binom{k}{j}\binom{l}{j} j!(-2)^j r^{l+k-2j}. \end{aligned} \tag{4.6}$$

[To see this, use (4.5) repeatedly in the definition (4.1) of J_{kl}, together with Leibniz's formula for computing $D^k(fg)$.] Thus $J_{k,k}$ is a *real* radial polynomial of degree $2k$. The orthogonality relations (4.3) show that $T_{k,l}$ is related to the Laguerre polynomials, $L_k^\alpha(t)$, e.g., Szegö (1939), by

$$T_{k, k+\alpha}(r) = \gamma_k r^\alpha L_k^\alpha(r^2/2), \qquad \alpha \ge 0, \tag{4.7}$$

where $\gamma_k = (-2)^k k!$ This yields the alternative representations

$$\begin{aligned} J_{kl}(z, \bar{z}) &= \gamma_k z^\alpha L_k^\alpha(|z|^2/2) \qquad (\alpha = l - k \ge 0) \\ &= \gamma_k e^{i\alpha\theta} r^\alpha L_k^\alpha(r^2/2), \end{aligned}$$

which are occasionally more revealing than (4.6) or (4.7).

REMARK. It turns out that related forms of the J_{kl} basis have also been exploited by workers in tomography, plasma physics and holography. Deans (1983) (Sections 3.8 and 7.7) has some development and references.

Differentiation. As is already apparent from (4.4), the $\{J_{kl}\}$ basis affords a convenient representation for various differentiation operations discussed in the

next section. The formulas, collected below, are simple consequences of (4.4) and (4.6):

$$D_x J_{kl} = kJ_{k-1,l} + lJ_{k,l-1}, \qquad D_y J_{kl} = ilJ_{k,l-1} - ikJ_{k-1,l},$$
$$\Delta J_{kl} = 4D_z D_{\bar{z}} J_{kl} = 4klJ_{k-1,l-1}, \tag{4.7'}$$
$$D_\theta J_{kl} = \frac{\partial}{\partial\theta} J_{kl} = i(l-k)J_{kl}.$$

Expansions. For $k \geq 0$, $l \geq 0$ and $k + l \leq m$, the $J_{kl}(z, \bar{z})$ form a basis for the space of polynomials in z and $\bar{z}$ (equivalently x and y) of degree at most m. [This follows from the orthogonality relations (4.3).] Consequently $\{J_{kl},\ k, l \geq 0\}$ form a complete orthogonal basis and any (complex-valued) function $f \in L^2(\Phi_2)$ may be written

$$f = \sum_{k,l\geq 0} c_{kl} J_{kl},$$

where c_{kl} and $J_{kl} \in \mathbb{C}$.

Our primary interest lies in real-valued functions, so we add the restriction $c_{lk} = \bar{c}_{kl}$, which implies that $\overline{c_{lk}J_{lk}} = c_{kl}J_{kl}$ and hence guarantees that f is real. Then the real and imaginary parts $J^0_{kl} + iJ^1_{kl}$ of $\{J_{kl},\ k + l = m,\ k \leq l\}$ form a basis for $\mathscr{P}^0_m$, the space of (real-valued) polynomials of pure Hermite degree m defined in 2. With these conventions,

(a) f is radial iff $c_{kl} = 0$ for $k \neq l$;
(b) f is harmonic iff $c_{kl} = 0$ whenever both $k \neq 0$, $l \neq 0$;
(c) f is polynomial of degree at most m iff $c_{kl} = 0$ for $k + l \geq m$.

A projection formula for J_{kl}. The following projection formula is fundamental. It shows that the angular part of J_{kl} is preserved under projection onto a line, while the radial part maps onto the Hermite polynomial of matching degree. Let $\theta = (\cos\theta, \sin\theta)$ and $\mathbf{X} = (X, Y)$.

LEMMA 4.1. $E(J_{kl}(z, \bar{z})|\theta \cdot \mathbf{X} = t) = e^{i(l-k)\theta} H_{k+l}(t)$.

A short proof can be given using the generating function (4.2): $G(s, t) = \exp\{s\bar{z} + tz - 2st\} = \exp\{\boldsymbol{\beta} \cdot \mathbf{x} - \frac{1}{2}|\beta|^2\}$, where $\beta = (s + t, i(t - s)) \in \mathbb{C}^2$. For real vectors $\beta \in \mathbb{R}^2$, $\mathscr{L}(\boldsymbol{\beta} \cdot \mathbf{X}|\theta\mathbf{X}) = N((\theta\beta)(\theta\mathbf{X}), |\beta|^2 - (\theta\beta)^2)$ and hence

$$\begin{aligned} E(G(s,t)|\theta \cdot \mathbf{X}) &= E\left(e^{\boldsymbol{\beta}\cdot\mathbf{X} - |\beta|^2/2}|\theta \cdot \mathbf{X}\right) \\ &= e^{(\theta\cdot\beta)(\theta\cdot\mathbf{X}) - (\theta\cdot\beta)^2/2} \\ &= \sum_m \frac{(\theta \cdot \beta)^m}{m!} H_m(\theta \cdot \mathbf{X}). \end{aligned}$$

The above identities extend to $\beta \in \mathbb{C}^2$ by analytic continuation. Finally, since $\theta \cdot \boldsymbol{\beta} = se^{-i\theta} + te^{i\theta}$, the result follows by expanding the powers, rearranging and equating coefficients.

COROLLARY 4.2. $E[Z^m|\theta \cdot \mathbf{X} = t] = e^{im\theta}H_m(t)$.

Illustration: Best approximation by a single ridge polynomial. The role of harmonic functions as worst cases for projection pursuit approximation appears in simple form already at the initial stage of fitting the first ridge function. For a square integrable function $f \in L^2(\Phi_2)$ the best ridge function approximation in direction $\theta = (\cos\theta, \sin\theta)$ is $P_\theta f(\mathbf{x}) = E(f(\mathbf{X})|\theta \cdot \mathbf{X} = \theta \cdot \mathbf{x})$. The best direction is then found by maximizing $S(f, \theta) = \|P_\theta f\|^2 = E[P_\theta f(\mathbf{X})]^2$. The percentage variation explained by the first ridge function is, therefore, $A(f) = \max_\theta S(f, \theta)/\|f\|^2$. A virtue of the basis $\{J_{kl}\}$ is that $A(J^0_{kl})$ and $A(J^1_{kl})$ can be readily calculated from Proposition 4.1:

$$A(J^0_{kl}) = A(J^1_{kl}) = \max_\theta \cos^2[(l-k)\theta]\, E[H^2_{k+l}]/E[J^0_{k,l}]^2$$

$$= \frac{(k+l)!}{k!l!}\begin{cases} 2^{1-k-l}, & k \neq l, \\ 2^{-k-l}, & k = l. \end{cases} \tag{4.8}$$

The worst cases amongst the basis functions of a given degree $k + l = m$ are, therefore, the harmonic polynomials $J^0_{m0} = \mathrm{Re}(z^m)$ and $J^1_{m0} = \mathrm{Im}(z^m)$.

This calculation can be extended to show that harmonic polynomials of degree m are worst, first amongst all polynomials of pure Hermite degree m, and then amongst all polynomials of degree at most m.

5. Smoothness classes for $L^2(\Phi_d)$. Because the differentiation operators have such simple representations in the J_{kl} basis, it is easy to define and to study Sobolev-type measures of smoothness in $L^2(\Phi_d)$. This section describes three such measures: ordinary (or Cartesian) smoothness, angular smoothness and Laplacian smoothness. These can all be expressed as simple quadratic forms in the coefficients of a function's J_{kl} expansion. Moreover, relations between these are easily derived. Cartesian smoothness implies a certain amount of angular or Laplacian smoothness. However, the reverse implications do not hold. Moreover, for a given amount of Cartesian smoothness, the angular and Laplacian smoothness are complementary. Also, radials have the most angular, and least Laplacian, smoothness for a given Cartesian smoothness, while harmonics have the most Laplacian, and least angular smoothness. It is the notions of angular and Laplacian smoothness that are crucial for generalizing the results of Section 3.

Three notions of smoothness. Our definitions of smoothness are of the traditional Sobolev-type. Let $D^{\mathbf{k}}$ denote the partial derivative operator with index $\mathbf{k} = (k_1, \ldots, k_d)$. ($d > 2$ occurs in Section 7.) Say that f has p (local) weak derivatives if for every multiindex $\mathbf{k}$ of order $|\mathbf{k}| = k_1 + \cdots + k_d = p$ there is a locally integrable function $D^{\mathbf{k}}f$ such that $\int \varphi\, D^{\mathbf{k}}f = (-1)^p \int f\, D^{\mathbf{k}}\varphi$ for all test functions φ.

The function $f \in L^2(\Phi_d)$ will be said to have Cartesian smoothness of order p if it has p weak derivatives, and these derivatives are all in $L^2(\Phi_d)$. It will be said to have angular smoothness of order q if it has q weak derivatives and

$D_\theta^q f \in L^2(\Phi_d)$. Finally, it will be said to have Laplacian smoothness of order r if it has $2r$ weak derivatives, and $\Delta^r f \in L^2(\Phi_d)$.

Moment representation of smoothness conditions. These definitions are equivalent to simpler moment conditions on the expansion of a function in the J_{kl} basis. Of the proofs we present only the computational aspects that relate to Gaussian measure, as everything else involves standard arguments [as in, for example, Gilbarg and Trudinger (1977), Chapter 7]. Let $f \in L^2(\Phi_d)$ have the expansion $\Sigma c_{kl} J_{kl}$. Because

$$\| f \|^2 = \sum |c_{kl}|^2 \| J_{kl} \|^2,$$

one can think of $|c_{kl}|^2 \| J_{kl} \|^2$ as representing the distribution of energy in f among the different modes of oscillation J_{kl}. Let $p_{kl}(f)$ denote the probability distribution on the upper half of the positive integer quadrant $\{(k, l) \in \mathbb{Z}^2 : 0 \leq k, l\}$ defined by

$$p_{kl} \propto |c_{kl}|^2 \| J_{kl} \|^2$$

and let E denote expectation with respect to this distribution. (Of course, when f is real, the distribution is symmetric about the line $k = l$.)

PROPOSITION 5.1. (a) *f has p ordinary $L^2(\Phi_d)$ derivatives if $E(K + L)^p < \infty$.*
(b) *f has q angular $L^2(\Phi_d)$ derivatives iff $E|K - L|^{2q} < \infty$.*
(c) *f has r Laplacians in $L^2(\Phi_d)$ iff $E[(KL)^r | K \wedge L \geq r] < \infty$.*

Thus the measures of smoothness correspond to conditions on the distribution of a function's energy. As $k + l$ represents the sum of the radial and angular oscillation of the basis element J_{kl}, Cartesian smoothness requires small energy at terms of high total oscillation. In order for a function to have angular smoothness, its energy should concentrate at low values of $k - l$: low angular oscillation.

PROOF. (a) Using the multiplier representation (4.4), we have [if $\alpha + \beta = p$ and $(l)_\alpha = l(l-1)\cdots(l-\alpha+1)$ for $l \geq \alpha$ and 0, otherwise] that $D_z^\alpha D_{\bar{z}}^\beta f = \Sigma c_{kl} (l)_\alpha (k)_\beta J_{k-\beta, l-\alpha}$. Hence

$$\begin{aligned} \| D_z^\alpha D_{\bar{z}}^\beta f \|^2 &= \sum |c_{kl}|^2 (l)_\alpha^2 (k)_\beta^2 (k-\beta)!(l-\alpha)! 2^{k-\beta+l-\alpha} \\ &= 2^{-p} \sum |c_{kl}|^2 (l)_\alpha (k)_\beta \| J_{kl} \|^2. \end{aligned}$$

Applying the hypergeometric sampling identity, we get

$$\sum_{\alpha=0}^{p} \binom{p}{\alpha} \| D_z^\alpha D_{\bar{z}}^{p-\alpha} f \|^2 = 2^{-p} \sum |c_{kl}|^2 (l+k)_p \| J_{kl} \|^2. \tag{5.1}$$

From this, we see that f has p Cartesian derivatives in $L^2(\Phi_d)$ if $E[(K + L)_p] < \infty$. Since

$$\frac{p^p}{(p)_p} \geq \frac{(k+l)^p}{(k+l)_p} \geq 1, \qquad k + l \geq p,$$

this condition is equivalent to $E(K + L)^p < \infty$.

(b) From (4.7'), one finds

$$\|D_\theta^q f\|^2 = \|\sum c_{k,l} D_\theta^q J_{kl}\|^2 = \sum |k-l|^{2q} |c_{kl}|^2 \|J_{kl}\|^2. \tag{5.2}$$

(c) Since $\Delta = 4D_z D_{\bar{z}}$, $\Delta J_{kl} = 4kl J_{k-1,l-1}$, and so

$$\Delta^r f = 4^r \sum_{k,l \geq r} c_{kl} (k)_r (l)_r J_{k-r,l-r},$$

$$\|\Delta^r f\|^2 = 4^{2r} \sum_{k,l \geq r} |c_{kl}|^2 (k)_r^2 (l)_r^2 (k-r)!(l-r)! 2^{k-r+l-r} \tag{5.3}$$

$$= 4^r \sum |c_{kl}|^2 (k)_r (l)_r \|J_{kl}\|^2.$$

Thus f has r L^2 Laplacians if $E[(K)_r (L)_r] < \infty$. Because

$$\frac{r^{2r}}{(r)_r (r)_r} \geq \frac{(kl)^r}{(k)_r (l)_r} \geq 1$$

for $k, l \geq r$, this last condition is equivalent to $E[(KL)^r | K \wedge L \geq r] < \infty$. □

Relations between smoothness measures. It is now easy to establish certain relations between the different smoothness measures. To begin with, Cartesian smoothness implies a certain amount of angular smoothness. Since $E(K+L)^p \geq E|K-L|^{2q}$ if $p \geq 2q$, a function f with p Cartesian derivatives has at least $q = p/2$ angular derivatives. However, if f is harmonic, so that the energy in its expansion concentrates on the axes $k = 0$ and $l = 0$, then $E|K-L|^{2q} = E(K+L)^{2q}$. Thus, harmonic functions can attain this lower bound on angular smoothness for a given Cartesian smoothness: $q = p/2$. In other words, for harmonic functions, if $q = p/2$, then $E|K-L|^{2q} < \infty$ iff $E(K+L)^p < \infty$. Of course, radial functions have the most angular smoothness for a given Cartesian smoothness ($q \geq p$ and $q = \infty$ if f is C^∞).

Cartesian smoothness also implies a certain degree of Laplacian smoothness. Since $(k+l)^p \geq (kl)^r$ if $p \geq 2r$, a function with p Cartesian derivatives $L^2(\Phi_d)$ must have at least $r = p/2$ Laplacians. If f is radial so that its energy concentrates on the diagonal $k = l$, then $E(K+L)^{2r} = 2^{2r} E(KL)^r$. Thus radial functions can attain the least Laplacian smoothness for a given Cartesian smoothness: $r = p/2$. Of course, a harmonic function has the most Laplacian smoothness for a given Cartesian smoothness: $\Delta f = 0$, so $\Delta^k f = 0$ for each $k \geq 1$. (Thus $r \geq p/2$ and $r = \infty$ if f is C^∞.)

These results illustrate the extreme nature of harmonics and radials, but they also suggest the complementary roles of angular and Laplacian smoothness. Formally, we have

PROPOSITION 5.2. *If $f \in L^2(\Phi_d)$ has exactly p ordinary derivatives in L^2, then f cannot have both more than $p/2$ angular derivatives in L^2 and more than $p/2$ Laplacians in $L^2(\Phi_d)$.*

PROOF. The assumption states that for the probability measure $\{p_{kl}\}$ corresponding to f,

$$\sup\{r: E(K+L)^r < \infty\} = p.$$

Now from the convexity of the function $x \to x^q$ (for $q \geq 1$), we have

$$E(K+L)^{2q} = E\{(K-L)^2 + 4KL\}^q$$
$$\leq 2^{q-1}E\{|K-L|^{2q} + 4^q(KL)^q\}.$$

Thus if $E(K+L)^{2q} = \infty$, then at most one of the terms on the right side can be finite. It is possible that $E(KL)^q = \infty$, but $E[(KL)^q | K \wedge L \geq q] < \infty$. In this case, for some $r < q$ we must have either $E[L^q | K = r] = \infty$ or $E[K^q | L = r] = \infty$. This however forces $E|K-L|^q$ and hence $E|K-L|^{2q}$ to diverge. □

Smoothness classes and approximation rates. Consider now the smoothness classes mentioned in the Introduction. The first is the class of $\mathscr{F}_p$ of functions with Cartesian smoothness p. The second is the class $\mathscr{A}_{pq}$ of functions with Cartesian smoothness p and angular smoothness q. The third is the class $\mathscr{L}_{pr}$ of functions with Cartesian smoothness p and Laplacian smoothness r. These are linear subspaces of $L^2(\Phi_d)$. Proposition 5.2 shows that $\mathscr{F}_p = \mathscr{A}_{p,\,p/2} = \mathscr{L}_{p,\,p/2}$ and that for $q \wedge r \geq p/2$,

$$\mathscr{A}_{pq} \cap \mathscr{L}_{pr} \subset \mathscr{F}_{2(q \wedge r)}.$$

Figure 3 in the Introduction gave a schematic of the relations between the subspaces in geometric terms.

The importance of smoothness classes comes in characterizing rates of convergence of approximation procedures. For example, *a function $f \in \mathscr{F}_p$ can be approximated by polynomials of degree m with a squared error of order m^{-p}.* This was asserted in Lemma 3.2; the proof is simple, using (5.1):

$$m^p \| f - P_m f \|^2 \leq \frac{m^p}{(m)_p} \sum_{k+l \geq m} |c_{kl}|^2 (k+l)_p \| J_{kl} \|^2$$
$$\leq \frac{m^p}{(m)_p} 2^p \sum_{\alpha=0}^{p} \binom{p}{\alpha} \| D_z^\alpha D_{\bar{z}}^{p-\alpha} f \|^2 \leq C(p, f).$$

As a second example, if $\| D_\theta^q f \| < \infty$, then f can be approximated by functions with angular frequency less than or equal to μ at a rate μ^{-2q}. This fact will be used in Section 6. Let $f = \sum c_{kl} J_{kl}$ and define the operator $Q_\nu f = \sum_{|k-l|=\nu} c_{kl} J_{kl}$. Write $f = f(r, \theta) = \sum_\nu e^{i\nu\theta} h_\nu(r)$ and note that

$$\| D_\theta^q f(r, \theta) \|^2 = \left\| \sum_{\nu} (i\nu)^q e^{i\nu\theta} h_\nu(r) \right\|^2$$
$$= \sum_{\nu} \nu^{2q} \| e^{i\nu\theta} h_\nu(r) \|^2$$

and hence

$$\left\| \sum_{\mu+1}^{\infty} Q_\nu f \right\|^2 = \sum_{|\nu| > \mu}^{\infty} \| e^{i\nu\theta} h_\nu(r) \|^2 \leq \mu^{-2q} \| D_\theta^q f(r, \theta) \|^2. \tag{5.4}$$

The next two sections will show that $\mathscr{A}_{pq}$ and $\mathscr{L}_{pr}$ have similar significance for the rate of approximation by RA and KA. Functions in $\mathscr{A}_{pq}$ can be very well approximated by RA when $q > p/2$; those in $\mathscr{L}_{pr}$ by KA when $r > p/2$.

6. Ridge function approximation and angular smoothness. This section uses the notion of angular smoothness to extend the results of Section 3. The basic result is as follows. Let $R_{n,m}f$ denote the best ridge approximation to f using polynomials of degree m in the equally spaced directions $\theta_j = j\pi/n$, $j = 0, \ldots, n-1$. As before, $P_m f$ is the best $L^2(\Phi_2)$ approximation to f by a bivariate polynomial of degree m. $\tilde{\mathscr{A}}_{pq}$ denotes a weakly compact subset of $\mathscr{A}_{pq}$ of functions satisfying $\|f\| \le B_0, \sum_{k=0}^{p}\binom{p}{k}\|D_x^{p-k}D_y^k f\|^2 \le B_p, \|D_\theta^q f\| \le B_q$.

THEOREM 6.1. *If $p/2 \le q \le p$, there is a sequence $m(n) = O(n^{-2q/p})$ $[O(n^2/\log n)$ if $q = p]$ such that the best equispaced ridge approximation*

$$\sup_{\tilde{\mathscr{A}}_{pq}} \|f - R_{n,m(n)}f\|^2 \le C\begin{cases} N^{-p/d(q)}, & \text{if } p/2 \le q < p, \\ (N/\log^2 N)^{-p/1.5}, & \text{if } q = p, \end{cases} \tag{6.1}$$

where $N = m(n)n$ is the total complexity of approximation, and $d(q) = 1 + (p/2q) \in (1.5, 2]$ for $q \in [p/2, p)$.

The theorem says that if f has p ordinary derivatives and angular smoothness beyond the amount $p/2$ guaranteed by section 5, then approximation by a sum of ridge functions leads to an improvement $N^{-p/d(q)}$ over the (usual) $N^{-p/2}$ rate of approximation. Indeed, if $p = q$, we have $d = 1.5$ rather than the usual 2 (up to logarithm terms). On the other hand, other methods of approximation are not able to take advantage of increased angular smoothness. For ordinary bivariate polynomial approximation of complexity N, we have

$$\sup_{\tilde{\mathscr{A}}_{pq}} \|f - P_{\sqrt{2N}}f\|^2 \asymp CN^{-p/2}. \tag{6.2}$$

Indeed, one has this behavior at any radial function $f \in \tilde{\mathscr{A}}_{pq}$. Results of Section 7 will show that KA is unable to take advantage of enhanced angular smoothness.

It is possible to show that when $p = q$, the attained dimensionality 1.5 is the best possible, at least for approximations based on projections onto linear subspaces of $L^2(\Phi_2)$. This follows from the theory of n-widths [e.g., Pinkus (1985)] applied to $\mathscr{A}_{pq}$. So not only does RA take advantage of angular smoothness, it does so in a nearly optimal fashion. In this sense, RA is a natural procedure to use for approximating functions in $\tilde{\mathscr{A}}_{pq}$.

The intuition behind the proof has already been indicated in the Introduction. Those basis elements J_{kl} that can be well approximated by sums of ridge functions are those where $k + l$ is small or $|k - l|$ is small. Membership of f in the angular smoothness class $\tilde{\mathscr{A}}_{pq}$ forces the expansion of f to concentrate at just such basis elements. Thus, the proof breaks the approximation error into two parts: one due to high angular and radial frequencies and the other at low frequencies. Membership in $\tilde{\mathscr{A}}_{pq}$ provides an a priori bound on the high frequency terms; the low frequency terms are studied directly. The basic tool is Lemma 6.2, which gives the form of the best approximation to a single J_{kl} using n equispaced ridge functions. This tool can be used, for example, to compute the

data for Figure 2a in the Introduction. The lemma shows that the best approximation to J_{kl} can be expressed as a sum of $J_{k',l'}$ all of the same degree, $k' + l' = k + l$. As an equispaced sum of n ridge functions must have an angular wavenumber n, the $J_{k',l'}$ which enter in this representation are all the aliased terms $k' - l' \equiv k - l\ [n]$. Armed with that lemma, the proof proceeds fairly directly. Let $b_{mr} = \binom{m}{r}2^{-m}$.

LEMMA 6.2. $R_{n,m}J_{r_0,m-r_0} = \sum_{r=0}^{m}\beta_r J_{r,m-r}$, *where*

$$\beta_r = \begin{cases} \dfrac{P(S_m = m - 2r)}{P(S_m \equiv m - 2r_0\ [2n])} \triangleq \dfrac{b_{mr}}{[b_{mr_0}]}, & \text{if } r \equiv r_0\ [n], \\ 0, & \text{otherwise}, \end{cases}$$

and the percentage residual variance

$$\frac{\|R_{n,m}J_{r_0,m-r_0} - J_{r_0,m-r_0}\|^2}{\|J_{r_0,m-r_0}\|^2} = P(S_m \neq m - 2r_0 | S_m \equiv m - r_0\ [2n]) = 1 - \frac{b_{mr_0}}{[b_{mr_0}]}. \tag{6.2'}$$

REMARK 1. For nonnegative integers α and k set $r_0 = m/2 - \alpha$ and $n = m/2 + k$ in (6.2') to find that the percentage of variance explained using n directions is

$$P(S_m = 2\alpha | S_m \equiv 2\alpha\ [m + 2k]),$$

which equals 1 if and only if $k > \alpha$. An interesting consequence is that a degree m basis polynomial $J_{m/2-\alpha,m/2+\alpha}$ of angular oscillation $\alpha \in \{0, 1, \ldots, m/2\}$ can be exactly represented using $m/2 + \alpha + 1$ (or more) *equally spaced* directions, but no fewer. Thus a radial polynomial requires $m/2 + 1$ directions and a harmonic $m + 1$ and the remaining basis polynomials interpolate between these extremes.

REMARK 2. Let $J^0_{r_0,m-r_0} = (J_{r_0,m-r_0} + J_{m-r_0,r_0})/2$ denote $\mathrm{Re}(J_{r_0,m-r_0})$. The analog of (6.2') for a real basis polynomial, namely,

$$\mathrm{PVE}_n\left(J^0_{r_0,m-r_0}\right) = \frac{\|R_{n,m}J^0_{r_0,m-r_0}\|^2}{\|J^0_{r_0,m-r_0}\|^2} = P(S_m = \pm(m - 2r_0) | S_m \equiv \pm(m - 2r_0)\ [2n])$$

involves a slight subtlety in that its derivation from (6.2') splits into two cases. In the second of these, the real polynomial $J^0_{r_0,m-r_0}$ is approximated twice as well (for given n) as its complex counterpart. If $m - r_0 \not\equiv r_0\ [n]$, then the lattices $m - r_0 + n\mathbb{Z}$ and $r_0 + n\mathbb{Z}$ are disjoint, so that $R_{n,m}J_{r_0,m-r_0}$ and $R_{n,m}J_{m-r_0,r_0}$ are orthogonal. If on the other hand $m - 2r_0 \equiv 0\ [n]$, then the lattices coincide, and $R_{n,m}J_{r_0,m-r_0} = R_{n,m}J_{m-r_0,r_0}$. In the following calculation, the factor $\{2\}$ is

present only in the case $m - 2r_0 \equiv 0\ [n]$:

$$\begin{aligned}\mathrm{PVE}_n\left(J^0_{r_0,\,m-r_0}\right) &= \frac{\|R_{n,m}\left(J_{r_0,\,m-r_0} + J_{m-r_0,\,r_0}\right)\|^2}{\|J_{r_0,\,m-r_0} + J_{m-r_0,\,r_0}\|^2} \\ &= \{2\}\frac{\|R_{n,m}\left(J_{r_0,\,m-r_0}\right)\|^2}{\|J_{r_0,\,m-r_0}\|^2} \\ &= \frac{2P(S_m = m - 2r_0)}{2P(S_m \equiv m - 2r_0\ [2n])/\{2\}} \\ &= \frac{P(S_m = \pm(m - 2r_0))}{P(S_m \equiv \pm(m - 2r_0)\ [2n])}.\end{aligned}$$

Proposition 1.1 is an immediate corollary (set $r_0 = m/2$ and 0, respectively).

PROOF OF LEMMA 6.2. (1) The linear span (over $\mathbb{C}$) of the ridge functions $H_m(\theta_j^t \cdot)$ for the equally-spaced directions $\theta_j = j\pi/n$, $j = 0, 1, \ldots, n-1$, is equal to the space of functions $\sum_{r=0}^{m}\gamma_r b_{mr} J_{r,\,m-r}$ having periodic coefficients γ_r with period n.

To see this, consider a ridge sum $g = \sum_0^{n-1} c_j H_m(\theta_j^t \cdot)$ and express it in the $\{J_{r,\,m-r}\}$ basis: The coefficients $\gamma_r(g)b_{mr}$ are found from $\langle g, J_{r,\,m-r}\rangle/\|J_{r,\,m-r}\|^2$, and so from Proposition 4.1,

$$\begin{aligned}\gamma_r(g) &= \frac{1}{m!}\sum_0^{n-1} c_j E H_m\left(\theta_j^t \mathbf{X}\right)\bar{J}_{r,\,m-r} \\ &= \sum_{j=0}^{n-1} c_j e^{-i(m-2r)\theta_j}.\end{aligned}$$

From this the periodicity of γ_r is clear, and since the map from $\{c_0, \ldots, c_{n-1}\}$ to $\{\gamma_0, \ldots, \gamma_{n-1}\}$ is a twisting followed by a discrete Fourier transform, it is invertible; so our two spaces must be one and the same.

(2) Given an mth degree polynomial f, to find the best ridge sum approximation g of the form above, we must minimize

$$\|f - g\|^2 = \sum_r b_{mr}^2 |\gamma_r(f) - \gamma_r(g)|^2 \|J_{r,\,m-r}\|^2$$

over g. If $f = J_{r_0,\,m-r_0}$, then $\gamma_r(f) = \delta_{r0}^r/b_{mr_0}$ and the task is to find the n-periodic (complex-valued) function γ minimizing

$$RSS = m!\sum_r |\gamma_r - \delta_{r_0}^r/b_{mr_0}|^2 b_{mr}.$$

The solution $\hat{\gamma}$ must have the form

$$\hat{\gamma}_r = \begin{cases} 0, & \text{if } r \not\equiv r_0[n], \\ c, & \text{if } r \equiv r_0[n], \end{cases}$$

and RSS reduces to the quadratic

$$RSS/m! = b_{mr_0}|c - 1/b_{mr_0}|^2 + |c|^2\left(\left[b_{mr_0}\right] - b_{mr_0}\right).$$

Minimization shows that $c_{\min} = 1/[b_{mr_0}]$ and leads to the claimed expressions for $\beta_r = \gamma_r b_{mr}$ and $RSS/\|J_{r_0, m-r_0}\|^2$. □

PROOF OF THEOREM 6.1. Let $f \in \tilde{A}_{pq}$ and define

$$Q_\nu^0 f = \sum_{|l-k|=\nu} c_{kl} J_{kl}, \qquad Q_\mu f = \sum_{\nu=0}^{\mu} Q_\nu^0 f, \qquad \tilde{Q}_\mu f = (I - Q_\mu) f.$$

Thus, Q_μ is projection on the space of functions of angular degree at most μ. We have

$$\|R_{n,m} f - f\|^2 \le 2\|R_{n,m} Q_\mu f - Q_\mu f\|^2 + 2\|R_{n,m} \tilde{Q}_\mu f - \tilde{Q}_\mu f\|^2.$$

Since $R_{n,m}$ is a projection, the second term is bounded by

$$2\|\tilde{Q}_\mu f\|^2 \le 2B_q \mu^{-2q}, \tag{6.3}$$

where we have used (5.4). Now write $f_{m\mu} = P_m Q_\mu f$ for the projection of $Q_\mu f$ onto polynomials of degree at most m. We have, again since $R_{n,m}$ is a projection,

$$\|R_{n,m} Q_\mu f - Q_\mu f\|^2 \le 2\|R_{n,m} f_{m\mu} - f_{m\mu}\|^2 + 2\|(I - P_m) Q_\mu f\|^2.$$

Since P_m and Q_μ commute and $\|Q_\mu\| \le 1$, we may apply Lemma 3.2 to find that the second term above is bounded by twice

$$\|Q_\mu\|^2 \|f - P_m f\|^2 \le B_p m^{-p}. \tag{6.4}$$

Equations (6.3) and (6.4) establish that terms in the J_{kl} expansion of $R_{n,m} f - f$ are small outside the shaded region in Figure 4.

To attack $f_{m\mu}$, note that $R_{n,m} J_{r_0, m-r_0}$ contains only frequencies which differ from those of $J_{r_0, m-r_0}$ by integer multiples of $2n$. If $\mu < n$, there is then no overlap in frequencies when $R_{n,m} - I$ is applied to the terms in

$$f_{m\mu} = \sum_{\substack{|l-k| \le \mu \\ l+k \le m}} c_{kl} J_{kl}.$$

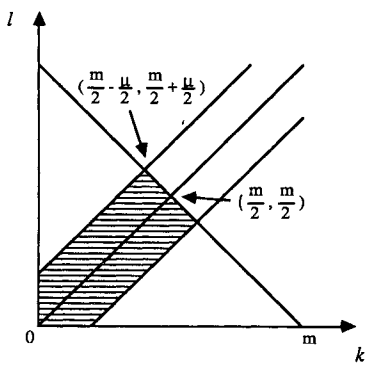

FIG. 4. *Hatched area: Set of J_{kl} basis elements on which quality of ridge approximation (by $R_{n,m}$) is studied.*

Hence from Lemma 6.2 and then Lemma 3.1,

$$\begin{aligned}\|R_{n,m}f_{m\mu} - f_{m\mu}\|^2 &= \sum |c_{kl}|^2 \|R_{n,m}J_{kl} - J_{kl}\|^2 \\ &= \sum |c_{kl}|^2 P(S_{k+l} \neq l - k | S_{k+l} \equiv l - k [2n]) \|J_{kl}\|^2 \\ &\leq C_1 \sum_{\substack{|l-k| \leq \mu \\ |k+l| \leq m}} \sqrt{k+l}\, e^{-2n(n-|l-k|)/(k+l)} |c_{kl}|^2 \|J_{kl}\|^2 \\ &\leq C_1 \sqrt{m}\, e^{-2n(n-\mu)/m} \|f_{m\mu}\|^2.\end{aligned}$$

Let now $m(n) = n^2/\delta \log n$ and $\mu = n\varepsilon$, for $\varepsilon < \frac{1}{2}$. On collecting terms we have

$$\|R_{n,m}f - f\|^2 \leq B_q (n\varepsilon)^{-2q} + B_q \left(\frac{n}{\sqrt{\delta \log n}} \right)^{-2p} + C_1 B_0 \frac{n^{1-2\delta(1-\varepsilon)}}{\sqrt{\delta \log n}}.$$

So for δ large enough,

$$\|R_{n,m}f - f\|^2 \leq C \left(\frac{n}{\log n} \right)^{-2(q \wedge p)} \leq C' \left(\frac{N}{\log^2 N} \right)^{-q \wedge p/1.5},$$

where $C = C(B_q, B_p, B_0)$ and $N = nm(n)$ is the number of coefficients used. For the case $q \in [p/2, p)$, the better bound given in the theorem is obtained by going through the above paragraph with $m(n) = n^{2q/p}$. □

7. Kernel smoothing and Laplacian smoothness. This section presents results on KA to complement those on RA. There are two basic ideas. First, that the roles played by harmonics and radials for RA are reversed for KA. Harmonics are well approximated by kernel procedures; indeed, because of the mean-value property of harmonics, convolution with an isotropic kernel leaves a harmonic unchanged. On the other hand, radials are a sort of worst case for the kernel method. The following formula gives the relative squared error for smoothing J_{kl} with a Gaussian kernel of bandwith σ:

$$\frac{\|J_{kl} * \Phi_\sigma - J_{kl}\|^2}{\|J_{kl}\|^2} = \sum_{j=1}^{k \wedge l} \binom{l}{j} \binom{k}{j} \sigma^{4j}. \tag{7.0}$$

This formula provides useful insight. Notice that if $l = 0$, so that $k \wedge l = 0$, the sum is empty: There is no error at harmonics. On the other hand, for a given (even) degree $m = k + l$, the sum will be largest at $k = l = m/2$: Radials are a worst case amongst all degree m polynomials for smoothing with a Gaussian kernel. The formula was used in preparing the contour plot Figure 2b.

The second main idea is that via smoothness classes, one can generalize results for radials and harmonics. For kernel smoothing, the key notion is *Laplacian smoothness*. For an appropriately chosen kernel ψ, Laplacian smoothness beyond that guaranteed by ordinary Cartesian smoothness will provide a better rate of approximation than that provided by ordinary Cartesian smoothness.

The results obtained are thus in direct analogy to those for RA. In fact the theory below is developed for more general classes of isotropic kernels, which for technical convenience are assumed to have compact support.

Let $\psi(|x|)$ be a radial (isotropic) kernel on $\mathbb{R}^d$ with the properties:

$$\text{(i)} \qquad \psi = 0, \quad \text{if } |x| > 1,$$

$$\text{(7.1) (ii)} \qquad \int \psi(|x|)\, dx = 1,$$

$$\text{(iii)} \qquad \int |x|^{2i}\psi(|x|)\, dx = 0, \quad i = 1, \ldots, r-1.$$

We shall denote by $\mathscr{K}_r$ the class of such kernels. Variation diminishing properties of the kernel t^i show that $t \to \psi(t)$ must have at least $t-1$ 0's for $t \in [0, \infty)$ and hence at least $[t/2]$ sidelobes [Karlin (1968) and Brown, Johnstone and McGibbon (1981)]. For functions $f \in L^2(\Phi_d)$, we wish to study the bias of kernel smoothers based on ψ as a function of bandwidth σ. So set $\psi_\sigma(|x|) = \sigma^{-d}\psi(|x|/\sigma)$ and

$$f *_\sigma (x) = f * \psi_\sigma(x) = \int f(x - \sigma z)\psi(|z|)\, dz.$$

In fact, the choice to work within $L^2(\Phi_d)$ below is made only to facilitate comparison with the results for two-dimensional sums of ridge functions in earlier sections: The conclusions below pertain to both approximation of smooth functions and to approximation in $L^2(\mathbb{R}^d)$. A bothersome technical point arises because Φ is not shift-invariant: $f *_\sigma$ need not have finite norm even though $f \in L^2(\Phi)$, as is illustrated by $f(x) = [\phi(x)(1 + |x|^{d+1})]^{-1/2}$. Let S_u denote a shift operator: $(S_u f)(x) = f(x - u)$. Writing $\mathscr{H} = L^2(\Phi)$, we introduce a (non-closed) subspace $\mathscr{H}'$ [also denoted $SL^2(\Phi)$] whose members have integrable local shifts:

$$\mathscr{H}' = \{ f \in \mathscr{H} : \text{for some } h > 0,\ S_{\pm he_i} f \in \mathscr{H},\ i = 1, \ldots, d \}.$$

Let $\| f \|_h^2 = \| f \|^2 + \Sigma_{u \in hI} \| S_u f \|^2$, where $I = \{ \pm e_i, i = 1, \ldots, d \}$. Thus $f \in \mathscr{H}'$ if and only if $\| f \|_h^2 < \infty$ for some positive h. Corresponding subspaces $\mathscr{F}_p'$, $\mathscr{A}_{pq}'$, $\mathscr{L}_{pr}'$ of the smoothness classes $\mathscr{F}_p$, $\mathscr{A}_{pq}$, $\mathscr{L}_{pr}$ are defined in the obvious way. The following lemma bounds the norm of a kernel smooth of a member of $\mathscr{H}'$.

LEMMA 7.1. *If ψ has support in $\{x: |x| \le 1\}$, if $\int |\psi(z)|\, dz = B$ and $h \ge \sigma\sqrt{d}$, then*

$$\text{(7.2)} \qquad \| f *_\sigma \|^2 \le c_d B^2 e^{h^2/2} \| f \|_h^2.$$

We defer an interpretation of the following results to samples drawn from Φ until Section 9.

Stated for $L^2(\Phi_d)$, our main conclusions can be summarized as

THEOREM 7.2. (i) *If f has r L^2 Laplacians in $SL^2(\Phi)$ and $\psi \in \mathscr{K}_r$, then*

$$\| f *_\sigma - f \|^2 = C(d, \psi, r)\sigma^{4r}\|\Delta^r f\|^2(1 + o(1)).$$

(ii) *If $\tilde{\mathscr{P}}_r' = \{f \in \mathscr{H}' : \|\Delta^r f\|_1^2 \le B_0\}$, then*

$$\sup_{f \in \tilde{\mathscr{P}}_r'} \| f *_\sigma - f \|^2 \le C'(d, \psi, r, B_0)\sigma^{4r}.$$

The following simple corollary is an analog, for kernel approximation, of Theorem 6.1.

COROLLARY 7.3. *Let $\tilde{\mathscr{F}}_p' = \{f \in \mathscr{F}_p' : \Sigma_k \|D_x^k D_y^{p-k} f\|^2 < B_1\}$, $\tilde{\mathscr{P}}_{p,r} = \{f \in \tilde{\mathscr{F}}_p' : \|\Delta^r f\|_1^2 \le B_0\}$ and $\psi \in \mathscr{K}_{p/2}$. Then*

(i) $\sup_{\tilde{\mathscr{F}}_p'} \| f *_\sigma - f \|^2 \le C\sigma^{2p}$, $C = C(\psi, p, B_1)$ *and*

(ii) *if $\psi \in \mathscr{K}_{p/2} \setminus \mathscr{K}_{p/2+1}$ and $\|\Delta^{p/2} f\|^2 > 0_{\tilde{\mathscr{F}}_p'}$, then*

$$\lim_{\sigma \to 0} \sigma^{-2p} \| f *_\sigma - f \|^2 = \left(\int \psi_{p/2,1}(x)\, dx \right)^2 \|\Delta^{p/2} f\|^2 > 0,$$

where $\psi_{p/2,1}$ is defined before Corollary 7.5 below.

Together (i) *and* (ii) *imply*

$$\sup_{\tilde{\mathscr{F}}_p'} \| f_\sigma - f \|^2 \asymp C\sigma^{2p}.$$

(iii) *If $p/2 \le r \le p$ and $\psi \in \mathscr{K}_r$, then*

$$\sup_{\tilde{\mathscr{P}}_{p,r}} \| f *_\sigma - f \|^2 \le C\sigma^{4r}, \qquad C = C(\psi, r, B_0, B_1).$$

REMARK. Part (i) of the corollary follows from (ii) of the theorem because Cartesian smoothness of order p guarantees Laplacian smoothness of order $p/2$.

Part (ii) says (roughly) that if $\psi(t)$ has no more than $p/2 - 1$ 0's in $(0, \infty]$, then the rate of convergence over $\tilde{\mathscr{F}}_p'$ is sharp, for example, a radial polynomial J_{kk} with degree $k > p/2$ has $\Delta^{p/2} J_{kk} = c_{kp} J_{k-p/2,\, k-p/2} \not\equiv 0$ and is thus approximated at rate exactly σ^{2p}.

Part (iii) represents the analog of the ridge function approximation result (6.1). It recovers a faster rate, σ^{4r}, for functions in the subclass $\tilde{\mathscr{P}}_{pr}'$ of $\tilde{\mathscr{F}}_p'$ having extra Laplacian smoothness. To obtain these higher rates however, kernels ψ having more (at least $r - 1$) 0's are also needed.

In Section 6, the usual rate of approximation was expressed in terms of best approximation using a polynomial with N free coefficients. Since kernel smoothing is not a projection, this benchmark is not available, and is replaced by worst case statements of parts (i) and (ii) of the corollary.

PROOF OF FORMULA (7.0). Although (7.0) provides useful information, it is somewhat tangential to the main development in this section, so its derivation is

merely sketched. It is a remarkable fact (proved by applying the Fourier transform twice) that Gaussian convolution on $\mathscr{P}_m^0$ is equivalent to rescaling:

$$J_{kl} * \Phi_\sigma(w, \overline{w}) = \left(\sqrt{1-\sigma^2}\right)^{k+l} J_{kl}\left(\frac{w}{\sqrt{1-\sigma^2}}, \frac{\overline{w}}{\sqrt{1-\sigma^2}}\right).$$

To expand the right side terms of $J_{k',l'}(w, \overline{w})$, we exploit the connection (4.7) between $T_{k,l}(r)$ of (4.6) and the Laguerre polynomials $L_k^{l-k}(r^2/2)$ and use a scaling formula for the latter [Szegö (1939), page 387]. The result is

$$J_{kl} * \Phi_\sigma = \sum_{j=0}^{k \wedge l} \binom{k}{j}\binom{l}{j} j!(2\sigma^2)^j J_{k-j,\, l-j},$$

from which (7.0) follows. □

PROOF OF LEMMA 7.1. As in previous arguments, and via the Cauchy–Schwarz inequality,

$$\int f^2 *_\sigma \phi \le B \int dx\, \phi(x) \int_{|z| \le 1} |\psi(z)| f^2(x - \sigma z)\, dz$$

$$= B \int_{|z| \le 1} dz |\psi(z)| \int dy\, \phi(y + \sigma z) f^2(y).$$

Now exploit the inequality $\sigma|yz| \le \sigma|z|\,|y| \le \sigma\sqrt{d}\, \max_i |y_i| \le h \max_i |y_i|$ to find

$$\phi(y + \sigma z) \le c_d e^{-|y|^2/2 + h \max |y_i|}$$

$$\le c_d e^{h^2/2} \sum_{u \in hI} \phi(y + u).$$

Substitution into the previous inequality yields the result. □

To prepare for the proof of Theorem 7.1, we establish some lemmas.

LEMMA 7.4. *Suppose that $\psi(|z|)$ has compact support and $\int \psi(|z|)\, dz = I(\psi)$. If f has two continuous derivatives, then*

$$f *_\sigma(x) - f(x) I(\psi) = \sigma^2 \Delta f * \psi_{1,\sigma}(x)$$

$$= \sigma^2 \int \Delta f(x - \sigma z) \psi_{1,\sigma}(|z|)\, dz, \tag{7.3}$$

where

$$\psi_1(s) = (T\psi)(s) = \int_s^\infty dr\, r^{d-1} \psi(r)[\mathscr{G}(r) - \mathscr{G}(s)]$$

$$= \int_{\{|z| \ge s\}} dz\, \psi(|z|)[\mathscr{G}(|z|) - \mathscr{G}(s)], \tag{7.4}$$

$$\mathscr{G}(r) = \begin{cases} r^{2-d}/(2-d), & d > 2, \\ \log r, & d = 2, \end{cases}$$

$T\psi$ has compact support and $\psi_{1,\sigma}(|z|) = \sigma^{-d}\psi_1(|z|/\sigma)$.

If $\psi(|x|)$ vanishes for $|x| > R$, then so does $T\psi$.

If $\psi(|x|)$ is absolutely integrable, then so is $(T\psi)(|x|)$.

REMARK. If we formally take Fourier transforms of both sides of (7.3), we get

$$\widehat{f\psi_\sigma - \delta} = -4\pi^2\sigma^2|\zeta|^2\hat{f}\hat{\psi}_{1,\sigma},$$

where δ is the delta function. On cancelling $\hat{f}$, we are led to the definition,

$$\psi_1 = \mathfrak{F}^{-1}\big(-\widehat{\psi - \delta}/4\pi^2|\zeta|^2\big) = (-\Delta)^{-1}(\psi - \delta),$$

where $\mathfrak{F}^{-1}$ is inverse Fourier transform. This approach can be made to work: Δ^{-1} is a Riesz transform [Stein (1970), Chapter 3], but we shall use the direct approach given below.

PROOF OF LEMMA 7.4. This is essentially a consequence of Green's theorem. First, set $\sigma = 1$ and write

$$f * \psi(x) - f(x)I(\psi) = \int_0^\infty dr\, r^{d-1}\psi(r)\int_{|\omega|=1} d\omega\,[f(x - r\omega) - f(x)],$$

where $z = r\omega$, $\omega \in S^{d-1}$. Let $\nu = \nu(x)$ denote the (outward) unit normal vector to S^{d-1} at $x \in S^{d-1}$. Using Green's theorem and the assumed smoothness of f and denoting by $B_\rho = B_\rho(x)$ the ball of radius ρ centered at x, we get

$$\begin{aligned}
\int d\omega\,[f(x - r\omega) - f(\omega)] &= \int_0^r d\rho \int_{|\omega|=1} \frac{\partial}{\partial\rho} f(x - \rho\omega)\, d\omega \\
&= \int_0^r d\rho\, \rho^{1-d} \int_{\partial B_\rho} \frac{\partial f}{\partial \nu} \\
&= \int_0^r d\rho\, \rho^{1-d} \int_{B_\rho} \Delta f \\
&= \int_{B_r} dz\, \Delta f(z)[\mathcal{G}(r) - \mathcal{G}(|z - x|)].
\end{aligned}$$

Integrate over r and interchange orders of integration: The end product reduces to (7.3). For general σ, it is easy to show directly from (7.4) that $(T\psi_\sigma) = \sigma^2(T\psi)_\sigma$, which yields the general form of (7.3). □

That $T\psi$ vanishes outside the support of ψ is evident from the definition. Now $T\psi(|x|)$ is continuous, except possibly at 0, and so to check integrability, it will suffice to bound the growth of $T\psi$ at 0. If $\int|\psi| = A$, then

$$\begin{aligned}
|T\psi(s)| &\le \int_s^\infty \frac{dv}{v^{d-1}} \int_{|z|>v} |\psi|(|z|)\, dz \\
&\le A\mathcal{G}(s),
\end{aligned}$$

which is integrable in a neighborhood of 0.

Denote by T^r the rth iterate of the map defined in (7.4) above, and write $\psi_{r,\sigma}(x) = \sigma^{-d}(T^r\psi)(x/\sigma)$.

COROLLARY 7.5. *If $f \in C^{2r}$ and $\psi \in \mathscr{K}_r$, then*

$$f *_\sigma - f = \sigma^{2r}\Delta^r f * \psi_{r,\sigma}.$$

PROOF. If in Lemma 7.4 we substitute $f(x) = |x|^{\alpha+2}$, $\sigma = 1$, and evaluate (7.3) at $x = 0$, we obtain

$$(d + \alpha)(2 + \alpha)\int T\psi(|z|)|z|^\alpha \, dz = \int \psi(|z|)|z|^{\alpha+2} \, dz. \tag{7.5}$$

When iterated, this identity shows that $\int T^i\psi(|z|)\,dz = 0$ for $i = 1, \ldots, r-1$ [from (iii) of (7.1)].

We now apply induction, the preceding lemma establishing the case $r = 1$. Writing out the result for $r - 1$, using (7.5) to check that $I(\psi_{r-1}) = 0$ and then applying (7.3) to ψ_{r-1}, we find

$$\begin{aligned} f * \psi_\sigma - f &= \sigma^{2r-2}\int \Delta^{r-1} f(x - \sigma z)\psi_{r-1}(|z|)\,dz \\ &= \sigma^{2r-2}\sigma^2 \int \Delta(\Delta^{r-1} f)(x - \sigma z) T\psi_{r-1}(|z|)\,dz \\ &= \sigma^{2r}\Delta^r f * \psi_{r,\sigma}. \end{aligned}$$

□

PROOF OF THEOREM 7.2. For the first part, we begin by noting that the equality

$$\| f *_\sigma - f \|^2 = \sigma^{4r} \| \Delta^r f * \psi_{r,\sigma} \|^2 \tag{7.6}$$

extends from $f \in C^{2r}$ to f having r L^2 Laplacians by the device of regularization: Let k be a C^∞ function of compact support, $k_\varepsilon(x) = \varepsilon^{-d}k(x/\varepsilon)$, and study the convergence of the smooth approximants $f * k_\varepsilon$ and their derivatives to f [see, e.g., Gilbarg and Trudinger (1977), Lemma 7.2, or Stein (1970), Section 3.2.2].

The details of this standard method are explained in the references cited, so we make only comments on the modifications necessitated by the use of a nonshift invariant measure (Φ_d). The two useful tools are:

(i) (convergence of regularizations) If $f(x) \in \mathscr{H}'$ and $k(x)$ has compact support with $\int k(x)\,dx = 1$, $\int |k(x)|\,dx < \infty$, then $\| f * k_\varepsilon - f \|^2 \to 0$.

A standard proof [see, e.g., Stein–Weiss (1971), Chapter 1] uses the bound

$$\| f * k_\varepsilon - f \| \le \int w_f(s\varepsilon)|k(s)|\,ds, \tag{1}$$

where $w_f(t) = \| S_t f - f \| \le \| S_t f \| + \| f \|$. In $\mathscr{H}'$, the shift $S_t f$ is continuous at $t = 0$, and the assumption that k has compact support is a convenient way to ensure that the integrand in (1) can be dominated by $(1 + \eta)|k(s)|$ for small ε. These facts together imply the claimed convergence.

(ii) (continuity of convolution) Let $k(x)$ have $\int k = 1$, $\int |k| < \infty$ and support contained in $\{x: |x| \le h/\sqrt{d}\}$. If $\| f_n - f \|_h \to 0$, then $\| f_n * k - f * k \| \to 0$. (This is an immediate consequence of Lemma 7.1.)

To prove the theorem, it now suffices to apply the regularization argument to (7.6) as $\sigma \to 0$. Lemma 7.4 shows that $\psi_{r,1}$ is an absolutely integrable kernel, and hence property (i) above shows that

$$\| \Delta^r f * \psi_{r,\sigma} \|^2 \to \left(\int \psi_{r,1}(x)\, dx \right)^2 \| \Delta^r f \|^2,$$

which implies part (i) of the theorem. [Note that the constant can be evaluated, if desired, from (7.5).] Part (ii) of the theorem can be read off from (7.6) and Lemma 7.1. □

8. Formal interpretation. The results given so far admit of interesting formal interpretations. We shall pursue these here: Difficulties and cautions are postponed to Section 9.

Three kinds of approximation procedures. The paper has studied procedures based on local averaging (KA) and on reconstruction from projections (RA). The paper has also considered, implicitly, a third approximation scheme—polynomial series approximation, represented by the operator P_N.

The differences between these procedures are perhaps clearer if we regard them as (approximately) orthogonal series procedures. The proof of Theorem 6.1 shows that RA behaves somewhat like an orthogonal series scheme using the J_{kl} basis, where terms corresponding to large values of $k - l$ or $k + l$ are dropped from the expansion. Similarly, KA behaves like an orthogonal series scheme where terms corresponding to large values of kl are dropped. Finally, the operator P_N drops terms with large values of $k + l$.

These three truncation strategies are appropriate in different circumstances. The first is appropriate when the function to be recovered has most of its energy concentrated at basis elements with small values of $k - l$; the second, when the energy concentrates at small values of kl. Membership in $\mathscr{A}_{pq}$ results in the first condition; membership in $\mathscr{L}_{pr}$ results in the second. The third strategy is appropriate when the energy is only known to concentrate near the origin. Unlike the other two strategies, it makes no judgement about whether the energy is more likely to concentrate near radial terms or near harmonic terms. Membership in $\mathscr{F}_p$ makes this strategy appropriate.

Such differences between procedures do not occur in one-dimensional smoothing. The main one-dimensional procedures all have a representation as local averaging procedures (delta-sequence representation). Equivalently, they can be represented in the frequency domain as cutting off all high frequency terms beyond a certain point. Thus, the orthogonal series estimate is very nearly a kernel estimate in one dimension. The spline estimate is nearly a kernel estimate, and so on.

By contrast, in high dimensions there are several notions of oscillation (Cartesian, angular and radial) and several strategies for truncating high

oscillation terms. In this sense, the high-dimensional smoothing theory can be much richer than the one-dimensional theory.

Complementarity of the procedures. A basic problem in communicating results on rates of convergence is their use of the smoothness index p. In a given case, how does one know if the relevant value of p is 1, 2, or ∞? The notion of effective dimensionality alluded to in the Introduction provides a way to compactly summarize these results without using p. To develop these ideas, we need the following definitions. Let $p^*(f)$ denote the smoothness of f as follows:

$$p^*(f) = \sup\left\{p: \sum (k+l)^p |c_{kl}(f)|^2 \|J_{kl}\|^2 < \infty\right\}.$$

Let $r^*(f, M)$ denote the rate of convergence of method M at f as follows:

$$r^*(f, M) = \sup\left\{r: N^r \|f - M_N f\|^2 = O(1)\right\},$$

where $M_N f$ symbolizes the approximation to f via method M using complexity N. Now typically, for p-smooth functions one has the minimax rate of convergence for $L^2(\Phi)$,

$$\|f - \hat{f}_N\|^2 \sim N^{-p/d}.$$

Therefore, we can define the effective dimensionality of a function f for a method M by

$$d^*(f, M) \equiv p^*(f)/r^*(f, M).$$

In terms of effective dimensionalities we have from Theorem 6.1 that when $f \in \mathscr{A}_{pp}$, RA can be tuned to take advantage of the enhanced angular smoothness, so that (ignoring logarithmic terms)

$$d^*(\mathscr{A}_{pp}, \mathrm{RA}) = 1.5.$$

Similarly, the kernel method can take advantage of enhanced Laplacian smoothness:

$$d^*(\mathscr{L}_{pp}, \mathrm{KA}) = 1.$$

On the other hand, RA cannot be tuned to take advantage of enhanced angular smoothness nor can KA be tuned to take advantage of enhanced Laplacian smoothness:

$$d^*(\mathscr{F}_p, \mathrm{RA}) = d^*(\mathscr{L}_{pr}, \mathrm{RA}) = 2,$$
$$d^*(\mathscr{F}_p, \mathrm{KA}) = d^*(\mathscr{A}_{pq}, \mathrm{KA}) = 2.$$

These two relations say that functions with Laplacian smoothness (e.g., harmonics) are least favorable for RA, while functions with angular smoothness (e.g., radials) are least favorable for KA, among functions with a certain ordinary smoothness.

Finally, the orthogonal series procedure based on P_N is unable to take advantage of ancillary smoothness to improve over the minimax rate. Indeed, considerations such as Lemma 3.2 will show that for a function with exactly p derivatives in $\mathscr{L}_{pr}$ (and no more), the rate of approximation of orthogonal series

approximation (PA), when well-tuned, is always $p/2$; the effective dimensionality of the procedure is therefore 2, whatever kind of ancillary smoothness we consider:

$$2 = d^*(\mathscr{F}_p, \mathrm{PA}) = d^*(\mathscr{A}_{pq}, \mathrm{PA}) = d^*(\mathscr{L}_{pr}, \mathrm{PA}).$$

The authors know of no comparable results for smoothing in one dimension. There is only one type of smoothness there, and methods are not complementary: They are similar.

Analogies to shrinkage and Bayes estimation. Consider estimating a multivariate normal mean under squared-error loss. The sample mean, while minimax, can be improved upon over regions of the parameter space, and different estimators can be created with different regions of improvement. For the analogy, think of d^* as a measure of risk. The analog of the mean is the orthogonal series estimate which has a constant effective dimensionality equal to 2. On the other hand RA improves on the minimax procedure over $\mathscr{A}_{pq}$ while KA obtains a rate improvement over $\mathscr{L}_{pr}$.

Another helpful analogy is with the Bayes point of view. Compare Wahba (1978, 1983). Let the f be chosen at random from a certain function space and consider approximation of f using a complexity no greater than N. A Bayes procedure in this second setting minimizes the risk, which is now expected approximation error; expectation being taken under the distribution of f.

Without going into details, polynomial approximation is Bayes for the distribution on f, where Z_{kl} are i.i.d. standard Gaussian, $\tilde{J}_{kl} = J_{kl}/\|J_{kl}\|$ and

$$f = \sum (k + l)^{-p/2} Z_{kl} \tilde{J}_{kl},$$

at least for complexities of the form $N = \binom{m}{2}$. A prior for which KA is approximately Bayes is where

$$f = \sum (kl)^{-r/2} Z_{kl} \tilde{J}_{kl},$$

and division by 0 at $k \wedge l = 0$ is interpreted as saying that the prior is improper on the subspace of harmonics. A prior for which RA can be tuned to be approximately Bayes is generated by

$$f = \sum \left((k + l)^{p/2} + \lambda |k - l|^q \right)^{-1} Z_{kl} \tilde{J}_{kl}.$$

However, neither KA nor RA will be exactly Bayes in these cases. This means that for a given complexity a slightly better procedure is available in each case. As in the shrinkage literature, the insight such representations should give is that on the region where the prior is large, one should do well with the Bayes procedure. The hyperparameter λ in the prior affects the number of directions that are used by the ridge approximation and thus is related to the amount of shrinkage from the minimax rule.

9. Cautionary notes. The interpretations of the last section depend in an essential way on formal manipulations possible for the Gaussian distribution.

While one is accustomed to such manipulations in other cases (e.g., trigonometric series over $L^2[0, 2\pi]$), caution is needed (at least at present) in extrapolating our results to broader situations.

Rate improvement classes (RIC's). Our results show that for RA and KA, there are classes of functions on which a particular method achieves better than minimax rates of convergence. Especially in view of the Bayes analogy, there is a sense in which RIC's can be thought of as nonparametric models: When the function of interest is a member of a certain procedure's RIC, it is natural to apply that procedure. However, several comments about such interpretations are apt.

(i) RIC's need not always exist: A method may be simply minimax without being "Bayes" [relative to the risk measure $d^*(f, M)$ of Section 8]. The orthogonal series approximation operator P_N is of this type, as described in Section 7.

(ii) RIC's need not take on a mathematically elegant form. For example, tree-structured regression [Breiman, Friedman, Olshen and Stone (1984)] seems naturally adapted to dealing with functions whose gradients tend to be parallel to one of the coordinate axes. The class of such functions does not seem amenable to description in usual function space terms.

(iii) Such models need not be unique. Although information that $f \in \mathscr{A}_{pq}$ might lead one to use PPR, quite a different sort of information can also lead one to seek nonparametric function estimates made up from sums of ridge functions. Suppose that $\mathbf{Z}$ represents an infinite sequence of unobservable independent Gaussian random variables. Suppose that in terms of this unobservable $\mathbf{Z}$, the function f has the representation

$$f(Z) = \sum g_i(Z_i),$$

where the g_i are nonlinear functions of their argument. Now suppose that we can observe $\mathbf{X} = (\beta_1^t Z, \ldots, \beta_p^t Z)$ and want to construct an estimate of f from this partial information. The Gaussian distribution on Z ensures that the best estimate

$$\hat{f}(\mathbf{X}) = E\{f(\mathbf{Z})|\mathbf{X}\}$$

has the form $\hat{f} = \sum_k h_k(\alpha_k^t X)$. That is, the conditional expectation of the unobservable function is a sum of ridge functions.

Plausibility of ancillary smoothness. We have shown that RA and KA are able to take advantage of the ancillary smoothness represented by membership in $\mathscr{A}_{pq}$ and $\mathscr{L}_{pr}$. Why should such smoothness occur in cases of practical interest?

It was hinted earlier that functions with good tail behavior are angularly smooth. To see this, note that

$$D_\theta f|_{\substack{x = r\cos\theta \\ y = r\sin\theta}} = r(-\sin\theta, \cos\theta)^t \nabla f,$$

so that

$$|D_\theta f| \leq r|\nabla f|.$$

Thus, if f is small for large values of r, for example, if $f = 0$ for $r > R$, then it will have as many angular $L^2(\Phi)$ derivatives as ordinary $L^2(\Phi)$ derivatives. Thus the functions in $\mathscr{F}_p$ of compact support are actually in $\mathscr{A}_{pp}$. More generally, we should expect functions with decent tail behavior to belong to $\mathscr{A}_{pp}$ whenever they belong to $\mathscr{F}_p$. For example, if ∇f grows no faster than a polynomial, then the last display shows that $D_\theta f$ is in $L^2(\Phi)$.

When should a function have more Laplacian derivatives than strictly guaranteed by the number of ordinary derivatives? The authors have no simple intuitive condition which guarantees this. One can construct such functions explicitly as solutions to the heat equation

$$\Delta f = h,$$

where h is a strategically chosen element of $L^2(\Phi)$. In geophysics, meteorology and oceanography, one has, of course, observations on functions of precisely this form. Compare Wahba and Wendelberger (1980). In those cases, one observes such data at sites scattered more or less uniformly on a sphere, so the present analysis does not apply directly. However, the qualitative superiority of kernel-based methods over projection-based methods should carry through.

In the authors' opinion, the two classes have very different plausibilities. In many real cases one expects the function to be estimated to have rather simple behavior in the tails. Then a PPR-type representation of the function makes sense. In physical problems, one might expect to see a function satisfying a heat equation, and then a kernel method may be appropriate. However, the tail condition seems more likely to hold in most applications of high-dimensional smoothing.

Smoothness condition as boundary conditions. Care is needed with the preceding story. The integral smoothness measures discussed in this paper have two components. The first requires the existence of ordinary continuous derivatives to some order and the second calls for finiteness of certain L^2 norms of these derivatives. It is the latter aspect that the quantitative theory of rates of convergence addresses. For example, a harmonic function is locally $\mathbf{C}^\infty$, so it possesses local Cartesian and angular derivatives of all orders. However, these derivatives may fail to be in $L^2(\Phi)$, because harmonic functions can grow so rapidly at ∞. For example, if $f \in \mathscr{L}_{pr}$, $r > p/2$, and has exactly p Cartesian derivatives, then some derivative of f necessarily grows faster than any polynomial. This follows from Proposition 5.2: Indeed, as described above, slow growth at ∞ certainly suffices for membership in $\mathscr{A}_{pp}$.

As a result of this feature, our theory is more sensitive to boundary behavior than one would like. However similar problems occur whenever there is a boundary, even for the unit disc. Furthermore, this phenomenon is likely to become more important in higher dimensions, since more points will be close to some part of the boundary.

Use of global L^2 error. At least three objections can be made against global L^2 error as a measure of approximation. The first, related to the discussion of previous subsections, is due to Rice and Rosenblatt (1983), who pointed out that in many cases the global L^2 error only measures what happens at the boundary; on a compact set interior to the domain the rate of convergence of the error can be much better. The second is that L^2 convergence is very weak; Askey and Wainger have shown that Hermite polynomials converge in L^2 but not in L^1 or L^∞, for example. The third is that L^2 error can be a poor indicator of the visual quality of approximation. For example, if f and the kernel are radial, an estimate of f obtained by kernel regression will have a radial shape. By contrast the quantitatively better estimate obtained using n_d equispaced directions will not have a radial average shape: It will be a sum of n_d ridge functions. For visual interpretation, this ridge approximation may contain misleading artifacts. Indeed, the authors have made plots of ridge approximants to radial functions and found unpleasant ripples in the best $L^2(\Phi)$ approximation. Larry Shepp has also stressed this undesirable feature of best L^2 approximation in personal communications. Of course, if there is no spatial interpretation of the x variables or if numerical prediction is a goal, then L^2 measures of error may well be very reasonable.

Translating approximation rates into estimation rates. Our interest in $L^2(\Phi)$ comes from the fact that with (X_i, Y_i) data randomly distributed according to

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i,$$

where the $(\mathbf{X}_i)$ are normally distributed, the $L^2(\Phi_2)$ norm (squared) has interpretations in terms of the predictive mean-squared error and the integrated squared bias. To translate our approximation results to the estimation setting, unfortunately, takes some effort.

For PPR there are several aspects to this translation problem. The first is that for PPR, the projection directions depend on the data, whereas our approximation results use equispaced sets of directions. Since optimally chosen directions for a given function will do better than equispaced directions, our results give an upper bound on the approximation error of an optimal procedure which actually does a directional pursuit. On the other hand, it is not clear that straightforward implementations of PPR—using the so called greedy algorithm—do as well as equally-spaced directions would do. Also, for radials and some harmonics, equally-spaced directions are optimal; so our results should give an accurate picture in that case.

The second problem of translation for PPR is that of specifying which one-dimensional smoother is being applied to the projections. For technical reasons, it is easiest for us to discuss smoothing in each projection via an orthogonal series estimate. Because many one-dimensional smoothers are so similar, our results probably carry through, for example, to local linear fits as used in the super smoother routine employed by Friedman and Stuetzle (although, the cross-validation used in that procedure produces additional complications).

A third remark, which we owe to Chuck Stone, is that even for fairly general assumptions on the distribution of the independent variables $\mathbf{X}_i$, the variance term in (9.1) below will be inflated due to boundary effects. For example, if the density $\{X_i\}$ is uniform on the unit disc in $\mathbb{R}^2$, the density of any projection $\alpha^t X$ is proportional to $(1 - (\alpha^t x)^2)^{1/2}$, so that there will be relatively little data for $\alpha^t x$ near ± 1.

In any event, the approximation-theoretic results presented here should translate into sampling-theoretic phenomena. To see how this works, note that if the directions used are equally spaced and chosen independently of the data, then the approximation errors used in this paper are essentially the bias term in the usual expansion of the mean-squared error:

$$\text{Integrated-MSE} = \text{bias}^2 + \text{variance}. \tag{9.1}$$

In our setting, the variance term is essentially proportional to N/n, where N is the number of degrees of freedom used by the prediction function (e.g., number of coefficients) and n is the number of observations. The best rate ρ at which the mean-squared error can go to 0 is derived from a tradeoff between the bias^2 and the variance terms. Formally, one has

$$n^{-\rho} \approx \min_{N_*(n)} \{N_*^{-r} + N_*(n)/n\}.$$

This tradeoff is controlled by the approximation rate r; one has the formal result

$$r = p/d \to \rho = \frac{p}{p + d}.$$

Approximation rates go over into estimation rates. In fact, this translation can be rigorously established, at least in the cases when the measure on X is known to be Gaussian.

PROPOSITION 9.1. *Let $Y_i = f(X_i) + \varepsilon_i$, where f is bounded and X_i and ε_i are independent and the pairs (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d. with $X_i \sim \Phi_2$ and $E\varepsilon_i = 0$. Let an estimate $R_{n_d, m} f$ of f be obtained from the data by fitting orthogonal polynomial ridge functions (each of degree m) in n_d equally spaced directions (so that $N = mn_d$ coefficients are used). Then*

$$\|\hat{R}_{n_d, m} f - R_{n_d, m} f\|^2 = O_p(N/n)$$

uniformly amongst all functions f bounded by M (say). Thus, if $d^*(f, \text{PA})$ is as defined in Section 8, there exist choices $m(n), n_d(n)$ making the rate of convergence (in probability) to 0 of the integrated mean-squared error $\rho = p/(p + d^*(f, \text{PA}))$.

In short, improvements in the effective dimensionality of the approximation problem transform to improvements in the effective dimensionality of the estimation problem. A proposition of wider validity would of course use a smoothing method not relying so heavily on Gaussian measure.

Hall (1987) has recently obtained a number of convergence rate results for the sampling theory of PPR, creating an opportunity for a synthesis of the sampling

and approximation aspects.

Translating results on KA. There are difficulties translating results on KA into results on estimation. In the first place, it is not clear what the estimation procedure analogous to kernel approximation is. Arguments can be made in favor of either local polynomial fits or kernel regression. Also, an important aspect of the approximation setting—the ability of isotropic kernel smoothing to reproduce harmonics—may be lost in the estimation case. Indeed, the straightforward KR estimate

$$\hat{f}(x) = \sum_i Y_i K_\lambda(x - X_i) \Big/ \sum_i K_\lambda(x - X_i)$$

corresponds to the approximation operator

$$f^{(\lambda)}(x) = \int K_\lambda(x - x') f(x') g(x')\, dx' \Big/ \int K_\lambda(x - x') g(x')\, dx',$$

where g is the density of x. This operator does not generally reproduce harmonics, unless g is uniform. For nonuniform g, one can obtain a KR-type estimator formally similar to KA via a sort of prewhitening to adjust for the nonuniformity. However, the kernel weighting used must then be position-dependent, as in

$$\hat{f}(x) = \frac{\Sigma_i Y_i K_\lambda(x - X_i)/g(X_i)}{\Sigma_i K_\lambda(x - X_i)/g(X_i)}.$$

This weighting is not reasonable in practice unless one knows g to very high accuracy.

Problems in translating results on KA into results on KR are not new. Indeed, there is a sense in which KR is not the right analog for sampling situations of KA in approximation situations. Actually, weighted local polynomial fitting behaves very like KR in good cases, but makes better sense in general. For example, Stone (1982) remarks that kernel procedures can have problems when the x density is not sufficiently smooth. In his case, one can use local polynomial fits when the density is not smooth and still get good performance. In this sense, local polynomial fits are perhaps a more appropriate analog of KA for sampling situations. However, we have not checked whether local polynomial fits provide the sort of dimensionality reductions under Laplacian smoothness that the KA results suggest may be possible.

Conclusion. These results indicate that, in large samples, Gaussian PPR should produce function estimates which behave well when the underlying function is angularly smooth. The results may be even better than our upper bounds suggest, if an optimal set of directions is obtained via a good pursuit algorithm. The angular smoothness condition will be met if the function to be estimated has tame tails.

On the other hand, kernel regression might seem to be appropriate for harmonic functions and solutions to the inhomogeneous heat equation. However, the actual performance in practice will not be as spectacular as a naive reading of

the approximation results might suggest. In fact we doubt that the function class $\mathscr{L}_{pr}$ for $r > p/2$ admits of any local smoother with good global L^2 properties. Functions in $\mathscr{L}_{pr}$ grow exponentially fast and so sampled data from such functions have very sporadic behavior. Occasionally an extreme X observation will give rise to an extreme Y observation swamping the rest of the data in terms of contributions to the total sum of squares.

Generalization to other measures. Our techniques of calculation rest heavily on the use of Gaussian measure on X. Consider $L^2(D)$, D denoting uniform measure on the unit disc in $\mathbb{R}^2$. There are Zernike polynomials analogous to the $\{J_{kl}\}$ basis, and Logan and Shepp (1975) and Hamaker and Solmon (1978) have developed projection formulas and many further results. From these formulas, it does not seem that harmonics and radials play the extreme roles in $L^2(D)$ that they assumed for $L^2(\Phi)$. Some progress, based on results from the tomography literature [e.g., Davison and Grunbaum (1981)] is described without proof in Donoho and Johnstone (1986). Further work is needed here, as well as in the situation of heavier tailed distributions than Φ: We do not know, in this case, if the differences between radial and harmonic functions become even more pronounced.

Effective dimensionality. The effective dimensionality approach of Section 8 requires care: The effective dimension is not an invariant. Its definition depends on the measure placed on X. For measures such as the uniform which have compact support, one has minimax approximation rates of the form

$$r = 2p/d$$

rather than $r = p/d$ as encountered in the Gaussian case. Presumably, with heavy-tailed measures one encounters cases where the minimax approximation rates are even of the form

$$r = cp/d$$

for $c < 1$. In each case one is forced into a different definition of effective dimensionality: $d^* = 2p/r^*$, $d^* = p/r^*$ and $d^* = cp/r^*$. The authors prefer to think of d^* as a quantity to be compared only within cases pertaining to the same measure on X.

10. Further and related work.

Extensions. The paper raises many questions that we hope to pursue further.

(i) We have not quantified the extent to which behavior at ∞ dominates the rates reported here.

(ii) How is L^2 approximation error distributed with distance from the origin?

(iii) What happens for other X-distributions (compare discussion in Section 9)?

Study of faster rates. We have described RIC's for dimensionalities in the range $[1.5, 2)$. What is the structure of RIC's for effective dimensionalities $d^* < 1.5$? We have two observations. First, the results obtained in this paper were obtained without choosing the directions to depend on the function: This will be necessary to get faster rates. Also, one will not be able to get dimensionality better than 1.5 on any class including the radials. [This follows from Beurling-type theorems on orthogonally invariant subspaces of $L^2(\Phi)$.] It seems that functions at which faster rates are possible will have to be nonisotropic, involving specific directional biases. Second, we believe that faster rates will require a certain kind of lacunarity of the function, namely that the Fourier transform of the function will, for high frequencies, concentrate on sets of small angular measure. For example, the rate $d^* = 1$ can be obtained for functions which are sums of infinitely many ridge functions in distinct directions, so long as the norms of the terms decline exponentially fast.

Higher dimensions. What happens when d is 3 or 4? We believe that radials are still well approximated and harmonics poorly. However, the equally-spaced sets of directions, so essential in the analysis of this paper, are not generally available in higher dimensions. It is at least possible to compute the analogs of Tables 1 and 2 for dimensions 3 and 4 in some special situations (directions given by vertices of regular and semiregular polytopes).

Connections to other work.

Tomography. Some connections of PPR to tomography have been described in Huber (1985). Some of our results for RA, based on equally-spaced directions, are closely related in approach to the results obtained by tomographers. Indeed the results in Section 2 above correspond in a direct way to results of Logan and Shepp (1975) for the unit disc. See also Logan (1975).

Additive models. Stone (1985) has also found classes of functions on which improvements over minimax rates are possible. His conditions are structural, requiring that the regression model be additive, namely a sum of functions of individual coordinates. In such cases, the effective dimensionality is one. However, the procedures are not consistent if the additive model does not hold, instead they approach the closest additive approximation to the true function. By contrast, the angular differentiability conditions we use pick out functions with a certain sort of smoothness, and RA achieves a dimensionality reduction (but only to $d^* = 1.5$, not $d^* = 1$) on these functions, while still being globally consistent.

Acknowledgments. The authors would like to thank Persi Diaconis, Charles Stone and Grace Wahba for various discussions and the referees and Associate Editor for their comments. The idea that PPR and biharmonic splines might

be complementary methods arose in conversations with Wahba at the AMS Summer Conference on Multivariate Analysis held at Bowdoin College in 1984. The first author would like to thank the Statistics Department at the Hebrew University for hospitality and the use of computing facilities (used, for example, to compute Table 1).

REFERENCES

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, Calif.

BRILLINGER, D. R. (1981). *Time Series Data Analysis and Theory*. Holden-Day, San Francisco.

BROWN, L. D., JOHNSTONE, I. M. and MCGIBBON, K. B. (1981). Variation diminishing transformations: A direct approach to total positivity and some of its statistical applications. *J. Amer. Statist. Assoc.* **76** 824–832.

COLLOMB, G. (1981). Estimation non-parametique de la régression: Revue bibliographique. *Internat. Statist. Rev.* **49** 75–93.

DAVISON, M. E. and GRUNBAUM, F. A. (1981). Tomographic reconstruction with arbitrary directions. *Comm. Pure Appl. Math.* **34** 77–120.

DEANS, S. R. (1983). *The Radon Transform and Some of Its Applications*. Wiley, New York.

DIACONIS, P. and SHAHSHAHANI, M. (1984). On non-linear functions of linear combinations. *SIAM J. Sci. Statist. Comput.* **5** 175–191.

DONOHO, D. L. and JOHNSTONE, I. M. (1986). Regression approximation using projections and isotropic kernels. In *Function Estimates* (J. S. Marron, ed.). *Contemp. Math* **59** 153–167. Amer. Math. Soc., Providence, R.I.

FRIEDMAN, J. H. (1985). Classification and multiple regression through projection pursuit. L.C.S. Technical Report 12, Dept. Statistics, Stanford Univ.

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

GILBARG, D. and TRUDINGER, N. S. (1977). *Elliptic Partial Differential Equations of Second Order*. Springer, Berlin.

HALL, P. G. (1987). On projection pursuit regression. Unpublished.

HAMAKER, C. and SOLMON, D. C. (1978). The angles between null spaces of X-rays. *J. Math. Anal. Appl.* **62** 1–23.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

HUBER, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13** 435–525.

KARLIN, S. (1968). *Total Positivity*. Stanford Univ. Press, Stanford, Calif.

LOGAN, B. F. (1975). The uncertainty principle in reconstructing functions from projections. *Duke Math. J.* **42** 661–706.

LOGAN, B. F. and SHEPP, L. A. (1975). Optimal reconstruction of a function from its projections. *Duke Math J.* **42** 645–660.

LORENTZ, G. G. (1966). *Approximation of Functions*. Holt, Rinehart and Winston, New York.

MCKEAN, H. P., JR. (1973). Geometry of differential space. *Ann. Probab.* **1** 197–206.

PINKUS, A. (1985). *n-Widths in Approximation Theory*. Springer, Berlin.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

RICE, J. and ROSENBLATT, M. (1983). Smoothing splines: Regression, derivatives and deconvolution. *Ann. Statist.* **11** 141–156.

STEIN, E. M. (1970). *Singular Integrals and Differentiability Properties of Functions*. Princeton Univ. Press, Princeton, N.J.

STEIN, E. M. and WEISS, G. (1971). *Introduction to Fourier Analysis on Euclidean Space*. Princeton Univ. Press, Princeton, N.J.

STONE, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 595–645.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

SZEGÖ, G. (1939). *Orthogonal Polynomials. Amer. Math. Soc. Colloq. Publ.* **23**. Amer. Math. Soc., New York.

WAHBA, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372.

WAHBA, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150.

WAHBA, G. and WENDELBERGER, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Rev.* **108** 1122–1143.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305