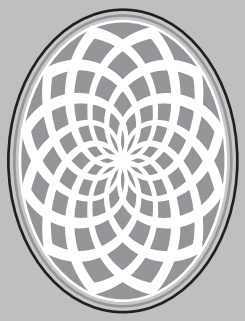

Industrial Mathematics Institute

2003:20

The gridge algorithm in Gaussian weighted L^2

R. Kozarev

IMI

Preprint Series

Department of Mathematics
University of South Carolina

The Gridge Algorithm in Gaussian Weighted L^2

Roumen Kozarev
USC

November 25, 2003

Abstract

Greedy algorithms in ridge approximation (gridge algorithms) are considered. Functions from the Gaussian weighted Hilbert space L^2 are approximated by linear combinations of ridge functions. The construction is iterative. On each step one more ridge function is added to the preceeding combination. This ridge function is selected greedily from the dictionary of ridge functions. The convergence rate of the gridge approximant is estimated in terms of the best approximations by algebraic polynomials.

Introduction

The main result of this paper is a constructive estimate of the efficiency of the greedy algorithm approximation of a multivariate function by linear combinations of ridge functions (plane waves) in a Gaussian weighted Hilbert space metric. A ridge function is a multivariate function whose level curves are straight lines. The paper was inspired by and follows closely [4] where the authors answered a similar question about ridge approximation of functions, supported on the unit ball in the non-weighted L^2- space.

This method of approximation is not new and have been used in Statistics since the 80's. It is called Projection pursuit regression. Generally, statistical methods of fitting a multidimensional regression model suffer from the so called "curse of dimensionality", which roughly means that in most cases the price for the better fit of a multidimensional model is choosing enormously large samples. The Projection pursuit regression seems to overcome this drawback of the traditional statistical methods. A very detailed discussion of the pros and cons of this method is given in [2]. The same approach of decomposing a signal into a linear expansion of waves is used also in Signal processing under the name of Matching pursuit and was developed by Mallath and Zhang [5]. The question about the greedy algorithm approximation efficiency, however, has not been studied until the appearance of some recent papers [4], [6].

The motivation for this is clear. It is important to estimate quantitatively the efficiency of this algorithm. Another reason is that this method of approximation is an application of the recently developed general theory of Greedy Algorithms to some particular sets of functions and dictionaries. While we cannot say too much in the most abstract setting, it is worth to see what happens in some concrete situations.

Preliminaries

As we mentioned before we shall use linear combinations of ridge functions to approximate a given multivariate function. This linear combination will be constructed by using the Pure Greedy Algorithm (PGA). This explains the new term "gridge", which is a combination of "greedy" and "ridge". The PGA is a step-wise method. On each step we expand the existing linear combination by one more ridge function. This new ridge function is selected greedily, i.e. in a certain optimal way. The method generates an infinite series of ridge functions whose partial sums approximate the given function.

As usual we shall denote the dot product of the elements $\mathbf{x}$ and $\mathbf{y}$ of $\mathbf{R}^d$ by $\mathbf{x}\cdot\mathbf{y}$ and the length of $\mathbf{x}$ by $|\mathbf{x}|$. The unit ball and the unit sphere in $\mathbf{R}^d$ will be denoted by $\mathbf{B}^d$ and $\mathbf{S}^{d-1}$, respectively. The measure on the sphere $\mathbf{S}^{d-1}$ will be assumed normalized, i.e.

$$\int_{\mathbf{S}^{d-1}} d\boldsymbol{\theta} = 1 \ .$$

Here is a brief description of the ridge functions and the properties of the PGA. We will refer the reader interested in more details to [4].

The function $W(\mathbf{x})$ is called a ridge function if it admits the special form $W(\mathbf{x}) = w(\mathbf{x}\cdot\boldsymbol{\theta})$, where w is a univariate function (called the profile of W) and $\boldsymbol{\theta}$ is a fixed unit vector. Then we say that W is a plane wave propagating in the direction $\boldsymbol{\theta}$.

If H is a Hilbert space and D is a subset of H whose span is dense in H, then D is called a (non-normalized) dictionary. Let $f \in H$ be a fixed element and let $\|f\|_H$ be its Hilbert space norm. The following notations are more or less standard in the Greedy Algorithm theory.

$$\begin{aligned} g_0(f) &= 0 \ , \\ G_{N-1}(f) &= \sum_{j=0}^{N-1} g_j(f) \ , \ N = 1, 2, \dots \ , \\ g_N(f) &= \arg\left(\min_{g \in D} \| \left(f - G_{N-1}(f)\right) - g \|_H \right) \ . \end{aligned}$$

We write "min" only for simplicity, because neither the existence nor the uniqueness of the extremal element are guaranteed in the general case. However for the particular space and dictionary considered in the paper we have existence, and g_N denotes any of the minimizers from D. It may happen

that the minimizer is not unique and the process "branches". In such a case we shall follow an arbitrary branch. If we denote

$$f_N := f - G_{N-1}(f) \ , \ N = 1, 2, \dots$$

then we have the following relation

$$\|f_N\|_H = \min_{g \in D} \|f_{N-1} - g\|_H \ .$$

We should mention the so called Parseval property of the PGA, that is

$$\|f - G_N(f)\|_H^2 = \|f\|_H^2 - \sum_{j=0}^{N-1} \|g_j(f)\|_H^2 \ .$$

The strong convergence of the PGA for a general Hilbert space and a general dictionary was proved by Jones [3]. The immediate corollaries from this result are the equalities

$$\begin{aligned} \|f - G_N(f)\|_H^2 &= \sum_{j=N+1}^{\infty} \|g_j(f)\|_H^2 \ , \\ \|f\|_H^2 &= \sum_{j=0}^{\infty} \|g_j(f)\|_H^2 \ , \\ f &\stackrel{H}{=} \sum_{j=0}^{\infty} g_j(f) \ . \end{aligned}$$

In the paper we shall approximate functions from the Gaussian weighted L^2 by means of ridge functions.

For a multivariate complex-valued function f we denote by $\|f\|$ the weighted norm

$$\|f\| := \left(\int_{\mathbf{R}^d} e^{-\pi |\mathbf{x}|^2} |f(\mathbf{x})|^2 d\mathbf{x} \right)^{\frac{1}{2}} \ ,$$

and define

$$L^2(\mathbf{R}^d, e^{-\pi |\mathbf{x}|^2}) := \{ f : \|f\| < \infty \} \ .$$

Next, define the dictionary D as the set of all ridge functions from $L^2(\mathbf{R}^d, e^{-\pi |\mathbf{x}|^2})$.

We can give another characterization of the elements of the dictionary based on their profiles. Assume that $\mathbf{e_d} = (0, 0, \ldots, \overset{d}{1})$ and $\mathbf{e_d} = B\boldsymbol{\theta}$ for some orthogonal matrix B. Then set

$$B\mathbf{x} = \mathbf{y} = (y_1, y_2, \ldots, y_d) \ .$$

Then we have

$$\begin{aligned}
\|W\|^2 &= \int_{\mathbf{R}^d} e^{-\pi|\mathbf{x}|^2} |W(\mathbf{x})|^2 d\mathbf{x} = \int_{\mathbf{R}^d} e^{-\pi|\mathbf{x}|^2} |w(\mathbf{x}\cdot\boldsymbol{\theta})|^2 d\mathbf{x} \\
&= \int_{\mathbf{R}^d} e^{-\pi|\mathbf{y}|^2} |w(\mathbf{y}\cdot\mathbf{e_d})|^2 d\mathbf{y} = \int_{\mathbf{R}^d} e^{-\pi|\mathbf{y}|^2} |w(y_d)|^2 d\mathbf{y} \\
&= \int_{-\infty}^{\infty} e^{-\pi y_d^2} |w(y_d)|^2 \left(\int_{\mathbf{R}^{d-1}} e^{-\pi(y_1^2+y_2^2+\cdots+y_{d-1}^2)} dy_1 dy_2 \ldots dy_{d-1} \right) dy_d \\
&= \int_{-\infty}^{\infty} e^{-\pi y_d^2} |w(y_d)|^2 dy_d \ .
\end{aligned}$$

This shows that D consists of all ridge functions whose profiles $w(t) \in L^2(\mathbf{R}, e^{-\pi t^2})$.

Fourier-Hermite Analysis in $L^2(\mathbf{R}^d, e^{-\pi|\mathbf{x}|^2})$

Let $\mathcal{P}^{m,d}$ be the space of all d - variate algebraic polynomials of degree not exceeding m. By $H_n(t)$ we denote the normalized univariate Hermite polynomials in $L^2(\mathbf{R}, e^{-\pi\mathbf{t}^2})$, i.e. $H_n \in \mathcal{P}^{n,1}$ and H_n satisfies the equalities

$$\int_{-\infty}^{\infty} e^{-\pi t^2} H_n(t) P(t) dt = 0 \ , \ \text{for } P \in \mathcal{P}^{n-1,1} \ ,$$

$$\int_{-\infty}^{\infty} e^{-\pi t^2} H_n^2(t) dt = 1 \ , \ n = 1, 2, \dots \ .$$

For every fixed $g \in L^2(\mathbf{R}^d, e^{-\pi|\mathbf{x}|^2})$ introduce the following functions of $\boldsymbol{\theta} \in \mathbf{S}^{d-1}$.

$$a_n(g, \boldsymbol{\theta}) := \int_{\mathbf{R}^d} e^{-\pi|\mathbf{x}|^2} g(\mathbf{x}) H_n(\mathbf{x}\cdot\boldsymbol{\theta}) d\mathbf{x} \ , \ n = 0, 1, 2, \dots \ .$$

Clearly, $a_n(g, \boldsymbol{\theta})$ is a spherical polynomial of degree n for every fixed nonnegative integer n.
Let us notice as well that

$$a_n(g, -\boldsymbol{\theta}) = (-1)^n a_n(g, \boldsymbol{\theta}) \ .$$

The latter follows from the well-known fact that $H_n(-t) = (-1)^n H_n(t)$. It is natural to call the spherical polynomials $a_n(g, \boldsymbol{\theta})$ $n = 0, 1, 2, \dots$ the Hermite momenta of g of degree n.
The following characterizes the expansion of a function $g \in L^2(\mathbf{R}^d, e^{-\pi|\mathbf{x}|^2})$ in terms of the ridge functions $H_n(\mathbf{x}\cdot\boldsymbol{\theta})$, $n = 0, 1, 2, \dots$. It seems to be a folk theorem, but as we can not find an explicit reference we include a short proof given in the appendix.

Lemma 1. *For every function $g \in L^2(\mathbf{R}^d, e^{-\pi|\mathbf{x}|^2})$ there exist unique spherical polynomials $b_n(g, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbf{S}^{d-1}$ such that*

$$g(\mathbf{x}) \overset{L^2(\mathbf{R}^d, e^{-\pi|\mathbf{x}|^2})}{=} \sum_{n=0}^{\infty} \int_{\mathbf{S}^{d-1}} b_n(g, \boldsymbol{\theta}) H_n(\mathbf{x}\cdot\boldsymbol{\theta}) d\boldsymbol{\theta} \ , \ n = 1, 2, \dots$$

The spherical polynomials $a_n(g, \boldsymbol{\theta})$ and $b_n(g, \boldsymbol{\theta})$ are related by

$$a_n(g,\boldsymbol{\theta}) = \int_{\mathbf{S}^{d-1}} (\boldsymbol{\theta}\cdot\boldsymbol{\phi})^n b_n(g,\boldsymbol{\phi}) d\boldsymbol{\phi} \ .$$

Moreover

$$\begin{aligned} \|g\|^2 &= \sum_{n=0}^{\infty} \int_{\mathbf{S}^{d-1}} b_n(g,\boldsymbol{\theta}) \overline{a_n(g,\boldsymbol{\theta})} d\boldsymbol{\theta} \\ &= \sum_{n=0}^{\infty} \iint_{\mathbf{S}^{d-1}\times\mathbf{S}^{d-1}} (\boldsymbol{\theta}\cdot\boldsymbol{\phi})^n b_n(g,\boldsymbol{\theta}) \overline{b_n(g,\boldsymbol{\phi})} d\boldsymbol{\phi} d\boldsymbol{\theta} \ . \end{aligned} \tag{1}$$

It is easy to see that the spherical polynomial $b_n(g,\boldsymbol{\theta})$ also satisfies the condition

$$b_n(g,-\boldsymbol{\theta}) = (-1)^n b_n(g,\boldsymbol{\theta}) \ .$$

Now we introduce the notation $\|\cdot\|_{\langle n\rangle}$, defined by

$$\|f\|^2_{\langle n\rangle} := \iint_{\mathbf{S}^{d-1}\times\mathbf{S}^{d-1}} (\boldsymbol{\theta}\cdot\boldsymbol{\phi})^n f(\boldsymbol{\theta}) \overline{f(\boldsymbol{\phi})} d\boldsymbol{\phi} d\boldsymbol{\theta}$$

for an arbitrary function f with integrable square on $\mathbf{S}^{d-1}$ and we shall use it throughout the paper. Then (1) becomes

$$\|g\|^2 = \sum_{n=0}^{\infty} \|b_n(g)\|^2_{\langle n\rangle} \ .$$

Theorem 1. *If* $f \in L^2(\mathbf{R}^2, e^{-\pi|\mathbf{x}|^2})$ *and* $E_N(f) = O(N^{-r})$, $N \to \infty$ *then*

$$\|f - G_N(f)\| = O(\log^{-r} N) \ , \ N \to \infty$$

Remark. The proof of the theorem follows the ideas from [4]. In order to find the optimal ridge approximant we have to find its profile and direction.

First, we find the optimal profile of a fixed direction. This is a linear approximation problem and we can even give the explicit solution in terms of the direct Radon transform.
Second, we find the optimal direction. This is a non-linear approximation problem, which cannot be solved explicitly in the general case. The approach used here is to estimate the error in the optimal direction by means of the average of the errors in all possible directions.

The key moment is to compare the two functionals $\|b_n(g)\|^2_{\langle n \rangle}$ and $\int_{\mathbf{S}^{d-1}} |a_n(g, \boldsymbol{\theta})|^2 d\boldsymbol{\theta}$.
Obtaining an inequality of the form

$$\|b_n(g)\|^2_{\langle n \rangle} \leq c_n \int_{\mathbf{S}^{d-1}} |a_n(g, \boldsymbol{\theta})|^2 d\boldsymbol{\theta} \tag{2}$$

where c_n is a constant (depending also on d) will result in the following estimate

$$\|f_{j+1}\|^2 \leq \sum_{n=1}^{\infty} \frac{c_n - c_{n-1}}{c_n c_{n-1}} \min(E_n^2(f), \|f_j\|^2) \ .$$

From this estimate the asymptotical order of the error can be obtained by using the general estimation lemma from [4].
In this paper we obtain an inequality of the form (2) for the case $d = 2$ and thus the case of an arbitrary dimension is still an open question.

Proof
Let $g \in L^2(\mathbf{R}^d, e^{-\pi|\mathbf{x}|^2})$. We shall estimate the error of the best ridge approximant of the function g. Since we can write

$$\min_{W \in D} \|g - W\| = \min_{\boldsymbol{\theta} \in \mathbf{S}^{d-1}} \left(\min_{W \in D_{\boldsymbol{\theta}}} \|g - W\| \right) ,$$

where

$$D_{\boldsymbol{\theta}} = \{W \in D : W(\mathbf{x}) = w(\mathbf{x}\cdot\boldsymbol{\theta})\}$$

we can estimate the error in two steps.

Lemma 2. *Let $g \in L^2(\mathbf{R}^d, e^{-\pi|\mathbf{x}|^2})$ and $\boldsymbol{\theta} \in \mathbf{S}^{d-1}$ be fixed. Then*

$$\min_{W \in D_{\boldsymbol{\theta}}} \|g - W\| = \|g\|^2 - \sum_{n=0}^{\infty} |a_n(g, \boldsymbol{\theta})|^2 , \tag{3}$$

and the best ridge approximant of g has profile

$$w^*_{\boldsymbol{\theta}}(t) = e^{\pi t^2} \int_{\mathbf{x}\cdot\boldsymbol{\theta}=t} e^{-\pi|\mathbf{x}|^2} g(\mathbf{x}) d\mathbf{x}' ,$$

where $d\mathbf{x}'$ stands for the Lebesgue measure on the hyperplane $\mathbf{x}\cdot\boldsymbol{\theta} = t$.

Proof. It is not hard to see that even in a more general setting the minimization of $\|g - W\|$ in D_1, a fixed subset of D is solved by the orthogonal projection of g onto Span (D_1). Thus by a well-known property of the linear approximation we have that the best approximant W^* satisfies

$$\langle g - W^* , W \rangle = 0 ,$$

for all $W \in D_1$, where the inner product $\langle f , g \rangle$ of f and g is defined as usual by $\langle f , g \rangle = \int_{\mathbf{R}^d} e^{-\pi|\mathbf{x}|^2} f(\mathbf{x}) \overline{g(\mathbf{x})} d\mathbf{x}$. Arguments similar to those, presented in [2] show that the optimal profile $w^*_{\boldsymbol{\theta}}$ is

$$w^*_{\boldsymbol{\theta}}(t) = \frac{\displaystyle\int_{\mathbf{x}\cdot\boldsymbol{\theta}=t} e^{-\pi|\mathbf{x}|^2} g(\mathbf{x}) d\mathbf{x}'}{\displaystyle\int_{\mathbf{x}\cdot\boldsymbol{\theta}=t} e^{-\pi|\mathbf{x}|^2} d\mathbf{x}'} .$$

Next we can see that the integral in the denominator is equal to $e^{-\pi t^2}$. Indeed, since $e^{-\pi|\mathbf{x}|^2}$ is orthogonal invariant we can assume $\boldsymbol{\theta} = \mathbf{e_d} = (0, 0, \ldots, \overset{d}{1})$ and then if we set $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ we evaluate

$$\int_{\mathbf{x}\cdot\boldsymbol{\theta}=t} e^{-\pi|\mathbf{x}|^2} d\mathbf{x}' = \int_{x_d=t} e^{-\pi(x_1^2+x_2^2+\cdots+x_{d-1}^2+t^2)} dx_1 dx_2 \ldots dx_{d-1} = e^{-\pi t^2} \ .$$

This proves the second part of Lemma 1. To prove (3) we make use of the above mentioned orthogonality $\langle g - W^*, W\rangle = 0$ and we obtain

$$\|g - W^*_{\boldsymbol{\theta}}\|^2 = \|g\|^2 - \|W^*_{\boldsymbol{\theta}}\|^2 \ .$$

We already mentioned the equality

$$\|W^*_{\boldsymbol{\theta}}\|^2 = \int_{-\infty}^{\infty} e^{-\pi y_d^2} |w^*_{\boldsymbol{\theta}}(y_d)|^2 dy_d \ .$$

On the other hand

$$w^*_{\boldsymbol{\theta}}(t) = e^{\pi t^2} \int_{\mathbf{x}\cdot\boldsymbol{\theta}=t} e^{-\pi|\mathbf{x}|^2} g(\mathbf{x}) d\mathbf{x}'$$

and therefore

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-\pi t^2} w^*_{\boldsymbol{\theta}}(t) H_n(t) dt &= \int_{-\infty}^{\infty} H_n(t) \left(\int_{\mathbf{x}\cdot\boldsymbol{\theta}=t} e^{-\pi|\mathbf{x}|^2} g(\mathbf{x}) d\mathbf{x}' \right) dt \\ &= \int_{\mathbf{R}^d} e^{-\pi|\mathbf{x}|^2} g(\mathbf{x}) H_n(\mathbf{x}\cdot\boldsymbol{\theta}) d\mathbf{x} \\ &= a_n(g, \boldsymbol{\theta}) \ . \end{aligned}$$

Hence

$$w^*_{\boldsymbol{\theta}}(t) = \sum_{n=0}^{\infty} a_n(g, \boldsymbol{\theta}) H_n(t)$$

and by the Parseval equality

$$\int_{-\infty}^{\infty} e^{-\pi t^2} |w^*_{\boldsymbol{\theta}}(t)|^2 dt = \sum_{n=0}^{\infty} |a_n^2(g, \boldsymbol{\theta})|^2 \ .$$

Thus we obtain (3), which can be rewritten as

$$\|g - W^*_{\boldsymbol{\theta}}\|^2 = \|g\|^2 - \sum_{n=0}^{\infty} |a_n(g, \boldsymbol{\theta})|^2 = \sum_{n=0}^{\infty} \left(\|b_n(g)\|^2_{\langle n\rangle} - |a_n(g, \boldsymbol{\theta})|^2 \right) \ .$$

Lemma 3. *Let $g \in L^2(\mathbf{R}^d, e^{-\pi|\mathbf{x}|^2})$ and $W^*_{\boldsymbol{\theta}}$ is the best ridge function approximating g in the dictionary $D_{\boldsymbol{\theta}}$ for every fixed $\boldsymbol{\theta} \in \mathbf{S}^{d-1}$. Then*

$$\min_{\boldsymbol{\theta}\in \mathbf{S}^{d-1}} \|g - W^*_{\boldsymbol{\theta}}\| \leq \sum_{n=0}^{\infty} \left(\|b_n(g)\|^2_{\langle n\rangle} - \int_{\mathbf{S}^{d-1}} |a_n(g,\boldsymbol{\theta})|^2 d\boldsymbol{\theta} \right) \ .$$

Proof. The result of the previous lemma shows that for an arbitrary fixed $\boldsymbol{\theta} \in \mathbf{S}^{d-1}$

$$\|g - W^*_{\boldsymbol{\theta}}\|^2 = \|g\|^2 - \sum_{n=0}^{\infty} |a_n(g,\boldsymbol{\theta})|^2 = \sum_{n=0}^{\infty} \left(\|b_n(g)\|^2_{\langle n\rangle} - |a_n(g,\boldsymbol{\theta})|^2 \right) \ .$$

Therefore

$$\begin{aligned} \min_{\boldsymbol{\theta}\in \mathbf{S}^{d-1}} \|g - W^*_{\boldsymbol{\theta}}\| &= \|g\|^2 - \max_{\boldsymbol{\theta}\in \mathbf{S}^{d-1}} \sum_{n=0}^{\infty} |a_n(g,\boldsymbol{\theta})|^2 \\ &\leq \|g\|^2 - \int_{\mathbf{S}^{d-1}} \sum_{n=0}^{\infty} |a_n(g,\boldsymbol{\theta})|^2 d\boldsymbol{\theta} \\ &= \sum_{n=0}^{\infty} \left(\|b_n(g)\|^2_{\langle n\rangle} - \int_{\mathbf{S}^{d-1}} |a_n(g,\boldsymbol{\theta})|^2 d\boldsymbol{\theta} \right) \ . \end{aligned}$$

Lemma 4. *Let $a_n(\theta)$ and $b_n(\theta)$ be trigonometric polynomials of degree n satisfying*

(i) $b_n(\theta + \pi) = (-1)^n b_n(\theta)$

(ii) $a_n(\theta) = \int_0^{2\pi} \cos^n(\theta - \phi) b_n(\phi) d\phi$,

where the measure on the interval $[0, 2\pi]$ is normalized by $\int_0^{2\pi} d\theta = 1$. Then

$$\int_0^{2\pi}\int_0^{2\pi} \cos^n(\theta - \phi) b_n(\theta) \overline{b_n(\phi)} d\phi d\theta \leq 2^n \int_0^{2\pi} |a_n(\theta)|^2 d\theta \ . \tag{4}$$

Proof. Let

$$b_n(\phi) = \sum_{|m|\leq n} \hat{b}_m e^{im\phi} \ .$$

Then $\hat{b}_m = 0$ if $m \not\equiv n \pmod 2$, because of the relation $b_n(\theta+\pi) = (-1)^n b_n(\theta)$. So keeping in mind that m and n are of the same parity we get

$$\begin{aligned}
a_n(\theta) &= \int_0^{2\pi} \cos^n(\theta-\phi) b_n(\phi) d\phi = \int_0^{2\pi} \cos^n\phi\, b_n(\theta-\phi) d\phi \\
&= \int_0^{2\pi} \left(\frac{e^{i\phi}+e^{-i\phi}}{2}\right)^n \sum_{|m|\le n} \hat{b}_m e^{im(\theta-\phi)} d\phi \\
&= \frac{1}{2^n}\int_0^{2\pi} \sum_{k=0}^{n} \binom{n}{k} e^{i(2k-n)\phi} \sum_{|m|\le n} \hat{b}_m e^{im(\theta-\phi)} d\phi \\
&= \frac{1}{2^n}\int_0^{2\pi} \sum_{k=0}^{n} \binom{n}{k} \sum_{|m|\le n} \hat{b}_m e^{i(2k-n-m)\phi} e^{im\theta} d\phi \\
&= \frac{1}{2^n} \sum_{|m|\le n} \hat{b}_m e^{im\theta} \sum_{k=0}^{n} \binom{n}{k} \int_0^{2\pi} e^{i(2k-n-m)\phi} d\phi \\
&= \frac{1}{2^n} \sum_{|m|\le n} \hat{b}_m e^{im\theta} \binom{n}{\frac{n+m}{2}} .
\end{aligned}$$

Thus

$$\int_0^{2\pi} |a_n(\theta)|^2 d\theta = \frac{1}{2^{2n}} \sum_{|m|\le n} |\hat{b}_m|^2 \binom{n}{\frac{n+m}{2}}^2 .$$

Moreover

$$\begin{aligned}
\int_0^{2\pi}\int_0^{2\pi} \cos^n(\theta-\phi) b_n(\theta)\overline{b_n(\phi)} d\phi d\theta &= \int_0^{2\pi} b_n(\theta) \left(\int_0^{2\pi} \cos^n(\theta-\phi)\overline{b_n(\phi)} d\phi\right) d\theta \\
&= \int_0^{2\pi} b_n(\theta)\overline{a_n(\theta)} d\theta \\
&= \frac{1}{2^n} \sum_{|m|\le n} |\hat{b}_m|^2 \binom{n}{\frac{n+m}{2}} .
\end{aligned}$$

Clearly, this implies (4).

Proof of the Theorem. From the first two lemmas we have the following estimate for the best ridge approximation of an arbitrary function $g \in L^2(\mathbf{R}^d, e^{-\pi|\mathbf{x}|^2})$.

$$\min_{W\in D} \|g-W\|^2 \le \sum_{n=0}^{\infty} \left(\|b_n(g)\|^2_{\langle n\rangle} - \int_{\mathbf{S}^{d-1}} |a_n(g,\boldsymbol{\theta})|^2 d\boldsymbol{\theta} \right) .$$

Let $d = 2$. We shall use the usual parametrization of the unit circle $\mathbf{S}^1$, that is we set $\boldsymbol{\theta} = (\cos\theta, \sin\theta)$, $\theta \in [0, 2\pi)$. Then (with a slight abuse of the notation) we can denote

$$b_n(\theta) = b_n\left(g, \boldsymbol{\theta}\right) \ ,$$
$$a_n(\theta) = a_n\left(g, \boldsymbol{\theta}\right) \ ,$$

and we have

$$\begin{aligned} \|b_n(g)\|^2_{\langle n\rangle} &= \int_0^{2\pi}\!\!\int_0^{2\pi} \cos^n(\theta-\phi) b_n(\theta)\overline{b_n(\phi)} d\phi \ , \\ \int_{\mathbf{S}^{d-1}} |a_n(g,\boldsymbol{\theta})|^2 d\boldsymbol{\theta} &= \int_0^{2\pi} |a_n(\theta)|^2 d\theta \ . \end{aligned}$$

Now from Lemma 3 we see that

$$\begin{aligned} \min_{W\in D} \|g - W\|^2 &\le \sum_{n=0}^{\infty} \left(\|b_n(g)\|^2_{\langle\rangle} - \frac{1}{2^n}\|b_n(g)\|^2_{\langle n\rangle} \right) \\ &= \sum_{n=1}^{\infty} \left(1 - \frac{1}{2^n}\right) \|b_n(g)\|^2_{\langle n\rangle} \ . \end{aligned}$$

Substituting $\|b_n(g)\|^2_{\langle n\rangle} = E^2_{n-1}(g) - E^2_n(g)$ and applying the Abel summation formula yields

$$\min_{W\in D} \|g - W\|^2 \le \sum_{n=1}^{\infty} \frac{1}{2^n} E^2_n(g) \ . \tag{5}$$

Now, applying (5) to the consecutive gridge approximants we get

$$\|f_{j+1}\|^2 \le \sum_{n=1}^{\infty} \frac{1}{2^n} E^2_n(f_j) \ , \quad j = 0, 1, 2, \dots \ . \tag{6}$$

Denote by $\mathbf{E}$ the infinite matrix of the best polynomial approximation of the gridge approximants of the function f.

$$\mathbf{E} := \begin{bmatrix} E_0(f_0) & E_1(f_0) & E_2(f_0) & \cdots & E_n(f_0) & \cdots \\ E_0(f_1) & E_1(f_1) & E_2(f_1) & \cdots & E_n(f_1) & \cdots \\ E_0(f_2) & E_1(f_2) & E_2(f_2) & \cdots & E_n(f_2) & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ E_0(f_j) & E_1(f_j) & E_2(f_j) & \cdots & E_n(f_j) & \cdots \\ \vdots & \vdots & \vdots & & \vdots & \ddots \end{bmatrix}$$

Then **E** is row and column monotonous. Indeed the following two inequalities hold true

$$E_n(f_j) \geq E_{n+1}(f_j) \ ,$$

$$E_n(f_j) \geq E_n(f_{j+1}) \ .$$

The first inequality follows from the definition of the best polynomial approximation while the second one is a consequence of Lemmas 1 and 2. Taking into account that $E_n(f_j) \leq \|f_j\|$, $n = 1, 2, \ldots$, $j = 0, 1, \ldots$ and the column monotonicity we get the estimate

$$\|f_{j+1}\|^2 \leq \sum_{n=1}^{\infty} \frac{1}{2^n} \min(E_n^2(f), \|f_j\|^2) \ , \quad j = 0, 1, 2, \ldots \ . \tag{7}$$

In order to complete the proof we shall use the following estimation lemma (Lemma 4) from [4].

Estimation Lemma. *Let $F(\xi)$ be a non-decreasing function defined on $(0, 1]$ and let the sequence $\{A_n\}_{n=0}^{\infty}$ be defined by*

$$A_{n+1} = \int_0^1 \min(F(\xi), A_n) d\xi \ , n = 0, 1, 2, \ldots \ .$$

Then

$$A_N \leq 2F\left(H^{-1}\left(\frac{N}{2}\right)\right) \ , \ N = 1, 2, \ldots \ ,$$

where

$$H(\xi) := \frac{2}{\xi} \sup_{1/2 \geq \eta \geq \xi} \left(\log \frac{F(2\eta)}{F(\eta)}\right)$$

and H^{-1} is the inverse of H.

Since we want to find the order of $\|f_j\|$ as $n \to \infty$ we can "replace" $E_n(f)$ by n^{-r} for simplicity. So instead of (7) we have now

$$\|f_{j+1}\|^2 \leq \sum_{n=1}^{\infty} \frac{1}{2^n} \min(n^{-2r}, \|f_j\|^2) \ , \quad j = 0, 1, 2, \ldots \ .$$

Define the function $F(\xi)$ as a step function on $(0,1]$.

$$F(\xi) := n^{-2r}, \quad \text{if } \frac{1}{2^n} < \xi \le \frac{1}{2^{n-1}} \ , \ n = 1, 2, \dots \ .$$

Note that $n = \left[\log_2 \frac{1}{\xi}\right] + 1$ for every $\xi \in (0,1]$. We denote also $\|f_j\|^2$ by A_j. Then it is not hard to see that we can rewrite the inequality

$$A_{j+1} \le \sum_{n=1}^{\infty} \frac{1}{2^n} \min(n^{-2r}, A_j) \tag{8}$$

as

$$A_{j+1} = \int_0^1 \min(F(\xi), A_j) d\xi \ . \tag{9}$$

Indeed, if m is such that $(m+1)^{-2r} < A_j \le m^{-2r}$, then (8) is equivalent to

$$A_{j+1} \le \sum_{k=1}^{m} \frac{A_j}{2^k} + \sum_{k=m+1}^{\infty} \frac{k^{-2r}}{2^k} = A_j \left(1 - \frac{1}{2^m}\right) + \sum_{k=m+1}^{\infty} \frac{k^{-2r}}{2^k}$$

and (9) is equivalent to

$$A_{j+1} \le \int_U \min(F(\xi), A_j) d\xi + \int_{[0,1]\setminus U} \min(F(\xi), A_j) d\xi = \int_U F(\xi) d\xi + A_j(1 - \operatorname{meas} U) \ ,$$

where $U = \{\xi : \xi \in (0,1] \text{ and } F(\xi) < A_j\}$, and $\operatorname{meas} U$ is the Lebesgue measure of U. The definition of F implies that $U = \left(0, \frac{1}{2^m}\right]$ and that $\int_U F(\xi) d\xi = \sum_{k=m+1}^{\infty} \frac{k^{-2r}}{2^k}$ and this proves the equivalency of (6) and (7). Finally, if $F(\eta) = n^{-2r}$ then $F(2\eta) = (n-1)^{-2r}$ where $\eta \in (0, 1/2]$. We have then

$$\frac{F(2\eta)}{F(\eta)} = \left(1 + \frac{1}{n-1}\right)^{2r} = \left(1 + \frac{1}{[\log_2(\eta^{-1})]}\right)^{2r} \ .$$

This means that $\dfrac{F(2\eta)}{F(\eta)}$ is a non-decreasing function and so is $\log \dfrac{F(2\eta)}{F(\eta)}$. Hence $H(\xi) = \dfrac{c}{\xi}$ for some constant c. Therefore $H^{-1}(\xi) = \dfrac{c}{\xi}$. Clearly this implies the estimate

$$\|f - G_N(f)\| = \mathrm{O}(\log^{-r} N) \ , \ N \to \infty \ .$$

Appendix

Proof of Lemma 1.

At first we will prove that there exist unique spherical polynomials $b_n(g,\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbf{S}^{d-1}$, $n = 0, 1, 2, \ldots$ such that

$$g(\mathbf{x}) \overset{L^2(\mathbf{R}^d, e^{-\pi|\mathbf{x}|^2})}{=} \sum_{n=0}^{\infty} \int_{\mathbf{S}^{d-1}} b_n(g,\boldsymbol{\theta}) H_n(\mathbf{x}\cdot\boldsymbol{\theta}) d\boldsymbol{\theta} \ , \ n = 1, 2, \ldots \ . \tag{10}$$

Since the set of all algebraic polynomials is dense in $L^2(\mathbf{R}^d, e^{-\pi|\mathbf{x}|^2})$ it suffices to show that this representation holds true when $g(\mathbf{x})$ is an algebraic polynomial. It is proven in [6, Theorem 3.1] that every function $f \in L^2(\mathbf{B}^d)$ can be represented uniquely as

$$f(\mathbf{x}) \overset{L^2(\mathbf{B}^d)}{=} \sum_{n=0}^{\infty} \int_{\mathbf{S}^{d-1}} A_n(f,\boldsymbol{\theta}) U_n(\mathbf{x}\cdot\boldsymbol{\theta}) d\boldsymbol{\theta} \ , \ n = 1, 2, \ldots \ ,$$

where U_n is the Gegenbauer polynomial $C_n^{d/2}$ and $A_n(f,\boldsymbol{\theta})$ is a spherical polynomial of degree n. Therefore if $g(\mathbf{x})$ is an algebraic polynomial of degree m we have

$$g(\mathbf{x}) = \sum_{n=0}^{m} \int_{\mathbf{S}^{d-1}} A_n(g,\boldsymbol{\theta}) U_n(\mathbf{x}\cdot\boldsymbol{\theta}) d\boldsymbol{\theta} \ , \ n = 1, 2, \ldots \ .$$

Since both U_n and H_n are algebraic polynomials of degree n, there exist constants c_{nj} , $n = 0, 1, 2, \ldots, m$, $j = 0, 1, 2, \ldots, n$ such that

$$U_n(t) = \sum_{j=0}^{n} c_{nj} H_j(t) \ , \ t \in \mathbf{R} \ .$$

Then

$$\begin{aligned} g(\mathbf{x}) &= \sum_{n=0}^{m} \int_{\mathbf{S}^{d-1}} A_n(g,\boldsymbol{\theta}) \sum_{j=0}^{m} c_{nj} H_j(\mathbf{x}\cdot\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \sum_{j=0}^{m} \int_{\mathbf{S}^{d-1}} \left(\sum_{n=j}^{m} c_{nj} A_n(g,\boldsymbol{\theta}) \right) H_j(\mathbf{x}\cdot\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \sum_{j=0}^{m} \int_{\mathbf{S}^{d-1}} b_j(g,\boldsymbol{\theta}) H_j(\mathbf{x}\cdot\boldsymbol{\theta}) d\boldsymbol{\theta} \ , \end{aligned}$$

where $b_j(g,\boldsymbol{\theta}) = \sum_{n=j}^{m} c_{nj} A_n(g,\boldsymbol{\theta})$ is a spherical polynomial of degree m. It is well-known that every spherical polynomial is a sum of spherical harmonics [1] so then we can rewrite $b_j(g,\boldsymbol{\theta})$ as

$$b_j(g,\boldsymbol{\theta}) = \sum_{k=0}^{m} S_k(\boldsymbol{\theta}) \ , \ S_k(\boldsymbol{\theta}) \text{ - spherical harmonic of degree } k \ . \tag{11}$$

Then observe that

$$\begin{aligned} \int_{\mathbf{S}^{d-1}} b_j(g,\boldsymbol{\theta}) H_j(\mathbf{x}\cdot\boldsymbol{\theta}) d\boldsymbol{\theta} &= \int_{\mathbf{S}^{d-1}} \left(\sum_{k=0}^{m} S_k(\boldsymbol{\theta}) \right) H_j(\mathbf{x}\cdot\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_{\mathbf{S}^{d-1}} \left(\sum_{k=0}^{j} S_k(\boldsymbol{\theta}) \right) H_j(\mathbf{x}\cdot\boldsymbol{\theta}) d\boldsymbol{\theta} \ , \end{aligned}$$

because $\int_{\mathbf{S}^{d-1}} S_k(\boldsymbol{\theta}) H_j(\mathbf{x}\cdot\boldsymbol{\theta}) d\boldsymbol{\theta} = 0$ if $k > j$.
This shows that we can cut off the senior spherical harmonics (namely the terms $S_k(\boldsymbol{\theta})$ with $k > j$) without changing the value of the integral. So, keeping the same notation $b_j(g,\boldsymbol{\theta})$ for the truncated sum we prove the expansion (10) in $\mathbf{B}^d$. Then clearly (10) holds true in $\mathbf{R}^d$ as well.
Now we prove the relation between $a_n(g,\boldsymbol{\theta})$ and $b_n(g,\boldsymbol{\theta})$. Using (10) and the definition of $a_n(g,\boldsymbol{\theta})$ we obtain

$$\begin{aligned} a_n(g,\boldsymbol{\theta}) &= \int_{\mathbf{R}^d} e^{-\pi|\mathbf{x}|^2} \left(\sum_{n=0}^{\infty} \int_{\mathbf{S}^{d-1}} b_n(g,\boldsymbol{\phi}) H_n(\mathbf{x}\cdot\boldsymbol{\phi}) d\boldsymbol{\phi} \right) H_n(\mathbf{x}\cdot\boldsymbol{\theta}) d\mathbf{x} \\ &= \sum_{n=0}^{\infty} \int_{\mathbf{S}^{d-1}} b_n(g,\boldsymbol{\phi}) \left(\int_{\mathbf{R}^d} e^{-\pi|\mathbf{x}|^2} H_n(\mathbf{x}\cdot\boldsymbol{\phi}) H_n(\mathbf{x}\cdot\boldsymbol{\theta}) d\mathbf{x} \right) d\boldsymbol{\phi} \\ &= \sum_{n=0}^{\infty} \int_{\mathbf{S}^{d-1}} (\boldsymbol{\phi}\cdot\boldsymbol{\theta})^n b_n(g,\boldsymbol{\phi}) d\boldsymbol{\phi} \ . \end{aligned}$$

Here we used the identity

$$\int_{\mathbf{R}^d} e^{-\pi|\mathbf{x}|^2} H_n(\mathbf{x}\cdot\boldsymbol{\phi}) H_n(\mathbf{x}\cdot\boldsymbol{\theta}) d\mathbf{x} = (\boldsymbol{\phi}\cdot\boldsymbol{\theta})^n \ .$$

The proof of the last identity is not hard and uses the orthogonal invariance of the weight function and the normalization of the Hermite polynomials. Finally, the proof of (1) follows easily from this relation between $a_n(g, \boldsymbol{\theta})$ and $b_n(g, \boldsymbol{\theta})$, and the Fubini Theorem.

Acknowledgement

The author is grateful to K. Oskolkov for his constant support.

References

[1] S. Axler, P. Bourdon, and W. Ramey, *Harmonic Function Theory, 2nd ed.,* Springer-Verlag New York 2001, ISBN 0-387-95218-7.

[2] P.J. Huber, *Projection Pursuit,* The Annals of Statistics, **13** (1985), pp. 435 – 475.

[3] L. Jones, *On a Conjecture of Huber Concerning the Convergence of Projection Pursuit Regression,* The Annals of Statistics, **15** (1987), pp. 880 – 882.

[4] V. Maiorov, K. Oskolkov, and V. Temlyakov, *Gridge Approximation and Radon Compass,* The IMI Research Reports, **00:09** (2000).

[5] S. Mallath, Z. Zhang, *Matching Pursuit with Time-Frequency Dictionaries,* IEEE Transactions in Signal Processing, **vol. 41, issue 12** (1993), pp. 3397 – 3415.

[6] P. Petrushev, *Approximation by Ridge Functions and Neural Networks,* SIAM J. Math. Anal., **30** (1998), pp. 155 – 189.