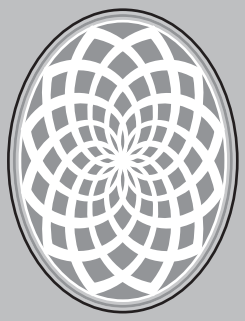

Industrial Mathematics Institute

2000:09

Gridge approximation and Radon compass

V.E. Maiorov, K.I. Oskolkov and V.N. Temlyakov

IMI

Preprint Series

Department of Mathematics
University of South Carolina

Gridge approximation and Radon compass

V.E. Maiorov, K.I. Oskolkov and V.N. Temlyakov

Abstract

Gridge approximation compiles *greedy algorithms* and *ridge approximation*. It is a class of algorithmic constructions of ridge functions – finite linear combinations of planar waves. The goal is to approximate a given target which is a multivariate function. On each step, a new planar wave is added to the preceeding linear combination. This wave is selected *greedily*, i. e. optimally with regard to both the direction of propagation and the profile. ln Mathematical Statistics, gridge approximation is known as projection pursuit regression. We consider gridge approximation in weighted Hilbert functional spaces on d-dimensional Euclidean space.

The notion of Radon compass is introduced, which is a tool of search of the optimal direction of propagation on each step of the algorithm.

The main quantitative result concerns error estimates for gridge processes in the norm of Hilbert space of functions supported on the unit ball, with regard to Lebesgue measure. Fourier analysis of Radon transformation, in terms of Chebyshev – Gegenbauer polynomials, provides the crucial tool in such case.

For a rather wide class of target functions whose polynomial approximations do not decrease "too rapidly", gridge approximation is equally efficient as classical algebraic polynomial. In particular, gridge approximation is not order-saturated.

1

[1]The research was supported by the following agencies and grants.
V. E. Maiorov: The Center for Absorption in Science, Ministry of Immigrant Absorption, State of Israel.
K.I. Oskolkov: National Science Foundation Grant DMS-9706883, ONR Grant N00014-

0.1 Ridge and polynomial approximation. Greedy algorithm

Let $\mathbb{R}^d,\ d = 1, 2, \ldots$ denote the real d-dimensional Euclidean space of vectors $\mathbf{x} = \langle x_1, x_2, \ldots, x_d \rangle, \mathbf{x} \cdot \mathbf{y} := x_1 y_1 + \cdots + x_d y_d,\ |\mathbf{x}| := \sqrt{\mathbf{x} \cdot \mathbf{x}}$; further, $\mathcal{B}^d := \{\mathbf{x} : \ |\mathbf{x}| \le 1\}$ and $\mathcal{S}^{d-1} := \{\mathbf{x} : \ |\mathbf{x}| = 1\}$ – the unit ball and the unit sphere in $\mathbb{R}^d$; $|\mathcal{B}^d|, \ |\mathcal{S}^{d-1}|$ – the volume of $\mathcal{B}^d$ and the surface area of $\mathcal{S}^{d-1}$; $\mu(d\mathbf{x}) := \frac{d\mathbf{x}}{|\mathcal{B}^d|}, \mu(d\boldsymbol{\theta}) := \frac{d\boldsymbol{\theta}}{|\mathcal{S}^{d-1}|}$ – normalized Lebesgue measures on $\mathcal{B}^d$ and $\mathcal{S}^{d-1}$, respectively. Let us fix d, for breviety denote $\mathcal{B} := \mathcal{B}^d$, $\mathcal{S} := \mathcal{S}^{d-1}$ and in the usual fashion introduce the Hilbert space $\mathcal{L}^2(\mathcal{B})$ of functions $f(\mathbf{x})$ supported in $\mathcal{B}$:

$$\mathcal{L}^2(\mathcal{B}) := \left\{ f(\mathbf{x}) : \ \|f\| = \left\| f, \mathcal{L}^2(\mathcal{B}) \right\| := \sqrt{\int_{\mathcal{B}} |f(\mathbf{x})|^2 \, \mu(d\mathbf{x})} \ < \infty \right\}.$$

For $M \ge 0$, a natural N and $f(\mathbf{x}) \in \mathcal{L}^2(\mathcal{B})$ let

$$\mathcal{P}^{d,M} := \mathrm{Span} \left\{ x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d} \right\}_{k_1 + k_2 + \cdots + k_d \le M}, E_M[f] := \min_{P \in \mathcal{P}^{d,M}} \|f - P\|;$$

$$\mathcal{R}^N := \left\{ S(\mathbf{x}) = \sum_{j=1}^{N} W_j(\mathbf{x} \cdot \boldsymbol{\theta}_j) \right\}, \sigma_N[f, \mathcal{R}] = \sigma_N[f] := \inf_{S \in \mathcal{R}^N} \|f - S\|.$$

$\mathcal{P}^{d,M}$ is the subspace of algebraic polynomials of degree $\le M$, in d variables. The quantity $E_M[f]$ is the classical best M-th degree polynomial approximation of f.

In the definition of the set $\mathcal{R}^N$, $W_j(x)$ are arbitrary single-variate functions, and $\boldsymbol{\theta}_j$ – arbitrary (unit) vectors. This set consists of all N-term linear combinations $S(\mathbf{x})$ of functions of the type planar wave. In the sequel, we call $S(\mathbf{x}) \in \mathcal{R}^N$ *ridge functions* of N-th order; the quantity $\sigma_N[f]$ is known as *best free ridge approximation* of f in $\mathcal{L}^2(\mathcal{B})$.

In the modern terminology, the set of all planar waves constitutes the *dictionary* $\mathcal{R}$. The quantity $\sigma_N[f, \mathcal{R}]$ characterizes the best non-linear N-term approximation with regard to $\mathcal{R}$, see [1]. Here, wave profiles $W_j(x)$ and

91-J-1076, ONR/ARO DEPSCoR Grant DAAG55-98-1-0002;
V.N.Temlyakov: National Science Foundation Grant DMS-9970326 and ONR Grant N00014-91-J1343.
Partially, the results of this paper were obtained in March 1999, during the visit of V.E. Maiorov to the Department of Mathematics of the University of South Carolina. This visit was supported by the Grant DoD-N00014-97-1-0806.

wave vectors (directions of propagation) $\boldsymbol{\theta}_j \in \mathcal{S}$, are subjects of optimization. This problem is indeed rather non-linear with respect to optimization of wave vectors.

Let us note that the existence of the best N-term linear combination of planar waves for $N \geq 2$ is non-trivial, due to the possible *collapse effect* of wave vectors. For more details, see [8]– [10].

The *gridge approximation process* also generates ridge functions

$$S(\mathbf{x}) = G_N[f, \mathbf{x}] = \sum_{j=0}^{N-1} W_j \left(\mathbf{x} \cdot \boldsymbol{\theta}_j\right),$$

but now they are constructed step-wise, algorithmically.

By definition, the non-constrained version of such process (without restrictions on the wave profiles $W(x)$), is described by the following iterations:

$$\begin{aligned} &G_0[f, \mathbf{x}] := 0, \quad f_N(\mathbf{x}) := f(\mathbf{x}) - G_N[f, \mathbf{x}]; \\ &(\boldsymbol{\theta}_N,\, W_N(x)) := \arg \min_{\boldsymbol{\theta} \in \mathcal{S}^{d-1}} \min_{\{W(x)\}} \left\| f_N(\mathbf{x}) - W(\mathbf{x} \cdot \boldsymbol{\theta}) \right\|, \\ &G_{N+1}[f, \mathbf{x}] := G_N[f, \mathbf{x}] + W_N \left(\mathbf{x} \cdot \boldsymbol{\theta}_N\right), \quad N = 0, 1, \ldots \end{aligned} \tag{1}$$

This algorithm (projection pursuit) was proposed in [4].

Clearly, gridge approximation may be a branching process, due to non-uniqueness of $\arg\min$ in $\boldsymbol{\theta}$ on a certain step. In this case, we do not make any preferences among optimal wave vectors. We fix a branch of the gridge process and measure its effectiveness by $\mathcal{R}_N^{\mathrm{gr}}[f] := \|f_N\|$.

Our primary goal is the following statement.

Theorem 1 *Assume that the sequence of best polynomial approximations of a function $f(\mathbf{x})$ satisfies the estimate $E_M[f] = O\left(E_{2M}[f]\right),\ M \to \infty$. Then*

$$\mathcal{R}_N^{\mathrm{gr}}[f] = O\left(E_{N^{\frac{1}{d-1}}}[f]\right), \quad N \to \infty \tag{2}$$

for any branch of the gridge approximation process.

Before turning to the proof, let us recall some related facts from the theory of greedy algorithms.

The problem of gridge approximation is a particular case of a general setting, see [2] and [1] where an arbitrary (normalized) dictionary $\mathcal{D}$ was

considered, and efficiency of pure greedy algorithm with regard to $\mathcal{D}$ studied. A set $\mathcal{D}$ of elements from a Hilbert space H is called a dictionary if each $g \in \mathcal{D}$ has norm one ($\|g\| = 1$), and $\overline{\operatorname{Span}}\,\mathcal{D} = H$. For $f \in H$, we denote $g(f) := \arg\max_{g\in\mathcal{D}} |\langle f \cdot g\rangle|$ one of the elements from $\mathcal{D}$ which maximizes the absolute value of the inner product (we make an additional assumption that such maximizer exists). Let

$$G[f] = G[f;\mathcal{D}] := \langle f \cdot g(f)\rangle g(f); \quad R[f] = R[f;\mathcal{D}] := f - G[f].$$

Then the pure greedy algorithm (PGA) with regard to the dictionary $\mathcal{D}$ is inductively defined by the relations: $R_0[f] := f$, $G_0[f] := 0$ and for $m \geq 1$

$$G_m[f] := G_{m-1}[f] + G\left[R_{m-1}[f]\right]; \quad R_m[f] := f - G_m[f] = R\left[R_{m-1}[f]\right].$$

There are some general results on efficiency of PGA in the case of a redundant dictionary, see [2]. Redundancy means that $\mathcal{D}$ is not a minimal system. The set $\mathcal{R}$ of all (normalized in $\mathcal{L}^2(\mathcal{B})$) ridge functions constitutes such a redundant dictionary. Another classical example of redundant dictionary for Hilbert space $\mathcal{L}^2\left([0,1]^2\right)$ is the set $\Pi := \mathcal{L}^2\left([0,1]\right) \times \mathcal{L}^2\left([0,1]\right)$, with normalization in $\mathcal{L}^2\left([0,1]^2\right)$. In the latter case, an m-term approximant looks as follows:

$$\sum_{j=1}^{m} c_j u_j(x_1) v_j(x_2).$$

A pioneering work in this direction was done by E. Schmidt [12]. It was understood later that in the case of the dictionary Π, the pure greedy algorithm *always*, i. e. for each function $f \in \mathcal{L}^2\left([0,1]^2\right)$, realizes the best m-term approximation with regard to Π. This is a strong argument in favor of PGA in nonlinear m-term approximation. However, [2] contains an example of a redundant dictionary $\mathcal{D}$ that is an orthonormal basis $\{h_j\}_{j=1}^{\infty}$ with one extra element added. For this dictionary and $f = h_1 + h_2$ one has

$$\|f - G_m[f,\mathcal{D}]\| \geq cm^{-1/2}$$

where c is a positive constant. This means that in general, PGA has a *saturation property*. On the contrary, theorem 1 above shows that in the case of the dictionary $\mathcal{D}$, the PGA is not saturated.

Before turning to the proof, let us also comment on relations between $E_M[f]$, $\sigma_N[f]$ and $\mathcal{R}_N^{\mathrm{gr}}[f]$.

First of all, obviously $\sigma_N[f] \le \mathcal{R}_N^{\rm gr}[f]$.

Second, there are no a priori restrictions on the profiles of the waves in the setting of ridge approximation problem. Thus, one has to be sure that the terms of best polynomial approximations $E_M[f]$ are natural in the estimates of efficiency of ridge approximations.

To justify this, one needs to exhibit sufficiently wide classes of functions $f(\mathbf{x})$ where the corresponding lower estimates are typical. Here, partial answers were obtained in [8], [9] and [6], [13]. For $d = 2$ and for *each radial function*, i. e. $f(\mathbf{x}) = f(|\mathbf{x}|)$ the following estimates hold true, cf. [8]

$$\sigma_N[f] \ge \frac{1}{3} E_{3N}[f], \quad N = 1, 2, \dots .$$

In [6], [13] Nikol'skii – Sobolev type spaces $H^r = H^r(\mathcal{L}^2(\mathcal{B}))$ were considered. For a fixed $r > 0$, let us define H^r as the collection of all functions $f(\mathbf{x})$ whose polynomial approximations satisfy the estimate $E_M[f] = O(M^{-r}),\ M \to \infty$. Then (cf. [6]) there exists a function $f(\mathbf{x}) \in H^r$ such that

$$\sigma_N[f] \ge N^{-\frac{r}{d-1}}, \quad N = 1, 2, \dots$$

Third, the upper estimates

$$\sigma_N[f] \le E_{N-1}[f],\ d = 2; \quad \sigma_N[f] \le E_{cN^{\frac{1}{d-1}}}[f],\ d \ge 3,\ c = c_d > 0, \qquad (3)$$

for free ridge approximations are true, without any restrictions concerning the order of decay of the sequence $\{E_M[f]\}$.

Indeed, let $\mathcal{R}^N \bigcap \mathcal{P}^{1,M}$ denote the subset of $\mathcal{R}^N$ consisting of linear combinations of planar wave polynomials of degree $\le M$:

$$\mathcal{R}^N \bigcap \mathcal{P}^{1,M} := \left\{ P(\mathbf{x}) = \sum_{j=1}^{N} p_j(\mathbf{x} \cdot \boldsymbol{\theta}_j), \quad p_j(x) \in \mathcal{P}^{1,M} \right\}.$$

Obviously, $\mathcal{R}^N \bigcap \mathcal{P}^{1,M} \subset \mathcal{P}^{d,M}$. On the other hand, it is also known, see e. g. [11], that

$$\mathcal{P}^{d,M} = \mathcal{R}^{N_M} \bigcap \mathcal{P}^{1,M}, \quad \text{where } N_M = O\left(M^{d-1}\right),\ M \to \infty,$$

that is, every polynomial $P(\mathbf{x}) \in \mathcal{P}^{d,M}$ can be represented as a ridge polynomial of order N_M and degree M, and the number N_M of planar wave

polynomials required for the representation satisfies the indicated estimate [2] from above. In the particular case of $d = 2$, see e.g. [5], $N_M \leq M + 1$. These purely algebraic facts imply estimates (3).

We will also prove an analogue of Theorem 1 concerning rates of *gridge polynomial approximation*. The corresponding algorithm is defined by the relations:

$$\begin{aligned} &G_0^{\mathrm{pol}}(\mathbf{x}) := 0; \quad R_N(\mathbf{x}) := f(\mathbf{x}) - G_N^{\mathrm{pol}}[f, \mathbf{x}], \\ &(\boldsymbol{\theta}_N,\ p_N(x)) := \arg \min_{\boldsymbol{\theta} \in \mathcal{S},\, p \in \mathcal{P}^{1,N}} \left\| \left(f(\mathbf{x}) - G_N^{\mathrm{pol}}(\mathbf{x}) \right) - p(\mathbf{x} \cdot \boldsymbol{\theta}) \right\|; \\ &G_{N+1}^{\mathrm{pol}}(\mathbf{x}) := G_N^{\mathrm{pol}}(\mathbf{x}) + p_N(\mathbf{x} \cdot \boldsymbol{\theta}_N), \quad N = 0, 1, \ldots . \end{aligned} \tag{4}$$

Such algorithm is, in general, also branching, because of non-uniqueness of the minimizer in the wave vector $\boldsymbol{\theta}$. If we follow any of such branches, after $N \geq 1$ steps the resulting approximant $G_N^{\mathrm{pol}}(\mathbf{x})$ of $f(\mathbf{x})$ will be a sum of planar wave polynomials,

$$G_N^{\mathrm{pol}}(\mathbf{x}) = \sum_{j=0}^{N-1} p_j \left(\mathbf{x} \cdot \boldsymbol{\theta}_j \right), \quad p_j(x) \in \mathcal{P}^{1,j},$$

so that $G_N^{\mathrm{pol}}(\mathbf{x}) \in \mathcal{R}^N \bigcap \mathcal{P}^{1,N-1}$.

Since the profiles of the waves are single-variate algebraic polynomials, this algorithm may represent a bigger interest in applications.

For a fixed branch, let $\mathcal{R}_N^{\mathrm{gp}}[f] := \|R_N\|$.

Theorem 2 *Assume that the sequence of best polynomial approximations of a function $f(\mathbf{x})$ satisfies the estimate $E_M[f] = O\left(E_{2M}[f]\right),\ M \to \infty$. Then*

$$\mathcal{R}_N^{\mathrm{gp}}[f] = O\left(E_{N^{\frac{1}{d-1}}}[f] \right), \quad N \to \infty, \tag{5}$$

for any branch of the gridge polynomial approximation process.

Let us outline the subsequent contents of the paper.

In the next section, Radon compass $C[f, \boldsymbol{\theta}],\ \boldsymbol{\theta} \in \mathcal{S}$, is introduced. For $f \in \mathcal{L}^2(\mathcal{B}),\ C[f, \boldsymbol{\theta}]$ is a non-negative continuous function on the sphere $\mathcal{S}$, and the maximizer(s)

$$\boldsymbol{\theta}_N := \arg \max_{\boldsymbol{\theta} \in \mathcal{S}} C\left[f_N, \boldsymbol{\theta}\right], \quad f_N := f - G_N[f], \tag{6}$$

[2] The exact values of the numbers $\nu(M, d) := \min\left\{N :\ \mathcal{P}^{d,M} = \mathcal{R}^N \bigcap \mathcal{P}^{1,M}\right\}$ seem to be unknown.

indicates the optimal wave vector on the N-th step of the gridge approximation. The definition of Radon compass is rather geometrical. It is applicable to a wider class of gridge approximation processes in Hilbert spaces than just $\mathcal{L}^2(\mathcal{B})$. However, in the case of $\mathcal{L}^2(\mathcal{B})$ the sequence $\{C\,[f_N, \boldsymbol{\theta}]\}_{N=0}^{\infty}$ is uniformly continuous on $\mathcal{S}$ (cf. lemma 3 below). The latter property seems important in applications.

After it we address to Fourier – Chebyshev analysis in $\mathcal{L}^2(\mathcal{B})$ which is crucial in the error estimates. Basing on the corresponding Parceval identities, we derive the important property of *polynomial shrinkage*. The essence is that optimization in profiles " improves $\mathcal{L}^2(\mathcal{B})$-smoothness" along the sequences $f_N,\ R_N$. This property is expressed by monotony relations of the type $E_M\,[f_{N+1}] \le E_M\,[f_N]\,,\ N = 0, 1, \ldots$. In fact, using shrinkage on each step, we substitute exact maximization of Radon compass by averaging over the sphere: $\max_{\boldsymbol{\theta}\in\mathcal{S}} C\,[f_N, \boldsymbol{\theta}] \ \longrightarrow\ \int_{\mathcal{S}} C\,[f_N, \boldsymbol{\theta}]\ \mu(d\boldsymbol{\theta})$. In the other words, the strict optimization of the couple $(\boldsymbol{\theta}, W(x))$ is replaced by a "partially stochastic", as follows:

$$\min_{\boldsymbol{\theta}\in\mathcal{S}} \min_{\{W(x)\}} \|f - W(\mathbf{x}\cdot\boldsymbol{\theta})\|^2 \ \longrightarrow\ \int_{\mathcal{S}} \min_{\{W(x)\}} \|f - W(\mathbf{x}\cdot\boldsymbol{\theta})\|^2\,\mu(d\boldsymbol{\theta}).$$

In the end we reduce the estimation problem to that of convergence rates of the sequence of truncated itegrals, of the type

$$\alpha_{N+1} \le \int_0^1 \min(\alpha_N, \varepsilon(\xi))\,d\xi, \quad N = 0, 1, \ldots,$$

where $\varepsilon(\xi)$ is a positive integrable function.

0.2 Optimization of profiles. Radon compass

Let $\omega(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^d$ be an integrable and, for simplicity sake, radial weight function i. e. $\omega(\mathbf{x}) = \omega(|\mathbf{x}|) \ge 0,\ \int_{\mathbb{R}^d} \omega(|\mathbf{x}|)\,d\mathbf{x} = |\mathcal{S}^{d-1}| \int_0^\infty x^{d-1}\omega(x)\,dx < \infty$. Denote $\mathcal{L}^2_\omega\left(\mathbb{R}^d\right)$ the Hilbert space of functions $f(\mathbf{x})$ square-integrable with regard to ω:

$$\mathcal{L}^2_\omega\left(\mathbb{R}^d\right) := \left\{ f(\mathbf{x}) :\ \sqrt{\int_{\mathbb{R}^d} |f(\mathbf{x})|^2 \omega(\mathbf{x})\,d\mathbf{x}} := \left\| f, \mathcal{L}^2_\omega\left(\mathbb{R}^d\right)\right\| < \infty \right\}.$$

For a fixed wave vector $\boldsymbol{\theta}$ and a function $f \in \mathcal{L}^2_\omega\left(\mathbb{R}^d\right)$, let us consider the problem of best approximation of f by planar waves propagating in the direction of $\boldsymbol{\theta}$:

$$\left\| f - W(\mathbf{x}\cdot\boldsymbol{\theta}), \mathcal{L}^2_\omega\left(\mathbb{R}^d\right)\right\| \longrightarrow \min \text{ in profiles } \ W(x). \tag{7}$$

It is not hard to solve this problem using the direct Radon transformation. Namely, for a function $g(\mathbf{x}) \in \mathcal{L}^1\left(\mathbb{R}^d\right)$ and $\boldsymbol{\theta} \in \mathcal{S}^{d-1}$, $x \in \mathbb{R}^1$, let $\mathrm{Rad}[g;\boldsymbol{\theta},x] := \int_{\mathbf{x}\cdot\boldsymbol{\theta}=x} g(\mathbf{x})\, d\mathbf{x}'$ where the integration is taken with respect to $(d-1)$-dimensional Lebesgue measure on the hyperplane $\mathbf{x}\cdot\boldsymbol{\theta} = x$.

Note that the Radon transform $\mathrm{Rad}[\omega;\boldsymbol{\theta},x]$ of the weight does not depend on $\boldsymbol{\theta}$ because $\omega(\mathbf{x}) = \omega(|\mathbf{x}|)$, and in fact

$$\begin{aligned}
&\mathrm{Rad}[\omega;\boldsymbol{\theta},x] = \int_{\mathbb{R}^{d-1}} \omega\left(\sqrt{x^2 + x_1^2 + \cdots + x_{d-1}^2}\right)\, dx_1 \cdots dx_{d-1} \\
&= |\mathcal{S}^{d-2}| \int_0^\infty t^{d-2}\omega\left(\sqrt{x^2+t^2}\right)\, dt \\
&= \frac{(d-1)|\mathcal{B}^{d-1}|}{2} \int_{x^2}^\infty \omega(\sqrt{\xi})(\xi - x^2)^{\frac{d-3}{2}}\, d\xi := w_{\omega,d}(x)\,.
\end{aligned}$$

The weight $\omega(\mathbf{x}) = \frac{1}{|\mathcal{B}^d|}$, $\mathbf{x} \in \mathcal{B}^d$; $\omega(\mathbf{x}) = 0$, $|\mathbf{x}| > 1$ (normalized characteristic function of $\mathcal{B}^d$), corresponds to the space $\mathcal{L}^2(\mathcal{B})$. For this particular weight, we have

$$w_d(x) = \frac{|\mathcal{B}^{d-1}|}{|\mathcal{B}^d|}\,(1-x^2)_+^{\frac{d-1}{2}}, \quad x_+ := \max(x,0). \tag{8}$$

Lemma 1 *The optimal profile* $\overline{W}(\boldsymbol{\theta},x) := \arg\min_{\{W(x)\}}$ *in the problem (7) is defined for* $x \in \mathcal{B}_\omega := \operatorname{supp} w_{\omega,d}$ *by*

$$\overline{W}(\boldsymbol{\theta},x) = \mathcal{E}_\omega[f;\boldsymbol{\theta},x] := \frac{\mathrm{Rad}[f\omega;\boldsymbol{\theta},x]}{\mathrm{Rad}[\omega;\boldsymbol{\theta},x]} = \frac{\mathrm{Rad}[f\omega;\boldsymbol{\theta},x]}{w_{\omega,d}(x)}, \tag{9}$$

and

$$\begin{aligned}
&\min_{\{W(x)\}} \left\| f - W(\mathbf{x}\cdot\boldsymbol{\theta}), \mathcal{L}^2_\omega\left(\mathbb{R}^d\right)\right\|^2 = \left\| f, \mathcal{L}^2_\omega\left(\mathbb{R}^d\right)\right\|^2 - \mathcal{C}_\omega[f,\boldsymbol{\theta}], \\
&\mathcal{C}_\omega[f,\boldsymbol{\theta}] := \left\|\overline{W}(\boldsymbol{\theta},\mathbf{x}\cdot\boldsymbol{\theta}), \mathcal{L}^2_\omega\left(\mathbb{R}^d\right)\right\|^2 = \int_{\mathcal{B}_\omega} |\mathcal{E}_\omega[f;\boldsymbol{\theta},x]|^2\ w_{\omega,d}(x)\, dx
\end{aligned} \tag{10}$$

Indeed, if $U(x)$ is a single-variate function such that the corresponding planar wave $U(\mathbf{x}\cdot\boldsymbol{\theta})$ belongs to $\mathcal{L}^2_\omega\left(\mathbb{R}^d\right)$, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d}\left(f(\mathbf{x})-\overline{W}(\boldsymbol{\theta},\mathbf{x}\cdot\boldsymbol{\theta})\right)U(\mathbf{x}\cdot\boldsymbol{\theta})\,\omega(\mathbf{x})d\mathbf{x}\\
&=\int_{\mathcal{B}_\omega}\left(\int_{\mathbf{x}\cdot\boldsymbol{\theta}=x}\left(f(\mathbf{x})-\overline{W}(\boldsymbol{\theta},x)\right)U(x)\,\omega(\mathbf{x})d\mathbf{x}'\right)\,dx\\
&=\int_{\mathcal{B}_\omega}\left(\mathrm{Rad}[f\omega;\boldsymbol{\theta},x]-\mathrm{Rad}[\omega;\boldsymbol{\theta},x]\mathcal{E}_\omega[f;\boldsymbol{\theta},x]\right)U(x)\,dx=0
\end{aligned}$$

which means that the difference $f(\mathbf{x})-\overline{W}(\boldsymbol{\theta},\mathbf{x}\cdot\boldsymbol{\theta})$ is orthogonal in $\mathcal{L}^2_\omega\left(\mathbb{R}^d\right)$ to the subspace of all planar waves propagating in the direction $\boldsymbol{\theta}$. This proves the extremal property of the profile function (9) in the problem (7), and (10) also easily follows.

We call the function $\mathcal{C}_\omega[f,\boldsymbol{\theta}]$, $\boldsymbol{\theta}\in\mathcal{S}^{d-1}$, *Radon compass.* It is non-negative, and as it is not hard to see, continuous on $\mathcal{S}^{d-1}$.

The special role of Radon compass is seen from (10):

$$\min_{\boldsymbol{\theta}\in\mathcal{S}^{d-1},\,\{W(x)\}}\left\|f-W(\mathbf{x}\cdot\boldsymbol{\theta}),\mathcal{L}^2_\omega\left(\mathbb{R}^d\right)\right\|^2=\left\|f,\mathcal{L}^2_\omega\left(\mathbb{R}^d\right)\right\|^2-\max_{\boldsymbol{\theta}\in\mathcal{S}^{d-1}}\mathcal{C}_\omega[f,\boldsymbol{\theta}]. \tag{11}$$

The maximizer $\boldsymbol{\theta}^{(0)}:=\arg\max_{\boldsymbol{\theta}\in\mathcal{S}^{d-1}}\mathcal{C}_\omega[f,\boldsymbol{\theta}]$ indicates the optimal direction on a typical step of the algorithm (1). In loose words, Radon compass serves as a navigational tool in gridge approximation process:

$$\begin{aligned}
&G_0[f,\mathbf{x}]=0;\quad f_N=f-G_N[f]\\
&\boldsymbol{\theta}_N=\arg\max_{\boldsymbol{\theta}\in\mathcal{S}}\mathcal{C}\left[f_N;\boldsymbol{\theta}\right],\quad W_N(x)=\mathcal{E}\left[f_N;\boldsymbol{\theta}_N,x\right],\\
&G_{N+1}[f,\mathbf{x}]=G_N[f,\mathbf{x}]+W_N\left(\mathbf{x}\cdot\boldsymbol{\theta}_N\right),\quad N=0,1,\dots.
\end{aligned} \tag{12}$$

Appropriate versions of Radon compass can be also constructed for gridge processes with constrained profiles. In the special case of polynomial gridge approximation (4), this construction is based on best algebraic approximation of $\mathcal{E}[f;\boldsymbol{\theta},x]$, cf. (4), (8), (9):

$$\begin{aligned}
&Q_N[f;\boldsymbol{\theta},x]:=\arg\min_{p(x)\in\mathcal{P}^{1,N}}\int_{-1}^{1}\left|\mathcal{E}\left[f;\boldsymbol{\theta},x\right]-p(x)\right|^2\,w_d(x)\,dx\,,\\
&\mathcal{C}_N[f;\boldsymbol{\theta}]:=\int_{-1}^{1}\left|Q_N[f;\boldsymbol{\theta},x]\right|^2\,w_d(x)\,dx,\\
&\boldsymbol{\theta}_N=\arg\max_{\boldsymbol{\theta}\in\mathcal{S}^{d-1}}\mathcal{C}_N[R_N;\boldsymbol{\theta}],\quad p_N(x)=Q_N\left[R_N;\boldsymbol{\theta}_N,x\right]\,.
\end{aligned} \tag{13}$$

0.3 Fourier – Chebyshev analysis. Shrinkage

The quantitative results of this paper concern only approximation in the metric of the Hilbert space $\mathcal{L}^2(\mathcal{B})$. It should be noted that proper analogs for approximation in metrics different from $\mathcal{L}^2(\mathcal{B})$ are not known. Such analogs represent an interesting circle of open problems, even for weighted Hilbert spaces $\mathcal{L}^2_\omega\left(\mathbb{R}^d\right)$; a particular example is $\mathcal{L}^2_\omega\left(\mathbb{R}^d\right)$ with the Gauss' weight $\omega(\mathbf{x}) = e^{-\pi|\mathbf{x}|^2}$.

From now on, for the sake of breviety, we will apply the notations

$$\mathcal{B} = \mathcal{B}^d,\ \|f\| = \|f, \mathcal{L}^2(\mathcal{B})\|, \quad \mathcal{S} = \mathcal{S}^{d-1},\ \|a\|_{\mathcal{S}} := \sqrt{\int_{\mathcal{S}} |a(\boldsymbol{\theta})|^2\, \mu(d\boldsymbol{\theta})},$$

$$w(x) := w_d(x) \frac{|\mathcal{B}^{d-1}|}{|\mathcal{B}^d|} (1-x^2)_+^{\frac{d-1}{2}}, \quad \|V\|_w := \sqrt{\int_{-1}^{1} |V(x)|^2\, w(x) dx},$$

$\{u_n^d(x)\}_0^\infty = \{u_n(x)\}_0^\infty$ – the system of Chebyshev – Gegenbauer polynomials orthonormal on $(-1,1)$ with the weight $w(x)$, see (8).

Let us temporarily put on hold the optimization of wave vectors in the definitions of algorithms (1) and (4). Instead, let us fix an arbitrary sequence $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_N\}_0^\infty \subset \mathcal{S}$ and consider the *profile-greedy* process, in which only profiles of the waves are optimized on each step. Such a process consists in the iterations: $f^{(0)} = R^{(0)} := f$

$$\begin{gathered} W_N(x) := \arg\min_W \|f_N - W(\mathbf{x}\cdot\boldsymbol{\theta}_N)\|\,; \\ p_N(x) := \arg\min_{p\in\mathcal{P}^{1,N}} \|R_N - p(\mathbf{x}\cdot\boldsymbol{\theta}_N)\|\,; \\ f_{N+1} := f_N - W_N(\mathbf{x}\cdot\boldsymbol{\theta}_N),\ R_{N+1} := R_N - p_N(\mathbf{x}\cdot\boldsymbol{\theta}_N). \end{gathered} \tag{14}$$

Theorem 3 *For every sequence of wave vectors $\boldsymbol{\Theta}$, the profile-greedy processes (14) are polynomial shrinkages: the matrices of best approximations are double-monotone* [3]

$$E_n\left[f_{N+1}\right] \le E_n\left[f_N\right], \quad E_n\left[R_N\right] \le E_n\left[R_N\right]. \tag{15}$$

This statement is a part of the proof of Theorems 1 and 2. We prove it here, using Chebyshev – Fourier expansion [4] and the corresponding Parceval

[3] The inequalities $E_{n+1}\left[f_N\right] \le E_n\left[f_N\right]$, $E_{n+1}\left[R_N\right] \le E_n\left[R_N\right]$ are trivial.

[4] In fact, this expansion represents the operator of *inverse Radon transformation* $R^{-1}[w\cdot]$ restricted on the functions supported in $\mathcal{B}$. Polynomials $u_n(x)$ are eigen-functions of this operator, and λ_n – the multiples of the eigen-values: $|\mathcal{B}^d| R^{-1}[wu_n](x) = \lambda_n u_n(x)$.

identity, cf. e.g. [11]

$$a_n[f,\boldsymbol{\theta}] := \int_{\mathcal{B}} f(\mathbf{x}) u_n(\mathbf{x}\cdot\boldsymbol{\theta})\,\mu(d\mathbf{x}), \quad \lambda_n = \binom{n+d-1}{n},$$
$$f(\mathbf{x}) \overset{\mathcal{L}^2(\mathcal{B})}{=} \int_{\mathcal{S}} \left(\sum_{n=0}^{\infty} \lambda_n a_n[f,\boldsymbol{\theta}] u_n(\mathbf{x}\cdot\boldsymbol{\theta})\right)\,\mu(d\boldsymbol{\theta}),$$
$$\|f\|^2 = \sum_{n=0}^{\infty} \lambda_n \int_{\mathcal{S}} |a_n[f,\boldsymbol{\theta}]|^2\,\mu(d\boldsymbol{\theta}) = \sum_{n=0}^{\infty} \lambda_n \|a_n[f]\|_{\mathcal{S}}^2\,. \tag{16}$$

The operator of orthogonal projection $P_M[f,\mathbf{x}]$ in $\mathcal{L}^2(\mathcal{B})$ onto the subspace of algebraic polynomials $\mathcal{P}^{d,M}$ and the values $E_M[f]$ of best polynomial approximation are given by

$$P_M[f,\mathbf{x}] = \int_{\mathcal{S}} \left(\sum_{n=0}^{M} \lambda_n a_n[f,\boldsymbol{\theta}] u_n(\mathbf{x}\cdot\boldsymbol{\theta})\right)\,\mu(d\boldsymbol{\theta}),$$
$$E_M[f] = \|f - P_M[f]\| = \sqrt{\sum_{n=M+1}^{\infty} \lambda_n \|a_n[f]\|_{\mathcal{S}}^2}\,. \tag{17}$$

The coefficient $a_n[f,\boldsymbol{\theta}]$ in the expansion (16) is called n-th *Chebyshev momentum* of f.

For profile-greedy processes (14), all Chebyshev momenta are shrinking in $\mathcal{L}^2(\mathcal{S})$, which can be seen from the following statement.

Lemma 2 *For* $n, N = 0, 1, \ldots$

$$\|a_n\,[f_{N+1}]\|_{\mathcal{S}} \le \|a_n\,[f_N]\|_{\mathcal{S}} \le \|a_n\,[f]\|_{\mathcal{S}}\,,$$
$$\|a_n\,[R_N]\|_{\mathcal{S}} \le \|a_n\,[R_N]\|_{\mathcal{S}} \le \|a_n\,[f]\|_{\mathcal{S}}\,. \tag{18}$$

Indeed, as a function of $\boldsymbol{\theta} \in \mathcal{S}^{d-1}$, $a_n[f,\boldsymbol{\theta}]$ is a spherical polynomial of degree n, satisfying $a_n[f,-\boldsymbol{\theta}] \equiv (-1)^n a_n[f,\boldsymbol{\theta}]$. Let us denote $\mathcal{T}_n^{\pm} = \mathcal{T}_{n,d}^{\pm}$ the subspace of all spherical polynomials with this property, and denote $K_n(x) = K_{n,d}(x)$ the *Dirichlet kernel* for $\mathcal{T}_n^{\pm}$. This kernel is the unique algebraic polynomial of degree n that satisfies $K_n(-x) \equiv (-1)^n K_n(x)$ and represents the identity operator on $\mathcal{T}_n^{\pm}$ by convolution on the sphere $\mathcal{S}$:

$$a(\boldsymbol{\varphi}) = \int_{\mathcal{S}} a(\boldsymbol{\theta}) K_n(\boldsymbol{\theta}\cdot\boldsymbol{\varphi})\,\mu(d\boldsymbol{\theta}), \quad \forall a \in \mathcal{T}_n^{\pm},\ \boldsymbol{\varphi} \in \mathcal{S} \tag{19}$$

(see [11]). It follows from the definition that for each fixed $\boldsymbol{\varphi} \in \mathcal{S}$

$$\int_{\mathcal{S}} \left(K_n(\boldsymbol{\theta} \cdot \boldsymbol{\varphi})\right)^2 \, \mu(d\boldsymbol{\theta}) = K_n(1) = \dim \mathcal{T}_{n,d}^{\pm} = \binom{n+d-1}{n} = \lambda_n. \qquad (20)$$

Further, let $V(x)$ be a single variate function, $V \in \mathcal{L}_w^2(\mathcal{B}^1)$, with the single-variate Chebyshev – Fourier representation

$$V(x) \stackrel{\mathcal{L}_w^2(\mathcal{B}^1)}{=} \sum_{n=0}^{\infty} \hat{V}_n u_n(x), \quad \hat{V}_n = \int_{-1}^{1} V(x) u_n(x) w(x)\, dx, \; n = 0, 1, \ldots,$$

and let $\boldsymbol{\varphi} \in \mathcal{S}$. For each fixed $\mathbf{x}$, $u_n(\mathbf{x} \cdot \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta} \in \mathcal{S}$ is a spherical polynomial of the class $\mathcal{T}_n^{\pm}$, and it follows from (19) that

$$V(\mathbf{x} \cdot \boldsymbol{\varphi}) \stackrel{\mathcal{L}^2(\mathcal{B})}{=} \sum_{n=0}^{\infty} \hat{V}_n u_n(\mathbf{x} \cdot \boldsymbol{\varphi}) = \int_{\mathcal{S}} \left(\sum_{n=0}^{\infty} \lambda_n \frac{\hat{V}_n}{\lambda_n} K_n(\boldsymbol{\theta} \cdot \boldsymbol{\varphi}) u_n(\mathbf{x} \cdot \boldsymbol{\theta}) \right) \mu(d\boldsymbol{\theta}). \qquad (21)$$

Comparing (21) and (16), we see that Chebyshev momenta of a planar wave function $V(\mathbf{x} \cdot \boldsymbol{\varphi})$ are multiples of shifted Dirichlet kernels:

$$a_n \left[V(\mathbf{x} \cdot \boldsymbol{\varphi}), \boldsymbol{\theta}\right] = \frac{\hat{V}_n}{\lambda_n} K_n(\boldsymbol{\theta} \cdot \boldsymbol{\varphi}). \qquad (22)$$

Now fix a wave vector $\boldsymbol{\varphi} \in \mathcal{S}$ and consider the Chebyshev – Fourier expansion of the optimal profile $\overline{W}(x) = \mathcal{E}[f; \boldsymbol{\varphi}, x]$ in the direction $\boldsymbol{\varphi}$, cf. (9). We have

$$\begin{aligned} &\mathcal{E}[f; \boldsymbol{\varphi}, x] \stackrel{\mathcal{L}_w^2(\mathcal{B}^1)}{=} \sum_{n=0}^{\infty} \hat{\mathcal{E}}_n[f, \boldsymbol{\varphi}] u_n(x), \\ &\quad \hat{\mathcal{E}}_n[f, \boldsymbol{\varphi}] = \int_{-1}^{1} \mathcal{E}[f; \boldsymbol{\varphi}, x] u_n(x) w(x)\, dx = \int_{-1}^{1} R[f; \boldsymbol{\varphi}, x] u_n(x)\, dx \\ &= \int_{\mathcal{B}} f(\mathbf{x}) u_n(\mathbf{x} \cdot \boldsymbol{\varphi})\, \mu(d\mathbf{x}) = a_n[f, \boldsymbol{\varphi}], \quad n = 0, 1, \ldots, \end{aligned} \qquad (23)$$

and it follows from (22) that the momenta of the corresponding planar wave $\mathcal{E}[f; \boldsymbol{\varphi}, \mathbf{x} \cdot \boldsymbol{\varphi}]$ and the difference $f^{(1)}(\mathbf{x}) = f(\mathbf{x}) - \mathcal{E}[f; \boldsymbol{\varphi}, \mathbf{x} \cdot \boldsymbol{\varphi}]$ are given by

$$\begin{aligned} a_n \left[\mathcal{E}[f; \boldsymbol{\varphi}, \mathbf{x} \cdot \boldsymbol{\varphi}], \boldsymbol{\theta}\right] &= \frac{a_n[f, \boldsymbol{\varphi}]}{\lambda_n} K_n(\boldsymbol{\theta} \cdot \boldsymbol{\varphi}); \\ a_n[f^{(1)}, \boldsymbol{\theta}] &= a_n[f, \boldsymbol{\theta}] - \frac{a_n[f, \boldsymbol{\varphi}]}{\lambda_n} K_n(\boldsymbol{\theta} \cdot \boldsymbol{\varphi}). \end{aligned} \qquad (24)$$

Thus, applying (19), (20) and (24) we see that

$$\|a_n[f^{(1)}]\|_{\mathcal{S}}^2 = \|a_n[f]\|_{\mathcal{S}}^2 - \frac{|a_n[f,\boldsymbol{\varphi}]|^2}{\lambda_n}\,. \tag{25}$$

Analogously, a typical step of the polynomial profile-greedy process (14) consists in the best polynomial approximation $\min_{p\in\mathcal{P}^{N,1}} \|\mathcal{E}[f;\boldsymbol{\varphi}] - p\|_w$. The partial sum $Q_N[f;\boldsymbol{\varphi},x] = \sum_{n=0}^{N} a_n[f,\boldsymbol{\varphi}]u_n(x)$ of the expansion of $\mathcal{E}[f;\boldsymbol{\varphi},x]$ provides the minimizing polynomial in the latter problem.

Let $R_N(\mathbf{x}) := f(\mathbf{x}) - Q_N[f;\boldsymbol{\varphi},\mathbf{x}\cdot\boldsymbol{\varphi}]$. Then the relations (24), (25) are modified as follows:

$$\begin{aligned} a_n[R_N,\boldsymbol{\theta}] &= a_n[f,\boldsymbol{\theta}] - \chi_N(n)\frac{a_n[f,\boldsymbol{\varphi}]}{\lambda_n}K_n(\boldsymbol{\theta}\cdot\boldsymbol{\varphi});\\ \|a_n[R_N]\|_{\mathcal{S}}^2 &= \|a_n[f]\|_s^2 - \chi_N(n)\frac{|a_n[f,\boldsymbol{\varphi}]|^2}{\lambda_n} \end{aligned} \tag{26}$$

where $\chi_N(n) = 1$ for $n \le N$ and $\chi_N(n) = 0$ for $n > N$. Relations (25) and (26) complete the proof of the lemma, and theorem 3 also follows in view of (17).

The Radon compass (10), (13) can be rewritten in terms of the Chebyshev momenta, cf. (23):

$$\begin{aligned} \mathcal{C}[f_N,\boldsymbol{\theta}] &= \|\mathcal{E}[\mathcal{C}[f_N,\boldsymbol{\theta}]\|_w^2 = \sum_{n=0}^{\infty} |a_n\left[f_N,\boldsymbol{\theta}\right]|^2\,,\\ \mathcal{C}_N\left[R_N,\boldsymbol{\theta}\right] &= \sum_{n=0}^{N} |a_n\left[R_N,\boldsymbol{\theta}\right]|^2\,. \end{aligned} \tag{27}$$

A useful property in applications is *the uniform continuity* of the sequences $\{\mathcal{C}[f_N]\}$, $\{\mathcal{C}_N[R_N]\}$ for $f \in \mathcal{L}^2(\mathcal{B})$. This property is a corollary of the shrinkage, cf. lemma 2. For a function $B(\boldsymbol{\theta})$ continuous on the sphere $\mathcal{S}$, and $M \ge 0$, let $E_M\left[\mathcal{C},\mathcal{L}^\infty(\mathcal{S})\right]$ denote the value of best approximation of B by spherical polynomials of degree $\le M$ in the uniform metric on $\mathcal{S}$:

$$E_M\left[B,\mathcal{L}^\infty(\mathcal{S})\right] := \min_{\deg T\le M}\max_{\boldsymbol{\theta}\in\mathcal{S}} |B(\boldsymbol{\theta}) - T(\boldsymbol{\theta})|.$$

Lemma 3 *The following inequalities are true*

$$\max\left(E_{2M}\left[\mathcal{C}[f_N],\mathcal{L}^\infty(\mathcal{S})\right]\,, E_{2M}\left[\mathcal{C}_N[R_N],\mathcal{L}^\infty(\mathcal{S})\right]\right) \le \Big(E_M[f,\mathcal{L}^2(\mathcal{B})]\Big)^2\,.$$

Indeed, let (cf. (27))

$$T_{M,N}(\boldsymbol{\theta}) := \sum_{n=0}^{M} |a_n[f_N, \boldsymbol{\theta}]|^2, \quad \tau_{M,N}(\boldsymbol{\theta}) := \sum_{n=0}^{\min(M,N)} |a_n[R_N, \boldsymbol{\theta}]|^2.$$

Since $a_n[f] \in \mathcal{T}_n^{\pm}$, we have $T_{M,N}$, $\tau_{M,N} \in \mathcal{T}_{2M}^{\pm}$. Further, $\|a, \mathcal{L}^{\infty}(\mathcal{S})\|^2 \leq \lambda_n \|a, \mathcal{L}^2(\mathcal{S})\|^2$ for each polynomial $a(\boldsymbol{\theta}) \in \mathcal{T}_n^{\pm}$, cf. (25). Consequently, according to (17)

$$\max\left(\|a_n\left[f_N\right], \mathcal{L}^{\infty}(\mathcal{S})\|^2\ , \|a_n\left[R_N\right], \mathcal{L}^{\infty}(\mathcal{S})\|^2\right) \leq \lambda_n \|a_n[f]\|_{\mathcal{S}}^2,$$
$$\|\mathcal{C}[f_N] - T_{M,N}, \mathcal{L}^{\infty}(\mathcal{S})\| \leq \sum_{n=M+1}^{\infty} \lambda_n \|a_n[f]\|_{\mathcal{S}}^2 = \left(E_M[f, \mathcal{L}^2(\mathcal{B})]\right)^2,$$

and exactly by the same reason

$$\|\mathcal{C}_N[R_N] - \tau_{M,N}, \mathcal{L}^{\infty}(\mathcal{S})\| \leq \left(E_M[f, \mathcal{L}^2(\mathcal{B})]\right)^2 .$$

This completes the proof of the lemma.

Now we turn to estmates of errors in gridge processes(1), (4). The maxima of $\mathcal{C}$ can be obviously estimated from below by the averages over $\mathcal{S}$:

$$\max_{\boldsymbol{\theta}} \mathcal{C}\left[f, \boldsymbol{\theta}\right] \geq \int_{\mathcal{S}} \mathcal{C}\left[f, \boldsymbol{\theta}\right] \mu(d\boldsymbol{\theta}) = \sum_{n=0}^{\infty} \|a_n\left[f\right]\|_{\mathcal{S}}^2 ,$$
$$\max_{\boldsymbol{\theta}} \mathcal{C}^{(N)}\left[f, \boldsymbol{\theta}\right] \geq \sum_{n=0}^{N} \|a_n\left[R_N\right]\|_{\mathcal{S}}^2 . \tag{28}$$

Thus, by (11), (13) and Parceval's identity (16), we obtain the following recursive estimates of errors in (1), (4) via Chebyshev momenta (note that $\lambda_0 = 1$):

$$\|f_{N+1}\|^2 \leq \sum_{n=1}^{\infty} \left(\lambda_n - 1\right) \|a_n\left[f_N\right]\|_{\mathcal{S}}^2 ,$$
$$\|R_N\|^2 \leq \sum_{n=1}^{N} \left(\lambda_n - 1\right) \|a_n\left[R_N\right]\|_{\mathcal{S}}^2 + \sum_{n=N+1}^{\infty} \lambda_n \|a_n\left[R_N\right]\|_{\mathcal{S}}^2 . \tag{29}$$

Further, by (17) we have

$$\lambda_n \|a_n[f]\|_{\mathcal{S}}^2 = \left(E_{n-1}[f]\right)^2 - \left(E_n[f]\right)^2 ,$$

so that applying Abel's transformation we can rewrite the estimates (29) in terms of best polynomial approximations:

$$\|f_{N+1}\|^2 \le \sum_{n=0}^{\infty} \Delta\xi_n \left(E_n\left[f_N\right]\right)^2 ,$$
$$\|R_N\|^2 \le \sum_{n=0}^{N-1} \Delta\xi_n \left(E_n\left[R_N\right]\right)^2 + \xi_N \left(E_N\left[R_N\right]\right)^2 \qquad (30)$$

where $\xi_n := 1/\lambda_n,\ \Delta\xi_n := \xi_n - \xi_{n+1}$. Now we make use of the polynomial shrinkage. According to Theorem 3, we can estimate the best polynomial approximations in (30) as follows:

$$E_n\left[f_N\right] \le \min\left(\|f_N\|, E_n[f]\right), \quad E_n\left[R_N\right] \le \min\left(\|R_N\|, E_n[f]\right) .$$

Consequently, the following iterative estimates are true

$$\left(\mathcal{R}^{\rm gr}_{N+1}[f]\right)^2 \le \sum_{n=0}^{\infty} \Delta\xi_n \min\left(\left(\mathcal{R}^{\rm gr}_{N}[f]\right)^2, E_n^2[f]\right) ,$$
$$\left(\mathcal{R}^{\rm gp}_{N+1}[f]\right)^2 \le \sum_{n=0}^{N-1} \Delta\xi_n \min\left(\left(\mathcal{R}^{\rm gp}_{N}[f]\right)^2, E_n^2[f]\right) + \xi_N \left(E_N[f]\right)^2 . \qquad (31)$$

0.4 Recursive truncations and difference equations

Let $\varepsilon(\xi) = \varepsilon_n := \left(E_n[f]\right)^2, \quad \xi \in (\xi_{n+1}, \xi_n],\ n = 0, 1, \ldots$, where as above, $\xi_n = 1/\lambda_n$. Obviously, $\varepsilon(\xi)$ is a non-decreasing step function on $(0, 1]$, $\varepsilon(\xi) \to 0,\ \xi \to 0$.

Consider the sequences $\{a_N\}, \{b_N\}$ defined by the integrals of successively truncated ε:

$$a_0 = b_0 := \varepsilon_0; \quad a_{N+1} = \int_0^1 \min\left(\varepsilon(\xi), a_N\right)\, d\xi,$$
$$b_{N+1} = \xi_N \varepsilon(\xi_N) + \int_{\xi_N}^1 \min\left(\varepsilon(\xi), b_N\right)\, d\xi. \qquad (32)$$

It follows from (31) that

$$\mathcal{R}^{\rm gr}_N[f] \le \sqrt{a_N}, \quad \mathcal{R}^{\rm gp}_N[f] \le \sqrt{b_N}, \quad N = 0, 1, \ldots, \qquad (33)$$

so that to finish the proofs of Theorems 1 and 2, it suffices to establish the appropriate upper estimates of the numbers $a_N,\ b_N$, defined by the recursive truncations (32).

An estimate sufficient for this purpose is provided by the next statement.

Lemma 4 *Let*

$$H(\xi) := \frac{2}{\xi} \sup_{\eta \geq \xi} \ln \frac{\varepsilon(2\eta)}{\varepsilon(\eta)}\,,\ \xi \in (0,1];$$
$$\delta(z) := \inf\{\xi \in (0,1):\ H(\xi) \leq z\},\ z > 0.$$

Then the following estimates hold for a_N, b_N defined by (32)

$$a_N \leq 2\varepsilon\left(\delta\left(\frac{N}{2}\right)\right);\quad b_N \leq 2\left(\varepsilon_{\frac{N}{2}} + \varepsilon\left(\delta\left(\frac{N}{8}\right)\right)\right)\,,\quad N = 0,1,\ldots\,. \tag{34}$$

Indeed, denote $m(y) := \mathrm{meas}\{\xi \in [0,1):\ \varepsilon(\xi) \leq y\},\ y \in [0,\varepsilon_0)$, the distribution function for $\varepsilon(\xi)$. Then $m(y) = \xi_{n+1},\quad y \in [\varepsilon_{n+1},\varepsilon_n),\ n = 0,1,\ldots$ and

$$\xi\varepsilon(\xi) + \int_\xi^1 \min\left(\varepsilon(\eta), a\right)\,d\eta = a - \int_{\varepsilon(\xi)}^a m(y)dy,\quad \varepsilon(\xi) \leq a,$$

which easily follows by consideration of the corresponding areas.

Let $M(y) := \int_0^y m(z)\,dz$. If we take $\xi = 0,\ a = a_N$, or, respectively, $\xi = \xi_n,\ a = b_N$ we see that $\{a_N\}, \{b_N\}$ in (32) coincide with the solutions of the non-linear difference equations

$$a_N - a_{N+1} = M(a_N),\quad b_N - b_{N+1} = M(b_N) - M(\varepsilon_N),\quad N = 0,1,\ldots \tag{35}$$

Since $M(y)$ is an increasing function, we have $M(y) \leq M(a_k),\ y \in (a_{k+1}, a_k)$. Thus,

$$\int_{a_N}^{a_0} \frac{dy}{M(y)} = \sum_{k=0}^{N-1} \int_{a_{k+1}}^{a_k} \frac{dy}{M(y)} \geq \sum_{k=0}^{N-1} \frac{a_k - a_{k+1}}{M(a_k)} = N. \tag{36}$$

We have $M(y) \geq \int_{\frac{y}{2}}^y m(z)\,dz \geq \frac{y}{2} m\left(\frac{y}{2}\right)$, because $m(y)$ is also an increasing function. Consequently, it follows from (36) that

$$\int_{\frac{a_N}{2}}^{\frac{a_0}{2}} \frac{dy}{ym(y)} = \frac{1}{2}\int_{a_N}^{a_0} \frac{dy}{\frac{y}{2}m\left(\frac{y}{2}\right)} \geq \frac{1}{2}\int_{a_N}^{a_0} \frac{dy}{M(y)} \geq \frac{N}{2}\,. \tag{37}$$

Further, let

$$h(\xi) := \int_{\xi}^{1} \frac{d(\ln \varepsilon(\eta))}{\eta}, \quad \xi \in (0,1] \setminus \bigcup_{0}^{\infty} \xi_n \,.$$

The function $h(\xi)$ is decreasing, piece-wise constant, and $h(\xi) \to \infty,\ \xi \to 0^+$. Moreover,

$$h(\xi) = \int_{\varepsilon(\xi)}^{\varepsilon_0} \frac{dy}{ym(y)}; \quad h(\xi) \le H(\xi). \tag{38}$$

Indeed, for $\xi_{n+1} < \xi < \xi_n$ we have

$$\int_{\varepsilon(\xi)}^{\varepsilon_0} \frac{dy}{ym(y)} = \sum_{j=0}^{n-1} \int_{\varepsilon_{j+1}}^{\varepsilon_j} \frac{dy}{ym(y)} = \sum_{j=0}^{n-1} \frac{1}{\varepsilon_{j+1}} \ln \frac{\varepsilon_j}{\varepsilon_{j+1}} = \int_{\xi}^{1} \frac{d(\ln \varepsilon(\eta))}{\eta},$$

and the inequality $h(\xi) \le H(\xi)$ easily follows by subdivision of the domain $\eta \ge \xi$ into intervals $\left[\xi 2^k, \xi 2^{k+1}\right),\ k = 0, 1, \ldots$

(37) and (38) imply:

$$\begin{aligned} & h\left(m\left(\frac{a_N}{2}\right)\right) \ge \frac{N}{2} \Longrightarrow H\left(m\left(\frac{a_N}{2}\right)\right) \ge \frac{N}{2} \\ & \Longrightarrow m\left(\frac{a_N}{2}\right) \le \delta\left(\frac{N}{2}\right) \Longrightarrow a_N \le 2\varepsilon\left(\delta\left(\frac{N}{2}\right)\right). \end{aligned} \tag{39}$$

This completes the proof of the estimate (34) for a_N.

To estimate $\{b_N\}$, we need to somewhat modify the above arguments. Let

$$A_N := \{k:\ 0 \le k < N,\ b_k \le 2\varepsilon_k\}, \quad B_N := \{k:\ 0 \le k < N,\ b_k > 2\varepsilon_k\},$$

and let ν be an integer, $0 \le \nu < N$. Then either a) $\mathrm{card} B_N \ge \nu$, or b) $\mathrm{card} A_N \ge N - \nu$. In the latter case, $b_N \le 2\varepsilon_{N-\nu}$, simply in view of monotonicity of the sequences $\{b_k\},\ \{\varepsilon_k\}$.

In the case a) for $k \in B_N$

$$b_k - b_{k+1} = \int_{\varepsilon_k}^{b_k} m(y)\,dy \ge \int_{\frac{b_k}{2}}^{b_k} m(y)\,dy \ge \frac{1}{2}\int_{0}^{b_k} m(y)\,dy = \frac{M(b_k)}{2}\,,$$

so that the estimates (37), (39) are modified as follows:

$$\begin{aligned} & \int_{b_N}^{b_0} \frac{dy}{M(y)} = \sum_{k=0}^{N-1} \int_{b_{k+1}}^{b_k} \frac{dy}{M(y)} \ge \sum_{k \in B_N} \frac{b_k - b_{k+1}}{M(b_k)} \\ & \ge \frac{\mathrm{card} B_N}{2} \ge \frac{\nu}{2}\,, \quad b_N \le 2\varepsilon\left(\delta\left(\frac{\nu}{4}\right)\right)\,. \end{aligned}$$

Summarizing, we see that

$$b_N \le 2 \min_{0\le\nu<N} \left(\varepsilon_{N-\nu} + \varepsilon\left(\delta\left(\frac{\nu}{4}\right)\right)\right), \quad b_N \le 2\left(\varepsilon_{\frac{N}{2}} + \varepsilon\left(\delta\left(\frac{N}{8}\right)\right)\right),$$

and the estimates (34) for b_N follow.

Now recall the definition of the function $\varepsilon(\xi)$ and that

$$\xi_n = \frac{1}{\lambda_n} = \frac{(d-1)!}{(n+1)\cdots(n+d-1)} \sim \frac{(d-1)!}{n^{d-1}}, \quad n \to \infty.$$

It is easy to see that the condition $E_n[f] = O\left(E_{2n}[f]\right)$ implies that $\varepsilon(2\xi) = O(\varepsilon(\xi))$, and further, $H(\xi) = O(1/\xi),\ \xi \to 0;\ \delta(v) = O(1/v),\ v \to \infty$. The latter estimate, (33) and (34) complete the proofs of theorems 1 and 2, because $\varepsilon(O(1/N)) = O(\varepsilon(1/N)) = O\left(E^2_{N^{\frac{1}{d-1}}}\right),\ N \to \infty$.

Remark 1. Obviously, we can reformulate theorems in terms of majorizing sequences, instead of exact values of best approximations.

If a sequence of of positive numbers $\{\varepsilon_n\}_0^\infty$ monotonically tends to 0 for $n \to \infty$ and satisfies the condition $\varepsilon_n = O(\varepsilon_{2n}),\ n \to \infty$ and if the estimate $E_n[f] = O(\varepsilon_n),\ n \to \infty$ holds for best polynomial approximations of a function $f \in \mathcal{L}^2(\mathcal{B})$, then the errors of gridge processes (1), (4) satisfy the estimates

$$\mathcal{R}^{\mathrm{gr}}_N[f] = O\left(\varepsilon_{N^{\frac{1}{d-1}}}\right), \quad \mathcal{R}^{\mathrm{gp}}_N[f] = O\left(\varepsilon_{N^{\frac{1}{d-1}}}\right), \quad N \to \infty.$$

Remark 2. Let us consider briefly also some classes of functions with very fast decreasing polynomial approximations, say, satisfying the $E_n[f] = O\left(e^{-n^\alpha}\right)$, where α is a positive number. In such case we have $\varepsilon(\xi) \sim \exp\left(-c\xi^{-\beta}\right),\ \beta = \frac{\alpha}{d-1},\ \xi \to 0$ and consequently

$$H(\xi) = O\left(\frac{1}{\xi^{\beta+1}}\right),\ \xi \to 0; \quad \delta(v) = O\left(v^{-\frac{1}{\beta+1}}\right),\ v \to \infty.$$

Therefore, by (33) and (34)

$$E_n[f] = O\left(e^{-n^\alpha}\right) \Longrightarrow \mathcal{R}^{\mathrm{gr}}_N[f] + \mathcal{R}^{\mathrm{gp}}_N[f] = O\left(e^{-c_1 N^\gamma}\right),\ \gamma = \frac{\alpha}{\alpha+d-1}$$

where c, c_1 are positive constants depending only on α, d.

In particular, for functions of two variables in the disc $\mathcal{B}^2$

$$E_n[f] = O\left(e^{-n}\right) \Longrightarrow \mathcal{R}^{\mathrm{gr}}_N[f] + \mathcal{R}^{\mathrm{gp}}_N[f] = O\left(e^{-c\sqrt{N}}\right).$$

References

[1] R.A. DeVore, *Nonlinear approximation,* Acta Numerica, **?**(1998), pp. 51–150.

[2] R.A. DeVore and V.N. Temlyakov, *Some remarks on greedy algorithms,* Advances in Computational Mathematics, **5**(1996), pp. 489 – 508.

[3] R.A. DeVore, K.I. Oskolkov and P.P. Petrushev, *Approximation by feed-forward neural networks,* Annals of Numerical Mathematics, **4** (1997), pp. 261 – 287.

[4] J.H. Friedman and W. Stuetzle, *Projection persuit regression,* J. Amer. Statist. Assoc.,**76**(1981), pp. 817 – 823.

[5] B. Logan, L. Schepp, *Optimal reconstruction of a function from its projections,* Duke Mathematical Journal, **42**(1975), pp. 645 – 659.

[6] V.E. Maiorov, *On best approximation by ridge functions,* Preprint, Department of Mathematics, Technion, Haifa, Israel, January 14, 1998. 27 pp.

[7] V.E. Maiorov, *On best approximation by ridge functions,* Preprint, Department of Mathematics, Technion, Haifa, Israel, 1997. 16 pp.

[8] K.I. Oskolkov, *Ridge approximation, Chebyshev – Fourier analysis and optimal quadrature formulas,* Proc. Steklov Inst. Math., **219** (1997), pp. 265 – 280 (Translation into English from *Trudy Matematicheskogo Instituta imeni V. A. Steklova*, **219** (1997), pp. 269 – 285).

[9] K.I. Oskolkov, *Ridge approximation and Kolmogorov – Nikol'skii problem,* Doklady Mathematics,**360**, issue 4 (1999), pp. 445 – 448 (in Russian); English version – Research Report 1998:06, IMI Series, USC.

[10] K.I. Oskolkov, *Non-linearity versus linearity in ridge approximation,* in Metric theory of functions and related topics of analysis, Collection of papers dedicated to the 70-th anniversary of Piotr Lavrent'evich Ul'yanov, Actuary and Finance Center Publ. (1999), ISBN 5-93379-002-8, pp. 165 – 195 (in Russian); English version – Research Report 1998:06, IMI Series, USC.

[11] P.P. Petrushev, *Approximation by ridge functions and neural networks,* SIAM J. Math. Anal., **30** (1998), pp. 155-189.

[12] E. Schmidt,*Zur Theorie der linearen und nichtlinearen Integralgleichungen,* I. Math. Annalen, **63**(1906-1907), pp. 433 – 476.

[13] V.N. Temlyakov, *On approximation by ridge functions,* Preprint, Department of Mathematics, University of South Carolina, 1996. 12 pp.