Entropy and sampling numbers of classes of ridge functions

Sebastian Mayer[a], Tino Ullrich[a,*] and Jan Vybíral[b]

August 21, 2014

[a] Hausdorff-Center for Mathematics, Endenicher Allee 62, 53115 Bonn, Germany
[b] Department of Mathematical Analysis, Charles University, Sokolovska 83, 186 00 Prague 8, Czech Republic

Abstract

We study properties of ridge functions $f(x) = g(a \cdot x)$ in high dimensions d from the viewpoint of approximation theory. The considered function classes consist of ridge functions such that the profile g is a member of a univariate Lipschitz class with smoothness $\alpha > 0$ (including infinite smoothness), and the ridge direction a has p-norm $\|a\|_p \leq 1$. First, we investigate entropy numbers in order to quantify the compactness of these ridge function classes in L_∞. We show that they are essentially as compact as the class of univariate Lipschitz functions. Second, we examine sampling numbers and face two extreme cases. In case $p = 2$, sampling ridge functions on the Euclidean unit ball suffers from the curse of dimensionality. Moreover, it is as difficult as sampling general multivariate Lipschitz functions, which is in sharp contrast to the result on entropy numbers. When we additionally assume that all feasible profiles have a first derivative uniformly bounded away from zero in the origin, then the complexity of sampling ridge functions reduces drastically to the complexity of sampling univariate Lipschitz functions. In between, the sampling problem's degree of difficulty varies, depending on the values of α and p. Surprisingly, we see almost the entire hierarchy of tractability levels as introduced in the recent monographs by Novak and Woźniakowski.

Keywords ridge functions · sampling numbers · entropy numbers · rate of convergence · information based complexity · curse of dimensionality

Mathematics Subject Classifications (2000) 41A10; 41A25; 41A50; 41A63; 46E35; 65D05; 65D15

1 Introduction

Functions depending on a large number of variables (or even infinitely many variables) naturally appear in many real-world applications. Since analytical representations are

*Corresponding author. Email: tino.ullrich@hcm.uni-bonn.de, Tel: +49 228 73 62224

rarely available, there is a need to compute approximations to such functions or at least functionals thereof. Examples include parametric and stochastic PDEs [7, 34] data analysis and learning theory [1, 8, 17], quantum chemistry [11], and mathematical finance [29].

It is a very well-known fact that approximation of smooth multivariate functions suffers from the so-called *curse of dimensionality* in many cases. Especially, for fixed smoothness, the order of approximation decays rapidly with increasing dimension [9, 23]. A recent result [27] from the area of *information-based complexity* states that on the unit cube, even uniform approximation of infinitely differentiable functions is intractable in high dimensions. These results naturally lead to the search for other assumptions than smoothness which would allow for tractable approximation, but would still be broad enough to include real-world applications. There are many different conditions of this kind. Usually, they require additional structure; for example, that the functions under consideration are tensor products or belong to some sort of weighted function space. We refer to [35] for an introduction to information-based complexity and [26, 28] for a detailed discussion of (in)tractability of high-dimensional problems.

In this work, we are interested in functions which take the form of a *ridge*. This means that for each function f there is direction a along which f may vary; along lines perpendicular to a the function is constant. In other words, the function is of the form $f(x) = g(a \cdot x)$, where g is a univariate function called the profile. Ridge functions provide a simple, coordinate-independent model, which describes inherently one-dimensional structures hidden in a high-dimensional ambient space.

That the unknown functions take the form of a ridge is a frequent assumption in statistics, for instance, in the context of *single index models*. For several of such statistical problems, minimax bounds have been studied on the basis of algorithms which exploit the ridge structure [15, 20, 32]. Another point of view on ridge functions, which has attracted attention for more than 30 years, is to approximate *by* ridge functions. An early work in this direction is [22], which took motivations from computerized tomography, and in which the term "ridge function" was actually coined. Another seminal paper is [14], which introduced *projection pursuit regression* for data analysis. More recent works include the mathematical analysis of neural networks [3, 31], and wavelet-type analysis [4]. For a survey on further approximation-theoretical results, we refer the reader to [30].

For classical setups in statistics and data analysis, it is typical that we have no influence on the choice of sampling points. In contrast, problems of *active learning* allow to *freely* choose a limited number of samples from which to recover the function. Such a situation occurs, for instance, if sampling the unknown function at a point is realized by a (costly) PDE solver. In this context, ridge functions have appeared only recently as function models. The papers [6] and [12, 37] provide several algorithms and upper bounds for the approximation error, the latter two even for the more general situation that $f(x) = g(Ax)$ with A a $(k \times d)$ matrix.

In the present paper, the central objective is to determine the complexity of approximating ridge functions in case that the only available information is a limited amount

of function values. We assume to have the following prior knowledge: the ridge functions' domain is the d-dimensional Euclidean unit ball; the profiles are Lipschitz of order $\alpha > 0$ (including infinite smoothness $\alpha = \infty$); the ridge vectors are contained in a ℓ_p^d-ball with $0 < p \leq 2$. Additionally, we study the situation that one additionally knows that $|g'(0)| \geq \kappa$ for all admissible profiles g and some prescribed $0 < \kappa \leq 1$ (of course, this makes only sense in case of $\alpha > 1$). For the function classes given by these a-priori assumptions, we prove lower and upper bounds for the deterministic worst-case error with regard to standard information. Following [25], we use the term *sampling numbers* for this worst-case error.

For given Lipschitz smoothness α, the ridge function classes are contained in the unit ball of the space of general multivariate Lipschitz functions of order α. The latter, in turn, is related to isotropic d-variate Besov spaces. For those spaces, it is known that their *entropy numbers*, which quantify the compactness in L_∞, provide a fair indicator for the behaviour of sampling numbers, see [25]. We investigate whether or not this is still the case for the ridge function classes. It turns out that they are essentially as compact as the class of univariate Lipschitz functions of the same order for all possible parameter values. For the sampling problem, however, we find a much more diverse picture. At first glance, the simple structure of ridge functions suggests that approximating them should not be too much harder than approximating a univariate function. But this is far from true in general. In fact, the sampling problem's degree of difficulty crucially depends on the constraint $|g'(0)| \geq \kappa$ in our setting. If $\kappa > 0$, then it becomes possible to first recover the ridge direction efficiently. What remains then is only the one-dimensional problem of sampling the profile. Thus, the ridge structure has a sweeping impact in this scenario and leads to a *polynomially tractable* problem. Moreover, the behaviour of entropy and sampling numbers is similar. But without the constraint on first derivatives the picture is completely different. Sampling ridge functions is now essentially as hard as sampling general Lipschitz functions over the same domain, given that all vectors in the domain may occur as ridge direction ($p = 2$). It even suffers from the *curse of dimensionality* as long as we only have finite smoothness of profiles. Supposing that $p < 2$, which can be interpreted as imposing a sparsity constraint on the ridge vectors, mitigates the situation to some extent. To our surprise, we see almost the entire spectrum of degrees of tractability as introduced in the recent monographs by Novak and Woźniakowski. In any case, however, entropy and sampling numbers behave totally different.

The paper is organized as follows. In Section 2, we define the setting in a precise way and introduce central concepts. Section 3 then is dedicated to the study of entropy numbers for the ridge function classes. Lower and upper bounds on sampling numbers are found in Section 4. Finally, in Section 5, we interpret our findings on sampling numbers in the language of information based-complexity.

2 Preliminaries

Notation For $x \in \mathbb{R}^d$, recall the (quasi-)norms $\|x\|_p = \left(\sum_{j=1}^d |x_j|^p\right)^{1/p}$ for $0 < p < \infty$, and $\|x\|_\infty = \max\{|x_1|, \ldots, |x_d|\}$. When X denotes a (quasi-)Banach space, equipped

with the (quasi-)norm $\|\cdot\|_X$, we write $B_X = \{f \in X : \ \|f\|_X < 1\}$ for the open unit ball and $\bar{B}_X$ for its closure. In the special case that $X = \ell_p^d(\mathbb{R}) = (\mathbb{R}^d, \|\cdot\|_p)$ we additionally use the notation B_p^d for the open unit ball and $\mathbb{S}_p^{d-1}$ for the unit sphere in ℓ_p^d.

The notation $f \lesssim g$ means that $f \leq Cg$ for some constant $C > 0$. Likewise, we write $f \gtrsim g$ if $f \geq cg$ for some constant $c > 0$, and $f \asymp g$ if both $f \lesssim g$ and $f \gtrsim g$.

2.1 Ridge function classes

The specific form of ridge functions suggests to describe a class of such functions in terms of two parameters: one to determine the smoothness of profiles, the other to restrict the norm of ridge directions.

Regarding smoothness, we require that ridge profiles are Lipschitz of some order. For the reader's convenience, let us briefly recall this notion. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and s be a natural number. The function space $C^s(\Omega)$ consists of those functions over the domain Ω which have partial derivatives up to order s in the interior $\mathring{\Omega}$ of Ω, and these derivatives are moreover bounded and continuous in Ω. Formally,

$$C^s(\Omega) = \big\{f : \Omega \to \mathbb{R} : \quad \|f\|_{C^s} := \max_{|\gamma| \leq s} \|D^\gamma f\|_\infty < \infty\big\},$$

where, for any multi-index $\gamma = (\gamma_1, \ldots, \gamma_d) \in \mathbb{N}_0^d$, the partial differential operator D^γ is given by

$$D^\gamma f := \frac{\partial^{|\gamma|} f}{\partial x_1^{\gamma_1} \cdots \partial x_d^{\gamma_d}}\,.$$

Here we have written $|\gamma| = \sum_{i=1}^d \gamma_i$ for the order of D^γ. For the vector of first derivatives we use the usual notation $\nabla f = (\partial f/\partial x_1, \ldots, \partial f/\partial x_d)$. Beside $C^s(\Omega)$ we further need the space of infinitely differentiable functions $C^\infty(\Omega)$ defined by

$$C^\infty(\Omega) = \big\{f : \Omega \to \mathbb{R} : \quad \|f\|_{C^\infty} := \sup_{\gamma \in \mathbb{N}_0^d} \|D^\gamma f\|_\infty < \infty\big\}\,. \tag{2.1}$$

For a function $f : \Omega \to \mathbb{R}$ and any positive number $0 < \beta \leq 1$, the *Hölder constant* of order β is given by

$$|f|_\beta := \sup_{\substack{x,y \in \Omega \\ x \neq y}} \frac{|f(x) - f(y)|}{2 \min\{1, \|x - y\|_1\}^\beta}\,.$$

This definition immediately implies the relation

$$|f|_\beta \leq |f|_{\beta'} \text{ if } 0 < \beta < \beta' \leq 1. \tag{2.2}$$

Now, for any $\alpha > 0$, we can define the *Lipschitz space* $\mathrm{Lip}_\alpha(\Omega)$. If we let $s = \lfloor\!\lfloor \alpha \rfloor\!\rfloor$ be the largest integer *strictly less* than α, it contains those functions in $C^s(\Omega)$ which have partial derivatives of order s which are moreover Hölder-continuous of order $\beta = \alpha - s > 0$. Formally,

$$\mathrm{Lip}_\alpha(\Omega) = \big\{f \in C^s(\Omega) : \quad \|f\|_{\mathrm{Lip}_\alpha(\Omega)} := \max\{\|f\|_{C^s},\ \max_{|\gamma| = s} |D^\gamma f|_\beta\} < \infty\big\}.$$

For $s \in \mathbb{N}_0$ and $1 \geq \beta_2 > \beta_1 > 0$ the following embeddings hold true

$$C^\infty(\Omega) \subset \mathrm{Lip}_{s+\beta_2}(\Omega) \subset \mathrm{Lip}_{s+\beta_1}(\Omega) \subset C^s(\Omega) \subset \mathrm{Lip}_s(\Omega)\,, \tag{2.3}$$

where the respective identity operators are of norm one. In other words, the respective unit balls satisfy the same relation. Note that the fourth inclusion only makes sense if $s \geq 1$. The third embedding is a trivial consequence of the definition. The second embedding follows from the third, and (2.2). The fourth embedding and the second imply the first. So it remains to establish the fourth embedding. We have to show that for every $\gamma \in \mathbb{N}_0^d$ with $|\gamma| = s-1$ it holds $|D^\gamma f|_1 \leq \|f\|_{C^s}$. On the one hand, Taylor's formula in $\mathbb{R}^d$ gives for some $0 < \theta < 1$

$$\begin{aligned} |D^\gamma f(x) - D^\gamma f(y)| &= |\nabla(D^\gamma f)(x + \theta(y-x)) \cdot (x-y)| \\ &\leq \max_{|\beta|=s} \|D^\beta f\|_\infty \cdot \|x-y\|_1 \\ &\leq \|f\|_{C^s} \|x-y\|_1\,. \end{aligned}$$

On the other hand, we have $|D^\gamma f(x) - D^\gamma f(y)| \leq 2\|f\|_{C^s}$. Both estimates together yield $|D^\gamma f|_1 \leq \|f\|_{C^s}$.

Having introduced Lipschitz spaces, we can give a formal definition of our ridge functions classes. For the rest of the paper, we fix as function domain the closed unit ball

$$\Omega = \bar{B}_2^d = \{x \in \mathbb{R}^d \ : \ \|x\|_2 \leq 1\}.$$

As before, let $\alpha > 0$ denote the order of Lipschitz smoothness. Further, let $0 < p \leq 2$. We define the class of ridge functions with Lipschitz profiles as

$$\mathcal{R}_d^{\alpha,p} = \left\{ f : \Omega \to \mathbb{R} \ : \ f(x) = g(a \cdot x),\ \|g\|_{\mathrm{Lip}_\alpha[-1,1]} \leq 1,\ \|a\|_p \leq 1 \right\}. \tag{2.4}$$

In addition, we define the class of ridge functions with infinitely differentiable profiles by

$$\mathcal{R}_d^{\infty,p} = \left\{ f : \Omega \to \mathbb{R} \ : \ f(x) = g(a \cdot x),\ \|g\|_{C^\infty[-1,1]} \leq 1,\ \|a\|_p \leq 1 \right\}.$$

Let us collect basic properties of these classes.

Lemma 2.1. *For any $\alpha > 0$ and $0 < p \leq 2$ the class $\mathcal{R}_d^{\alpha,p}$ is contained in $\bar{B}_{\mathrm{Lip}_\alpha(\Omega)}$ and $\mathcal{R}_d^{\infty,p}$ is contained in $\bar{B}_{C^\infty(\Omega)}$.*

Proof. Let $f \in \mathcal{R}_d^{\alpha,p}$ and $s = \lfloor\!\lfloor \alpha \rfloor\!\rfloor$. Furthermore, let $\gamma \in \mathbb{N}_0^d$ be such that $|\gamma| \leq s$. Then, there exists $g \in \mathrm{Lip}_\alpha([-1,1])$ with

$$D^\gamma f(x) = D^{|\gamma|} g(a \cdot x) a^\gamma\,, \quad x \in \Omega\,,$$

where we used the convention $a^\gamma = \prod_{i=1}^d a_i^{\gamma_i}$. Therefore, we have

$$\|D^\gamma f\|_\infty \leq \|D^{|\gamma|} g\|_\infty \|a\|_\infty^{|\gamma|} \leq \|a\|_p^{|\gamma|} \leq 1\,.$$

If we let $s \to \infty$ this immediately implies $\mathcal{R}_d^{\infty,p} \subset \bar{B}_{C^\infty(\Omega)}$. Moreover, if $|\gamma| = s$ and $\beta = \alpha - s$ we obtain by Hölder's inequality for $x, y \in \Omega$

$$\begin{aligned} |D^\gamma f(x) - D^\gamma f(y)| &= |a^\gamma| \cdot |D^s g(a \cdot x) - D^s g(a \cdot y)| \\ &\leq \|a\|_p^s \cdot |D^s g|_\beta \cdot 2 \min\{1, \|a\|_p \cdot \|x - y\|_1\}^\beta \\ &\leq 2 \min\{1, \|x - y\|_1\}^\beta . \end{aligned}$$

Consequently, we have $\|f\|_{\mathrm{Lip}_\alpha(\Omega)} \leq 1$ and hence $\mathcal{R}_d^{\alpha,p} \subset \bar{B}_{\mathrm{Lip}_\alpha(\Omega)}$. □

Note that in the special case $\alpha = 1$, we have Lipschitz-continuous profiles. Whenever $0 < \alpha_1 < \alpha_2 \leq \infty$, we have $\mathcal{R}_d^{\alpha_2,p} \subset \mathcal{R}_d^{\alpha_1,p}$, which is an immediate consequence of (2.3). Likewise, for $p < q$ we have the relation $\mathcal{R}_d^{\alpha,p} \subset \mathcal{R}_d^{\alpha,q}$.

Finally, for Lipschitz smoothness $\alpha > 1$, we want to introduce a restricted version of $\mathcal{R}_d^{\alpha,p}$, where profiles obey the additional constraint $|g'(0)| \geq \kappa > 0$. See Section 4.2 and in particular Remark 4.9 for an explanation why we study this additional constraint. We define

$$\mathcal{R}_d^{\alpha,p,\kappa} = \{g(a\cdot) \in \mathcal{R}_d^{\alpha,p} : \quad |g'(0)| \geq \kappa\}. \tag{2.5}$$

Whenever we say in the sequel that we consider ridge functions with first derivatives bounded away from zero in the origin, we mean that they are contained in the class $\mathcal{R}_d^{\alpha,p,\kappa}$ for some $0 < \kappa \leq 1$.

Taylor expansion. We introduce a straight-forward, multivariate extension of Taylor's expansion on intervals to ridge functions in $\mathcal{R}_d^{\alpha,p}$ and functions in $\mathrm{Lip}_\alpha(\Omega)$. For $x, x^0 \in \mathring{\Omega}$ we define the function $\Phi_x(\cdot)$ by

$$\Phi_x(t) := f(x^0 + t(x - x^0)) , \quad t \in [0, 1] .$$

Lemma 2.2. *Let $\alpha > 1$ and $\alpha = s + \beta$, $s \in \mathbb{N}$, $0 < \beta \leq 1$. Let further $f \in \mathrm{Lip}_\alpha(\Omega)$ and $x, x^0 \in \mathring{\Omega}$. Then there is a real number $\theta \in (0, 1)$ such that*

$$f(x) = T_{s,x^0} f(x) + R_{s,x^0} f(x) ,$$

where the Taylor polynomial $T_{s,x^0} f(x)$ is given by

$$T_{s,x^0} f(x) = \sum_{j=0}^{s} \frac{\Phi_x^{(j)}(0)}{j!} = \sum_{|\gamma| \leq s} \frac{D^\gamma f(x^0)}{\gamma!} (x - x^0)^\gamma$$

and the remainder

$$\begin{aligned} R_{s,x^0} f(x) &= \frac{1}{s!} \Big(\Phi_x^{(s)}(\theta) - \Phi_x^{(s)}(0) \Big) && (2.6) \\ &= \sum_{|\gamma| = s} \frac{D^\gamma f(x^0 + \theta(x - x^0)) - D^\gamma f(x^0)}{\gamma!} (x - x^0)^\gamma . && (2.7) \end{aligned}$$

The previous lemma has a nice consequence for the approximation of functions from $\mathcal{R}_d^{\alpha,p}$ in case $\alpha > 1$ and $0 < p \leq 2$. Let p' denote the dual index of p given by $1/\max\{p,1\} + 1/p' = 1$.

Lemma 2.3. *Let $\alpha = s + \beta > 1$ and $\Omega = \bar{B}_2^d$.*
(i) *For $f \in \mathrm{Lip}_\alpha(\Omega)$ and $x, x^0 \in \mathring{\Omega}$ we have*

$$|f(x) - T_{s,x^0} f(x)| \leq 2\|f\|_{\mathrm{Lip}_\alpha(\Omega)} \frac{\|x - x^0\|_1^\alpha}{s!}\,.$$

(ii) *Let $0 < p \leq 2$. Then for $f \in \mathcal{R}_d^{\alpha,p}$ we have the slightly better estimate*

$$|f(x) - T_{s,x^0} f(x)| \leq \frac{2}{s!}\|x - x^0\|_{p'}^\alpha\,.$$

Proof. To prove (i) we use (2.7) and the definition of $\mathrm{Lip}_\alpha(\Omega)$ and estimate as follows

$$\begin{aligned}|f(x) - T_{s,x^0} f(x)| &\leq \sum_{|\gamma|=s} \frac{|D^\gamma f(x^0 + \theta(x - x^0)) - D^\gamma f(x^0)|}{\gamma!} |(x - x^0)^\gamma| \\ &\leq 2\|f\|_{\mathrm{Lip}_\alpha(\Omega)} \min\{1, \|x - x^0\|_1\}^\beta \cdot \sum_{|\gamma|=s} \frac{\prod_{i=1}^d |x_i - x_i^0|^{\gamma_i}}{\gamma!}\,.\end{aligned}$$

Using mathematical induction it is straight-forward to verify the multinomial identity

$$(a_1 + \cdots + a_d)^s = \sum_{|\gamma|=s} \frac{s!}{\gamma!} a_1^{\gamma_1} \cdot \cdots \cdot a_d^{\gamma_d}\,.$$

Hence, choosing $a_i = |x_i - x_i^0|$ we can continue estimating

$$|f(x) - T_{s,x^0} f(x)| \leq 2\|f\|_{\mathrm{Lip}_\alpha(\Omega)} \min\{1, \|x - x^0\|_1\}^\beta \frac{\|x - x^0\|_1^s}{s!}$$

and obtain the assertion in (i).
For showing the improved version (ii) for functions of type $f(x) = g(a \cdot x)$ we use formula (2.6) of the Taylor remainder. We easily see that for $t \in (0,1)$ it holds

$$\Phi_x^{(s)}(t) = g^{(s)}\Big(a \cdot (x^0 + t(x - x^0))\Big) \cdot [a \cdot (x - x^0)]^s\,.$$

Using Hölder continuity of $g^{(s)}$ of order β and Hölder's inequality we see that

$$\begin{aligned}|f(x) - T_{s,x^0} f(x)| &\leq \frac{1}{s!}\Big|[a \cdot (x - x^0)]^s \cdot \Big\{g^{(s)}\Big(a \cdot (x^0 + \theta(x - x^0))\Big) - g^{(s)}(a \cdot x^0)\Big\}\Big| \\ &\leq \frac{1}{s!}\|a\|_p^s \cdot \|x - x^0\|_{p'}^s \cdot 2\min\{1, |\theta a \cdot (x - x^0)|^\beta\} \\ &\leq \frac{2}{s!}\|x - x^0\|_{p'}^\alpha\,.\end{aligned}$$

The proof is complete. □

2.2 Information complexity and tractability

In this work, we want to approximate ridge functions from $\mathcal{F} = \mathcal{R}_d^{\alpha,p}$ or $\mathcal{F} = \mathcal{R}_d^{\alpha,p,\kappa}$ by means of deterministic sampling algorithms, using a limited amount of function values. Any allowed algorithm S consists of an *information map* $N_S^{\mathrm{ada}} : \mathcal{F} \to \mathbb{R}^n$, and a *reconstruction map* $\varphi_S : \mathbb{R}^n \to L_\infty(\Omega)$. For given $f \in \mathcal{F}$, the former provides function values $f(x_1), \dots, f(x_n)$ at points $x_1, \dots, x_n \in \Omega$, which are allowed to be chosen *adaptively*. Adaptivity here means that x_i may depend on the preceding values $f(x_1), \dots, f(x_{i-1})$. According to [26], we speak of *standard information*. The reconstruction map then builds an approximation to f based on those function values provided by the information map.

Formally, we consider the class of deterministic, adaptive sampling algorithms $\mathcal{S}^{\mathrm{ada}} = \bigcup_{n\in\mathbb{N}} \mathcal{S}_n^{\mathrm{ada}}$, where

$$\mathcal{S}_n^{\mathrm{ada}} = \Big\{ S : \mathcal{F} \to L_\infty(\bar{B}_2^d) \ : \\ S = \varphi \circ N^{\mathrm{ada}}, \varphi : \mathbb{R}^m \to L_\infty \ N^{\mathrm{ada}} : \mathcal{F} \to \mathbb{R}^m, \ m \le n \Big\}.$$

The *n-th minimal worst-case error*

$$g_{n,d}^{\mathrm{ada}}(\mathcal{F}, L_\infty) := \mathrm{err}_{n,d}(\mathcal{F}, \mathcal{S}^{\mathrm{ada}}, L_\infty) = \inf \big\{ \sup_{f\in\mathcal{F}} \|f - S(f)\|_\infty : S \in \mathcal{S}_n^{\mathrm{ada}} \big\},$$

describes the approximation error of the best possible algorithm. Stressing that function values are the only available information, we refer to $g_{n,d}^{\mathrm{ada}}(\mathcal{F}, L_\infty)$ as the n-th *(adaptive) sampling number*. To reveal the effect of adaption, it is useful to compare adaptive algorithms with the subclass $\mathcal{S} \subset \mathcal{S}^{\mathrm{ada}}$ of *non-adaptive*, deterministic algorithms; that is, for each algorithm $S \in \mathcal{S}$ the information map is now of the form $N_S = (\delta_{x_1}, \dots, \delta_{x_n})$, with $n \in \mathbb{N}$ and $x_1, \dots, x_n \in \bar{B}_2^d$. This corresponds to *non-adaptive standard information* in [26]. The associated n-th worst-case error

$$g_{n,d}(\mathcal{F}, L_\infty) := \inf_{S\in\mathcal{S}_n} \sup_{f\in\mathcal{F}} \big\| f - S(f) \big\|_\infty = \mathrm{err}_{n,d}(\mathcal{F}, \mathcal{S}_n, L_\infty)$$

coincides with the standard n-th *sampling number* as known in approximation theory [25]. As a third restriction, let us introduce the n-th *linear* sampling number $g_{n,d}^{\mathrm{lin}}(\mathcal{F}, L_\infty)$; here, only algorithms from $\mathcal{S}$ with linear reconstruction map are allowed. Clearly,

$$g_{n,d}^{\mathrm{ada}}(\mathcal{F}, L_\infty) \le g_{n,d}(\mathcal{F}, L_\infty) \le g_{n,d}^{\mathrm{lin}}(\mathcal{F}, L_\infty).$$

Remark 2.4. Studying adaptive algorithms makes sense since the considered ridge functions classes are not *convex*. Hence, the general results on linear problems [26, Section 4.2] do not apply here. Nevertheless, the analysis in Section 4 will reveal that neither adaptivity nor non-linearity lead to any substantial improvement in the approximation of ridge functions defined on a Euclidean ball.

Whenever we speak of sampling of ridge functions in the following, we refer to the problem of approximating ridge functions in $\mathcal{F}$ by sampling algorithms from $\mathcal{S}^{\mathrm{ada}}$, the L_∞-approximation error measured in the worst-case. Its *information complexity* $n(\varepsilon, d)$ is given for $0 < \varepsilon \le 1$ and $d \in \mathbb{N}$ by

$$n(\varepsilon, d) := \min\{ n \in \mathbb{N} : g_{n,d}^{\mathrm{ada}}(\mathcal{F}, L_\infty) \le \varepsilon \}.$$

2.3 Entropy numbers

The concept of entropy numbers is central to this work. They can be understood as a measure to quantify the compactness of a set with respect to some reference space. For a detailed discussion and historical remarks, we refer to the monographs [5, 10]. The k-th entropy number $e_k(K,X)$ of a subset K of a (quasi-)Banach space X is defined as

$$e_k(K,X) = \inf\left\{\varepsilon > 0 : K \subset \bigcup_{j=1}^{2^{k-1}} (x_j + \varepsilon \bar{B}_X) \text{ for some } x_1,\dots,x_{2^{k-1}} \in X\right\}.$$

Note that $e_k(K,X) = \inf\{\varepsilon > 0 : N_\varepsilon(K,X) \le 2^{k-1}\}$ holds true, where

$$N_\varepsilon(K,X) := \min\left\{n \in \mathbb{N} : \quad \exists x_1,\dots,x_n \in X : \ K \subset \bigcup_{j=1}^{n} (x_j + \varepsilon \bar{B}_X)\right\}$$

denotes the *covering number* of the set K in the space X, which is the minimal natural number n such that there is an ε-net of K in X of n elements. We can introduce entropy numbers for operators, as well. The k-th entropy number $e_k(T)$ of an operator $T : X \to Y$ between two quasi-Banach spaces X and Y is defined by

$$e_k(T) = e_k(T(\bar{B}_X), Y).$$

The results in Section 3 and 4 rely to a great degree on entropy numbers of the identity operator between the two finite dimensional spaces $X = \ell_p^d(\mathbb{R})$, and $Y = \ell_q^d(\mathbb{R})$. Their behavior is understood very well, see [10, 21, 33, 36]. For the reader's convenience, we restate the result.

Lemma 2.5. *Let $0 < p \le q \le \infty$ and let k and d be natural numbers. Then,*

$$e_k(\bar{B}_p^d, \ell_q^d) \asymp \begin{cases} 1 & : \ 1 \le k \le \log(d), \\ \left(\frac{\log(1+d/k)}{k}\right)^{1/p-1/q} & : \ \log(d) \le k \le d, \\ 2^{-k/d} d^{1/q-1/p} & : \ k \ge d\,. \end{cases}$$

The constants behind "$\asymp$" do neither depend on k nor on d. They only depend on the parameters p and q.

If we consider entropy numbers of ℓ_p^d-spheres instead of ℓ_p^d-balls in ℓ_q^d, the situation is quite similar. We are not aware of a reference where this has already been formulated thoroughly.

Lemma 2.6. *Let $d \in \mathbb{N}$, $d \ge 2$, $0 < p \le q \le \infty$, and $\bar{p} = \min\{1,p\}$. Then,*

(i)
$$2^{-k/(d-1)} d^{1/q-1/p} \lesssim e_k(\mathbb{S}_p^{d-1}, \ell_q^d) \lesssim 2^{-k/(d-\bar{p})} d^{1/q-1/p}, \quad k \ge d.$$

(ii)

$$e_k(\mathbb{S}_p^{d-1}, \ell_q^d) \asymp \begin{cases} 1 & : \ 1 \le k \le \log(d), \\ \left(\frac{\log(1+d/k)}{k}\right)^{1/p-1/q} & : \ \log(d) \le k \le d. \end{cases}$$

The constants behind "$\asymp$" only depend on p and q.

Proof. For given $\varepsilon > 0$, an ε-covering $\{y_1, \dots, y_N\}$ of $\mathbb{S}_p^{d-1}$ in ℓ_p^d fulfils

$$(1+\varepsilon)\bar{B}_p^d \setminus (1-\varepsilon)\bar{B}_p^d \subseteq \bigcup_{i=1}^{N} (y_i + 2^{1/\bar{p}}\varepsilon \bar{B}_p^d). \tag{2.8}$$

Let $\bar{q} = \min\{1, q\}$. For given $\varepsilon > 0$, a maximal set $\{x_1, \dots, x_M\} \subset \mathbb{S}_p^{d-1}$ of vectors with mutual distance greater ε obeys

$$\bigcup_{i=1}^{M} (x_i + 2^{-1/\bar{q}}\,\varepsilon \bar{B}_q^d) \subseteq (1+\varepsilon_d^{\bar{p}})^{1/\bar{p}} \bar{B}_p^d \setminus (1-\varepsilon_d^{\bar{p}})^{1/\bar{p}} \bar{B}_p^d, \tag{2.9}$$

where $\varepsilon_d = 2^{-1/\bar{q}}\,\varepsilon\, d^{1/p-1/q}$.
(i). A standard volume argument applied to (2.8) yields $h(\varepsilon) \le N\varepsilon^d 2^{d/\bar{p}}$, where $h(\varepsilon) = (1+\varepsilon)^d - (1-\varepsilon)^d$. First-order Taylor expansion in ε allows to estimate $h(\varepsilon) \ge d\varepsilon$. Solving for N yields a lower bound for covering numbers in case $p = q$. The lower bound in case $p \ne q$ follows from the trivial estimate $e_k(\mathbb{S}_p^{d-1}, \ell_q^d) \ge d^{1/q-1/p}\, e_k(\mathbb{S}_p^{d-1}, \ell_p^d)$.

For the upper bound in case $p = q$ a standard volume argument applied to (2.9) yields $M\varepsilon^d 2^{-d/\bar{p}} \le h_p(\varepsilon^{\bar{p}}/2)$ with $h_p(x) = (1+x)^{d/\bar{p}} - (1-x)^{d/\bar{p}}$. The mean value theorem gives $h_p(x) \le d/\bar{p}\, 2^{d/\bar{p}} x$ if $0 < x \le 1$. Hence, we get $h_p(\varepsilon^{\bar{p}}/2) \le d/\bar{p}\, 2^{d/\bar{p}} \varepsilon^{\bar{p}}/2$. Solving for M gives an upper bound for packing numbers and hence also for covering numbers. In case $p \ne q$ we again use (2.9) and pass to volumes. This time the quotient $\mathrm{vol}(B_p^d)/\mathrm{vol}(B_q^d)$ remains in the upper bound for M. The given bounds now easily translate to the stated bounds on entropy numbers. In case $p \ne q$ one has to take

$$\Big[\frac{\mathrm{vol}(B_p^d)}{\mathrm{vol}(B_q^d)}\Big]^{1/(d-\bar{p})} \asymp d^{1/q-1/p}$$

into account to get the additional factor in d.
(ii). The proof by Kühn [21] immediately gives the lower bound. The upper bound follows trivially from $\mathbb{S}_p^{d-1} \subset \bar{B}_p^d$.

□

Remark 2.7. Note, that in case $p \ge 1$ we have the sharp bounds

$$e_k(\mathbb{S}_p^{d-1}, \ell_q^d) \asymp \begin{cases} 1 & : \ 1 \le k \le \log(d), \\ \left(\frac{\log(1+d/k)}{k}\right)^{1/p-1/q} & : \ \log(d) \le k \le d, \\ 2^{-\frac{k}{d-1}} d^{1/q-1/p} & : \ k \ge d\,. \end{cases}$$

In case $p < 1$ there remains a gap between the upper and lower estimate for $e_k(\mathbb{S}_p^{d-1}, \ell_q^d)$ if $k \ge d$. However, this gap can be closed by using a different proof technique, see [18].

3 Entropy numbers of ridge functions

This section is devoted to the study of entropy numbers of the classes $\mathcal{R}_d^{\alpha,p}$ and $\mathcal{R}_d^{\alpha,p,\kappa}$. Specifically, we want to relate their behavior to that of entropy numbers of uni- and multivariate Lipschitz functions. This will give us an understanding how "large" the ridge function classes are. Let us stress that we are interested in the dependence of the entropy numbers on the underlying dimension d, as it is usually done in the area of information-based complexity.

To begin with, we examine uni- and multivariate Lipschitz functions from $\mathrm{Lip}_\alpha[-1,1]$ and $\mathrm{Lip}_\alpha(\Omega)$. Recall the notations $B_\alpha := B_{\mathrm{Lip}_\alpha[-1,1]}$ and $B_{\mathrm{Lip}_\alpha(\Omega)}$ for the respective open unit balls. The behavior of entropy numbers of univariate Lipschitz functions is well-known, see for instance [23, Chap. 15, §2, Thm. 2.6].

Lemma 3.1. *For $\alpha > 0$ there exist two constants $0 < c_\alpha < C_\alpha$ such that*

$$c_\alpha k^{-\alpha} \le e_k(\bar{B}_\alpha, L_\infty([-1,1])) \le C_\alpha k^{-\alpha}\,, \quad k \in \mathbb{N}\,.$$

This behavior does not change if we consider only functions with first derivative in the origin bounded away from zero, as we do with the profiles in the class $\mathcal{R}_d^{\alpha,p,\kappa}$.

Proposition 3.2. *Let $\alpha > 1$ and $0 < \kappa \le 1$. Consider the class*

$$\mathrm{Lip}_\alpha^\kappa([-1,1]) = \{f \in \mathrm{Lip}_\alpha([-1,1]) : \|f\|_{\mathrm{Lip}_\alpha[-1,1]} \le 1\ ,\ \left|f'(0)\right| \ge \kappa\}.$$

For the entropy numbers of this class we have two constants $0 < c_\alpha < C_\alpha$, such that

$$c_\alpha k^{-\alpha} \le e_k(\mathrm{Lip}_\alpha^\kappa([-1,1]), L_\infty([-1,1])) \le C_\alpha k^{-\alpha}\,, \quad k \in \mathbb{N}\,.$$

Proof. The upper bound is immediate by Lemma 3.1. The lower bound is proven in the same way as for general univariate Lipschitz functions of order α except that we have to adapt the "bad" functions such that they meet the constraint on the first derivative in the origin. Put again $s = \lfloor\!\lfloor \alpha \rfloor\!\rfloor$ and $\beta = \alpha - s > 0$. Consider the standard smooth bump function

$$\varphi(x) = \begin{cases} e^{-\frac{1}{1-x^2}} & :\ |x| < 1, \\ 0 & :\ |x| \ge 1. \end{cases}$$

Let

$$\psi_{k,b}(x) = \frac{c_\alpha \cdot \varphi(5k(x-b))}{k^\alpha}, \quad k \in \mathbb{N},\ b \in \mathbb{R},$$

where $c_\alpha = 1/(5^\alpha \|\varphi\|_{\mathrm{Lip}_\alpha})$. The scaling factor $c_\alpha k^{-\alpha}$ assures $\psi_{k,b} \in \mathrm{Lip}_\alpha([-1,1])$. Let $a = \pi/4 - 1/5$ and $I = [a, a + 2/5] \subset (0,1)$. We put $h(x) = \sin(x)$ and

$$\gamma = \sup_{j \in \mathbb{N}_0} \max_{x \in I} |h^{(j)}(x)| = \max_{x \in I} \max\{\cos(x), \sin(x)\} < 1. \tag{3.1}$$

For any multi-index $\theta = (\theta_1, \dots, \theta_k) \in \{0,1\}^k$ let

$$g_\theta = (1-\gamma) \sum_{j=1}^{k} \theta_j \psi_{k,b_j}, \qquad b_j = a + \frac{2j-1}{5k}.$$

Observe, that $\operatorname{supp} g_\theta \subset I$.

There are 2^k such multi-indices and for two different multi-indices $\hat{\theta}$ and $\tilde{\theta}$ we have

$$\left\| g_{\hat{\theta}} - g_{\tilde{\theta}} \right\|_\infty = (1-\gamma) \left\| \psi_{k,0} \right\|_\infty = c_\alpha (1-\gamma) e^{-1} k^{-\alpha}.$$

Put $f_\theta = h + g_\theta$. Because of the scaling factors, it is assured that $f_\theta \in \operatorname{Lip}_\alpha^\kappa([-1,1])$. On the other hand, $f'_\theta(0) = \cos(0) = 1$. Obviously, $\left\| f_{\tilde{\theta}} - f_{\hat{\theta}} \right\|_\infty = \left\| g_{\tilde{\theta}} - g_{\hat{\theta}} \right\|_\infty$. We conclude

$$e_k(\operatorname{Lip}_\alpha^\kappa([-1,1]), L_\infty) \geq c'_\alpha k^{-\alpha}$$

for $c'_\alpha = (1-\gamma) e^{-1} c_\alpha$.

□

Considering multivariate Lipschitz functions, decay rates of entropy numbers change dramatically compared to those of univariate Lipschitz functions; they depend exponentially on $1/d$. This is known if the domain is a cube $\Omega = I^d$, see [23, Chap. 15, §2]. We provide an extension to our situation where the domain is $\Omega = \bar{B}_2^d$.

Proposition 3.3. *Let $\alpha > 0$. For natural numbers n and k such that $2^{k-1} < n \leq 2^k$ we have*

$$e_n(\bar{B}_{\operatorname{Lip}_\alpha(\bar{B}_2^d)}, L_\infty(\bar{B}_2^d)) \geq c_\alpha e_{k+1}(id : \ell_2^d \to \ell_2^d)^\alpha.$$

In particular, we have $e_n(id : \operatorname{Lip}_\alpha(\bar{B}_2^d) \to L_\infty(\bar{B}_2^d)) \gtrsim n^{-\alpha/d}$.

Proof. Consider the radial bump function $\varphi(x)$ given by

$$\varphi(x) = \begin{cases} e^{-\frac{1}{1-\|x\|_2^2}} & : \ \|x\|_2 < 1, \\ 0 & : \ \|x\|_2 \geq 1. \end{cases}$$

Let $s = \lfloor\!\lfloor \alpha \rfloor\!\rfloor$. With $c_\alpha := (\|\varphi\|_{\operatorname{Lip}_\alpha})^{-1}$ the rescaling

$$\varphi_\varepsilon^\alpha(x) := c_\alpha \varepsilon^\alpha \varphi(x/\varepsilon)$$

is contained in the closed unit ball of $\operatorname{Lip}_\alpha(\Omega)$.

For $0 < \varepsilon < e_{k+1}(\bar{B}_2^d, \ell_2^d)$, let $\{x_1, \dots, x_m\}$ be a maximal set of 2ε-separated points in the Euclidean ball $\bar{B}_2^d$, the distance measured in ℓ_2^d. Clearly, every closed ball of radius ε contains at most one x_i, and consequently every covering of $\bar{B}_2^d$ by balls of radius ε contains at least m elements. The choice of ε implies $m > 2^k \geq n$. For every multi-index $\theta \in \{0,1\}^m$, we define

$$f_\theta(x) := \sum_{j=1}^{m} \theta_j \varphi_\varepsilon^\alpha(x - x_j).$$

By construction of $\varphi_\varepsilon^\alpha$, it is assured that $f_\theta \in \mathrm{Lip}_\alpha(\Omega)$ and $\|f_\theta\|_{\mathrm{Lip}_\alpha} \leq 1$. Moreover, we see immediately that $\|f_\theta\|_\infty = c_\alpha e^{-1}\varepsilon^\alpha$, and

$$\|f_\theta - f_{\theta'}\|_\infty \geq c_\alpha e^{-1}\varepsilon^\alpha =: \varepsilon_1$$

for $\theta \neq \theta'$. Therefore, the set $\{f_\theta : \theta \in \{0,1\}^m\}$ consists of 2^m functions with mutual distances greater than or equal to ε_1. This implies

$$2^n < 2^m < N_{\varepsilon_1/2}(\bar{B}_{\mathrm{Lip}_\alpha(\Omega)}, L_\infty).$$

Hence, $e_n(id : \mathrm{Lip}_\alpha(\Omega) \to L_\infty(\Omega)) > \varepsilon_1/2$, and by the choice of ε also

$$e_n(\bar{B}_{\mathrm{Lip}_\alpha(\Omega)}, L_\infty(\Omega)) > c'_\alpha e_k(id : \ell_2^d \to \ell_2^d)^\alpha$$

for $c'_\alpha = c_\alpha/(4e)$. Now it follows immediately from the estimate above and Lemma 2.5 that

$$e_n(\bar{B}_{\mathrm{Lip}_\alpha(\Omega)}, L_\infty(\Omega)) \gtrsim 2^{-\alpha k/d} \gtrsim n^{-\alpha/d}.$$

□

Now consider ridge functions with Lipschitz profile as given by the class $\mathcal{R}_d^{\alpha,p}$.

Theorem 3.4. *Let d be a natural number, $\alpha > 0$, and $0 < p \leq 2$. Then, for any $k \in \mathbb{N}$,*

$$\frac{1}{2}\max\{e_{2k}(\bar{B}_p^d, \ell_2^d), e_{2k}(\bar{B}_\alpha, L_\infty)\} \leq e_{2k}(\mathcal{R}_d^{\alpha,p}, L_\infty) \leq e_k(\bar{B}_p^d, \ell_2^d)^{\min\{\alpha,1\}} + e_k(\bar{B}_\alpha, L_\infty)\,.$$

Proof. Lower bounds: For $\varepsilon > 0$ let $g_1, \ldots, g_n$ be a maximal set of univariate Lipschitz functions in $\bar{B}_\alpha$ with mutual distances $\|g_i - g_j\|_\infty > \varepsilon$ for $i \neq j$. Now, let $a = (1, 0, \ldots, 0)$ and put $f_i(x) = g_i(a \cdot x)$ for $i = 1, \ldots, n$. Then, of course, we have $f_i \in \mathcal{R}_d^{\alpha,p}$, and

$$\|f_i - f_j\|_\infty = \|g_i - g_j\|_\infty > \varepsilon\,.$$

Consequently, the functions $f_1, \ldots, f_n$ are ε-separated, as well. This implies

$$e_{2k}(\mathcal{R}_d^{\alpha,p}, L_\infty) \geq \frac{1}{2}\, e_{2k}(\bar{B}_\alpha, L_\infty)\,.$$

On the other hand, for $\varepsilon > 0$, let $a_1, \ldots, a_n$ be a maximal set of vectors in $\bar{B}_p^d$ with pairwise distances $\|a_i - a_j\|_2 > \varepsilon$. Furthermore, let $g(t) = t$ and put $\tilde{f}_i(x) = g(a_i \cdot x)$ for $i = 1, \ldots, n$. Then $\tilde{f}_i \in \mathcal{R}_d^{\alpha,p}$ and

$$\begin{aligned}\|\tilde{f}_i - \tilde{f}_j\|_\infty &= \sup_{x \in \bar{B}_2^d} |\tilde{f}_i(x) - \tilde{f}_j(x)| = \sup_{x \in \bar{B}_2^d} |g(a_i \cdot x) - g(a_j \cdot x)| \\ &= \sup_{x \in \bar{B}_2^d} |(a_i - a_j) \cdot x| = \|a_i - a_j\|_2 > \varepsilon.\end{aligned}$$

Thus, the functions $\tilde{f}_1, \ldots, \tilde{f}_n$ are ε-separated w.r.t. the L_∞-norm. This implies

$$e_{2k}(\mathcal{R}_d^{\alpha,p}, L_\infty) \geq \frac{1}{2}\, e_{2k}(\bar{B}_p^d, \ell_2^d)\,.$$

Upper bound: We use the shorthand $\bar{\alpha} = \min\{\alpha, 1\}$. Let $1/2 > \varepsilon_1, \varepsilon_2 > 0$ be fixed and put $\varepsilon := \varepsilon_1^{\bar{\alpha}} + \varepsilon_2$. Let $\mathcal{N} = \{g_1, \dots, g_n\}$ be a minimal ε_1-net of $\bar{B}_\alpha$ in the L_∞-norm. Further, let $\mathcal{M} = \{a_1, \dots, a_m\}$ be a minimal ε_2-net of $\bar{B}_p^d$ in the ℓ_2^d-norm.

Now, fix some ridge function $f : x \mapsto g(a \cdot x)$ in $\mathcal{R}_d^{\alpha,p}$, i.e. $\|g\|_{\mathrm{Lip}_\alpha} \le 1$ and $\|a\|_p \le 1$. Then there is a function $g_i \in \mathcal{N}$ with $\|g - g_i\|_\infty \le \varepsilon_1$ and a vector $a_j \in \mathcal{M}$ with $\|a - a_j\|_2 \le \varepsilon_2$. We obtain

$$\begin{aligned}\|g(a \cdot x) - g_i(a_j \cdot x)\|_\infty &\le \sup_{x \in \bar{B}_2^d} |g(a \cdot x) - g(a_j \cdot x)| + |g(a_j \cdot x) - g_i(a_j \cdot x)| \\ &\le \sup_{x \in \bar{B}_2^d} |g|_{\bar{\alpha}} \cdot |a \cdot x - a_j \cdot x|^{\bar{\alpha}} + \|g - g_i\|_\infty \\ &\le \|a - a_j\|_2^{\bar{\alpha}} + \|g - g_i\|_\infty \le \varepsilon_1^{\bar{\alpha}} + \varepsilon_2 = \varepsilon.\end{aligned}$$

Hence, the set $\{x \to g(a \cdot x) : g \in \mathcal{N}, a \in \mathcal{M}\}$ is an ε-net of $\mathcal{R}_d^{\alpha,p}$ in $L_\infty(\Omega)$ with cardinality

$$\#\mathcal{N} \cdot \#\mathcal{M} = N_{\varepsilon_1}(\bar{B}_\alpha, L_\infty) \cdot N_{\varepsilon_2}(\bar{B}_p^d, \ell_2^d)\,.$$

Consequently, $N_\varepsilon(\mathcal{R}_d^{\alpha,p}, L_\infty) \le \#\mathcal{N} \cdot \#\mathcal{M}$ and we conclude that

$$e_{2k}(\mathcal{R}_d^{\alpha,p}, L_\infty) \le e_k(\bar{B}_p^d, \ell_2^d)^{\bar{\alpha}} + e_k(\bar{B}_\alpha, L_\infty)\,.$$

□

Remark 3.5. In view of Proposition 3.2, it is easy to see that Theorem 3.4 keeps valid if we replace the class $\mathcal{R}_d^{\alpha,p}$ by $\mathcal{R}_d^{\alpha,p,\kappa}$.

We exemplify the consequences of Theorem 3.4 by considering the case $p = 2$; for $0 < p < 2$ estimates would be similar. As the corollary below shows, entropy numbers of ridge functions asymptotically decay as fast as those of their profiles. In contrast to multivariate Lipschitz functions on Ω, the dimension d does not appear in the decay rate's exponent. It only affects how long we have to wait until the asymptotic decay becomes visible.

Corollary 3.6. *Let d be a natural number and $\alpha > 0$. For the entropy numbers of $\mathcal{R}_d^{\alpha,2}$ in $L_\infty(\Omega)$ we have*

$$\max(k^{-\alpha}, 2^{-k/d}) \lesssim e_k(\mathcal{R}_d^{\alpha,2}, L_\infty) \lesssim \begin{cases} 1 & : k \le c_\alpha d \log d, \\ k^{-\alpha} & : k \ge c_\alpha d \log d\,, \end{cases} \tag{3.2}$$

for some universal constant $c_\alpha > 0$ which does not depend on d.

Before we turn to the proof, let us note that (3.2) implies that

$$e_k(\mathcal{R}_d^{\alpha,2}, L_\infty) \asymp 1 \quad \text{if} \quad k \le d,$$

and

$$e_k(\mathcal{R}_d^{\alpha,2}, L_\infty) \asymp k^{-\alpha} \quad \text{if} \quad k \ge c_\alpha d \ln d.$$

Hence, entropy numbers of ridge functions are guaranteed to decay like those of their profiles for $k \ge c_\alpha d \log d$—and surely behave differently for $k \le d$.

Proof of Corollary 3.6. The lower bound in (3.2) follows from Theorem 3.4 combined with Lemma 2.5, and Lemma 3.1. The upper bounds are proven in the same manner, using the simple fact that for every $\alpha > 0$ there are two constants $c_\alpha, c'_\alpha > 0$, such that $k \geq c_\alpha d \log d$ implies that $2^{-\min\{\alpha,1\}k/d} \leq c'_\alpha k^{-\alpha}$. □

Summarizing this section, the classes of ridge functions with Lipschitz profiles of order α are essentially as compact as the class of univariate Lipschitz functions of order α. Consequently, when speaking in terms of metric entropy, these classes of functions must be much smaller than the class of multivariate Lipschitz functions of order α.

Remark 3.7. The reader who is interested in results on entropy numbers of other classes of ridge functions, is referred to the recent work [24]. There, classes of sums of ridge functions are studied such that each sum of ridge functions forms a multivariate polynomial of some maximal degree.

4 Sampling numbers of ridge functions

In light of Section 3, one is led to think that efficient sampling of ridge functions should be feasible. Moreover, their simple, two-component structure naturally suggests a two-step procedure: first, use a portion of the available function samples to identify either the profile or the direction; then, use the remaining samples to unveil the other component.

However, in Subsection 4.1, we learn that for ridge functions in the class $\mathcal{R}_d^{\alpha,p}$, sampling is almost as hard as sampling of general multivariate Lipschitz functions on the Euclidean unit ball. In particular, such two-step procedures as sketched above cannot work in an efficient manner. It needs additional assumptions on the ridge profiles or directions. We discuss this in Subsection 4.2.

4.1 Sampling of functions in $\mathcal{R}_d^{\alpha,p}$

As usual, throughout the section let $\alpha > 0$ be the Lipschitz smoothness of profiles, $s = \lfloor\!\lfloor \alpha \rfloor\!\rfloor$ the order up to which derivatives exist, and let $0 < p \leq 2$ indicate the p-norm such that ridge directions are contained in the closed ℓ_p^d-ball.

The algorithms we use to derive upper bounds are essentially the same as those which are known to be optimal for general multivariate Lipschitz functions. Albeit, the ridge structure allows a slightly improved analysis at least in case $p < 2$.

Proposition 4.1. *Let $\alpha > 0$ and $0 < p \leq 2$. For $n \geq \binom{d+s}{s}$ sampling points the n-th sampling number is bounded from above by*

$$g_{n,d}^{lin}(\mathcal{R}_d^{\alpha,p}, L_\infty) \leq e_{k-\Delta}(\bar{B}_2^d, \ell_{p'}^d)^\alpha \,,$$

where $k = \lfloor \log n \rfloor + 2$, $\Delta = 1 + \lceil \log \binom{d+s}{s} \rceil$, and p' is the dual index of p.

Proof. Case $\alpha \leq 1$: In this case, $s = 0$ and $\Delta = 1$. We choose sampling points $x_1, \dots, x_{2^{k-2}}$ such that they form an ε-covering of $\bar{B}_2^d$ in $\ell_{p'}^d$. Given this covering, we construct (measurable) sets $U_1, \dots, U_{2^{k-2}}$ such that $U_i \subseteq x_i + \varepsilon \bar{B}_{p'}^d$ for $i = 1, \dots, 2^{k-2}$ and

$$\bigcup_{i=1}^{2^{k-2}} \left(x_i + \varepsilon \bar{B}_{p'}^d \right) = \bigcup_{i=1}^{2^{k-2}} U_i, \qquad U_i \cap U_j = \emptyset \text{ for } i \neq j.$$

Now we use piecewise constant interpolation: we approximate $f = g(a\cdot) \in \mathcal{R}_d^{\alpha,p}$ by $Sf := \sum_{i=1}^{2^{k-2}} f(x_i) \mathbb{1}_{U_i}$. Then,

$$\begin{aligned} \|f - Sf\|_\infty &= \sup_{i=1,\dots,2^{k-2}} \sup_{x \in U_i} \big| f(x) - f(x_i) \big| \\ &\leq \sup_{i=1,\dots,2^{k-2}} \sup_{x \in U_i} \|g\|_{\mathrm{Lip}_\alpha} \|a\|_p^\alpha \|x - x_i\|_{p'}^\alpha \leq \varepsilon^\alpha. \end{aligned}$$

The smallest ε is determined by the $(k-1)$-st entropy number $e_{k-1}(\bar{B}_2^d, \ell_{p'}^d)$. Consequently,

$$g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{\alpha,p}, L_\infty) \leq g_{2^{k-2},d}^{\mathrm{lin}}(\mathcal{R}_d^{\alpha,p}, L_\infty) \leq e_{k-1}(\bar{B}_2^d, \ell_{p'}^d)^\alpha.$$

Case $\alpha > 1$: Let the sampling points $x_1, \dots, x_{2^{k-\Delta-1}}$ and the sets $U_1, \dots, U_{2^{k-\Delta-1}}$ be as above. However, instead of piecewise constant interpolation we apply on each of the sets $U_i \subseteq x_i + \varepsilon \bar{B}_{p'}^d$ a Taylor formula of order s around the center x_i.

That is, to approximate a given $f = g(a\cdot) \in \mathcal{R}_d^{\alpha,p}$ we set $Sf := \sum_{i=1}^{2^{k-\Delta-1}} T_{x_i,s} f \mathbb{1}_{U_i}$. Then, by Lemma 2.3 (ii), we have

$$\|f - Sf\|_\infty = \sup_{i=1,\dots,2^{k-\Delta-1}} \sup_{x \in U_i} \big| f(x) - T_{x_i,s} f(x) \big| \leq \frac{1}{s!} \|x - x_i\|_{p'}^\alpha \leq \varepsilon^\alpha.$$

It takes $2^{k-\Delta-1} \binom{d+s}{s} \leq n$ function values to approximate all the $T_{x_i,s}$ above up to arbitrary precision by finite-order differences, cf. [38].

The smallest ε is now determined by the $(k-\Delta)$-th entropy number $e_{k-\Delta}(\bar{B}_2^d, \ell_{p'}^d)$. We conclude

$$g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{\alpha,p}, L_\infty) \leq g_{2^{k-\Delta-1},d}^{\mathrm{lin}}(\mathcal{R}_d^{\alpha,p}, L_\infty) \leq e_{k-\Delta}(\bar{B}_2^d, \ell_{p'}^d)^\alpha.$$

□

We turn to an analysis of lower bounds for the classes $\mathcal{R}_d^{\alpha,p}$. Our strategy is to find "bad" directions which map, for a given budget $n \in \mathbb{N}$, all possible choices of n sampling points to a small range of $[-1, 1]$. There, we let the "fooling" profiles be zero; outside of that range, we let the profiles climb as steep as possible. Proposition 4.2 below states the lower bound that results from this strategy, provided that the "bad" directions are given by some $\mathcal{M} \subseteq \bar{B}_p^d \setminus \{0\}$. We discuss appropriate choices of $\mathcal{M}$ later. In the sequel, we use the mapping $\Psi : \mathbb{R}^d \setminus \{0\} \to \mathbb{S}_2^{d-1}$ defined by $x \mapsto x/\|x\|_2$.

Proposition 4.2. *Let $\alpha > 0$, $0 < p \le 2$, and $\mathcal{M} \subseteq \bar{B}_p^d \setminus \{0\}$. Then, for all natural numbers k and n with $n \le 2^{k-1}$, we have*

$$g_{n,d}^{ada}(\mathcal{R}_d^{\alpha,p}, L_\infty) \ge c_\alpha \inf_{a\in\mathcal{M}} \|a\|_2^\alpha \cdot e_k(\Psi(\mathcal{M}), \ell_2^d)^{2\alpha}\,.$$

The constant c_α depends only on α.

Proof. Let us first describe the "fooling" profiles in detail. For each $a \in \mathcal{M}$ and $\varepsilon < 1$, we define a function

$$g_{a,\varepsilon}(t) = \vartheta_\alpha \big[(t - \|a\|_2(1-\varepsilon^2/2))_+\big]^\alpha \tag{4.1}$$

on the interval $[-1,1]$. The factor ϑ_α assures that $\|g_{a,\varepsilon}\|_{\mathrm{Lip}_\alpha[-1,1]} = 1$. Put $f_{a,\varepsilon}(x) = g_{a,\varepsilon}(a\cdot x)$. By construction, we have that $f_{a,\varepsilon} \in \mathcal{R}_d^{\alpha,p}$. Moreover, whenever $x \in \bar{B}_2^d$ and $a \in \mathcal{M}$ is such that

$$\varepsilon^2 < \big\|x - \Psi(a)\big\|_2^2 \tag{4.2}$$

then $\varepsilon^2 \le 2 - 2(x\cdot\Psi(a))$ and hence

$$x\cdot a = \|a\|_2(x\cdot\Psi(a)) < \|a\|_2(1-\varepsilon^2/2)\,.$$

Therefore, (4.2) implies $f_{a,\varepsilon}(x) = 0$.

Now, let $n \le 2^{k-1}$ and $S \in \mathcal{S}_n^{\mathrm{ada}}$ be an adaptive algorithm with a budget of n sampling points. Clearly, the first sampling point x_1 must have been fixed by S in advance. Then, let $x_2, \dots, x_n$ be the sampling points which S would choose when applied to the zero function. Furthermore, let $F(x_1,\dots,x_n) \subseteq \mathcal{R}_d^{\alpha,p}$ denote the set of functions that make S choose the very points $x_1,\dots,x_n$. Obviously, we have $f_{a,\varepsilon} \in F(x_1,\dots,x_n)$ if (4.2) holds for every x_i, $i = 1,...,n$. This is true for some $a \in \mathcal{M}$ if we choose $\varepsilon < e_k(\Psi(\mathcal{M}), \ell_2^d)$. For the respective function $f_{a,\varepsilon}$, we have in particular $N_S^{\mathrm{ada}}(f_{a,\varepsilon}) = 0$ and hence $S[f_{a,\varepsilon}] = S[-f_{a,\varepsilon}]$. Consequently,

$$\max\big\{\|f_{a,\varepsilon} - S[f_{a,\varepsilon}]\|_\infty, \|-f_{a,\varepsilon} - S[-f_{a,\varepsilon}]\|_\infty\big\} \ge \|f_{a,\varepsilon}\|_\infty = g_{a,\varepsilon}(\|a\|_2) = c_\alpha \|a\|_2^\alpha \varepsilon^{2\alpha}\,, \tag{4.3}$$

where $c_\alpha := 2^{-\alpha}\vartheta_\alpha$. Since ε has been chosen arbitrarily but less than $e_k(\Psi(\mathcal{M}), \ell_2^d)$, we are allowed to replace ε by $e_k(\Psi(\mathcal{M}), \ell_2^d)$ in (4.3) and get

$$\sup_{f\in\mathcal{R}_d^{\alpha,p}} \|f - S(f)\|_\infty \ge c_\alpha \inf_{a\in\mathcal{M}} \|a\|_2^\alpha \cdot e_k(\Psi(\mathcal{M}), \ell_2^d)^{2\alpha}.$$

Taking the infimum over all algorithms $S \in \mathcal{S}_n^{\mathrm{ada}}$ yields

$$g_{n,d}^{\mathrm{ada}}(\mathcal{R}_d^{\alpha,p}, L_\infty) \ge c_\alpha \inf_{a\in\mathcal{M}} \|a\|_2^\alpha\, e_k(\Psi(\mathcal{M}), \ell_2^d)^{2\alpha}.$$

□

Theorem 4.3. *Let $\alpha > 0$, $s = \lfloor\!\lfloor \alpha \rfloor\!\rfloor$, and $0 < p \le 2$. For the classes $\mathcal{R}_d^{\alpha,p}$, we have the following bounds:*
(i) *The n-th (linear) sampling number is bounded from above by*

$$g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{\alpha,p}, L_\infty) \le C_{p,\alpha} \begin{cases} 1 & : n \le 2d\binom{d+s}{s}, \\ \left[\frac{\log(1+d/\log n_1)}{\log n_1}\right]^{\alpha(1/\max\{1,p\}-1/2)} & : 2d\binom{d+s}{s} < n \le 2^{d+1}\binom{d+s}{s}, \\ n^{-\alpha/d}\, d^{-\alpha(1/\max\{p,1\}-1/2)} & : n > 2^{d+1}\binom{d+s}{s}, \end{cases}$$

where $n_1 = n/[2\binom{d+s}{s}]$, and the constant $C_{p,\alpha}$ depends only on α and p.

(ii) *The n-th (adaptive) sampling number is bounded from below by*

$$g_{n,d}^{\mathrm{ada}}(\mathcal{R}_d^{\alpha,p}, L_\infty) \ge c_{p,\alpha} \begin{cases} 1 & : n < d, \\ \left[\frac{\log_2\left(1+d/(2+\log_2 n)\right)}{2+\log_2 n}\right]^{\alpha(1/p-1/2)} & : d \le n < 2^{d-1}, \\ n^{-2\alpha/(d-1)}\, d^{-\alpha(1/p-1/2)} & : n \ge 2^{d-1}. \end{cases}$$

The constant $c_{p,\alpha}$ depends only on α and p.

Proof. (i) The upper bound is a direct consequence of Proposition 4.1 and Lemma 2.5. Note that, for k and Δ as in Proposition 4.1, it holds true that $k-\Delta-2 \le \log n_1 \le k-\Delta$. Note also that

$$\binom{d+s}{s}^{\alpha/d} \le (1+s)^{s\alpha/d} d^{s\alpha/d} \le ((1+s)e)^{s\alpha}$$

ensures that the constant $C_{p,\alpha}$ can be chosen independently of d and n.

(ii) Case $n < d$. Let $\mathcal{M} = \{\pm e_1, \dots, \pm e_d\}$ be the set of positive and negative canonical unit vectors. Clearly, we have $\sharp\mathcal{M} = 2d$ and every two distinct vectors in $\mathcal{M}$ have mutual ℓ_2^d-distance equal to or greater than $\sqrt{2}$. Let k be the smallest integer such that $n \le 2^{k-1}$; this implies $2^{k-1} < 2d$. Hence, whenever 2^{k-1} balls of radius ε cover the set $\mathcal{M}$, there is at least one ε-ball which contains two elements from $\mathcal{M}$. In consequence, we have $2\varepsilon \ge \sqrt{2}$ and hence $e_k(\mathcal{M}, \ell_2^d) \ge \sqrt{2}/2$. By Proposition 4.2 and the fact that $\mathcal{M} = \Psi(\mathcal{M})$, we obtain

$$g_{n,d}^{\mathrm{ada}}(\mathcal{R}_d^{\alpha,p}, L_\infty) \ge c_\alpha e_k(\mathcal{M}, \ell_2^d)^{2\alpha} \ge c_\alpha 2^{-\alpha}.$$

Case $d \le n < 2^{d-1}$. For $m \le d$, consider the subset of m-sparse vectors of the p-sphere,

$$\mathfrak{S}_{m,p}^{d-1} = \big\{x \in \mathbb{S}_p^{d-1} : \sharp\,\mathrm{supp}\,(x) = m\big\}.$$

Using the combinatorial construction of [16], cf. also [13], we know that there exist at least $(d/(4m))^{m/2}$ vectors in $\Psi(\mathfrak{S}_{m,p}^{d-1}) = \mathfrak{S}_{m,2}^{d-1}$ having mutual ℓ_2^d-distance greater than $1/\sqrt{2}$. Therefore, we have

$$\ell \le m/2\log(d/(4m)) \quad \Longrightarrow \quad e_\ell(\Psi(\mathfrak{S}_{m,p}^{d-1}), \ell_2^d) \ge \sqrt{2}/4. \tag{4.4}$$

Let k again be the smallest integer such that $n \le 2^{k-1}$. Hence, $k \le d$. Choose

$$m^* := \big\lfloor \min\{4k/\log(d/(4k)), k\} \big\rfloor \le k\,.$$

Because of $k > \log d$, we have $\min\{\log d, 4\} \le m^* \le d$. Put $\mathcal{M} = \mathfrak{S}^{d-1}_{m^*,p}$. If $k \le d/64$, then $\log(d/(4k)) \ge 4$ and $k \le m^* \log(d/(4k))/2 \le m^* \log(d/(4m^*))/2$. Hence, by (4.4), one has $e_k(\Psi(\mathfrak{S}^{d-1}_{m^*,p}), \ell_2^d) \ge \sqrt{2}/4$. Consequently, by Proposition 4.2, it follows that

$$\begin{aligned} g^{\mathrm{ada}}_{n,d}(\mathcal{R}^{\alpha,d}_d, L_\infty) &\ge c_\alpha (m^*)^{\alpha(1/2-1/p)} e_k(\Psi(\mathfrak{S}^{d-1}_{m^*,p}), \ell_2^d)^{2\alpha} \\ &\ge c_\alpha 8^{-\alpha} 4^{-\alpha(1/p-1/2)} \Big[\frac{\log(d/(4k))}{k}\Big]^{\alpha(1/p-1/2)} \\ &\ge c_\alpha 8^{-\alpha} 8^{-\alpha(1/p-1/2)} \left(\frac{\log(1+d/k)}{k}\right)^{\alpha(1/p-1/2)} \\ &\ge c_{p,\alpha} \left(\frac{\log(1+d/k)}{k}\right)^{\alpha(1/p-1/2)}\,. \end{aligned}$$

On the other hand, if $d/64 < k \le d$, then $m^* = k$. By $\mathbb{S}_2^{k-1} \subset \Psi(\mathfrak{S}^{d-1}_{m^*,p}) \subset \mathbb{S}_2^{d-1}$ and Lemma 2.6, we have $e_k(\Psi(\mathfrak{S}^{d-1}_{m^*,p}), \ell_2^d) \asymp 1$. Proposition 4.2, together with $\log(1+d/k) < 8$ for $k > d/64$, implies

$$\begin{aligned} g^{\mathrm{ada}}_{n,d}(\mathcal{R}^{\alpha,p}_d, L_\infty) \ge c'_\alpha k^{-\alpha(1/p-1/2)} &\ge c'_\alpha 8^{-\alpha(1/p-1/2)} \left(\frac{\log(1+d/k)}{k}\right)^{\alpha(1/p-1/2)} \\ &= c'_{p,\alpha} \left(\frac{\log(1+d/k)}{k}\right)^{\alpha(1/p-1/2)}\,. \end{aligned}$$

Case $n \ge 2^{d-1}$. Again, k is chosen such that $2^{k-2} < n \le 2^{k-1}$, which implies $k \ge d$. In this case, we choose $\mathcal{M} = \mathbb{S}_p^{d-1}$. By Lemma 2.6 and Proposition 4.2, we obtain

$$\begin{aligned} g^{\mathrm{ada}}_{n,d}(\mathcal{R}^{\alpha,p}_d, L_\infty) &\ge c_\alpha\, d^{-\alpha(1/p-1/2)} e_k(\mathbb{S}_2^{d-1}, \ell_2^d)^{2\alpha} \\ &\ge c_\alpha d^{-\alpha(1/p-1/2)}\, (4n)^{-2\alpha/(d-1)} \\ &\ge c_\alpha 4^{-2\alpha} d^{-\alpha(1/p-1/2)} n^{-2\alpha/(d-1)}\,. \end{aligned}$$

This completes the proof. □

Remark 4.4. Consider the situation $p = 2$. For sampling numbers with $n \le 2^{d-1}$, we have

$$g^{\mathrm{ada}}_{n,d}(\mathcal{R}^{\alpha,2}_d, L_\infty) \asymp 1.$$

For sampling numbers with $n \ge 2^{d+1}\binom{d+s}{s}$, we have

$$n^{-2\alpha/(d-1)} \lesssim g^{\mathrm{ada}}_{n,d}(\mathcal{R}^{\alpha,2}_d, L_\infty) \lesssim n^{-\alpha/d}. \tag{4.5}$$

The upper estimate on sampling numbers is exactly the same as for multivariate Lipschitz functions from $\mathrm{Lip}_\alpha(\Omega)$. Although there is a gap between lower and upper bound in

(4.5), the factor $1/(d-1)$ in the exponent of the lower bound allows us to conclude that sampling of ridge functions in $\mathcal{R}_d^{\alpha,2}$ is nearly as hard as sampling of general Lipschitz functions from $\mathrm{Lip}_\alpha(\Omega)$. Hence, we have the opposite situation to Section 3, where ridge functions in $\mathcal{R}_d^{\alpha,2}$ behave similar to univariate Lipschitz functions.

Remark 4.5. Let us consider the modified ridge function classes $\tilde{\mathcal{R}}_d^{\alpha,p}$ and $\bar{\mathcal{R}}_d^{\alpha,p}$ defined by

$$\tilde{\mathcal{R}}_d^{\alpha,p} := \big\{ f : [0,1]^d \to \mathbb{R} \,:\, f(x) = g(a \cdot x),\ \|g\|_{\mathrm{Lip}_\alpha[0,1]} \le 1,\ \|a\|_p \le 1,\ a \ge 0 \big\}, \quad (4.6)$$

for $0 < p \le 1$, and

$$\bar{\mathcal{R}}_d^{\alpha,p} := \big\{ f : \bar{B}_2^d \cap [0,1]^d \to \mathbb{R} \,:\, f(x) = g(a \cdot x),\ \|g\|_{\mathrm{Lip}_\alpha[0,1]} \le 1,\ \|a\|_p \le 1,\ a \ge 0 \big\}. \quad (4.7)$$

for $0 < p \le 2$. Here, $a \ge 0$ means, that all coordinates of a are non-negative.

(i) In the recent paper [6] it has been shown that there is an adaptive algorithm which attains a decay rate of $n^{-\alpha}$ for the worst-case L_∞-approximation error with respect to the class $\tilde{\mathcal{R}}_d^{\alpha,1}$, provided that $n \ge d$. In terms of adaptive sampling numbers (such that the feasible algorithms are adjusted to the domain $[0,1]^d$), this reads as

$$g_{n,d}^{\mathrm{ada}}(\tilde{\mathcal{R}}_d^{\alpha,1}, L_\infty) \le C_\alpha n^{-\alpha}, \quad n \ge d. \quad (4.8)$$

At the same time, a careful inspection of the proofs of Propositions 4.1, 4.2, and Theorem 4.3 shows that the results can be carried over to the classes $\bar{\mathcal{R}}_d^{\alpha,p}$ for all $0 < p \le 2$. In particular, for $0 < p \le 1$, we have the lower bound

$$g_{n,d}^{\mathrm{ada}}(\bar{\mathcal{R}}_d^{\alpha,p}, L_\infty) \ge c_{p,\alpha} n^{-2\alpha/(d-1)} d^{\alpha(1/2-1/p)}, \quad n \in \mathbb{N}\,. \quad (4.9)$$

The estimates (4.8) and (4.9) look conflicting at first glance. We encounter the rather surprising phenomenon that enlarging the domain of the class of functions under consideration leads to better approximation rates. To understand this, let us briefly sketch the adaptive algorithm of [6]. For $f = g(a\cdot) \in \tilde{\mathcal{R}}_d^{\alpha,p}$ not the zero function, the idea is to first sample along the diagonal of the first orthant, that is, at points $x = t(1,\dots,1)$ with $t \in [0,1]$. Importantly, it is guaranteed that we can take samples from the whole relevant range $[0, \|a\|_1]$ of the profile g of f. This in turn assures that, by sampling adaptively along the diagonal, we find a small range in $[0, \|a\|_1]$ where the absolute value of g' is strictly larger than 0. Then, the ridge direction a can be recovered in a similar way as we do in Subsection 4.2.

On the other hand, for the classes $\bar{\mathcal{R}}_d^{\alpha,p}$, this adaptive algorithm will not work. Assume we sample again along the (rescaled) diagonal. This time, we can be sure that we are able to reach every point in the intervall $[0, \|a\|_1/\sqrt{d}]$. But this interval is in most cases strictly included in the relevant interval $[0, \|a\|_2]$ for g. Hence, it is not guaranteed anymore that we sample the whole relevant range of g and find an interval on which g' is not zero.

(ii) Admittedly, the domain $\Omega = [0,1]^d \cap B_2^d$ in (4.7) is a somewhat artificial choice in case of $p \le 1$, whereas the cube $\Omega = [0,1]^d$ seems natural. Conversely, the definition

in (4.6) is not reasonable in case $p > 1$, since then $a \cdot x$ might exceed the domain interval for g. However, $\Omega = [0,1]^d \cap B_2^d$ is the natural choice for $p = 2$ in (4.7). In this situation, we suffer from the curse of dimensionality for adaptive algorithms using standard information, see Remark 4.4 and Theorem 5.1,(1) below. This shows that the condition $p \leq 1$ is essential in the setting of [6] and that (4.8) can not be true for the class $\bar{\mathcal{R}}_d^{\alpha,2}$.

Remark 4.6. We are not aware of any results on the approximation of ridge functions when arbitrary *bounded, linear functionals* are admitted in the information map, see Section 2.2. It seems to be an open problem whether or not such *linear information* would lead to substantially better bounds for the worst-case error.

4.2 Recovery of ridge directions

At the beginning of Section 4,we have sketched two-step procedures for the recovery of ridge functions. In this section, we discuss under which conditions these two-step procedures are feasible within our setting. The adaptive algorithm of [6], which we have already discussed in Remark 4.5, first approximates the profile g. Unfortunately, we could already argue that this algorithm cannot work in our setting. There is an opposite approach in Fornasier et al. [12], which first tries to recover the ridge direction and conforms to our setting. Following the ideas of [2], the authors developed an efficient scheme using Taylor's formula to approximate ridge functions with C^s profile obeying certain integral condition on the modulus of its derivative. This condition was satisfied for example if $\left|g'(0)\right| \geq \kappa > 0$. In their approach, the smoothness parameter s had to be at least 2. Using a slightly different analysis, this scheme turns out to work for Lipschitz profiles of order $\alpha > 1$.

Before we turn to the analysis, let us sketch the Taylor-based scheme in more detail. As transposes of matrices and vectors appear frequently, for reasons of convenience, we write $a \cdot x = a^T x$ for the remainder of this subsection. Now, Taylor's formula in direction e_i yields

$$\begin{aligned} f(he_i) &= f(0) + h\nabla f(\xi_h^{(i)} e_i)^T e_i \\ &= g(0) + hg'(\xi_h^{(i)} a_i) a_i \,. \end{aligned}$$

Hence, we can expose the vector a, distorted by a diagonal matrix with components

$$\xi_h = (g'(\xi_h^{(1)} a_1), \ldots, g'(\xi_h^{(d)} a_d))$$

on the diagonal. In total, we have to spend only $d+1$ function evaluations for that. Moreover, each of ξ_h's components can be pushed arbitrarily close to $g'(0)$. This gives an estimate $\hat{a}$ of $a/\|a\|_2$, along which we can now conduct classical univariate approximation. Effectively, one samples a distorted version of g given by

$$\tilde{g} : [-1,1] \to \mathbb{R},\ t \mapsto f(t\hat{a}) = g(ta^T\hat{a}).$$

The approximation $\hat{g}$ obtained in this way, together with $\hat{a}$, forms the sampling approximation to f,

$$\hat{f}(x) = \hat{g}(\hat{a}^T x).$$

Observe that $\tilde{g}(\hat{a}^T x) = g(a^T \hat{a}\hat{a}^T x)$, so it is crucial that $\hat{a}\hat{a}^T$ spans a subspace which is close to the one-dimensional subspace spanned by aa^T, in the sense that

$$\left\| a^T (I_d - \hat{a}\hat{a}^T) \right\|_2$$

has to be small. Importantly, this gives the freedom to approximate a only up to a sign. Finally, let us note that if the factor $g'(0)$ can become arbitrary small, the information we get through Taylor's scheme about a becomes also arbitrarily bad. Hence, for this approach to work, it is necessary to require $\left|g'(0)\right| \geq \kappa$.

Lemma 4.7. *Let $0 < \beta \leq 1$, $0 < \kappa \leq 1$, and $\varepsilon > 0$. Further let $\delta = \frac{\varepsilon \cdot \kappa}{2+\varepsilon}$ and $h = (\delta/2)^{1/\beta}$. For any $g \in \mathrm{Lip}^{\kappa}_{1+\beta}([-1,1])$ and $a \in \bar{B}^d_2$ with $a \neq 0$ let $f = g(a\cdot)$. Put*

$$\tilde{a}_i = \frac{f(he_i) - f(0)}{h}, \quad i = 1, \ldots, d \tag{4.10}$$

and $\hat{a} = \tilde{a}/\|\tilde{a}\|_2$. Then

$$\left\| \mathrm{sign}\,(g'(0))\hat{a} - a/\|a\|_2 \right\|_2 \leq \varepsilon.$$

Proof. By the mean value theorem of calculus there exist $\xi_h^{(i)} \in [0, h]$ such that

$$\tilde{a}_i = g'(\xi_h^{(i)} a_i) a_i.$$

By Hölder continuity we get

$$|g'(\xi_h^{(i)} a_i) - g'(0)| < 2|g'|_\beta |a_i|^\beta |h|^\beta \leq \delta$$

for all $i = 1, \ldots, d$. Let us observe that $\delta < \kappa$ and, therefore, $\tilde{a} \neq 0$ and $\hat{a}$ is well defined. Put $\xi = (g'(\xi_h^{(i)} a_i))_{i=1}^d$. Then we can write $\tilde{a} = \mathrm{diag}(\xi) a$. For the norm of $\tilde{a}$ we get

$$\begin{aligned}
\|\tilde{a}\|_2 &\leq \|\mathrm{diag}(\xi)a - g'(0)a\|_2 + |g'(0)|\|a\|_2 \\
&\leq \max_{i=1,\ldots,d} |g'(\xi_h^{(i)} a_i) - g'(0)| \|a\|_2 + |g'(0)|\|a\|_2 \\
&\leq (\delta + |g'(0)|)\|a\|_2.
\end{aligned}$$

Analogously, by the inverse triangle inequality $\|\tilde{a}\|_2 \geq (|g'(0)| - \delta)\|a\|_2$. In particular,

$$\left| \|\tilde{a}\|_2/\|a\|_2 - |g'(0)| \right| \leq \delta.$$

Now, writing $\gamma = \mathrm{sign}\,(g'(0))$, we observe

$$\begin{aligned}
\left\| \gamma\hat{a} - a/\|a\|_2 \right\|_2 &\leq \left\| \gamma\hat{a} - |g'(0)|a/\|\tilde{a}\|_2 \right\|_2 + \left\| |g'(0)|a/\|\tilde{a}\|_2 - a/\|a\|_2 \right\|_2 \\
&= \|\tilde{a}\|_2^{-1} \left(\|(\mathrm{diag}(\xi) - g'(0) I_d)\ a\|_2 + \left| |g'(0)| - \|\tilde{a}\|_2/\|a\|_2 \right| \|a\|_2 \right) \\
&\leq 2\delta \|a\|_2/\|\tilde{a}\|_2 \leq 2\delta/(|g'(0)| - \delta) \leq 2\delta/(\kappa - \delta) = \varepsilon.
\end{aligned}$$

□

Having recovered the ridge direction, we manage to unveil the one-dimensional structure from the high-dimensional ambient space. In other words, recovery of the ridge direction is a *dimensionality reduction* step. What remains is the problem of sampling the profile, which can be done using standard techniques. In combination, this leads to the following result:

Theorem 4.8. *Let $\alpha > 1$ and $0 < \kappa \leq 1$.*

(i) Let $n \leq d-1$. Then $g_{n,d}(\mathcal{R}_d^{\alpha,2,\kappa}, L_\infty) = g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{\alpha,2,\kappa}, L_\infty) = 1$.

(ii) Let $n \geq d+1$. Then

$$c_\alpha \cdot n^{-\alpha} \leq g_{n,d}(\mathcal{R}_d^{\alpha,2,\kappa}, L_\infty) \leq g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{\alpha,2,\kappa}, L_\infty) \leq C_\alpha (n-d)^{-\alpha}$$

with constants c_α and C_α, which depend on α only.

Proof. (i) It is enough to show that $g_{n,d}(\mathcal{R}_d^{\alpha,2,\kappa}, L_\infty) \geq 1$ for $n \leq d-1$. Let us assume that a given (adaptive) approximation method samples at $x_1, \ldots, x_n$ and let us denote by L their linear span. Then $\dim L \leq n < d$ and we may find $a \in \mathbb{R}^d$ with $\|a\|_2 = 1$ orthogonal to all $x_1, \ldots, x_n$. Finally, if we define $g(t) = t$, we obtain

$$\begin{aligned} 1 = \|g(a^T\cdot)\|_\infty &\leq \frac{1}{2} \cdot \Big\{ \|g(a^T\cdot) - S_n(g(a^T\cdot))\|_\infty + \| - g(a^T\cdot) - S_n(-g(a^T\cdot))\|_\infty \Big\} \\ &\leq g_{n,d}(\mathcal{R}_d^{\alpha,2,\kappa}, L_\infty). \end{aligned}$$

(ii) Fix some $0 < \varepsilon < 1$. Let $\hat{a}$ denote the reconstruction of a obtained by Lemma 4.7, which uses $d+1$ sampling points of f. We estimate g by sampling the distorted version

$$\tilde{g} : [-1,1] \to \mathbb{R},\ t \mapsto f(t\hat{a}) = g(ta^T\hat{a}).$$

Re-using the value $g(0)$ which we have already employed for the recovery of a, we spend $k = n - d \geq 1$ sampling points and obtain a function $\hat{g}$ with $\|\hat{g} - \tilde{g}\|_\infty \leq \varepsilon := C'_\alpha k^{-\alpha} \|\tilde{g}\|_{\mathrm{Lip}_\alpha}$.

Now put $\hat{f}(x) = \hat{g}(\hat{a}^T x)$ as our approximation to f. To control the total approximation error, observe that

$$|\hat{f}(x) - f(x)| \leq |\hat{g}(\hat{a}^T x) - \tilde{g}(\hat{a}^T x)| + |\tilde{g}(\hat{a}^T x) - g(a^T x)| =: E_1 + E_2.$$

For the first error term E_1, we immediately get

$$E_1 \leq \|\hat{g} - \tilde{g}\|_\infty \leq \varepsilon = C'_\alpha \|\tilde{g}\|_{\mathrm{Lip}_\alpha} k^{-\alpha} \leq C'_\alpha k^{-\alpha}$$

as $\|\tilde{g}\|_{\mathrm{Lip}_\alpha} \leq \|a\|_2 \, \|g\|_{\mathrm{Lip}_\alpha} \leq 1$.

For the second error term, note that

$$\begin{aligned} E_2 = |g(a^T\hat{a}\hat{a}^T x) - g(a^T x)| &\leq \|g\|_{\mathrm{Lip}_\alpha} \, \|a^T(I_d - \hat{a}\hat{a}^T)\|_2 \, \|x\|_2 \\ &\leq \|g\|_{\mathrm{Lip}_\alpha} \, \|x\|_2 \, \|a\|_2 \, \big\|a^T/\|a\|_2 \ (I_d - \hat{a}\hat{a}^T)\big\|_2. \end{aligned}$$

We do not know the exact value of the subspace stability term $\|a^T/\|a\|_2\ (I_d - \hat{a}\hat{a}^T)\|_2$. But because $\hat{a}\hat{a}^T$ is the identity in direction of $\hat{a}$, we have the estimate

$$\begin{aligned}\left\|a^T/\|a\|_2\ (I_d - \hat{a}\hat{a}^T)\right\|_2 &= \left\|\left(a/\|a\|_2 - \operatorname{sign}(g'(0))\hat{a}\right)^T\ (I_d - \hat{a}\hat{a}^T)\right\|_2 \\ &\leq \|I_d - \hat{a}\hat{a}^T\|_{2\to 2}\ \left\|a/\|a\|_2 - \operatorname{sign}(g'(0))\hat{a}\right\|_2 \\ &\leq \varepsilon.\end{aligned}$$

For the last inequality, we have used Lemma 4.7 and the fact that $\|I_d - \hat{a}\hat{a}^T\|_{2\to 2} \leq 1$. As a consequence,

$$E_2 \leq \|x\|_2\, \|a\|_2\, \|g\|_{\mathrm{Lip}_\alpha}\, \varepsilon \leq \varepsilon.$$

Putting everything together, we conclude

$$\|\hat{f} - f\|_\infty \leq 2\varepsilon \leq 2C'_\alpha k^{-\alpha}.$$

Let us turn to the lower bound. Assume we are given a feasible approximation method S_n that samples at points $\{x_1, \ldots, x_n\} \subset \Omega$. Let $\psi_{k,b}$ be as in the proof of Proposition 3.2. There is an interval $I' \subset I = [\pi/4 - 1/5, \pi/4 + 1/5]$ of length $|I'| = 1/(5n)$ such that I' does not contain any of the first coordinates of $x_1, \ldots, x_n$; in other words, it is disjoint with $\{x_1 \cdot e_1, \ldots, x_n \cdot e_1\}$, where $e_1 = (1, 0, \ldots, 0)$ is the first canonical unit vector. Furthermore, let b be the center of I', put $\psi = \psi_{2n,b}$, and $a = e_1$. Finally, with γ as in (3.1), we write

$$\begin{aligned}f(x) &= \sin(x \cdot e_1), \\ f_+(x) &= \sin(x \cdot e_1) + (1-\gamma)\psi(x \cdot e_1), \\ f_-(x) &= \sin(x \cdot e_1) - (1-\gamma)\psi(x \cdot e_1).\end{aligned}$$

As $S_n(f) = S_n(f_+) = S_n(f_-)$ and all the three functions are in $\mathcal{R}_d^{\alpha,2,\kappa}$, we may use the triangle inequality

$$\begin{aligned}\|(1-\gamma)\psi\|_\infty &= \|(1-\gamma)\psi(e_1\cdot)\|_\infty \\ &\leq \frac{1}{2}\Big\{\|(1-\gamma)\psi(e_1\cdot) + f - S_n(f)\|_\infty + \|(1-\gamma)\psi(e_1\cdot) - [f - S_n(f)]\|_\infty\Big\} \\ &= \frac{1}{2}\Big\{\|f_+ - S_n(f_+)\|_\infty + \|f_- - S_n(f_-)\|_\infty\Big\},\end{aligned}$$

to conclude that

$$g_{n,d}(\mathcal{R}_d^{\alpha,2,\kappa}, L_\infty) \gtrsim n^{-\alpha},$$

with a constant depending only on α. □

Remark 4.9. Let us briefly comment why we assume $g'(0) \geq \kappa$ and not $g'(t_0) \geq \kappa$ for some arbitrary, but known $t_0 \in [-1, 1]$. Let $x \in \bar{B}_2^d$ be some arbitrary sampling point (taken e.g. uniformly at random in $\bar{B}_2^d$). Since the only a-priori information is $\|a\|_2 \leq 1$, by the concentration of measure phenomenon the inner product $a \cdot x$ will most likely be close to zero when d is large. Hence, to exploit $g'(t_0) \geq \kappa$ for some $t_0 \neq 0$ we effectively have to know the vector a beforehand.

Remark 4.10. Once we have control on the derivative in the origin, recovery of the ridge direction and approximation of the ridge profile can be addressed independently. Formula (4.10) is based on the simple observation that

$$\frac{\partial f}{\partial x_i}(0) = g'(0)a_i = g'(0)\langle a, e_i\rangle$$

might be well approximated by first order differences. Furthermore, this holds also for every other direction $\varphi \in \mathbb{S}_2^{d-1}$, i.e.,

$$\frac{\partial f}{\partial \varphi}(0) = g'(0)\langle a, \varphi\rangle$$

can be approximated by differences

$$\frac{f(h\varphi) - f(0)}{h}.$$

Taking the directions $\varphi_1, \dots, \varphi_{m_\Phi}$ at random (and appropriately normalized), one can approximate the scalar products $\{\langle a, \varphi_i\rangle\}_{i=1}^{m_\Phi}$. Finally, if one assumes that $a \in \bar{B}_p^d$ for $0 < p \le 1$, one can recover a good approximation to a by the *sparse recovery* methods of the modern area of *compressed sensing*. This approach has been investigated in [12].

Although the algorithms of compressed sensing involve random matrices, once a random matrix with good sensing properties (typically with small constants of their Restricted Isometry Property) is fixed, the algorithms become fully deterministic. This allows to transfer the estimates of [12] into an upper bound for the deterministic worst-case error $g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{2,p,\kappa}, L_\infty)$.

Let $0 < p \le 1$ and

$$c\kappa^{-\frac{2p}{2-p}} \log d \le m_\Phi \le Cd,$$

for two universal positive constants c, C. It follows from the results of [12] that drawing the directions $\varphi_1, \dots, \varphi_{m_\Phi}$ *once* yields with high probability a deterministic algorithm that needs $n > m_\Phi$ sampling points to recover any function $f \in \mathcal{R}_d^{2,p,\kappa}$ up to precision

$$\Big[\frac{m_\Phi}{\log(d/m_\Phi)}\Big]^{1/2-1/p} + (n - m_\Phi)^{-2}.$$

If $1/p \le 5/2$ and $c'\kappa^{-\frac{2p}{2-p}} \log d \le n \le C'd$, this implies that

$$g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{2,p,\kappa}, L_\infty) \lesssim \Big[\frac{n}{\log(d/n)}\Big]^{1/2-1/p}$$

and the same estimate holds if $1/p > 5/2$ and $c'\kappa^{-\frac{2p}{2-p}} \log d \le n \le c''(\log d)^{\frac{1/p-1/2}{1/p-5/2}}$. Finally, if $c''(\log d)^{\frac{1/p-1/2}{1/p-5/2}} \le n \le C'd$, we obtain

$$g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{2,p,\kappa}, L_\infty) \lesssim n^{-2}.$$

5 Tractability results

The field of information-based complexity [26] provides a family of notions of so-called *tractability*, which allow to classify ridge function sampling by degrees of difficulty. In regard of these notions, the studied ridge function classes are surprisingly rich. We run across almost the whole hierarchy of degrees of tractability if we vary the problem parameters α and p, or add the constraint on the profiles' first derivative in the origin.

Let us briefly introduce the standard notions of tractability. We say that a problem is *polynomially tractable* if its information complexity $n(\varepsilon, d)$ is bounded polynomially in ε^{-1} and d, i.e. there exist numbers $C, r, q > 0$ such that

$$n(\varepsilon, d) \leq C\, \varepsilon^{-r}\, d^q \text{ for all } 0 < \varepsilon < 1 \text{ and all } d \in \mathbb{N}.$$

A problem is called *quasi-polynomially tractable* if there exist two constants $C, t > 0$ such that

$$n(\varepsilon, d) \leq C \exp(t(1 + \ln(1/\varepsilon))(1 + \ln d))\,. \tag{5.1}$$

It is called *weakly tractable* if

$$\lim_{1/\varepsilon + d \to \infty} \frac{\log n(\varepsilon, d)}{1/\varepsilon + d} = 0\,, \tag{5.2}$$

i.e., the information complexity $n(\varepsilon, d)$ neither depends exponentially on $1/\varepsilon$ nor on d.

We say that a problem is *intractable*, if (5.2) does not hold. If for some fixed $0 < \varepsilon < 1$ the number $n(\varepsilon, d)$ is an exponential function in d then a problem is, of course, intractable. In that case, we say that the problem suffers from *the curse of dimensionality*. To make it precise, we face the curse if there exist positive numbers c, ε_0, γ such that

$$n(\varepsilon, d) \geq c(1 + \gamma)^d\,, \quad \text{for all } 0 < \varepsilon \leq \varepsilon_0 \text{ and infinitely many } d \in \mathbb{N}\,.$$

In the language of IBC, Theorems 4.3 and 4.8 now read as follows:

Theorem 5.1. *Consider the problem of ridge function sampling as defined in Subsection 2.2. Assume that ridge profiles have at least Lipschitz smoothness $\alpha > 0$; further, assume that ridge directions are contained in the closed ℓ_p^d-unit ball for $p \in (0, 2]$. Then sampling of ridge functions in the class $\mathcal{R}_d^{\alpha,p}$*

(1) suffers from the curse of dimensionality if $p = 2$ and $\alpha < \infty$,

(2) never suffers from the curse of dimensionality if $p < 2$,

(3) is intractable if $p < 2$ and $\alpha \leq \frac{1}{1/p - 1/2}$,

(4) is weakly tractable if $p < 2$ and $\alpha > \frac{1}{1/\max\{1,p\} - 1/2}$,

(5) is quasi-polynomially tractable if $\alpha = \infty$,

(6) and with positive first derivatives of the profiles in the origin it is polynomially tractable, no matter what the values of α and p are.

To prove Theorem 5.1, we translate Theorem 4.3 into bounds on the information complexity

$$n(\varepsilon, d) = \min\{n \in \mathbb{N} : \ g_{n,d}(\mathcal{R}_d^{\alpha,p}, L_\infty) \le \varepsilon\}.$$

Lemma 5.2. *Let $p < 2$ and $\alpha > 0$. Set $\eta = \alpha(1/2 - 1/p') = \alpha(1/\max\{1,p\} - 1/2)$ and define*

$$\varepsilon_1^U := C_{p,\alpha}\left[\frac{\log(1 + d/\log d)}{\log d}\right]^\eta, \qquad \varepsilon_2^U := C_{p,\alpha}\left(\frac{1}{d}\right)^\eta.$$

Then there are positive constants C_0 and C_1 such that

$$\log n(\varepsilon, d) \le C_0 + C_1 \begin{cases} \log d & : \varepsilon_1^U \le \varepsilon \le 1, \\ \log d \cdot (1/\varepsilon)^{1/\eta} & : \varepsilon_2^U \le \varepsilon < \varepsilon_1^U, \\ \log(1/\varepsilon) \cdot (1/\varepsilon)^{1/\eta} & : \varepsilon < \varepsilon_2^U. \end{cases}$$

The constants depend only on p and α.

Lemma 5.3. *Let $p < 2$ and $\alpha > 0$. Put*

$$\varepsilon_1^L := c_{p,\alpha}\left[\frac{\log(1 + d/\log d)}{\log d}\right]^{\alpha(1/p-1/2)}, \quad \varepsilon_2^L := c_{p,\alpha}\left(\frac{1}{d}\right)^{\alpha(1/p-1/2)}, \quad \varepsilon_3^L := 4^{-\alpha}\varepsilon_2^L.$$

Then there are universal constants c_0, c_1, which depend only on p and α, such that

$$\log n(\varepsilon, d) \ge c_0 + c_1(1/\varepsilon)^{\alpha^{-1}(1/p-1/2)^{-1}}$$

for $\varepsilon_3^L \le \varepsilon < \varepsilon_1^L$.

Proof of Theorem 5.1. (1). For $n \le 2^{d-2}$, the lower bound in Theorem 4.3 gives

$$g_{n,d}(\mathcal{R}_d^{\alpha,2}, L_\infty) \ge c_{p,\alpha} =: \varepsilon_0.$$

Hence, $n(\varepsilon, d) \ge 2^{d-2}$ for all $\varepsilon < \varepsilon_0$ and we have the curse of dimensionality.

(2). Since $\alpha_1 > \alpha_2$ implies $\mathcal{R}_d^{\alpha_1,p} \subseteq \mathcal{R}_d^{\alpha_2,p}$, we can w.l.o.g. assume $\alpha \le 1$. We choose an arbitrary $\varepsilon_2^U \le \varepsilon \le 1$. By Lemma 5.2,

$$n(\varepsilon, d) \le 2^{C_0} d^{C_1 \varepsilon^{\alpha^{-1}(1/\max\{1,p\}-1/2)^{-1}}}.$$

By our assumption $\varepsilon \ge \varepsilon_2^U$, this is true for all natural $d > (C_{p,\alpha}/\varepsilon)^{\alpha^{-1}(1/\max\{1,p\}-1/2)^{-1}}$. Hence, the curse of dimensionality does not occur.

(3). Put $\gamma = \alpha(1/p - 1/2)$. Assume $d \to \infty$ and $\varepsilon_3^L \le \varepsilon < \varepsilon_2^L$. The latter implies

$$\left(\frac{c_{p,\alpha}}{4^\alpha}\right)^{1/\gamma} (1/\varepsilon)^{1/\gamma} \le d < c_{p,\alpha}^{1/\gamma}(1/\varepsilon)^{1/\gamma}.$$

This yields

$$\frac{\log_2 n(\varepsilon,d)}{d+1/\varepsilon} \geq \frac{c_0}{d+1/\varepsilon} + c_1 \frac{(1/\varepsilon)^{1/\gamma}}{c_{p,\alpha}^{1/\gamma}(1/\varepsilon)^{1/\gamma}+1/\varepsilon}.$$

Assuming that $\alpha \leq 1/(1/p-1/2)$, we have $\gamma \leq 1$ and thus $1/\varepsilon \leq (1/\varepsilon)^{1/\gamma}$. We conclude that

$$\frac{\log n(\varepsilon,d)}{d+1/\varepsilon} \geq \frac{c_1}{c_{p,\alpha}^{1/\gamma}+1} > 0.$$

Consequently, the problem is not weakly tractable; and thus intractable.

(4). Put $x = 1/\varepsilon + d$. By Lemma 5.2 and $1/\varepsilon \leq x$, $d \leq x$, we have

$$\log n(\varepsilon,d) \leq C_0 + C_1 \log(x) x^{\alpha^{-1}(1/\max\{1,p\}-1/2)^{-1}}.$$

Now, if $\alpha > \frac{1}{1/\max\{1,p\}-1/2}$, then $\lim_{x\to\infty}\ x^{-1}\log n(\varepsilon,d) = 0$.

(5). By embedding arguments it is enough to consider the class $\mathcal{R}_d^{\infty,2}$. We approximate the function $f \in \mathcal{R}_d^{\infty,2}$ via the Taylor polynomial $T_{s,0}f(x)$ in $x^0 = 0$. Lemma 2.3, (ii) gives for every $s \in \mathbb{N}$ the bound

$$\|f - T_{s,0}f\|_\infty \leq \frac{2}{s!}\,.$$

Let $\varepsilon > 0$ be given and let $s \in \mathbb{N}$ be the smallest integer such that $2/s! \leq \varepsilon$. Then $(s-1)! \leq 2/\varepsilon$ and therefore $[(s-1)/e]^{s-1} \leq (s-1)! \leq 2/\varepsilon$. This gives

$$(s-1)\ln((s-1)/e) \leq \ln(2/\varepsilon)\,. \tag{5.3}$$

We know from [38] that it requires $\binom{s+d}{s}$ function values to approximate the Taylor polynomial up to arbitrary (but fixed) precision. Hence, using (5.3), we see that there is a constant $t > 0$ such that

$$\ln n(\varepsilon,d) \leq s\ln(e(d+1)) \leq t(1+\ln(1/\varepsilon))(1+\ln d),$$

which is (5.1).

(6). From Theorem 4.8 we can immediately conclude $\varepsilon^{-1/\alpha} \lesssim n(\varepsilon,d) \lesssim \varepsilon^{-1/\alpha}$, where the constants behind "$\lesssim$" behave polynomially in d. Consequently, sampling of ridge functions in $\mathcal{R}_d^{\alpha,2,\kappa}$ is polynomially tractable. □

By Lemma 2.1, we know that $\mathcal{R}_d^{\infty,2}$ is a subclass of the unit ball in $C^\infty(\Omega)$. Besides, we know that approximation using function values is quasi-polynomially tractable in $\mathcal{R}_d^{\infty,2}$, see Theorem 5.1. What is the respective tractability level in $C^\infty(\Omega)$? Or, to put it differently: how much do we gain by imposing a ridge structure in $C^\infty(\Omega)$? The seminal paper [27] tells us that approximation in $C^\infty([0,1]^d)$ suffers from the curse of dimensionality when norming the space in the way as we did in (2.1). In contrast, we will show that sampling in $C^\infty(\Omega)$ is still weakly tractable. This is not too much of a surprise. Due to the concentration of measure phenomenon, the Euclidean unit ball's volume gets "very small" in high dimensions d; its measure scales like $(2\pi e/d)^{d/2}$. Anyhow, the result

suggests that one still benefits from supposing a ridge structure; infinitely differentiable ridge functions from $\mathcal{R}_d^{\infty,2}$ probably can be approximated easier than general functions from the unit ball of $C^\infty(\Omega)$. This is not guaranteed, however, because we do not show that one cannot get anything better than weak tractability for the sampling of functions in the unit ball of $C^\infty(\Omega)$.

Theorem 5.4. *The sampling problem for $C^\infty(\Omega)$, where the error is measured in $L_\infty(\Omega)$, is weakly tractable.*

Proof. Applying Lemma 2.3, (i) together with (2.3) we obtain for any $f \in C^\infty(\Omega)$ with $\|f\|_{C^\infty(\Omega)} \le 1$ and every $s \in \mathbb{N}$ the relation

$$\begin{aligned} |f(x) - T_{s,0}f(x)| &\le \frac{2}{(s-1)!}\|x\|_1^s \quad , \quad x \in \Omega\,, \\ &\le \frac{2d^{s/2}}{(s-1)!}\,. \end{aligned}$$

Let $s \in \mathbb{N}$ be the smallest integer such that $2d^{s/2}/(s-1)! \le \varepsilon$. This leads to

$$\frac{1}{\sqrt{d}}\Big(\frac{s-2}{e\sqrt{d}}\Big)^{s-2} \le \frac{(s-2)!}{d^{\frac{s-1}{2}}} \le \frac{2}{\varepsilon}$$

which implies

$$(s-2)\ln\Big(\frac{s-2}{e\sqrt{d}}\Big) \le \ln(2/\varepsilon) + \frac{1}{2}\ln(d)\,. \tag{5.4}$$

To approximate the Taylor polynomial $T_{s,0}f$ with arbitrary precision (uniformly in f) we need $\binom{d+s}{s}$ function values, see [38, p. 4]. Let us distinguish two cases. If $(s-2) \le e^2\sqrt{d}$ we obtain

$$\ln n(\varepsilon, d) \le s\ln(e(d+1)) \le (e^2\sqrt{d}+2)\cdot\ln(e(d+1))$$

and hence (5.2). If $s - 2 > e^2\sqrt{d}$ then (5.4) yields $s - 2 \le \ln(2/\varepsilon) + \ln(d)$. Thus,

$$\ln n(\varepsilon, d) \le s\ln(e(d+1)) \le (\ln(2/\varepsilon) + \ln(d) + 2)\cdot\ln(e(d+1))$$

and again (5.2) holds true. This establishes weak tractability. □

Remark 5.5. (i) The result in Theorem 5.4 is also a consequence of the arguments in [19, Sections 5.2, 5.3, and Section 6] by putting $L_{j,d} = d^{j/2}$.
(ii) Recently, Vybíral [38] showed that there is quasi-polynomial tractability if one replaces the classical norm $\sup_{\gamma\in\mathbb{N}_0^d}\|D^\gamma f\|_\infty$ by $\sup_{k\in\mathbb{N}_0}\sum_{|\gamma|=k}\|D^\gamma f\|_\infty/\gamma!$ in $C^\infty([0,1]^d)$. In contrast to that, Theorem 5.4 shows weak tractability for the classical norm on the unit ball.

Acknowledgments The authors would like to thank Aicke Hinrichs, Erich Novak, and Mario Ullrich for pointing out relations to the paper [19], as well as Sjoerd Dirksen, Thomas Kühn, and Winfried Sickel for useful comments and discussions. The last author acknowledges the support by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin. The last author was supported by the ERC CZ grant LL1203 of the Czech Ministry of Education.

References

[1] P. Bühlmann, S. van de Geer. Statistics for high-dimensional data. Springer, Heidelberg (2011).

[2] M. D. Buhmann, A. Pinkus. Identifying linear combinations of ridge functions. Adv. in Appl. Math., 22(1999), pp. 103–118.

[3] E. J. Candés. Harmonic analysis of neural networks. Appl. Comput. Harmon. Anal., 6(1999), pp. 197–218.

[4] E. J. Candés, D. L. Donoho. Ridgelets: a key to higher-dimensional intermittency?. Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 357(1999), pp. 2495–2509.

[5] B. Carl, I. Stefani. Entropy, compactness and the approximation of operators. Cambridge Tracts in Mathematics, vol. 98, Cambridge University Press, Cambridge(1990).

[6] A. Cohen, I. Daubechies, R. A. DeVore, G. Kerkyacharian, D. Picard. Capturing ridge functions in high dimensions from point queries. Constr. Approx., 35(2012), pp. 225–243.

[7] J. Creutzig, S. Dereich, T. Müller-Kronbach, K. Ritter. Infinite-dimensional quadrature and approximation of distributions. Found. Comp. Math., 9(2009), pp. 391–429.

[8] F. Cucker, D.-X. Zhou. Learning theory: an approximation theory viewpoint. Cambridge Monographs on Applied and Computational Mathematics, vol. 24, Cambridge University Press, Cambridge (2007).

[9] R. A. DeVore, G. G. Lorentz. Constructive approximation. Springer, Berlin (1993).

[10] D.E. Edmunds, H. Triebel. Function Spaces, Entropy Numbers, Differential Operators. Cambridge Tracts in Mathematics, vol. 120, Cambridge University Press, Cambridge (1996).

[11] H. J. Flad, W. Hackbusch, B. N. Khoromskij, R. Schneider. Concepts of data-sparse tensor-product approximation in many-particle modeling. In V. Olshevsky, E. Tyrtyshnikov (eds.). Matrix Methods: Theory, Algorithms and Applications. World Scientific (2010).

[12] M. Fornasier, K. Schnass, J. Vybíral. Learning functions of few arbitrary linear parameters in high dimensions. Found. Comput. Math., 12(2012), pp. 229–262.

[13] S. Foucart, A. Pajor, H. Rauhut, T. Ullrich. The Gelfand widths of lp-balls for $0 < p \leq 1$. J. Complexity, 26(2010), pp. 629–640.

[14] J. H. Friedman, W. Stuetzle. Projection Pursuit Regression. J. Am. Stat. Assoc., 76(1981), pp. 817–823.

[15] G. K. Golubev. Asymptotically minimax estimation of a regression function in an additive model. Problemy Peredachi Informatsii, 28(1992), pp. 101–112.

[16] R. Graham, N. Sloane. Lower bounds for constant weight codes. IEEE Trans. Inform. Theory, 26(1980), pp. 37–43.

[17] T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical learning. Springer, New York (2001).

[18] A. Hinrichs, S. Mayer. Entropy numbers of spheres in Banach and quasi-Banach spaces. Work in progress.

[19] A. Hinrichs, E. Novak, M. Ullrich, H. Woźniakowski. The curse of dimensionality for numerical integration of smooth functions II. J. Complexity, 30(2014), pp. 117–143.

[20] M. Hristache, A. Juditsky, V. Spokoiny. Direct estimation of the index coefficient in a single-index model. Ann. Statist., 29(2001), pp. 595–623.

[21] T. Kühn. A lower estimate for entropy numbers. J. Approx. Theory 110(2001), pp. 120–124.

[22] B. P. Logan, L. A. Shepp. Optimal reconstruction of a function from its projections. Duke Math. J. 42(1975), pp. 645–659.

[23] G. Lorentz, M. von Golitschek, Y. Makovoz. Constructive Approximation: Advanced Problems. Volume 304 of Grundlehren der Mathematischen Wissenschaften, Springer, Berlin(1996).

[24] V. Maiorov. Geometric properties of the ridge manifold. Adv. Comp. Math., 32(2010), pp. 239–253.

[25] E. Novak, H. Triebel. Function Spaces in Lipschitz Domains and Optimal Rates of Convergence for Sampling. Constr. Approx., 23(2006), pp. 325–350.

[26] E. Novak, H. Woźniakowski. Tractability of Multivariate Problems, Volume I: Linear Information. EMS Tracts in Mathematics, Vol. 6, Eur. Math. Soc. Publ. House, Zürich(2008).

[27] E. Novak, H. Woźniakowski. Approximation of infinitely differentiable multivariate functions is intractable. J. Compl., 25(2009), pp. 398–404.

[28] E. Novak, H. Woźniakowski. Tractability of Multivariate Problems, Volume II: Standard Information for Functionals. EMS Tracts in Mathematics, Vol. 12, Eur. Math. Soc. Publ. House, Zürich (2010).

[29] S. Paskov, J. Traub. Faster evaluation of financial derivatives. J. Portfolio Manage. 22(1995), pp. 113–120.

[30] A. Pinkus. Approximating by ridge functions. In A. Le Méhauté, C. Rabut, L. L. Schumaker (eds.). Surface Fitting and Multiresolution Methods. Vanderbilt University Press(1997), pp. 279–292.

[31] A. Pinkus. Approximation theory of the MLP model in neural networks. Acta Numerica, 8(1999), pp. 143–195.

[32] G. Raskutti, M. J. Wainwright, B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. J. Mach. Learn. Res., 13(2012), pp. 389–427.

[33] C. Schütt. Entropy numbers of diagonal operators between symmetric Banach spaces. J. Approx. Theory, 40(1984), pp. 121–128.

[34] C. Schwab, C. J. Gittelson. Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. Acta Numerica, 20(2011), pp. 291–467.

[35] J. Traub, G. Wassilkowski, H. Wozniakowski. Information-Based Complexity. Academic Press, New York (1988).

[36] H. Triebel. Fractals and Spectra. Birkhäuser, Basel (1997).

[37] H. Tyagi, V. Cevher. Active learning of multi-index function models. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (eds.). Advances in Neural Information Processing Systems 25. Curran Associates (2012), pp. 1475–1483.

[38] J. Vybíral. Weak and quasi-polynomial tractability of approximation of infinitely differentiable functions. J. Complexity 30(2014), pp. 48–55.