The recovery of ridge functions on the hypercube suffers from the curse of dimensionality

Benjamin Doerr, Sebastian Mayer

The recovery of ridge functions on the hypercube suffers from the curse of dimensionality

Benjamin Doerr* Sebastian Mayer†

March 26, 2019

Abstract

A multivariate ridge function is a function of the form $f(x) = g(a^T x)$, where g is univariate and $a \in \mathbb{R}^d$. We show that the recovery of an unknown ridge function defined on the hypercube $[-1,1]^d$ with Lipschitz-regular profile g suffers from the curse of dimensionality when the recovery error is measured in the L_∞-norm, even if we allow randomized algorithms. If a limited number of components of a is substantially larger than the others, then the curse of dimensionality is not present and the problem is weakly tractable provided the profile g is sufficiently regular.

1 Introduction

In the *uniform recovery problem* (or L_∞*-recovery problem*), the aim is to compute an approximation $\widehat{f}$ of an unknown function $f : D \subseteq \mathbb{R}^d \to \mathbb{R}$ such that the approximation error $\|f - \widehat{f}\|_\infty$ is small. The only available information about f is a sequence of samples $f(x_1), \dots, f(x_n)$ and that f belongs to some class of functions F_d, which describes the a priori model assumptions. The sampling points $x_1, \dots, x_n$ may be freely chosen. A measure for the difficulty of the recovery problem is the so-called *information complexity* $n(\varepsilon, F_d)$. It is the smallest number n such that there is an algorithm which evaluates at most n samples and achieves an error $\|f - \widehat{f}\|_\infty \leq \varepsilon$, irrespective of which $f \in F_d$ is presented as input to the algorithm. The general question is what properties of the function class F_d make the recovery problem efficiently solvable.

For functions depending only on a few variables, regularity has proved to be sufficient for the existence of efficient algorithms. This is nicely demonstrated by the classic notion of *generalized Lipschitz regularity*. For $r > 0$ and $m = \lceil r - 1 \rceil$, consider the following

*Laboratoire d'Informatique (LIX), École Polytechnique, CS35003, 91120 Palaiseau, France email: doerr@lix.polytechnique.fr

†Corresponding author, Fraunhofer Center for Machine Learning and Fraunhofer-Institute for Algorithms and Scientific Computing SCAI, Schloss Birlinghoven, 53754 Sankt Augustin, Germany, email: sebastian.mayer@scai.fraunhofer.de

class of univariate functions $B^{\mathrm{Lip}(r)}$ defined on $[-1,1]$. Every $g \in B^{\mathrm{Lip}(r)}$ is m-times continuously differentiable, we have $\max\{\|g\|_\infty, \|g^{(1)}\|_\infty, \dots, \|g^{(m)}\|_\infty\} \leq 1$, and the m-th derivative is Hölder continuous with exponent $\beta = r - m$. Then, it is well-known that the worst-case approximation error of the optimal algorithm decays polynomially in the number of samples,

$$c_r\, n^{-r} \leq \mathrm{err}(n, B^{\mathrm{Lip}(r)}) \leq C_r\, n^{-r}, \tag{1}$$

with positive constants c_r, C_r depending only on r. The optimal algorithm is given by a spline. Since $\mathrm{err}(n, B^{\mathrm{Lip}(r)})$ is inverse to the information complexity $n(\varepsilon, B^{\mathrm{Lip}(r)})$, this implies $n(\varepsilon, B^{\mathrm{Lip}(r)}) \simeq \varepsilon^{-1/r}$.

Using the same notion of regularity for d-variate functions, the picture changes dramatically for large d. Let $B_d^{\mathrm{Lip}(r)}$ be the counterpart[1] of $B^{\mathrm{Lip}(r)}$ for d-variate functions defined on the cube $[-1,1]^d$. It is a classical result from approximation theory [3] that

$$c_{r,d}\, n^{-r/d} \leq \mathrm{err}(n, B_d^{\mathrm{Lip}(r)}) \leq C_{r,d}\, n^{-r/d},$$

where $c_{r,d}, C_{r,d}$ denote positive constants depending on r and d. This shows that the asymptotic decay of the error is extremely slow in large dimensions and that for small error thresholds ε, we certainly have $n(\varepsilon, B_d^{\mathrm{Lip}(r)}) \simeq_d (1/\varepsilon)^{d/r}$. That indeed any algorithm needs exponentially many samples to guarantee a non-trivial error has been shown only recently by Novak and Woźniakowski [29], who proved that

$$n(\varepsilon, B_d^{\mathrm{Lip}(r)}) \geq 2^{\lfloor d/2 \rfloor}$$

for all $\varepsilon \in (0,1)$ and $d \in \mathbb{N}$. Hence, this recovery problem is *intractable* and suffers from the *curse of dimensionality* in the strict sense of Information-based Complexity (IBC).

The considerations made so far clearly demonstrate that if we want the uniform recovery problem to be efficiently solvable in high dimensions, then we need a priori assumptions stronger than just regularity. In this paper, we study the assumption that the unknown function is a *ridge function*

$$f(x) = g(a^T x), \quad \text{with } g \in B^{\mathrm{Lip}(r)} \text{ and } \|a\|_1 \leq 1. \tag{2}$$

It is common to call the univariate function g the ridge functions's *profile* and to call the d-dimensional vector a the *ridge vector*. Like a linear function, a ridge function is constant along hyperplanes and so we hope that this prior knowledge greatly reduces the complexity of the recovery problem. This idea is not new. In statistics, models based on ridge functions have been used since the early 1980s to avoid the typical issues occurring in nonparametric regression problems over high-dimensional domains. We give a more detailed overview of research on ridge functions in Section 6. In the context of the uniform recovery problem, ridge functions have first been studied by Cohen et al. [7]. Additionally to (2), they assumed that

$$a_i \geq 0, \quad i = 1, \dots, d, \tag{3}$$

[1] See, e.g., the monograph [9] for a formal definition.

and that, for some $0 < p \leq 1$ and $S \in \{1, \ldots, d-1\}$,

$$\|a\|_p \leq 1 \quad \text{and} \quad \|a\|_1 \geq \min\{1, 4S^{1-1/p}\}. \tag{4}$$

Assumption (3) is equivalent to knowing the signs of the ridge vector's components in advance. If $0 < p < 1$, assumption (4) implies that roughly S components of the ridge vector have to be substantially larger than the others[2]. We call this *approximate sparsity* and note that it is a stronger condition than *compressibility*, which only asks for $\|a\|_p \leq 1$, see [13, p. 42]. Given a ridge function f such that (2), (3), and (4) are fulfilled, Cohen et al. [7] employ spline approximation and compressive sensing techniques [13] to obtain an approximation $\widehat{f}$ with error bound

$$\|f - \widehat{f}\|_\infty \lesssim n^{-r} + \begin{cases} \left(\frac{\log(ed/n)}{n}\right)^{1/p-1} & , n < d, \\ 0 & , n \geq d. \end{cases} \tag{5}$$

So under the given conditions, the recovery of a multivariate ridge function is polynomially tractable and almost as easy as the univariate problem, see (1).

Assumption (3) is rather restrictive as it does not allow to model situations where some of the variables $x_1, \ldots, x_d$ may have inhibitory effects but it is not clear which ones. Hence, we investigate in this paper consequences for the complexity if we drop assumption (3) and allow ridge vectors with negative entries, that is, we study the recovery of ridge functions from the class

$$R_d^{r,(p,S)} = \left\{ f : [-1,1]^d \to \mathbb{R} : f(x) = g(a^T x),\ g \in B^{\mathrm{Lip}(r)},\ a \text{ fulfills } (4) \right\}, \tag{6}$$

where $r > 0$, $0 < p \leq 1$, and $S \in \{1, \ldots, d-1\}$. Moreover, we allow algorithms to use randomness, e.g., to use random sampling points. The quantity of interest, for which we wish to prove lower and upper bounds, is then the nth minimal worst-case error in the randomized setting,

$$\mathrm{err}^{\mathrm{ran}}(n, R_d^{r,(p,S)}) = \inf_{S_n} \sup_{f \in R_d^{r,(p,S)}} \left(\mathbb{E}[\|f - S_n(f)\|_\infty^2]\right)^{1/2},$$

where the infimum is taken over all admissible randomized algorithms using at most n function evaluations. See Section 2 for a formal definition of the randomized setting. Note that

$$\mathrm{err}^{\mathrm{ran}}(n, R_d^{r,(p,S)}) \leq \mathrm{err}(n, R_d^{r,(p,S)}).$$

In turns out that dropping assumption (3) leads to a drastic change in the complexity. For $p = 1$, i.e., if ridge vectors are not approximately sparse, we show that

$$\mathrm{err}^{\mathrm{ran}}(n, R_d^{r,(1,S)}) \gtrsim 1$$

[2]Strictly speaking, the paper [7] assumes that $\|a\| = 1$ and $\|a\|_p \leq M$ for some positive constant M. This is equivalent to assuming (4). The latter formulation is more convenient for our considerations.

as long as $1 \leq n \lesssim e^{d/8}$, with an equivalence constant in the estimate that depends only on r. We conclude that the recovery of an unknown ridge function from the class $R_d^{r,(1,S)}$ suffers from the curse of dimensionality, even if we allow sampling points to be chosen adaptively and at random.

When $0 < p < 1$, the answer that we can give is not final. We show the upper bound

$$\mathrm{err}(n, R_d^{r,(p,S)}) \leq C_{r,p,S} \begin{cases} 1 & , 1 \leq n \leq d, \\ \left(\frac{1}{\log(n)}\right)^{r(1/p-1)} & , d \leq n \leq 2^d d^{1/p-1}, \\ 2^{rd}\, n^{-r} & , n \geq 2^d d^{1/p-1} \end{cases}$$

for $r > 1$, $0 < p < 1$, and $S \in \{1, \ldots, d-1\}$, see Theorem 20. The algorithm establishing the upper bound is an extension of the algorithm used in [7], which we have augmented by a search for the most important signs of the ridge vector to compensate for the dropped assumption (3). Although the extended algorithm reaches asymptotically an error decay of n^{-r}, it is important to note that we can establish this rate only for exponentially many sampling points $n > 2^d d^{1/p-1}$. In the *preasymptotic range*, we can only guarantee an error decay that is logarithmic in the number of samples. This implies that the recovery of an unknown ridge function with approximately sparse ridge vector is at least *weakly tractable*, provided

$$r > \frac{1}{1/p - 1}.$$

Unfortunately, it is unclear whether the constructed algorithm is optimal. We are only able to prove a lower bound for a different function class, namely

$$R_d^{r,p} = \left\{ f : [-1,1]^d \to \mathbb{R} : f(x) = g(a^T x),\ g \in B^{\mathrm{Lip}(r)},\ \|a\|_p = 1 \right\}, \tag{7}$$

where $r > 0$, and $0 < p \leq 1$. For this class, we obtain the lower bound

$$\mathrm{err}^{\mathrm{ran}}(n, R_d^{r,p}) \gtrsim \left(\frac{1}{\log(2n)}\right)^{1/p-1},$$

for $r > 0$ and $0 < p \leq 1$, provided $n \lesssim \exp(d/8)$, see Theorem 11. Note that only for $p = 1$, we have

$$R_d^{r,1} = R_d^{r,(1,S)}$$

for all $r > 0$ and $S \in \{1, \ldots, d-1\}$. Otherwise, the function classes are different so that it remains an open problem to prove lower bounds for $\mathrm{err}^{\mathrm{ran}}(n, R_d^{r,(p,S)})$ when $0 < p < 1$.

Outline. The paper is organized as follows. We begin with a thorough definition of the complexity-theoretical setup in Section 2, where we give a definition what we consider to be a deterministic and a randomized algorithm. Then, in Section 3, we prove lower bounds for the worst-case error for both deterministic and randomized algorithms. Section 4 is dedicated to the description of our algorithm, followed by a detailed error analysis in Section 5, which leads to upper bounds on the worst-case error. Finally, we discuss related work in Section 6, in particular, relations to the regression problem in semiparametric statistics.

Acknowledgement. This work was partially developed in the Fraunhofer Cluster of Excellence "Cognitive Internet Technologies".

2 Preliminaries

As usual, we denote by $\mathbb{N}$ the natural numbers $1, 2, 3, \ldots$. Throughout this paper, let $d \in \mathbb{N}$. For sequences $(f_n)_{n\in\mathbb{N}}$, $(g_n)_n n \in \mathbb{N}$, we write $f_n \lesssim g_n$ whenever there is a constant $C > 0$ such that $f_n \leq C g_n$ for all $n \in \mathbb{N}$. Note that the constant need not be absolute. Where necessary, we indicate on what parameters the constant depends.

Let us recall some basic notions. For $p > 0$ and $x \in \mathbb{R}^d$, the quasi-norm $\|x\|_p$ is given by

$$\|x\|_p := \big(\sum_{i=1}^{d} |x_i|^p\big)^{1/p}.$$

For $D \subset \mathbb{R}^d$, consider a function $f : D \to \mathbb{R}$. The *uniform norm* $\|f\|_\infty$ is given by

$$\|f\|_\infty := \sup_{x\in D} |f(x)|.$$

For any positive number $0 < \beta \leq 1$, the *Hölder constant* of order β is given by

$$|f|_\beta := \sup_{\substack{x,y\in[-1,1]^d \\ x\neq y}} \frac{|f(x) - f(y)|}{2\min\{1, \|x-y\|_1\}^\beta}\,. \tag{8}$$

We say that f is *Hölder-continuous* of order β if $|f|_\beta < \infty$. This definition immediately implies the relation

$$|f|_\beta \leq |f|_{\beta'} \text{ if } 0 < \beta < \beta' \leq 1. \tag{9}$$

Let $C([-1,1]^d)$ be the space of continuous functions defined on $[-1,1]^d$, equipped with the norm $\|\cdot\|_\infty$.

2.1 Deterministic algorithms

Understanding the worst-case complexity of the uniform recovery problem for a given function class F_d means to understand how *any* possible algorithm performs in the worst-case on the given class. This requires a rigorous definition of what we consider to be a feasible algorithm. The field of Information-based Complexity (IBC) provides a well-established framework that we follow in this work. Let us first consider the *deterministic setting*, where algorithms are only allowed to acquire information about a function in a deterministic, i.e. non-random, fashion. Note that since all ridge functions in the classes (6) and (7) are continuous, it is sufficient to define the concept of algorithm with respect to a general function class $F_d \subset C([-1,1]^d)$. Moreover, we restrict our considerations to algorithms that only use function evaluations as information operations

and not more general linear functionals. For a general definition of the deterministic setting, discussions and references, we refer to [28].

A deterministic algorithm using at most n function evaluations is a mapping S_n that maps a function $f \in F_d$ to an approximant $S_n(f) \in C([-1,1]^d)$. More specifically, the approximant is given by

$$S_n(f) = \phi(f(x_1), \dots, f(x_n)),$$

where $fx_1, \dots, x_n \in [-1,1]^d$ are sampling points and $\phi : \mathbb{R}^n \to C([-1,1]^d)$. The first sampling point x_1 is completely independent of the input f, while the choice of the remaining points can be adaptive, that is, x_i may functionally depend on the function values $f(x_1), \dots, f(x_{i-1})$. Formally, this means that there are functions

$$\psi_i \colon \mathbb{R}^{i-1} \to [-1,1]^d, \quad i = 2, \dots, n$$

that recursively define the sampling points via

$$x_i = \psi_i(f(x_1), \dots, f(x_{i-1})).$$

We do not make any computational assumptions, e.g., that the functions ϕ and ψ_i describing the algorithm are efficiently computable in some specific model of computation. Beside the a priori information that f is in the class F_d, the algorithm S_n has no other information than the function values $f(x_1), \dots, f(x_n)$.

It remains to define precisely how we quantify the complexity of the recovery problem given a function class F_d. Let us first introduce the *nth minimal worst-case error*

$$\operatorname{err}(n, F_d) := \inf_{S_n} \sup_{f \in F_d} \|f - S_n(f)\|_\infty,$$

where the infimum is taken over all deterministic algorithms that use at most n function values. Then, we define the *information complexity* as the inverse of the minimal worst-case error,

$$n(\varepsilon, F_d) := \min\{n \in \mathbb{N} : \operatorname{err}(n, F_d) \le \varepsilon\}, \quad \varepsilon > 0.$$

Remark 1. The information complexity neglects any computational cost. This is justified as in function recovery problems, the information cost are usually dominating. In particular, for the algorithm studied in Section 4, the computational cost are proportional to the number of used function samples.

2.2 Randomized algorithms

In this paper, we also wish to study algorithms that use randomness in the choice of the sampling points $x_1, \dots, x_n$ and the mapping ϕ. While there is a clear agreement in IBC what to consider a deterministic algorithm, the situation is less settled when it comes to randomized algorithms. The crux are measurability assumptions. We follow the common approach in IBC and assume just as much measurability as required in our proofs. As a result, our definition of randomized algorithm will be less general than in [28], but closely resemble [21].

We first give a precise definition of what we consider to be a sequence of adaptively chosen random sampling points.

Definition 2. Let $n \in \mathbb{N}$. A sequence of n adaptively chosen sampling points is a sequence of random variables $X_1, \dots, X_n$ defined over a common probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ that take values in $[-1,1]^d$ and fulfill the following. For every $i = 2, \dots, n$, there is a mapping

$$L_i : \Omega \times \mathbb{R}^{i-1} \to [-1,1]^d$$

such that, for all $\omega \in \Omega$,

$$(x_1, \dots, x_{i-1}) \mapsto L_i(\omega, x_1, \dots, x_{i-1})$$

is Borel measurable and

$$X_i(\omega) = L_i(\omega, f(X_1(\omega)), \dots, f(X_{i-1}(\omega))).$$

Definition 3. A randomized algorithm using at most n information operations is given by a probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ and a mapping

$$S_n : \Omega \times F \to C([-1,1]^d)$$

such that

$$S_n(\omega, f) = \phi(\omega, f(X_1(\omega)), \dots, f(X_n(\omega))),$$

where

- $\phi : \Omega \times \mathbb{R}^n \to C([-1,1])$ is measurable in the first argument w.r.t. $\mathfrak{A}$ and Borel measurable in the last n arguments,
- $X_1, \dots, X_n$ is a sequence of adaptively chosen random sampling points according to Definition 2.

It remains to define the nth *minimal worst-case error* in the randomized setting as

$$\mathrm{err}^{\mathrm{ran}}(n, F_d) := \inf_{S_n} \sup_{f \in F_d} \left(\mathbb{E}[\|f - S_n(f)\|_\infty^2]\right)^{1/2},$$

where the infimum is taken over all admissible randomized algorithms using at most n function evaluations. The information complexity in the randomized setting is given by

$$n^{\mathrm{ran}}(\varepsilon, F_d) = \inf\{n \in \mathbb{N} : \mathrm{err}^{\mathrm{ran}}(n, F_d) \le \varepsilon\}, \quad \varepsilon > 0.$$

2.3 Complexity classes

In IBC research, various complexity classes for continuous problems have been introduced. Let us introduce those that we encounter in this work. A problem is said to *suffer from the curse of dimensionality* in the deterministic setting if there are $C > 0$ and $\gamma > 1$ such that

$$n(\varepsilon, F_d) \ge C\gamma^d$$

holds for all $\varepsilon > 0$ and infinitely many $d \in \mathbb{N}$. Furthermore, a problem is said to be *weakly tractable* if

$$\lim_{\varepsilon^{-1}+d\to\infty} \frac{\log n(\varepsilon, F)}{\varepsilon^{-1} + d} = 0.$$

Finally, a problem is *polynomially tractable* if there are $C, p, q > 0$ sucht that for all $\varepsilon > 0$ and all $d \in \mathbb{N}$, we have

$$n(\varepsilon, F_d) \leq C(1/\varepsilon)^p d^q.$$

The same notions of tractability can be introduced in the randomized setting by replacing $n(\varepsilon, F_d)$ by $n^{\mathrm{ran}}(\varepsilon, F)$ in the above definitions. For further levels of tractability, we refer to [15, 28, 30, 31, 38, 39].

2.4 Approximation of univariate Lipschitz functions

Let $r > 0$ and $a < b$. By $m = \lfloor\!\lfloor r \rfloor\!\rfloor$ we denote the largest integer strictly less than r. The *Lipschitz space* $\mathrm{Lip}^r([a, b])$ is given by all univariate functions $g\colon [a, b] \to \mathbb{R}$ such that the *Lipschitz norm*

$$\|g\|_{\mathrm{Lip}(r)} = \max\left\{\|g\|_\infty, \|g^{(1)}\|_\infty, \ldots, \|g^{(s)}\|_\infty, |g^{(s)}|_\beta\right\}$$

is finite, where $g^{(i)}$ is the ith derivative and

$$|g^{(m)}|_\beta = \sup_{u,v\in[-1,1]} \frac{|g^{(s)}(u) - g^{(s)}(v)|}{2\min\{1,|u - v|\}^\beta}, \quad \beta = r - m \in (0, 1] \tag{10}$$

the*Hölder constant*. By $B^{\mathrm{Lip}(r)}$ we denote the closed unit ball of $\mathrm{Lip}^r([-1, 1])$.

We recapitulate some basic facts of univariate spline approximation, see [9, Chap. 12] for further background. For $r_0, n \in \mathbb{N}$ and $h = 1/n$, let $\mathcal{P}_h$ be the space of piecewise polynomials of degree $r_0 - 1$ over equidistant intervals, determined by the points ih, $i \in \{\pm 1, \ldots, \pm n\}$, such that on each of these interval the piecewise polynomials have continuous derivatives of order $r_0 - 2$. A *quasi-interpolant* Q_h with step-size h is a linear operator mapping from the continuous functions on $[-1, 1]$ to $\mathcal{P}_h$ such that the application of Q_h uses only the function values at the points

$$ih, \quad i \in \{\pm 1, \ldots, \pm n\}.$$

The resulting spline has the following approximation properties.

Lemma 4. *For $0 < r \leq r_0$ and $g \in B^{\mathrm{Lip}(r)}$ it holds that*

$$\|g - Q_h g\|_{L_\infty} \leq c_r h^r \tag{11}$$

with a constant c_r depending only on r, and

$$\|Q_h g\|_{L_\infty} \leq c_r \max_{i\in\{\pm 1,\ldots,\pm n_g\}} \left|g(ih)\right|, \tag{12}$$

with again a constant c_r depending only on r.

We also need extrapolation in our error analysis. That is, we need error estimates for points outside the interval that has been sampled. Although Lemma 4 does not directly apply then, the reader familiar with spline approximation knows that $Q_h g$ can also be used for extrapolation and that properties similar to (11) and (12) hold true. However, we could not find an explicit statement suitable for our needs in the literature. Hence, let us collect what we need later on in Section 5 in terms of Taylor polynomials.

Given reals $a < b$, let $g \in \mathrm{Lip}^r([a,b])$ and consider the Taylor polynomial

$$T_{m,t_0} g(t) = g(t_0) + \sum_{i=1}^{m} \frac{g^{(i)}(t_0)}{i!}(t-t_0)^i,$$

where $m = \lfloor\!\lfloor r \rfloor\!\rfloor$. Let $\beta = r - m$. For any $t_0, t_1 \in [a,b]$, the standard error estimate for Taylor polynomials gives

$$|g(t_1) - T_{m,t_0} g(t_1)| \le \frac{2}{m!} |g^{(m)}|_\beta (t_1 - t_0)^r, \tag{13}$$

which is the required counterpart of (11).

As a counterpart to (12), we need that if the approximations of the derivatives $g^{(i)}$ are small in absolute value for all $i = 1, \dots, m$, then the polynomial $|T_{m,t_0} g(\cdot) - g(t_0)|$ has to be small in a neighborhood of t_0. To prove this, we have to consider divided differences. For $m \in \mathbb{N}$, the mth *difference* with stepsize $h \in \mathbb{R}$ in the point $t \in \mathbb{R}$ of a univariate function $g : \mathbb{R} \to \mathbb{R}$ is defined as

$$\Delta_h^m(g,t) := \sum_{j=0}^{m} \binom{m}{j} (-1)^{m-j} g(t + jh).$$

The mth *divided difference* is given by $D_h^m(g,t) := h^{-m} \Delta_h^m(g,t)$. For our purposes, it is convenient to work with the representation

$$D_h^m(g,t) = h^{-m+1} \sum_{j=0}^{m-1} \binom{m-1}{j} (-1)^{m-1-j} D_h^1(g, t + jh), \tag{14}$$

which easily follows from the definition of $\Delta_h^m(g,t)$. If g is m times continuously differentiable, then an iterative application of the mean value theorem of calculus gives

$$D_h^m(g,t) = g^{(m)}(t + \xi), \quad \text{for some } \xi \in [t, t + mh]. \tag{15}$$

Lemma 5 (Counterpart of (12) for extrapolation)**.** *Let $g \in B_{\mathrm{Lip}^r([a,b])}$ for $r > 1$ and $m = \lfloor\!\lfloor r \rfloor\!\rfloor$. If, for some $t_0 \in [a,b]$ and $h > 0$, we have*

$$\left| D_{-h}^i(g, t_0) \right| \le 2^{i-1} h^{r-i+1} \quad \text{for } i = 1, \dots, m,$$

then

$$g^{(i)}(t_0) \le C_i h^{r-i} \qquad \text{for } i = 1, \dots, m,$$

with constants $C_i \le 2^m m!$. Consequently, for any $t_1 \in [a,b]$,

$$|T_{m,t_0} g(t_1) - g(t_0)| \le 2^m m! \max\{h, |t_1 - t_0|\}^r.$$

Proof. See Appendix A. □

3 Lower bounds

The subject of this section is to prove strong worst-case error bounds from below for the classes of ridge functions $R_d^{r,p}$ in the deterministic and randomized setting. We establish the lower bounds by constructing suitable "fooling" ridge functions, which force any sampling algorithm to produce a large error. The general idea thereby is as follows. Suppose we find a ridge vector a such that for all n sampling points $x_1, \dots, x_n \in [-1,1]^d$ the inner products fulfill $a^T x_i < \lambda/2$. Then the algorithm cannot distinguish any profile in $B^{\mathrm{Lip}(r)}$ which is supported only on $[\lambda/2, 1]$ from the the profile that is constantly zero. A good fooling profile in this respect is the truncated power

$$g_{r,\lambda}(t) = \max\{0, t - \lambda/2\}^r. \tag{16}$$

3.1 A lower bound in the deterministic setting

Instead of an explicit construction of fooling ridge vectors, we will derive their existence employing a standard method known in discrete mathematics as *Erdős's probabilistic method* [11]. We define a suitable finite subset of the possible ridge vectors and show that a random one of them satisfies our needs with positive probability. This, in particular, implies the existence of such a ridge vector. To control probabilities, we need a standard concentration inequality, which is known as *Hoeffding's inequality*. For a proof, see, e.g., [13].

Lemma 6. *Let $Z_1, \dots, Z_m$ be a sequence of independent random variables with expectation $\mathbb{E}[X_i] = 0$ and $|X_i| \le B_i$ almost surely for $i \in \{1, \dots, m\}$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^m X_i \ge t\right) \le \exp\left(-\frac{t^2}{2\sum_{i=1}^m B_i}\right).$$

Using Hoeffding's inequality, we can prove the following lemma, which is the core ingredient of the probabilistic construction.

Lemma 7. *Let $0 < p \le 1$, $s \in \{1, \dots, d\}$, and $n \in \mathbb{N}$ such that $n < e^{s/8}$. Consider points $z^1, \dots, z^n \in [-1,1]^d$. Then there exists an $a \in \mathbb{R}^d$ with $\|a\|_p = 1$ such that*

$$a^T z^i < \|a\|_1/2 = s^{1-1/p}/2 \qquad \text{for} \quad i = 1, \dots, n.$$

Proof. Let $(\mathfrak{a}_i)_{i=1,\dots,s}$ be a sequence of i.i.d. random variables such that

$$\mathbb{P}(\mathfrak{a}_i = 1/s^{1/p}) = \mathbb{P}(\mathfrak{a}_i = -1/s^{1/p}) = 1/2.$$

By $\mathfrak{a}$ we denote the d-dimensional random vector given by

$$\mathfrak{a} = (\mathfrak{a}_1, \dots, \mathfrak{a}_s, 0, \dots, 0).$$

Note that $\|\mathfrak{a}\|_p = 1$ with probability one. For fixed z^i, we have

$$\mathfrak{a}^T z^i = \sum_{j=1}^{s} Z_j,$$

where $Z_j = \mathfrak{a}_j z_j^i$. The random variables $Z_1, \dots, Z_s$ are independent, each taking values in $[-1/s^{1/p}, 1/s^{1/p}]$. Since $\mathbb{E}[Z_j] = 0$ for all $j \in [s]$, we obtain by Lemma 6 that

$$\mathbb{P}(\mathfrak{a}^T z^i \geq s^{1-1/p}/2) \leq e^{-s/8}.$$

Taking a union bound, we obtain

$$\mathbb{P}\left(\bigcap_{i=1}^{n} \{\mathfrak{a}^T z^i < s^{1-1/p}/2\}\right) \geq 1 - n\, e^{-s/8}.$$

The right hand side of the previous inequality is strictly positive if $n < e^{s/8}$. Thus, for any $n < e^{s/8}$ there exists a realization a of $\mathfrak{a}$ such that

$$a^T z^i < s^{1-1/p}/2$$

for all $1 \leq i \leq n$. By construction of $\mathfrak{a}$ we have $\|a\|_p = 1$. □

Given points $x_1, \dots, x_n \in [-1,1]^d$, Lemma 7 guarantees the existence of an s-sparse $a \in \mathbb{S}_p^{d-1}$ such that all inner products $a^T x_1, \dots, a^T x_n$ are small. To derive a lower bound for the worst-case recovery error from this finding, we have to take into account that the sampling points are not fixed beforehand as in Lemma 7, but may be chosen adaptively by the algorithm given the current input. This is possible and leads to the following lower bound.

Theorem 8. *Let $r > 0$ and $0 < p \leq 1$. Consider the class of ridge functions $R_d^{r,p}$ defined in (7). For any $n \in \mathbb{N}$ with $n < e^{d/8}$ we have*

$$\operatorname{err}(n, R_d^{r,p}), L_\infty) \geq c_r \left(\frac{1}{8\log(2n)}\right)^{r(1/p-1)},$$

where $c_r = 2^{-r}\,\frac{\Gamma(r+1-s)}{\Gamma(r+1)}$ and $s = \lceil\!\lceil r \rceil\!\rceil$. Here, Γ denotes the gamma function given by

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x$$

for $z > 0$. Note that we have $c_r = 2^{-r}/r!$ for $r \in \mathbb{N}$.

Proof. Let S_n be an arbitrary deterministic, adaptive algorithm having a budget of n sampling points. Further, we denote by $z_1, \dots, z_n \in [-1,1]^d$ those sampling points which

are successively used by S_n if the input is the zero function. Choose $s \in \{1, \dots, d\}$ to be the smallest s such that

$$e^{s/8}/2 \leq n < e^{s/8}$$

and let $a^* \in \mathbb{R}^d$ be given by Lemma 7 depending on the points $z_1, \dots, z_n$. Set $\lambda = \|a^*\|_1$ and put $g^*(t) = g_\lambda(t)/\|g_\lambda\|_{\mathrm{Lip}(r)}$, where g_λ is defined in (16). The normalization assures that for the ridge function $f^*(x) = g^*(a^{*T}x)$ we have $\pm f^* \in R_d^{r,p}$.

Let $x_1, \dots, x_n \in [-1,1]^d$ be the sampling points successively chosen by S_n if the input is the ridge function f^*. Since S_n is deterministic, we necessarily have $x_1 = z_1$. We also have $f^*(x_1) = 0$ by construction. It follows inductively that $x_i = z_i$ and $f^*(x_i) = 0$ for all $i = 1, \dots, n$. Consequently, we have $S_n(f^*) = S_n(-f^*) = S_n(0)$ and

$$\begin{aligned}
\sup_{h \in R_d^{r,p}} \big\|h - S_n(h)\big\|_\infty &\geq \max\Big\{\big\|f^* - S_n(f^*)\big\|_\infty, \big\|f^* + S_n(-f^*)\big\|_\infty\Big\} \\
&\geq \|f^*\|_\infty = \|g_\lambda\|_{\mathrm{Lip}(r)}^{-1} \sup_{t \in [-\|a^*\|_1, \|a^*\|_1]} \big|g_\lambda(t)\big| \\
&= \frac{2^{-r}}{\|g_\lambda\|_{\mathrm{Lip(r)}}} \|a^*\|_1^r.
\end{aligned}$$

Standard calculations show that $c_r := 2^{-r}\,\frac{\Gamma(r+1-s)}{\Gamma(r+1)} \leq 2^{-r}/\|g_\lambda\|_{\mathrm{Lip(r)}}$ with c_r independent of λ. Moreover, $\|a^*\|_1 = s^{1-1/p} \geq (8\log(2n))^{1-1/p}$. Since S_n was arbitrary the desired lower bound follows. □

The lower bound allows us to draw an immediate conclusion on the tractability of the uniform recovery problem.

Corollary 9. *Let $S \in \{1, \dots, d\}$ and $r > 0$. The L_∞-recovery of ridge functions from the class $R_d^{r,1}$ or the class $R_d^{r,(1,S)}$ using deterministic sampling algorithms suffers from the curse of dimensionality.*

Proof. Since

$$R_d^{r,1} = R_d^{r,(1,S)},$$

the result is an immediate consequence of Theorem 8 and the definition of the curse of dimensionality. □

3.2 A lower bound for randomized algorithms

The probabilistic constructed described in the previous section can be generalized to work in the randomized setting, as well. We begin with the counterpart to Lemma 7.

Lemma 10. *Let $0 < p \leq 1$ and $0 < \delta < 1$. For $n \in \mathbb{N}$ such that $n < (1-\delta)e^{d/8}$, let $Z^1, \dots, Z^n$ be random vectors defined on a common probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ that take values in $[-1,1]^d$. Then, for any $s \in \{1, \dots, d\}$ that fulfills $n < (1-\delta)e^{s/8}$, there exists an s-sparse vector $a \in \mathbb{R}^d$ such that*

$$\|a\|_p = 1, \quad \|a\|_1 = s^{1-1/p},$$

and

$$\mathbb{P}\big(\bigcap_{i=1}^{n}\{a^T Z^i < \|a\|_1/2\}\big) > \delta.$$

Proof. Let $\lambda = s^{1-1/p}$, $\Omega' = \{-s^{-1/p}, s^{-1/p}\}^s \times \{0\}^{d-s}$ and let $\mathbb{P}'$ denote the uniform distribution on Ω'. The projections

$$\mathfrak{a}_i : \Omega' \ni a \mapsto a_i, \quad i = 1, \dots, s,$$

form a sequence of i.i.d. random variables with

$$\mathbb{P}'(\mathfrak{a}_i = -s^{-1/p}) = \mathbb{P}'(\mathfrak{a}_i = s^{-1/p}) = 1/2.$$

For the d-dimensional random vector $\mathfrak{a} = (\mathfrak{a}_1, \dots, \mathfrak{a}_s, 0, \dots, 0)$ (which is the identity on Ω'), Hoeffding's inequality yields, for all $\omega \in \Omega$, that

$$\mathbb{P}'(\mathfrak{a}^T Z^i(\omega) \geq \lambda/2) \leq e^{-s/8}.$$

By Fubini's theorem, the same estimate holds true with respect to the product probability measure $\widetilde{\mathbb{P}} = \mathbb{P} \otimes \mathbb{P}'$. Namely, for the event

$$\widetilde{\Omega}_i := \{(\omega, a) \in \Omega \times \Omega' \mid a^T X_i(\omega) \geq \lambda/2\},$$

we have

$$\widetilde{\mathbb{P}}(\widetilde{\Omega}_i) = \int_\Omega \mathbb{P}'(\mathfrak{a}^T Z^i(\omega) \geq \lambda/2)\mathbb{P}(d\omega) \leq e^{-s/8}.$$

With the convention $\widetilde{\Omega}_0 = \Omega \times \Omega'$, we can derive from the above estimate that

$$\begin{aligned}\widetilde{\mathbb{P}}\big(\bigcap_{i=1}^{n}\widetilde{\Omega}_i^c\big) &= 1 - \sum_{i=1}^{n}\widetilde{\mathbb{P}}\big(\widetilde{\Omega}_i \cap \bigcap_{j=0}^{i-1}\widetilde{\Omega}_j^c\big) \\ &\geq 1 - \sum_{i=1}^{n}\widetilde{\mathbb{P}}\big(\widetilde{\Omega}_i\big) \geq 1 - ne^{-s/8} > \delta,\end{aligned}$$

where the last estimate is due to the choice of s. Applying Fubini's theorem once again and noting that $\lambda = \|a\|_1$ for all $a \in \Omega'$, we obtain

$$\begin{aligned}\max_{a\in\Omega'}\mathbb{P}\big(\bigcap_{i=1}^{n}\{a^T Z^i < \|a\|_1/2\}\big) &\geq |\Omega'|^{-1}\sum_{a\in\Omega'}\mathbb{P}\big(\bigcap_{i=1}^{n}\{a^T Z^i < \|a\|_1/2\}\big) \\ &= \widetilde{\mathbb{P}}\big(\bigcap_{i=1}^{n}\widetilde{\Omega}_i^c\big).\end{aligned}$$

Hence, we have

$$\max_{a\in\Omega'}\mathbb{P}\big(\bigcap_{i=1}^{n}\{a^T Z^i < \|a\|_1/2\}\big) > \delta,$$

which guarantees the existence of some $a \in \Omega'$ which has the desired properties. □

We come to the main result of this section, a counterpart to Theorem 8 for the randomized setting. Up to this point, we have only considered finite probability spaces such that measurability was not an issue. This is different now that we are considering algorithms that employ random functions.

Theorem 11. *Let $r > 0$, $0 < p \leq 1$, and $0 < \delta \leq 1$. For any $n \in \mathbb{N}$ with $n + 1 < e^{d/8}/2$ we have*

$$\mathrm{err}^{\mathrm{ran}}(n, R_d^{r,p}([-1,1]^d), L_\infty) \geq c_r' \left(\frac{1}{8 \log(4(n+1))} \right)^{r(1/p-1)},$$

with a constant c_r' depending only on the smoothness parameter r.

Proof. Let $\delta = 1/2$. For $s \in \{1, \ldots, d\}$ being the smallest integer such that

$$0.5(1-\delta)e^{s/8} \leq n + 1 < (1-\delta)e^{s/8},$$

let

$$\Omega' := \{-1/s^{1/p}, 1/s^{1/p}\}^s \times \{0\}^{d-s}.$$

Further, we define a ridge function

$$f_a(x) = g_\lambda(a^T x)/\|g_\lambda\|_{\mathrm{Lip}(r)}$$

for each $a \in \Omega'$, where $\lambda = \|a\|_1 = s^{1-1/p}$ and g_λ as defined in (16). Note that $\|f_a\|_\infty = \varepsilon_0$ for all $a \in \Omega'$, where $\varepsilon_0 = 2^{-r}\|a\|_1^r/\|g_\lambda\|_{\mathrm{Lip}(r)}$.

Let $(\Omega, \mathfrak{A}, \mathbb{P})$ be the probability space and let S_n be a randomized sampling algorithm according to Definition 3. That is, the algorithm is given by

$$S_n(f) = \phi(f(X_1), \ldots, f(X_n)),$$

where $X_1, \ldots, X_n$ are adaptively chosen, $[-1,1]^d$-valued random variables according to Definition 2 and ϕ is a Borel measurable random function taking values in the space of all mappings $\mathbb{R}^n \to C([-1,1]^d)$. We show in the following that there is at least one f_a such that

$$\|f_a - S_n(f_a)\|_{L_\infty} \geq \varepsilon_0/2$$

with probability at least δ. We first consider the situation that the algorithm has observed only the function value 0, i.e., the event

$$\Omega_{n,a} := \bigcap_{i=1}^n \{\omega \in \Omega \mid f_a(X_i(\omega)) = 0\}$$

has occurred. By ϕ^0 we denote the random function used by S_n if

$$f(X_1) = \cdots = f(X_n) = 0,$$

i.e., $\phi^0(\omega) = \phi(\omega)(0, \ldots, 0)$, $\omega \in \Omega$. Assuming w.l.o.g. that Ω' is ordered, we define the d-dimensional random vector A^0 by

$$A^0(\omega) = \begin{cases} \min\{a \in \Omega' \mid \|f_a - \phi^0(\omega)\|_{L_\infty} \leq \varepsilon_0/2\}, & \text{if the minimum exists,} \\ 0, & \text{otherwise.} \end{cases}$$

For any $a \in \Omega'$, consider the event $\Omega_a := \{\omega \in \Omega : a^T V(\omega) < \lambda/2\}$, where $V := \operatorname{sgn} A^0$. Conditionally on Ω_a we have $\|f_a - \phi^0\| \geq \varepsilon_0/2$. Namely, for those $\omega \in \Omega_a$ such that $A^0(\omega) = 0$, there is nothing to prove. Otherwise, on $\Omega_a \setminus \{A^0 = 0\}$, we have

$$\|f_a - f_{A^0}\|_{L_\infty} \geq |f_a(V) - f_{A^0}(V)| = |f_{A^0}(V)| = \varepsilon_0,$$

and thus,

$$\|f_a - \phi^0\|_{L_\infty} \geq \|f_a - f_{A^0}\|_{L_\infty} - \|f_{A^0} - \phi^0\|_{L_\infty} \geq \varepsilon_0/2.$$

The considerations made so far show that

$$\begin{aligned}\mathbb{P}(\|f_a - S_n(f_a)\|_{L_\infty} \geq \varepsilon_0/2) &\geq \mathbb{P}(\{\|f_a - S_n(f_a)\|_{L_\infty} \geq \varepsilon_0/2\} \cap \Omega_{n,a} \cap \Omega_a) \\ &= \mathbb{P}(\Omega_{n,a} \cap \Omega_a)\end{aligned}$$

for any $a \in \Omega'$. Hence, what remains is to show that $\mathbb{P}(\Omega_{n,a} \cap \Omega_a) > \delta$ for some $a \in \Omega'$. To this end, consider the sequence of random sampling points $Z_1, \ldots, Z_n$ which the algorithm S_n uses in case that the input function is identical to zero. Obviously, we have $X_1(\omega) = Z_1(\omega)$. Furthermore, if $f(X_1(\omega)) = \cdots = f(X_{j-1}(\omega)) = 0$ for some $j \geq 2$, then, it follows by induction that $X_j(\omega) = Z_j(\omega)$. Consequently, for any $a \in \Omega'$,

$$\Omega_{n,a} = \bigcap_{i=1}^{n} \{f_a(Z^i) = 0\} = \bigcap_{i=1}^{n} \{a \cdot Z^i < \lambda/2\},$$

where the last equality obviously follows from the definition of f_a. Since the random variables $Z_1, \ldots, Z_n$ and V are completely independent of a and take values in $[-1, 1]^d$, we may apply Lemma 10 with $Z_{n+1} = V$ to obtain $\mathbb{P}(\Omega_{n,a^*} \cap \Omega_{a^*}) > \delta$ for some $a^* \in \Omega'$.

Now estimate

$$\mathbb{E}[\|f_{a^*} - S_n(f_{a^*})\|_{L_\infty}^2]^{1/2} \geq \sqrt{\mathbb{P}(\|f_a - S_n(f_a)\|_{L_\infty} \geq \varepsilon_0/2)}\varepsilon_0/2 \geq 2^{-3/2}\varepsilon_0.$$

and since S_n is arbitrary we conclude that

$$\begin{aligned}\operatorname{err}^{\mathrm{ran}}(n, R_d^{r,p}) &\geq 2^{-3/2}\varepsilon_0 \\ &\geq c_r' s^{r(1-1/p)} \\ &= c_r' \left(\frac{1}{8\log(4(n+1))}\right)^{r(1/p-1)},\end{aligned}$$

where $c_r' = 2^{-3/2} c_r$ and c_r is the constant given in Theorem 8. □

4 An algorithm for ridge functions on the cube

For $r > 1, 0 < p \leq 1$, and $S \in \{1, \ldots, d-1\}$, consider the class of ridge functions $R_d^{r,(p,S)}$ defined in (6). If $p = 1$, we know by Corollary 9 that the recovery of an unknown ridge function suffers from the curse of dimensionality. On the other side, if we additionally

know in advance the signs of the unknown ridge vector, then the problem is polynomially tractable, which follows from (5). So the question that remains is how hard is it to recover an unknown ridge function if the ridge vector is approximately sparse in the sense of (4) but we do not know its signs in advance.

In this section, we design an algorithm that tries to exploit this a priori knowledge by finding the signs of the largest components of the unknown ridge vector. This algorithm extends the adaptive algorithm developed in [7, Section 3]. The error analysis will be done in Section 5, where we also discuss consequences for the tractability of the recovery problem.

4.1 Recap: recovery given the signs of the ridge vector

To better understand our idea how to compensate for dropping assumption (3) it is instructive to recapitulate the basic ingredients of the adaptive algorithm described in [7, Section 3]. Let f be a ridge function given by $f(x) = g(a^T x)$ with unknown profile $g \in B^{\mathrm{Lip}(r)}$ and unknown ridge vector a such that $\|a\| \leq 1$ and the signs

$$\operatorname{sgn}(a) = (\operatorname{sgn}(a_1), \dots, \operatorname{sgn}(a_d))$$

are given to us in advance. First note that g and a are not uniquely determined since

$$f(x) = g(a^T x) = g_{\operatorname{sgn}(a)}(\bar{a}^T x),$$

where $\bar{a} = a/\|a\|_1$ and

$$g_{\operatorname{sgn}(a)} : [-1,1] \to \mathbb{R}, \quad t \mapsto f(t \operatorname{sgn}(a)) = g(t\|a\|_1).$$

Since $\max_{x \in [-1,1]^d} |a \cdot x| = a \cdot \operatorname{sgn}(a) = \|a\|_1$, only g restricted to $[-\|a\|_1, \|a\|_1]$ contributes to f and thus $g_{\operatorname{sgn}(a)}$ comprises all relevant information about g. As we know the sign vector $\operatorname{sgn}(a)$, we can access $g_{\operatorname{sgn}(a)}(t)$ via $f(t \operatorname{sgn}(a))$ for any $t \in [-1,1]$. Hence, we can use univariate splines to first approximate $g_{\operatorname{sgn}(a)}$.

To approximate the ridge direction $\bar{a}$, we first search the interval $[-1,1]$ for a point t^* such that $|g'_{\operatorname{sgn}(a)}(t^*)|$ is sufficiently large. Then, we consider the vector $x^* = t^* \operatorname{sgn}(a)$ and use first-order differences to approximate

$$\nabla f(x^*) = g'(\|a\|_1 t^*)\, a = g'_{\operatorname{sgn}(a)}(t^*)\, \bar{a}$$

and exploit $\nabla f(x^*)/\|\nabla f(x^*)\|_1 = \bar{a}$. If there is no t^* such that $g'_{\operatorname{sgn}(a)}(\bar{a}^T x^*)$ is sufficiently large, then $g_{\operatorname{sgn}(a)}$ must be approximately constant and we can approximate f by a constant function.

4.2 The algorithm

Not knowing $\operatorname{sgn}(a)$ in advance, it is still possible to select some $v \in \{-1,1\}^d$ and search the univariate function

$$g_v : [-1,1] \to \mathbb{R}, \quad t \mapsto f(tv) = g(t\, a^T v), \tag{17}$$

for a point t^* such that $g_v'(t^*)$ is sufficiently large, i.e., $g_v'(t^*) > n_g^{-r}$. If such a point is found, then one can proceed very similar as in [7]. The crucial difference comes to light when we do not find t^*. The function g_v is the rescaled restriction of the profile g to the interval $[-|a \cdot v|, |a \cdot v|]$. If $v \neq \operatorname{sgn}(a)$, then $|a \cdot v| < \|a\|_1$ and there is a relevant part of g which we cannot observe. Consequently, not finding t^* does not imply that $g_{\operatorname{sgn}(a)}$ is constant within the approximation accuracy. If we manage to control the difference $|\|a\|_1 - a \cdot v|$, however, then the regularity of g guarantees that $g_{\operatorname{sgn}(a)}$ cannot be too different from g_v.

The key to find a v such that $|a \cdot v|$ is close to $\|a\|_1$ is the approximate sparsity of a. We argue that it suffices that v agrees with $\operatorname{sgn}(a)$ on the s most relevant components of a. To make this latter point precise, let

$$I_{s,a} \subseteq \{1, \dots, d\}$$

be the set of indices of the s in absolute value largest components of a (breaking ties arbitrarily). In Lemma 16, we prove that for every vertex v from the set

$$\mathrm{HIT}_{s,a} := \Big\{ v \in \{-1,1\}^d : v_i = \operatorname{sgn}(a_i) \text{ for } i \in I_{s,a} \Big\}. \tag{18}$$

the difference is at most

$$|\|a\|_1 - a \cdot v| \leq 2s^{1-1/p}.$$

How can we find a vector that is in the set $\mathrm{HIT}_{s,a}$? We sketch in the following how to construct a vertex set $\mathcal{V}$ such that while iterating over $\mathcal{V}$, we are guaranteed to find an element from $\mathrm{HIT}_{s,a}$, see Section 5 for details. It is relatively easy to establish an absolute guarantee using basic combinatorics. However, if we content ourself with finding a suitable vertex with high probability, then a vertex set $\mathcal{V}$ of much smaller cardinality will suffice. Assume that we simply draw v uniformly at random from $\{-1,1\}^d$. The event

$$\{v \in \mathrm{HIT}_{s,a}\}$$

then says that we have guessed the s most important signs of a correctly. Since a random v is in $\mathrm{HIT}_{s,a}$ with probability 2^{-s}, we can ensure that the event $\{v \in \mathrm{HIT}_{s,a}\}$ occurs at least once with probability $1-(1-2^{-s})^{n_v}$ by taking a large enough number n_v of i.i.d. samples. It remains to choose s appropriately with regard to the desired approximation error. By a union bound argument, this construction can be derandomized. With high probability, we construct a set of vertices $\mathcal{V}^{\mathrm{det}} \subseteq \{-1,1\}^d$ with

$$|\mathcal{V}^{\mathrm{det}}| \leq 2^s s \log_2(2d/s)$$

such that for all $a \in \mathbb{R}^d$ with $\|a\|_p \leq 1$ there is $v \in \mathcal{V}^{\mathrm{det}} \cap \mathrm{HIT}_{s,a}$.

The considerations made so far lead to the following procedure to recover an unknown ridge function f from the class $R_d^{r,(p,S)}$, which has three parameters:

- a vertex set $\mathcal{V} \subset \{-1,1\}^d$ that determines in which orthants we search for a large derivative of the ridge functions's profile;

- the number of sampling points $n_g \in \mathbb{N}$ spent for every approximation of the profile;
- the number of refinement steps $n_b \in \mathbb{N}$ used to narrow down the interval where the profile has a large derivative.

Step 1: Until a stopping criterion is met, do the following for every $v \in \mathcal{V}$. Compute the samples $f(jhv)$ where $h = 1/n_g$ and $j \in \{0, \pm 1, \ldots, \pm n_g\}$. Let

$$L_v = \max_{-n_g \le j < n_g} \frac{\big|f((j+1)hv) - f(jhv)\big|}{h}. \tag{19}$$

If

$$L_v > n_g^{-r}, \tag{20}$$

then go to Step 2. If ultimately

$$\max_{v \in \mathcal{V}} L_v \le n_g^{-r}, \tag{21}$$

then let

$$f_{\max} := \max_{v \in \mathcal{V}; j = 0, \pm 1, \ldots, \pm n_g} f(x^{v,j}), \tag{22}$$

$$f_{\min} := \min_{v \in \mathcal{V}; j = 0, \pm 1, \ldots, \pm n_g} f(x^{v,j}). \tag{23}$$

and return $\widehat{f} = 1/2(f_{\min} + f_{\max})$.

Step 2: Let $[t_0, t_1]$ be the interval for which the maximum L_v in Step 1 is attained. Using bisection as in [7], refine this interval to an interval $[t_{\text{mid}} - \delta, t_{\text{mid}} + \delta]$ with interval length $2\delta = h/2^{n_b}$ such that

$$2\frac{|f((t_{\text{mid}} + \delta)v) - f((t_{\text{mid}} - \delta)v)|}{\delta} > n_g^{-r}.$$

Step 3: Let $z_0 = (t_{\text{mid}} - \delta)v$, $z_1 = (t_{\text{mid}} + \delta)v$, $x_0 = t_{\text{mid}}v$, and

$$x_i = t_{\text{mid}}v + \delta e_i, \qquad i = 1, \ldots, d,$$

where $e_1, \ldots, e_d$ are the canonical unit vectors. Compute the vector $\tilde{a}$ with components

$$\tilde{a}_i = 2\frac{f(x_i) - f(x_0)}{f(z_1) - f(z_0)}. \tag{24}$$

Set $\widehat{a} = \tilde{a}/\|\tilde{a}\|$. This is our approximation of the ridge direction $\bar{a}$.

Step 4: Let $h = n_g^{-1}$. Evaluate f at the points $jh\,\mathrm{sgn}(\hat{a})$, where $j \in \{0, \pm 1, \ldots, \pm n_g\}$, which yields the samples $g_{\mathrm{sgn}(\hat{a})}(jh)$. Use these to compute the quasi-interpolant

$$\widehat{g} = Q_h g_{\mathrm{sgn}(\hat{a})},$$

which is our approximation of the profile.

Step 5: Return the final approximation $\widehat{f}$ which is given by $\widehat{f}(x) = \widehat{g}(\widehat{a}^T x)$.

We clarify the choice of the parameters in Section 5. Note that Step 1 is the crucial part where our algorithm differs essentially from the original algorithm studied in [7]. Step 2 corresponds to QSTEP2 in [7], Steps 3 and 4 basically combine QSTEP3 and RSTEP2 in [7].

Remark 12. As in [7], we could in principle refine the scheme by techniques from compressed sensing to further exploit the approximate sparsity of the ridge vector. However, our subsequent analysis shows that for a ridge function $f \in R^{r,(p,S)}([-1,1])$, the overwhelming fraction of function samples is spent for Step 1 so that sparse recovery methods have no effects in this setting in terms of tractability.

5 Error analysis

The analysis of the algorithm described in Section 4.2 is rather lengthy and technical. Basically, we have to distinguish two scenarios:

(A) Step 1 finds an interval of large deviation, i.e, Eq. (20) is fulfilled;

(B) Step 1 does not find such an interval, i.e., Eq. (21) holds true.

This case distinction is analogous to the proof of [7, Thm. 3.2]. In particular, if Scenario (A) occurs, then the error analysis differs from [7] only at some minor technical details. The crucial difference comes to light when Scenario (B) occurs. Then, the error analysis is much more involved than in the proof of [7, Thm. 3.2] since we are not guaranteed to have sampled the complete relevant part of the profile. As we have already sketched in the previous section, choosing the vertex set $\mathcal{V}$ appropriately is crucial in Scenario (B).

5.1 Error analysis for Scenario (A)

Since the error analysis is analogous to [7, Thm. 3.2], we present only the results in this section. The interested reader will find the proofs in Section A in the appendix. We begin with analyzing the recovery of the ridge direction $\bar{a} = a/\|a\|_1$ in Step 3. Note that in Scenario (A) the distinction between approximate sparsity and compressibility is irrelevant.

Lemma 13 (Error analysis for Step 3)**.** *For $r > 1$ and $0 < p \leq 1$, let f be a rige function given by $f(x) = g(a^T x)$ with $g \in B^{\mathrm{Lip}(r)}$ and $\|a\|_p \leq 1$. Set $\rho = \min\{r-1, 1\}$. Given $n_g \in \mathbb{N}$ and $\varepsilon > 0$, choose*

$$n_b := \lceil \rho^{-1} \log_2(4 n_g^{r-\rho}(3+\varepsilon)\varepsilon^{-1}) \rceil \tag{25}$$

for the bisection performed in Step 2. Let $v \in \{-1,1\}^d$ such that (20) holds true. Then, for the approximation $\widehat{a}$ computed by Step 3, we have

$$\| \operatorname{sgn}(a^T v)\widehat{a} - a/\|a\|_1\|_1 \leq \varepsilon/3.$$

Next, we combine the previous lemma with an error analysis for Step 4, which recovers the profile $g_{\operatorname{sgn}(\widehat{a})}$. This leads to the following recovery guarantee for any ridge function with profile in $B^{\mathrm{Lip}(r)}$ and ridge vector a such that $\|a\|_p \leq 1$ for given $0 < p \leq 1$. This completes the analysis of Scenario (A).

Theorem 14. *For $r > 1$ and $0 < p \leq 1$, let f be as in Lemma 13. Given $\varepsilon > 0$, choose*

$$n_g := \lceil (10\, c_r/\varepsilon)^{1/r} \rceil, \tag{26}$$

and n_b as in (25). Let $v \in \{-1,1\}^d$ such that (20) holds true. Let $\widehat{f}$ be the ridge function given by $\widehat{f}(x) = \widehat{g}(\widehat{a}^T x)$, where $\widehat{a}$ is computed in Step 3 and $\widehat{g}$ is computed in Step 4. Then, we have

$$\|f - \widehat{f}\|_\infty \leq \varepsilon.$$

5.2 Error analysis for Scenario (B)

In this section, we show that Scenario (B) implies that the unknown ridge function f can be approximated sufficiently well by a constant function, provided the set $\mathcal{V}$ has certain properties. We start with a simple observation.

Lemma 15. *Let $n_g \in \mathbb{N}$, $v \in \{-1,1\}^d$, and $f \in R_d^{r,(p,S)}$. For the given ridge function f, consider the function g_v as defined in (17). Put*

$$f^v_{\max} := \max_{i \in \{0,\pm 1,\ldots,\pm n_g\}} g_v(ih), \qquad f^v_{\min} := \min_{i \in \{0,\pm 1,\ldots,\pm n_g\}} g_v(ih),$$

and $f^v := (f^v_{\min} + f^v_{\max})/2$. If $L_v \leq n_g^{-r}$, where L_v is defined by (19), then there is a constant $c_r > 0$ such that

$$\|g_v - f^v\|_\infty \leq 2 c_r n_g^{-r},$$

Proof. The proof is analogous to [7, Proof of Thm. 3.2]). First note that

$$|g_v(t_i) - f^v| \leq L \leq h^r$$

for all $i \in \{0, \pm 1, \ldots, \pm n_g\}$, where $h = n_g^{-1}$. Moreover, by properties (11) and (12) from Lemma 4 we obtain

$$\begin{aligned} \|g_v - f^v\|_\infty &\leq \|g_v - f^v + Q_h(g_v - f^v)\|_\infty + \|Q_h(g_v - f^v)\|_\infty \\ &\leq c_r \left(\|g\|_{\mathrm{Lip}(r)} h^r + \max_{i \in \{0,\pm 1,\ldots,\pm n_g\}} |g_v(t_i) - f^v| \right) \leq 2 c_r h^r. \end{aligned}$$

□

Lemma 15 implies that for every $v \in \mathcal{V}$, the profile segment given by g_v is approximately constant. We have to clarify now when this implies that $g_{\mathrm{sgn}(a)}$ is approximately constant, as well. A first step is to control the difference $\|a\|_1 - |a^T v|$.

Lemma 16. *Let $0 < p < 1$ and $a \in \mathbb{R}^d$ with $\|a\|_p \leq 1$. For $s \in \{1, \dots, d-1\}$, consider the set $\mathrm{HIT}_{s,a}$ defined in (18). For all $v \in \mathrm{HIT}_{s,a}$, we have*

$$0 \leq \|a\|_1 - |a^T v| \leq 2s^{1-1/p}.$$

If $v \in \mathrm{HIT}_{d,a}$, then obviously $\|a\|_1 = |a \cdot v|$.

Proof. Let $\pi \colon \{1, \dots, d\} \to \{1, \dots, d\}$ denote a permutation which determines the *non-increasing rearrangement*, say a^*, of a. This means we have

$$a^* = (a_{\pi(1)}, \dots, a_{\pi(d)}) \quad \text{and} \quad a_{\pi(1)} \geq \cdots \geq a_{\pi(d)}.$$

Put $\tilde{v} = \mathrm{sgn}(a^T v)v$. By definition of the set $\mathrm{HIT}_{r,a}$, see (18), and the fact that $\sigma_s(a) = \sum_{i=s+1}^d |a_i^*|$, we have

$$|a^T v| = a^T \tilde{v} = \sum_{i=1}^{s} |a_i^*| + \sum_{i=s+1}^{d} a_i^* \tilde{v}_{\pi(i)} = \|a\|_1 - \sigma_s(a) + \sum_{i=s+1}^{d} a_i^* \tilde{v}_{\pi(i)}.$$

Hence

$$0 \leq \|a\|_1 - |a \cdot v| = \sum_{i=1}^{s} a_i^* (\mathrm{sgn}(a_i^*) - \tilde{v}_{\pi(i)}) \leq 2\sigma_s(a),$$

where

$$\sigma_s(a) := \inf\{\|a - z\|_1 : z \in \mathbb{R}^d \text{ is } s\text{-sparse}\}$$

is the error of the best s-term approximation. The claim now follows from the well-known estimate $\sigma_s(a) \leq s^{1-1/p}$ which holds for all a with $\|a\|_p \leq 1$, see [13, Prop. 2.3]. □

In the course of the proof of the following theorem, it will become clear that, in order to control the error $\|g_{\mathrm{sgn}(a)} - g_v\|_\infty$, we have to bound the quotient

$$\frac{|\|a\|_1 - a^T v|}{|a^T v|}$$

from above. Hence, we need a lower bound on $|a^T v|$. This is the reason why we have to require that the unknown ridge vector a is not only compressible, but approximately sparse. Consequently, we have to assume that the unknown ridge function f is from the class $R_d^{r,(p,S)}$, where $r > 1$, $0 < p < 1$ and $S \in \mathbb{N}$ with $S < d$. Then, we obtain the following result, which is the centerpiece of the analysis of Scenario (B).

Theorem 17. *For $r > 1$, $0 < p < 1$, and $S < d$, let $f \in R_d^{r,(p,S)}$ with $f(x) = g(a^T x)$. Given $n_g \in \mathbb{N}$ and $\mathcal{V} \subseteq \{-1, 1\}^d$, assume that (21) is true and let s be the largest integer $s \in \{S, \dots, d\}$ such that*

$$\mathcal{V} \cap \mathrm{HIT}_{s,a} \neq \emptyset.$$

Define

$$\widehat{f}(x) = \frac{f_{\min} + f_{\max}}{2}, \quad x \in [-1, 1]^d,$$

where $f_{\max}$ and $f_{\min}$ are given by (22) and (23). Let c_r be the constant appearing in Lemma 4 and $\tilde{c}_r = (2 + 4c_r + 2^{\lfloor\!\lfloor r \rfloor\!\rfloor} \lfloor\!\lfloor r \rfloor\!\rfloor!)$. If $s < d$, then

$$\|f - \widehat{f}\|_\infty \leq \tilde{c}_r \max\{(s/S)^{1-1/p}, n_g^{-1}\}^r,$$

whereas if $s = d$, then

$$\|f - \widehat{f}\|_\infty \leq 4c_r n_g^{-r}.$$

Proof. Let $v \in \mathcal{V}$ be one of the vectors for which the scalar product with a is maximized, i.e.,

$$v := \arg\max_{v' \in \mathcal{V}} |a^T v'|.$$

W.l.o.g. we may assume $\mathrm{sgn}(a^T v) = 1$ (since otherwise we can simply replace v by $-v$ in the following arguments).

Let

$$f_{\max}^v := \max_{i \in \{0, \pm 1, \dots, \pm n_g\}} g_v(t_i), \qquad f_{\min}^v := \min_{i \in \{0, \pm 1, \dots, \pm n_g\}} g_v(t_i),$$

and $f^v := (f_{\min}^v + f_{\max}^v)/2$. By Lemma 15, we have that g_v is approximately constant,

$$\|g_v - f^v\|_\infty \leq 2c_r h^r. \tag{27}$$

Next we show that the constant f^v is close to the constant $\widehat{f}$. According to the choice of v, observe that there are indices $i_1, i_2 \in \{-n_g + 1, \dots, n_g\}$ and $\xi_1 \in [t_{i_1 - 1}, t_{i_1}]$, $\xi_2 \in [t_{i_2 - 1}, t_{i_2}]$ such that

$$g_v(\xi_1) = f_{\min}, \quad g_v(\xi_2) = f_{\max}.$$

Hence, by (27),

$$\|\widehat{f} - f^v\|_\infty \leq 1/2 |g_v(\xi_1) - f^v| + 1/2 |g_v(\xi_2) - f^v| \leq 2c_r h^r.$$

If $s = d$, then $v \in \mathrm{HIT}_{d,a}$ and $g_{\mathrm{sgn}(a)} = g_v$ such that the statement follows.

Otherwise, if $S < s < d$, then we have to control

$$\|g_{\mathrm{sgn}(a)} - \widehat{f}\| = \|g_v(\|a\|_1 / |a^T v| \cdot) - \widehat{f}\|,$$

with g_v now considered as a function on $[-1/|a^T v|, 1/|a^T v|]$. For

$$|t| \|a\|_1 / |a^T v| \leq 1,$$

we are in the interval which we have sampled, and thus as before,

$$|g_{\mathrm{sgn}(a)}(t) - \widehat{f}| \leq 4c_r h^r.$$

The crucial case is $|t|\|a\|_1/|a^T v| > 1$. Now we have to extrapolate. Henceforth assume $t > 0$ and put $t_1 = t\|a\|_1/|a^T v|$ (the arguments for $t < 0$ are completely analogous). For $m = \lfloor\!\lfloor r \rfloor\!\rfloor$, let $T_{m,1} g_v$ be the order-m Taylor expansion of g_v in the point 1. By the triangle inequality, we have

$$|g_v(t_1) - \widehat{f}| \leq |g_v(t_1) - T_{m,1} g_v(t_1)| + |T_{m,1} g_v(t_1) - g_v(1)| + |g_v(1) - \widehat{f}|.$$

By (13), we have

$$|g_v(t_1) - T_{k,1} g_v(t_1)| \leq 2|t_1 - 1|^r.$$

Furthermore, since (21) holds true, we can compute from the representation formula (14) that the divided difference

$$|D^i_{-h}(g_v, 1)| \leq 2^{i-1} h^{r-i+1}$$

for all $i = 1, \ldots, s$. Thus, by Lemma 5,

$$|T_{k,1} g_v(t_1) - g_v(1)| \leq 2^k k! \max\{h, |t_1 - 1|\}^r.$$

It remains to estimate $|t_1 - 1| \leq |a^T v|^{-1} |\|a\|_1 - a^T v|$. Since

$$\mathcal{V} \cap \mathrm{HIT}_{s,a} \neq \emptyset$$

by assumption and by definition of v, there is $v' \in \mathrm{HIT}_{s,a}$ such that Lemma 16 gives

$$|\|a\|_1 - a^T v| \leq |\|a\|_1 - a^T v'| \leq 2s^{1-1/p}.$$

Consequently, $|a^T v| \geq \|a\|_1 - 2s^{1-1/p} \geq 2S^{1-1/p}$ and

$$|T_{k,t} g_v(t_1) - g_v(t)| \leq 2^k k! \max\{h, (s/S)^{1-1/p}\}^r.$$

We conclude

$$|g_{\mathrm{sgn}(a)}(t) - \widehat{f}| \leq (2 + 4c_r + 2^k k!) \max\{h, (s/S)^{1-1/p}\}^r.$$

□

5.3 Choice of parameters and vertex set

We now clarify how to choose the parameters $\mathcal{V}$, n_g, and n_b of the algorithm described in Section 4.2 such that, for given $0 < \varepsilon < 1$, an approximation error of at most ε can be guaranteed. We first consider the case in which we randomly draw vertices.

Theorem 18. *Assume*

$$f \in R_d^{r,(p,S)}, \quad f(x) = g(a^T x),$$

where $r > 1$, $0 < p \leq 1$, and $S \in \mathbb{N}$ with $S < d$. Let $C_r = \max\{c_r, \tilde{c}_r\}$ where c_r is the constant from Lemma 4 and $\tilde{c}_r$ is defined in Theorem 17.

Given $0 < \varepsilon < 1$ and a failure probability $0 < \delta < 1$, choose

$$\begin{aligned} s &:= \min\{S\lceil (C_r/\varepsilon)^{1/(r(1/p-1))}\rceil, d\}, \\ n_v &:= 2^s \lceil \log(1/\delta)\rceil, \\ n_g &:= \lceil (C_r/\varepsilon)^{1/r}\rceil, \\ n_b &\text{ as in Lemma 13.} \end{aligned}$$

If $n_v < 2^d$, let $\mathcal{V} = \{v_1, \ldots, v_{n_v}\}$ be n_v vertices drawn independently and uniformly at random. If $n_v = 2^d$, then let $\mathcal{V} = \{-1, 1\}^d$. Given these parameter choices, the approximation $\widehat{f}$ computed by the algorithm described in Section 4.2 fulfills

$$\mathbb{P}(\|f - \widehat{f}\|_\infty \leq \varepsilon) \geq \begin{cases} 1 & n_v \geq 2^d \\ 1 - \delta & n_v < 2^d. \end{cases}$$

Proof. Case $n_v \geq 2^d$: We have $\mathcal{V} = \{-1, 1\}^d$ and

$$\mathcal{V} \cap \mathrm{HIT}_{d,a} = \{\mathrm{sgn}(a), -\mathrm{sgn}(a)\}.$$

If (21) is true, then Theorem 17 yields

$$\|f - \widehat{f}\|_\infty \leq 4c_r n_g^{-r} \leq \varepsilon$$

by the choice of n_g. Otherwise, if (21) is not true, then Theorem 14 gives

$$\|f - \widehat{f}\|_\infty \leq 4c_r n_g^{-r} \leq \varepsilon$$

by the choice of n_g.

Case $n_v < 2^d$: Consider the set of random vertices $\mathcal{V}$. If (21) is not true, then Theorem 14 gives

$$\|f - \widehat{f}\|_\infty \leq 4c_r n_g^{-r} \leq \varepsilon$$

by the choice of n_g. The fact that $\mathcal{V}$ was chosen at random is irrelevant in this case.

Assume, (21) is true. By the definition of $\mathrm{HIT}_{s,a}$, see (18), it is clear that $\mathbb{P}(v \in \mathrm{HIT}_{s,a}) = 2^{-s+1}$ for any $v \in \mathcal{V}$. Consequently, the probability that $\mathcal{V} \cap \mathrm{HIT}_r = \emptyset$ is at most $(1 - 2^{-s+1})^{n_v}$. Since

$$-2/x \leq \log(1 - 1/x) \leq -1/x,$$

we have $(1 - 2^{-s+1})^{n_v} \leq \delta$ by our choice of n_v. Hence, with probability at least $1 - \delta$, $\mathcal{V} \cap \mathrm{HIT}_{r,a} \neq \emptyset$. Then, by Theorem 17 and our choice of parameters,

$$\|f - \widehat{f}\| \leq \tilde{c}_r \max\{(s/S)^{1-1/p}, n_g^{-1}\}^r \leq \varepsilon$$

□

If we are willing to spend a few more samples, a randomly constructed set of vertices $\mathcal{V}$ will be good for all possible ridge vectors *simultaneously*. It requires just a simple union bound argument to prove this. In this way, we use randomness to construct a deterministic version of the algorithm, which uses for all possible inputs f the same set of vertices $\mathcal{V}$. Since $\mathcal{V}$ has been randomly constructed, we only have control over the error of this deterministic algorithm with a certain probability.

Theorem 19. *Assume*

$$f \in R^{r,(p,S)}([-1,1]^d), \quad f(x) = g(a^T x),$$

where $r > 1$, $0 < p \leq 1$, and $S \in \mathbb{N}$ with $S < d$. Given $0 < \varepsilon < 1$ and a desired failure probability $0 < \delta < 1$, choose s, n_g, and n_b as in Theorem 18. Further, choose

$$n_v := 2^s \lceil s \log(d/s)) + \log(1/\delta) \rceil$$

and let $\mathcal{V}$ be as in Theorem 18. Let $\widehat{f}$ be the approximation computed by the procedure introduced in Section 4.2 given the inputs $f, \mathcal{V}, n_g$, and n_b. Then, we have

$$\mathbb{P}\left(\sup_{f \in R^{r,p,S}([-1,1]^d)} \|f - \widehat{f}\|_\infty \leq \varepsilon \right) \geq \begin{cases} 1 & n_v \geq 2^d \\ 1-\delta & n_v < 2^d. \end{cases}$$

Proof. The proof is identical to that of Theorem 18, except for one point. Before, we had to control

$$\sup_{\substack{f \in R^{r,p,S}([-1,1]^d), \\ f(x) = g(a^T x)}} \mathbb{P}(\mathcal{V} \cap \mathrm{HIT}_{r,a} \neq \emptyset) = \max_{\substack{I \subseteq \{1,\dots,d\}, |I| = s \\ u \in \{-1,1\}^s}} \mathbb{P}(\exists v \in \mathcal{V} : v_I = u \vee (-v)_I = u).$$

Now, we use a union bound argument to see that

$$\begin{aligned} &\mathbb{P}\Big((\forall I \subseteq \{1,\dots,d\}, |I| = s)\ (u \in \{-1,1\}^s)\ (\exists v \in \mathcal{V}) :\ v_I = u \vee (-v)_I = u \Big) \\ &\geq 1 - 2^s \binom{d}{s} (1 - 2^{-s+1})^{n_v} \end{aligned}$$

Since $\log \binom{d}{s} \leq s \log(d/s)$, our choice of n_v yields

$$1 - 2^s \binom{d}{s} (1 - 2^{-s+1})^{n_v} \geq 1 - \delta.$$

□

5.4 Upper bounds for the worst-case error

In this section, we translate the results from the previous section into upper bounds for the worst-case recovery error. We begin with the deterministic setting.

Theorem 20. *Let $r > 1$, $0 < p \leq 1$, and $0 < S < d$. For constants $c_{p,S}, C_{r,p,S} > 0$ independent of n and d, we have*

$$\mathrm{err}(n, R_d^{r,(p,S)}) \leq C_{r,p,S} \begin{cases} 1 & , 1 \leq n \leq 4d, \\ \left(\frac{1}{\log(n)}\right)^{r(1/p-1)} & , 4d \leq n \leq c_{p,S} 2^d d^{1/p-1}, \\ 2^{rd}\, n^{-r} & , n \geq c_{p,S} 2^d d^{1/p-1}. \end{cases}$$

Proof. Fix $\delta = 1/2$. Let n_v, n_g, n_b be as in Theorem 19 and put

$$n' = n'(\varepsilon) = n_v n_g + n_g + n_b + d.$$

For every possible n_v, there is a set $\mathcal{V} \subseteq \{-1, 1\}^d$ of cardinality n_v such that the procedure introduced in Section 4.2 yields a mapping $S_{n'}(f)$ which achieves a recovery error

$$\|f - S_{n'}(f)\|_\infty \leq \varepsilon$$

at worst-case information cost n'.

Case $4d \leq n \leq c_{p,S} 2^d d^{1/p-1}$. Let $0 < \varepsilon_0 < 1$ be the smallest ε such that

$$S\lceil (C_r/\varepsilon)^{\frac{1}{r(1/p-1)}} \rceil < d.$$

Then,

$$n' \geq 2^s n_g \geq \frac{1}{4} S^{1-1/p} 2^d d^{1/p-1} = c_{p,S} 2^d d^{1/p-1}.$$

Consequently, if

$$4d \leq n \leq c_{p,S} 2^d d^{1/p-1},$$

then there is $\varepsilon \geq \varepsilon_0$ such that $n \leq n'(\varepsilon)$ and $s < d$. Moreover, we may assume $s \geq \log\log d$, since otherwise $n' \leq 4d$. Now, since

$$\begin{aligned} n_v &\leq 2^{s+2} s \log(d/s) \leq 2^{3s+2} \leq 2 \cdot 16^{S(C_r/\varepsilon)^{\frac{1}{r(1/p-1)}}} \\ n_g &\leq 2(C_r/\varepsilon)^{1/r} \leq 2^{1+1/r(C_r/\varepsilon)^{\frac{1}{r(1/p-1)}}}, \end{aligned}$$

we have

$$n' - d \leq 4 n_v n_g \leq 16 \cdot 2^{(4S+1/r)(C_r/\varepsilon)^{\frac{1}{r(1/p-1)}}}.$$

Using the assumption $n \geq 4d$, we find a constant $c > 1$ such that

$$\mathrm{err}(n, R_d^{r,(p,S)}) \leq \varepsilon \leq c C_r (4S + 1/r)^{1/p-1} \left(\frac{1}{\log(n)}\right)^{r(1/p-1)}.$$

Case $n \geq c_{p,S} 2^d d^{1/p-1}$. Choose $0 < \varepsilon < \varepsilon_0$ such that $n \leq n'(\varepsilon)$. Now

$$n' - d \leq 4 n_v n_g \leq 8 \cdot 2^d (C_r/\varepsilon)^{1/r}$$

and thus

$$\mathrm{err}(n, R_d^{r,(p,S)}) \le \varepsilon \le C_r 16^r 2^d n^{-r}. \tag{28}$$

Case $1 \le n \le 4d$. The trivial algorithm gives

$$\mathrm{err}(n, R_d^{r,(p,S)}) \le \sup_{f \in R^{r,(p,S)}([-1,1]^d)} \|f\|_\infty = 1.$$

□

Corollary 21. *Let* $0 < p < 1$, $S \in \{1, \dots, d-1\}$, *and*

$$r > \frac{1}{1/p - 1}.$$

Then the L_∞*-recovery of an unknown ridge function from the class* $R_d^{r,(p,S)}$ *is at least weakly tractable.*

Proof. Let $C_{r,p,S}$ and $c_{p,S}$ be the constants defined in Theorem 20 and put

$$\varepsilon_1 = C_{r,p,S} \left(\frac{1}{\log(4d)} \right)^{r(1/p-1)}.$$

Then, it follows from Theorem 20 that there are constants C_0 and C_1 that are independent of ε and d such that

$$\log n(\varepsilon, R_d^{r,(p,S)}) \le C_0 + C_1 \begin{cases} \log(d) & , \varepsilon_1 \le \varepsilon \le 1, \\ (1/\varepsilon)^{\frac{1}{r(1/p-1)}} & , \varepsilon < \varepsilon_1. \end{cases}$$

Put $x = 1/\varepsilon + d$. Then, it follows that

$$\log n(\varepsilon, R_d^{r,(p,S)}) \le C_0 + C_1 \log(x) x^{\frac{1}{r(1/p-1)}}$$

and $\lim_{x \to \infty} x^{-1} \log n(\varepsilon, R_d^{r,(p,S)}) = 0$. By definition of weak tractability, the desired result follows. □

For completeness, let us also consider the randomized version of the algorithm described in Section 4.2. Although the randomized version is less costly than its deterministic counterpart, the following result shows that we basically have the same upper bounds as in the deterministic setting.

Theorem 22. *Let* $r > 1$, $0 < p \le 1$, *and* $0 < S < d$. *For* $c_{p,S}$ *as in Theorem 20 and a constant* $C_{r,p,S} > 0$ *independent of* n *and* d, *we have*

$$\mathrm{err}^{\mathrm{ran}}(n, R_d^{r,(p,S)}) \le C_{r,p,S} \begin{cases} 1 & , 1 \le n \le 2d, \\ \left(\frac{1}{\log(n)} \right)^{r(1/p-1)} & , 2d \le n \le c_{p,S} 2^d d^{1/p-1}, \\ 2^{rd}\, n^{-r} & , n \ge c_{p,S} 2^d d^{1/p-1}. \end{cases}$$

Proof. Let n_v, n_b, n_b, and $\mathcal{V}$ as in Theorem 18. With these choices, the procedure introduce in Section 4.2 yields a mapping $S_{n'}(f)$ with information cost

$$n' = n'(\varepsilon) = n_v n_g + n_g + n_b + d.$$

Case $n > c_{p,S} 2^d d^{1/p-1}$. We find $0 < \varepsilon \leq 1$ such that $n \leq n'$ and $s = d$. Then $S_{n'}$ is deterministic and the argumentation is identical to the proof of Theorem 19; by (28), we obtain

$$\mathrm{err}^{\mathrm{ran}}(n, R_d^{r,(p,S)}) \leq C_{r,p,S} 2^{rd} n^{-r}.$$

Case $2d \leq n \leq c_{p,S} 2^d d^{1/p-1}$. Choose $0 < \varepsilon \leq 1$ such that with the choice $\delta = \varepsilon^2$, we have $n \leq n'$ and $s < d$. By Theorem 18, we have

$$\mathbb{P}(\|f - S_{n'}(f)\|_\infty > \varepsilon) \leq \varepsilon^2,$$

which leads, in combination with $\|f - \hat{S}_{n'}(f)\|_\infty \leq 2$ a.s., to the estimate

$$\begin{aligned}
&\mathrm{err}^{\mathrm{ran}}(n, R_d^{r,(p,S)}) \leq \sqrt{\mathbb{E}\|f - S_{n'}(f)\|_\infty^2} \\
&= \sqrt{\int_{\{\|f-S_{n'}(f)\|_\infty > \varepsilon\}} \|f - S_{n'}(f)\|_\infty^2 + \int_{\{\|f-S_{n'}(f)\|_\infty \leq \varepsilon\}} \|f - S_{n'}(f)\|_\infty^2} \\
&\leq \sqrt{5}\varepsilon.
\end{aligned}$$

Now, since

$$\begin{aligned}
n_v &\leq 2^{s+2} \log(1/\varepsilon) \leq 2^{2s+2} \leq 2 \cdot 8^{S(C_r/\varepsilon)^{\frac{1}{r(1/p-1)}}}, \\
n_g &\leq 2(C_r/\varepsilon)^{1/r} \leq 2^{1+1/r(C_r/\varepsilon)^{\frac{1}{r(1/p-1)}}},
\end{aligned}$$

we have

$$n' - d \leq 4 n_v n_g \leq 16 \cdot 2^{(3S+1/r)(C_r/\varepsilon)^{\frac{1}{r(1/p-1)}}}.$$

Using the assumption $n \geq 2d$, we find a constant $c > 1$ such that

$$\mathrm{err}^{\mathrm{ran}}(n, R_d^{r,(p,S)}) \leq \sqrt{5}\varepsilon \leq \sqrt{5} c C_r (3S + 1/r)^{1/p-1} \left(\frac{1}{\log(n)}\right)^{r(1/p-1)}.$$

Case $1 \leq n \leq 4d$. the trivial algorithm gives

$$\mathrm{err}^{\mathrm{ran}}(n, R_d^{r,(p,S)}) \leq \sup_{f \in R^{r,p,S}([-1,1]^d)} \|f\|_\infty = 1.$$

□

Remark 23. In the case $p = 1$, all results hold still true if we replace the class $R_d^{r,(S,1)}$ by $R_d^{r,1}$ since we only have to consider the case $s = d$ in Theorem 17 then. We do not now whether the obtained upper bounds are optimal when $0 < p < 1$.

6 Related Work

There is a vast body of literature that is concerned with ridge functions. Since various mathematical communities have contributed to the research on ridge functions, we find it of value to close this work with a broader overview of related research. This overview does by no means claim to be exhaustive.

6.1 Further work on uniform recovery

An alternative to the component-wise positivity of the ridge vector (3) that also guarantees polynomial tractability is to assume that

$$g'(0) > \kappa \tag{29}$$

for some given $\kappa > 0$. This assumption works both for ridge functions defined on the hypercube and for ridge functions defined on the Euclidean ball

$$B_2^d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\},$$

whereas (3) does not lead to a polynomially tractable problem for ridge functions defined on the hypercube, see [26]. Assumption (29) has been studied in [12, 25, 26], where [25] studies also the effect of noisy measurements. In [26], it has been shown that recovery of ridge functions defined on the Euclidean ball in general suffers from the curse of dimensionality. This finding is based on novel two-sided estimates that reduce the decay behavior of the worst-case recovery error to the decay behavior of entropy numbers of ℓ_p^d-balls. whereas $0 < p < 2$ implies weak tractability for sufficiently large $\alpha > 0$.

There is an obvious generalization of the ridge function model, namely functions of the form

$$f(x) = g(Ax), \quad g : \mathbb{R}^m \to \mathbb{R}, \quad A \in \mathbb{R}^{m \times d},$$

where m is supposed to be much smaller than the ambient dimension d. In [34], such functions are called *generalized ridge functions*. Following the ideas of [4], the paper [12] develops an efficient algorithm for the recovery of generalized ridge functions defined on Euclidean balls, provided the function g fulfills certain integral conditions and the rows of the matrix are compressible. In the case $m = 1$, the integral conditions are fulfilled, e.g., if we assume (29). Instead of compressibility assumptions on the rows of A, the work [41] assumes that the matrix A is a low-rank tensor and obtains an algorithms that requires only polynomially many function samples.

A rank-1 tensor is a multivariate function of the form $f(x_1, \ldots, x_d) = \prod_{j=1}^d f(x_j)$. The recent works [2, 27] study efficient methods and tractability aspects. The proof techniques show interesting resemblances to the techniques used in the context of ridge functions.

6.2 Ridge functions in semi-parametric statistics

The phenomenon "curse of dimensionality" is also known in statistics. In the *regression problem*, one has stochastically independent observations

$$(X^{(1)}, Y_1), \ldots, (X^{(n)}, Y_n),$$

which are assumed to be related by

$$Y_i = f(X^{(i)}) + \epsilon_i, \quad i = 1, \ldots, n,$$

where the ϵ_i are noise terms. The goal is to derive from these observations a reconstruction $\hat{f}$ of the unknown function f such that the least squares error $\|f - \hat{f}\|_2$ is small. In this context, curse of dimensionality refers to the fact that the random sampling points $X^{(1)}, \ldots, X^{(n)}$ are sparsely scattered when they take values in high-dimensional metric spaces. This has the unpleasant consequence that standard nonparametric regression techniques such as kernel estimation, nearest-neighbor, and spline smoothing work poorly in high dimensions since they are based on local averaging.

It has been a prominent idea in statistics to allow only specific functional dependencies in models to mitigate the burden of high-dimensionality. In this way, one seeks to find a compromise between linear models, that scale rather well with the dimension, and fully nonparametric models, which face the issues mentioned above in high-dimensional settings. *Projection pursuit regression (PPR)* [14, 23] is one possible semiparametric approach used since the early 1980s to face the problem of sparsely scattered data. The key assumption is that the unknown regression surface f can be approximated well by a *sum of ridge functions*, i.e.

$$f(x) \approx \sum_{j=1}^{m} g_j(a_j^T x) \tag{30}$$

with $a_j \in \mathbb{R}^d$ and univariate functions g_j. This can be interpreted as a non-linear generalization of *principal component analysis (PCA)* [20]. A widely used simplification of (30) are *additive models* [20, 37], where the a_j are assumed to be coordinate directions.

For $m = 1$ in (30), a closely related semiparametric model is popular in econometrics under the name *single-index model* [19, 24]; it assumes that f is a ridge function,

$$f(x) = g(a^T x).$$

These simple ridge-based regression models have been successfully applied to high-dimensional real-world data, for instance, to identify the variables that influence income [8], the severity of side impact accidents [18], or air pollution [14]. As a particular family of estimation methods, we mention *average gradient estimation (ADE)* [22, 36, 40], which also rests on the idea to exploit $\nabla f(x) = g(a^T x)a$. Concerning theoretical error bounds, root-n-consistency for various estimation methods has been shown, assuming that g is two-times differentiable and $\|a\|_2 = 1$; see, e.g., [8, 17, 18, 22]. If a method is root-n-consistent, then this implies

$$\mathbb{E}\|f - \widehat{f}\|_2 = \mathcal{O}(n^{-1/2}), \tag{31}$$

where f is the unknown ridge function, $\widehat{f}$ the computed estimate and n the number of samples used. There is also a work that proves asymptotically optimal minimax bounds [16]. It seems that all the afore-mentioned results are only of asymptotic nature and hide constants which potentially depend on the dimension d. We further note that the decay rate in (31) mainly reflects the assumptions that have been made for the noise of the samples.

6.3 One-bit compressed sensing

In 1-bit compressed sensing, the aim is to recover a compressible signal $a \in \mathbb{R}^d$ from 1-bit measurements $y_i = \operatorname{sgn}(a^T x_i)$, $i = 1, \ldots, n$, given that $\mathbb{E}y_i = g(a^T x)$ for some unknown, univariate $g : \mathbb{R} \to [-1, 1]$ such that

$$\mathbb{E}[g(X)X] = \lambda > 0 \tag{32}$$

for a standard normal random variable X, see [35] and the references there. Note that the goal here is only to recover the vector a and not the non-linearity g. Further, note the similarity between (32) and the integral condition discussed in [12]. In particular, it is clear that (32) is fulfilled if g is a continuous function with $g(0) > \kappa > 0$.

6.4 Ridge functions as atoms for approximation

The afore-mentioned PPR provides an example for approximating an unknown function *by* a sum of ridge functions. The recent monograph [34] gives a detailed overview of what is known about approximation by sums of ridge functions. This includes, among other aspects, uniqueness of representation, density properties (i.e., what functions can be approximated by sums of ridge functions), degree of approximation, best approximation, and greedy methods. We also recommend the older work [32] by the same author, which is a well-written introduction to this topic.

Closely related to PPR in spirit and in terms of algorithmic approaches are neural network models [1]. For instance, in *single hidden-layer feedforward networks*, the given data is fitted to a function of the form

$$f(x) = \sum_{i=1}^{m} \beta_i \sigma(a_i^T x + b_i).$$

In contrast to projection pursuit regression, the univariate σ, which is called *activation function*, is chosen in advance. Approximation-theoretical properties of these models and also the more general *multilayer feedforward perceptrons (MLP)* have been surveyed in [33]. A new approximation-theoretical approach towards neural networks models—and more generally, learning based on dictionaries—has been established by [5]. This work introduces an analogon to the well-known Fourier and wavelet transforms based on ridge functions, the so-called *ridgelet transform*. This transform provides representations with frame properties that are particularly suited to represent functions with singularities along hypersurfaces. Further statistical properties of ridgelets have been studied in [6]. The paper [10] constructs an orthonormal basis based on ridgelets.

A Further proofs

Proof of Lemma 5. There is $\xi_m \in [1-sh, 1]$ such that $D^m_{-h}(g,1) = g^{(m)}(\xi_m)$. Further, for $\beta = r - s$, the derivative $g^{(m)}$ is Hölder-continuous with

$$|g^{(m)}| \le |a^T v|^m |g^{(m)}|_\beta \le |a^T v|^m.$$

Hence, for all $\xi \in [1 - sh, 1]$ we obtain

$$\begin{aligned} |g^{(m)}(\xi)| &\le |g^{(m)}(\xi_m)| + |\xi - \xi_m|^\beta \\ &\le \min\{|a^T v|^m, 2^{m-1}h^\beta\} + |a^T v|^m (mh)^\beta = (2^{m-1} + s^\beta) h^\beta = C_m h^\beta. \end{aligned}$$

Considering the derivative $g^{(m-1)}$, there is $\xi_{m-1} \in [1-(m-1)h, 1]$ such that

$$D^{m-1}_{-h}(g,1) = g^{(m-1)}(\xi_{m-1}).$$

By the mean value theorem, there is for all $\xi \in [1-(m-1)h, 1]$ a

$$\xi'_m \in [1-(m-1)h, 1]$$

such that

$$g^{(m-1)}(\xi) = g^{(m-1)}(\xi_{m-1}) + g^{(m)}(\xi'_m)(\xi - \xi_{m-1}).$$

By assumption and the previous considerations we conclude

$$|g^{(m-1)}(\xi)| \le 2^{m-2} h^{\beta+1} + C_m (m-1) h^{\beta+1} = C_{m-1} h^{\beta+1},$$

where $C_{m-1} = 2^{m-2} + 2^{m-1}(m-1) + s^\beta (s-1)$.

Iteratively repeating this argument for the remaining derivatives, we obtain

$$|g^{(i)}(\xi)| \le C_i h^{r-i}, \quad \xi \in [1-ih, 1],$$

with $C_i = \sum_{j=i}^m 2^{j-1} \prod_{l=i}^{j-1} l + \prod_{l=i}^{s-1} s^\beta$. It is easy to see that $C_i \le 2^k s!$.

Proof of Lemma 13. We can assume that $a^T v \neq 0$, otherwise the profile segment g_v given by (17) is constant on $[-1,1]$ and consequently, there is nothing to prove. Let $[t_{\text{mid}} - \delta, t_{\text{mid}} + \delta]$ be the refined interval computed in Step 2 using n_b samples. Recall that $\widehat{a} = \widetilde{a}/\|\widetilde{a}\|_1$, where the ith coordinate of $\widetilde{a}$ is given by (24) for $i = 1, \dots, d$. By the fact that

$$\operatorname{sgn}(\|a\|_1^{-1} / (v^T a)) = \operatorname{sgn}(1/(v^T a))$$

and [25, Lemma 3.1], we have

$$\|\operatorname{sgn}(1/(v^T a))\, \widehat{a} - a/\|a\|_1\|_1 \le 2\, \frac{\|\widetilde{a} - a/(a^T v)\|_1}{\|\widetilde{a}\|_1}. \tag{33}$$

Let us prove an upper bound for the right-hand side in (33). Extending the definition in (17), let

$$g_v : [-|a^T v|^{-1}, |a^T v|^{-1}] \to \mathbb{R}, \quad t \mapsto g(ta^T v)$$

denote the stretched profile of which Step 3 observed function values. By the mean value theorem, there is a real number ξ_0 satisfying

$$|\xi_0 - t_{\text{mid}}| \le \delta,$$

and real numbers ξ_i for $i \in \{1, \dots, d\}$ satisfying $|\xi_i - t_{\text{mid}}| \le \delta |a_i|/|a^T v|$ such that

$$\begin{aligned} g_v'(\xi_0) &= \frac{g_v(t_{\text{mid}} + \delta) - g_v(t_{\text{mid}} - \delta)}{2\delta}, \\ g_v'(\xi_i) &= \frac{g_v(t_{\text{mid}} + \frac{\delta a_i}{a^T v}) - g_v(t_{\text{mid}})}{\delta}^T \frac{a^T v}{a_i}. \end{aligned}$$

This implies

$$\widetilde{a}_i = \frac{a_i}{a^T v}^T \frac{g_v'(\xi_i)}{g_v'(\xi_0)} = \frac{a_i}{a^T v} \left(1 + \frac{g_v'(\xi_i) - g_v'(\xi_0)}{g_v'(\xi_0)}\right).$$

Hence

$$\frac{\|\widetilde{a} - a/(v^T a)\|_1}{\|\widetilde{a}\|_1} = \frac{\sum_{i=1}^d |g_v'(\xi_i) - g_v'(\xi_0)||a_i|}{\sum_{i=1}^d |g_v'(\xi_i)||a_i|}.$$

Now, since g_v' is Hölder continuous on $[-|a^T v|^{-1}, |a^T v|^{-1}]$ with exponent ρ, we obtain by (8) that

$$\begin{aligned} |g_v'(\xi_i) - g_v'(\xi_0)| &\le 2|g_v'|_\rho \min\{1, |\xi_i - \xi_0|\}^\rho \\ &\le 2\|g\|_{\text{Lip}(r)} |a^T v| \left(|\xi_i - t_{\text{mid}}|^\rho + |\xi_0 - t_{\text{mid}}|^\rho\right) \\ &\le 2\delta^\rho \|g\|_{\text{Lip}(r)} \left(|a_i|^\rho |a^T v|^{1-\rho} + |a^T v|\right). \end{aligned}$$

By $|v^T a| \le \|v\|_\infty \|a\|_1$ and $|a_i| \le \|a\|_1$, it follows that

$$|g_v'(\xi_i) - g_v'(\xi_0)| \le 4\|g\|_{\text{Lip}(r)} \|a\|_1 \delta^\beta \le 4\delta^\rho.$$

Using $|g_v'(\xi_i)| \ge |g_v'(\xi_0)| - |g_v'(\xi_i) - g_v'(\xi_0)| \ge L - 4\delta^\rho$, we obtain

$$\frac{\|\widetilde{a} - a/v^T a\|_1}{\|\widetilde{a}\|_1} \le \frac{\delta^\rho}{L/4 - \delta^\rho}.$$

The choice of n_b guarantees that

$$\delta = 2^{-n_b} |I_0| = 2^{-n_b}/n_g \le \left(\frac{L\epsilon}{4(6+\varepsilon)}\right)^{1/\rho}. \tag{34}$$

which in turn yields

$$\frac{\delta^\rho}{L/4 - \delta^\rho} \le \varepsilon/3.$$

This proves the statement of this lemma.

Proof of Theorem 14. Let $\gamma := \operatorname{sgn}(v^T a)$. Recall that

$$f(x) = g(a^T x) = g_{\operatorname{sgn}(a)}(\bar{a}^T x),$$

where $\bar{a} = a/\|a\|_1$. Let Q_h denote a quasi-interpolant as introduced in Section 2. For any $x \in [-1,1]^d$, the approximation error can be decomposed into three components,

$$\begin{aligned}
|\widehat{f}(x) - f(x)| &= |(Q_h g_{\operatorname{sgn}(\widehat{a})})(\widehat{a}^T x) - g_{\operatorname{sgn}(a)}(\bar{a}^T x)| \\
&\le |(Q_h g_{\operatorname{sgn}(\widehat{a})})(\widehat{a}^T x) - g_{\operatorname{sgn}(\widehat{a})}(\widehat{a}^T x)| \\
&\qquad + |g_{\operatorname{sgn}(\widehat{a})}(\widehat{a}^T x) - g_{\operatorname{sgn}(a)}(\gamma \widehat{a}^T x)| \\
&\qquad + |g_{\operatorname{sgn}(a)}(\gamma \widehat{a}^T x) - g_{\operatorname{sgn}(a)}(\bar{a}^T x)|.
\end{aligned}$$

The first part is because we can only approximate $g_{\operatorname{sgn}(\widehat{a})}$, the second component is due to the uncertainty regarding the orthant (the signs of the ridge vector), and the third one is due to the uncertainty regarding the ridge vector. By Lemma 4, the choice of n_g gives

$$|(Q_h g_{\operatorname{sgn}(\widehat{a})})(\widehat{a}^T x) - g_{\operatorname{sgn}(\widehat{a})}(\widehat{a}^T x)| \le \|Q_h g_{\operatorname{sgn}(\widehat{a})} - g_{\operatorname{sgn}(\widehat{a})}\|_\infty \le c_r n_g^{-r} \le \varepsilon/3.$$

To treat the second term we need some preliminary calculations. Namely, as in [25, Eq. (3.10)] we have

$$\begin{aligned}
\left|\bar{a}^T(\operatorname{sgn}(\gamma\widehat{a}) - \operatorname{sgn}(\bar{a}))\right| &= \left|\|\bar{a}\|_1 - \|\widehat{a}\|_1 - (\bar{a} - \gamma\widehat{a})^T(\operatorname{sgn}(\gamma\widehat{a}))\right| \\
&\le \|\bar{a} - \gamma\widehat{a}\|_1 \left\|\operatorname{sgn}(\gamma\widehat{a})\right\|_\infty \le \|\bar{a} - \gamma\widehat{a}\|_1,
\end{aligned}$$

since $\|\widehat{a}\|_1 = \|\bar{a}\|_1 = 1$. Then, for the second term we obtain

$$\begin{aligned}
|g_{\operatorname{sgn}(\widehat{a})}(\widehat{a}^T x) - g_{\operatorname{sgn}(a)}(\gamma\widehat{a}^T x)| &= |g\left((a^T \operatorname{sgn}(\widehat{a}))\,(\widehat{a}^T x)\right) - g\left(\|a\|_1(\gamma\widehat{a}^T x)\right)| \\
&\le \|g\|_{\mathrm{Lip}(r)}\,|(\gamma\widehat{a}^T x)|\,|a^T(\operatorname{sgn}(\gamma\widehat{a}) - \operatorname{sgn}(a))| \\
&\le \|a\|_1 \left\|\gamma\widehat{a} - a/\|a\|_1\right\|_1 \\
&\le \left\|\gamma\widehat{a} - a/\|a\|_1\right\|_1
\end{aligned}$$

and for the third term we have

$$\begin{aligned}
|g_{\operatorname{sgn}(\gamma a)}(\widehat{a}^T x) - g(a^T x)| &= |g(\gamma\|a\|_1\widehat{a}^T x) - g(a^T x)| \\
&\le \|g\|_{\mathrm{Lip}(r)}\|a\|_1 \left\|\gamma\widehat{a} - a/\|a\|_1\right\|_1 \\
&\le \left\|\gamma\widehat{a} - a/\|a\|_1\right\|_1.
\end{aligned}$$

By Lemma 13, we have $\left\|\gamma\widehat{a} - a/\|a\|_1\right\|_1 \le \varepsilon/3$, which proves the statement.

References

[1] Martin Anthony and Peter L. Bartlett. *Neural network learning: theoretical foundations*. Cambridge University Press, Cambridge, 1999.

[2] Markus Bachmayr, Wolfgang Dahmen, Ronald DeVore, and Lars Grasedyck. Approximation of high-dimensional rank one tensors. *Constr. Approx.*, 39(2):385–395, 2014.

[3] Nikolai Sergeevich Bakhvalov. On the approximate calculation of multiple integrals. *J. Complexity*, 31(4):502–516, 2015. [English translation; the original appeared in *Vestnik MGU Ser. Mat. Meh. Astr. Fiz. Him.*, 4:3-18, 1959].

[4] Martin D. Buhmann and Allan Pinkus. Identifying linear combinations of ridge functions. *Adv. in Appl. Math.*, 22(1):103–118, 1999.

[5] Emmanuel J. Candès. Harmonic analysis of neural networks. *Appl. Comput. Harmon. Anal.*, 6(2):197–218, 1999.

[6] Emmanuel J. Candès. Ridgelets: estimating with ridge functions. *Ann. Statist.*, 31(5):1561–1599, 2003.

[7] Albert Cohen, Ingrid Daubechies, Ronald DeVore, Gerard Kerkyacharian, and Dominique Picard. Capturing ridge functions in high dimensions from point queries. *Constr. Approx.*, 35(2):225–243, 2012.

[8] Xia Cui, Wolfgang Karl Härdle, and Lixing Zhu. The EFM approach for single-index models. *Ann. Statist.*, 39(3):1658–1688, 2011.

[9] Ronald A. DeVore and George G. Lorentz. *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993.

[10] David L. Donoho. Orthonormal ridgelets and linear singularities. *SIAM J. Math. Anal.*, 31(5):1062–1099, 2000.

[11] Paul Erdos and Joel Spencer. Probabilistic methods in combinatorics. *AMC*, 10:12, 1974.

[12] Massimo Fornasier, Karin Schnass, and Jan Vybíral. Learning functions of few arbitrary linear parameters in high dimensions. *Found. Comput. Math.*, 12(2):229–262, 2012.

[13] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[14] Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *J. Amer. Statist. Assoc.*, 76(376):817–823, 1981.

[15] Michael Gnewuch and Henryk Woźniakowski. Quasi-polynomial tractability. *Journal of Complexity*, 27(3):312–330, 2011.

[16] G. K. Golubev. Asymptotically minimax estimation of a regression function in an additive model. *Problemy Peredachi Informatsii*, 28(2):3–15, 1992.

[17] Peter Hall. On projection pursuit regression. *Ann. Statist.*, 17(2):573–588, 1989.

[18] Wolfgang Härdle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *Ann. Statist.*, 21(1):157–178, 1993.

[19] Wolfgang Härdle, Marlene Müller, Stefan Sperlich, and Axel Werwatz. *Nonparametric and semiparametric models*. Springer Series in Statistics. Springer-Verlag, New York, 2004.

[20] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[21] S. Heinrich. Lower bounds for the complexity of Monte Carlo function approximation. *J. Complexity*, 8(3):277–300, 1992.

[22] Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, 29(3):595–623, 2001.

[23] Peter J. Huber. Projection pursuit. *Ann. Statist.*, 13(2):435–525, 1985. With discussion.

[24] Hidehiko Ichimura. *Estimation of single-index models*. Massachusetts Institute of Technology, 1988.

[25] Anton Kolleck and Jan Vybíral. On some aspects of approximation of ridge functions. *J. Approx. Theory*, 194:35–61, 2015.

[26] Sebastian Mayer, Tino Ullrich, and Jan Vybíral. Entropy and sampling numbers of classes of ridge functions. *Constr. Approx.*, 42(2):231–264, 2015.

[27] Erich Novak and Daniel Rudolf. Tractability of the approximation of high-dimensional rank one tensors. *Constr. Approx.*, 43(1):1–13, 2016.

[28] Erich Novak and Henryk Woźniakowski. *Tractability of multivariate problems. Vol. 1: Linear information*, volume 6 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2008.

[29] Erich Novak and Henryk Woźniakowski. Approximation of infinitely differentiable multivariate functions is intractable. *J. Complexity*, 25(4):398–404, 2009.

[30] Erich Novak and Henryk Woźniakowski. *Tractability of multivariate problems. Volume II: Standard information for functionals*, volume 12 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2010.

[31] Erich Novak and Henryk Woźniakowski. *Tractability of multivariate problems. Volume III: Standard information for operators*, volume 18 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2012.

[32] Allan Pinkus. Approximating by ridge functions. In *Surface fitting and multiresolution methods (Chamonix–Mont-Blanc, 1996)*, pages 279–292. Vanderbilt Univ. Press, Nashville, TN, 1997.

[33] Allan Pinkus. Approximation theory of the MLP model in neural networks. In *Acta numerica, 1999*, volume 8 of *Acta Numer.*, pages 143–195. Cambridge Univ. Press, Cambridge, 1999.

[34] Allan Pinkus. *Ridge functions*, volume 205 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2015.

[35] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, Jan 2013.

[36] James L. Powell, James H. Stock, and Thomas M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, 57(6):1403–1430, 1989.

[37] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13:389–427, 2012.

[38] PawełSiedlecki. Uniform weak tractability. *J. Complexity*, 29(6):438–453, 2013.

[39] Paweł Siedlecki and Markus Weimar. Notes on (s, t)-weak tractability: a refined classification of problems with (sub)exponential information complexity. *J. Approx. Theory*, 200:227–258, 2015.

[40] Thomas M. Stoker. Consistent estimation of scaled coefficients. *Econometrica*, 54(6):1461–1481, 1986.

[41] Hemant Tyagi and Volkan Cevher. Learning non-parametric basis independent models from point queries via low-rank methods. *Appl. Comput. Harmon. Anal.*, 37(3):389–412, 2014.