Learning Functions of Few Arbitrary Linear Parameters in High Dimensions

Massimo Fornasier*, Karin Schnass† and Jan Vybiral‡

Dedicated to Ronald A. DeVore for his 70^{th} birthday

Abstract

Let us assume that f is a continuous function defined on the unit ball of $\mathbb{R}^d$, of the form $f(x) = g(Ax)$, where A is a $k \times d$ matrix and g is a function of k variables for $k \ll d$. We are given a budget $m \in \mathbb{N}$ of possible point evaluations $f(x_i)$, $i = 1, \dots, m$, of f, which we are allowed to query in order to construct a uniform approximating function. Under certain smoothness and variation assumptions on the function g, and an *arbitrary* choice of the matrix A, we present in this paper

1. a sampling choice of the points $\{x_i\}$ drawn at random for each function approximation;

2. algorithms (Algorithm 1 and Algorithm 2) for computing the approximating function, whose complexity is at most polynomial in the dimension d and in the number m of points.

Due to the arbitrariness of A, the choice of the sampling points will be according to suitable random distributions and our results hold with overwhelming probability. Our approach uses tools taken from the *compressed sensing* framework, recent Chernoff bounds for sums of positive-semidefinite matrices, and classical stability bounds for invariant subspaces of singular value decompositions.

AMS subject classification (MSC 2010): 65D15, 03D32, 68Q30, 60B20, 60G50

Key Words: high dimensional function approximation, compressed sensing, Chernoff bounds for sums of positive-semidefinite matrices, stability bounds for invariant subspaces of singular value decompositions.

Communicated by Emmanuel Candes

*Technical University of Munich, Faculty of Mathematics, Boltzmannstraße 3, D-85748 Garching, Germany, email: massimo.fornasier@ma.tum.de.

†Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria, email: karin.schnass@oeaw.ac.at.

‡Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria, email: jan.vybiral@oeaw.ac.at.

1 Introduction

1.1 Learning high dimensional functions from few samples

In large scale data analysis and learning, several real-life problems can be formulated as capturing or approximating a function defined on $\Omega \subset \mathbb{R}^d$ with dimension d very large, from relatively few given samples or queries. The usual assumption on the class of functions to be recovered is smoothness. The more regular a function is, the more accurately and the more efficiently it can be numerically approximated. However, in the field of *information based complexity* it has been clarified that such a problem is in general *intractable*, i.e., it does not have polynomial complexity. To clarify this poor approximation phenomenon, assume

$$\mathcal{F}_d := \{f : [0,1]^d \to \mathbb{R}, \|D^\alpha f\|_\infty \leq 1, \alpha \in \mathbb{N}_0^d\},$$

to be the class of smooth functions we would like to approximate. We define the sampling operator $S_n = \phi \circ N$, where $N : \mathcal{F}_d \to \mathbb{R}^n$ is a suitable measurement operator and $\phi : \mathbb{R}^n \to L_\infty([0,1]^d)$ a recovery map. For example N can take n samples $f(x_i)$, $i = 1, \ldots, n$ of f and ϕ can be a suitable interpolation operator. The approximation error provided by such a sampling operator is given by

$$e(S_n) := \sup_{f \in \mathcal{F}_d} \|f - S_n(f)\|_\infty.$$

With this notion we further define the approximation numbers

$$e(n,d) := \inf_{S_n} e(S_n),$$

indicating the performance of the best sampling method, and

$$n(\varepsilon, d) := \inf\{n : e(n,d) \leq \varepsilon\}, \tag{1}$$

which is the minimal number of samples we need for the best sampling method to achieve a uniform accuracy $\varepsilon \in (0,1)$.

1.2 Intractability results

Recent results by Novak and Woźniakowski [24] state that for a uniform approximation over $\mathcal{F}_d$ we have $e(n,d) = 1$ for all $n \leq 2^{\lfloor d/2 \rfloor} - 1$ or $n(\varepsilon, d) \geq 2^{\lfloor d/2 \rfloor}$ for all $\varepsilon \in (0,1)$. Hence, the number of samples to approximate even a C^∞-function grows exponentially with the dimension d. This result seems to obliterate any hope for an efficient solution of the learning problem in high dimension, and this phenomenon is sometimes referred to as the *curse of dimensionality*.

Nevertheless, very often the high dimensional functions which we can expect as solutions to real-life problems exhibit more structure and eventually are much better behaved with respect to the approximation problem. There are several models currently appearing in the literature for which the approximation problem is *tractable*, i.e., the approximation error does not grow exponentially with respect to the dimension d.

According to the behavior of the *information complexity* $n(\varepsilon, d)$, cf. (1), for small $\varepsilon > 0$ and large $d \in \mathbb{N}$, one speaks about

- *polynomial tractability*: if $n(\varepsilon, d)$ depends polynomially on ε^{-1} and d,
- *strong polynomial tractability*: if $n(\varepsilon, d)$ depends polynomially only on ε^{-1},
- *weak tractability*: if $\lim_{\varepsilon^{-1}+d\to\infty} \frac{\log n(\varepsilon, d)}{\varepsilon^{-1} + d} = 0.$

We point to [23, Chapters 1 and 2] for further notions of tractability and many references.

In the next two subsections we will recount a few relevant approaches leading in some cases to (some sort of) tractability.

1.3 Functions of few variables

A function $f : [0,1]^d \to \mathbb{R}$ of d variables (d large) may be a sum of functions, which only depend on k variables (k small):

$$f(x_1, \ldots, x_d) = \sum_{\ell=1}^{m} g_\ell(x_{i_1}, \ldots, x_{i_k}). \tag{2}$$

In optimization such functions are called *partially separable*. This model arises for instance in physics, when we consider problems involving interaction potentials, such as the Coulomb potential in electronic structure computations, or in social and economical models describing multiagent dynamics. Once k is fixed and $d \to \infty$, the learning problem of such functions is tractable, even if the g_ℓ are not very smooth. We specifically refer to the recent work of DeVore, Petrova, and Wojtaszczyk [13] which describes an adaptive method for the recovery of high dimensional functions in this class, for $m = 1$.

This model can be extended to functions which are only approximatively depending on few variables, by considering the unit ball $\mathcal{H}_{d,\gamma}$ of the weighted Sobolev space of functions $f : [0,1]^d \to \mathbb{R}$ with

$$\|f\|_{d,\gamma}^2 := \sum_{u \subset [d]} \gamma_{d,u}^{-1} \int_{[0,1]^d} \left(\frac{\partial^{|u|}}{\partial x_u} f(x) \right)^2 dx \leq 1, \tag{3}$$

where $[d] := \{1, \ldots, d\}$, and $\gamma := \{\gamma_{d,u}\}$ are non-negative weights; the definition $\frac{0}{0} := 0$ and the choice of $\gamma_{d,u} = 0$ leads us again to the model (2). A study of the tractability of this class, for various weights, can be found in [23].

1.4 Functions of one linear parameter in high dimensions

One of the weaknesses of the model classes introduced above is that they are very coordinate biased. It would be desirable to have results for a class of basis changes

which would make the model basis-independent. A general model assumes that,

$$f(x) = g(Ax), \tag{4}$$

for A an arbitrary $k \times d$ matrix. While solution to this unconstrained problems have so far been elusive, the special case of

$$f(x) = g(a \cdot x), \tag{5}$$

where a is a stochastic vector, i.e., $a = (a_1, \ldots, a_d)$, $a_j \geq 0$, $\sum_{j=1}^{d} a_j = 1$, and $g : [0,1] \to \mathbb{R}$ is a $\mathcal{C}^s$ function for $s > 1$ has been fully addressed with an optimal recovery method in [11].

The aim of this work is to find an appropriate formulation of the general model (4), which generalizes both the model of k active coordinates as well as the model of one stochastic vector, and to analyze the tractability of the corresponding approximation problem. The rest of the paper is organized as follows. After introducing some basic notations, the next section is dedicated to the motivation and discussion of the generalized model. As an introduction to our formulation and solution approach, we then proceed to analyze the simple case of one active direction in Section 3, under milder assumptions on the vector $a = (a_1, \ldots, a_d)$, before finally addressing the fully generalized problem in Section 4. The last section is dedicated to the discussion of further extensions of our approach, to be addressed in successive papers.

1.5 Notations

In the following we will deal exclusively with real matrices and we denote the space of $n \times m$ real matrices by $M_{n\times m}$. The entries of a matrix X are denoted by lower case letters and the corresponding indices, i.e., $X_{ij} = x_{ij}$. The transposed matrix $X^T \in M_{m\times n}$ of a matrix $X \in M_{n\times m}$ is the matrix with entries $x_{ij}^T = x_{ji}$. For $X \in M_{n\times m}$ we can write its (reduced) *singular value decomposition* [19] as

$$X = U\Sigma V^T$$

with $U \in M_{n\times p}$, $V \in M_{m\times p}$, $p \leq \min(n, m)$, matrices with orthonormal columns and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \in M_{p\times p}$ a diagonal matrix where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ are the *singular values*. For specific matrices X we write the singular value decomposition

$$X = U(X)\Sigma(X)V(X)^T = U_X\Sigma_X V_X^T.$$

For symmetric, positive semidefinite matrices, i.e., $X = X^T$ and $v^T X v \geq 0$ for all vectors v, we can take $V = U$ and the singular value decomposition is equivalent to the eigenvalue decomposition. Note also that $\sigma_i(X) = \sqrt{\lambda_i(X^T X)}$, where $\lambda_i(X^T X)$ is the i^{th} largest eigenvalue of the matrix $X^T X$ (actually, this holds for $n \geq m$, whereas we may want to consider XX^T instead of $X^T X$ if $m > n$). The rank of $X \in M_{n\times m}$

denoted by $\operatorname{rank}(X)$ is the number of nonzero singular values. We define the Frobenius norm of a matrix X as

$$\|X\|_F := \left(\sum_{ij} |x_{ij}|^2 \right)^{1/2}.$$

It is also convenient to introduce the ℓ_p^n vector norms

$$\|x\|_{\ell_p^n} := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad 0 < p < \infty.$$

We denote by $I_n \in M_{n\times n}$ the identity matrix. The symbol $B_{\mathbb{R}^n}$ stands for the unit ball and $B_{\mathbb{R}^n}(r)$ for the ball of radius $r > 0$ in $\mathbb{R}^n$. The unit sphere in $\mathbb{R}^n$ is denoted by $\mathbb{S}^{n-1}$. Finally, $\mathcal{L}^n$ indicates the Lebesgue measure in $\mathbb{R}^n$.

2 The General Model $f(x) = g(Ax)$ and Its Simplifications

The first approach one may be tempted to consider to a generalization of (5) is to ask that $f : [0,1]^d \to \mathbb{R}$ is of the form $f(x) = g(Ax)$, where A is a $k \times d$ stochastic matrix with orthonormal rows, i.e., $a_{ij} \geq 0$, $\sum_{j=1}^{d} a_{ij} = 1$ for all $i = 1, \ldots, k$, $AA^T = I_k$, and $g : A([0,1]^d) \to \mathbb{R}$ is a $\mathcal{C}^s$ function for $s > 1$. There are however two main problems with this formulation. The conditions of stochasticity and orthonormality of the rows of A together are very restrictive - the only matrices satisfying both of them are those having only one non-negative entry per column - and the domain of g cannot be chosen generically as $[0,1]^k$ but depends on A, i.e., it is the k-dimensional polytope $A([0,1]^d)$. Thus we will at first return to the unconstrained model in (4) and give up the conditions of stochasticity and orthonormality. This introduces rotational invariance for the rows of A and the quadrant defined by $[0,1]^d$ is no longer set apart as search space. In consequence and to avoid the complications arising with the polytope $A([0,1]^d)$ we will therefore focus on functions defined on the Euclidean ball.
To be precise, we consider functions $f : B_{\mathbb{R}^d}(1+\bar{\epsilon}) \to \mathbb{R}$ of the form (4), where A is an arbitrary $k \times d$ matrix whose rows are in ℓ_q^d, for some $0 < q \leq 1$,

$$\left(\sum_{j=1}^{d} |a_{ij}|^q \right)^{1/q} \leq C_1.$$

Further, we assume, that the function g is defined on the image of $B_{\mathbb{R}^d}(1+\bar{\epsilon})$ under the matrix A and is twice continuously differentiable on this domain, i.e., $g \in C^2(AB_{\mathbb{R}^d}(1+\bar{\epsilon}))$, and

$$\max_{|\alpha|\leq 2} \|D^\alpha g\|_\infty \leq C_2.$$

For $\mu_{\mathbb{S}^{d-1}}$ the uniform surface measure on the sphere $\mathbb{S}^{d-1}$ we define the matrix

$$H^f := \int_{\mathbb{S}^{d-1}} \nabla f(x) \nabla f(x)^T d\mu_{\mathbb{S}^{d-1}}(x). \tag{6}$$

From the identity $\nabla f(x) = A^T \nabla g(Ax)$ we get that

$$H^f = A^T \cdot \int_{\mathbb{S}^{d-1}} \nabla g(Ax) \nabla g(Ax)^T d\mu_{\mathbb{S}^{d-1}}(x) \cdot A, \tag{7}$$

and therefore that the rank of H^f is k or less. We will require H^f to be well conditioned, i.e., that its singular values satisfy $\sigma_1(H^f) \geq \cdots \geq \sigma_k(H^f) \geq \alpha > 0$.
The parameters in our model are the dimension d (large), the linear parameter dimension k (small), the nonnegative constants C_1, C_2, $0 < q \leq 1$, and $0 < \alpha \leq kC_2^2$.
We now show that such a model can be simplified as follows. First of all we see that giving up the orthonormality condition on the rows of A was actually unnecessary. Let us consider the singular value decomposition of $A = U\Sigma V^T$, hence we rewrite

$$f(x) = g(Ax) = \tilde{g}(\tilde{A}x), \quad \tilde{A}\tilde{A}^T = I_k,$$

where $\tilde{g}(y) = g(U\Sigma y)$ and $\tilde{A} = V^T$. In particular, by simple direct computations,

- $\sup_{|\alpha|\leq 2} \|D^\alpha \tilde{g}\|_\infty \leq \sup_{|\alpha|\leq 2} \|D^\alpha g\|_\infty \cdot \max\{\sqrt{k}\sigma_1(A), k\sigma_1(A)^2\}$, and
- $\left(\sum_{j=1}^d |\tilde{a}_{ij}|^q\right)^{1/q} \leq C_1 \sigma_k(A)^{-1} k^{1/q-1/2}$.

Hence, by possibly considering different constants $\tilde{C}_1 = k^{1/q-1/2}\sigma_k(A)^{-1}C_1$ and $\tilde{C}_2 = \max\{\sqrt{k}\sigma_1(A), k\sigma_1(A)^2\}C_2$, we can always assume that $AA^T = I_k$, meaning A is row-orthonormal. Note that for a row-orthonormal matrix A, equation (7) tells us that the singular values of H^f are the same as those of H_g, where

$$H_g := \int_{\mathbb{S}^{d-1}} \nabla g(Ax) \nabla g(Ax)^T d\mu_{\mathbb{S}^{d-1}}(x).$$

The following simple result states that our model is almost well-defined. As we will see later, the conditions on A and f will be sufficient for the unique identification of f by approximation up to any accuracy, but not necessarily for the unique identification of A and g.

Lemma 2.1. *Assume that $f(x) = g(Ax) = \tilde{g}(\tilde{A}x)$ with $A, \tilde{A}$ two $k \times d$ matrices such that $AA^T = I_k = \tilde{A}\tilde{A}^T$ and that H^f has rank k. Then $\tilde{A} = \mathcal{O}A$ for some $k \times k$ orthonormal matrix $\mathcal{O}$.*

Proof. Because A and $\tilde{A}$ are row-orthonormal the singular values of H_g and $H_{\tilde{g}}$ are the same as those of H^f, i.e., we have $H_g = U\Sigma U^T$ and $H_{\tilde{g}} = \tilde{U}\Sigma\tilde{U}^T$, where Σ is a $k \times k$ diagonal matrix containing the singular values of H^f in nonincreasing order and $U, \tilde{U}$ are orthonormal $k \times k$ matrices. Inserting this into (7) we get

$$\begin{aligned} H^f &= A^T H_g A = A^T U \Sigma U^T A \\ &= \tilde{A}^T H_{\tilde{g}} \tilde{A} = \tilde{A}^T \tilde{U} \Sigma \tilde{U}^T \tilde{A}. \end{aligned}$$

$U^T A$ and $\tilde{U}^T \tilde{A}$ are both row-orthonormal, so we have two singular value decompositions of H^f. Because the singular vectors are unique up to an orthonormal transform, we have $\tilde{U}^T \tilde{A} = VU^T A$ for some orthonormal matrix V or $\tilde{A} = \mathcal{O}A$ for $\mathcal{O} = \tilde{U}VU^T$, which is by construction orthonormal. □

With the above observations in mind, let us now restate the problem we are addressing and summarize our requirements. We restrict the learning problem to functions $f : B_{\mathbb{R}^d}(1+\bar{\epsilon}) \to \mathbb{R}$ of the form $f(x) = g(Ax)$, where $A \in M_{k\times d}$ and $AA^T = I_k$. As we are interested in recovering f from a small number of samples, the accuracy will depend on the smoothness of g. In order to get simple convergence estimates, we require $g \in C^2(B_{\mathbb{R}^k}(1+\bar{\epsilon}))$. These choices determine two positive constants C_1, C_2 for which

$$\left(\sum_{j=1}^{d} |a_{ij}|^q\right)^{1/q} \leq C_1, \tag{8}$$

and

$$\sup_{|\alpha|\leq 2} \|D^\alpha g\|_\infty \leq C_2. \tag{9}$$

For the problem to be well-conditioned we need that the matrix H^f is positive definite

$$\sigma_1(H^f) \geq \cdots \geq \sigma_k(H^f) \geq \alpha, \tag{10}$$

for a fixed constant $\alpha > 0$ (actually later we may simply choose $\alpha = \sigma_k(H^f)$).

Remark 1. *Let us shortly comment on condition* (10) *in the most simple case* $k = 1$*, by showing that such condition is actually necessary in order to formulate a* tractable *algorithm for the uniform approximation of* f *from point evaluations.*
The optimal choice of α *is given by*

$$\alpha = \int_{\mathbb{S}^{d-1}} |g'(a\cdot x)|^2 d\mu_{\mathbb{S}^{d-1}}(x) = \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \int_{-1}^{1} (1-|y|^2)^{\frac{d-3}{2}} dy, \tag{11}$$

cf. Theorem 3.7. Furthermore, we consider the function $g \in C^2([-1-\bar{\epsilon}, 1+\bar{\epsilon}])$ *given by* $g(y) = 8(y-1/2)^3$ *for* $y \in [1/2, 1+\bar{\epsilon}]$ *and zero otherwise. Notice that, for every* $a \in \mathbb{R}^d$ *with* $\|a\|_{\ell_2^d} = 1$*, the function* $f(x) = g(a\cdot x)$ *vanishes everywhere on* $\mathbb{S}^{d-1}$ *outside of the cap* $\mathcal{U}(a, 1/2) := \{x \in \mathbb{S}^{d-1} : a\cdot x \geq 1/2\}$*, see Figure 1. The* $\mu_{\mathbb{S}^{d-1}}$ *measure of* $\mathcal{U}(a, 1/2)$ *obviously does not depend on* a *and is known to be exponentially small in* d *[21], see also Section 3.3. Furthermore, it is known, that there is a constant* $c > 0$ *and unit vectors* $a^1, \ldots, a^K$*, such that the sets* $\mathcal{U}(a^1, 1/2), \ldots, \mathcal{U}(a^K, 1/2)$ *are mutually disjoint and* $K \geq e^{cd}$*. Finally, we observe that* $\max_{x\in\mathbb{S}^{d-1}} |f(x)| = f(a) = g(1) = 1$.

We conclude that any *algorithm making only use of the structure of* $f(x) = g(a\cdot x)$ *and the condition* (9) *needs to use exponentially many sampling points in order to distinguish between* $f(x) \equiv 0$ *and* $f(x) = g(a^i \cdot x)$ *for some of the* a^i*'s as constructed above. Hence, some additional conditions like* (8) *and* (10) *are actually necessary to avoid the curse of dimensionality and to achieve at least some sort of tractability. Let us observe that* $\alpha = \alpha(d)$ *decays exponentially with* d *for the function* g *considered above. We shall further discuss the role of* α *in Section 3.3.*

Contrary to the approach in [11] our strategy used to learn functions of the type (4) is to first find an approximation $\hat{A}$ to A. Once this is known, we will give a pointwise

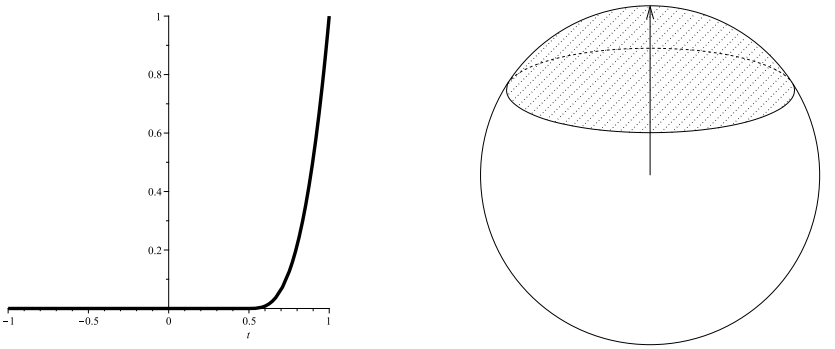

Figure 1: The function g and the spherical cap $\mathcal{U}(a, 1/2)$.

definition of the function $\hat{g}$ on $B_{\mathbb{R}^k}(1)$ such that $\hat{f}(x) := \hat{g}(\hat{A}x)$ is a good approximation to f on $B_{\mathbb{R}^d}(1)$. This will be in a way such that the evaluation of $\hat{g}$ at one point will require only one function evaluation of f. Consequently, an approximation of $\hat{g}$ on its domain $B_{\mathbb{R}^k}(1)$ using standard techniques, like sampling on a regular grid and spline-type approximations, will require a number of function evaluations of f depending only on the desired accuracy and k, but not on d. We will therefore restrict our analysis to the problem of finding $\hat{A}$, defining $\hat{g}$, and the amount of queries necessary to do that.

3 The One Dimensional Case $k = 1$

For the sake of an easy introduction, we start by addressing our recovery method again in the simplest case of a *ridge function*

$$f(x) = g(a \cdot x), \tag{12}$$

where $a = (a_1, \ldots, a_d) \in \mathbb{R}^d$ is a row vector, $\|a\|_{\ell_2^d} = 1$, and g is a function from the image of $B_{\mathbb{R}^d}(1 + \bar{\epsilon})$ under a to $\mathbb{R}$, i.e., $g : B_{\mathbb{R}}(1 + \bar{\epsilon}) \to \mathbb{R}$.
The ridge function terminology was introduced in the 1970's by Logan and Shepp [22] in connection with the mathematics of computer tomography. However these functions have been considered for some time, but under the name of *plane waves*. See, for example, [12, 20]. Ridge functions and ridge function approximation are studied in statistics. There they often go under the name of projection pursuit. Projection pursuit algorithms approximate a function of d variables by functions of the form

$$f(x) \approx \sum_{j=1}^{\ell} g_j(a_j \cdot x). \tag{13}$$

Hence the recovery of f in (12) from few samples can be seen as an instance of the projection pursuit problem. For a survey on some approximation-theoretic questions

concerning ridge functions and their connections to neural networks, see [27] and references therein, and the work of Candès and Donoho on ridgelet approximation [5, 6, 7]. For further clarity of notations, in the following we will assume a to be a row vector, i.e., a $1 \times d$ matrix, while other vectors, $x, \xi, \varphi \ldots$, are always assumed to be column vectors. Hence the symbol $a \cdot x$ stands for the product of the $1 \times d$ matrix a with the $d \times 1$ vector x.

3.1 The Algorithm

As in [11] a basic ingredient of the algorithm is a version of Taylor's theorem giving access to the vector a. For $\xi \in B_{\mathbb{R}^d}$, $\varphi \in B_{\mathbb{R}^d}(r)$, $\epsilon, r \in \mathbb{R}_+$, with $r\epsilon \leq \bar{\epsilon}$, we have, by Taylor expansion, the identity

$$\begin{aligned} [g'(a \cdot \xi)a] \cdot \varphi &= \frac{\partial f}{\partial \varphi}(\xi) \\ &= \frac{f(\xi + \epsilon\varphi) - f(\xi)}{\epsilon} - \frac{\epsilon}{2}[\varphi^T \nabla^2 f(\zeta)\varphi], \end{aligned} \tag{14}$$

for a suitable $\zeta(\xi, \varphi) \in B_{\mathbb{R}^d}(1 + \bar{\epsilon})$. Thanks to our assumptions (8) and (9), the term $[\varphi^T \nabla^2 f(\zeta)\varphi]$ is uniformly bounded as soon as φ is bounded. We will consider the above equality for several directions φ_i and at several sampling points ξ_j.

To be more precise we define two sets $\mathcal{X}, \Phi$ of points. The first

$$\mathcal{X} = \{\xi_j \in \mathbb{S}^{d-1} : j = 1, \ldots, m_{\mathcal{X}}\}, \tag{15}$$

contains the $m_{\mathcal{X}}$ sampling points and is drawn at random in $\mathbb{S}^{d-1}$ according to the probability measure $\mu_{\mathbb{S}^{d-1}}$. For the second, containing the m_Φ derivative directions, we have

$$\begin{aligned} \Phi = \Bigg\{ \varphi_i \in B_{\mathbb{R}^d}(\sqrt{d}/\sqrt{m_\Phi}) : \varphi_{i\ell} = \frac{1}{\sqrt{m_\Phi}} \begin{cases} 1, & \text{with probability } 1/2, \\ -1, & \text{with probability } 1/2, \end{cases} \\ i = 1, \ldots, m_\Phi, \text{ and } \ell = 1, \ldots, d \Bigg\}. \end{aligned} \tag{16}$$

Actually we identify Φ with the $m_\Phi \times d$ matrix whose rows are the vectors φ_i. To write the $m_{\mathcal{X}} \times m_\Phi$ instances of (14) in a concise way we collect the directional derivatives $g'(a \cdot \xi_j)a$, $j = 1, \ldots, m_{\mathcal{X}}$ as columns in the $d \times m_{\mathcal{X}}$ matrix X, i.e.,

$$X = (g'(a \cdot \xi_1)a^T, \ldots, g'(a \cdot \xi_{m_{\mathcal{X}}})a^T), \tag{17}$$

and we define the $m_\Phi \times m_{\mathcal{X}}$ matrices Y and $\mathcal{E}$ entrywise by

$$y_{ij} = \frac{f(\xi_j + \epsilon\varphi_i) - f(\xi_j)}{\epsilon}, \tag{18}$$

and

$$\varepsilon_{ij} = \frac{\epsilon}{2}[\varphi_i^T \nabla^2 f(\zeta_{ij})\varphi_i]. \tag{19}$$

We denote by y_j the columns of Y and by ε_j the columns of $\mathcal{E}$, $j = 1, \dots, m_{\mathcal{X}}$. With these matrices we can write the following factorization

$$\Phi X = Y - \mathcal{E}. \tag{20}$$

The algorithm we propose to approximate the vector a is now based on the fact that the matrix X has a very special structure, i.e., $X = a^T \mathcal{G}^T$, where $\mathcal{G} = (g'(a \cdot \xi_1), \dots, g'(a \cdot \xi_{m_{\mathcal{X}}}))^T$. In other words every column x_j is a scaled copy of the vector a^T and *compressible* if a is *compressible*. We define a vector a compressible informally by saying that it can be well approximated in ℓ_p-norm by a sparse vector. Actually, any vector a with small ℓ_q-norm can be approximated in ℓ_p by its best K-term approximation $a_{[K]}$ according to the following well-known estimate

$$\sigma_K(x)_{\ell_p^d} := \|a - a_{[K]}\|_{\ell_p^d} \leq \|a\|_{\ell_q^d} K^{1/p - 1/q}, \quad p \geq q. \tag{21}$$

Thus by changing view point to get

$$Y = \Phi X + \mathcal{E}$$

we see that due to the random construction of Φ we actually have a *compressed sensing* problem and known theory tells us that we can recover a stable approximation $\hat{x}_j$ to x_j via ℓ_1-minimization (see Theorem 3.2 for the precise statement). To get an approximation of a we then simply have to set $\hat{a} = \hat{x}_j / \|\hat{x}_j\|_{\ell_2^d}$ for j such that $\|\hat{x}_j\|_{\ell_2^d}$ is maximal. From these informal ideas we derive the following algorithm.

Algorithm 1:

- *Given $m_\Phi, m_{\mathcal{X}}$, draw at random the sets Φ and $\mathcal{X}$ as in* (15) *and* (16)*, and construct Y according to* (18)*.*
- *Set $\hat{x}_j = \Delta(y_j) := \arg\min_{y_j = \Phi z} \|z\|_{\ell_1^d}$.*
- *Find*
$$j_0 = \arg \max_{j=1,\dots,m_{\mathcal{X}}} \|\hat{x}_j\|_{\ell_2^d}. \tag{22}$$
- *Set $\hat{a} = \hat{x}_{j_0} / \|\hat{x}_{j_0}\|_{\ell_2^d}$.*
- *Define $\hat{g}(y) := f(\hat{a}^T y)$ and $\hat{f}(x) := \hat{g}(\hat{a} \cdot x)$.*

The quality of the final approximation clearly depends on the error between $\hat{x}_j$ and x_j, which can be controlled through the number of *compressed sensing measurements* m_Φ, and the size of $\hat{a} \approx \max_j \|x_j\|_{\ell_2^d} = \max_j |g'(a \cdot \xi_j)|$, which is related to the number of random samples $m_{\mathcal{X}}$. If (11) is satisfied with α large, we shall show in Lemma 3.6 with help of Hoeffding's inequality that also $\max_j \|x_j\|_{\ell_2^d} = \max_j |g'(a \cdot \xi_j)|$ is large with high probability. If the value of α is unknown and small, the values of $\|\hat{x}_j\|_{\ell_2^d}$ produced by Algorithm 1 could be small as well and, as discussed after the formula (11), no reliable and tractable approximation procedure is possible.

To be exact we will in the next section prove the following approximation result.

Theorem 3.1. *Let $0 < s < 1$ and $\log d \leq m_\Phi \leq [\log 6]^{-2} d$. Then there is a constant c_1' such that using $m_\mathcal{X} \cdot (m_\Phi + 1)$ function evaluations of f, Algorithm 1 defines a function $\hat{f} : B_{\mathbb{R}^d}(1+\bar{\epsilon}) \to \mathbb{R}$ that, with probability*

$$1 - \left(e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}} + 2e^{-\frac{2m_\mathcal{X} s^2 \alpha^2}{C_2^4}} \right), \tag{23}$$

will satisfy

$$\|f - \hat{f}\|_\infty \leq 2C_2(1+\bar{\epsilon}) \frac{\nu_1}{\sqrt{\alpha(1-s)} - \nu_1}, \tag{24}$$

where

$$\nu_1 = C' \left(\left[\frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2 - 1/q} + \frac{\epsilon}{\sqrt{m_\Phi}} \right) \tag{25}$$

and C' depends only on C_1 and C_2 from (8) *and* (9).

Remark 2. *1. We shall fix ν_1 as defined by* (25) *for the rest of this section. Furthermore, we suppose that the selected parameters (s, ϵ and m_Φ) are such that $\nu_1 < \sqrt{\alpha(1-s)}$ holds. See Remark 4 (ii) for knowing how we can circumvent in practice the case that this condition may not hold, clearly invalidating the approximation* (24).

2. In order to show a concrete application of the previous result, let us assume for simplicity that we consider a class of functions g such that $|g'(0)| \neq 0$; hence, by Proposition 3.8, $\alpha = \alpha(g) > 0$ is independent of the dimension d. If additionally we choose $q = 1$, $m_\Phi < d$, and $\epsilon > 0$ such that $m_\Phi(\epsilon + \log(d/m_\Phi))^{-2} = \mathcal{O}(\delta^{-2}\alpha^{-1})$, $1/\log(d) > \delta > 0$, for $\delta, \alpha \to 0$ and $m_\mathcal{X} = \mathcal{O}(\alpha^{-2})$ for $\alpha \to 0$, then, according to Theorem 3.1, we obtain the uniform error estimate

$$\|f - \hat{f}\|_\infty = \mathcal{O}(\delta), \quad \delta \to 0,$$

with high probability. Notice that, in this case, the number of evaluation points $m_\mathcal{X} \cdot (m_\Phi + 1) = \mathcal{O}(\delta^{-3}\alpha^{-3})$, for $\delta, \alpha \to 0$ and $d \to \infty$, is actually independent of the dimension d.

3.2 The Analysis

We will first show that $\hat{x}_j$ is a good approximation to x_j for all j. This follows by the results from the framework of *compressed sensing* [3, 8, 10, 14, 16, 18, 17]. In particular, we state the following useful result which is a specialization of Theorem 1.2 from [36], to the case of Bernoulli matrices.

Theorem 3.2. *Assume that Φ is an $m \times d$ random matrix with all entries being independent Bernoulli variables scaled with $1/\sqrt{m}$, see, e.g.,* (16).

(i) Let $0 < \delta < 1$. Then there are two positive constants $c_1, c_2 > 0$, such that the matrix Φ has the Restricted Isometry Property

$$(1-\delta)\|x\|_{\ell_2^d}^2 \leq \|\Phi x\|_{\ell_2^m}^2 \leq (1+\delta)\|x\|_{\ell_2^d}^2 \tag{26}$$

for all $x \in \mathbb{R}^d$ *such that* $\#\operatorname{supp}(x) \leq c_2 m/\log(d/m)$ *with probability at least*

$$1 - e^{-c_1 m}. \tag{27}$$

(ii) Let us suppose that $d > [\log 6]^2 m$. *Then there are positive constants* $C, c_1', c_2' > 0$, *such that, with probability at least*

$$1 - e^{-c_1' m} - e^{-\sqrt{md}}, \tag{28}$$

the matrix Φ *has the following property. For every* $x \in \mathbb{R}^d$, $\varepsilon \in \mathbb{R}^m$ *and every natural number* $K \leq c_2' m/\log(d/m)$ *we have*

$$\|\Delta(\Phi x + \varepsilon) - x\|_{\ell_2^d} \leq C \left(K^{-1/2} \sigma_K(x)_{\ell_1^d} + \max\{\|\varepsilon\|_{\ell_2^m}, \sqrt{\log d}\|\varepsilon\|_{\ell_\infty^m}\} \right), \tag{29}$$

where

$$\sigma_K(x)_{\ell_1^d} := \inf\{\|x - z\|_{\ell_1^d} : \#\operatorname{supp} z \leq K\}$$

is the best K-term approximation of x.

Remark 3. *(i) The first part of Theorem 3.2 is well known, see, e.g., [3] or [16, Page 15] and references therein.*
(ii) The second part of Theorem 3.2 is relatively new. It follows from Theorem 2.3 of [36] combined with Theorem 3.5 of [13], and the first part of Theorem 3.2. Without the explicit bound of the probability (28)*, it appears also as Theorem 1.2 in [36].*

Applied to the situation at hand we immediately derive the following corollary.

Corollary 3.3. *(i) Let* $d > [\log 6]^2 m_\Phi$. *Then with probability at least*

$$1 - (e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}})$$

all the vectors $\hat{x}_j = \Delta(y_j)$, $j = 1, \ldots, m_{\mathcal{X}}$ *calculated in Algorithm 1 satisfy*

$$\|x_j - \hat{x}_j\|_{\ell_2^d} \leq C \left(\left[\frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2 - 1/q} + \max\{\|\varepsilon_j\|_{\ell_2^{m_\Phi}}, \sqrt{\log d}\|\varepsilon_j\|_{\ell_\infty^{m_\Phi}}\} \right) \tag{30}$$

where C *depends only on* C_1 *and* C_2 *from* (8) *and* (9).

(ii) If furthermore $m_\Phi \geq \log d$ *holds, then with the same probability also*

$$\|x_j - \hat{x}_j\|_{\ell_2^d} \leq C' \left(\left[\frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2 - 1/q} + \frac{\epsilon}{\sqrt{m_\Phi}} \right) \tag{31}$$

where C' *depends again only on* C_1 *and* C_2 *from* (8) *and* (9).

Proof. We apply Theorem 3.2 to the equation $y_j = \Phi x_j + \varepsilon_j$ and $K \leq c_2' m_\Phi / \log(d/m_\Phi)$. To do so, we have to estimate the best K-term approximation error of $\sigma_K(x_j)_{\ell_1^d}$ and the size of the errors ε_j. We start by bounding $\sigma_K(x_j)_{\ell_1^d}$. Recall that due to the construction of X every column is a scaled copy of the vector a^T, i.e., $x_j = g'(a \cdot \xi_j) a^T$, so we have by (21)

$$K^{-1/2} \sigma_K(x_j)_{\ell_1^d} \leq |g'(a \cdot \xi_j)| \cdot \|a\|_{\ell_q^d} \cdot K^{1/2-1/q} \leq C_1 C_2 \left[\frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2-1/q}. \quad (32)$$

This finishes the proof of the first part.

To prove the second part, we estimate the size of the errors using (19),

$$\begin{aligned} \|\varepsilon_j\|_{\ell_\infty^{m_\Phi}} &= \frac{\epsilon}{2} \cdot \max_{i=1,\dots,m_\Phi} |\varphi_i^T \nabla^2 f(\zeta_{ij}) \varphi_i| \\ &= \frac{\epsilon}{2m_\Phi} \cdot \max_{i=1,\dots,m_\Phi} \left| \sum_{k,l=1}^{d} a_k a_l g''(a \cdot \zeta_{ij}) \right| \quad (33) \\ &\leq \frac{\epsilon \|g''\|_\infty}{2m_\Phi} \left(\sum_{k=1}^{d} |a_k| \right)^2 \leq \frac{\epsilon \|g''\|_\infty}{2m_\Phi} \left(\sum_{k=1}^{d} |a_k|^q \right)^{2/q} \leq \frac{C_1^2 C_2}{2m_\Phi} \epsilon, \\ \|\varepsilon_j\|_{\ell_2^{m_\Phi}} &\leq \sqrt{m_\Phi} \|\varepsilon_j\|_{\ell_\infty^{m_\Phi}} \leq \frac{C_1^2 C_2}{2\sqrt{m_\Phi}} \epsilon, \quad (34) \end{aligned}$$

leading to

$$\max\{\|\varepsilon_j\|_{\ell_2^{m_\Phi}}, \sqrt{\log d} \|\varepsilon_j\|_{\ell_\infty^{m_\Phi}}\} \leq \frac{C_1^2 C_2}{2\sqrt{m_\Phi}} \epsilon \cdot \max\left\{1, \sqrt{\frac{\log d}{m_\Phi}}\right\}.$$

Together with our assumption $m_\Phi \geq \log d$ this finishes the proof. □

Next we need a technical lemma to relate the error between the normalized version of $\hat{x}_j$ and a to the size of $\|\hat{x}_j\|_{\ell_2^d}$.

Lemma 3.4 (Stability of subspaces - one dimensional case)**.** *Let us fix $\hat{x} \in \mathbb{R}^d$, $a \in \mathbb{S}^{d-1}$, $0 \neq \gamma \in \mathbb{R}$, and $n \in \mathbb{R}^d$ with norm $\|n\|_{\ell_2^d} \leq \nu_1 < |\gamma|$. If we assume $\hat{x} = \gamma a + n$ then*

$$\left\| \operatorname{sign} \gamma \frac{\hat{x}}{\|\hat{x}\|_{\ell_2^d}} - a \right\|_{\ell_2^d} \leq \frac{2\nu_1}{\|\hat{x}\|_{\ell_2^d}}. \quad (35)$$

Proof. Applying the triangular inequality and its reverse form several times and using that $a \in \mathbb{S}^{d-1}$ we get

$$\begin{aligned} \left\| \operatorname{sign} \gamma \frac{\hat{x}}{\|\hat{x}\|_{\ell_2^d}} - a \right\|_{\ell_2^d} &\leq \left\| \operatorname{sign} \gamma \frac{\hat{x}}{\|\hat{x}\|_{\ell_2^d}} - \frac{|\gamma| a}{\|\hat{x}\|_{\ell_2^d}} \right\|_{\ell_2^d} + \left\| \frac{|\gamma| a}{\|\hat{x}\|_{\ell_2^d}} - a \right\|_{\ell_2^d} \\ &\leq \frac{\nu_1}{\|\hat{x}\|_{\ell_2^d}} + \left| \frac{|\gamma|}{\|\hat{x}\|_{\ell_2^d}} - 1 \right| \leq \frac{2\nu_1}{\|\hat{x}\|_{\ell_2^d}}. \end{aligned}$$

□

Applied to our situation where $\hat{x}_j = g'(a \cdot \xi_j)a^T + n_j$ we see that the bound in (35) is best for $\|\hat{x}_j\|_{\ell_2^d}$ maximal which justifies our definition of $\hat{a}$ in Algorithm 1.

As a last ingredient for the proof of Theorem 3.1 we need a lower bound for $\max_{j=1,\dots,m_{\mathcal{X}}} \|\hat{x}\|_{\ell_2^d}$. Since we have $\max_j \|\hat{x}_j\|_{\ell_2^d} \geq \max_j |g'(a \cdot \xi_j)| - \max_j \|\hat{x}_j - x_j\|_{\ell_2^d} \geq \max_j |g'(a \cdot \xi_j)| - \nu_1$ we just have to show that, with high probability, our random sampling of the gradient via the ξ_j provided a good maximum. To do this we will use Hoeffding's inequality, which we recall below for reader's convenience.

Proposition 3.5 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_m$ be independent random variables. Assume that the X_j are almost surely bounded, i.e., there exist finite scalars a_j, b_j such that*

$$\mathbb{P}\{X_j - \mathbb{E}X_j \in [a_j, b_j]\} = 1,$$

for $j = 1, \dots, m$. Then we have

$$\mathbb{P}\left\{\left|\sum_{j=1}^{m} X_j - \mathbb{E}\left(\sum_{j=1}^{m} X_j\right)\right| \geq t\right\} \leq 2e^{-\frac{2t^2}{\sum_{j=1}^{m}(b_j - a_j)^2}}.$$

Let us now apply Hoeffding's inequality to the random variables $X_j = |g'(a \cdot \xi_j)|^2$.

Lemma 3.6. *Let us fix $0 < s < 1$. Then with probability $1 - 2e^{-\frac{2m_{\mathcal{X}} s^2 \alpha^2}{C_2^4}}$ we have*

$$\max_{j=1,\dots,m_{\mathcal{X}}} |g'(a \cdot \xi_j)| \geq \sqrt{\alpha(1-s)},$$

where $\alpha := \mathbb{E}_\xi(|g'(a \cdot \xi_j)|^2)$.

Proof. By our assumptions (10) and (9) we have

$$\mathbb{E}X_j = \mathbb{E}_\xi(|g'(a \cdot \xi_j)|^2) = \int_{\mathbb{S}^{d-1}} |g'(a \cdot \xi)|^2 d\mu_{\mathbb{S}^{d-1}}(\xi) \geq \alpha > 0,$$

and

$$X_j - \mathbb{E}X_j \in [-\alpha, C_2^2 - \alpha].$$

Hence, by Hoeffding's inequality we have

$$\mathbb{P}\left\{\left|\sum_{j=1}^{m_{\mathcal{X}}} |g'(a \cdot \xi_j)|^2 - m_{\mathcal{X}}\alpha\right| \geq s m_{\mathcal{X}}\alpha\right\} \leq 2e^{-\frac{2m_{\mathcal{X}} s^2 \alpha^2}{C_2^4}}. \tag{36}$$

Using (36) we immediately obtain

$$\frac{1}{m_{\mathcal{X}}} \sum_{j=1}^{m_{\mathcal{X}}} |g'(a \cdot \xi_j)|^2 \geq \alpha(1-s), \tag{37}$$

with probability $1 - 2e^{-\frac{2m_{\mathcal{X}} s^2 \alpha^2}{C_2^4}}$. If $|g'(a \cdot \xi_j)|^2 < \alpha(1-s)$ for all $j = 1, \dots, m_{\mathcal{X}}$ then (37) would be violated. Hence for the maximum we have

$$\max_{j=1,\dots,m_{\mathcal{X}}} |g'(a \cdot \xi_j)| \geq \sqrt{\alpha(1-s)}.$$

□

Finally we have all the tools ready to prove Theorem 3.1.

Proof of Theorem 3.1:

Proof. Lemma 3.6 ensures that

$$|g'(a \cdot \xi_{j_0})| \geq \sqrt{\alpha(1-s)}$$

with probability $1 - 2e^{-\frac{2m_{\mathcal{X}} s^2 \alpha^2}{C_2^4}}$. Therefore, Corollary 3.3 together with Lemma 3.4 show that with probability at least

$$1 - \left(e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}} + 2e^{-\frac{2m_{\mathcal{X}} s^2 \alpha^2}{C_2^4}} \right),$$

$\hat{a}$ as defined in Algorithm 1 satisfies

$$\left\| \operatorname{sign}(g'(a \cdot \xi_{j_0}))\hat{a} - a \right\|_{\ell_2^d} \leq \frac{2\nu_1}{\sqrt{\alpha(1-s)} - \nu_1} \tag{38}$$

for the unknown sign of $g'(a \cdot \xi_{j_0})$.
Using this estimate we can prove that $\hat{f}$ as defined in Algorithm 1 is a good approximation to f. For $x \in B_{\mathbb{R}^d}(1+\bar{\epsilon})$ we have,

$$\begin{aligned} |f(x) - \hat{f}(x)| &= |g(a \cdot x) - \hat{g}(\hat{a} \cdot x)| \\ &= |g(a \cdot x) - f(\hat{a}^T \cdot \hat{a} \cdot x)| \\ &= |g(a \cdot x) - g(a \cdot \hat{a}^T \cdot \hat{a} \cdot x)| \\ &\leq C_2 |a \cdot x - a \cdot [\hat{a}^T \hat{a}] \cdot x| \\ &= C_2 |a \cdot (I_d - \hat{a}^T \hat{a}) x|. \end{aligned}$$

Because $\hat{a}(I_d - \hat{a}^T \hat{a}) = 0$ and therefore $\operatorname{sign}(g'(a \cdot \xi_{j_0}))\hat{a}(I_d - \hat{a}^T \hat{a}) = 0$, we can further estimate

$$\begin{aligned} |f(x) - \hat{f}(x)| &\leq C_2 |a \cdot (I_d - \hat{a}^T \hat{a}) x| \\ &= C_2 |(a - \operatorname{sign}(g'(a \cdot \xi_{j_0}))\hat{a}) \cdot (I_d - \hat{a}^T \hat{a}) x| \\ &\leq C_2 \|a - \operatorname{sign}(g'(a \cdot \xi_{j_0}))\hat{a}\|_{\ell_2^d} \cdot \|x\|_{\ell_2^d} \\ &\leq 2C_2(1+\bar{\epsilon}) \frac{\nu_1}{\sqrt{\alpha(1-s)} - \nu_1}. \end{aligned}$$

□

Remark 4. *We collect here a few comments about this result.*
(i) Our recovery method differs from the one proposed by Cohen, Daubechies, DeVore, Kerkyacharian, Picard [11]. In their approach, the domain is taken to be $[0,1]^d$ *and they make heavy use of the additional assumption* $\sum_{j=1}^d a_j = 1$ *and* $a_j \geq 0$*. This allows them to derive an almost completely deterministic and adaptive strategy for sampling the function* f *in order to find first an approximation to* g *and only then addressing the approximation to* a*. Here we follow somehow the opposite order, first approximating* a *and then finding a uniform approximation to* g *and, eventually, to* f *as well. Notice further that not having at disposal additional information on* a*, which is fully arbitrary in our case, we need to use a random sampling scheme which eventually gives a result holding with high probability.*
(ii) Note that Theorem 3.1 gives an a priori estimate of the success probability and approximation error of Algorithm 1. If the problem parameters q, C_1, C_2*, and* α *are known, they can be used to choose* m_Φ *and* $m_\mathcal{X}$ *big enough to have, say, a prescribed desired accuracy* δ *with probability at least* $1-p$*.*
However once Algorithm 1 has been run we have the following a posteriori *estimate. With probability at least* $1-(e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}})$ *we have that*

$$\|f - \hat{f}\|_\infty \leq C_2(1+\bar{\epsilon})\frac{2\nu_1}{\|x_{j_0}\|_{\ell_2^d}}.$$

Hence, the ratio $\frac{2\nu_1}{\|x_{j_0}\|_{\ell_2^d}} \ll 1$ *defines an a posteriori indicator that the number of samples* $m_\mathcal{X}$ *and* m_Φ *has been properly calibrated, otherwise just more points will be drawn until such a condition is obtained.*
(iii) The parameter ϵ *is chosen at the very beginning in the Taylor expansion* (14) *and, from a purely theoretical point of view, could be chosen arbitrarily small. Unfortunately, this may affect the numerical stability in the approximation in* (14) *of the derivative* $\frac{\partial f}{\partial \varphi}(\xi)$ *by means of a finite difference. Hence, the parameter* ϵ *should not be taken too small in practice. Up to some extent this may be compensated by choosing a larger number of points* m_Φ *in* (25)*, as in our expression for* ν_1 *in* (25) ϵ *appears in a ratio of the form* $\frac{\epsilon}{\sqrt{m_\Phi}}$*. We return in more detail to this point later in Section 5.1. In recent numerical experiments associated to the work [31], we have been experiencing very stable reconstructions with reasonable choices, e.g.,* $\epsilon \approx 0.1$*. Hence we do not consider this issue of any practical relevance or difficulty.*

3.3 Discussion on tractability

The approximation performances of our learning strategy are basically determined by the optimal value of α (see, e.g., (10)), which is achieved by the choice

$$\alpha := \int_{\mathbb{S}^{d-1}} |g'(a \cdot x)|^2 d\mu_{\mathbb{S}^{d-1}}(x). \tag{39}$$

Due to symmetry reasons this quantity does not depend on the particular choice of a.

The rotation invariant probability measure $\mu_{\mathbb{S}^{d-1}}$ on $\mathbb{S}^{d-1}$ is induced on the sphere by the (left) Haar measure on the Lie group of all orientation preserving rotations. For a given $k \times d$ matrix U such that $UU^T = I_k$ (i.e., with orthonormal rows) we define the measure μ_k on the unit ball $B_{\mathbb{R}^k}$ in $\mathbb{R}^k$ induced by the projection of $\mu_{\mathbb{S}^{d-1}}$ via U, i.e., for any Borel set $B \subset B_{\mathbb{R}^k}$ we define

$$\mu_k(B) := U_{\#}\mu_{\mathbb{S}^{d-1}}(B) := \mu_{\mathbb{S}^{d-1}}(U^{\leftarrow}(B)). \tag{40}$$

Since $\mu_{\mathbb{S}^{d-1}}$ is rotation invariant, μ_k does not depend on the particular matrix U, and is itself a rotation invariant measure on $B_{\mathbb{R}^k}$. Hence for any summable function $h : B_{\mathbb{R}^k} \to \mathbb{R}$, for any $k \times k$ orthogonal matrix $\mathcal{O}$ such that $\mathcal{O}\mathcal{O}^T = I_k = \mathcal{O}^T\mathcal{O}$, and for any $k \times d$ matrix U such that $UU^T = I_k$, we have the identities

$$\int_{B_{\mathbb{R}^k}} h(\mathcal{O}y)d\mu_k(y) = \int_{B_{\mathbb{R}^k}} h(y)d\mu_k(y) = \int_{\mathbb{S}^{d-1}} h(Ux)d\mu_{\mathbb{S}^{d-1}}(x). \tag{41}$$

The following result is well known. We refer to [30, Section 1.4.4] for the case of $\mathbb{C}^n$. The proof given there works literally also in the real case.

Theorem 3.7. *Let $1 \leq k < d$ be natural numbers. Then the measure μ_k defined in* (40) *is given by*

$$d\mu_k(y) = \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)}(1 - \|y\|^2_{\ell_2^k})^{\frac{d-2-k}{2}}dy.$$

Notice that as $d \to \infty$, and for fixed k, the measure μ_k becomes more and more concentrated around 0, in the sense that, for $\varepsilon > 0$ fixed

$$\mu_k(B_{\mathbb{R}^k}(\varepsilon)) \to 1, \text{ for } d \to \infty,$$

very rapidly (typically exponentially). By using the explicit form of the measure μ_k we can compute

$$\begin{aligned}
\mu_k(B_{\mathbb{R}^k}(\varepsilon)) &= 1 - \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)} \int_{B_{\mathbb{R}^k} \setminus B_{\mathbb{R}^k}(\varepsilon)} (1 - \|y\|^2_{\ell_2^k})^{\frac{d-2-k}{2}} dy \\
&= 1 - \frac{2\Gamma(d/2)}{\Gamma(k/2)\Gamma((d-k)/2)} \int_{\varepsilon}^{1} (1 - r^2)^{\frac{d-2-k}{2}} r^{k-1} dr \\
&\geq 1 - \frac{2\Gamma(d/2)}{\Gamma(k/2)\Gamma((d-k)/2)} e^{-\frac{d-2-k}{2}\varepsilon^2}.
\end{aligned}$$

By Stirling's approximation $\frac{2\Gamma(d/2)}{\Gamma(k/2)\Gamma((d-k)/2)} \approx \sqrt{\frac{d^{d-1}}{\pi k^{k-1}(d-k)^{d-k-1}}}$, thus for k and ε constant

$$\mu_k(B_{\mathbb{R}^k}(\varepsilon)) \to 1$$

exponentially fast as $d \to \infty$. For $k = 1$, this phenomenon can be summarized informally by saying that the surface measure of the unit sphere in high dimension is

concentrated around the equator [21]. Hence in case $d \gg k$ we may want to take into account possible rescaling, i.e., working with spheres of larger radii, in order to eventually consider properties of g (actually the matrix H_g) on larger subsets of $\mathbb{R}^k$, see also Remark 4. Without loss of generality, by keeping in mind this possible rescaling, we can therefore assume to work with the unit sphere.

For $k = 1$, we observe, that α as in (39) is determined by the interplay between the variation properties of g and the measure μ_1. As just mentioned above, the most relevant feature of μ_1 is that it concentrates around zero exponentially fast as $d \to \infty$. Hence, the asymptotic behavior of α exclusively depends on the behavior of the function g' in a neighborhood of 0.

To illustrate this phenomenon more precisely, we present the following result.

Proposition 3.8. *Let us fix $M \in \mathbb{N}$ and assume that $g : B_{\mathbb{R}} \to \mathbb{R}$ is C^{M+2}-differentiable in an open neighborhood $\mathcal{U}$ of 0 and $\frac{d^\ell}{dx^\ell} g(0) = 0$ for $\ell = 1, \dots, M$. Then*

$$\alpha(d) \;=\; \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \int_{-1}^{1} |g'(y)|^2 (1-y^2)^{\frac{d-3}{2}} dy = \mathcal{O}(d^{-M}), \text{ for } d \to \infty.$$

Proof. First of all, we compute the ℓ^{th} moment of the measure $\frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)}(1-y^2)^{\frac{d-3}{2}}\mathcal{L}^1$:

$$\frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \int_{-1}^{1} y^\ell (1-y^2)^{\frac{d-3}{2}} dy = \frac{[1+(-1)^\ell]\Gamma(d/2)\Gamma((1+\ell)/2)}{2\sqrt{\pi}\Gamma((d+\ell)/2)}. \tag{42}$$

Notice that all the odd moments vanish. By Taylor expansion of g' around 0 and by taking into account that $\frac{d^\ell}{dx^\ell} g(0) = 0$ for $\ell = 1, \dots, M$, we obtain

$$g'(y) = \sum_{\ell=1}^{M+1} \frac{1}{(\ell-1)!} \frac{d^\ell}{dx^\ell} g(0) y^{\ell-1} + \mathcal{O}(y^{M+1}) = \frac{1}{M!} \frac{d^{M+1}}{dx^{M+1}} g(0) y^M + \mathcal{O}(y^{M+1}).$$

Hence,

$$|g'(y)|^2 = \left(\frac{1}{M!} \frac{d^{M+1}}{dx^{M+1}} g(0) \right)^2 y^{2M} + \mathcal{O}(y^{2M+1}),$$

and

$$\begin{aligned}
\alpha(d) \;&=\; \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \int_{-1}^{1} |g'(y)|^2 (1-y^2)^{\frac{d-3}{2}} dy \\
&=\; \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \left(\int_{\mathcal{U}} |g'(y)|^2 (1-y^2)^{\frac{d-3}{2}} dy + \int_{B_{\mathbb{R}} \setminus \mathcal{U}} |g'(y)|^2 (1-y^2)^{\frac{d-3}{2}} dy \right) \\
&=\; \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \left(\left(\frac{1}{M!} \frac{d^{M+1}}{dx^{M+1}} g(0) \right)^2 \int_{\mathcal{U}} y^{2M} (1-y^2)^{\frac{d-3}{2}} dy \right. \\
&\qquad \left. + \int_{\mathcal{U}} \mathcal{O}(y^{2M+2}) (1-y^2)^{\frac{d-3}{2}} dy + \int_{B_{\mathbb{R}} \setminus \mathcal{U}} |g'(y)|^2 (1-y^2)^{\frac{d-3}{2}} dy \right).
\end{aligned}$$

Notice that we consider the $(2M+2)^{th}$ moment in the expression above because the previous one is odd and therefore vanishes. Now, the term $\int_{B_{\mathbb{R}}\setminus\mathcal{U}} |g'(y)|^2(1-y^2)^{\frac{d-3}{2}}dy$ goes to zero exponentially fast for $d\to 0$. By using (42) we immediately obtain

$$\begin{aligned}\alpha(d) &= \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)}\int_{-1}^{1}|g'(y)|^2(1-y^2)^{\frac{d-3}{2}}dy\\ &= \mathcal{O}\left(\frac{\Gamma(d/2)\Gamma((1+2M)/2)}{\Gamma((d+2M)/2)}\right), \quad d\to\infty.\end{aligned}$$

By Stirling's approximation, for which $\Gamma(z)=\sqrt{\frac{2\pi}{z}}\left(\frac{z}{e}\right)^z+\mathcal{O}(1+1/z)$, for $z\to\infty$, we obtain

$$\frac{\Gamma(d/2)\Gamma((1+2M)/2)}{\Gamma((d+2M)/2)} \approx d^{(d-1)/2}(1+2M)^M(d+2M)^{-(\frac{d+1}{2}+M)}, \quad d\to\infty.$$

This eventually yields

$$\alpha(d)=\frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)}\int_{-1}^{1}|g'(y)|^2(1-y^2)^{\frac{d-3}{2}}dy=\mathcal{O}\left(d^{-M}\right), \quad d\to\infty.$$

□

The number $m_{\mathcal{X}}\times(m_\Phi+1)$ of points we need in order to achieve a prescribed accuracy in the error estimate (24) of Theorem 3.1 depends on α. Proposition 3.8 ensures that, if $g'(y)$ does not vanish for $y\to 0$ super-polynomially, then the dependence of α (therefore of the error estimate and the number $m_{\mathcal{X}}\times(m_\Phi+1)$ of points) on d is at most polynomial. According to this observation we distinguish three classes of ridge functions:

(1) For $0<q\le 1$, $C_1>1$ and $C_2\ge\alpha_0>0$, we define

$$\begin{aligned}\mathcal{F}_d^1 &:= \mathcal{F}_d^1(\alpha_0,q,C_1,C_2):=\{f:B_{\mathbb{R}^d}\to\mathbb{R}:\\ &\quad \exists a\in\mathbb{R}^d, \|a\|_{\ell_2^d}=1, \|a\|_{\ell_q^d}\le C_1 \quad\text{and}\\ &\quad \exists g\in C^2(B_{\mathbb{R}}),\ |g'(0)|\ge\alpha_0>0 : f(x)=g(a\cdot x)\,\}.\end{aligned}$$

(2) For a neighborhood $\mathcal{U}$ of 0, $0<q\le 1$, $C_1>1$, $C_2\ge\alpha_0>0$ and $N\ge 2$, we define

$$\begin{aligned}\mathcal{F}_d^2 &:= \mathcal{F}_d^2(\mathcal{U},\alpha_0,q,C_1,C_2,N):=\{f:B_{\mathbb{R}^d}\to\mathbb{R}:\\ &\quad \exists a\in\mathbb{R}^d, \|a\|_{\ell_2^d}=1, \|a\|_{\ell_q^d}\le C_1 \quad\text{and}\quad \exists g\in C^2(B_{\mathbb{R}})\cap C^N(\mathcal{U})\\ &\quad \exists 0\le M\le N-1,\ |g^{(M)}(0)|\ge\alpha_0>0 : f(x)=g(a\cdot x)\,\}.\end{aligned}$$

(3) For a neighborhood $\mathcal{U}$ of 0, $0<q\le 1$, $C_1>1$ and $C_2\ge\alpha_0>0$, we define

$$\begin{aligned}\mathcal{F}_d^3 &:= \mathcal{F}_d^3(\mathcal{U},\alpha_0,q,C_1,C_2):=\{f:B_{\mathbb{R}^d}\to\mathbb{R}:\\ &\quad \exists a\in\mathbb{R}^d, \|a\|_{\ell_2^d}=1, \|a\|_{\ell_q^d}\le C_1 \quad\text{and}\quad \exists g\in C^2(B_{\mathbb{R}})\cap C^\infty(\mathcal{U})\\ &\quad |g^{(M)}(0)|=0 \quad\text{for all}\quad M\in\mathbb{N} : f(x)=g(a\cdot x)\,\}.\end{aligned}$$

Theorem 3.1 and Proposition 3.8 immediately imply the following tractability result for these function classes.

Corollary 3.9. *The problem of learning functions f in the classes $\mathcal{F}_d^1$ and $\mathcal{F}_d^2$ from point evaluations is* strongly polynomially tractable *and* polynomially tractable *respectively.*

Unfortunately, our learning method and approximation estimates in Theorem 3.1 do not provide any information about the tractability of the problem for functions in the class $\mathcal{F}_d^3$.

4 The General Case $k \geq 1$

In this section we generalize our approach to the case $k \geq 1$, i.e., we consider *k-ridge functions*

$$f(x) = g(Ax). \tag{43}$$

Obviously, the sum of k ridge functions (as appearing for example in (13)) is a k-ridge function and the same holds true also for the product.

We will proceed as in the one-dimensional case, giving first the basic ideas, which motivate the recovery algorithm and then stating and proving our main theorem. Remember that we assume, that A is a $k \times d$ matrix such that $AA^T = I_k$, and $g : B_{\mathbb{R}^k}(1+\bar{\epsilon}) \to \mathbb{R}$ is a C^2 function.

4.1 The Algorithm

As before we consider a version of Taylor's theorem giving access to the matrix A. For $\xi \in B_{\mathbb{R}^d}$, $\varphi \in B_{\mathbb{R}^d}(r)$, $\epsilon, r \in \mathbb{R}_+$, with $r\epsilon \leq \bar{\epsilon}$, we have the identity

$$[\nabla g(A\xi)^T A]\varphi \;=\; \frac{f(\xi+\epsilon\varphi)-f(\xi)}{\epsilon} - \frac{\epsilon}{2}[\varphi^T \nabla^2 f(\zeta)\varphi], \tag{44}$$

for a suitable $\zeta(\xi,\varphi) \in B_{\mathbb{R}^d}(1+\bar{\epsilon})$ and thanks to (9) the term $[\varphi^T \nabla^2 f(\zeta)\varphi]$ is again uniformly bounded as soon as φ is bounded.
As in the one-dimensional case we now consider (44) for the m_Φ directions in the set Φ and at the $m_\mathcal{X}$ sampling points in the set $\mathcal{X}$, where $\mathcal{X}, \Phi$ are defined as in (15) and (16) respectively. Again we collect the directional derivatives $\nabla g(A\xi_j)^T A$, $j = 1, \ldots, m_\mathcal{X}$ as columns in the $d \times m_\mathcal{X}$ matrix X, i.e.,

$$X = (A^T \nabla g(A\xi_1), \ldots, A^T \nabla g(A\xi_{m_\mathcal{X}})), \tag{45}$$

and using the matrices Y and $\mathcal{E}$ as defined in (18) and (19), we can write the following factorization

$$\Phi X = Y - \mathcal{E}. \tag{46}$$

Similarly to the one-dimensional case we find that the matrix X has a special structure, which we will exploit for the algorithm, i.e., $X = A^T \mathcal{G}^T$, where $\mathcal{G} =$

$(\nabla g(A\xi_1)^T|\ldots|\nabla g(A\xi_{m_\mathcal{X}})^T)^T$. The columns of X are now no longer scaled copies of one compressible vector but they are linear combinations of k compressible vectors, i.e., the rows of the matrix A. Thus compressed sensing theory again tells us that we can stably recover the columns of X from the columns of Y via ℓ_1-minimization and in consequence get a good approximation $\hat{X}$ to X.
Furthermore, since A has rank k, as long as $\mathcal{G}^T$ has full rank, also X will have rank k and moreover the column span of the right singular vectors of $X^T = USV^T$ will coincide with the row span of A, i.e., $A^TA = VV^T$. Moreover, V^T gives us an alternative representation of f as follows, i.e.,

$$f(x) = g(Ax) = g(AA^TAx) = g(AVV^Tx) =: \tilde{g}(V^Tx),$$

where $\tilde{g}(y) := g(AVy) = f(Vy)$. If $\hat{X}$ is a good approximation to X, then we can expect that the first k right singular vectors of $\hat{X}$ have almost the same span as those of X and thus of A, which inspires the following algorithm.

Algorithm 2:

- *Given $m_\Phi, m_\mathcal{X}$, draw at random the sets Φ and $\mathcal{X}$ as in* (15) *and* (16)*, and construct Y according to* (18)*.*
- *Set $\hat{x}_j = \Delta(y_j) := \arg\min_{y_j=\Phi z} \|z\|_{\ell_1^d}$, for $j = 1,\ldots,m_\mathcal{X}$, and $\hat{X} = (\hat{x}_1,\ldots,\hat{x}_{m_\mathcal{X}})$.*
- *Compute the singular value decomposition of*

$$\hat{X}^T = \begin{pmatrix} \hat{U}_1 & \hat{U}_2 \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \hat{V}_1^T \\ \hat{V}_2^T \end{pmatrix}, \tag{47}$$

  where $\hat{\Sigma}_1$ contains the k largest singular values.
- *Set $\hat{A} = V_1^T$.*
- *Define $\hat{g}(y) := f(\hat{A}^Ty)$ and $\hat{f}(x) := \hat{g}(\hat{A}x)$.*

The quality of the final approximation of f by means of $\hat{f}$ depends on two kinds of accuracies:

1. The error between $\hat{X}$ and X, which can be controlled through the number of compressed sensing measurements m_Φ;
2. The stability of the span of V^T, simply characterized by how well the singular values of X or equivalently $\mathcal{G}$ are separated from 0, which is related to the number of random samples $m_\mathcal{X}$.

To be precise, in the next section we will prove the following approximation result.

Theorem 4.1. *Let* $\log d \le m_\Phi \le [\log 6]^2 d$. *Then there is a constant* c_1' *such that using* $m_\mathcal{X} \cdot (m_\Phi + 1)$ *function evaluations of* f, *Algorithm 2 defines a function* $\hat{f} : B_{\mathbb{R}^d}(1+\bar{\epsilon}) \to \mathbb{R}$ *that, with probability*

$$1 - \left(e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}} + k e^{\frac{-m_\mathcal{X} \alpha s^2}{2kC_2^2}} \right), \tag{48}$$

will satisfy

$$\|f - \hat{f}\|_\infty \le 2C_2 \sqrt{k}(1+\bar{\epsilon}) \frac{\nu_2}{\sqrt{\alpha(1-s)} - \nu_2}, \tag{49}$$

where

$$\nu_2 = C \left(k^{1/q} \left[\frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2-1/q} + \frac{\epsilon k^2}{\sqrt{m_\Phi}} \right),$$

and C *depends only on* C_1 *and* C_2 *(cf.* (8) *and* (9)*).*

4.2 The Analysis

We will first show that $\hat{X}$ is a good approximation to X by applying Theorem 3.2 columnwise. This leads to the following corollary.

Corollary 4.2. *Let* $\log d \le m_\Phi < [\log 6]^2 d$. *Then with probability*

$$1 - (e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}})$$

the matrix $\hat{X}$ *as calculated in Algorithm 2 satisfies*

$$\|X - \hat{X}\|_F \le C\sqrt{m_\mathcal{X}} \left(k^{1/q} \left[\frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2-1/q} + \frac{\epsilon k^2}{\sqrt{m_\Phi}} \right), \tag{50}$$

where C *depends only on* C_1 *and* C_2 *(cf.* (8) *and* (9)*).*

Proof. The proof works essentially like that of Corollary 3.3. We decompose

$$\|X - \hat{X}\|_F^2 = \sum_{j=1}^{m_\mathcal{X}} \|x_j - \hat{x}_j\|_{\ell_2^d}^2.$$

The best K-term approximation of x_j may be estimated using

$$\|x_j\|_{\ell_q^d} = \|A^T \nabla g(A\xi_j)\|_{\ell_q^d} \le C_2 \left(\sum_{v=1}^{d} \left(\sum_{u=1}^{k} |a_{uv}| \right)^q \right)^{1/q} \le C_1 C_2 k^{1/q},$$

which leads to

$$K^{-1/2} \sigma_K(x_j)_{\ell_1^d} \le \|x_j\|_{\ell_q^d} K^{1/2-1/q} \le C_1 C_2 k^{1/q} K^{1/2-1/q}.$$

The norms of ε_j may be estimated similarly to the proof of Corollary 3.3 as

$$\|\varepsilon_j\|_{\ell_2^{m_\Phi}} \leq \frac{C_1^2\, C_2\, k^2 \epsilon}{2\sqrt{m_\Phi}} \quad \text{and} \quad \|\varepsilon_j\|_{\ell_\infty^{m_\Phi}} \leq \frac{C_1^2\, C_2\, k^2 \epsilon}{2m_\Phi}.$$

Putting all these estimates (with the choice $K \approx m_\Phi / \log(d/m_\Phi)$) into Theorem 3.2 we obtain the result. □

Remark 5. *The construction* $\hat{x}_j = \Delta(y_j) := \arg\min_{y_j = \Phi z} \|z\|_{\ell_1^d}$, *for* $j = 1, \ldots, m_{\mathcal{X}}$, *and* $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_{m_{\mathcal{X}}})$ *and Corollary 4.2 are not the unique possible approach to approximate* X. *As we are expecting* X *to be a* k*-rank matrix for* $k \ll \min\{d, m_{\mathcal{X}}\}$, *one might want to consider also* nuclear norm minimization, *i.e., the minimization of the* ℓ_1*-norm of singular values, as a possible way of accessing* X *from* m_Φ *random measurements, as in the work [15, 26, 28]. However, presently no estimates of the type* (29) *are available in this context, hence we postpone an analysis based on these methods fully tailored to matrices to further research.*

Next we need the equivalent of Lemma 3.4 to relate the error between the subspaces defined by the largest right singular values of $\hat{X}$ and X respectively to the error $\|X - \hat{X}\|_F$. We will develop the necessary tools in the following subsection.

4.2.1 Stability of the singular value decomposition

Given two matrices B and $\hat{B}$ with corresponding singular value decompositions

$$B = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix}$$

and

$$\hat{B} = \begin{pmatrix} \hat{U}_1 & \hat{U}_2 \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \hat{V}_1^T \\ \hat{V}_2^T \end{pmatrix},$$

where it is understood that two corresponding submatrices, e.g., $U_1, \hat{U}_1$, have the same size, we would like to bound the difference between V_1 and $\hat{V}_1$ by the error $\|B - \hat{B}\|_F$. As a consequence of Wedin's perturbation bound [34], see also [32, Section 7], we have the following useful result.

Theorem 4.3 (Stability of subspaces - Wedin's bound)**.** *If there is an* $\bar{\alpha} > 0$ *such that*

$$\min_{\ell, \hat{\ell}} |\sigma_{\hat{\ell}}(\hat{\Sigma}_1) - \sigma_\ell(\Sigma_2)| \geq \bar{\alpha}, \tag{51}$$

and

$$\min_{\hat{\ell}} |\sigma_{\hat{\ell}}(\hat{\Sigma}_1)| \geq \bar{\alpha}, \tag{52}$$

then

$$\|V_1 V_1^T - \hat{V}_1 \hat{V}_1^T\|_F \leq \frac{2}{\bar{\alpha}} \|B - \hat{B}\|_F. \tag{53}$$

The conditions (51) and (52) are separation conditions. The first says that the singular values of Σ_1 are separated from those of Σ_2. Actually, strictly speaking the separation is between Σ_1 and $\hat{\Sigma}_2$. However, if $\|B - \hat{B}\|_F$ is sufficiently small compared to $\bar{\alpha}$, then Weyl's inequality [35]

$$|\sigma_\ell(B) - \sigma_\ell(\hat{B})| \leq \|B - \hat{B}\|_F,$$

guarantees that the two separations are essentially equivalent. The second condition says that the singular values of Σ_1 or $\hat{\Sigma}_1$ have to be far away from 0.
Applied to our situation, where X has rank k and thus $\Sigma_2 = 0$, we get

$$\|V_1 V_1^T - \hat{V}_1 \hat{V}_1^T\|_F \leq \frac{2\sqrt{m_{\mathcal{X}}}\nu_2}{\sigma_k(\hat{X}^T)}, \tag{54}$$

and further since $\sigma_k(\hat{X}^T) \geq \sigma_k(X^T) - \|X - \hat{X}\|_F$, that

$$\|V_1 V_1^T - \hat{V}_1 \hat{V}_1^T\|_F \leq \frac{2\sqrt{m_{\mathcal{X}}}\nu_2}{\sigma_k(X^T) - \sqrt{m_{\mathcal{X}}}\nu_2}. \tag{55}$$

As final ingredient we need to estimate the k-th singular value of X. The next subsection will provide us with a generalization of Hoeffding's inequality, that can be used to show that with high probability on the random draw of the sampling points ξ_j the k-th singular value of X is separated from zero.

4.2.2 Spectral estimates and sums of random semidefinite matrices

The following theorem generalizes Hoeffding's inequality to sums of random semidefinite matrices and was recently proved by Tropp in [33, Corollary 5.2 and Remark 5.3], improving over results in [1], and using techniques from [29] and [25].

Theorem 4.4 (Matrix Chernoff)**.** *Consider $X_1, \ldots, X_m$ independent random, positive-semidefinite matrices of dimension $k \times k$. Moreover suppose*

$$\sigma_1(X_j) \leq C, \tag{56}$$

almost surely. Compute the singular values of the sum of the expectations

$$\mu_{\max} = \sigma_1\left(\sum_{j=1}^m \mathbb{E}X_j\right) \text{ and } \mu_{\min} = \sigma_k\left(\sum_{j=1}^m \mathbb{E}X_j\right), \tag{57}$$

then

$$\mathbb{P}\left\{\sigma_1\left(\sum_{j=1}^m X_j\right) - \mu_{\max} \geq s\mu_{\max}\right\} \leq k\left(\frac{(1+s)}{e}\right)^{-\frac{\mu_{\max}(1+s)}{C}}, \tag{58}$$

for all $s > (e-1)$, and

$$\mathbb{P}\left\{\sigma_k\left(\sum_{j=1}^m X_j\right) - \mu_{\min} \leq -s\mu_{\min}\right\} \leq ke^{-\frac{\mu_{\min} s^2}{2C}}, \tag{59}$$

for all $s \in (0,1)$.

Applied to the matrix X^T the above theorem leads to the following estimate of the singular values of X^T.

Lemma 4.5. *For any $s \in (0,1)$ we have that*

$$\sigma_k(X^T) \geq \sqrt{m_{\mathcal{X}}\alpha(1-s)} \tag{60}$$

with probability $1 - ke^{\frac{-m_{\mathcal{X}}\alpha s^2}{2kC_2^2}}$.

Proof. The proof is based on an application of Theorem 4.4. First of all note that

$$X^T = \mathcal{G}A = U_{\mathcal{G}}\Sigma_{\mathcal{G}}[V_{\mathcal{G}}^T A],$$

hence $\Sigma_{X^T} = \Sigma_{\mathcal{G}}$. Moreover

$$\sigma_i(\mathcal{G}) = \sqrt{\sigma_i(\mathcal{G}^T\mathcal{G})}, \quad \text{for all } i = 1, \dots, k.$$

Thus, to get information about the singular values of X^T it is sufficient to study that of

$$\mathcal{G}^T\mathcal{G} = \sum_{j=1}^{m_{\mathcal{X}}} \nabla g(A\xi_j)\nabla g(A\xi_j)^T.$$

We further notice that

$$\sigma_1(\nabla g(A\xi_j)\nabla g(A\xi_j)^T) \leq \left(\sum_{\ell,\ell'=1}^{k} |\nabla g(A\xi_j)_\ell \nabla g(A\xi_j)_{\ell'}|^2\right)^{1/2} \leq kC_2^2 := C.$$

Hence $X_j = \nabla g(A\xi_j)\nabla g(A\xi_j)^T$ is a random positive-semidefinite matrix, that is almost surely bounded. Moreover

$$\mathbb{E}X_j = \mathbb{E}_\xi \nabla g(A\xi_j)\nabla g(A\xi_j)^T = \int_{\mathbb{S}^{d-1}} \nabla g(Ax)\nabla g(Ax)^T d\mu_{\mathbb{S}^{d-1}}(x) = H_g.$$

Hence, remembering that the singular values of H_g are equivalent to that of H^f, by condition (10) we have $\mu_{\max} = m_{\mathcal{X}}\sigma_1(H_g) \leq m_{\mathcal{X}}kC_2^2$ and $\mu_{\min} = m_{\mathcal{X}}\sigma_k(H_g) \geq m_{\mathcal{X}}\alpha > 0$. In particular

$$m_{\mathcal{X}}k^2C_2 \geq \mu_{\max} \geq \mu_{\min} \geq m_{\mathcal{X}}\alpha > 0.$$

By an application of Theorem 4.4 we conclude that

$$\sigma_k(X^T) = \sigma_k(\mathcal{G}) = \sqrt{\sigma_k\left(\sum_{j=1}^{m_{\mathcal{X}}} \nabla g(A\xi_j)\nabla g(A\xi_j)^T\right)} \geq \sqrt{\mu_{\min}(1-s)} \geq \sqrt{m_{\mathcal{X}}\alpha(1-s)},$$

with probability

$$1 - ke^{-\frac{\mu_{\min}s^2}{2kC_2^2}} \geq 1 - ke^{\frac{-m_{\mathcal{X}}\alpha s^2}{2kC_2^2}},$$

for all $s \in (0,1)$. ∎

Finally we have collected all the results necessary to prove Theorem 4.1.

Proof of Theorem 4.1:

Proof. Combining Corollary 4.2, Theorem 4.3, and Lemma 4.5 shows that with probability at least

$$1 - \left(e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}} + k e^{\frac{-m_{\mathcal{X}} \alpha s^2}{2kC_2^2}} \right),$$

for the first k right singular vectors of $\hat{X}$ and X we have

$$\|V_1 V_1^T - \hat{V}_1 \hat{V}_1^T\|_F \leq \frac{2\nu_2}{\sqrt{\alpha(1-s)} - \nu_2}.$$

Recalling from the proof of Lemma 4.5 that the (first k) right singular vectors V_1^T of X^T have the form $V_1^T = V_{\mathcal{G}}^T A$ then shows that $\hat{A}$ as defined in Algorithm 2 satisfies

$$\|A^T A - \hat{A}^T \hat{A}\|_F = \|A^T V_{\mathcal{G}} V_{\mathcal{G}}^T A - \hat{V}_1 \hat{V}_1^T\|_F = \|V_1 V_1^T - \hat{V}_1 \hat{V}_1^T\|_F \leq \frac{2\nu_2}{\sqrt{\alpha(1-s)} - \nu_2},$$

Using this estimate we can prove that $\hat{f}$ as defined in Algorithm 2 is a good approximation to f. Since A is row-orthogonal we have $A = AA^T A$ and therefore

$$\begin{aligned}
|f(x) - \hat{f}(x)| &= |g(Ax) - \hat{g}(\hat{A}x)| \\
&= |g(Ax) - g(A\hat{A}^T \hat{A}x)| \\
&\leq C_2 \sqrt{k} \|Ax - A\hat{A}^T \hat{A}x\|_{\ell_2^k} \\
&= C_2 \sqrt{k} \|A(A^T A - \hat{A}^T \hat{A})x\|_{\ell_2^k} \\
&\leq C_2 \sqrt{k} \|(A^T A - \hat{A}^T \hat{A})\|_F \|x\|_{\ell_2^d} \\
&\leq 2C_2 \sqrt{k}(1+\bar{\epsilon}) \frac{\nu_2}{\sqrt{\alpha(1-s)} - \nu_2}.
\end{aligned}$$

□

Remark 6. *(i) Note that Theorem 4.1 is again an a priori estimate of the success probability and approximation error of Algorithm 2. Once Algorithm 2 has been run we have the following a posteriori estimate. With probability at least* $1 - (e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}})$ *we have that*

$$\|f - \hat{f}\|_\infty \leq 2C_2 \sqrt{k m_{\mathcal{X}}}(1+\bar{\epsilon}) \frac{\nu_2}{\sigma_k(\hat{X}^T)}.$$

(ii) We further observe that Theorem 4.1 does not straightforwardly reduce to Theorem 3.1 for $k = 1$*, because in the one-dimensional case we used the simpler maximum strategy as in* (22) *instead of the singular value decomposition* (47)*.*

4.3 Discussion on tractability

Recall, that the push-forward measure $\mu_k = \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)}(1-\|y\|_{\ell_2^k}^2)^{\frac{d-2-k}{2}}\mathcal{L}^k$ of $\mu_{\mathbb{S}^{d-1}}$ on the unit ball $B_{\mathbb{R}^k}$ was determined in Theorem 3.7 as the measure, for which

$$\begin{aligned} H_g &= \int_{\mathbb{S}^{d-1}} \nabla g(Ax)\nabla g(Ax)^T d\mu_{\mathbb{S}^{d-1}}(x) \\ &= \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)} \int_{B_{\mathbb{R}^k}} \nabla g(y)\nabla g(y)^T (1-\|y\|_{\ell_2^k}^2)^{\frac{d-2-k}{2}} dy. \end{aligned}$$

As an instructive example, let us apply this formula to the case when g is a radial function, i.e.,

$$g(y) = g_0(\|y\|_{\ell_2^k}),$$

for a function $g_0 : [0,1] \to \mathbb{R}$ sufficiently smooth, and $g_0'(0) = 0$.

A direct calculation shows, that $\nabla g(y) = \frac{g_0'(r)}{r} \cdot y$, where $r = \|y\|_{\ell_2^k}$, and

$$\nabla g(y)\nabla g(y)^T = \frac{g_0'(r)^2}{r^2} yy^T.$$

Hence,

$$(H_g)_{ij} = \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)} \int_{B_{\mathbb{R}^k}} \frac{g_0'(\|y\|_{\ell_2^k})^2}{\|y\|_{\ell_2^k}^2} y_i y_j (1-\|y\|_{\ell_2^k}^2)^{\frac{d-2-k}{2}} dy.$$

If $i \neq j$, the integral vanishes due to the symmetry of $B_{\mathbb{R}^k}$. If $i = j$, we get again by symmetry

$$\begin{aligned} (H_g)_{ii} &= \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)} \int_{B_{\mathbb{R}^k}} \frac{g_0'(\|y\|_{\ell_2^k})^2}{\|y\|_{\ell_2^k}^2} y_i^2 (1-\|y\|_{\ell_2^k}^2)^{\frac{d-2-k}{2}} dy \\ &= \frac{\Gamma(d/2)}{k\pi^{k/2}\Gamma((d-k)/2)} \int_{B_{\mathbb{R}^k}} g_0'(\|y\|_{\ell_2^k})^2 (1-\|y\|_{\ell_2^k}^2)^{\frac{d-2-k}{2}} dy \\ &= \frac{2\Gamma(d/2)}{k\Gamma((d-k)/2)\Gamma(k/2)} \int_0^1 g_0'(r)^2 (1-r^2)^{\frac{d-2-k}{2}} r^{k-1} dr =: \alpha(k,d). \end{aligned}$$

Hence, $H_g = \alpha(k,d) I_k$. Similarly to Proposition 3.8, we can expand g_0' into a Taylor series

$$g_0'(r) = \sum_{\ell=2}^{N-1} \frac{g_0^{(\ell)}(0)}{(\ell-1)!} r^{\ell-1} + \mathcal{O}(r^N).$$

If we assume that $g_0^{(\ell)}(0) = 0$, for all $\ell = 1, \dots, M$, but $g_0^{(M+1)}(0) \neq 0$, then we obtain

$$g_0'(r)^2 = \left(\frac{g_0^{(M+1)}(0)}{M!}\right)^2 r^{2M} + \mathcal{O}(r^{2M+1}),$$

and, by Stirling's approximation,

$$\begin{aligned}\alpha(k,d) &= \mathcal{O}\left(\frac{\Gamma(d/2)}{\Gamma((d-k)/2)}\int_0^1 r^{2M+k-1}(1-r^2)^{\frac{d-k-2}{2}}dr\right)\\ &= \mathcal{O}\left(\frac{\Gamma(d/2)}{\Gamma(d/2+M)}\right)\\ &= \mathcal{O}\left(d^{-M}\right), \quad d\to\infty.\end{aligned}$$

From these computations, we deduce that learning functions $f(x) = g(Ax)$, where g is radial (or nearly radial), using our method has usually polynomial complexity with respect to the dimension d.

5 Extensions and Generalizations

We assumed throughout the paper that the function f is defined on the unit ball $B_{\mathbb{R}^d}$ of $\mathbb{R}^d$. To be able to approximate the derivatives of f even on the boundary of $B_{\mathbb{R}^d}$, we actually supposed, that f is defined also on an $\bar{\epsilon}$ neighborhood of the unit ball. Furthermore, we assumed that the function values may be measured exactly without any error. The main aim of this section is to discuss the possibilities and limitations of our method. Firstly, we discuss the numerical stability of our approach with respect to noise. Secondly, we deal with functions defined on a convex body $\Omega \subset \mathbb{R}^d$. As it is our intention here only to sketch, still rigorously, further interesting research directions, we limit our discussion to the case of $k = 1$.

5.1 Stability under noisy measurements

Let us assume that the function evaluation in (14) can be performed only with certain precision. We again collect the $m_{\mathcal{X}} \times m_\Phi$ instances of (14) as

$$\Phi X = Y - \mathcal{E} + \frac{\mathcal{W}}{\epsilon}, \tag{61}$$

where the (i,j) entry of $\mathcal{W}$ (denoted by w_{ij}) is the difference between the exact value of $f(\xi_j + \epsilon\varphi_i) - f(\xi_j)$ and its value measured with noise. This leads to a compressed sensing setting

$$Y = \Phi X + \mathcal{E} - \frac{\mathcal{W}}{\epsilon}. \tag{62}$$

Applying Theorem 3.2 we obtain a substitute for Corollary 3.3 with $\mathcal{E}$ replaced by $\mathcal{E} - \mathcal{W}/\epsilon$. Therefore we would like to estimate the norm of w_j (the j-th column of $\mathcal{W}$) in $\ell_2^{m_\Phi}$ and $\ell_\infty^{m_\Phi}$. If we merely assume that the noise is bounded (i.e. $|w_{ij}| \le \nu$), the best possible estimate is $\|w_j\|_{\ell_2^{m_\Phi}} \le \nu\sqrt{m_\Phi}$. We observe that the more sampling points we take the greater is the level of noise. This effect of amplification of the noise is actually known under the name of *noise folding* [2] and, unfortunately, corrupts the estimate (31), see also [11, Section 4] for a discussion in a related context.

Let us therefore sketch a different approach. We make the rather natural assumption that w_{ij} is a random noise.

The analogue of Theorem 3.2 for the recovery of x from noisy measurements $y = \Phi x + \omega$, where $\omega = (\omega_1, \dots, \omega_m)$ are independent identically distributed (i.i.d.) Gaussian variables with mean zero and variance σ^2, was given in the work of Candès and Tao [9]. They proposed a certain ℓ_1-regularization problem, whose solution (called the *Dantzig selector*) satisfies

$$\|x - \hat{x}\|_{\ell_2^d}^2 \leq C^2 \cdot 2 \log d \cdot \left(\sigma^2 + \sum_{i=1}^{d} \min(x_i^2, \sigma^2) \right).$$

Especially, if x is a k-sparse vector, then $\|x - \hat{x}\|_{\ell_2^d} \leq C \cdot \sqrt{2 \log d} \cdot \sqrt{k+1} \cdot \sigma$. We observe that this estimate scales very favorably with d (only as $\sqrt{\log d}$) and, moreover, does depend only on the sparsity of x, and not anymore on the number of measurements m_Φ. Therefore, there is no noise folding in this case.

The equation (62) requires a combination of Theorem 3.2 and the result of Candès and Tao. Namely, we would like to reconstruct x if $y = \Phi x + \varepsilon + \omega$ is given, where ε is a deterministic error and ω is a vector of i.i.d. Gaussian variables. Obviously, the detailed analysis of this issue goes beyond the scope of this paper. Nevertheless, let us present some numerical evidence of the numerical stability of our approach in the presence of random noise.

We consider the function

$$f(x) = \max\left(\left[1 - 5\sqrt{(x_3 - 1/2)^2 + (x_4 - 1/2)^2} \right]^3, 0 \right), \quad x \in \mathbb{R}^{1000} \tag{63}$$

in dimension $d = 1000$. We use a variant of Algorithm 1 based on ℓ_1 minimization to identify the active coordinates of f, cf. [31] for details. We suppose that function evaluations were distorted by Gaussian error $\nu\omega$ with $\omega \approx \mathcal{N}(0,1)$ and $\nu \in \{0.1, 0.01, 0.001\}$. We chose $\epsilon = 0.1$ in the approximation (14). For each number of points $m_{\mathcal{X}} \in \{6\ell, \ell = 1, \dots, 10\}$ (x-axis) and each number of directions $m_\Phi \in \{20\ell, \ell = 1, \dots, 10\}$ (y-axis) we produced one hundred trials. The success rates of recovery go from white color (no success) to black (100 successful recoveries).

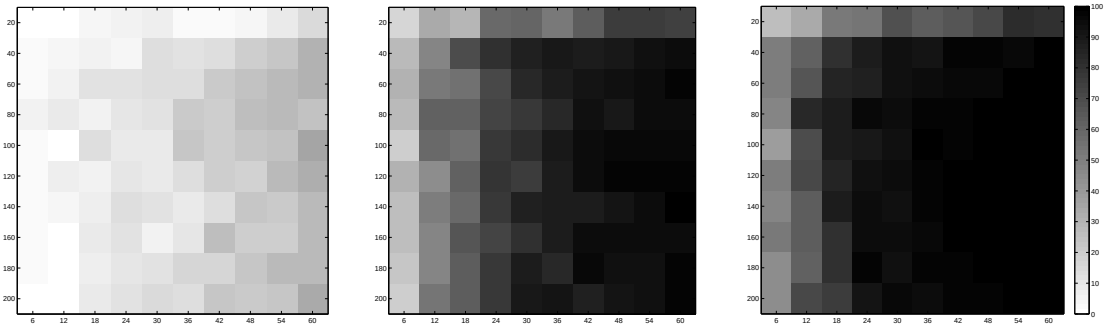

Figure 2: Recovery of active coordinates of $f(x)$ given by (63) with $\nu = 0.1$, $\nu = 0.01$

and $\nu = 0.001$, from left to right respectively. Let us mention that the success rates of recovery for noise-free setting are hardly distinguishable from the last picture above ($\nu = 0.001$).

We conclude from Figure 1 that there is a smooth increase of the rate of successful recovery with decreasing noise power and a fully stable recovery behavior.

5.2 Convex bodies

A careful inspection of our method shows, that it may be generalized to arbitrary convex bodies. Let us describe the necessary modifications and give an overview of the results for the case $k = 1$. First of all, one has to replace (6) by

$$H^f := \int_\Omega \nabla f(x) \nabla f(x)^T d\mu_\Omega(x). \tag{64}$$

Here, μ_Ω is a probability measure on Ω and the points in $\mathcal{X}$ (cf. (15)) are selected at random with respect to μ_Ω. For $\Omega = B_{\mathbb{R}^d}$, we simply selected $\mu_\Omega = \mu_{\mathbb{S}^{d-1}}$ to be the normalized surface measure on $\mathbb{S}^{d-1}$. This corresponded to the fact, that $a \in \mathbb{S}^{d-1}$ was arbitrary and therefore a-priori no direction was preferred. To be able to evaluate the derivatives of f even on the boundary of Ω, we suppose, that f is actually defined on an $\bar{\epsilon}$ neighborhood of Ω, namely on the set $\Omega + \bar{\epsilon} := \{x \in \mathbb{R}^d : \mathrm{dist}(\Omega, x) \leq \bar{\epsilon}\}$. The function g is supposed to be defined on the image of $\Omega + \bar{\epsilon}$ under the mapping $x \to a \cdot x$, i.e., on an interval. We assume again (9).
Surprisingly enough, these are all the modifications necessary to proceed with the identification of $\hat{a}$ and (38) holds true under these circumstances.

The proof of Theorem 3.1 was based on the fact, that for every $y \in B_{\mathbb{R}}$, we can easily find an element $x_y \in B_{\mathbb{R}^d}$, such that $\hat{a} \cdot x_y = y$. It is enough to consider $x_y = \hat{a}^T y$. In the case of a general convex set Ω, we first need to define for any $\hat{a} \in \mathbb{S}^{d-1}$ fixed, a function $x_\cdot : \hat{a}(\Omega + \bar{\epsilon}) \to \Omega + \bar{\epsilon}$ given by $y \mapsto x_y$, and such that

$$\hat{a} \cdot x_y = y.$$

In particular, for all $y \in \hat{a}(\Omega + \bar{\epsilon})$ we need to find

$$x_y \in \Omega + \bar{\epsilon} \cap \{x \in \mathbb{R}^d : \hat{a} \cdot x = y\}.$$

Since both $\Omega + \bar{\epsilon}$ and the solution space $\{x \in \mathbb{R}^d : \hat{a} \cdot x = y\}$ are closed convex sets in $\mathbb{R}^d$, one could use an alternating projection algorithm for finding x_y [4]. Thus, we can assume that, at least algorithmically, this map can be computed. Moreover, and alternatively, since the operation described above, i.e., finding $x_y \in B_{\mathbb{R}^d}$, such that $\hat{a} \cdot x_y = y$, has to be executed as many times as we need to define, e.g., an appropriate spline approximation of $\hat{g}$, we may proceed as follows: we find first $x_{\max}, x_{\min} \in B_{\mathbb{R}^d}$, such that $\hat{a} \cdot x_{\max} = \max_{x \in B_{\mathbb{R}^d}} \hat{a} \cdot x$ and $\hat{a} \cdot x_{\min} = \min_{x \in B_{\mathbb{R}^d}} \hat{a} \cdot x$. Then any other x_y

such that $y = \hat{a} \cdot x_y$ is computed very fast by $x_y = \lambda_y x_{\min} + (1 - \lambda_y) x_{\max}$ for some $\lambda_y \in [0, 1]$.

With this modification, also Theorem 3.1 holds true, with the definition of $\hat{g}$ given in Algorithm 1 replaced now by

$$\hat{g}(y) := f(x_y), \quad y \in \hat{a}(\Omega + \bar{\epsilon})$$

and (24) replaced by

$$\|f - \hat{f}\|_\infty \leq 2C_2(\mathrm{diam}(\Omega) + 2\bar{\epsilon}) \frac{\nu_1}{\sqrt{\alpha(1-s)} - \nu_1}.$$

Unfortunately, and this seems to be the main drawback of this approach, the diameter of Ω, $\mathrm{diam}(\Omega) = \max_{x,x' \in \Omega} \|x - x'\|_{\ell_2^d}$ may grow with d. This is especially the case, when $\Omega = [-1, 1]^d$, which gives $\mathrm{diam}(\Omega) = \sqrt{2d}$.

5.3 An approach through Minkowski functional

To get better results for specific convex bodies (i.e. $\Omega = [-1, 1]^d$), we propose another approach. We stress very clearly that up to now this is only to be understood as an open direction, which is a subject of further research.
We assume, that Ω is a closed convex set, which is *absorbing* and *balanced*, i.e.

- for every $x \in \mathbb{R}^d$, there is a $t = t(x) > 0$, such that $tx \in \Omega$,
- $\alpha\Omega := \{\alpha x : x \in \Omega\} \subset \Omega$ for every $\alpha \in [-1, 1]$.

Then we can define its Minkowski functional as

$$p_\Omega(x) := \inf\{r > 0 : x/r \in \Omega\}, \quad x \in \mathbb{R}^d.$$

It is well known, that this expression is actually a norm and Ω is its unit ball. Hence

$$\sup_{x,x' \in \Omega} p_\Omega(x - x') \leq 2. \tag{65}$$

This allows us to replace the inequality

$$|(a - \hat{a}) \cdot (x_y - x)| \leq \|a - \hat{a}\|_2 \cdot \|x_y - x\|_2$$

by

$$|(a - \hat{a}) \cdot (x_y - x)| \leq \|a - \hat{a}\|'_\Omega \cdot \|x_y - x\|_\Omega.$$

Here, $\|\cdot\|_\Omega = p_\Omega(\cdot)$ and $\|\cdot\|'_\Omega$ is its dual norm. According to (65), this solves the problem of the factor $\mathrm{diam}(\Omega)$ - the diameter of Ω with respect to $\|\cdot\|_\Omega$ is always bounded by 2. Unfortunately the problem is transferred to the second factor, namely $\|a - \hat{a}\|'_\Omega$. For this, one would need the analogue of Theorem 3.2 with the ℓ_2^d-norm in (29) replaced by $\|\cdot\|'_\Omega$. While any treatment of this general case is clearly beyond the scope of this paper and remains a subject of further investigation, we can shortly sketch what happens in the special case $\Omega = [-1, 1]^d$. Then we simply have $\|\cdot\|_\Omega = \|\cdot\|_{\ell_\infty^d}$ and $\|\cdot\|'_\Omega = \|\cdot\|_{\ell_1^d}$. To estimate $\|a - \hat{a}\|_{\ell_1^d}$ we would have to combine Lemma 3.1 in [11] with (38) and would get again a result that does not depend on the dimension d.

Acknowledgments

Massimo Fornasier would like to thank Ronald A. DeVore for his kind and warm hospitality at Texas A&M University and the very exciting daily joint discussions which later inspired part of this work. We acknowledge the financial support provided by the START-award "Sparse Approximation and Optimization in High Dimensions" of the Fonds zur Förderung der wissenschaftlichen Forschung (FWF, Austrian Science Foundation). We would like to thank the anonymous referees for their very valuable comments and remarks.

References

[1] R. Ahlswede and A. Winter, *Strong converse for indentification via quantum channels*, IEEE Trans. Inform. Theory **48** (2002), no. 3, 569–579.

[2] E. Arias–Castro and Y. C. Eldar, *Noise folding in compressed sensing*, IEEE Signal Processing Letters **18** (2011), no. 8, 478–481.

[3] R. G. Baraniuk, M. Davenport, R. A. DeVore, and M. Wakin, *A simple proof of the restricted isometry property for random matrices*, Constr. Approx. **28** (2008), no. 3, 253–263.

[4] H. Bauschke and H. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Review **38** (1996), no. 3, 367–426.

[5] E. J. Candès, *Harmonic analysis of neural networks.*, Appl. Comput. Harmon. Anal. **6** (1999), no. 2, 197–218.

[6] _____, *Ridgelets: Estimating with ridge functions.*, Ann. Stat. **31** (2003), no. 5, 1561–1599.

[7] E. J. Candès and D. L. Donoho, *Ridgelets: a key to higher-dimensional intermittency?*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **357** (1999), no. 1760, 2495–2509.

[8] E. J. Candès, J. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math. **59** (2006), no. 8, 1207–1223.

[9] E. J. Candès and T. Tao, *The Dantzig selector: statistical estimation when p is much larger than n*, Ann. Statist. **35** (2007), no. 6, 2313–2351.

[10] A. Cohen, W. Dahmen, and R. A. DeVore, *Compressed sensing and best k-term approximation*, J. Amer. Math. Soc. **22** (2009), no. 1, 211–231.

[11] A. Cohen, I. Daubechies, R. A. DeVore, G. Kerkyacharian, and D. Picard, *Capturing ridge functions in high dimensions from point queries*, Constr. Approx. (to appear).

[12] R. Courant and D. Hilbert, *Methods of Mathematical Physics, II*, New York: Interscience Publishers, 1962.

[13] R. A. DeVore, G. Petrova, and P. Wojtaszczyk, *Instance optimality in probability with an ℓ_1-minimization decoder*, Appl. Comput. Harmon. Anal. **27** (2009), no. 3, 275–288.

[14] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory **52** (2006), no. 4, 1289–1306.

[15] M. Fazel, *Matrix Rank Minimization with Applications*, Ph.D. thesis, Stanford University, 2002.

[16] M. Fornasier, *Numerical Methods for Sparse Recovery*, Theoretical Foundations and Numerical Methods for Sparse Recovery (M. Fornasier, ed.), Radon Series on Computational and Applied Mathematics, De Gruyter Verlag, 2010.

[17] M. Fornasier and H. Rauhut, *Compressive Sensing*, Handbook of Mathematical Methods in Imaging (O. Scherzer, ed.), vol. 1, Springer, 2010, pp. 187–229.

[18] S. Foucart, *A note on ensuring sparse recovery via ℓ_1-minimization*, Appl. Comput. Harmon. Anal. **29** (2010), no. 1, 97–103.

[19] G. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, 1996.

[20] F. John, *Plane waves and spherical means applied to partial differential equations*, New York: Interscience Publishers, 1955.

[21] M. Ledoux, *The Concentration of Measure Phenomenon*, American Mathematical Society, Providence, Rhode Island, 2001.

[22] B. F. Logan and L. A. Shepp, *Optimal reconstruction of a function from its projections*, Duke Math. J. **42** (1975), no. 4, 645–659.

[23] E. Novak and H. Woźniakowski, *Tractability of Multivariate Problems, Volume I: Linear Information*, EMS Tracts in Mathematics, Vol. 6, Eur. Math. Soc., Zürich, 2008.

[24] ———, *Approximation of infinitely differentiable multivariate functions is intractable*, Journal of Complexity **25** (2009), 398–404.

[25] R. I. Oliveira, *Sums of random Hermitian matrices and an inequality by Rudelson*, Electron. Commun. Probab. (2010), no. 15, 203–212.

[26] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi, *A simplified approach to recovery conditions for low-rank matrices*, Proc. Intl. Sympo. Information Theory (ISIT), 2011.

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

[28] B. Recht, M. Fazel, and P. Parillo, *Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization*, SIAM Rev. **52** (2010), no. 3, 471–501.

[29] M. Rudelson and R. Vershynin, *Sampling from large matrices: An approach through geometric functional analysis*, J. ACM **54** (2007), no. 4, Art. 21, 19 pp.

[30] W. Rudin, *Function theory in the unit ball of* $\mathbb{C}^n$, Springer-Verlag, New York-Berlin, 1980.

[31] K. Schnass and J. Vybíral, *Compressed learning of high-dimensional sparse functions*, Proc. ICASSP11, 2011.

[32] G. W. Stewart, *Perturbation theory for the singular value decomposition*, SVD and Signal Processing, II (R. J. Vacarro, ed.), Elsevier, 1991.

[33] J. Tropp, *User-friendly tail bounds for sums of random matrices*, Found. Comput. Math. (to appear).

[34] P.-A. Wedin, *Perturbation bounds in connection with singular value decomposition*, BIT **12** (1972), 99–111.

[35] H. Weyl, *Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)*, Mathematische Annalen **71** (1912), 441–479.

[36] P. Wojtaszczyk, ℓ_1 *minimisation with noisy data*, preprint (2011).