Complexity of Approximation of Functions of Few Variables in High Dimensions

P. Wojtaszczyk *
Institut of Applied Mathematics
University of Warsaw
ul. Banacha 2, 02-097 Warszawa
Poland
e-mail: wojtaszczyk@mimuw.edu.pl

October 18, 2010

Abstract

In [7] we considered smooth functions on $[0,1]^N$ which depends on a much smaller number of variables ℓ or continuous functions which can be approximated by such functions. We were interested in approximating those functions when we can calculate point values at points of our choice. The number of points we needed for non-adaptive algorithm was higher then in the adaptive case. In this paper we improve on [7] and show that in the non-adaptive case one can use the same number of points (up to a multiplicative constant depending on ℓ) that we need in the adaptive case.

1 Introduction

The numerical solution of many scientific problems can be reformulated as the approximation of a function f, defined on a domain in $\mathbb{R}^N$. When N is large this problem very often becomes untractable. This is the so-called *curse of dimensionality.* On the other hand, the functions f that arise as solutions to real world problems are thought to be much better behaved than a general N-variate function. Very often it turns out that essentially they depend on only a few parameters. This has led to a concerted effort to develop a theory and algorithms which approximate such functions well without suffering the effect of the curse of dimensionality. There are many impressive approaches (see [2, 6, 17, 11, 12, 15, 18] as representative) which are being developed in a variety of settings. There is also the active literature in compressed sensing which is based on the model that real world functions are sparsely represented in a suitable basis (see e.g. [3, 8, 4] and the references in these papers).

In [7] we considered a version of this problem, namely we considered a continuous function f defined on $[0,1]^N$ but depending only on ℓ variables $x_{i_1},\dots,x_{i_\ell}$ where $i_1,\dots,i_\ell$ are unknown to us (this is the exact case). Under some smoothness assumptions we gave an approximation to such f from point values. We also considered a situation when f is not a function of ℓ variables but can be approximated by such a function to some tolerance ϵ (this is the approximate case). We considered both adaptive and non-adaptive choices of points. Our benchmark was $(L+1)^\ell$ points which is the number of points in the uniform grid of $[0,1]^\ell$ and we wanted the level of approximation which can be achived for functions on $[0,1]^\ell$ using those points under our smoothness assumption. The numbers of points we needed to prove our results are sumarised in this table

	Exact	Approximate
Adaptive	$C(\ell)(L+1)^\ell \log N$	$C(\ell)(L+1)^\ell \log N + C'(\ell)\log^2 N$
Non-adaptive	$C(\ell)(L+1)^{\ell+1}\log^2 N$	$C(\ell)(L+1)^{\ell+1}\log^2 N$

*This research was partially supported by Polish MNiSW grant N201 269335.

We see that we needed substantially more points in the non-adaptive setting. In this paper we show (see Corollary 4.7 and Theorem 5.3) that up to a multiplicative constant $C(\ell)$ in all cases we need the same number of points i.e. $C(\ell)(L+1)^{\ell}\ln N$. Our algorithms are theoretical; we are interested in a real complexity of the problem so we aimed at minimising the number of point evaluations used. It is known that for natural linear problem adaptivity does not help much (see [16, 17] for detailed discussion). However our problem is non-linear so the question remains.

In information based complexity related problems are studied in the framework of weighted spaces. In particular finite order weights deal with *linear* spaces of functions f of N variables which can be represented (in a way unknown to us) as a finite sum $f=\sum_j g_j$ where each g_j depends only on $\ell < N$ variables (see [19, 17]). There are two differences between our approach and the existing theory of finite order weights: we work in the sup norm while they work mostly in the Hilbert space setting and we deal with one function of ℓ variables which makes our problem non-linear.

The paper is organized as follows. In the next Section 2 we give the necessary combinatorial background. In particular we introduce sets with **Determining property** which is used in Section 5. In Section 3 we recall our approximation setup from [7]. In Section 4 we discuss the exact case and in Section 5 we discuss the approximate case. In the last Section we present some remarks and open problems.

1.1 Notation

Before we proceed let us explain some notation used throughout the paper. N and ℓ are integers $\ell < N$; we think about N as large and about ℓ as rather small. We also use an integer L. We denote by $\mathcal{L}$ the set

$$\{0, \tfrac{1}{N}, \tfrac{2}{N}, \ldots, \tfrac{N-1}{N}, 1\} \subset [0,1]$$

and $h = 1/L$. For an integer R we will denote by $[R]$ the set of integers $\{1,2,\ldots,R\}$. For any finite set Γ the symbol $\#\Gamma$ denotes the cardinality of Γ. For a finite set A by $\mathcal{L}^A$ ($[0,1]^A$ resp.) we will denote the set of all vectors indexed by A with values in $\mathcal{L}$ ($[0,1]$ resp.) Formally they are functions from A into $\mathcal{L}$ ($[0,1]$ resp.). We will use this point of view by using a functional notation for points from $[0,1]^N$ (which formally is $[0,1]^{[N]}$) and $\mathcal{L}^{\ell}$ (which formally is $\mathcal{L}^{[\ell]}$). In particular for a point $x \in [0,1]^N$ and a set $A \subset [N]$ the symbol $\chi_A x$ denotes an element from $[0,1]^N$ which has coordinates outside A equal to zero and coordinates from A equal to corresponding coordinates of x. We will also use $x|A$ to denote an element from $[0,1]^A$ which has on A the same coordinates as x. As is customary the symbol $C(X)$ will denote the space of all continuous functions on the set X. We will use this symbol also when X is finite when it means all functions on X.

2 Combinatorial background

Let $\mathcal{A}$ be a collection of partitions $\mathbf{A}$ of $[N]$ such that each $\mathbf{A}$ consists of ℓ disjoint sets $A_1, A_2, \ldots, A_\ell$. We say that the collection $\mathcal{A}$ is ℓ-separating if for any distinct integers $i_1, \ldots, i_\ell \in [N]$ there exists a partition $\mathbf{A} \in \mathcal{A}$ such that each set in $\mathbf{A}$ contains precisely one of integers $i_1, \ldots, i_\ell$. In [7] this property was termed Partition Assumption. It appears in theoretical computer science as a perfect hashing (c.f. [9, 13]). It is known [9, 13] and also explained in [7] that there exist ℓ-separating families of partitions with small cardinality. Random constructions give families $\mathcal{A}$ with

$$\#\mathcal{A} \le 2\ell e^{\ell} \ln N. \tag{2.1}$$

The lower estimates for the cardinality of ℓ separating families are the main subject of [9, 13].

If we look at an ℓ-separating family of partitions $\mathcal{A}$ then in the notation of Friedman and Komlos [9] the minimal cardinality of $\mathcal{A}$ is $Y(l,l,N)$. An easy estimate (7) in [9] gives a lower estimate $\#\mathcal{A} \ge \log N/\log l$. The main result (Theorem 2) of [9] (see also 1.1 in [13]) is

$$\#\mathcal{A} \ge \frac{l^{l-1}\log N}{l!\log 2}.$$

From Stirling formula we obtain

$$\#\mathcal{A} \ge \frac{e^{l-\theta/12l}}{\sqrt{2\pi}l^{3/2}}\log N \tag{2.2}$$

so the random result (2.1) is very precise.

A set $\mathfrak{A} \subset \mathcal{L}^N$ is called an ℓ-projection set if for every set $A \subset [N]$ of cardinality ℓ and every vector $w \in \mathcal{L}^A$ there exists $v \in \mathfrak{A}$ such that $v|A = w$.

Here we provide a simple random estimate of the cardinality of an ℓ-projection set.

Proposition 2.1. *A random subset of $\mathcal{L}^N$ of cardinality $2\ell[\ln(L+1)+\ln N](L+1)^\ell$ with overwhelming probability is an ℓ-projection set.*

Proof. First we describe our randomness. Let $(\gamma_j)_{j=1}^\infty$ be a sequence of independent, identically distributed, random variables each taking values in $\mathcal{L}$, each value with the same probability $(L+1)^{-1}$. We define random vectors $x_j(\omega) \in \mathcal{L}^N$ as $x_j(\omega) = (\gamma_{(j-1)N+1}, \gamma_{(j-1)N+2}, \ldots, \gamma_{jN})$ for $j = 1, 2, \ldots$. We define the set

$$\mathfrak{A} = \mathfrak{A}_r(\omega) = \{x_1(\omega), \ldots, x_r(\omega)\}.$$

Obviously $\mathfrak{A}$ *is not an ℓ-projection set* if there exists $A \subset [N]$ with $\#A = \ell$ and $w \in \mathcal{L}^A$ such that $x_j|A \neq w$ for $j = 1, 2, \ldots, r$. So

$$\begin{aligned}
\Delta &:= \mathbb{P}(\mathfrak{A} \text{ is not } \ell\text{-projection set })\\
&\leq (L+1)^\ell \binom{N}{\ell} \mathbb{P}(x_j|A \neq w \text{ for } j = 1, 2, \ldots, r)\\
& \text{ since } x_j\text{'s are independent}\\
&= (L+1)^\ell \binom{N}{\ell} \left[\mathbb{P}(x_1|A \neq w)\right]^r\\
&= (L+1)^\ell \binom{N}{\ell} \left[1-(L+1)^{-\ell}\right]^r.
\end{aligned}$$

Since $\left[1-(L+1)^{-\ell}\right]^r \leq \exp -\frac{r}{(L+1)^\ell}$ we get

$$\Delta \leq \exp\left(\ell\ln(L+1) + \ell\ln N - \frac{r}{(L+1)^\ell}\right)$$

so for $r \geq 2\ell[\ln(L+1)+\ln N](L+1)^\ell$ we get

$$\Delta \leq \exp -\ell[\ln(L+1)+\ln N] = [N(L+1)]^{-\ell}.$$

■

The question what is the smallest cardinality of projection sets for various combinations of parameters seems to be unsolved. Some results are given in [14, 10]. We will need a somewhat stronger property

Determining Property: *The set $\mathfrak{B} \subset \mathcal{L}^N$ is said to have this property if it satisfies ℓ-projection property and for each A, B of cardinality $\leq \ell$ and any pair $a, b \in \mathcal{L}^A$ there exist $P, P' \in \mathfrak{B}$ such that $P|A = a$ and $P'|A = b$ and $P|(B \setminus A) = P'|(B \setminus A)$.*

There are at least two ways to build sets with Determining Property of rather small cardinality:

(i) Each 2ℓ-projection set in $\mathcal{L}^N$ has Determining Property; this is clear. Proposition 2.1 gives such sets of cardinality $\leq 4\ell(\ln L + \ln N)(L+1)^{2\ell}$.

(ii) Let us fix a 2ℓ-separating family $\mathcal{B}$ of partitions of $[N]$ and let us define

$$\mathfrak{B} = \bigcup_{V\subset[2\ell]} \bigcup_{\mathbf{B}\in\mathcal{B}} \Big\{ \sum_{j\in V} \alpha_j h \chi_{B_j} \ : \ \alpha_j = 0, 1, \ldots, N\Big\}. \tag{2.3}$$

where $\mathbf{B} = (B_1, \ldots, B_{2\ell}) \in \mathcal{B}$ and $V \subset [2\ell]$ is a set with $\#V = \ell$. It is easy to check that this set satisfies Determining Property. Using (2.1) we see that there are such sets with cardinality $\leq 2\binom{2\ell}{\ell}\ell e^\ell \ln N(L+1)^\ell$.

3 Approximation background

For any function ϕ defined on a set $\mathcal{D}$ (it may be defined on a bigger set $\mathcal{E} \supset \mathcal{D}$) we put

$$\|\phi\|_{\mathcal{D}} = \sup\{|\phi(P)| \ : \ P \in \mathcal{D}\} \tag{3.1}$$

$\mathcal{L}$ is the lattice of equally spaced points (spacing h) on $[0,1]$ and $\mathcal{L}^\ell$ is the lattice of equally spaced points (spacing h) in $[0,1]^\ell$.

We assume that we have a sequence of linear operators $A_h : C(\mathcal{L}^\ell) \to C([0,1]^\ell)$ (defined for $L = 1, 2, \dots$) such that

(i) There is an absolute constant C_0 such that $\|A_h(g)\|_{[0,1]^\ell} \le C_0 \|g\|_{\mathcal{L}^\ell}$ for all h.

(ii) If g depends only on the variables from the set $A \subset [\ell]$ with $\#A \le \ell$ then $A_h(g)$ also depends only on those variables.

(iii) $A_h(g) \equiv g$ if g is a constant.

(iv) If π is any permutation of the variables $x = (x_1, \dots, x_\ell)$ (which can respectively be thought of as a permutation of the indices $\{1, \dots, \ell\}$), then $A_h(g(\pi(\cdot))(\pi^{-1}(x)) = A_h(g)(x),\ x \in [0,1]^\ell$.

We define the following approximation class:

$$\mathcal{A}^s := \mathcal{A}^s((A_h)) = \{g \in C([0,1]^\ell) : \ \|g - A_h(g|\mathcal{L}^\ell)\|_{[0,1]^\ell} \le Ch^s,\ h = 1/L,\ L = 1, 2, \dots\}, \tag{3.2}$$

with semi-norm

$$|g|_{\mathcal{A}^s} := \sup_h \left\{ h^{-s} \|g - A_h(g|\mathcal{L}^\ell)\|_{[0,1]^\ell} \right\}. \tag{3.3}$$

Let us note that this class depends on the whole sequence of operators (A_h) not on just one operator. We obtain the norm on $\mathcal{A}^s$ by adding $\|\cdot\|_{[0,1]^\ell}$ to the semi-norm. As was discussed in [7], there is typically a range $0 < s \le S$, where the approximation classes can be characterized as smoothness spaces.

We need the following simple fact from [7] about $\mathcal{A}^s$ functions.

Lemma 3.1. *Suppose $g \in \mathcal{A}^s$ and $\|g(x)\|_{\mathcal{L}^{[\ell]}} \le \epsilon$. Then,*

$$\|g\|_{[0,1]^\ell} \le C_0 \epsilon + |g|_{\mathcal{A}^s} h^s, \tag{3.4}$$

where C_0 is the constant in (i).

Proof. This follows directly from the triangle inequality

$$\|g\|_{[0,1]^\ell} \le \|g - A_h(g|\mathcal{L}^N)\|_{[0,1]^\ell} + \|A_h(g|\mathcal{L}^N)\|_{[0,1]^\ell} \le |g|_{\mathcal{A}^s} h^s + C_0 \epsilon.$$

∎

With slight abuse of the notation, for $g \in C([0,1]^\ell)$ we will denote by $A_h(g)$ the operator $A_h(g|\mathcal{L}^\ell)$.

We say that a function $f \in C([0,1]^N)$ depends on variable $j \in [N]$ if there are two points $P, P' \in [0,1]^N$ which differ only at coordinate j such that $f(P) \ne f(P')$. We call such a variable a *change variable*. Those are variables on which our function depends. For a given sequence $\mathcal{J} = (j_1, \dots, j_\ell)$ of distinct integers from $[N]$ and a function g defined on $[0,1]^\ell$ we define an embedding operator

$$I_{\mathcal{J}}(g)(x_1, \dots, x_N) = g(x_{j_1}, \dots, x_{j_\ell}).$$

Clearly every function on $[0,1]^N$ which depends on ℓ variables is of such form for some $\mathcal{J}$ and g defined on $[0,1]^\ell$. We will consider two cases:

Exact case: We assume that the function f on $C([0,1]^N)$ equals $I_{\mathcal{J}}(g)$ for some sequence $\mathcal{J}$ of ℓ distinct integers from $[N]$ and some $g \in \mathcal{A}^s$ for some $s > 0$. We do not know neither g nor $\mathcal{J}$.

Approximate case: We assume that $f \in C([0,1]^N)$ and there exists a function $g \in \mathcal{A}^s$ for some $s > 0$ and $\epsilon \ge 0$ and a sequence $\mathcal{J}$ of ℓ distinct integers from $[N]$ such that

$$\|f - I_{\mathcal{J}}(g)\|_{[0,1]^N} \le \epsilon.$$

We do not know $\mathcal{J}$, g nor $\epsilon \ge 0$.

4 Exact case

In this section we consider the case when the function $f \in C([0,1]^N)$ depends on at most ℓ variables. We decided to present this case separately (despite the fact that it formally follows from the more general approximate case) because we can present a different, simpler and more efficient algorithm which uses fewer points (but only by a multiplicative constant). Also this section can serve as an introduction to Section 5.

We fix an $(\ell+1)$-separating family $\mathcal{A}$ of partitions of the set $[N]$. For any $\mathbf{A} = (A_1, \ldots, A_{\ell+1}) \in \mathcal{A}$ and any $s = 1, \ldots, \ell+1$ we define a family of points

$$\mathcal{P}(\mathbf{A}, s) = \Big\{ \sum_{j \neq s} \alpha_j h \chi_{A_j} \ : \ \alpha_j = 0, 1, \ldots, L \Big\} \subset \mathcal{L}^N. \tag{4.1}$$

We denote $\mathfrak{A}_\mathbf{A} = \bigcup_{s=1}^{\ell+1} \mathcal{P}(\mathbf{A}, s)$ and define the set $\mathfrak{A}$ of *base points* as

$$\mathfrak{A} = \bigcup_{\mathbf{A} \in \mathcal{A}} \mathfrak{A}_\mathbf{A} = \bigcup_{\mathbf{A} \in \mathcal{A}} \bigcup_{s=1}^{\ell+1} \mathcal{P}(\mathbf{A}, s) \subset \mathcal{L}^N \tag{4.2}$$

We see that $\#\mathfrak{A} \leq (\ell+1)(L+1)^\ell \#\mathcal{A}$. Now we evaluate f at points from $\mathfrak{A}$. We say that the set $A_j \in \mathbf{A} \in \mathcal{A}$ is a *change set* if there exist two points P and P' in $\mathfrak{A}_\mathbf{A}$ which differ only on A_j such that $f(P) \neq f(P')$. Clearly each change set contains at least one change variable of f. We look at $\mathbf{A}$'s with maximal number of change sets; there may be many of them — call them *maximal.* Since the function f depends on at most ℓ variables each partition $\mathbf{A} \in \mathcal{A}$ contains a nonchange set. Actually there may be change variables that are not reflected by values of f on $\mathcal{L}^N$. Let us call a variable $j \in [N]$ *visible at the scale* h if there exist $P, P' \in \mathcal{L}^N$ which differ only at coordinate j such that $f(P) \neq f(P')$.

Lemma 4.1. *(i) Each change set contains at least one change variable visible at scale h.*

(ii) In each maximal partition each change set contains exactly one change variable visible at scale h.

(iii) For every maximal partition, each change variable visible at scale h belongs to some change set.

Proof. The first statement is clear as the change set is defined via points from $\mathcal{L}^N$. Let $\mu \leq \ell$ be the number of variables visible at scale h. Thus the number of change sets in any partition is $\leq \mu$. Let $\mathbf{A} = (A_1, \ldots, A_{\ell+1}) \in \mathcal{A}$ be a partition such that $\#\mathcal{J} \cap A_j \leq 1$ for $j = 1, 2, \ldots, \ell+1$. Such an $\mathbf{A}$ exists because $\mathcal{A}$ is $(\ell+1)$-separating family. Let j be visible at scale h and let $P, P' \in \mathcal{L}^N$ differ only at coordinate j and $f(P) \neq f(P')$. One easily see that there exist points $\tilde{P}$ and $\tilde{P}'$ in $\mathfrak{A}_\mathbf{A}$ such that $\tilde{P}|\mathcal{J} = P|\mathcal{J}$ and $\tilde{P}'|\mathcal{J} = P'|\mathcal{J}$. Thus we get $f(\tilde{P}) = f(P) \neq f(P') = f(\tilde{P}')$. This means that the set A_s such that $j \in A_s$ is a change set. Repeating this argument for other variables visible at scale h we see that the number of change sets in partition $\mathbf{A}$ equals μ. Thus $\mathbf{A}$ is maximal and each maximal partition has μ change sets. From this the remaining statements follow. ■

For a maximal $\mathbf{A}$ let $U_\mathbf{A}$ be the union of non-change sets in $\mathbf{A}$. Consider the set

$$W =: [N] \setminus \bigcup_{\mathbf{A} \text{ maximal}} U_\mathbf{A}.$$

Clearly every variable visible at scale h is in W. Suppose that $j \in [N]$ is not visible at scale h. There exists a partition $\mathbf{A} = (A_1, \ldots, A_{\ell+1}) \in \mathcal{A}$ such that each set A_s contains at most one variable visible at scale h and there exists a set A_{s_0} which contains j but no variable visible at scale h. From Lemma 4.1 we infer that $\mathbf{A}$ is a maximal partition and that A_{s_0} is not a change set, so $j \in U_\mathbf{A}$. Thus W is the set of change variables visible at scale h.

Once the set of change variables visible at the scale h have been identified we can (and for the sake of completeness of exposition we will) repeat arguments from [7].

Our function $f = I_\mathcal{J}(g)$. Clearly $W \subset \mathcal{J}$ but we may be missing some variables, so first we add(if needed) arbitrary coordinaes to W to get a sequence $\mathcal{J}' = (j'_1, \ldots, j'_\ell)$ with $1 \leq j'_1 < j'_2 < \ldots, < j'_\ell \leq N$.

We fix a partition $\mathbf{A} \in \mathcal{A}$ such that each $A_j \in \mathbf{A}$ contains at most one coordinate from $\mathcal{J}'$. We will assume (by property (iv) of operators (A_h) in our underlying approximation scheme this will not change the function $\hat{f}$ we are going to define) that $j_i' \in A_i$ for $i = 1, \ldots, \ell$, so $s = \ell+1$. The function $f|\mathcal{P}(\mathbf{A}, \ell+1)$ we naturally identify with a function on $\mathcal{L}^\ell$ and we define

$$\hat{g} = A_h(f|\mathcal{P}(\mathbf{A}, \ell+1)) \in C([0,1]^\ell).$$

Now we define $\hat{f} \in C([0,1]^N)$ as

$$\hat{f}(x_1, \ldots, x_N) = I_{\mathcal{J}'}(\hat{g})(x_1, \ldots, x_N) = \hat{g}(x_{j_1'}, \ldots, x_{j_\ell'}). \tag{4.3}$$

Now we are ready to state the main result of this section

Theorem 4.2. *If $f = I_{\mathcal{J}}(g)$ with $g \in \mathcal{A}^s$, then the function $\hat{f}$ defined in (4.3) satisfies*

$$\|f - \hat{f}\|_{[0,1]^N} \leq |g|_{\mathcal{A}^s} h^s. \tag{4.4}$$

To define $\hat{f}$ we use $\leq (\ell+1)(L+1)^\ell \#\mathcal{A}$ points chosen non-adaptively.

Proof. The number of points was already calculated. To prove tha bound (4.4) we define $S = A_h(g)$ and write

$$f - \hat{f} = (I_{\mathcal{J}}(g) - I_{\mathcal{J}}(S)) + (I_{\mathcal{J}}(S) - I_{\mathcal{J}'}(\hat{g}))\,. \tag{4.5}$$

The first term on the right side satisfies

$$\|I_{\mathcal{J}}(g) - I_{\mathcal{J}}(S)\|_{[0,1]^N} = \|g - A_h(g)\|_{[0,1]^\ell} \leq |g|_{\mathcal{A}^s} h^s. \tag{4.6}$$

From properties of operators (A_h) and the fact that both $\mathcal{J}$ and $\mathcal{J}'$ contain the set W of all coordinates visible at scale h we see that $I_{\mathcal{J}}(S) = I_{\mathcal{J}'}(\hat{g})$, so the second term on the right side of (4.5) is identically zero. ∎

If we use separating sets of partitions given in (2.1) we get

Corollary 4.3. *If $f = I_{\mathcal{J}}(g)$ with $g \in \mathcal{A}^s$, then the function $\hat{f}$ defined in (4.3) satisfies*

$$\|f - \hat{f}\|_{[0,1]^N} \leq |g|_{\mathcal{A}^s} h^s. \tag{4.7}$$

To define $\hat{f}$ we use $\leq 2(\ell+1)^2 e^\ell (L+1)^\ell \ln N$ point values chosen non-adaptively.

5 Approximate case

In this section we assume that we are given a function $f \in C([0,1]^N)$ such that

$$\|f - I_{\mathcal{J}}(g)\|_{[0,1]^N} \leq \epsilon$$

for certain sequence $\mathcal{J}$, function $g \in \mathcal{A}^s$ and $\epsilon \geq 0$. We do not know any of those. As in the previous section the argument splits into two part. First combinatorically we identify a good set of coordinates (which may be different from $\mathcal{J}$) and then prove the approximation estimate.

We switch notation somewhat and assume that $f \in C([0,1]^N)$ is a function of the form $f = I_{\mathcal{J}}(g) + \eta$. To simplify the notation we put $I_{\mathcal{J}}(g) =: \tilde{g}$ so $\tilde{g}$ depends only on variables from a set $B \subset [N]$ with $\#B = \ell$ and $\|\eta\|_{\mathcal{L}^N} \leq \|\eta\|_{[0,1]^N} \leq \epsilon$.

We define the set $\mathfrak{B}$ of *base points* as any subset of $\mathcal{L}^N$ with the **Determining property**.

For a set $A \subset [N]$, $\#A = \ell$ and $x \in \mathcal{L}^N$ and ϕ a function on $\mathcal{L}^N$ (or on $[0,1]^N$) we define

$$\begin{aligned} \alpha(\phi, A, x) &= \max\{\phi(P) \ : \ P \in \mathfrak{B} \ \text{ and } \ P|A = x|A\} \\ \beta(\phi, A, x) &= \min\{\phi(P) \ : \ P \in \mathfrak{B} \ \text{ and } \ P|A = x|A\}. \end{aligned}$$

We define

$$h_A(x) = \frac{\alpha(f, A, x) + \beta(f, A, x)}{2}. \tag{5.1}$$

Clearly each h_A is a function on $\mathcal{L}^N$ which depends only on variables from A.

We define set A_0 as $\operatorname{argmin}_A \|f - h_A\|_{\mathfrak{B}}$. From the very definition $\|f - h_{A_0}\|_{\mathfrak{B}} \le \|f - h_B\|_{\mathfrak{B}}$.

Note that

$$\begin{aligned} \alpha(f, A, x) &= \max\{\tilde{g}(P) + \eta(P) \ : \ P \in \mathfrak{B} \ \text{ and } \ P|A = x|A\} \\ &\le \epsilon + \max\{\tilde{g}(P) \ : \ P \in \mathfrak{B} \ \text{ and } \ P|A = x|A\} \\ &= \epsilon + \alpha(\tilde{g}, A, x) \end{aligned}$$

Analogously we have

$$\begin{aligned} \alpha(f, A, x) &\ge \alpha(\tilde{g}, A, x) - \epsilon \\ \beta(f, A, x) &\le \beta(\tilde{g}, A, x) + \epsilon \\ \beta(f, A, x) &\ge \beta(\tilde{g}, A, x) - \epsilon \end{aligned}$$

For simplicity of notation we put $h := h_{A_0}$. Now we see that

$$\begin{aligned} \|f - h\|_{\mathfrak{B}} &\le \|f - h_B\|_{\mathfrak{B}} \le \frac{1}{2}\left(\|f - \alpha(f, B, \cdot)\|_{\mathfrak{B}} + \|f - \beta(f, B, \cdot)\|_{\mathfrak{B}}\right) \\ &\le \epsilon + \frac{1}{2}\left(\|\tilde{g} - \alpha(f, B, \cdot)\|_{\mathfrak{B}} + \|\tilde{g} - \beta(f, B, \cdot)\|_{\mathfrak{B}}\right) \\ &\le 2\epsilon + \frac{1}{2}\left(\|\tilde{g} - \alpha(\tilde{g}, B, \cdot)\|_{\mathfrak{B}} + \|\tilde{g} - \beta(\tilde{g}, B, \cdot)\|_{\mathfrak{B}}\right). \end{aligned}$$

But for $\tilde{P} \in \mathfrak{B}$ we have

$$\alpha(\tilde{g}, B, \tilde{P}) = \max\{\tilde{g}(P) \ : \ P \in \mathfrak{B} \ \text{ and } \tilde{P}|B = P|B\} = \tilde{g}(\tilde{P})$$

and

$$\beta(\tilde{g}, B, \tilde{P}) = \min\{\tilde{g}(P) \ : \ P \in \mathfrak{B} \ \text{ and } \tilde{P}|B = P|B\} = \tilde{g}(\tilde{P})$$

so we get $\|f - h\|_{\mathfrak{B}} \le 2\epsilon$ which gives $\|\tilde{g} - h\|_{\mathfrak{B}} \le 3\epsilon$.

For a set $V \subset [N]$ we define

$$\operatorname{osc}(V) = \max\{|\tilde{g}(P) - \tilde{g}(P')| \ : \ P, P' \in \mathcal{L}^N \ \text{ and } \ P|([N] \setminus V) = P'|([N] \setminus V)\}.$$

Proposition 5.1. *We have* $\operatorname{osc}(B \setminus A_0) \le 6\epsilon$.

Proof: If $\operatorname{osc}(B \setminus A_0) > 6\epsilon$ we can fix $P, P' \in \mathcal{L}^N$ such that P and P' differ only on $B \setminus A_0$ and $6\epsilon < |\tilde{g}(P) - \tilde{g}(P')|$. Since $\#B, \#A_0 \le \ell$ we fix $Q, Q' \in \mathfrak{B}$ such that $Q|B = P|B$ and $Q'|B = P'|B$ and $Q|A_0 \setminus B = Q'|A_0 \setminus B$. Note that $Q|A_0 = Q'|A_0$. Now we have

$$\begin{aligned} 6\epsilon &< |\tilde{g}(P) - \tilde{g}(P')| = |\tilde{g}(Q) - \tilde{g}(Q')| \\ &\le |h(Q) - h(Q')| + 6\epsilon = 6\epsilon \end{aligned}$$

This contradiction completes the proof. ∎

Remark 5.2. *The algorithm to identify A_0 described above is not very efficient. It requires to search through all ℓ element subsets of $[N]$ which is very time consuming.*

Now that we have A_0 we follow ideas from [7]. We fix a subset $\mathcal{V} \subset \mathfrak{B}$ such that for each $x \in \mathcal{L}^{A_0}$ there exists unique $v_x \in \mathcal{V}$ such that $v_x|A_0 = x$. There may be many such $\mathcal{V}$'s but our arguments works for any of them. For a function $\phi \in C([0,1]^N)$ we define an operator R by the formula $R(\phi)(x) = \phi(v_x)$ for $x \in \mathcal{L}^{A_0}$. We fix a sequence $\mathcal{J}' = (j'_1, \ldots, j'_\ell)$ such that $\{j'_1, \ldots, j'_\ell\} = A_0$ i.e. this sequence identifies A_0 with $[\ell]$. Using this identification we treat R as a linear operator from $C([0,1]^N)$ into $C(\mathcal{L}^{[\ell]})$ and define

$$\hat{f} = I_{\mathcal{J}'} A_h R(f). \tag{5.2}$$

Now we are ready to formulate the main result of this note.

Theorem 5.3. *Let $\hat{f}$ be defined in (5.2). We have*

$$\|f - \hat{f}\|_{[0,1]^N} \leq (13C_0 + 1)\epsilon + |g|_{\mathcal{A}^s} \tag{5.3}$$

If we use the set $\mathfrak{B}$ defined in (2.3) then the definition of $\hat{f}$ uses $\leq 2\binom{2\ell}{\ell}\ell e^\ell (L+1)^\ell \ln N$ point values chosen non-adaptively.

Proof. Clearly $\hat{f} = I_{\mathcal{J}'} A_h R(\tilde{g}) + I_{\mathcal{J}'} A_h R(\eta)$ and each of those functions depends on variables from A_0. We obtain

$$\begin{aligned} \|f - \hat{f}\|_{[0,1]^N} &= \|\tilde{g} + \eta - I_{\mathcal{J}'} A_h R(\tilde{g}) - I_{\mathcal{J}'} A_h R(\eta)\|_{[0,1]^N} \\ &\leq \|\tilde{g} - I_{\mathcal{J}'} A_h R(\tilde{g})\|_{[0,1]^N} + \|\eta\|_{[0,1]^N} + \|I_{\mathcal{J}'} A_h R(\eta)\|_{[0,1]^N} \\ &\leq \|\tilde{g} - I_{\mathcal{J}'} A_h R(\tilde{g})\|_{[0,1]^N} + (C_0 + 1)\epsilon \end{aligned}$$

Now let us recall that we also have a sequence $\mathcal{J}$ which identifies $[\ell]$ with B. From property (iv) of our approximation scheme we may assume that for $\nu = \#A_0 \cap B$ we have $j_s = j'_s$ for $s \leq \nu$ and $A_0 \cap B = \{j_1, \ldots, j_\mu\}$. Note that $R\tilde{g}$ is a function on $\mathcal{L}^\ell$ and $R\tilde{g} = g|\mathcal{L}^\ell$. Now we define $g_1(x) = g(x_1, \ldots, x_\nu, 0, \ldots, 0)$ for $x \in \mathcal{L}^\ell$. From Proposition 5.1 we infer that $\|g - g_1\|_{\mathcal{L}^\ell} \leq 6\epsilon$. Now we define $\Phi = I_{\mathcal{J}} A_h g_1$. Since g_1 depends on coordinates $1, \ldots, \nu$ we see that

$$\Phi = I_{\mathcal{J}} A_h g_1 = I_{\mathcal{J}'} A_h g_1 \tag{5.4}$$

and depends only on coordinates from $B \cap A_0$. If we put $\zeta = g - g_1$ we get

$$\begin{aligned} \|\tilde{g} - \Phi\|_{[0,1]^N} &= \|I_{\mathcal{J}} g - I_{\mathcal{J}} A_h g_1\|_{[0,1]^N} = \|I_{\mathcal{J}} g - I_{\mathcal{J}} A_h g + I_{\mathcal{J}} A_h \zeta\|_{[0,1]^N} && (5.5) \\ &\leq \|g - A_h g\|_{[0,1]^\ell} + \|A_h \zeta\|_{[0,1]^\ell} \leq |g|_{\mathcal{A}^s} h^s + C_0 \|\zeta\|_{\mathcal{L}^\ell} && (5.6) \\ &\leq |g|_{\mathcal{A}^s} h^s + 6C_0\epsilon && (5.7) \end{aligned}$$

Using (5.4) we obtain

$$\|\Phi - I_{\mathcal{J}'} A_h R(\tilde{g})\|_{[0,1]^N} = \|A_h(\zeta)\|_{[0,1]^\ell} \leq 6C_0\epsilon \tag{5.8}$$

So

$$\begin{aligned} \|\tilde{g} - I_{\mathcal{J}'} A_h R(\tilde{g})\|_{[0,1]^N} &\leq \|\tilde{g} - \Phi\|_{[0,1]^N} + \|\Phi - I_{\mathcal{J}'} A_h R(\tilde{g})\|_{[0,1]^N} && (5.9) \\ &\leq |g|_{\mathcal{A}^s} h^s + 12C_0\epsilon && (5.10) \end{aligned}$$

which proves the theorem. ∎

6 Notes and remarks

In [7] and in this note we assume that our function f depends on some of the variables x_j. However in practical application we may not know the correct variables. In a recent preprint [5] authors study the case of ridge functions. Those are functions $f(x_1, \ldots, x_N) = g(\langle \mathbf{b}, \mathbf{x} \rangle)$ for some function of one variable g. Thus f is essentially a function of one variable but this variable is not one of the coordinate variables.

The basic problem left open in our study is the construction of numerically feasible algorithm which uses $CL^\ell \ln N$ point values and produces the approximation error $\leq C(\epsilon + |g|_{\mathcal{A}^s})$.

Another problem is the lower estimate of the number of points needed in various situations. Let us concentrate on the exact case and let us assume that we consider Lipschitz functions and our operators are interpolation operators. If we want an estimate like (4.7) then we easily see that even for fixed coordinates we need $\sim (L+1)^\ell$ point values. The factor $C(\ell)\ln N$ is the price we pay to find the right coordinates. However we do not know if it is the right factor. Comparing (2.1) and (2.2) we see that we use the best possible set of separating partitions. However it is possible that there is a better construction of base points. Since it is plausible that we always need something like ℓ-projection set, this raises the question of a lower estimate for the cardinality of an ℓ-projection set in $\mathcal{L}^N$. We also do not know lower estimates for the cardinality of sets with Determining property. Let me also point out that we do not know if the optimal number of point values in the adaptive and non-adaptive situation is the same.

References

[1] Ashok K. Chandra, Lawrence T. Kou, George Markovsky, Shmuel Zaks *On sets of Boolean n-vectors with all k-projections surjective*, Acta Informatica **20** (1983) 103-111

[2] M. Belkin and P. Niyogi, *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*. Neural Computation , **15**(2003),1373–1396.

[3] E. Candès, J. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure and Appl. Math., **59**(2006), 1207–1223.

[4] A. Cohen, W. Dahmen and R. DeVore, *Compressed sensing and best k-term approximation*, J.Amer. Math. Soc. **22**(1), 211-231 (2009),

[5] A. Cohen, I. Daubechies, R. DeVore, G. Kerkyacharian, and D. Picard, *Capturing Ridge Functions in High Dimensions from Point Queries*, preprint.

[6] Coifman, R.R., M. Maggioni, *Diffusion wavelets*, Appl. Comp. Harm. Anal., **21**(1), (2006), 53-94.

[7] Ronald DeVore, Guergana Petrova, Przemyslaw Wojtaszczyk *Approximation of Functions of Few Variables in High Dimensions*, Constructive Approximation, to appear

[8] D. Donoho, *Compressed Sensing*, IEEE Trans. Information Theory, **52**(2006), 1289–1306.

[9] M.Fredman and J.Komlos *On the size of separating systems and families of perfect hash functions* SIAM J. Alg. Disc. Meth. **5** (1984), 61–68

[10] L. Gargano, J. Körner, U. Vaccaro, *Sperner capacities* Graphs Combin. **9** (1993), no. 1, 31–46.

[11] Eitan Greenshtein, *Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1 constraint,* Ann. Stat., **34** (2006), 2367–2386.

[12] M. Kolountzakis, E. Markakis and A. Mehta, *Learning Symmetric k-juntas in time $n^{o(k)}$*, Proceedings of "Interface between Harmonic Analysis and Number Theory" (2005)

[13] J. Körner and K.Marton *New bounds for Perfect Hashing via Information Theory*, Europ. J. Combinatorics **9** (1988), 523–530

[14] Daniel J. Kleitman, Joel Spencer, *Families of k-independent sets.* Discrete Math. **6** (1973), 255–262.

[15] E. Mossel, R. O'Donnell and R. Servedio *Learning Juntas*, Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC) San-Diego (2003), 206–212

[16] E. Novak, *On the power of adaption*, J. Complexity, **12** (1996), 199237

[17] E. Novak and H. Woźniakowski, **Tractability of Multivariate Problems vol. I: Linear Information**, European Math. Soc. 2008

[18] Todor, R. A., C. Schwab, *Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients*, ETH Report, 2007.

[19] G. W. Wasilkowski, H. Woźniakowski, *Finite-order weights imply tractability of linear multivariate problems*, J. Approximation, **130**, (2004) 57-77