Learning general sparse additive models from point queries in high dimensions

Hemant Tyagi, Jan Vybiral

Learning general sparse additive models from point queries in high dimensions

Hemant Tyagi[*] Jan Vybiral[†]
hemant.tyagi@inria.fr jan.vybiral@fjfi.cvut.cz

April 29, 2019

Abstract

We consider the problem of learning a d-variate function f defined on the cube $[-1,1]^d \subset \mathbb{R}^d$, where the algorithm is assumed to have black box access to samples of f within this domain. Denote $\mathcal{S}_r \subset \binom{[d]}{r}; r = 1, \dots, r_0$ to be sets consisting of unknown r-wise interactions amongst the coordinate variables. We then focus on the setting where f has an additive structure, i.e., it can be represented as

$$f = \sum_{\mathbf{j} \in \mathcal{S}_1} \phi_{\mathbf{j}} + \sum_{\mathbf{j} \in \mathcal{S}_2} \phi_{\mathbf{j}} + \cdots + \sum_{\mathbf{j} \in \mathcal{S}_{r_0}} \phi_{\mathbf{j}},$$

where each $\phi_{\mathbf{j}}; \mathbf{j} \in \mathcal{S}_r$ is at most r-variate for $1 \leq r \leq r_0$. We derive randomized algorithms that query f at carefully constructed set of points, and exactly recover each $\mathcal{S}_r$ with high probability. In contrary to the previous work, our analysis does not rely on numerical approximation of derivatives by finite order differences.

Key words: Sparse additive models, sampling, hash functions, sparse recovery

Mathematics Subject Classifications (2010): 41A25, 41A63, 65D15

1 Introduction

Approximating a function from its samples is a fundamental problem with rich theory developed in areas such as numerical analysis and statistics, and which also has numerous practical applications such as in systems biology [19], solving PDEs [11], control systems [50], optimization [39] etc. Concretely, for an unknown d-variate function $f : \mathcal{G} \to \mathbb{R}$, one is given information about f in the form of samples $(\mathbf{x}_i, f(\mathbf{x}_i))_{i=1}^n$. Here, the $\mathbf{x}_i$'s belong to a compact subset $\mathcal{G} \subset \mathbb{R}^d$. The goal is to construct a smooth estimate $\widehat{f} : \mathcal{G} \to \mathbb{R}$ such that the error between $\widehat{f}$ and f is small. In this paper we focus on the high dimensional setting where d is large. We will consider the scenario where the algorithm has black box access to the function, and can query it at any point within $\mathcal{G}$. This setting appears for instance in materials science [18], where $\mathbf{x}$ represents some material and $f(\mathbf{x})$ some of

[*] INRIA Lille-Nord Europe, France. This work was done by the author while affiliated to the Alan Turing Institute, London, United Kingdom, and the School of Mathematics, University of Edinburgh, United Kingdom; this author was supported by EPSRC grant EP/N510129/1 at the Alan Turing Institute.

[†] Dept. of Mathematics FNSPE, Czech Technical University in Prague, Trojanova 13, 12000 Prague, Czech Republic; this author was supported by the grant P201/18/00580S of the Grant Agency of the Czech Republic, by the grant 8X17028 of the Czech Ministry of Education, and by the Neuron Fund for Support of Science.

its properties of interest (like thermal or electric conductivity). The local-density approximations in density functional theory can be used to compute to high accuracy such properties of a given material. The sampling then corresponds to running a costly numerical PDE-solver. Since such simulations are typically expensive to run, one would like to minimize the number of queries made. This setting is different from the regression setting typically considered in statistics wherein the $\mathbf{x}_i$'s are generated apriori from some unknown distribution over $\mathcal{G}$.

Curse of dimensionality. It is well known that provided we only make smoothness assumptions on f (such as differentiability or Lipschitz continuity), then the problem is intractable, i.e., has exponential complexity (in the worst case) with respect to the dimension d. For instance if $f \in C^s(\mathcal{G})$, then any algorithm needs in the worst case $n = \Omega(\delta^{-d/s})$ samples[1] to uniformly approximate f with error $\delta \in (0, 1)$, cf. [32, 45]. Furthermore, the constants behind the Ω-notation may also depend on d. A detailed study of the dependence on d was performed in the field of *Information Based Complexity* for $f \in C^\infty(\mathcal{G})$ in a more recent work [33]. The authors show that even here, $n = \Omega(2^{\lfloor d/2 \rfloor})$ samples are needed in the worst case for uniform approximation within an error $\delta \in (0, 1)$ (with no additional dependence on d hidden behind the Ω-notation). This exponential dependence on d is commonly referred to as the *curse of dimensionality.* The above results suggest that in order to get tractable algorithms in the high dimensional regime, one needs to make additional assumptions on f. To this end, a growing line of work over the past decade has focused on the setting where f possesses an intrinsic, albeit unknown, low dimensional structure, with much smaller intrinsic dimension than the ambient dimension d. The motivation is that one could now hope to design algorithms with complexity at most exponential in the intrinsic dimension, but with mild dependence on d.

1.1 Sparse additive models (SPAMs)

A popular class of functions with an intrinsic low dimensional structure are the so-called sparse additive models (SPAMs). These are functions that are decomposable as the sum of a small number of lower dimensional functions. To give the formal definition, we denote $[d] = \{1, \dots, d\}$ and by $\binom{[d]}{r}$ we mean the collection of all ordered r-tuples from $[d]$. Then, for $\mathcal{S}_r \subset \binom{[d]}{r}; r = 1, \dots, r_0$, the function $f : \mathcal{G} \to \mathbb{R}$ is of the form

$$f = \sum_{j \in \mathcal{S}_1} \phi_j(x_j) + \sum_{(j_1,j_2) \in \mathcal{S}_2} \phi_{(j_1,j_2)}(x_{j_1}, x_{j_2}) + \cdots + \sum_{(j_1,\dots,j_{r_0}) \in \mathcal{S}_{r_0}} \phi_{(j_1,\dots,j_{r_0})}(x_{j_1}, \dots, x_{j_{r_0}}) \qquad (1.1)$$

with each $|\mathcal{S}_r| \ll \binom{d}{r}$, and $r_0 \ll d$. We can interpret the tuples in $\mathcal{S}_r$ as r^{th} order interactions terms. Let us remark, that usually the terminology Sparse additive models is used for the case $r_0 = 1$, but we prefer to use it here in the general sense of (1.1).

These models appear in optimization under the name *partially separable* models (cf., [20]). They also arise in electronic structure computations in physics (cf., [4]), and problems involving multiagent systems represented as decentralized partially observable Markov decision processes (cf., [16]). There exists a rich line of work that mostly study special cases of the model (1.1). We review them briefly below, leaving a detailed comparison with our results to Section 8.

The case $r_0 = 1$. In this setting, (1.1) reduces to a sparse sum of univariate functions. This model has been studied extensively in the non parametric statistics literature with a range of

[1] This means that there exists a constant $c > 0$ such that $n \geq c\delta^{-d/s}$ when d is sufficiently large. See Section 2 for a formal definition.

results on estimation of f (cf., [23, 25, 26, 28, 35, 37]) and also on variable selection, i.e., identifying the support $\mathcal{S}_1$ (cf., [23, 37, 47]). The basic idea behind these approaches is to approximately represent each ϕ_j in a suitable basis of finite size (for eg., splines or wavelets) and then to find the coefficients in the basis expansion by solving a least squares problem with smoothness and sparsity penalty constraints. Koltchinskii et al. [26] and Raskutti et al. [35] proposed a convex program for estimating f in the Reproducing kernel Hilbert space (RKHS) setting, and showed that f lying in a Sobolev space with smoothness parameter $\alpha > 1/2$ can be estimated at the L_2 rate $\frac{k \log d}{n} + kn^{-\frac{2\alpha}{2\alpha+1}}$. This rate was shown to be optimal in [35]. There also exist results for variable selection, i.e., identifying the support $\mathcal{S}_1$. These results in non parametric statistics are typically asymptotic in the limit of large n, also referred to as sparsistency [23, 37, 47]. Recently, Tyagi et al. [41] derived algorithms that query f, along with non-asymptotic sampling bounds for identifying $\mathcal{S}_1$. They essentially estimate the (sparse) gradient of f using results from compressed sensing (CS), at few carefully chosen locations in $\mathcal{G}$.

The case $r_0 = 2$. This setup has received relatively less attention than the aforementioned setting. Radchenko et al. [34] proposed an algorithm VANISH, and showed that it is sparsistent, i.e., recovers $\mathcal{S}_1, \mathcal{S}_2$ in the limit of large n. The ACOSSO algorithm [40] can handle this setting, with theoretical guarantees (sparsistency, convergence rates) shown when $r_0 = 1$. Recently, Tyagi et al. [42, 43] derived algorithms that query f, and derived non-asymptotic sampling bounds for recovering $\mathcal{S}_1, \mathcal{S}_2$. Their approach for recovering $\mathcal{S}_2$ was based on estimating the (sparse) Hessian of f using results from CS, at carefully chosen points in $\mathcal{G}$. The special case where f is *multilinear* has been studied considerably; there exist algorithms that recover $\mathcal{S}_1, \mathcal{S}_2$, along with convergence rates for estimating f in the limit of large n [10, 34, 3]. There also exist non-asymptotic sampling bounds for identifying $\mathcal{S}_1, \mathcal{S}_2$ in the noiseless setting (cf., [30, 24]); these works essentially make use of the CS framework.

The general case. Much less is known about the general setup where $r_0 \geq 2$ is possible. Lin et al. [27] were the first to introduce learning SPAMs of the form (1.1), and proposed the COSSO algorithm. Recently, Dalalyan et al. [14] and Yang et al. [49] studied (1.1) in the regression setting and derived non-asymptotic error rates for estimating f. In particular, Dalalyan et al. studied this in the Gaussian white noise model,while Yang et al. considered the Bayesian setup wherein a Gaussian process (GP) prior is placed on f. When f is multilinear, the work of Nazer et al. [30], which is in the CS framework, gives non-asymptotic sampling bounds for recovering $\mathcal{S}_r$, $r = 1, \dots, r_0$.

1.2 Our contributions and main idea

Before proceeding, we will briefly mention our problem setup to put our results in the context; it is described more formally later on in Section 2. We consider $f : [-1,1]^d \rightarrow \mathbb{R}$ of the form (1.1) and denote by $\mathcal{S}_j^{(1)}$ the variables occurring in $\mathcal{S}_j$. We assume, that $\mathcal{S}_j^{(1)}$ are disjoint[2] for $1 \leq j \leq r_0$. Each component ϕ is assumed to be Hölder smooth, and is also assumed to be "sufficiently large" at some point within its domain. Our goal is to query f at few locations in $\mathcal{G} = [-1,1]^d$, and recover the underlying sets of interactions $\mathcal{S}_r$, for each $r = 1, \dots, r_0$.

Our results. To our knowledge, we provide the first non-asymptotic sampling bounds for exact identification of $\mathcal{S}_r$, for each $r = 1, \dots, r_0$, for SPAMs of the form (1.2). In particular, we derive a

[2] For $r_0 = 2$, this represents no additional assumption. See discussion after Proposition 2.

randomized algorithm that with high probability recovers each $\mathcal{S}_r$, $r = 1, \dots, r_0$ with

$$\Omega\left(\sum_{i=3}^{r_0}\Big[\underbrace{c_i^i i^2|\mathcal{S}_i|^2 \log^2 d}_{\text{Identifying } \mathcal{S}_i}\Big] + \underbrace{c_2|\mathcal{S}_2| \log\left(\frac{d^2}{|\mathcal{S}_2|}\right) \log d}_{\text{Identifying } \mathcal{S}_2} + \underbrace{c_1|\mathcal{S}_1| \log\left(\frac{d}{|\mathcal{S}_1|}\right)}_{\text{Identifying } \mathcal{S}_1}\right) \tag{1.2}$$

noiseless queries of f within $[-1, 1]^d$. The same bound holds when the queries are corrupted with arbitrary bounded noise provided the noise magnitude is sufficiently small (see Theorem 8 and Remark 9). Here, the c_i's depend on the smoothness parameters of the ϕ_i's and scale as $\sqrt{i}$ with i. In the setting of i.i.d. Gaussian noise, which we handle by resampling each query sufficiently many times, and averaging, we obtain a similar sample complexity as (1.2) with additional factors depending on the variance of the noise (see Theorem 9).

We improve on the recent work of Tyagi et al. [42, 43], wherein SPAMs with $r_0 = 2$ were considered, by being able to handle general $r_0 \geq 1$. Moreover, we only require f to be Hölder smooth while the algorithms in [42, 43] necessarily require f to be continuously differentiable. Finally, our bounds improve upon those in [42, 43] when the noise is i.i.d. Gaussian. In this scenario, our bounds are linear in the sparsity $|\mathcal{S}_2| + |\mathcal{S}_1|$ while those in [42, 43] are polynomial in the sparsity.

The sampling scheme that we employ to achieve these bounds is novel, and is specifically tailored to the additive nature of f. We believe this scheme to be of independent interest for other problems involving additive models, such as in optimization of high dimensional functions with partially separable structure.

Main idea. We identify each set $\mathcal{S}_i$ in a sequential "top down" manner by first identifying $\mathcal{S}_{r_0}$. Once we find $\mathcal{S}_{r_0}$, the same procedure is repeated on the remaining set of variables (excluding those found in $\mathcal{S}_{r_0}$) to identify $\mathcal{S}_{r_0-1}$, and consequently, each remaining $\mathcal{S}_i$. We essentially perform the following steps for recovering $\mathcal{S}_{r_0}$. Consider some given partition of $[d]$ into r_0 disjoint subsets $\mathcal{A} = (\mathcal{A}_1, \dots, \mathcal{A}_{r_0})$, a Rademacher vector $\boldsymbol{\beta} \in \{-1, 1\}^d$ and some given $\mathbf{x} \in [-1, 1]^d$. We generate 2^{r_0} query points $(\mathbf{x}_i)_{i=1}^{2^{r_0}}$, where each $\mathbf{x}_i$ is constructed using $\boldsymbol{\beta}, \mathbf{x}$ and $\mathcal{A}$. Then, for some fixed sequence of signs $s_1, s_2, \dots, s_{2^{r_0}} \in \{-1, 1\}$ (depending only on r_0), we show for the anchored-ANOVA representation of f (see Section 2 and Lemma 4) that

$$\sum_{i=1}^{2^{r_0}} s_i f(\mathbf{x}_i) = \sum_{(j_1,\dots,j_{r_0}) \in \mathcal{A} \cap \mathcal{S}_{r_0}} \beta_{j_1} \dots \beta_{j_{r_0}} \phi_{(j_1,\dots,j_{r_0})}(x_{j_1}, \dots, x_{j_{r_0}}). \tag{1.3}$$

Observe, that (1.3) corresponds to a *multilinear* measurement of a *sparse* vector with entries $\phi_{(j_1,\dots,j_{r_0})}(x_{j_1}, \dots, x_{j_{r_0}})$, indexed by the tuple $(j_1, \dots, j_{r_0})$. Indeed, this vector is $|\mathcal{S}_{r_0}|$ sparse. This suggests that by repeating the above process at sufficiently many random $\boldsymbol{\beta}$'s, we can recover an estimate of this $|\mathcal{S}_{r_0}|$ vector by using known results from CS. Thereafter, we repeat the above process for each $\mathcal{A}$ corresponding to a family of perfect hash functions (see Definition 1). The size of this set is importantly at most exponential in r_0, and only logarithmic in d. The $\mathbf{x}$'s are then chosen to be points on a uniform r_0 dimensional grid constructed using $\mathcal{A}$. This essentially enables us to guarantee that we are able to sample each $\phi_{(j_1,\dots,j_{r_0})}$ sufficiently fine within its domain, and thus identify $(j_1, \dots, j_{r_0})$ by thresholding.

Organization of paper. The rest of the paper is organized as follows. In Section 2, we set up the notation and also define the problem formally. In Section 3, we begin with the case $r_0 = 1$ as warm up, and describe the sampling scheme, along with the algorithm for this setting. Section 4 considers the bivariate case $r_0 = 2$, while Section 5 consists of the most general setting wherein $r_0 \geq 2$ is possible. Section 6 contains (mostly) known results from compressed sensing for estimating sparse multilinear functions from random samples. Section 7 then puts together the content from the earlier sections, wherein we derive our final theorems. Section 8 consists of a comparison of our results with closely related work, along with some directions for future work.

2 Notation and problem setup

Notation. Scalars will be usually denoted by plain letters (e.g. d), vectors by lowercase boldface letters (e.g., $\mathbf{x}$), matrices by uppercase boldface letters (e.g., $\mathbf{A}$) and sets by uppercase calligraphic letters (e.g., $\mathcal{S}$), with the exception of $[n]$, which denotes the index set $\{1, \dots, n\}$ for any natural number $n \in \mathbb{N}$. For a (column) vector $\mathbf{x} = (x_1 \dots x_d)^T$ and an ordered r-tuple $\mathbf{j} = (j_1, j_2, \dots, j_r) \in \binom{[d]}{r}$ with $1 \leq j_1 < j_2 < \cdots < j_r \leq d$, we denote $\mathbf{x}_\mathbf{j} = (x_{j_1} \dots x_{j_r})^T \in \mathbb{R}^r$ to be the restriction of $\mathbf{x}$ on $\mathbf{j}$. For any finite set $\mathcal{A}$, $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$. Moreover, if $\mathcal{A} \subseteq [d]$, then $\Pi_\mathcal{A}(\mathbf{x})$ denotes the projection of $\mathbf{x}$ on $\mathcal{A}$ where

$$(\Pi_\mathcal{A}(\mathbf{x}))_i = \begin{cases} x_i \ ; & i \in \mathcal{A}, \\ 0 \ ; & i \notin \mathcal{A}, \end{cases} \qquad i \in [d]. \tag{2.1}$$

The ℓ_p norm of a vector $\mathbf{x} \in \mathbb{R}^d$ is defined as $\|\mathbf{x}\|_p := \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$. A random variable β is called Rademacher variable if $\beta = +1$ with probability 1/2 and $\beta = -1$ with probability 1/2. A vector $\boldsymbol{\beta} \in \{-1, +1\}^d$ of independent Rademacher variables is called Rademacher vector. Similarly, a matrix $\mathbf{B} \in \{-1, +1\}^{n \times d}$ is called Rademacher matrix, if all its entries are independent Rademacher variables. For non-negative functions f, g we write $f(x) = \Theta(g(x))$ if there exist constants $c_1, c_2, x_0 > 0$ such that $c_1 g(x) \leq f(x) \leq c_2 g(x)$ for all $x \geq x_0$. Similarly, if there exist constants $c, x_0 > 0$ such that

- $f(x) \leq c g(x)$ for all $x \geq x_0$, then we write $f(x) = O(g(x))$;
- $f(x) \geq c g(x)$ for all $x \geq x_0$, then we write $f(x) = \Omega(g(x))$.

Sparse Additive Models. For an unknown $f : \mathbb{R}^d \to \mathbb{R}$, our aim will be to approximate f uniformly from point queries within a compact domain $\mathcal{G} \subset \mathbb{R}^d$. From now on, we will assume $\mathcal{G} = [-1, 1]^d$. The sets $\mathcal{S}_r \subset \binom{[d]}{r}; r = 1, \dots, r_0$, will represent the interactions amongst the coordinates, with $\mathcal{S}_r$ consisting of r-wise interactions. Our interest will be in the setting where each $\mathcal{S}_r$ is sparse, i.e., $|\mathcal{S}_r| \ll d^r$. Given this setting, we assume to have the following structure

$$f = \sum_{\mathbf{j} \in \mathcal{S}_1} \phi_\mathbf{j} + \sum_{\mathbf{j} \in \mathcal{S}_2} \phi_\mathbf{j} + \cdots + \sum_{\mathbf{j} \in \mathcal{S}_{r_0}} \phi_\mathbf{j}. \tag{2.2}$$

It is important to note here that the components in $\mathcal{S}_r$ will be assumed to be truly r-variate, in the sense that they cannot be written as the sum of lower dimensional functions. For example, we assume that the components in $\mathcal{S}_2$ cannot be expressed as the sum of univariate functions.

Model Uniqueness and ANOVA-decompositions. We note now that the representation of f in (2.2) is not necessarily unique and some additional assumptions are needed to ensure uniqueness. For instance, one could add constants to each ϕ that sum up to zero, thereby giving the same f.

Moreover, if $\mathcal{S}_2$ contains overlapping pairs of variables, then for each such variable – call it p – one could add/subtract functions of the variable x_p to each corresponding $\phi_{\mathbf{j}}$ such that f remains unaltered. To obtain unique representation of f, we will work with the so-called Anchored ANOVA-decomposition of f. We recall its notation and results in the form needed later and refer to [22] for more details.

The usual notation of an ANOVA-decomposition works with functions indexed by subsets of $[d]$, instead of tuples from $[d]$. As there is an obvious one-to-one correspondence between ordered r-tuples and subsets of $[d]$ with r elements, we prefer to give the ANOVA-decomposition in its usual form.

Let $\mu_j, j = 1, \dots, d$ be measures defined on all Borel subsets of $[-1,1]$ and let $U \subseteq [d]$. We let $d\mu_U(\mathbf{x}_U) = \prod_{j\in U} d\mu_j(x_j)$ be the product measure. We define

$$P_U f(\mathbf{x}_U) = \int_{[-1,1]^{d-|U|}} f(\mathbf{x}) d\mu_{[d]\setminus U}(\mathbf{x}_{[d]\setminus U}).$$

The ANOVA-decomposition of f is then given as

$$f(\mathbf{x}) = f_\emptyset + \sum_{i=1}^{d} f_i(x_i) + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} f_{i,j}(x_i, x_j) + \cdots + f_{1,\dots,d}(x_1, \dots, x_d) = \sum_{U\subseteq[d]} f_U(\mathbf{x}_U),$$

where

$$f_U(\mathbf{x}_U) = \sum_{V\subseteq U} (-1)^{|U|-|V|} P_V f(\mathbf{x}_V). \tag{2.3}$$

In the case of $d\mu_j(x_j) = \delta(x_j)dx_j$, where δ is the Dirac distribution, we obtain the Anchored-ANOVA decomposition

$$f(\mathbf{x}) = \sum_{U\subseteq[d]} f_U(\mathbf{x}_U),$$

where $f_\emptyset = f(0)$ and $f_U(\mathbf{x}_U) = 0$ if $x_j = 0$ for some $j \in U$.

The standard theory of ANOVA decompositions is usually based on Hilbert space theory. As we prefer to work with continuous functions, we give the following representation theorem. The proof can be found in the Appendix.

Proposition 1. *Let $f \in C([-1,1]^d)$. Then the collection of $(f_U)_{U\subseteq[d]}$ defined in* (2.3) *is the unique system such that the following holds.*

a) $f_U \in C([-1,1]^{|U|})$;

b) f can be represented as

$$f(\mathbf{x}) = \sum_{U\subseteq[d]} f_U(\mathbf{x}_U), \quad \mathbf{x} \in [-1,1]^d, \tag{2.4}$$

where $\mathbf{x}_U \in [-1,1]^{|U|}$ is the restriction of $\mathbf{x}$ onto indices included in U;

c) $f_U(\mathbf{x}_U) = 0$ if $x_j = 0$ for some $j \in U$.

The Anchored ANOVA-decomposition (2.4) can be used to ensure uniqueness of representation of f of the form (2.2). For the clarity of presentation, we will later distinguish between three settings. The first one is univariate with $r_0 = 1$, the second one with $r_0 = 2$ allows also for bivariate interactions between the variables. Finally, in the multivariate case $r_0 > 2$, arbitrary higher-order interactions can occur. We will present a detailed proposition about the corresponding ANOVA-decomposition in each of the sections separately.

Assumptions. We will specify the assumptions in each of the settings discussed later in more detail. But, in general, we will work with two groups of conditions.

1. *Smoothness.* We will assume throughout the paper that the components of the ANOVA-decomposition are Hölder smooth with exponent $\alpha \in (0,1]$ and constant $L > 0$, i.e.,

$$|\phi(\mathbf{x}) - \phi(\mathbf{y})| \le L\|\mathbf{x} - \mathbf{y}\|_2^\alpha$$

for all admissible $\mathbf{x}, \mathbf{y}$.

2. *Identifiability.* Furthermore, our aim is the identification of the possible interactions between the variables. We are therefore not only interested in the approximation of f but also on the identification of the sets $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{r_0}$. Naturally, this is only possible if the non-zero functions in the Anchored-ANOVA decomposition are significantly large at some point. We will therefore assume that

$$\|\phi\|_\infty = \sup_{\mathbf{x}} |\phi(\mathbf{x})| > D$$

for some $D > 0$.

Problem parameters and goal. Based on the above setup, we will consider our problem specific parameters to be

(a) smoothness parameters: $L > 0$, $\alpha \in (0,1]$,

(b) identifiability parameters: D,

(c) intrinsic/extrinsic dimensions: $d, r_0, |\mathcal{S}_1|, \dots, |\mathcal{S}_{r_0}|$.

These parameters will be assumed to be known by the algorithm. The goal of the algorithm will then be to query f within $[-1,1]^d$, and to identify the sets $\mathcal{S}_1, \dots, \mathcal{S}_{r_0}$ exactly. Using standard methods of approximation theory and sampling along canonical subspaces, one may recover also the components in (2.2). We give some more details on this issue in Section 8.

3 The univariate case

As a warm up, we begin with the relatively simple setting where $r_0 = 1$, meaning that f is a sum of only univariate components. It means that f admits the representation

$$f = \mu + \sum_{p \in \mathcal{S}_1} \phi_p(x_p). \tag{3.1}$$

To ensure the uniqueness of this decomposition, we set $\mu = f(0)$ and assume that $\phi_p(0) = 0$ for all $p \in \mathcal{S}_1$.

Assumptions. We will make the following assumptions on the model (3.1).

1. *Smoothness.* The terms in (3.1) are Hölder continuous with parameters $L > 0, \alpha \in (0,1]$, i.e.,

$$|\phi_p(x) - \phi_p(y)| \le L|x - y|^\alpha \quad \text{for all} \quad p \in \mathcal{S}_1 \quad \text{and all} \quad x, y \in [-1,1].$$

2. *Identifiability.* For every $p \in \mathcal{S}_1$ there is an $x_p^* \in [-1,1]$, such that $|\phi_p(x_p^*)| > D_1$.

Sampling scheme. Our sampling scheme is motivated by the following simple observation. For any fixed $\mathbf{x} \in [-1,1]^d$, and some $\boldsymbol{\beta} \in \{-1,+1\}^d$, consider the points $\mathbf{x}^+, \mathbf{x}^- \in [-1,1]^d$ defined as

$$x_i^+ = \begin{cases} x_i \; ; & \beta_i = +1, \\ 0 \; ; & \beta_i = -1 \end{cases} \quad \text{and} \quad x_i^- = \begin{cases} 0 \; ; & \beta_i = +1, \\ x_i \; ; & \beta_i = -1, \end{cases} \quad i \in [d]. \tag{3.2}$$

Upon querying f at $\mathbf{x}^+, \mathbf{x}^-$, we obtain the noisy samples

$$\tilde{f}(\mathbf{x}^+) = f(\mathbf{x}^+) + \eta^+, \quad \tilde{f}(\mathbf{x}^-) = f(\mathbf{x}^-) + \eta^-,$$

where $\eta^+, \eta^- \in \mathbb{R}$ denotes the noise. One can then easily verify that the following identity holds on account of the structure of f

$$\tilde{f}(\mathbf{x}^+) - \tilde{f}(\mathbf{x}^-) = \sum_{i \in \mathcal{S}_1} \beta_i \underbrace{\phi_i(x_i)}_{z_i^*(x_i)} + \eta^+ - \eta^- = \langle \boldsymbol{\beta}, \mathbf{z}^*(\mathbf{x}) \rangle + \eta^+ - \eta^-. \tag{3.3}$$

Note that $\mathbf{z}^*(\mathbf{x}) = (z_1^*(x_1) \dots z_d^*(x_d))^T$ is $|\mathcal{S}_1|$ sparse, and $\tilde{f}(\mathbf{x}^+) - \tilde{f}(\mathbf{x}^-)$ corresponds to a noisy linear measurement of $\mathbf{z}^*(\mathbf{x})$, with $\boldsymbol{\beta}$. From standard compressive sensing results, we know that a sparse vector can be recovered *stably*, from only a few noisy linear measurements with random vectors, drawn from a suitable distribution. In particular, it is well established that random Rademacher measurements satisfy this criteria. We discuss this separately later on, for now it suffices to assume that we have at hand an appropriate sparse recovery algorithm: SPARSE-REC.

We thus generate independent Rademacher vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_n \in \{-1,+1\}^d$. For each $\boldsymbol{\beta}_i$, we create $\mathbf{x}_i^+, \mathbf{x}_i^-$ as described in (3.2) (for some fixed $\mathbf{x}$), and obtain $\tilde{f}(\mathbf{x}_i^+), \tilde{f}(\mathbf{x}_i^-)$. Then, (3.3) gives us the linear system

$$\underbrace{\begin{pmatrix} \tilde{f}(\mathbf{x}_1^+) - \tilde{f}(\mathbf{x}_1^-) \\ \vdots \\ \vdots \\ \tilde{f}(\mathbf{x}_n^+) - \tilde{f}(\mathbf{x}_n^-) \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} \boldsymbol{\beta}_1^T \\ \vdots \\ \vdots \\ \boldsymbol{\beta}_n^T \end{pmatrix}}_{\mathbf{B}} \mathbf{z}^*(\mathbf{x}) + \underbrace{\begin{pmatrix} \eta_1^+ - \eta_1^- \\ \vdots \\ \vdots \\ \eta_n^+ - \eta_n^- \end{pmatrix}}_{\boldsymbol{\eta}}. \tag{3.4}$$

SPARSE-REC will take as input $\mathbf{y}, \mathbf{B}$, and will output an estimate $\widehat{\mathbf{z}^*}(\mathbf{x})$ to $\mathbf{z}^*(\mathbf{x})$. As will be shown formally in Section 6, one can choose SPARSE-REC as a ℓ_1 minimization (convex) program for which it is well known from the compressive sensing literature that if n is sufficiently large, then we will have for some $\epsilon \geq 0$ depending on $\|\boldsymbol{\eta}\|_\infty$ that $\|\widehat{\mathbf{z}^*}(\mathbf{x}) - \mathbf{z}^*(\mathbf{x})\|_\infty \leq \epsilon$ holds. In such a case, we will refer to SPARSE-REC as being "ϵ-accurate" at $\mathbf{x}$. Also, we remark that the choice for SPARSE-REC that we consider in Section 6 will need an upper bound estimate of the noise level $\boldsymbol{\eta}$ (in a suitable norm).

Given the above, we now describe how to choose $\mathbf{x} \in [-1,1]^d$. To this end, we adopt the approach of [41], where the following grid on the diagonal of $[-1,1]^d$ was considered

$$\chi := \left\{ \mathbf{x} = (x \; x \; \cdots \; x)^T \in \mathbb{R}^d : x \in \left\{ -1, -\frac{m-1}{m}, \dots, \frac{m-1}{m}, 1 \right\} \right\}. \tag{3.5}$$

Our aim will be to obtain the estimate $\widehat{\mathbf{z}^*}(\mathbf{x})$ at each $\mathbf{x} \in \chi$. Note that this gives us estimates to $\phi_p(x_p)$ for $p = 1, \dots, d$, with x_p lying on a uniform one dimensional grid in $[-1,1]$. Thus we can see, at least intuitively, that provided ϵ is small enough, and the grid is fine enough (so that we are close to $\phi_p(x_p^*)$ for each $p \in \mathcal{S}_1$), we will be able to detect each $p \in \mathcal{S}_1$ by thresholding.

Algorithm 1 Algorithm for estimating $\mathcal{S}_1$

```
Input: $d$, $|\mathcal{S}_1|$, $m$, $n$, $\epsilon$.
Initialization: $\widehat{\mathcal{S}_1} = \emptyset$.
Output: $\widehat{\mathcal{S}_1}$.

Construct $\chi$ as defined in (3.5) with $|\chi| = 2m+1$.
Generate Rademacher vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_n \in \{-1,+1\}^d$.
Form $\mathbf{B} \in \mathbb{R}^{n \times d}$ as in (3.4).
for $\mathbf{x} \in \chi$ do
    Generate $\mathbf{x}_i^+, \mathbf{x}_i^- \in [-1,1]^d$, as in (3.2), using $\mathbf{x}, \boldsymbol{\beta}_i$ for each $i \in [n]$.
    Using the samples $(\tilde{f}(\mathbf{x}_i^+), \tilde{f}(\mathbf{x}_i^-))_{i=1}^n$, form $\mathbf{y}$ as in (3.4).
    Obtain $\widehat{\mathbf{z}}^*(\mathbf{x}) = \texttt{SPARSE-REC}(\mathbf{y}, \mathbf{B})$.
    Update $\widehat{\mathcal{S}_1} = \widehat{\mathcal{S}_1} \cup \{p \in [d] : |(\widehat{\mathbf{z}}^*(\mathbf{x}))_p| > \epsilon\}$.
end for
```

Algorithm outline and guarantees The discussion above is outlined formally in the form of Algorithm 1. Lemma 1 below provides formal guarantees for exact recovery of support $\mathcal{S}_1$.

Lemma 1. *Let* SPARSE-REC *be ϵ-accurate for each $\mathbf{x} \in \chi$ with $\epsilon < D_1/3$, which uses n linear measurements. Then for $m \geq (3L/D_1)^{1/\alpha}$, Algorithm 1 recovers $\mathcal{S}_1$ exactly, i.e., $\widehat{\mathcal{S}_1} = \mathcal{S}_1$. Moreover, the total number of queries of f is $2(2m+1)n$.*

Proof. Recall that we denote $\mathbf{z}^*(\mathbf{x}) = (\phi_1(x_1) \dots \phi_d(x_d))^T$, and $z_i^*(x_i) = \phi_i(x_i)$. For any given $p \in \mathcal{S}_1$, we know that there exists $x_p^* \in [-1,1]$ such that $|\phi_p(x_p^*)| > D_1$. Also, on account of the construction of χ, there exists $\mathbf{x} = (x \dots x)^T \in \chi$ such that $|x - x_p^*| \leq 1/m$. Then starting with the fact that SPARSE-REC is ϵ accurate at $\mathbf{x}$, we obtain

$$
\begin{aligned}
|\widehat{z_p^*}(x)| &\geq |\phi_p(x)| - \epsilon \geq |\phi_p(x_p^*)| - |\phi_p(x_p^*) - \phi_p(x)| - \epsilon \\
&\geq D_1 - \frac{L}{m^\alpha} - \epsilon \geq \frac{2D_1}{3} - \epsilon.
\end{aligned}
$$

We used the reverse triangle inequality and the identifiability and smoothness assumptions on ϕ_p. On the other hand, since SPARSE-REC is ϵ accurate at each point in χ, therefore for every $q \notin \mathcal{S}_1$ and $(c\ c\ \dots c) \in \chi$, we know that $|\widehat{z_q^*}(c)| \leq \epsilon$. It then follows readily for the stated choice of m, ϵ that $\widehat{\mathcal{S}_1}$ contains each variable in $\mathcal{S}_1$, and none from $\mathcal{S}_1^c$. □

4 The bivariate case

Next, we consider the scenario where $r_0 = 2$, i.e., f can be written as a sum of univariate and bivariate functions. We denote by $\mathcal{S}_2^{\text{var}} = \mathcal{S}_2^{(1)}$ the set of variables which are part of a 2-tuple in $\mathcal{S}_2$. Inserting this restriction into Proposition 1, we derive the following uniqueness result (its proof is postponed to the Appendix).

Proposition 2. *Let $f \in C([-1,1]^d)$ be of the form*

$$
f = \mu + \sum_{p \in \mathcal{S}_1} \phi_p(x_p) + \sum_{\mathbf{j} \in \mathcal{S}_2} \phi_{\mathbf{j}}(x_{\mathbf{j}}) + \sum_{l \in \mathcal{S}_2^{\text{var}}} \phi_l(x_l), \tag{4.1}
$$

where $\mathcal{S}_1 \cap \mathcal{S}_2^{\text{var}} = \emptyset$. Moreover, let *a)* $\mu = f(0)$,

b) $\phi_j(0) = 0$ *for all* $j \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$,

c) $\phi_{\mathbf{j}}(x_{\mathbf{j}}) = 0$ *if* $\mathbf{j} = (j_1, j_2) \in \mathcal{S}_2$ *and* $x_{j_1} = 0$ *or* $x_{j_2} = 0$.

Then the representation (4.1) *of* f *is unique in the sense that each component in* (4.1) *is uniquely identifiable.*

Remark 1. *In* (4.1)*, we could have "collapsed" the terms corresponding to variables* l *in* $\sum_{l \in \mathcal{S}_2^{\text{var}}} \phi_l(x_l)$ *– for* l *occurring exactly once in* $\mathcal{S}_2$ *– uniquely into the corresponding component* $\phi_{\mathbf{j}}(x_{\mathbf{j}})$*. A similar approach was adopted in [43], and the resulting model was shown to be uniquely identifiable. Yet here, we choose to represent* f *in the form* (4.1) *for convenience, and clarity of notation. This also leads to a less cumbersome expression, when we work with general interaction terms later.*

Assumptions. We now make the following assumptions on the model (4.1).

1. *Smoothness.* We assume each term in (4.1) to be Hölder continuous with parameters $L > 0, \alpha \in (0, 1]$, i.e.,

$$|\phi_p(x) - \phi_p(y)| \leq L|x - y|^\alpha \quad \text{for all } p \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}} \text{ and for all } x, y \in [-1, 1],$$
$$|\phi_{\mathbf{j}}(\mathbf{x}) - \phi_{\mathbf{j}}(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2^\alpha \quad \text{for all } \mathbf{j} \in \mathcal{S}_2 \text{ and for all } \mathbf{x}, \mathbf{y} \in [-1, 1]^2.$$

2. *Identifiability of* $\mathcal{S}_1, \mathcal{S}_2$*.* We assume that for each $p \in \mathcal{S}_1$, there exists $x_p^* \in [-1, 1]$ so that $|\phi_p(x_p^*)| > D_1$ for some constant $D_1 > 0$. Furthermore, we assume that for each $\mathbf{j} \in \mathcal{S}_2$ there exists $\mathbf{x}_{\mathbf{j}}^* \in [-1, 1]^2$ such that $|\phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}^*)| > D_2$.

Remark 2. *We consider the same* α, L *for all components in* $\mathcal{S}_1, \mathcal{S}_2$ *for the ease of exposition only. One can also consider the parameters* α_i, L_i *for the components in* $\mathcal{S}_i$*. This also applies to the general multivariate setting in Section 5.*

Before describing our sampling scheme, we need some additional notation. For any $\beta \in \{-1, 1\}$, we denote $\bar{\beta} = (-\beta)$. Moreover, $\mathbb{1}_\beta$ denotes the indicator variable of β, i.e., $\mathbb{1}_\beta = 1$ if $\beta = 1$, and $\mathbb{1}_\beta = 0$ if $\beta = -1$. Overall, our scheme proceeds in two stages. We first identify $\mathcal{S}_2$, and only then identify $\mathcal{S}_1$.

Sampling lemma for identifying $\mathcal{S}_2$**.** We begin by providing the motivation behind our sampling scheme for identifying $\mathcal{S}_2$. Consider some fixed mapping $h : [d] \to \{1, 2\}$ that partitions $[d]$ into $\mathcal{A}_1 = \{i \in [d] : h(i) = 1\}$ and $\mathcal{A}_2 = \{i \in [d] : h(i) = 2\}$. Then for a given Rademacher vector $\boldsymbol{\beta} \in \{-1, 1\}^d$ and $\mathbf{x} = (x_1 \dots x_d)^T \in [-1, 1]^d$, consider the points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4 \in [-1, 1]^d$ defined as

$$\begin{array}{ll} x_{1,i} = \begin{cases} \mathbb{1}_{\beta_i} x_i \ ; & i \in \mathcal{A}_1, \\ \mathbb{1}_{\beta_i} x_i \ ; & i \in \mathcal{A}_2, \end{cases} & x_{2,i} = \begin{cases} \mathbb{1}_{\bar{\beta}_i} x_i \ ; & i \in \mathcal{A}_1, \\ \mathbb{1}_{\beta_i} x_i \ ; & i \in \mathcal{A}_2, \end{cases} \\ x_{3,i} = \begin{cases} \mathbb{1}_{\beta_i} x_i \ ; & i \in \mathcal{A}_1, \\ \mathbb{1}_{\bar{\beta}_i} x_i \ ; & i \in \mathcal{A}_2, \end{cases} & x_{4,i} = \begin{cases} \mathbb{1}_{\bar{\beta}_i} x_i \ ; & i \in \mathcal{A}_1, \\ \mathbb{1}_{\bar{\beta}_i} x_i \ ; & i \in \mathcal{A}_2, \end{cases} \end{array} \quad i \in [d]. \tag{4.2}$$

The following lemma is the key motivation behind our sampling scheme.

Lemma 2. *Denote* $\mathcal{A} = \left\{\mathbf{j} \in \binom{[d]}{2} : \mathbf{j} \in \{\mathcal{A}_1 \times \mathcal{A}_2\} \cup \{\mathcal{A}_2 \times \mathcal{A}_1\}\right\}$. *Then for functions* f *of the form* (4.1)*, we have that*

$$\begin{aligned} f(\mathbf{x}_1) - f(\mathbf{x}_2) - f(\mathbf{x}_3) + f(\mathbf{x}_4) &= \sum_{\mathbf{j} \in \mathcal{S}_2 : j_1 \in \mathcal{A}_1, j_2 \in \mathcal{A}_2} \beta_{j_1}\beta_{j_2}\phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}) + \sum_{\mathbf{j} \in \mathcal{S}_2 : j_1 \in \mathcal{A}_2, j_2 \in \mathcal{A}_1} \beta_{j_1}\beta_{j_2}\phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}) \quad (4.3) \\ &= \sum_{\mathbf{j} \in \mathcal{A} \cap \mathcal{S}_2} \beta_{j_1}\beta_{j_2}\phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}). \end{aligned}$$

Proof. For any $\mathbf{j} \in \mathcal{S}_2$, let us first consider the case where j_1, j_2 lie in different sets. For example, let $j_1 \in \mathcal{A}_1$ and $j_2 \in \mathcal{A}_2$. Then the contribution of $\phi_{\mathbf{j}}$ to the left-hand side of (4.3) turns out to be for all possible values of $\beta_{j_1}, \beta_{j_2} \in \{-1, +1\}$ equal to

$$\begin{aligned} &\phi_{\mathbf{j}}(\mathbf{x}_{1,\mathbf{j}}) - \phi_{\mathbf{j}}(\mathbf{x}_{2,\mathbf{j}}) - \phi_{\mathbf{j}}(\mathbf{x}_{3,\mathbf{j}}) + \phi_{\mathbf{j}}(\mathbf{x}_{4,\mathbf{j}}) \\ &= \phi_{\mathbf{j}}(\mathbb{1}_{\beta_{j_1}} x_{j_1}, \mathbb{1}_{\beta_{j_2}} x_{j_2}) - \phi_{\mathbf{j}}(\mathbb{1}_{\bar{\beta}_{j_1}} x_{j_1}, \mathbb{1}_{\beta_{j_2}} x_{j_2}) - \phi_{\mathbf{j}}(\mathbb{1}_{\beta_{j_1}} x_{j_1}, \mathbb{1}_{\bar{\beta}_{j_2}} x_{j_2}) + \phi_{\mathbf{j}}(\mathbb{1}_{\bar{\beta}_{j_1}} x_{j_1}, \mathbb{1}_{\bar{\beta}_{j_2}} x_{j_2}) \\ &= \beta_{j_1}\beta_{j_2}(\phi_{\mathbf{j}}(x_{j_1}, x_{j_2}) - \phi_{\mathbf{j}}(x_{j_1}, 0) - \phi_{\mathbf{j}}(0, x_{j_2}) + \phi_{\mathbf{j}}(0,0)) = \beta_{j_1}\beta_{j_2}\phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}). \end{aligned}$$

In case $j_1 \in \mathcal{A}_2$ and $j_2 \in \mathcal{A}_1$, then the contribution of $\phi_{\mathbf{j}}$ turns out to be the same as above. Since f is additive over $\mathbf{j} \in \mathcal{S}_2$, thus the total contribution of $\mathcal{S}_2$ is given by the right-hand side of (4.3).

Now, for all $\mathbf{j} \in \mathcal{S}_2$ with j_1, j_2 lying in the same set, the contribution of $\phi_{\mathbf{j}}$ turns out to be zero. Indeed, if $j_1, j_2 \in \mathcal{A}_1$, the contribution of $\phi_{\mathbf{j}}$ is

$$\begin{aligned} &\phi_{\mathbf{j}}(\mathbf{x}_{1,\mathbf{j}}) - \phi_{\mathbf{j}}(\mathbf{x}_{2,\mathbf{j}}) - \phi_{\mathbf{j}}(\mathbf{x}_{3,\mathbf{j}}) + \phi_{\mathbf{j}}(\mathbf{x}_{4,\mathbf{j}}) \\ &= \phi_{\mathbf{j}}(\mathbb{1}_{\beta_{j_1}} x_{j_1}, \mathbb{1}_{\beta_{j_2}} x_{j_2}) - \phi_{\mathbf{j}}(\mathbb{1}_{\bar{\beta}_{j_1}} x_{j_1}, \mathbb{1}_{\bar{\beta}_{j_2}} x_{j_2}) - \phi_{\mathbf{j}}(\mathbb{1}_{\beta_{j_1}} x_{j_1}, \mathbb{1}_{\beta_{j_2}} x_{j_2}) + \phi_{\mathbf{j}}(\mathbb{1}_{\bar{\beta}_{j_1}} x_{j_1}, \mathbb{1}_{\bar{\beta}_{j_2}} x_{j_2}) = 0. \end{aligned}$$

The same is easily verified if $j_1, j_2 \in \mathcal{A}_2$. Lastly, let us verify that the contribution of ϕ_p for each $p \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$ is zero. Indeed, when $p \in \mathcal{A}_1$, we get

$$\phi_p(x_{1,p}) - \phi_p(x_{2,p}) - \phi_p(x_{3,p}) + \phi_p(x_{4,p}) = \phi_p(\mathbb{1}_{\beta_p} x_p) - \phi_p(\mathbb{1}_{\bar{\beta}_p} x_p) - \phi_p(\mathbb{1}_{\beta_p} x_p) + \phi_p(\mathbb{1}_{\bar{\beta}_p} x_p) = 0$$

and the same is true also for $p \in \mathcal{A}_2$. This completes the proof. □

Denoting $z^*_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}) = \phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}})$ if $\mathbf{j} \in \mathcal{S}_2$ and 0 otherwise, let $\mathbf{z}^*(\mathbf{x}) \in \mathbb{R}^{\binom{[d]}{2}}$ be the corresponding ($|\mathcal{S}_2|$ sparse) vector. For $\mathcal{A} \subseteq \binom{[d]}{2}$ we denote $\mathbf{z}^*(\mathbf{x}; \mathcal{A}) \in \mathbb{R}^{\binom{[d]}{2}}$ to be the *projection* of $\mathbf{z}^*(\mathbf{x})$ onto $\mathcal{A}$. Clearly $\mathbf{z}^*(\mathbf{x}; \mathcal{A})$ is at most $|\mathcal{S}_2|$ sparse too – it is in fact $|\mathcal{S}_2 \cap \mathcal{A}|$ sparse. For a Rademacher vector $\boldsymbol{\beta} \in \{-1, +1\}^d$, let $\boldsymbol{\beta}^{(2)} \in \{-1, +1\}^{\binom{[d]}{2}}$, where $\beta^{(2)}_{\mathbf{j}} = \beta_{j_1}\beta_{j_2}$ for each $\mathbf{j} = (j_1, j_2)$. Hence we see that (4.3) corresponds to a linear measurement of $\mathbf{z}^*(\mathbf{x}; \mathcal{A})$ with the Rademacher vector $\boldsymbol{\beta}^{(2)}$.

Sampling scheme for identifying $\mathcal{S}_2$. We first generate independent Rademacher vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n \in \{-1, 1\}^d$. Then for some fixed $\mathbf{x} \in [-1, 1]^d$ and a mapping $h : [d] \to \{1, 2\}$ – the choice of both to be made clear later – we obtain the samples $\tilde{f}(\mathbf{x}_{i,p}) = f(\mathbf{x}_{i,p}) + \eta_{i,p}$, $i \in [n]$ and $p \in \{1, 2, 3, 4\}$. Here, $\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \mathbf{x}_{i,3}, \mathbf{x}_{i,4}$ are generated using $\mathbf{x}, \boldsymbol{\beta}_i, h$ as outlined in (4.2). As a direct implication of Lemma 2, we obtain the linear system

$$\underbrace{\begin{pmatrix} \tilde{f}(\mathbf{x}_{1,1}) - \tilde{f}(\mathbf{x}_{1,2}) - \tilde{f}(\mathbf{x}_{1,3}) + \tilde{f}(\mathbf{x}_{1,4}) \\ \vdots \\ \vdots \\ \tilde{f}(\mathbf{x}_{n,1}) - \tilde{f}(\mathbf{x}_{n,2}) - \tilde{f}(\mathbf{x}_{n,3}) + \tilde{f}(\mathbf{x}_{n,4}) \end{pmatrix}}_{\mathbf{y} \in \mathbb{R}^n} = \underbrace{\begin{pmatrix} {\boldsymbol{\beta}_1^{(2)}}^T \\ \vdots \\ \vdots \\ {\boldsymbol{\beta}_n^{(2)}}^T \end{pmatrix}}_{\mathbf{B} \in \mathbb{R}^{n \times \binom{d}{2}}} \mathbf{z}^*(\mathbf{x}; \mathcal{A}) + \underbrace{\begin{pmatrix} \eta_{1,1} - \eta_{1,2} - \eta_{1,3} + \eta_{1,4} \\ \vdots \\ \vdots \\ \eta_{n,1} - \eta_{n,2} - \eta_{n,3} + \eta_{n,4} \end{pmatrix}}_{\boldsymbol{\eta} \in \mathbb{R}^n}. \quad (4.4)$$

By feeding $\mathbf{y}, \mathbf{B}$ as input to `SPARSE-REC`, we then obtain the estimate $\widehat{\mathbf{z}^*}(\mathbf{x}; \mathcal{A})$ to $\mathbf{z}^*(\mathbf{x}; \mathcal{A})$. Assuming `SPARSE-REC` to be ϵ-accurate at $\mathbf{x}$, we will have that $\|\widehat{\mathbf{z}^*}(\mathbf{x}; \mathcal{A}) - \mathbf{z}^*(\mathbf{x}; \mathcal{A})\|_\infty \leq \epsilon$ holds. Let us mention that $\mathcal{A}$ from Lemma 2 is completely determined by h but we avoid denoting this explicitly for clarity of notation.

At this point, it is natural to ask, how one should choose $\mathbf{x}$ and the mapping h. To this end, we borrow the approach of [15], which involves choosing h from a family of hash functions, and creating for each h in the family a uniform grid. To begin with, we introduce the following definition of a family of hash functions.

Definition 1. *For some $t \in \mathbb{N}$ and $j = 1, 2, \dots,$ let $h_j : [d] \to \{1, 2, \dots, t\}$. We call the family of hash functions $\mathcal{H}_t^d = (h_1, h_2, \dots)$ a (d, t)-hash family if for any distinct $i_1, i_2, \dots, i_t \in [d]$, there exists $h \in \mathcal{H}_t^d$ such that h is an injection when restricted to $i_1, i_2, \dots, i_t$.*

Hash functions are commonly used in theoretical computer science and are widely used in finding juntas [29]. One can construct $\mathcal{H}_t^d$ of size $O(te^t \log d)$ using a standard probabilistic argument. The reader is for instance referred to Section 5 in [15], where for any constant $C_1 > 1$ the probabilistic construction yields $\mathcal{H}_t^d$ of size $|\mathcal{H}_t^d| \leq (C_1 + 1)te^t \log d$ with probability at least $1 - d^{-C_1 t}$, in time linear in the output size.

Focusing on the setting $t = 2$ now, say we have at hand a family $\mathcal{H}_2^d$ of size $O(\log d)$. Then for any $(i, j) \in \binom{[d]}{2}$, there exists $h \in \mathcal{H}_2^d$ so that $h(i) \neq h(j)$. For each $h \in \mathcal{H}_2^d$, let us define $\mathbf{e}_1(h), \mathbf{e}_2(h) \in \mathbb{R}^d$, where

$$(\mathbf{e}_i(h))_q := \begin{cases} 1\ ; & h(q) = i, \\ 0\ ; & \text{otherwise} \end{cases} \qquad \text{for } i = 1, 2 \text{ and } q \in [d].$$

Then we create a two dimensional grid with respect to h

$$\chi(h) := \left\{ \mathbf{x} \in [-1, 1]^d : \mathbf{x} = c_1 \mathbf{e}_1(h) + c_2 \mathbf{e}_2(h); c_1, c_2 \in \left\{ -1, -\frac{m-1}{m}, \dots, \frac{m-1}{m}, 1 \right\} \right\}. \tag{4.5}$$

Equipped with $\chi(h)$ for each $h \in \mathcal{H}_2^d$, we now possess the following approximation property. For any $\mathbf{j} \in \binom{[d]}{2}$ and any $(x_{j_1}^*, x_{j_2}^*) \in [-1, 1]^2$, there exists $h \in \mathcal{H}_2^d$ with $h(j_1) \neq h(j_2)$ and a corresponding $\mathbf{x} \in \chi(h)$ so that $|x_{j_1}^* - x_{j_1}|, |x_{j_2}^* - x_{j_2}| \leq 1/m$.

Informally speaking, our idea is the following. Assume that `SPARSE-REC` is ϵ-accurate for each $h \in \mathcal{H}_2^d$, $\mathbf{x} \in \chi(h)$. Also, say m, ϵ are sufficiently large and small respectively. Hence, if we estimate $\mathbf{z}^*(\mathbf{x}; \mathcal{A})$ at each $h \in \mathcal{H}_2^d$ and $\mathbf{x} \in \chi(h)$, then for every $\mathbf{j} \in \mathcal{S}_2$, we are guaranteed to have a point $\mathbf{x}$ at which the estimate $|\widehat{z_{\mathbf{j}}^*}(\mathbf{x_j})|$ is sufficiently large. Moreover, for every $\mathbf{j} \notin \mathcal{S}_2$, we would always (i.e., for each $h \in \mathcal{H}_2^d$ and $\mathbf{x} \in \chi(h)$) have $|\widehat{z_{\mathbf{j}}^*}(\mathbf{x_j})|$ sufficiently small; more precisely, $|\widehat{z_{\mathbf{j}}^*}(\mathbf{x_j})| \leq \epsilon$ since $\phi_{\mathbf{j}} \equiv 0$. Consequently, we will be able to identify $\mathcal{S}_2$ by thresholding, via a suitable threshold.

Sampling scheme for identifying $\mathcal{S}_1$. Assuming $\mathcal{S}_2$ is identified, the model (4.1) reduces to the univariate case on the reduced set $\mathcal{P} := [d] \setminus \widehat{\mathcal{S}_2^{\text{var}}}$ with $\mathcal{S}_1 \subset \mathcal{P}$. We can therefore apply Algorithm 1 on $\mathcal{P}$ by setting the coordinates in $\mathcal{P}^c = \widehat{\mathcal{S}_2^{\text{var}}}$ to zero. Indeed, we first construct for some $m \in \mathbb{N}$ the following set

$$\chi = \left\{ (c\ c\ \dots\ c)^T \in \mathbb{R}^{\mathcal{P}} : c \in \left\{ -1, -\frac{m-1}{m}, \dots, \frac{m-1}{m}, 1 \right\} \right\} \subset [-1, 1]^{\mathcal{P}}. \tag{4.6}$$

Then, for any given $\boldsymbol{\beta} \in \{-1, 1\}^{\mathcal{P}}$, and $\mathbf{x} \in \chi$, we construct $\mathbf{x}^+, \mathbf{x}^- \in \mathbb{R}^d$ using $\boldsymbol{\beta}, \mathbf{x}$ as follows

$$x_i^+ = \begin{cases} x_i\ ; & \beta_i = +1 \text{ and } i \in \mathcal{P}, \\ 0\ ; & \text{otherwise,} \end{cases} \qquad x_i^- = \begin{cases} x_i\ ; & \beta_i = -1 \text{ and } i \in \mathcal{P}, \\ 0\ ; & \text{otherwise,} \end{cases} \qquad i \in [d]. \tag{4.7}$$

Note that $x_i^+, x_i^- = 0$ for $i \notin \mathcal{P}$. Then, similarly to (3.3), we have that

$$\tilde{f}(\mathbf{x}^+) - \tilde{f}(\mathbf{x}^-) = \sum_{i \in \mathcal{P}} \beta_i \underbrace{\phi_i(x_i)}_{z_i^*(x_i)} + \eta^+ - \eta^- = \langle \boldsymbol{\beta}, \mathbf{z}^*_{\mathcal{P}}(\mathbf{x}) \rangle + \eta^+ - \eta^-, \tag{4.8}$$

where $\mathbf{z}^*_{\mathcal{P}}(\mathbf{x}) \in \mathbb{R}^{\mathcal{P}}$ is the restriction of $\mathbf{z}^*(\mathbf{x})$ onto $\mathcal{P}$, and is $|\mathcal{S}_1|$ sparse. Thereafter, we proceed as in Algorithm 1 by forming a linear system as in (3.4) (where now $\mathbf{B} \in \mathbb{R}^{n \times |\mathcal{P}|}$) at each $\mathbf{x} \in \chi$, and employing an ϵ-accurate SPARSE-REC to estimate $\mathbf{z}^*_{\mathcal{P}}(\mathbf{x})$.

Algorithm outline and guarantees. Our scheme for identifying $\mathcal{S}_2$ is outlined formally as the first part of Algorithm 2. The second part involves the estimation of $\mathcal{S}_1$. Lemma 3 provides exact recovery guarantees for $\mathcal{S}_2$ and $\mathcal{S}_1$ by Algorithm 2.

Algorithm 2 Algorithm for estimating $\mathcal{S}_2, \mathcal{S}_1$

```
Input: $d$, $|\mathcal{S}_2|$, $m_2$, $n_2$, $\epsilon_2$.                // Estimation of $\mathcal{S}_2$
Initialization: $\widehat{\mathcal{S}_2} = \emptyset$.
Output: $\widehat{\mathcal{S}_2}$.

Generate independent Rademacher vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_{n_2} \in \{-1,+1\}^d$.
Form $\mathbf{B} \in \mathbb{R}^{n_2 \times \binom{d}{2}}$ as in (4.4).
Construct a $(d,2)$ hash family: $\mathcal{H}_2^d$.
for $h \in \mathcal{H}_2^d$ do
    Construct $\chi(h)$ as defined in (4.5) with $|\chi(h)| = (2m_2+1)^2$.
    for $\mathbf{x} \in \chi(h)$ do
        Generate $\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \mathbf{x}_{i,3}, \mathbf{x}_{i,4} \in [-1,1]^d$, as in (4.2), using $\mathbf{x}, \boldsymbol{\beta}_i$ for each $i \in [n_2]$.
        Using the samples $(\tilde{f}(\mathbf{x}_{i,1}), \tilde{f}(\mathbf{x}_{i,2}), \tilde{f}(\mathbf{x}_{i,3}), \tilde{f}(\mathbf{x}_{i,4}))_{i=1}^{n_2}$, form $\mathbf{y}$ as in (4.4).
        Obtain $\widehat{\mathbf{z}^*}(\mathbf{x}; \mathcal{A}) = \text{SPARSE-REC}_2(\mathbf{y}, \mathbf{B})$.
        Update $\widehat{\mathcal{S}_2} = \widehat{\mathcal{S}_2} \cup \left\{ \mathbf{j} \in \mathcal{A} : |\widehat{z^*_{\mathbf{j}}}(\mathbf{x}_{\mathbf{j}})| > \epsilon_2 \right\}$.
    end for
end for

Input: $d$, $|\mathcal{S}_1|$, $\widehat{\mathcal{S}_2^{\text{var}}}$, $m_1$, $n_1$, $\epsilon_1$.      // Estimation of $\mathcal{S}_1$
Initialization: $\widehat{\mathcal{S}_1} = \emptyset$, $\mathcal{P} = [d] \setminus \widehat{\mathcal{S}_2^{\text{var}}}$.
Output: $\widehat{\mathcal{S}_1}$.
Construct $\chi \subset [-1,1]^{\mathcal{P}}$ with $|\chi| = 2m_1 + 1$, as in (4.6).
Generate independent Rademacher vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_{n_1} \in \{-1,+1\}^{\mathcal{P}}$.
Form $\mathbf{B} \in \mathbb{R}^{n_1 \times |\mathcal{P}|}$ as in (3.4).
for $\mathbf{x} \in \chi$ do
    Generate $\mathbf{x}_i^+, \mathbf{x}_i^- \in [-1,1]^d$, as in (4.7), using $\mathbf{x}, \boldsymbol{\beta}_i$ for each $i \in [n_1]$.
    Using the samples $(\tilde{f}(\mathbf{x}_i^+), \tilde{f}(\mathbf{x}_i^-))_{i=1}^{n_1}$, form $\mathbf{y}$ as in (3.4).
    Obtain $\widehat{\mathbf{z}^*}(\mathbf{x}) = \text{SPARSE-REC}_1(\mathbf{y}, \mathbf{B})$ where $\widehat{\mathbf{z}^*}(\mathbf{x}) \in \mathbb{R}^{\mathcal{P}}$.
    Update $\widehat{\mathcal{S}_1} = \widehat{\mathcal{S}_1} \cup \left\{ p \in \mathcal{P} : |(\widehat{\mathbf{z}^*}(\mathbf{x}))_p| > \epsilon_1 \right\}$.
end for
```

Lemma 3. *Let $\mathcal{H}_2^d$ be a $(d,2)$ hash family, and let SPARSE-REC$_2$ be ϵ_2-accurate for each $h \in \mathcal{H}_2^d$, $\mathbf{x} \in \chi(h)$ with $\epsilon_2 < D_2/3$, which uses n_2 linear measurements. If $m_2 \geq \sqrt{2}\left(\frac{3L}{D_2}\right)^{1/\alpha}$, then Algorithm*

2 recovers $\mathcal{S}_2$ exactly, i.e., $\widehat{\mathcal{S}_2} = \mathcal{S}_2$. Moreover, assuming $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ holds, and **SPARSE-REC**$_1$ *is ϵ_1-accurate (using n_1 measurements), then if m_1, n_1, ϵ_1 satisfy the conditions of Lemma 1, we have $\widehat{\mathcal{S}_1} = \mathcal{S}_1$. Lastly, the total number of queries of f made is $4(2m_2+1)^2 n_2 |\mathcal{H}_2^d| + 2(2m_1+1)n_1$.*

Proof. For any given $\mathbf{j} \in \mathcal{S}_2$ there exists $\mathbf{x}_\mathbf{j}^* \in [-1,1]^2$ with $|\phi_\mathbf{j}(\mathbf{x}_\mathbf{j}^*)| \geq D_2$. Moreover, since $\mathcal{H}_2^d$ is a $(d,2)$ hash family, there exists $h \in \mathcal{H}_2^d$ that is an injection on $\mathbf{j}$. Consequently, there exists $\mathbf{x} \in \chi(h)$ such that $\|\mathbf{x}_\mathbf{j} - \mathbf{x}_\mathbf{j}^*\|_2 \leq \frac{\sqrt{2}}{m_2}$. This in turn implies by Hölder continuity of $\phi_\mathbf{j}$ that

$$|\phi_\mathbf{j}(\mathbf{x}_\mathbf{j}) - \phi_\mathbf{j}(\mathbf{x}_\mathbf{j}^*)| \leq L\frac{2^{\alpha/2}}{m_2^\alpha}. \tag{4.9}$$

Since SPARSE-REC$_2$ is ϵ_2-accurate for each $h \in \mathcal{H}_2^d$, $\mathbf{x} \in \chi(h)$, we know that at the aforementioned $\mathbf{x}$, the following holds via reverse triangle inequality

$$|\widehat{z_\mathbf{j}^*}(\mathbf{x}_\mathbf{j})| \geq |\phi_\mathbf{j}(\mathbf{x}_\mathbf{j})| - \epsilon_2. \tag{4.10}$$

Using (4.9), (4.10) and the reverse triangle inequality, we get by the choice of ϵ_2 and m_2

$$|\widehat{z_\mathbf{j}^*}(\mathbf{x}_\mathbf{j})| \geq |\phi_\mathbf{j}(\mathbf{x}_\mathbf{j}^*)| - L\frac{2^{\alpha/2}}{m_2^\alpha} - \epsilon_2 \geq D_2 - L\frac{2^{\alpha/2}}{m_2^\alpha} - \epsilon_2 \geq \frac{2D_2}{3} - L\frac{2^{\alpha/2}}{m_2^\alpha} \geq \frac{D_2}{3}.$$

Also, for any $\mathbf{j} \notin \mathcal{S}_2$, we have for all $h \in \mathcal{H}_2^d$, $\mathbf{x} \in \chi(h)$ that $|\widehat{z_\mathbf{j}^*}(\mathbf{x}_\mathbf{j})| \leq \epsilon_2 < D_2/3$ (since $\phi_\mathbf{j} \equiv 0$). Hence, the stated choice of ϵ_2 guarantees identification of each $\mathbf{j} \in \mathcal{S}_2$, and none from $\binom{[d]}{2} \setminus \mathcal{S}_2$. The proof for recovery of $\mathcal{S}_1$ is identical to Lemma 1, and hence omitted. □

Remark 3. *On a top level, Algorithm 2 is similar to [43, Algorithms 3,4] in the sense that they all involve solving ℓ_1 minimization problems at base points lying in $\chi(h)$ defined in* (4.5) *(for identification of $\mathcal{S}_2$), and χ defined in* (4.6) *(for identification of $\mathcal{S}_1$). The difference however lies in the nature of the sampling schemes. The scheme in [43, Algorithms 3,4] relies on estimating sparse Hessians, gradients of f via their linear measurements, through random samples in the neighborhood of the base point. In contrast, the sampling scheme in Algorithm 2 is not local; for instance during the identification of $\mathcal{S}_2$, at each base point $\mathbf{x} \in \chi(h)$, the points $\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \mathbf{x}_{i,3}, \mathbf{x}_{i,4}$ for any given $i \in [n_2]$ can be arbitrarily far from each other. The same is true during the identification of $\mathcal{S}_1$.*

5 The multivariate case

Finally, we treat also the general case where f consists of at most r_0-variate components, where $r_0 > 2$ is possible. To begin with, let $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{r_0}$ be such that $\mathcal{S}_r \subset \binom{[d]}{r}$ for $r \in [r_0]$. Here $\mathcal{S}_r$ represents the r wise interaction terms. We now need some additional notation.

1. For $r \geq 1$, let $\mathcal{S}_r^{(1)}$ denote the set of variables occurring in $\mathcal{S}_r$, with $\mathcal{S}_1^{(1)} = \mathcal{S}_1$. Hence, $\mathcal{S}_r^{(1)} = \mathcal{S}_2^{\text{var}}$ in the bivariate case $r = 2$.

2. For each $1 \leq i < r \leq r_0$, denote $\mathcal{S}_r^{(i)} = \binom{\mathcal{S}_r^{(1)}}{i}$ to be the sets of i^{th} order tuples induced by $\mathcal{S}_r$.

The multivariate analogue of Proposition 2 is provided by the following result.

Proposition 3. *Let $1 \leq r_0 \leq d$ and let $f \in C([-1,1]^d)$ be of the form*

$$f(\mathbf{x}) = \mu + \sum_{j \in \bigcup_{r=2}^{r_0} \mathcal{S}_r^{(1)} \cup \mathcal{S}_1} \phi_j(x_j) + \sum_{\mathbf{j} \in \bigcup_{r=3}^{r_0} \mathcal{S}_r^{(2)} \cup \mathcal{S}_2} \phi_{\mathbf{j}}(x_{\mathbf{j}}) + \cdots + \sum_{\mathbf{j} \in \mathcal{S}_{r_0}^{(r_0-1)} \cup \mathcal{S}_{r_0-1}} \phi_{\mathbf{j}}(x_{\mathbf{j}}) + \sum_{\mathbf{j} \in \mathcal{S}_{r_0}} \phi_{\mathbf{j}}(x_{\mathbf{j}}), \tag{5.1}$$

where all the functions ϕ_j are not identically zero. Moreover, let

(a) $\mu = f(0)$.

(b) For each $1 \leq l \leq r_0 - 1$, $\phi_{\mathbf{j}}(x_{\mathbf{j}}) = 0$ if $\mathbf{j} = (j_1, \ldots, j_l) \in \bigcup_{r=l+1}^{r_0} \mathcal{S}_r^{(l)} \cup \mathcal{S}_l$, and $x_{j_i} = 0$ for some $i \in [l]$.

(c) $\phi_j(0) = 0$ if $\mathbf{j} = (j_1, \ldots, j_{r_0}) \in \mathcal{S}_{r_0}$, and $x_{j_i} = 0$ for some $i \in [r_0]$.

Then the representation (5.1) *of f is unique in the sense that each component in* (5.1) *is uniquely identifiable.*

The proof of this result is similar to the proof of Proposition 2, and we leave it to the reader.

Remark 4. *Let us note that for the special cases $r_0 \in \{1,2\}$, the statement of Proposition 3 reduces to that of Proposition 2 for univariate/bivariate SPAMs.*

We now generalize the sampling scheme given before for bivariate components to the setting of multivariate components. For this sake, we denote by $\mathtt{digit}(a,b) \in \{0,1\}$ the b^{th} digit of the dyadic decomposition of a for $a, b \in \mathbb{N}_0$. and put $\mathtt{digit}(a)$ to be the sum of digits of $a \in \mathbb{N}_0$, i.e.

$$a = \sum_{i=0}^{\infty} \mathtt{digit}(a,i) \cdot 2^i, \qquad \mathtt{digit}(a) = \sum_{i=0}^{\infty} \mathtt{digit}(a,i).$$

Let us fix some mapping $h : [d] \to [r_0]$ that partitions $[d]$ into $\mathcal{A}_1 = \{i \in [d] : h(i) = 1\}$, $\mathcal{A}_2 = \{i \in [d] : h(i) = 2\}, \ldots, \mathcal{A}_{r_0} = \{i \in [d] : h(i) = r_0\}$. Let us fix a Rademacher vector $\boldsymbol{\beta} \in \{-1,1\}^d$ and $\mathbf{x} = (x_1 \ldots x_d)^T \in [-1,1]^d$. For $z \in [2^{r_0}]$ and $i \in [d]$, we define

$$(\mathbf{x}_z)_i = x_{z,i} = \begin{cases} x_i & \text{if } \beta_i = (-1)^{\mathtt{digit}(z-1,h(i)-1)}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

Remark 5. *It $r_0 = 1$, it is easily verified, that the points $(\mathbf{x}_z)_{z=1}^2$ in* (5.2) *coincide with the points $\mathbf{x}^+, \mathbf{x}^-$ defined in* (3.2) *for univariate SPAMs. Similarly, for $r_0 = 2$, the points $(\mathbf{x}_z)_{z=1}^4$ from* (5.2) *agree with those defined in* (4.2) *for bivariate SPAMs. In the same way, the following lemma is a generalization of* (3.3) *and Lemma 2 for* (5.1)*. Indeed, if $r_0 = 1$, there is only one mapping $h : [d] \to \{1\}$, and so $\mathcal{A} = [d]$.*

Lemma 4. *Denote $\mathcal{A} = \left\{\mathbf{j} \in \binom{[d]}{r_0} : h \text{ is injective on } \{j_1, \ldots, j_{r_0}\}\right\}$. Then for functions f of the form* (5.1)*, we have that*

$$\sum_{z=1}^{2^{r_0}} (-1)^{\mathtt{digit}(z-1)} f(\mathbf{x}_z) = \sum_{\mathbf{j} \in \mathcal{A} \cap \mathcal{S}_{r_0}} \beta_{j_1} \ldots \beta_{j_{r_0}} \phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}). \tag{5.3}$$

Proof. We plug (5.1) into the left-hand side of (5.3) and obtain

$$\begin{aligned}
&\sum_{z=1}^{2^{r_0}}(-1)^{\texttt{digit}(z-1)}f(\mathbf{x}_z)\\
&=\sum_{z=1}^{2^{r_0}}(-1)^{\texttt{digit}(z-1)}\Big[\mu+\sum_{j\in\bigcup_{r=2}^{r_0}\mathcal{S}_r^{(1)}\cup\mathcal{S}_1}\phi_j(x_{z,j})+\sum_{\mathbf{j}\in\bigcup_{r=3}^{r_0}\mathcal{S}_r^{(2)}\cup\mathcal{S}_2}\phi_{\mathbf{j}}((\mathbf{x}_z)_{\mathbf{j}})+\cdots+\sum_{\mathbf{j}\in\mathcal{S}_{r_0}}\phi_{\mathbf{j}}((\mathbf{x}_z)_{\mathbf{j}})\Big]\\
&=\mu\sum_{z=1}^{2^{r_0}}(-1)^{\texttt{digit}(z-1)}+\sum_{j\in\bigcup_{r=2}^{r_0}\mathcal{S}_r^{(1)}\cup\mathcal{S}_1}\sum_{z=1}^{2^{r_0}}(-1)^{\texttt{digit}(z-1)}\phi_j(x_{z,j})+\cdots+\sum_{\mathbf{j}\in\mathcal{S}_{r_0}}\sum_{z=1}^{2^{r_0}}(-1)^{\texttt{digit}(z-1)}\phi_{\mathbf{j}}((\mathbf{x}_z)_{\mathbf{j}})\\
&=I_0+I_1+\cdots+I_{r_0}.
\end{aligned}$$

We show first that $I_0=I_1=\cdots=I_{r_0-1}=0$. Indeed,

$$I_0=\mu\sum_{z=1}^{2^{r_0}}(-1)^{\texttt{digit}(z-1)}=\mu\sum_{z=1}^{2^{r_0-1}}\Big((-1)^{\texttt{digit}(2z-2)}+(-1)^{\texttt{digit}(2z-1)}\Big)$$

and the last expression vanishes as $\texttt{digit}(2z-1)=\texttt{digit}(2z-2)+1$ for every $z\in[2^{r_0-1}]$.

If $j\in\bigcup_{r=2}^{r_0}\mathcal{S}_r^{(1)}\cup\mathcal{S}_1$, we define the set $U_j=\big\{z\in[2^{r_0}]:\beta_j=(-1)^{\texttt{digit}(z-1,h(j)-1)}\big\}$ and write

$$\begin{aligned}
I_1&=\sum_{j\in\bigcup_{r=2}^{r_0}\mathcal{S}_r^{(1)}\cup\mathcal{S}_1}\sum_{z=1}^{2^{r_0}}(-1)^{\texttt{digit}(z-1)}\phi_j(x_{z,j})=\sum_{j\in\bigcup_{r=2}^{r_0}\mathcal{S}_r^{(1)}\cup\mathcal{S}_1}\phi_j(x_j)\sum_{\substack{z\in[2^{r_0}]\\ \beta_j=(-1)^{\texttt{digit}(z-1,h(j)-1)}}}(-1)^{\texttt{digit}(z-1)}\\
&=\sum_{j\in\bigcup_{r=2}^{r_0}\mathcal{S}_r^{(1)}\cup\mathcal{S}_1}\phi_j(x_j)\sum_{z\in U_j}(-1)^{\texttt{digit}(z-1)}.
\end{aligned}$$

The definition of U_j fixes one digit of $z-1$ (namely the one at position $h(j)-1$). The sums over U_j contain 2^{r_0-1} number of summands. Looking at their digit on a position different from $h(j)-1$, we see that half of the summands is equal to 1 and the other half to -1. Therefore also $I_1=0$.

Similarly, if $\mathbf{j}=(j_1,j_2)\in\bigcup_{r=3}^{r_0}\mathcal{S}_r^{(2)}\cup\mathcal{S}_2$, we set

$$U_{\mathbf{j}}=\Big\{z\in[2^{r_0}]:\beta_{j_1}=(-1)^{\texttt{digit}(z-1,h(j_1)-1)}\text{ and }\beta_{j_2}=(-1)^{\texttt{digit}(z-1,h(j_2)-1)}\Big\}$$

and obtain

$$I_2=\sum_{\mathbf{j}\in\bigcup_{r=3}^{r_0}\mathcal{S}_r^{(2)}\cup\mathcal{S}_2}\phi_{\mathbf{j}}((\mathbf{x}_z)_{\mathbf{j}})\sum_{z\in U_{\mathbf{j}}}(-1)^{\texttt{digit}(z-1)}.$$

If now $h(j_1)=h(j_2)$ and $\beta_{j_1}\neq\beta_{j_2}$, then $U_{\mathbf{j}}$ is empty and the sum over $U_{\mathbf{j}}$ is zero. If $h(j_1)=h(j_2)$ and $\beta_{j_1}=\beta_{j_2}$, then $U_{\mathbf{j}}=\{z\in[r_0]:\beta_{j_1}=(-1)^{\texttt{digit}(z-1,h(j_1)-1)}\}=U_{j_1}$ contains 2^{r_0-1} elements and the sum over $U_{\mathbf{j}}$ is again zero by the same argument as above. Finally, if $h(j_1)\neq h(j_2)$, the definition of $U_{\mathbf{j}}$ fixes two digits of $z-1$. Therefore, $U_{\mathbf{j}}$ has 2^{r_0-2} elements. Then we consider an index $l\in\{0,1,\ldots,r_0-1\}$ different from $h(j_1)$ and $h(j_2)$, and observe that $z\in U_{\mathbf{j}}$ can have dyadic digit on l equal to zero or one. Hence the sum over $U_{\mathbf{j}}$ is again equal to zero and $I_2=0$. The same argument can be applied as long as $\{h(j_1),\ldots,h(j_r)\}$ is a proper subset of $[r_0]$ leading to $I_0=I_1=\cdots=I_{r_0-1}=0$.

Finally, if $\mathbf{j} = (j_1, \dots, j_{r_0}) \in \mathcal{S}_{r_0}$, we define

$$U_{\mathbf{j}} = \left\{ z \in [2^{r_0}] : \beta_{j_i} = (-1)^{\mathtt{digit}(z-1, h(j_i)-1)} \text{ for all } i \in [r_0] \right\}.$$

If h is an injection on $\{j_1, \dots, j_{r_0}\}$, we get that $\{h(j_1), \dots, h(j_{r_0})\} = [r_0]$ and $U_{\mathbf{j}} = \left\{z^{\mathbf{j}}\right\}$ is a singleton with

$$\sum_{z \in U_{\mathbf{j}}} (-1)^{\mathtt{digit}(z-1)} = (-1)^{\mathtt{digit}(z^{\mathbf{j}}-1)} = \prod_{i=1}^{r_0} (-1)^{\mathtt{digit}(z^{\mathbf{j}}-1, i-1)} = \prod_{i=1}^{r_0} (-1)^{\mathtt{digit}(z^{\mathbf{j}}-1, h(j_i)-1)} = \prod_{i=1}^{r_0} \beta_{j_i}.$$

If, on the other hand, h is no injection on $\{j_1, \dots, j_{r_0}\}$, $U_{\mathbf{j}}$ has even number of elements and using the same argument as above we obtain

$$\sum_{z \in U_{\mathbf{j}}} (-1)^{\mathtt{digit}(z-1)} = 0.$$

We conclude that

$$\begin{aligned} \sum_{z=1}^{2^{r_0}} (-1)^{\mathtt{digit}(z-1)} f(\mathbf{x}_z) = I_0 + I_1 + \dots + I_{r_0} &= \sum_{\mathbf{j} \in \mathcal{S}_{r_0}} \phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}) \sum_{z \in U_{\mathbf{j}}} (-1)^{\mathtt{digit}(z-1)} \\ &= \sum_{\mathbf{j} \in \mathcal{A} \cap \mathcal{S}_{r_0}} \beta_{j_1} \dots \beta_{j_{r_0}} \phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}). \end{aligned}$$ □

We denote again $z^*_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}) = \phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}})$ if $\mathbf{j} \in \mathcal{S}_{r_0}$ and 0 otherwise. Similarly, $\mathbf{z}^*(\mathbf{x}) \in \mathbb{R}^{\binom{[d]}{r_0}}$ stands for the corresponding $|\mathcal{S}_{r_0}|$-sparse vector and $\mathbf{z}^*(\mathbf{x}; \mathcal{A}) \in \mathbb{R}^{\binom{[d]}{r_0}}$ for the projection of $\mathbf{z}^*(\mathbf{x})$ onto $\mathcal{A}$. Again, $\mathbf{z}^*(\mathbf{x}; \mathcal{A})$ is $|\mathcal{S}_{r_0} \cap \mathcal{A}|$-sparse. Finally, for a Rademacher vector $\boldsymbol{\beta} \in \{-1, +1\}^d$, let $\boldsymbol{\beta}^{(r_0)} \in \{-1, +1\}^{\binom{[d]}{r_0}}$ where $\beta^{(r_0)}_{\mathbf{j}} = \beta_{j_1} \beta_{j_2} \dots \beta_{j_{r_0}}$ for each $\mathbf{j} = (j_1, j_2, \dots, j_{r_0})$. Hence, (5.3) corresponds to a linear measurement of $\mathbf{z}^*(\mathbf{x}; \mathcal{A})$ with $\boldsymbol{\beta}^{(r_0)}$.

Assumptions. We will make the following assumptions on the model (5.1).

1. *Smoothness.* Each term in (5.1) is Hölder continuous with parameters $L > 0, \alpha \in (0, 1]$, i.e., for each $i \in [r_0]$,

$$|\phi_{\mathbf{j}}(\mathbf{x}) - \phi_{\mathbf{j}}(\mathbf{y})| \leq L \|\mathbf{x} - \mathbf{y}\|_2^\alpha \quad \text{for all } \mathbf{j} \in \mathcal{S}_i \cup \bigcup_{l=i+1}^{r_0} \mathcal{S}_l^{(i)} \text{ and for all } \mathbf{x}, \mathbf{y} \in [-1, 1]^i. \tag{5.4}$$

2. *Identifiability of $\mathcal{S}_i$, $i \in [r_0]$.* We assume that for each $i \in [r_0]$ there exists a constant $D_i > 0$, such that for every $\mathbf{j} \in \mathcal{S}_i$ there exists $\mathbf{x}^*_{\mathbf{j}} \in [-1, 1]^i$ with $|\phi_{\mathbf{j}}(\mathbf{x}^*_{\mathbf{j}})| > D_i$.

3. *Disjointness.* We assume that $\mathcal{S}_p^{(1)} \cap \mathcal{S}_q^{(1)} = \emptyset$ for all $p \neq q \in [r_0]$. If $r_0 = 2$, we observed earlier in Section 4 that the assumption $\mathcal{S}_1 \cap \mathcal{S}_2^{(1)} = \emptyset$ can be made without loss of generality. However, for $r_0 > 2$, this is an additional assumption. It will allow to structure the recovery algorithm into recursive steps. For eg., if $r_0 = 3$, then the following configuration does not satisfy the disjointness assumption

$$\mathcal{S}_1 = \{1, 2, 3\}, \quad \mathcal{S}_2 = \{(4, 5), (5, 6), (6, 7)\}, \quad \mathcal{S}_3 = \{(6, 8, 9)\}.$$

In this case, $\mathcal{S}_2^{(1)} = \{4, 5, 6, 7\}$ and $\mathcal{S}_3^{(1)} = \{6, 8, 9\}$ are not disjoint.

Sampling scheme for identifying $\mathcal{S}_{r_0}$. Similarly to the sampling scheme for identifying $\mathcal{S}_2$ in the bivariate case, we generate independent Rademacher vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n \in \{-1,1\}^d$. For fixed $\mathbf{x} \in [-1,1]^d$ and $h : [d] \to [r_0]$, we obtain the samples $\tilde{f}(\mathbf{x}_{i,z}) = f(\mathbf{x}_{i,z}) + \eta_{i,z}$, where $i \in [n]$ and $z \in [2^{r_0}]$. Here, $\mathbf{x}_{i,z}$ are generated using $\mathbf{x}_i, \boldsymbol{\beta}_i, h$ as outlined in (5.2).

As a direct implication of Lemma 4, we obtain the linear system

$$\underbrace{\begin{pmatrix} \sum_{z=1}^{2^{r_0}} (-1)^{\texttt{digit}(z-1)} \tilde{f}(\mathbf{x}_{1,z}) \\ \vdots \\ \vdots \\ \sum_{z=1}^{2^{r_0}} (-1)^{\texttt{digit}(z-1)} \tilde{f}(\mathbf{x}_{n,z}) \end{pmatrix}}_{\mathbf{y} \in \mathbb{R}^n} = \underbrace{\begin{pmatrix} \boldsymbol{\beta}_1^{(r_0)^T} \\ \vdots \\ \vdots \\ \boldsymbol{\beta}_n^{(r_0)^T} \end{pmatrix}}_{\mathbf{B} \in \mathbb{R}^{n \times \binom{d}{r_0}}} \mathbf{z}^*(\mathbf{x}; \mathcal{A}) + \underbrace{\begin{pmatrix} \sum_{z=1}^{2^{r_0}} (-1)^{\texttt{digit}(z-1)} \eta_{1,z} \\ \vdots \\ \vdots \\ \sum_{z=1}^{2^{r_0}} (-1)^{\texttt{digit}(z-1)} \eta_{n,z} \end{pmatrix}}_{\boldsymbol{\eta} \in \mathbb{R}^n}. \tag{5.5}$$

By feeding $\mathbf{y}, \mathbf{B}$ as input to `SPARSE-REC`, we will obtain the estimate $\widehat{\mathbf{z}^*}(\mathbf{x}; \mathcal{A})$ to $\mathbf{z}^*(\mathbf{x}; \mathcal{A})$. Assuming `SPARSE-REC` to be ϵ-accurate at $\mathbf{x}$, we will have that $\|\widehat{\mathbf{z}^*}(\mathbf{x}; \mathcal{A}) - \mathbf{z}^*(\mathbf{x}; \mathcal{A})\|_\infty \leq \epsilon$ holds.

The choice of $\mathbf{x}, h$ is along similar lines as in the previous section. Indeed we first construct a (d, r_0) hash family $\mathcal{H}_{r_0}^d$ so that for any $\mathbf{j} = (j_1, \ldots, j_{r_0}) \in \binom{[d]}{r_0}$, there exists $h \in \mathcal{H}_{r_0}^d$ which is injective on $[\mathbf{j}]$. For each $h \in \mathcal{H}_{r_0}^d$, let us define $\mathbf{e}_1(h), \mathbf{e}_2(h), \ldots, \mathbf{e}_{r_0}(h) \in \mathbb{R}^d$, where

$$(\mathbf{e}_i(h))_q := \left\{ \begin{array}{ll} 1 \ ; & h(q) = i, \\ 0 \ ; & \text{otherwise} \end{array} \right. \quad \text{for } i \in [r_0] \text{ and } q \in [d].$$

We then create the following r_0 dimensional grid with respect to h.

$$\chi(h) := \left\{ \mathbf{x} \in [-1,1]^d : \mathbf{x} = \sum_{i=1}^{r_0} c_i \mathbf{e}_i(h); c_1, c_2, \ldots, c_{r_0} \in \left\{ -1, -\frac{m-1}{m}, \ldots, \frac{m-1}{m}, 1 \right\} \right\}. \tag{5.6}$$

Equipped with $\chi(h)$ for each $h \in \mathcal{H}_{r_0}^d$, we now possess the following approximation property. For any $\mathbf{j} \in \binom{[d]}{r_0}$ and any $(x_{j_1}^*, x_{j_2}^*, \ldots, x_{j_{r_0}}^*) \in [-1,1]^{r_0}$, there exists $h \in \mathcal{H}_{r_0}^d$ and a corresponding $\mathbf{x} \in \chi(h)$ so that $|x_{j_1}^* - x_{j_1}|, |x_{j_2}^* - x_{j_2}|, \ldots, |x_{j_{r_0}}^* - x_{j_{r_0}}| \leq 1/m$.

Here on, our idea for estimating $\mathcal{S}_{r_0}$ is based on the same principle that we followed in the preceding section. Assume that `SPARSE-REC` is ϵ accurate for each $h \in \mathcal{H}_{r_0}^d$, $\mathbf{x} \in \chi(h)$, and that m, ϵ are sufficiently large and small respectively. Hence, if we estimate $\mathbf{z}^*(\mathbf{x}; \mathcal{A})$ at each $h \in \mathcal{H}_{r_0}^d$ and $\mathbf{x} \in \chi(h)$, then for every $\mathbf{j} \in \mathcal{S}_{r_0}$ we are guaranteed to have a point $\mathbf{x}$ at which the estimate $|\widehat{z_{\mathbf{j}}^*}(\mathbf{x}_{\mathbf{j}})|$ is sufficiently large. Moreover, for every $\mathbf{j} \notin \mathcal{S}_{r_0}$, we would always (i.e., for each $h \in \mathcal{H}_{r_0}^d$, $\mathbf{x} \in \chi(h)$) have that $|\widehat{z_{\mathbf{j}}^*}(\mathbf{x}_{\mathbf{j}})|$ is sufficiently small; more precisely, $|\widehat{z_{\mathbf{j}}^*}(\mathbf{x}_{\mathbf{j}})| \leq \epsilon$ since $\phi_{\mathbf{j}} \equiv 0$. Consequently, we will be able to identify $\mathcal{S}_{r_0}$ by thresholding, via a suitable threshold.

Sampling scheme for identifying $\mathcal{S}_{r_0-1}$. Say we have an estimate for $\mathcal{S}_{r_0}$, lets call it $\widehat{\mathcal{S}_{r_0}}$, and assume $\widehat{\mathcal{S}_{r_0}}$ was identified correctly, so $\widehat{\mathcal{S}_{r_0}} = \mathcal{S}_{r_0}$. Then, we now have a SPAM of order $r_0 - 1$ on the reduced set of variables $\mathcal{P} = [d] \setminus \widehat{\mathcal{S}_{r_0}^{(1)}}$. Therefore, in order to estimate $\mathcal{S}_{r_0-1}$, we simply repeat the above procedure on the reduced set $\mathcal{P}$ by freezing the variables in $\widehat{\mathcal{S}_{r_0}^{(1)}}$ to 0. More precisely, we have the following steps.

- We will construct a $(\mathcal{P}, r_0 - 1)$ hash family $\mathcal{H}_{r_0-1}^{\mathcal{P}}$, hence each $h \in \mathcal{H}_{r_0-1}^{\mathcal{P}}$ is a mapping $h : \mathcal{P} \to [r_0 - 1]$.

- For each $h \in \mathcal{H}_{r_0-1}^{\mathcal{P}}$, define $\mathbf{e}_1(h), \mathbf{e}_2(h), \dots, \mathbf{e}_{r_0-1}(h) \in \mathbb{R}^{\mathcal{P}}$, where

  $$(\mathbf{e}_i(h))_q := \begin{cases} 1 \ ; & h(q) = i \text{ and } q \in \mathcal{P}, \\ 0 \ ; & \text{otherwise}, \end{cases} \qquad \text{for } i \in [r_0 - 1] \text{ and } q \in \mathcal{P},$$

  and use $(\mathbf{e}_i(h))_{i=1}^{r_0-1}$ to create a $r_0 - 1$ dimensional grid $\chi(h) \subset [-1,1]^{\mathcal{P}}$ in the same manner as in (5.6).

- For $h \in \mathcal{H}_{r_0-1}^{\mathcal{P}}$, a Rademacher vector $\boldsymbol{\beta} \in \{-1,1\}^{\mathcal{P}}$ and $\mathbf{x} \in [-1,1]^{\mathcal{P}}$, we define $\mathbf{x}_z \in \mathbb{R}^d$ in (5.2) as follows

  $$(\mathbf{x}_z)_i = x_{z,i} = \begin{cases} x_i \ ; & \text{if } \beta_i = (-1)^{\texttt{digit}(z-1, h(i)-1)} \text{ and } i \in \mathcal{P}, \\ 0 \ ; & \text{otherwise}, \end{cases} \qquad \text{for } i \in [d] \quad \text{and} \quad z \in [2^{r_0-1}]. \tag{5.7}$$

  Hence denoting $\mathcal{A} = \left\{ \mathbf{j} \in \binom{\mathcal{P}}{r_0-1} : h \text{ is injective on } \{j_1, \dots, j_{r_0-1}\} \right\}$, since $\widehat{\mathcal{S}_{r_0}} = \mathcal{S}_{r_0}$, we obtain as a result of Lemma 4 that

  $$\sum_{z=1}^{2^{r_0-1}} (-1)^{\texttt{digit}(z-1)} f(\mathbf{x}_z) = \sum_{\mathbf{j} \in \mathcal{A} \cap \mathcal{S}_{r_0-1}} \beta_{j_1} \dots \beta_{j_{r_0-1}} \phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}). \tag{5.8}$$

  Consequently, in the linear system in (5.5), we have $\mathbf{B} \in \mathbb{R}^{n \times \binom{|\mathcal{P}|}{r_0-1}}$ where the i^{th} row of $\mathbf{B}$ is $\boldsymbol{\beta}^{(r_0-1)} \in \{-1,+1\}^{\binom{\mathcal{P}}{r_0-1}}$ with $\beta_{\mathbf{j}}^{(r_0-1)} = \beta_{j_1}\beta_{j_2}\dots\beta_{j_{r_0-1}}$ for each $\mathbf{j} = (j_1, j_2, \dots, j_{r_0-1})$. Note that $\mathbf{z}^*(\mathbf{x}; \mathcal{A}) \in \mathbb{R}^{\binom{\mathcal{P}}{r_0-1}}$ is the $|\mathcal{S}_{r_0-1}|$ sparse vector to be estimated.

- Finally, we will estimate $\mathbf{z}^*(\mathbf{x}; \mathcal{A})$ at each $h \in \mathcal{H}_{r_0-1}^{\mathcal{P}}$ and $\mathbf{x} \in \chi(h)$. If SPARSE-REC is ϵ accurate, with ϵ sufficiently small, then by choosing the number of points m to be sufficiently large, we will be able to identify $\mathcal{S}_{r_0-1}$ via thresholding.

By repeating the above steps for all $i = r_0, r_0 - 1, \dots, 1$, we arrive at a procedure for estimating the supports $\mathcal{S}_i, i \in [r_0]$; this is outlined formally in the form of the Algorithm 3 below. Lemma 5 below provides sufficient conditions on the sampling parameters in Algorithm 3 for exact recovery of all $\mathcal{S}_i$'s.

Lemma 5. *For each $i \in [r_0]$ assume that the following hold:*

1. $m_i \geq \sqrt{i}\left(\frac{3L}{D_i}\right)^{1/\alpha}$.

2. *SPARSE-REC$_i$ is ϵ_i accurate with $\epsilon_i < D_i/3$ for all $h \in \mathcal{H}_i^{\mathcal{P}_i}$, $\mathbf{x} \in \chi(h)$, where $\mathcal{P}_i$ denotes the set $\mathcal{P}$ at the beginning of iteration i (so $\mathcal{P}_{r_0} = [d]$). The number of measurements used by SPARSE-REC$_i$ is denoted by n_i.*

3. *$\mathcal{H}_i^{\mathcal{P}_i}$ is a $(\mathcal{P}_i, i)$ hash family.*

Then $\widehat{\mathcal{S}}_i = \mathcal{S}_i$ for all $i = r_0, r_0 - 1, \dots, 1$ in Algorithm 3. Moreover, the total number of queries of f made is

$$\sum_{i=1}^{r_0} 2^i (2m_i + 1)^i n_i |\mathcal{H}_i^{\mathcal{P}_i}|.$$

Algorithm 3 Algorithm for estimating $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{r_0}$

```
Input: $d$, $|\mathcal{S}_i|$, $(m_i, n_i, \epsilon_i)$ for $i = 1, \dots, r_0$.
Initialization: $\widehat{\mathcal{S}}_i = \emptyset$ for $i = 1, \dots, r_0$. $\mathcal{P} = [d]$.
Output: $\widehat{\mathcal{S}}_i$ for $i = 1, \dots, r_0$.

for $i = r_0, r_0 - 1, \dots, 1$ do        // Estimation of $\mathcal{S}_i$
    Generate Rademacher random vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_{n_i} \in \{-1, 1\}^{\mathcal{P}}$.
    Form $\mathbf{B} \in \mathbb{R}^{n_i \times \binom{|\mathcal{P}|}{i}}$ as in (5.5).
    Construct a $(\mathcal{P}, i)$ hash family $\mathcal{H}_i^{\mathcal{P}}$.
    for $h \in \mathcal{H}_i^{\mathcal{P}}$ do
        Construct $\chi(h) \subset [-1, 1]^{\mathcal{P}}$ in the same manner as in (5.6) with $|\chi(h)| = (2m_i + 1)^i$.
        for $\mathbf{x} \in \chi(h)$ do
            Generate $\mathbf{x}_z \in [-1, 1]^d$, with $z \in [2^i]$ as in (5.7), using $\mathbf{x}, \boldsymbol{\beta}_u$ for each $u \in [n_i]$.
            Using the samples $(\tilde{f}(\mathbf{x}_z))_{z=1}^{2^i}$, form $\mathbf{y} \in \mathbb{R}^{n_i}$ as in (5.5).
            Obtain $\widehat{\mathbf{z}^*}(\mathbf{x}; \mathcal{A}) = \texttt{SPARSE-REC}_i(\mathbf{y}, \mathbf{B})$.
            Update $\widehat{\mathcal{S}}_i = \widehat{\mathcal{S}}_i \cup \left\{ \mathbf{j} \in \mathcal{A} : |\widehat{z^*_{\mathbf{j}}}(\mathbf{x_j})| > \epsilon_i \right\}$.
        end for
    end for
    Update $\mathcal{P} = \mathcal{P} \setminus \widehat{\mathcal{S}}_i^{(1)}$.
end for
```

Proof. The proof outline builds on what we have seen in the preceding sections. Say we are at the beginning of iteration $i \in [r_0]$ with $\widehat{\mathcal{S}}_l = \mathcal{S}_l$ holding true for each $l > i$. Hence, the model has reduced to an order i sparse additive model on the set $\mathcal{P}_i \subset [d]$, with $\mathcal{S}_i^{(1)}, \mathcal{S}_{i-1}^{(1)}, \dots, \mathcal{S}_1 \subset \mathcal{P}_i$.

By identifiability assumption, we know that for any given $\mathbf{j} \in \mathcal{S}_i$, there exists $\mathbf{x}^*_{\mathbf{j}} \in [-1, 1]^i$ such that $|\phi_{\mathbf{j}}(\mathbf{x}^*_{\mathbf{j}})| \geq D_i$ holds. Moreover, since $\mathcal{H}_i^{\mathcal{P}_i}$ is a $(\mathcal{P}_i, i)$ hash family, there exists a $h \in \mathcal{H}_i^{\mathcal{P}_i}$ that is an injection on $\mathbf{j}$. Consequently, there exists $\mathbf{x} \in \chi(h)$ such that $\|\mathbf{x_j} - \mathbf{x}^*_{\mathbf{j}}\|_2 \leq (\sum_{p=1}^{i} \frac{1}{m_i^2})^{1/2} = \sqrt{i}/m_i$. By Hölder continuity of $\phi_{\mathbf{j}}$, this means

$$|\phi_{\mathbf{j}}(\mathbf{x_j}) - \phi_{\mathbf{j}}(\mathbf{x}^*_{\mathbf{j}})| \leq L \frac{i^{\alpha/2}}{m_i^{\alpha}}. \tag{5.9}$$

Since $\texttt{SPARSE-REC}_i$ is ϵ_i accurate for each $h \in \mathcal{H}_i^{\mathcal{P}_i}$, $\mathbf{x} \in \chi(h)$, we know that at the aforementioned $\mathbf{x}$, the following holds via reverse triangle inequality

$$|\widehat{z^*_{\mathbf{j}}}(\mathbf{x_j})| \geq |\phi_{\mathbf{j}}(\mathbf{x_j})| - \epsilon_i. \tag{5.10}$$

Using (5.9), (5.10), reverse triangle inequality and the choice of ϵ_i and m_i, we obtain

$$|\widehat{z^*_{\mathbf{j}}}(\mathbf{x_j})| \geq |\phi_{\mathbf{j}}(\mathbf{x}^*_{\mathbf{j}})| - L \frac{i^{\alpha/2}}{m_i^{\alpha}} - \epsilon_i \geq D_i - L \frac{i^{\alpha/2}}{m_i^{\alpha}} - \epsilon_i \geq \frac{2D_i}{3} - L \frac{i^{\alpha/2}}{m_i^{\alpha}} \geq \frac{D_i}{3}.$$

For any $\mathbf{j} \notin \mathcal{S}_i$, we have for all $h \in \mathcal{H}_i^{\mathcal{P}_i}$, $\mathbf{x} \in \chi(h)$ that $|\widehat{z^*_{\mathbf{j}}}(\mathbf{x_j})| \leq \epsilon_i < D_i/3$ (since $\phi_{\mathbf{j}} \equiv 0$). Hence clearly, the stated choice of ϵ_i guarantees identification of each $\mathbf{j} \in \mathcal{S}_i$, and none from $\binom{\mathcal{P}_i}{i} \setminus \mathcal{S}_i$. This means that we will recover $\mathcal{S}_i$ exactly. As this is true for each $i \in [r_0]$, it also completes the proof for exact recovery of $\mathcal{S}_i$ for each $i \in [r_0]$.

The expression for the total number of queries made follows from a simple calculation where we note that at iteration i, and corresponding to each $\mathbf{x} \in \cup_{h \in \mathcal{H}_i^{\mathcal{P}_i}} \chi(h)$, we make $2^i n_i$ queries of f. □

6 Estimating sparse multilinear functions from few samples

In this section, we provide results from the sparse recovery literature for estimating sparse multilinear forms from random samples. In particular, these results cover arbitrary bounded noise and i.i.d. Gaussian noise models.

6.1 Sparse linear functions

Consider a linear function $g : \mathbb{R}^d \to \mathbb{R}$, where $g(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \mathbf{a}$. Our interest is in recovering the unknown coefficient vector $\mathbf{a}$ from n noisy samples $y_i = g(\boldsymbol{\beta}_i) + \eta_i$, $i \in [n]$, where η_i refers to the noise in the i^{th} sample. Arranging the samples together, we arrive at the linear system $\mathbf{y} = \mathbf{B}\mathbf{a} + \boldsymbol{\eta}$, where

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}\in\mathbb{R}^n} = \underbrace{\begin{pmatrix} {\boldsymbol{\beta}_1}^T \\ \vdots \\ \vdots \\ {\boldsymbol{\beta}_n}^T \end{pmatrix}}_{\mathbf{B}\in\mathbb{R}^{n\times d}} \mathbf{a} + \underbrace{\begin{pmatrix} \eta_1 \\ \vdots \\ \vdots \\ \eta_n \end{pmatrix}}_{\boldsymbol{\eta}}. \tag{6.1}$$

Denoting by $\mathcal{S} := \{j \in [d] : a_j \neq 0\}$ the support of $\mathbf{a}$, our interest is in the setting where $\mathbf{a}$ is sparse, i.e., $|\mathcal{S}| = k \ll d$, and consequently to estimate $\mathbf{a}$ from a small number of samples n. To begin with, we will require $\mathbf{B}$ in (6.1) to satisfy the so called ℓ_2/ℓ_2 RIP, defined below.

Definition 2. *A matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ is said to satisfy the ℓ_2/ℓ_2 Restricted Isometry Property (RIP) of order k with constant $\delta_k \in (0,1)$ if*

$$(1-\delta_k)\|\mathbf{x}\|_2^2 \ \leq\ \frac{1}{n}\|\mathbf{A}\mathbf{x}\|_2^2 \ \leq\ (1+\delta_k)\|\mathbf{x}\|_2^2$$

holds for all k-sparse $\mathbf{x}$.

Bounded noise model. Let us consider the scenario where the noise is bounded in the ℓ_2 norm, i.e., $\|\boldsymbol{\eta}\|_2 \leq \nu$. We will recover an estimate $\widehat{\mathbf{a}}$ to $\mathbf{a}$ as a solution of the following quadratically constrained ℓ_1 minimization program [7]

$$\text{(P1)} \quad \min_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{z}\|_1 \quad \text{s.t} \quad \|\mathbf{y} - \mathbf{B}\mathbf{z}\|_2 \leq \nu. \tag{6.2}$$

The following result provides a bound on the estimation error $\|\widehat{\mathbf{a}} - \mathbf{a}\|_2$ for (P1).

Theorem 1. *Consider the sampling model in (6.1), where $\mathbf{B} \in \{-1,+1\}^{n\times d}$ is a Rademacher matrix. Then the following hold.*

1. *([2]) For any constant $\delta \in (0,1)$, there exist constants $c_1, c_2 > 0$ depending on δ such that if*

$$n \geq c_1 k \log(d/k),$$

then with probability at least $1 - 2\exp(-c_2 n)$, the matrix $\mathbf{B}$ satisfies ℓ_2/ℓ_2 RIP of order k, with $\delta_k \leq \delta$.

2. *([7, Theorem 1.2]) Let $\mathbf{B}$ satisfy the ℓ_2/ℓ_2 RIP with $\delta_{2k} < \sqrt{2} - 1$. Then there exist constants $C_1, C_2 > 0$ such that, simultaneously for all vectors $\mathbf{a} \in \mathbb{R}^d$, any solution $\widehat{\mathbf{a}}$ to (P1) satisfies*

$$\|\widehat{\mathbf{a}} - \mathbf{a}\|_2 \leq C_1 \frac{\|\mathbf{a} - \mathbf{a}_k\|_1}{\sqrt{k}} + C_2 \frac{\nu}{\sqrt{n}}.$$

Here, $\mathbf{a}_k$ denotes the best k-term approximation of $\mathbf{a}$.

Gaussian noise model. We now consider the scenario where the noise samples are i.i.d. Gaussian with variance σ^2, i.e, $\eta_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. for all $i \in [n]$. Using standard concentration inequalities for sub-exponential random variables (see Proposition 6(1) in Appendix B), one can show that $\|\boldsymbol{\eta}\|_2 = \Theta(\sigma\sqrt{n})$ with high probability. This leads to the following straightforward corollary of Theorem 1.

Corollary 1. *Consider the sampling model in (6.1) for some given vector $\mathbf{a} \in \mathbb{R}^n$, and let $\eta_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. for all $i \in [n]$. Say $\mathbf{B}$ satisfies ℓ_2/ℓ_2 RIP with $\delta_{2k} < \sqrt{2} - 1$. For some $\varepsilon \in (0, 1)$, let $\widehat{\mathbf{a}}$ be a solution to (P1) with $\nu = (1 + \varepsilon)\sigma\sqrt{n}$. Then there exists a constant $c_3 > 0$ so that any solution $\widehat{\mathbf{a}}$ to (P1) satisfies*

$$\|\widehat{\mathbf{a}} - \mathbf{a}\|_2 \leq C_1 \frac{\|\mathbf{a} - \mathbf{a}_k\|_1}{\sqrt{k}} + C_2(1 + \varepsilon)\sigma$$

with probability at least $1 - 2\exp(-c_3\varepsilon^2 n)$. Here C_1, C_2 are the constants from Theorem 1.

Proof. Use Proposition 6(1) from Appendix B with Theorem 1. □

6.2 Sparse bilinear functions

Let $\mathbf{a} \in \mathbb{R}^{\binom{d}{2}}$ be a vector of length $\binom{d}{2}$ with entries indexed by $\binom{[d]}{2}$ (sorted in lexicographic order). The entry of $\mathbf{a}$ at index $\mathbf{j} = (j_1, j_2) \in \binom{[d]}{2}$ will be denoted by $a_{\mathbf{j}}$. We now consider the setting where $g : \mathbb{R}^d \to \mathbb{R}$ is a second order multilinear function, i.e., $g(\boldsymbol{\beta}) = \langle \boldsymbol{\beta}^{(2)}, \mathbf{a} \rangle$ with $\boldsymbol{\beta}^{(2)}, \mathbf{a} \in \mathbb{R}^{\binom{d}{2}}$, and $\beta^{(2)}_{(j_1,j_2)} = \beta_{j_1}\beta_{j_2}$. As before, our goal is to recover $\mathbf{a}$ from n noisy samples $y_i = g(\boldsymbol{\beta}_i) + \eta_i$, $i \in [n]$ resulting in the linear system

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y} \in \mathbb{R}^n} = \underbrace{\begin{pmatrix} {\boldsymbol{\beta}_1^{(2)}}^T \\ \vdots \\ \vdots \\ {\boldsymbol{\beta}_n^{(2)}}^T \end{pmatrix}}_{\mathbf{B} \in \mathbb{R}^{n \times \binom{d}{2}}} \mathbf{a} + \underbrace{\begin{pmatrix} \eta_1 \\ \vdots \\ \vdots \\ \eta_n \end{pmatrix}}_{\boldsymbol{\eta}}. \tag{6.3}$$

Observe that

$$\langle \boldsymbol{\beta}^{(2)}, \mathbf{a} \rangle = \boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta} = \langle \boldsymbol{\beta}\boldsymbol{\beta}^T, \mathbf{A} \rangle, \tag{6.4}$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a symmetric matrix with zero on the diagonal and $A_{i,j} = a_{(i,j)}/2$ if $i \neq j$. This simple observation allows us to rewrite (6.3) as

$$y_i = \langle \boldsymbol{\beta}_i \boldsymbol{\beta}_i^T, \mathbf{A} \rangle + \eta_i, \quad i \in [n]. \tag{6.5}$$

Chen et al. [9] recently showed that a sparse symmetric matrix – not necessarily being zero on the diagonal – can be recovered from its measurements of the form (6.5), provided that β_j is sampled in an i.i.d. manner from a distribution satisfying

$$\mathbb{E}[\beta_i] = 0, \quad \mathbb{E}[\beta_i^2] = 1 \quad \text{and} \quad \mathbb{E}[\beta_i^4] > 1. \tag{6.6}$$

Their recovery program[3] is essentially constrained ℓ_1 minimization ([9, Eq.(4)]), the guarantees for which rely on the ℓ_2/ℓ_1 RIP for sparse symmetric matrices ([9, Def. 2]).

Since our matrix $\mathbf{A}$ in (6.5) is sparse and symmetric, it is natural for us to use their scheme. We do this, albeit with some technical changes:

[3] Note that one can remove the positive semi-definite constraint in their program and replace it with a symmetry enforcing constraint, the result remains unchanged (cf. remark in [9, Section E])

- We will see that if $\mathbf{A}$ is known to be zero on the diagonal (which is the case here), the fourth order moment condition in (6.6) is not needed. Hence one could, for instance, sample β from the symmetric Rademacher distribution.

- Instead of optimizing over the set of symmetric matrices with zeros on the diagonal, we will perform ℓ_1 minimization over the upper triangular entries of $\mathbf{A}$, represented by $\mathbf{a}$. These approaches are equivalent, but the latter has the computational advantage of having fewer constraints.

The analysis is based on the notion of ℓ_2/ℓ_1 RIP of the matrix $\mathbf{B}$ in (6.3), which is defined as follows.

Definition 3. *A matrix $\mathbf{B} \in \mathbb{R}^{n \times N}$ is said to satisfy the ℓ_2/ℓ_1 Restricted Isometry Property (RIP) of order k with constants $\gamma_k^{\text{lb}} \in (0,1)$ and $\gamma_k^{\text{ub}} > 0$ if*

$$(1-\gamma_k^{\text{lb}})\|\mathbf{x}\|_2 \ \leq \ \frac{1}{n}\|\mathbf{B}\mathbf{x}\|_1 \ \leq \ (1+\gamma_k^{\text{ub}})\|\mathbf{x}\|_2$$

holds for all k-sparse $\mathbf{x} \in \mathbb{R}^N$.

The above definition is analogous to the one in [9, Def. 2] for sparse symmetric matrices.

Bounded noise model. Let us first consider the setting where the noise is bounded in the ℓ_1 norm, i.e., $\|\boldsymbol{\eta}\|_1 \leq \nu$. We recover the estimate $\widehat{\mathbf{a}}$ as a solution to the following program

$$\text{(P2)} \qquad \min_{\mathbf{z} \in \mathbb{R}^{\binom{d}{2}}} \|\mathbf{z}\|_1 \quad \text{s.t} \quad \|\mathbf{y} - \mathbf{B}\mathbf{z}\|_1 \leq \nu. \tag{6.7}$$

The next result shows that $\mathbf{B}$ satisfies ℓ_2/ℓ_1 RIP with high probability if the rows of $\mathbf{B}$ are formed by independent Rademacher vectors. Consequently, the above program stably recovers $\mathbf{a}$.

Theorem 2. *Consider the sampling model in (6.3), where the rows of $\mathbf{B}$ are formed by independent Rademacher vectors. Then the following hold.*

1. *There exist absolute constants $c_1, c_2, c_3 > 0$, such that the following is true. Let $\mathbf{a} \in \mathbb{R}^{\binom{d}{2}}$. Then*
$$c_1\|\mathbf{a}\|_2 \leq \frac{1}{n}\|\mathbf{B}\mathbf{a}\|_1 \leq c_2\|\mathbf{a}\|_2 \tag{6.8}$$
holds with probability at least $1 - \exp(-c_3 n)$.

2. *With constants c_1, c_2 from (6.8), there exist constants $c_1', c_2', c_3' > 0$ such that if $n > c_3' k \log(d^2/k)$, then $\mathbf{B}$ satisfies ℓ_2/ℓ_1 RIP of order k with probability at least $1 - c_1' \exp(-c_2' n)$ with constants γ_k^{lb} and γ_k^{ub}, which fulfill*
$$1 - \gamma_k^{\text{lb}} \geq \frac{c_1}{2} \quad \text{and} \quad 1 + \gamma_k^{\text{ub}} \leq 2c_2. \tag{6.9}$$

3. *If there exists a number $K > 2k$ such that $\mathbf{B}$ satisfies*
$$\frac{1-\gamma_{k+K}^{\text{lb}}}{\sqrt{2}} - (1+\gamma_K^{\text{ub}})\sqrt{\frac{k}{K}} \geq \beta > 0$$

for some $\beta > 0$*, then the solution* $\widehat{\mathbf{a}}$ *to (P2) satisfies*

$$\|\widehat{\mathbf{a}} - \mathbf{a}\|_2 \leq \Big(\frac{\tilde{C}_1}{\beta} + \tilde{C}_3\Big)\frac{\|\mathbf{a} - \mathbf{a}_k\|_1}{\sqrt{K}} + \frac{\tilde{C}_2}{\beta}\frac{\nu}{n},$$

where $\mathbf{a}_k$ *denotes the best* k*-term approximation of* $\mathbf{a}$ *and* $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3$ *are universal positive constants.*

4. *There exist absolute constants* $\tilde{c}_3, \tilde{c}_4, C_3, C_4 > 0$ *such that if* $n \geq \tilde{c}_3 k \log(d^2/k)$*, then the solution* $\widehat{\mathbf{a}}$ *to (P2) satisfies*

$$\|\widehat{\mathbf{a}} - \mathbf{a}\|_2 \leq C_3\frac{\|\mathbf{a} - \mathbf{a}_k\|_1}{\sqrt{k}} + C_4\frac{\nu}{n},$$

simultaneously for all $\mathbf{a} \in \mathbb{R}^{\binom{d}{2}}$ *with probability at least* $1 - \exp(-\tilde{c}_4 n)$*.*

We sketch the proof of Theorem 2, which is essentially based on [9], in Appendix C.

Gaussian noise model. We now consider the scenario where the noise samples are i.i.d. Gaussian with variance σ^2, i.e, $\eta_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. for all $i \in [n]$. Using standard concentration inequalities for sub-Gaussian random variables (see Proposition 6(2) in Appendix B), one can show that $\|\boldsymbol{\eta}\|_1 = \Theta(\sigma n)$ with high probability. This leads to the following straightforward corollary of Theorem 2.

Corollary 2. *For constants* $\tilde{c}_3, \tilde{c}_4, C_3, C_4 > 0$ *defined in Theorem 2, the following is true. Consider the sampling model in* (6.3) *for a given* $\mathbf{a} \in \mathbb{R}^{\binom{d}{2}}$*, where the rows of* $\mathbf{B}$ *are formed by independent Rademacher vectors, and* $\eta_i \sim \mathcal{N}(0, \sigma^2)$ *i.i.d. for all* $i \in [n]$ *with* $n \geq \tilde{c}_3 k \log(d^2/k)$*. For some* $\varepsilon \in (0, 1)$*, let* $\widehat{\mathbf{a}}$ *be a solution to (P2) with* $\nu = (1 + \varepsilon)\sigma n$*. Then*

$$\|\widehat{\mathbf{a}} - \mathbf{a}\|_2 \leq C_3\frac{\|\mathbf{a} - \mathbf{a}_k\|_1}{\sqrt{k}} + C_4\sqrt{\frac{2}{\pi}}(1 + \varepsilon)\sigma$$

with probability at least $1 - \exp(-\tilde{c}_4 n) - e \cdot \exp\left(-\tilde{c}_5 \varepsilon^2 n\right)$*, for some constant* $\tilde{c}_5 > 0$*. Here,* $\mathbf{a}_k$ *denotes the best* k*-term approximation of* $\mathbf{a}$.

Proof. Use Proposition 6(2) from Appendix B with Theorem 2. □

Remark 6. *Both (P1) and (P2) are convex programs where (P1) can be cast as a second order cone program (SOCP) (eg., [8]) and (P2) can be easily written as a linear program (LP). Hence they can be solved to arbitrary accuracy in time polynomial in n and d, using for instance interior point algorithms (eg., [31, 1]). In practice, one often considers non-convex alternatives such as Iterative Hard Thresholding (IHT) (eg. [5, 6]) which typically have low computational cost.*

6.3 Sparse multilinear functions

For $p, d \in \mathbb{N}$ with $p \leq d$, let $\mathbf{a} \in \mathbb{R}^{\binom{d}{p}}$ be a vector with entries indexed by $\binom{[d]}{p}$ (sorted in lexicographic order). We denote again the entry of $\mathbf{a}$ at index $\mathbf{j} = (j_1, \dots, j_p) \in \binom{[d]}{p}$ by $a_{\mathbf{j}}$ and consider a multilinear function $g : \mathbb{R}^d \to \mathbb{R}$ in d variables $\boldsymbol{\beta} = (\beta_1 \dots \beta_d)^T$ such that

$$g(\boldsymbol{\beta}) = \sum_{\mathbf{j}=(j_1,\dots,j_p)\in\binom{[d]}{p}} \beta_{j_1}\beta_{j_2}\cdots\beta_{j_p} a_{\mathbf{j}}. \tag{6.10}$$

We will refer to (6.10) as a multilinear function of order p. For clarity of notation, we will write (6.10) as $g(\boldsymbol{\beta}) = \langle \boldsymbol{\beta}^{(p)}, \mathbf{a} \rangle$, where $\boldsymbol{\beta}^{(p)} \in \mathbb{R}^{\binom{d}{p}}$, and the entry of $\boldsymbol{\beta}^{(p)}$ at index $\mathbf{j} = (j_1, \dots, j_p) \in \binom{[d]}{p}$ being $\beta_{j_1}\beta_{j_2}\cdots\beta_{j_p}$.

We are again interested in recovering the unknown coefficient vector $\mathbf{a}$ from n noisy samples $y_i = g(\boldsymbol{\beta}_i) + \eta_i, i \in [n]$, where η_i refers to the noise in the i^{th} sample. Arranging the samples together, we arrive at the linear system $\mathbf{y} = \mathbf{B}\mathbf{a} + \boldsymbol{\eta}$, where

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}\in\mathbb{R}^n} = \underbrace{\begin{pmatrix} {\boldsymbol{\beta}_1^{(p)}}^T \\ \vdots \\ \vdots \\ {\boldsymbol{\beta}_n^{(p)}}^T \end{pmatrix}}_{\mathbf{B}\in\mathbb{R}^{n\times\binom{d}{p}}} \mathbf{a} + \underbrace{\begin{pmatrix} \eta_1 \\ \vdots \\ \vdots \\ \eta_n \end{pmatrix}}_{\boldsymbol{\eta}}. \tag{6.11}$$

We denote by $\mathcal{S} = \left\{(j_1, j_2, \dots, j_p) \in \binom{[d]}{p} : a_{(j_1,j_2,\dots,j_p)} \neq 0\right\}$ the support of $\mathbf{a}$ and we are especially interested in the setting where $\mathbf{a}$ is sparse, i.e., $|\mathcal{S}| = k \ll \binom{d}{p}$. Our aim is to estimate $\mathbf{a}$ with a small number of samples n.

Bounded noise model. Let us consider the scenario where the noise is bounded in the ℓ_2 norm, i.e., $\|\boldsymbol{\eta}\|_2 \leq \nu$. We will recover an estimate $\widehat{\mathbf{a}}$ to $\mathbf{a}$ as a solution of (P1). The following result provides a bound on the estimation error $\|\widehat{\mathbf{a}} - \mathbf{a}\|_2$.

Theorem 3. *([30, Theorem 4; Lemmas 4, 5]) Let $D = \binom{d}{p}$ and let $\mathbf{B}$ be defined as in* (6.11) *with rows formed by independent Rademacher vectors. Then, for any $\delta \in (0,1)$, there exist constants $c_6, c_7 > 0$ depending on δ such that the matrix $\mathbf{B}$ satisfies ℓ_2/ℓ_2 RIP of order k, with $\delta_k \leq \delta$,*

(a) with probability at least $1 - \exp(-c_7 n/k^2)$ *if* $n \geq c_6 k^2 \log D$

(b) and with probability at least $1 - \exp(-c_7 \min\{n/3^{2p}, n/k\})$ *if*

$$n \geq c_6 \max\left\{3^{2p} k \log(D/k), k^2 \log(D/k)\right\}.$$

The bounds on n in the theorem are obtained via the application of two very different methodologies. The bound in part (a) is a consequence of bounding the eigenvalues of the $k \times k$ Gram matrices $\frac{1}{n}\mathbf{B}_{\mathcal{S}}^T\mathbf{B}_{\mathcal{S}}$ for all $\mathcal{S}$ (here $\mathbf{B}_{\mathcal{S}}$ is the submatrix of $\mathbf{B}$ with column indices in $\mathcal{S}$) using Gershgorin's disk theorem, along with standard concentration inequalities [30, Theorem 4]. The bound in part (b) involves the usage of tail estimates for Rademacher chaos variables and follows from [30, Lemmas 4, 5]. We also note that in Theorem 3, the number of measurements n scales quadratically with the sparsity parameter k. However when $p = 2$, we can see that Theorem 2 is stronger since n therein scales linearly with k.

Remark 7. *The analysis in [30, Section A] derives RIP bounds in terms of the so-called combinatorial dimension. However in our opinion, this analysis has several inaccuracies because of which we are not sure if the corresponding bounds are correct. Hence we do not state those bounds here.*

Gaussian noise model. In the scenario where the noise in the samples is i.i.d. Gaussian, we arrive at a statement similar to Corollary 1, hence we do not discuss this further.

7 Putting it together: final theoretical guarantees

We are now in a position to combine our efforts from the preceding sections and to state the final results for recovery of the support sets. As before, we state this separately for the univariate, bivariate, and general multivariate cases.

7.1 Univariate case

We begin with the univariate case considered in Section 3. Recall Algorithm 1 for recovering the support $\mathcal{S}_1$. The ensuing Lemma 1 gave sufficient conditions for exact recovery provided `SPARSE-REC` is ϵ-accurate at each $\mathbf{x} \in \chi$, for small enough ϵ. Instantiating `SPARSE-REC` with (P1) in (6.2) gives the final results below. Let us start with the bounded noise model.

Bounded noise model. In this noise model, querying f at $\mathbf{x}$ returns $f(\mathbf{x}) + \eta$, where $|\eta| \leq \triangle$. For the linear system (3.4), this means that $|\eta_i^+|, |\eta_i^-| \leq \triangle$, $i \in [n]$, and hence $\|\boldsymbol{\eta}\|_\infty \leq 2\triangle$. The following theorem shows that if $\triangle$ is sufficiently small, then Algorithm 1 recovers $\mathcal{S}_1$ exactly provided the parameters m, n, ϵ are chosen in a suitable way.

Theorem 4. *For the bounded noise model with the noise uniformly bounded by $\triangle$, consider Algorithm 1 with `SPARSE-REC` instantiated with $(P1)$ in (6.2). If $\nu = 2\triangle\sqrt{n}$ in $(P1)$,*

$$\triangle < \frac{D_1}{6C_2}, \quad m \geq (3L/D_1)^{1/\alpha}, \quad \textit{and} \quad n \geq \tilde{c}_1 |\mathcal{S}_1| \log(d/|\mathcal{S}_1|)$$

are satisfied, it follows for the choice $\epsilon = 2C_2\triangle$ that $\widehat{\mathcal{S}_1} = \mathcal{S}_1$, with probability at least $1 - 2\exp(-\tilde{c}_2 n)$. The total number of queries made is $2(2m+1)n = \Omega\left(\bar{c}_1 |\mathcal{S}_1| \log\left(\frac{d}{|\mathcal{S}_1|}\right)\right)$ where $\bar{c}_1 > 1$ depends on L, D_1, α.

Proof. The proof follows by combining Lemma 1 with Theorem 1. Since $|\eta_i^+|, |\eta_i^-| \leq \triangle$ in (3.4) for $i \in [n]$, we obtain $\|\boldsymbol{\eta}\|_2 \leq \sqrt{n}\|\boldsymbol{\eta}\|_\infty \leq 2\triangle\sqrt{n}$. Therefore we set `SPARSE-REC` $= (P1)$ with $\nu = 2\triangle\sqrt{n}$.

As a consequence of Theorem 1, there exist constants $\tilde{c}_1, \tilde{c}_2 > 0$ (depending only on $c_1, c_2 > 0$ defined therein) so that, for $n \geq \tilde{c}_1 |\mathcal{S}_1| \log(d/|\mathcal{S}_1|)$, $\mathbf{B}$ satisfies ℓ_2/ℓ_2 RIP with $\delta_{2|\mathcal{S}_1|} < \sqrt{2} - 1$ with probability at least $1 - 2\exp(-\tilde{c}_2 n)$. Conditioning on this event, it follows from Theorem 1 that

$$\|\widehat{\mathbf{z}^*}(\mathbf{x}) - \mathbf{z}^*(\mathbf{x})\|_\infty \ \leq \ \|\widehat{\mathbf{z}^*}(\mathbf{x}) - \mathbf{z}^*(\mathbf{x})\|_2 \ \leq \ 2C_2\triangle \quad \text{for all} \quad \mathbf{x} \in \chi.$$

Hence, we see that with probability at least $1 - 2\exp(-\tilde{c}_2 n)$, `SPARSE-REC` is $\epsilon = 2C_2\triangle$ accurate at each $\mathbf{x} \in \chi$. Now invoking Lemma 1, it follows that if

$$2C_2\triangle < \frac{D_1}{3} \Leftrightarrow \triangle < \frac{D_1}{6C_2}$$

holds, then the stated choice of ϵ and m ensures exact recovery. This completes the proof. □

Gaussian noise model. We now move to the Gaussian noise model, wherein querying f at $\mathbf{x}$ returns $f(\mathbf{x}) + \eta$; $\eta \sim \mathcal{N}(0, \sigma^2)$. Moreover, the noise samples are independent across the queries. For the linear system (3.4), this means that $\eta_i \sim \mathcal{N}(0, 2\sigma^2), i \in [n]$ are i.i.d. random variables. The following theorem essentially shows that if the noise variance σ^2 is sufficiently small, then Algorithm 1 recovers $\mathcal{S}_1$ exactly provided the parameters m, n, ϵ are chosen properly. The reduction in the variance is handled via re-sampling each query N times and averaging the values. Essentially, this leads to the same sampling model with i.i.d. $\eta_i \sim \mathcal{N}(0, 2\sigma^2/N), i \in [n]$.

Theorem 5. *For the Gaussian noise model with i.i.d. noise samples with variance σ^2 and $N \in \mathbb{N}$, we resample each query N times, and average the values. Consider Algorithm 1 with* **`SPARSE-REC`** *instantiated with $(P1)$ in (6.2), wherein $\nu = \sqrt{2}(1+\varepsilon)\sigma\sqrt{n/N}$ for some $\varepsilon \in (0,1)$, and with*

$$N \geq \left\lfloor \frac{18C_2^2(1+\varepsilon)^2\sigma^2}{D_1^2} \right\rfloor + 1, \ m \geq \left(\frac{3L}{D_1}\right)^{1/\alpha}, \quad n \geq \max\left\{ \tilde{c}_1|\mathcal{S}_1|\log\left(\frac{d}{|\mathcal{S}_1|}\right), \frac{2\log(2m+1)}{c_3\varepsilon^2} \right\} \tag{7.1}$$

being satisfied. Then $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ with probability at least $1-2\exp(-\tilde{c}_2 n)-2\exp(-\frac{c_3\varepsilon^2 n}{2})$. The constants $\tilde{c}_1, \tilde{c}_2 > 0$ are as in Theorem 4; $C_2, c_3 > 0$ come from Theorem 1 and Corollary 1, respectively. The total number of queries made is $2(2m+1)nN = \Omega\left(\tilde{c}_1'|\mathcal{S}_1|\log\left(\frac{d}{|\mathcal{S}_1|}\right)\right)$ where $\tilde{c}_1' > 1$ depends on L, D_1, α, σ.

Proof. First note that in (3.4), as a consequence of resampling N times and averaging, we have $\eta_i^+, \eta_i^- \sim \mathcal{N}(0, \sigma^2/N)$, and $\eta_i = \eta_i^+ - \eta_i^- \sim \mathcal{N}(0, 2\sigma^2/N), i \in [n]$. Hence we set `SPARSE-REC` $= (P1)$ with $\nu = \sqrt{2}(1+\varepsilon)\sigma\sqrt{n/N}$.

From Theorem 1, there exist constants $\tilde{c}_1, \tilde{c}_2 > 0$ (depending only on $c_1, c_2 > 0$ defined therein) so that for the choice $n \geq \tilde{c}_1|\mathcal{S}_1|\log(d/|\mathcal{S}_1|)$, $\mathbf{B}$ satisfies ℓ_2/ℓ_2 RIP with $\delta_{2|\mathcal{S}_1|} < \sqrt{2}-1$ with probability at least $1-2\exp(-\tilde{c}_2 n)$. Conditioning on this event, and invoking Corollary 1 for the stated choice of ν, we have for any given $\mathbf{x} \in \chi$ that with probability at least $1-2\exp(-c_3\varepsilon^2 n)$, the following holds

$$\|\widehat{\mathbf{z}^*}(\mathbf{x}) - \mathbf{z}^*(\mathbf{x})\|_\infty \leq \|\widehat{\mathbf{z}^*}(\mathbf{x}) - \mathbf{z}^*(\mathbf{x})\|_2 \leq \sqrt{2}C_2(1+\varepsilon)\sigma/\sqrt{N}. \tag{7.2}$$

By the union bound over the $2m+1$ elements of χ, it follows that (7.2) holds for all $\mathbf{x} \in \chi$ with probability at least

$$1 - 2|\chi|\exp(-c_3\varepsilon^2 n) = 1 - 2\exp(\log(2m+1) - c_3\varepsilon^2 n) \geq 1 - 2\exp\Big(-\frac{c_3\varepsilon^2 n}{2}\Big)$$

if $n \geq \frac{2\log(2m+1)}{c_3\varepsilon^2}$ holds. By (7.1), this condition is indeed satisfied and furthermore, we see that `SPARSE-REC` is $\epsilon = \sqrt{2}C_2(1+\varepsilon)\sigma/\sqrt{N}$ accurate for all $\mathbf{x} \in \chi$ with probability at least $1-2\exp(-\tilde{c}_2 n) - 2\exp(-\frac{c_3\varepsilon^2 n}{2})$. Now invoking Lemma 1, and using (7.2), it follows that if

$$\frac{\sqrt{2}C_2(1+\varepsilon)\sigma}{\sqrt{N}} < \frac{D_1}{3} \tag{7.3}$$

holds, then the stated choice of ϵ and m ensures exact recovery. Finally, (7.1) ensures that (7.3) holds and this completes the proof. □

7.2 Bivariate case

In the bivariate case from Section 4, we use Algorithm 2 for recovering the supports $\mathcal{S}_2, \mathcal{S}_1$ and with the Lemma 3 giving sufficient conditions for their exact recovery. Instantiating `SPARSE-REC`$_2$ with (P2) in (6.7), and `SPARSE-REC`$_1$ with (P1) in (6.2) gives then the results below.

Bounded noise model. The following theorem shows that if $\triangle$ is sufficiently small in the bounded noise model described before, then Algorithm 2 recovers $\mathcal{S}_2$ and $\mathcal{S}_1$ exactly provided the parameters m_i, n_i, ϵ_i, $i = 1, 2$ are chosen in a suitable way.

Theorem 6. *For the bounded noise model with the noise uniformly bounded by $\triangle$, consider Algorithm 2 with $\texttt{SPARSE-REC}_2$ instantiated with $(P2)$ with $\nu_2 = 4\triangle n_2$ in (6.7), and $\texttt{SPARSE-REC}_1$ realized by $(P1)$ with $\nu_1 = 2\triangle\sqrt{n_1}$ in (6.2), respectively. Let $\mathcal{H}_2^d$ be a $(d, 2)$ hash family. If*

$$\triangle < \min\left\{\frac{D_2}{12C_4}, \frac{D_1}{6C_2}\right\}, \ m_2 \geq \sqrt{2}\left(\frac{3L}{D_2}\right)^{1/\alpha}, \ n_2 \geq \tilde{c}_3|\mathcal{S}_2|\log(d^2/|\mathcal{S}_2|),$$

$$m_1 \geq (3L/D_1)^{1/\alpha} \quad \text{and} \quad n_1 \geq \tilde{c}_1|\mathcal{S}_1|\log\left(\frac{d - |\widehat{\mathcal{S}_2^{\text{var}}}|}{|\mathcal{S}_1|}\right)$$

are satisfied, then $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ and $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ with probability at least $1 - \exp(-\tilde{c}_4 n_2) - 2\exp(-\tilde{c}_2 n_1)$. Here, the constants $C_4, \tilde{c}_3, \tilde{c}_4 > 0$ are from Theorem 2, while $\tilde{c}_1, \tilde{c}_2, C_2 > 0$ are as defined in Theorem 4. The total number of queries made is

$$4(2m_2+1)^2 n_2|\mathcal{H}_2^d| + 2(2m_1+1)n_1 = \Omega\left(\bar{c}_2|\mathcal{S}_2|\log\left(\frac{d^2}{|\mathcal{S}_2|}\right)|\mathcal{H}_2^d| + \bar{c}_1|\mathcal{S}_1|\log\left(\frac{d - |\widehat{\mathcal{S}_2^{\text{var}}}|}{|\mathcal{S}_1|}\right)\right).$$

Here $\bar{c}_i > 1$ depends on L, D_i, α with $\bar{c}_1$ as in Theorem 4.

Proof. We focus on the proof of the exact recovery of $\mathcal{S}_2$. Once $\mathcal{S}_2$ is recovered exactly, the model reduces to a univariate SPAM on the variable set $\mathcal{P} = [d] \setminus \mathcal{S}_2^{\text{var}}$, with $\mathcal{S}_1 \subset \mathcal{P}$. Thereafter, the proof of the exact recovery of $\mathcal{S}_1$ is identical to the proof of Theorem 4.

Since $|\eta_{i,1}|, |\eta_{i,2}|, |\eta_{i,3}|, |\eta_{i,4}| \leq \triangle$ in (4.4) for $i \in [n_2]$, we obtain $\|\boldsymbol{\eta}\|_1 \leq n_2\|\boldsymbol{\eta}\|_\infty \leq 4n_2\triangle$. Therefore we set $\texttt{SPARSE-REC}_2 = (P2)$ with $\nu = 4\triangle n_2$.

Now, as a consequence of Theorem 2, there exist constants $\tilde{c}_3, \tilde{c}_4, C_4 > 0$ such that if $n_2 \geq \tilde{c}_3|\mathcal{S}_2|\log(d^2/|\mathcal{S}_2|)$, then with probability at least $1 - \exp(-\tilde{c}_4 n_2)$,

$$\|\widehat{\mathbf{z}^*}(\mathbf{x};\mathcal{A}) - \mathbf{z}^*(\mathbf{x};\mathcal{A})\|_\infty \ \leq \ \|\widehat{\mathbf{z}^*}(\mathbf{x};\mathcal{A}) - \mathbf{z}^*(\mathbf{x};\mathcal{A})\|_2 \ \leq 4C_4\triangle \quad \text{for all } \mathbf{x} \in \bigcup_{h\in\mathcal{H}_2^d} \chi(h).$$

This holds since $\mathbf{z}^*(\mathbf{x};\mathcal{A})$ is always at most $|\mathcal{S}_2|$ sparse. Thus, with probability at least $1 - \exp(-\tilde{c}_4 n_2)$, $\texttt{SPARSE-REC}_2$ is $\epsilon_2 = 4C_4\triangle$ accurate for each $h \in \mathcal{H}_2^d$, $\mathbf{x} \in \chi(h)$. Now invoking Lemma 3 reveals that if

$$4C_4\triangle < \frac{D_2}{3} \Leftrightarrow \triangle < \frac{D_2}{12C_4}$$

holds, then the stated choice of ϵ_2 and m_2 ensures $\widehat{\mathcal{S}_2} = \mathcal{S}_2$.

In order to derive the lower bound on the success probability of identifying $\mathcal{S}_1, \mathcal{S}_2$, we will make use of the following simple union bound inequality. For events $\mathcal{A}, \mathcal{B}$, it holds that

$$\mathbb{P}(\mathcal{A} \cup \mathcal{B}) = \mathbb{P}(\mathcal{A} \cup \{\mathcal{A}^c \cap \mathcal{B}\}) = \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{A}^c \cap \mathcal{B}) \leq \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B} \mid \mathcal{A}^c). \tag{7.4}$$

Hence with $\mathcal{A} = \left\{\widehat{\mathcal{S}_2} \neq \mathcal{S}_2\right\}$, $\mathcal{B} = \left\{\widehat{\mathcal{S}_1} \neq \mathcal{S}_1\right\}$, we readily arrive at the lower bound on the success probability.

Lastly, the bound on the total number of queries follows from Lemma 3 by plugging in the stated bounds on m_i, n_i; $i = 1, 2$. □

Gaussian noise model. Next, we consider the Gaussian noise model with the noise samples i.i.d. Gaussian $(\sim \mathcal{N}(0, \sigma^2))$ across queries. Similarly to Theorem 5, we show that if the noise variance σ^2 is sufficiently small, then Algorithm 2 recovers $\mathcal{S}_2$ and $\mathcal{S}_1$ for a careful choice of m_i, n_i, $i = 1, 2$. We again reduce the variance via re-sampling each query either N_2 times (during the estimation of $\mathcal{S}_2$) or N_1 times (during the estimation of $\mathcal{S}_1$), and averaging the values.

Theorem 7. *For the Gaussian noise model with i.i.d noise samples* $\sim \mathcal{N}(0,\sigma^2)$*, say we resample each query* N_2 *times (during estimation of* $\mathcal{S}_2$*) or* N_1 *times (during estimation of* $\mathcal{S}_1$*), and average the values. Consider Algorithm 2, where* $\texttt{SPARSE-REC}_2$ *and* $\texttt{SPARSE-REC}_1$ *are instantiated with* $(P2)$ *with* $\nu_2 = 2\sigma(1+\varepsilon)n_2/\sqrt{N_2}$ *in (6.7), and* $(P1)$ *with* $\nu_1 = \sqrt{2}(1+\varepsilon)\sigma\sqrt{n_1/N_1}$ *in (6.2), respectively, for some* $\varepsilon \in (0,1)$*. Let* $\mathcal{H}_2^d$ *be a* $(d,2)$ *hash family. If*

$$N_2 \geq \left\lfloor \frac{72C_4^2(1+\varepsilon)^2\sigma^2}{\pi D_2^2} \right\rfloor + 1, \; m_2 \geq \sqrt{2}\left(\frac{3L}{D_2}\right)^{1/\alpha}, \; n_2 \geq \max\left\{\tilde{c}_3|\mathcal{S}_2|\log\left(\frac{d^2}{|\mathcal{S}_2|}\right), \frac{2\log[(2m_2+1)^2 e|\mathcal{H}_2^d|]}{\tilde{c}_5\varepsilon^2}\right\}$$

$$N_1 \geq \left\lfloor \frac{18C_2^2(1+\varepsilon)^2\sigma^2}{D_1^2} \right\rfloor + 1, \; m_1 \geq \left(\frac{3L}{D_1}\right)^{1/\alpha}, \quad n_1 \geq \max\left\{c_1|\mathcal{S}_1|\log\left(\frac{d - |\widehat{\mathcal{S}_2^{\text{var}}}|}{|\mathcal{S}_1|}\right), \frac{2\log(2m_1+1)}{c_3\varepsilon^2}\right\}$$

hold, then $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ *and* $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ *with probability at least*

$$1 - \exp(-\tilde{c}_4 n_2) - \exp\left(-\frac{\tilde{c}_5 n_2\varepsilon^2}{2}\right) - 2\exp(-\tilde{c}_2 n_1) - 2\exp\left(-\frac{c_3\varepsilon^2 n_1}{2}\right).$$

The total number of queries made is

$$4(2m_2+1)^2 n_2 N_2|\mathcal{H}_2^d| + 2(2m_1+1)n_1 N_1 = \Omega\left(\bar{c}_2'|\mathcal{S}_2|\log\left(\frac{d^2}{|\mathcal{S}_2|}\right)|\mathcal{H}_2^d| + \bar{c}_1'|\mathcal{S}_1|\log\left(\frac{d-|\widehat{\mathcal{S}_2^{\text{var}}}|}{|\mathcal{S}_1|}\right)\right).$$

Here $\bar{c}_i' > 1$ *depends on* L, D_i, α, σ *with* $\bar{c}_1'$ *as in Theorem 5.*

Proof. As in the bounded noise model, we only prove the exact recovery of $\mathcal{S}_2$. First, we note that in (4.4), as a consequence of resampling each query N_2 times and averaging, $\eta_{i,1}, \eta_{i,2}, \eta_{i,3}, \eta_{i,4} \sim \mathcal{N}(0, \frac{\sigma^2}{N_2})$, $i \in [n_2]$ are independent. Therefore $\eta_i \sim \mathcal{N}(0, \frac{4\sigma^2}{N_2}), i \in [n_2]$ are also independent and we set $\texttt{SPARSE-REC}_2 = (P2)$ with $\nu_2 = 2\sigma(1+\varepsilon)n_2/\sqrt{N_2}$.

From Corollary 2, we know that there exist constants $\tilde{c}_3, \tilde{c}_4, \tilde{c}_5, C_4 > 0$ such that if $n_2 \geq \tilde{c}_3|\mathcal{S}_2|\log(d^2/|\mathcal{S}_2|)$, then for any given $h \in \mathcal{H}_2^d$, $\mathbf{x} \in \chi(h)$, we have for the stated choice of ν_2 that

$$\|\widehat{\mathbf{z}^*}(\mathbf{x};\mathcal{A}) - \mathbf{z}^*(\mathbf{x};\mathcal{A})\|_\infty \;\leq\; \|\widehat{\mathbf{z}^*}(\mathbf{x};\mathcal{A}) - \mathbf{z}^*(\mathbf{x};\mathcal{A})\|_1 \;\leq\; C_4\sqrt{\frac{2}{\pi}}(1+\varepsilon)\cdot\frac{2\sigma}{\sqrt{N_2}} =: \epsilon_2, \tag{7.5}$$

with probability at least $1 - \exp(-\tilde{c}_4 n_2) - e\cdot\exp\left(-\tilde{c}_5\varepsilon^2 n_2\right)$. This is true since $\mathbf{z}^*(\mathbf{x};\mathcal{A})$ is always at most $|\mathcal{S}_2|$ sparse. Therefore, by taking the union bound, (7.5) holds uniformly for all $h \in \mathcal{H}_2^d$, $\mathbf{x} \in \chi(h)$, with probability at least

$$\begin{aligned} &1 - \exp(-\tilde{c}_4 n_2) - (2m_2+1)^2|\mathcal{H}_2^d|e\cdot\exp\left(-\tilde{c}_5\varepsilon^2 n_2\right) \\ =&1 - \exp(-\tilde{c}_4 n_2) - \exp\left(\log[(2m_2+1)^2|\mathcal{H}_2^d|e] - \tilde{c}_5\varepsilon^2 n_2\right) \\ \geq&1 - \exp(-\tilde{c}_4 n_2) - \exp\left(-\frac{\tilde{c}_5 n_2\varepsilon^2}{2}\right) \end{aligned} \tag{7.6}$$

if $n_2 \geq \frac{2\log[(2m_2+1)^2 e|\mathcal{H}_2^d|]}{\tilde{c}_5\varepsilon^2}$ holds. We conclude that $\texttt{SPARSE-REC}_2$ is ϵ_2-accurate for each $h \in \mathcal{H}_2^d$, $\mathbf{x} \in \chi(h)$ with probability at least (7.6).

By the stated choice of N_2, we have $\epsilon_2 < \frac{D_2}{3}$ and using Lemma 3 and the condition on m_2, we obtain $\widehat{\mathcal{S}_2} = \mathcal{S}_2$. The lower bound on the success probability of identifying $\mathcal{S}_1, \mathcal{S}_2$ follows via the same argument as in the proof of Theorem 6. Lastly, the bound on the total number of queries follows from the expression in Lemma 3 (taking the resampling into account) by plugging in the stated bounds on m_i, n_i, N_i; $i = 1,2$. □

Remark 8. *As discussed in Section 4, we can construct $\mathcal{H}_2^d$ via a simple randomized method (in time linear in output size) (eg., [15, Section 5]) where $|\mathcal{H}_2^d| = O(\log d)$, with probability at least $1 - d^{-\Omega(1)}$. Plugging this into Theorem 6 leads to a (worst case) sample complexity of*

$$\Omega\left(\bar{c}_2|\mathcal{S}_2|\log(d^2/|\mathcal{S}_2|)\log d + \bar{c}_1|\mathcal{S}_1|\log(d/|\mathcal{S}_1|)\right)$$

in the setting of arbitrary bounded noise.

7.3 Multivariate case

Finally, we consider the most general multivariate setting of Section 5. Based on Lemma 5, we analyze Algorithm 3 recovering the supports $\mathcal{S}_i$; $i \in [r_0]$. The recovery routines $\texttt{SPARSE-REC}_i$ are realized by (P1) in (6.2) for $i = 1$ and for each $3 \le i \le r_0$ and $\texttt{SPARSE-REC}_2$ is instantiated with (P2) in (6.7). This leads to the results below.

Bounded noise model. This is the same noise model as described in the preceding subsections. The following theorem shows that if $\triangle$ is sufficiently small, then Algorithm 3 recovers $\mathcal{S}_i$ exactly for all $i \in [r_0]$ provided the parameters $(m_i, n_i)_{i=1}^{r_0}$ are well chosen.

Theorem 8. *For the bounded noise model with noise uniformly bounded by $\triangle$, consider Algorithm 3 with*

(a) $\texttt{SPARSE-REC}_i$ instantiated with (P1) with $\nu_i = 2^i\triangle\sqrt{n_i}$ in (6.2) for $i \in \{3,\dots,r_0\}$,

(b) $\texttt{SPARSE-REC}_2$ instantiated with (P2) with $\nu_2 = 4\triangle n_2$ in (6.7),

(c) $\texttt{SPARSE-REC}_1$ instantiated with (P1) with $\nu_1 = 2\triangle\sqrt{n_1}$ in (6.2), respectively.

Let $\mathcal{H}_i^{\mathcal{P}_i}$ be a $(\mathcal{P}_i, i)$ hash family for each $2 \le i \le r_0$ and let $\mathcal{P}_i$ denote the set $\mathcal{P} \subseteq [d]$ at the beginning of i^{th} iteration (with $\mathcal{P}_{r_0} = [d]$). If

$$\triangle < \min\Big\{\min_{i\in\{3,\dots,r_0\}} \frac{D_i}{2^i 3C_2}, \frac{D_2}{12C_4}, \frac{D_1}{6C_2}\Big\}, \quad m_i \ge \left(\frac{3L(\sqrt{i})^\alpha}{D_i}\right)^{1/\alpha}; \quad i \in [r_0],$$
$$n_i \ge \tilde{c}_6|\mathcal{S}_i|^2 \log\binom{|\mathcal{P}_i|}{i}; \quad i \in \{3,\dots,r_0\},$$
$$n_2 \ge \tilde{c}_3|\mathcal{S}_2|\log(|\mathcal{P}_2|^2/|\mathcal{S}_2|) \quad \textit{and} \quad n_1 \ge \tilde{c}_1|\mathcal{S}_1|\log(|\mathcal{P}_1|/|\mathcal{S}_1|)$$

are satisfied, then $\widehat{\mathcal{S}}_i = \mathcal{S}_i$ for all $i \in [r_0]$ with probability at least

$$1 - \sum_{i=3}^{r_0}\exp\left(-\frac{\tilde{c}_7 n_i}{|\mathcal{S}_i|^2}\right) - \exp(-\tilde{c}_4 n_2) - 2\exp(-\tilde{c}_2 n_1).$$

Here, the constants $C_4, \tilde{c}_3, \tilde{c}_4 > 0$ are from Theorem 2, while $\tilde{c}_1, \tilde{c}_2, C_2 > 0$ are as defined in Theorem 4. The constants $\tilde{c}_6, \tilde{c}_7 > 0$ depend only on the constants $c_6, c_7 > 0$ defined in Theorem 3. The total number of queries made is

$$\sum_{i=1}^{r_0} 2^i(2m_i+1)^i n_i|\mathcal{H}_i^{\mathcal{P}_i}|$$
$$= \Omega\left(\sum_{i=3}^{r_0}\Big[c_i^i i|\mathcal{S}_i|^2\log(|\mathcal{P}_i|)|\mathcal{H}_i^{\mathcal{P}_i}|\Big] + \bar{c}_2|\mathcal{S}_2|\log\left(\frac{|\mathcal{P}_2|^2}{|\mathcal{S}_2|}\right)|\mathcal{H}_2^d| + \bar{c}_1|\mathcal{S}_1|\log\left(\frac{|\mathcal{P}_1|}{|\mathcal{S}_1|}\right)\right)$$

where for $i = 3,\dots,r_0$, each $c_i > 1$ depends on D_i,L, α, i, and $\bar{c}_1, \bar{c}_2 > 1$ are as in Theorem 6.

Proof. Say we are at the beginning of i^{th} iteration with $3 \le i \le r_0$ and $\widehat{\mathcal{S}}_l = \mathcal{S}_l$ holds true for each $l > i$. Hence, the model has reduced to an order i sparse additive model on the set $\mathcal{P}_i \subseteq [d]$, with $\mathcal{S}_i^{(1)}, \mathcal{S}_{i-1}^{(1)}, \dots, \mathcal{S}_1 \subset \mathcal{P}_i$.

From (5.5), we see for the noise vector $\boldsymbol{\eta} \in \mathbb{R}^{n_i}$ that

$$\eta_s = \sum_{z=1}^{2^i} (-1)^{\texttt{digit}(z-1)} \eta_{s,z} \quad \text{for all } s \in [n_i]. \tag{7.7}$$

Since $|\eta_{s,z}| \le \triangle$, this implies $\|\boldsymbol{\eta}\|_\infty \le 2^i \triangle$ and thus $\|\boldsymbol{\eta}\|_2 \le 2^i \triangle \sqrt{n_i}$. This bound holds uniformly for each $\mathbf{x}$ at which the linear system is formed. So we now instantiate $\texttt{SPARSE-REC}_i$ with $(P1)$ in (6.2), with $\nu_i = 2^i \triangle \sqrt{n_i}$.

As a consequence of part 1 of Theorem 3, there exists constants $\tilde{c}_6, \tilde{c}_7 > 0$ depending on $c_6, c_7 > 0$ (as defined in Theorem 3) so that if $n_i \ge \tilde{c}_6 |\mathcal{S}_i|^2 \log \binom{|\mathcal{P}_i|}{i}$, then with probability at least $1 - \exp\left(-\frac{\tilde{c}_7 n_i}{|\mathcal{S}_i|^2}\right)$, the matrix $\mathbf{B} \in \mathbb{R}^{n_i \times \binom{|\mathcal{P}_i|}{i}}$ satisfies ℓ_2/ℓ_2 RIP with $\delta_{2|\mathcal{S}_i|} < \sqrt{2} - 1$. Conditioning on this event, it follows from Theorem 3 that

$$\|\widehat{\mathbf{z}^*}(\mathbf{x}; \mathcal{A}) - \underbrace{\mathbf{z}^*(\mathbf{x}; \mathcal{A})}_{|\mathcal{S}_i| \text{ sparse}}\|_\infty \le \ \|\widehat{\mathbf{z}^*}(\mathbf{x}; \mathcal{A}) - \mathbf{z}^*(\mathbf{x}; \mathcal{A})\|_2 \ \le 2^i \triangle C_2 =: \epsilon_i \text{ for all } \mathbf{x} \in \bigcup_{h \in \mathcal{H}_i^{\mathcal{P}_i}} \chi(h).$$

Thus, with probability at least $1 - \exp(-\frac{\tilde{c}_7 n_i}{|\mathcal{S}_i|^2})$, $\texttt{SPARSE-REC}_i$ is ϵ_i-accurate for each $h \in \mathcal{H}_i^{\mathcal{P}_i}$, $\mathbf{x} \in \chi(h)$. The assumption on $\triangle$ ensures that $\epsilon_i < D_i/3$ and by Lemma 5 it follows that the stated choice of (m_i, ϵ_i) ensures exact recovery of $\mathcal{S}_i$.

Hence if

$$\triangle < \min_{i \in \{3, \dots, r_0\}} \frac{D_i}{2^i 3 C_2}$$

is satisfied and m_i, n_i, ϵ_i satisfy their stated bounds (for $3 \le i \le r_0$), then we can lower bound the probability that $\widehat{\mathcal{S}}_i = \mathcal{S}_i$ holds for all $i = 3, \dots, r_0$ via the following simple generalization of (7.4). For events $\mathcal{A}_3, \dots, \mathcal{A}_{r_0}$ it holds that

$$\mathbb{P}(\cup_{i=3}^{r_0} \mathcal{A}_i) \le \mathbb{P}(\mathcal{A}_{r_0}) + \mathbb{P}(\mathcal{A}_{r_0-1} \mid \mathcal{A}_{r_0}^c) + \mathbb{P}(\mathcal{A}_{r_0-2} \mid \mathcal{A}_{r_0}^c \cap \mathcal{A}_{r_0-1}^c) + \cdots + \mathbb{P}(\mathcal{A}_3 \mid \cap_{i=4}^{r_0} \mathcal{A}_i^c). \tag{7.8}$$

Therefore plugging $\mathcal{A}_i = \left\{\widehat{\mathcal{S}}_i \neq \mathcal{S}_i\right\}$ in (7.8), we readily have that $\widehat{\mathcal{S}}_i = \mathcal{S}_i$ holds for all $3 \le i \le r_0$ with probability at least

$$1 - \sum_{i=3}^{r_0} \exp\left(-\frac{\tilde{c}_7 n_i}{|\mathcal{S}_i|^2}\right).$$

Once $\mathcal{S}_i$ are identified exactly for all $3 \le i \le r_0$, we are left with a bivariate SPAM on the set $\mathcal{P}_2$, with $\mathcal{S}_1, \mathcal{S}_2^{(1)} \subset \mathcal{P}_2$. Therefore by invoking Theorem 6, we see that if furthermore $\triangle < \min\left\{\frac{D_2}{12 C_4}, \frac{D_1}{6 C_2}\right\}$ holds, and m_1, n_1, m_2, n_2 satisfy their respective stated conditions, then the stated instantiations of $\texttt{SPARSE-REC}_1, \texttt{SPARSE-REC}_2$ (along with the stated choices of ν_1, ν_2) ensures $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ and $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ with probability at least $1 - \exp(-\tilde{c}_4 n_2) - 2\exp(-\tilde{c}_2 n_1)$. The stated lower bound on the success probability of identifying each $\mathcal{S}_i$ ($i = 1, \dots, r_0$) follows readily from (7.4) by plugging $\mathcal{A} = \cap_{i=3}^{r_0} \left\{\widehat{\mathcal{S}}_i \neq \mathcal{S}_i\right\}$ and $\mathcal{B} = \cap_{i=1}^{2} \left\{\widehat{\mathcal{S}}_i \neq \mathcal{S}_i\right\}$. This completes the proof for exact recovery of $\mathcal{S}_i$'s for all $i \in [r_0]$.

Finally, the stated sample complexity bound for the total number of queries made by Algorithm 3 follows in a straightforward manner by plugging in

$$m_i = \Omega\left(3^{1/\alpha} \sqrt{i} L^{1/\alpha} D_i^{-1/\alpha}\right), \quad i \in [r_0]; \quad n_i = \Omega(i |\mathcal{S}_i|^2 \log |\mathcal{P}_i|), \quad 3 \le i \le r_0$$

along with the complexity bounds for n_1, n_2 into the expression for total number of samples from Lemma 5. This completes the proof. $\square$

Gaussian noise model. In the Gaussian noise model with noise samples i.i.d. Gaussian ($\sim \mathcal{N}(0, \sigma^2)$) across queries, we again reduce the variance via re-sampling each query N_i times (during the estimation of $\mathcal{S}_i$) for every $i \in [r_0]$ and averaging the values. We show that if the noise variance σ^2 is sufficiently small, then Algorithm 3 recovers $\mathcal{S}_i$ exactly for each $i \in [r_0]$, provided the parameters m_i, n_i; $i \in [r_0]$, are well chosen.

Theorem 9. *For the Gaussian noise model with i.i.d. noise samples $\sim \mathcal{N}(0, \sigma^2)$, consider Algorithm 3 wherein we resample each query N_i times during estimation of $\mathcal{S}_i$ and average the values. Let*

(a) SPARSE-REC$_i$ = (P1) with $\nu_i = 2^{i/2}(1+\varepsilon)\sigma\sqrt{n_i/N_i}$ in (6.2) for $3 \leq i \leq r_0$,

(b) SPARSE-REC$_2$ = (P2) with $\nu_2 = 2(1+\varepsilon)\sigma n_2/\sqrt{N_2}$ in (6.7) and

(c) SPARSE-REC$_1$ = (P1) with $\nu_1 = \sqrt{2}(1+\varepsilon)\sigma\sqrt{n_1/N_1}$ in (6.2), respectively,

for some $\varepsilon \in (0,1)$. Let $\mathcal{H}_i^{\mathcal{P}_i}$ be a $(\mathcal{P}_i, i)$ hash family for each $2 \leq i \leq r_0$, and denote $\mathcal{P}_i$ to be the set $\mathcal{P} \subseteq [d]$ at the beginning of i^{th} iteration with $\mathcal{P}_{r_0} = [d]$. If

$$N_i \geq \left\lfloor \frac{9C_2^2(1+\varepsilon)^2 2^i \sigma^2}{D_i^2} \right\rfloor + 1, \quad n_i \geq \max\left\{ \tilde{c}_6 |\mathcal{S}_i|^2 \log\binom{|\mathcal{P}_i|}{i}, \frac{2\log[(2m_i+1)^i |\mathcal{H}_i^{\mathcal{P}_i}|]}{c_3\varepsilon^2} \right\}; \quad 3 \leq i \leq r_0$$

$$m_i \geq \left(\frac{3L(\sqrt{i})^\alpha}{D_i}\right)^{1/\alpha}; \quad i \in [r_0],$$

$$N_2 \geq \left\lfloor \frac{72C_4^2(1+\varepsilon)^2\sigma^2}{\pi D_2^2} \right\rfloor + 1, \ n_2 \geq \max\left\{ \tilde{c}_3 |\mathcal{S}_2| \log\left(\frac{|\mathcal{P}_2|^2}{|\mathcal{S}_2|}\right), \frac{2\log[(2m_2+1)^2 e |\mathcal{H}_2^{\mathcal{P}_2}|]}{\tilde{c}_5\varepsilon^2} \right\},$$

$$N_1 \geq \left\lfloor \frac{18C_2^2(1+\varepsilon)^2\sigma^2}{D_1^2} \right\rfloor + 1, \ n_1 \geq \max\left\{ \tilde{c}_1 |\mathcal{S}_1| \log\left(\frac{|\mathcal{P}_1|}{|\mathcal{S}_1|}\right), \frac{2\log(2m_1+1)}{c_3\varepsilon^2} \right\}$$

hold, then $\widehat{\mathcal{S}_i} = \mathcal{S}_i$ for all $i \in [r_0]$ with probability at least

$$1 - \sum_{i=3}^{r_0} \left[\exp\left(-\frac{\tilde{c}_7 n_i}{|\mathcal{S}_i|^2}\right) + 2\exp\left(-\frac{c_3\varepsilon^2 n_i}{2}\right) \right]$$
$$- \exp(-\tilde{c}_4 n_2) - \exp\left(-\frac{\tilde{c}_5 n_2 \varepsilon^2}{2}\right) - 2\exp(-\tilde{c}_2 n_1) - 2\exp\left(-\frac{c_3\varepsilon^2 n_1}{2}\right). \tag{7.9}$$

The constants $C_4, \tilde{c}_3, \tilde{c}_4, \tilde{c}_1, \tilde{c}_2, C_2, \tilde{c}_6, \tilde{c}_7 > 0$ are as explained in Theorem 8, while $\tilde{c}_5, c_3 > 0$ come from Corollaries 2, 1 respectively. The total number of queries made is

$$\sum_{i=1}^{r_0} N_i 2^i (2m_i+1)^i n_i |\mathcal{H}_i^{\mathcal{P}_i}|$$
$$= \Omega\left(\sum_{i=3}^{r_0} \left[\bar{c}_i' \bar{c}_i^i i |\mathcal{S}_i|^2 \log(|\mathcal{P}_i|) |\mathcal{H}_i^{\mathcal{P}_i}| \right] + \bar{c}_2' |\mathcal{S}_2| \log\left(\frac{|\mathcal{P}_2|^2}{|\mathcal{S}_2|}\right) |\mathcal{H}_2^{\mathcal{P}_2}| + \bar{c}_1' |\mathcal{S}_1| \log\left(\frac{|\mathcal{P}_1|}{|\mathcal{S}_1|}\right) \right)$$

where for $i = 3, \ldots, r_0$, $\bar{c}_i' > 1$ depends on σ, D_i, and $\bar{c}_i > 1$ depends on L, D_i, α, i. Moreover, $\bar{c}_1', \bar{c}_2' > 1$ are as in Theorem 7.

Proof. Again, in the beginning of i^{th} iteration $3 \le i \le r_0$ with $\widehat{\mathcal{S}_l} = \mathcal{S}_l$ for each $l > i$, the model has reduced to an order i sparse additive model on the set $\mathcal{P}_i \subseteq [d]$, with $\mathcal{S}_i^{(1)}, \mathcal{S}_{i-1}^{(1)}, \dots, \mathcal{S}_1 \subset \mathcal{P}_i$.

The noise vector is again given by (7.7) As a consequence of resampling each query point N_i times and averaging, we get $\eta_{s,z} \sim \mathcal{N}(0, \frac{\sigma^2}{N_i})$ i.i.d. for all s, z and, therefore, $\eta_s \sim \mathcal{N}(0, \frac{2^i\sigma^2}{N_i})$ i.i.d. for each s. From part 1 of Theorem 3, we know that if $n_i \ge \tilde{c}_6 |\mathcal{S}_i|^2 \log \binom{|\mathcal{P}_i|}{i}$, then with probability at least $1 - \exp\left(-\frac{\tilde{c}_7 n_i}{|\mathcal{S}_i|^2}\right)$ (with $\tilde{c}_6, \tilde{c}_7$ depending only on c_6, c_7), the matrix $\mathbf{B} \in \mathbb{R}^{n_i \times \binom{|\mathcal{P}_i|}{i}}$ satisfies ℓ_2/ℓ_2 RIP with $\delta_{2|\mathcal{S}_i|} < \sqrt{2} - 1$. Let us condition on this event. Then by setting $\texttt{SPARSE-REC}_i = (P1)$ with $\nu_i = (1+\varepsilon)\left(\frac{2^{i/2}\sigma\sqrt{n_i}}{\sqrt{N_i}}\right)$, and invoking Corollary 1, it follows for any given $h \in \mathcal{H}_i^{\mathcal{P}_i}$, $\mathbf{x} \in \chi(h)$ that

$$\|\widehat{\mathbf{z}^*}(\mathbf{x};\mathcal{A}) - \underbrace{\mathbf{z}^*(\mathbf{x};\mathcal{A})}_{|\mathcal{S}_i| \text{ sparse}}\|_\infty \ \le \ \|\widehat{\mathbf{z}^*}(\mathbf{x};\mathcal{A}) - \mathbf{z}^*(\mathbf{x};\mathcal{A})\|_2 \ \le \ 2^{i/2} C_2 (1+\varepsilon)\sigma\sqrt{n_i/N_i} =: \epsilon_i \tag{7.10}$$

with probability at least $1 - 2\exp(-c_3\varepsilon^2 n_i)$. By the union bound, it follows that (7.10) holds for all $h \in \mathcal{H}_i^{\mathcal{P}_i}$, $\mathbf{x} \in \chi(h)$, with probability at least

$$\begin{aligned} 1 - 2(2m_i+1)^i |\mathcal{H}_i^{\mathcal{P}_i}| \exp(-c_3\varepsilon^2 n_i) &= 1 - 2\exp[\log[(2m_i+1)^i |\mathcal{H}_i^{\mathcal{P}_i}|] - c_3\varepsilon^2 n_i] \\ &\ge 1 - 2\exp\left(-\frac{c_3\varepsilon^2 n_i}{2}\right) \end{aligned}$$

if $n_i \ge \frac{2\log[(2m_i+1)^i|\mathcal{H}_i^{\mathcal{P}_i}|]}{c_3\varepsilon^2}$ holds. This gives us the stated condition on n_i for $3 \le i \le r_0$. Hence for the aforementioned choice of n_i, $\texttt{SPARSE-REC}_i$ is ϵ_i-accurate for each $h \in \mathcal{H}_i^{\mathcal{P}_i}$, $\mathbf{x} \in \chi(h)$, with probability at least $1 - 2\exp\left(-\frac{c_3\varepsilon^2 n_i}{2}\right) - \exp\left(-\frac{\tilde{c}_7 n_i}{|\mathcal{S}_i|^2}\right)$. By the condition on N_i, we obtain $\epsilon_i < D_i/3$ and from Lemma 5, it follows that for the stated choice of m_i and ϵ_i, we have $\widehat{\mathcal{S}_i} = \mathcal{S}_i$. Thus we conclude by the union bound in (7.8) that $\widehat{\mathcal{S}_i} = \mathcal{S}_i$ holds for all $3 \le i \le r_0$, with high probability, for the stated choice of $m_i, n_i, \epsilon_i, N_i$.

Finally, say $\widehat{\mathcal{S}_i} = \mathcal{S}_i$ holds for all $3 \le i \le r_0$. Then we are left with a bivariate SPAM on the set $\mathcal{P}_2$, with $\mathcal{S}_1, \mathcal{S}_2^{(1)} \subset \mathcal{P}_2$. Thereafter, we only need to invoke Theorem 7, which guarantees that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ and $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ holds with high probability for the stated choices of $\texttt{SPARSE-REC}_i, \nu_i, m_i, n_i, \epsilon_i, N_i$; $i = 1, 2$. The lower bound on the success probability of identifying each $\mathcal{S}_i$ ($i = 1, \dots, r_0$) then follows via the same argument as in the proof of Theorem 8. This completes the proof for the exact recovery of $\mathcal{S}_i$'s.

Since each query is resampled N_i times, with $N_i = 1 + \Omega(2^i\sigma^2/D_i^2)$, we obtain the stated sample complexity bound by proceeding as in the proof of Theorem 8. □

Remark 9. *As discussed in Section 4, we can construct $\mathcal{H}_t^d$ (for $t \ge 2$) via a simple randomized method (in time linear in output size) (eg., [15, Section 5]) where $|\mathcal{H}_t^d| = O(te^t \log d)$, with probability at least $1 - d^{-\Omega(t)}$. Plugging this into Theorem 8 leads to a (worst case) sample complexity of*

$$\Omega\left(\sum_{i=3}^{r_0}\left[c_i^i e^i i^2 |\mathcal{S}_i|^2 (\log(|\mathcal{P}_i|))^2\right] + \bar{c}_2|\mathcal{S}_2|\log\left(\frac{|\mathcal{P}_2|^2}{|\mathcal{S}_2|}\right)\log(|\mathcal{P}_2|) + \bar{c}_1|\mathcal{S}_1|\log\left(\frac{|\mathcal{P}_1|}{|\mathcal{S}_1|}\right)\right)$$

in the setting of arbitrary bounded noise. Since $|\mathcal{P}_i| = O(d)$, this leads to the expression in (1.2).

8 Discussion

We start by providing a brief summary of our results with remarks concerning certain aspects of our algorithm. We then discuss how the components ϕ can be identified, and also compare our results with closely related work. Finally, we discuss an alternative approach described in [43] in more detail.

Summary of our results. Let us recall that in the setting where the queries are corrupted with arbitrary bounded noise, Algorithm 3 succeeds with high probability in identifying each $\mathcal{S}_i$ for $i = 1, \dots, r_0$ provided the noise level is sufficiently small, and makes

$$\Omega\left(\sum_{i=3}^{r_0} \Big[\underbrace{c_i^i i^2 |\mathcal{S}_i|^2 \log^2 d}_{\text{Identifying } \mathcal{S}_i}\Big] + \underbrace{c_2 |\mathcal{S}_2| \log\left(\frac{d^2}{|\mathcal{S}_2|}\right) \log d}_{\text{Identifying } \mathcal{S}_2} + \underbrace{c_1 |\mathcal{S}_1| \log\left(\frac{d}{|\mathcal{S}_1|}\right)}_{\text{Identifying } \mathcal{S}_1}\right)$$

queries. This is the same expression in (1.2) and is a consequence of Theorem 8 (see Remark 9). For each $3 \leq i \leq r_0$, the term $c_i^i i^2 |\mathcal{S}_i|^2 \log^2 d$ represents the sample complexity of identifying $\mathcal{S}_i$. In particular, c_i^i arises from the size of the i-dimensional grid $\chi(h)$ for a given hash function h in Algorithm 3, with c_i depending on the smoothness parameters L, α, D_i and scaling as $\sqrt{i}$ with i. The term $i|\mathcal{S}_i|^2 \log d = \Theta(|\mathcal{S}_i|^2 \log \binom{d}{i})$ arises from the sample complexity of estimating a i^{th} order sparse multilinear function in d variables (with $|\mathcal{S}_i|$ terms) and follows from Theorem 3. Finally, the term $ie^i \log d$ arises from the size of a (d, i) hash family (see Remark 9) where the e^i factor is subsumed by c_i. The term $c_2 |\mathcal{S}_2| \log\left(\frac{d^2}{|\mathcal{S}_2|}\right) \log d$ is the sample complexity for identification of $\mathcal{S}_2$. Here, c_2 arises from the size of a two-dimensional grid in $[-1, 1]^2$; $|\mathcal{S}_2| \log\left(\frac{d^2}{|\mathcal{S}_2|}\right)$ is the sample complexity of estimating a sparse bilinear function (see Theorem 2), and $\log d$ arises from the size of a $(d, 2)$ hash family (see Remark 8). Finally, the term $c_1 |\mathcal{S}_1| \log\left(\frac{d}{|\mathcal{S}_1|}\right)$ is the sample complexity for identification of $\mathcal{S}_1$ where c_1 arises from the size of a grid in $[-1, 1]$ and $|\mathcal{S}_1| \log\left(\frac{d}{|\mathcal{S}_1|}\right)$ is the sample complexity of estimating a sparse linear function (see Theorem 1).

The following points are worth noting for our algorithms.

- In general, we do not need to know the exact values of the smoothness parameters α, L, D_i for $i = 1, 2, ..., r_0$. It suffices to use lower bounds for α, D_i and an upper bound for L. Similarly, it is also possible to work with just the upper bound estimates of r_0 and $|\mathcal{S}_1|, \dots, |\mathcal{S}_{r_0}|$.
- The computational cost of Algorithm 3 is typically dominated by `SPARSE-REC`. At iteration $i = r_0$, for a given hash function h and base point $\mathbf{x} \in \chi(h)$, the computational complexity of `SPARSE-REC` is at most polynomial in (n_{r_0}, d^{r_0}) (recall Remark 6). Since there are $O(m_{r_0}^{r_0} |\mathcal{H}_{r_0}^d|)$ base points, the total cost (over $i = 1, \dots, r_0$) is at most polynomial in the number of queries and d^{r_0}.

Identification of $\phi_{\mathbf{j}}$'s. Once the sets $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{r_0}$ are known, then one can identify each component in the representation (5.1) by querying f along the corresponding canonical subspaces. Indeed, for a given $1 \leq p \leq r_0$, and $1 \leq r \leq p$, let us see how we can identify the r-variate component $\phi_{\mathbf{j}}$ for a given $\mathbf{j} \in \mathcal{S}_p^{(r)}$. Consider any $\mathbf{x} = (x_1 \dots x_d)^T \in [-1, 1]^d$ which is supported on $\mathcal{S}_p^{(1)}$, i.e., $x_l = 0$ if $l \notin \mathcal{S}_p^{(1)}$. We then have from (5.1) that

$$f(\mathbf{x}) = \mu + \sum_{u \in \mathcal{S}_p^{(1)}} \phi_u(x_u) + \sum_{\mathbf{u} \in \mathcal{S}_p^{(2)}} \phi_{\mathbf{u}}(x_{\mathbf{u}}) \quad + \dots + \sum_{\mathbf{u} \in \mathcal{S}_p^{(p-1)}} \phi_{\mathbf{u}}(x_{\mathbf{u}}) + \sum_{\mathbf{u} \in \mathcal{S}_p} \phi_{\mathbf{u}}(x_{\mathbf{u}}).$$

In particular, it follows from (2.3) that

$$\sum_{\mathbf{i} \subseteq \mathbf{j}} (-1)^{|\mathbf{j}|-|\mathbf{i}|} f(\Pi_{\mathbf{i}}(\mathbf{x})) = \phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}), \tag{8.1}$$

with $\Pi_{\mathbf{i}}(\mathbf{x})$ denoting the projection of $\mathbf{x}$ on the set of variables $\mathbf{i}$. Hence by querying f at the $2^{|\mathbf{j}|}$ points $\{\Pi_{\mathbf{i}}(\mathbf{x}) : \mathbf{i} \subseteq \mathbf{j}\}$, we can obtain the sample $\phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}})$ via (8.1). Consequently, by choosing $\mathbf{x}_j$ from a regular grid on $[-1,1]^{\mathbf{j}}$, we can estimate $\phi_{\mathbf{j}}$ from its corresponding samples via standard quasi interpolants (in the noiseless case) or tools from non parametric regression (in the noisy case).

SPAMs. To begin with, we note that our work generalizes the recent results of Tyagi et al. [42, 43] in two fundamental ways. Firstly, we provide an algorithm for the general case where $r_0 \geq 2$ is possible, the results in [42, 43] are for the case $r_0 = 2$. Secondly, our results only require f to be Hölder continuous and not continuously differentiable as is the case in [42, 43]. We also mention that our sampling bounds for the case $r_0 = 2$ are linear in the sparsity $|\mathcal{S}_1| + |\mathcal{S}_2|$ even when the noise samples are i.i.d. Gaussian. However the algorithms in [42, 43] have super-linear dependence on the sparsity in this noise model. This is unavoidable in [42, 43] due to the localized nature of the sampling scheme, wherein finite difference operations are used to obtain linear measurements of the sparse gradient and Hessian of f. In the presence of noise, this essentially leads to the noise level getting scaled up by the step size parameter, and thus reducing the variance of noise necessarily leads to a resampling factor which is super linear in sparsity.

Dalalyan et al. [14] recently studied models of the form (1.1) with a set $\mathcal{S}_{r_0}$ of r_0-wise interaction terms. They considered the Gaussian white noise model, which while not the same as the usual regression setup, is known to be asymptotically equivalent to the same. They derived non-asymptotic L_2 error rates in expectation for an estimator, with f lying in a Sobolev space, and showed the rates to be minimax optimal. However, they do not consider the problem of identification of the interaction terms. Moreover, as noted in [14], the computational cost of their method typically scales exponentially in $|\mathcal{S}_{r_0}|, r_0, d$. Yang et al. [49] also studied models of the form (1.1) in a Bayesian setting, wherein they place a Gaussian prior (GP) on f and can carry out inference via the posterior probability distribution. They derive an estimator and provide error rates in the empirical L_2 norm for Hölder smooth f, but do not address the problem of identifying the interaction terms.

Functions with few active variables. There has been a fair amount of work in the literature on functions which intrinsically depend on a small subset of $k \ll d$ variables [15, 38, 13, 12]. To our knowledge, this model was first considered in [15], and in fact, our idea of using a family of hash functions is essentially motivated from [15] wherein such a family was used to construct the query points. A prototypical result in [15] is an algorithm that identifies the set of active variables with $(m+1)^k|\mathcal{H}_k^d| + k \log d$ queries, with $m > 0$ being the number of points on a uniform grid along a coordinate. The randomized construction of $\mathcal{H}_k^d$ yields $|\mathcal{H}_k^d| = O(ke^k \log d)$ (see [15, Section 5]) which results in a sample complexity of $m^k e^k k \log d$. The exponential dependence on k is unavoidable in the worst case, and indeed our bounds are also exponential in r_0 (see (1.2)).

- When $r_0 \geq k$, (1.1) is clearly a generalization of this model.
- In general, the model (1.1) is also a function of few active variables (those part of $\mathcal{S}_i$'s); more precisely, at most $\sum_{i=1}^{r_0} i|\mathcal{S}_i|$ variables. However, using a method that is designed generically for learning intrinsically k-variate functions would typically have sample complexity scaling exponentially with $\sum_{i=1}^{r_0} i|\mathcal{S}_i|$. This is clearly suboptimal; our bounds in general depend at most polynomially on the size of $\mathcal{S}_i$'s. This dependence is actually linear for the case $r_0 = 2$.

An alternative approach. Next, we discuss an alternative approach for learning the model (5.1) which was mentioned already in [43]. It is based on a truncated expansion in a bounded orthonormal system. For simplicity, we assume that f takes the form

$$f(\mathbf{x}) = \sum_{\mathbf{j}\in\mathcal{S}_{r_0}} \phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}), \quad \mathbf{x} \in [-1,1]^d. \tag{8.2}$$

Let $\{\psi_k\}_{k\in\mathbb{Z}}$ be an orthonormal basis in $L_2([-1,1])$ with respect to the normalized Lebesgue measure on $[-1,1]$. We assume that $\psi_0 \equiv 1$ and we define $\{\psi_{\mathbf{i}}\}_{\mathbf{i}\in\mathbb{Z}^d}$ to be the tensor product orthonormal basis in $L_2([-1,1]^d)$, where

$$\psi_{\mathbf{i}}(\mathbf{x}) = \bigotimes_{l=1}^{d} \psi_{i_l}(x_l), \quad \mathbf{x} \in [-1,1]^d.$$

For the components of (8.2) we obtain (for each $\mathbf{j} \in \mathcal{S}_{r_0}$) the decomposition

$$\phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}) = \sum_{\mathbf{i}\in\mathbb{Z}^d} a_{\mathbf{j},\mathbf{i}}\psi_{\mathbf{i}}(\mathbf{x}_{\mathbf{j}}) = \sum_{\mathbf{i}\in\mathbb{Z}^d:\operatorname{supp}\mathbf{i}\subseteq\mathbf{j}} a_{\mathbf{j},\mathbf{i}}\psi_{\mathbf{i}}(\mathbf{x}_{\mathbf{j}}). \tag{8.3}$$

In the last identity, we used that $\int_{-1}^{1}\psi_k(t)dt = 0$ for every $k \in \mathbb{Z}\setminus\{0\}$ and, therefore, $a_{\mathbf{j},\mathbf{i}} = \langle\phi_{\mathbf{j}},\psi_{\mathbf{i}}\rangle = 0$ if $i_l \neq 0$ for some $l \notin \mathbf{j}$.

Using the smoothness of $\phi_{\mathbf{j}}$, we can truncate (8.3) at level $N \in \mathbb{N}$ (which we will determine later) and obtain

$$\phi_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}}) = \sum_{\substack{\mathbf{i}\in\mathbb{Z}^d:\|\mathbf{i}\|_\infty\leq N\\ \operatorname{supp}\mathbf{i}\subseteq\mathbf{j}}} a_{\mathbf{j},\mathbf{i}}\psi_{\mathbf{i}}(\mathbf{x}_{\mathbf{j}}) + r_{\mathbf{j}}(\mathbf{x}_j) \quad \text{for all} \quad \mathbf{j}\in\mathcal{S}_{r_0}. \tag{8.4}$$

Summing up over $\mathbf{j}\in\mathcal{S}_{r_0}$ we arrive at

$$f(\mathbf{x}) = \sum_{\mathbf{j}\in\mathcal{S}_{r_0}}\Bigg(\sum_{\substack{\mathbf{i}\in\mathbb{Z}^d:\|\mathbf{i}\|_\infty\leq N\\ \operatorname{supp}\mathbf{i}\subseteq\mathbf{j}}} a_{\mathbf{j},\mathbf{i}}\psi_{\mathbf{i}}(\mathbf{x}_{\mathbf{j}})\Bigg) + r(\mathbf{x}) \quad \text{with} \quad r(\mathbf{x}) = \sum_{\mathbf{j}\in\mathcal{S}_{r_0}} r_{\mathbf{j}}(\mathbf{x}_j). \tag{8.5}$$

The worst-case error of the uniform approximation of Hölder continuous functions with exponent $\alpha > 0$ (i.e., functions from the unit ball of C^α) is bounded from below by the Kolmogorov numbers of the embedding of C^α into L_∞, cf. [46],

$$\|r_{\mathbf{j}}(\mathbf{x}_{\mathbf{j}})\|_\infty \approx [(2N)^{r_0}]^{-\alpha/r_0} = (2N)^{-\alpha} \quad \text{and} \quad \|r(\mathbf{x})\|_\infty \approx |\mathcal{S}_{r_0}|(2N)^{-\alpha}.$$

Thus for any $\epsilon\in(0,1)$, we typically require $N \gtrsim \left(\frac{|\mathcal{S}_{r_0}|}{\epsilon}\right)^{1/\alpha}$ to ensure $\|r(\mathbf{x})\|_\infty \lesssim \epsilon$.

Furthermore, the number of degrees of freedom D in (8.5) is lower bounded by

$$D \geq \binom{d}{r_0}(2N)^{r_0}.$$

If the basis functions $\psi_{\mathbf{i}}$ are uniformly bounded (i.e. they form a Bounded Orthonormal System (BOS)), it was shown in [17, Theorem 12.31] or [36, Theorem 4.4], that one can recover $\mathbf{a} = (a_{\mathbf{j},\mathbf{i}})$ in (8.5) from $m \gtrsim |\mathcal{S}_{r_0}|(2N)^{r_0}\log^4(D)$ random samples of f by ℓ_1-minimization. Plugging in our estimates on D and N, we arrive at $m \gtrsim |\mathcal{S}_{r_0}|^{1+r_0/\alpha}\cdot\log^4(d)$. This bound is always superlinear in $|\mathcal{S}_{r_0}|$ and (if $\alpha\in(0,1]$) with the power of dependence at least $1+r_0$.

References

[1] F. Alizadeh and D. D. Goldfarb. Second-order cone programming. *Math. Program., Ser. B*, 95(1):3–51, 2003.

[2] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008.

[3] J. Bien, J. Taylor, and R. Tibshirani. A Lasso for hierarchical interactions. *Ann. Statist.*, 41(3):1111–1141, 2013.

[4] P. E. Blöchl. Generalized separable potentials for electronic-structure calculations. *Phys. Rev. B*, 41:5414–5416, 1990.

[5] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comp. Harm. Anal.*, 27(3):265 – 274, 2009.

[6] T. Blumensath and M. E. Davies. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE J. Selected Topics in Signal Proc.*, 4(2):298–309, 2010.

[7] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, 2008.

[8] E.J. Candès and J. Romberg. ℓ_1-magic: Recovery of sparse signals via convex programming. (2005), available at `http://www.acm.caltech.edu/l1magic/`.

[9] Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Trans. Inf. Theory*, 61(7):4034–4059, 2015.

[10] N. H. Choi, W. Li, and J. Zhu. Variable selection with the strong heredity constraint and its oracle property. *J. Amer. Statist. Assoc.*, 105(489):354–364, 2010.

[11] A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best n-term Galerkin approximations for a class of elliptic spdes. *Found. Comp. Math.*, 10(6):615–646, 2010.

[12] L. Comminges and A. S. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.*, 40(5):2667–2696, 2012.

[13] L. Comminges and A. S. Dalalyan. Tight conditions for consistent variable selection in high dimensional nonparametric regression. *J. Mach. Learn. Res.*, 19:187–206, 2012.

[14] A. Dalalyan, Y. Ingster, and A. B. Tsybakov. Statistical inference in compound functional models. *Probability Theory and Related Fields*, 158(3-4):513–532, 2014.

[15] R. DeVore, G. Petrova, and P. Wojtaszczyk. Approximation of functions of few variables in high dimensions. *Constr. Approx.*, 33:125–143, 2011.

[16] S.J. Dibangoye, C. Amato, O. Buffet, and F. Charpillet. Exploiting separability in multiagent planning with continuous-state mdps. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pages 1281–1288. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[17] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing.* Birkhäuser/Springer (New York), 2013.

[18] L. M. Ghiringhelli, J. Vybíral, S. V. Levchenko, C. Draxl, and M. Scheffler. Big data of materials science: Critical role of the descriptor. *Phys. Rev. Lett.*, 114(10):105503, 2015.

[19] G. Goel, I.-C. Chou, and E. O. Voit. System estimation from metabolic time-series data. *Bioinformatics*, 24(21):2505–2511, 2008.

[20] A. Griewank and P. L. Toint. On the unconstrained optimization of partially separable functions. In *Nonlinear Optimization 1981*, pages 301–312. Academic Press, 1982.

[21] D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Stat.*, 42(3):1079–1083, 1971.

[22] M. Holtz. *Sparse grid quadrature in high dimensions with applications in finance and insurance*, volume 77. Springer Science & Business Media, 2010.

[23] J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 2010.

[24] V. Kekatos and G.B. Giannakis. Sparse volterra and polynomial regression models: Recoverability and estimation. *Trans. Sig. Proc.*, 59(12):5907–5920, 2011.

[25] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *21st Annual Conference on Learning Theory (COLT)*, pages 229–238, 2008.

[26] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *Ann. Statist.*, 38(6):3660–3695, 2010.

[27] Y. Lin and H.H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, 34(5):2272–2297, 2006.

[28] L. Meier, S. Van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821, 2009.

[29] E. Mossel, R. O'Donnell, and R. Servedio. Learning juntas. In *35th Annual ACM Symposium on Theory of Computing (STOC)*, pages 206–212, 2003.

[30] B. Nazer and R. D. Nowak. Sparse interactions: Identifying high-dimensional multilinear systems via compressed sensing. In *48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1589–1596, 2010.

[31] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming.* Society for Industrial and Applied Mathematics, 1994.

[32] E. Novak and H. Triebel. Function spaces in lipschitz domains and optimal rates of convergence for sampling. *Constr. Approx.*, 23(3):325–350, 2006.

[33] E. Novak and H. Woźniakowski. Approximation of infinitely differentiable multivariate functions is intractable. *J. Compl.*, 25:398–404, 2009.

[34] P. Radchenko and G. M. James. Variable selection using adaptive nonlinear interaction structures in high dimensions. *J. Amer. Statist. Assoc.*, 105:1541–1553, 2010.

[35] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13(1):389–427, 2012.

[36] H. Rauhut. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9:1–92, 2010.

[37] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *J. Royal Statist. Soc.: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

[38] K. Schnass and J. Vybíral. Compressed learning of high-dimensional sparse functions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3924–3927, 2011.

[39] S. Shan and G. G. Wang. Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions. *Struct. Multidiscip. Optim.*, 41(2):219–241, 2010.

[40] C. B. Storlie, H. D. Bondell, B. J. Reich, and H. H. Zhang. Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica*, 21(2):679–705, 2011.

[41] H. Tyagi, A. Krause, and B. Gärtner. Efficient sampling for learning sparse additive models in high dimensions. In *Advances in Neural Information Processing Systems (NIPS) 27*, pages 514–522. 2014.

[42] H. Tyagi, A. Kyrillidis, B. Gärtner, and A. Krause. Learning sparse additive models with interactions in high dimensions. In *19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 111–120, 2016.

[43] H. Tyagi, A. Kyrillidis, B. Gärtner, and A. Krause. Algorithms for learning sparse additive models with interactions in high dimensions. *Information and Inference: A Journal of the IMA*, page iax008, 2017.

[44] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

[45] J. Vybíral. Sampling numbers and function spaces. *J. Compl.*, 23(4-6):773–792, 2007.

[46] J. Vybíral. Widths of embeddings in function spaces. *J. Compl.*, 24(4):545–570, 2008.

[47] M. Wahl. Variable selection in high-dimensional additive models based on norms of projections. ArXiv e-prints, arXiv:1406.0052, 2015.

[48] A. Winkelbauer. Moments and absolute moments of the normal distribution. ArXiv e-prints, arXiv:1209.4340.v2, 2014.

[49] Y. Yang and S. T. Tokdar. Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.*, 43(2):652–674, 2015.

[50] P. Zhu, J. Morelli, and S. Ferrari. Value function approximation for the control of multiscale dynamical systems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 5471–5477, 2016.

A Proofs for Section 2

Proof of Proposition 1. For $U \subseteq [d]$, we define

$$(P_U f)(\mathbf{x}_U) = f\Big(\sum_{j\in U} x_j \mathbf{e}_j\Big) \quad \text{and} \quad f_U(\mathbf{x}_U) = \sum_{V\subseteq U} (-1)^{|U|-|V|}(P_V f)(\mathbf{x}_V).$$

Here, $\{\mathbf{e}_1, \dots, \mathbf{e}_d\}$ is the canonical basis of $\mathbb{R}^d$.

By its definition, f_U is a continuous function. Furthermore,

$$\begin{aligned}\sum_{U\subseteq[d]} f_U(\mathbf{x}_U) &= \sum_{U\subseteq[d]}\sum_{V\subseteq U}(-1)^{|U|-|V|}(P_V f)(\mathbf{x}_V) = \sum_{V\subseteq[d]}(P_V f)(\mathbf{x}_V)\sum_{U\supseteq V}(-1)^{|U|-|V|}\\ &= \sum_{V\subseteq[d]}(P_V f)(\mathbf{x}_V)\sum_{W\subseteq[d]\setminus V}(-1)^{|W|}.\end{aligned}$$

The last sum can be rewritten as

$$\sum_{W\subseteq[d]\setminus V}(-1)^{|W|} = \sum_{k=0}^{d-|V|}(-1)^k\binom{d-|V|}{k},$$

which is equal to $(1-1)^{d-|V|} = 0$ for all $V \subseteq [d]$ except $V = [d]$. This leads to (2.4).

If $x_j = 0$ for some $j \in U$, we get

$$\begin{aligned}f_U(\mathbf{x}_U) &= \sum_{V\subseteq U}(-1)^{|U|-|V|}(P_V f)(\mathbf{x}_V)\\ &= \sum_{V\subseteq U\setminus\{j\}}\Big[(-1)^{|U|-|V|}(P_V f)(\mathbf{x}_V) + (-1)^{|U|-|V\cup\{j\}|}(P_{V\cup\{j\}}f)(\mathbf{x}_{V\cup\{j\}})\Big] = 0\end{aligned}$$

as all the terms in the last sum are equal to zero.

Finally, the uniqueness follows by induction. The statement is obvious for $d = 1$. Let now $d > 1$ and let us assume that a given function $f \in C([-1,1]^d)$ allows a decomposition

$$f(\mathbf{x}) = \sum_{U\subseteq[d]} f_U(\mathbf{x}_U), \quad \mathbf{x}\in[-1,1]^d,$$

which satisfies the properties a)-c) of Proposition 1.

Let W be a proper subset of $[d]$ and put for $\mathbf{x}_W \in [-1,1]^{|W|}$

$$g_W(\mathbf{x}_W) = f\Big(\sum_{j\in W} x_j\mathbf{e}_j\Big).$$

Then $g_W \in C([-1,1]^{|W|})$ and using c) of Proposition 1 we obtain

$$g_W(\mathbf{x}_W) = \sum_{U\subseteq[d]} f_U\Big(\Big(\sum_{j\in W}x_j\mathbf{e}_j\Big)_U\Big) = \sum_{U\subseteq[d]} f_U\Big(\sum_{j\in W}x_j(\mathbf{e}_j)_U\Big) = \sum_{U\subseteq W} f_U(\mathbf{x}_U).$$

This decomposition of g_W satisfies a)-c) of Proposition 1 with $|W| \le d-1$. By the induction assumption, this decomposition is therefore unique. We conclude, that f_U is uniquely determined for all $U \subset [d]$. Finally, also

$$f_{[d]}(\mathbf{x}) = f(\mathbf{x}) - \sum_{U\subset[d]} f_U(\mathbf{x}_U)$$

is uniquely determined. □

Proof of Proposition 2. Let f be given by (4.1) and let us assume that it satisfies all the assumptions of Proposition 2. We show that (4.1) coincides with its Anchored-ANOVA decomposition as described in Proposition 1 and (4.1) is therefore unique.

Let $U \subseteq [d]$ with $|U| \geq 3$. Then

$$\begin{aligned}
f_U(\mathbf{x}_U) &= \sum_{V \subseteq U} (-1)^{|U|-|V|} (P_V f)(\mathbf{x}_V) \\
&= \sum_{V \subseteq U} (-1)^{|U|-|V|} \Big\{ \mu + \sum_{j \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}} \phi_j((x_V)_j) + \sum_{\mathbf{j}=(j_1,j_2) \in \mathcal{S}_2} \phi_{\mathbf{j}}((x_V)_{j_1}, (x_V)_{j_2}) \Big\} \\
&= \mu \sum_{V \subseteq U} (-1)^{|U|-|V|} + \sum_{j \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}} \phi_j(x_j) \sum_{\substack{V \subseteq U \\ j \in V}} (-1)^{|U|-|V|} \\
&\quad + \sum_{j \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}} \phi_j(0) \sum_{\substack{V \subseteq U \\ j \notin V}} (-1)^{|U|-|V|} + \sum_{\mathbf{j}=(j_1,j_2) \in \mathcal{S}_2} \sum_{V \subseteq U} (-1)^{|U|-|V|} \phi_{\mathbf{j}}((x_V)_{j_1}, (x_V)_{j_2}).
\end{aligned}$$

It is easy to see, that the first three terms are zero. Finally, the last term can be split into a sum of four terms depending on if $j_1 \in V$ or $j_2 \in V$, i.e. terms of the kind

$$\sum_{\mathbf{j}=(j_1,j_2) \in \mathcal{S}_2} \phi_{\mathbf{j}}(x_{j_1}, x_{j_2}) \sum_{\substack{V \subseteq U \\ j_1, j_2 \in V}} (-1)^{|U|-|V|},$$

which also vanish. Therefore, $f_U(\mathbf{x}_U) = 0$.

If $U = \{j_1, j_2\}$ with $1 \leq j_1 < j_2 \leq d$, then $\emptyset, \{j_1\}, \{j_2\}$ and $\{j_1, j_2\}$ are the only subsets of U and we obtain

$$\begin{aligned}
f_U(\mathbf{x}_U) &= \sum_{V \subseteq U} (-1)^{|U|-|V|} (P_V f)(\mathbf{x}_V) \\
&= (P_\emptyset f)(0) - (P_{j_1} f)(x_{j_1}) - (P_{j_2} f)(x_{j_2}) + (P_{\{j_1,j_2\}} f)(x_{j_1}, x_{j_2}) \\
&= f(0) - f(x_{j_1}\mathbf{e}_{j_1}) - f(x_{j_2}\mathbf{e}_{j_2}) + f(x_{j_1}\mathbf{e}_{j_1} + x_{j_2}\mathbf{e}_{j_2}) \\
&= \sum_{j \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}} [\phi_j(0) - \phi_j((x_{j_1}\mathbf{e}_{j_1})_j) - \phi_j((x_{j_2}\mathbf{e}_{j_2})_j) + \phi_j((x_{j_1}\mathbf{e}_{j_1} + x_{j_2}\mathbf{e}_{j_2})_j)] \\
&\quad + \sum_{\mathbf{j} \in \mathcal{S}_2} [\phi_{\mathbf{j}}(0) - \phi_{\mathbf{j}}((x_{j_1}\mathbf{e}_{j_1})_{\mathbf{j}}) - \phi_{\mathbf{j}}((x_{j_2}\mathbf{e}_{j_2})_{\mathbf{j}}) + \phi_{\mathbf{j}}((x_{j_1}\mathbf{e}_{j_1} + x_{j_2}\mathbf{e}_{j_2})_{\mathbf{j}})].
\end{aligned}$$

The first sum vanishes for all $j \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$, which can be easily observed by considering the options $j \notin \{j_1, j_2\}, j = j_1$, or $j = j_2$. If $(j_1, j_2) \notin \mathcal{S}_2$, then also the second sum vanishes. If $(j_1, j_2) \in \mathcal{S}_2$, then

$$f_{j_1,j_2}(x_{j_1}, x_{j_2}) = \phi_{j_1,j_2}(x_{j_1}, x_{j_2}) - \phi_{j_1,j_2}(x_{j_1}, 0) - \phi_{j_1,j_2}(0, x_{j_2}) + \phi_{j_1,j_2}(0,0) = \phi_{j_1,j_2}(x_{j_1}, x_{j_2}).$$

Finally, if $l \in [d]$ then $\emptyset$ and $\{l\}$ are the only subsets of $\{l\}$. If furthermore $l \notin \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$, we get

$$\begin{aligned}
f_{\{l\}}(x_l) &= -f(0) + f(x_l \mathbf{e}_l) \\
&= -\Big(\mu + \sum_{j \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}} \phi_j(0) + \sum_{\mathbf{j} \in \mathcal{S}_2} \phi_{\mathbf{j}}(0)\Big) + \Big(\mu + \sum_{j \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}} \phi_j(0) + \sum_{\mathbf{j} \in \mathcal{S}_2} \phi_{\mathbf{j}}(0)\Big) = 0.
\end{aligned}$$

Similarly, $f_{\{l\}}(x_l) = \phi_l(x_l)$ if $l \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$. □

B Some standard concentration results

First, we recall that the sub-Gaussian norm of a random variable X is defined as

$$\|X\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}$$

and X is called sub-Gaussian random variable if $\|X\|_{\psi_2}$ is finite. Similarly, the sub-exponential norm of a random variable is the quantity

$$\|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}$$

and X is called sub-exponential if $\|X\|_{\psi_1}$ is finite. Next, we recall the following concentration results for sums of i.i.d. sub-Gaussian and sub-exponential random variables.

Proposition 4. *[44, Proposition 5.16] Let $X_1, \ldots, X_N$ be independent centered sub-exponential random variables, and $K = \max_i \|X_i\|_{\psi_1}$. Then for every $\mathbf{a} = (a_1, \ldots, a_N) \in \mathbb{R}^N$ and every $t \geq 0$, we have*

$$\mathbb{P}\left(\Big|\sum_i a_i X_i\Big| \geq t\right) \leq 2\exp\left[-c\min\left\{\frac{t^2}{K^2\|\mathbf{a}\|_2^2}, \frac{t}{K\|\mathbf{a}\|_\infty}\right\}\right],$$

where $c > 0$ is an absolute constant.

Proposition 5. *[44, Proposition 5.10] Let $X_1, \ldots, X_N$ be independent centered sub-Gaussian random variables, and $K = \max_i \|X_i\|_{\psi_2}$. Then for every $\mathbf{a} = (a_1, \ldots, a_N) \in \mathbb{R}^N$ and every $t \geq 0$, we have*

$$\mathbb{P}\left(\Big|\sum_i a_i X_i\Big| \geq t\right) \leq e \cdot \exp\left[-\frac{ct^2}{K^2\|\mathbf{a}\|_2^2}\right],$$

where $c > 0$ is an absolute constant.

Let $\boldsymbol{\eta} = (\eta_1 \ \eta_2 \ \ldots \ \eta_n)^T$, where $\eta_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. for each i. The Proposition below is a standard concentration result, stating that $\|\boldsymbol{\eta}\|_2 = \Theta(\sigma\sqrt{n})$ and $\|\boldsymbol{\eta}\|_1 = \Theta(\sigma n)$, with high probability. We provide proofs for completeness.

Proposition 6. *Let $\boldsymbol{\eta} = (\eta_1 \ \eta_2 \ \ldots \ \eta_n)^T$ where $\eta_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d for each i. Then, there exists constants $c_1, c_2 > 0$ so that for any $\epsilon \in (0,1)$, we have:*

1. $\mathbb{P}\left(\|\boldsymbol{\eta}\|_2 \in [(1-\epsilon)\sigma\sqrt{n}, (1+\epsilon)\sigma\sqrt{n}\ ]\right) \geq 1 - 2\exp\left(-c_1\epsilon^2 n\right)$*, and*

2. $\mathbb{P}\left(\|\boldsymbol{\eta}\|_1 \in \left[(1-\epsilon)n\sigma\sqrt{\frac{2}{\pi}}, (1+\epsilon)n\sigma\sqrt{\frac{2}{\pi}}\right]\right) \geq 1 - e \cdot \exp\left(-\frac{2c_2\epsilon^2 n}{\pi}\right)$.

Proof. 1. Note that η_i are i.i.d. sub-Gaussian random variables with $\|\eta_i\|_{\psi_2} \leq C_1\sigma$. Hence η_i^2 are i.i.d. sub-exponential[4] with

$$\|\eta_i^2\|_{\psi_1} \leq 2\|\eta_i\|_{\psi_2}^2 \leq C_2\sigma^2$$

for some constant $C_2 > 0$. This implies that $\eta_i^2 - \mathbb{E}[\eta_i^2]$ are i.i.d. sub-exponential with

$$\|\eta_i^2 - \mathbb{E}[\eta_i^2]\|_{\psi_1} \leq 2\|\eta_i^2\|_{\psi_1} \leq C_3\sigma^2$$

[4] A random variable X is sub-Gaussian iff X^2 is sub-exponential. Moreover, $\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2$ (cf. [44, Lemma 5.14]).

for some constant $C_3 > 0$. Using Proposition 4 with $t = n\epsilon\sigma^2$ for $0 < \epsilon < 1$, we obtain for some constant $c_1 > 0$

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(\eta_i^2 - \mathbb{E}[\eta_i^2])\right| \leq n\epsilon\sigma^2\right) \geq 1 - 2\exp\left[-c_1 \min\left\{\epsilon^2, \epsilon\right\} n\right].$$

Together with some standard manipulations this completes the proof.

2. Note that $|\eta_i| - \mathbb{E}[|\eta_i|]$ is sub-Gaussian with

$$\||\eta_i| - \mathbb{E}[|\eta_i|]\|_{\psi_2} \leq 2\|\eta_i\|_{\psi_2} \leq C\sigma,$$

for some constant $C > 0$. Using Proposition 5 with $t = n\epsilon\mathbb{E}[|\eta_1|]$ for $\epsilon > 0$, we hence obtain

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(|\eta_i| - \mathbb{E}[|\eta_i|])\right| \leq n\epsilon\mathbb{E}[|\eta_1|]\right) \geq 1 - e \cdot \exp\left[-\frac{c_2 n\epsilon^2(\mathbb{E}[|\eta_1|])^2}{\sigma^2}\right]$$

for some constant $c_2 > 0$. Finally, observing that $\mathbb{E}[|\eta_1|] = \sigma\sqrt{\frac{2}{\pi}}$ (cf., [48]) completes the proof.

□

C Proof of Theorem 2

The proof of part (1) follows in the same manner as [9, Proposition 1], with minor differences in calculation at certain parts. For completeness, we outline the main steps below.

Invoking the Hanson-Wright inequality for quadratic forms [21], we get for some constant $c > 0$ and all $t > 0$

$$\mathbb{P}(|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta} - \mathbb{E}[\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}]| > t) \leq 2\exp\left[-c\min\left\{\frac{t^2}{K^4\|\mathbf{A}\|_F^2}, \frac{t}{K^2\|\mathbf{A}\|}\right\}\right], \tag{C.1}$$

where $\|\beta_i\|_{\psi_2} \leq K$. For a Rademacher random variable β_i, $K = 1$ and

$$\mathbb{E}[\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}] = \mathbb{E}\sum_{i\neq j}\beta_i\beta_j A_{ij} = 0.$$

Using $\|\mathbf{A}\| \leq \|\mathbf{A}\|_F$, (C.1) implies for every $t > 0$

$$\mathbb{P}(|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}| > t) \leq 2\exp\left[-c\min\left\{\frac{t^2}{\|\mathbf{A}\|_F^2}, \frac{t}{\|\mathbf{A}\|_F}\right\}\right]. \tag{C.2}$$

We will now find upper and lower bounds on $\mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|]$. The upper bound is easy since (C.2) implies

$$\mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|] = \int_0^\infty \mathbb{P}(|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}| > t)dt \leq c'\|\mathbf{A}\|_F. \tag{C.3}$$

In order to find the lower bound, we have via repeated application of Cauchy Schwartz inequality

$$\left(\mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|^2]\right)^2 \leq \mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|] \cdot \mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|^3] \leq \mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|] \cdot \left(\mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|^2]\right)^{1/2} \cdot \left(\mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|^4]\right)^{1/2}$$

and

$$\mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|] \geq \sqrt{\frac{(\mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|^2])^3}{\mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|^4]}}.$$

Since $\boldsymbol{\beta}$ consists of i.i.d. Rademacher random variables, we obtain

$$\mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|^2] = 2\|\mathbf{A}\|_F^2 = \|\mathbf{a}\|_2^2.$$

Moreover, an argument similar to (C.3) gives $\mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|^4] \leq c''\|\mathbf{A}\|_F^4$. Hence,

$$\mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|] \geq \sqrt{\frac{8\|\mathbf{A}\|_F^6}{c''\|\mathbf{A}\|_F^4}} = \tilde{c}\|\mathbf{A}\|_F. \tag{C.4}$$

Eqs. (C.3), (C.4) give us upper and lower bounds for $\mathbb{E}[|\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}|]$. As a last step, we consider the zero mean random variables $X_1, \ldots, X_n$, where $X_i = |\boldsymbol{\beta}_i^T\mathbf{A}\boldsymbol{\beta}_i| - \mathbb{E}|\boldsymbol{\beta}_i^T\mathbf{A}\boldsymbol{\beta}_i|$. A simple modification of (C.3) shows that they are sub-exponential with $\|X_i\|_{\psi_1} \leq c\|\mathbf{A}\|_F$. We can therefore apply a standard concentration bound (see [44, Proposition 5.16]) to bound the deviation

$$\left|\frac{1}{n}\|\mathbf{B}\mathbf{a}\|_1 - \frac{1}{n}\mathbb{E}[\|\mathbf{B}\mathbf{a}\|_1]\right| = \frac{1}{n}\left|\sum_{i=1}^{n} X_i\right|.$$

A straightforward calculation then yields the statement of part (1) of the Theorem.

Part (2) follows from standard arguments based on ϵ-nets, detailed for instance in [2].

The proof of part (3) copies that of [9, Theorem 3], which again is inspired by [7]. We sketch the main steps below for completeness. Denoting $\widehat{\mathbf{a}} = \mathbf{a} + \mathbf{h}$, we have by feasibility of $\mathbf{a}$ that $\frac{1}{n}\|\mathbf{B}\mathbf{h}\|_1 \leq \frac{2\nu}{n}$. Denoting Ω_0 to be set of indices corresponding to the k largest entries of $\mathbf{a}$, we get $\mathbf{a} = \mathbf{a}_{\Omega_0} + \mathbf{a}_{\Omega_0^c}$. For a suitable positive integer K, we define Ω_1 as the set of indices of K largest entries of $\mathbf{h}_{\Omega_0^c}$, Ω_2 as the set of indices of K largest entries of $\mathbf{h}$ on $(\Omega_0 \cup \Omega_1)^c$ and so on. Following this argument of [7] gives the proof.

The proof of part (4) follows by choosing $K = 4\big(\frac{4c_2}{c_1}\big)^2 k$. Indeed, (6.9) gives

$$\frac{1-\gamma_{k+K}^{\text{lb}}}{\sqrt{2}} - (1+\gamma_K^{\text{ub}})\sqrt{\frac{k}{K}} \geq \frac{c_1}{2\sqrt{2}} - 2c_2 \cdot \frac{c_1}{8c_2} = \frac{(\sqrt{2}-1)c_1}{4} = \beta > 0$$

if $n > c_3'(k+K)\log(d^2/(k+K))$, with probability at least $1 - e^{-C_4 n}$ for some constant $C_4 > 0$.