Algorithms for Learning Sparse Additive Models with Interactions in High Dimensions*

Hemant Tyagi† htyagi@turing.ac.uk
Anastasios Kyrillidis‡ anastasios@utexas.edu
Bernd Gärtner§ gaertner@inf.ethz.ch
Andreas Krause¶ krausea@ethz.ch

May 9, 2017

Abstract

A function $f : \mathbb{R}^d \to \mathbb{R}$ is a Sparse Additive Model (SPAM), if it is of the form $f(\mathbf{x}) = \sum_{l \in \mathcal{S}} \phi_l(x_l)$ where $\mathcal{S} \subset [d]$, $|\mathcal{S}| \ll d$. Assuming ϕ's, $\mathcal{S}$ to be unknown, there exists extensive work for estimating f from its samples. In this work, we consider a generalized version of SPAMs, that also allows for the presence of a sparse number of *second order interaction terms*. For some $\mathcal{S}_1 \subset [d], \mathcal{S}_2 \subset \binom{[d]}{2}$, with $|\mathcal{S}_1| \ll d, |\mathcal{S}_2| \ll d^2$, the function f is now assumed to be of the form: $\sum_{p \in \mathcal{S}_1} \phi_p(x_p) + \sum_{(l,l') \in \mathcal{S}_2} \phi_{(l,l')}(x_l, x_{l'})$. Assuming we have the freedom to query f anywhere in its domain, we derive efficient algorithms that provably recover $\mathcal{S}_1, \mathcal{S}_2$ with *finite sample bounds*. Our analysis covers the noiseless setting where exact samples of f are obtained, and also extends to the noisy setting where the queries are corrupted with noise. For the noisy setting in particular, we consider two noise models namely: i.i.d Gaussian noise and arbitrary but bounded noise. Our main methods for identification of $\mathcal{S}_2$ essentially rely on estimation of sparse Hessian matrices, for which we provide two novel compressed sensing based schemes. Once $\mathcal{S}_1, \mathcal{S}_2$ are known, we show how the individual components ϕ_p, $\phi_{(l,l')}$ can be estimated via additional queries of f, with uniform error bounds. Lastly, we provide simulation results on synthetic data that validate our theoretical findings.

1 Introduction

Many scientific problems involve estimating an unknown function f, defined over a compact subset of $\mathbb{R}^d$, with d large. Such problems arise for instance, in modeling complex physical processes [35, 32, 58]. Information about f is typically available in the form of point values $(x_i, f(x_i))_{i=1}^n$, which are then used for learning f. It is well known that the problem suffers from the curse of dimensionality, if only smoothness assumptions are placed on f. For example, if f is C^s smooth (s times continuously differentiable), then for uniformly approximating f within error $\delta \in (0,1)$, one needs $n = \Omega(\delta^{-d/s})$ samples [51].

A popular line of work in recent times, considers the setting where f possesses an intrinsic low dimensional structure, *i.e.*, depends on only a small subset of d variables. There exist algorithms for estimating such f – tailored to the underlying structural assumption – along with attractive theoretical guarantees, that do not suffer from the curse of dimensionality (cf., [15, 9, 53, 18]). One such assumption leads to the class of sparse additive models (SPAMs) wherein $f = \sum_{l \in \mathcal{S}} \phi_l$ for some unknown $S \subset \{1, \ldots, d\}$ with $|\mathcal{S}| = k \ll d$. There exist several algorithms for learning these models (cf. [45, 33, 23, 43, 54]). Here we focus on a generalized SPAM model, where f can also contain a small number of *second order interaction terms*, *i.e.*,

$$f(x_1, x_2, \ldots, x_d) = \sum_{p \in \mathcal{S}_1} \phi_p(x_p) + \sum_{(l,l') \in \mathcal{S}_2} \phi_{(l,l')}(x_l, x_{l'}); \quad \mathcal{S}_1 \subset [d], \mathcal{S}_2 \subset \binom{[d]}{2}, \tag{1.1}$$

with $|\mathcal{S}_1| \ll d, |\mathcal{S}_2| \ll d^2$. Here, $\phi_{(l,l')}(x_l, x_{l'}) \not\equiv g_l(x_l) + h_{l'}(x_{l'})$ for some univariates $g_l, h_{l'}$ meaning that $\frac{\partial^2}{\partial_l \partial_{l'}} \phi_{(l,l')} \not\equiv 0$. As opposed to SPAMs, the problem is significantly harder now – allowing interactions leads to an additional $d(d-1)/2$ unknowns out of which of only a few terms (*i.e.*, those in $\mathcal{S}_2$) are relevant. In the sequel, we will denote $\mathcal{S}$ to be the support of f consisting of variables that are part of $\mathcal{S}_1$ or $\mathcal{S}_2$, and k to be the size of $\mathcal{S}$. Moreover, we will denote ρ_m to be the maximum number of occurrences of a variable in $\mathcal{S}_2$ – this parameter captures the underlying *complexity* of the interactions.

There exist relatively few results for learning models of the form (1.1), with the existing work being mostly in the *regression framework* in statistics (cf., [31, 42, 49]). Here, $(x_i, f(x_i))_{i=1}^n$ are typically samples from an unknown probability measure $\mathbb{P}$,

*A preliminary version of this paper appeared in the proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016 [55]. The present draft is an expanded version containing additional results.

†School of Mathematics, University of Edinburgh, Edinburgh, United Kingdom; The Alan Turing Institute, London, United Kingdom
‡Department of Electrical and Computer Engineering, The University of Texas at Austin
§Department of Computer Science, Institute of Theoretical Computer Science, ETH Zürich, CH-8092 Zürich
¶Department of Computer Science, ETH Zürich, CH-8092 Zürich

with the samples moreover assumed to be corrupted with (i.i.d) stochastic noise. In this paper, we consider the *approximation theoretic* setting where we have the freedom to query f at any desired set of points (cf. [15, 18, 54]). We propose strategies for querying f, along with efficient recovery algorithms, which leads to much stronger guarantees than known in the regression setting. In particular, we provide the first *finite sample bounds* for exactly recovering $\mathcal{S}_1$ and $\mathcal{S}_2$. This is shown for (i) the noiseless setting where exact samples are observed, as well as (ii) the noisy setting, where the samples are corrupted with noise (either i.i.d Gaussian or arbitrary but bounded noise models).

Once $\mathcal{S}_1$, $\mathcal{S}_2$ are identified, we show in Section 6 how the individual components: $\phi_p, \phi_{(l,l')}$ of the model can be estimated, with *uniform* error bounds. This is shown for both the noiseless and noisy query settings. It is accomplished by additionally sampling f along the identified one/two dimensional subspaces corresponding to $\mathcal{S}_1, \mathcal{S}_2$ respectively, and by employing standard estimators from approximation theory and statistics.

1.1 Our contributions

We make the following contributions for learning models of the form (1.1).

1. Firstly, we provide an efficient algorithm, namely Algorithm 3, which provably recovers $\mathcal{S}_1, \mathcal{S}_2$ exactly with high probability[1] (w.h.p), with $O(k\rho_m(\log d)^3)$ noiseless queries. When the point queries are corrupted with (i.i.d) Gaussian noise, we show that Algorithm 3 identifies $\mathcal{S}_1$, $\mathcal{S}_2$ w.h.p, with $O(\rho_m^5 k^2(\log d)^4)$ noisy queries of f. We also analyze the setting of arbitrary but bounded noise, and derive sufficient conditions on the noise magnitude that enable recovery of $\mathcal{S}_1, \mathcal{S}_2$.

2. Secondly, we provide another efficient algorithm namely Algorithm 4, which provably recovers $\mathcal{S}_1, \mathcal{S}_2$ exactly w.h.p, with (i) $O(k\rho_m(\log d)^2)$ noiseless queries and, (ii) $O(\rho_m^5 k^5(\log d)^3)$ noisy queries (i.i.d Gaussian noise). We also analyze the setting of arbitrary but bounded noise.

3. We provide an algorithm tailored to the special case where the underlying interaction graph corresponding to $\mathcal{S}_2$ is known to be a *perfect matching*, *i.e.*, each variable interacts with at most one variable (so $\rho_m = 1$). We show that the algorithm identifies $\mathcal{S}_1, \mathcal{S}_2$ w.h.p, with (i) $O(k(\log d)^2)$ noiseless queries and, (ii) $O(k^2(\log d)^3)$ noisy queries (i.i.d Gaussian noise). We also analyze the setting of arbitrary but bounded noise.

4. An important part of Algorithms 3, 4 are two novel compressive sensing based methods, for estimating *sparse*, $d \times d$ Hessian matrices. These might be of independent interest.

We also provide simulation results on synthetic data, that validate our theoretical findings concerning the identification of $\mathcal{S}_1, \mathcal{S}_2$. Algorithm 3 appeared in AISTATS 2016 [55], in a preliminary version of this paper. The results in Section 6 (estimating individual components of f) were part of the supplementary material in [55].

1.2 Related work

We now provide a brief overview of related work, followed by an outline of our main contributions and an overview of the methods. A more detailed comparison with related work is provided in Section 8.

Learning SPAMs. This model was introduced in the nonparametric regression setting by Lin et al. [31] who proposed the COSSO (Component selection and smoothing) method – an extension of the lasso to the reproducing kernel Hilbert space (RKHS) setting. It essentially performs least squares minimization with a sparsity inducing penalty term involving the sum of norms of the function components. In fact, this method is designed to handle the more general smoothing spline analysis of variance (SS-ANOVA) model [56, 21]. It has since been studied extensively in the regression framework with a multitude of results involving: estimation of f (cf.,[25, 33, 45, 43, 26, 23]) and/or variable selection, *i.e.*, identifying the support $\mathcal{S}$ (cf., [23, 45, 57]).

A common theme behind (nearly all of) these approaches is to first (approximately) represent each ϕ_j; $1 \leq j \leq d$, in a suitable basis of finite size. This is done for example via: B-splines (cf. [23, 33]), finite combination of kernel functions (cf. [43, 26]) etc. Thereafter, the problem reduces to a finite dimensional one, that involves finding the values of the coefficients in the corresponding basis representation. This is accomplished by performing least squares minimization subject to sparsity and smoothness inducing penalty terms – the optimization problem is convex on account of the choice of the penalty terms, and hence can be solved efficiently.

With regards to the problem of estimating f, Koltchinskii et al. [26], Raskutti et al. [43] proposed a convex program for estimating f in the RKHS setting along with L_2 error rates. These error rates were shown to be minimax optimal by Raskutti et al. [43]. For example, f lying in a Sobolev space with smoothness parameter $\alpha > 1/2$, are estimated at the optimal L_2 rate: $\frac{k \log d}{n} + kn^{-\frac{2\alpha}{2\alpha+1}}$ where n denotes the number of samples. There also exist results for the *variable selection* problem, *i.e.*, for estimating the support $\mathcal{S}$. In contrast to the setting of sparse linear models, for which non-asymptotic sample complexity bounds are known [59, 58], the corresponding results in the nonparametric setting are usually *asymptotic*, *i.e.*, derived in the limit of large n. This property is referred to as *sparsistency* in the statistics literature; an estimator is called *sparsistent* if $\widehat{\mathcal{S}} = \mathcal{S}$ with

[1] With probability $1 - O(d^{-c})$ for some constant $c > 0$.

probability approaching one as $n \to \infty$. Variable selection results for SPAMs in the nonparametric regression setting can be found for instance in [45, 23, 57]. Recently, Tyagi et al. [54] considered this problem in the approximation theoretic setting; they proposed a method that identifies $\mathcal{S}$ w.h.p with sample complexities $O(k \log d)$, $O(k^3(\log d)^2)$ in the absence/presence of Gaussian noise, respectively.

While there exists a significant amount of work in the literature for SPAMs, the aforementioned methods are designed for specifically learning SPAMs, and cannot handle generalized SPAMs of the form (1.1) containing interaction terms.

Learning generalized SPAMs. There exist fewer results for generalized SPAMs of the form (1.1), in the regression setting. The COSSO algorithm [31] can handle (1.1), however its convergence rates are shown only for the case of no interactions. Radchenko et al. [42] proposed the VANISH algorithm – a least squares method with sparsity constraints and show that their method is sparsistent. Storlie et al. [49] proposed ACOSSO – an adaptive version of the COSSO algorithm – which can also handle (1.1). They derived convergence rates and sparsistency results for their method, albeit for the case of no interactions. Recently, Dalalayan et al. [13], Yang et al. [61] studied a generalization of (1.1) that allows for the presence of a sparse number of m-wise interaction terms for some additional sparsity parameter m. While they derive non-asymptotic L_2 error rates for estimating f in such generic setting, they do not guarantee unique identification of the interaction terms for any value of m.

A special case of (1.1) – where ϕ_p's are linear and each $\phi_{(l,l')}$ is of the form $x_l x_{l'}$ – has been studied considerably. Within this setting, there exist algorithms that recover $\mathcal{S}_1, \mathcal{S}_2$, along with convergence rates for estimating f in the limit of large n [8, 42, 3]. There also exist non-asymptotic sampling bounds for identifying the interaction terms in the noiseless setting (cf., [37, 24]). However finite sample bounds for the non-linear model (1.1) are not known in general.

Other low-dimensional function models. There exist results for other, more general classes of intrinsically low dimensional functions, that we now mention starting with the approximation theoretic setting. Devore et al. [15] consider functions depending on an unknown subset $\mathcal{S}$ of the variables with $|\mathcal{S}| = k \ll d$. The functions do not necessarily possess an additive structure, so the function class is more general than (1.1). They provide algorithms that recover $\mathcal{S}$ exactly w.h.p, with $O(c^k k \log d)$ noiseless queries of f, for some constant $c > 0$. Schnass et al. [46] derived a simpler algorithm for this problem in the noiseless setting. This function class was also studied by Comminges et al. [12, 11] in the nonparametric regression setting wherein they analyzed an estimator that identifies $\mathcal{S}$ w.h.p, with $O(c^k k \log d)$ samples of f. Fornasier et al. [18], Tyagi et al. [53] considered a generalization of the above function class where f is now of the form $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$, for unknown $\mathbf{A} \in \mathbb{R}^{k \times d}$. They derived algorithms that approximately recover the row-span of $\mathbf{A}$, with sample complexities typically polynomial in k, d. While the above methods could possibly recover the underlying support $\mathcal{S}$ for the SPAM model (1.1), their sample complexities are either exponential in k [15, 12, 11] or polynomial in d [18, 53]. As explained in Section 8, the algorithm of Schnass et al. [46] would recover $\mathcal{S}$ w.h.p, with $O(\rho_m^4 k (\log d)^2)$ noiseless queries, with potentially large constants (depending on smoothness of f) within the $O(\cdot)$ term. Moreover, we note that the aforementioned methods are not designed for identifying *interactions* among the variables.

1.3 Overview of methods used

We now describe the main underlying ideas behind the algorithms described in this paper, for identifying $\mathcal{S}_1, \mathcal{S}_2$. On a top level, our methods are based on two simple observations for the model (1.1), namely that for any $\mathbf{x} \in \mathbb{R}^d$:

- The gradient $\nabla f(\mathbf{x}) \in \mathbb{R}^d$ is k sparse.
- The Hessian $\nabla^2 f(\mathbf{x}) \in \mathbb{R}^{d \times d}$ is at most $k(\rho_m + 1)$ sparse. In particular, it has k non zero rows, with each such row having at most $\rho_m + 1$ non zero entries.

For the special case of *no overlap*, *i.e.*, $\rho_m = 1$, we proceed in two phases. In the first phase – outlined as Algorithm 1 – we identify all variables in $\mathcal{S}$ by estimating $\nabla f(\mathbf{x})$ via ℓ_1 minimization[2], for each $\mathbf{x}$ lying within a carefully constructed finite set $\chi \in \mathbb{R}^d$. The set χ in particular is constructed[3] so that it provides a uniform discretization of all possible two dimensional canonical subspaces in $\mathbb{R}^d$. In the second phase – outlined as Algorithm 2 – we identify the sets $\mathcal{S}_1, \mathcal{S}_2$ via a simple (deterministic) binary search based procedure, over the rows of the corresponding $k \times k$ sub-matrix of the Hessian of f.

For the general case however where $\rho_m \geq 1$, the above scheme does not guarantee identification of $\mathcal{S}$; see discussion at beginning of Section 4.1. Therefore now, we consider a different "two phase" approach where in the first phase, we query f with the goal of identifying the set of interactions $\mathcal{S}_2$. This in fact entails estimating the sparse Hessian $\nabla^2 f(\mathbf{x})$, at each $\mathbf{x}$ lying within χ. We propose two different methods for estimating $\nabla^2 f(\mathbf{x})$, utilizing tools from compressive sensing (CS).

- The first method is a part of Algorithm 3 where we estimate each row of $\nabla^2 f(\mathbf{x})$ separately, via a "difference of gradients" approach. This is motivated by the following identity, based on the Taylor expansion of ∇f at $\mathbf{x}$, for suitable $\mathbf{v}' \in \mathbb{R}^d$, $\mu_1 > 0$:

$$\frac{\nabla f(\mathbf{x} + \mu_1 \mathbf{v}') - \nabla f(\mathbf{x})}{\mu_1} = \nabla^2 f(\mathbf{x}) \mathbf{v}' + O(\mu_1). \tag{1.2}$$

[2] We note that the idea of estimating a sparse gradient via ℓ_1 minimization is motivated from Fornasier et al. [18]; their algorithm however is for a more general function class than ours.

[3] see Definition 1 and ensuing discussion.

We can see from (1.2), that a difference of gradient vectors corresponds to obtaining a perturbed linear measurement of *each* $\rho_m + 1$ sparse row of $\nabla^2 f(\mathbf{x})$. CS theory tells us that by collecting $O(\rho_m \log d)$ such "gradient differences" – each difference term corresponding to a random choice of $\mathbf{v}'$ from a suitable distribution – we can estimate each row of $\nabla^2 f(\mathbf{x})$ via ℓ_1 minimization. Since ∇f is k sparse, it can also be estimated via $O(k \log d)$ queries of f – this leads to obtaining an estimate of $\nabla^2 f(\mathbf{x})$ with $O(k\rho_m (\log d)^2)$ queries of f in total.

- The second method is a part of Algorithm 4 where we estimate all entries of $\nabla^2 f(\mathbf{x})$ in "one go". This is motivated by the following identity, based on the Taylor expansion of f at $\mathbf{x}$, for suitable $\mathbf{v} \in \mathbb{R}^d$, $\mu > 0$:

$$\frac{f(\mathbf{x} + 2\mu\mathbf{v}) + f(\mathbf{x} - 2\mu\mathbf{v}) - 2f(\mathbf{x})}{4\mu^2} = \langle \mathbf{v}\mathbf{v}^T, \nabla^2 f(\mathbf{x}) \rangle + O(\mu). \tag{1.3}$$

We see from (1.3) that the L.H.S corresponds to a perturbed linear measurement of the Hessian, with a rank one matrix. By leveraging recent results in CS – most notably the work of Chen et al. [7] – we recover an estimate of $\nabla^2 f(\mathbf{x})$ through ℓ_1 minimization, by choosing $\mathbf{v}$'s randomly from a suitable distribution. As described in detail in Section 5, this requires us to make $O(k\rho_m \log d)$ queries of f.

Once $\mathcal{S}_2$ is estimated, we estimate $\mathcal{S}_1$ by invoking (a slightly improved version of) the method of Tyagi et al. [54] for learning SPAMs, on the reduced variables set.

Outline of the paper. The rest of the paper is organized as follows. Section 2 contains a formal description of the problem along with notation used. We begin by analyzing the special case of *no overlap* between the elements of $\mathcal{S}_2$ (*i.e.*, $\rho_m = 1$), in Section 3. Section 4 then considers the general setting where $\rho_m \geq 1$. In particular, it describes Algorithm 3 wherein the underlying sparse Hessian of f is estimated via a difference of sparse gradients mechanism. Section 5 also handles the general overlap setting, albeit with a different method for estimating the sparse Hessian of f. Once $\mathcal{S}_1, \mathcal{S}_2$ are estimated, we describe how the individual components of f can be estimated via standard tools from approximation theory and statistics, in Section 6. Section 7 contains simulation results on synthetic examples. We provide a detailed discussion of related work in Section 8, and conclude with directions for future work in Section 9. All proofs are deferred to the appendix.

2 Notation and problem setup

Notation. Scalars are mostly denoted by plain letters (e.g. k_1, k_2, d), vectors by lowercase boldface letters (e.g., $\mathbf{x}$) or by lowercase Greek letters (*e.g.*, ζ), matrices by uppercase boldface letters (e.g. $\mathbf{A}$) and sets by uppercase calligraphic letters (e.g. $\mathcal{S}$), with the exception of $[d]$ which denotes the index set $\{1, \ldots, d\}$. Given a set $\mathcal{S} \subseteq [d]$, we denote its complement by $\mathcal{S}^c := [d] \setminus \mathcal{S}$ and for vector $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, $(\mathbf{x})_{\mathcal{S}}$ denotes the restriction of $\mathbf{x}$ onto $\mathcal{S}$, *i.e.*, $((\mathbf{x})_{\mathcal{S}})_l = x_l$ if $l \in \mathcal{S}$ and 0 otherwise. We use $|\mathcal{S}|$ to denote the cardinality of a set $\mathcal{S}$. The ℓ_p norm of a vector $\mathbf{x} \in \mathbb{R}^d$ is defined as $\| \mathbf{x} \|_p := \left(\sum_{l=1}^d |x_i|^p \right)^{1/p}$. Let g be a function of n variables, $g(x_1, \ldots, x_n)$. $\mathbb{E}_p[g]$, $\mathbb{E}_{(l,l')}[g]$ denote expectation w.r.t uniform distributions over x_p and $(x_l, x_{l'})$ respectively. $\mathbb{E}[g]$ denotes expectation w.r.t. uniform distribution over $(x_1, \ldots, x_n)$. For any compact $\Omega \subset \mathbb{R}^n$, we denote by $\| g \|_{L_\infty(\Omega)}$, the L_∞ norm of g in Ω. The partial derivative operator $\frac{\partial}{\partial x_i}$ is denoted by ∂_i, for $i = 1, \ldots, n$. So for instance, $\frac{\partial^3 g}{\partial x_1^2 \partial x_2}$ will be denoted by $\partial_1^2 \partial_2 g$.

We are interested in the problem of approximating functions $f : \mathbb{R}^d \to \mathbb{R}$ from point queries. For some unknown sets $\mathcal{S}_1 \subset [d], \mathcal{S}_2 \subset \binom{[d]}{2}$, the function f is assumed to have the following form.

$$f(x_1, \ldots, x_d) = \sum_{p \in \mathcal{S}_1} \phi_p(x_p) + \sum_{(l,l') \in \mathcal{S}_2} \phi_{(l,l')}(x_l, x_{l'}). \tag{2.1}$$

Hence f is considered to be a sum of a sparse number of univariate and bivariate functions, denoted by ϕ_p and $\phi_{(l,l')}$ respectively. Here, $\phi_{(l,l')}$ is considered to be "truly bivariate" meaning that $\partial_l \partial_{l'} \phi_{(l,l')} \not\equiv 0$. The set of coordinate variables that are in $\mathcal{S}_2$, is denoted by

$$\mathcal{S}_2^{\text{var}} := \{ l \in [d] : \exists l' \in [d] \text{ s.t } (l, l') \in \mathcal{S}_2 \text{ or } (l', l) \in \mathcal{S}_2 \}. \tag{2.2}$$

For each $l \in \mathcal{S}_2^{\text{var}}$, we refer to the number of occurrences of l in $\mathcal{S}_2$, as the *degree* of l, formally denoted as follows.

$$\rho(l) := |\{ l' \in \mathcal{S}_2^{\text{var}} : (l, l') \in \mathcal{S}_2 \text{ or } (l', l) \in \mathcal{S}_2 \}|; \quad l \in \mathcal{S}_2^{\text{var}}. \tag{2.3}$$

Model Uniqueness. We first note that representation (2.1) is not unique. Firstly, we could add constants to each $\phi_l, \phi_{(l,l')}$, which sum up to zero. Furthermore, for each $l \in \mathcal{S}_2^{\text{var}}$ with $\rho(l) > 1$ we could add univariates that sum to zero. We can do the same for $l \in \mathcal{S}_1 \cap \mathcal{S}_2^{\text{var}} : \rho(l) = 1$. These ambiguities are thankfully avoided by re-writing (2.1) *uniquely* in the following ANOVA form.

$$f(x_1, \ldots, x_d) = c + \sum_{p \in \mathcal{S}_1} \phi_p(x_p) + \sum_{(l,l') \in \mathcal{S}_2} \phi_{(l,l')}(x_l, x_{l'}) + \sum_{q \in \mathcal{S}_2^{\text{var}} : \rho(q) > 1} \phi_q(x_q); \quad \mathcal{S}_1 \cap \mathcal{S}_2^{\text{var}} = \emptyset. \tag{2.4}$$

Here, $c = \mathbb{E}[f]$ and $\mathbb{E}_p[\phi_p] = \mathbb{E}_{(l,l')}[\phi_{(l,l')}] = 0$; $\forall p \in \mathcal{S}_1, (l,l') \in \mathcal{S}_2$, with expectations being over uniform distributions w.r.t. variable range $[-1,1]$. In addition, certain bivariate components have zero marginal mean with respect to either l or l'. In particular, $\mathbb{E}_l[\phi_{(l,l')}] = 0$ if $\rho(l') > 1$ and $\mathbb{E}_{l'}[\phi_{(l,l')}] = 0$ if $\rho(l) > 1$. The univariate ϕ_q corresponding to $q \in \mathcal{S}_2^{\text{var}}$ with $\rho(q) > 1$, represents the net marginal effect of the variable and has $\mathbb{E}_q[\phi_q] = 0$. We note that $\mathcal{S}_1, \mathcal{S}_2^{\text{var}}$ are disjoint in (2.4). This is due to the fact that each $p \in \mathcal{S}_1 \cap \mathcal{S}_2^{\text{var}}$ with $\rho(p) = 1$ can be merged with its bivariate form, while each $p \in \mathcal{S}_1 \cap \mathcal{S}_2^{\text{var}}$ with $\rho(p) > 1$ can be merged with its net marginal univariate form. The uniqueness of (2.4) is shown formally in the appendix.

We assume the setting $|\mathcal{S}_1| = k_1 \ll d$, $|\mathcal{S}_2| = k_2 \ll d$. Clearly, $|\mathcal{S}_2^{\text{var}}| \leq 2k_2$ with equality iff elements in $\mathcal{S}_2$ are pairwise disjoint. The set of *all* active variables, *i.e.*, $\mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$ will be denoted by $\mathcal{S}$. We then define $k := |\mathcal{S}| = k_1 + |\mathcal{S}_2^{\text{var}}|$ to be the *total sparsity* of the problem. The largest degree of a variable in $\mathcal{S}_2^{\text{var}}$, is defined to be $\rho_m := \max_{l \in \mathcal{S}_2^{\text{var}}} \rho(l)$. Clearly, $1 \leq \rho_m \leq k_2$.

Goals. Assuming that we have the freedom to query f within its domain, our goal is now two fold.

- Firstly, we would like to exactly recover the unknown sets $\mathcal{S}_1, \mathcal{S}_2$.
- Secondly, we would like to estimate c as well as each: (i) $\phi_p; p \in \mathcal{S}_1$, (ii) $\phi_{(l,l')}; (l,l') \in \mathcal{S}_2$ and (iii) $\phi_q; q \in \mathcal{S}_2^{\text{var}}, \rho(q) > 1$, in (2.4). In particular, we would like to estimate the univariate and bivariate components within compact domains $[-1,1]$, $[-1,1]^2$ respectively.

If $\mathcal{S}_1, \mathcal{S}_2$ were known beforehand, then one can estimate f via standard results from approximation theory or nonparametric regression [4]. Hence our primary focus in the paper is to recover $\mathcal{S}_1, \mathcal{S}_2$. Our main assumptions for this problem are listed below.

Assumption 1. *We assume that f can be queried from the slight enlargement: $[-(1+r),(1+r)]^d$ of $[-1,1]^d$ for some small $r > 0$. As will be seen later, the enlargement r can be made arbitrarily close to 0.*

Assumption 2. *We assume each $\phi_{(l,l')}, \phi_p$ to be three times continuously differentiable, within $[-(1+r),(1+r)]^2$ and $[-(1+r),(1+r)]$ respectively. Since these domains are compact, there then exist constants $B_m \geq 0$; $m = 0,1,2,3$, so that*

$$\| \partial_l^{m_1} \partial_{l'}^{m_2} \phi_{(l,l')} \|_{L_\infty[-(1+r),(1+r)]^2} \leq B_m; \quad (l,l') \in \mathcal{S}_2,\ m_1 + m_2 = m, \tag{2.5}$$

$$\| \partial_p^m \phi_p \|_{L_\infty[-(1+r),(1+r)]} \leq B_m; \quad p \in \mathcal{S}_1 \text{ or, } p \in \mathcal{S}_2^{var} \ \&\ \rho(p) > 1. \tag{2.6}$$

Our next assumption is for the purpose of identification of active variables, *i.e.*, the elements of $\mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$.

Assumption 3. *For some constants $D_1, \lambda_1 > 0$, we assume that for each $(l,l') \in \mathcal{S}_2$, $\exists$ connected $\mathcal{I}_{l,1}, \mathcal{I}_{l',1}, \mathcal{I}_{l,2}, \mathcal{I}_{l',2} \subset [-1,1]$, each of Lebesgue measure at least $\lambda_1 > 0$, so that*

$$|\partial_l \phi_{(l,l')}(x_l, x_{l'})| > D_1, \quad \forall (x_l, x_{l'}) \in \mathcal{I}_{l,1} \times \mathcal{I}_{l',1}, \tag{2.7}$$

$$|\partial_{l'} \phi_{(l,l')}(x_l, x_{l'})| > D_1, \quad \forall (x_l, x_{l'}) \in \mathcal{I}_{l,2} \times \mathcal{I}_{l',2}. \tag{2.8}$$

Similarly, we assume that for each $p \in \mathcal{S}_1$, $\exists$ connected $\mathcal{I}_p \subset [-1,1]$, of Lebesgue measure at least $\lambda_1 > 0$, such that $|\partial_p \phi_p(x_p)| > D_1$, $\forall x_p \in \mathcal{I}_p$. These assumptions essentially serve to distinguish an active variable from a non-active one, and are also in a sense necessary. For instance, if say $\partial_l \phi_{(l,l')}$ was zero throughout $[-1,1]^2$, then it equivalently means that $\partial_l \phi_{(l,l')}$ is only a function of $x_{l'}$. If $\partial_l \phi_{(l,l')} = \partial_{l'} \phi_{(l,l')} = 0$ in $[-1,1]^2$, then $\phi_{(l,l')} \equiv 0$ in $[-1,1]^2$. The same reasoning applies for ϕ_p's.

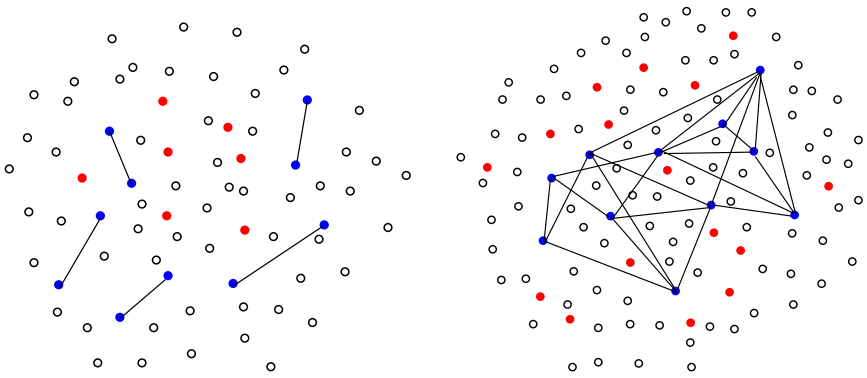
Our last assumption concerns the identification of $\mathcal{S}_2$.

Assumption 4. *For some constants $D_2, \lambda_2 > 0$, we assume that for each $(l,l') \in \mathcal{S}_2$, $\exists$ connected $\mathcal{I}_l, \mathcal{I}_{l'} \subset [-1,1]$, each interval of Lebesgue measure at least $\lambda_2 > 0$, such that $|\partial_l \partial_{l'} \phi_{(l,l')}(x_l, x_{l'})| > D_2, \quad \forall (x_l, x_{l'}) \in \mathcal{I}_l \times \mathcal{I}_{l'}$.*

Our problem specific parameters are: (i) B_i; $i = 0, \ldots, 3$, (ii) D_j, λ_j; $j = 1,2$ and, (iii) k, ρ_m. We do not assume k_1, k_2 to be known but instead assume that k is known. Furthermore it suffices to use estimates for the problem parameters instead of exact values. In particular, we can use upper bounds for: $k, \rho_m, B_i; i = 0, \ldots, 3$ and lower bounds for: $D_j, \lambda_j; j = 1,2$.

Underlying interaction graph. One might intuitively guess that the underlying "structure" of interactions between the elements in $\mathcal{S}_2^{\text{var}}$, shapes the difficulty of the problem. More formally, consider the graph $G = (V, E)$ where $V = [d]$ and $E = \mathcal{S}_2 \subset \binom{V}{2}$ denote the set of vertices and edges, respectively. We refer to the induced subgraph $I_G = (\mathcal{S}_2^{\text{var}}, \mathcal{S}_2)$ of G, as the *interaction graph*. We consider not only the general setting – where no assumption is made on I_G – but also a special case where I_G is a perfect matching. This is illustrated in Figure 1. In Fig. 1a, I_G is a perfect matching meaning that each vertex is of degree one. In other words, there is *no overlap* between the elements of $\mathcal{S}_2$. In terms of the difficulty of interactions, this corresponds to the easiest setting. Fig. 1b corresponds to the general setting where no structural assumption is placed on I_G. Therefore, we can now potentially *have overlaps* between the elements of $\mathcal{S}_2$, since each element in $\mathcal{S}_2^{\text{var}}$ can be paired with up to ρ_m other elements. This corresponds to the hardest setting as far as the difficulty of interactions is concerned.

[4] This is discussed later.



(a) I_G is a perfect matching (b) I_G has arbitrary structure

Figure 1: Blue (resp. red) disks denote elements of $\mathcal{S}_2^{\text{var}}$ (resp. $\mathcal{S}_1$). Circles denote elements of $[d] \setminus \{\mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}\}$. On the left, we have the special setting where I_G is a perfect matching. On the right, we have the most general setting where no assumption is made on I_G.

3 Sampling scheme for the non-overlap case

In this section we consider the special case where all elements in $\mathcal{S}_2$ are pair-wise disjoint. In other words, $\rho(i) = 1$, for each $i \in \mathcal{S}_2^{\text{var}}$. We first treat the noiseless setting in Section 3.1, wherein the exact function values are obtained at each query. We then handle the noisy setting in Section 3.2, where the function values are corrupted with external noise.

3.1 Analysis for noiseless setting

Our approach essentially consists of two phases. In the first phase, we sample the function f appropriately, and recover the *complete set* of active variables $\mathcal{S}$. In the second phase, we focus on the reduced k dimensional subspace corresponding to $\mathcal{S}$. We sample f at appropriate points in this subspace, and consequently identify $\mathcal{S}_1$ as well as $\mathcal{S}_2$. Let us now elaborate on these two phases in more detail.

3.1.1 First Phase: Recovering all active variables

The crux of this phase is based on the following observation. On account of the structure of f, we see that at any $\mathbf{x} \in \mathbb{R}^d$, the gradient $\nabla f(\mathbf{x}) \in \mathbb{R}^d$ has the following form:

$$(\nabla f(\mathbf{x}))_q = \begin{cases} \partial_q \phi_q(x_q) & ; \quad q \in \mathcal{S}_1 \\ \partial_q \phi_{(q,q')}(x_q, x_{q'}) & ; \quad (q,q') \in \mathcal{S}_2 \\ \partial_q \phi_{(q',q)}(x_{q'}, x_q) & ; \quad (q',q) \in \mathcal{S}_2 \\ 0 & ; \quad \text{otherwise} \end{cases} \quad ; \quad q = 1, \ldots, d.$$

Hence $\nabla f(\mathbf{x})$ is at most k-sparse, *i.e.*, has at most k non zero entries, for any $\mathbf{x}$. Note that the q^{th} component of $\nabla f(\mathbf{x})$ is zero if $q \notin \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$. Say we somehow recover $\nabla f(\mathbf{x})$ at sufficiently many $\mathbf{x}$'s within $[-1,1]^d$. Then, we would also have suitably many samples of the functions: $\partial_q \phi_q, \partial_l \phi_{(l,l')}, \partial_{l'} \phi_{(l,l')}, \forall\ p \in \mathcal{S}_1, (l,l') \in \mathcal{S}_2$. Specifically, if the number of samples is large enough, then we would have sampled each of $\partial_q \phi_q, \partial_l \phi_{(l,l')}, \partial_{l'} \phi_{(l,l')}$, within their respective "critical intervals", as defined in Assumption 3. Provided that the estimation noise is sufficiently small enough, this suggests that we should then, via a threshold operation, be able to detect all variables in $\mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$. We now proceed to formalize our above discussion, in a systematic manner.

Compressive sensing formulation. We begin by discussing how a sparse gradient ∇f can be estimated at any point $\mathbf{x}$, via compressive sensing (CS). As f is $\mathcal{C}^3$ smooth, therefore the Taylor's expansion of f at $\mathbf{x}$, along $\mathbf{v}, -\mathbf{v} \in \mathbb{R}^d$, with step size $\mu > 0$, and $\zeta = \mathbf{x} + \theta \mathbf{v}$, $\zeta' = \mathbf{x} - \theta' \mathbf{v}$; $\theta, \theta' \in (0, \mu)$ gives us:

$$f(\mathbf{x} + \mu \mathbf{v}) = f(\mathbf{x}) + \mu \langle \mathbf{v}, \nabla f(\mathbf{x}) \rangle + \frac{1}{2} \mu^2 \mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} + R_3(\zeta), \tag{3.1}$$

$$f(\mathbf{x} - \mu \mathbf{v}) = f(\mathbf{x}) + \mu \langle -\mathbf{v}, \nabla f(\mathbf{x}) \rangle + \frac{1}{2} \mu^2 \mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} + R_3(\zeta'). \tag{3.2}$$

Subtracting the above, and dividing by 2μ leads to the standard "central difference" estimate of $\langle \mathbf{v}, \nabla f(\mathbf{x}) \rangle$.

$$\frac{f(\mathbf{x} + \mu \mathbf{v}) - f(\mathbf{x} - \mu \mathbf{v})}{2\mu} = \langle \mathbf{v}, \nabla f(\mathbf{x}) \rangle + \underbrace{\frac{R_3(\zeta) - R_3(\zeta')}{2\mu}}_{O(\mu^2)}. \tag{3.3}$$

Notice that in (3.3), the expression on the left hand side corresponds to a noisy-linear measurement of $\nabla f(\mathbf{x})$, with $\mathbf{v}$. The "noise" here arises on account of the third order terms $R_3(\zeta), R_3(\zeta') = O(\mu^3)$, in the Taylor expansion. Now let the $\mathbf{v}$'s be chosen from the set:

$$\mathcal{V} := \left\{ v_j \in \mathbb{R}^d : v_{j,q} = \pm \frac{1}{\sqrt{m_v}} \text{ w.p. } 1/2 \text{ each; } j = 1, \ldots, m_v \text{ and } q = 1, \ldots, d \right\}. \tag{3.4}$$

Then, employing (3.3) at each $\mathbf{v}_j \in \mathcal{V}$ gives us the linear system:

$$\underbrace{\frac{f(\mathbf{x}+\mu\mathbf{v}_j) - f(\mathbf{x}-\mu\mathbf{v}_j)}{2\mu}}_{y_j} = \langle \mathbf{v}_j, \nabla f(\mathbf{x}) \rangle + \underbrace{\frac{R_3(\zeta_j) - R_3(\zeta_j')}{2\mu}}_{n_j}; \quad j = 1, \ldots, m_v. \tag{3.5}$$

Denoting $\mathbf{y} = [y_1 \ldots y_{m_v}]$, $\mathbf{n} = [n_1 \ldots n_{m_v}]$ and $\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_{m_v}]^T \in \mathbb{R}^{m_v \times d}$, we can re-write (3.5) succinctly as:

$$\mathbf{y} = \mathbf{V}\nabla f(\mathbf{x}) + \mathbf{n}. \tag{3.6}$$

As we know $\mathbf{y}, \mathbf{V}$, therefore we can estimate the unknown k-sparse vector $\nabla f(\mathbf{x})$ via standard ℓ_1 minimization [6, 16]:

$$\widehat{\nabla} f(\mathbf{x}) := \underset{\mathbf{y}=\mathbf{V}\mathbf{z}}{\operatorname{argmin}} \parallel \mathbf{z} \parallel_1 . \tag{3.7}$$

Remark 1. *Estimating sparse gradients via compressive sensing was – to the best of our knowledge – first considered by Fornasier et al. [18] for learning functions of the form: $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$. It was then also employed by Tyagi et al. [54] for learning SPAMs (without interaction terms). However, [18, 54] consider a "forward difference" estimate of $\langle \mathbf{v}, \nabla f(\mathbf{x}) \rangle$, via $(f(\mathbf{x}+\mu\mathbf{v}) - f(\mathbf{x}))/\mu$, resulting in $O(\mu)$ perturbation error in* (3.3)*.*

Remark 2. *The above sampling mechanism is related to the "simultaneous perturbation" gradient approximation method of [48]. Specifically in [48], for a random $\mathbf{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, $\widehat{\nabla} f(\mathbf{x})$ is defined to be:*

$$\left(\frac{f(\mathbf{x}+\mu\mathbf{v}) - f(\mathbf{x}-\mu\mathbf{v})}{2\mu v_1}, \ldots, \frac{f(\mathbf{x}+\mu\mathbf{v}) - f(\mathbf{x}-\mu\mathbf{v})}{2\mu v_d} \right)^T \tag{3.8}$$

The bias of the above estimate can be shown to be $O(\mu^2)$ for C^3 smooth f.

The following theorem from [18] provides guarantees for stable recovery via ℓ_1 minimization: $\triangle(\mathbf{y}) := \underset{\mathbf{y}=\mathbf{V}\mathbf{z}}{\operatorname{argmin}} \parallel \mathbf{z} \parallel_1$. While the first part is by now standard (see for example [2]), the second result was stated in [18] as a specialization of Theorem 1.2 from [60] to the case of Bernoulli measurement matrices.

Theorem 1 ([60, 18])**.** *Let $\mathbf{V}$ be a $m_v \times d$ random matrix with all entries being Bernoulli i.i.d random variables scaled with $1/\sqrt{m_v}$. Then the following results hold.*

1. *Let $0 < \kappa < 1$. Then there are two positive constants $c_1, c_2 > 0$, such that the matrix $\mathbf{V}$ has the Restricted Isometry Property*

$$(1-\kappa) \parallel \mathbf{w} \parallel_2^2 \leq \parallel \mathbf{V}\mathbf{w} \parallel_2^2 \leq (1+\kappa) \parallel \mathbf{w} \parallel_2^2 \tag{3.9}$$

   for all $\mathbf{w} \in \mathbb{R}^d$ such that $|supp(\mathbf{w})| \leq c_2 m_v / \log(d/m_v)$ with probability at least $1 - e^{-c_1 m_v}$.

2. *Let us suppose $d > (\log 6)^2 m_v$. Then there are positive constants $C, c_1', c_2' > 0$ such that with probability at least $1 - e^{-c_1' m_v} - e^{-\sqrt{m_v d}}$ the matrix $\mathbf{V}$ has the following property. For every $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{n} \in \mathbb{R}^{m_v}$ and every natural number $K \leq c_2' m_v / \log(d/m_v)$, we have*

$$\parallel \triangle(\mathbf{V}\mathbf{w} + \mathbf{n}) - \mathbf{w} \parallel_2 \leq C \left(K^{-1/2} \sigma_K(\mathbf{w})_1 + \max \left\{ \parallel \mathbf{n} \parallel_2, \sqrt{\log d} \parallel \mathbf{n} \parallel_\infty \right\} \right), \tag{3.10}$$

   where

$$\sigma_K(\mathbf{w})_1 := \inf \{ \parallel \mathbf{w} - \mathbf{z} \parallel_1 : |supp(\mathbf{z})| \leq K \}$$

   is the best K-term approximation of $\mathbf{w}$.

Remark 3. *The proof of the second part of Theorem 1 requires* (3.9) *to hold, which is the case in our setting with high probability.*

Remark 4. *Since $m_v \geq K$ is necessary, note that $K \leq c_2' m_v / \log(d/m_v)$ is satisfied if $m_v > (1/c_2') K \log(d/K)$. Also note that $K \log(d/K) > \log d$ in the regime[5] $K \ll d$. As pointed out by a reviewer, a slight improvement over Theorem 1 is given by [19, Theorem 11.10] where the $\log d$ term in* (3.10) *is replaced with $\log(d/m_v)$.*

[5] More precisely, if $d > K^{\frac{K}{K-1}}$.

Estimating sufficiently many gradients. Given the discussion above, the next natural question is - how should one choose the points $\mathbf{x}$, where the gradient $\nabla f(\mathbf{x})$ should be estimated? Note that f is composed of the sum of univariate and bivariate functions, residing on mutually orthogonal 1 or 2 dimensional canonical subspaces of $\mathbb{R}^d$. Therefore, this suggests that it is sufficient if our set of points – let us call it χ – has the property that it provides a 2-dimensional discretization of *any canonical* 2 *dimensional subspace of* $\mathbb{R}^d$. In order to construct χ we will make use of hash functions or more specifically - a family of hash functions, defined as follows.

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

Hash functions are common in theoretical computer science, and are widely used such as in finding juntas [34]. There exists a fairly simple probabilistic method using which one can construct $\mathcal{H}_t^d$ of size $O(te^t \log d)$ with high probability. 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. We note that the size of $\mathcal{H}_t^d$ is *nearly optimal* - it is known that the size of any such family is $\Omega(e^t \log d/\sqrt{t})$ [20, 27, 40]. There also exist efficient *deterministic* constructions for such families of partitions, with the size of the family being $O(t^{O(\log t)} e^t \log d)$ and which take time linear in the output size [36]. For our purposes, we consider the probabilistic construction of the family due to its smaller resulting size. Specifically, we consider the family $\mathcal{H}_2^d$ so that for any distinct i, j, there exists $h \in \mathcal{H}_2^d$ s.t $h(i) \neq h(j)$. Let us first define for any $h \in \mathcal{H}_2^d$, the vectors $\mathbf{e}_1(h), \mathbf{e}_2(h) \in \mathbb{R}^d$ where:

$$(\mathbf{e}_i(h))_q := \begin{cases} 1 & ; \quad h(q) = i, \\ 0 & ; \quad \text{otherwise} \end{cases} \qquad \text{for } i = 1, 2 \text{ and } q = 1, \ldots, d. \tag{3.11}$$

Given at hand $\mathcal{H}_2^d$, we construct our set χ using the procedure[6] in [15]. Specifically, for some integer $m_x > 0$, we construct for each $h \in \mathcal{H}_2^d$ the set $\chi(h)$ as:

$$\chi(h) := \left\{ \mathbf{x}(h) \in [-1, 1]^d : \mathbf{x}(h) = \sum_{i=1}^{2} c_i \mathbf{e}_i(h); c_1, c_2 \in \left\{ -1, -\frac{m_x - 1}{m_x}, \ldots, \frac{m_x - 1}{m_x}, 1 \right\} \right\}. \tag{3.12}$$

Note that $\chi(h)$ consists of $(2m_x + 1)^2$ points that discretize: $\text{span}(\mathbf{e}_1(h), \mathbf{e}_2(h))$, within $[-1, 1]^d$, with a spacing of $1/m_x$ along each $\mathbf{e}_i$. Given this, we obtain the complete set as $\chi = \cup_{h \in \mathcal{H}_2^d} \chi(h)$ so that $|\chi| \leq (2m_x + 1)^2 |\mathcal{H}_2^d|$. Clearly, χ discretizes *any* 2-dimensional canonical subspace, within $[-1, 1]^d$.

Recovering set of active variables. Our scheme for recovering the set of active variables is outlined formally in the form of Algorithm 1. At each $\mathbf{x} \in \chi$, we obtain the estimate $\widehat{\nabla} f(\mathbf{x})$ via ℓ_1 minimization. We then perform a thresholding operation, *i.e.*, set to zero those components of $\widehat{\nabla} f(\mathbf{x})$, whose magnitude is below a certain threshold. All indices then corresponding to non zero components are identified as active variables.

Algorithm 1 Sub-routine for estimating $\mathcal{S}$

1: Construct $(d, 2)$-hash family $\mathcal{H}_2^d$ and the set $\mathcal{V}$ for suitable $m_v \in \mathbb{Z}^+$. Choose suitable $\mu \in \mathbb{Z}^+$ and initialize $\widehat{\mathcal{S}} = \emptyset$.
2: Choose suitable $m_x \in \mathbb{Z}^+$. For each $h \in \mathcal{H}_2^d$ do:

1. Create the set $\chi(h)$. For $\mathbf{x}_i \in \chi(h); i = 1, \ldots, (2m_x + 1)^2$ do:
   (a) Construct $\mathbf{y}_i$ where $(\mathbf{y}_i)_j = \frac{f(\mathbf{x}_i + \mu \mathbf{v}_j) - f(\mathbf{x}_i - \mu \mathbf{v}_j)}{2\mu}; j = 1, \ldots, m_v$.
   (b) Set $\widehat{\nabla} f(\mathbf{x}_i) := \underset{\mathbf{y}_i = \mathbf{V}\mathbf{z}}{\text{argmin}} \parallel \mathbf{z} \parallel_1$. For suitable $\tau > 0$, update:

$$\widehat{\mathcal{S}} = \widehat{\mathcal{S}} \cup \left\{ q \in \{1, \ldots, d\} : |(\widehat{\nabla} f(\mathbf{x}_i))_q| > \tau \right\}.$$

The following Lemma provides sufficient conditions on the sampling parameters: m_x, m_v, μ and the threshold τ, which guarantee that $\widehat{\mathcal{S}} = \mathcal{S}$ holds.

Lemma 1. *Let $\mathcal{H}_2^d$ be of size $|\mathcal{H}_2^d| \leq 2(C_1 + 1)e^2 \log d$ for some constant $C_1 > 1$. Then there exist constants $c_3' \geq 1$ and $C, c_1' > 0$ such that for any m_x, m_v, μ satisfying*

$$c_3' k \log(d/k) < m_v < d/(\log 6)^2, \quad m_x \geq \lambda_1^{-1} \quad \textit{and} \quad \mu < \left(\frac{3D_1 m_v}{4CB_3 k} \right)^{1/2}, \tag{3.13}$$

the choice $\tau = \frac{2C\mu^2 B_3 k}{3m_v}$ implies that $\widehat{\mathcal{S}} = \mathcal{S}$ holds with probability at least $1 - e^{-c_1' m_v} - e^{-\sqrt{m_v d}} - d^{-2C_1}$. Here $\lambda_1, D_1, B_3 > 0$ are problem specific constants defined in Section 2.

[6] Such sets were used in [15] for a more general problem involving functions that are intrinsically k variate, and do not necessarily have an additive structure.

Query complexity. We estimate ∇f at $(2m_x+1)^2|\mathcal{H}_2^d|$ many points. For each such estimate, we query f at $2m_v$ points, leading to a total of $2m_v(2m_x+1)^2|\mathcal{H}_2^d|$ queries. From Lemma 1, we then obtain a query complexity of $O(k(\log d)^2\lambda_1^{-2})$ for exact recovery of the set of active variables, *i.e.*, $\mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$.

Computational complexity. The family $\mathcal{H}_2^d$ can be constructed[7] in time polynomial in d. Step 1b involves solving a linear program in $O(d)$ variables, which can be done efficiently up to arbitrary accuracy, in time polynomial in (m_v, d) (using for instance, interior point methods (cf., [39]). Since we solve $O(\lambda_1^{-2}\log d)$ such linear programs, hence the overall computation time is polynomial in the number of queries and dimension d.

Remark 5. *It is worth noting that in practice, it might be preferable to replace the ℓ_1 minimization step with a non-convex algorithm such as "Iterative hard thresholding" (IHT) (cf., [4, 5, 28, 29, 30]). Such methods consider solving the non-convex optimization problem:*

$$\min_{\mathbf{z}} \| \mathbf{V}\mathbf{z} - \mathbf{y} \|^2 \quad \textit{s.t.} \quad \| \mathbf{z} \|_0 \leq K$$

for finding a K-sparse solution to an under-determined linear system of equations, and generally have a lower computational complexity than their convex analogues. Moreover, provided $\mathbf{V}$ also satisfies the Restricted Isometry Property (as stated in 3.9), they then also enjoy strong theoretical guarantees, similar to that for convex approaches.

Remark 6. *Algorithm 1 essentially estimates ∇f at $O(\log d)$ points. The method of Fornasier et al. [18] is designed for a more general function class than ours and hence involves estimating ∇f on points sampled uniformly at random from the unit sphere $\mathbb{S}^{d-1}$ – the size of such a set is typically polynomial in d. The method of Tyagi et al. [54] is tailored towards SPAMs without interactions; it essentially estimates ∇f along a uniform one-dimensional grid (hence at constantly many points). Hence conceptually, Algorithm 1 is a simple generalization of the scheme of Tyagi et al. [54].*

3.1.2 Second Phase: Recovering individual sets

Given that we have recovered $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$, we now proceed to see how we can recover the individual sets: $\mathcal{S}_1$ and $\mathcal{S}_2$. Let us denote w.l.o.g, $\mathcal{S}$ to be $\{1, 2, \ldots, k\}$ and also denote $g : \mathbb{R}^k \to \mathbb{R}$ to be

$$g(x_1, x_2, \ldots, x_k) = c + \sum_{p \in \mathcal{S}_1} \phi_p(x_p) + \sum_{(l,l') \in \mathcal{S}_2} \phi_{(l,l')}(x_l, x_{l'}). \tag{3.14}$$

Here $\mathcal{S}_2 \subset \binom{[k]}{2}$ with $\mathcal{S}_2^{\text{var}} \cap \mathcal{S}_1 = \emptyset$. We have reduced our problem to that of querying some unknown function k-variate function g, of the form (3.14), with queries $\mathbf{x} \in \mathbb{R}^k$. Indeed, this is equivalent to querying f at $(\mathbf{x})_{\mathcal{S}}$, *i.e.*, the restriction of $\mathbf{x}$ onto $\mathcal{S}$.

In order to identify $\mathcal{S}_1$ and $\mathcal{S}_2$, let us recall the discussion in Assumption 4 : for any $(l, l') \in \mathcal{S}_2$, we will have that $\exists (x_l, x_{l'}) \in [-1, 1]^2$ such that $\partial_l \partial_{l'} g(\mathbf{x}) = \partial_l \partial_{l'} \phi_{(l,l')}(x_l, x_{l'}) \neq 0$. Furthermore for $p \in \mathcal{S}_1$ and any $p' \neq p$, we know that $\partial_p \partial_{p'} g(\mathbf{x}) \equiv 0$, $\forall \mathbf{x} \in \mathbb{R}^k$. In light of this, our goal will be now to query g in order to estimate the *off-diagonal* entries of its Hessian $\nabla^2 g$. This is a natural approach as these entries contain information about the mixed second order partial derivatives of g. We now proceed towards motivating our sampling scheme.

Motivation behind sampling scheme. At any $\mathbf{x} \in \mathbb{R}^k$ the Hessian $\nabla^2 g(\mathbf{x})$ is a $k \times k$ symmetric matrix with the following structure.

$$(\nabla^2 g(\mathbf{x}))_{i,j} = \begin{cases} \partial_i^2 \phi_i(x_i) & ; \quad i \in \mathcal{S}_1,\ i = j \\ \partial_i^2 \phi_{(i,i')}(x_i, x_{i'}) & ; \quad (i, i') \in \mathcal{S}_2,\ j = i \\ \partial_i^2 \phi_{(i',i)}(x_{i'}, x_i) & ; \quad (i', i) \in \mathcal{S}_2,\ j = i \\ \partial_i \partial_j \phi_{(i,j)}(x_i, x_j) & ; \quad (i, j) \in \mathcal{S}_2 \\ \partial_i \partial_j \phi_{(j,i)}(x_j, x_i) & ; \quad (j, i) \in \mathcal{S}_2 \\ 0 & ; \quad \text{otherwise} \end{cases}.$$

Note that each row of $\nabla^2 g$ has at most 2 non zero entries. If $i \in \mathcal{S}_1$, then the non zero entry can only be the $(i,i)^{th}$ entry of $\nabla^2 g$. If $i \in \mathcal{S}_2^{\text{var}}$, then the i^{th} row can have two non zero entries. In this case, the non zero entries will be the $(i,i)^{th}$ and $(i,j)^{th}$ entries of $\nabla^2 g$, if $(i,j) \in \mathcal{S}_2$ or $(j,i) \in \mathcal{S}_2$.

Now, for $\mathbf{x}, \mathbf{v} \in \mathbb{R}^k$, $\mu_1 > 0$, consider the Taylor expansion of ∇g at $\mathbf{x}$ along $\mathbf{v}$, with step size μ_1. For $\zeta_i = \mathbf{x} + \theta_i \mathbf{v}$, for some $\theta_i \in (0, \mu_1)$; $i = 1, \ldots, k$, we have:

$$\frac{\nabla g(\mathbf{x} + \mu_1 \mathbf{v}) - \nabla g(\mathbf{x})}{\mu_1} = \nabla^2 g(\mathbf{x})\mathbf{v} + \frac{\mu_1}{2} \begin{pmatrix} \mathbf{v}^T \nabla^2 \partial_1 g(\zeta_1) \mathbf{v} \\ \vdots \\ \mathbf{v}^T \nabla^2 \partial_k g(\zeta_k) \mathbf{v} \end{pmatrix}. \tag{3.15}$$

Alternately, we have the following identity for each individual $\partial_i g$.

$$\frac{\partial_i g(\mathbf{x} + \mu_1 \mathbf{v}) - \partial_i g(\mathbf{x})}{\mu_1} = \langle \nabla \partial_i g(\mathbf{x}), \mathbf{v} \rangle + \frac{\mu_1}{2} \mathbf{v}^T \nabla^2 \partial_i g(\zeta_i) \mathbf{v}; \quad i = 1, \ldots, k. \tag{3.16}$$

[7] Recall discussion following Definition 1.

Say we estimate $\partial_i g(\mathbf{x}), \partial_i g(\mathbf{x}+\mu_1\mathbf{v})$ with $\widehat{\partial}_i g(\mathbf{x}), \widehat{\partial}_i g(\mathbf{x}+\mu_1\mathbf{v})$ respectively, using finite differences with step size parameter $\beta > 0$. Then we can write

$$\widehat{\partial}_i g(\mathbf{x}) = \partial_i g(\mathbf{x}) + \eta_i(\mathbf{x},\beta), \quad \widehat{\partial}_i g(\mathbf{x}+\mu_1\mathbf{v}) = \partial_i g(\mathbf{x}+\mu_1\mathbf{v}) + \eta_i(\mathbf{x}+\mu_1\mathbf{v},\beta) \tag{3.17}$$

with $\eta_i(\mathbf{x},\beta), \eta_i(\mathbf{x}+\mu_1\mathbf{v},\beta) = O(\beta^2)$ being the corresponding estimation errors. Plugging these estimates in (3.16), we finally obtain the following.

$$\frac{\widehat{\partial}_i g(\mathbf{x}+\mu_1\mathbf{v}) - \widehat{\partial}_i g(\mathbf{x})}{\mu_1} = \langle \nabla\partial_i g(\mathbf{x}), \mathbf{v}\rangle + \underbrace{\frac{\mu_1}{2}\mathbf{v}^T\nabla^2\partial_i g(\zeta_i)\mathbf{v} + \frac{\eta_i(\mathbf{x}+\mu_1\mathbf{v},\beta) - \eta_i(\mathbf{x},\beta)}{\mu_1}}_{\text{Error term}}. \tag{3.18}$$

We see in (3.18) that the L.H.S can be viewed as taking a noisy linear measurement of the i^{th} row of $\nabla^2 g(\mathbf{x})$ with measurement vector $\mathbf{v}$. Hence for any $i \in \mathcal{S}$ we can via (3.18) hope to recover the 2 sparse vector: $\nabla\partial_i g(\mathbf{x}) \in \mathbb{R}^k$. In fact, we are only interested in estimating the *off-diagonal* entries of $\nabla^2 g$. Therefore while testing for $i \in \mathcal{S}$, we can fix the i^{th} component of $\mathbf{v}$ to be zero. This means that $\nabla\partial_i g$ can in fact be considered as a 1 sparse vector, and our task is to find the location of the non zero entry. We now describe our sampling scheme that accomplishes this, by performing a binary search over $\nabla\partial_i g$.

Sampling scheme. Say that we are currently testing for variable $i \in \mathcal{S}$, *i.e.*, we would like to determine whether it is in $\mathcal{S}_1$ or $\mathcal{S}_2^{\text{var}}$. Denote $\mathcal{T}$ as the set of variables that have been classified so far. We will first create our set of points χ_i at which $\nabla\partial_i g$ will be estimated, as follows. Consider $\mathbf{e}_1(i), \mathbf{e}_2(i) \in \mathbb{R}^k$ where for $j = 1,\ldots,k$:

$$(\mathbf{e}_1(i))_j := \begin{cases} 1 & ; \quad j = i, \\ 0 & ; \quad \text{otherwise} \end{cases}, \quad (\mathbf{e}_2(i))_j := \begin{cases} 0 & ; \quad j = i \text{ or } j \in \mathcal{T}, \\ 1 & ; \quad \text{otherwise} \end{cases}. \tag{3.19}$$

We then form the following set of points which corresponds to a discretization of the 2-dimensional space spanned by $\mathbf{e}_1(i), \mathbf{e}_2(i)$, within $[-1,1]^k$.

$$\chi_i := \left\{ \mathbf{x} \in [-1,1]^k : \mathbf{x} = c_1\mathbf{e}_1(i) + c_2\mathbf{e}_2(i); c_1, c_2 \in \left\{ -1, -\frac{m_x'-1}{m_x'}, \ldots, \frac{m_x'-1}{m_x'}, 1 \right\} \right\}. \tag{3.20}$$

Now for each $\mathbf{x} \in \chi_i$ and suitable step size parameter $\beta > 0$, we will obtain the samples $g(\mathbf{x}+\beta\mathbf{e}_1(i)), g(\mathbf{x}-\beta\mathbf{e}_1(i))$. Then, we obtain via *central differences*, the estimate: $\widehat{\partial}_i g(\mathbf{x}) = (g(\mathbf{x}+\beta\mathbf{e}_1(i)) - g(\mathbf{x}-\beta\mathbf{e}_1(i)))/(2\beta)$. For our choice of $\mathbf{v}$ and parameter $\mu_1 > 0$, we can similarly obtain $\widehat{\partial}_i g(\mathbf{x}+\mu_1\mathbf{v})$. We now describe how the measurement vectors $\mathbf{v}$ can be chosen in an adaptive fashion, in order to identify $\mathcal{S}_1, \mathcal{S}_2$.

Firstly, we create a vector $\mathbf{v}_0(i)$ that enables us to test, whether there exists a variable $j \neq i$ such that $(i,j) \in \mathcal{S}_2$ (if $i > j$) or $(j,i) \in \mathcal{S}_2$ (if $j > i$). To this end, we set $\mathbf{v}_0(i) = \mathbf{e}_2(i)$. Clearly, $i \in \mathcal{S}_2^{\text{var}}$ iff there exists $\mathbf{x} \in [-1,1]^k$ such that $\langle \nabla\partial_i g(\mathbf{x}), \mathbf{v}_0(i)\rangle \neq 0$. This suggests the following strategy. For each $\mathbf{x} \in \chi_i$, we compute $(\widehat{\partial}_i g(\mathbf{x}+\mu_1\mathbf{v}_0(i)) - \widehat{\partial}_i g(\mathbf{x}))/(\mu_1)$ – this will be a noisy estimate of $\langle \nabla\partial_i g(\mathbf{x}), \mathbf{v}_0(i)\rangle$. Provided that the number of points is large enough and the noise is made suitably small, we see that via a threshold based procedure as in the previous phase, one would be able to correctly classify the other variable as either belonging to $\mathcal{S}_1$ or $\mathcal{S}_2$. In case the above procedure classifies i as being a part of $\mathcal{S}_2^{\text{var}}$, then we would still need to identify the other variable $j \in \mathcal{S}_2^{\text{var}}$, forming the pair. This can be handled via a binary search based procedure, as follows.

The measurement vectors $\mathbf{v}_1(i), \mathbf{v}_2(i), \ldots$ are chosen adaptively, meaning that the choice of $\mathbf{v}_j(i)$ depends on the past choices: $\mathbf{v}_1(i), \ldots, \mathbf{v}_{j-1}(i)$. $\mathbf{v}_1(i)$ is constructed as follows. We construct an equipartition $\mathcal{P}_1(i), \mathcal{P}_2(i) \subset \mathcal{S} \setminus \{\mathcal{T} \cup \{i\}\}$ such that: $\mathcal{P}_1(i) \cup \mathcal{P}_2(i) = \mathcal{S} \setminus \{\mathcal{T} \cup \{i\}\}$, $\mathcal{P}_1(i) \cap \mathcal{P}_2(i) = \emptyset$, $|\mathcal{P}_1(i)| = \lfloor \frac{k-1-|\mathcal{T}|}{2} \rfloor$ and $|\mathcal{P}_2(i)| = k-1-|\mathcal{T}|-|\mathcal{P}_1(i)|$. Then $\mathbf{v}_1(i)$ is chosen to be such that:

$$(\mathbf{v}_1(i))_l := \begin{cases} 1 & ; \quad l \in \mathcal{P}_1(i), \\ 0 & ; \quad \text{otherwise} \end{cases} \quad ; \quad l = 1,\ldots,k. \tag{3.21}$$

Let $\mathbf{x}^* \in \chi_i$ be the point, at which $\mathbf{v}_0(i)$ detects i. We now find: $(\widehat{\partial}_i g(\mathbf{x}^* + \mu_1\mathbf{v}_1(i)) - \widehat{\partial}_i g(\mathbf{x}^*))/\mu_1$, and test whether it is larger then a certain threshold. This tells us whether the other active variable j belongs to $\mathcal{P}_1(i)$ or to $\mathcal{P}_2(i)$. Then, we create $\mathbf{v}_2(i)$ by partitioning the identified subset, in the same manner as $\mathbf{v}_1(i)$ and perform the same tests again. It is clear that we would need at most $\lceil \log(k-|\mathcal{T}|) \rceil$ many $\mathbf{v}(i)$'s in this process. Hence, if $i \in \mathcal{S}_2^{\text{var}}$ then we would need at most $\lceil \log(k-|\mathcal{T}|) \rceil + 1$ measurement vectors in order to find the other member of the pair in $\mathcal{S}_2^{\text{var}}$. In case $i \in \mathcal{S}_1$, then $\mathbf{v}_0(i)$ by itself suffices. The above procedure is outlined formally in Algorithm 2.

We now provide sufficient conditions on the parameters $m_x' > 0, \beta$ and $\mu_1 > 0$, along with a corresponding threshold, that together guarantee recovery of $\mathcal{S}_1$ and $\mathcal{S}_2$. This is stated in the following lemma.

Lemma 2. *Let $m_x' > 0, \beta$ and $\mu_1 > 0$ be chosen to satisfy:*

$$m_x' \geq \lambda_2^{-1}, \quad \beta < \frac{\sqrt{3}D_2}{4\sqrt{2}B_3}, \quad \mu_1 \in \left(\frac{D_2 - \sqrt{D_2^2 - (32/3)\beta^2 B_3^2}}{8B_3}, \frac{D_2 + \sqrt{D_2^2 - (32/3)\beta^2 B_3^2}}{8B_3} \right). \tag{3.22}$$

Then for the choice $\tau' = \frac{\beta^2 B_3}{3\mu_1} + 2\mu_1 B_3$, we have for Algorithm 2 that $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ and $\widehat{\mathcal{S}_2} = \mathcal{S}_2$. Here, $B_3, D_2, \lambda_2 > 0$ are problem specific constants, defined in Section 2.

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

1: Initialize $\widehat{\mathcal{S}_1}, \widehat{\mathcal{S}_2} = \emptyset$.
2: **while** $\mathcal{S} \setminus \left\{\widehat{\mathcal{S}_1} \cup \widehat{\mathcal{S}_2^{\text{var}}}\right\} \neq \emptyset$ **do**
3: Choose $i \in \mathcal{S} \setminus \left\{\widehat{\mathcal{S}_1} \cup \widehat{\mathcal{S}_2^{\text{var}}}\right\}$. For suitable $m_x' \in \mathbb{Z}^+$, construct χ_i as in (3.20). Set $\mathbf{v}_0(i) = \mathbf{e}_2(i)$.
4: Choose $\mathbf{x} \in \chi_i$ that has not yet been chosen.
1. Obtain estimates: $\widehat{\partial}_i g(\mathbf{x}), \widehat{\partial}_i g(\mathbf{x} + \mu_1 \mathbf{v}_0(i))$ via central differences, for suitable $\mu_1, \beta > 0$.
2. If $\frac{|\widehat{\partial}_i g(\mathbf{x}+\mu_1 \mathbf{v}_0(i)) - \widehat{\partial}_i g(\mathbf{x})|}{\mu_1} > \tau'$, then denote $\mathbf{x}^* \leftarrow \mathbf{x}$ and go to 6. Else goto 4.
5: Update $\widehat{\mathcal{S}_1} = \widehat{\mathcal{S}_1} \cup \{i\}$ and go to 2.
6: Set $\mathcal{R} = \mathcal{S} \setminus \left\{\{i\} \cup \widehat{\mathcal{S}_1} \cup \widehat{\mathcal{S}_2^{\text{var}}}\right\}$.
7: **while** $|\mathcal{R}| > 1$ **do**
8: Initialize $\mathcal{P}_1(i), \mathcal{P}_2(i)$ as equipartition of $\mathcal{R}$. Construct $\mathbf{v}(i)$ w.r.t. $\mathcal{P}_1(i), \mathcal{P}_2(i)$ as defined in (3.21).
9: Obtain: $\widehat{\partial}_i g(\mathbf{x}^* + \mu_1 \mathbf{v}(i))$. If $\frac{|\widehat{\partial}_i g(\mathbf{x}^* + \mu_1 \mathbf{v}(i)) - \widehat{\partial}_i g(\mathbf{x}^*)|}{\mu_1} > \tau'$, then $\mathcal{R} \leftarrow \mathcal{P}_1(i)$ else $\mathcal{R} \leftarrow \mathcal{P}_2(i)$.
10: **end while**
11: Denote $\mathcal{R} = \{j\}$. If $i < j$ then $\widehat{\mathcal{S}_2} = \widehat{\mathcal{S}_2} \cup \{(i,j)\}$, else $\widehat{\mathcal{S}_2} = \widehat{\mathcal{S}_2} \cup \{(j,i)\}$.
12: **end while**

Query complexity. Note that for each $i \in \mathcal{S}_1$ we make at most $4{m_x'}^2$ queries. This is clear from Step 4: four queries are made for estimating the two partial derivatives and this is done at most ${m_x'}^2$ times. If $i \in \mathcal{S}_2^{\text{var}}$, then we notice that in Step 9, we make two queries for each $\mathbf{v}(i)$ leading to at most $2\lceil \log k \rceil$ queries during Steps 8–9. In addition, we still make at most $4{m_x'}^2$ queries during Step 4, as discussed earlier. Hence the total number of queries made is at most:

$$k_1 \cdot 4{m_x'}^2 + k_2 \cdot \left(4{m_x'}^2 + 2\lceil \log k \rceil\right) < k(4{m_x'}^2 + 2\lceil \log k \rceil). \tag{3.23}$$

Since $m_x' \geq \lambda_2^{-1}$, the query complexity for this phase is $O(k(\lambda_2^{-2} + \log k))$.

Computational complexity. It is clear that the overall computation time is linear in the the number of queries and hence at most polynomial in k.

3.2 Analysis for noisy setting

We now analyse the noisy setting where at each query $\mathbf{x}$, we observe: $f(\mathbf{x}) + z'$, where $z' \in \mathbb{R}$ denotes external noise. In order to see how this affects Algorithm 1, (3.6) now changes to $\mathbf{y} = \mathbf{V}\nabla f(\mathbf{x}) + \mathbf{n} + \mathbf{z}$, where $z_j = (z'_{j,1} - z'_{j,2})/(2\mu)$. Therefore while the Taylor's remainder term $|n_j| = O(\mu^2)$, the external noise term $|z_j|$ scales as μ^{-1}. Hence in contrast to Lemma 1 the step-size μ needs to be chosen carefully now – a value which is too small would blow up the external noise component while a large value would increase perturbation due to higher order Taylor's terms.

A similar problem would occur in the next phase when we try to identify $\mathcal{S}_1, \mathcal{S}_2$. Indeed, due to the introduction of noise, we now observe $g(\mathbf{x} + \beta \mathbf{e}_1(i)) + z'_{i,1}$, $g(\mathbf{x} - \beta \mathbf{e}_1(i)) + z'_{i,2}$. This changes the expression for $\widehat{\partial}_i g(\mathbf{x})$ in (3.17) to: $\widehat{\partial}_i g(\mathbf{x}) = \partial_i g(\mathbf{x}) + \eta_i(\mathbf{x}, \beta) + z_i(\mathbf{x}, \beta)$ where $z_i(\mathbf{x}, \beta) = (z'_{i,1} - z'_{i,2})/(2\beta)$. Recall that $\eta_i(\mathbf{x}, \beta) = O(\beta^2)$ corresponds to the Taylor's remainder term. Hence we again see that in contrast to Lemma 2, the step β cannot be chosen too small now, as it would blow up the external noise component.

Arbitrary bounded noise. In this scenario, we assume the external noise to be arbitrary and bounded, meaning that $|z'| < \varepsilon$, for some finite $\varepsilon \geq 0$. Clearly, if ε is too large, then we would expect recovery of $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$ to be impossible, as the structure of f would be destroyed. However we show that if $\varepsilon = O(\frac{D_1^{3/2}}{\sqrt{B_3 k}})$, then Algorithm 1 recovers the total support $\mathcal{S}$, with appropriate choice of sampling parameters. Furthermore, assuming $\mathcal{S}$ is recovered exactly, and provided ε additionally satisfies $\varepsilon = O(\frac{D_2^3}{B_3^2})$, then with proper choice of sampling parameters, Algorithm 2 identifies $\mathcal{S}_1, \mathcal{S}_2$. This is stated formally in the following Theorem.

Theorem 2. *Let the constants c_3', C, c_1', C_1 and $\mathcal{H}_2^d, m_x, m_v$ be as defined in Lemma 1. Say $\varepsilon < \varepsilon_1 = \frac{D_1^{3/2}}{3C\sqrt{4B_3 kC}}$. Then for $\theta_1 = \cos^{-1}(-\varepsilon/\varepsilon_1)$, let μ be chosen to satisfy:*

$$\mu \in \left(2\sqrt{\frac{D_1 m_v}{4B_3 k}}\cos(\theta_1/3 - 2\pi/3), 2\sqrt{\frac{D_1 m_v}{4B_3 k}}\cos(\theta_1/3)\right) \tag{3.24}$$

We then have in Algorithm 1 for the choice: $\tau = C\left(\frac{2\mu^2 B_3 k}{3m_v} + \frac{\varepsilon\sqrt{m_v}}{\mu}\right)$ *that* $\widehat{\mathcal{S}} = \mathcal{S}$ *holds with probability at least* $1 - e^{-c_1' m_v} - e^{-\sqrt{m_v d}} - d^{-2C_1}$. *Given that* $\widehat{\mathcal{S}} = \mathcal{S}$, *let* m_x' *be as defined in Lemma 2. Assuming* $\varepsilon < \frac{D_2^3}{384\sqrt{2}B_3^2} = \varepsilon_2$ *holds, then for* $\theta_2 = \cos^{-1}(-\varepsilon/\varepsilon_2)$ *let* β, μ_1 *be chosen to satisfy:*

$$\mu_1 \in \left(\frac{D_2 - \sqrt{D_2^2 - \frac{32}{3\beta}B_3(\beta^3 B_3 + 6\varepsilon)}}{8B_3}, \frac{D_2 + \sqrt{D_2^2 - \frac{32}{3\beta}B_3(\beta^3 B_3 + 6\varepsilon)}}{8B_3}\right), \tag{3.25}$$

$$\beta \in \left(\frac{D_2}{2\sqrt{2}B_3}\cos(\theta_2/3 - 2\pi/3), \frac{D_2}{2\sqrt{2}B_3}\cos(\theta_2/3)\right). \tag{3.26}$$

Then the choice $\tau' = \frac{\beta^2 B_3}{3\mu_1} + 2\mu_1 B_3 + \frac{2\varepsilon}{\beta\mu_1}$ *implies in Algorithm 2 that* $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ *and* $\widehat{\mathcal{S}_2} = \mathcal{S}_2$.

Stochastic noise. We now assume that the point queries are corrupted with i.i.d Gaussian noise, so that $z' \sim \mathcal{N}(0, \sigma^2)$ for $\sigma^2 < \infty$. In order to reduce σ, we consider resampling each point query a sufficient number of times, and averaging the values. In Algorithm 1, *i.e.*, during the estimation of $\mathcal{S}$, we resample each query N_1 times so that $z' \sim \mathcal{N}(0, \sigma^2/N_1)$. For any $\varepsilon > 0$, if N_1 is chosen large enough, then we can obtain a uniform bound $|z'| < \varepsilon$ – via standard tail bounds for Gaussian's – over all noise samples, with high probability. Consequently, the noise model transforms to a bounded noise one which means that by choosing $\varepsilon < \varepsilon_1$, we can use the result of Theorem 2 for estimating $\mathcal{S}$. Similarly in Algorithm 2, we resample each query N_2 times so that now $z' \sim \mathcal{N}(0, \sigma^2/N_2)$. For any $\varepsilon' > 0$, and N_2 large enough, we can again uniformly bound $|z'| < \varepsilon'$ with high probability. By now choosing $\varepsilon' < \varepsilon_2$, we can then use the result of Theorem 2 for estimating $\mathcal{S}_1, \mathcal{S}_2$. These conditions are stated formally in the following Theorem.

Theorem 3. *Let the constants* c_3', C, c_1', C_1 *and* $\mathcal{H}_2^d, m_x, m_v$ *be as defined in Lemma 1. For any* $\varepsilon < \varepsilon_1 = \frac{D_1^{3/2}}{3C\sqrt{4B_3 kC}}$, $0 < p_1 < 1$, $\theta_1 = \cos^{-1}(-\varepsilon/\varepsilon_1)$, *say we resample each query in Algorithm 1,* $N_1 > \frac{\sigma^2}{\varepsilon^2}\log(\frac{2}{p_1}m_v(2m_x+1)^2|\mathcal{H}_2^d|)$ *times, and average the values. Then by choosing* μ *and* τ *as in Theorem 2, we have that* $\widehat{\mathcal{S}} = \mathcal{S}$ *holds with probability at least* $1 - p_1 - e^{-c_1' m_v} - e^{-\sqrt{m_v d}} - d^{-2C_1}$.

Given that $\widehat{\mathcal{S}} = \mathcal{S}$, *let* m_x' *be as defined in Lemma 2. For any* $\varepsilon' < \frac{D_2^3}{384\sqrt{2}B_3^2} = \varepsilon_2$, $0 < p_2 < 1$, $\theta_2 = \cos^{-1}(-\varepsilon'/\varepsilon_2)$, *say we resample each query in Algorithm 2,* $N_2 > \frac{\sigma^2}{\varepsilon'^2}\log\left(\frac{2}{p_2}(k(2{m_x'}^2 + \lceil\log k\rceil))\right)$ *times. Then by choosing* β, μ_1, τ' *as in Theorem 2, we have that* $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ *and* $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, *with probability at least* $1 - p_2$.

We now analyze the query complexity for the i.i.d Gaussian noise case. One can verify that $\varepsilon_1 = O(k^{-1/2})$. Since $m_v = O(k\log d)$, $|\mathcal{H}_2^d| = O(\log d), m_x = O(\lambda_1^{-1})$, then by choosing $p_1 = O(d^{-\delta})$ for any constant $\delta > 0$, we arrive at $N_1 = O(k\log((d^\delta)(k\log d)(\lambda_1^{-2}\log d))) = O(k\log d)$. This leads to a total sample complexity of $O(N_1 k(\log d)^2\lambda_1^{-2}) = O(k^2(\log d)^3\lambda_1^{-2})$ for guaranteeing $\widehat{\mathcal{S}} = \mathcal{S}$, with high probability. Next, we see that $\varepsilon' = O(1)$ and thus $N_2 = O(\log(k(\lambda_2^{-2} + \log k)/p_2))$. Therefore with an additional $O(N_2 k(\lambda_2^{-2} + \log k)) = O(k(\lambda_2^{-2} + \log k)\log(k/p_2))$ samples, we are guaranteed with probability at least $1 - p_2$ that $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ and $\widehat{\mathcal{S}_2} = \mathcal{S}_2$.

4 Sampling scheme for the general overlap case

We now analyze the general scenario where overlaps can occur amongst the elements of $\mathcal{S}_2$. Therefore the degrees of the variables occurring in $\mathcal{S}_2^{\text{var}}$, can be greater than one. Contrary to the non-overlap case, we now sample f in order to directly estimate its $d \times d$ Hessian $\nabla^2 f$, at suitably chosen points. In particular, this enables us to subsequently identify $\mathcal{S}_2$. Once $\mathcal{S}_2$ is identified, we are left with a SPAM – with no variable interactions – on the set $[d] \setminus \mathcal{S}_2$. We then identify $\mathcal{S}_1$ by employing the sampling scheme from [54] on this reduced space.

4.1 Analysis for noiseless setting

In this section, we consider the noiseless scenario, *i.e.*, we assume the exact sample $f(\mathbf{x})$ is obtained for any query $\mathbf{x}$. To begin with, we explain why the sampling scheme for the non overlap case does not directly apply here. To this end, note that the gradient of f has the following structure for each $q \in [d]$.

$$(\nabla f(\mathbf{x}))_q = \begin{cases} \partial_q\phi_q(x_q) & ; \quad q \in \mathcal{S}_1 \\ \partial_q\phi_{(q,q')}(x_q, x_{q'}) & ; \quad (q,q') \in \mathcal{S}_2 \ \&\ \rho(q) = 1, \\ \partial_q\phi_{(q',q)}(x_{q'}, x_q) & ; \quad (q',q) \in \mathcal{S}_2 \ \&\ \rho(q) = 1, \\ \partial_q\phi_q(x_q) + \sum\limits_{(q,q')\in\mathcal{S}_2} \partial_q\phi_{(q,q')}(x_q, x_{q'}) & \\ \quad + \sum\limits_{(q',q)\in\mathcal{S}_2} \partial_q\phi_{(q',q)}(x_{q'}, x_q) & ; \quad q \in \mathcal{S}_2^{\text{var}} \ \&\ \rho(q) > 1, \\ 0 & ; \quad \text{otherwise.} \end{cases}$$

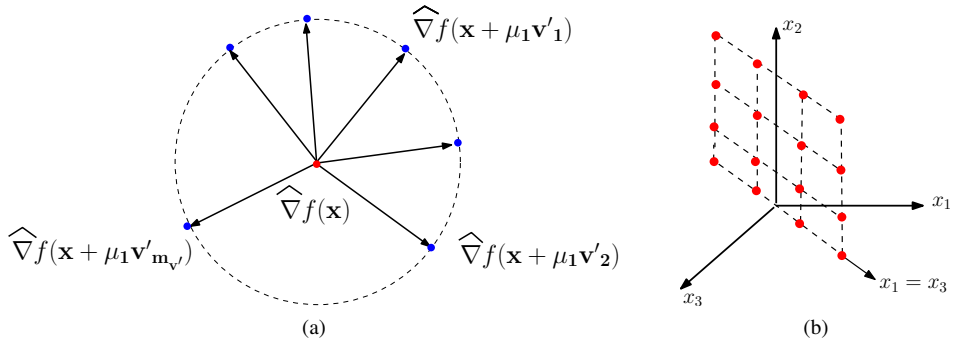


Figure 2: (a) $\nabla^2 f(\mathbf{x})$ estimated using: $\widehat{\nabla} f(\mathbf{x})$ (at red disk) and neighborhood gradient estimates (at blue disks) (b) Geometric picture: $d = 3$, $h \in \mathcal{H}_2^3$ with $h(1) = h(3) \neq h(2)$. Red disks are points in $\chi(h)$.

Therefore, for any $q \in \mathcal{S}_2^{\text{var}}$ with $\rho(q) > 1$, we notice that $(\nabla f(\mathbf{x}))_q$ is by itself the sum of $\rho(q)$ many bivariate functions, and $\partial_q \phi_q$. This causes an issue as far as identifying q – via estimating ∇f followed by thresholding – is concerned, as was done for the non-overlap case. While we assume the magnitudes of $\partial_q \phi_{(q,q')}$ to be sufficiently large within respective subsets of $[-1,1]^2$, it is not clear what that implies for $|(\nabla f(\mathbf{x}))_q|$. Note that $(\nabla f(\mathbf{x}))_q \not\equiv 0$ since q is an active variable. However a lower bound on: $|(\nabla f(\mathbf{x}))_q|$, and also on the measure of the interval where it is attained, appears to be non-trivial to obtain.

Estimating sparse Hessian matrices In light of the above discussion, we consider an alternative approach, wherein we directly estimate the Hessian $\nabla^2 f(\mathbf{x}) \in \mathbb{R}^{d\times d}$, at suitably chosen $\mathbf{x} \in [-1,1]^d$. Observe that $\nabla^2 f(\mathbf{x})$ has the following structure for $i \in \mathcal{S}_2^{\text{var}}$ and $j = 1, \ldots, d$:

$$(\nabla^2 f(\mathbf{x}))_{i,j} = \begin{cases} \partial_i^2 \phi_{(i,i')}(x_i, x_{i'}) & ; \quad \rho(i) = 1, (i,i') \in \mathcal{S}_2, i = j, \\ \partial_i^2 \phi_{(i',i)}(x_{i'}, x_i) & ; \quad \rho(i) = 1, (i',i) \in \mathcal{S}_2, i = j, \\ \partial_i^2 \phi_i(x_i) + \sum\limits_{(i,i')\in\mathcal{S}_2} \partial_i^2 \phi_{(i,i')}(x_i, x_{i'}) & \\ \quad + \sum\limits_{(i',i)\in\mathcal{S}_2} \partial_i^2 \phi_{(i',i)}(x_{i'}, x_i) & ; \quad \rho(i) > 1, i = j, \\ \partial_i \partial_j \phi_{(i,j)}(x_i, x_j) & ; \quad (i,j) \in \mathcal{S}_2, \\ \partial_i \partial_j \phi_{(j,i)}(x_j, x_i) & ; \quad (j,i) \in \mathcal{S}_2, \\ 0 & ; \quad \text{otherwise} \end{cases},$$

while if $i \in \mathcal{S}_1$, we have for $j = 1, \ldots, d$:

$$(\nabla^2 f(\mathbf{x}))_{i,j} = \begin{cases} \partial_i^2 \phi_i(x_i) & ; \quad i = j, \\ 0 & ; \quad \text{otherwise} \end{cases}.$$

The l^{th} row of $\nabla^2 f(\mathbf{x})$ can be denoted by $\nabla \partial_l f(\mathbf{x})^T \in \mathbb{R}^d$. If $l \in \mathcal{S}_1$, then $\nabla \partial_l f(\mathbf{x})^T$ has at most one non-zero entry, namely the l^{th} entry, and has all other entries equal to zero. In other words, $\nabla \partial_l f(\mathbf{x})^T$ is 1-sparse for $l \in \mathcal{S}_1$. If $l \in \mathcal{S}_2^{\text{var}}$, then we see that $\nabla \partial_l f(\mathbf{x})^T$ will have at most $(\rho(l)+1)$ non-zero entries, implying that it is $(\rho(l)+1) \leq (\rho_m + 1)$-sparse.

At suitably chosen $\mathbf{x}$'s, our aim specifically is to detect the non-zero off diagonal entries of $\nabla^2 f(\mathbf{x})$ since they correspond precisely to $\mathcal{S}_2$. To this end, we consider the "difference of gradients" based approach used in Section 3.1.2. Contrary to the setting in Section 3.1.2 however, we now have a $d \times d$ Hessian and have *no knowledge* about the set of active variables: $\mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$. Therefore, the Hessian estimation problem is harder now, and requires a different sampling scheme.

Sampling scheme for estimating $\mathcal{S}_2$. For $\mathbf{x}, \mathbf{v}' \in \mathbb{R}^d$, $\mu_1 > 0$, consider the Taylor expansion of ∇f at $\mathbf{x}$ along $\mathbf{v}'$, with step size μ_1. For $\zeta_i = \mathbf{x} + \theta_i \mathbf{v}'$, for some $\theta_i \in (0, \mu_1)$; $i = 1, \ldots, d$, we obtain the following identity.

$$\frac{\nabla f(\mathbf{x}+\mu_1\mathbf{v}') - \nabla f(\mathbf{x})}{\mu_1} = \nabla^2 f(\mathbf{x})\mathbf{v}' + \frac{\mu_1}{2}\begin{pmatrix} \mathbf{v}'^T \nabla^2 \partial_1 f(\zeta_1)\mathbf{v}' \\ \vdots \\ \mathbf{v}'^T \nabla^2 \partial_d f(\zeta_d)\mathbf{v}' \end{pmatrix} = \begin{pmatrix} \langle \nabla \partial_1 f(\mathbf{x}), \mathbf{v}' \rangle \\ \vdots \\ \langle \nabla \partial_d f(\mathbf{x}), \mathbf{v}' \rangle \end{pmatrix} + \frac{\mu_1}{2}\begin{pmatrix} \mathbf{v}'^T \nabla^2 \partial_1 f(\zeta_1)\mathbf{v}' \\ \vdots \\ \mathbf{v}'^T \nabla^2 \partial_d f(\zeta_d)\mathbf{v}' \end{pmatrix}. \quad (4.1)$$

We see from (4.1) that the l^{th} entry of $(\nabla f(\mathbf{x}+\mu_1\mathbf{v}') - \nabla f(\mathbf{x}))/\mu_1$, corresponds to a linear measurement of the l^{th} row of $\nabla^2 f(\mathbf{x})$ with $\mathbf{v}'$. From the preceding discussion, we also know that each row of $\nabla^2 f(\mathbf{x})$ is at most $(\rho_m + 1)$-sparse. This suggests the following idea: for any $\mathbf{x}$, if we obtain sufficiently many linear measurements of each row of $\nabla^2 f(\mathbf{x})$, then we can *estimate each row separately* via ℓ_1 minimization. To this end, we first need an efficient way for estimating $\nabla f(\mathbf{x}) \in \mathbb{R}^d$, at any point $\mathbf{x}$. Note that $\nabla f(\mathbf{x})$ is k-sparse, therefore we can estimate it via the randomized scheme, explained in Section 3.1.1, with

$O(k \log d)$ queries of f. This gives us: $\widehat{\nabla} f(\mathbf{x}) = \nabla f(\mathbf{x}) + \mathbf{w}(\mathbf{x})$, where $\mathbf{w}(\mathbf{x}) \in \mathbb{R}^d$ denotes the estimation noise. Plugging this in (4.1) results in the following identity.

$$\frac{\widehat{\nabla} f(\mathbf{x} + \mu_1 \mathbf{v}') - \widehat{\nabla} f(\mathbf{x})}{\mu_1} = \begin{pmatrix} \langle \nabla \partial_1 f(\mathbf{x}), \mathbf{v}' \rangle \\ \vdots \\ \langle \nabla \partial_d f(\mathbf{x}), \mathbf{v}' \rangle \end{pmatrix} + \underbrace{\frac{\mu_1}{2} \begin{pmatrix} \mathbf{v}'^T \nabla^2 \partial_1 f(\zeta_1) \mathbf{v}' \\ \vdots \\ \mathbf{v}'^T \nabla^2 \partial_d f(\zeta_d) \mathbf{v}' \end{pmatrix} + \frac{\mathbf{w}(\mathbf{x} + \mu_1 \mathbf{v}') - \mathbf{w}(\mathbf{x})}{\mu_1}}_{\text{“Noise”}}. \tag{4.2}$$

Now let $\mathbf{v}'$ be chosen from the set:

$$\mathcal{V}' := \left\{ \mathbf{v}'_j \in \mathbb{R}^d : v'_{j,q} = \pm \frac{1}{\sqrt{m_{v'}}} \text{ w.p. } 1/2 \text{ each; } j = 1, \ldots, m_{v'} \text{ and } q = 1, \ldots, d \right\}. \tag{4.3}$$

Then, employing (4.2) at each $\mathbf{v}'_j \in \mathcal{V}'$, and denoting $\mathbf{V}' = [\mathbf{v}'_1 \ldots \mathbf{v}'_{m_{v'}}]^T \in \mathbb{R}^{m_{v'} \times d}$, we obtain d linear systems for $q = 1, \ldots, d$:

$$\underbrace{\frac{1}{\mu_1} \begin{pmatrix} (\widehat{\nabla} f(\mathbf{x} + \mu_1 \mathbf{v}'_1) - \widehat{\nabla} f(\mathbf{x}))_q \\ \vdots \\ (\widehat{\nabla} f(\mathbf{x} + \mu_1 \mathbf{v}'_{m_{v'}}) - \widehat{\nabla} f(\mathbf{x}))_q \end{pmatrix}}_{\mathbf{y}_q} = \mathbf{V}' \nabla \partial_q f(\mathbf{x}) + \underbrace{\frac{\mu_1}{2} \begin{pmatrix} \mathbf{v}'^T_1 \nabla^2 \partial_q f(\zeta_1) \mathbf{v}'_1 \\ \vdots \\ \mathbf{v}'^T_{m_{v'}} \nabla^2 \partial_q f(\zeta_{m_{v'}}) \mathbf{v}'_{m_{v'}} \end{pmatrix}}_{\eta_{\mathbf{q},\mathbf{1}}} + \underbrace{\frac{1}{\mu_1} \begin{pmatrix} w_q(\mathbf{x} + \mu_1 \mathbf{v}'_1) - w_q(\mathbf{x}) \\ \vdots \\ w_q(\mathbf{x} + \mu_1 \mathbf{v}'_{m_{v'}}) - w_q(\mathbf{x}) \end{pmatrix}}_{\eta_{\mathbf{q},\mathbf{2}}}. \tag{4.4}$$

Given the measurement vector $\mathbf{y}_q$, we can obtain the estimate $\widehat{\nabla} \partial_q f(\mathbf{x})$ individually for each q, via ℓ_1 minimization:

$$\widehat{\nabla} \partial_q f(\mathbf{x}) := \underset{\mathbf{y}_q = \mathbf{V}' \mathbf{z}}{\operatorname{argmin}} \| \mathbf{z} \|_1; \quad q = 1, \ldots, d. \tag{4.5}$$

Hence, we have obtained an estimate $\widehat{\nabla^2} f(\mathbf{x}) := [\widehat{\nabla} \partial_1 f(\mathbf{x}) \cdots \widehat{\nabla} \partial_d f(\mathbf{x})]^T$ of the Hessian $\nabla^2 f(\mathbf{x})$, at the point $\mathbf{x}$. Next, we would like to have a suitable set of points $\mathbf{x}$, in the sense that it provides a sufficiently fine discretization, of any canonical 2-dimensional subspace of $\mathbb{R}^d$. To this end, we can simply consider the set χ as defined in (3.12), for the same reasons as before.

Sampling scheme for estimating $\mathcal{S}_1$. While the above sampling scheme enables us to recover $\mathcal{S}_2$, we can recover $\mathcal{S}_1$ as follows. Let $\widehat{\mathcal{S}_2^{\text{var}}}$ denote the set of variables in the estimated set $\widehat{\mathcal{S}_2}$, and let $\mathcal{P} := [d] \setminus \widehat{\mathcal{S}_2^{\text{var}}}$. Assuming $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, we have $\mathcal{S}_1 \subset \mathcal{P}$. Therefore the model we are left with now is a SPAM with no variable interactions on the *reduced* variable set $\mathcal{P}$. For identification of $\mathcal{S}_1$, we employ the sampling scheme of [54], wherein the gradient of f is estimated along a discrete set of points on the line: $\{(x, \ldots, x) \in \mathbb{R}^d : x \in [-1, 1]\}$. For some $m'_x \in \mathbb{Z}^+$, we denote this discrete set by:

$$\chi_{\text{diag}} := \left\{ \mathbf{x} = (x \; x \; \cdots \; x) \in \mathbb{R}^d : x \in \left\{ -1, -\frac{m'_x - 1}{m'_x}, \ldots, \frac{m'_x - 1}{m'_x}, 1 \right\} \right\}. \tag{4.6}$$

Note that $|\chi_{\text{diag}}| = 2m'_x + 1$. The motivation for estimating ∇f at $\mathbf{x} \in \chi_{\text{diag}}$ is that we obtain estimates of $\partial_p \phi_p$ at equispaced points within $[-1, 1]$, for $p \in \mathcal{S}_1$. With a sufficiently fine discretization, we would “hit” the critical regions associated with each $\partial_p \phi_p$, as defined in Assumption 3. By applying a thresholding operation, we would then be able to identify each $p \in \mathcal{S}_1$. Let us denote $\mathcal{V}''$ to be the set of sampling directions in $\mathbb{R}^d$ – analogous to $\mathcal{V}, \mathcal{V}'$ defined in (3.4), (4.3) respectively – with $|\mathcal{V}''| = m_{v''}$:

$$\mathcal{V}'' := \left\{ \mathbf{v}''_j \in \mathbb{R}^d : v''_{j,q} = \pm \frac{1}{\sqrt{m_{v''}}} \text{ w.p. } 1/2 \text{ each; } j = 1, \ldots, m_{v''} \text{ and } q = 1, \ldots, d \right\}. \tag{4.7}$$

For each $\mathbf{x} \in \chi_{\text{diag}}$, we will query f at points $(\mathbf{x} + \mu' \mathbf{v}''_j)_{\mathcal{P}}, (\mathbf{x} - \mu' \mathbf{v}''_j)_{\mathcal{P}}$; $\mathbf{v}''_j \in \mathcal{V}''$, restricted to $\mathcal{P}$. Then by obtaining the measurements: $y_j = (f((\mathbf{x} + \mu' \mathbf{v}''_j)_{\mathcal{P}}) - f((\mathbf{x} - \mu' \mathbf{v}''_j)_{\mathcal{P}}))/(2\mu')$; $j = 1, \ldots, m_{v''}$, and denoting $(\mathbf{V}'')_{\mathcal{P}} = [(\mathbf{v}''_1)_{\mathcal{P}} \cdots (\mathbf{v}''_{m_{v''}})_{\mathcal{P}}]^T$, we obtain the estimate $(\widehat{\nabla} f((\mathbf{x})_{\mathcal{P}}))_{\mathcal{P}} := \underset{\mathbf{y} = (\mathbf{V}'')_{\mathcal{P}} (\mathbf{z})_{\mathcal{P}}}{\operatorname{argmin}} \| (\mathbf{z})_{\mathcal{P}} \|_1$. This notation simply means that we search over $\mathbf{z} \in \mathbb{R}^{\mathcal{P}}$, to form the estimate $(\widehat{\nabla} f((\mathbf{x})_{\mathcal{P}}))_{\mathcal{P}}$.

The complete procedure for estimating $\mathcal{S}_1, \mathcal{S}_2$, is described formally in Algorithm 3. Next, we provide sufficient conditions on our sampling parameters that guarantee exact recovery of $\mathcal{S}_1, \mathcal{S}_2$ by the algorithm. This is stated in the following Theorem.

Theorem 4. *Let $\mathcal{H}_2^d$ be of size $|\mathcal{H}_2^d| \leq 2(C+1)e^2 \log d$ for some constant $C > 1$. Then $\exists$ constants $c'_1, c'_2 \geq 1$ and $C_1, C_2, c'_4, c'_5 > 0$, such that the following is true. Let $m_x, m_v, m_{v'}$ satisfy*

$$m_x \geq \lambda_2^{-1}, \quad c'_1 k \log\left(\frac{d}{k}\right) < m_v < \frac{d}{(\log 6)^2}, \quad c'_2 \rho_m \log\left(\frac{d}{\rho_m}\right) < m_{v'} < \frac{d}{(\log 6)^2}.$$

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

Input: $m_v, m_{v'}, m_x, m'_x \in \mathbb{Z}^+$; $\mu, \mu_1, \mu' > 0$; $\tau' > 0, \tau'' > 0$.
Initialization: $\widehat{\mathcal{S}_1}, \widehat{\mathcal{S}_2} = \emptyset$.
Output: Estimates $\widehat{\mathcal{S}_2}$, $\widehat{\mathcal{S}_1}$.

Construct $(d, 2)$-hash family $\mathcal{H}_2^d$ and sets $\mathcal{V}, \mathcal{V}'$.
for $h \in \mathcal{H}_2^d$ **do**
  Construct the set $\chi(h)$.
  for $i = 1, \ldots, (2m_x + 1)^2$ and $\mathbf{x}_i \in \chi(h)$ **do**
    $(\mathbf{y}_i)_j = \frac{f(\mathbf{x}_i + \mu \mathbf{v}_j) - f(\mathbf{x}_i - \mu \mathbf{v}_j)}{2\mu}$; $j = 1, \ldots, m_v$; $\mathbf{v}_j \in \mathcal{V}$.
    $\widehat{\nabla} f(\mathbf{x}_i) := \underset{\mathbf{y}_i = \mathbf{V}\mathbf{z}}{\operatorname{argmin}} \parallel \mathbf{z} \parallel_1$.
    for $p = 1, \ldots, m_{v'}$ **do**
      $(\mathbf{y}_{i,p})_j = \frac{f(\mathbf{x}_i + \mu_1 \mathbf{v}'_p + \mu \mathbf{v}_j) - f(\mathbf{x}_i + \mu_1 \mathbf{v}'_p - \mu \mathbf{v}_j)}{2\mu}$; $j = 1, \ldots, m_v$; $\mathbf{v}'_p \in \mathcal{V}'$. // ESTIMATION OF $\mathcal{S}_2$
      $\widehat{\nabla} f(\mathbf{x}_i + \mu_1 \mathbf{v}'_p) := \underset{\mathbf{y}_{i,p} = \mathbf{V}\mathbf{z}}{\operatorname{argmin}} \parallel \mathbf{z} \parallel_1$.
    end for
    for $q = 1, \ldots, d$ **do**
      $(\mathbf{y}_q)_j = \frac{(\widehat{\nabla} f(\mathbf{x}_i + \mu_1 \mathbf{v}'_j) - \widehat{\nabla} f(\mathbf{x}_i))_q}{\mu_1}$; $j = 1, \ldots, m_{v'}$.
      $\widehat{\nabla} \partial_q f(\mathbf{x}_i) := \underset{\mathbf{y}_q = \mathbf{V}'\mathbf{z}}{\operatorname{argmin}} \parallel \mathbf{z} \parallel_1$.
      $\widehat{\mathcal{S}_2} = \widehat{\mathcal{S}_2} \cup \left\{ (q, q') : q' \in \{q+1, \ldots, d\} \ \& \ |(\widehat{\nabla} \partial_q f(\mathbf{x}_i))_{q'}| > \tau' \right\}$.
    end for
  end for
end for

Construct the sets $\chi_{\text{diag}}, \mathcal{V}''$ and initialize $\mathcal{P} := [d] \setminus \widehat{\mathcal{S}_2^{\text{var}}}$.
for $i = 1, \ldots, (2m'_x + 1)$ and $\mathbf{x}_i \in \chi_{\text{diag}}$ **do**
  $(\mathbf{y}_i)_j = \frac{f((\mathbf{x}_i + \mu' \mathbf{v}''_j)_{\mathcal{P}}) - f((\mathbf{x}_i - \mu' \mathbf{v}''_j)_{\mathcal{P}})}{2\mu'}$; $j = 1, \ldots, m_{v''}$; $\mathbf{v}_j \in \mathcal{V}''$.
  $(\widehat{\nabla} f((\mathbf{x}_i)_{\mathcal{P}}))_{\mathcal{P}} := \underset{\mathbf{y}_i = (\mathbf{V}'')_{\mathcal{P}}(\mathbf{z})_{\mathcal{P}}}{\operatorname{argmin}} \parallel (\mathbf{z})_{\mathcal{P}} \parallel_1$. // ESTIMATION OF $\mathcal{S}_1$
  $\widehat{\mathcal{S}_1} = \widehat{\mathcal{S}_1} \cup \left\{ q \in \mathcal{P} : |((\widehat{\nabla} f((\mathbf{x}_i)_{\mathcal{P}})_q| > \tau'' \right\}$.
end for

Denoting $a = \frac{(4\rho_m + 1)B_3}{2\sqrt{m_{v'}}}$, $b = \frac{C_1 \sqrt{m_{v'}}((4\rho_m + 1)k)B_3}{3m_v}$, *let* μ, μ_1 *satisfy*

$$\mu^2 < \frac{D_2^2}{16abC_2^2},\ \mu_1 \in \left((D_2/(4aC_2)) - \sqrt{(D_2/(4aC_2))^2 - (b\mu^2/a)}, (D_2/(4aC_2)) + \sqrt{(D_2/(4aC_2))^2 - (b\mu^2/a)} \right).$$

We then have that the choice

$$\tau' = C_2 \left(\frac{\mu_1 (4\rho_m + 1)B_3}{2\sqrt{m_{v'}}} + \frac{C_1 \sqrt{m_{v'}} \mu^2 ((4\rho_m + 1)k)B_3}{3\mu_1 m_v} \right),$$

implies $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ *with probability at least* $1 - e^{-c'_4 m_v} - e^{-\sqrt{m_v d}} - e^{-c'_5 m_{v'}} - e^{-\sqrt{m_{v'} d}} - d^{-2C}$.

Given that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, *then* $\exists$ *constants* $c'_3 \geq 1$ *and* $C_3, c'_6 > 0$, *such that for* $m'_x, m_{v''}, \mu'$ *satisfying*

$$m'_x \geq \lambda_1^{-1}, \quad c'_3 (k - |\widehat{\mathcal{S}_2^{var}}|) \log \left(\frac{|\mathcal{P}|}{k - |\widehat{\mathcal{S}_2^{var}}|} \right) < m_{v''} < \frac{|\mathcal{P}|}{(\log 6)^2}, \quad {\mu'}^2 < \frac{3 m_{v''} D_1}{C_3 (k - |\widehat{\mathcal{S}_2^{var}}|) B_3},$$

the choice: $\tau'' = \frac{C_3 (k - |\widehat{\mathcal{S}_2^{var}}|) {\mu'}^2 B_3}{6 m_{v''}}$, *implies* $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ *with probability at least* $1 - e^{-c'_6 m_{v''}} - e^{-\sqrt{m_{v''} |\mathcal{P}|}}$.

Query complexity. Estimating $\nabla f(\mathbf{x})$ at some fixed $\mathbf{x}$ requires $2m_v = O(k \log d)$ queries. Estimating $\nabla^2 f(\mathbf{x})$ involves the estimation of $\nabla f(\mathbf{x})$ – along with an additional $m_{v'}$ gradient vectors in a neighborhood of $\mathbf{x}$ – implying $O(m_v m_{v'}) = O(k \rho_m (\log d)^2)$ point queries of f. Since $\nabla^2 f$ is estimated at all points in χ in the worst case, this consequently implies a total query complexity of $O(k \rho_m (\log d)^2 |\chi|) = O(\lambda_2^{-2} k \rho_m (\log d)^3)$, for estimating $\mathcal{S}_2$. We make an additional $O(\lambda_1^{-1}(k - |\widehat{\mathcal{S}_2^{\text{var}}}|) \log(d - |\widehat{\mathcal{S}_2^{\text{var}}}|))$ queries of f, in order to estimate $\mathcal{S}_1$. Therefore, the overall query complexity for estimating $\mathcal{S}_1, \mathcal{S}_2$ is $O(\lambda_2^{-2} k \rho_m (\log d)^3)$.

Computational complexity. The family $\mathcal{H}_2^d$ can be constructed[8] in time polynomial in d. For each $\mathbf{x} \in \chi$, we first solve $m_{v'}+1$ linear programs in $O(d)$ variables (Steps 10, 13), each solvable in time polynomial in (m_v, d). We then solve d linear programs in $O(d)$ variables (Step 17), each of which takes time polynomial in $(m_{v'}, d)$. Since this is done at $|\chi| = O(\lambda_2^{-2}\log d)$ many points, hence the overall computation time for estimation of $\mathcal{S}_2$ (and subsequently $\mathcal{S}_1$) is polynomial in the number of queries, and in d.

Remark 7. *In Algorithm 4, we could have optimized the procedure for identifying $\mathcal{S}_1$ as follows. Observe that for each $h \in \mathcal{H}_2^d$, we always have a subset of points (i.e., $\subset \chi(h)$) that discretize $\{(x,\dots,x) \in \mathbb{R}^d : x \in [-1,1]\}$. Therefore for each $\mathbf{x}$ lying in this subset, we could go through $\widehat{\nabla} f(\mathbf{x})$, and check via a thresholding operation, whether there exists a variables(s) in $\mathcal{S}_1$. If m_x is large enough ($\geq \lambda_1^{-1}$), then it would also enable us to recover $\mathcal{S}_1$ completely. A downside of this approach is that we would require additional, stronger conditions on the step size parameter μ to guarantee identification of $\mathcal{S}_1$. Since the estimation procedure for $\mathcal{S}_1$ in Algorithm 3 comes at the same order-wise sampling cost, therefore we choose to query f again, in order to identify $\mathcal{S}_1$.*

Remark 8. *We also note that the condition on μ' is less strict than in [54] for identifying $\mathcal{S}_1$. This is because in [54], the gradient is estimated via a forward difference procedure, while we perform a central difference procedure in* (3.3).

4.2 Analysis for noisy setting

We now consider the case where at each query $\mathbf{x}$, we observe $f(\mathbf{x})+z'$, with $z' \in \mathbb{R}$ denoting external noise. In order to estimate $\nabla f(\mathbf{x})$, we obtain the samples : $f(\mathbf{x}+\mu\mathbf{v}_j)+z'_{j,1}$ and $f(\mathbf{x}-\mu\mathbf{v}_j)+z'_{j,2}$; $j=1,\dots,m_v$. This changes (3.6) to the linear system $\mathbf{y} = \mathbf{V}\nabla f(\mathbf{x}) + \mathbf{n} + \mathbf{z}$, where $z_j = (z'_{j,1} - z'_{j,2})/(2\mu)$.

Arbitrary bounded noise. In this scenario, we assume the external noise to be arbitrary and bounded, meaning that $|z'| < \varepsilon$, for some finite $\varepsilon \geq 0$. Theorem 5 shows that Algorithm 3 recovers $\mathcal{S}_1, \mathcal{S}_2$ with appropriate choice of sampling parameters, provided ε is not too large.

Theorem 5. *Assuming the notation in Theorem 4, let $a, b, m_x, m_v, m_{v'}, \mathcal{H}_2^d$ be as defined in Theorem 4. Say $\varepsilon < \varepsilon_1 = \frac{D_2^3}{192\sqrt{3}C_1C_2^3\sqrt{a^3 b m_{v'} m_v}}$. Then for $\theta_1 = \cos^{-1}(-\varepsilon/\varepsilon_1)$, let μ, μ_1 satisfy:*

$$\mu \in \left(\sqrt{\frac{D_2^2}{12abC_2^2}}\cos(\theta_1/3 - 2\pi/3), \sqrt{\frac{D_2^2}{12abC_2^2}}\cos(\theta_1/3) \right), \tag{4.8}$$

$$\mu_1 \in \left(\frac{D_2}{4aC_2} - \sqrt{\left(\frac{D_2}{4aC_2}\right)^2 - \left(\frac{b\mu^2 + 2C_1\sqrt{m_v m_{v'}}\varepsilon}{a}\right)}, \frac{D_2}{4aC_2} + \sqrt{\left(\frac{D_2}{4aC_2}\right)^2 - \left(\frac{b\mu^2 + 2C_1\sqrt{m_v m_{v'}}\varepsilon}{a}\right)} \right). \tag{4.9}$$

We then have in Algorithm 3 for the choice

$$\tau' = C_2\left(\frac{\mu_1(4\rho_m+1)B_3}{2\sqrt{m_{v'}}} + \frac{C_1\sqrt{m_{v'}}\mu^2((4\rho_m+1)k)B_3}{3\mu_1 m_v} + \frac{2C_1\varepsilon\sqrt{m_{v'}m_v}}{\mu\mu_1} \right), \tag{4.10}$$

that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ with probability at least $1 - e^{-c'_4 m_v} - e^{-\sqrt{m_v d}} - e^{-c'_5 m_{v'}} - e^{-\sqrt{m_{v'} d}} - d^{-2C}$. Given that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, let $m'_x, m_{v''}$ be chosen as in Theorem 4. Let $a_1 = \frac{(k-|\widehat{\mathcal{S}_2^{var}}|)B_3}{6m_{v''}}$, $b_1 = \sqrt{m_{v''}}$ and assume $\varepsilon < \varepsilon_2 = \frac{D_1^{3/2}}{3\sqrt{6a_1C_3^3 b_1^2}}$. For $\theta_2 = \cos^{-1}(-\varepsilon/\varepsilon_2)$, let $\mu' \in (2\sqrt{D_1/(6a_1C_3)}\cos(\theta_2/3 - 2\pi/3), 2\sqrt{D_1/(6a_1C_3)}\cos(\theta_2/3))$. We then have in Algorithm 3 for the choice $\tau'' = C_3\left(\frac{(k-|\widehat{\mathcal{S}_2^{var}}|)\mu'^2 B_3}{6m_{v''}} + \frac{b_1\varepsilon}{\mu}\right)$ that $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ with probability at least $1 - e^{-c'_6 m_{v''}} - e^{-\sqrt{m_{v''}|\mathcal{P}|}}$.

We see that in contrast to Theorem 4, the step sizes: μ, μ' cannot be chosen too small now, on account of external noise. Also note that the parameters $\pi/2 \leq \theta_1, \theta_2 \leq \pi$ arising due to ε, affect the size of the intervals from which μ, μ' can be chosen respectively. One can verify that plugging $\varepsilon = 0$ in Theorem 5 (implying $\theta_1, \theta_2 = \pi/2$), gives us the sampling conditions of Theorem 4.

Stochastic noise. We now consider i.i.d Gaussian noise, so that $z' \sim \mathcal{N}(0, \sigma^2)$ for variance $\sigma^2 < \infty$. As in Section 3.2, we resample each point query a sufficient number of times and average, in order to reduce σ. Doing this N_1 times in Steps 9,12, and N_2 times in Step 25, for N_1, N_2 large enough, we can recover $\mathcal{S}_1, \mathcal{S}_2$ as shown formally in the following theorem.

Theorem 6. *Assuming the notation in Theorem 4, let $a, b, m_x, m_v, m_{v'}, \mathcal{H}_2^d$ be as defined in Theorem 4. For any $\varepsilon < \varepsilon_1 = \frac{D_2^3}{192\sqrt{3}C_1C_2^3\sqrt{a^3 b m_{v'} m_v}}$, $0 < p_1 < 1$ and $\theta_1 = \cos^{-1}(-\varepsilon/\varepsilon_1)$, say we resample each query in Steps 9-12 of Algorithm 3, $N_1 >$*

[8] Recall discussion following Definition 1.

$\frac{\sigma^2}{\varepsilon^2}\log(\frac{2}{p_1}m_v(m_{v'}+1)(2m_x+1)^2|\mathcal{H}_2^d|)$ *times, and average the values. Let* μ, μ_1, τ' *be chosen to satisfy* (4.8), (4.9) *and* (4.10) *respectively. We then have in Algorithm 3, that* $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ *with probability* $1-p_1-e^{-c_4'm_v}-e^{-\sqrt{m_v d}}-e^{-c_5'm_{v'}}-e^{-\sqrt{m_{v'}d}}-d^{-2C}$.

Given that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, *let* $m_x', m_{v''}, a_1, b_1$ *be as stated in Theorem 5. For any* $\varepsilon' < \varepsilon_2 = \frac{D_1^{3/2}}{\sqrt{6a_1C_3^3b_1^2}}$, $0 < p_2 < 1$, *and* $\theta_2 = \cos^{-1}(-\varepsilon'/\varepsilon_2)$, *say we resample each query in Step 25 of Algorithm 3,* $N_2 > \frac{\sigma^2}{\varepsilon'^2}\log(\frac{2(2m_x'+1)m_{v''}}{p_2})$ *times. Furthermore, let* μ', τ'' *be chosen as stated in Theorem 5. We then have in Algorithm 3 that* $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ *with probability at least* $1 - p_2 - e^{-c_6'm_{v''}} - e^{-\sqrt{m_{v''}|\mathcal{P}|}}$.

Query complexity. Let us analyze the query complexity when the noise is i.i.d Gaussian. For estimating $\mathcal{S}_2$, we have $\varepsilon = O(\rho_m^{-2}k^{-1/2})$. Furthermore: $(2m_x+1)^2 = \lambda_2^{-2}$, $|\mathcal{H}_2^d| = O(\log d)$, $m_v = O(k\log d)$ and $m_{v'} = O(\rho_m \log d)$. Choosing $p_1 = d^{-\delta}$ for any constant $\delta > 0$ gives us

$$N_1 = O(\rho_m^4 k \log((d^\delta)k\rho_m(\log d)^3)) = O(\rho_m^4 k \log d)$$

. This means that our total sample complexity for estimating $\mathcal{S}_2$ is:

$$O(N_1 k\rho_m(\log d)^2|\chi|) = O(\rho_m^5 k^2(\log d)^4\lambda_2^{-2}).$$

This ensures $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ with high probability. Next, for estimating $\mathcal{S}_1$, we have $\varepsilon' = O((k-|\mathcal{S}_2^{\text{var}}|)^{-1/2})$. Choosing $p_2 = ((d-|\mathcal{S}_2^{\text{var}}|)^{-\delta})$ for any constant $\delta > 0$, we get $N_2 = O((k-|\mathcal{S}_2^{\text{var}}|)\log(d-|\mathcal{S}_2^{\text{var}}|))$. This means the total sample complexity for estimating $\mathcal{S}_1$ is $O(N_2\lambda_1^{-1}(k-|\widehat{\mathcal{S}_2^{\text{var}}}|)\log(d-|\widehat{\mathcal{S}_2^{\text{var}}}|)) = O(\lambda_1^{-1}(k-|\widehat{\mathcal{S}_2^{\text{var}}}|)^2(\log(d-|\widehat{\mathcal{S}_2^{\text{var}}}|))^2)$. Putting it together, we have that in case of i.i.d Gaussian noise, the sampling complexity of Algorithm 3 for estimating $\mathcal{S}_1, \mathcal{S}_2$ is $O(\rho_m^5k^2(\log d)^4)$.

Remark 9. *We saw above that* $O(k^2(\log d)^2)$ *samples are sufficient for estimating* $\mathcal{S}_1$ *in presence of i.i.d Gaussian noise. This improves the corresponding bound in [54] by a* $O(k)$ *factor, and is due to the less strict condition on* μ' *(cf., Remark 8).*

5 Alternate sampling scheme for the general overlap case

We now derive an alternate algorithm for estimating the sets $\mathcal{S}_1, \mathcal{S}_2$, for the general overlap case. This algorithm differs from Algorithm 3 with respect to the scheme for estimating $\mathcal{S}_2$ – the procedure for estimating $\mathcal{S}_1$ is the same as Algorithm 3. In order to estimate $\mathcal{S}_2$, we now make use of recent results from CS, for recovering sparse symmetric matrices from few *linear measurements*. More precisely, we leverage these results for estimating the sparse Hessian $\nabla^2 f(\mathbf{x})$ at any fixed $\mathbf{x} \in \mathbb{R}^d$. This is in stark contrast to the approaches we proposed so far, wherein, each row of the Hessian $\nabla^2 f(\mathbf{x})$ was approximated separately. As we will show, this results in slightly improved sampling bounds for estimating $\mathcal{S}_2$ in the noiseless setting as opposed to those stated in Theorem 4.

5.1 Analysis for noiseless setting

We begin with the setting of noiseless point queries, and show how the problem of estimating $\nabla^2 f(\mathbf{x})$ at any $\mathbf{x} \in \mathbb{R}^d$ can be formulated as one of recovering an unknown sparse, symmetric matrix from linear measurements. To this end, first note that for $\mathbf{x}, \mathbf{v} \in \mathbb{R}^d$, step size $\mu > 0$, and $\zeta = \mathbf{x} + \theta\mathbf{v}$, $\zeta' = \mathbf{x} - \theta'\mathbf{v}$; $\theta, \theta' \in (0, 2\mu)$, one obtains via Taylor expansion of the C^3 smooth f, the following identity:

$$\frac{f(\mathbf{x}+2\mu\mathbf{v}) + f(\mathbf{x}-2\mu\mathbf{v}) - 2f(\mathbf{x})}{4\mu^2} = \mathbf{v}^T\nabla^2 f(\mathbf{x})\mathbf{v} + \underbrace{\frac{R_3(\zeta)+R_3(\zeta')}{4\mu^2}}_{O(\mu)}. \tag{5.1}$$

Here $R_3(\zeta), R_3(\zeta') = O(\mu^3)$ denote the third order Taylor terms. Importantly, (5.1) corresponds to a "noisy" *linear measurement* of $\nabla^2 f(\mathbf{x})$ *i.e.*, $\mathbf{v}^T\nabla f(\mathbf{x})\mathbf{v} = \langle \mathbf{v}\mathbf{v}^T, \nabla^2 f(\mathbf{x})\rangle$, via the measurement matrix $\mathbf{v}\mathbf{v}^T$. The noise arises on account of the Taylor remainder terms. We now present a recent result for recovering sparse symmetric matrices [7], that we leverage for estimating $\nabla^2 f(\mathbf{x})$.

Recovering sparse symmetric matrices via ℓ_1 minimization. Let $\mathbf{v}$ be composed of i.i.d sub-Gaussian entries with $v_i = a_i/\sqrt{m_v}$, and the a_i's drawn in an i.i.d manner from a distribution satisfying:

$$\mathbb{E}[a_i] = 0,\ \mathbb{E}[a_i^2] = 1 \text{ and } \mathbb{E}[a_i^4] > 1. \tag{5.2}$$

For concreteness, we will consider the following set whose elements clearly meet these moment conditions:

$$\mathcal{V} := \left\{ \mathbf{v}_j \in \mathbb{R}^d : v_{j,q} = \left\{ \begin{array}{rl} \pm\sqrt{\frac{3}{m_v}}; & \text{w.p } 1/6 \text{ each,} \\ 0; & \text{w.p } 2/3 \end{array} \right\}; \; j = 1, \ldots, m_v \text{ and } q = 1, \ldots, d \right\}. \tag{5.3}$$

Note that a symmetric Bernoulli distribution does not meet the aforementioned fourth order moment condition. Furthermore, let $\mathcal{M} : \mathbb{R}^{d \times d} \to \mathbb{R}^{m_v}$ denote a linear operator acting on square matrices, with

$$\mathcal{M}(\mathbf{H}) := [\langle \mathbf{v}_1 \mathbf{v}_1^T, \mathbf{H} \rangle \cdots \langle \mathbf{v}_{m_v} \mathbf{v}_{m_v}^T, \mathbf{H} \rangle]^T; \quad \mathbf{H} \in \mathbb{R}^{d \times d}. \tag{5.4}$$

For an unknown symmetric matrix $\mathbf{H}_0 \in \mathbb{R}^{d \times d}$, say we have at hand m_v linear measurements

$$\mathbf{y} = \mathcal{M}(\mathbf{H}_0) + \mathbf{n}; \quad \mathbf{y}, \mathbf{n} \in \mathbb{R}^{m_v}; \| \mathbf{n} \|_1 \leq \eta. \tag{5.5}$$

Then as shown in [7, Section C], we can recover an estimate $\widehat{\mathbf{H}_0}$ to $\mathbf{H}_0$ via ℓ_1 minimization, by solving:

$$\widehat{\mathbf{H}}_0 = \underset{\mathbf{H}}{\operatorname{argmin}} \| \mathbf{H} \|_1 \quad \text{s.t} \quad \mathbf{H}^T = \mathbf{H}, \quad \| \mathbf{y} - \mathcal{M}(\mathbf{H}) \|_1 \leq \eta. \tag{5.6}$$

Remark 10. (5.6) *was proposed in [7, Section C] for recovering sparse covariance matrices (which are positive semidefinite (PSD)) with the symmetry constraint replaced by a PSD constraint. However as noted in the discussion in [7, Section E], one can replace the PSD constraint by a symmetry constraint, in order to recover more general symmetric matrices (which are not necessarily PSD).*

Remark 11. *Note that* (5.6) *can be reformulated as a linear program in $O(d^2)$ variables, and hence can be solved efficiently up to arbitrary accuracy (using for instance, interior point methods (cf., [39])).*

The estimation property of (5.6) is captured in the following Theorem.

Theorem 7. *[7, Theorem 3] Consider the sampling model in* (5.5) *with $\mathbf{v}_i$'s satisfying* (5.2)*, and let $(\mathbf{H}_0)_\Omega$ denote the best K term approximation of $\mathbf{H}_0$. Then there exist constants $c_1, c_1', c_2, C_1, C_2 > 0$ such that with probability exceeding $1 - c_1 e^{-c_2 m_v}$, the solution $\widehat{\mathbf{H}}_0$ to* (5.6) *satisfies*

$$\| \widehat{\mathbf{H}}_0 - \mathbf{H}_0 \|_F \leq C_2 \frac{\| \mathbf{H}_0 - (\mathbf{H}_0)_\Omega \|_1}{\sqrt{K}} + C_1 \eta, \tag{5.7}$$

simultaneously for all (symmetric) $\mathbf{H}_0 \in \mathbb{R}^{d \times d}$, provided $m_v > c_1' K \log(d^2/K)$.

The proof of Theorem 7 relies on the ℓ_2/ℓ_1 Restricted Isometry Property (RIP) for sparse symmetric matrices, introduced by Chen et al. [7]:

Definition 2. *[7] For the set of symmetric K sparse matrices, the operator $\mathcal{B}$ is said to satisfy the ℓ_2/ℓ_1 Restricted Isometry Property (RIP) with constants $\gamma_1, \gamma_2 > 0$, if for all such matrices $\mathbf{X}$:*

$$(1 - \gamma_1) \| \mathbf{X} \|_F \leq \| \mathcal{B}(\mathbf{X}) \|_1 \leq (1 + \gamma_2) \| \mathbf{X} \|_F \, .$$

While the operator $\mathcal{M}$ defined in (5.4) does not satisfy ℓ_2/ℓ_1 RIP (since each $\mathbf{v}_i \mathbf{v}_i^T$ has non-zero mean), one could consider instead a set of debiased measurement matrices $\mathbf{B}_i := \mathbf{v}_{2i-1} \mathbf{v}_{2i-1}^T - \mathbf{v}_{2i} \mathbf{v}_{2i}^T$, with $\mathcal{B}_i(\mathbf{X}) := \langle \mathbf{B}_i, \mathbf{X} \rangle$ for $i = 1, \ldots, m$. Chen et al. [7, Corollary 2] then show that the linear map $\mathcal{B} : \mathbb{R}^{d \times d} \to \mathbb{R}^m$ satisfies ℓ_2/ℓ_1 RIP, for $\mathbf{v}_i$'s satisfying (5.2), provided $m > K \log(d^2/K)$.

Remark 12. *Observe that the ℓ_1 norm constraint in* (5.6) *arises due to the ℓ_2/ℓ_1 RIP in Definition 2. It is unclear whether the linear map $\mathcal{B}$ also satisfies the conventional ℓ_2/ℓ_2 RIP*[9]*. However assuming it were do so, the ℓ_1 norm constraint in* (5.6) *could then be replaced by $\| \mathbf{y} - \mathcal{M}(\mathbf{H}) \|_2 \leq \eta$. In particular, it might then be possible to use faster non-convex IHT based methods (cf., Remark 5).*

Estimating $\mathcal{S}_2, \mathcal{S}_1$. Given the linear program defined in (5.6), we can estimate $\nabla^2 f(\mathbf{x})$ in a straightforward manner, at any fixed $\mathbf{x} \in \mathbb{R}^d$. Indeed, for some suitable step size $\mu > 0$, we first collect the samples: $f(\mathbf{x}), \{f(\mathbf{x} - 2\mu \mathbf{v}_j)\}_{j=1}^{m_v}, \{f(\mathbf{x} + 2\mu \mathbf{v}_j)\}_{j=1}^{m_v}$, with $\mathbf{v}_j \in \mathcal{V}$. Then, we form the linear system $\mathbf{y} = \mathcal{M}(\nabla^2 f(\mathbf{x})) + \mathbf{n}$, where

$$y_j = \frac{f(\mathbf{x} + 2\mu \mathbf{v}_j) + f(\mathbf{x} - 2\mu \mathbf{v}_j) - 2f(\mathbf{x})}{4\mu^2}, \quad n_j = \frac{R_3(\zeta_j) + R_3(\zeta_j')}{4\mu^2}; j = 1, \ldots, m_v. \tag{5.8}$$

Since $\nabla^2 f(\mathbf{x})$ is at most $k(\rho_m + 1)$ sparse, therefore we obtain an estimate $\widehat{\nabla^2 f}(\mathbf{x})$ to $\nabla^2 f(\mathbf{x})$ with $2m_v + 1$ queries of f with $m_v > c_1' k \rho_m \log(\frac{d^2}{k\rho_m})$. Thereafter, we proceed as in Section 4, *i.e.*, we estimate $\nabla^2 f$ at each $\mathbf{x} \in \chi = \cup_{h \in \mathcal{H}_2^d} \chi(h)$, with $\chi(h)$ as defined in (3.12).

[9] We are not aware of a formal proof of this fact in the literature.

Remark 13. *Note that $\nabla^2 f(\mathbf{x})$ actually has at most $k+2|\mathcal{S}_2|$ non-zero entries. Therefore, if we had assumed $|\mathcal{S}_2|$ to be known as part of our problem setup (in Section 2), then the choice $m_v > c_1'(k+2|\mathcal{S}_2|)\log(\frac{d^2}{k+2|\mathcal{S}_2|})$ would suffice for estimating $\nabla^2 f(\mathbf{x})$. We can bound $2|\mathcal{S}_2| \leq k\rho_m$ – this is also tight in the worst case – however in certain settings this would be pessimistic*[10]

Once $\mathcal{S}_2$ is identified, we can simply reuse the procedure in Algorithm 3, for estimating $\mathcal{S}_1$. The above discussion for identifying $\mathcal{S}_1, \mathcal{S}_2$ is formally outlined in Algorithm 4. The following Theorem provides sufficient conditions on the sampling

Algorithm 4 Algorithm for estimating $\mathcal{S}_1, \mathcal{S}_2$

1: **Input:** $m_v, m_x \in \mathbb{Z}^+$; $\mu > 0$; $\eta, \tau > 0$.
2: **Initialization:** $\widehat{\mathcal{S}_1}, \widehat{\mathcal{S}_2} = \emptyset$.
3: **Output:** Estimates $\widehat{\mathcal{S}_2}$, $\widehat{\mathcal{S}_1}$.
4:
5: // ESTIMATION OF $\mathcal{S}_2$
6: Construct $(d,2)$-hash family $\mathcal{H}_2^d$ and sets $\mathcal{V}$.
7: **for** $h \in \mathcal{H}_2^d$ **do**
8: Construct the set $\chi(h)$.
9: **for** $i = 1, \ldots, (2m_x+1)^2$ and $\mathbf{x}_i \in \chi(h)$ **do**
10: $(\mathbf{y}_i)_j = \frac{f(\mathbf{x}_i + 2\mu\mathbf{v}_j) + f(\mathbf{x}_i - 2\mu\mathbf{v}_j) - 2f(\mathbf{x}_i)}{4\mu^2}$; $j = 1, \ldots, m_v$; $\mathbf{v}_j \in \mathcal{V}$.
11: $\widehat{\nabla^2} f(\mathbf{x}_i) := \underset{\mathbf{H}}{\operatorname{argmin}} \parallel \mathbf{H} \parallel_1$ s.t. $\mathbf{H}^T = \mathbf{H}$, $\parallel \mathbf{y}_i - \mathcal{M}(\mathbf{H}) \parallel_1 \leq \eta$.
12: $\widehat{\mathcal{S}_2} = \widehat{\mathcal{S}_2} \cup \left\{(q,q') \in \binom{[d]}{2} : |(\widehat{\nabla^2} f(\mathbf{x}_i))_{q,q'}| > \tau\right\}$.
13: **end for**
14: **end for**
15:
16: // ESTIMATION OF $\mathcal{S}_1$
17: Estimate $\mathcal{S}_1$ as in Algorithm 3.

parameters in Algorithm 4, that guarantee $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ with high probability.

Theorem 8. *Let $\mathcal{H}_2^d$ be of size $|\mathcal{H}_2^d| \leq 2(C+1)e^2 \log d$ for some constant $C > 1$. Then $\exists$ constants $c_1, c_1', c_2, C_1 > 0$, such that the following is true. Let m_x, m_v, μ satisfy*

$$m_x \geq \lambda_2^{-1},\ m_v > c_1' k\rho_m \log\left(\frac{d^2}{k\rho_m}\right),\ \mu < \frac{\sqrt{m_v} D_2}{2\sqrt{6} C_1 B_3 (4\rho_m + 1)k}. \tag{5.9}$$

We then have for the choices $\eta = \frac{2\sqrt{3}\mu B_3(4\rho_m+1)k}{\sqrt{m_v}}$, $\tau = C_1\eta$ that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ with probability at least $1 - c_1 e^{-c_2 m_v} - d^{-2C}$. Given that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, the sampling conditions for estimating $\widehat{\mathcal{S}_1}$ are identical to Theorem 4.

Query complexity. Estimating $\nabla^2 f(\mathbf{x})$ at some fixed $\mathbf{x}$ requires $2m_v + 1 = O(k\rho_m \log(\frac{d^2}{k\rho_m}))$ queries. Since $\nabla^2 f$ is estimated at all points in χ in the worst case, this consequently implies a total query complexity of $O(k\rho_m \log(\frac{d^2}{k\rho_m})|\chi|) = O(\lambda_2^{-2} k\rho_m (\log d)^2)$, for estimating $\mathcal{S}_2$. As seen in Theorem 4, we make an additional $O(\lambda_1^{-1}(k - |\widehat{\mathcal{S}_2^{\text{var}}}|)\log(d - |\widehat{\mathcal{S}_2^{\text{var}}}|))$ queries of f, in order to estimate $\mathcal{S}_1$. Therefore, the overall query complexity for estimating $\mathcal{S}_1, \mathcal{S}_2$ is $O(\lambda_2^{-2} k\rho_m (\log d)^2)$. Observe that this is better by a $\log d$ factor as compared to the sampling bound for Algorithm 3 (in the noiseless setting).

Computational complexity. The family $\mathcal{H}_2^d$ can be constructed[11] in time polynomial in d. At each $\mathbf{x} \in \chi$, we solve a linear program (Step 11) in $O(d^2)$ variables, which can be done up to arbitrary accuracy in time polynomial in (m_v, d). Since this is done at $|\chi| = O(\lambda_2^{-2} \log d)$ many points, hence the overall computation time for estimation of $\mathcal{S}_2$ (and subsequently $\mathcal{S}_1$) is polynomial in the number of queries, and in d.

5.2 Analysis for noisy setting

We now consider the case where at each query $\mathbf{x}$, we observe $f(\mathbf{x}) + z'$, with $z' \in \mathbb{R}$ denoting external noise. In order to estimate $\nabla^2 f(\mathbf{x})$, we obtain the samples : $f(\mathbf{x} + 2\mu\mathbf{v}_j) + z'_{j,1}$, $f(\mathbf{x} - 2\mu\mathbf{v}_j) + z'_{j,2}$ and $f(\mathbf{x}) + z'_3$; $j = 1, \ldots, m_v$. This changes (5.8) to the linear system $\mathbf{y} = \mathcal{M}(\nabla^2 f(\mathbf{x})) + \mathbf{n} + \mathbf{z}$, where $z_j = (z'_{j,1} + z'_{j,2} - 2z'_3)/(4\mu^2)$.

[10] For example when $O(1)$ variables have degree ρ_m, and the remaining variables have degree 1 leading to $|\mathcal{S}_2| = O(k + \rho_m)$.
[11] Recall discussion following Definition 1.

Arbitrary bounded noise. Assuming the external noise to be arbitrary and bounded, meaning that $|z'| < \varepsilon$, Theorem 9 shows that Algorithm 4 recovers $\mathcal{S}_1, \mathcal{S}_2$ with appropriate choice of sampling parameters provided ε is not too large.

Theorem 9. *Assuming the notation in Theorem 8, let m_x, m_v and $\mathcal{H}_2^d$ be as defined in Theorem 8. Denoting $a = \frac{\sqrt{6}B_3(4\rho_m+1)k}{\sqrt{m_v}}$, say ε satisfies $\varepsilon < \varepsilon_1 = \frac{\sqrt{2}D_2^3}{54a^2C_1^3m_v}$ and $\theta_1 = \cos^{-1}\left(1 - \frac{2\varepsilon}{\varepsilon_1}\right)$. Let*

$$\mu \in \left(-\frac{D_2}{3aC_1}\cos\left(\frac{\theta_1}{3}+\frac{\pi}{3}\right)+\frac{D_2}{6aC_1}, \frac{D_2}{3aC_1}\cos\left(\frac{\theta_1}{3}\right)+\frac{D_2}{6aC_1}\right). \tag{5.10}$$

We then have in Algorithm 4 for the choices $\eta = \left(\frac{2\sqrt{3}\mu B_3(4\rho_m+1)k}{\sqrt{m_v}} + \frac{\varepsilon m_v}{\mu^2}\right)$, $\tau = C_1\eta$, that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ with probability at least $1 - c_1e^{-c_2m_v} - d^{-2C}$. Given that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, the sampling conditions for estimating $\widehat{\mathcal{S}_1}$ are identical to Theorem 5.

Stochastic noise. We now consider i.i.d Gaussian noise, so that $z' \sim \mathcal{N}(0, \sigma^2)$ for variance $\sigma^2 < \infty$. As in Sections 3.2, 4.2, we reduce σ via resampling and averaging. Doing this N_1 times in Step 10, and N_2 times during estimation of $\mathcal{S}_1$, for N_1, N_2 large enough, we can recover $\mathcal{S}_1, \mathcal{S}_2$ as shown formally in the following Theorem.

Theorem 10. *Assuming the notation in Theorem 8, let m_x, m_v and $\mathcal{H}_2^d$ be as defined in Theorem 8. For any $\varepsilon < \varepsilon_1 = \frac{\sqrt{2}D_2^3}{54a^2C_1^3m_v}$, $0 < p_1 < 1$, say we resample each query in Step 10 of Algorithm 4, $N_1 > \frac{3\sigma^2}{4\varepsilon^2}\log(\frac{2}{p_1}m_v(2m_x+1)^2|\mathcal{H}_2^d|)$ times, and average the values. We then have in Algorithm 4 for the choices of η, τ, μ as in Theorem 9, that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ with probability at least $1 - c_1e^{-c_2m_v} - d^{-2C} - p_1$. Given that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, the sampling conditions for estimating $\widehat{\mathcal{S}_1}$ are identical to Theorem 6.*

Query complexity. We now analyze the query complexity for Algorithm 4, when the noise is i.i.d Gaussian. For estimating $\mathcal{S}_2$, we have $\varepsilon = O(\rho_m^{-2}k^{-2})$. Furthermore: $(2m_x+1)^2 = \lambda_2^{-2}$, $|\mathcal{H}_2^d| = O(\log d)$, $m_v = O(k\rho_m \log d)$. Choosing $p_1 = d^{-\delta}$ for any constant $\delta > 0$ gives us

$$N_1 = O(\rho_m^4k^4\log(d^\delta(k\rho_m\log d)\lambda_2^{-2}\log d)) = O(\rho_m^4k^4\log d).$$

This means that our total sample complexity for ensuring $\widehat{\mathcal{S}_2} = \mathcal{S}_2$ with high probability is:

$$O(N_1k\rho_m\log d|\chi|) = O(\rho_m^5k^5(\log d)^3\lambda_2^{-2}).$$

Lastly, by noting the sample complexity for estimating $\mathcal{S}_1$ from Theorem 6, we conclude that the overall sample complexity for ensuring $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ and $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, in the presence of i.i.d Gaussian noise, is $O(\rho_m^5k^5(\log d)^3\lambda_2^{-2})$. Observe that this bound has a relatively worse scaling w.r.t ρ_m compared to that for Algorithm 3 (derived after Theorem 6); specifically, by a factor of k^3. On the other hand, the scaling w.r.t d is better by a logarithmic factor, compared to that for Algorithm 3.

6 Learning individual components of model

Recall from (2.4) the unique representation of the model:

$$f(x_1,\ldots,x_d) = c + \sum_{p\in\mathcal{S}_1}\phi_p(x_p) + \sum_{(l,l')\in\mathcal{S}_2}\phi_{(l,l')}(x_l,x_{l'}) + \sum_{q\in\mathcal{S}_2^{\text{var}}:\rho(q)>1}\phi_q(x_q), \tag{6.1}$$

where $\mathcal{S}_1 \cap \mathcal{S}_2^{\text{var}} = \emptyset$. Having estimated the sets $\mathcal{S}_1$ and $\mathcal{S}_2$, we now show how the individual univariate and bivariate functions in the model can be estimated. We will see this for the settings of noiseless, as well as noisy (arbitrary, bounded noise and stochastic noise) point queries.

6.1 Noiseless queries

In this scenario, we obtain the exact value $f(\mathbf{x})$ at each query $\mathbf{x} \in \mathbb{R}^d$. Let us first see how each ϕ_p; $p \in \mathcal{S}_1$ can be estimated. For some $-1 = t_1 < t_2 < \cdots < t_n = -1$, consider the set

$$\chi_p := \left\{\mathbf{x}_i \in \mathbb{R}^d : (\mathbf{x}_i)_j = \left\{\begin{array}{ll} t_i; & j = p, \\ 0; & j \neq p \end{array}\right\}; 1 \le i \le n; 1 \le j \le d\right\}; \quad p \in \mathcal{S}_1. \tag{6.2}$$

We obtain the samples $\{f(\mathbf{x}_i)\}_{i=1}^n$; $\mathbf{x}_i \in \chi_p$. Here $f(\mathbf{x}_i) = \phi_p(t_i) + C$ with C being a constant that depends on the other components in the model. Given the samples, one can then employ spline based "quasi interpolant operators" [14], to obtain an estimate $\tilde{\phi}_p : [-1,1] \to \mathbb{R}$, to $\phi_p + C$. Construction of such operators can be found for instance in [14] (see also [22]). One

can suitably choose the t_i's and construct quasi interpolants that approximate any C^m smooth univariate function with optimal $L_\infty[-1,1]$ error rate $O(n^{-m})$ [14, 22]. Having obtained $\tilde{\phi}_p$, we then define

$$\widehat{\phi}_p := \tilde{\phi}_p - \mathbb{E}_p[\tilde{\phi}_p]; \quad p \in \mathcal{S}_1, \tag{6.3}$$

to be the estimate of ϕ_p. The bivariate components corresponding to each $(l,l') \in \mathcal{S}_2$ can be estimated in a similar manner as above. To this end, for some strictly increasing sequences: $(-1 = t'_1, t'_2, \ldots, t'_{n_1} = 1)$, $(-1 = t_1, t_2, \ldots, t_{n_1} = 1)$, consider the set

$$\chi_{(l,l')} := \left\{ \mathbf{x}_{i,j} \in \mathbb{R}^d : (\mathbf{x}_{i,j})_q = \left\{ \begin{array}{ll} t'_i; & q = l, \\ t_j; & q = l', \\ 0; & q \neq l, l' \end{array} \right\}; 1 \leq i,j \leq n_1; 1 \leq q \leq d \right\}; \quad (l,l') \in \mathcal{S}_2. \tag{6.4}$$

We then obtain the samples $\{f(\mathbf{x}_{i,j})\}_{i,j=1}^{n_1}; \mathbf{x}_{i,j} \in \chi_{(l,l')}$ where

$$\begin{aligned} f(\mathbf{x}_{i,j}) &= \phi_{(l,l')}(t'_i, t_j) + \sum_{\substack{l_1:(l,l_1)\in\mathcal{S}_2 \\ l_1 \neq l'}} \phi_{(l,l_1)}(t'_i, 0) + \sum_{\substack{l_1:(l_1,l)\in\mathcal{S}_2 \\ l_1 \neq l'}} \phi_{(l_1,l)}(0, t'_i) \\ &\quad + \sum_{\substack{l'_1:(l',l'_1)\in\mathcal{S}_2 \\ l'_1 \neq l}} \phi_{(l',l'_1)}(t_j, 0) + \sum_{\substack{l'_1:(l'_1,l')\in\mathcal{S}_2 \\ l'_1 \neq l}} \phi_{(l'_1,l')}(0, t_j) + \phi_l(t'_i) + \phi_{l'}(t_j) + C, \qquad (6.5) \\ &= g_{(l,l')}(t'_i, t_j) + C, \qquad (6.6) \end{aligned}$$

with C being a constant. (6.5) is a general expression – if for example $\rho(l) = 1$, then the terms $\phi_l, \phi_{(l,l_1)}, \phi_{(l_1,l)}$ will be zero. Given this, we can again obtain estimates $\tilde{\phi}_{(l,l')} : [-1,1]^2 \to \mathbb{R}$ to $g_{(l,l')} + C$, via spline based quasi interpolants. Let us denote $n = n_1^2$ to be the total number of samples of f. For an appropriate choice of (t'_i, t_j)'s, one can construct bivariate quasi interpolants that approximate any C^m smooth bivariate function, with optimal $L_\infty[-1,1]^2$ error rate $O(n^{-m/2})$ [14, 22]. Subsequently, we define the final estimates $\widehat{\phi}_{(l,l')}$ to $\phi_{(l,l')}$ as follows.

$$\widehat{\phi}_{(l,l')} := \left\{ \begin{array}{rl} \tilde{\phi}_{(l,l')} - \mathbb{E}_{(l,l')}[\tilde{\phi}_{(l,l')}]; & \rho(l), \rho(l') = 1, \\ \tilde{\phi}_{(l,l')} - \mathbb{E}_{l}[\tilde{\phi}_{(l,l')}]; & \rho(l) = 1, \rho(l') > 1, \\ \tilde{\phi}_{(l,l')} - \mathbb{E}_{l'}[\tilde{\phi}_{(l,l')}]; & \rho(l) > 1, \rho(l') = 1, \\ \tilde{\phi}_{(l,l')} - \mathbb{E}_{l}[\tilde{\phi}_{(l,l')}] - \mathbb{E}_{l'}[\tilde{\phi}_{(l,l')}] + \mathbb{E}_{(l,l')}[\tilde{\phi}_{(l,l')}]; & \rho(l) > 1, \rho(l') > 1. \end{array} \right. \tag{6.7}$$

Lastly, we require to estimate the univariate's : ϕ_l for each $l \in \mathcal{S}_2^{\text{var}}$ such that $\rho(l) > 1$. As above, for some strictly increasing sequences: $(-1 = t'_1, t'_2, \ldots, t'_{n_1} = 1)$, $(-1 = t_1, t_2, \ldots, t_{n_1} = 1)$, consider the set

$$\chi_l := \left\{ \mathbf{x}_{i,j} \in \mathbb{R}^d : (\mathbf{x}_{i,j})_q = \left\{ \begin{array}{ll} t'_i; & q = l, \\ t_j; & q \neq l \;\&\; q \in \mathcal{S}_2^{\text{var}}, \\ 0; & q \notin \mathcal{S}_2^{\text{var}}, \end{array} \right\}; 1 \leq i,j \leq n_1; 1 \leq q \leq d \right\}; \quad l \in \mathcal{S}_2^{\text{var}} : \rho(l) > 1. \tag{6.8}$$

We obtain $\{f(\mathbf{x}_{i,j})\}_{i,j=1}^{n_1}; \mathbf{x}_{i,j} \in \chi_l$ where this time

$$\begin{aligned} f(\mathbf{x}_{i,j}) &= \phi_l(t'_i) + \sum_{\rho(l')>1, l' \neq l} \phi_{l'}(t_j) + \sum_{l':(l,l')\in\mathcal{S}_2} \phi_{(l,l')}(t'_i, t_j) \qquad (6.9) \\ &\quad + \sum_{l':(l',l)\in\mathcal{S}_2} \phi_{(l',l)}(t_j, t'_i) + \sum_{(q,q')\in\mathcal{S}_2 : q,q' \neq l} \phi_{(q,q')}(t_j, t_j) + C \qquad (6.10) \\ &= g_l(t'_i, t_j) + C \qquad (6.11) \end{aligned}$$

for a constant, C. Denoting $n = n_1^2$ to be the total number of samples of f, we can again obtain an estimate $\tilde{\phi}_l(x_l, x)$ to $g_l(x_l, x) + C$, with $L_\infty[-1,1]^2$ error rate $O(n^{-3/2})$. Then with $\tilde{\phi}_l$ at hand, we define the estimate $\widehat{\phi}_l : [-1,1] \to \mathbb{R}$ as

$$\widehat{\phi}_l := \mathbb{E}_x[\tilde{\phi}_l] - \mathbb{E}_{(l,x)}[\tilde{\phi}_l]; \quad l \in \mathcal{S}_2^{\text{var}} : \rho(l) > 1. \tag{6.12}$$

The following proposition formally describes the error rates for the aforementioned estimates.

Proposition 1. *For C^3 smooth components $\phi_p, \phi_{(l,l')}, \phi_l$, let $\widehat{\phi}_p$, $\widehat{\phi}_{(l,l')}, \widehat{\phi}_l$ be the respective estimates as defined in* (6.3)*,* (6.7) *and* (6.12) *respectively. Also, let n denote the number of queries (of f) made per component. We then have that:*

1. $\| \widehat{\phi}_p - \phi_p \|_{L_\infty[-1,1]} = O(n^{-3}); \forall p \in \mathcal{S}_1$,
2. $\| \widehat{\phi}_{(l,l')} - \phi_{(l,l')} \|_{L_\infty[-1,1]^2} = O(n^{-3/2}); \forall (l,l') \in \mathcal{S}_2$, *and*
3. $\| \widehat{\phi}_l - \phi_l \|_{L_\infty[-1,1]} = O(n^{-3/2}); \forall l \in \mathcal{S}_2^{var} : \rho(l) > 1$.

6.2 Noisy queries

We now look at the case where for each query $\mathbf{x} \in \mathbb{R}^d$, we obtain a noisy value $f(\mathbf{x}) + z'$.

Arbitrary bounded noise. We begin with the scenario where z'_i is arbitrary and bounded with $|z'_i| < \varepsilon;\ \forall i$. Since the noise is arbitrary in nature, therefore we simply proceed *as in the noiseless case*, *i.e.*, by approximating each component via a quasi-interpolant. As the magnitude of the noise is bounded by ε, it results in an additional $O(\varepsilon)$ term in the approximation error rates of Proposition 1.

To see this for the univariate case, let us denote $Q : C(\mathbb{R}) \to \mathcal{H}$ to be a quasi-interpolant operator. This a linear operator, with $C(\mathbb{R})$ denoting the space of continuous functions defined over $\mathbb{R}$ and $\mathcal{H}$ denoting a univariate spline space. Consider $u \in C^m[-1,1]$ for some positive integer m, and let $g : [-1,1] \to \mathbb{R}$ be an arbitrary continuous function with $\| g \|_{L_\infty[-1,1]} < \varepsilon$. Denote $\widehat{u} = u + g$ to be the "corrupted" version of u, and let n be the number of samples of $\widehat{u}$ used by Q. We then have by linearity of Q that:

$$\| Q(\widehat{u}) - u \|_{L_\infty[-1,1]} = \| Q(u) + Q(g) - u \|_{L_\infty[-1,1]} \leq \underbrace{\| Q(u) - u \|_{L_\infty[-1,1]}}_{=O(n^{-m})} + \| Q \| \underbrace{\| g \|_{L_\infty[-1,1]}}_{\leq \|Q\|\varepsilon}, \tag{6.13}$$

with $\| Q \|$ being the operator norm of Q. One can construct Q with $\| Q \|$ bounded[12] from above by a constant depending only on m. The above argument can be extended easily to the multivariate case. We state this for the bivariate case for completeness. Denote $Q_1 : C(\mathbb{R}^2) \to \mathcal{H}$ to be a quasi-interpolant operator, with $\mathcal{H}$ denoting a bivariate spline space. Consider $u_1 \in C^m[-1,1]^2$ for some positive integer m, and let $g_1 : [-1,1] \to \mathbb{R}$ be an arbitrary continuous function with $\| g_1 \|_{L_\infty[-1,1]^2} < \varepsilon$. Let $\widehat{u}_1 = u_1 + g_1$ and let n be the number of samples of $\widehat{u_1}$ used by Q_1. We then have by linearity of Q_1 that:

$$\| Q_1(\widehat{u_1}) - u_1 \|_{L_\infty[-1,1]^2} = \| Q_1(u_1) + Q_1(g_1) - u_1 \|_{L_\infty[-1,1]^2} \leq \underbrace{\| Q_1(u_1) - u_1 \|_{L_\infty[-1,1]^2}}_{=O(n^{-m/2})} + \| Q_1 \| \underbrace{\| g_1 \|_{L_\infty[-1,1]^2}}_{\leq \|Q_1\|\varepsilon}, \tag{6.14}$$

with $\| Q_1 \|$ being the operator norm of Q_1. As for the univariate case, one can construct Q_1 with $\| Q_1 \|$ bounded[12] from above by a constant depending only on m.

Let us define our final estimates $\widehat{\phi}_p$, $\widehat{\phi}_{(l,l')}$ and $\widehat{\phi}_l$ as in (6.3), (6.7) and (6.12), respectively. The following proposition formally states the error bounds, for this particular noise model.

Proposition 2 (Arbitrary bounded noise)**.** *For C^3 smooth components $\phi_p, \phi_{(l,l')}, \phi_l$, let $\widehat{\phi}_p$, $\widehat{\phi}_{(l,l')}$, $\widehat{\phi}_l$ be the respective estimates as defined in* (6.3)*,* (6.7) *and* (6.12) *respectively. Also, let n denote the number of noisy queries (of f) made per component with the external noise magnitude being bounded by ε. We then have that*

1. $\| \widehat{\phi}_p - \phi_p \|_{L_\infty[-1,1]} = O(n^{-3}) + O(\varepsilon); \forall p \in \mathcal{S}_1$,
2. $\| \widehat{\phi}_{(l,l')} - \phi_{(l,l')} \|_{L_\infty[-1,1]^2} = O(n^{-3/2}) + O(\varepsilon); \forall (l,l') \in \mathcal{S}_2$, *and*
3. $\| \widehat{\phi}_l - \phi_l \|_{L_\infty[-1,1]} = O(n^{-3/2}) + O(\varepsilon); \forall l \in \mathcal{S}_2^{var} : \rho(l) > 1$.

The proof is similar to that of Proposition 1 and hence skipped.

Stochastic noise. We now consider the setting where $z'_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d Gaussian random variables. Similar to the noiseless case, estimating the individual components again involves sampling f along the subspaces corresponding to $\mathcal{S}_1$, $\mathcal{S}_2$. Due to the presence of stochastic noise however, we now make use of *nonparametric regression* techniques to compute the estimates. While there exist a number of methods that could be used for this purpose (cf. [52]), we only discuss a specific one for clarity of exposition.

To elaborate, we again construct the sets defined in (6.2),(6.4) and(6.8). In particular, we uniformly discretize the domains $[-1,1]$ and $[-1,1]^2$, by choosing the respective t_i's and (t'_i, t_j)'s accordingly. This is the so called "fixed design" setting in nonparametric statistics. Upon collecting the samples $\{f(\mathbf{x}_i) + z'_i\}_{i=1}^n$ one can then derive estimates $\tilde{\phi}_p$, $\tilde{\phi}_{(l,l')}$, $\tilde{\phi}_l$, to $\phi_p + C$, $g_{(l,l')} + C$ and $g_l + C$ respectively, by using *local polynomial estimators* (cf. [52, 17] and references within). It is known that these estimators achieve the (minimax optimal) L_∞ error rate: $\Omega((n^{-1}\log n)^{\frac{m}{2m+d}})$, for estimating d-variate, C^m smooth functions over compact domains[13]. Translated to our setting, we then have that the functions: $\phi_p + C$, $g_{(l,l')} + C$ and $g_l + C$ are estimated at the rates: $O((n^{-1}\log n)^{\frac{3}{7}})$ and $O((n^{-1}\log n)^{\frac{3}{8}})$ respectively.

Denoting the above intermediate estimates by $\tilde{\phi}_p$, $\tilde{\phi}_{(l,l')}$, $\tilde{\phi}_l$, we define our final estimates $\widehat{\phi}_p$, $\widehat{\phi}_{(l,l')}$ and $\widehat{\phi}_l$ as in (6.3), (6.7) and (6.12), respectively. The following Proposition describes the error rates of these estimates.

[12] For instance, see Theorems 14.4, 15.2 in [22]

[13] See [52] for $d = 1$, and [38] for $d \geq 1$

Proposition 3 (i.i.d Gaussian noise)**.** *For C^3 smooth components $\phi_p, \phi_{(l,l')}, \phi_l$, let $\widehat{\phi}_p$, $\widehat{\phi}_{(l,l')}, \widehat{\phi}_l$ be the respective estimates as defined in* (6.3)*,* (6.7) *and* (6.12) *respectively. Let n denote the number of noisy queries (of f) made per component, with noise samples $z'_1, z'_2, \ldots, z'_n$ being i.i.d Gaussian. Furthermore, let $\mathbb{E}_z[\cdot]$ denote expectation w.r.t the joint distribution of $z'_1, z'_2, \ldots, z'_n$. We then have that*

1. $\mathbb{E}_z[\| \widehat{\phi}_p - \phi_p \|_{L_\infty[-1,1]}] = O((n^{-1}\log n)^{\frac{3}{7}}); \forall p \in \mathcal{S}_1$,
2. $\mathbb{E}_z[\| \widehat{\phi}_{(l,l')} - \phi_{(l,l')} \|_{L_\infty[-1,1]^2}] = O((n^{-1}\log n)^{\frac{3}{8}}); \forall (l,l') \in \mathcal{S}_2$, *and*
3. $\mathbb{E}_z[\| \widehat{\phi}_l - \phi_l \|_{L_\infty[-1,1]}] = O((n^{-1}\log n)^{\frac{3}{8}}); \forall l \in \mathcal{S}_2^{var} : \rho(l) > 1$.

7 Simulation results

We now provide some simulation results for our methods on synthetic examples. The main goal of our experiments is to provide a proof of concept, validating some of the theoretical results that were derived earlier. We consider both non-overlapping (Section 7.1) and overlapping settings (Section 7.2). In our experiments, we use the ALPS algorithm [28] as our CS solver – an efficient first-order method.

Starting with the non-overlapping case, we present phase transition results and also show the dependence of d on the number of samples, for recovery of $\mathcal{S}_1$, $\mathcal{S}_2$. We then empirically demonstrate the dependence of the number of samples on k. In both cases, our findings support our theory for sample complexities. We conduct similar experiments for the overlapping case, and also additionally demonstrate empirically the dependence of the number of samples on the parameter ρ_m.

7.1 Non-overlapping setting

We consider the following experimental setup: $\mathcal{S}_1 = \{1,2\}$ and $\mathcal{S}_2 = \{(3,4),(5,6)\}$, which implies $k_1 = 2$, $k_2 = 2$ and $k = 6$. Moreover, we consider three different types of f namely:

(i) $f_1(\mathbf{x}) = 2x_1 - 3x_2^2 + 4x_3x_4 - 5x_5x_6$,

(ii) $f_2(\mathbf{x}) = 10\sin(\pi \cdot x_1) + 5e^{-2x_2} + 10\sin(\pi \cdot x_3x_4) + 5e^{-2x_5x_6}$,

(iii) $f_3(\mathbf{x}) = \frac{10}{3}\cos(\pi \cdot x_1) + 8x_1^2 + 5(x_2^4 - x_2^2 + \frac{4}{5}x_4) + \frac{10}{3}\cos(\pi \cdot x_3x_4) + 8(x_3x_4)^2 + 5((x_5x_6)^4 - (x_5x_6)^2 + \frac{4}{5}x_5x_6)$.

For all cases, we use Algorithms 1 and 2. For f_1, the problem parameters are set to $\lambda_1 = 0.3$, $\lambda_2 = 1$, $D_1 = 2$, $D_2 = 3$, $B_3 = 6$, while for f_2, f_3: $\lambda_1 = \lambda_2 = 0.3$, $D_1 = 8$, $D_2 = 4$, $B_3 = 35$. Given these constants, we obtain $m_x = 1$, $m'_x = 4$ for f_1 and $m_x = m'_x = 4$ for f_2, f_3. We use constant $\widetilde{C}$ (to be defined next) when we set $m_v := \widetilde{C}k\log(d/k)$. For the construction of the hash functions, we set the size to $|\mathcal{H}_2^d| = C'\log d$ with $C' = 1.7$, leading to $|\mathcal{H}_2^d| \in [8,12]$ for $10^2 \leq d \leq 10^3$. For the noiseless setting, we choose step sizes: μ, μ_1, β and thresholds: τ', τ as in Lemma 1 and Lemma 2.

For the noisy setting, we consider the function values to be corrupted with i.i.d. Gaussian noise. We reduce the noise variance by repeating each query N_1 and N_2 times respectively, and averaging. The noise variance values considered are $\sigma^2 \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ for which we choose:

$$\begin{aligned} & (N_1,N_2) \in \{(40,15),(75,31),(80,35)\} \quad \text{for} \quad f_1, \\ & (N_1,N_2) \in \{(60,30),(85,36),(90,40)\} \quad \text{for} \quad f_2, \\ \text{and} \quad & (N_1,N_2) \in \{(59,30),(85,35),(90,40)\} \quad \text{for} \quad f_3. \end{aligned}$$

Moreover, we now choose parameters $\mu, \mu_1, \beta, \tau', \tau$ as in Theorem 2.

Dependence on d. We see in Fig. 3, that for $\widetilde{C} \approx 3.8$ the probability of successful identification (noiseless case) undergoes a phase transition and becomes close to 1, for different values of d. This validates the statements of Lemmas 1-2. Fixing $\widetilde{C} = 3.8$, we then see that with the total number of queries growing slowly with d, we have successful identification. For the noisy case, the total number of queries is roughly 10^2 times that in the noiseless setting, however the scaling with d is similar to that for noiseless case. Focusing on the function models f_2 and f_3, observe that the number of queries is seen to be slightly larger than that for f_1 in the noisy settings; this fact becomes more obvious in the overlapping case later on.

Dependence on k. We now demonstrate the scaling of the total number of queries versus the sparsity k for identification of $\mathcal{S}_1, \mathcal{S}_2$. Consider the model

$$f(\mathbf{x}) = \sum_{i=1}^{T} \left(\alpha_1 \mathbf{x}_{(i-1)5+1} - \alpha_2 \mathbf{x}_{(i-1)5+2}^2 + \alpha_3 \mathbf{x}_{(i-1)5+3}\mathbf{x}_{(i-1)5+4} - \alpha_4 \mathbf{x}_{(i-1)5+5}\mathbf{x}_{(i-1)5+6} \right) \tag{7.1}$$

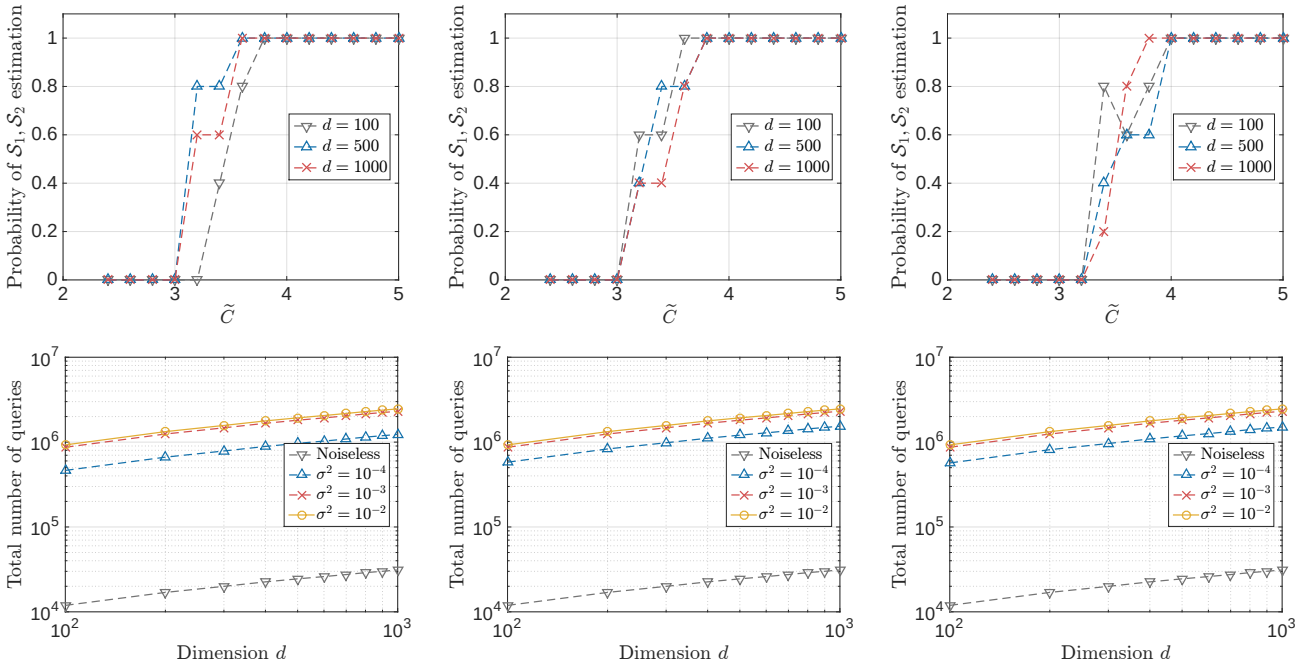


Figure 3: First (resp. second and third) column is for f_1 (resp. f_2 and f_3). Top row depicts the success probability of identifying exactly $\mathcal{S}_1, \mathcal{S}_2$, in the noiseless case. x-axis represent the constant $\widetilde{C}$. The bottom panel depicts total queries vs. d for exact recovery, with $\widetilde{C} = 3.8$ and various noise settings. All results are over 5 independent Monte Carlo trials.

where $\mathbf{x} \in \mathbb{R}^d$ for $d = 500$. Here, $\alpha_i \in [2, 5], \forall i$; *i.e.*, we randomly selected α_i's within range and kept the values fixed for all 5 Monte Carlo iterations. Note that sparsity $k = 6T$; we consider $T \in \{1, 2, \ldots, 10\}$. We set $\lambda_1 = 0.3$, $\lambda_2 = 1$, $D_1 = 2$, $D_2 = 3$, $B_3 = 6$ and $\widetilde{C} = 3.8$, *i.e.*, the same setting with model f_1 above. For the noisy cases, we consider σ^2 as before, and choose the same values for (N_1, N_2) as for f_1. In Figure 4, we see that the number of queries scales as $\sim k \log(d/k)$, and is roughly 10^2 more in the noisy case as compared to the noiseless setting.

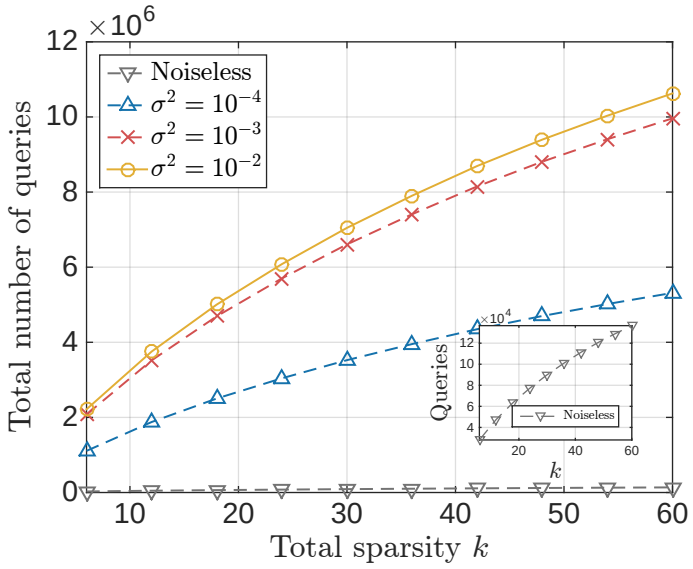


Figure 4: Total number of queries versus different sparsity values k, for (7.1). This is for both noiseless and noisy cases (i.i.d Gaussian) with variances $\sigma^2 \in \{10^{-4}, 10^{-3}, 10^{-2}\}$.

7.2 Overlapping setting

For the overlapping case, we set $\mathcal{S}_1 = \{1, 2\}$ and $\mathcal{S}_2 = \{(3, 4), (4, 5)\}$, which implies $k_1 = 2$, $k_2 = 2$, $\rho_m = 2$ and $k = 5$. Due to the presence of overlap between the elements of $\mathcal{S}_2$, we now employ Algorithm 3 for identifying $\mathcal{S}_1, \mathcal{S}_2$.

Remark 14. *We deliberately avoid using Algorithm 4 on account of Remark 12 – it is unclear to us whether IHT based methods could be employed for solving* (5.6)*, with provable recovery guarantees. While we could instead use standard interior point solvers, they will be slow, especially for the range of values of dimension d that we will be considering.*

For an easier comparison with the non-overlapping case, we consider similar models as the previous subsection; observe that there are now common variables across the components of f.

(i) $f_1(\mathbf{x}) = 2x_1 - 3x_2^2 + 4x_3x_4 - 5x_4x_5$,

(ii) $f_2(\mathbf{x}) = 10\sin(\pi \cdot x_1) + 5e^{-2x_2} + 10\sin(\pi \cdot x_3x_4) + 5e^{-2x_4x_5}$,

(iii) $f_3(\mathbf{x}) = \frac{10}{3}\cos(\pi \cdot x_1) + 8x_1^2 + 5(x_2^4 - x_2^2 + \frac{4}{5}x_4) + \frac{10}{3}\cos(\pi \cdot x_3x_4) + 8(x_3x_4)^2 + 5((x_4x_5)^4 - (x_4x_5)^2 + \frac{4}{5}x_4x_5)$.

Parameters $\lambda_1, \lambda_2, D_1, D_2, B_3$ are set as in the previous subsection. For a constant $\widetilde{C}$ (chosen later), we set $m_v := \widetilde{C}k\log(d/k)$, $m_{v'} := \widetilde{C}\rho_m \log(d/\rho_m)$, and $m_{v''} := \widetilde{C}(k - |\widehat{\mathcal{S}_2^{\text{var}}}|)\log(\frac{|\mathcal{P}|}{k-|\widehat{\mathcal{S}_2^{\text{var}}}|})$. The size of the hash family $|\mathcal{H}_2^d|$ for different values of d is set as before for the non-overlapping setting. For the noiseless setting, we choose step sizes: μ, μ_1, μ' and thresholds: τ', τ'' as in Theorem 4.

For the noisy setting, we consider the function values to be corrupted with i.i.d. Gaussian noise. We reduce the noise variance by repeating each query N_1 and N_2 times respectively, and averaging. The noise variance values considered are: $\sigma^2 \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ for which we choose:

$$\begin{aligned} & (N_1, N_2) \in \{(50, 20), (85, 36), (90, 40)\} \quad \text{for} \quad f_1, \\ & (N_1, N_2) \in \{(60, 30), (90, 40), (95, 43)\} \quad \text{for} \quad f_2, \\ \text{and} \quad & (N_1, N_2) \in \{(59, 30), (89, 40), (93, 43)\} \quad \text{for} \quad f_3. \end{aligned}$$

Moreover, we now choose the parameters: $\mu, \mu_1, \mu', \tau', \tau''$ as in Theorem 5.

Dependence on d. We see in Fig. 5, that for $\widetilde{C} \approx 5.6$ the probability of successful identification (noiseless case) undergoes a phase transition and becomes close to 1, for different values of d, as in the non-overlapping case. This validates the statement of Theorem 4. As in the non-overlapping case, in the presence of noise, the total number of queries is roughly 10^2 times that in the noiseless setting, however the scaling with d is similar to that for the noiseless setting.

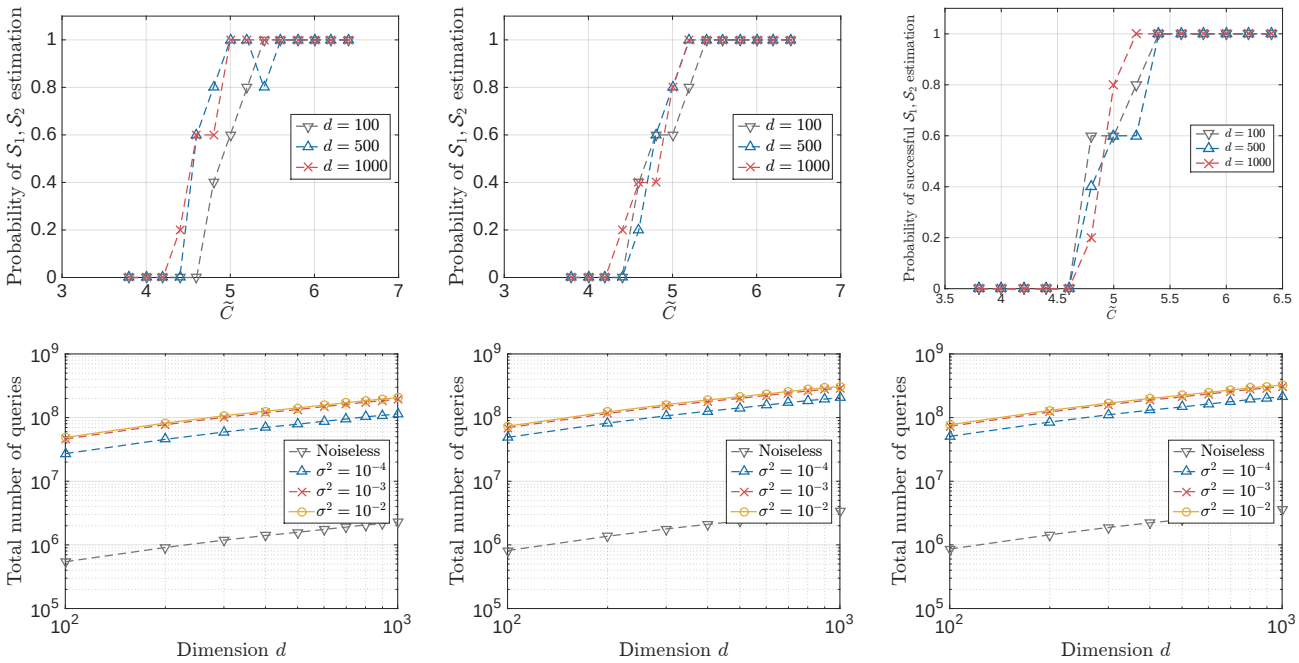


Figure 5: First (resp. second and third) column is for f_1 (resp. f_2 and f_3). Top row depicts the success probability of identifying exactly $\mathcal{S}_1, \mathcal{S}_2$, in the noiseless case. x-axis represent the constant $\widetilde{C}$. The bottom panel depicts total queries vs. d for exact recovery, with $\widetilde{C} = 5.6$ and various noise settings. All results are over 5 independent Monte Carlo trials.

Dependence on k. We now demonstrate the scaling of the total number of queries versus the sparsity k for identification of $\mathcal{S}_1, \mathcal{S}_2$. Consider the model

$$f(\mathbf{x}) = \sum_{i=1}^{T} \left(\alpha_1 \mathbf{x}_{(i-1)5+1} - \alpha_2 \mathbf{x}^2_{(i-1)5+2} + \alpha_3 \mathbf{x}_{(i-1)5+3} \mathbf{x}_{(i-1)5+4} - \alpha_4 \mathbf{x}_{(i-1)5+4} \mathbf{x}_{(i-1)5+5} \right) \tag{7.2}$$

where $\mathbf{x} \in \mathbb{R}^d$ for $d = 500$. Here, $\alpha_i \in [2, 5], \forall i$; *i.e.*, we randomly selected α_i's within range and kept the values fixed for all 5 Monte Carlo iterations. Note that $\rho_m = 2$ and the sparsity $k = 5T$; we consider $T \in \{1, 2, \ldots, 10\}$. We set $\lambda_1 = 0.3$, $\lambda_2 = 1$, $D_1 = 2$, $D_2 = 3$, $B_3 = 6$ and $\widetilde{C} = 5.6$. For the noisy cases, we consider σ^2 as before, and choose the same values for (N_1, N_2) as for f_1. In Figure 6(Left panel), we again see that the number of queries scales as $\sim k \log(d/k)$, and is roughly 10^2 more in the noisy case as compared to the noiseless setting.

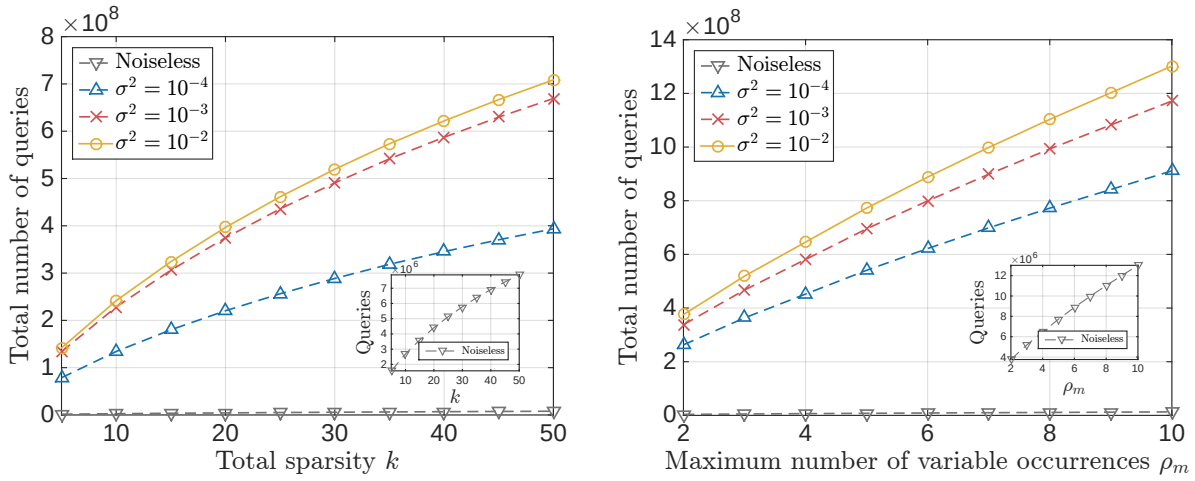


Figure 6: Left panel: Total number of queries versus different sparsity values k, for (7.2). Right panel: Total number of queries versus ρ_m for (7.3). This is for both noiseless and noisy cases (i.i.d Gaussian) with variances $\sigma^2 \in \{10^{-4}, 10^{-3}, 10^{-2}\}$.

Dependence on ρ_m. We now demonstrate the scaling of the total queries versus the maximum degree ρ_m for identification of $\mathcal{S}_1, \mathcal{S}_2$. Consider the model $f(\mathbf{x}) =$

$$\alpha_1 \mathbf{x}_1 - \alpha_2 \mathbf{x}_2^2 + \sum_{i=1}^{T} (\alpha_{3,i} \mathbf{x}_3 \mathbf{x}_{i+3}) + \sum_{i=1}^{5} (\alpha_{4,i} \mathbf{x}_{2+2i} \mathbf{x}_{3+2i}) \,. \tag{7.3}$$

We choose $d = 500$, $\widetilde{C} = 6$, $\alpha_i \in [2, \ldots, 5], \forall i$ (as earlier) and set $\lambda_1 = 0.3$, $\lambda_2 = 1$, $D_1 = 2$, $D_2 = 3$, $B_3 = 6$. For $T \geq 2$, we have $\rho_m = T$; we choose $T \in \{2, 3, \ldots, 10\}$. Also note that $k = 13$ throughout. For the noisy cases, we consider σ^2 as before, and choose $(N_1, N_2) \in \{(70, 40), (90, 50), (100, 70)\}$. In Figure 6(Right panel), we see that the number of queries scales as $\sim \rho_m \log(d/\rho_m)$, and is roughly 10^2 more in the noisy case as compared to the noiseless setting.

8 Discussion

We now provide a more detailed discussion with respect to related work, starting with results for learning SPAMs.

Learning SPAMs. Ravikumar et al. [45], Meier et al. [33] proposed methods based on least squares loss regularized with sparsity and smoothness constraints. While Ravikumar et al. show their method to be sparsistent for second order Sobolev smooth f, one can obtain a rough estimate of how the number of samples n behaves with respect to k, d. Indeed, from Corollary 1 of Theorem 2 in [45], we see that the probability of incorrect identification of $\mathcal{S}$ approximately scales[14] as: $\frac{\log d}{(\log n)^2} + \frac{k}{\log n} + \frac{\log d \sqrt{k}}{n^{1/6}}$. This means that n roughly scales as $\max\{k^3 (\log d)^6, e^k, e^{\sqrt{\log d}}\}$, for a constant probability of error. In contrast, our $O(k^2 (\log d)^2)$ bound (recall Theorem 6) has a clearly better scaling.

Meier et al. [33] derive error rates of $O(k(\log d/n)^{2/5})$ for estimating C^2 smooth f in the empirical $L_2(\mathbb{P}_n)$ norm. They also show conditions under which their method is guaranteed to recover $\widehat{\mathcal{S}} \subset \mathcal{S}$.

Huang et al. [23] proposed a method based on the adaptive group Lasso, and show that it is sparsistent. In contrast to [45], it is unclear here how exactly n scales with k, d. They also derive L_2 error rates for estimating the individual components of the SPAM.

[14] Here, we set the term ρ_n^* capturing the minimum magnitude of the univariate components (as defined in [45, Theorem 2]) to O(1).

Wahl [57] consider the variable selection problem for SPAMs. They propose an estimator that essentially involves looking at all subsets of $\{1, \ldots, d\}$ of size k, and hence is practically infeasible. They show that for the periodic Sobolev class of functions (with smoothness parameter $\alpha > 1/2$), their estimator recovers $\mathcal{S}$ w.h.p with $O(k^{\frac{2\alpha+1}{2\alpha}}(\log d)^4)$ samples [57, Corollary 3]. Consequently, they are also able to estimate each individual component of the model in the $L_2(\mathbb{P})$ norm. We observe that the dependency of their bound on d is worse than ours by a factor of $(\log d)^2$, however the scaling with k is better for all $\alpha > 1/2$.

Learning generalized SPAMs. Radchenko et al. [42] proposed the VANISH algorithm – a least squares method with sparsity constraints. Assuming f to be second order Sobolev smooth, they show their method to be sparsistent. They also show a consistency result for estimating f, similar to [45]. One can obtain a rough estimate of how their sampling bounds scale with $d, |\mathcal{S}_1|, |\mathcal{S}_2|$ for exact identification of $\mathcal{S}_1, \mathcal{S}_2$. Denoting $m = |\mathcal{S}_1| + |\mathcal{S}_2|$, and n to be the number of samples, we see from Corollary 1 of [42, Theorem 2] that the probability of failure, i.e., incorrect identification of $\mathcal{S}_1, \mathcal{S}_2$, approximately scales[15] as $\frac{\sqrt{m}}{\log n} + \frac{(\log d)^3}{n^{3/5}}$. This implies that n roughly scales as $\max\{e^m, (\log d)^5\}$ for a constant probability of error. In contrast, as seen from Theorems 6,10, our bounds are polynomial in m, and have a better scaling with dimension d.

Dalalyan et al. [13] studied a generalization of (1.1) that allows for the presence of a sparse number (m) of s-wise interaction terms for some additional sparsity parameter s. Specifically, they studied this in the Gaussian white noise model[16]. Assuming f to lie in a Sobolev space with smoothness parameters $\beta, L > 0$, and some $\epsilon \in (0, 1)$[17], they derive a non-asymptotic L_2 error rate (in expectation) of: $\max\{mL^{\frac{s}{2\beta+s}}\epsilon^{\frac{4\beta}{2\beta+s}}, ms\epsilon^2 \log(d/(sm^{1/s}))\}$, which is also shown to be minimax optimal. However, they do not guarantee unique identification of the interaction terms for any value of s. Furthermore, the computational complexity of their estimator is exponential in d, s, m, although they discuss possible ways to reduce this complexity.

The above model was also recently studied by Yang et al. [61]; they consider a Bayesian estimation of f in the Gaussian process (GP) setting wherein a GP prior is placed on f, and inference on f is carried out by summarizing the resulting posterior probability given the data. They derived minimax estimation rates for Hölder smooth f in the L_2 norm, along with a method that nearly achieves the optimal estimation rate (modulo some log factors) in the empirical $L_2(\mathbb{P}_n)$ norm. However they do not guarantee unique identification of the interaction terms. Suzuki [50] studied a special case where $[d]$ is pre-divided into m disjoint subsets, with an additive component[18] defined on each subset. Assuming a sparse number of components, they derived PAC Bayesian bounds for estimation of f in the $L_2(\mathbb{P}_n)$ norm.

A special case of (1.1) – where ϕ_p's are linear and each $\phi_{(l,l')}$ is of the form $x_l x_{l'}$ – has been studied considerably. Within this setting, there exist algorithms that recover $\mathcal{S}_1, \mathcal{S}_2$, along with convergence rates for estimating f, in the limit of large n [8, 42, 3]. Kekatos et al. [24] show that exact recovery is possible (w.h.p) via ℓ_1 minimization with $O((|\mathcal{S}_1| + |\mathcal{S}_2|)(\log d)^4)$ noiseless point queries. This is based on the Restricted Isometry Property (RIP) for structured random matrices as developed in [44]. Nazer et al. [37] generalized this to the setting of sparse multilinear systems – albeit in the noiseless setting – and derived non-asymptotic sampling bounds for identifying the interaction terms, via ℓ_1 minimization. Upon translating Theorem 1.1 from their paper into our setting, with general overlap (so $\rho_m \geq 1$), we obtain a sample complexity[19] of $O((|\mathcal{S}_1| + |\mathcal{S}_2|)^2 \log(d/(|\mathcal{S}_1| + |\mathcal{S}_2|))) = O(k^2\rho_m^2 \log(d/(k\rho_m)))$. On the other hand, for the case of no overlap, their sample complexity turns out to be $O((|\mathcal{S}_1| + |\mathcal{S}_2|) \log(d/(|\mathcal{S}_1| + |\mathcal{S}_2|))) = O(k \log(d/k))$ for recovering $\mathcal{S}_1, \mathcal{S}_2$ w.h.p. However finite sample bounds for the non-linear model (1.1) are not known in general.

We also note that it is common in the statistics literature to impose a heredity constraint on the interactions, wherein an interaction term is present only if the corresponding main effect terms (i.e. those in $\mathcal{S}_1$) are present (cf., [8, 42, 3]). This is typically done to make the model interpretable, as interaction terms are difficult to interpret compared to main effect terms.

Other low-dimensional function models. We now provide a comparison with existing work related to other low dimensional models from the literature, starting with the approximation theoretic setting. Devore et al. [15] consider functions depending on a small subset $\mathcal{S}$ of the variables. The functions do not necessarily possess an additive structure, thus the setting is more general than (1.1). They provide algorithms that recover $\mathcal{S}$ exactly w.h.p, with $O(c^k k \log d)$ noiseless queries of f, for some constant $c > 1$. Their methods essentially make use of a (d, k)-hash family: $\mathcal{H}_k^d$ (cf. Definition 1). for constructing their sampling sets, and while these methods could be used for identifying $\mathcal{S}$, the sample complexity would be exponential in k.

Schnass et al. [46] consider the same model for f in the noiseless setting, and derive a simple algorithm that recovers $\mathcal{S}$ w.h.p, with $O(\frac{C_1^4}{\alpha^4} k (\log d)^2)$ noiseless queries. Here, $C_1 = \max_{i \in \mathcal{S}} \parallel \partial_i f \parallel_\infty$ and $\alpha = \min_{i \in \mathcal{S}} \parallel \partial_i f \parallel_1$ with $\parallel \cdot \parallel_1$ denoting the L_1 norm. While the C_1 term is a constant depending on the smoothness of f, one can construct examples of f for which $\alpha = c^{-k}$, for some constant $c > 1$. This implies that the sample bounds could be exponential in k for general k variate functions (as one would expect). This method could be applied to (1.1), to learn the set of active variables. In particular, for the general overlap case ($\rho_m \geq 1$), their algorithm will identify the support $\mathcal{S}$ w.h.p, with $O(\frac{{C_1'}^4 \rho_m^4}{\alpha^4} k (\log d)^2)$ noiseless queries where now: $C_1 = \max_{i \in \mathcal{S}} \parallel \partial_i f \parallel_\infty \leq C_1' \rho_m$, with C_1' a constant depending on the smoothness of f. For the general overlap case, we see that their bounds in the noiseless setting are worse by a ρ_m^3 factor compared to those for Algorithms 3, 4, however better by a $\log d$ factor compared to Algorithm 3. Moreover, it is not clear how the α term scales with respect to ρ_m here. For the non-overlap

[15] Here, we set the term b capturing the minimum magnitude of the univariate and bivariate components (as defined in [42, Section 3.2]) to O(1).

[16] This is known to be asymptotically equivalent to the nonparametric regression model as the number of samples $n \to \infty$.

[17] ϵ corresponds to $\sigma^2/\sqrt{n}$ in regression, where σ^2 denotes variance of noise.

[18] Thus for $m = d$, we obtain a Sparse additive model (SPAM).

[19] This sample complexity implies exact recovery of $\mathcal{S}_1, \mathcal{S}_2$ w.h.p

case, the scaling of their sampling bounds with respect to k, d matches ours for the noiseless setting, up to an additional $\frac{C_1^4}{\alpha^4}$ term. While α does not depend on k now, their sampling bound increases for small values of α or large values of C_1. The dependence of the sampling bound on the parameters C_1, α is not necessary in the noiseless setting, as seen from our sampling bounds that (in the noiseless case) depend on the *measure* of the region where $\partial_i f$ ($i \in \mathcal{S}$) and/or $\partial_l \partial_{l'} f$ ($(l, l') \in \mathcal{S}_2$) are large.

This model was considered by Comminges et al. [12, 11] in the regression setting. Assuming f to be differentiable, and the joint density of the covariates to be known, they propose an estimator that identifies the unknown subset $\mathcal{S}$ w.h.p, with sample complexity $O(c^k k \log d)$. This bound is shown to be tight although the estimator that achieves it is impractical – in the worst case it looks at all subsets of $\{1, \ldots, d\}$ of size k.

Fornasier et al. [18], Tyagi et al. [53] generalized this model class to functions f of the form $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$, for unknown $\mathbf{A} \in \mathbb{R}^{k \times d}$. They derive algorithms that approximately recover the row-span of $\mathbf{A}$, with sample complexities[20] typically *polynomial* in k, d. Specifically, [18] considers the setting where the rows of $\mathbf{A}$ are sparse. They propose a method that essentially estimates the gradient of f – via ℓ_1 minimization – at suitably (typically polynomially in d) many points on the unit sphere $\mathbb{S}^{d-1}$. [53] generalized this result to the setting where $\mathbf{A}$ is not necessarily sparse, by making use of low rank matrix recovery techniques.

Estimation of sparse Hessian matrices. There exists related work for estimating sparse Hessian matrices in the optimization literature. Powell et al. [41] and Coleman et al. [10] consider the setting where the sparsity structure of $\nabla^2 f(\mathbf{x})$ *is known*, and aim to estimate $\nabla^2 f(\mathbf{x})$ via gradient differences. Their aim is to minimize the number of gradient evaluations, needed for this purpose. In particular, Coleman et al. [10] approach the problem from a graph theoretic point of view and provide a graph coloring interpretation. Bandeira et al. [1] consider derivative free optimization (DFO) problems, wherein they approximate the underlying objective function f, by a quadratic polynomial interpolation model. Specifically, they build such a model by assuming $\nabla^2 f$ to be sparse, but do not assume the sparsity pattern to be known. Their approach is to minimize the ℓ_1 norm of the entries of the model Hessian, subject to interpolation conditions. As they do not assume ∇f to be sparse, they arrive at a sampling bound of $O(d(\log d)^4)$ [1, Corollary 4.1], for recovering $\nabla f(\mathbf{x})$, $\nabla^2 f(\mathbf{x})$, with high probability. In case ∇f were also sparse, one can verify that their bound changes to $O((|\mathcal{S}| + 2|\mathcal{S}_2|)(\log(|\mathcal{S}| + 2|\mathcal{S}_2|))^2(\log d)^2) = O(k\rho_m(\log(k\rho_m))^2(\log d)^2)$. They essentially make use of the Restricted Isometry Property (RIP) for structured random matrices as outlined in Theorem 4.4 of [44].

Bounded orthonormal systems. One of the reviewers pointed out another interesting approach that could be used for identifying $\mathcal{S}_1, \mathcal{S}_2$, that we now discuss. Note that this is only a rough sketch and verifying the details is left for future work. Let $\psi_k(\mathbf{x})$ be a bounded orthonormal system[21] in $L_2([-1,1]^d)$, for $k = 0, 1, \ldots, N$, consisting of univariate and bivariate functions. This could for example be constructed using a subset of the real trigonometric basis functions (see [13, Section 1.2]), with the ψ_k's satisfying the zero (marginal) mean conditions. In our model, there are a total of d univariate and $\binom{d}{2}$ bivariate functions. Say we take N_1 basis functions per coordinate, and N_2 basis functions per coordinate-tuple, so that $N = dN_1 + \binom{d}{2}N_2$.

Now, $f(\mathbf{x}) = \sum_{k=0}^{N} \alpha_k \psi_k(\mathbf{x}) + r(\mathbf{x})$ where r denotes the remainder term. Since f is C^3 smooth, we can uniformly approximate each univariate and bivariate ϕ with error rates: $N_1^{-p_1}$ (for some $p_1 > 0$) and $N_2^{-p_2}$ (for some $p_2 > 0$) respectively. Using triangle inequality, we then obtain for any $\mathbf{x} \in [-1,1]^d$ the bound:

$$|r(\mathbf{x})| \lesssim |\mathcal{S}_1| N_1^{-p_1} + |\mathcal{S}_2| N_2^{-p_2}. \tag{8.1}$$

So for bounding $|r(\mathbf{x})|$ by a sufficiently small constant, we require $N_1 \sim |\mathcal{S}_1|^{\frac{1}{p_1}}$ and $N_2 \sim |\mathcal{S}_2|^{\frac{1}{p_1}}$. By querying f at $\mathbf{x}_1, \ldots, \mathbf{x}_m$ (sampled uniformly at random), we get $y_l = f(\mathbf{x}_l) + z_l$; $l = 1, \ldots, m$, which in matrix form can be written as $y = \mathbf{A}\alpha + \mathbf{e}$. Here, $e_l = z_l + r_l(\mathbf{x})$ and, $\alpha \in \mathbb{R}^N$ is $N_1|\mathcal{S}_1| + N_2|\mathcal{S}_2|$ sparse. Since the rows of $\mathbf{A}$ correspond to a bounded orthonormal system (BOS), one can recover α via ℓ_1 minimization[22]; using the RIP result for BOS [44, Theorem 4.4], we obtain the bound:

$$m \geq C_1(|\mathcal{S}_1|N_1 + |\mathcal{S}_2|N_2)\log^2(|\mathcal{S}_1|N_1 + |\mathcal{S}_2|N_2)\log^2(dN_1 + \binom{d}{2}N_2) \tag{8.2}$$

$$\gtrsim (|\mathcal{S}_1|^{\frac{1}{p_1}+1} + |\mathcal{S}_2|^{\frac{1}{p_2}+1})\log^2(|\mathcal{S}_1|^{\frac{1}{p_1}+1} + |\mathcal{S}_2|^{\frac{1}{p_2}+1})\log^2(d). \tag{8.3}$$

Note that the above bound is super-linear in the sparsity: $|\mathcal{S}_1| + |\mathcal{S}_2|$ and this would be the case even when the samples are noiseless. In contrast, our bounds for Algorithms 1-4 are linear in sparsity, for the noiseless and bounded noise case. Also, observe that α is actually *block sparse*: it has $\binom{d}{2}$ "blocks", each of length N_2, out of which exactly $|\mathcal{S}_2|$ blocks are non-zero. Moreover, there are d blocks, each of length N_1, out of which $|\mathcal{S}_1|$ blocks are non-zero. While we are not aware of a RIP result for BOS with block sparsity[23], we would nevertheless still require $m \gtrsim (|\mathcal{S}_1|N_1 + |\mathcal{S}_2|N_2) \sim |\mathcal{S}_1|^{\frac{1}{p_1}+1} + |\mathcal{S}_2|^{\frac{1}{p_2}+1}$, which is super-linear in sparsity. For the setting of Gaussian noise however, it is possible that the above approach might give a better scaling with k, ρ_m compared to our results.

[20] These were derived predominantly in the noiseless setting, with some discussion in [53] about handling Gaussian noise via resampling and averaging.

[21] $\psi_0 \equiv 1$, i.e., it is the constant function.

[22] Consequently, we would be able to identify $\mathcal{S}_1, \mathcal{S}_2$ by thresholding.

[23] The existing ones seem to be only for matrices with i.i.d sub-Gaussian entries.

9 Concluding remarks

In this paper, we considered a generalization of Sparse Additive Models, of the form (1.1), now also allowing for the presence of a small number of bivariate components. We started with the special case where each variable interacts with at most one other variable, and then moved on to the general setting where variables can possibly be part of more than interaction term. For each of these settings, we derived algorithms with sample complexity bounds – both in the noiseless as well as the noisy query settings. For the general overlap case, the identification of the interaction set $\mathcal{S}_2$ essentially involved the estimation of the $d \times d$ Hessian of f at carefully chosen points. In fact, these points were simply part of a collection of canonical two dimensional uniform grids, within $[-1, 1]^d$. Upon identifying $\mathcal{S}_2$, the estimation of $\mathcal{S}_1$ was subsequently performed by employing the sampling scheme of Tyagi et al. [54] on the reduced set of variables. Furthermore, once $\mathcal{S}_1, \mathcal{S}_2$ are identified, we showed how one can recover uniform approximations to the individual components of the model, by additionally querying f along the one/two dimensional subspaces corresponding to $\mathcal{S}_1, \mathcal{S}_2$.

For the setting of noiseless queries, we observed that the sample complexity of Algorithm 4 is close to optimal. However for the noisy setting – in particular the setting of Gaussian noise – we saw that the sample complexity of Algorithm 4 has a worse dependency in terms of k, ρ_m compared to Algorithm 3. In general, the sample complexity bounds of our algorithms, in the presence of Gaussian noise, have a sub optimal dependence on k, ρ_m. This is mainly due to the localized nature of our sampling schemes – the external noise gets scaled by the step size parameter leading to the noise variance scaling up. Hence the number of samples required to reduce the noise variance (by resampling and averaging) increases, leading to an increase in the total sample complexity. An interesting direction for future work would be to consider alternate – possibly non localized sampling schemes – with improved non-asymptotic sampling bounds for identifying $\mathcal{S}_1, \mathcal{S}_2$ in the setting of Gaussian noise.

Another limitation of our analysis is that it is restricted to C^3 smooth functions. It would be interesting to extend the results to more general C^r smooth functions $r \geq 1$ and also to other smoothness classes such as Hölder/Lipschitz continuous functions. Lastly, we only consider pairwise interactions between the variables; a natural generalization would be to consider a model that can include components which are at most m-variate. The goal would then be to query f, in order to identify all interaction terms.

Acknowledgments. This research was supported in part by SNSF grant CRSII2_147633 and by The Alan Turing Institute under the EPSRC grant EP/N510129/1. This work was mostly done while H.T was affiliated to the Department of Computer Science, ETH Zürich. H.T would like to thank: Yuxin Chen for helpful discussions related to the recovery of sparse symmetric matrices in Section 5.1; Jan Vybiral for helpful discussions related to bounded orthonormal systems in Section 8. The authors would like to thank the anonymous reviewers for helpful comments and suggestions that greatly helped to improve a preliminary version of the manuscript.

References

[1] A.S. Bandeira, K. Scheinberg, and L.N. Vicente. Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization. *Mathematical Programming*, 134(1):223–257, 2012.

[2] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 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] T. Blumensath and M.E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265 – 274, 2009.

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

[6] E.J. Candès, J.K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

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

[8] N.H. Choi, W. Li, and J. Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364, 2010.

[9] A. Cohen, I. Daubechies, R.A. DeVore, G. Kerkyacharian, and D. Picard. Capturing ridge functions in high dimensions from point queries. *Constr. Approx.*, pages 1–19, 2011.

[10] T.F. Coleman and J.J. Moré. Estimation of sparse Hessian matrices and graph coloring problems. *Mathematical Programming*, 28(3):243–270, 1984.

[11] 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.

[12] 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.

[13] 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.

[14] C. de Boor. *A practical guide to splines*. Springer Verlag (New York), 1978.

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

[16] D.L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[17] J. Fan and I. Gijbels. *Local polynomial modeling and its applications*. Chapman & Hall, London, New York, 1996.

[18] M. Fornasier, K. Schnass, and J. Vybíral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 12(2):229–262, 2012.

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

[20] M. Fredman and J. Komlos. On the size of separating systems and families of perfect hash functions. *SIAM J. Algebr. Discrete Methods*, 5:61–68, 1984.

[21] C. Gu. *Smoothing Spline ANOVA Models*. Springer (New York), 2002.

[22] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

[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] J. Korner and K. Martin. New bounds for perfect hashing via information theory. *Eur. J. Combin.*, 9:523–530, 1988.

[28] A. Kyrillidis and V. Cevher. Recipes on hard thresholding methods. In 4^{th} *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 353–356, 2011.

[29] A. Kyrillidis and V. Cevher. Combinatorial selection and least absolute shrinkage via the CLASH algorithm. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2216–2220, 2012.

[30] A. Kyrillidis, G. Puy, and V. Cevher. Hard thresholding with norm constraints. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3645–3648, 2012.

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

[32] M.H. Maathuis, M. Kalisch, and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *Ann. Statist.*, 37(6A):3133–3164, 2009.

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

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

[35] Th. Muller-Gronbach and K. Ritter. Minimal errors for strong and weak approximation of stochastic differential equations. *Monte Carlo and Quasi-Monte Carlo Methods*, pages 53–82, 2008.

[36] M. Naor, L.J. Schulman, and A. Srinivasan. Splitters and near-optimal derandomization. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science, 1995.*, pages 182–191, 1995.

[37] 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.

[38] A. Nemirovski. *Topics in non-parametric statistics*. In Ecole d'Et'e de Probabilitès de Saint-Flour XVIII, 1998, 85-277, Springer, New York, 2000.

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

[40] A. Nilli. Perfect hashing and probability. *Combinatorics, Probability and Computing*, 3:407–409, 1994.

[41] M.J.D. Powell and Ph. L. Toint. On the estimation of sparse Hessian matrices. *SIAM Journal on Numerical Analysis*, 16(6):pp. 1060–1074, 1979.

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

[43] 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.

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

[45] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

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

[47] V.I Smirnov. *A course of higher mathematics*. Addison-Wesley, Reading, MA, 1964.

[48] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

[49] 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.

[50] T. Suzuki. PAC-Bayesian bound for Gaussian process regression and multiple kernel additive model. In *25th Annual Conference on Learning Theory (COLT)*, pages 8.1–8.20, 2012.

[51] J.F. Traub, G.W. Wasilkowski, and H. Wozniakowski. *Information-Based Complexity*. Academic Press, New York, 1988.

[52] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.

[53] H. Tyagi and V. Cevher. Active learning of multi-index function models. In *Advances in Neural Information Processing Systems (NIPS) 25*, pages 1475–1483. 2012.

[54] 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.

[55] 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.

[56] G. Wahba. An introduction to (smoothing spline) ANOVA models in RKHS, with examples in geographical data, medicine, atmospheric science and machine learning. *13th IFAC Symposium on System Identification, Rotterdam*, pages 549–559, 2003.

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

[58] M.J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory*, 55(12):5728–5741, 2009.

[59] M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theor.*, 55(5):2183–2202, May 2009.

[60] P. Wojtaszczyk. ℓ_1 minimization with noisy data. *SIAM J. Numer. Anal.*, 50(2):458–467, 2012.

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

A Model uniqueness

We show here that the model representation (2.4) is a unique representation for f of the form (2.1). We first note that any measurable $f : [-1,1]^d \to \mathbb{R}$ admits a unique ANOVA decomposition [21, 56] of the form:

$$f(x_1,\ldots,x_d) = c + \sum_{\alpha} f_\alpha(x_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta} + \sum_{\alpha<\beta<\gamma} f_{\alpha\beta\gamma} + \cdots \tag{A.1}$$

Indeed, for any probability measure μ_α on $[-1,1]$, let $\mathcal{E}_\alpha$ denote the averaging operator, defined as

$$\mathcal{E}_\alpha(f)(\mathbf{x}) := \int_{[-1,1]} f(x_1,\ldots,x_d)d\mu_\alpha. \tag{A.2}$$

Then the components of the model can be written as: $c = (\prod_\alpha \mathcal{E}_\alpha)f$, $f_\alpha = (I-\mathcal{E}_\alpha)\prod_{\beta\neq\alpha}\mathcal{E}_\beta f$, $f_{\alpha\beta} = ((I-\mathcal{E}_\alpha)(I-\mathcal{E}_\beta)\prod_{\gamma\neq\alpha,\beta}\mathcal{E}_\gamma)f$, and so on. For our purpose, μ_α is taken to be the uniform probability measure on $[-1,1]$. Given this, we now find the ANOVA decomposition of f defined in (2.1). As a sanity check, let us verify that $f_{\alpha\beta\gamma} \equiv 0$ for all $\alpha<\beta<\gamma$. Indeed if $p \in \mathcal{S}_1$, then at least two of $\alpha<\beta<\gamma$ will not be equal to p. Similarly for any $(l,l') \in \mathcal{S}_2$, at least one of α,β,γ will not be equal to l and l'. This implies $f_{\alpha\beta\gamma} \equiv 0$. The same reasoning trivially applies for high order components of the ANOVA decomposition.

That $c = \mathbb{E}[f] = \sum_{p\in\mathcal{S}_1}\mathbb{E}_p[\phi_p] + \sum_{(l,l')\in\mathcal{S}_2}\mathbb{E}_{(l,l')}[\phi_{(l,l')}]$ is readily seen. Next, we have that

$$(I-\mathcal{E}_\alpha)\prod_{\beta\neq\alpha}\mathcal{E}_\beta\phi_p = \left\{\begin{array}{rcl} 0 & ; & \alpha\neq p, \\ \phi_p - \mathbb{E}_p[\phi_p] & ; & \alpha = p \end{array}\right\}; \quad p\in\mathcal{S}_1. \tag{A.3}$$

$$(I-\mathcal{E}_\alpha)\prod_{\beta\neq\alpha}\mathcal{E}_\beta\phi_{(l,l')} = \left\{\begin{array}{rcl} \mathbb{E}_{l'}[\phi_{(l,l')}] - \mathbb{E}_{(l,l')}[\phi_{(l,l')}] & ; & \alpha = l, \\ \mathbb{E}_l[\phi_{(l,l')}] - \mathbb{E}_{(l,l')}[\phi_{(l,l')}] & ; & \alpha = l', \\ 0 & ; & \alpha\neq l,l', \end{array}\right\}; \quad (l,l')\in\mathcal{S}_2. \tag{A.4}$$

(A.3), (A.4) give us the first order components of $\phi_p, \phi_{(l,l')}$ respectively. One can next verify using the same arguments as earlier that for any $\alpha<\beta$:

$$(I-\mathcal{E}_\alpha)(I-\mathcal{E}_\beta)\prod_{\gamma\neq\alpha,\beta}\mathcal{E}_\gamma\phi_p = 0; \quad \forall p\in\mathcal{S}_1. \tag{A.5}$$

Lastly, we have for any $\alpha<\beta$ that the corresponding second order component of $\phi_{(l,l')}$ is given by:

$$(I-\mathcal{E}_\alpha)(I-\mathcal{E}_\beta)\prod_{\gamma\neq\alpha,\beta}\mathcal{E}_\gamma\phi_{(l,l')} = \left\{\begin{array}{rcl} \phi_{(l,l')} - \mathbb{E}_l[\phi_{(l,l')}] & & \\ -\mathbb{E}_{l'}[\phi_{(l,l')}] + \mathbb{E}_{(l,l')}[\phi_{(l,l')}] & ; & \alpha = l, \beta = l', \\ 0 & ; & \text{otherwise} \end{array}\right\}; \quad (l,l')\in\mathcal{S}_2. \tag{A.6}$$

We now make the following observations regarding the variables in $\mathcal{S}_1 \cap \mathcal{S}_2^{\text{var}}$.

1. For each $l \in \mathcal{S}_1 \cap \mathcal{S}_2^{\text{var}}$ such that: $\rho(l) = 1$, and $(l,l') \in \mathcal{S}_2$, we can simply merge ϕ_l with $\phi_{(l,l')}$. Thus l is no longer in $\mathcal{S}_1$.
2. For each $l \in \mathcal{S}_1 \cap \mathcal{S}_2^{\text{var}}$ such that: $\rho(l) > 1$, we can add the first order component for ϕ_l with the total first order component corresponding to all $\phi_{(l,l')}$'s and $\phi_{(l',l)}$'s. Hence again, l will no longer be in $\mathcal{S}_1$.

Therefore all $q \in \mathcal{S}_1 \cap \mathcal{S}_2^{\text{var}}$ can essentially be merged with $\mathcal{S}_2$. Keeping this re-arrangement in mind, we can to begin with, assume in (2.1) that $\mathcal{S}_1 \cap \mathcal{S}_2^{\text{var}} = \emptyset$. Then with the help of (A.3), (A.4), (A.5), (A.6), we have that any f of the form (2.1) (with $\mathcal{S}_1 \cap \mathcal{S}_2^{\text{var}} = \emptyset$), can be uniquely written as:

$$f(x_1,\ldots,x_d) = c + \sum_{p\in\mathcal{S}_1}\tilde{\phi}_p(x_p) + \sum_{(l,l')\in\mathcal{S}_2}\tilde{\phi}_{(l,l')}(x_l,x_{l'}) + \sum_{q\in\mathcal{S}_2^{\text{var}}:\rho(q)>1}\tilde{\phi}_q(x_q); \quad \mathcal{S}_1\cap\mathcal{S}_2^{\text{var}} = \emptyset, \tag{A.7}$$

where

$$c = \sum_{p\in\mathcal{S}_1}\mathbb{E}_p[\phi_p] + \sum_{(l,l')\in\mathcal{S}_2}\mathbb{E}_{(l,l')}[\phi_{(l,l')}], \tag{A.8}$$

$$\tilde{\phi}_p = \phi_p - \mathbb{E}_p[\phi_p]; \quad \forall p\in\mathcal{S}_1, \tag{A.9}$$

$$\tilde{\phi}_{(l,l')} = \left\{\begin{array}{rl} \phi_{(l,l')} - \mathbb{E}_{(l,l')}[\phi_{(l,l')}]; & \rho(l),\rho(l') = 1, \\ \phi_{(l,l')} - \mathbb{E}_l[\phi_{(l,l')}]; & \rho(l) = 1, \rho(l') > 1, \\ \phi_{(l,l')} - \mathbb{E}_{l'}[\phi_{(l,l')}]; & \rho(l) > 1, \rho(l') = 1, \\ \phi_{(l,l')} - \mathbb{E}_l[\phi_{(l,l')}] - \mathbb{E}_{l'}[\phi_{(l,l')}] + \mathbb{E}_{(l,l')}[\phi_{(l,l')}]; & \rho(l),\rho(l') > 1, \end{array}\right. \tag{A.10}$$

$$\text{and} \quad \tilde{\phi}_q = \sum_{q':(q,q')\in\mathcal{S}_2}(\mathbb{E}_{q'}[\phi_{(q,q')}] - \mathbb{E}_{q,q'}[\phi_{(q,q')}]) + \sum_{q':(q',q)\in\mathcal{S}_2}(\mathbb{E}_{q'}[\phi_{(q',q)}] - \mathbb{E}_{q',q}[\phi_{(q',q)}]); \quad \forall q\in\mathcal{S}_2^{\text{var}} : \rho(q) > 1. \tag{A.11}$$

B Real roots of a cubic equation in trigonometric form

Before proceeding with the proofs, we briefly recall the conditions under which a cubic equation possesses real roots, along with expressions for the same. The material in this section is taken from [47, Chapter 18 (Secs. 191,192)]. To begin with, given any cubic equation:

$$y^3 + a_1 y^2 + a_2 y + a_3 = 0, \tag{B.1}$$

one can make the substitution $x = y - (a_1/3)$ to change (B.1) to the form:

$$x^3 + px + q = 0 \quad \text{where} \quad p = a_2 - \frac{a_1^2}{3} \text{ and } q = \frac{2a_1^3}{27} - \frac{a_1 a_2}{3} + a_3. \tag{B.2}$$

If p, q are real (which is the case if a_1, a_2, a_3 are real), then (B.2) has three real and distinct roots if its discriminant: $(q^2/4) + (p^3/27) < 0$. Denoting

$$r = \sqrt{-\frac{p^3}{27}}, \quad \cos\phi = -\frac{q}{2r}, \tag{B.3}$$

we then have that the real roots of (B.2) are given by

$$x = 2\sqrt[3]{r}\cos\frac{\phi + 2j\pi}{3} = 2\sqrt{-\frac{p}{3}}\cos\frac{\phi + 2j\pi}{3}; \quad j = 0, 1, 2. \tag{B.4}$$

Consequently, the roots of (B.1) are then given by:

$$y = 2\sqrt{-\frac{p}{3}}\cos\frac{\phi + 2j\pi}{3} - \frac{a_1}{3}; \quad j = 0, 1, 2. \tag{B.5}$$

C Proofs for Section 3

C.1 Proof of Lemma 1

Recall that for $\mathbf{x} \in \chi$, we recover a stable approximation $\widehat{\nabla} f(\mathbf{x})$ to $\nabla f(\mathbf{x})$ via ℓ_1 minimization [6, 16]:

$$\widehat{\nabla} f(\mathbf{x}) = \triangle(\mathbf{y}) := \underset{\mathbf{y}=\mathbf{V}\mathbf{z}}{\operatorname{argmin}} \parallel \mathbf{z} \parallel_1 . \tag{C.1}$$

Applying Theorem 1 to our setting yields the following Corollary.

Corollary 1. *There exist constants $c_3' \geq 1$ and $C, c_1' > 0$ such that for m_v satisfying $c_3' k \log(d/k) < m_v < d/(\log 6)^2$ we have with probability at least $1 - e^{-c_1' m_v} - e^{-\sqrt{m_v d}}$ that $\widehat{\nabla} f(\mathbf{x})$ satisfies for all $\mathbf{x} \in \chi$:*

$$\parallel \widehat{\nabla} f(\mathbf{x}) - \nabla f(\mathbf{x}) \parallel_2 \leq \frac{2C\mu^2 B_3 k}{3m_v}, \tag{C.2}$$

where $B_3 > 0$ is the constant defined in Assumption 2.

Proof. Since $\nabla f(\mathbf{x})$ is at most k-sparse for any $\mathbf{x} \in \mathbb{R}^d$ we immediately have from (3.10) that

$$\parallel \widehat{\nabla} f(\mathbf{x}) - \nabla f(\mathbf{x}) \parallel_2 \leq C \max\left\{ \parallel \mathbf{n} \parallel_2, \sqrt{\log d} \parallel \mathbf{n} \parallel_\infty \right\}; \quad \forall \mathbf{x} \in \chi. \tag{C.3}$$

It remains to bound $\parallel \mathbf{n} \parallel_2, \parallel \mathbf{n} \parallel_\infty$. To this end, recall that $\mathbf{n} = [n_1 \dots n_{m_v}]$ where $n_j = \frac{R_3(\zeta_j) - R_3(\zeta_j')}{2\mu}$, for some $\zeta_j, \zeta_j' \in \mathbb{R}^d$. Here $R_3(\zeta)$ denotes the third order Taylor remainder term. By taking the structure of f into account, we can uniformly bound $|R_3(\zeta_j)|$ as follows (so the same bound holds for $|R_3(\zeta_j')|$).

$$\begin{aligned}
|R_3(\zeta_j)| &= \frac{\mu^3}{6} | \sum_{p \in \mathcal{S}_1} \partial_p^3 \phi_p(\zeta_{j,p}) v_p^3 + \sum_{(l,l') \in \mathcal{S}_2} (\partial_l^3 \phi_{(l,l')}(\zeta_{j,l}, \zeta_{j,l'}) v_l^3 + \partial_{l'}^3 \phi_{(l,l')}(\zeta_{j,l}, \zeta_{j,l'}) v_{l'}^3) && \text{(C.4)} \\
&\quad + \sum_{(l,l') \in \mathcal{S}_2} (3\partial_l \partial_{l'}^2 \phi_{(l,l')}(\zeta_{j,l}, \zeta_{j,l'}) v_l v_{l'}^2 + 3\partial_l^2 \partial_{l'} \phi_{(l,l')}(\zeta_{j,l}, \zeta_{j,l'}) v_l^2 v_{l'}) |, \\
&\leq \frac{\mu^3}{6} \left[\left(\frac{1}{\sqrt{m_v}} \right)^3 k_1 B_3 + \left(\frac{1}{\sqrt{m_v}} \right)^3 k_2 (2B_3) + \left(\frac{1}{\sqrt{m_v}} \right)^3 k_2 (6B_3) \right], && \text{(C.5)} \\
&= \frac{\mu^3 B_3 (k_1 + 8k_2)}{6 m_v^{3/2}}. && \text{(C.6)}
\end{aligned}$$

Using the fact that $k_1 + 8k_2 \leq 4(k_1 + 2k_2) = 4k$, we consequently obtain

$$\| \mathbf{n} \|_\infty = \max_j |n_j| \leq \frac{\mu^2 B_3 (k_1 + 8k_2)}{6 m_v^{3/2}} \leq \frac{2\mu^2 B_3 k}{3 m_v^{3/2}}, \tag{C.7}$$

$$\text{and} \quad \| \mathbf{n} \|_2 \leq \sqrt{m_v} \, \| \mathbf{n} \|_\infty \leq \frac{2\mu^2 B_3 k}{3 m_v}. \tag{C.8}$$

Using (C.7),(C.8) in (C.3), we finally obtain for the stated choice of m_v (cf. Remark 4), the bound in (C.2). □

Let us denote $\tau = \frac{2C\mu^2 B_3 k}{3m_v}$. In order to prove the lemma, we first observe that (C.2) trivially implies that

$$\widehat{\partial}_q f(\mathbf{x}) \in [\partial_q f(\mathbf{x}) - \tau, \partial_q f(\mathbf{x}) + \tau]; \quad q = 1, \ldots, d. \tag{C.9}$$

Now, in case $q \notin \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$, then $\partial_q f(\mathbf{x}) = 0 \, \forall \mathbf{x} \in \mathbb{R}^d$, meaning that $\widehat{\partial}_q f(\mathbf{x}) \in [-\tau, \tau]$. If $m_x \geq \lambda_1^{-1}$ then for every $q \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$, $\exists h \in \mathcal{H}_2^d$ and at least one $\mathbf{x} \in \chi(h)$, so that $|\partial_q f(\mathbf{x})| > D_1$. Indeed, this follows from the definition of $\mathcal{H}_2^d$, and by construction of $\chi(h)$ for $h \in \mathcal{H}_2^d$. Furthermore, for such $\mathbf{x}$, we have from (C.9) that $|\widehat{\partial}_q f(\mathbf{x})| \geq D_1 - \tau$. Therefore if $\tau < \frac{D_1}{2}$ holds, then clearly we would have $|\widehat{\partial}_q f(\mathbf{x})| > \frac{D_1}{2} > \tau$, meaning that we will be able to identify q.

Lastly, we observe that the condition $\tau < \frac{D_1}{2}$ translates to an equivalent condition on the step size μ as follows.

$$\tau < \frac{D_1}{2} \Leftrightarrow \frac{2C\mu^2 B_3 k}{3m_v} < \frac{D_1}{2} \Leftrightarrow \mu < \left(\frac{3 D_1 m_v}{4 C B_3 k} \right)^{1/2} \tag{C.10}$$

C.2 Proof of Lemma 2

We proceed by first bounding the error term that arises in the estimation of $\partial_i g(\mathbf{x})$. As g is $\mathcal{C}^3$ smooth, consider the Taylor's expansion of g at $\mathbf{x}$, along $\mathbf{e}_1(i), -\mathbf{e}_1(i) \in \mathbb{R}^k$, with step size $\beta > 0$. For some $\zeta = \mathbf{x} + \theta \mathbf{e}_1(i)$, $\zeta' = \mathbf{x} - \theta' \mathbf{e}_1(i)$ with $\theta, \theta' \in (0, \beta)$, we obtain the identities:

$$g(\mathbf{x} + \beta \mathbf{e}_1(i)) = g(\mathbf{x}) + \beta \langle \mathbf{e}_1(i), \nabla g(\mathbf{x}) \rangle + \frac{\beta^2}{2} \mathbf{e}_1(i)^T \nabla^2 g(\mathbf{x}) \mathbf{e}_1(i) + R_3(\zeta), \tag{C.11}$$

$$g(\mathbf{x} - \beta \mathbf{e}_1(i)) = g(\mathbf{x}) - \beta \langle \mathbf{e}_1(i), \nabla g(\mathbf{x}) \rangle + \frac{\beta^2}{2} \mathbf{e}_1(i)^T \nabla^2 g(\mathbf{x}) \mathbf{e}_1(i) + R_3(\zeta'), \tag{C.12}$$

with $R_3(\zeta), R_3(\zeta') = O(\beta^3)$ being the third order remainder terms. Subtracting the above leads to the following identity.

$$\underbrace{\frac{g(\mathbf{x} + \beta \mathbf{e}_1(i)) - g(\mathbf{x} - \beta \mathbf{e}_1(i))}{2\beta}}_{\widehat{\partial}_i g(\mathbf{x})} = \underbrace{\langle \mathbf{e}_1(i), \nabla g(\mathbf{x}) \rangle}_{\partial_i g(\mathbf{x})} + \underbrace{\frac{R_3(\zeta) - R_3(\zeta')}{2\beta}}_{\eta_i(\mathbf{x}, \beta) = O(\beta^2)} \tag{C.13}$$

We now uniformly bound $|R_3(\zeta)|$, so the same bound holds for $|R_3(\zeta')|$. Due to the structure of g, we have that

$$|R_3(\zeta)| = \frac{\beta^3}{6} | \sum_{p \in \mathcal{S}_1} \partial_p^3 \phi_p(\zeta_p) (\mathbf{e}_1(i))_p^3 + \sum_{(l,l') \in \mathcal{S}_2} (\partial_l^3 \phi_{(l,l')}(\zeta_l, \zeta_{l'}) (\mathbf{e}_1(i))_l^3 + \partial_{l'}^3 \phi_{(l,l')}(\zeta_l, \zeta_{l'}) (\mathbf{e}_1(i))_{l'}^3) \tag{C.14}$$

$$+ \sum_{(l,l') \in \mathcal{S}_2} (3 \partial_l^2 \partial_{l'} \phi_{(l,l')}(\zeta_l, \zeta_{l'}) (\mathbf{e}_1(i))_l^2 (\mathbf{e}_1(i))_{l'} + 3 \partial_{l'}^2 \partial_l \phi_{(l,l')}(\zeta_l, \zeta_{l'}) (\mathbf{e}_1(i))_{l'}^2 (\mathbf{e}_1(i))_l) | \tag{C.15}$$

$$= \begin{cases} \frac{\beta^3}{6} |\partial_i^3 \phi_i(\zeta_i)|; & i \in \mathcal{S}_1, \\ \frac{\beta^3}{6} |\partial_i^3 \phi_{i,j}(\zeta_i, \zeta_j)|; & i \in \mathcal{S}_2^{\text{var}}, (i,j) \in \mathcal{S}_2, \\ \frac{\beta^3}{6} |\partial_i^3 \phi_{j,i}(\zeta_j, \zeta_i)|; & i \in \mathcal{S}_2^{\text{var}}, (j,i) \in \mathcal{S}_2 \end{cases} \tag{C.16}$$

$$\leq \frac{\beta^3 B_3}{6}. \tag{C.17}$$

The above consequently implies that $|\eta_i(\mathbf{x}, \beta)| \leq \frac{\beta^2 B_3}{6}$. This in turn means, for any $\mathbf{v} \in \mathbb{R}^k, \mu_1 > 0$, that

$$| \frac{\eta_i(\mathbf{x} + \mu_1 \mathbf{v}, \beta) - \eta_i(\mathbf{x}, \beta)}{\mu_1} | \leq \frac{\beta^2 B_3}{3 \mu_1}. \tag{C.18}$$

Thus we have a uniform bound on the magnitude of one of the contributors of the error term in (3.18). We can bound the magnitude of the other term as follows. For $\mathbf{v} \in \mathbb{R}^k$ and $\zeta = \mathbf{x} + \theta \mathbf{v}$; $\theta \in (0, \mu_1)$, we have

$$\mathbf{v}^T \nabla^2 \partial_i g(\zeta) \mathbf{v} = \begin{cases} v_i^2 \partial_i^3 \phi_i(\zeta_i); & i \in \mathcal{S}_1, \\ v_i^2 \partial_i^3 \phi_{i,i'}(\zeta_i, \zeta_{i'}) + v_{i'}^2 \partial_{i'}^2 \partial_i \phi_{i,i'}(\zeta_i, \zeta_{i'}) + 2 v_i v_{i'} \partial_{i'} \partial_i^2 \phi_{i,i'}(\zeta_i, \zeta_{i'}); & i \in \mathcal{S}_2^{\text{var}}, (i,i') \in \mathcal{S}_2 \end{cases} \tag{C.19}$$

Since in our scheme we employ only $\mathbf{v} \in \{0,1\}^k$, this leads to the following uniform bound.

$$|\frac{\mu_1}{2}\mathbf{v}^T\nabla^2\partial_i g(\zeta)\mathbf{v}| \leq 4B_3\frac{\mu_1}{2} = 2\mu_1 B_3; \quad \forall i \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}. \tag{C.20}$$

Denoting by τ', the upper bound on the magnitude of the error term in (3.18), we thus obtain:

$$\tau' = 2\mu_1 B_3 + \frac{\beta^2 B_3}{3\mu_1}. \tag{C.21}$$

Now in case $i \in \mathcal{S}_1$, we have $\langle \nabla\partial_i g(\mathbf{x}), \mathbf{v}_0(i)\rangle = 0, \forall \mathbf{x} \in \mathbb{R}^k$. This in turn implies that

$$|\frac{\widehat{\partial}_i g(\mathbf{x} + \mu_1\mathbf{v}_0(i)) - \widehat{\partial}_i g(\mathbf{x})}{\mu_1}| \leq \tau'; \quad \forall \mathbf{x} \in \mathbb{R}^k. \tag{C.22}$$

If $i \in \mathcal{S}_2^{\text{var}}$ with $(i,i') \in \mathcal{S}_2$, then

$$\langle \nabla\partial_i g(\mathbf{x}), \mathbf{v}_0(i)\rangle = \partial_i\partial_{i'} g(\mathbf{x}) = \partial_i\partial_{i'}\phi_{(i,i')}(x_i, x_{i'}). \tag{C.23}$$

For the choice $m'_x > \lambda_2^{-1}$, $\exists \mathbf{x}^* \in \chi_i$ such that $|\partial_i\partial_{i'}\phi_{(i,i')}(x_i^*, x_{i'}^*)| > D_2$. This is clear from the construction of χ_i, and on account of Assumption 3. If we guarantee that $\tau' < D_2/2$ holds, then consequently

$$|\frac{\widehat{\partial}_i g(\mathbf{x}^* + \mu_1\mathbf{v}_0(i)) - \widehat{\partial}_i g(\mathbf{x}^*)}{\mu_1}| > D_2 - \tau' > \tau' \tag{C.24}$$

meaning that the pair (i,i') can be identified. Lastly, it is easily verifiable, that the requirement $\tau' < D_2/2$, equivalently translates to the stated conditions on β, μ.

C.3 Proof of Theorem 2

We begin by first establishing the conditions that guarantee $\widehat{\mathcal{S}} = \mathcal{S}$, and then derive conditions that guarantee exact recovery of $\mathcal{S}_1, \mathcal{S}_2$.

Estimation of $\mathcal{S}$. We first note that (3.6) now changes to $\mathbf{y} = \mathbf{V}\nabla f(\mathbf{x}) + \mathbf{n} + \mathbf{z}$ where $z_j = (z'_{j,1} - z'_{j,2})/(2\mu)$ represents the external noise component, for $j = 1, \ldots, m_v$. Since $\| \mathbf{z} \|_\infty \leq \varepsilon/\mu$, therefore using the bounds on $\| \mathbf{n} \|_\infty$ from Section C.1 one can verify that (C.2) in Corollary 1 changes to

$$\| \widehat{\nabla} f(\mathbf{x}) - \nabla f(\mathbf{x}) \|_2 \leq C\left(\frac{2\mu^2 B_3 k}{3m_v} + \frac{\varepsilon\sqrt{m_v}}{\mu}\right). \tag{C.25}$$

Following the same arguments mentioned in Section C.1, we observe that if $\tau < D_1/2$ holds, then it implies that $\widehat{\mathcal{S}} = \mathcal{S}$. Now, $\tau < D_1/2$ is equivalent to

$$\underbrace{\frac{2\mu^2 B_3 k}{3m_v}}_{a\mu^2} + \underbrace{\frac{\varepsilon\sqrt{m_v}}{\mu}}_{\frac{b\varepsilon}{\mu}} < \frac{D_1}{2C} \quad \Leftrightarrow \quad \mu^3 - \frac{D_1}{2aC}\mu + \frac{b\varepsilon}{a} < 0. \tag{C.26}$$

(C.26) is a cubic inequality. Recall from Section B that a cubic equation of the form: $y^3 + py + q = 0$, has 3 distinct real roots if its discriminant $\frac{p^3}{27} + \frac{q^2}{4} < 0$. Note that for this to be possible, p must be negative, which is the case in (C.26). Applying the discriminant condition on (C.26) leads to

$$-\frac{D_1^3}{27 \cdot 8a^3C^3} + \frac{b^2}{4a^2}\varepsilon^2 < 0 \quad \Leftrightarrow \quad \varepsilon < \frac{D_1^{3/2}}{3bC\sqrt{6aC}}. \tag{C.27}$$

Also, recall from (B.4) that the 3 distinct real roots of the cubic equation are then given by:

$$y_1 = 2\sqrt{-p/3}\cos(\theta/3),\ y_2 = -2\sqrt{-p/3}\cos(\theta/3 + \pi/3),\ y_3 = -2\sqrt{-p/3}\cos(\theta/3 - \pi/3) \tag{C.28}$$

where $\theta = \cos^{-1}\left(\frac{-q/2}{\sqrt{-p^3/27}}\right)$. In particular, if $q > 0$, then one can verify that $y^3 + py + q < 0$ holds if $y \in (y_2, y_1)$. Applying this to the cubic equation corresponding to (C.26), we consequently obtain:

$$\mu \in \left(2\sqrt{\frac{D_1}{6aC}}\cos(\theta_1/3 - 2\pi/3), 2\sqrt{\frac{D_1}{6aC}}\cos(\theta_1/3)\right). \tag{C.29}$$

where $\theta_1 = \cos^{-1}(-\varepsilon/\varepsilon_1)$.

Estimation of $\mathcal{S}_1, \mathcal{S}_2$. On account of noise, we first note that (C.13) changes to

$$\underbrace{\frac{g(\mathbf{x}+\beta\mathbf{e}_1(i))-g(\mathbf{x}-\beta\mathbf{e}_1(i))}{2\beta}}_{\widehat{\partial}_i g(\mathbf{x})} = \underbrace{\langle \mathbf{e}_1(i), \nabla g(\mathbf{x})\rangle}_{\partial_i g(\mathbf{x})} + \underbrace{\frac{R_3(\zeta)-R_3(\zeta')}{2\beta}}_{\eta_i(\mathbf{x},\beta)=O(\beta^2)} + \underbrace{\frac{z'_{i,1}-z'_{i,2}}{2\beta}}_{z_i(\mathbf{x},\beta)}. \tag{C.30}$$

This in turn results in (3.18) changing to

$$\frac{\widehat{\partial}_i g(\mathbf{x}+\mu_1\mathbf{v})-\widehat{\partial}_i g(\mathbf{x})}{\mu_1} = \langle \nabla\partial_i g(\mathbf{x}), \mathbf{v}\rangle + \underbrace{\frac{\mu_1}{2}\mathbf{v}^T\nabla^2\partial_i g(\zeta_i)\mathbf{v} + \frac{\eta_i(\mathbf{x}+\mu_1\mathbf{v},\beta)-\eta_i(\mathbf{x},\beta)}{\mu_1} + \frac{z_i(\mathbf{x}+\mu_1\mathbf{v},\beta)-z_i(\mathbf{x},\beta)}{\mu_1}}_{\text{Error term}}. \tag{C.31}$$

Using (C.18), (C.20) and noting that $|(z_i(\mathbf{x}+\mu_1\mathbf{v},\beta)-z_i(\mathbf{x},\beta))/\mu_1| \leq 2\varepsilon/(\beta\mu_1)$, then by denoting τ' to be an upper bound on the magnitude of the error term in (C.31), we have that $\tau' = 2\mu_1 B_3 + \frac{\beta^2 B_3}{3\mu_1} + \frac{2\varepsilon}{\beta\mu_1}$. Following the same argument as in Section C.2, we have that $\tau' < D_2/2$ implies $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ and $\widehat{\mathcal{S}_2} = \mathcal{S}_2$. The condition $\tau' < D_2/2$ is equivalent to

$$2\mu_1 B_3 + \frac{\beta^2 B_3}{3\mu_1} + \frac{2\varepsilon}{\beta\mu_1} < \frac{D_2}{2} \tag{C.32}$$

$$\Leftrightarrow 6B_3\beta\mu_1^2 - \frac{3D_2}{2}\beta\mu_1 + (\beta^3 B_3 + 6\varepsilon) < 0. \tag{C.33}$$

Solving (C.33) in terms of μ_1 leads to

$$\mu_1 \in \left(\frac{\frac{3D_2}{2}\beta - \sqrt{\frac{9D_2^2}{4}\beta^2 - 24\beta B_3(\beta^3 B_3+6\varepsilon)}}{12B_3\beta}, \frac{\frac{3D_2}{2}\beta + \sqrt{\frac{9D_2^2}{4}\beta^2 - 24\beta B_3(\beta^3 B_3+6\varepsilon)}}{12B_3\beta}\right) \tag{C.34}$$

$$\Leftrightarrow \mu_1 \in \left(\frac{D_2 - \sqrt{D_2^2 - \frac{32}{3\beta}B_3(\beta^3 B_3+6\varepsilon)}}{8B_3}, \frac{D_2 + \sqrt{D_2^2 - \frac{32}{3\beta}B_3(\beta^3 B_3+6\varepsilon)}}{8B_3}\right) \tag{C.35}$$

Now in order for the above condition on μ_1 to be meaningful, we require

$$D_2^2 - \frac{32}{3\beta}B_3(\beta^3 B_3+6\varepsilon) > 0 \quad \Leftrightarrow \quad \beta^3 - \frac{3D_2^2}{32B_3^2}\beta + \frac{6\varepsilon}{B_3} < 0. \tag{C.36}$$

Since (C.36) is a cubic inequality, therefore by following the steps described earlier (for identification of $\mathcal{S}$), one readily obtains the stated conditions on μ_1, ε and β.

C.4 Proof of Theorem 3

We first derive conditions for estimating $\mathcal{S}$, and then for estimating $\mathcal{S}_1, \mathcal{S}_2$.

Estimating $\mathcal{S}$. Upon resampling N_1 times and averaging, we have for the noise vector $\mathbf{z} \in \mathbb{R}^{m_v}$ that

$$\mathbf{z} = \left[\frac{(z'_{1,1}-z'_{1,2})}{2\mu} \cdots \frac{(z'_{m_v,1}-z'_{m_v,2})}{2\mu}\right], \tag{C.37}$$

where $z'_{j,1}, z'_{j,2} \sim \mathcal{N}(0, \sigma^2/N_1)$ are i.i.d. Our aim is to guarantee that $|z'_{j,1}-z'_{j,2}| < 2\varepsilon$ holds $\forall j = 1, \ldots, m_v$, and across all points where ∇f is estimated. Indeed, we then obtain a bounded noise model and can simply use the analysis for the setting of arbitrary bounded noise.

To this end, note that $z'_{j,1} - z'_{j,2} \sim \mathcal{N}(0, \frac{2\sigma^2}{N_1})$. It can be shown that for any $X \sim \mathcal{N}(0,1)$ we have:

$$\mathbb{P}(|X| > t) \leq 2e^{-t^2/2}, \quad \forall t > 0. \tag{C.38}$$

Since $z'_{j,1} - z'_{j,2} = \sigma\sqrt{\frac{2}{N_1}}X$ therefore for any $\varepsilon > 0$ we have that:

$$\mathbb{P}(|z'_{j,1}-z'_{j,2}| > 2\varepsilon) = \mathbb{P}\left(|X| > \frac{2\varepsilon}{\sigma}\sqrt{\frac{N_1}{2}}\right) \tag{C.39}$$

$$\leq 2\exp\left(-\frac{\varepsilon^2 N_1}{\sigma^2}\right). \tag{C.40}$$

Now to estimate $\nabla f(\mathbf{x})$ we have m_v many "difference" terms: $z'_{j,1} - z'_{j,2}$. As this is done for each $\mathbf{x} \in \chi$, therefore we have a total of $m_v(2m_x+1)^2|\mathcal{H}_2^d|$ many difference terms. Taking a union bound over all of them, we have for any $p_1 \in (0,1)$ that the choice $N_1 > \frac{\sigma^2}{\varepsilon^2}\log(\frac{2}{p_1}m_v(2m_x+1)^2|\mathcal{H}_2^d|)$ implies that the magnitudes of all difference terms are bounded by 2ε, with probability at least $1-p_1$.

Estimating $\mathcal{S}_1, \mathcal{S}_2$. In this case, we resample each query N_2 times and average – therefore the variance of the noise terms gets scaled by N_2. Note that for each $i \in \mathcal{S}$ and $\mathbf{x} \in \chi_i$, we have two difference terms corresponding to external noise – one corresponding to $\widehat{\partial}_i g(\mathbf{x})$ and the other corresponding to $\widehat{\partial}_i g(\mathbf{x}+\mu_1\mathbf{v})$. This means that in total we have at most $k(2{m'_x}^2+\lceil\log k\rceil)$ many difference terms arising.

Therefore, taking a union bound over all of them, we have for any $p_2 \in (0,1)$ that the choice $N_2 > \frac{\sigma^2}{\varepsilon'^2}\log(\frac{2k(2{m'_x}^2+\lceil\log k\rceil)}{p_2})$ implies that the magnitudes of all difference terms are bounded by $2\varepsilon'$, with probability at least $1-p_2$.

D Proofs for Section 4

D.1 Proof of Theorem 4

The proof is divided into the following steps.

Bounding the $\eta_{\mathbf{q},\mathbf{2}}$ term. The proof of this step is similar to that of Corollary 1. Since $\nabla f(\mathbf{x})$ is at most k sparse, therefore for any $\mathbf{x} \in \mathbb{R}^d$ we immediately have from Theorem 1, (3.10), the following. $\exists C_1, c'_4 > 0, c'_1 \geq 1$ such that for $c'_1 k\log(\frac{d}{k}) < m_v < \frac{d}{(\log 6)^2}$ we have with probability at least $1-e^{-c'_4 m_v} - e^{-\sqrt{m_v d}}$ that

$$\| \widehat{\nabla} f(\mathbf{x}) - \nabla f(\mathbf{x}) \|_2 \leq C_1 \max\left\{ \| \mathbf{n} \|_2, \sqrt{\log d}\, \| \mathbf{n} \|_\infty \right\}. \tag{D.1}$$

Recall from (3.5) that $\mathbf{n} = [n_1 \dots n_{m_v}]$ where $n_j = \frac{R_3(\zeta_j) - R_3(\zeta'_j)}{2\mu}$, for some $\zeta_j, \zeta'_j \in \mathbb{R}^d$. Here $R_3(\zeta)$ denotes the third order Taylor remainder terms of f. By taking the structure of f into account, we can uniformly bound $|R_3(\zeta_j)|$ as follows (so the same bound holds for $|R_3(\zeta'_j)|$). Let us define $\alpha := |\{q \in \mathcal{S}_2^{\text{var}} : \rho(q) > 1\}|$, to be the number of variables in $\mathcal{S}_2^{\text{var}}$, with degree greater than one.

$$\begin{aligned}
|R_3(\zeta_j)| &= \frac{\mu^3}{6}\Big|\sum_{p\in\mathcal{S}_1}\partial_p^3\phi_p(\zeta_{j,p})v_p^3 + \sum_{(l,l')\in\mathcal{S}_2}(\partial_l^3\phi_{(l,l')}(\zeta_{j,l},\zeta_{j,l'})v_l^3 + \partial_{l'}^3\phi_{(l,l')}(\zeta_{j,l},\zeta_{j,l'})v_{l'}^3) \\
&\quad + \sum_{(l,l')\in\mathcal{S}_2}(3\partial_l\partial_{l'}^2\phi_{(l,l')}(\zeta_{j,l},\zeta_{j,l'})v_l v_{l'}^2 + 3\partial_l^2\partial_{l'}\phi_{(l,l')}(\zeta_{j,l},\zeta_{j,l'})v_l^2 v_{l'}) + \sum_{q\in\mathcal{S}_2^{\text{var}}:\rho(q)>1}\partial_q^3\phi_q(\zeta_{j,q})v_q^3\Big| && \text{(D.2)} \\
&\leq \frac{\mu^3}{6}\left(\frac{k_1 B_3}{m_v^{3/2}} + \frac{2k_2 B_3}{m_v^{3/2}} + \frac{\alpha B_3}{m_v^{3/2}} + \frac{6k_2 B_3}{m_v^{3/2}}\right) && \text{(D.3)} \\
&= \frac{\mu^3}{6}\frac{(k_1+\alpha+8k_2)B_3}{m_v^{3/2}}. && \text{(D.4)}
\end{aligned}$$

Using the fact $2k_2 = \sum_{l\in\mathcal{S}_2^{\text{var}}:\rho(l)>1}\rho(l) + (|\mathcal{S}_2^{\text{var}}| - \alpha)$, we can observe that $2k_2 \leq \rho_m\alpha + (|\mathcal{S}_2^{\text{var}}| - \alpha) = |\mathcal{S}_2^{\text{var}}| + (\rho_m - 1)\alpha$. Plugging this in (D.4), and using the fact $\alpha \leq k$ (since we do not assume α to be known), we obtain

$$|R_3(\zeta_j)| \leq \frac{\mu^3}{6}\frac{(k_1+\alpha+4|\mathcal{S}_2^{\text{var}}| + 4(\rho_m-1)\alpha)B_3}{m_v^{3/2}} \leq \frac{\mu^3(4k+(4\rho_m-3)\alpha)B_3}{6m_v^{3/2}} \leq \frac{\mu^3((4\rho_m+1)k)B_3}{6m_v^{3/2}}. \tag{D.5}$$

This in turn implies that $\| \mathbf{n} \|_\infty \leq \frac{\mu^2((4\rho_m+1)k)B_3}{6m_v^{3/2}}$. Using the fact $\| \mathbf{n} \|_2 \leq \sqrt{m_v}\, \| \mathbf{n} \|_\infty$, we thus obtain for the stated choice of m_v (cf. Remark 4) that

$$\| \widehat{\nabla} f(\mathbf{x}) - \nabla f(\mathbf{x}) \|_2 \leq \frac{C_1\mu^2((4\rho_m+1)k)B_3}{6m_v}, \quad \forall \mathbf{x} \in [-(1+r), 1+r]^d. \tag{D.6}$$

Recall that $[-(1+r), 1+r]^d, r > 0$, denotes the enlargement around $[-1,1]^d$, in which the smoothness properties of $\phi_p, \phi_{(l,l')}$ are defined in Section 2 (as Assumption 1). Also recall $\mathbf{w}(\mathbf{x}) \in \mathbb{R}^d, \eta_{\mathbf{q},\mathbf{2}} \in \mathbb{R}^{m_{v'}}$ from (4.4). Since $\| \mathbf{w}(\mathbf{x}) \|_\infty \leq \| \widehat{\nabla} f(\mathbf{x}) - \nabla f(\mathbf{x}) \|_2$, this then implies that $\| \eta_{\mathbf{q},\mathbf{2}} \|_\infty \leq \frac{C_1\mu^2((4\rho_m+1)k)B_3}{3m_v\mu_1}$.

Bounding the $\eta_{\mathbf{q},\mathbf{1}}$ term. We will bound $\| \eta_{\mathbf{q},\mathbf{1}} \|_\infty$. To this end, we see from (4.4) that it suffices to uniformly bound $|\mathbf{v'}^T \nabla^2 \partial_q f(\zeta) \mathbf{v'}|$, over all: $q \in \mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}$, $\mathbf{v'} \in \mathcal{V'}$, $\zeta \in [-(1+r),(1+r)]^d$. Note that

$$\mathbf{v'}^T \nabla^2 \partial_q f(\zeta) \mathbf{v'} = \sum_{l=1}^{d} v_l'^2 (\nabla^2 \partial_q f)_{l,l} + \sum_{i \neq j=1}^{d} v_i' v_j' (\nabla^2 \partial_q f)_{i,j}. \tag{D.7}$$

We have the following three cases, depending on the type of q.

1. $\mathbf{q} \in \mathcal{S}_\mathbf{1}$.

$$\mathbf{v'}^T \nabla^2 \partial_q f(\zeta) \mathbf{v'} = v_q'^2 \partial_q^3 \phi_q(\zeta_q) \Rightarrow |\mathbf{v'}^T \nabla^2 \partial_q f(\zeta) \mathbf{v'}| \leq \frac{B_3}{m_{v'}}. \tag{D.8}$$

2. $(\mathbf{q}, \mathbf{q'}) \in \mathcal{S}_\mathbf{2}$, $\rho(\mathbf{q}) = \mathbf{1}$.

$$\mathbf{v'}^T \nabla^2 \partial_q f(\zeta) \mathbf{v'} = v_q'^2 \partial_q^3 \phi_{(q,q')}(\zeta_q, \zeta_{q'}) + v_{q'}'^2 \partial_{q'}^2 \partial_q \phi_{(q,q')}(\zeta_q, \zeta_{q'}) + 2 v_q' v_{q'}' \partial_{q'} \partial_q^2 \phi_{(q,q')}(\zeta_q, \zeta_{q'}), \tag{D.9}$$

$$\Rightarrow |\mathbf{v'}^T \nabla^2 \partial_q f(\zeta) \mathbf{v'}| \leq \frac{4B_3}{m_{v'}}. \tag{D.10}$$

3. $\mathbf{q} \in \mathcal{S}_\mathbf{2}^{\text{var}}$, $\rho(\mathbf{q}) > \mathbf{1}$.

$$\begin{aligned}
\mathbf{v'}^T \nabla^2 \partial_q f(\zeta) \mathbf{v'} = v_q'^2 (\partial_q^3 \phi_q(\zeta_q) &+ \sum_{(q,q') \in \mathcal{S}_2} \partial_q^3 \phi_{(q,q')}(\zeta_q, \zeta_{q'}) + \sum_{(q',q) \in \mathcal{S}_2} \partial_q^3 \phi_{(q',q)}(\zeta_{q'}, \zeta_q)) \\
&+ \sum_{(q,q') \in \mathcal{S}_2} v_{q'}'^2 \partial_{q'}^2 \partial_q \phi_{(q,q')}(\zeta_q, \zeta_{q'}) + \sum_{(q',q) \in \mathcal{S}_2} v_{q'}'^2 \partial_{q'}^2 \partial_q \phi_{(q',q)}(\zeta_{q'}, \zeta_q) \\
&+ 2 \sum_{(q,q') \in \mathcal{S}_2} v_q' v_{q'}' \partial_{q'} \partial_q^2 \phi_{(q,q')}(\zeta_q, \zeta_{q'}) + 2 \sum_{(q',q) \in \mathcal{S}_2} v_q' v_{q'}' \partial_{q'} \partial_q^2 \phi_{(q',q)}(\zeta_{q'}, \zeta_q),
\end{aligned} \tag{D.11}$$

$$\Rightarrow |\mathbf{v'}^T \nabla^2 \partial_q f(\zeta) \mathbf{v'}| \leq \frac{1}{m_{v'}} ((\rho_m + 1) B_3 + \rho_m B_3 + 2\rho_m B_3) = \frac{(4\rho_m + 1) B_3}{m_{v'}}. \tag{D.12}$$

We can now uniformly bound $\| \eta_{\mathbf{q},\mathbf{1}} \|_\infty$ as follows.

$$\| \eta_{\mathbf{q},\mathbf{1}} \|_\infty := \max_{j=1,\ldots,m_{v'}} \frac{\mu_1}{2} |\mathbf{v}_j'^T \nabla^2 \partial_q f(\zeta_j) \mathbf{v}_j'| \leq \frac{\mu_1 (4\rho_m + 1) B_3}{2 m_{v'}}. \tag{D.13}$$

Estimating $\mathcal{S}_2$. We now proceed towards estimating $\mathcal{S}_2$. To this end, we estimate $\nabla \partial_q f(\mathbf{x})$ for each $q = 1, \ldots, d$ and $\mathbf{x} \in \chi$. Since $\nabla \partial_q f(\mathbf{x})$ is at most $(\rho_m + 1)$-sparse, therefore Theorem 1, (3.10), immediately yield the following. $\exists C_2, c_5' > 0, c_2' \geq 1$ such that for $c_2' \rho_m \log(\frac{d}{\rho_m}) < m_{v'} < \frac{d}{(\log 6)^2}$ we have with probability at least $1 - e^{-c_5' m_{v'}} - e^{-\sqrt{m_{v'} d}}$ that

$$\| \widehat{\nabla} \partial_q f(\mathbf{x}) - \nabla \partial_q f(\mathbf{x}) \|_2 \leq C_2 \max \left\{ \| \eta_{\mathbf{q},\mathbf{1}} + \eta_{\mathbf{q},\mathbf{2}} \|_2, \sqrt{\log d} \, \| \eta_{\mathbf{q},\mathbf{1}} + \eta_{\mathbf{q},\mathbf{2}} \|_\infty \right\}. \tag{D.14}$$

Since $\| \eta_{\mathbf{q},\mathbf{1}} + \eta_{\mathbf{q},\mathbf{2}} \|_\infty \leq \| \eta_{\mathbf{q},\mathbf{1}} \|_\infty + \| \eta_{\mathbf{q},\mathbf{2}} \|_\infty$, therefore using the bounds on $\| \eta_{\mathbf{q},\mathbf{1}} \|_\infty, \| \eta_{\mathbf{q},\mathbf{2}} \|_\infty$ and noting that $\| \eta_{\mathbf{q},\mathbf{1}} + \eta_{\mathbf{q},\mathbf{2}} \|_2 \leq \sqrt{m_{v'}} \| \eta_{\mathbf{q},\mathbf{1}} + \eta_{\mathbf{q},\mathbf{2}} \|_\infty$, we obtain for the stated choice of $m_{v'}$ (cf. Remark 4) that

$$\| \widehat{\nabla} \partial_q f(\mathbf{x}) - \nabla \partial_q f(\mathbf{x}) \|_2 \leq \underbrace{C_2 \left(\frac{\mu_1 (4\rho_m + 1) B_3}{2\sqrt{m_{v'}}} + \frac{C_1 \sqrt{m_{v'}} \mu^2 ((4\rho_m + 1)k) B_3}{3 m_v \mu_1} \right)}_{\tau'}; \quad q = 1, \ldots, d, \ \forall \mathbf{x} \in [-1,1]^d. \tag{D.15}$$

We next note that (D.15) trivially leads to the bound

$$\widehat{\partial_q \partial_{q'}} f(\mathbf{x}) \in [\partial_q \partial_{q'} f(\mathbf{x}) - \tau', \partial_q \partial_{q'} f(\mathbf{x}) + \tau']; \quad q, q' = 1, \ldots, d. \tag{D.16}$$

Now if $q \notin \mathcal{S}_2^{\text{var}}$ then clearly $\widehat{\partial_q \partial_{q'}} f(\mathbf{x}) \in [-\tau', \tau']$; $\forall \mathbf{x} \in [-1,1]^d, q \neq q'$. On the other hand, if $(q, q') \in \mathcal{S}_2$ then

$$\widehat{\partial_q \partial_{q'}} f(\mathbf{x}) \in [\partial_q \partial_{q'} \phi_{(q,q')}(x_q, x_{q'}) - \tau', \partial_q \partial_{q'} \phi_{(q,q')}(x_q, x_{q'}) + \tau']. \tag{D.17}$$

If furthermore $m_x \geq \lambda_2^{-1}$, then due to the construction of χ, $\exists \mathbf{x} \in \chi$ so that $|\widehat{\partial_q \partial_{q'}} f(\mathbf{x})| \geq D_2 - \tau'$. Hence if $\tau' < D_2/2$ holds, the we would have $|\widehat{\partial_q \partial_{q'}} f(\mathbf{x})| > D_2/2$, leading to the identification of (q, q'). Since this is true for each $(q, q') \in \mathcal{S}_2$, hence it

follows that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$. Now, $\tau' < D_2/2$ is equivalent to

$$\underbrace{\frac{(4\rho_m+1)B_3}{2\sqrt{m_{v'}}}}_{a}\mu_1 + \underbrace{\left(\frac{C_1\sqrt{m_{v'}}((4\rho_m+1)k)B_3}{3m_v}\right)}_{b}\frac{\mu^2}{\mu_1} < \frac{D_2}{2C_2} \Leftrightarrow a\mu_1^2 - \frac{D_2}{C_2}\mu_1 + b\mu^2 < 0 \tag{D.18}$$

$$\Leftrightarrow \mu_1 \in \left((D_2/(4aC_2)) - \sqrt{(D_2/(4aC_2))^2 - (b\mu^2/a)}, (D_2/(4aC_2)) + \sqrt{(D_2/(4aC_2))^2 - (b\mu^2/a)}\right). \tag{D.19}$$

Lastly, we see that the bounds in (D.19) are valid if:

$$\mu^2 < \frac{D_2^2}{16abC_2^2} = \frac{3D_2^2 m_v}{8C_1C_2^2B_3^2(4\rho_m+1)((4\rho_m+1)k)}. \tag{D.20}$$

Estimating $\mathcal{S}_1$. With $\mathcal{P} := [d] \setminus \widehat{\mathcal{S}_2^{\text{var}}}$, we have via Taylor's expansion of f:

$$\frac{f((\mathbf{x}+\mu'\mathbf{v}_j'')_{\mathcal{P}}) - f((\mathbf{x}-\mu'\mathbf{v}_j'')_{\mathcal{P}})}{2\mu'} = \langle (\mathbf{v}_j'')_{\mathcal{P}}, (\nabla f((\mathbf{x})_{\mathcal{P}}))_{\mathcal{P}}\rangle + \underbrace{\frac{R_3((\zeta_j)_{\mathcal{P}}) - R_3((\zeta_j')_{\mathcal{P}})}{2\mu'}}_{n_j}; \quad j = 1, \ldots, m_{v''}. \tag{D.21}$$

(D.21) corresponds to linear measurements of the $(k - |\widehat{\mathcal{S}_2^{\text{var}}}|)$ sparse vector: $(\nabla f((\mathbf{x})_{\mathcal{P}}))_{\mathcal{P}}$. We now proceed similar to the proof of Corollary 1. Note that we effectively perform ℓ_1 minimization over $\mathbb{R}^{|\mathcal{P}|}$. Therefore for any $\mathbf{x} \in \mathbb{R}^d$ we immediately have from Theorem 1, (3.10), the following. $\exists C_3, c_6' > 0, c_3' \geq 1$ such that for $c_3'(k - |\widehat{\mathcal{S}_2^{\text{var}}}|)\log(\frac{|\mathcal{P}|}{k-|\widehat{\mathcal{S}_2^{\text{var}}}|}) < m_{v''} < \frac{|\mathcal{P}|}{(\log 6)^2}$, we have with probability at least $1 - e^{-c_6' m_{v''}} - e^{-\sqrt{m_{v''}|\mathcal{P}|}}$ that

$$\| (\widehat{\nabla} f((\mathbf{x})_{\mathcal{P}}))_{\mathcal{P}} - (\nabla f((\mathbf{x})_{\mathcal{P}}))_{\mathcal{P}} \|_2 \leq C_3 \max\left\{\| \mathbf{n} \|_2, \sqrt{\log|\mathcal{P}|}\, \| \mathbf{n} \|_\infty\right\}. \tag{D.22}$$

We now uniformly bound $R_3((\zeta_j)_{\mathcal{P}})$ for all $j = 1, \ldots, m_{v''}$ and $\zeta_j \in [-(1+r), 1+r]^d$ as follows.

$$R_3((\zeta_j)_{\mathcal{P}}) = \frac{{\mu'}^3}{6}\sum_{p\in\mathcal{S}_1\cap\mathcal{P}} \partial_p^3\phi_p(\zeta_{j,p}){v_{j,p}''}^3 \quad \Rightarrow |R_3((\zeta_j)_{\mathcal{P}})| \leq \frac{(k - |\widehat{\mathcal{S}_2^{\text{var}}}|){\mu'}^3 B_3}{6m_{v''}^{3/2}}. \tag{D.23}$$

This in turn implies that $\| \mathbf{n} \|_\infty \leq \frac{(k-|\widehat{\mathcal{S}_2^{\text{var}}}|){\mu'}^2 B_3}{6m_{v''}^{3/2}}$ and $\| \mathbf{n} \|_2 \leq \sqrt{m_{v''}}\, \| \mathbf{n} \|_\infty \leq \frac{(k-|\widehat{\mathcal{S}_2^{\text{var}}}|){\mu'}^2 B_3}{6m_{v''}}$. Plugging these bounds in (D.22), we obtain for the stated choice of $m_{v''}$ (cf. Remark 4) that

$$\| (\widehat{\nabla} f((\mathbf{x})_{\mathcal{P}}))_{\mathcal{P}} - (\nabla f((\mathbf{x})_{\mathcal{P}}))_{\mathcal{P}} \|_2 \leq \underbrace{\frac{C_3(k - |\widehat{\mathcal{S}_2^{\text{var}}}|){\mu'}^2 B_3}{6m_{v''}}}_{\tau''}; \quad \mathbf{x} \in [-1, 1]^d. \tag{D.24}$$

Finally, using the same arguments as before, we have that $\tau'' < D_1/2$ or equivalently ${\mu'}^2 < \frac{3m_{v''}D_1}{C_3(k-|\widehat{\mathcal{S}_2^{\text{var}}}|)B_3}$ is sufficient to recover $\mathcal{S}_1$. This completes the proof.

D.2 Proof of Theorem 5

We begin by establishing the conditions pertaining to the estimation of $\mathcal{S}_2$. Then we prove the conditions for estimation of $\mathcal{S}_1$.

Estimation of $\mathcal{S}_2$. We first note that the linear system (3.6) now has the form: $\mathbf{y} = \mathbf{V}\nabla f(\mathbf{x}) + \mathbf{n} + \mathbf{z}$ where $z_j = (z_{j,1}' - z_{j,2}')/(2\mu)$ represents the external noise component, for $j = 1, \ldots, m_v$. Observe that $\| \mathbf{z} \|_\infty \leq \varepsilon/\mu$. Using the bounds on $\| \mathbf{n} \|_\infty, \| \mathbf{n} \|_2$ from Section D.1, we then observe that (D.6) changes to:

$$\| \widehat{\nabla} f(\mathbf{x}) - \nabla f(\mathbf{x}) \|_2 \leq C_1\left(\frac{\mu^2((4\rho_m+1)k)B_3}{6m_v} + \frac{\varepsilon\sqrt{m_v}}{\mu}\right), \quad \forall \mathbf{x} \in [-(1+r), 1+r]^d. \tag{D.25}$$

As a result, we then have that

$$\| \eta_{\mathbf{q},\mathbf{2}} \|_\infty \leq C_1\left(\frac{\mu^2((4\rho_m+1)k)B_3}{3m_v\mu_1} + \frac{2\varepsilon\sqrt{m_v}}{\mu\mu_1}\right). \tag{D.26}$$

Now note that the bound on $\| \eta_{\mathbf{q},\mathbf{1}} \|_\infty$ is unchanged from Section D.1, *i.e.*, $\| \eta_{\mathbf{q},\mathbf{1}} \|_\infty \leq \frac{\mu_1(4\rho_m+1)B_3}{2m_{v'}}$. As a consequence, we see that (D.15) changes to:

$$\| \widehat{\nabla}\partial_q f(\mathbf{x}) - \nabla\partial_q f(\mathbf{x}) \|_2 \leq \underbrace{C_2\left(\frac{\mu_1(4\rho_m+1)B_3}{2\sqrt{m_{v'}}} + C_1\frac{\sqrt{m_{v'}}\mu^2((4\rho_m+1)k)B_3}{3m_v\mu_1} + \frac{2C_1\varepsilon\sqrt{m_v m_{v'}}}{\mu\mu_1}\right)}_{\tau'}. \tag{D.27}$$

With a and b as stated in the Theorem, we then see that $\tau' < D_2/2$ is equivalent to

$$a\mu_1^2 - \frac{D_2}{2C_2}\mu_1 + \left(b\mu^2 + \frac{2C_1\varepsilon\sqrt{m_v m_{v'}}}{\mu}\right) < 0. \tag{D.28}$$

which in turn is equivalent to

$$\mu_1 \in \left(\frac{D_2}{4aC_2} - \sqrt{\left(\frac{D_2}{4aC_2}\right)^2 - \left(\frac{b\mu^3 + 2C_1\varepsilon\sqrt{m_v m_{v'}}}{a\mu}\right)}, \frac{D_2}{4aC_2} + \sqrt{\left(\frac{D_2}{4aC_2}\right)^2 - \left(\frac{b\mu^3 + 2C_1\varepsilon\sqrt{m_v m_{v'}}}{a\mu}\right)}\right). \tag{D.29}$$

For the above bound to be valid, we require

$$\frac{b\mu^2}{a} + \frac{2C_1\varepsilon\sqrt{m_v m_{v'}}}{a\mu} < \frac{D_2^2}{16a^2C_2^2} \tag{D.30}$$

$$\Leftrightarrow \mu^3 - \frac{D_2^2}{16abC_2^2}\mu + \frac{2C_1\varepsilon\sqrt{m_v m_{v'}}}{b} < 0 \tag{D.31}$$

to hold. (D.31) is a cubic inequality. Recall from Section B that a cubic equation of the form: $y^3 + py + q = 0$, has 3 distinct real roots if its discriminant $\frac{p^3}{27} + \frac{q^2}{4} < 0$. Note that for this to be possible, p must be negative, which is the case in (D.31). Applying this to (D.31) leads to the condition: $\varepsilon < \frac{D_2^3}{192\sqrt{3}C_1C_2^3\sqrt{a^3bm_{v'}m_v}} = \varepsilon_1$. Furthermore, as stated in (B.4), the 3 distinct real roots are given by:

$$y_1 = 2\sqrt{-p/3}\cos(\theta/3),\ y_2 = -2\sqrt{-p/3}\cos(\theta/3 + \pi/3),\ y_3 = -2\sqrt{-p/3}\cos(\theta/3 - \pi/3) \tag{D.32}$$

where $\theta = \cos^{-1}\left(\frac{-q/2}{\sqrt{-p^3/27}}\right)$. Applying this to (D.31) then leads to $\theta_1 = \cos^{-1}(-\varepsilon/\varepsilon_1)$. For $0 < \varepsilon < \varepsilon_1$ we have $\pi/2 < \theta_1 < \pi$ which implies $0 < y_2 < y_1$ and $y_3 < 0$. In particular if $q > 0$, then one can verify that $y^3 + py + q < 0$ holds if $y \in (y_2, y_1)$. Applying this to (D.31), we consequently obtain:

$$\mu \in \left(\sqrt{\frac{D_2^2}{12abC_2^2}}\cos(\theta_1/3 - 2\pi/3), \sqrt{\frac{D_2^2}{12abC_2^2}}\cos(\theta_1/3)\right). \tag{D.33}$$

Estimation of $\mathcal{S}_1$. We now prove the conditions for estimation of $\mathcal{S}_1$. First note that (D.21) now changes to:

$$\frac{f((\mathbf{x} + \mu'\mathbf{v}_j'')_{\mathcal{P}}) - f((\mathbf{x} - \mu'\mathbf{v}_j'')_{\mathcal{P}})}{2\mu'} = \langle (\mathbf{v}_j'')_{\mathcal{P}}, (\nabla f((\mathbf{x})_{\mathcal{P}}))_{\mathcal{P}}\rangle + \underbrace{\frac{R_3((\zeta_j)_{\mathcal{P}}) - R_3((\zeta_j')_{\mathcal{P}})}{2\mu'}}_{n_j} + \underbrace{\frac{z'_{j,1} - z'_{j,2}}{2\mu'}}_{z_j}, \tag{D.34}$$

for $j = 1, \ldots, m_{v''}$. Denoting $\mathbf{z} = [z_1 \cdots z_{m_{v''}}]$, we have $\| \mathbf{z} \|_\infty \leq \varepsilon/\mu'$. As the bounds on $\| \mathbf{n} \|_2, \| \mathbf{n} \|_\infty$ are unchanged, therefore (D.35) now changes to:

$$\| (\widehat{\nabla} f((\mathbf{x})_{\mathcal{P}}))_{\mathcal{P}} - (\nabla f((\mathbf{x})_{\mathcal{P}}))_{\mathcal{P}} \|_2 \leq \underbrace{C_3\left(\frac{(k - |\widehat{\mathcal{S}_2^{\text{var}}}|)\mu'^2 B_3}{6m_{v''}} + \frac{\varepsilon\sqrt{m_{v''}}}{\mu'}\right)}_{\tau''};\quad \mathbf{x} \in [-1, 1]^d. \tag{D.35}$$

Denoting $a_1 = \frac{(k - |\widehat{\mathcal{S}_2^{\text{var}}}|)B_3}{6m_{v''}}$, $b_1 = \sqrt{m_{v''}}$, we then see from (D.35) that the condition $\tau'' < D_1/2$ is equivalent to

$$\mu'^3 - \frac{D_1}{2a_1C_3}\mu' + \frac{b_1\varepsilon}{a_1} < 0. \tag{D.36}$$

As discussed earlier for estimation of $\mathcal{S}_2$, the cubic equation corresponding to (D.36) has 3 distinct real roots if its discriminant is negative. This then leads to the condition $\varepsilon < \frac{D_1^{3/2}}{3\sqrt{6a_1C_3^3b_1^2}} = \varepsilon_2$. Then by using the expressions for the roots of the cubic from (D.32), one can verify that (D.36) holds if

$$\mu' \in (2\sqrt{D_1/(6a_1C_3)}\cos(\theta_2/3 - 2\pi/3), 2\sqrt{D_1/(6a_1C_3)}\cos(\theta_2/3)) \tag{D.37}$$

with $\theta_2 = \cos^{-1}(\varepsilon/\varepsilon_2)$. This completes the proof.

D.3 Proof of Theorem 6

We first derive conditions for estimating $\mathcal{S}_2$, and then for $\mathcal{S}_1$. The outline is essentially the same as the proof of Theorem 3 in Section C.4, so we omit the details.

Estimating $\mathcal{S}_2$. Upon resampling N_1 times and averaging, we have for the noise vector $\mathbf{z} \in \mathbb{R}^{m_v}$ that

$$\mathbf{z} = \left[\frac{(z'_{1,1} - z'_{1,2})}{2\mu} \cdots \frac{(z'_{m_v,1} - z'_{m_v,2})}{2\mu} \right], \tag{D.38}$$

where $z'_{j,1}, z'_{j,2} \sim \mathcal{N}(0, \sigma^2/N_1)$ are i.i.d. Our aim is to guarantee that $|z'_{j,1} - z'_{j,2}| < 2\varepsilon$ holds $\forall j = 1, \ldots, m_v$, and across all points where ∇f is estimated. Indeed, we then obtain a bounded noise model and can simply use the analysis for the setting of arbitrary bounded noise.

Now to estimate $\nabla f(\mathbf{x})$ we have m_v many "difference" terms: $z'_{j,1} - z'_{j,2}$. We additionally estimate $m_{v'}$ many gradients at each $\mathbf{x}$ implying a total of $m_v(m_{v'}+1)$ difference terms. As this is done for each $\mathbf{x} \in \chi$, therefore we have a total of $m_v(m_{v'}+1)(2m_x+1)^2|\mathcal{H}_2^d|$ many difference terms. Taking a union bound over all of them, we have for any $p_1 \in (0,1)$ that the choice $N_1 > \frac{\sigma^2}{\varepsilon^2}\log(\frac{2}{p_1}m_v(m_{v'}+1)(2m_x+1)^2|\mathcal{H}_2^d|)$ implies that the magnitudes of all difference terms are bounded by 2ε, with probability at least $1-p_1$.

Estimating $\mathcal{S}_1$. In this case, we resample each query N_2 times and average – therefore the variance of the noise terms gets scaled by N_2. We now have $|\chi_{\text{diag}}|m_{v''} = (2m'_x+1)m_{v''}$ many "difference" terms corresponding to Gaussian noise. Therefore, taking a union bound over all of them, we have for any $p_2 \in (0,1)$ that the choice $N_2 > \frac{\sigma^2}{\varepsilon'^2}\log(\frac{2(2m'_x+1)m_{v''}}{p_2})$ implies that the magnitudes of all difference terms are bounded by $2\varepsilon'$, with probability at least $1-p_2$.

E Proofs for Section 5

E.1 Proof of Theorem 8

We only prove the part concerning the identification of $\mathcal{S}_2$, as the proof for identifying $\mathcal{S}_1$ is identical to that of Theorem 4 (see Section D.1). Consider the linear system defined in (5.8) at some $\mathbf{x} \in [-1,1]^d$. We begin by uniformly bounding the magnitude of the remainder terms: $|R_3(\zeta_j)|$, $|R_3(\zeta'_j)|$ where $\zeta_j, \zeta'_j \in [-(1+r), 1+r]^d$ for some $r > 0$; $j = 1, \ldots, m_v$. Let us define $\alpha := |\{q \in \mathcal{S}_2^{\text{var}} : \rho(q) > 1\}|$, to be the number of variables in $\mathcal{S}_2^{\text{var}}$, with degree greater than one. By taking the structure of f into account, we can uniformly bound $|R_3(\zeta_j)|$ as follows.

$$\begin{aligned}
|R_3(\zeta_j)| = \frac{\mu^3}{6}| &\sum_{p\in\mathcal{S}_1} \partial_p^3\phi_p(\zeta_{j,p})(2v_p)^3 + \sum_{(l,l')\in\mathcal{S}_2} (\partial_l^3\phi_{(l,l')}(\zeta_{j,l},\zeta_{j,l'})(2v_l)^3 + \partial_{l'}^3\phi_{(l,l')}(\zeta_{j,l},\zeta_{j,l'})(2v_{l'})^3) \\
&+ \sum_{(l,l')\in\mathcal{S}_2} (3\partial_l\partial_{l'}^2\phi_{(l,l')}(\zeta_{j,l},\zeta_{j,l'})(2v_l)(2v_{l'})^2 + 3\partial_l^2\partial_{l'}\phi_{(l,l')}(\zeta_{j,l},\zeta_{j,l'})(2v_l)^2(2v_{l'})) + \sum_{q\in\mathcal{S}_2^{\text{var}}:\rho(q)>1} \partial_q^3\phi_q(\zeta_{j,q})(2v_q)^3| \qquad \text{(E.1)} \\
&\leq \frac{\mu^3}{6}\left(\frac{8k_1B_3(\sqrt{3})^3}{m_v^{3/2}} + \frac{16k_2B_3(\sqrt{3})^3}{m_v^{3/2}} + \frac{8\alpha B_3(\sqrt{3})^3}{m_v^{3/2}} + \frac{48k_2B_3(\sqrt{3})^3}{m_v^{3/2}}\right) \qquad \text{(E.2)} \\
&= \frac{4\sqrt{3}\mu^3B_3}{m_v^{3/2}}(k_1+\alpha+8k_2). \qquad \text{(E.3)}
\end{aligned}$$

By observing $2k_2 = \sum_{l\in\mathcal{S}_2^{\text{var}}:\rho(l)>1}\rho(l) + (|\mathcal{S}_2^{\text{var}}| - \alpha)$, we obtain $2k_2 \leq \rho_m\alpha + (|\mathcal{S}_2^{\text{var}}| - \alpha) = |\mathcal{S}_2^{\text{var}}| + (\rho_m - 1)\alpha$. Plugging this in (E.3), and using the fact $\alpha \leq k$, we obtain

$$|R_3(\zeta_j)| \leq \frac{4\sqrt{3}\mu^3B_3}{m_v^{3/2}}(4\rho_m+1)k. \tag{E.4}$$

Since the same bound holds also for $|R_3(\zeta'_j)|$, we thus obtain:

$$\| \mathbf{n} \|_\infty \leq \frac{1}{2\mu^2}\left(\frac{4\sqrt{3}\mu^3B_3}{m_v^{3/2}}(4\rho_m+1)k\right) = \frac{2\sqrt{3}\mu B_3}{m_v^{3/2}}(4\rho_m+1)k, \tag{E.5}$$

$$\Rightarrow \| \mathbf{n} \|_1 \leq m_v\frac{2\sqrt{3}\mu B_3}{m_v^{3/2}}(4\rho_m+1)k = \frac{2\sqrt{3}\mu B_3}{m_v^{1/2}}(4\rho_m+1)k. \tag{E.6}$$

Therefore by setting $\eta = \frac{2\sqrt{3}\mu B_3}{m_v^{1/2}}(4\rho_m+1)k$, and for the stated choice of m_v, we obtain via Theorem 7 that

$$\| \widehat{\nabla^2 f}(\mathbf{x}) - \nabla^2 f(\mathbf{x}) \|_F \leq C_1\eta = \underbrace{C_1\frac{2\sqrt{3}\mu B_3}{m_v^{1/2}}(4\rho_m+1)k}_{\tau}; \quad \forall \mathbf{x} \in [-1,1]^d. \tag{E.7}$$

We next note that (E.7) leads to

$$\widehat{\partial_q \partial_{q'}} f(\mathbf{x}) \in [\partial_q \partial_{q'} f(\mathbf{x}) - \frac{\tau}{\sqrt{2}}, \partial_q \partial_{q'} f(\mathbf{x}) + \frac{\tau}{\sqrt{2}}]; \quad (q, q') \in \binom{[d]}{2}. \tag{E.8}$$

Now if $(q, q') \notin \mathcal{S}_2$ then clearly $\widehat{\partial_q \partial_{q'}} f(\mathbf{x}) \in [-\frac{\tau}{\sqrt{2}}, \frac{\tau}{\sqrt{2}}]; \forall \mathbf{x} \in [-1, 1]^d$. On the other hand, if $(q, q') \in \mathcal{S}_2$ then

$$\widehat{\partial_q \partial_{q'}} f(\mathbf{x}) \in [\partial_q \partial_{q'} \phi_{(q,q')}(x_q, x_{q'}) - \frac{\tau}{\sqrt{2}}, \partial_q \partial_{q'} \phi_{(q,q')}(x_q, x_{q'}) + \frac{\tau}{\sqrt{2}}]. \tag{E.9}$$

If furthermore $m_x \geq \lambda_2^{-1}$, then due to the construction of χ, $\exists \mathbf{x} \in \chi$ so that $|\widehat{\partial_q \partial_{q'}} f(\mathbf{x})| \geq D_2 - \frac{\tau}{\sqrt{2}}$. Hence if $\frac{\tau}{\sqrt{2}} < D_2/2$ holds, the we would have $|\widehat{\partial_q \partial_{q'}} f(\mathbf{x})| > D_2/2$, leading to the identification of (q, q'). Since this is true for each $(q, q') \in \mathcal{S}_2$, hence it follows that $\widehat{\mathcal{S}_2} = \mathcal{S}_2$. Lastly, we easily see that $\frac{\tau}{\sqrt{2}} < D_2/2$ is equivalent to the stated condition on μ.

E.2 Proof of Theorem 9

We only prove the part concerning the identification of $\mathcal{S}_2$, as the proof for identifying $\mathcal{S}_1$ is identical to that of Theorem 5 (see Section D.2). To this end, note that (5.8) now changes to the linear system $\mathbf{y} = \mathcal{M}(\nabla^2 f(\mathbf{x})) + \mathbf{n} + \mathbf{z}$, where $z_j = (z'_{j,1} + z'_{j,2} - 2z'_3)/(4\mu^2)$ for $j = 1, \ldots, m_v$. Since $\| \mathbf{z} \|_\infty \leq \frac{\varepsilon}{\mu^2}$, therefore using the bound on $\| \mathbf{n} \|_1$ in (E.6), we readily obtain

$$\| \mathbf{n} + \mathbf{z} \|_1 \leq \underbrace{\frac{2\sqrt{3}\mu B_3}{m_v^{1/2}}(4\rho_m + 1)k + \frac{\varepsilon m_v}{\mu^2}}_{\eta}, \tag{E.10}$$

which in conjunction with Theorem 7 readily implies that

$$\| \widehat{\nabla^2} f(\mathbf{x}) - \nabla^2 f(\mathbf{x}) \|_F \leq C_1 \eta = \underbrace{C_1 \left(\frac{2\sqrt{3}\mu B_3}{m_v^{1/2}}(4\rho_m + 1)k + \frac{\varepsilon m_v}{\mu^2} \right)}_{\tau}; \quad \forall \mathbf{x} \in [-1, 1]^d. \tag{E.11}$$

As shown in Section E.1, it is sufficient to guarantee $\tau/\sqrt{2} < D_2/2$ for exact identification of $\mathcal{S}_2$. This is equivalent to saying that

$$\underbrace{\frac{\sqrt{6}\mu B_3}{m_v^{1/2}}(4\rho_m + 1)k}_{a\mu} + \underbrace{\frac{\varepsilon m_v}{\mu^2\sqrt{2}}}_{b/\mu^2} < \frac{D_2}{2C_1} \Leftrightarrow \mu^3 - \frac{D_2}{2aC_1}\mu^2 + \frac{b}{a} < 0. \tag{E.12}$$

(E.12) is a cubic inequality. Recall from Section B that a cubic equation of the form: $x^3 + Ax^2 + C = 0$, has 3 distinct real roots if its discriminant $\frac{p^3}{27} + \frac{q^2}{4} < 0$ where $p = -\frac{A^2}{3}$ and $q = \frac{27C + 2A^3}{27}$. Assuming the discriminant to be negative (which means $p < 0$), and denoting $\theta_1 = \cos^{-1}(-\frac{q/2}{\sqrt{-p^3/27}})$, the three roots are given as in (B.5):

$$x_1 = 2\sqrt{-\frac{p}{3}} \cos\left(\frac{\theta_1}{3}\right) - \frac{A}{3} = -\frac{2A}{3} \cos\left(\frac{\theta_1}{3}\right) - \frac{A}{3}, \tag{E.13}$$

$$x_2 = 2\sqrt{-\frac{p}{3}} \cos\left(\frac{\theta_1}{3} + \frac{2\pi}{3}\right) - \frac{A}{3} = \frac{2A}{3} \cos\left(\frac{\theta_1}{3} - \frac{\pi}{3}\right) - \frac{A}{3}, \tag{E.14}$$

$$x_3 = 2\sqrt{-\frac{p}{3}} \cos\left(\frac{\theta_1}{3} + \frac{4\pi}{3}\right) - \frac{A}{3} = \frac{2A}{3} \cos\left(\frac{\theta_1}{3} + \frac{\pi}{3}\right) - \frac{A}{3}. \tag{E.15}$$

For $0 < \theta_1 < \pi$ one can verify that $x_2 < 0$ and $0 < x_3 < x_1$. Moreover, since $A < 0$ and $C > 0$, it is not hard to verify that $x^3 + Ax^2 + C < 0$ for $x \in (x_3, x_1)$.

Translated to our setting, we have $A = -D_2/(2aC_1)$, $C = b/a$ which gives us $p = -\frac{D_2^2}{12a^2C_1^2}$ and $q = \left(\frac{27b}{a} - \frac{D_2^3}{4a^3C_1^3}\right)/27$. The cubic equation corresponding to (E.12) has three distinct real roots if

$$|q|/2 < (-p^3/27)^{1/2} = \frac{D_2^3}{216a^3C_1^3}, \tag{E.16}$$

$$\Leftrightarrow \frac{27b}{a} - \frac{D_2^3}{4a^3C_1^3} < \frac{D_2^3}{4a^3C_1^3}, \tag{E.17}$$

$$\Leftrightarrow \varepsilon < \frac{\sqrt{2}D_2^3}{54a^2C_1^3m_v} = \frac{D_2^3}{162\sqrt{2}C_1^3B_3^2} \frac{1}{(4\rho_m + 1)^2k^2} = \varepsilon_1. \tag{E.18}$$

Furthermore, we have:

$$\theta_1 = \cos^{-1}\left(\frac{-q/2}{\sqrt{-p^3/27}}\right) = \cos^{-1}\left(\frac{\frac{-27b}{a} + \frac{D_2^3}{4a^3C_1^3}}{\frac{D_2^3}{4a^3C_1^3}}\right) = \cos^{-1}\left(1 - \frac{2\varepsilon}{\varepsilon_1}\right). \tag{E.19}$$

Lastly, (E.12) is satisfied for $\mu \in (x_3, x_1)$. Substituting the expression for A in (E.13),(E.15), we arrive at the stated condition on μ. This completes the proof.

E.3 Proof of Theorem 10

Let the external noise vector be denoted by $\mathbf{z} \in \mathbb{R}^{m_v}$ where $z_j = (z'_{j,1} + z'_{j,2} - 2z'_3)/(4\mu^2)$. Upon resampling N_1 times and averaging, we have $z'_{j,1}$, $z'_{j,2}$, $z'_3 \sim \mathcal{N}(0, \sigma^2/N_1)$, which in turn implies $z'_{j,1} + z'_{j,2} - 2z'_3 \sim \mathcal{N}(0, 6\sigma^2/N_1)$. Our aim is to guarantee that $|z'_{j,1} + z'_{j,2} - 2z'_3| < 4\varepsilon$ holds $\forall j = 1, \ldots, m_v$, and across all points where $\nabla^2 f$ is estimated. Indeed, we then obtain a bounded noise model and can simply use the analysis for the setting of arbitrary bounded noise.

To this end, we proceed as in the proof of Theorem 3 in Section C.4. Denoting $X \sim \mathcal{N}(0,1)$, we first have $z'_{j,1} + z'_{j,2} - 2z'_3 = \sigma\sqrt{\frac{6}{N_1}}X$. Using the tail bound for standard Gaussian random variables, we then obtain

$$\mathbb{P}(|z'_{j,1} + z'_{j,2} - 2z'_{j,3}| > 4\varepsilon) \leq 2\exp\left(-\frac{4\varepsilon^2 N_1}{3\sigma^2}\right).$$

At each $\mathbf{x} \in \chi$, we have m_v many terms of the form: $z'_{j,1}+z'_{j,2}-2z'_3$, meaning that we have a total of $m_v(2m_x+1)^2|\mathcal{H}_2^d|$ such terms. Taking a union bound over all of them, we have for any $p_1 \in (0,1)$ that the choice $N_1 > \frac{3\sigma^2}{4\varepsilon^2}\log(\frac{2}{p_1}m_v(2m_x+1)^2|\mathcal{H}_2^d|)$ implies that the magnitudes of all such terms are bounded by 4ε, with probability at least $1 - p_1$.

F Proofs for Section 6

F.1 Proof of Proposition 1

1. $\mathbf{p} \in \mathcal{S}_\mathbf{1}$.

   We have for $\tilde{\phi}_p$ that $\| \tilde{\phi}_p - (\phi_p + C) \|_{L_\infty[-1,1]} = O(n^{-3})$. Denoting $\tilde{\phi}_p(x_p) - (\phi_p(x_p) + C) = z_p(x_p)$, this means $|z_p(x_p)| = O(n^{-3})$, $\forall x_p \in [-1,1]$. Now $|\mathbb{E}_p[\tilde{\phi}_p - (\phi_p + C)]| = |\mathbb{E}_p[\tilde{\phi}_p] - C| = |\mathbb{E}_p[z_p]| \leq \mathbb{E}_p[|z_p|] = O(n^{-3})$.

   Lastly, we have that:

$$\begin{aligned}
\| \widehat{\phi}_p - \phi_p \|_{L_\infty[-1,1]} &= \| \tilde{\phi}_p - \mathbb{E}_p[\tilde{\phi}_p] - \phi_p \|_{L_\infty[-1,1]} && \text{(F.1)}\\
&= \| \tilde{\phi}_p - (\phi_p + C) - (\mathbb{E}_p[\tilde{\phi}_p] - C) \|_{L_\infty[-1,1]} && \text{(F.2)}\\
&= O(n^{-3}). && \text{(F.3)}
\end{aligned}$$

2. $(\mathbf{l}, \mathbf{l'}) \in \mathcal{S}_\mathbf{2}$.

   We only consider the case where $\rho(l), \rho(l') > 1$ as proofs for the other cases are similar. Now for $\tilde{\phi}_{(l,l')}$ we have that $\| \tilde{\phi}_{(l,l')} - (g_{(l,l')} + C) \|_{L_\infty[-1,1]^2} = O(n^{-3/2})$. Denoting $\tilde{\phi}_{(l,l')}(x_l, x_{l'}) - (g_{(l,l')}(x_l, x_{l'}) + C) = z_{(l,l')}(x_l, x_{l'})$, this means $|z_{(l,l')}(x_l, x_{l'})| = O(n^{-3/2})$, $\forall (x_l, x_{l'}) \in [-1,1]^2$. Consequently, one can easily verify that:

$$\begin{aligned}
\| \mathbb{E}_l[\tilde{\phi}_{(l,l')}] - (\mathbb{E}_l[g_{(l,l')}] + C) \|_{L_\infty[-1,1]} &= O(n^{-3/2}), && \text{(F.4)}\\
\| \mathbb{E}_{l'}[\tilde{\phi}_{(l,l')}] - (\mathbb{E}_{l'}[g_{(l,l')}] + C) \|_{L_\infty[-1,1]} &= O(n^{-3/2}), && \text{(F.5)}\\
\| \mathbb{E}_{(l,l')}[\tilde{\phi}_{(l,l')}] - (\mathbb{E}_{(l,l')}[g_{(l,l')}] + C) \|_{L_\infty} &= O(n^{-3/2}). && \text{(F.6)}
\end{aligned}$$

   Now note that using the form for $g_{(l,l')}$ from (6.5), we have that

$$\begin{aligned}
\mathbb{E}_l[g_{(l,l')}] = &\sum_{\substack{l_1:(l,l_1)\in\mathcal{S}_2\\ l_1\neq l'}} \mathbb{E}_l[\phi_{(l,l_1)}(x_l, 0)] + \sum_{\substack{l_1:(l_1,l)\in\mathcal{S}_2\\ l_1\neq l'}} \mathbb{E}_l[\phi_{(l_1,l)}(0, x_l)] + \sum_{\substack{l'_1:(l',l'_1)\in\mathcal{S}_2\\ l'_1\neq l}} \phi_{(l',l'_1)}(x_{l'}, 0)\\
&+ \sum_{\substack{l'_1:(l'_1,l')\in\mathcal{S}_2\\ l'_1\neq l}} \phi_{(l'_1,l')}(0, x_{l'}) + \phi_{l'}(x_{l'}) + C, \quad \text{and} && \text{(F.7)}
\end{aligned}$$

$$\mathbb{E}_{l'}[g_{(l,l')}] = \sum_{\substack{l_1:(l,l_1)\in\mathcal{S}_2\\ l_1\neq l'}} \phi_{(l,l_1)}(x_l,0) + \sum_{\substack{l_1:(l_1,l)\in\mathcal{S}_2\\ l_1\neq l'}} \phi_{(l_1,l)}(0,x_l) + \sum_{\substack{l'_1:(l',l'_1)\in\mathcal{S}_2\\ l'_1\neq l}} \mathbb{E}_{l'}[\phi_{(l',l'_1)}(x_{l'},0)]$$
$$+ \sum_{\substack{l'_1:(l'_1,l')\in\mathcal{S}_2\\ l'_1\neq l}} \mathbb{E}_{l'}\phi_{(l'_1,l')}(0,x_{l'}) + \phi_l(x_l) + C, \quad \text{and} \tag{F.8}$$

$$\mathbb{E}_{(l,l')}[g_{(l,l')}] = \sum_{\substack{l_1:(l,l_1)\in\mathcal{S}_2\\ l_1\neq l'}} \mathbb{E}_l[\phi_{(l,l_1)}(x_l,0)] + \sum_{\substack{l_1:(l_1,l)\in\mathcal{S}_2\\ l_1\neq l'}} \mathbb{E}_l[\phi_{(l_1,l)}(0,x_l)]$$
$$+ \sum_{\substack{l'_1:(l',l'_1)\in\mathcal{S}_2\\ l'_1\neq l}} \mathbb{E}_{l'}[\phi_{(l',l'_1)}(x_{l'},0)] + \sum_{\substack{l'_1:(l'_1,l')\in\mathcal{S}_2\\ l'_1\neq l}} \mathbb{E}_{l'}\phi_{(l'_1,l')}(0,x_{l'}) + C. \tag{F.9}$$

We then have from (6.5), (F.7), (F.8), (F.9) that

$$g_{(l,l')} - \mathbb{E}_l[g_{(l,l')}] - \mathbb{E}_{l'}[g_{(l,l')}] + \mathbb{E}_{(l,l')}[g_{(l,l')}] = \phi_{(l,l')}. \tag{F.10}$$

Using (F.4), (F.5), (F.6), (F.10), and (6.7) it then follows that:

$$\| \widehat{\phi}_{(l,l')} - \phi_{(l,l')} \|_{L_\infty[-1,1]^2} = O(n^{-3/2}). \tag{F.11}$$

3. $\mathbf{l} \in \mathcal{S}_{\mathbf{2}}^{\text{var}} : \rho(\mathbf{l}) > \mathbf{1}$.

In this case, for $\tilde{\phi}_l : [-1,1]^2 \to \mathbb{R}$, we have that $\| \tilde{\phi}_l - (g_l + C) \|_{L_\infty[-1,1]^2} = O(n^{-3/2})$, with

$$g_l(x_l,x) = \phi_l(x_l) + \sum_{\rho(l')>1, l'\neq l} \phi_{l'}(x) + \sum_{l':(l,l')\in\mathcal{S}_2} \phi_{(l,l')}(x_l,x)$$
$$+ \sum_{l':(l',l)\in\mathcal{S}_2} \phi_{(l',l)}(x,x_l) + \sum_{(q,q')\in\mathcal{S}_2: q,q'\neq l} \phi_{(q,q')}(x,x). \tag{F.12}$$

From (F.12), we see that:

$$\mathbb{E}_x[g_l(x_l,x)] = \phi_l(x_l) + \sum_{(q,q')\in\mathcal{S}_2: q,q'\neq l} \mathbb{E}_x[\phi_{(q,q')}(x,x)], \tag{F.13}$$

$$\text{and} \quad \mathbb{E}_{(l,x)}[g_l(x_l,x)] = \sum_{(q,q')\in\mathcal{S}_2: q,q'\neq l} \mathbb{E}_x[\phi_{(q,q')}(x,x)]. \tag{F.14}$$

Hence clearly, $\mathbb{E}_x[g_l(x_l,x)] - \mathbb{E}_{(l,x)}[g_l(x_l,x)] = \phi_l(x_l)$. One can also easily verify that

$$\| \mathbb{E}_x[\tilde{\phi}_l] - (\mathbb{E}_x[g_l] + C) \|_{L_\infty[-1,1]} = O(n^{-3/2}), \tag{F.15}$$
$$\| \mathbb{E}_{(l,x)}[\tilde{\phi}_l] - (\mathbb{E}_{(l,x)}[g_l] + C) \|_{L_\infty} = O(n^{-3/2}). \tag{F.16}$$

Therefore it follows that

$$\| \widehat{\phi}_l - \phi_l \|_{L_\infty[-1,1]} = \| (\mathbb{E}_x[\tilde{\phi}_l] - \mathbb{E}_{(l,x)}[\tilde{\phi}_l]) - (\mathbb{E}_x[g_l] - \mathbb{E}_{(l,x)}[g_l]) \|_{L_\infty[-1,1]} \tag{F.17}$$
$$\leq \| \mathbb{E}_x[\tilde{\phi}_l] - (\mathbb{E}_x[g_l] + C) \|_{L_\infty[-1,1]} + \| \mathbb{E}_{(l,x)}[\tilde{\phi}_l] - (\mathbb{E}_{(l,x)}[g_l] + C) \|_{L_\infty} \tag{F.18}$$
$$= O(n^{-3/2}). \tag{F.19}$$

This completes the proof.

F.2 Proof of Proposition 3

Although the proof is again very similar to that of Proposition 1, there are some technical differences. Hence we provide a brief sketch of the proof, avoiding details already highlighted in the proof of Proposition 1.

1. $\mathbf{p} \in \mathcal{S}_{\mathbf{1}}$.

We have for $\tilde{\phi}_p$ that $\mathbb{E}_z[\| \tilde{\phi}_p - (\phi_p + C) \|_{L_\infty[-1,1]}] = O((n^{-1}\log n)^{\frac{3}{7}})$. Denoting $\tilde{\phi}_p(x_p) - (\phi_p(x_p) + C) = b_p(x_p)$, this means $\mathbb{E}_z[|b_p(x_p)|] = O((n^{-1}\log n)^{\frac{3}{7}})$. Now,

$$\mathbb{E}_z[|\mathbb{E}_p[\tilde{\phi}_p - (\phi_p + C)]|] = \mathbb{E}_z[|\mathbb{E}_p[b_p]|] \leq \mathbb{E}_z[\mathbb{E}_p[|b_p|]] = \mathbb{E}_p[\mathbb{E}_z[|b_p(x_p)|]] = O((n^{-1}\log n)^{\frac{3}{7}}). \tag{F.20}$$

The penultimate equality above involves swapping the order of expectations, which is possible by Tonelli's theorem (since $|b_p| > 0$). Then using triangle inequality, it follows that $\mathbb{E}_z[\| \widehat{\phi}_p - \phi_p \|_{L_\infty[-1,1]}] = O((n^{-1}\log n)^{\frac{3}{7}})$.

2. $(\mathbf{l}, \mathbf{l}') \in \mathcal{S}_{\mathbf{2}}$.

We only consider the case where $\rho(l), \rho(l') > 1$ as proofs for the cases are similar. For $\tilde{\phi}_{(l,l')}$, we have that $\mathbb{E}_z[\| \tilde{\phi}_{(l,l')} - (g_{(l,l')} + C) \|_{L_\infty[-1,1]^2}] = O((n^{-1}\log n)^{\frac{3}{8}})$. Denoting $\tilde{\phi}_{(l,l')}(x_l, x_{l'}) - (g_{(l,l')}(x_l, x_{l'}) + C) = b_{(l,l')}(x_l, x_{l'})$, this means $\mathbb{E}_z[|b_{(l,l')}(x_l, x_{l'})|] = O((n^{-1}\log n)^{\frac{3}{8}})$, $\forall (x_l, x_{l'}) \in [-1,1]^2$. Using Tonelli's theorem as earlier, one can next verify that:

$$\mathbb{E}_z[\| \mathbb{E}_l[\tilde{\phi}_{(l,l')}] - (\mathbb{E}_l[g_{(l,l')}] + C) \|_{L_\infty[-1,1]}] = O((n^{-1}\log n)^{\frac{3}{8}}), \tag{F.21}$$

$$\mathbb{E}_z[\| \mathbb{E}_{l'}[\tilde{\phi}_{(l,l')}] - (\mathbb{E}_{l'}[g_{(l,l')}] + C) \|_{L_\infty[-1,1]}] = O((n^{-1}\log n)^{\frac{3}{8}}), \tag{F.22}$$

$$\mathbb{E}_z[|\mathbb{E}_{(l,l')}[\tilde{\phi}_{(l,l')}] - (\mathbb{E}_{(l,l')}[g_{(l,l')}] + C)|] = O((n^{-1}\log n)^{\frac{3}{8}}). \tag{F.23}$$

As in the proof of Proposition 1, we obtain from (F.21), (F.22), (F.23), (F.10), (6.7) (via triangle inequality):

$$\mathbb{E}_z[\| \widehat{\phi}_{(l,l')} - \phi_{(l,l')} \|_{L_\infty[-1,1]^2}] = O((n^{-1}\log n)^{\frac{3}{8}}). \tag{F.24}$$

3. $\mathbf{l} \in \mathcal{S}_{\mathbf{2}}^{\text{var}} : \rho(\mathbf{l}) > \mathbf{1}$.

In this case, for $\tilde{\phi}_l : [-1,1]^2 \to \mathbb{R}$, we have that $\mathbb{E}_z[\| \tilde{\phi}_l - (g_l + C) \|_{L_\infty[-1,1]^2}] = O((n^{-1}\log n)^{\frac{3}{8}})$, with $g_l(x_l, x)$ as defined in (F.12). Using Tonelli's theorem as earlier, one can verify that

$$\mathbb{E}_z[\| \mathbb{E}_x[\tilde{\phi}_l] - (\mathbb{E}_x[g_l] + C) \|_{L_\infty[-1,1]}] = O((n^{-1}\log n)^{\frac{3}{8}}), \tag{F.25}$$

$$\mathbb{E}_z[|\mathbb{E}_{(l,x)}[\tilde{\phi}_l] - (\mathbb{E}_{(l,x)}[g_l] + C)|] = O((n^{-1}\log n)^{\frac{3}{8}}). \tag{F.26}$$

Then using the fact $\mathbb{E}_x[g_l(x_l, x)] - \mathbb{E}_{(l,x)}[g_l(x_l, x)] = \phi_l(x_l)$, we obtain via triangle inequality the bound: $\mathbb{E}_z[\| \widehat{\phi}_l - \phi_l \|_{L_\infty[-1,1]}] = O((n^{-1}\log n)^{\frac{3}{8}})$. This completes the proof.