Learning Sparse Additive Models with Interactions in High Dimensions

Hemant Tyagi
ETH Zürich

Anastasios Kyrillidis
UT Austin, Texas

Bernd Gärtner
ETH Zürich

Andreas Krause
ETH Zürich

Abstract

A function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is referred to as a Sparse Additive Model (SPAM), if it is of the form

\[
    f(x) = \sum_{i \in S} \phi_i(x_i),
\]

where \( S \subset [d] \), \( |S| \ll d \). Assuming \( \phi_i \)'s and \( S \) to be unknown, the problem of estimating \( f \) from its samples has been studied extensively. In this work, we consider a generalized SPAM, allowing for second order interaction terms. For some \( S_1 \subset [d] \), \( S_2 \subset \binom{[d]}{2} \), the function \( f \) is assumed to be of the form:

\[
    f(x) = \sum_{p \in S_1} \phi_p(x_p) + \sum_{(l,l') \in S_2} \phi(l,l')(x_l,x_{l'}). \]

Assuming \( \phi_p, \phi(l,l') \), \( S_1 \) and, \( S_2 \) to be unknown, we provide a randomized algorithm that queries \( f \) and exactly recovers \( S_1, S_2 \). Consequently, this also enables us to estimate the underlying \( \phi_p, \phi(l,l') \). We derive sample complexity bounds for our scheme and also extend our analysis to include the situation where the queries are corrupted with noise – either stochastic, or arbitrary but bounded. 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 [1, 2, 3]. 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 \( C^\alpha \) smooth, then for uniformly approximating \( f \) within error \( \delta \in (0, 1) \), one needs \( n = \Omega(\delta^{-d/\alpha}) \) samples [4].

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; see [5, 6, 7, 8]. One such assumption leads to the class of sparse additive models (SPAMs), wherein:

\[
    f(x_1, \ldots, x_d) = \sum_{l \in S} \phi_l(x_l),
\]

for some unknown \( S \subset \{1, \ldots, d\} \) with \( |S| = k \ll d \). There exist several algorithms for learning these models; we refer to [9, 10, 11, 12, 13] and references therein.

In this paper, we focus on a generalized SPAM model, where \( f \) can also contain a small number of second order interaction terms, i.e.,

\[
    f(x_1, \ldots, x_d) = \sum_{p \in S_1} \phi_p(x_p) + \sum_{(l,l') \in S_2} \phi(l,l')(x_l,x_{l'});
\]

(1.1)

\( S_1 \subset [d], S_2 \subset \binom{[d]}{2} \), with \( |S_1| \ll d, |S_2| \ll d^2 \). There exist relatively few results for learning models of the form (1.1), with the existing work being in the regression framework [14, 15, 16]. Here, \( (x_i, f(x_i))_{i=1}^n \) are typically samples from an unknown probability measure \( P \).

We consider the setting where we have the freedom to query \( f \) at any desired set of points. We propose a strategy for querying \( f \), along with an efficient recovery algorithm, which leads to much stronger guarantees, compared to those known in the regression setting. In particular, we provide the first finite sample bounds for exactly recovering sets \( S_1 \) and \( S_2 \). Subsequently, we uniformly estimate the individual components: \( \phi_p, \phi(l,l') \) via additional queries of \( f \) along the subspaces corresponding to \( S_1, S_2 \).

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

(i) Firstly, we provide a randomized algorithm which provably recovers \( S_1, S_2 \) exactly, with \( O(k\rho_m(\log d)^3) \) noiseless point queries. Here, \( \rho_m \) denotes the maximum number of occurrences of a
variable in $S_2$, and captures the underlying complexity of the interactions.

(ii) An important tool in our analysis is a compressive sensing based sampling scheme, for recovering each row of a sparse Hessian matrix, for functions that also possess sparse gradients. This might be of independent interest.

(iii) We theoretically analyze the impact of additive noise in the point queries on the performance of our algorithm, for two noise models: arbitrary bounded noise and independent, identically distributed (i.i.d.) noise. In particular, for additive Gaussian noise, we show that with $O(p_m^2 k^2 (\log d)^4)$ noisy point queries, our algorithm recovers $S_1, S_2$ exactly. We also provide simulation results on synthetic data that validate our theoretical findings.

Notation. For any vector $x \in \mathbb{R}^d$, we denote its $\ell_p$-norm by $|x|_p := \left(\sum_{i=1}^{d}|x_i|^p\right)^{1/p}$. For a set $S$, $(x)_S$ denotes the restriction of $x$ onto $S$, i.e., $(x)_S = x_i$ if $i \in S$ and 0 otherwise. For a function $g : \mathbb{R}^m \to \mathbb{R}$ of $m$ variables, $E_p[g], E_{(l',l)}[g], E[l,g]$ denote expectation with respect to uniform distributions over $x_{p}, (x_l, x_{l'})$ and $(x_1, \ldots, x_m)$, respectively. For any compact $\Omega \subset \mathbb{R}^d$, $\|g\|_{L^\infty(\Omega)}$ denotes the $L_\infty$ norm of $g$ in $\Omega$. The partial derivative operator $\partial l/\partial x_i$ is denoted by $\partial_i$. For instance, $\partial_l^2 f g$ denotes $\partial_l^2 g / \partial x_i^2 \partial x_j$.

2 Problem statement

We are interested in the problem of approximating functions $f : \mathbb{R}^d \to \mathbb{R}$ from point queries. For some unknown sets $S_1 \subset \{0,1\}^d, S_2 \subset \{0,1\}^d$, the function $f$ is assumed to have the following form.

$$f(x_1, \ldots, x_d) = \sum_{p \in S_1} \phi_p(x_p) + \sum_{(l',l) \in S_2} \phi_{(l',l)}(x_l, x_{l'}).$$

Here, $\phi_{(l',l)}$ is considered to be “truly bivariate” meaning that $\partial_l \partial_{l'} \phi_{(l',l)} \neq 0$. The set of all variables that occur in $S_2$, is denoted by $S_2^{var}$. For each $l \in S_2^{var}$, we refer to $\rho(l)$ as the degree of $l$, i.e., the number of occurrences of $l$ in $S_2$, formally defined as:

$$\rho(l) := |\{l' \in S_2^{var} : (l',l) \in S_2 \text{ or } (l',l) \in S_2\}|: l \in S_2^{var}.$$ \hspace{1cm} (2.1)

The largest such degree is denoted by $\rho_m := \max_{l \in S_2^{var}} \rho(l)$.

Our goal is to query $f$ at suitably chosen points in its domain, in order to estimate it within the compact region $[-1,1]^d$. To this end, note that representation (2.1) is not unique\footnote{One could more generally consider the region $[\alpha, \beta]^d$ and transform the variables to $[-1,1]^d$ via scaling and transformation.} This is avoided by re-writing (2.1) in the following unique ANOVA form [17]:

$$f(x_1, \ldots, x_d) = c + \sum_{p \in S_1} \phi_p(x_p) + \sum_{(l',l) \in S_2} \phi_{(l',l)}(x_l, x_{l'}) + \sum_{q \in S_2^{var}} \phi_q(x_q),$$

where $S_1 \cap S_2^{var} = \emptyset$. Here, $c = E[f]$ and $E_p[\phi_p] = E_{(l',l)}[\phi_{(l',l)}] = 0; \forall p \in S_1, (l',l) \in S_2$, with expectations being over uniform distributions with respect to variable range $[-1,1]$. In addition, $E[\phi_{(l',l)}] = 0$ if $\rho(l) = 1$. The univariate $\phi_q$ corresponding to $q \in S_2^{var}$ with $\rho(q) > 1$, represents the net marginal effect of the variable and has $E_q[\phi_q] = 0$. We note that $S_1, S_2^{var}$ are disjoint in (2.2) as each $p \in S_1 \cap S_2^{var}$ can be merged with their bivariate counterparts, uniquely. The uniqueness of (2.2) is shown formally in the appendix.

We assume the setting $|S_1| = k_1 \ll d, |S_2| = k_2 \ll d^2$. The set of all active variables i.e., $S_1 \cup S_2^{var}$ is denoted by $S$, with $k := |S| = k_1 + |S_2^{var}|$ being the total sparsity of the problem.

Due to the special structure of $f$ in (2.2), we note that if $S_1, S_2$ were known beforehand, then one can estimate $f$ via standard results from approximation theory or from regression.\footnote{We note that $ho_{\max} = \sum_{l \in S_1} \rho(l) > 0$. However, for each $l \in S_1$, we refer to $\rho_l$ as zero throughout $[-1,1]$, then it implies that $\phi_l \equiv 0$, since each $\phi_l$ has zero mean in (2.2).}

Our main assumptions for this problem are listed below.

Assumption 1. $f$ can be queried from the slight enlargement: $[-(1+r), (1+r)]^d$, for some small $r > 0$.

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

$$\|\partial_l^{m_1} \partial_{l'}^{m_2} \phi_{(l',l)}\|_{L_\infty[-(1+r), (1+r)]^2} \leq B_m; m_1 + m_2 = m, \quad \text{where } (l',l) \in S_2, \quad \|\partial_l^{m} \phi_p\|_{L_\infty[-(1+r), (1+r)]} \leq B_m,$$

where $p \in S_1$ or, $p \in S_2^{var}$ and $\rho(p) > 1$.

Our next assumption is for identifying $S_1$.

Assumption 3. For some constants $D_1, \lambda_1 > 0$, we assume that for each $p \in S_1$, $\exists$ connected $I_p \subset [-1,1]$ of Lebesgue measure at least $\lambda_1 > 0$, such that $|\partial_{l'} \phi_p(x_p)| > D_1, \forall x_p \in I_p$. This assumption is in a sense necessary. If say $\partial_{l'} \phi_p$ was zero throughout $[-1,1]$, then it implies that $\phi_p \equiv 0$, since each $\phi_p$ has zero mean in (2.2).
Our last assumption concerns the identification of $S_2$.

**Assumption 4.** For some constants $D_2, \lambda_2 > 0$, we assume that for each $(l, l') \in S_2$, $\exists$ connected $I_l, I_{l'} \subset [-1, 1]$, each interval of Lebesgue measure at least $\lambda_2 > 0$, such that $|\partial_l \partial_{l'} \phi(l, l')(x_1, x_2)| > D_2$, $\forall (x_1, x_2) \in I_l \times I_{l'}$.

Given the above, our problem specific parameters are: (i) $B_i$, $i = 0, \ldots, 3$, (ii) $D_j$, $\lambda_j$, $j = 1, 2$ and, (iii) $k, \rho_m$. 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: $D_j, \lambda_j; j = 1, 2$.

### 3 Our sampling scheme and algorithm

We start by explaining our sampling scheme, followed by our algorithm for identifying $S_1, S_2$. Our algorithm proceeds in two phases – we first estimate $S_2$ and then $S_1$. Its theoretical properties for the noiseless query setting are described in Section[4] Section[5] then analyzes how the sampling conditions can be adapted to handle the noisy query setting.

#### 3.1 Sampling scheme for estimating $S_2$

Our main idea for estimating $S_2$ is to estimate the off-diagonal entries of the Hessian of $f$, at appropriately chosen points. The motivation is the observation that for any $(l, l') \in S_2^{[2]}$:

$$\partial_l \partial_{l'} f = \begin{cases} \partial_l \partial_{l'} \phi(l, l') & \text{if } (l, l') \in S_2, \\ 0 & \text{otherwise.} \end{cases}$$

To this end, consider the Taylor expansion of the gradient $\nabla f$, at $x \in \mathbb{R}^d$, along the direction $v' \in \mathbb{R}^d$, with step size $\mu_1$. Since $f$ is $C^3$ smooth, we have for $\zeta_q = x + \theta_q v'$, for some $\theta_q \in (0, \mu_1)$, $q = 1, \ldots, d$:

$$\nabla f(x + \mu_1 v') - \nabla f(x) \approx \frac{\mu_1}{2} \begin{pmatrix} v'^T \nabla^2 \partial_l f(\zeta_q) v' \\ \vdots \\ v'^T \nabla^2 \partial_{l'} f(\zeta_q) v' \end{pmatrix}. \tag{3.1}$$

We see from (3.1) that the $l^{th}$ entry of $(\nabla f(x + \mu_1 v') - \nabla f(x))/\mu_1$, corresponds to a “noisy” linear measurement of the $l^{th}$ row of $\nabla^2 f(x)$ with $v'$. The noise corresponds to the third order Taylor remainder terms of $f$.

Denoting the $l^{th}$ row of $\nabla^2 f(x)$ by $\nabla \partial_l f(x)$, we make the following crucial observation: if $l \in S_2^{[2]}$ then $\nabla \partial_l f(x)$ has at most $\rho_m$ non-zero off-diagonal entries, implying that it is $(\rho_m + 1)$ sparse. This follows on account of the structure of $f$ (2.2). Furthermore, if $l \in S_1$ then $\nabla \partial_l f(x)$ has at most one non zero entry (namely the diagonal entry), while if $l \notin S$, then $\nabla \partial_l f(x) \equiv 0$.

**Compressive sensing based estimation.** Assuming now that we have access to an oracle that provides us with gradient estimates of $f$, this suggests the following idea. We can obtain random linear measurements, for each row of $\nabla^2 f(x)$ via gradient differences, as in (3.1). As each row is sparse, it is known from compressive sensing (CS) [18] [19] that it can be recovered with only a few measurements.

Inspired by this observation, consider an oracle that provides us with the estimates: $\hat{\nabla} f(x), \{\hat{\nabla} f(x + \mu_1 v'_j)\}_{j=1}^{m_\nu}$, where $v'_j$ belong to the set:

$$V' = \{v'_j \in \mathbb{R}^d : v'_{j,q} = \pm 1/\sqrt{m_\nu} \text{ w.p. } 1/2 \text{ each;} \quad j = 1, \ldots, m_\nu \text{ and } q = 1, \ldots, d\}.$$ 

Let $\hat{\nabla} f(x) = \nabla f(x) + w(x)$, where $w(x) \in \mathbb{R}^d$ denotes the gradient estimation noise. Denoting $V' = [v'_1 \ldots v'_m]_\nu^T$, we obtain $d$ linear systems, by employing (3.1) at each $v'_j \in V'$:

$$y_q = V' \nabla \partial_q f(x) + \eta_{q,1} + \eta_{q,2}; \quad q = 1, \ldots, d. \tag{3.2}$$

$y_q \in \mathbb{R}^{m_\nu}$ represents the measurement vector for the $q^{th}$ row, with

$$(y_q)_j = (\hat{\nabla} f(x + \mu_1 v'_j) - \nabla f(x))/\mu_1$$

while $\eta_{q,1}, \eta_{q,2} \in \mathbb{R}^{m_\nu}$ represent noise with $(\eta_{q,1})_j = (\mu_1/2)v'_j \nabla^2 \partial_q f(\zeta_q) v'_j$ and $(\eta_{q,2})_j = (w_q(x + \mu_1 v'_j) - w_q(x))/\mu_1$. Given the measurement vector $y_q$, we can then obtain the estimate $\hat{\nabla} \partial_q f(x)$ individually for each $q = 1, \ldots, d$, via $\ell_1$ minimization [18] [19] [20].

**Estimating sufficiently many Hessian’s.** Having estimated each row of $\nabla^2 f$ at some fixed $x$, we have at hand an estimate of the set: $\{\partial_l \partial_{l'} f(x) : (l, l') \in S_2^{[2]}\}$. Our next goal is to repeat the process, at sufficiently many $x$’s within $[-1, 1]^d$.

We will denote the set of such points as $\chi$. This will then enable us to sample each underlying $\partial_l \partial_{l'} \phi(l, l')$ within its respective critical interval, as defined in Assumption 4.

Roughly speaking, since $|\partial_l \partial_{l'} \phi(l, l')|$ is “suitably large” in such an interval, we will consequently be able to detect each $(l, l') \in S_2$, via a thresholding procedure. To this end, we make use of a family of hash functions, defined as follows.

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

Hash functions are widely used in theoretical computer science, such as in finding juntas [21]. There exists a simple probabilistic method for constructing such a family, so that
for any constant \( C > 1 \), \(|\mathcal{H}_d^2| \leq (C+1)te^t \log d\) with high probability (w.h.p)\(^5\). For our purposes, we consider the family \( \mathcal{H}_d^2 \) so that for any distinct \( i, j \), there exists \( h \in \mathcal{H}_d^2 \) such that \( h(i) \neq h(j) \).

For any \( h \in \mathcal{H}_d^2 \), let us now denote the vectors \( \mathbf{e}_1(h), \mathbf{e}_2(h) \in \mathbb{R}^d \) where

\[
(\mathbf{e}_1(h))_q = \begin{cases} 1 & \text{if } h(q) = i, \\ 0 & \text{otherwise}, \end{cases}
\]

for \( i = 1, 2 \) and \( q = 1, \ldots, d \). Given at hand \( \mathcal{H}_d^2 \), we construct our set \( \chi \) using the procedure\(^6\). Specifically, for some \( m_x \in \mathbb{Z}^+ \), we construct for each \( h \in \mathcal{H}_d^2 \) the set:

\[
\chi(h) := \left\{ x(h) \in [-1,1]^d : x(h) = \sum_{i=1}^2 c_i \mathbf{e}_i(h) ; \right\}
\]

\[
c_1, c_2 \in \left\{ -\frac{m_x-1}{m_x}, \ldots, \frac{m_x-1}{m_x}, 1 \right\}.
\]

Then, we obtain \( \chi = \bigcup_{h \in \mathcal{H}_d^2} \chi(h) \) as the set of points at which we will recover \( \nabla^2 f \). Observe that \( \chi \) has the property of discretizing any 2-dimensional canonical subspace, within \([-1,1]^d\) with \(|\chi| \leq (2m_x + 1)^2 |\mathcal{H}_d^2| = O(\log d)\).

**Estimating sparse gradients.** Note that \( \nabla f \) is at most \( k \) sparse, due to the structure of \( f \). We now describe the oracle that we use, for estimating sparse gradients. As \( f \) is \( C^3 \) smooth, therefore the third order Taylor’s expansion of \( f \) at \( x \), along \( v, -v \in \mathbb{R}^d \), with step size \( \mu > 0 \), and \( \zeta = x + \theta v, \zeta' = x - \theta' v; \theta, \theta' \in (0, \mu) \) leads to

\[
\nabla f(x + \mu v) = \nabla f(x) + (R_3(\zeta) - R_3(\zeta'))/(2\mu).
\]

\[(3.3)\] corresponds to a noisy-linear measurement of \( \nabla f(x) \), with \( v \). The “noise” here arises on account of the third order terms \( R_3(\zeta), R_3(\zeta') = O(\mu^3) \), in the Taylor expansion.

Let \( \mathcal{V} \) denote the set of measurement vectors:

\[
\mathcal{V} := \{ v_j \in \mathbb{R}^d : v_j,q = \pm 1/\sqrt{m_x} \text{ w.p. } 1/2 \text{ each}; \]

\[
j = 1, \ldots, m_x \text{ and } q = 1, \ldots, d \}.
\]

Employing \[(3.3)\] at each \( v_j \in \mathcal{V} \), we obtain:

\[
y = V \nabla f(x) + n.
\]

Here, \( y \in \mathbb{R}^{m_x} \) denotes the measurement vector with \( (y)_j = (f(x + \mu v_j) - f(x - \mu v_j))/(2\mu) \). Also, \( V = \lfloor v_1 \ldots v_{m_x} \rfloor \in \mathbb{R}^{m_x \times d} \) denotes the measurement matrix and \( n \in \mathbb{R}^{m_x} \) denotes the noise terms. We then estimate \( \nabla f(x) \) via standard \( \ell_1 \) minimization\(^7\). Estimating sparse gradients via CS, has been considered previously in\(^8\), albeit using second order Taylor expansions, for different function models.

### 3.2 Sampling scheme for estimating \( S_1 \)

Having obtained an estimate \( \hat{S}_2 \) of \( S_2 \) we now proceed to estimate \( S_1 \). Let \( \hat{S}_2^{\text{var}} \) denote the set of variables in \( \hat{S}_2 \) and \( \mathcal{P} := [d] \setminus \hat{S}_2^{\text{var}} \). Assuming \( \hat{S}_2 = S_2 \), we are now left with a SPAM on the reduced variable set \( \mathcal{P} \). Consequently, we employ the sampling scheme of\(^9\), wherein the gradient of \( f \) is estimated at equispaced points, along a diagonal of \([-1,1]^d \). For \( m_x' \in \mathbb{Z}^+ \), this set is defined as:

\[
\chi_{\text{diag}} := \left\{ x = (x \ x \ldots x) \in \mathbb{R}^d : \right\}
\]

\[
x \in \left\{ -1, -\frac{m_x' - 1}{m_x'}, \ldots, \frac{m_x' - 1}{m_x'}, 1 \right\}.
\]

Note that \(|\chi_{\text{diag}}| = 2m_x' + 1 \). The motivation for estimating \( \nabla f \) at \( x \in \chi_{\text{diag}} \) is that we obtain estimates of \( \partial_p \partial_p \) at equispaced points within \([-1,1]\), for \( p \in S_1 \). With a sufficiently fine discretization, we would “hit” the critical regions associated with each \( \partial_p \partial_p \), as defined in Assumption\(^10\). By applying a thresholding operation, we would then be able to identify each \( p \in S_1 \).

To this end, consider the set of sampling directions:

\[
\mathcal{V}' := \{ v_{j,q}' \in \mathbb{R}^d : v_{j,q}' = \pm 1/\sqrt{m_x'} \text{ w.p. } 1/2 \text{ each}; \]

\[
j = 1, \ldots, m_x' \text{ and } q = 1, \ldots, d \},
\]

and let \( \mu' > 0 \) denote the step size. For each \( x \in \chi_{\text{diag}} \), we will query \( f \) at points: \((x + \mu v_j)'_p, (x - \mu v_j)'_p; v_j' \in \mathcal{V}'\), restricted to \( \mathcal{P} \). Then, as described earlier, we can form a linear system consisting of \( m_x' \) equations, and solve it via \( \ell_1 \) minimization to obtain the gradient estimate. The complete procedure for estimating \( S_1, S_2 \), is described formally in Algorithm\(^1\).
Algorithm 1 Algorithm for estimating $S_1, S_2$

1: **Input:** $m_v, m_w, m_x, m_x' \in \mathbb{Z}^+$; $\mu, \mu_1, \mu' > 0$; $\tau' > 0$, $\tau'' > 0$.
2: **Initialization:** $\widehat{S}_1, \widehat{S}_2 = \emptyset$.
3: **Output:** Estimates $\widehat{S}_2, \widehat{S}_1$.
4: $\quad$
5: Construct the $(d, 2)$-hash family $\mathcal{H}_d^2$ and sets $\mathcal{V}, \mathcal{V}'$.
6: for $h \in \mathcal{H}_d^2$ do
7: Construct the set $\chi(h)$.
8: for $i = 1, \ldots, (2m_x + 1)$ and $x \in \chi(h)$ do
9: $(y_i)_j = (f(x_i + \mu v_j) - f(x_i - \mu v_j))_j$; $j = 1, \ldots, m_v$; $v_j \in \mathcal{V}$.
10: $\nabla f(x_i) := \text{argmin} \| \mathbf{z} \|_1$.
11: for $p = 1, \ldots, m_v$ do
12: $(y_{i,p}) = (f(x_i + \mu_1 v_p) - f(x_i - \mu_1 v_p))_j$; $j = 1, \ldots, m_v$; $v_p \in \mathcal{V}'$.
13: $\nabla f(x_i + \mu_1 v_p) := \text{argmin} \| \mathbf{z} \|_1$.
end for
14: for $q = 1, \ldots, d$ do
15: $(y_{i,q}) = (f(x_i + \mu v_j) - f(x_i - \mu v_j))_j$; $j = 1, \ldots, m_v$.
16: $\nabla f(x_i) := \text{argmin} \| \mathbf{z} \|_1$.
17: $\widehat{S}_2 = \widehat{S}_2 \cup \left\{ (q, q') : q' \in \{q + 1, \ldots, d\} \text{ and } |(\nabla f(x_i))_q'| > \tau' \right\}$.
end for
end for
21: end for
22: Construct the sets $\chi_{\text{diag}}, \mathcal{V}''$ and initialize $\mathcal{P} := [d] \setminus \widehat{S}_2^{\text{var}}$.
23: for $i = 1, \ldots, (2m_x + 1)$ and $x \in \chi_{\text{diag}}$ do
24: $(y_i)_j = (f(x_i + \mu v_j - \mu v_j))_j$; $j = 1, \ldots, m_v$; $v_j \in \mathcal{V}''$.
25: $(\nabla f((x_i)_{\mathcal{P}}))_p := \text{argmin} \| (\mathbf{z})_p \|_1$.
26: $\widehat{S}_1 = \widehat{S}_1 \cup \left\{ q \in \mathcal{P} : |(\nabla f((x_i)_{\mathcal{P}}))_q| > \tau'' \right\}$.
27: end for
28: end for

4 Theoretical guarantees for noiseless case

Next, we provide sufficient conditions on our sampling parameters that guarantee exact recovery of $S_1, S_2$, in the noiseless query setting. This is stated in the following Theorem. All proofs are deferred to the appendix.

**Theorem 1.** There exist positive constants $\{c'_3\}_{i=1}^3, \{C_1\}_{i=1}^3$ so that if: $m_x \geq \lambda_2^{-1}, m_v > c'_3 k \log(d/k)$, and $m_w > c'_2 \rho_m \log(d/\rho_m)$, then the following holds. Denoting $a = 2^{d+1} \rho_m^{1/2}$, $b = C_2 \sqrt{\lambda_2^{-1} d/m_v}, a' = D_2 \sqrt{d/m_v}$, let $\mu, \mu_1$ satisfy: $\mu'^2 < (a'^2/a)/b$ and $\mu_1 \in (a' - \sqrt{a'^2 - (b/\mu_1)^2}, a' + \sqrt{a'^2 - (b/\mu_1)^2})$.

We then have that for $\tau' = C_2 (\mu_1 + b/\mu_1)$, that $\widehat{S}_2 = S_2$ w.h.p. Provided $\widehat{S}_2 = S_2$, if $m_x \geq \lambda_2^{-1}, m_v > c'_3 (k - |\widehat{S}_2^{\text{var}}|)$ and $\mu'^2 < \frac{3m_v D_2}{C_3 k (k - |\widehat{S}_2^{\text{var}}|)B_2}$, then $\tau'' = C_3 (k - |\widehat{S}_2^{\text{var}}|)^{1/2} b_{2/3}$, implies $\widehat{S}_1 = S_1$ w.h.p.

**Remark 1.** We note that the condition on $\mu'$ is less strict than in [12] for identifying $S_1$. This is because in [12], the gradient is estimated via a forward difference procedure, while we perform a central difference procedure in [3.3].

**Query complexity.** Estimating $\nabla f(x)$ at some fixed $x$ requires $2m_v = O(k \log d)$ queries. Estimating $\nabla^2 f(x)$ involves computing an additional $m_v$ queries for $\nabla f(x)$ in a neighborhood of $x$ – implying $O(m_v m_w) = O(\rho_m \log d)$ point queries. This consequently implies a total query complexity of $O(k \log^2 d)$ for estimating $S_1$ with high probability. We make an additional $O(\lambda_2^{-2} k \rho_m \log d)$ point queries of $f$, in order to estimate $S_1$. Therefore, the overall query complexity for estimating $S_1, S_2$ is $O(\lambda_2^{-2} k \rho_m \log d)$. 


Learning Sparse Additive Models with Interactions in High Dimensions

$\mathcal{H}_d^f$ can be constructed in poly($d$) time. For each $x \in \chi$, we first solve $m_v + 1$ linear programs (Steps 10-13), each solvable in poly($m_v$, $d$) time. We then solve $d$ linear programs (Step 17), with each taking poly($m_v$, $d$) time. This is done at $|\chi| = O(\lambda_2^{-2} \log d)$ points, hence the overall computation cost for estimation of $S_2$ (and later $S_1$) is polynomial in: the number of queries, and $d$. Lastly, we note that [23] also estimates sparse Hessians via CS, albeit for the function optimization problem. Their scheme entails a sample complexity$\mathcal{O}(k \log(k m_v)^2(\log d)^2)$ for estimating $\nabla^2 f(x)$; this is worse by a $O((\log(k m_v)^2)^2)$ term compared to our method.

Recovering the components of the model. Having estimated $S_1, S_2$, we can now estimate each underlying component in (2.2) by sampling $f$ along the subspace corresponding to the component. Using these samples, one can then construct via standard techniques, a spline based quasi interpolant [23] that uniformly approximates the component. This is shown formally in the appendix.

5 Impact of noise

We now consider the case where the point queries are corrupted with external noise. This means that at query $x$, we observe $f(x) + z'$, where $z' \in \mathbb{R}$ denotes external noise.

In order to estimate $\nabla f(x)$, we obtain the samples: $f(x + \mu v_j) + z_{j,1}'$ and $f(x - \mu v_j) + z_{j,2}'; \ j = 1, \ldots, m_v$. This changes (3.4) to the linear system $y = \nabla f(x) + n + z$, where $z_j = (z_{j,1}' - z_{j,2}')/(2\mu)$. Hence, the step-size $\mu$ needs to be chosen carefully now – a small value would blow up the external noise component, while a large value would increase perturbation due to the higher order Taylor’s terms.

Arbitrary bounded noise. In this scenario, we assume the external noise to be arbitrary and bounded, meaning that $|z'| < \varepsilon$, for some finite $\varepsilon \geq 0$. If $\varepsilon$ is too large, then we would expect recovery of $S_1, S_2$ to be impossible, since $f$ would be destroyed.

We show in Theorem 2 that if $\varepsilon < \varepsilon_1 = O \left( \frac{d^2}{B_1^2 \rho_m \sqrt{\chi}} \right)$, then Algorithm 1 recovers $S_2$ with appropriate choice of sampling parameters. Furthermore, assuming $S_2$ is recovered exactly, and provided $\varepsilon$ additionally satisfies $\varepsilon < \varepsilon_2 = O \left( \frac{d^2}{\sqrt{B_1 \rho_m \sqrt{\chi}}} \right)$, then the algorithm also recovers $S_1$ exactly. In contrast to Theorem 1, the step size $\mu$ cannot be chosen arbitrarily small now, due to external noise.

Theorem 2. Let $m_x, m_v, m_v', m_v''$, be as defined in Theorem 1. Say $\varepsilon < \varepsilon_1 = O \left( \frac{d^2}{B_1^2 \rho_m \sqrt{\chi}} \right)$. Denoting $b' = 2C_1 \sqrt{\rho_m/m_v}$, we have $S_2 = S_2$ w.h.p. Given $S_2 = S_2$, denote $a_1 = (k - |S_2'|)B_3/(6m_v)$, $b_1 = \sqrt{m_v}$ and say $\varepsilon < \varepsilon_2 = O \left( \frac{d^2}{\sqrt{(|k - |S_2'|)|B_3}} \right)$. If $0 < A_5 < A_6$ so that $\mu' \in (A_5, A_6)$, then $\tau'' = C_3(a_1 \mu'^2 + \frac{b_1 \varepsilon}{\rho_m})$ implies $S_1 = S_1$ w.h.p.

Stochastic noise. We now assume the point queries to be corrupted with i.i.d. Gaussian noise, so that $z' \sim \mathcal{N}(0, \sigma^2)$ with variance $\sigma^2$. We consider resampling each point query a sufficient number of times, and averaging the values. During the $S_2$ estimation phase, we resample each query $N_1$ times so that $z' \sim \mathcal{N}(0, \sigma^2/N_1)$. For any $0 < \varepsilon < \varepsilon_1$, if $N_1$ is suitably large, then we can uniformly bound $|z'| < \varepsilon$ – via standard tail bounds for Gaussians – over all noise samples, with high probability. Consequently, we can use the result of Theorem 2 for estimating $S_2$. The same reasoning applies to Step 25 i.e., the $S_1$ estimation phase, where we resample each query $N_2$ times.

Theorem 3. Let $m_x', m_v', m_v', m_v''$, be as defined in Theorem 4. For any $\varepsilon < \varepsilon_1, 0 < p_1 < 1$, say we resample each query in Steps 9, 12 $N_1 > \frac{d^2}{\sigma^2} \log \left( \frac{\sigma^2}{\epsilon'} \right)(m_v + 1)/(2m_x + 1)^2 |S_2'|$ times, and take the average. For $\mu, \mu'$, $\tau'$ as in Theorem 2 we have $S_2 = S_2$ with probability at least $1 - p_1 - o(1)$. Given $S_2 = S_2$, with $\varepsilon' < \varepsilon_2$, we resample each query in Step 25 $N_2 \geq \frac{\sigma^2}{\mu'} \log \left( \frac{\sigma^2}{\epsilon'} \right)^2(p_2')m_v''$ times, and take the average. Then for $\mu', \tau''$ as in Theorem 2 (with $\varepsilon$ replaced by $\varepsilon'$), we have $S_1 = S_1$ with probability at least $1 - p_2 - o(1)$.

Query complexity. In the case of arbitrary, but bounded noise, the query complexity remains the same as for the noiseless case. In case of i.i.d. Gaussian noise, for estimating $S_2$, we have $\varepsilon = O(\rho_m^{-2}k^{-1/2})$. Choosing $p_1 = d^{-\delta}$ for any constant $\delta > 0$ gives us $N_1 = O(\rho_m^{-2}k \log d)$. This means that with $O(N_1 k \rho_m \log d) |\chi| = O(\rho_m^{-2}k \log d)^2 \lambda_2^2$ queries, $S_2 = S_2$ holds w.h.p. Next, for estimating $S_1$, we have $\varepsilon' = O((k - |S_2'|)^{-1/2})$. Choosing $p_2 = ((d - |S_2'|)^{-\delta})$ for any constant $\delta > 0$, we get $N_2 = O((k - |S_2'|)^{-1/2}) \log (d - |S_2'|)^2 \lambda_2^2$ queries.

Remark 1. We have stated above that $O(k^2 \log d)$ samples are sufficient for estimating $S_1$ in presence of i.i.d Gaussian noise. This improves the corresponding bound in [12] by a $O(k)$ factor, and is due to the less strict condition on $\mu'$.

Recovering the components of the model. Having identified $S_1, S_2$, we can estimate the underlying components in (2.2), via standard nonparametric regression for ANOVA type models [25]. Alternately, for each component, we could also sample $f$ along the subspace corresponding to
the component and then perform regression, to obtain its estimate with uniform error bounds. This is shown formally in the appendix.

6 Related work

Learning SPAMs. We begin with an overview of results for learning SPAMs, in the regression setting. [14] proposed the COSSO algorithm, that extends the Lasso to the reproducing kernel Hilbert space (RKHS) setting. [26] generalizes the non negative garrote to the nonparametric setting. [27, 9, 10] consider least squares methods, regularized by sparsity inducing penalty terms, for learning such models. [12, 23] propose a convex program for estimating \( f \) (in the RKHS setting) that achieves the minimax optimal error rates. [11] proposes a method based on the adaptive group Lasso. These methods are designed for learning SPAMs and cannot handle models of the form (1.1).

Learning generalized SPAMs. There exist fewer results for generalized SPAMs of the form (1.1), in the regression setting. The COSSO algorithm [14] can handle (1.1), however its convergence rates are shown only for the case of no interactions. [15] proposes the VANISH algorithm – a least squares method with sparsity constraints. It is shown to be sparsistent, i.e., it asymptotically recovers \( S_1, S_2 \) for \( n \to \infty \). They also show a consistency result for estimating \( f \), similar to [9]. [16] proposes the ACOSSO method, an adaptive version of the COSSO algorithm, which can also handle (1.1). They derive convergence rates and sparsistency results for their method, albeit for the case of no interactions. [29] studies 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 \( f \), they do not guarantee unique identification of the interaction terms for any value of \( m \). A special case of (1.1) – where \( \phi_y \)'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 \( S_1, S_2 \), along with convergence rates for estimating \( f \), but only in the limit of large \( n \). [30, 15, 31], [32] 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. However finite sample bounds for the non-linear model (1.1) are not known in general.

Learning generic low-dimensional function models. There exists related work in approximation theory – which is also the setting considered in this paper – wherein one assumes freedom to query \( f \) at any desired set of points within its domain. [5] considers functions depending on an unknown subset \( S (|S| = k) \) of the variables – a more general model than (1.1). They provide a choice of query points of size \( O(k^c \log d) \) for some constant \( c > 1 \), and algorithms that recover \( S \) w.h.p.\(^{33}\) derived a simpler algorithm with sample complexity \( O((C_1^c \rho^c)k \log d^2) \) for recovering \( S \) w.h.p., where \( C_1, \rho \) depend\(^*\) on smoothness of \( f \). For general \( k \)-variate \( f: \alpha = c k^{\alpha} \) for some constant \( c > 1 \), while for our model (1.1): \( C_1 = O(\rho_m) \). This model was also studied in [34, 35] in the regression setting – they proposed an estimator that recovers \( S \) w.h.p, with sample complexity \( O(k \log d) \). [8, 7] generalize this model to functions \( f(x) = g(Ax) \), for unknown \( A \in \mathbb{R}^{k \times d} \). They derive algorithms that approximately recover the row-span of \( A \) w.h.p, with sample complexities typically polynomial in \( d \).

While the above methods could possibly recover \( S \), they are not designed for identifying interactions among the variables. Specifically, their sample complexities exhibit a worse dependence on \( k, \rho_m \) and/or \( d \). [13] provides a sampling scheme that specifically learns SPAMs, with sample complexities \( O(k \log d), O(k^c (\log d)^2) \), in the absence/presence of Gaussian noise, respectively.

7 Simulation results

Dependence on \( d \). We first consider the following experimental setup: \( S_1 = \{1, 2\} \) and \( S_2 = \{3, 4, 5\} \), which implies \( k_1 = 2, k_2 = 2, \rho_m = 2 \) and \( k = 5 \). We consider two models:

\[(i) \ f_1(x) = 2x_1 - 3x_2^2 + 4x_3x_4 - 5x_4x_5, \]
\[(ii) \ f_2(x) = 10 \sin(\pi \cdot x_1) + 5e^{-2x_2} + 10 \sin(\pi \cdot x_3x_4) + 5e^{-2x_4x_5}. \]

We begin with the relatively simple model \( f_1 \), for which the problem parameters are set to: \( \lambda_1 = 0.3, \lambda_2 = 1, D_1 = 2, D_2 = 3, B_3 = 6 \). We obtain \( m_x = 1, m \rho = 4 \). We use the same constant \( C \) when we set \( m_v := \tilde{C} \rho m \log (d/k), m_v' := \tilde{C} \rho m \log (d/\rho_m), \) and \( m_v'' := \tilde{C} (k - |S_2'|) \log (\frac{1}{k - |S_2'|}) \).

For the construction of the hash functions, we set the size to \( |\mathcal{H}_3| = C' \log d \) with \( C' = 1.7 \), leading to \( |\mathcal{H}_3| \in [8, 12] \) for \( 10^2 \leq d \leq 10^3 \). We choose step sizes: \( \mu, \mu_1, \mu' \) and thresholds: \( \tau, \tau'' \) as in Theorem 2. As CS solver, we use the ALPS algorithm [36], an efficient first-order method.

For the noisy setting, we consider the function values to be corrupted with i.i.d. Gaussian noise. The noise variance values considered are: \( \sigma^2 \in \{10^{-4}, 10^{-3}, 10^{-2}\} \) for which we choose resampling factors: \( (N_1, N_2) \in \{(50, 20), (85, 36), (90, 40)\} \). We see in Fig. 2 that for \( C \approx 5.6 \) the probability of successful identification (noiseless case) undergoes a phase transition and becomes close to 1, for different values of \( d \). This validates Theorem 1.

Fixing \( C = 5.6 \), 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

\[^{33}C_1 = \max_{i \in S} ||\partial_i f||_\infty \quad \text{and} \quad \alpha = \min_{i \in S} ||\partial_i f||_1 \]
roughly $10^2$ times that in the noiseless setting, however the scaling with $d$ is similar to the noiseless case.

We next consider the relatively harder model: $f_2$, where the problem parameters are set to: $\lambda_1 = \lambda_2 = 0.3$, $D_1 = 8$, $D_2 = 4$, $B_3 = 35$ and, $m_x = m_y = 4$. We see in Fig. 2 a phase transition (noiseless case) at $C = 5.6$ thus validating Theorem 1. For noisy cases, we consider $\sigma^2$ as before, and $(N_1, N_2) \in \{(60, 30), (90, 40), (95, 43)\}$. The number of queries is seen to be slightly larger than that for $f_1$.

**Dependence on $k$.** We now demonstrate the scaling of the total number of queries versus the sparsity $k$ for identification of $S_1, S_2$. Consider the model $f_3(x) = \sum_{i=1}^{T} (\alpha_1 x_{(i-1)5+1} - \alpha_2 x_{(i-1)5+2} + \alpha_3 x_{(i-1)5+3} x_{(i-1)5+4} - \alpha_4 x_{(i-1)5+4} x_{(i-1)5+5})$ where $x \in \mathbb{R}^d$ for $d = 500$. Here, $\alpha_i \in [2, 5], \forall i$; i.e., we randomly selected $\alpha_i$’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 $C = 5.6$. For the noisy cases, we consider $\sigma^2$ as before, and choose the same values for $(N_1, N_2)$ as for $f_1$. In Figure 3 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.

**Dependence on $\rho_m$.** We now demonstrate the scaling of the total queries versus the maximum degree $\rho_m$ for identification of $S_1, S_2$. Consider the model $f_4(x) = \alpha_1 x_1 - \alpha_2 x_2^2 + \sum_{i=1}^{T} (\alpha_3 x_i x_{i+3}) + \sum_{i=1}^{T} (\alpha_4 x_{2i+2} x_{3i+2})$. We choose $d = 500$, $C = 6$, $\alpha_i \in [2, \ldots, 5], \forall i$ (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 $\sigma^2$ as before, and choose $(N_1, N_2) \in \{(70, 40), (90, 50), (100, 70)\}$. In Figure 4 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 Concluding remarks.**

We proposed a sampling scheme for learning a generalized SPAM and provided finite sample bounds for recovering the underlying structure of such models. We also considered the setting where the point queries are corrupted with noise and analyzed sampling conditions for the same. It would be interesting to improve the sampling bounds that we obtained, and under similar assumptions. We leave this for future work.

**Acknowledgements.** This research was supported in part by SNSF grant CRSII2_147633.
References


