Learning Sparse Additive Models with Interactions in High Dimensions

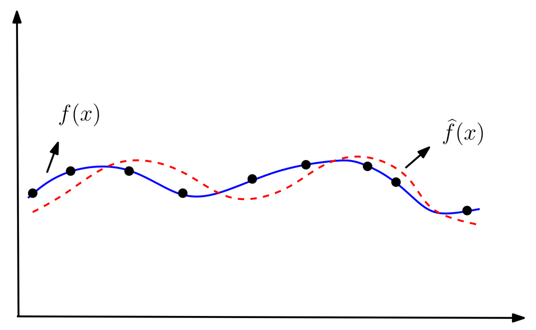
H. Tyagi, A. Kyrillidis, B. Gärtner, A. Krause
Emails: anastasios@utexas.edu; krausea@ethz.ch; {htyagi,gaertner}@inf.ethz.ch

ETH zürich
Department of Computer Science
Universitätstrasse 6, CH-8092 Zürich

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Introduction

- Unknown smooth function $f : \mathbb{R}^d \to \mathbb{R}$.
- **Given:** $\{(\mathbf{x}_1, f(\mathbf{x}_1)), \dots, (\mathbf{x}_n, f(\mathbf{x}_n))\}$; $\mathbf{x}_i \in G$, where compact $G \subset \mathbb{R}^d$.
- **Goal:** Using $\{(\mathbf{x}_i, f(\mathbf{x}_i))\}_{i=1}^n$, construct $\widehat{f} : G \to \mathbb{R}$.
- **Applications:** Biological systems, Solving PDE's etc.



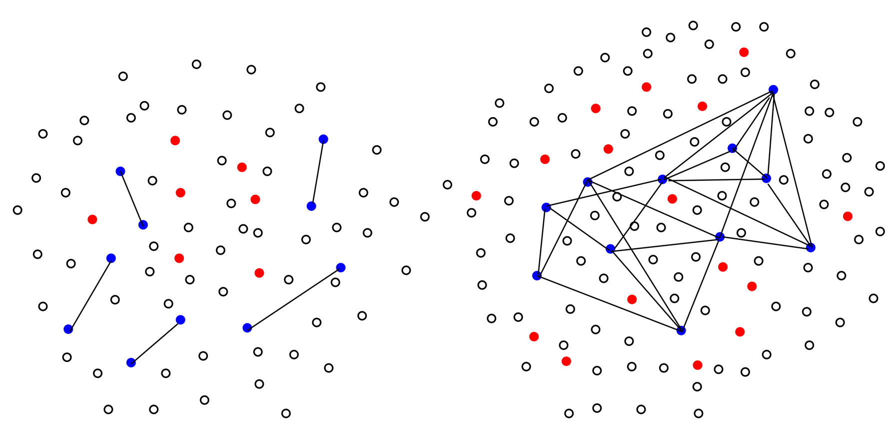
Curse of dimensionality: For C^r smooth f, $n = \Omega(\delta^{-d/r})$ samples needed to ensure $\|f - \widehat{f}\|_\infty \le \delta$ for any $\delta \in (0,1)$ (Traub et al. '88).

- Additional assumption on f – low intrinsic dimension. Examples:
  - $f(\mathbf{x}) = g(\mathbf{x}_\mathcal{S})$; k active vars.
  - $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$; k dim. subspace.
  - $f(\mathbf{x}) = \sum_{p \in \mathcal{S}} \phi_p(x_p)$; Sparse additive models (SPAMs).

SPAMs with pairwise interactions

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

- l and l' interact $\Leftrightarrow \partial_l \partial_{l'} \phi_{(l,l')} \not\equiv 0$.



- Existing work:
  - Identify $\mathcal{S}_1, \mathcal{S}_2$ as $n \to \infty$ (Radchenko et al.'10).
  - Estimating f in L_2 norm. (Dalalyan et al.'14)
  - Special case: ϕ is multi-linear. (Nazer et al.'10)

Problem Setup

- **Setting:** Freedom to query f within $[-1,1]^d$.
- $|\mathcal{S}_1 \cup \mathcal{S}_2^{\text{var}}| = k$ and ρ_m – maximum degree of a variable in interaction graph.
- Unique ANOVA rep. for f:

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

Goal: Identify $\mathcal{S}_1, \mathcal{S}_2$ from few queries; then uniformly estimate each ϕ.

- If $\mathcal{S}_1, \mathcal{S}_2$ known, estimate ϕ's by additionally querying f along corresponding one/two dim. subspaces.

Identify $\mathcal{S}_1, \mathcal{S}_2$: Noiseless setting

First identify $\mathcal{S}_2$; then identify $\mathcal{S}_1$ on reduced SPAM (Tyagi et al. '14). **Identifying $\mathcal{S}_2$:**

- Observation – For any $(l,l') \in \binom{[d]}{2}$:

$$\partial_l \partial_{l'} f = \begin{cases} \partial_l \partial_{l'} \phi_{(l,l')} & \text{if } (l,l') \in \mathcal{S}_2, \\ 0 & \text{otherwise.} \end{cases}$$

- $\nabla^2 f(\mathbf{x})$ *sparse* – k non-zero rows; at most $(\rho_m + 1)$ non-zero entries per row.

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

- Choose $\mathbf{v}'$ randomly; compute $O(\rho_m \log d)$ gradient differences to obtain $\{\nabla^2 f(\mathbf{x})\mathbf{v}'_i + \mathbf{z}_i\}_{i=1}^{m_{v'}}$.

Estimate k-sparse gradients from $O(k \log d)$ queries via ℓ_1 min.:

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

Query f at $\{f(\mathbf{x} \pm \mu \mathbf{v}_i)\}_{i=1}^{m_v}$.

- Estimate each row of $\nabla^2 f(\mathbf{x})$ via ℓ_1 minimization; this gives estimates $\{\widehat{\partial_i \partial_j} f(\mathbf{x}) : (i,j) \in \binom{[d]}{2}\}$ with $O(k \rho_m (log d)^2)$ queries.
- **How to choose $\mathbf{x}$?** Create (d,t) hash family: $\mathcal{H}_2^d = \{h_1, h_2, \dots\}$ with $h_j : [d] \to \{1,2\}$. Construct $\chi = \cup_{h \in \mathcal{H}_2^d} \chi(h)$ where

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

$|\chi| \le (2m_x + 1)^2 |\mathcal{H}_2^d| = O(m_x^2 \log d)$; uniformly discretizes all canonical 2-dim subspaces.

- Estimate $\nabla^2 f(\mathbf{x})$ at each $\mathbf{x} \in \chi$. Identify $\mathcal{S}_2$ via thresholding.

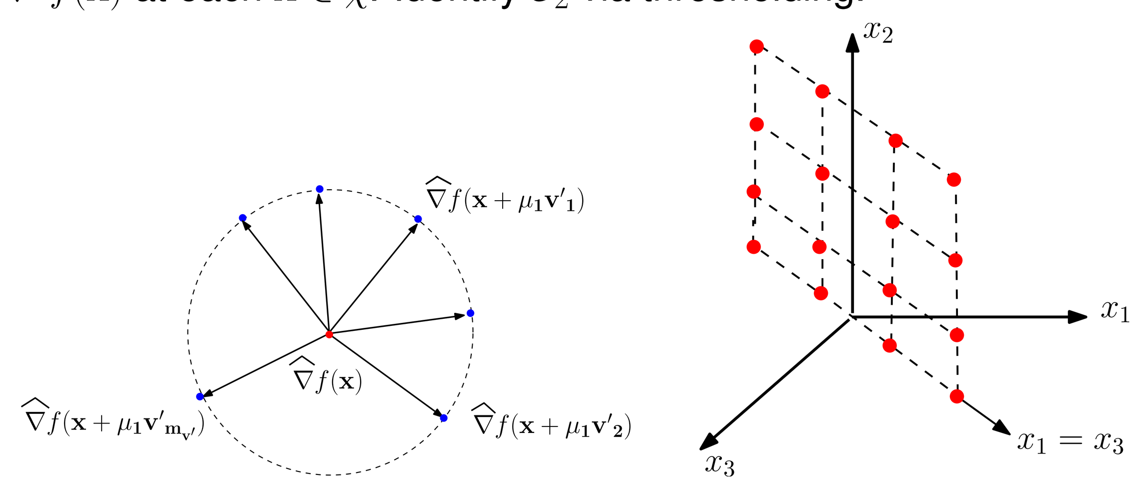


Figure 1: (a) $\nabla^2 f(\mathbf{x})$ estimated using: $\widehat{\nabla} f(\mathbf{x})$ and neighborhood gradient estimates (b) Geometric picture: $d = 3$, $h \in \mathcal{H}_2^3$ with $h(1) = h(3) \neq h(2)$. Red disks are points in $\chi(h)$.

Identifying $\mathcal{S}_1$: Apply scheme of Tyagi et al. '14 on $[d] \setminus \widehat{\mathcal{S}_2^{\text{var}}}$. Recovers $\mathcal{S}_1$ with $O((k - |\mathcal{S}_2^{\text{var}}|) \log d)$ queries.

Algorithm for identifying $\mathcal{S}_2, \mathcal{S}_1$:

- Construct $\chi \subset [-1,1]^d$ using $\mathcal{H}_2^d$.

  At each $\mathbf{x} \in \chi$:
- Estimate $\nabla^2 f(\mathbf{x})$ to obtain $\widehat{\partial_i \partial_j} f(\mathbf{x})$ for all $(i,j) \in \binom{[d]}{2}$.
- For threshold parameter $\tau' > 0$ update $\widehat{\mathcal{S}_2} = \widehat{\mathcal{S}_2} \cup \left\{ (i,j) \in \binom{[d]}{2} : |\widehat{\partial_i \partial_j} f(\mathbf{x})| > \tau' \right\}$

Apply scheme of Tyagi et al. '14 on $[d] \setminus \widehat{\mathcal{S}_2^{\text{var}}}$ to obtain $\widehat{\mathcal{S}_1}$.

Theorem 1. *For suitable choice of step sizes and thresholds, we have $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ w.h.p. Total number of queries made is $O(k \rho_m (\log d)^3)$.*

Identify $\mathcal{S}_1, \mathcal{S}_2$: Noisy setting

- Two noise models: Arbitrary bounded noise and i.i.d Gaussian noise.
- Arbitrary bounded noise: Observe $f(\mathbf{x}) + z'$ with $|z'| < \varepsilon$, and ε known.

  Theorem 2. *If $\varepsilon = O(\rho_m^{-2} k^{-1/2})$, then for suitable choice of step sizes and and thresholds, we have $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ w.h.p.*
- i.i.d Gaussian noise: Observe $f(\mathbf{x}) + z'$ with $z' \sim \mathcal{N}(0, \sigma^2)$.

  Theorem 3. *If we resample each query $O(\rho_m^4 k \log d)$ times and average, then for suitable choice of step sizes and and thresholds, we have $\widehat{\mathcal{S}_2} = \mathcal{S}_2$, $\widehat{\mathcal{S}_1} = \mathcal{S}_1$ w.h.p.*
  - Total number of queries made: $O(\rho_m^5 k^2 (\log d)^4)$.

Simulation results

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

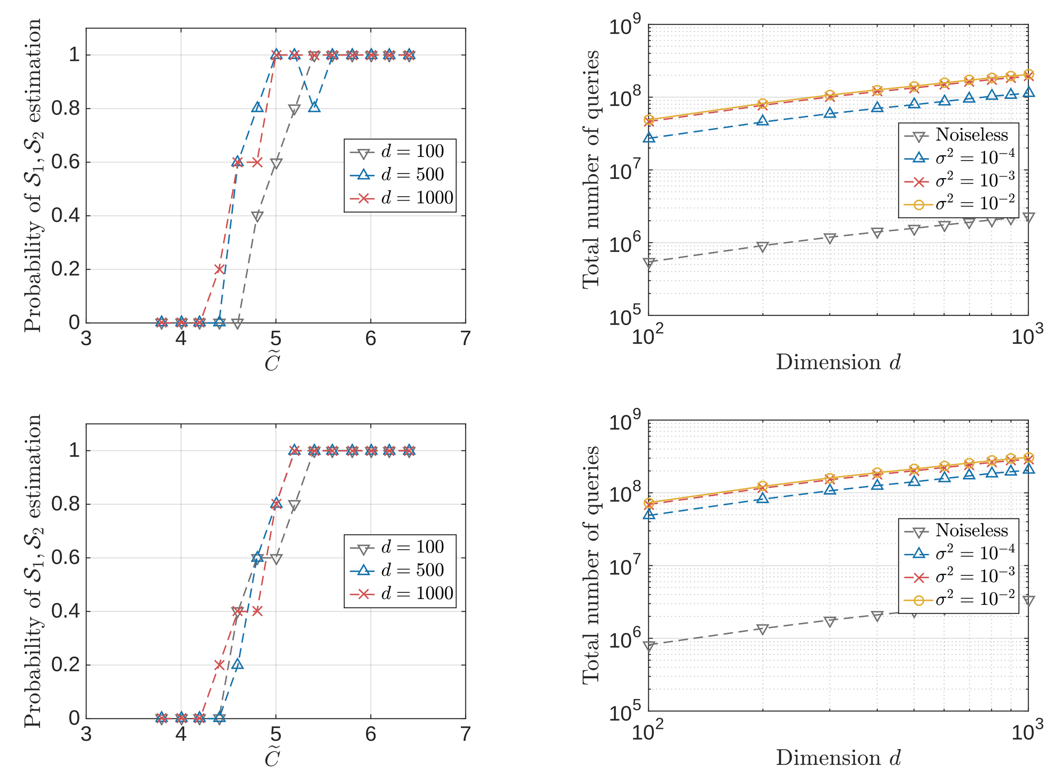


Figure 2: First (resp. second) row is for f_1 (resp. f_2). 5 independent Monte Carlo trials.

- $\widetilde{C}$ is a constant such that $m_v := \widetilde{C} k \log(d/k)$, $m_{v'} := \widetilde{C} \rho_m \log(d/\rho_m)$.

Acknowledgements. Supported in part by SNSF grant CRSII2_147633.