Active Learning of Multi-Index Function Models

Hemant Tyagi and Volkan Cevher
LIONS – EPFL

Abstract

We consider the problem of actively learning *multi-index* functions of the form $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x}) = \sum_{i=1}^{k} g_i(\mathbf{a}_i^T\mathbf{x})$ from point evaluations of f. We assume that the function f is defined on an ℓ_2-ball in $\mathbb{R}^d$, g is twice continuously differentiable almost everywhere, and $\mathbf{A} \in \mathbb{R}^{k\times d}$ is a rank k matrix, where $k \ll d$. We propose a randomized, active sampling scheme for estimating such functions with uniform approximation guarantees. Our theoretical developments leverage recent techniques from low rank matrix recovery, which enables us to derive an estimator of the function f along with sample complexity bounds. We also characterize the noise robustness of the scheme, and provide empirical evidence that the high-dimensional scaling of our sample complexity bounds are quite accurate.

1 Introduction

Learning functions f: $\mathbf{x} \to y$ based on training data $(y_i, \mathbf{x}_i)_{i=1}^m$: $\mathbb{R} \times \mathbb{R}^d$ is a fundamental problem with many scientific and engineering applications. Often, the function f has a parametric model, as in linear regression when $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$, and hence, learning the function amounts to learning the model parameters. In this setting, obtaining an approximate model $\widehat{f}$ when $d \gg 1$ is challenging due to the curse-of-dimensionality. Fortunately, low-dimensional parameter models, such as sparsity and low-rank models, enable successful learning from dimensionality reduced or incomplete data [1, 2].

Since any parametric form is at best an approximation, non-parametric models remain as important alternatives where we also attempt to learn the structure of the mapping f from data [3–15]. Unfortunately, the curse-of-dimensionality problem in non-parametric function learning in high-dimensions is particularly difficult even with smoothness assumptions on f [16–18]. For instance, learning functions $f \in \mathcal{C}^s$ (i.e., the derivatives $f', \ldots, f^{(s)}$ exist and are continuous), defined over compact supports, require $m = \Omega((1/\delta)^{d/s})$ samples for a uniform approximation guarantee of $\delta \ll 1$ (i.e., $\|f - \widehat{f}\|_{L_\infty} \leq \delta$) [17]. Surprisingly, even infinitely differentiable functions ($s = \infty$) are not immune to this problem ($m = \Omega(2^{\lfloor d/2 \rfloor})$) [18]. Therefore, further assumptions on the multivariate functions beyond smoothness are needed for the tractability of successful learning [13, 14, 16, 19].

To this end, we seek to learn low-dimensional function models $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$ that decompose as

$$\text{Model 1: } f(\mathbf{x}) = \sum_{i=1}^{k} g_i(\mathbf{a}_i^T\mathbf{x}) \mid \text{Model 2: } f(\mathbf{x}) = \mathbf{a}_1^T\mathbf{x} + \sum_{i=2}^{k} g_i(\mathbf{a}_i^T\mathbf{x}), \tag{1}$$

thereby constraining f to effectively live on k-dimensional subspaces, where $k \ll d$. The models in (1) have several important machine learning applications, and are known as the multi-index models in statistics and econometrics, and multi-ridge functions in signal processing [4–7, 20–24].

In stark contrast to the classical regression setting where $(y_i, \mathbf{x}_i)_{i=1}^m$ is given *a priori*, we posit the active learning setting where we can query the function to obtain first an explicit approximation of $\mathbf{A}$ and subsequently of f. As a stylized example of the active learning setting, consider numerical solutions of parametric partial differential equations (PDE). Given PDE$(f, \mathbf{x}) = 0$, where $f(\mathbf{x})$:

$\Omega \to \mathbb{R}$ is the implicit solution, obtaining a function sample typically requires running a computationally expensive numerical solver. As we have the ability to choose the samples, we can minimize the number of queries to the PDE solver in order to learn an explicit approximation of f [13].

Background: To set the context for our contributions, it is necessary to review the (rather extensive) literature that revolve around the models (1). We categorize the earlier works by how the samples are obtained (regression (passive) vs. active learning), what the underlying low-dimensional model is (low-rank vs. sparse), and how the smoothness is encoded (kernels vs. $\mathcal{C}^s$).

Regression/low-rank [3–7]: We consider the function model $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$ to be kernel smooth or $\mathcal{C}^s$. Noting the differentiability of f, we observe that the gradients $\nabla f(\mathbf{x}) = \mathbf{A}^T \nabla g(\mathbf{A}\mathbf{x})$ live within the low-dimensional subspaces of $\mathbf{A}^T$. *Assuming that* ∇g has sufficient richness to span k-dimensional subspaces of the rows of $\mathbf{A}$, we use the given samples to obtain Hessian estimates via local smoothing techniques, such as kernel estimates, nearest-neighbor, or spline methods. We then use the k-principal vectors of the estimated Hessian to approximate $\mathbf{A}$. In some cases, we can even establish asymptotic distribution of $\mathbf{A}$ estimates but not finite sample complexity bounds.

Regression/sparse [8–12]: We add sparsity restrictions on the function models: for instance, we assume only one coordinate is active per additive term in (1). To encode smoothness, we restrict f to a particular functional space, such as the reproducing kernel Hilbert or Sobolev spaces. We employ greedy algorithms, back-fitting approaches, or convex regularizers to not only estimate the active coordinates but also the function itself. We can then establish finite sample complexity rates with guarantees of the form $\|f - \widehat{f}\|_{L_2} \leq \delta$, which grow logarithmically with d as well as match the minimax bounds for the learning problem. Moreover, the function estimation incurs a linear cost in k since the problem formulation affords a rotation-free structure between $\mathbf{x}$ and g_i's.

Active learning [13–15]: The majority of the (rather limited) literature on active non-parametric function learning makes sparsity assumptions on $\mathbf{A}$ to obtain guarantees of the form $\|f - \widehat{f}\|_{L_\infty} \leq \delta$, where $f \in \mathcal{C}^s$ with $s > 1$.[1] For instance, we consider the form $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$, where the rows of A live in a weak ℓ_q ball with $q < 2$ (i.e., they are approximately sparse).[2] We then leverage a prescribed random sampling, and prove that the sample complexity grows logarithmically with d and is inversely proportional to the k-th singular value of a "Hessian" matrix H^f (for a precise definition of H^f, see (7)). Thus far, the only known characterization for the k-th singular value of H^f is for radial basis functions, i.e., $f(x) = g(\|\mathbf{A}\mathbf{x}\|_2)$. Just recently, we also see a low-rank model to handle $f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})$ for a general $\mathbf{a}$ ($k = 1$) with a sample complexity proportional to d [15].

Our contributions: In this paper, which is a summary of [26], we take the active learning perspective via low-rank methods, where we have a general $\mathbf{A}$ with only $\mathcal{C}^s$ assumptions on g_i's.

Our main contributions are as follows:

1. *k-th singular value of H^f [14, 15]*: Based on the random sampling schemes of [14, 15], we rigorously establish the first high-dimensional scaling characterization of the k-th singular value of H^f, which governs the sample complexity in both sparse and general $\mathbf{A}$ for the multi-index models in (1). To achieve this result, we introduce an easy-to-verify, new analysis tool based on Lipschitz continuous second order partial derivatives.
2. *Generalization of [13–15]*: We derive the first sample complexity bound for the $\mathcal{C}^s$ functions in (1) with arbitrary number of linear parameters k without the compressibility assumptions on the rows of $\mathbf{A}$. Along the way, we leverage the conventional low-rank models in regression approaches and bridge them with the recent low-rank recovery algorithms. Our result also lifts the sparse additive models in regression [8–12] to a basis-free setting.
3. *Impact of additive noise*: We analytically show how additive white Gaussian noise in the function queries impacts the sample complexity of our low-rank approach.

[1] Not to be confused with the online active learning approaches, which "optimize" a function, such as finding its maximum [25]. In contrast, we would like to obtain uniform approximation guarantees on f, which might lead to redundant samples if we truly are only interested in finding a critical point of the function.

[2] As having one *known* basis to sparsify all k-dimensions in order to obtain a sparse $\mathbf{A}$ is rather restrictive, this model *does not* provide a basis-free generalization of the sparse additive models in regression [8–12].

2 A recipe for active learning of low-dimensional non-parametric models

This section provides the preliminaries for our low-rank active learning approach for multi-index models in (1). We first introduce our sampling scheme (based on [14, 15]), summarize our main observation model (based on [6, 7, 14, 15]), and explain our algorithmic approach (based on [15]). This discussion sets the stage for our main theoretical contributions, as described in Section 4.

Our sampling scheme: Our sampling approach relies on a specific interaction of two sets: sampling centers and an associated set of directions for each center. We denote the set of sampling centers as $\mathcal{X} = \{\xi_j \in \mathbb{S}^{d-1}; j = 1, \dots, m_{\mathcal{X}}\}$. We form $\mathcal{X}$ by sampling points uniformly at random in $\mathbb{S}^{d-1}$ (the unit sphere in d-dimensions) according to the uniform measure $\mu_{\mathbb{S}^{d-1}}$. Along with each $\xi_j \in \mathcal{X}$, we define a directions vector $\Phi_j = [\phi_{1,j}| \dots |\phi_{m_\Phi,j}]^T$, and construct the sampling directions operator $\mathbf{\Phi}$ for $j = 1, \dots, m_{\mathcal{X}}, i = 1, \dots, m_\Phi,$ and $l = 1, \dots, d$ as

$$\mathbf{\Phi} = \left\{ \phi_{i,j} \in B_{\mathbb{R}^d}\left(\sqrt{d/m_\Phi}\right) : [\phi_{i,j}]_l = \pm \frac{1}{\sqrt{m_\Phi}} \text{with probability } 1/2 \right\}, \tag{2}$$

where $B_{\mathbb{R}^d}\left(\sqrt{d/m_\Phi}\right)$ is the ℓ_2-ball with radius $r = \sqrt{d/m_\Phi}$.

Our low-rank observation model: We first write the Taylor series approximation of f as follows

$$f(\mathbf{x} + \epsilon\phi) = f(\mathbf{x}) + \epsilon \langle \phi, \nabla f(\mathbf{x}) \rangle + \epsilon E(\mathbf{x}, \epsilon, \phi); \quad E(\mathbf{x}, \epsilon, \phi) = \frac{\epsilon}{2} \phi^T \nabla^2 f(\zeta(\mathbf{x}, \phi))\phi, \tag{3}$$

where $\epsilon \ll 1$, $\epsilon E(\mathbf{x}, \epsilon, \phi)$ is the curvature error, and $\zeta(\mathbf{x}, \phi) \in [\mathbf{x}, \mathbf{x} + \epsilon\phi] \in B_{\mathbb{R}^d}(1 + \epsilon r)$. Substituting $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$ into (3), we obtain a perturbed observation model ($\nabla g(\cdot)$ is a $k \times 1$ vector):

$$\left\langle \phi, \mathbf{A}^T \nabla g(\mathbf{A}\mathbf{x}) \right\rangle = \frac{1}{\epsilon} \left(f(\mathbf{x} + \epsilon\phi) - f(\mathbf{x}) \right) - E(\mathbf{x}, \epsilon, \phi). \tag{4}$$

We then introduce a matrix $\mathbf{X} := \mathbf{A}^T\mathbf{G}$ with $\mathbf{G} := [\nabla g(\mathbf{A}\xi_1)|\nabla g(\mathbf{A}\xi_2)| \cdots |\nabla g(\mathbf{A}\xi_{m_{\mathcal{X}}})]_{k \times m_{\mathcal{X}}}$. Based on (4), we then derive the following linear system via the operator $\Phi : \mathbb{R}^{d \times m_{\mathcal{X}}} \to \mathbb{R}^{m_\Phi}$

$$\mathbf{y} = \Phi(\mathbf{X}) + \varepsilon; \quad y_i = \epsilon^{-1} \sum_{j=1}^{m_{\mathcal{X}}} \left[f(\xi_j + \epsilon\phi_{i,j}) - f(\xi_j) \right], \tag{5}$$

where $\mathbf{y} \in \mathbb{R}^{m_\Phi}$ are the perturbed measurements of $\mathbf{X}$ with $[\Phi(\mathbf{X})]_j = \text{trace}\left(\Phi_j^T \mathbf{X}\right)$, and $\varepsilon = E(\mathcal{X}, \epsilon, \mathbf{\Phi})$ is the curvature perturbations. The formulation (5) motivates us to leverage affine rank-minimization algorithms [27–29] for low-rank matrix recovery since $\text{rank}(\mathbf{X}) \leq k \ll d$.

Our active low-rank learning algorithm Algorithm 1 outlines the main steps involved in our approximation scheme. Step 1 constructs the operator $\mathbf{\Phi}$ and the measurements $\mathbf{y}$, given m_Φ, $m_{\mathcal{X}}$, and ϵ. Step 2 revolves around the affine-rank minimization algorithms. Step 3 maps the recovered low-rank matrix to $\widehat{\mathbf{A}}$ using the singular value decomposition (SVD) and rank-k approximation. Given $\widehat{\mathbf{A}}$, step 4 constructs $\widehat{f}(\mathbf{x}) := \widehat{g}(\widehat{\mathbf{A}}\mathbf{x})$ as our estimator, where $\widehat{g}(\mathbf{y}) := f(\widehat{\mathbf{A}}^T\mathbf{y})$.

Algorithm 1: Active learner algorithm for the non-parametric model $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$

1: Choose m_Φ, $m_{\mathcal{X}}$, and ϵ and construct the sets $\mathcal{X}$ and $\mathbf{\Phi}$, and the measurements $\mathbf{y}$.
2: Obtain $\widehat{\mathbf{X}}$ via a stable low-rank recovery algorithm (see Section 3 for an example).
3: Compute SVD$(\widehat{\mathbf{X}}) = \widehat{\mathbf{U}}\widehat{\Sigma}\widehat{\mathbf{V}}^T$ and set $\widehat{\mathbf{A}}^T = \widehat{\mathbf{U}}^{(k)}$, corresponding to k largest singular values.
4: Obtain an approximation $\widehat{f}(\mathbf{x}) := \widehat{g}(\widehat{\mathbf{A}}\mathbf{x})$ via quasi interpolants where $\widehat{g}(\mathbf{y}) := f(\widehat{\mathbf{A}}^T\mathbf{y})$.

Remark 1. *We uniformly approximate the function $\widehat{g}$ by first sampling it on a rectangular grid: $h\mathbb{Z}^k \cap (-(1+\bar{\epsilon}), (1+\bar{\epsilon}))^k$ with uniformly spaced points in each direction (step size h). We then use quasi-interpolants to interpolate in between the points thereby obtaining the approximation $\hat{g}_h$, where the complexity is exponential in k (see the tractability discussion in the introduction). We refer the reader to Chapter 12 of [17] regarding the construction of these operators.*

3 Stable low-rank recovery algorithms within our learning scheme

By stable low-rank recovery in Algorithm 1, we mean any algorithm that returns an $\widehat{\mathbf{X}}$ with the following guarantee: $\|\mathbf{X}-\widehat{\mathbf{X}}\|_F \leq c_1\|\mathbf{X}-\mathbf{X}_k\|_F + c_2\|\varepsilon\|_2$, where $c_{1,2}$ are constants, and $\mathbf{X}_k$ is the best rank-k approximation of $\mathbf{X}$. Since there exists a vast set of algorithms with such guarantees, we use the matrix Dantzig selector [29] as a running example. This discussion is intended to expose the reader to the key elements necessary to re-derive the sample complexity of our scheme in Section 4 for different algorithms, which might offer additional computational trade-offs.

Stable embedding: We first explain an elementary result stating that our sampling mechanism satisfies the restricted isometry property (RIP) for all rank-k matrices with overwhelming probability. That is, $(1-\kappa_k)\|\mathbf{X}_k\|_F^2 \leq \|\Phi(\mathbf{X}_k)\|_{\ell_2}^2 \leq (1+\kappa_k)\|\mathbf{X}_k\|_F^2$, where κ_k is the RIP constant [29]). This property can be used in establishing stability of virtually all low-rank recovery algorithms.

As Φ in (5) is a Bernoulli random measurement ensemble, it follows from standard concentration inequalities [30,31] that for any rank-k $\mathbf{X} \in \mathbb{R}^{d\times m_{\mathcal{X}}}$, we have $\mathbb{P}(|\,\|\Phi(\mathbf{X})\|_{\ell_2}^2 - \|\mathbf{X}\|_F^2\,| > t\,\|\mathbf{X}\|_F^2) \leq 2e^{-\frac{m_\Phi}{2}(t^2/2-t^3/3)}, \quad t\in(0,1)$. By using a standard covering argument, as shown in Theorem 2.3 of [29], we can verify that our Φ satisfies RIP with isometry constant $0<\kappa_k<\kappa<1$ with probability at least $1-2e^{-m_\Phi q(\kappa)+k(d+m_{\mathcal{X}}+1)u(\kappa)}$, where $q(\kappa)=\frac{1}{144}\left(\kappa^2-\frac{\kappa^3}{9}\right)$ and $u(\kappa)=\log\left(\frac{36\sqrt{2}}{\kappa}\right)$.

Recovery algorithm and its tuning parameters: The Dantzig selector criteria is given by

$$\widehat{\mathbf{X}}_{DS} = \arg\min_{M} \|M\|_* \text{ s.t. } \|\Phi^*(y-\Phi(M))\| \leq \lambda, \tag{6}$$

where $\|\cdot\|_*$ and $\|\cdot\|$ are the nuclear and spectral norms, respectively, and λ is a tuning parameter. We require the true $\mathbf{X}$ to be feasible, i.e., $\|\Phi^*(\varepsilon)\| \leq \lambda$. Hence, the parameter λ can be obtained via

Proposition 1. *In* (5)*, we have* $\|\varepsilon\|_{\ell_2^{m_\Phi}} \leq \frac{C_2\epsilon d m_{\mathcal{X}} k^2}{2\sqrt{m_\Phi}}$*. Moreover, it holds that* $\|\Phi^*(\varepsilon)\| \leq \lambda = \frac{C_2\epsilon d m_{\mathcal{X}} k^2}{2\sqrt{m_\Phi}}(1+\kappa)^{1/2}$*, with probability at least* $1-2e^{-m_\Phi q(\kappa)+(d+m_{\mathcal{X}}+1)u(\kappa)}$.

Proposition 1 is a new result that provides the typical low-rank recovery algorithm tuning parameters for the random sampling scheme in Section 2. We prove Proposition 1 in [26]. Note that the dimension d appears in the bound as we do not make any compressibility assumption on $\mathbf{A}$. If the rows of $\mathbf{A}$ are compressible, that is $(\sum_{j=1}^d |a_{ij}|^q)^{1/q} \leq D_1 \;\forall\; i=1,\ldots,k$ for some $0<q<1,\ D_1>0$, we can then remove the explicit d-dependence in the bound here.

Stability of low-rank recovery: We first restate a stability result from [29] for bounded noise in Theorem 1. We then exploit this result in Corollary 1 along with Proposition 1 in order to obtain the error bound for the rank-k approximation $\widehat{\mathbf{X}}_{DS}^{(k)}$ to $\mathbf{X}$ in step 4 of our Algorithm 1:

Theorem 1 (Theorem 2.4 in [29])**.** *Let rank(*$\mathbf{X}$*)* $\leq k$ *and let* $\widehat{\mathbf{X}}_{DS}$ *be the solution to* (6)*. If* $\kappa_{4k} < \kappa < \sqrt{2}-1$ *and* $\|\Phi^*(\varepsilon)\| \leq \lambda$*, then we have with probability at least* $1-2e^{-m_\Phi q(\kappa)+4k(d+m_{\mathcal{X}}+1)u(\kappa)}$

$$\left\|\widehat{\mathbf{X}}_{DS}-\mathbf{X}\right\|_F^2 \leq C_0 k\lambda^2,$$

where C_0 *depends only on the isometry constant* κ_{4k}*.*

Corollary 1. *Denoting* $\widehat{\mathbf{X}}_{DS}$ *to be the solution of* (6)*, if* $\widehat{\mathbf{X}}_{DS}^{(k)}$ *is the best rank-k approximation to* $\widehat{\mathbf{X}}_{DS}$ *in the sense of* $\|\cdot\|_F$*, and if* $\kappa_{4k} < \kappa < \sqrt{2}-1$*, then we have*

$$\left\|\mathbf{X}-\widehat{\mathbf{X}}_{DS}^{(k)}\right\|_F^2 \leq 4C_0 k\lambda^2 = \frac{C_0 C_2^2 k^5 \epsilon^2 d^2 m_{\mathcal{X}}^2}{m_\Phi}(1+\kappa),$$

with probability at least $1-2e^{-m_\Phi q(\kappa)+4k(d+m_{\mathcal{X}}+1)u(\kappa)}$.

Corollary 1 is the main result of this section, which is proved in [26]. The approximation guarantee in Corollary 1 can be tightened if other low-rank recovery algorithms are employed in estimation of $\mathbf{X}$. However, we note again that the Dantzig selector enables us to highlight the key steps that lead to the sample complexity of our approach in the next section.

4 Main results

Overview: Below, we study m_Φ, $m_\mathcal{X}$, and ϵ that together achieve and balance three objectives:

- $m_\mathcal{X}$: Sampling centers $\mathcal{X}$ are chosen so that the matrix $\mathbf{G}$ has rank-k. This is critical in ensuring that $\mathbf{G}$ explores the full k-dimensional subspaces as spanned by $\mathbf{A}^T$ lest $\mathbf{X}$ is rank deficient.
- m_Φ: Sampling directions $\mathbf{\Phi}$ (2) are designed to satisfy the RIP for rank-k matrices (cf., Section 3). This property is typically key in proving low-rank recovery guarantees.
- ϵ: The step-size ϵ in (3) manages the impact of the curvature effects E in the linear system (5). Unfortunately, this leads to a collateral damage of amplifying the impact of noise if the queries are corrupted. We provide a remedy below based on sampling the same data points.

Assumptions: We explicitly mention our assumptions here. Without loss of generality, we assume $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_k]^T$ is an arbitrary $k \times d$ matrix with orthogonal rows so that $\mathbf{A}\mathbf{A}^T = \mathbf{I}_k$, and the function f is defined over the unit ball, i.e., f: $B_{\mathbb{R}^d}(1) \to \mathbb{R}$.[3] For simplicity, we carry out our analysis by assuming g to be a $\mathcal{C}^2$ function. By our set up, g also lives over a compact set, hence all its partial derivatives till the order of two are bounded as a result of the Stone-Weierstrass theorem:

$$\sup_{|\beta|\leq 2} \left\|D^\beta g\right\|_\infty \leq C_2; \quad D^\beta g \;=\; \frac{\partial^{|\beta|}}{\partial y_1^{\beta_1} \dots \partial y_k^{\beta_k}}\,; \quad |\beta| = \beta_1 + \dots + \beta_k,$$

for some constant $C_2 > 0$. Finally, the effectiveness of our sampling approach depends on whether or not the following "Hessian" matrix H^f is well-conditioned:

$$H^f := \int_{\mathbb{S}^{d-1}} \nabla f(\mathbf{x}) \nabla f(\mathbf{x})^T d\mu_{\mathbb{S}^{d-1}}(\mathbf{x}). \tag{7}$$

That is, for the singular values of H^f, we assume $\sigma_1(H^f) \geq \dots \geq \sigma_k(H^f) \geq \alpha > 0$ for some α. This assumption ensures $\mathbf{X}$ has full rank-k so that $\mathbf{A}$ can be successfully learned.

Restricted singular values of multi-index models: Our first main technical contribution provides a local condition in Proposition 2 that fully characterizes α for multi-index models in (1) below. We prove Proposition 2 and the ensuing Proposition 3 in [26].

Proposition 2. *Assume that $g \in \mathcal{C}^2 : B_{\mathbb{R}^k} \to \mathbb{R}$ has Lipschitz continuous second order partial derivatives in an open neighborhood of the origin, $\mathcal{U}_\theta = B_{\mathbb{R}^k}(\theta)$ for some fixed $\theta = \mathcal{O}\left(d^{-(s+1)}\right)$, and for some $s > 0$:*

$$\frac{\left|\dfrac{\partial^2 g}{\partial y_i \partial y_j}(\mathbf{y}_1) - \dfrac{\partial^2 g}{\partial y_i \partial y_j}(\mathbf{y}_2)\right|}{\|\mathbf{y}_1 - \mathbf{y}_2\|_{\ell_2^k}} \leq L_{i,j} \quad \forall \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{U}_\theta, \mathbf{y}_1 \neq \mathbf{y}_2,\ i,j = 1, \dots, k.$$

Denote $L = \max_{1\leq i,j\leq k} L_{i,j}$. Also assume, $\dfrac{\partial^2 g(\mathbf{y})}{\partial y_i^2}\Big|_{\mathbf{y}=\mathbf{0}} \neq 0$ $\forall i = 1, \dots, k$ for Model 1 and $\forall i = 2, \dots, k$ for Model 2 in (1)*. Then, we have $\alpha = \Theta(1/d)$ as $d \to \infty$.*

The proof of Proposition 2 also leads to the following proposition for tractability of learning the general set $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$ without the particular modular decomposition as in (1):

Proposition 3. *With the same Lipschitz continuous second order partial derivative assumption as in Proposition 2, if $\nabla^2 g(\mathbf{0})$ is rank-k, then we have $\alpha = \Theta(1/d)$ as $d \to \infty$*

Sampling complexity of active multi-index model learning: The importance of Proposition 2 and Proposition 3 is further made explicit in our second main technical contribution as Theorem 2 below, which characterizes the sample complexity of our low-rank learning recipe in Section 2 for non-parametric models along with the Dantzig selector algorithm. Its proof can be found in [26].

[3] Unless further assumptions are made on f or g_i's, we can only identify the subspace spanned by the rows of $\mathbf{A}$ up to a rotation. Hence, while we discuss approximation results on $\mathbf{A}$, the reader should keep in mind that our final guarantees only apply to the function f and not necessarily for $\mathbf{A}$ and g individually. Moreover, if f lives in some other convex body other than $B_{\mathbb{R}^d}(1)$, say L_∞-ball, our analysis can be extended in a straightforward fashion (cf., the concluding discussion in [14]). We also assume that an enlargement of the unit ball $B_{\mathbb{R}^d}(1)$ on the domain of f for a sufficiently small $\bar{\epsilon} > 0$ is allowed. This is not a restriction, but is a consequence of our scheme as we work with directional derivatives of f at points on the unit sphere $\mathbb{S}^{d-1}$.

Theorem 2. *[Sample complexity of Algorithm 1] Let $\delta \in \mathbb{R}^+$, $\rho \ll 1$, and $\kappa < \sqrt{2}-1$ be fixed constants. Choose $m_{\mathcal{X}} \geq \frac{2kC_2^2}{\alpha\rho^2}\log(k/p_1)$, $m_\Phi \geq \frac{\log(2/p_2)+4k(d+m_{\mathcal{X}}+1)u(\kappa)}{q(\kappa)}$, and $\epsilon \leq \frac{\delta}{C_2 k^{5/2} d(\delta+2C_2\sqrt{2k})}\left(\frac{(1-\rho)m_\Phi\alpha}{(1+\kappa)C_0 m_{\mathcal{X}}}\right)^{1/2}$. Then, given $m = m_{\mathcal{X}}(m_\Phi+1)$ samples, our function estimator $\widehat{f}$ in step 4 of Algorithm 1 obeys $\left\|f-\widehat{f}\right\|_{L_\infty} \leq \delta$ with probability at least $1-p_1-p_2$.*

Theorem 2 characterizes the necessary scaling of the sample complexity for our active learning scheme in order to obtain uniform approximation guarantees on f with overwhelming probability: $m_{\mathcal{X}} = \mathcal{O}\left(\frac{k\log k}{\alpha}\right)$, $m_\Phi = \mathcal{O}(k(d+m_{\mathcal{X}}))$, and $\epsilon = \mathcal{O}\left(\frac{\alpha}{\sqrt{d}}\right)$. Note the important role played by α in the sample complexity. Finally, we also mention that the sample complexity can be written differently to trade-off δ among $m_{\mathcal{X}}$, m_Φ, and ϵ. For instance, we can remove δ dependence in the sampling bound for ϵ: let $\delta < 1$, then we just need to scale $m_{\mathcal{X}}$ by δ^{-2}, and m_Φ by δ^{-4}.

Remark 2. *Note that the sample complexity in [14] for learning compressible $\mathbf{A}$ is $m = \mathcal{O}\left(k^{\frac{4-q}{2-q}} d^{\frac{2}{2-q}}\log(k)\right)$ with uniform approximation guarantees on $f \in \mathcal{C}^2$. However, the authors are able to obtain this result only for a restricted set of radial basis functions. Surprisingly, our sample complexity for multi-index models* (1) *not only generalizes this result for general $\mathbf{A}$ but features a better dimensional dependence for $q \in (1,2)$: $m = \mathcal{O}\left(k^3 d^2 (\log(k))^2\right)$. Of course, we require more computation since we use low-rank recovery as opposed to sparse recovery methods.*

Impact of noisy queries: Here, we focus on how α impacts ϵ in particular. Our motivation is to understand how additive noise in function queries, a realistic assumption in many applications, can impact our learning scheme, which will form the basis of our third main technical contribution.

Let us assume that the evaluation of f at a point $\mathbf{x}$ yields: $f(\mathbf{x})+Z$, where $Z \sim \mathcal{N}(0,\sigma^2)$. Thus under this noise model, (5) changes to $\mathbf{y} = \Phi(\mathbf{X}) + \varepsilon + \mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^{m_\Phi}$ and $z_i = \sum_{j=1}^{m_{\mathcal{X}}} \frac{z_{ij}}{\epsilon}$. Assuming independent and identically distributed (iid) noise, we have $z_{ij} \sim \mathcal{N}(0,2\sigma^2)$, and $z_i \sim \mathcal{N}\left(0, \frac{2m_{\mathcal{X}}\sigma^2}{\epsilon^2}\right)$. Therefore, the noise variance gets amplified by a factor of $\frac{2m_{\mathcal{X}}}{\epsilon^2}$.

In our analysis in Section 3, recall that we require the true matrix $\mathbf{X}$ to be feasible. Then, from Lemma 1.1 in [29] and Proposition 1, it follows that the bound below holds with high probability.

$$\|\Phi^*(\varepsilon+\mathbf{z})\| \leq \frac{2\gamma\sigma}{\epsilon}\sqrt{2(1+\kappa)m_{\mathcal{X}}m_\Phi} + \frac{C_2\epsilon d m_{\mathcal{X}}k^2}{2\sqrt{m_\Phi}}(1+\kappa)^{1/2}, \quad (\gamma > 2\sqrt{\log 12}). \tag{8}$$

Unfortunately, we cannot control the upper bound λ on $\|\Phi^*(\varepsilon+\mathbf{z})\|$ by simply choosing smaller ϵ, due to the appearance of the $(1/\epsilon)$ term. Hence, unless σ is $\mathcal{O}(\epsilon)$ or less, (e.g., σ reduces with d), we can only declare that our learning scheme with the matrix Dantzig selector is sensitive to noise *unless we resample the same data points $\mathcal{O}(\epsilon^{-1})$-times and average*. If the noise variance σ^2 is constant, this would keep the impact of noise below a constant times the impact of the curvature errors, which our scheme can handle. The sample complexity then becomes $m = \mathcal{O}\left(\sqrt{d}/\alpha\right)m_{\mathcal{X}}(m_\Phi+1)$, since we choose $m_{\mathcal{X}}(m_\Phi+1)$ *unique* points, and then re-query and average the same points $\mathcal{O}\left(\sqrt{d}/\alpha\right)$-times. Unfortunately, we cannot qualitatively improve the $\mathcal{O}\left(\sqrt{d}/\alpha\right)$-expansion for noise robustness by simply changing the low-rank recovery algorithm since it depends on the relative ratio of the curvature errors $\|\varepsilon\|_2$ to the norm of the noise vector $\|\mathbf{z}\|$. As $\mathbf{\Phi}$ satisfies the RIP assumption, we can verify that this relative ratio is approximately preserved in (8) for iid Gaussian noise.

5 Numerical Experiments

We present simulation results on toy examples to empirically demonstrate the tightness of the sampling bounds. In the sequel, we assume $\mathbf{A}$ to be row orthonormal and concern ourselves only with the recovery of $\mathbf{A}$ upto an orthonormal transformation. Therefore, we seek a guaranteed lower

bound on $\left\|\mathbf{A}\widehat{\mathbf{A}}^T\right\|_F \geq (k\eta)^{1/2}$ for some $0 < \eta < 1$. Then it is possible to show, along the lines of the proof for Theorem 2 (see [26]), we would need to pick ϵ as follows:

$$\epsilon \leq \frac{1}{C_2 k^2 d(\sqrt{k(1-\eta)}+\sqrt{2})} \left(\frac{(1-\rho)m_\Phi \alpha(1-\eta)}{(1+\kappa)C_0 m_\mathcal{X}}\right)^{1/2}. \tag{9}$$

Logistic function ($k = 1$) We first consider $f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})$ where $g(y) = (1+e^{-y})^{-1}$ is the logistic function. This case allows us to explicitly calculate all the necessary parameters within our paper. For instance, we can easily verify that $C_2 = \sup_{|\beta|\leq 2} \left|g^{(\beta)}(y)\right| = 1$. Furthermore we compute the value of α through the approximation: $\alpha = \int \left|g'(\mathbf{a}^T\mathbf{x})\right|^2 d\mu_{\mathbb{S}^{d-1}} \approx |g'(0)|^2 = (1/16)$, which holds for large d. We require $|\langle \hat{\mathbf{a}}, \mathbf{a}\rangle|$ to be greater then $\eta = 0.99$. We fix values of $\kappa < \sqrt{2}-1$, $\rho \in (0,1)$ and $\epsilon = 10^{-3}$. The value of $m_\mathcal{X}$ (number of points sampled on $\mathbb{S}^{d-1}$) is fixed at 20 and we vary d over the range 200–3000. For each value of d, we increase m_Φ till $|\langle \hat{\mathbf{a}}, \mathbf{a}\rangle|$ reaches the specified performance criteria of η. We remark that for each value of d and m_Φ, we choose ϵ according to the derived equation (9) for the specified performance criteria given by η.

Figure 1 depicts the scaling of m_Φ with the dimension d. The results are obtained by selecting $\mathbf{a}$ uniformly at random on $\mathbb{S}^{d-1}$ and averaging the value of $|\langle \hat{\mathbf{a}}, \mathbf{a}\rangle|$ over 10 independent trials using the Danzig selector. We observe that for large values of d, the minimum number of directional derivatives needed to achieve the performance bound on $|\langle \hat{\mathbf{a}}, \mathbf{a}\rangle|$ scales approximately linearly with d, with a scaling factor of around 1.45.

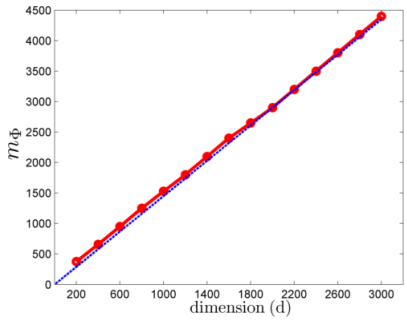
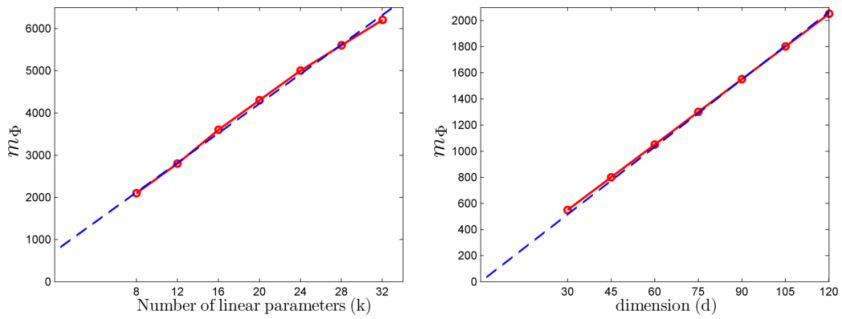


Figure 1: Plot of $\frac{m_\Phi}{d}$ versus d for $m_\mathcal{X} = 20$, with m_Φ chosen to be minimum value needed to achieve $|\langle \hat{\mathbf{a}}, \mathbf{a}\rangle| \geq 0.99$. ϵ is fixed at 10^{-3}. m_Φ scales approximately linearly with d where the constant is 1.45.

Sum of Gaussian functions ($k > 1$) We next consider functions of the form $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x}+\mathbf{b}) = \sum_{i=1}^k g_i(a_i^T\mathbf{x}+b_i)$, where $g_i(y) = (2\pi\sigma_i^2)^{-1/2}\exp\left(-(y+b_i)^2/(2\sigma_i^2)\right)$. We fix $d = 100$, $\epsilon = 10^{-3}$, $m_\mathcal{X} = 100$ and vary k from 8 to 32 in steps of 4. For each value of k we are interested in the minimum value of m_Φ needed to achieve $\frac{1}{k}\left\|\mathbf{A}\widehat{\mathbf{A}}\right\|_F^2 \geq 0.99$. In Figure 2(a), we see that m_Φ scales approximately linearly with the number of Gaussian atoms k. The results are averaged over 10 trials. In each trial, we select the rows of $\mathbf{A}$ over the left Haar measure on $\mathbb{S}^{d-1}$, and the parameter $\mathbf{b}$ uniformly at random on $\mathbb{S}^{k-1}$ scaled by a factor 0.2. Furthermore we generate the standard deviations of the individual Gaussian functions uniformly over the range $[0.1, 0.5]$.

Impact of Noise ($k > 1$) We now consider quadratic forms, i.e. $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x}) = \|\mathbf{A}\mathbf{x}-b\|^2$ with the point queries corrupted with Gaussian noise. Here, we take α to be $1/d$. We fix $k = 5$, $m_\mathcal{X} = 30$, $\epsilon = 10^{-1}$ and vary d from 30 to 120 in steps of 15. For each d we perturb the point queries with Gaussian noise of standard deviation: $0.01/d^{3/2}$. This is the same as repeatedly sampling each random location approximately $d^{3/2}$ times followed by averaging. We then compute the minimum value of m_Φ needed to achieve $\frac{1}{k}\left\|\mathbf{A}\widehat{\mathbf{A}}\right\|_F^2 \geq 0.99$. We average the results over 10 trials, and in each trial, we select the rows of $\mathbf{A}$ over the left Haar measure on $\mathbb{S}^{d-1}$. The parameter $\mathbf{b}$ is chosen uniformly at random on $\mathbb{S}^{k-1}$. In Figure 2(b), we see that m_Φ scales approximately linearly with d, which follows our sample complexity bound for m_Φ in Theorem 2.



(a) $k > 1$ (Gaussian) (b) $k > 1$ (quadratic) with noise

Figure 2: The empirical performance of our oracle-based low-rank learning scheme (circles) agrees well with the theoretical scaling (dashed). Section 5 has further details.

6 Conclusions

In this work, we consider the problem of learning non-parametric low-dimensional functions $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$, which can also have a modular decomposition as in (1), for arbitrary $\mathbf{A} \in \mathbb{R}^{k \times d}$ where rank($\mathbf{A}$) = k. The main contributions of the work are three-fold. By introducing a new analysis tool based on Lipschitz property on the second order derivatives, we provide the first rigorous characterization of the dimension dependence of the k-restricted singular value of the "Hessian" matrix H^f for general multi-index models. We establish the first sample complexity bound for learning non-parametric multi-index models with low-rank recovery algorithms and also analyze the impact of additive noise to the sample complexity of the scheme. Lastly, we provide empirical evidence on toy examples to show the tightness of the sampling bounds. Finally, while our active learning scheme ensures the tractability of learning non-parametric multi-index models, it does not establish a lowerbound on the sample complexity, which is left for future work.

7 Acknowledgments

This work was supported in part by the European Commission under Grant MIRG-268398, ERC Future Proof, SNF 200021-132548, ARO MURI W911NF0910383, and DARPA KeCoM program #11-DARPA-1055. VC also would like to acknowledge Rice University for his Faculty Fellowship. The authors thank Jan Vybiral for useful discussions and Anastasios Kyrillidis for his help with the low-rank matrix recovery simulations.

References

[1] P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer-Verlag New York Inc, 2011.

[2] L. Carin, R.G. Baraniuk, V. Cevher, D. Dunson, M.I. Jordan, G. Sapiro, and M.B. Wakin. Learning low-dimensional signal models. *Signal Processing Magazine, IEEE*, 28(2):39–51, 2011.

[3] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566, 2001.

[4] K.C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, pages 316–327, 1991.

[5] P. Hall and K.C. Li. On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, pages 867–889, 1993.

[6] Y. Xia, H. Tong, WK Li, and L.X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.

[7] Y. Xia. A multiple-index model and dimension reduction. *Journal of the American Statistical Association*, 103(484):1631–1640, 2008.

[8] Y. Lin and H.H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006.

[9] L. Meier, S. Van De Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

[10] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Technical Report*, UC Berkeley, Department of Statistics, August 2010.

[11] 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.

[12] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 2010.

[13] 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.

[14] M. Fornasier, K. Schnass, and J. Vybíral. Learning functions of few arbitrary linear parameters in high dimensions. *Preprint*, 2010.

[15] H. Tyagi and V. Cevher. Learning ridge functions with randomized sampling in high dimensions. In *ICASSP*, 2011.

[16] J.F. Traub, G.W Wasilkowski, and H. Wozniakowski. Information-based complexity. Academic Press, New York, 1988.

[17] R. DeVore and G.G. Lorentz. Constructive approximation. vol. 303, Grundlehren, Springer Verlag, N.Y., 1993.

[18] E.Novak and H.Woniakowski. Approximation of infinitely differentiable multivariate functions is intractable. *J. Complex.*, 25:398–404, August 2009.

[19] W. Hardle. *Applied nonparametric regression*, volume 26. Cambridge Univ Press, 1990.

[20] J.H. Friedman and W. Stuetzel. Projection pursuit regression. *J. Amer. Statist. Assoc.*, 76:817–823, 1981.

[21] D.L. Donoho and I.M. Johnstone. Projection based regression and a duality with kernel methods. *Ann. Statist.*, 17:58–106, 1989.

[22] P.J. Huber. Projection pursuit. *Ann. Statist.*, 13:435–475, 1985.

[23] A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[24] E.J Candès. Harmonic analysis of neural networks. *Appl. Comput. Harmon. Anal.*, 6(2):197–218, 1999.

[25] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *To appear in the IEEE Trans. on Information Theory*, 2012.

[26] Hemant Tyagi and Volkan Cevher. Learning non-parametric basis independent models from point queries via low-rank methods. *Technical Report*, Infoscience EPFL, 2012.

[27] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[28] E.J. Candès and T. Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inf. Theor.*, 56:2053–2080, May 2010.

[29] E.J. Candès and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *CoRR*, abs/1001.0339, 2010.

[30] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM REVIEW*, 52:471–501, 2010.

[31] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338.