LEARNING RIDGE FUNCTIONS WITH RANDOMIZED SAMPLING IN HIGH DIMENSIONS

Hemant Tyagi and Volkan Cevher

Ecole Polytechnique Federale de Lausanne
Laboratory for Information and Inference Systems

ABSTRACT

We study the problem of learning *ridge* functions of the form $f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^d$, from random samples. Assuming g to be a twice continuously differentiable function, we leverage techniques from low rank matrix recovery literature to derive a uniform approximation guarantee for estimation of the ridge function f. Our new analysis removes the *de facto* compressibility assumption on the parameter $\mathbf{a}$ for learning in the existing literature. Interestingly the price to pay in high dimensional settings is not major. For example, when g is thrice continuously differentiable in an open neighbourhood of the origin, the sampling complexity changes from $\mathcal{O}(\log d)$ to $\mathcal{O}(d)$ or from $\mathcal{O}(d^{2+\frac{q}{2-q}})$ to $\mathcal{O}(d^4)$, depending on the behaviour of g' and g'' at the origin, with $0 < q < 1$ characterizing the sparsity of $\mathbf{a}$.

Index Terms— Ridge functions, high dimensional function approximation, low rank recovery

1. INTRODUCTION

Several important problems in learning theory, statistics, modeling physical systems, neural networks, and stochastic PDE's involve approximating a function f, defined on a compact domain $\Omega \subset \mathbb{R}^d$, from its point values (cf., [1] and the references therein).

In general, if the only assumption we make on the function f is its smoothness with an order $s > 0$ (i.e., loosely speaking, it has s continuous derivatives), then the best approximation one can achieve is $\mathcal{O}(n^{-s/d})$, where n is the number of points at which the function is queried. In other words, the problem has exponential complexity. Therefore, in order to even attempt learning, we need to consider other restrictions on the functions, especially in high dimensions, that hold in real world settings.

In this paper, we are interested in approximating a particular class of functions known as *ridge functions*. A ridge function is a multivariate function $f : \mathbb{R}^d \to \mathbb{R}$ of the following form

$$f(x_1, \ldots, x_d) = \sum_{i=1}^{m} g_i(\mathbf{a}_i^T\mathbf{x}). \tag{1}$$

The name "ridge function" was first introduced by Logan and Shepp in 1975 [2]. Ridge function approximations are studied in Statistics under the name of "Projection Pursuit" [3, 4, 5]. In short, projection pursuit algorithms approximate a function of d variables by functions of the form (1). The idea here is to reduce dimension by projecting $\mathbf{x}$ along $\mathbf{a}_i$'s to pick out the salient features. Ridge functions also appear in neural networks [6, 7, 8, 9].

This work was supported in part by the European Commission under Grant MIRG-268398, SNF 200021_132548, and DARPA KeCoM program #11-DARPA-1055. VC also would like to acknowledge Rice University for his Faculty Fellowship.

Previous work: Recently, Cohen et al. [1] proposed a recovery method for estimating functions of the form $f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})$ from point queries, where $g : [0, 1] \to \mathbb{R}$ is a C^s function for $s > 1$. However they made a rather restrictive assumption that $\mathbf{a} \succeq 0, \mathbf{1}^T\mathbf{a} = 1$. In order to establish estimation guarantees on f, the authors leverage a compressive sensing twist: the parameter $\mathbf{a}$ must be compressible. That is, $\mathbf{a}$ can be well-approximated by a sparse set of its coefficients.

In [10], the authors extend this work of Cohen et al. to also capture functions of the form $f(\mathbf{x}) = g(A\mathbf{x})$ with A being an arbitrary $k \times d$ matrix of rank k, with each row of A being compressible and g being a C^2 function. They exploit the second tenet of compressive sensing: randomized sampling. As a result, for the class of $\mathcal{C}^2$ smooth ridge functions which are $\mathcal{C}^3$ differentiable in an open neighbourhood of the origin, their sampling complexity comes out to be $\mathcal{O}(\log d)$ or $\mathcal{O}(d^{2+\frac{q}{2-q}})$ (depending on the behaviour of g' and g'' at the origin) with $0 < q < 1$ characterizing the sparsity of the linear parameter $\mathbf{a}$.

Our contribution: In this paper, we also consider learning functions of the form $f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})$ with randomized sampling with g being a C^2 function and $\|a\|_{l_2^d} = 1$, similar to Fornasier et al. [10]. However, compared to [10], we remove the assumption that $\mathbf{a}$ is compressible, in order to generalize the results to arbitrary $\mathbf{a}$. Although we only consider the simplest form of a ridge function with a single parameter $\mathbf{a}$, our setting can be extended in a straightforward manner to functions of the form (1) with $m > 1$. Our main result is a concatenation of a few existing ideas: we first leverage the matrix Dantzig selector from [11] to recover an approximation $\hat{\mathbf{a}}$ to $\mathbf{a}$. We then use $\hat{\mathbf{a}}$ to obtain a uniform approximation to f.

Organization: Section 2 delineates the mathematical set up along with the notations and assumptions used throughout in the paper. Section 3 describes our analysis, which explains our ridge function estimation ideas in three steps. Section 4 provides a concluding discussion along with comparisons of the sampling complexities.

2. PROBLEM SETUP AND ASSUMPTIONS

We borrow our notation from [10]. We consider estimating functions $f : B_{\mathbb{R}^d}(1 + \bar{\epsilon}) \to \mathbb{R}$ of the form

$$f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x}); \quad \|\mathbf{a}\|_2 = 1,$$

where $B_{\mathbb{R}^d}$ denotes the unit ball and $B_{\mathbb{R}^d}(r)$ the ball of radius $r > 0$ in $\mathbb{R}^d$. We assume g is a C^2 function so that for some $C_2 > 0$, we have

$$\sup_{|\beta| \leq 2} \left\| D^\beta g \right\|_\infty \leq C_2,$$

where D is the derivative operator. Denoting $\mu_{\mathbb{S}^{d-1}}$ to be the uniform measure on the d-dimensional unit sphere $\mathbb{S}^{d-1}$, we assume

that the matrix

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

is well conditioned. That is,

$$\int_{\mathbb{S}^{d-1}} \left| g'(\mathbf{a}^T \mathbf{x}) \right|^2 d\mu_{\mathbb{S}^{d-1}}(x) = \alpha > 0.$$

Note, however, that we depart from [10] by making no compressibility assumption on $\mathbf{a}$. Therefore, the parameters in the model can be summarized as follows: the dimension d of $\mathbf{x}$, the smoothness constant C_2, and the matrix conditioning parameter $0 < \alpha < C_2^2$.

Since g is a C^2 function, we have the following identity by Taylors expansion of g at ξ:

$$[g'(\mathbf{a}^T \xi)\mathbf{a}] \cdot \phi = \frac{f(\xi + \epsilon\phi) - f(\xi)}{\epsilon} - \frac{\epsilon}{2}[\phi^T \nabla^2 f(\zeta)\phi], \quad (2)$$

for $\xi \in B_{\mathbb{R}^d}, \phi \in B_{\mathbb{R}^d}(r)$, $\epsilon, r \in \mathbb{R}_+$ with $r\epsilon \leq \bar{\epsilon}$, and for a suitable $\zeta(\xi, \phi) \in B_{\mathbb{R}^d}(1 + \bar{\epsilon})$. We consider two sets of points $\mathcal{X}$ and $\mathbf{\Phi}$ defined as follows.

$$\mathcal{X} = \left\{ \xi_j \in \mathbb{S}^{d-1} : j = 1, \ldots, m_{\mathcal{X}} \right\},$$

$$\mathbf{\Phi} = \{ \phi_{i,j} \in B_{\mathbb{R}^d}\left(\sqrt{d/m_\Phi}\right) : [\phi_{i,j}]_k = \pm \frac{1}{\sqrt{m_\Phi}} \text{w.p. } 1/2,$$
$$i = 1, \ldots, m_\Phi, j = 1, \ldots, m_{\mathcal{X}} \quad \text{and} \quad k = 1, \ldots, d \}.$$

Hence by using (2), we can obtain the following factorization:

$$\Phi(X) = \mathbf{y} + \varepsilon, \quad (3)$$

where $X = \mathbf{a}\mathcal{G}^T = [g'(\mathbf{a}^T\xi_1)\mathbf{a} \ldots g'(\mathbf{a}^T\xi_{m_{\mathcal{X}}})\mathbf{a}]$ is a $d \times m_{\mathcal{X}}$ matrix of rank 1, and $\mathbf{y}, \varepsilon \in \mathbb{R}^{m_\Phi}$ are defined as follows ($i = 1, \ldots, m_\Phi$):

$$y_i = \sum_{j=1}^{m_{\mathcal{X}}} \left[\frac{f(\xi_j + \epsilon\phi_{i,j}) - f(\xi_j)}{\epsilon} \right],$$

$$\varepsilon_i = \sum_{j=1}^{m_{\mathcal{X}}} \left[\frac{-\epsilon}{2} \phi_{i,j}^T \nabla^2 f(\zeta(\xi_j, \phi_{i,j}))\phi_{i,j} \right].$$

Similar to [10], we choose $\Phi : \mathbb{R}^{d \times m_{\mathcal{X}}} \to \mathbb{R}^{m_\Phi}$ to be a random linear measurement operator. The i-th entry of $\Phi(X)$ is denoted as $[\Phi(X)]_i = \langle \Phi_i, X \rangle$ where $\Phi_i = [\phi_{i,1} \ldots \phi_{i,m_{\mathcal{X}}}]$ is of dimensions $d \times m_{\mathcal{X}}$ and $\langle \Phi_i, X \rangle = \text{trace}(\Phi_i^T X)$ is the standard inner product.

We present Proposition 1 below, which upperbounds the $l_2^{m_\Phi}$-norm of the noise ε:

Proposition 1. *Using the factorization equality* (3) *we obtain* $\|\varepsilon\|_{l_2^{m_\Phi}} \leq \frac{C_2 K m_{\mathcal{X}}}{2\sqrt{m_\Phi}}$ *where* $K = \epsilon d$.

Proof. We can express the noise norm as follows:

$$\|\varepsilon\|_{l_2^{m_\Phi}} = \frac{\epsilon}{2} \left(\sum_{i=1}^{m_\Phi} \left| \sum_{j=1}^{m_{\mathcal{X}}} \left[\phi_{i,j}^T \nabla^2 f(\zeta_{ij}) \phi_{i,j} \right] \right|^2 \right)^{\frac{1}{2}}.$$

Now, $\phi_{i,j}^T \nabla^2 f(\zeta_{ij}) \phi_{i,j} \leq \frac{|g''(\mathbf{a} \cdot \zeta_{ij})|}{m_\Phi} \left(\sum_{l=1}^{d} |a_l| \right)^2 \leq \frac{C_2 d}{m_\Phi}$

$$\Rightarrow \|\varepsilon\|_{l_2^{m_\Phi}} \leq \frac{C_2 K m_{\mathcal{X}}}{2\sqrt{m_\Phi}}. \quad (4)$$

□

Remark 1. *Note that the dimension d appears in the bound (within K) as we do not make any compressibility assumption on $\mathbf{a}$. If $\mathbf{a}$ is compressible, that is $(\sum_{i=1}^{d} |a_i|^q)^{1/q} \leq D_1$ for some $0 < q < 1$ and some non-negative constant D_1, the bound becomes independent of d, which would be replaced by D_1.*

3. THE ANALYSIS

Our goal now is to recover the rank 1 matrix X from a few random linear measurements m_Φ. We proceed in Section 3.1 by first solving a nuclear norm minimization based convex program, namely the *matrix Dantzig selector* to obtain an approximation $\hat{X}_{DS}$ to X with a guaranteed upper bound on approximation error. We then take the best rank 1 approximation $\hat{X}_{DS}^{(1)}$ to $\hat{X}_{DS}$, which doubles the constant in the previous error bound. In Section 3.2 we use $\hat{X}_{DS}^{(1)}$ to recover an approximation $\hat{\mathbf{a}}$ to $\mathbf{a}$ with a guaranteed lower bound on $|\langle \mathbf{a}, \hat{\mathbf{a}} \rangle|$. Finally, in Section 3.3 we use this lower bound to derive a uniform approximation $\hat{f}$ to f.

3.1. Low-rank matrix recovery with Dantzig Selector

To recover X, we solve the nuclear norm minimization problem based on the following convex formulation [11]:

$$\hat{X}_{DS} = \arg\min \|M\|_* \text{ s.t. } \|\Phi^*(y - \Phi(M))\| \leq \lambda, \quad (5)$$

where the optimal solution is the estimate $\hat{X}_{DS}$, $\|\cdot\|$ is the operator norm and $\|\cdot\|_*$ is its dual, i.e. the nuclear norm and Φ^* is the adjoint of Φ. This convex program is referred to as the *matrix Dantzig selector* [11]. As in [11], we require the 'true' matrix X to be feasible, i.e. one should have $\|\Phi^*(\varepsilon)\| \leq \lambda$. In the case of bounded noise, this corresponds to $\lambda = C\frac{m_{\mathcal{X}}}{\sqrt{m_\Phi}}$ for some constant C as is mentioned in Lemma 1. Before proving this we first introduce the matrix version of the restricted isometry property (RIP), for linear mappings as defined in [11].

Definition 1. *For matrices of dimensions $n_1 \times n_2$, $n = \min(n_1, n_2)$, for each integer $r = 1, 2, \ldots, n$, the isometry constant δ_r of Φ is the smallest quantity such that*

$$(1 - \delta_r) \|X\|_F^2 \leq \|\Phi(X)\|_{l_2}^2 \leq (1 + \delta_r) \|X\|_F^2$$

holds for all matrices of rank at most r.

As Φ is a Bernoulli random measurement ensemble it follows from standard concentration inequalities [12, 13] that for any given $X \in \mathbb{R}^{d \times m_{\mathcal{X}}}$ and any fixed $0 < t < 1$,

$$\mathbb{P}(|\, \|\Phi(X)\|_{l_2}^2 - \|X\|_F^2 \,| < t \|X\|_F^2) \leq 2\exp\left(-\frac{m_\Phi}{2}(t^2/2 - t^3/3) \right).$$

By using a standard covering argument as shown in Theorem 2.3 of [11] it is easily verifiable that Φ satisfies RIP with isometry constant $0 < \delta_r < \delta < 1$ with probability at least

$$1 - 2\exp\left(-(m_\Phi q(\delta) - r(d + m_{\mathcal{X}} + 1)u(\delta)) \right),$$

where $q(\delta)$ and $u(\delta)$ are constants depending only on δ.

Lemma 1. *Given ε with a bounded norm, we have with probability at least* $1 - 2\exp\left(-(m_\Phi q(\delta_1) - (d + m_{\mathcal{X}} + 1)u(\delta_1)) \right)$ *that*

$$\|\Phi^*(\varepsilon)\| \leq \frac{C_2 K m_{\mathcal{X}}}{2\sqrt{m_\Phi}} (1 + \delta_1)^{1/2}.$$

Proof. Let $E = \Phi^*(\varepsilon)$. So, $\|\Phi^*(\varepsilon)\| = \sup_{v,w\in\mathbb{S}^{m_{\mathcal{X}}-1}} |\langle v, Ew\rangle|$.

$$\begin{aligned}\langle v, Ew\rangle &= \text{trace}(v^T Ew) = \text{trace}(Ewv^T)\\ &= \text{trace}(\Phi^*(\varepsilon)wv^T) = \left\langle vw^T, \Phi^*(\varepsilon)\right\rangle\\ &= \left\langle \Phi(vw^T), \varepsilon\right\rangle \leq \|\varepsilon\|_{l_2^{m_\Phi}} \left\|\Phi(vw^T)\right\|_{l_2^{m_\Phi}}.\end{aligned}$$

Using (4) and since $\left\|\Phi(vw^T)\right\|^2_{l_2^{m_\Phi}} \leq (1+\delta_1)$ we arrive at the bound on $\|\Phi^*(\varepsilon)\|$. □

We now present the error bound for the *matrix Dantzig selector* as was obtained in [11] in Theorem 1. In Corollary 1, we exploit this result in our setting for $r = 1$ in order to obtain the error bound for recovering the rank 1 approximation $\hat{X}^{(1)}_{DS}$ to X.

Theorem 1. *Let rank(X) ≤ r and let $\hat{X}_{DS}$ be the solution to* (5)*. If $\delta_{4r} < \delta < \sqrt{2}-1$ and $\|\Phi^*(\varepsilon)\| \leq \lambda$, then with probability at least $1 - 2\exp\left(-\left(m_\Phi q(\delta) - 4r(d+m_{\mathcal{X}}+1)u(\delta)\right)\right)$ we have*

$$\left\|\hat{X}_{DS} - X\right\|^2_F \leq C_0 r\lambda^2,$$

where C_0 depends only on the isometry constant δ_{4r}.

Corollary 1. *Let $\hat{X}^{(1)}_{DS}$ be the best rank 1 approximation (in the sense of $\|\cdot\|_F$) to $\hat{X}_{DS}$. If $\delta_4 < \delta < \sqrt{2}-1$ we have with probability at least $1 - 2\exp\left(-\left(m_\Phi q(\delta) - 4(d+m_{\mathcal{X}}+1)u(\delta)\right)\right)$ that*

$$\left\|X - \hat{X}^{(1)}_{DS}\right\|^2_F \leq \frac{C_0 C_2^2 K^2 m_{\mathcal{X}}^2}{m_\Phi}(1+\delta),$$

where C_0 is a constant depending only on δ.

Proof. Lemma 1 in conjunction with Theorem 1 gives us the following bound on $\left\|X - \hat{X}_{DS}\right\|^2_F$:

$$\left\|X - \hat{X}_{DS}\right\|^2_F \leq \frac{C_0 C_2^2 K^2 m_{\mathcal{X}}^2}{4m_\Phi}(1+\delta).$$

In general $\text{rank}(\hat{X}_{DS}) > 1$, thus we consider the best rank 1 approximation to $\hat{X}_{DS}$, in the sense of $\|\cdot\|_F$. We then obtain the following error bound:

$$\begin{aligned}\left\|X - \hat{X}^{(1)}_{DS}\right\|_F &\leq \left\|X - \hat{X}_{DS}\right\|_F + \left\|\hat{X}_{DS} - \hat{X}^{(1)}_{DS}\right\|_F,\\ &\leq 2\left\|X - \hat{X}_{DS}\right\|_F.\end{aligned}$$

Here $\left\|\hat{X}_{DS} - \hat{X}^{(1)}_{DS}\right\|_F \leq \left\|X - \hat{X}_{DS}\right\|_F$ as $\hat{X}^{(1)}_{DS}$ is the best rank 1 approximation to $\hat{X}_{DS}$ in the sense of $\|\cdot\|_F$. □

3.2. Approximation of a

In the previous section, we have found a rank 1 approximation $\hat{X}^{(1)}_{DS}$ to the original rank 1 matrix, X. Now, we let

$$X = \sigma \mathbf{a}\mathbf{g}^T,$$
$$\hat{X}^{(1)}_{DS} = \hat{\sigma}\hat{\mathbf{a}}\hat{\mathbf{g}}^T,$$

where $\sigma = (\sum_{j=1}^{m_{\mathcal{X}}} \left|g'(\mathbf{a}^T\xi_j)\right|^2)^{1/2}$, $\hat{\sigma} > 0$ and $\|\mathbf{a}\| = \|\mathbf{g}\| = \|\hat{\mathbf{a}}\| = \|\hat{\mathbf{g}}\| = 1$. We now show that if the bound on $\left\|X - \hat{X}^{(1)}_{DS}\right\|_F$ is driven to be lower than a certain value then it guarantees probabilistically a lower bound on $|\langle \mathbf{a}, \hat{\mathbf{a}}\rangle|$. This is stated precisely in Lemma 2

Lemma 2. *For a fixed $0 < \rho < 1$, $m_{\mathcal{X}} \geq 1$, $m_\Phi < m_{\mathcal{X}} d$, if $\epsilon < \frac{1}{d}\left(\frac{m_\Phi \alpha(1-\rho)}{C_0 C_2^2 m_{\mathcal{X}}(1+\delta)}\right)^{1/2}$, then we have with probability at least*

$$1 - 2\exp\left(\frac{-2m_{\mathcal{X}}\alpha^2\rho^2}{C_2^4}\right) - 2\exp\left(-\left(m_\Phi q(\delta) - 4(d+m_{\mathcal{X}}+1)u(\delta)\right)\right),$$

that $|\langle \mathbf{a}, \hat{\mathbf{a}}\rangle| \geq \left(\frac{\sqrt{m_{\mathcal{X}}\alpha(1-\rho)} - \tau}{\sqrt{m_{\mathcal{X}}\alpha(1+\rho)} + \tau}\right)$,

where $\tau^2 = \frac{C_0 C_2^2 K^2 m_{\mathcal{X}}^2}{m_\Phi}(1+\delta)$ is the error bound derived in Corollary 1.

Proof. $\left\|X - \hat{X}^{(1)}_{DS}\right\|^2_F = \sigma^2 + \hat{\sigma}^2 - 2\sigma\hat{\sigma}\,\langle \mathbf{a}, \hat{\mathbf{a}}\rangle\,\langle \mathbf{g}, \hat{\mathbf{g}}\rangle$. From Weyls inequality [14] we have $\left\|X - \hat{X}^{(1)}_{DS}\right\|_F \leq \tau \Rightarrow |\sigma - \hat{\sigma}| \leq \tau$. Hence we have

$$\begin{aligned}\langle \mathbf{a}, \hat{\mathbf{a}}\rangle\,\langle \mathbf{g}, \hat{\mathbf{g}}\rangle &\geq \frac{\sigma^2 + \hat{\sigma}^2 - \tau^2}{2\sigma\hat{\sigma}}\\ &\geq \frac{\sigma^2 + (\sigma-\tau)^2 - \tau^2}{2\sigma(\sigma+\tau)} = \frac{\sigma - \tau}{\sigma + \tau}\end{aligned}$$

From Hoeffdings inequality we have for any fixed $0 < \rho < 1$,

$$\mathbb{P}\left(\left|\frac{1}{m_{\mathcal{X}}}\sum_{j=1}^{m_{\mathcal{X}}}\left|g'(\mathbf{a}^T\xi_j)\right|^2 - \alpha\right| > \rho\alpha\right) \leq 2\exp\left(\frac{-2m_{\mathcal{X}}\alpha^2\rho^2}{C_2^4}\right)$$

So $\sigma \in [\sqrt{m_{\mathcal{X}}\alpha(1-\rho)}, \sqrt{m_{\mathcal{X}}\alpha(1+\rho)}]$ with probability at least $1 - 2\exp\left(\frac{-2m_{\mathcal{X}}\alpha^2\rho^2}{C_2^4}\right)$. Conditioning on this event, we see that $\tau < \sigma$ is ensured if $\epsilon < \frac{1}{d}\left(\frac{m_\Phi\alpha(1-\rho)}{C_0 C_2^2 m_{\mathcal{X}}(1+\delta)}\right)^{1/2}$. This completes the proof. □

3.3. Approximation of f

We now have the results necessary to state our main approximation result for the function f. Note that our estimation $\hat{f}$ is constructed in a manner similar to [10].

Theorem 2. *(Main approximation theorem) Let us fix $0 < \rho < 1, 0 < \delta < \sqrt{2}-1$. Under the assumptions and notations mentioned earlier, for a fixed $m_{\mathcal{X}} \geq 1$, $m_\Phi < m_{\mathcal{X}} d$ and $\epsilon < \frac{1}{d}\left(\frac{m_\Phi\alpha(1-\rho)}{C_0 C_2^2 m_{\mathcal{X}}(1+\delta)}\right)^{1/2}$ we have with probability at least*

$$1 - 2\exp\left(\frac{-2m_{\mathcal{X}}\alpha^2\rho^2}{C_2^4}\right) - 2\exp\left(-\left(m_\Phi q(\delta) - 4(d+m_{\mathcal{X}}+1)u(\delta)\right)\right)$$

that the function $\hat{f}(\mathbf{x}) = \hat{g}(\hat{\mathbf{a}}^T\mathbf{x})$ defined by means of

$$\hat{g}(y) := f(\hat{\mathbf{a}}y), \quad y \in (-(1+\bar{\epsilon}), (1+\bar{\epsilon})),$$

has the uniform approximation bound

$$\left\|f - \hat{f}\right\|_\infty \leq C_2(1+\bar{\epsilon})\sqrt{1 - \left(\frac{\sqrt{m_{\mathcal{X}}\alpha(1-\rho)} - \tau}{\sqrt{m_{\mathcal{X}}\alpha(1+\rho)} + \tau}\right)^2}.$$

Proof. For $\mathbf{x} \in B_{\mathbb{R}^d}(1+\bar{\epsilon})$ we have

$$\begin{aligned}\left|g(\mathbf{a}^T\mathbf{x}) - \hat{g}(\hat{\mathbf{a}}^T\mathbf{x})\right| &= \left|g(\mathbf{a}^T\mathbf{x}) - g((\mathbf{a}^T\hat{\mathbf{a}})(\hat{\mathbf{a}}^T\mathbf{x}))\right| \\ \leq C_2\left|\mathbf{a}^T\mathbf{x} - (\mathbf{a}^T\hat{\mathbf{a}})(\hat{\mathbf{a}}^T\mathbf{x})\right| &= C_2\left|(\mathbf{a} - \langle\mathbf{a},\hat{\mathbf{a}}\rangle\,\hat{\mathbf{a}})^T\mathbf{x}\right| \\ \leq C_2(1+\bar{\epsilon})\left\|\mathbf{a} - \langle\mathbf{a},\hat{\mathbf{a}}\rangle\,\hat{\mathbf{a}}\right\|_{l_2^d} &\leq C_2(1+\bar{\epsilon})\sqrt{1-\langle\mathbf{a},\hat{\mathbf{a}}\rangle^2}.\end{aligned}$$

The bound follows from the approximation result of Lemma 2. □

Remark 2. *Note that once* $\hat{\mathbf{a}}$ *has been obtained, one would uniformly sample the estimated function* $\hat{g}$ *on a grid* $h\mathbb{Z}\bigcap(-(1+\bar{\epsilon}),(1+\bar{\epsilon}))$*, with* $h>0$ *being the step size, and compute a suitable interpolation* $\hat{g}_h$ *(by using quasi interpolants, for example) with the following uniform approximation error bound:*

$$\|\hat{g}_h - \hat{g}\|_\infty \leq C_2 h^2.$$

Thus we would have the following:

$$\begin{aligned}\|g - \hat{g}_h\|_\infty &\leq \|g-\hat{g}\|_\infty + \|\hat{g}-\hat{g}_h\|_\infty \\ &\leq C_2(1+\bar{\epsilon})\sqrt{1-\left(\frac{\sqrt{m_{\mathcal{X}}\alpha(1-\rho)}-\tau}{\sqrt{m_{\mathcal{X}}\alpha(1+\rho)}+\tau}\right)^2} + C_2h^2.\end{aligned}$$

Remark 3. *Similar to [10], the approximation perfomance of our learning scheme is determined by* α*. It was shown in [10] that the measure* $\mu_{\mathbb{S}^{d-1}}$ *determines a push-forward measure* μ_1 *on the unit interval* $B_{\mathbb{R}}$ *which concentrates around* 0*, exponentially fast as* $d\to\infty$*. In other words the asymptotic behaviour of* α *is determined completely by the function* g' *in a neighbourhood of* 0*. In particular, when* g *is* $\mathcal{C}^3$ *differentiable in an open neighbourhood of the origin we can show that [10]:*

1. *If* $g'(0)\neq 0$*, then* $\alpha(d)=\mathcal{O}(1)$*,* $\epsilon=\mathcal{O}(1/\sqrt{d})$ *as* $d\to\infty$
2. *If* $g'(0)=0$ *and* $g''(0)\neq 0$ *then* $\alpha(d)=\mathcal{O}(1/d)$*,* $\epsilon=\mathcal{O}(1/(d\sqrt{d}))$ *as* $d\to\infty$

Remark 4. *Finally we see that for a given fixed* ϵ*, the sampling complexity of our learning scheme is* $m_{\mathcal{X}}(m_\Phi+1)$*, which is delineated in Table 1. Observe that the smoothness properties of* g *at the origin significantly impacts the sampling complexity of our learning scheme, by a factor of 3. Furthermore the difference in sampling complexities with the case when* $\mathbf{a}$ *is sparse (Fornasier et al. [10]), which has the same sampling scheme, can be observed clearly. This difference arises on account of the compressive sensing tools used due to the sparsity assumption made on* $\mathbf{a}$*.*

Fornasier et al.	$m_{\mathcal{X}}$	m_Φ	$m_{\mathcal{X}}\times m_\Phi$
$g'(0)=0,\ g''(0)\neq 0$	$\mathcal{O}(d^2)$	$\mathcal{O}(d^{\frac{q}{2-q}})$	$\mathcal{O}(d^{2+\frac{q}{2-q}})$
$g'(0)\neq 0$	$\mathcal{O}(1)$	$\mathcal{O}(\log d)$	$\mathcal{O}(\log d)$
Our work	$m_{\mathcal{X}}$	m_Φ	$m_{\mathcal{X}}\times m_\Phi$
$g'(0)=0,\ g''(0)\neq 0$	$\mathcal{O}(d^2)$	$\mathcal{O}(d^2)$	$\mathcal{O}(d^4)$
$g'(0)\neq 0$	$\mathcal{O}(1)$	$\mathcal{O}(d)$	$\mathcal{O}(d)$

Table 1. Comparison of sampling complexities when $\mathbf{a}$ is sparse (Fornasier et al [10]) with no sparsity assumption based recovery scheme (our work) for the cases when: (i) $g'(0)=0$, $g''(0)\neq 0$ and (ii) $g'(0)\neq 0$. Here g is assumed to be $\mathcal{C}^3$ differentiable in an open neighborhood of the origin.

4. CONCLUSIONS

In this paper, we consider the problem of learning ridge functions of the form $f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})$, for arbitrary $\mathbf{a}\in\mathbb{R}^d$ with $\|\mathbf{a}\|_{l_2^d}=1$. By removing the sparsity assumption on $\mathbf{a}$, we generalize the work done in [10]. Assuming g to be a C^2 function, our learning strategy leverages a low rank matrix recovery program [11] to first recover an approximation $\hat{\mathbf{a}}$ to $\mathbf{a}$, and then uses $\hat{\mathbf{a}}$ to form an approximation to f. We establish the sampling complexity of our approach to be polynomial in the dimension d. Without loss of generality, we treat the case with only a single parameter $\mathbf{a}$ as the results are easier to interpret, however the case when $m>1$ in (1) can also be treated in our setting, which is left for future work.

5. REFERENCES

[1] A. Cohen, I. Daubechies, R. A. DeVore, G. Kerkyacharian, and D. Picard, "Capturing ridge functions in high dimensions from point queries," *Constructive Approximation*, pp. 1–19, 2011.

[2] B.F. Logan and L.A. Shepp, "Optimal reconstruction of a function from its projections.," *Duke Math. J.*, vol. 42, pp. 645–659, 1975.

[3] J.H. Friedman and W. Stuetzel, "Projection pursuit regression.," *J. Amer. Statist. Assoc.*, vol. 76, pp. 817–823, 1981.

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

[5] P.J. Huber, "Projection pursuit.," *Ann. Statist.*, vol. 13, pp. 435–475, 1985.

[6] A. Pinkus, "Approximation theory of the MLP model in neural networks.," *Acta Numerica*, vol. 8, pp. 143–195, 1999.

[7] E.J. Candès, "Harmonic analysis of neural networks.," *Appl. Comput. Harmon. Anal.*, vol. 6, no. 2, pp. 197–218, 1999.

[8] E.J. Candès and D.L. Donoho, "Ridgelets: a key to higher dimensional intermittency?," *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, vol. 357, no. 1760, pp. 2495–2509, 1999.

[9] E.J. Candès, "Ridgelets: Estimating with ridge functions.," *Ann. Stat.*, vol. 31, no. 5, pp. 1561–1599, 2003.

[10] M. Fornasier, K. Schnass, and J. Vybíral, "Learning functions of few arbitrary linear parameters in high dimensions," *CoRR*, vol. abs/1008.3043, 2010.

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

[12] B. Recht, M. Fazel, and P.A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *Submitted to SIAM review*, 2007.

[13] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *The Annals of Statistics*, vol. 28, no. 5, pp. 1302–1338.

[14] H. Weyl, "Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung)," *Mathematische Annalen*, vol. 71, pp. 441–479, 1912.