

ACTIVE LEARNING OF SELF-CONCORDANT LIKE MULTI-INDEX FUNCTIONS

Ilija Bogunovic*

Volkan Cevher*

Jarvis Haupt†

Jonathan Scarlett*

* LIONS, EPFL, Lausanne, Switzerland

{ilija.bogunovic, volkan.cevher, jonathan.scarlett}@epfl.ch*

† Department of Electrical and Computer Engineering, University of Minnesota,
Twin Cities Minneapolis, MN 55414
jdhaupt@umn.edu

ABSTRACT

We study the problem of actively learning a *multi-index* function of the form $f(\mathbf{x}) = g_0(A_0\mathbf{x})$ from its point evaluations, where $A_0 \in \mathbb{R}^{k \times d}$ with $k \ll d$. We build on the assumptions and techniques of an existing approach based on low-rank matrix recovery (Tyagi and Cevher, 2012). Specifically, by introducing an additional *self-concordant like* assumption on g_0 and adapting the sampling scheme and its analysis accordingly, we provide a bound on the sampling complexity with a weaker dependence on d in the presence of additive Gaussian sampling noise. For example, under natural assumptions on certain other parameters, the dependence decreases from $\mathcal{O}(d^{3/2})$ to $\mathcal{O}(d^{3/4})$.

Index Terms— Function learning, multi-index functions, low-rank matrix recovery, Dantzig selector

1. INTRODUCTION

The problem of approximating a function $f : \Omega \rightarrow \mathbb{R}$ ($\Omega \subseteq \mathbb{R}^d$) from its point values arises frequently in machine learning and statistics. This problem is intractable in general; for example, for general functions belonging to $C^s[0, 1]^d$ for a fixed smoothness order $s > 0$, exponentially many point samples are needed to obtain a given uniform approximation error $\epsilon \in (0, 1)$ [1]. However, under further structural assumptions on f , the problem becomes tractable, requiring only polynomially many point samples [2, 3, 4]. In this paper, we consider a class of functions known as *multi-index functions* (studied in statistics under the name of “projection pursuit regression”, e.g. [5], [6] and [7]), which are of interest in numerous areas including neural networks [8], ridgelets [9], and econometrics [10].

Before formally stating our setup, we briefly outline some of the most relevant previous works. Cohen *et al.* [2] proposed a method for recovering functions of the form $f(\mathbf{x}) = g(\mathbf{a}^T \mathbf{x})$, where g is a C^s function for some $s > 1$, and \mathbf{a} is both stochastic (i.e. its entries are non-negative and sum to one) and compressible. Leveraging on the latter assumption, tools from compressive sensing were applied. Fornasier *et*

al. [3] extended this work to handle functions of the form $f(\mathbf{x}) = g(A\mathbf{x})$, where g is a C^2 function and A is a full-rank $k \times d$ matrix with compressible rows. The work most relevant to ours is that of Tyagi and Cevher [4] (see also [11]), who proposed methods based on low rank matrix recovery in order to drop the assumption that A has compressible rows.

A key limitation of the results presented in [4] is the dependence of the sampling complexity on d in the case of noisy samples. Specifically, under some natural assumptions, the bound therein on the sampling complexity is $\mathcal{O}(d^{3/2})$. The main result of this paper shows that, by a variation of the techniques in [4] and the introduction of a *self-concordant like* assumption [12], this can be improved to $\mathcal{O}(d^{3/4})$.

1.1. Problem Setup and Assumptions

Let $\bar{\epsilon} \in (0, 1)$ be a positive constant, and let $B_{\mathbb{R}^d}(r)$ be the ball of radius r in \mathbb{R}^d . We consider the approximation of a function $f : B_{\mathbb{R}^d}(1 + \bar{\epsilon}) \rightarrow \mathbb{R}$ of the form

$$f(\mathbf{x}) = g_0(A_0\mathbf{x}) \quad (1)$$

where $A_0 \in \mathbb{R}^{k \times d}$ is a full-rank matrix with $k \ll d$, and g_0 is a function on \mathbb{R}^k (both of which are unknown). The goal is to construct an approximation \hat{f} of f based on a number of samples whose location may be chosen freely. We consider noiseless samples in Section 3, and noisy samples in Section 4. We consider the trade-off between the number of samples and the *worst-case* approximation error $\|\hat{f} - f\|_{L_\infty}$. At a high level, this problem is tractable due to the reduction in dimensionality (from d to k).

We proceed by introducing the class of self-concordant like functions. The definition resembles the usual definition of self concordance, but it should be noted that neither class is a subset of the other. A notable example of a self-concordant like function is the logistic function, which appears frequently in neural network learning problems.

For a multivariate function $h(y)$ and a vector $\beta \in \mathbb{Z}^m$, we define the derivative operator $D^\beta h = \frac{\partial^{|\beta|} h}{\partial y_1^{\beta_1} \dots \partial y_m^{\beta_m}}$, where $|\beta| = \sum_{i=1}^m \beta_i$ (e.g. $D^2 h(y)[u, u] = u^T \nabla^2 h(y) u$).

Definition 1.1 A function $h : \text{dom}(h) \rightarrow \mathbb{R}$ defined on an open domain $\text{dom}(h) \subseteq \mathbb{R}^m$ is *self-concordant like* with parameter $M \geq 0$ with respect to a norm $\|\cdot\|$ on \mathbb{R}^m if

*This work is supported in part by the European Commission under the grants MIRG-268398 and ERC Future Proof, and by the Swiss Science Foundation under the grants SNF 200021-146750 and SNF CRSII2-147633

1. $h \in C^3(\text{dom}(h))$;
2. $|D^3h(\mathbf{x})[\mathbf{u}, \mathbf{v}, \mathbf{v}]| \leq M \|\mathbf{u}\| D^2h(\mathbf{x})[\mathbf{v}, \mathbf{v}]$ for all $\mathbf{x} \in \text{dom}(h)$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$.

Our assumptions on f and g_0 are given as follows:

1. The function g_0 belongs to C^2 , and hence there exists a constant $C_{0,2} > 0$ such that

$$\max_{|\beta| \leq 2} \|D^\beta g_0\|_\infty \leq C_{0,2}, \quad (2)$$

where the ℓ_∞ -norm is understood to act on the vectorization of the derivative matrix in the case that $\beta = 2$.

2. Letting $\mu_{\mathbb{S}^{d-1}}$ be the uniform measure on the unit sphere \mathbb{S}^{d-1} in d -dimensional space, the matrix

$$H^f := \int_{\mathbb{S}^{d-1}} \nabla f(\mathbf{x}) \nabla f(\mathbf{x})^T d\mu_{\mathbb{S}^{d-1}}(\mathbf{x}) \quad (3)$$

is well-conditioned in the sense that its singular values satisfy $\sigma_1(H^f) \geq \dots \geq \sigma_k(H^f) \geq \alpha > 0$ for some positive constant α . It should be noted that α may scale with d , typically as $\mathcal{O}(1/d)$; see [4] for examples.

3. The function g_0 is self-concordant like with respect to the ℓ_2 -norm, with some parameter $M_0 \geq 0$.

The final of these assumptions is the key difference here compared to [4].

As was noted in [3], we can reduce (1) to a simpler model with a *row-orthonormal* matrix A :

$$f(\mathbf{x}) = g(A\mathbf{x}). \quad (4)$$

Specifically, this follows from a singular value decomposition (SVD): Write $A_0 = U\Sigma V^T$, and set $A = V^T$ and $g(y) = g_0(U\Sigma y)$. By a direct differentiation, it is readily verified that the derivatives of g are bounded as in (2) with a constant $C_2 := \sigma_0^2 C_{0,2}$, where σ_0 is the spectral norm of A_0 .

Applying the chain rule to (1) and using the assumption 3., it can similarly be verified that f is also self-concordant like, with parameter $M := \sigma_0 M_0$. As will be seen in the later sections, this implies that the consideration of the simplified model (4) only affects our analysis up to multiplicative powers of σ_0 . We assume that σ_0 is uniformly bounded in d , implying that these factors do not affect the resulting scaling laws, i.e. C_2 and M are uniformly bounded.

2. SAMPLING SCHEME

In this section, we describe a method for taking samples and using them to construct a low-rank matrix recovery problem that will provide the starting point of our analysis.

2.1. Sampling Points

We describe a scheme taking $2m_\chi m_\Phi$ samples, where m_χ and m_Φ are integers. As in [4], we construct a set of *sampling centers* $\mathcal{X} = \{\xi_1, \xi_2, \dots, \xi_{m_\chi}\}$, drawn i.i.d. from the

unit sphere \mathbb{S}^{d-1} according to the uniform measure $\mu_{\mathbb{S}^{d-1}}$. For each sampling center ξ_j ($j = 1, \dots, m_\chi$), we randomly construct a set of *direction vectors* $\{\phi_{1,j}, \dots, \phi_{m_\Phi,j}\}$, where the entries of the vectors are i.i.d. and equiprobable on $\{\frac{-1}{\sqrt{m_\Phi}}, \frac{1}{\sqrt{m_\Phi}}\}$. These are collected into m_Φ matrices $\Phi_i = [\phi_{i,1}, \dots, \phi_{i,m_\chi}]_{d \times m_\chi}$. For each sampling center ξ and sampling direction ϕ , we take two samples of the function f , namely $f(\xi + \epsilon\phi)$ and $f(\xi - \epsilon\phi)$ (though as we will see shortly, it suffices to know only their difference).

2.2. Formulation of a Low-Rank Recovery Problem

Let $X := A^T [\nabla g(A\xi_1) \dots \nabla g(A\xi_{m_\chi})]_{k \times m_\chi}$ be a matrix containing the gradients of g at the sampling centers. Since A has rank k , the matrix X also has low rank (at most k). As we will see shortly, the low rank property of X allows us to use low rank matrix recovery techniques to approximate X and infer, up to a rotation, the subspace matrix A .

Since we cannot evaluate ∇g directly, the elements of the gradient matrix X are estimated via a linear approximation of f . We make use of the following Taylor expansion:

$$f(\xi + \epsilon\phi) = f(\xi) + \langle \nabla f(\xi), \phi \rangle \epsilon + \frac{\epsilon^2}{2} \phi^T \nabla^2 f(\xi) + \frac{\epsilon^3}{3!} \nabla^3 f(\zeta^+) [\phi, \phi, \phi] \quad (5)$$

for a suitable value of ζ^+ . By forming a similar expansion with $-\epsilon$ in place of ϵ , and taking the difference between the two expansions, we obtain

$$\langle \nabla f(\xi), \phi \rangle = \frac{1}{2\epsilon} (f(\xi + \epsilon\phi) - f(\xi - \epsilon\phi)) + E(\xi, \epsilon, \phi), \quad (6)$$

where the remainder term is given by

$$E(\xi, \epsilon, \phi) = \frac{\epsilon^2}{2 \cdot 3!} [\nabla^3 f(\zeta^-) [\phi, \phi, \phi] + \nabla^3 f(\zeta^+) [\phi, \phi, \phi]] \quad (7)$$

for suitably chosen $\zeta^- \in [\xi - \epsilon\phi, \xi]$ and $\zeta^+ \in [\xi, \xi + \epsilon\phi]$.

Since $f(\xi + \epsilon\phi) - f(\xi - \epsilon\phi)$ is known, (6) allows us to obtain information about the gradients up to the error term $E(\xi, \epsilon, \phi)$ and possible sampling noise. We observe that the second-order terms have canceled in (6); this is a key a difference in our analysis compared to [4].

Applying (6) for each sampling center and direction, we can obtain the following linear system:

$$\mathbf{y} = \mathcal{F}(X) + \mathbf{z} + \boldsymbol{\varepsilon} \quad (8)$$

where \mathbf{y} are the measurements, \mathbf{z} represents possible sampling noise, and $\boldsymbol{\varepsilon}$ represents the accumulated error:

$$y_i = \frac{1}{2\epsilon} \sum_{j=1}^{m_\chi} [f(\xi_j + \epsilon\phi_{i,j}) - f(\xi_j - \epsilon\phi_{i,j})]$$

$$\varepsilon_i = \sum_{j=1}^{m_\chi} E(\xi_j, \epsilon, \phi_i).$$

The linear measurement operator $\mathcal{F} : \mathbb{R}^{d \times m_\chi} \rightarrow \mathbb{R}^{m_\Phi}$ is defined as $\mathcal{F}(X)_i = \text{Tr}(\Phi_i X)$.

Proposition 2.1 *For the error term in (8), we have*

$$\|\varepsilon\|_{l_2^{m_\Phi}} \leq M \frac{\epsilon^2 k^2 C_2 d^{3/2} m_\chi}{6 m_\Phi}. \quad (9)$$

Proof. Let $\varphi_{i,j}^+$ denote $\nabla^3 f(\zeta_{i,j}^+) [\phi_{i,j}, \phi_{i,j}, \phi_{i,j}]$ and $\varphi_{i,j}^-$ denote $\nabla^3 f(\zeta_{i,j}^-) [\phi_{i,j}, \phi_{i,j}, \phi_{i,j}]$. By definition,

$$\|\varepsilon\|_{l_2^{m_\Phi}}^2 = \left(\frac{\epsilon^2}{12}\right)^2 \left(\sum_{i=1}^{m_\Phi} \left| \sum_{j=1}^{m_\chi} \varphi_{i,j}^+ + \varphi_{i,j}^- \right|^2 \right). \quad (10)$$

Since f is self-concordant like with parameter M , we have

$$|\varphi_{i,j}^+| \leq M \|\phi_{i,j}\|_{l_2} D^2 f(\zeta_{i,j}^+) [\phi_{i,j}, \phi_{i,j}].$$

Moreover, a direct differentiation yields

$$\begin{aligned} |D^2 f(\zeta_{i,j}^+) [\phi_{i,j}, \phi_{i,j}]| &= \left| \phi_{i,j}^T A^T \nabla^2 g(A \zeta_{i,j}^+) A \phi_{i,j} \right| \\ &\leq \|A \phi_{i,j}\|_{l_2}^2 \|\nabla^2 g(A \zeta_{i,j}^+)\|_F \\ &\leq \frac{d k^2 C_2}{m_\Phi}, \end{aligned}$$

since $\|\nabla^2 g(A \zeta_{i,j}^+)\|_F \leq k C_2$ by the definition of C_2 following (4). Handling $D^2 f(\zeta_{i,j}^-)$ similarly and applying the triangle inequality, we obtain

$$|\varphi_{i,j}^+ + \varphi_{i,j}^-| \leq 2M \frac{d^{3/2} k^2 C_2}{m_\Phi^{3/2}}.$$

Combining this with (10), we obtain the desired result.

3. NOISELESS OBSERVATIONS

We proceed along the same lines as in [4]. First, we look at the noise free setting of (8), i.e. $\mathbf{y} = \mathcal{F}(X) + \varepsilon$. We use low-rank recovery to recover an approximation of X , which is then used to obtain an approximate subspace matrix \hat{A} with a guaranteed lower bound on $\|A \hat{A}^T\|_F$. This is then used to obtain the final function approximation.

3.1. Stable Low Rank Recovery

As shown in [4], under the random construction of the sampling directions, the linear measurement operator \mathcal{F} satisfies the matrix restricted isometry property [13] with high probability. More precisely, for all rank- k matrices, it holds that

$$(1 - \kappa_k) \|X_k\|_F^2 \leq \|\mathcal{F}(X_k)\|_{l_2}^2 \leq (1 + \kappa_k) \|X_k\|_F^2$$

with probability at least $1 - 2e^{-m_\Phi q(\kappa) + (d+m_\chi+1)u(\kappa)}$, where the RIP constant κ_k satisfies $0 < \kappa_k \leq \kappa < 1$. The RIP property, together with the low-rank property of the matrix X , allows us to use stable low-rank recovery algorithms to obtain an approximation \hat{X} of X .

We make use of the following convex optimization problem, known as the matrix Dantzig selector [13]:

$$\hat{X}_{\text{DS}} = \arg \min \|M\|_* \text{ s.t. } \|\mathcal{F}^*(\mathbf{y} - \mathcal{F}(M))\| \leq \lambda, \quad (11)$$

where $\|\cdot\|_*$ and $\|\cdot\|$ are the nuclear and operator norms, $\mathcal{F}^* : \mathbb{R}^{m_\Phi} \rightarrow \mathbb{R}^{d \times m_\chi}$ is the adjoint operator of \mathcal{F} , and λ is a tuning parameter.

We seek to choose λ such that X is feasible i.e., $\|\mathcal{F}^*(\varepsilon)\| \leq \lambda$. The following lemma serves this purpose, and is proved using the steps in Appendix C of [4].

Lemma 3.1 *For any ε satisfying (9), we have*

$$\|\mathcal{F}^*(\varepsilon)\| \leq \lambda^* := M \frac{\epsilon^2 k^2 C_2 d^{3/2} m_\chi}{6 m_\Phi} (1 + \kappa)^{1/2}, \quad (12)$$

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

We now choose λ in (11) to equal λ^* in (12), and apply Corollary 1 from [4] (based on Theorem 2.4 in [13]) together with Lemma 3.1. The result is the following corollary.

Corollary 3.1 *Let $\hat{X}_{\text{DS}}^{(k)}$ be the best rank- k approximation to the solution \hat{X}_{DS} of (11) in Frobenius norm. If $\|\mathcal{F}^*(\varepsilon)\| \leq \lambda = \lambda^*$ and $\kappa_{4k} < \kappa < \sqrt{2} - 1$ then with probability at least $1 - 2e^{-m_\Phi q(\kappa) + 4k(d+m_\chi+1)u(\kappa)}$ we have*

$$\|X - \hat{X}_{\text{DS}}^{(k)}\|_F^2 \leq 4C_0 k \lambda^2 = C_0 M^2 \frac{\epsilon^4 k^5 C_2^2 d^3 m_\chi^2}{9 m_\Phi^2} (1 + \kappa).$$

3.2. Subspace Approximation

Next, we perform an SVD of $\hat{X}_{\text{DS}}^{(k)}$, namely $\hat{X}_{\text{DS}}^{(k)} = \hat{A} \hat{\Sigma} \hat{V}$. While our algorithm does not necessarily recover an accurate estimate of A , the following lemma provides conditions under which it does provide an accurate estimate up to a rotation. The proof follows Appendix E of [4], with the key tool being the matrix Chernoff bound.

Lemma 3.2 *Fix $m_\chi \geq 1$, $m_\Phi < m_\chi d$, and $0 < \rho < 1$. If*

$$\epsilon < \frac{1}{k} \left(\frac{3m_\Phi}{M C_2 (\sqrt{k} + \sqrt{2})} \right)^{1/2} \left(\frac{(1 - \rho)\alpha}{(1 + \kappa) C_0 d^3 m_\chi} \right)^{1/4}, \quad (13)$$

then with probability at least $1 - k \exp\left\{-\frac{m_\chi \alpha \rho^2}{2k C_2^2}\right\} - 2 \exp\{-m_\Phi q(\kappa) + 4k(d + m_\chi + 1)u(\kappa)\}$, we have

$$\|A \hat{A}^T\|_F \geq \left(k - \frac{2\tau^2}{(\sqrt{(1 - \rho m_\chi \alpha) - \tau^2})} \right)^{1/2} \quad (14)$$

where $\tau^2 = C_0 M^2 \frac{\epsilon^4 k^5 C_2^2 d^3 m_\chi^2}{9 m_\Phi^2} (1 + \kappa)$ is the error bound derived in Corollary 3.1, and α is defined following (3).

3.3. Function Approximation

We now form an approximation of f , namely $\tilde{f}(\mathbf{x}) = \tilde{g}(\hat{A}\mathbf{x})$ with $\tilde{g}(\mathbf{y}) := f(\hat{A}^T \mathbf{y})$. This should not be considered the final approximation, as its evaluation requires sampling f ; however, one can form the final estimate \hat{f} by uniformly approximating \tilde{g} via quasi-interpolants [4]. The resulting approximation error is bounded in a straightforward fashion via the triangle inequality. Since these arguments are well-known, we omit them here, and we focus our attention on bounding the error between f and \tilde{f} .

It was shown in Appendix F of [4] that

$$\|f - \tilde{f}\|_{L_\infty} \leq C_2 \sqrt{k} (1 + \bar{\epsilon}) (k - \|A \hat{A}^T\|_F^2). \quad (15)$$

Under the conditions of Lemma 3.2, we can combine this bound with (14) to deduce that

$$\|f - \tilde{f}\|_{L_\infty} \leq C_2 \sqrt{k} (1 + \bar{\epsilon}) \left(\frac{\sqrt{2}\tau}{\sqrt{(1-\rho)m_\chi \alpha} - \tau} \right) \quad (16)$$

with probability $1 - k \exp\left\{-\frac{m_\chi \alpha \rho^2}{2kC_2^2}\right\} - 2 \exp\{-m_\Phi q(\kappa) + 4k(d + m_\chi + 1)u(\kappa)\}$ or higher. Upper bounding $\bar{\epsilon}$ by one in (16) and performing some algebra, it follows that $\|f - \tilde{f}\|_{L_\infty} \leq \delta$ provided that

$$\tau \leq \frac{\delta \sqrt{(1-\rho)m_\chi \alpha}}{2C_2 \sqrt{2k} + \delta} \quad (17)$$

Combining this with the definition of τ , we see that this holds provided that

$$\epsilon \leq \left(\frac{3\delta m_\Phi}{MC_2(2C_2 \sqrt{2k} + \delta)} \right)^{1/2} \left(\frac{(1-\rho)\alpha}{(1+\kappa)C_0 k^5 d^3 m_\chi} \right)^{1/4}. \quad (18)$$

Finally, we fix p_1 and p_2 and choose m_χ and m_Φ as in [4]:

$$m_\Phi \geq \frac{\log(2/p_2) + 4k(d + m_\chi + 1)u(\kappa)}{q(\kappa)}, \quad (19)$$

$$m_\chi \geq \frac{2kC_2^2}{\alpha \rho^2} \log(k/p_1). \quad (20)$$

The former choice ensures that \mathcal{F} satisfies the RIP with high probability, and the latter ensures that the gradient matrix X has rank k . Putting things together, we obtain the following.

Theorem 3.1 *Fix the constants $\delta \in \mathbb{R}^+$, $\rho \in (0, 1)$, $\kappa < \sqrt{2} - 1$, $p_1 > 0$, and $p_2 > 0$, and suppose that ϵ , m_χ and m_Φ satisfy (18)–(20) and the conditions of Lemma 3.2. Then the function \tilde{f} satisfies $\|f - \tilde{f}\|_{L_\infty} \leq \delta$ with probability at least $1 - p_1 - p_2$.*

From Theorem 3.1, it is possible to obtain uniform approximation guarantees on \tilde{f} with high probability under the scalings $\epsilon = \mathcal{O}\left(\frac{\alpha^{1/2}}{d^{1/4}}\right)$, $m_\chi = \mathcal{O}\left(\frac{k \log k}{\alpha}\right)$, and $m_\Phi = \mathcal{O}(k(d + m_\chi))$. The latter two scalings coincide with those in [4], whereas the former is significantly different to the behavior $\epsilon = \mathcal{O}\left(\frac{\alpha}{d^{1/2}}\right)$ from [4]. We now proceed to the noisy case, where this is seen to have significant implications.

4. NOISY OBSERVATIONS

In practical applications, one generally cannot expect to acquire perfect function samples, and it is therefore imperative to understand the effects of noise. Here we consider the case that the samples are corrupted by $\mathcal{N}(0, \sigma^2)$ Gaussian noise. Since each entry z_i of \mathbf{z} in (8) is a sum of $2m_\chi$ noise terms normalized by 2ϵ , the resulting distribution is $z_i \sim \mathcal{N}(0, m_\chi \sigma^2 / 2\epsilon^2)$.

Once again, for the matrix X to be feasible in (11), we need to tune the parameter λ in (11). Using Lemma 1.1 in [13] and Lemma 3.1 of the present paper, it can be shown that with high probability,

$$\|\mathcal{F}^*(\epsilon + \mathbf{z})\| \leq \lambda^* := \frac{\sqrt{2}\gamma\sigma}{\epsilon} \sqrt{2(1+\kappa)m_\chi m_\Phi} + M \frac{\epsilon^2 k^2 C_2 d^{3/2} m_\chi}{6m_\Phi} (1+\kappa)^{1/2}, \quad (21)$$

where $\gamma > 2\sqrt{\log 12}$.

The analog of Corollary 3.1 holds true with this modified value of $\lambda = \lambda^*$, and we again seek to make the upper bound on the recovery error $\|X - \hat{X}_{\text{DS}}^{(k)}\|_F$ small by making λ small. However, as was observed in [4], we can no longer make λ smaller by decreasing ϵ , as now λ^* also depends on ϵ^{-1} . Assuming that σ is constant (i.e. it does not decay with d), the only immediate way to overcome this issue is to re-sample every point $\mathcal{O}(\epsilon^{-1})$ -times and take the average. By doing this, the sampling complexity becomes

$$m = \mathcal{O}\left(\frac{d^{1/4}}{\alpha^{1/2}} m_\chi m_\Phi\right). \quad (22)$$

As a concrete example, in the case that $\alpha = \Theta(1/d)$ (see [4] for examples), we have to re-sample each point for only $\mathcal{O}(d^{3/4})$ -times, which is a significant improvement compared to the $\mathcal{O}(d^{3/2})$ behavior derived in [4].

The latter result was used in [14] to derive the regret bound for the problem of optimizing an unknown function from noisy samples. In future work, we will investigate how the additional assumption and improved sampling complexity bound in this paper impact the regret bound therein.

5. CONCLUSIONS

We have presented a new scheme for approximating functions of the form (1), considering both noiseless and noisy point evaluations. Introducing the self-concordant like property and adapting the sampling scheme of [4], we derived a bound on the sampling complexity with a significantly weaker dependence on d compared to that in [4]. Moreover, we studied the interplay between the self-concordant like assumption and the matrix A_0 , allowing us to handle arbitrary matrices having a bounded spectral norm.

6. REFERENCES

- [1] Joseph F. Traub, Grzegorz W. Wasilkowski, and Henryk Wozniakowski, *Information-Based Complexity*, Academic Press, New York, 1988.
- [2] Albert Cohen, Ingrid Daubechies, Ronald DeVore, Gerard Kerkycharian, and Dominique Picard, “Capturing ridge functions in high dimensions from point queries,” *Constructive Approximation*, vol. 35, no. 2, pp. 225–243, April 2012.
- [3] Massimo Fornasier, Karin Schnass, and Jan Vybiral, “Learning functions of few arbitrary linear parameters in high dimensions,” *Foundations of Computational Mathematics*, vol. 12, no. 2, pp. 229–262, April 2012.
- [4] Hemant Tyagi and Volkan Cevher, “Learning non-parametric basis independent models from point queries via low-rank methods,” *Applied and Computational Harmonic Analysis*, vol. 37, no. 3, pp. 389–412, Jan. 2014.
- [5] Jerome H Friedman and Werner Stuetzle, “Projection pursuit regression,” *Journal of the American statistical Association*, vol. 76, no. 376, pp. 817–823, 1981.
- [6] Peter J Huber, “Projection pursuit,” *The annals of Statistics*, pp. 435–475, 1985.
- [7] David L Donoho and Iain M Johnstone, “Projection-based approximation and a duality with kernel methods,” *The Annals of Statistics*, pp. 58–106, 1989.
- [8] Allan Pinkus, “Approximation theory of the MLP model in neural networks,” *Acta Numerica*, vol. 8, pp. 143–195, Jan. 1999.
- [9] Emmanuel J. Candès, “Ridgelets: Estimating with ridge functions,” *Annals of Statistics*, vol. 31, pp. 1561–1599, Oct. 2003.
- [10] Yingcun Xia, “A multiple-index model and dimension reduction,” *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1631–1640, Dec. 2008.
- [11] Tyagi Hemant and Volkan Cevher, “Active learning of multi-index function models,” in *Advances in Neural Information Processing Systems* 25, 2012, pp. 1475–1483.
- [12] Francis Bach, “Self-concordant analysis for logistic regression,” *Electronic Journal of Statistics*, vol. 4, pp. 384–414, 2010.
- [13] Emmanuel J. Candès and Yaniv Plan, “Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements,” *IEEE Transactions on Information Theory*, vol. 57, pp. 2342–2359, April 2011.
- [14] Josip Djolonga, Andreas Krause, and Volkan Cevher, “High-dimensional gaussian process bandits,” in *Advances in Neural Information Processing Systems*, 2013, pp. 1025–1033.