LEARNING FUNCTIONS OF FEW ARBITRARY LINEAR PARAMETERS IN HIGH DIMENSIONS

Massimo Fornasier, Karin Schnass, and Jan Vybíral

Johann Radon Institute for Computational and Applied Mathematics,
Austrian Academy of Sciences,
Altenbergerstrasse 69, A-4040 Linz, Austria
Emails: {massimo.fornasier, karin.schnass, jan.vybiral}@oeaw.ac.at

ABSTRACT

In this talk we summarize the results of our recent work [4, 5]. Let us assume that f is a continuous function defined on a convex body in $\mathbb{R}^d$, of the form $f(x) = g(Ax)$, where A is a $k \times d$ matrix and g is a function of k variables for $k \ll d$. Using only a limited number of point evaluations $f(x_i)$, we would like to construct a uniform approximation of f. Under certain smoothness and variation assumptions on the function g, and an *arbitrary* choice of the matrix A, we present a randomized algorithm, where the sampling points $\{x_i\}$ are drawn at random and which recovers a uniform approximation of f with high probability.

We start with the case, when $f(x_1, \dots, x_d) = g(x_{i_1}, \dots, x_{i_k})$, where the indices $1 \le i_1 < i_2 < \dots < i_k \le d$ are unknown. Later on, we study the case, when $k = 1$, i.e. $f(x) = g(a \cdot x)$ and $a \in \mathbb{R}^d$ is *compressible*, and finally the problem as stated above with k arbitrary and A with compressible rows. Due to the arbitrariness of A, the choice of the sampling points will be according to suitable random distributions and our results hold with overwhelming probability. Our approach uses tools taken from the *compressed sensing* framework, recent Chernoff bounds for sums of positive-semidefinite matrices, and classical stability bounds for invariant subspaces of singular value decompositions.

Keywords— high dimensional function approximation, compressed sensing, Chernoff bounds for sums of positive-semidefinite matrices, stability bounds for invariant subspaces of singular value decompositions.

1. INTRODUCTION

We study the recovery of the function

$$f(x) = g(Ax), \quad x \in \mathbb{R}^d, \tag{1}$$

where A is a $k \times d$ matrix and g is a function of k variables for $k \ll d$. Important special cases include the following.

- A is a projection of $\mathbb{R}^d$ onto a linear span of $(e_{i_1}, \dots, e_{i_k})$, where e_{i_j} are the canonical vectors, i.e.

$$A = \begin{pmatrix} e_{i_1}^T \\ \vdots \\ e_{i_k}^T \end{pmatrix} \tag{2}$$

  and

$$f(x) = f(x_1, \dots, x_d) = g(x_{i_1}, \dots, x_{i_k}) = g(x_I). \tag{3}$$

  Here the set $I = \{i_1, \dots, i_k\} \subseteq \{1, \dots, d\}$ collects the k (unknown) active coordinates i_ℓ.

- $k = 1$, i.e.

$$f(x) = g(a \cdot x), \tag{4}$$

  where $a \in \mathbb{R}^d$ is a given vector.

First, let us give a brief overview of known results. Functions of type (3) were recently studied using deterministic algorithms in [3]. In particular, the authors of [3] describe, how to approximate f uniformly to accuracy $\|g\|_{\mathrm{Lip}} h$ by evaluating the function on $2(k+1)e^{k+1}h^{-k}\log_2 d$ adaptively chosen points. Here, $h > 0$ is the chosen precision and g is assumed to be Lipschitz with its Lipschitz norm denoted by $\|g\|_{\mathrm{Lip}}$. The non-adaptive choice of points was further treated in [7]. Furthermore, [2] studies the functions of type (4).

Our approach is different. We give a probabilistic algorithm, which gives a good approximation of f with high probability. It uses the ideas of *concentration of measure* and *compressed sensing* combined with recent Chernoff bounds for sums of positive-semidefinite matrices, and classical stability bounds for invariant subspaces of singular value decompositions.

2. ACTIVE COORDINATES

Let us start with functions of type (3) defined on $[0, 1]^d$, where A is given by (2). Similarly to the approach described in [2, 4], we rely on numerical approximations of directional derivatives $\frac{\partial f}{\partial \varphi}(x)$. For this reason, we assume that f is actually defined on a small neighborhood of $[0, 1]^d$, namely on $D = (-\bar{\epsilon}, 1 + \bar{\epsilon})^d$.

For $x \in [0,1]^d$, $\varphi \in \mathbb{R}^d$ with $\|\varphi\|_\infty := \max_i |\varphi_i| \le r$ and $\epsilon, r \in \mathbb{R}_+$, with $r\epsilon < \bar{\epsilon}$, we get by Taylor expansion the identity

$$\begin{aligned}\nabla g(Ax)^T A\varphi &= \frac{\partial f}{\partial \varphi}(x) \\ &= \frac{f(x+\epsilon\varphi) - f(x)}{\epsilon} - \frac{\epsilon}{2}[\varphi^T \nabla^2 f(\zeta)\varphi] \quad (5)\end{aligned}$$

for a suitable $\zeta(x,\varphi) \in D$. We apply (5) to the set of points $\mathcal{X} = \{x^j \in [0,1]^d : j = 1,\dots,m_{\mathcal{X}}\}$ drawn uniformly at random with respect to the Lebesgue measure and the set of directions $\Phi = \{\varphi^j \in \mathbb{R}^d : j = 1,\dots,m_\Phi\}$, where

$$\varphi^j_\ell = \begin{cases} 1/\sqrt{m_\Phi} & \text{with prob.} \quad 1/2, \\ -1/\sqrt{m_\Phi} & \text{with prob.} \quad 1/2 \end{cases}$$

for every $j \in \{1,\dots,m_\Phi\}$ and every $\ell \in \{1,\dots,d\}$. Actually we identify Φ with the $m_\Phi \times d$ matrix whose rows are the vectors $(\varphi^i)^T$. We rewrite the $m_{\mathcal{X}} \times m_\Phi$ instances of (5) in matrix notation as

$$\Phi X = Y + \mathcal{E}, \tag{6}$$

where Y and $\mathcal{E}$ are the $m_\Phi \times m_{\mathcal{X}}$ matrices defined entry-wise by

$$y_{ij} = \frac{f(x^j + \epsilon\varphi^i) - f(x^j)}{\epsilon}, \tag{7}$$

$$\varepsilon_{ij} = -\frac{\epsilon}{2}[(\varphi^i)^T \nabla^2 f(\zeta_{ij})\varphi^i], \tag{8}$$

and X is the $d \times m_{\mathcal{X}}$ matrix with i-th row

$$X^i := \left(\frac{\partial g}{\partial z_i}(Ax^1), \dots, \frac{\partial g}{\partial z_i}(Ax^{m_{\mathcal{X}}})\right),$$

for $i \in I$ and all other rows equal to zero.

Now we can already describe the idea, how to recover the (unknown) indices $i \in I$. The discussion above shows that it is enough to identify the non zero rows of X. Multiplying (6) with Φ^T from the left-hand side, we get

$$\Phi^T \Phi X = \Phi^T Y + \Phi^T \mathcal{E}. \tag{9}$$

This identity is crucial for our algorithm. Observe that Y is obtained by sampling f as described by (7), using $2m_{\mathcal{X}}m_\Phi$ function evaluations, and $\Phi^T Y$ can be calculated by a matrix product. Looking at the random construction of $\Phi^T\Phi$ we see that in expectation it is identical to the $d \times d$ identity matrix. Thus we can expect it to behave essentially like that when applied to the rank k matrix X, i.e. $\Phi^T\Phi X \approx X$. Finally, $\Phi^T\mathcal{E}$ should be small as long as ϵ was chosen small enough, leading to $\Phi^T Y \approx \Phi^T \Phi X$. Putting these pieces together we get that

$$\Phi^T Y \approx X,$$

meaning that to identify the active components of f, we just need to select the k largest rows of $\Phi^T Y$ in the maximum norm.

Expressed in a mathematical way, we need to estimate the probability that the k largest rows of $\Phi^T Y$ in the maximum norm coincide with the k non-vanishing rows of X. This was done in [5], where the following theorem was proved.

Theorem 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a sparse function as described in* (3) *that is defined and twice continuously differentiable on a small neighborhood of* $[0,1]^d$. *For $L \le d$, a positive real number, the randomized algorithm described above recovers the k unknown active coordinates of f with probability at least* $1 - 6\exp(-L)$ *using only*

$$\mathcal{O}(k(L + \log k)(L + \log d)) \tag{10}$$

samples of f.

3. ONE DIMENSIONAL CASE

We consider functions $f : B_{\mathbb{R}^d} \to \mathbb{R}$ of type (4) with $\|a\|_{\ell_2^d} = 1$ and

$$\|a\|_{\ell_q^d} := \left(\sum_{j=1}^d |a_j|^q\right)^{1/q} \le C_1 \tag{11}$$

for some $0 < q \le 1$. Here, $B_{\mathbb{R}^d}$ stands for the unit ball of $\mathbb{R}^d$. As before, we suppose that f us defined on some $\bar{\epsilon}$ neighborhood of B, i.e. $(1+\bar{\epsilon})B$. Furthermore, we assume that

$$\max_{0\le\alpha\le 2} \|D^\alpha g\|_\infty \le C_2 \tag{12}$$

and

$$\begin{aligned}\alpha &= \int_{\mathbb{S}^{d-1}} \|\nabla f(x)\|^2_{\ell_2^d} d\mu_{\mathbb{S}^{d-1}}(x) \\ &= \int_{\mathbb{S}^{d-1}} |g'(a \cdot x)|^2 d\mu_{\mathbb{S}^{d-1}}(x) > 0, \end{aligned} \tag{13}$$

where $\mathbb{S}^{d-1}$ is the sphere of $B_{\mathbb{R}^d}$ and $\mu_{\mathbb{S}^{d-1}}$ is the normalized surface measure on $\mathbb{S}^{d-1}$.

We modify the approach presented above. We consider again the Taylor expansion (5). This time, we choose the points $\mathcal{X} = \{x^j \in [0,1]^d : j = 1,\dots,m_{\mathcal{X}}\}$ generated at random on $\mathbb{S}^{d-1}$ with respect to $\mu_{\mathbb{S}^{d-1}}$. The matrix Φ is generated as before and we obtain (6) again.

However, the matrix X has a different structure determined by the form of A, namely $X = a^T \mathcal{G}^T$, where $\mathcal{G} = (g'(a \cdot x^1), \dots, g'(a \cdot x^{m_{\mathcal{X}}}))^T$. Let us observe that X and ΦX are now matrices with rank one. The assumptions (12) and (13) combined with the usual Hoeffding's inequality imply immediately that there exists at least one $j \in \{1,\dots,m_{\mathcal{X}}\}$ such that $|g'(a \cdot x^j)|$ is larger then $\sqrt{\alpha(1-s)}$, $0 < s < 1$ with high probability (depending on $m_{\mathcal{X}}, s, \alpha$ and C_2).

Let us now describe, how we use the techniques of compressed sensing to construct an approximation $\hat{a}$ of a. Each column of X has the form $X_j = g'(a \cdot x^j)a^T$ and for this (compressible) vector the theory of compressed sensing implies that if Φ was drawn at random as described above, an approximation $\hat{X}_j$ of X_j may be obtained through an ℓ_1 minimization problem with the error

$$\|X_j - \hat{X}_j\|_{\ell_2^d} \lesssim \left(\frac{m_\Phi}{\log(d/m_\Phi) + 1}\right)^{-\left(\frac{1}{q} - \frac{1}{2}\right)} + \frac{\epsilon}{\sqrt{m_\Phi}} \tag{14}$$

with high probability. Here the constants involved do not depend on m_Φ or d, but depend on C_1, C_2, q and other parameters. We refer to [4] for an extensive track of the constants.

It turns out that the estimate (14) transfers immediately into the estimate of $\|a-\hat a\|_{\ell_2^d}$ for $\hat a = \hat X_j/\|\hat X_j\|_{\ell_2^d}$, i.e. $\hat a$ is a good approximation of a. With these tools at hand we obtain the following result.

Theorem 2. *Let us fix $0<s<1$, $0<q\le 1$, $m_{\mathcal X}\ge 1$ and $1\le m_\Phi\le d$. Under the assumptions and notations fixed above, with high probability[1] there exists a vector $\hat X_j$ obtained by ℓ_1 minimization, such that for $\hat a = \hat X_j/\|\hat X_j\|_{\ell_2^d}$ the function*

$$\hat f(x) = \hat g(\hat a\cdot x), \tag{15}$$

defined by means of

$$\hat g(y) := f(\hat a^T y), \quad y\in(-(1+\bar\epsilon), 1+\bar\epsilon), \tag{16}$$

has the approximation property

$$\|f-\hat f\|_\infty \le 2C_2(1+\bar\epsilon)\frac{\hat\varepsilon}{\sqrt{\alpha(1-s)}-\hat\varepsilon}. \tag{17}$$

where $\hat\varepsilon$ is the right hand side of (14)*.*

Let us summarize the algorithm. We evaluate the function f as described in (7) and construct the matrix Y. Using the techniques of compressed sensing (i.e. with the help of ℓ_1 minimization) we recover the corresponding approximation $\hat X_j$ for each column X_j of X. We fix the j, for which $\|\hat X_j\|_{\ell_2^d}$ is maximal. Then we put $\hat a = \hat X_j/\|\hat X_j\|_{\ell_2^d}$ and define $\hat g$ by (16). The error estimate (17) then follows. Due to the randomness of Φ and corresponding concentration effects, in praxis it would be sufficient to choose the j to be the index of the largest row of Y.

The approximation performances of our learning strategy are basically determined by the constant

$$\alpha = \int_{\mathbb S^{d-1}} |g'(a\cdot x)|^2 d\mu_{\mathbb S^{d-1}}(x).$$

Due to symmetry reasons this quantity does not depend on the particular choice of a. As clarified in [4], under the legitimate assumption that $\|a\|_{\ell_2^d}=1$, the measure $\mu_{\mathbb S^{d-1}}$ determines a push-forward measure $\mu_1 = \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)}(1-y^2)^{\frac{d-3}{2}}\mathcal L^1$ on the unit interval $B_{\mathbb R}$, for which

$$\begin{aligned}\alpha &= \int_{\mathbb S^{d-1}} |g'(a\cdot x)|^2 d\mu_{\mathbb S^{d-1}}(x)\\ &= \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)}\int_{-1}^1 |g'(y)|^2(1-y^2)^{\frac{d-3}{2}}dy.\end{aligned}$$

We observe that α is determined by the interplay between the variation properties of g and the measure μ_1. The most important property of μ_1 is that it concentrates around zero exponentially fast as $d\to\infty$. Hence, the asymptotic behavior of α exclusively depends on the behavior of the function g' in a neighborhood of 0. To illustrate this phenomenon more precisely, we present the following result.

[1] the probability of failure decays exponentially if m_Φ and $m_{\mathcal X}$ are increasing.

Proposition 1. *Let us fix $M\in\mathbb N$ and assume that $g: B_{\mathbb R}\to\mathbb R$ is C^{M+2}-differentiable in an open neighborhood $\mathcal U$ of 0 and $\frac{d^\ell}{dx^\ell}g(0)=0$ for $\ell=1,\dots,M$. Then*

$$\begin{aligned}\alpha(d) &= \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)}\int_{-1}^1 |g'(y)|^2(1-y^2)^{\frac{d-3}{2}}dy\\ &= \mathcal O(d^{-M}),\ \text{for } d\to\infty.\end{aligned}$$

4. GENERAL DIMENSION

We describe briefly the modification necessary if $k>1$, namely if $f(x)=g(Ax)$ and A is a $k\times d$ matrix. We suppose that the rows of A are compressible

$$\left(\sum_{j=1}^d |a_{ij}|^q\right)^{1/q} \le C_1 \tag{18}$$

for every $i\in\{1,\dots,k\}$ and (without loss of generality) that AA^T is the identity operator on $\mathbb R^k$. The regularity condition (12) is replaced by

$$\sup_{|\alpha|\le 2}\|D^\alpha g\|_\infty \le C_2. \tag{19}$$

Instead of the condition (13), we consider the matrix

$$H^f := \int_{\mathbb S^{d-1}} \nabla f(x)\nabla f(x)^T d\mu_{\mathbb S^{d-1}}(x). \tag{20}$$

One may observe that H^f is a positive semi-definite k-rank matrix. For the problem to be well-conditioned we demand that the the singular values of the matrix H^f satisfy

$$\sigma_1(H^f)\ge\cdots\ge\sigma_k(H^f)\ge\alpha>0. \tag{21}$$

Using (5) with the same choice of $\mathcal X$ and Φ, we obtain again (6). The form of X is now $X = A^T\mathcal G^T$, where $\mathcal G = (\nabla g(Ax_1)^T|\dots|\nabla g(Ax_{m_{\mathcal X}})^T)^T$ collects again the derivatives of g.

Using again the techniques of compressed sensing applied to each column X_j of X separately, we obtain

$$\|X-\hat X\|_F \lesssim \sqrt{m_{\mathcal X}}\hat\varepsilon, \tag{22}$$

where

$$\hat\varepsilon = k\left(\frac{m_\Phi}{\log(d/m_\Phi)+1}\right)^{-\left(\frac1q-\frac12\right)} + \frac{k^2\epsilon}{\sqrt{m_\Phi}} \tag{23}$$

and $\|\cdot\|_F$ is the Frobenius norm of a matrix.

Hoeffding's inequality may be generalized to sums of random semidefinite matrices, cf. [1] and [6]. In combination with (21) it follows that $\sigma_k(X)\ge\sqrt{m_{\mathcal X}\alpha(1-s)}$ with high probability. The matrix $\hat A$ (which then serves as an approximation of A) is obtained as a part of the singular value decomposition of $\hat X$. This is then combined with results on stability of singular value decomposition to obtain an estimate for $\|A-\hat A\|_F$.

Finally, the main approximation results looks as follows.

Theorem 3. *Let us fix* $0 < s < 1$, $0 < q \leq 1$, $m_{\mathcal{X}} \geq 1$ *and* $1 \leq m_\Phi \leq d$*. Under the assumptions and notations fixed above, let* $\hat{X}$ *be the* $d \times m_{\mathcal{X}}$ *matrix whose columns are the vectors* $\hat{X}_j$ *obtained by* ℓ_1 *minimization and write the singular value decomposition of its transpose* $\hat{X}^T$ *as*

$$\hat{X}^T = \begin{pmatrix} \hat{U}_1 & \hat{U}_2 \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \hat{V}_1^T \\ \hat{V}_2^T \end{pmatrix},$$

where $\hat{\Sigma}_1$ *contains the largest* k *singular values. Then with high probability the matrix* $\hat{A} = \hat{V}_1^T$ *satisfies that the function*

$$\hat{f}(x) = \hat{g}(\hat{A}x), \tag{24}$$

defined by means of

$$\hat{g}(y) := f(\hat{A}^T y), \quad y \in B_{\mathbb{R}^k}(1 + \bar{\epsilon}), \tag{25}$$

has the approximation property

$$\|f - \hat{f}\|_\infty \leq 2C_2\sqrt{k}(1 + \bar{\epsilon})\frac{\hat{\varepsilon}}{\sqrt{\alpha(1-s)} - \hat{\varepsilon}}, \tag{26}$$

where $\hat{\varepsilon}$ *is as in* (23)*.*

The discussion on tractability can proceed exactly as in the case $k = 1$ with the push-forward measure $\mu_k = \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)}(1 - \|y\|_{\ell_2^k}^2)^{\frac{d-2-k}{2}}\mathcal{L}^k$ of $\mu_{\mathbb{S}^{d-1}}$ on the unit ball $B_{\mathbb{R}^k}$ instead of μ_1.

5. REFERENCES

[1] R. Ahlswede and A. Winter. Strong converse for indentification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, 2002.

[2] A. Cohen, I. Daubechies, R. A. DeVore, G. Kerkyacharian, and D. Picard. Capturing ridge functions in high dimensions from point queries. *preprint*, 2010.

[3] R. A. DeVore, G. Petrova, and P. Wojtaszcyzk. Approximation of functions of few variables in high dimensions. *Constructive Approximation*, 33(1):125–143, 2011.

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

[5] K. Schnass and J. Vybíral. Compressed learning of high-dimensional sparse functions. ICASSP. 2011.

[6] J Tropp. User-friendly tail bounds for matrix martingales. *arXiv:math.PR 1004.4389*, 2010.

[7] P Wojtaszczyk. Complexity of approximation of functions of few variables in high dimensions. *preprint*, 2010.