Learning Functions of Few Arbitrary Linear Parameters in High Dimensions

Jan Vybíral

Austrian Academy of Sciences
RICAM, Linz, Austria

Banff, Canada, March 2011

joint work with Massimo Fornasier and Karin Schnass (RICAM)

supported by START-award of FWF, Austria:
“Sparse Approximation and Optimization in High Dimensions”

Outline

- Introduction
  - Approximation of functions of many variables
  - Non-tractability results
  - Special structure, finite-order weights, ridge functions
  - State of the art

Outline

- Introduction
  - Approximation of functions of many variables
  - Non-tractability results
  - Special structure, finite-order weights, ridge functions
  - State of the art
- Algorithm
  - Numerical evaluation of directional derivatives
  - Points and directions chosen at random
  - Active coordinates (. . . concentration of measure . . .)
  - $k = 1$ (. . . compressed sensing . . .)
  - General case (. . . stability of SVD . . .)

Introduction

Let $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ be a function of many ($d \gg 1$) variables

We want to approximate f uniformly using only (a small number of) function values of f

Introduction

Let $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ be a function of many ($d \gg 1$) variables

We want to approximate f uniformly using only (a small number of) function values of f

The problem is known to be *intractable* (Novak & Woźniakowski, 2009) even for C^∞ functions

The number of sampling points must grow exponentially in d...

Let

$$\mathcal{F}_d := \{f : [0,1]^d \to \mathbb{R}, \|D^\alpha f\|_\infty \le 1, \alpha \in \mathbb{N}_0^d\}$$

Let

$$\mathcal{F}_d := \{f : [0,1]^d \to \mathbb{R}, \|D^\alpha f\|_\infty \leq 1, \alpha \in \mathbb{N}_0^d\}$$

Sampling operator $S_n = \phi \circ N$
Information map $N : \mathcal{F}_d \to \mathbb{R}^n$, $N(f) = (f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n$
Continuous recovery map $\phi : \mathbb{R}^n \to L_\infty([0,1]^d)$

Let

$$\mathcal{F}_d := \{f : [0,1]^d \to \mathbb{R}, \|D^\alpha f\|_\infty \le 1, \alpha \in \mathbb{N}_0^d\}$$

Sampling operator $S_n = \phi \circ N$
Information map $N : \mathcal{F}_d \to \mathbb{R}^n$, $N(f) = (f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n$
Continuous recovery map $\phi : \mathbb{R}^n \to L_\infty([0,1]^d)$
Approximation error

$$e(S_n) := \sup_{f \in \mathcal{F}_d} \|f - S_n(f)\|_\infty$$

Sampling numbers

$$e(n,d) := \inf_{S_n} e(S_n)$$

Let

$$\mathcal{F}_d := \{f : [0,1]^d \to \mathbb{R}, \|D^\alpha f\|_\infty \leq 1, \alpha \in \mathbb{N}_0^d\}$$

Sampling operator $S_n = \phi \circ N$
Information map $N : \mathcal{F}_d \to \mathbb{R}^n$, $N(f) = (f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n$
Continuous recovery map $\phi : \mathbb{R}^n \to L_\infty([0,1]^d)$
Approximation error

$$e(S_n) := \sup_{f \in \mathcal{F}_d} \|f - S_n(f)\|_\infty$$

Sampling numbers

$$e(n,d) := \inf_{S_n} e(S_n)$$

Novak, Woźniakowski (2009): $e(n,d) = 1$ for all $n \leq 2^{\lfloor d/2 \rfloor} - 1$

Conclusion: High smoothness does not help!

Conclusion: High smoothness does not help!

Way out: Inner structure of functions like

- *finite order Sobolev spaces*
- *partially separable functions*
- *k-ridge functions*

$$f(x) = g(Ax), \quad g : \mathbb{R}^k \to \mathbb{R}, \quad A \in \mathbb{R}^{k \times d}, \quad k \ll d$$

Special cases:

A is a projection, i.e.

$$f(x) = f(x_1, \ldots, x_d) = g(x_{i_1}, \ldots, x_{i_k})$$

The *active coordinates* $i_1, \ldots, i_k$ are unknown

Special cases:

A is a projection, i.e.

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

The *active coordinates* $i_1, \dots, i_k$ are unknown

$k = 1$

$$f(x) = g(a \cdot x), \quad a \in \mathbb{R}^d$$

Known results:

Unknown coordinates:

R. DeVore, G. Petrova, P. Wojtaszczyk: *Approximation of functions of few variables in high dimensions*

P. Wojtaszczyk: *Complexity of Approximation of Functions of Few Variables in High Dimensions*

Deterministic algorithms, $C(k)(L+1)^k \log d$ points (adaptively or non-adaptively chosen), uniform approximation of the order $1/L$

Known results:

Unknown coordinates:

R. DeVore, G. Petrova, P. Wojtaszczyk: *Approximation of functions of few variables in high dimensions*

P. Wojtaszczyk: *Complexity of Approximation of Functions of Few Variables in High Dimensions*

Deterministic algorithms, $C(k)(L+1)^k \log d$ points (adaptively or non-adaptively chosen), uniform approximation of the order $1/L$

$\mathbf{k} = \mathbf{1}$:

A. Cohen, I. Daubechies, R. DeVore, G. Kerkyacharian, D. Picard, *Capturing ridge functions in high dimensions from point queries*

$g \in C^s([0,1])$, $1 < s$, $\|g\|_{C^s} \leq M_0$, $\|a\|_{\ell_q^d} \leq M_1$

$$\|f - \hat{f}\|_{C(\Omega)} \leq CM_0 \left\{ L^{-s} + M_1 \left(\frac{1 + \log(d/L)}{L} \right)^{1/q-1} \right\}$$

using $3L + 2$ sampling points

Active coordinates

We assume, that

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

i.e.

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

where $f : [0,1]^d \to \mathbb{R}$ and $g : [0,1]^k \to \mathbb{R}$

We want to identify the active coordinates $i_1, \dots, i_k$. Then one can apply any usual k-dimensional approximation method...

Our algorithm chooses the sampling points at random, due to the *concentration of measure* effects, we get the right result with overwhelming probability.

We rely on numerical approximation of $\frac{\partial f}{\partial \varphi}$

$$\nabla g(Ax)^T A\varphi = \frac{\partial f}{\partial \varphi}(x) \qquad (*)$$
$$= \frac{f(x+\epsilon\varphi) - f(x)}{\epsilon} - \frac{\epsilon}{2}[\varphi^T \nabla^2 f(\zeta)\varphi]$$

We rely on numerical approximation of $\frac{\partial f}{\partial \varphi}$

$$\nabla g(Ax)^T A\varphi = \frac{\partial f}{\partial \varphi}(x) \qquad (*)$$
$$= \frac{f(x+\epsilon\varphi) - f(x)}{\epsilon} - \frac{\epsilon}{2}[\varphi^T \nabla^2 f(\zeta)\varphi]$$

$\mathcal{X} = \{x^j \in [0,1]^d : j = 1, \dots, m_X\}$ drawn uniformly at random with respect to the Lebesgue measure

$\Phi = \{\varphi^j \in \mathbb{R}^d, j = 1, \dots, m_\Phi\}$, where

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

for every $j \in \{1, \dots, m_\Phi\}$ and every $\ell \in \{1, \dots, d\}$

$\Phi \ldots m_\Phi \times d$ matrix, $X \ldots d \times m_X$ matrix with i-th row

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

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

$\Phi \ldots m_\Phi \times d$ matrix, $X \ldots d \times m_X$ matrix with i-th row

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

for $i \in I$ and all other rows equal to zero
The $m_X \times m_\Phi$ instances of $(*)$ in matrix notation as

$$\Phi X = Y + \mathcal{E} \qquad (**)$$

Y and $\mathcal{E}$ are $m_\Phi \times m_X$ matrices defined by

$$y_{ij} = \frac{f(x^j + \epsilon\varphi^i) - f(x^j)}{\epsilon},$$
$$\varepsilon_{ij} = -\frac{\epsilon}{2}[(\varphi^i)^T \nabla^2 f(\zeta_{ij})\varphi^i],$$

The algorithm is based on the identity

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

The algorithm is based on the identity

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

In expectation:
$\Phi^T \Phi \approx I_d : \mathbb{R}^d \to \mathbb{R}^d$
$\Phi^T \Phi X \approx X$ and
$\Phi^T \mathcal{E}$ is small $\implies \Phi^T Y \approx X$

The algorithm is based on the identity

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

In expectation:
$\Phi^T \Phi \approx I_d : \mathbb{R}^d \to \mathbb{R}^d$
$\Phi^T \Phi X \approx X$ and
$\Phi^T \mathcal{E}$ is small $\implies \Phi^T Y \approx X$

We select the k largest rows of $\Phi^T Y$ and estimate the probability, that their indices coincide with the indices of the non-zero rows of X.

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function of k active coordinates that is defined and twice continuously differentiable on a small neighbourhood 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

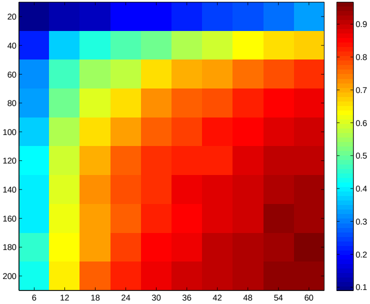
$$\mathcal{O}(k(L + \log k)(L + \log d))$$

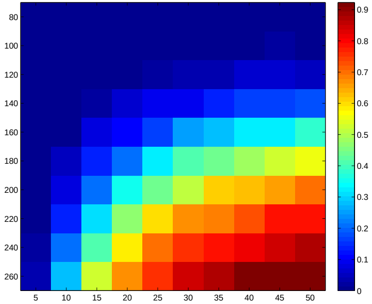
samples of f.

The constants involved in the $\mathcal{O}$ notation depend on smoothness properties of g, namely on

$$\frac{\max_{j=1,\dots,k} \|\partial_{i_j} g\|_\infty}{\min_{j=1,\dots,k} \|\partial_{i_j} g\|_1}$$

$d = 1000$





$$\max(1 - 5\sqrt{(x_3 - 1/2)^2 + (x_4 - 1/2)^2}, 0)^3$$

$$\sin\left(6\pi \sum_{i=21}^{40} x_i\right) + \sum_{i=21}^{40} \sin(6\pi x_i) + 5(x_i - 1/2)^2$$

$k = 1$

Let $f(x) = g(a \cdot x), f : B_{\mathbb{R}^d} \to \mathbb{R}$, where $a \in \mathbb{R}^d$

$\|a\|_2 = 1$ and $\|a\|_q \leq C_1$, $0 < q \leq 1$, $\max_{0 \leq \alpha \leq 2} \|D^\alpha g\|_\infty \leq C_2$

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

$k = 1$

Let $f(x) = g(a \cdot x), f : B_{\mathbb{R}^d} \to \mathbb{R}$, where $a \in \mathbb{R}^d$

$\|a\|_2 = 1$ and $\|a\|_q \le C_1$, $0 < q \le 1$, $\max_{0 \le \alpha \le 2} \|D^\alpha g\|_\infty \le C_2$

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

We consider again the Taylor expansion (*)

We choose the points $\mathcal{X} = \{x^j \in [0,1]^d : j = 1, \ldots, m_{\mathcal{X}}\}$ 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 (**) again.

$X = a^T \mathcal{G}^T$, where $\mathcal{G} = (g'(a \cdot x^1), \ldots, g'(a \cdot x^{m_{\mathcal{X}}}))^T$

X and ΦX are rank one matrices

$X = a^T \mathcal{G}^T$, where $\mathcal{G} = (g'(a \cdot x^1), \dots, g'(a \cdot x^{m_{\mathcal{X}}}))^T$

X and ΦX are rank one matrices

Hoeffding's inequality:
$\exists j \in \{1, \dots, m_{\mathcal{X}}\} : |g'(a \cdot x^j)| \geq \sqrt{\alpha(1-s)}$, $0 < s < 1$
with high probability (depending on $m_{\mathcal{X}}, s, \alpha$ and C_2).
X_j - the j-th column of X - is equal to $g'(a \cdot x^j) a^T$

$X = a^T \mathcal{G}^T$, where $\mathcal{G} = (g'(a \cdot x^1), \ldots, g'(a \cdot x^{m_\mathcal{X}}))^T$

X and ΦX are rank one matrices

Hoeffding's inequality:
$\exists j \in \{1, \ldots, m_\mathcal{X}\} : |g'(a \cdot x^j)| \geq \sqrt{\alpha(1-s)}$, $0 < s < 1$
with high probability (depending on $m_\mathcal{X}, s, \alpha$ and C_2).
X_j - the j-th column of X - is equal to $g'(a \cdot x^j) a^T$

Due to the construction of Φ, compressed sensing gives the approximation $\hat{X}_j$

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

...transfers 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.

Theorem

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, with high probability 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

$$\hat{f}(x) = \hat{g}(\hat{a} \cdot x), \tag{1}$$

defined by means of

$$\hat{g}(y) := f(\hat{a}^T y), \quad y \in (-(1+\bar{\epsilon}), 1+\bar{\epsilon}), \tag{2}$$

has the approximation property

$$\|f - \hat{f}\|_\infty \leq 2C_2(1+\bar{\epsilon}) \frac{\hat{\varepsilon}}{\sqrt{\alpha(1-s)} - \hat{\varepsilon}}. \tag{3}$$

where $\hat{\varepsilon}$ is the right hand side of ($\clubsuit$).

Key role is played by

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

Due to symmetry ... independent on a

Push-forward measure μ_1 on $[-1,1]$

$$\begin{aligned} \alpha &= \int_{-1}^{1} |g'(y)|^2 d\mu_1(y) \\ &= \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \int_{-1}^{1} |g'(y)|^2 (1-y^2)^{\frac{d-3}{2}} dy \end{aligned}$$

μ_1 concentrates around zero exponentially fast as $d \to \infty$

Proposition

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

$$\alpha(d) = \mathcal{O}(d^{-M}), \text{ for } d \to \infty.$$

$k \gg 1$

$f(x) = g(Ax)$, A is a $k \times d$ matrix

$k \gg 1$

$f(x) = g(Ax)$, A is a $k \times d$ matrix

Rows of A are compressible: $\max_i \|a_i\|_q \leq C_1$
AA^T is the identity operator on $\mathbb{R}^k$

The regularity condition: $\sup_{|\alpha| \leq 2} \|D^\alpha g\|_\infty \leq C_2$

$k \gg 1$

$f(x) = g(Ax)$, A is a $k \times d$ matrix

Rows of A are compressible: $\max_i \|a_i\|_q \leq C_1$
AA^T is the identity operator on $\mathbb{R}^k$

The regularity condition: $\sup_{|\alpha|\leq 2} \|D^\alpha g\|_\infty \leq C_2$

The matrix $H^f := \int_{\mathbb{S}^{d-1}} \nabla f(x) \nabla f(x)^T d\mu_{\mathbb{S}^{d-1}}(x)$ is a positive semi-definite k-rank matrix

We assume, that the singular values of the matrix H^f satisfy

$$\sigma_1(H^f) \geq \cdots \geq \sigma_k(H^f) \geq \alpha > 0.$$

$X = A^T \mathcal{G}^T$, where $\mathcal{G} = (\nabla g(Ax_1)^T | \dots | \nabla g(Ax_{m_{\mathcal{X}}})^T)^T$

$X = A^T \mathcal{G}^T$, where $\mathcal{G} = (\nabla g(Ax_1)^T | \dots | \nabla g(Ax_{m_{\mathcal{X}}})^T)^T$

Compressed sensing applied to each column X_j of X separately:

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

where

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

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

Theorem

Let $0 < s < 1$, $0 < q \leq 1$, $m_{\mathcal{X}} \geq 1$ and $1 \leq m_\Phi \leq d$. Under the 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 $\hat{f}(x) = \hat{g}(\hat{A}x)$ defined by means of

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

has the approximation property

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

References:

M. Fornasier, K. Schnass and J. Vybíral, *Learning functions of few arbitrary linear parameters in high dimensions*, submitted

K. Schnass and J. Vybíral, *Compressed Learning of High-Dimensional Sparse Functions*, to appear in Proc. ICASSP 2011

http://people.ricam.oeaw.ac.at/j.vybiral/