What is Information Based Complexity?

FSDONA

July 2016, Prague

Jan Vybíral

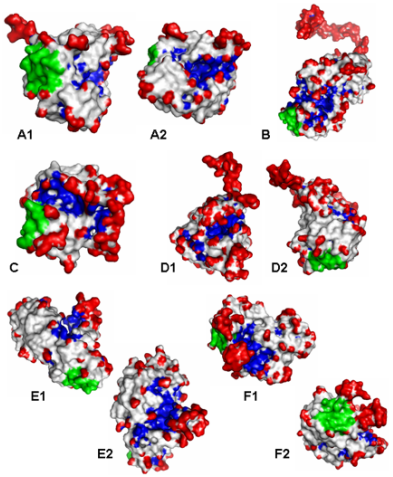
(Charles University, Prague, Czech Republic)

Introduction

- Multivariate functions
- Curse of dimension
- Sampling numbers
- IBC + Tractability
- Structural assumptions
- Ridge functions

Many real-life applications nowadays feature functions of d variables with $d \gg 1$, i.e. $d \approx 10^3$

Many real-life applications nowadays feature functions of d variables with $d \gg 1$, i.e. $d \approx 10^3$

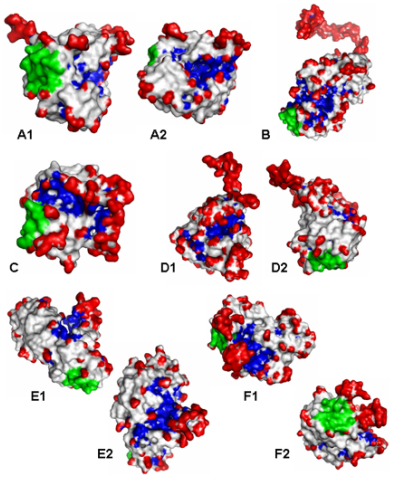


Searching for metastable states of proteins:

Describe the shape of the protein by hundreds of parameters

Find the local minima of the energy as a function of these parameters

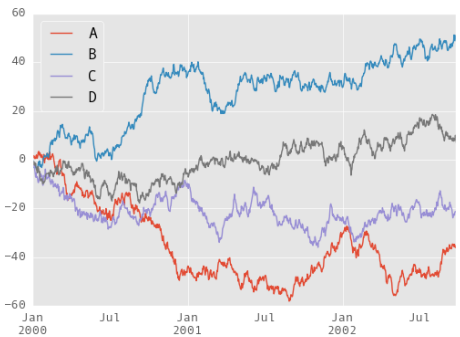
Many real-life applications nowadays feature functions of d variables with $d \gg 1$, i.e. $d \approx 10^3$



Searching for metastable states of proteins:

Describe the shape of the protein by hundreds of parameters

Find the local minima of the energy as a function of these parameters



Evaluating global parameters of financial markets

Classical problems of approximation theory

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

We may want to (with n small)

- approximate $\int_\Omega f$ using only function values $f(x_1), \ldots, f(x_n)$
- approximate f using only function values $f(x_1), \ldots, f(x_n)$
- approximate $\int_\Omega f$ using only values of linear functionals $L_1(f), \ldots, L_n(f)$
- approximate f using only values of linear functionals $L_1(f), \ldots, L_n(f)$

  $\vdots$

Questions

- Which functions (assumptions on f)?
- How to measure the error?
- Decay of the error with growing number of sampling points?
- Algorithms and optimality?
- Dependence on d?

. . . the nature of these problems is well understood for small d's. . .

Curse of dimension

Many classical problems suffer from exponential dependence of the results on d!

Example: Uniform approximation of smooth functions

- Let $f \in \mathcal{F}_d := \{f : [0,1]^d \to \mathbb{R}, \|D^\alpha f\|_\infty \leq 1, \alpha \in \mathbb{N}_0^d\}$
- Uniform approximation - the error is measured in $\|\cdot\|_\infty$
- uniform approximation of infinitely differentiable functions on $\Omega = [0,1]^d$
- initial error - approximation by zero function.

Curse of dimension

Many classical problems suffer from exponential dependence of the results on d!

Example: Uniform approximation of smooth functions

- Let $f \in \mathcal{F}_d := \{f : [0,1]^d \to \mathbb{R}, \|D^\alpha f\|_\infty \leq 1, \alpha \in \mathbb{N}_0^d\}$
- Uniform approximation - the error is measured in $\|\cdot\|_\infty$
- uniform approximation of infinitely differentiable functions on $\Omega = [0,1]^d$
- initial error - approximation by zero function.

Novak, Woźniakowski (2009): Initial error is the same as error of uniform approximation for $n \leq 2^{\lfloor d/2 \rfloor} - 1$

. . . curse of dimension!

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

Smoothness does not help!

Function spaces

In the classical setting of small d we usually consider

- $\mathcal{C}^k(\Omega)$ - k-times continuously differentiable functions
- $W_p^s(\Omega)$ - Sobolev spaces with smoothness s and integrability p
- $B_{p,q}^s(\Omega)$ - Besov spaces with smoothness s, integrability p and summability q
- $F_{p,q}^s(\Omega)$ - Triebel-Lizorkin spaces with smoothness s, integrability p and summability q
- Other classes adapted to problems under consideration (weighted spaces, boundary values, anisotropic spaces, ...)

Sampling numbers

. . . describe approximation of a function from limited number of function values

$\mathcal{F}_d$ - a class of d-variate functions on $\Omega_d \subset \mathbb{R}^d, \mathcal{F}_d \subset C(\Omega_d)$
(the continuity of f makes function values meaningful)

Sampling numbers

. . . describe approximation of a function from limited number of function values

$\mathcal{F}_d$ - a class of d-variate functions on $\Omega_d \subset \mathbb{R}^d, \mathcal{F}_d \subset C(\Omega_d)$
(the continuity of f makes function values meaningful)

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:** (linear or not) $\phi : \mathbb{R}^n \to L_\infty(\Omega_d)$

Sampling operator: $S_n = \phi \circ N : \mathcal{F}_d \to L_\infty(\Omega_d)$

Modifications:

- Linear information: $L_1(f), \ldots, L_n(f)$
- Adaptivity: x_j may depend on $x_1, \ldots, x_{j-1}$ and on $f(x_1), \ldots, f(x_{j-1})$
- Numerical integration: $S_n : \mathcal{F}_d \to \mathbb{R}$

Sampling numbers

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

. . . consider the worst function for your algorithm. . .

Sampling numbers: $g_{n,d}(\mathcal{F}_d, L_\infty) := \inf_{S_n} e(S_n)$

. . . the error of the "best algorithm" when using only n function values. . .

Its inverse function: $n(\varepsilon, d) = \min\{n \in \mathbb{N} : g_{n,d}(\mathcal{F}_d, L_\infty) \leq \varepsilon\}$
. . . minimal number of sampling points needed to achieve an approximation of the error at most $\varepsilon > 0$. . .

Sampling numbers: general setting

If $\mathcal{F}_d$ is a unit ball of a function space $X(\Omega_d) \subset C(\Omega_d)$ and the error is measured in $Y(\Omega_d)$, then

$$g_{n,d}(\mathcal{F}_d, Y(\Omega_d)) = \inf_{S_n} \sup_{\|f\|_{X(\Omega_d)} \le 1} \|f - S_n(f)\|_{Y(\Omega_d)}$$
$$= \inf_{S_n} \|I - S_n : \mathcal{L}(X(\Omega_d), Y(\Omega_d))\|$$

Well known for many classical function spaces, like Sobolev spaces, Besov spaces, Triebel-Lizorkin spaces, etc.

Typical decay: $n^{-s/d}$

Birman, Solomyak, Temlyakov, Kudryavtsev, Kashin, DeVore, Maiorov, Kruglyak, Heinrich, Novak, Triebel, and many others ...

Novak & Woźniakowski: Tractability of Multivariate Problems, Volumes II & III

Theorem (V.'07): Let $\Omega_d = [0,1]^d$ and

$$s_1 > \frac{d}{p_1} \quad \text{and} \quad s_1 - d\Big(\frac{1}{p_1} - \frac{1}{p_2}\Big)_+ > s_2 > 0.$$

Then

$$g_{n,d}(B^{s_1}_{p_1,q_1}(\Omega_d), B^{s_2}_{p_2,q_2}(\Omega_d)) \approx n^{-\frac{s_1-s_2}{d}+(\frac{1}{p_1}-\frac{1}{p_2})_+}.$$

If

$$s_1 > \frac{d}{p_1} \quad \text{and} \quad s_2 < 0,$$

then

$$g_{n,d}(B^{s_1}_{p_1,q_1}(\Omega_d), B^{s_2}_{p_2,q_2}(\Omega_d)) \approx n^{-\frac{s_1}{d}+(\frac{s_2}{d}-\frac{1}{p_2}+\frac{1}{p_1})_+}.$$

Information Based Complexity a.k.a. IBC

Studies

- **algorithms**
- **computational complexity**

for the continuous problems

- **very-high-dimensional integration**
- **path integration**
- **partial differential equations**
- **systems of ordinary differential equations**
- **nonlinear equations**
- **integral equations**
- **fixed points**

Tractability

Different notions to describe the growth of $n(\varepsilon, d)$

- **Polynomial tractability:** $\exists c, p, q > 0$

$$n(\varepsilon, d) \leq c\varepsilon^{-p} d^{q} \quad \text{for all} \quad 0 < \varepsilon < 1, d \in \mathbb{N}$$

- **Quasi-polynomial tractability:** $\exists C, t > 0$

$$n(\varepsilon, d) \leq C \exp(t(1 + \ln(\varepsilon^{-1}))(1 + \ln d))$$

- **Weak tractability:**

$$\lim_{\varepsilon^{-1}+d \to \infty} \frac{\ln n(\varepsilon, d)}{\varepsilon^{-1} + d} = 0$$

- **Curse of dimension:** $\exists c, \gamma > 0$

$$n(\varepsilon, d) \geq c(1 + \gamma)^{d} \quad \text{for all } 0 < \varepsilon < \varepsilon_0 \quad \text{and inf. many } d\text{'s}$$

Theorem (Novak, Woźniakowski '09)

Uniform approximation on the cube $[0,1]^d$ of functions from the class

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

suffers the *curse of dimension*:

$$g_{n,d}(F_d^1, L_\infty) = 1 \quad \text{for} \quad n \leq 2^{\lfloor d/2 \rfloor} - 1$$

or, equivalently,

$$n(\varepsilon, d) \geq 2^{\lfloor d/2 \rfloor} - 1 \quad \text{for all} \quad 0 < \varepsilon < 1.$$

Theorem (V. '14): Uniform approximation on the cube $[-1/2,1/2]^d$ of functions from the class

$$F_d^2 = \left\{ f \in C^\infty([-1/2,1/2]^d) : \sup_{k\in\mathbb{N}_0} \sum_{|\beta|=k} \frac{\|D^\beta f\|_\infty}{\beta!} \le 1 \right\}$$

is *weakly tractable*, ie.

$$\lim_{\varepsilon^{-1}+d\to\infty} \frac{\ln n(\varepsilon,d)}{\varepsilon^{-1}+d} = 0.$$

. . . similar results for other norms, domains, . . . based on Taylor's theorem or other "constructive" techniques

. . . A. Hinrichs, E. Novak, M. Ullrich, H. Woźniakowski, P. Kritzer, F. Pillichshammer, J. Dick, . . .

Discrepancy

. . . uniform distribution of points in the unit cube . . .
Sequence of points: $x_1, \ldots, x_n \in [0,1]^d$
Rectangle: $Q(t,t') = [t_1, t_1'] \times \cdots \times [t_d, t_d'] \subset [0,1]^d$

We expect (on the average)

$$\#\{i : x_i \in Q(t,t')\} = \sum_{i=1}^{n} \chi_{Q(t,t')}(x_i)$$
$$\sim n \cdot (t_1' - t_1) \cdot \cdots \cdot (t_d' - t_d) = n \,\mathrm{vol}(Q(t,t'))$$

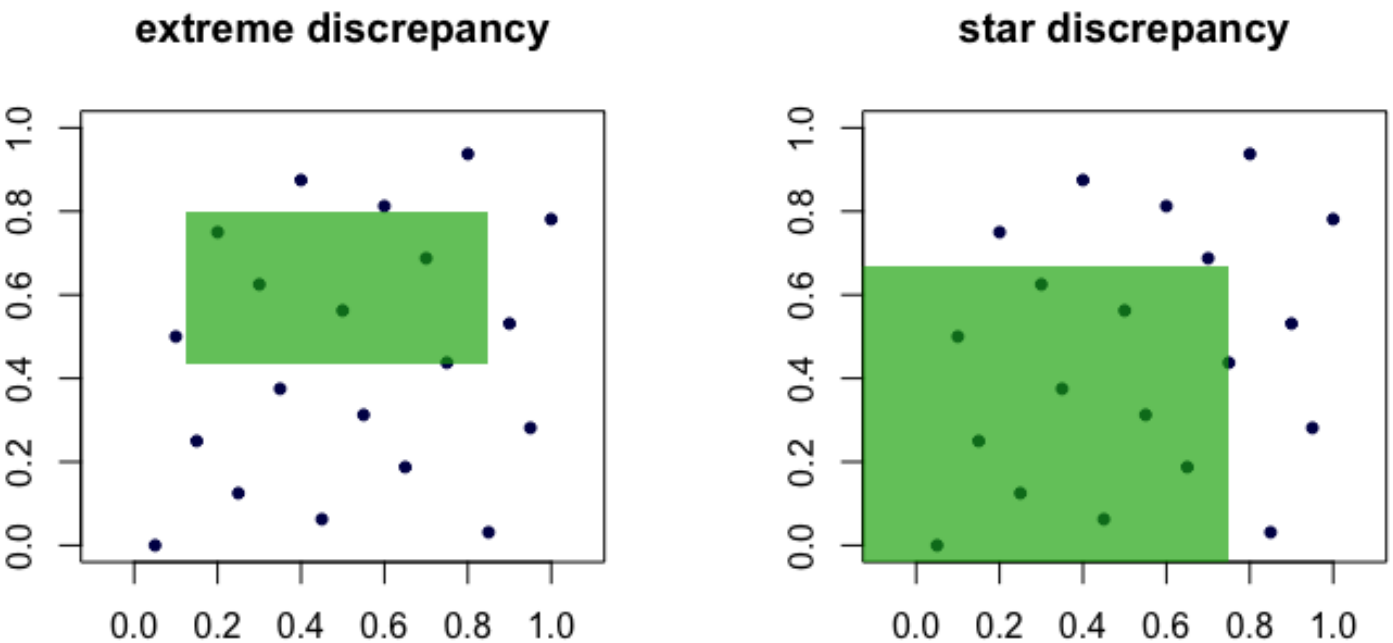
Discrepancy:

$$D_\infty(x_1, \ldots, x_n) = \sup_{Q \subset [0,1]^d} \left| \mathrm{vol}(Q) - \frac{1}{n} \sum_{i=1}^{n} \chi_Q(x_i) \right|$$

Star discrepancy

Star discrepancy ($t = 0$):

$$D_\infty^*(x_1, \ldots, x_n) = \sup_{t' \in [0,1]^d} \left| \mathrm{vol}(Q(0, t')) - \frac{1}{n} \sum_{i=1}^{n} \chi_{[0,t')}(x_i) \right|$$



Star discrepancy

Looking for a sequence with low discrepancy:

$$n(\varepsilon, d) = \min\{n \in \mathbb{N} : \exists x_1, \ldots, x_n : D_\infty^*(x_1, \ldots, x_n) \leq \varepsilon\}$$

Heinrich, Novak, Wasilkowski, Woźniakowski ('01):

$$n(\varepsilon, d) \leq Cd\varepsilon^{-2}$$

Hinrichs ('04):

$$n(\varepsilon, d) \geq cd\varepsilon^{-1}, \quad 0 < \varepsilon < \varepsilon_0$$

Dominating mixed smoothness

Function spaces defined through mixed derivatives
Usual scales: Sobolev, Besov, Triebel-Lizorkin, ...

For example, $f \in W_p^{1,\mathrm{mix}}(\mathbb{R}^2)$:

$$f \in L_p(\mathbb{R}^2), \quad \frac{\partial f}{\partial x_1} \in L_p(\mathbb{R}^2), \quad \frac{\partial f}{\partial x_2} \in L_p(\mathbb{R}^2)$$

but most importantly

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} \in L_p(\mathbb{R}^2)$$

Shortly: $D^\alpha f \in L_p$ for all $\|\alpha\|_\infty \leq 1$

Dominating mixed smoothness

Numerical integration:

$$e(A_n) = \sup_{f \in \mathcal{F}_d} \left| A_n(f) - \int_{\Omega_d} f \right|$$

$$e(n, \mathcal{F}_d) = \inf_{A_n} e(A_n) = \inf_{A_n} \sup_{f \in \mathcal{F}_d} \left| A_n(f) - \int_{\Omega_d} f \right|$$

Dominating mixed smoothness

Numerical integration:

$$e(A_n) = \sup_{f \in \mathcal{F}_d} \left| A_n(f) - \int_{\Omega_d} f \right|$$

$$e(n, \mathcal{F}_d) = \inf_{A_n} e(A_n) = \inf_{A_n} \sup_{f \in \mathcal{F}_d} \left| A_n(f) - \int_{\Omega_d} f \right|$$

Theorem (Bakhvalov, '71) If $\mathcal{F}_d$ is convex and balanced,

$$e(n, \mathcal{F}_d) = \inf_{x_1, \ldots, x_n \in \Omega_d} \sup_{\substack{f \in \mathcal{F}_d \\ N(f)=0}} \int_{\Omega_d} f$$

Dominating mixed smoothness

Numerical integration:

$$e(A_n) = \sup_{f \in \mathcal{F}_d} \left| A_n(f) - \int_{\Omega_d} f \right|$$

$$e(n, \mathcal{F}_d) = \inf_{A_n} e(A_n) = \inf_{A_n} \sup_{f \in \mathcal{F}_d} \left| A_n(f) - \int_{\Omega_d} f \right|$$

Theorem (Bakhvalov, '71) If $\mathcal{F}_d$ is convex and balanced,

$$e(n, \mathcal{F}_d) = \inf_{x_1, \ldots, x_n \in \Omega_d} \sup_{\substack{f \in \mathcal{F}_d \\ N(f)=0}} \int_{\Omega_d} f$$

$\Omega_d = [0,1]^d, 1 < p < \infty$

$$e(n, \{f : \Omega_d \to \mathbb{R} : \|D^\alpha f\|_\infty \leq 1, \|\alpha\|_1 \leq k\}) \approx n^{-k/d}$$

$$e(n, \{f : \Omega_d \to \mathbb{R} : \|D^\alpha f\|_p \leq 1, \|\alpha\|_\infty \leq k\}) \approx n^{-k} (\log n)^{(d-1)/2}$$

Hlawka-Zaremba-equality

$W_1^{1,\mathrm{mix}}([0,1]^d)$, zero on the boundary:

$$\mathcal{F}_d := \left\{ \left\| \frac{\partial^d f}{\partial x_1 \dots \partial x_d} \right\|_1 \leq 1, f(x) = 0 \text{ if } \exists i : x_i = 1 \right\}$$

Connection between star discrepancy and the error of numerical integration on $\mathcal{F}_d$

$$D_\infty^*(x_1, \dots, x_n) = \sup_{f \in \mathcal{F}_d} \left| \int_{[0,1]^d} f - \frac{1}{n} \sum_{j=1}^{n} f(x_j) \right|$$

Lower bounds in IBC

What to do if $f(x_1) = \cdots = f(x_n) = 0$?
... or ... $L_1(f) = \cdots = L_n(f) = 0$? ... or $N_n(f) = 0$?

Best choice of $\phi(0)$: the "center" of $\{f \in \mathcal{F}_d \ \& \ N_n(f) = 0\}$
and (for $\mathcal{F}_d \subset X$ symmetric and convex)

$$g_{n,d}(\mathcal{F}_d, Y) \geq \inf_{\substack{M \subset\subset X \\ \operatorname{codim} M \leq n}} \sup_{f \in \mathcal{F}_d \cap M} \|f\|_Y$$
$$= c_{n+1}(\mathcal{F}_d, Y)$$

Lower bounds on sections of (convex) bodies?

Gelfand numbers

For $T : X \to Y$ bounded linear operator

$$c_n(T) = \inf_{\substack{M \subset\subset X \\ \operatorname{codim} M < n}} \sup_{\substack{x \in M \\ \|x\|_X \leq 1}} \|Tx\|_Y$$

Carl's inequality (1981):

$$\sup_{1 \leq k \leq n} k^\alpha e_k(T) \leq c_\alpha \sup_{1 \leq k \leq n} k^\alpha s_k(T).$$

Entropy numbers:

$$e_n(T) = \inf\left\{\varepsilon > 0 : T(B_X) \subset \bigcup_{j=1}^{2^{n-1}} (y_j + \varepsilon B_Y)\right\}.$$

Hinrichs, Kolleck, V. (2016): Carl's inequality holds also for quasi-Banach spaces.

New classes of function spaces

Starting point:

Approximation theory in high dimensions is very specific!

New classes of function spaces

Starting point:

Approximation theory in high dimensions is very specific!

- Some problems are just intractable - and will always be:
  curse of dimension
- It is important to look for alternative formulations
- Probabilistic formulations//probabilistic algorithms
  blessing of dimension
- High-dimensional problems call for tailored settings - i.e. function spaces defined not (only) by integrability and smoothness
- Dimensionality reduction: the amount of information is usually much smaller than the dimension would suggest

Ridge functions - definition

Let $g : \mathbb{R} \to \mathbb{R}$ and $a \in \mathbb{R}^d \setminus \{0\}$. *Ridge function* with *ridge profile* g and *ridge vector* a is the function

$$f(x) := g(\langle a, x \rangle).$$

Constant along the hyperplane $a^\perp = \{y \in \mathbb{R}^d : \langle y, a \rangle = 0\}$ and its translates.

More general, if $g : \mathbb{R}^k \to \mathbb{R}$ and $A \in \mathbb{R}^{k \times d}$ with $k \ll d$ then

$$f(x) := g(Ax)$$

is a $k-$*ridge function*.

Ridge functions in mathematics

- Kernels of important transforms (Fourier, Radon)
- Plane waves in PDE's:
  Solutions to
  $$\prod_{i=1}^{r}\Big(b_i\frac{\partial}{\partial x}-a_i\frac{\partial}{\partial y}\Big)F=0$$
  are of the form
  $$F(x,y)=\sum_{i=1}^{r}f_i(a_ix+b_iy).$$
- Ridgelets, curvelets, shearlets,. . . : wavelet-like frames capturing singularities along curves and edges (Candes, Donoho, Kutyniok, Labate, . . .)

Ridge functions in approximation theory

Approximation of a function by functions from the dictionary

$$D_{\text{ridge}} = \{\varrho(\langle k, x\rangle - b) : k \in \mathbb{R}^d, b \in \mathbb{R}\}$$

- Fundamentality
- Greedy algorithms

Lin & Pinkus, Fundamentality of ridge functions, J. Approx. Theory 75 (1993), no. 3, 295–311

Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control Signals Systems 2 (1989), 303–314

Leshno, Lin, Pinkus & Schocken, Multilayer feedforward networks with a nonpolynomial activation function can approximate any function, Neural Networks 6 (1993), 861–867

Ridge functions in statistical learning

- Least squares & ridge regression
- Phase retrieval
- Neural networks
- Single index models
- Support vector machines

Approximation algorithms

1. Buhmann & Pinkus '99
2. Cohen, Daubechies, DeVore, Kerkyacharian, Picard '12
3. Fornasier, Schnass, V. '12
4. Tyagi & Cevher '12, '14
5. Kolleck, V. '15

Ridge functions: Approximation algorithms

$k = 1$: $f(x) = g(\langle a, x \rangle), \quad \|a\|_2 = 1, \quad g$ smooth

Approximation has two parts: approximation of g and of a

Recovery of a - from $\nabla f(x)$:

$$\nabla f(x) = g'(\langle a, x \rangle)a, \nabla f(0) = g'(0)a.$$

After recovering a, the problem becomes essentially one-dimensional and one can use arbitrary sampling method to approximate g.

$g'(0) \neq 0 \ldots g'(0) = 1$

Buhmann & Pinkus '99
Identifying linear combinations of ridge functions

Approximation of functions

$$f(x) = \sum_{i=1}^{m} g_i(\langle a_i, x \rangle), \quad x \in \mathbb{R}^d$$

- $g_i \in C^{2m-1}(\mathbb{R})$, $i = 1, \ldots, m$;
- $g_i^{(2m-1)}(0) \neq 0$, $i = 1, \ldots, m$;

$$(D_u^{2m-1-k} D_v^k f)(0) = \sum_{i=1}^{m} (\langle u, a_i \rangle)^{2m-1-k} (\langle v, a_i \rangle)^k g_i^{(2m-1)}(0)$$

for $k = 0, \ldots, 2m-1$ and $v_1, \ldots, v_d \in \mathbb{R}^d$ and solving this system of equations.

A.Cohen, I.Daubechies, R.DeVore, G.Kerkyacharian, D.Picard, *Capturing ridge functions in high dimensions from point queries*, Constr. Approx. (2012)

- $\boldsymbol{k = 1 : f(x) = g(\langle a, x \rangle)}$
- $f : [0,1]^d \to \mathbb{R}$
- $g \in C^s([0,1])$, $1 < s$
- $\|g\|_{C^s} \leq M_0$
- $\boldsymbol{\|a\|_{\ell_q^d} \leq M_1, 0 < q \leq 1}$
- $\boldsymbol{a \geq 0}$

Then

$$\|f - \hat{f}\|_\infty \leq C M_0 \left\{ L^{-s} + M_1 \left(\frac{1 + \log(d/L)}{L} \right)^{1/q-1} \right\}$$

using $3L + 2$ sampling points

- First sampling along the diagonal $\frac{i}{L}\mathbf{1} = \frac{i}{L}(1,\dots,1), i = 0,\dots,L$:

$$f\left(\frac{i}{L}\mathbf{1}\right) = g\left(\left\langle \frac{i}{L}\mathbf{1}, a\right\rangle\right) = g(i\|a\|_1/L)$$

- Recovery of g on a grid of $[0, \|a\|_1]$
- Finding i_0 with largest $g((i_0+1)\|a\|_1/L) - g(i_0\|a\|_1/L)$
- Approximating $D_{\varphi_j} f(i_0/L \cdot \mathbf{1}) = g'(i_0\|a\|_1/L)\langle a, \varphi_j\rangle$ by first order differences
- Then recovery of a from $\langle a, \varphi_1\rangle, \dots, \langle a, \varphi_m\rangle$ by methods of compressed sensing (CS)

Shortly on Compressed Sensing

Recover arbitrary $x \in \mathbb{R}^d$ from linear information about x:

$$x = \sum_{j=1}^{d} \langle x, \psi_j \rangle \psi_j$$

If $x \in \mathbb{R}^d$ is sparse (with small number s of non-zero coefficients on unknown positions), take ψ_j's independently at random and recover x *exactly* by LASSO (Tibshirani, 1996) from the linear information $y_j = \langle x, \psi_j \rangle$:

$$\arg\min_{\omega \in \mathbb{R}^d} \|\omega\|_1, \quad \text{s.t. } \langle \omega, \psi_j \rangle = y_j, \quad j = 1, \ldots, m,$$

where $m \approx s \log(d)$.

ℓ_1-minimization is robust (w.r.t. noise) and stable (defects of sparsity)

M. Fornasier, K. Schnass, J. V., *Learning functions of few arbitrary linear parameters in high dimensions*, Found. Comput. Math. (2012)

- $f : B(0,1) \to \mathbb{R}$
- $\|a\|_2 = 1$
- $0 < q \leq 1, \|a\|_q \leq c$
- $g \in C^2[-1,1]$

Put

$$y_j := \frac{f(h\varphi_j) - f(0)}{h} \approx g'(0)\langle a, \varphi_j \rangle, \quad j = 1, \dots, m_\Phi, \quad m_\Phi \leq d,$$

where $h > 0$ is small, and

$$\varphi_{j,k} = \pm \frac{1}{\sqrt{m_\Phi}}, \quad k = 1, \dots, d$$

y_j are scalar products $\langle a, \varphi_j \rangle$ corrupted by deterministic noise

$$\tilde{a} = \arg\min_{z \in \mathbb{R}^d} \|z\|_1, \quad \text{s.t. } \langle \varphi_j, z \rangle = y_j, \quad j = 1, \dots, m_\Phi.$$

$\hat{a} = \tilde{a}/\|\tilde{a}\|_2$ - approximation of a

$\hat{g}$ is obtained by sampling f along $\hat{a}$: $\hat{g}(y) := f(\hat{a} \cdot t), \ t \in (-1, 1)$.

Then

$$\hat{f}(x) := \hat{g}(\langle \hat{a}, x \rangle),$$

has the approximation property

$$\|f - \hat{f}\|_\infty \le C \Big[\left(\frac{m_\Phi}{\log(d/m_\Phi) + 1} \right)^{-\left(\frac{1}{q} - \frac{1}{2}\right)} + \frac{h}{\sqrt{m_\Phi}} \Big].$$

Active coordinates:

R. DeVore, G. Petrova, P. Wojtaszczyk, *Approximation of functions of few variables in high dimensions*, Constr. Appr. '11

K. Schnass, J.V., Compressed learning of high-dimensional sparse functions, Proceedings of ICASSP '11

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

Use of low-rank matrix recovery:

H. Tyagi, V. Cevher, *Learning non-parametric basis independent models from point queries via low-rank methods*, ACHA '14

A. Kolleck, J. V., *On some aspects of approximation of ridge functions*, JAT '15: **ridge functions on a cube**

- $f : [-1,1]^d \to \mathbb{R}$
- $h > 0$ small
- $\|a\|_1 = 1$, s-sparse
- $\tilde{b}_j := \dfrac{f(h\varphi_j) - f(0)}{h}$, $j = 1, \dots, m$; $m \geq cs\log(d)$;
- $\hat{a}$ obtained by ℓ_1 minimization;
- $\hat{g}(t) := f(t \cdot \operatorname{sign}(\hat{a}))$.
- $\hat{f}(x) := \hat{g}(\langle \hat{a}, x \rangle)$

Then

$$\|f - \hat{f}\|_\infty \leq 2c_0 \|\hat{a} - a\|_{I_1^d} \leq C \frac{h}{g'(0) - c_1 h}$$

$$\operatorname{sign}(x) := (\operatorname{sign}(x_i))_i \in \mathbb{R}^d.$$

- Discontinuous;
- $\langle a, \operatorname{sign}(a)\rangle = \|a\|_1$;
- The scalar product of $\operatorname{sign}(a)$ and $\operatorname{sign}(\hat{a})$ with a is nearly the same:

$$\begin{aligned}
&|\langle a, \operatorname{sign}(a) - \operatorname{sign}(\hat{a})\rangle| \\
&\qquad = |\langle a, \operatorname{sign}(a)\rangle - \langle \hat{a}, \operatorname{sign}(\hat{a})\rangle - \langle a - \hat{a}, \operatorname{sign}(\hat{a})\rangle| \\
&\qquad \le \|a - \hat{a}\|_{l_1^d} \|\operatorname{sign}(\hat{a})\|_{l_\infty^d} \\
&\qquad = \|a - \hat{a}\|_{l_1^d}.
\end{aligned}$$

S. Mayer, T. Ullrich, and J. V., *Entropy and sampling numbers of classes of ridge functions*, Constr. Approx. (2015)

Tractability of approximation of ridge functions

Positive results based on the formula $\nabla f(x) = g'(\langle a, x\rangle)a$

. . . is that optimal?

$$\mathcal{R}_d^{\alpha,p} = \Big\{ f : B(0,1) \subset \mathbb{R}^d \to \mathbb{R} \ : \ f(x) = g(\langle a, x\rangle), \\ \|g\|_{\mathrm{Lip}_\alpha[-1,1]} \le 1, \ \|a\|_p \le 1 \Big\}$$

$\alpha > 0$: Lipschitz smoothness of profiles

$\alpha = \infty$: infinitely differentiable profiles

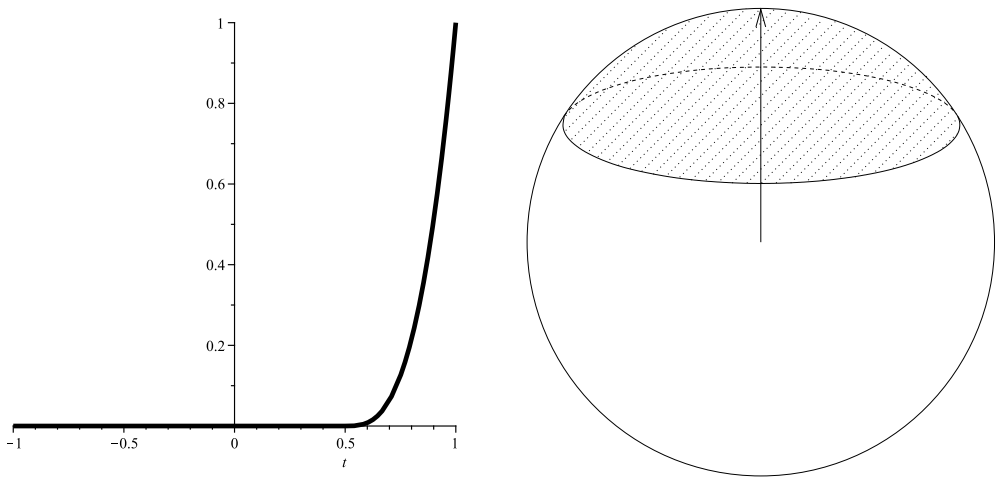
$0 < p \le 2$: restriction of ridge directions

$\mathcal{R}_d^{\alpha,p,\varkappa}$: functions from $\mathcal{R}_d^{\alpha,p}$ with $|g'(0)| \ge \varkappa > 0$

Uniform approximation of functions from $\mathcal{R}_d^{\alpha,p}$

- suffers from the curse of dimension if $p = 2$ and $\alpha < \infty$,
- never suffers from the curse of dimension if $p < 2$,
- is intractable if $p < 2$ and $\alpha \le \frac{1}{1/p-1/2}$,
- is weakly tractable if $p < 2$ and $\alpha > \frac{1}{1/\max\{1,p\}-1/2}$,
- is quasi-polynomially tractable if $\alpha = \infty$,
- for $\mathcal{R}_d^{\alpha,p,\varkappa}$ it is polynomially tractable, no matter what the values of α and p are.

Estimates from below are mainly based on "bad profiles" with $g'(0) = 0$:



$g(x) = 0$ for $-1 \le x \le 1/2$
$f(x) = 0$ for $\{x : \|x\|_2 \le 1 \ \& \ \langle x, a\rangle \le 1/2\}$

Concentration of measure: The measure of the cap $\{x : \|x\|_2 \le 1 \ \& \ \langle x, a\rangle > 1/2\}$ is exponentially small in d

Summary

- Non-linear, non-convex classes
- Sparsity and other structural assumptions
- Compressed Sensing, low-rank matrix recovery, ...
- Linear information for ridge functions?

Thank you for your attention!