Learning sums of ridge functions in high dimension: a nonlinear compressed sensing model

Massimo Fornasier

Winter School on Compressed Sensing
Technical University of Berlin
December 3-5, 2015
Collection of joint results with
Ingrid Daubechies, Karin Schnass, and Jan Vybíral
Introduction on ridge functions

- A ridge function - in its simplest form - is a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ of the type

$$f(x) = g(a^T x) = g(a \cdot x),$$

where $g : \mathbb{R} \rightarrow \mathbb{R}$ is a scalar univariate function and $a \in \mathbb{R}^d$ is the direction of the ridge function;
A ridge function - in its simplest form - is a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ of the type

$$f(x) = g(a^T x) = g(a \cdot x),$$

where $g : \mathbb{R} \rightarrow \mathbb{R}$ is a scalar univariate function and $a \in \mathbb{R}^d$ is the direction of the ridge function;

Ridge functions are constant along the hyperplanes $a \cdot x = \lambda$ for any given level $\lambda \in \mathbb{R}$ and are among the most simple form of multivariate functions;
Introduction on ridge functions

- A ridge function - in its simplest form - is a function $f : \mathbb{R}^d \to \mathbb{R}$ of the type
  \[ f(x) = g(a^T x) = g(a \cdot x), \]
  where $g : \mathbb{R} \to \mathbb{R}$ is a scalar univariate function and $a \in \mathbb{R}^d$ is the direction of the ridge function;

- Ridge functions are constant along the hyperplanes $a \cdot x = \lambda$ for any given level $\lambda \in \mathbb{R}$ and are among the most simple form of multivariate functions;

- They have been extensively studied in the past couple of decades as approximation building blocks for more complicated high dimensional functions.
Some origins of ridge functions

- In multivariate Fourier series, the basis functions are of the form $e^{in \cdot x}$ for $n \in \mathbb{Z}^d$ and $e^{ia \cdot x}$ for arbitrary directions $a \in \mathbb{R}^d$ in the Radon transform.
Some origins of ridge functions

- In multivariate Fourier series, the basis functions are of the form $e^{in \cdot x}$ for $n \in \mathbb{Z}^d$ and $e^{ia \cdot x}$ for arbitrary directions $a \in \mathbb{R}^d$ in the Radon transform;

- The term “ridge function” has been actually coined by Logan and Shepp in 1975 in their work on computer tomography where they show how ridge functions solve the corresponding $L_2$-minimum norm approximation problem.
Projection pursuit of the ’80s

Ridge function approximation has been as well extensively studied during the 80’s in mathematical statistics under the name of projection pursuit (Huber, 1985; Donoho-Johnston, 1989);
Ridge function approximation has been as well extensively studies during the 80’s in mathematical statistics under the name of projection pursuit (Huber, 1985; Donoho-Johnston, 1989);

Projection pursuit algorithms approximate a function of $d$ variables by functions of the form

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

for some functions $g_i : \mathbb{R} \to \mathbb{R}$ and some non-zero vectors $a_i \in \mathbb{R}^d$. 
Some relevant applications of the ’90s

In the early 90’s there has been an explosion of interest in the field of neural networks. One very popular model is the multilayer feed-forward neural network with input, hidden (internal), and output layers;
Some relevant applications of the ’90s

- In the early 90’s there has been an explosion of interest in the field of *neural networks*. One very popular model is the multilayer feed-forward neural network with input, hidden (internal), and output layers;

- the simplest case of such a network is described mathematically by a function of the form

\[ \sum_{i=1}^{m} \alpha_i \sigma \left( \sum_{j=1}^{m} w_{ij} x_j + \theta_i \right), \]

where \( \sigma : \mathbb{R} \to \mathbb{R} \) is somehow given and called the *activation function* and \( w_{ij} \) are suitable weights;
Ridge functions and approximation theory

- In the early 90’s the question of whether one can use sums of ridge functions to approximate well arbitrary functions has been at the center of the attention of the approximation theory community (overviews by Li 2002 and Pinkus 1997);
Ridge functions and approximation theory

- In the early 90’s the question of whether one can use sums of ridge functions to approximate well arbitrary functions has been at the center of the attention of the approximation theory community (overviews by Li 2002 and Pinkus 1997);
- the efficiency of such an approximation compared to, e.g., spline type approximation for smoothness classes of functions, has been extensively considered (DeVore et al. 1997; Petrushev, 1999);
Ridge functions and approximation theory

- In the early 90’s the question of whether one can use sums of ridge functions to approximate well arbitrary functions has been at the center of the attention of the approximation theory community (overviews by Li 2002 and Pinkus 1997);
- the efficiency of such an approximation compared to, e.g., spline type approximation for smoothness classes of functions, has been extensively considered (DeVore et al. 1997; Petrushev, 1999);
- The identification of a ridge function has also been thoroughly considered, in particular we mention the work of Pinkus, and, for what concerns multilayer neural networks, we refer to the work by Fefferman, 1994.
In the early 90’s the question of whether one can use sums of ridge functions to approximate well arbitrary functions has been at the center of the attention of the approximation theory community (overviews by Li 2002 and Pinkus 1997);

the efficiency of such an approximation compared to, e.g., spline type approximation for smoothness classes of functions, has been extensively considered (DeVore et al. 1997; Petrushev, 1999);

The identification of a ridge function has also been thoroughly considered, in particular we mention the work of Pinkus, and, for what concerns multilayer neural networks, we refer to the work by Fefferman, 1994.

Except for the work of Candès on ridglets, there has been less attention after 2000 on the problem of approximating functions by means of ridge functions;
In the early 90’s the question of whether one can use sums of ridge functions to approximate well arbitrary functions has been at the center of the attention of the approximation theory community (overviews by Li 2002 and Pinkus 1997);

the efficiency of such an approximation compared to, e.g., spline type approximation for smoothness classes of functions, has been extensively considered (DeVore et al. 1997; Petrushev, 1999);

The identification of a ridge function has also been thoroughly considered, in particular we mention the work of Pinkus, and, for what concerns multilayer neural networks, we refer to the work by Fefferman, 1994.

Except for the work of Candès on ridglets, there has been less attention after 2000 on the problem of approximating functions by means of ridge functions;
Capturing ridge functions from point queries

- The above results on the identification of such functions based on disposing of any possible output or even derivatives;
Capturing ridge functions from point queries

- The above results on the identification of such functions based on disposing of any possible output or even derivatives;
- this might be in certain practical situations very expensive, hazardous or impossible;

For $g \in C^s([0, 1])$, $1 < s$, $\|g\|_{C^s} \leq M_0$, $\|a\|_{\ell^dq} \leq M_1$, $0 < q \leq 1$ $\|f - \hat{f}\|_{C^\infty(\Omega)} \leq CM_0\{L^{-s} + M_1\left(1 + \log\left(d/L\right)\right)\}^{1/q - 1}$ using $3L + 2$ sampling points, deterministically and adaptively chosen.
Capturing ridge functions from point queries

- The above results on the identification of such functions based on disposing of any possible output or even derivatives;
- this might be in certain practical situations very expensive, hazardous or impossible;
- In a paper of 2012, Cohen, Daubechies, DeVore, Kerkyacharian, and Picard address the approximation of ridge functions by the minimal amount of sampling queries:
Capturing ridge functions from point queries

- The above results on the identification of such functions based on disposing of any possible output or even derivatives;
- this might be in certain practical situations very expensive, hazardous or impossible;
- In a paper of 2012, Cohen, Daubechies, DeVore, Kerkyacharian, and Picard address the approximation of ridge functions by the minimal amount of sampling queries:

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

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

using \( 3L + 2 \) sampling points, deterministically and adaptively chosen.
Capturing ridge functions from point queries: a nonlinear compressed sensing model

Compressed sensing: given $X \in \mathbb{R}^{m \times d}$ sensing matrix, for $m \ll d$, suitable matrix, we wish to identify a nearly sparse vector $a \in \mathbb{R}^d$ from its measurements

$$y \approx Xa,$$

by means of suitable algorithms ($\ell_1$-minimization, greedy algs) aware of $y$ and $X$. 
Capturing ridge functions from point queries: a nonlinear compressed sensing model

Compressed sensing: given $X \in \mathbb{R}^{m \times d}$ sensing matrix, for $m \ll d$, suitable matrix, we wish to identify a nearly sparse vector $a \in \mathbb{R}^d$ from its measurements

$$y \approx Xa,$$

by means of suitable algorithms ($\ell_1$-minimization, greedy algs) aware of $y$ and $X$. The data

$$y_i \approx x_i \cdot a = x_i^T a, \quad i = 1, \ldots, m$$

are linear measurements of $a$. 
Capturing ridge functions from point queries: a nonlinear compressed sensing model

Compressed sensing: given $X \in \mathbb{R}^{m \times d}$ sensing matrix, for $m \ll d$, suitable matrix, we wish to identify a nearly sparse vector $a \in \mathbb{R}^d$ from its measurements

$$y \approx Xa,$$

by means of suitable algorithms ($\ell_1$-minimization, greedy algs) aware of $y$ and $X$. The data

$$y_i \approx x_i \cdot a = x_i^T a, \quad i = 1, \ldots, m$$

are linear measurements of $a$. If now we assume $y_i$ to be the values of a ridge function at the points $x_i$
Capturing ridge functions from point queries: a nonlinear compressed sensing model

Compressed sensing: given $X \in \mathbb{R}^{m \times d}$ sensing matrix, for $m \ll d$, suitable matrix, we wish to identify a nearly sparse vector $a \in \mathbb{R}^d$ from its measurements

$$y \approx Xa,$$

by means of suitable algorithms ($\ell_1$-minimization, greedy algs) aware of $y$ and $X$. The data

$$y_i \approx x_i \cdot a = x_i^T a, \quad i = 1, \ldots, m$$

are linear measurements of $a$. If now we assume $y_i$ to be the values of a ridge function at the points $x_i$

$$y_i \approx g(a \cdot x_i), \quad i = 1, \ldots, m,$$

for some unknown or roughly given nonlinear function $g$. 
Capturing ridge functions from point queries: a nonlinear compressed sensing model

Compressed sensing: given \( X \in \mathbb{R}^{m \times d} \) sensing matrix, for \( m \ll d \), suitable matrix, we wish to identify a nearly sparse vector \( a \in \mathbb{R}^d \) from its measurements

\[
y \approx Xa,
\]

by means of suitable algorithms (\( \ell_1 \)-minimization, greedy algs) aware of \( y \) and \( X \). The data

\[
y_i \approx x_i \cdot a = x_i^T a, \quad i = 1, \ldots, m
\]

are linear measurements of \( a \). If now we assume \( y_i \) to be the values of a ridge function at the points \( x_i \)

\[
y_i \approx g(a \cdot x_i), \quad i = 1, \ldots, m,
\]

for some unknown or roughly given nonlinear function \( g \), the problem of identifying the ridge direction can be understood as a nonlinear compressed sensing model ...
Ridge functions and functions of data clustered around manifolds

Figure: Functions on data clustered around a manifold can be locally approximated by $k$-ridge functions.
Universal random sampling for a more general ridge model

M. Fornasier, K. Schnass, J. Vybíral, *Learning functions of few arbitrary linear parameters in high dimensions, FoCM, 2012*

\[ f(x) = g(Ax), \text{ } A \text{ is a } k \times d \text{ matrix} \]
Universal random sampling for a more general ridge model

M. Fornasier, K. Schnass, J. Vybíral, *Learning functions of few arbitrary linear parameters in high dimensions, FoCM, 2012*

\[ f(x) = g(Ax), \, A \text{ is a } k \times d \text{ matrix} \]

Rows of \( A \) are compressible: \( \max_i \| a_i \|_q \leq C_1, \, 0 < q \leq 1 \)
\( AA^T \) is the identity operator on \( \mathbb{R}^k \)

The regularity condition: \( \sup_{|\alpha| \leq 2} \| D^\alpha g \|_\infty \leq C_2 \)
Universal random sampling for a more general ridge model

M. Fornasier, K. Schnass, J. Vybíral, *Learning functions of few arbitrary linear parameters in high dimensions, FoCM, 2012*

\[ f(x) = g(Ax), \text{ } A \text{ is a } k \times d \text{ matrix} \]

Rows of \( A \) are compressible: \( \max_i \| a_i \|_q \leq C_1, \text{ } 0 < q \leq 1 \)
\( AA^T \) is the identity operator on \( \mathbb{R}^k \)

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

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

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

\[ \sigma_1(H^f) \geq \cdots \geq \sigma_k(H^f) \geq \alpha > 0. \]
How can we learn $k$-ridge functions from point queries?
MD. House’s differential diagnosis (or simply called ”sensitivity analysis”)

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

\[
\nabla g(Ax)^T A \varphi = \frac{\partial f}{\partial \varphi} (x) \\
= \frac{f(x + \epsilon \varphi) - f(x)}{\epsilon} - \frac{\epsilon}{2} \varphi^T \nabla^2 f(\zeta) \varphi, \quad \epsilon \leq \bar{\epsilon}
\]
MD. House’s differential diagnosis (or simply called "sensitivity analysis")

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

\[
\nabla g(Ax)^T A\varphi = \frac{\partial f}{\partial \varphi}(x) = \frac{f(x + \epsilon \varphi) - f(x)}{\epsilon} - \frac{\epsilon}{2} [\varphi^T \nabla^2 f(\zeta) \varphi], \quad \epsilon \leq \bar{\epsilon}
\]

$X = \{x^j \in \Omega : j = 1, \ldots, m_X\}$ drawn uniformly at random in $\Omega \subset \mathbb{R}^d$

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

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

for every $j \in \{1, \ldots, m_\Phi\}$ and every $\ell \in \{1, \ldots, d\}$
Sensitivity analysis

Figure: We perform at random, randomized sensitivity analysis
Collecting together the differential analysis

\( \Phi \ldots m_\Phi \times d \) matrix whose rows are \( \varphi^i \), \( X \ldots d \times m_X \) matrix

\[
X = (A^T \nabla g(Ax^1)) \ldots (A^T \nabla g(Ax^{m_X})).
\]

The \( m_X \times m_\Phi \) instances of (\( \ast \)) in matrix notation as

\[\Phi X = Y + \mathcal{E} \quad (\ast \ast)\]

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

\[
y_{ij} = \frac{f(x^j + \epsilon \varphi^i) - f(x^j)}{\epsilon},
\]

\[
\epsilon_{ij} = -\frac{\epsilon}{2} [(\varphi^i)^T \nabla^2 f(\zeta_{ij}) \varphi^i],
\]
Example of active coordinates: which factor does play a role?

We assume, that

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

i.e.

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

where $f : \Omega = [0, 1]^d \rightarrow \mathbb{R}$ and $g : [0, 1]^k \rightarrow \mathbb{R}$.
Example of active coordinates: which factor does play a role?

We assume, that

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

i.e.

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

where \( f : \Omega = [0, 1]^d \to \mathbb{R} \) and \( g : [0, 1]^k \to \mathbb{R} \).

We want to identify first the active coordinates \( i_1, \ldots, i_k \). Then one can apply any usual \( k \)-dimensional approximation method...
Example of active coordinates: which factor does play a role?

We assume, that

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

i.e.

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

where \( f : \Omega = [0, 1]^d \to \mathbb{R} \) and \( g : [0, 1]^k \to \mathbb{R} \)

We want to identify first the active coordinates \( i_1, \ldots, i_k \). Then one can apply any usual \( k \)-dimensional approximation method...

A possible algorithm chooses the sampling points at random, due to the concentration of measure effects, we get the right result with overwhelming probability.
A simple algorithm based on concentration of measure

The algorithm to identify the active coordinates $I$ is based on the identity

$$\Phi^T \Phi X = \Phi^T Y + \Phi^T \xi$$

where now $X$ has $i^{th}$-row

$$X_i = \left( \frac{\partial g}{\partial z_i}(Ax^1), \ldots, \frac{\partial g}{\partial z_i}(Ax^m) \right),$$

for $i \in I$, and all other row equal to zero.
A simple algorithm based on concentration of measure

The algorithm to identify the active coordinates $I$ is based on the identity

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

where now $X$ has $i^{th}$-row

$$X_i = \left( \frac{\partial g}{\partial z_i} (Ax^1), \ldots, \frac{\partial g}{\partial z_i} (Ax^m) \right),$$

for $i \in I$, and all other row equal to zero. In expectation:

$\Phi^T \Phi \approx I_d : \mathbb{R}^d \rightarrow \mathbb{R}^d$

$\Phi^T \Phi X \approx X$ and

$\Phi^T \mathcal{E}$ is small $\implies \Phi^T Y \approx X,$
A simple algorithm based on concentration of measure

The algorithm to identify the active coordinates $I$ is based on the identity

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

where now $X$ has $i^{th}$-row

$$X_i = \left( \frac{\partial g}{\partial z_i}(Ax^1), \ldots, \frac{\partial g}{\partial z_i}(Ax^m) \right),$$

for $i \in I$, and all other row equal to zero. 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 $X$. 
A first recovery result

**Theorem (Schnass and Vybíral 2011)**

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 \( L \leq d \), 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

\[
\frac{\max_{j=1,...,k} \| \partial_i g \|_\infty}{\min_{j=1,...,k} \| \partial_i g \|_1}
\]
Examples of active coordinate detection in dimension $d = 1000$

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

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

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

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

\[ \alpha = \int_{S^{d-1}} \|\nabla f(x)\|_2^2 d\mu_{S^{d-1}}(x) = \int_{S^{d-1}} |g'(a \cdot x)|^2 d\mu_{S^{d-1}}(x) > 0, \]
Learning ridge functions $k = 1$

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

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

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

We consider again the Taylor expansion (*) with $\Omega = S^{d-1}$

We choose the points $X = \{x^j \in S^{d-1} : j = 1, \ldots, m_X\}$ generated at random on $S^{d-1}$ with respect to $\mu_{S^{d-1}}$

The matrix $\Phi$ is generated as before and we obtain (**) again in the form

$$\Phi[g'(a \cdot x^j) a] = y_j + \varepsilon_j, j = 1, \ldots m_X.$$
Algorithm 1:

- Given $m_\Phi, m_\mathcal{X}$, draw at random the sets $\Phi$ and $\mathcal{X}$, and construct $Y$ according (*).
- Set $\hat{x}_j = \Delta(y_j) := \arg\min_{y_j=\Phi} \|z\|_{\ell^1}$.
- Find
  $$j_0 = \arg\max_{j=1,\ldots,m_\mathcal{X}} \|\hat{x}_j\|_{\ell^2}.$$
- Set $\hat{a} = \hat{x}_{j_0}/\|\hat{x}_{j_0}\|_{\ell^2}$.
- Define $\hat{g}(y) := f(\hat{a}^T y)$ and $\hat{f}(x) := \hat{g}(\hat{a} \cdot x)$. 
Recovery result

Theorem (F., Schnass, and Vybíral 2012)

Let $0 < s < 1$ and $\log d \leq m_\Phi \leq \lfloor \log 6 \rfloor d$. Then there is a constant $c_1'$ such that using $m_\chi \cdot (m_\Phi + 1)$ function evaluations of $f$, Algorithm 1 defines a function $\hat{f} : B_{\mathbb{R}^d}(1 + \bar{e}) \to \mathbb{R}$ that, with probability

$$1 - \left( e^{-c_1'm_\Phi} + e^{-\sqrt{m_\Phi}d} + 2e^{-\frac{2m_\chi s^2 \alpha^2}{c_1^2}} \right),$$

will satisfy

$$\|f - \hat{f}\|_\infty \leq 2C_2(1 + \bar{e}) \frac{\nu_1}{\sqrt{\alpha(1 - s) - \nu_1}},$$

where

$$\nu_1 = C' \left( \left[ \frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2 - 1/q} + \frac{e}{\sqrt{m_\Phi}} \right)$$

and $C'$ depends only on $C_1$ and $C_2$. 
Ingredients of the proof

- compressed sensing;
Ingredients of the proof

- compressed sensing;
- stability of one dimensional subspaces;
Ingredients of the proof

- compressed sensing;
- stability of one dimensional subspaces;
- concentration inequalities (Hoeffding’s inequality).
Compressed sensing

Theorem (Wojtaszczyk, 2011)

Assume that $\Phi$ is an $m \times d$ random matrix with all entries being independent Bernoulli variables scaled by $1/\sqrt{m}$.

Let us suppose that $d > \left\lfloor \log_2 6m \right\rfloor$. Then there are positive constants $C, c'_1, c'_2 > 0$, such that, with probability at least $1 - e^{-c'_1 m} - e^{-\sqrt{md}}$, the matrix $\Phi$ has the following property. For every $x \in \mathbb{R}^d$, $\varepsilon \in \mathbb{R}^m$ and every natural number $K \leq c'_2 m / \log (d/m)$ we have

$$
\|\Delta (\Phi x + \varepsilon) - x\|_{\ell^d_2} \leq C (K^{-1/2} \sigma_K(x)_{\ell^d_1} + \max\{\|\varepsilon\|_{\ell^m_2}, \sqrt{\log d} \|\varepsilon\|_{\ell^m_\infty}\}),
$$

where $\sigma_K(x)_{\ell^d_1} := \inf\{\|x - z\|_{\ell^d_1} : \#\text{supp } z \leq K\}$ is the best $K$-term approximation of $x$. 

Compressed sensing

Theorem (Wojtaszczyk, 2011)

Assume that $\Phi$ is an $m \times d$ random matrix with all entries being independent Bernoulli variables scaled by $1/\sqrt{m}$. Let us suppose that $d > [\log 6]^2 m$. 
Compressed sensing

Theorem (Wojtaszczyk, 2011)

Assume that $\Phi$ is an $m \times d$ random matrix with all entries being independent Bernoulli variables scaled by $1/\sqrt{m}$. Let us suppose that $d > [\log 6]^2 m$. Then there are positive constants $C, c_1', c_2' > 0$, such that, with probability at least

$$1 - e^{-c_1'm} - e^{-\sqrt{md}},$$

the matrix $\Phi$ has the following property.
Compressed sensing

Theorem (Wojtaszczyk, 2011)
Assume that $\Phi$ is an $m \times d$ random matrix with all entries being independent Bernoulli variables scaled by $1/\sqrt{m}$. Let us suppose that $d > \left\lceil \log 6 \right\rceil^2 m$. Then there are positive constants $C, c_1', c_2' > 0$, such that, with probability at least

$$1 - e^{-c_1' m} - e^{-\sqrt{md}},$$

the matrix $\Phi$ has the following property. For every $x \in \mathbb{R}^d$, $\varepsilon \in \mathbb{R}^m$ and every natural number $K \leq c_2' m / \log(d/m)$ we have

$$\|\Delta(\Phi x + \varepsilon) - x\|_{\ell_2^d} \leq C \left( K^{-1/2} \sigma_K(x)_{\ell_1^d} + \max\{\|\varepsilon\|_{\ell_2^m}, \sqrt{\log d} \|\varepsilon\|_{\ell_\infty^m}\} \right),$$

where

$$\sigma_K(x)_{\ell_1^d} := \inf\{\|x - z\|_{\ell_1^d} : \# \text{ supp } z \leq K\}$$

is the best $K$-term approximation of $x$. 
How does compressed sensing play a role?

For the $d \times m_X$ matrix $X$, i.e.,

$$X = (g'(a \cdot x^1)a^T | \ldots | g'(a \cdot x^{m_X})a^T),$$

$$\Phi[g'(a \cdot x^j)a] = y_j + \varepsilon_j, j = 1, \ldots m_X,$$

$$:=x_j$$
How does compressed sensing play a role?

For the $d \times m_X$ matrix $X$, i.e.,

$$X = (g'(a \cdot x^1)a^T | \ldots | g'(a \cdot x^{m_X})a^T),$$

$$\Phi[g'(a \cdot x^j)a] = y_j + \varepsilon_j, j = 1, \ldots, m_X,$$

and

$$\hat{x}_j = \Delta(y_j) := \arg \min_{y_j = \Phi z} ||z||_{\ell^q}$$
How does compressed sensing play a role?

For the \( d \times m_X \) matrix \( X \), i.e.,

\[
X = (g'(a \cdot x^1)a^T | \ldots | g'(a \cdot x^{m_X})a^T),
\]

\[
\Phi[g'(a \cdot x^j)a] = y_j + \varepsilon_j, \ j = 1, \ldots m_X,
\]

and

\[
\hat{x}_j = \Delta(y_j) := \arg \min_{y^j = \Phi z} \|z\|_{\ell_1}
\]

the previous result gives - with the probability provided there -

\[
\hat{x}_j = g'(a \cdot x^j)a^T + n_j,
\]

with \( n_j \) properly estimated by

\[
\|n_j\|_{\ell_2} \leq C \left( K^{-1/2} \sigma_K (g'(a \cdot x^j)a^T)_{\ell_1} + \max[\|\varepsilon_j\|_{\ell_2}, \sqrt{\log d}\|\varepsilon_j\|_{\ell_\infty}] \right).
\]
Some computations

Let us estimate the quantities. By Stechkin’s inequality for which

$$
\sigma_K(x)_{\ell_1} \leq \|x\|_{\ell_q} K^{1-1/q},
$$

for all $x \in \mathbb{R}^d$, 

$$
\|\epsilon_j\|_{\ell_{m \Phi}} \leq C_1^2 C_2^2 \sqrt{m \Phi} \epsilon,
$$

leading to

$$
\max\{\|\epsilon_j\|_{\ell_{m \Phi}}, \sqrt{\log d} \|\epsilon_j\|_{\ell_{m \Phi}}\} \leq C_2^2 C_1^2 \sqrt{m \Phi} \epsilon.
$$
Some computations

Let us estimate the quantities. By Stechkin’s inequality for which

$$\sigma_K(x)_{\ell_1} \leq \|x\|_{\ell_q} K^{1-1/q},$$

for all $x \in \mathbb{R}^d$, one obtains - for $x_j = g'(a \cdot x^j)a$

$$K^{-1/2} \sigma_K(x_j)_{\ell_1} \leq |g'(a \cdot x^j)| \cdot \|a\|_{\ell_q} \cdot K^{1/2-1/q} \leq C_1 C_2 \left[ \frac{m\Phi}{\log(d/m\Phi)} \right]^{1/2-1/q}.$$
Some computations

Let us estimate the quantities. By Stechkin’s inequality for which

\[ \sigma_K(x)_{\ell_1} \leq \|x\|_{\ell_q} K^{1-1/q}, \]

for all \( x \in \mathbb{R}^d \), one obtains - for \( x_j = g'(a \cdot x^j)a \)

\[ K^{-1/2} \sigma_K(x_j)_{\ell_1} \leq |g'(a \cdot x^j)| \|a\|_{\ell_q} K^{1/2-1/q} \leq C_1 C_2 \left[ \frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2-1/q}. \]

Moreover

\[ \|\varepsilon_j\|_{\ell_{\infty}^{m_\Phi}} = \frac{\varepsilon}{2} \cdot \max_{i=1,...,m_\Phi} |\varphi^T \nabla^2 f(\zeta_{ij}) \varphi^i| \]

\[ = \frac{\varepsilon}{2m_\Phi} \cdot \max_{i=1,...,m_\Phi} \left| \sum_{k,l=1}^d a_k a_l g''(a \cdot \zeta_{ij}) \right| \]
Some computations

Let us estimate the quantities. By Stechkin’s inequality for which
\[ \sigma_K(x)_{\ell_1^d} \leq \|x\|_{\ell_q^d} K^{1-1/q}, \]
for all \( x \in \mathbb{R}^d \), one obtains - for \( x_j = g'(a \cdot x^j) a \)
\[ K^{-1/2} \sigma_K(x_j)_{\ell_1^d} \leq |g'(a \cdot x^j)| \cdot \|a\|_{\ell_q^d} K^{1/2-1/q} \leq C_1 C_2 \left[ \frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2-1/q}. \]

Moreover
\[ \|\varepsilon_j\|_{\ell_\infty^m} = \frac{\varepsilon}{2} \cdot \max_{i=1,\ldots,m_\Phi} |\varphi^i \nabla^2 f(\zeta_{ij}) \varphi^i| \]
\[ = \frac{\varepsilon}{2m_\Phi} \cdot \max_{i=1,\ldots,m_\Phi} \left| \sum_{k,l=1}^d a_k a_l g''(a \cdot \zeta_{ij}) \right| \]
\[ \leq \frac{\varepsilon \|g''\|_{\ell_\infty^d}}{2m_\Phi} \left( \sum_{k=1}^d |a_k| \right)^2 \leq \frac{\varepsilon \|g''\|_{\ell_\infty^d}}{2m_\Phi} \left( \sum_{k=1}^d |a_k|^q \right)^{2/q} \leq \frac{C_1^2 C_2}{2m_\Phi} \varepsilon, \]
Some computations

Let us estimate the quantities. By Stechkin’s inequality for which

$$\sigma_K(x)_{\ell^d_1} \leq \|x\|_{\ell^d_q} K^{1-1/q},$$

for all $x \in \mathbb{R}^d$, one obtains - for $x_j = g'(a \cdot x^j) a$

$$K^{-1/2} \sigma_K(x_j)_{\ell^d_1} \leq |g'(a \cdot x^j)| \|a\|_{\ell^d_q} K^{1/2-1/q} \leq C_1 C_2 \left[ \frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2-1/q}.$$

Moreover

$$\|\varepsilon_j\|_{\ell^m_\Phi} = \frac{\varepsilon}{2} \max_{i=1,\ldots,m_\Phi} |\varphi^i T \nabla^2 f(\zeta_{ij}) \varphi^i|$$

$$= \frac{\varepsilon}{2m_\Phi} \max_{i=1,\ldots,m_\Phi} \left| \sum_{k,l=1}^d a_k a_l g''(a \cdot \zeta_{ij}) \right|$$

$$\leq \frac{\varepsilon \|g''\|_\infty}{2m_\Phi} \left( \sum_{k=1}^d |a_k| \right)^2 \leq \frac{\varepsilon \|g''\|_\infty}{2m_\Phi} \left( \sum_{k=1}^d |a_k|^q \right)^{2/q} \leq \frac{C_1^2 C_2}{2m_\Phi} \varepsilon,$$

$$\|\varepsilon_j\|_{\ell^m_\Phi} \leq \frac{C_1^2 C_2}{2m_\Phi} \varepsilon,$$
Some computations

Let us estimate the quantities. By Stechkin’s inequality for which
\[ \sigma_K(x)_{\ell_1^d} \leq \|x\|_{\ell_q^d} K^{1-1/q}, \]
for all \( x \in \mathbb{R}^d \), one obtains - for \( x_j = g'(a \cdot x^j)a \)

\[ K^{-1/2} \sigma_K(x_j)_{\ell_1^d} \leq |g'(a \cdot x^j)| \|a\|_{\ell_q^d} K^{1/2-1/q} \leq C_1 C_2 \left[ \frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2-1/q}. \]

Moreover
\[
\| \varepsilon_j \|_{\ell_{m_\Phi}^2} = \frac{\varepsilon}{2} \cdot \max_{i=1, \ldots, m_\Phi} |\varphi_i^T \nabla^2 f(\zeta_{ij}) \varphi^i| \\
= \frac{\varepsilon}{2m_\Phi} \cdot \max_{i=1, \ldots, m_\Phi} \left| \sum_{k,l=1}^{d} a_k a_l g''(a \cdot \zeta_{ij}) \right| \\
\leq \frac{\varepsilon \|g''\|_{\infty}}{2m_\Phi} \left( \sum_{k=1}^{d} |a_k| \right)^2 \leq \frac{\varepsilon \|g''\|_{\infty}}{2m_\Phi} \left( \sum_{k=1}^{d} |a_k|^q \right)^{2/q} \leq \frac{C_1^2 C_2}{2m_\Phi} \varepsilon,
\]
\[
\| \varepsilon_j \|_{\ell_{m_\Phi}^2} \leq \sqrt{m_\Phi} \| \varepsilon_j \|_{\ell_{m_\Phi}^\infty} \leq \frac{C_1^2 C_2}{2\sqrt{m_\Phi}} \varepsilon, \quad \text{leading to}
\]
\[
\max\{\| \varepsilon_j \|_{\ell_{m_\Phi}^2}, \sqrt{\log d} \| \varepsilon_j \|_{\ell_{m_\Phi}^\infty} \} \leq \frac{C_1^2 C_2}{2\sqrt{m_\Phi}} \varepsilon \cdot \max \left\{ 1, \sqrt{\frac{\log d}{m_\Phi}} \right\} = \frac{C_1^2 C_2}{2\sqrt{m_\Phi}} \varepsilon.
\]
With high probability

\[ \hat{x}_j = g'(a \cdot x^j)a^T + n_j, \]
With high probability

\[ \hat{x}_j = g'(a \cdot x^j)a^T + n_j, \]

where

\[ \|n_j\|_{\ell_2^d} \leq C \left( K^{-1/2} \sigma_K (g'(a \cdot x^j)a^T)_{\ell_1^d} + \max\{\|\varepsilon_j\|_{\ell_2^m}, \sqrt{\log d} \|\varepsilon_j\|_{\ell_\infty^m}\} \right) \]

\[ \leq C' \left( \left[ \frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2-1/q} + \frac{\epsilon}{\sqrt{m_\Phi}} \right) := \nu_1 \]
Stability of one dimensional subspaces

Lemma
Let us fix \( \hat{x} \in \mathbb{R}^d \), \( a \in \mathbb{S}^{d-1} \), \( 0 \neq \gamma \in \mathbb{R} \), and \( n \in \mathbb{R}^d \) with norm \( \|n\|_{\ell^2} \leq \nu_1 < |\gamma| \). If we assume \( \hat{x} = \gamma a + n \) then

\[
\left\| \text{sign} \gamma \frac{\hat{x}}{\|\hat{x}\|_{\ell^2}} - a \right\|_{\ell^2} \leq \frac{2\nu_1}{\|\hat{x}\|_{\ell^2}}.
\]
Stability of one dimensional subspaces

Lemma
Let us fix \( \hat{x} \in \mathbb{R}^d \), \( a \in \mathbb{S}^{d-1} \), \( 0 \neq \gamma \in \mathbb{R} \), and \( n \in \mathbb{R}^d \) with norm \( \| n \|_{\ell^2} \leq \nu_1 < |\gamma| \). If we assume \( \hat{x} = \gamma a + n \) then

\[
\left\| \operatorname{sign} \gamma \frac{\hat{x}}{\| \hat{x} \|_{\ell^2}} - a \right\|_{\ell^2} \leq \frac{2\nu_1}{\| \hat{x} \|_{\ell^2}}.
\]

We recall, that

\( \hat{x}_j = g'(a \cdot x^j)a^T + n_j \).

and

\[
\max_j \| \hat{x}_j \|_{\ell^2} \geq \max_j |g'(a \cdot x^j)| - \max_j \| \hat{x}_j - x^j \|_{\ell^2} \geq \max_j |g'(a \cdot x^j)| - \nu_1
\]

we need to estimate it.
Lemma (Hoeffding’s inequality)

Let $X_1, \ldots, X_m$ be independent random variables. Assume that the $X_j$ are almost surely bounded, i.e., there exist finite scalars $a_j, b_j$ such that

$$
\Pr\{X_j - \mathbb{E}X_j \in [a_j, b_j]\} = 1,
$$

for $j = 1, \ldots, m$. Then we have

$$
\Pr\left\{ \left| \sum_{j=1}^m X_j - \mathbb{E}\left( \sum_{j=1}^m X_j \right) \right| \geq t \right\} \leq 2e^{-\frac{2t^2}{\sum_{j=1}^m (b_j-a_j)^2}}.
$$
Lemma (Hoeffding’s inequality)

Let $X_1, \ldots, X_m$ be independent random variables. Assume that the $X_j$ are almost surely bounded, i.e., there exist finite scalars $a_j, b_j$ such that

$$\mathbb{P}\{X_j - \mathbb{E}X_j \in [a_j, b_j]\} = 1,$$

for $j = 1, \ldots, m$. Then we have

$$\mathbb{P}\left\{\left|\sum_{j=1}^{m} X_j - \mathbb{E}\left(\sum_{j=1}^{m} X_j\right)\right| \geq t\right\} \leq 2e^{-\frac{2t^2}{\Sigma_{j=1}^{m} (b_j - a_j)^2}}.$$

Let us now apply Hoeffding’s inequality to the random variables $X_j = |g'(a \cdot x^j)|^2$. 
Probabilistic estimates from below

By applying Hoeffding’s inequality to the random variables $X_j = |g'(a \cdot x^j)|^2$, we have

**Lemma**

Let us fix $0 < s < 1$. Then with probability $1 - 2e^{-\frac{2m_X s^2 \alpha^2}{C_1^2}}$ we have

$$\max_{j=1,\ldots,m_X} |g'(a \cdot x^j)| \geq \sqrt{\alpha(1-s)},$$

where $\alpha := \mathbb{E}_x(|g'(a \cdot x^j)|^2) = \int_{S^{d-1}} |g'(a \cdot x)|^2 d\mu_{S^{d-1}}(x) = \int_{S^{d-1}} \|\nabla f(x)\|_{\ell_2}^2 d\mu_{S^{d-1}}(x) > 0.$
Algorithm 1:

▶ Given \( m_\Phi \), \( m_\mathcal{X} \), draw at random the sets \( \Phi \) and \( \mathcal{X} \), and construct \( Y \) according (\*).

▶ Set \( \hat{x}_j = \Delta(y_j) := \arg \min_{y_j = \Phi} \| z \|_{\ell^1} \).

▶ Find

\[
j_0 = \arg \max_{j=1,...,m_\mathcal{X}} \| \hat{x}_j \|_{\ell^2}.
\]

▶ Set \( \hat{a} = \hat{x}_{j_0} / \| \hat{x}_{j_0} \|_{\ell^2} \).

▶ Define \( \hat{g}(y) := f(\hat{a}^T y) \) and \( \hat{f}(x) := \hat{g}(\hat{a} \cdot x) \).
Theorem (F., Schnass, and Vybíral 2012)

Let $0 < s < 1$ and $\log d \leq m_\Phi \leq [\log 6]^2 d$. Then there is a constant $c'_1$ such that using $m_X \cdot (m_\Phi + 1)$ function evaluations of $f$, Algorithm 1 defines a function $\hat{f} : B_{\mathbb{R}^d}(1 + \bar{\epsilon}) \to \mathbb{R}$ that, with probability

$$1 - \left( e^{-c'_1 m_\Phi} + e^{-\sqrt{m_\Phi} d} + 2e^{-\frac{2m_X s^2 \alpha^2}{c'_2}} \right),$$

will satisfy

$$\|f - \hat{f}\|_\infty \leq 2C_2(1 + \bar{\epsilon}) \frac{\nu_1}{\sqrt{\alpha(1 - s)} - \nu_1},$$

where

$$\nu_1 = C' \left( \left[ \frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2 - 1/q} + \frac{\epsilon}{\sqrt{m_\Phi}} \right)$$

and $C'$ depends only on $C_1$ and $C_2$. 
Concentration of measure phenomenon and risk of intractability

Key role is played by

\[ \alpha = \int_{S^{d-1}} |g'(a \cdot x)|^2 d\mu_{S^{d-1}}(x) \]
Concentration of measure phenomenon and risk of intractability

Key role is played by

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

Due to symmetry ... independent on \( a \)
Concentration of measure phenomenon and risk of intractability

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 \(\mu_1\) on \([-1, 1]\)

\[ \alpha = \int_{-1}^{1} |g'(y)|^2 d\mu_1(y) \]

\[ = \frac{\Gamma(d/2)}{\pi^{1/2} \Gamma(((d - 1)/2))} \int_{-1}^{1} |g'(y)|^2 (1 - y^2)^{d-3/2} dy \]

\(\mu_1\) concentrates around zero exponentially fast as \(d \to \infty\)
Dependence on the dimension $d$

**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 $\ell = 1, \ldots, M$. Then

$$\alpha(d) = \Theta(d^{-M}), \text{ for } d \to \infty.$$
Tractability classes

(1) For $0 < q \leq 1$, $C_1 > 1$ and $C_2 \geq \alpha_0 > 0$, we define

\[ F^1_d := F^1_d(\alpha_0, q, C_1, C_2) := \{ f : B_{\mathbb{R}^d} \to \mathbb{R} : \exists a \in \mathbb{R}^d, \|a\|_{\ell^q} = 1, \|a\|_{\ell^2} \leq C_1 \text{ and } \exists g \in C^2(B_{\mathbb{R}}), |g'(0)| \geq \alpha_0 > 0 : f(x) = g(a \cdot x) \} \].
Tractability classes

(1) For $0 < q \leq 1$, $C_1 > 1$ and $C_2 \geq \alpha_0 > 0$, we define

$$F_1^d := F_1^d(\alpha_0, q, C_1, C_2) := \{ f : B_{\mathbb{R}^d} \to \mathbb{R} : \exists a \in \mathbb{R}^d, \|a\|_2 = 1, \|a\|_q \leq C_1 \text{ and } \exists g \in C^2(B_{\mathbb{R}}), |g'(0)| \geq \alpha_0 > 0 : f(x) = g(a \cdot x) \}.$$

(2) For a neighborhood $U$ of 0, $0 < q \leq 1$, $C_1 > 1$, $C_2 \geq \alpha_0 > 0$ and $N \geq 2$, we define

$$F_2^d := F_2^d(U, \alpha_0, q, C_1, C_2, N) := \{ f : B_{\mathbb{R}^d} \to \mathbb{R} : \exists a \in \mathbb{R}^d, \|a\|_2 = 1, \|a\|_q \leq C_1 \text{ and } \exists g \in C^2(B_{\mathbb{R}}) \cap C^N(U), \exists 0 \leq M \leq N - 1, |g^{(M)}(0)| \geq \alpha_0 > 0 : f(x) = g(a \cdot x) \}.$$
Tractability classes

(1) For $0 < q \leq 1$, $C_1 > 1$ and $C_2 \geq \alpha_0 > 0$, we define

$$
\mathcal{F}_d^1 := \mathcal{F}_d^1(\alpha_0, q, C_1, C_2) := \{f : B_{\mathbb{R}^d} \to \mathbb{R} : \\
\exists a \in \mathbb{R}^d, \|a\|_{\ell_2} = 1, \|a\|_{\ell_q} \leq C_1 \quad \text{and} \\
\exists g \in C^2(B_{\mathbb{R}}) \cap C^\infty(\mathbb{U}), \ |g'(0)| \geq \alpha_0 > 0 : f(x) = g(a \cdot x) \}. 
$$

(2) For a neighborhood $\mathcal{U}$ of $0$, $0 < q \leq 1$, $C_1 > 1$, $C_2 \geq \alpha_0 > 0$ and $N \geq 2$, we define

$$
\mathcal{F}_d^2 := \mathcal{F}_d^2(\mathcal{U}, \alpha_0, q, C_1, C_2, N) := \{f : B_{\mathbb{R}^d} \to \mathbb{R} : \\
\exists a \in \mathbb{R}^d, \|a\|_{\ell_2} = 1, \|a\|_{\ell_q} \leq C_1 \quad \text{and} \quad \exists g \in C^2(B_{\mathbb{R}}) \cap C^N(\mathcal{U}) \\
\exists 0 \leq M \leq N - 1, \ |g^{(M)}(0)| \geq \alpha_0 > 0 : f(x) = g(a \cdot x) \}. 
$$

(3) For a neighborhood $\mathcal{U}$ of $0$, $0 < q \leq 1$, $C_1 > 1$ and $C_2 \geq \alpha_0 > 0$, we define

$$
\mathcal{F}_d^3 := \mathcal{F}_d^3(\mathcal{U}, \alpha_0, q, C_1, C_2) := \{f : B_{\mathbb{R}^d} \to \mathbb{R} : \\
\exists a \in \mathbb{R}^d, \|a\|_{\ell_2} = 1, \|a\|_{\ell_q} \leq C_1 \quad \text{and} \quad \exists g \in C^2(B_{\mathbb{R}}) \cap C^\infty(\mathcal{U}) \\
|g^{(M)}(0)| = 0 \quad \text{for all} \quad M \in \mathbb{N} : f(x) = g(a \cdot x) \}. 
$$
Corollary

The problem of learning functions $f$ in the classes $\mathcal{F}^1_d$ and $\mathcal{F}^2_d$ from point evaluations is strongly polynomially tractable (no poly dep. on $d$) and polynomially tractable (with poly dep. on $d$) respectively.
Intractability

On the one hand, let us notice that if in the class $F_d^3$ we remove the condition $\|a\|_{\ell_d} \leq C_1$, then the problem actually becomes *intractable*. 
Intractability

On the one hand, let us notice that if in the class $\mathcal{F}_d^3$ we remove the condition $\|a\|_{\ell_q} \leq C_1$, then the problem actually becomes intractable. Let $g \in C^2([-1 - \bar{\epsilon}, 1 + \bar{\epsilon}])$ given by $g(y) = 8(y - 1/2)^3$ for $y \in [1/2, 1 + \bar{\epsilon}]$ and zero otherwise.
**Intractability**

On the one hand, let us notice that if in the class $\mathcal{F}_d^3$ we remove the condition $\|a\|_{\ell^d} \leq C_1$, then the problem actually becomes *intractable*. Let $g \in C^2([-1 - \bar{\epsilon}, 1 + \bar{\epsilon}])$ given by $g(y) = 8(y - 1/2)^3$ for $y \in [1/2, 1 + \bar{\epsilon}]$ and zero otherwise. Notice that, for every $a \in \mathbb{R}^d$ with $\|a\|_{\ell^d} = 1$, the function $f(x) = g(a \cdot x)$ vanishes everywhere on $\mathbb{S}^{d-1}$ outside of the cap $\mathcal{U}(a, 1/2) := \{x \in \mathbb{S}^{d-1} : a \cdot x \geq 1/2\}$.

![Figure: The function $g$ and the spherical cap $\mathcal{U}(a, 1/2)$.](image-url)
The $\mu_{S^{d-1}}$ measure of $U(a, 1/2)$ obviously does not depend on $a$ and is known to be exponentially small in $d$. Furthermore, it is known, that there is a constant $c > 0$ and unit vectors $a^1, \ldots, a^K$, such that the sets $U(a^1, 1/2), \ldots, U(a^K, 1/2)$ are mutually disjoint and $K \geq e^{cd}$. Finally, we observe that $\max_{x \in S^{d-1}} |f(x)| = f(a) = g(1) = 1$. 
Intractability

The $\mu_{S^{d-1}}$ measure of $\mathcal{U}(a, 1/2)$ obviously does not depend on $a$ and is known to be exponentially small in $d$. Furthermore, it is known, that there is a constant $c > 0$ and unit vectors $a^1, \ldots, a^K$, such that the sets $\mathcal{U}(a^1, 1/2), \ldots, \mathcal{U}(a^K, 1/2)$ are mutually disjoint and $K \geq e^{cd}$. Finally, we observe that $\max_{x \in S^{d-1}} |f(x)| = f(a) = g(1) = 1$. We conclude that any algorithm making only use of the structure of $f(x) = g(a \cdot x)$ and the condition needs to use exponentially many sampling points in order to distinguish between $f(x) \equiv 0$ and $f(x) = g(a^i \cdot x)$ for some of the $a^i$’s as constructed above.
Truly $k$-ridge functions for $k \gg 1$

\[ f(x) = g(Ax), \ A \text{ is a } k \times d \text{ matrix} \]
Truly $k$-ridge functions for $k \gg 1$

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

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: $\sup_{|\alpha| \leq 2} \|D^\alpha g\|_{\infty} \leq C_2$
Truly $k$-ridge functions for $k \gg 1$

\[ f(x) = g(Ax), \ A \text{ is a } k \times d \text{ matrix} \]

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

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

$$\sigma_1(H^f) \geq \cdots \geq \sigma_k(H^f) \geq \alpha > 0.$$
MD. House’s differential diagnosis (or simply called ”sensitivity analysis”)

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

\[
\nabla g(Ax)^T A \varphi = \frac{\partial f}{\partial \varphi}(x)
= \frac{f(x + \epsilon \varphi) - f(x)}{\epsilon} - \frac{\epsilon}{2} [\varphi^T \nabla^2 f(\zeta) \varphi], \quad \epsilon \leq \bar{\epsilon}
\]
MD. House’s differential diagnosis (or simply called ”sensitivity analysis”)

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

$$\nabla g(Ax)^TA\varphi = \frac{\partial f}{\partial \varphi}(x) \quad \text{(*)}$$

$$= \frac{f(x + \epsilon \varphi) - f(x)}{\epsilon} - \frac{\epsilon}{2} [\varphi^T \nabla^2 f(\zeta) \varphi], \quad \epsilon \leq \bar{\epsilon}$$

$X = \{x^j \in \Omega : j = 1, \ldots, m_X\}$ drawn uniformly at random in $\Omega \subset \mathbb{R}^d$

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

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

for every $j \in \{1, \ldots, m_\Phi\}$ and every $\ell \in \{1, \ldots, d\}$
Sensitivity analysis

Figure: We perform at random, randomized sensitivity analysis
Collecting together the differential analysis

$\Phi \ldots m_\Phi \times d$ matrix whose rows are $\varphi^i$, $X \ldots d \times m_X$ matrix

$$X = (A^T \nabla g(Ax^1)| \ldots | A^T \nabla g(Ax^{m_X})) .$$

The $m_X \times m_\Phi$ instances of $(\ast)$ in matrix notation as

$$\Phi X = Y + \mathcal{E} \quad (\ast\ast)$$

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

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

$$\mathcal{E}_{ij} = -\frac{\epsilon}{2} [(\varphi^i)^T \nabla^2 f(\zeta_{ij}) \varphi^i] ,$$
Algorithm 2:

- Given $m_\Phi, m_\mathcal{X}$, draw at random the sets $\Phi$ and $\mathcal{X}$, and construct $Y$ according to (*).
- Set $\hat{x}_j = \Delta(y_j) := \arg \min_{y_j = \Phi} \|z\|_{\ell_1^q}$, for $j = 1, \ldots, m_\mathcal{X}$, and $\hat{X} = (\hat{x}_1|\ldots|\hat{x}_{m_\mathcal{X}})$ is again a $d \times m_\mathcal{X}$ matrix.
- Compute the singular value decomposition of

\[
\hat{X}^T = (\hat{U}_1 \hat{U}_2) \begin{pmatrix}
\hat{\Sigma}_1 & 0 \\
0 & \hat{\Sigma}_2
\end{pmatrix} \begin{pmatrix}
\hat{V}_1^T \\
\hat{V}_2^T
\end{pmatrix},
\]

where $\hat{\Sigma}_1$ contains the $k$ largest singular values.
- Set $\hat{A} = \hat{V}_1^T$.
- Define $\hat{g}(y) := f(\hat{A}^T y)$ and $\hat{f}(x) := \hat{g}(\hat{A}x)$. 
The control of the error

The quality of the final approximation of $f$ by means of $\hat{f}$ depends on two kinds of accuracies:

1. The error between $\hat{X}$ and $X$, which can be controlled through the number of compressed sensing measurements $m_\Phi$;
The control of the error

The quality of the final approximation of $f$ by means of $\hat{f}$ depends on two kinds of accuracies:

1. The error between $\hat{X}$ and $X$, which can be controlled through the number of compressed sensing measurements $m_\Phi$;

2. The stability of the span of $V^T$, simply characterized by how well the singular values of $X$ or equivalently $\mathcal{G}$ are separated from 0, which is related to the number of random samples $m_X$.

To be precise, we have
Recovery result

**Theorem (F., Schnass, and Vybíral)**

Let $\log d \leq m_\Phi \leq [\log 6]^2 d$. Then there is a constant $c_1'$ such that using $m_\chi \cdot (m_\Phi + 1)$ function evaluations of $f$, Algorithm 2 defines a function $\hat{f} : B_{\mathbb{R}^d}(1 + \bar{\epsilon}) \rightarrow \mathbb{R}$ that, with probability

$$1 - \left( e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi} d} + ke^{-\frac{m_\chi \alpha s^2}{2kC_2^2}} \right),$$

will satisfy

$$\|f - \hat{f}\|_\infty \leq 2C_2 \sqrt{k(1 + \bar{\epsilon})} \frac{\nu_2}{\sqrt{\alpha(1 - s)} - \nu_2},$$

where

$$\nu_2 = C \left( k^{1/q} \left[ \frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2 - 1/q} + \frac{\epsilon k^2}{\sqrt{m_\Phi}} \right),$$

and $C$ depends only on $C_1$ and $C_2$. 
Ingredients of the proof

- compressed sensing;
Ingredients of the proof

- compressed sensing;
- stability of the SVD;
Ingredients of the proof

- compressed sensing;
- stability of the SVD;
- concentration inequalities (Chernoff bounds for sums of positive-semidefinite matrices).
Compressed sensing

**Corollary (after Wojtaszczyk, 2011)**

Let \( \log d \leq m_\Phi < [\log 6]^2 d \). Then with probability

\[
1 - (e^{-c'_1 m_\Phi} + e^{-\sqrt{m_\Phi} d})
\]

the matrix \( \hat{X} \) as calculated in Algorithm 2 satisfies

\[
\|X - \hat{X}\|_F \leq C \sqrt{m_X} \left( k^{1/q} \left[ \frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2 - 1/q} + \frac{\epsilon k^2}{\sqrt{m_\Phi}} \right),
\]

where \( C \) depends only on \( C_1 \) and \( C_2 \).
Stability of SVD

Given two matrices $B$ and $\hat{B}$ with corresponding singular value decompositions

$$B = ( \begin{array}{c} U_1 \\ U_2 \end{array} ) \left( \begin{array}{cc} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{array} \right) \left( \begin{array}{c} V_1^T \\ V_2^T \end{array} \right)$$

and

$$\hat{B} = ( \begin{array}{c} \hat{U}_1 \\ \hat{U}_2 \end{array} ) \left( \begin{array}{cc} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{array} \right) \left( \begin{array}{c} \hat{V}_1^T \\ \hat{V}_2^T \end{array} \right),$$

we have:
Theorem (Stability of subspaces)

If there is an $\bar{\alpha} > 0$ such that

$$\min_{\ell, \hat{\ell}} |\sigma_{\ell}(\hat{\Sigma}_1) - \sigma_{\ell}(\Sigma_2)| \geq \bar{\alpha},$$

and

$$\min_{\hat{\ell}} |\sigma_{\hat{\ell}}(\hat{\Sigma}_1)| \geq \bar{\alpha},$$

then

$$\| V_1 V_1^T - \hat{V}_1 \hat{V}_1^T \|_F \leq \frac{2}{\bar{\alpha}} \| B - \hat{B} \|_F.$$
Wedin’s bound

Applied to our situation, where $X$ has rank $k$ and thus $\Sigma_2 = 0$, we get

$$\| V_1 V_1^T - \hat{V}_1 \hat{V}_1^T \|_F \leq \frac{2\sqrt{m_X \nu_2}}{\sigma_k(\hat{X}^T)},$$

and further since $\sigma_k(\hat{X}^T) \geq \sigma_k(X^T) - \|X - \hat{X}\|_F$, that

$$\| V_1 V_1^T - \hat{V}_1 \hat{V}_1^T \|_F \leq \frac{2\sqrt{m_X \nu_2}}{\sigma_k(X^T) - \sqrt{m_X \nu_2}}.$$
Wedin’s bound

Applied to our situation, where $X$ has rank $k$ and thus $\Sigma_2 = 0$, we get

$$\| V_1 V_1^T - \hat{V}_1 \hat{V}_1^T \|_F \leq \frac{2\sqrt{mX\nu_2}}{\sigma_k(\hat{X}^T)},$$

and further since $\sigma_k(\hat{X}^T) \geq \sigma_k(X^T) - \|X - \hat{X}\|_F$, that

$$\| V_1 V_1^T - \hat{V}_1 \hat{V}_1^T \|_F \leq \frac{2\sqrt{mX\nu_2}}{\sigma_k(X^T) - \sqrt{mX\nu_2}}.$$

Note that

$$X^T = \mathcal{G}A = U_\mathcal{G}\Sigma_\mathcal{G}[V_\mathcal{G}^T A],$$

for $\mathcal{G} = (\nabla g(Ax^1)|\ldots|\nabla g(Ax^{mx}))^T$, hence $\Sigma_{X^T} = \Sigma_\mathcal{G}$. Moreover

$$\sigma_i(\mathcal{G}) = \sqrt{\sigma_i(\mathcal{G}^T \mathcal{G})}, \quad \text{for all } i = 1, \ldots, k.$$
Concentration inequalities II

Theorem (Matrix Chernoff bounds)

Consider $X_1, \ldots, X_m$ independent random, positive-semidefinite matrices of dimension $k \times k$. Moreover suppose $\sigma_1(X_j) \leq C$, almost surely.

Compute the singular values of the sum of the expectations

$$\mu_{\text{max}} = \sigma_1 \left( \sum_{j=1}^{m} \mathbb{E}X_j \right) \quad \text{and} \quad \mu_{\text{min}} = \sigma_k \left( \sum_{j=1}^{m} \mathbb{E}X_j \right),$$

then

$$\mathbb{P} \left\{ \sigma_1 \left( \sum_{j=1}^{m} X_j \right) - \mu_{\text{max}} \geq s \mu_{\text{max}} \right\} \leq k \left( \frac{(1 + s)}{e} \right)^{\frac{\mu_{\text{max}} (1+s)}{C}},$$

for all $s > (e - 1)$, and

$$\mathbb{P} \left\{ \sigma_k \left( \sum_{j=1}^{m} X_j \right) - \mu_{\text{min}} \leq -s \mu_{\text{min}} \right\} \leq ke^{-\frac{\mu_{\text{min}} s^2}{2c}},$$

for all $s \in (0, 1)$. 
Note that
\[ \mathcal{G}^T \mathcal{G} = \sum_{j=1}^{m_X} \nabla g(Ax^j) \nabla g(Ax^j)^T. \]

and by applying the previous result to \( X_j = \nabla g(Ax^j) \nabla g(Ax^j)^T \), we have:

**Lemma**

*For any \( s \in (0, 1) \) we have that*

\[ \sigma_k(X^T) \geq \sqrt{m_X \alpha(1 - s)} \]

*with probability \( 1 - ke^{\frac{-m_X \alpha s^2}{2kC_2^2}} \).*
Proof of Theorem

with probability at least

$$1 - \left( e^{-c_1'm\Phi} + e^{-\sqrt{m\Phi}d} + ke^{\frac{-m\chi^2}{2kC^2}} \right),$$

we have

$$\|V_1 V_1^T - \hat{V}_1 \hat{V}_1^T\|_F \leq \frac{2\nu_2}{\sqrt{\alpha(1-s)} - \nu_2}.$$
with probability at least

\[
1 - \left( e^{-c_1' m \Phi} + e^{-\sqrt{m \Phi} d} + ke^{-\frac{m \chi \alpha s^2}{2kC_2^2}} \right),
\]

we have

\[
\| V_1 V_1^T - \hat{V}_1 \hat{V}_1^T \|_F \leq \frac{2\nu_2}{\sqrt{\alpha(1 - s) - \nu_2}}.
\]

and for \( \hat{A} = \hat{V}_1^T \) and \( V_9^T A = V_1^T \)

\[
\| A^T A - \hat{A}^T \hat{A} \|_F = \| A^T V_9 V_9^T A - \hat{V}_1 \hat{V}_1^T \|_F \leq \frac{2\nu_2}{\sqrt{\alpha(1 - s) - \nu_2}}.
\]
Proof of Theorem ... continue

Since $A$ is row-orthogonal we have $A = AA^T A$ and

$$|f(x) - \hat{f}(x)| = |g(Ax) - \hat{g}(\hat{A}x)|$$

$$= |g(Ax) - g(A\hat{A}^T \hat{A}x)|$$

$$\leq C_2 \sqrt{k} \|Ax - A\hat{A}^T \hat{A}x\|_{\ell^2_k}$$

$$= C_2 \sqrt{k} \|A(A^TA - \hat{A}^T \hat{A})x\|_{\ell^2_k}$$

$$\leq C_2 \sqrt{k} \|(A^TA - \hat{A}^T \hat{A})\|_F \|x\|_{\ell^2_d}$$

$$\leq 2C_2 \sqrt{k}(1 + \bar{e}) \frac{\nu_2}{\sqrt{\alpha(1 - s) - \nu_2}}.$$

where we used

$$\|A^TA - \hat{A}^T \hat{A}\|_F = \|A^T V_9 V_9^T A - \hat{V}_1 \hat{V}_1^T\|_F \leq \frac{2\nu_2}{\sqrt{\alpha(1 - s) - \nu_2}}.$$
\( k \)-ridge functions may be too simple!

\[ f(x) \approx g_1(A_1x) + g_2(A_2x) \]

**Figure**: Functions on data clustered around a manifold with multiple directions can be locally approximated by sums of \( k \)-ridge functions
Can we still be able to learn functions of the type

\[ f(x) = \sum_{i=1}^{m} g_i(a_i \cdot x), \quad x \in [-1, 1]^d? \]
Sums of ridge functions

Can we still be able to learn functions of the type

\[ f(x) = \sum_{i=1}^{m} g_i(a_i \cdot x), \quad x \in [-1, 1]^d? \]

Our approach (Daubechies, F., Vybíral) is essentially based on the formula

\[ D_{c_1}^{\alpha_1} \cdots D_{c_k}^{\alpha_k} f(x) = \sum_{i=1}^{m} g_i^{(\alpha_1 + \cdots + \alpha_k)}(a_i \cdot x)(a_i \cdot c_1)^{\alpha_1} \cdots (a_i \cdot c_k)^{\alpha_k}, \]

where \( k \in \mathbb{N}, \, c_i \in \mathbb{R}^d, \, \alpha_i \in \mathbb{N} \) for all \( i = 1, \ldots, k \) and \( D_{c_i}^{\alpha_i} \) is the \( \alpha_i \)-th derivative in the direction \( c_i \).
The recovery strategy: nearly orthonormal systems

We assume that vectors $a_1, \ldots, a_m \in \mathbb{R}^m$ are nearly orthonormal, to mean that

$$S(a_1, \ldots, a_m) = \inf \left\{ \left( \sum_{i=1}^m \| a_i - w_i \|_2^2 \right)^{1/2} : w_1, \ldots, w_m \text{ orthonormal basis in } \mathbb{R}^m \right\}$$

is small!
The recovery strategy: nearly orthonormal systems

We assume that vectors $a_1, \ldots, a_m \in \mathbb{R}^m$ are nearly orthonormal, to mean that

$$S(a_1, \ldots, a_m) = \inf \left\{ \left( \sum_{i=1}^{m} \| a_i - w_i \|_2^2 \right)^{1/2} : w_1, \ldots, w_m \text{ orthonormal basis in } \mathbb{R}^m \right\}$$

is small!

Furthermore, we denote by

$$L = \text{span}\{a_i \otimes a_i, i = 1, \ldots, m\} \subset \mathbb{R}^{m \times m}$$

the subspace of symmetric matrices generated by tensor products $a_i \otimes a_i = a_i a_i^T$. 
The recovery strategy: **nearly orthonormal** systems

We assume that vectors $a_1, \ldots, a_m \in \mathbb{R}^m$ are **nearly orthonormal**, to mean that

$$S(a_1, \ldots, a_m) = \inf \left\{ \left( \sum_{i=1}^{m} \| a_i - w_i \|_2^2 \right)^{1/2} : w_1, \ldots, w_m \text{ orthonormal basis in } \mathbb{R}^m \right\}$$

is small!

Furthermore, we denote by

$$L = \text{span}\{a_i \otimes a_i, i = 1, \ldots, m\} \subset \mathbb{R}^{m \times m}$$

the subspace of symmetric matrices generated by tensor products $a_i \otimes a_i = a_i a_i^T$.

We first recover an approximation of $L$, i.e. instead of $L$ we have then a subspace $\tilde{L}$ of symmetric matrices at our disposal, which is (in some sense) close to $L$. 
The recovery strategy: nearly orthonormal systems

We assume that vectors $a_1, \ldots, a_m \in \mathbb{R}^m$ are nearly orthonormal, to mean that

$$S(a_1, \ldots, a_m) = \inf \left\{ \left( \sum_{i=1}^m \| a_i - w_i \|_2^2 \right)^{1/2} : w_1, \ldots, w_m \text{ orthonormal basis in } \mathbb{R}^m \right\}$$

is small!

Furthermore, we denote by

$$L = \text{span}\{a_i \otimes a_i, i = 1, \ldots, m\} \subset \mathbb{R}^{m \times m}$$

the subspace of symmetric matrices generated by tensor products $a_i \otimes a_i = a_i a_i^T$.

We first recover an approximation of $L$, i.e. instead of $L$ we have then a subspace $\tilde{L}$ of symmetric matrices at our disposal, which is (in some sense) close to $L$. Finally, we propose the following algorithm

$$\arg \max \| M \|_\infty, \quad \text{s.t. } M \in \tilde{L}, \| M \|_F \leq 1$$

to recover $a_i$’s - or their good approximation $\hat{a}_i$ (which is of course possible only up to the sign).
Nonlinear programming to recover the $a_i \otimes a_i$’s

Figure: The $a_i \otimes a_i$ are the extremal points of the matrix operator norm!
On the ambiguity of learning for nonorthogonal profiles

Let $a_1 = (1, 0)^T$, $a_2 = (\sqrt{2}/2, \sqrt{2}/2)^T$ and $b = (a_1 + a_2)/\|a_1 + a_2\|_2$. 
Let $a_1 = (1, 0)^T$, $a_2 = (\sqrt{2}/2, \sqrt{2}/2)^T$ and $b = (a_1 + a_2)/\|a_1 + a_2\|_2$. We assume that $L = \text{span}\{a_1a_1^T, a_2a_2^T\}$.
On the ambiguity of learning for nonorthogonal profiles

Let $a_1 = (1, 0)^T$, $a_2 = (\sqrt{2}/2, \sqrt{2}/2)^T$ and $b = (a_1 + a_2)/\|a_1 + a_2\|_2$. We assume that $L = \text{span}\{a_1 a_1^T, a_2 a_2^T\}$ and that

$$
\tilde{L} = \text{span}\left\{ \begin{pmatrix} 1 & \epsilon \\ \epsilon & -\epsilon \end{pmatrix}, \begin{pmatrix} 0.5 & 0.5 + \epsilon \\ 0.5 + \epsilon & 0.5 - \epsilon \end{pmatrix} \right\}
$$
On the ambiguity of learning for nonorthogonal profiles

Let \( a_1 = (1, 0)^T \), \( a_2 = (\sqrt{2}/2, \sqrt{2}/2)^T \) and \( b = (a_1 + a_2)/\|a_1 + a_2\|_2 \). We assume that \( L = \text{span}\{a_1 a_1^T, a_2 a_2^T\} \) and that

\[
\tilde{L} = \text{span}\left\{ \begin{pmatrix} 1 & \epsilon \\ \epsilon & -\epsilon \end{pmatrix}, \begin{pmatrix} 0.5 & 0.5 + \epsilon \\ 0.5 + \epsilon & 0.5 - \epsilon \end{pmatrix} \right\}
\]

When choosing \( \epsilon = 0.05 \), we find out that

\[
\{\text{dist}(a_1 a_1^T, \tilde{L}), \text{dist}(a_2 a_2^T, \tilde{L}), \text{dist}(bb^T, \tilde{L})\} \subset [0.07, 0.08].
\]
On the ambiguity of learning for nonorthogonal profiles

Let \( a_1 = (1, 0)^T \), \( a_2 = (\sqrt{2}/2, \sqrt{2}/2)^T \) and \( b = (a_1 + a_2)/\|a_1 + a_2\|_2 \). We assume that \( L = \text{span}\{a_1 a_1^T, a_2 a_2^T\} \) and that

\[
\tilde{L} = \text{span}\left\{ \begin{pmatrix} 1 & \epsilon \\ \epsilon & -\epsilon \end{pmatrix}, \begin{pmatrix} 0.5 & 0.5 + \epsilon \\ 0.5 + \epsilon & 0.5 - \epsilon \end{pmatrix} \right\}
\]

When choosing \( \epsilon = 0.05 \), we find out that

\[
\{\text{dist}(a_1 a_1^T, \tilde{L}), \text{dist}(a_2 a_2^T, \tilde{L}), \text{dist}(bb^T, \tilde{L})\} \subseteq [0.07, 0.08].
\]

Hence, looking at \( \tilde{L} \) alone, every algorithm will have difficulties to decide, which two of the three rank-1 matrices above are the generators of the true \( L \).
Let $a_1 = (1, 0)^T$, $a_2 = (\sqrt{2}/2, \sqrt{2}/2)^T$ and $b = (a_1 + a_2)/\|a_1 + a_2\|_2$. We assume that $L = \text{span}\{a_1 a_1^T, a_2 a_2^T\}$ and that

$$\tilde{L} = \text{span}\left\{\begin{pmatrix} 1 & \epsilon \\ \epsilon & -\epsilon \end{pmatrix}, \begin{pmatrix} 0.5 & 0.5 + \epsilon \\ 0.5 + \epsilon & 0.5 - \epsilon \end{pmatrix}\right\}$$

When choosing $\epsilon = 0.05$, we find out that

$$\{\text{dist}(a_1 a_1^T, \tilde{L}), \text{dist}(a_2 a_2^T, \tilde{L}), \text{dist}(bb^T, \tilde{L})\} \subset [0.07, 0.08].$$

Hence, looking at $\tilde{L}$ alone, every algorithm will have difficulties to decide, which two of the three rank-1 matrices above are the generators of the true $L$. Nevertheless, $\|b - a_1\|_2 = \|b - a_2\|_2 \geq 0.39$. 
On the ambiguity of learning for nonorthogonal profiles

Let $a_1 = (1, 0)^T$, $a_2 = (\sqrt{2}/2, \sqrt{2}/2)^T$ and $b = (a_1 + a_2)/\|a_1 + a_2\|_2$. We assume that $L = \text{span}\{a_1 a_1^T, a_2 a_2^T\}$ and that

$$\tilde{L} = \text{span}\left\{\begin{pmatrix} 1 & \epsilon \\ \epsilon & -\epsilon \end{pmatrix}, \begin{pmatrix} 0.5 & 0.5 \epsilon \\ 0.5 + \epsilon & 0.5 - \epsilon \end{pmatrix}\right\}$$

When choosing $\epsilon = 0.05$, we find out that

$$\{\text{dist}(a_1 a_1^T, \tilde{L}), \text{dist}(a_2 a_2^T, \tilde{L}), \text{dist}(bb^T, \tilde{L})\} \subset [0.07, 0.08].$$

Hence, looking at $\tilde{L}$ alone, every algorithm will have difficulties to decide, which two of the three rank-1 matrices above are the generators of the true $L$. Nevertheless, $\|b - a_1\|_2 = \|b - a_2\|_2 \geq 0.39$. We see that although the level of noise was rather mild, we have difficulties to distinguish between well separated vectors.
The approximation to $L$

Define

$$\tilde{L} = \text{span}\{\Delta f(x_j), j = 1, \ldots, m_x\},$$

where

$$(\Delta f(x))_{j,k} = \frac{f(x + \epsilon(e_j + e_k)) - f(x + \epsilon e_j) - f(x + \epsilon e_k) + f(x)}{\epsilon^2},$$

for $j, k = 1, \ldots, m$, is an approximation to the Hessian of $f$ at $x$. For $x$ drawn at random and by applying in a suitable way the Chernoff matrix bounds, one derives a probabilistic error estimate, in the sense that

$$\|P_L - P_{\tilde{L}}\|_{F \to F} \leq Cm^{3/2}\epsilon,$$

with high probability.
A nonlinear operator towards a gradient ascent

Let us introduce first for a given parameter $\gamma > 1$ an operator acting on the singular values of a matrix $X = U\Sigma V^T$ as follows:

$$
\Pi_\gamma(X) = U \frac{\text{diag}(\gamma, 1, \ldots, 1) \times \Sigma}{\|\text{diag}(\gamma, 1, \ldots, 1) \times \Sigma\|_F} V^T,
$$

where

$$
\text{diag}(\gamma, 1, \ldots, 1) \times \Sigma = 
\begin{pmatrix}
\gamma \sigma_1 & 0 & \ldots & 0 \\
0 & \sigma_2 & 0 & \ldots \\
\vdots & \vdots & \vdots & \ddots \\
0 & \ldots & 0 & \sigma_m
\end{pmatrix}
$$

Notice that $\Pi_\gamma$ maps any matrix $X$ onto a matrix of unit Frobenius norm, simply exalting the first singular value and damping the others. It is not a linear operator.
The nonlinear programming

We propose a *projected gradient method* for solving

\[
\operatorname{arg\ max} \| M \|_\infty, \quad \text{s.t.} \quad M \in \tilde{L}, \| M \|_F \leq 1.
\]

**Algorithm 3:**

- *Fix a suitable parameter* \( \gamma > 1 \)
- *Assume to have identified a basis for* \( \tilde{L} \) *of semi-positive definite matrices, for instance, one can use the second order finite differences* \( \Delta f(x_j), j = 1, \ldots, m_X \) *to form such a basis;*
- *Generate an initial guess* \( X^0 = \sum_{j=1}^{m_X} \zeta_j \Delta f(x_j) \) *by choosing at random* \( \zeta_j \geq 0 \), *so that* \( X^0 \in \tilde{L} \) *and* \( \| X^0 \|_F = 1 \);  
- *For* \( \ell \geq 0 \):
  \[ X^{\ell+1} := P_{\tilde{L}} \Pi_\gamma(X^\ell); \]
Analysis of the algorithm for $\tilde{L} = L$

**Proposition** (Daubechies, F., Vybřal)

Assume that $\tilde{L} = L$ and that $a_1, \ldots, a_m$ are orthonormal. Let $\gamma > \sqrt{2}$ and let $\|X^0\|_\infty > 1/\sqrt{\gamma^2 - 1}$. Then there exists $\mu_0 < 1$ such that

$$|1 - \|X^{\ell+1}\|_\infty| \leq \mu_0 \left|1 - \|X^\ell\|_\infty\right|, \quad \text{for all } \ell \geq 0.$$

Being the sequence $(X^\ell)_\ell$ made of matrices with Frobenius norm bounded by 1, we conclude that any accumulation point of it has both unit Frobenius and spectral norm and therefore it has to coincide with one maximizer.
Analysis of the algorithm for $\tilde{L} = L$

**Proposition** (Daubechies, F., Vybřal)

Assume that $\tilde{L} = L$ and that $a_1, \ldots, a_m$ are orthonormal. Let $\gamma > \sqrt{2}$ and let $\|X^0\|_\infty > 1/\sqrt{\gamma^2 - 1}$. Then there exists $\mu_0 < 1$ such that

$$\left|1 - \|X^{\ell+1}\|_\infty\right| \leq \mu_0 \left|1 - \|X^{\ell}\|_\infty\right|, \quad \text{for all } \ell \geq 0.$$

Being the sequence $(X^{\ell})_\ell$ made of matrices with Frobenius norm bounded by 1, we conclude that any accumulation point of it has both unit Frobenius and spectral norm and therefore it has to coincide with one maximizer.

The proof is based on the following observation

$$\|X^{\ell+1}\|_\infty = \sigma_1(X^{\ell+1}) = \frac{\gamma \sigma_1(X^{\ell})}{\sqrt{\gamma^2 \sigma_1(X^{\ell})^2 + \sigma_2(X^{\ell})^2 + \cdots + \sigma_m(X^{\ell})^2}}$$

$$\geq \frac{\gamma \|X^{\ell}\|_\infty}{\sqrt{(\gamma^2 - 1)\|X^{\ell}\|_\infty^2 + 1}}.$$
Analysis of the algorithm for $\tilde{L} \approx L$

Theorem (Daubechies, F., Vybřal)

Assume for that $\|P_{\tilde{L}} - P_L\|_{F \rightarrow F} < \epsilon < 1$ and that $a_1, \ldots, a_m$ are orthonormal. Let $\|X^0\|_\infty > \max\{\frac{1}{\sqrt{\gamma^2 - 1}}, \frac{1}{\sqrt{2}} + \epsilon + \xi\}$ and $\sqrt{2} < \gamma$. Then for the iterations $(X^\ell)_\ell$ produced by Algorithm 3, there exists $\mu_0 < 1$ such that

$$\limsup_{\ell} |1 - \|X^\ell\|_\infty| \leq \frac{\mu_1(\gamma, t_0, \epsilon) + 2\epsilon}{1 - \mu_0} + \epsilon,$$

where $\mu_1(\gamma, \xi, \epsilon) \approx \epsilon$. The sequence $(X^\ell)_\ell$ is bounded and its accumulation points $\bar{X}$ satisfy simultaneously the following properties

$$\|\bar{X}\|_F \leq 1 \text{ and } \|\bar{X}\|_\infty \geq 1 - \frac{\mu_1(\gamma, t_0, \epsilon) + 2\epsilon}{1 - \mu_0} + \epsilon,$$

and

$$\|P_L \bar{X}\|_F \leq 1 \text{ and } \|P_L \bar{X}\|_\infty \geq 1 - \frac{\mu_1(\gamma, t_0, \epsilon) + 2\epsilon}{1 - \mu_0}.$$
A graphical explanation of the algorithm

Figure: Objective function $\| \cdot \|_\infty$ to be maximized and iterations of Algorithm 3 converging to one of the extremal points $a_i \otimes a_i$
Theorem (Daubechies, F., Vybřal)

Let $M$ be any local maximizer of

$$\arg \max \|M\|_\infty, \quad \text{s.t.} \quad M \in \tilde{L}, \|M\|_F \leq 1.$$

Then

$$u_j^T Xu_j = 0 \quad \text{for all} \quad X \in S_{\tilde{L}} \quad \text{with} \quad X \perp M$$

and all $j \in \{1, \ldots, m\}$ with $|\lambda_j(0)| = \|M\|_\infty$.

If furthermore the $a_i$’s are nearly orthonormal $S(a_1, \ldots, a_m) \leq \epsilon$ and

$$3 \cdot m \cdot \|P_L - P_{\tilde{L}}\| < (1 - \epsilon)^2,$$

then $\lambda_1 = \|M\|_\infty > \max\{|\lambda_2|, \ldots, |\lambda_m|\}$ and

$$2 \sum_{k=2}^m \frac{(u_1^T Xu_k)^2}{\lambda_1 - \lambda_k} \leq \lambda_1.$$
Nonlinear programming

Algorithm 4:

- Let $M$ be a local maximizer of the nonlinear programming
- Take its singular value decomposition $M = \sum_{j=1}^{m} \lambda_j u_j \otimes u_j$
- Put $\hat{a} := u_1$

Theorem (Daubechies, F., Vybřal)

Let $L = \tilde{L}$ and $S(a_1, \ldots, a_m) \leq \varepsilon$. Then there is $j_0 \in \{1, \ldots, m\}$, such that $\hat{a}$ found by Algorithm 4 satisfies $\|\hat{a} - a_{j_0}\|_2 \leq C \sqrt{\varepsilon}$. 
Nonlinear programming

Algorithm 4:
- Let $M$ be a local maximizer of the nonlinear programming
- Take its singular value decomposition $M = \sum_{j=1}^{m} \lambda_j u_j \otimes u_j$
- Put $\hat{a} := u_1$

Theorem (Daubechies, F., Vybráľ)

Let $L = \tilde{L}$ and $\delta(a_1, \ldots, a_m) \leq \varepsilon$. Then there is $j_0 \in \{1, \ldots, m\}$, such that $\hat{a}$ found by Algorithm 4 satisfies $\|\hat{a} - a_{j_0}\|_2 \leq C\sqrt{\varepsilon}$.

The proof is based on testing the optimality condtions for $X = X_j = a_j \otimes a_j$ and showing that $\lambda_1(M) \approx 1$. 
Learning sums of ridge functions

Algorithm 5:

- Let $\hat{a}_j$ are normalized approximations of $a_j, j = 1, \ldots, m$
- Let $(\hat{b}_j)_{j=1}^m$ be the dual basis to $(\hat{a}_j)_{j=1}^m$
- Assume, that $f(0) = g_1(0) = \cdots = g_m(0)$
- Put $\hat{g}_j(t) := f(t\hat{b}_j), t \in (-1/\|\hat{b}_j\|_2, 1/\|\hat{b}_j\|_2)$
- Put $\hat{f}(x) := \sum_{j=1}^m \hat{g}_j(\hat{a}_j \cdot x), \|x\|_2 \leq 1$

Theorem (Daubechies, F., Vybřal)

Let

- $S(a_1, \ldots, a_m) \leq \varepsilon$ and $S(\hat{a}_1, \ldots, \hat{a}_m) \leq \varepsilon'$;
- $\|a_j - \hat{a}_j\|_2 \leq \eta, j = 1, \ldots, m$.

Then

$$\|f - \hat{f}\|_{\infty} \leq c(\varepsilon, \varepsilon') m\eta.$$
Our literature

- S. Mayer, T. Ullrich, and J. Vybíral, *Entropy and sampling numbers of classes of ridge functions*, to appear in Constructive Approximation,