NEW (non-classical) METHODS in APPROXIMATION THEORY

BORIS KASHIN
(Steklov Math. Inst., MSU, MOSCOW)

ULM

September, 2013

Approximation theory — more than 100 years of very active research.

"The main problem of approximation consists in finding, for a complicated function f from a large space X, a close-by, simple function φ from a small subset Φ of X"

R. DeVore, G Lorentz
"Constructive Approximation", Springer, 1993

Two sources of development:

I) Needs of Applications;

II) Aspiration to the Ideal which leads to the attempts to solve the Extremal Problems

Sometimes I) $\Rightarrow$ II) and II) $\rightarrow$ I)

Classical period of A.T.:

large space X - space of functions

$(C(\Delta), L^p(\Delta), \ldots)$

small subset - the set $\mathcal{T}_n$ of all algebraic polynomials of degree $\le n$, or trigonometric polynomials..., or rational functions...

Definition 1 For $f \in C[-1,1]$, $n=1,2,\ldots$

$$E_n(f) \equiv \inf_{P \in \mathcal{T}_n} \|f-P\|_{C[-1,1]}$$

The solution $P = P_n$ of the corresponding extremal problem is called the polynomial of best uniform approximation;

$$R_{n,n}(f) \equiv \inf_{Q \in R_{n,n}} \|f-Q\|_{C[-1,1]}$$

P.L. Chebyshev (1855) trying to solve some applied problem gave a characterization of the polynomial of best uniform approximation to f from $C[-1,1]$ (Chebyshev's alternance)

In the case when $f(x) = x^n$ he founded P_{n-1} explicitely:

$$x^n - P_{n-1} = 2^{-n+1} \cdot \cos(n \arccos x)$$

$$= \frac{1}{2}\left\{ \left(x + \sqrt{x^2-1}\right)^n + \left(x - \sqrt{x^2-1}\right)^n \right\}$$

$n = 1, 2, \ldots$

– Chebyshev polynomials — the solution of the extremal problem

$$\inf_{\{a_{n-1}, \ldots a_0\}} \| x^n + a_{n-1} x^{n-1} + \ldots + a_0 \|_{C[-1,1]} \to \min$$

In the classical A.T. important results concerning approximation order of concrete functions and functional classes were obtained.

Some examples:

a) direct and inverse theorems for polynomial approximation

Th. (Jackson 1912, Bernstein 1912) For $0 < \alpha < 1$

$$f \in \mathrm{Lip}\,\alpha \iff E_n^T(f) = \underset{n \to \infty}{O}(n^{-\alpha})$$

where f is 2π-periodic and

$$\mathrm{Lip}\,\alpha = \{ f \in C(-\pi, \pi),\ \omega(f, \delta) \le C\delta^\alpha \}$$

b) Bernstein, 1913 : there exists $\beta > 0$ such that

$$n \cdot E_n(|x|, C[-1,1]) \to \beta \text{ if } n \to \infty$$

(Bernstein conjecture : $\beta = 1/2\sqrt{\pi}$ - false, $\beta = 0.2801$). The same result for $E_n^T(|x|, (-\pi, \pi))$.

c) Herbert Stahl (1942-2013), in Sbornik Mathematics v. 183, N8, 1992 :

$$e^{\pi\sqrt{n}} \cdot R_{n,n}(|x|, C[-1,1]) \to 8 \text{ if } n \to \infty.$$

Many more important extremal problems were studied and solved in the Classical Approximation Theory.

In the last 30-40 years the main stream of studies in A.T. moved away from Classical Analysis. The main reason is the needs of Applied mathematics.

Now we start to consider the problems of Approximation in much more general setting.

Let B be a normed space, $\Phi \subset B$ be subset of B — we will call it a dictionary, and f be any element of B

Definition 2 n-term approximation of $f \in B$ with respect to the dictionary Φ is

$$e_n(f, \Phi, B) \equiv \inf_{P \in \Sigma_n} \|f - P\|_B ,$$

where for $n = 1, 2, \ldots$

$$\Sigma_n \equiv \left\{ \sum_{j=1}^{n} a_j x_j , \; a_j \in \mathbb{R}, \; x_j \in \Phi \right\}$$

Further, if K is a subset of B, then

$$e_n(K, \Phi, B) = \sup_{f \in K} e_n(f, \Phi, B)$$

Nowdays the notion of n-term approximation is one of the basic notion of Approximation Theory.

Various dictionaries have been considered. Here are some examples important for application:

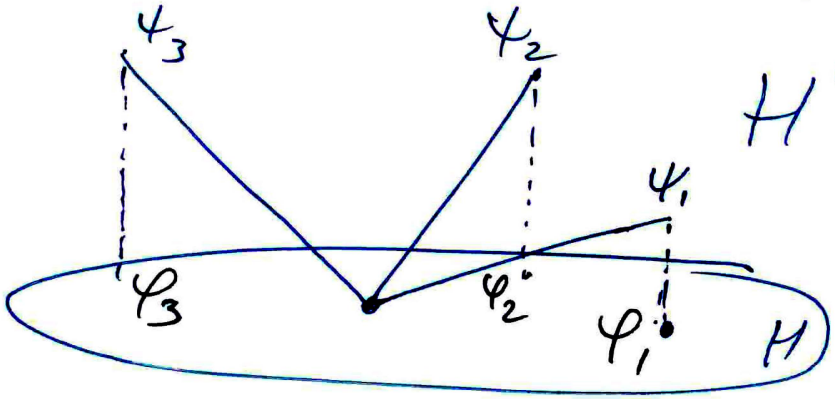
a) Φ is a complete orthonormal set of elements (O.N.S) in some Hilbert space H (H could be the space of functions $L^2(\Omega)$); in particular, Φ can be a trigonometric system

a') Φ is a frame in H

($\Phi = \{\varphi_i\}$ is a frame if for some constants $A > 0$, B and any $f \in H$

$$A\|f\|_H^2 \le \sum_{i=1}^{\infty} \langle f, \varphi_i \rangle^2 \le B\|f\|_H^2;$$

tight frame if $A = B$)



$\overline{\Pi}_{H' \to H}$ operator of orthogonal projection

$\{\psi_i\}$ - O.N.S. in H'; $H \subset H'$, $\varphi_i = \overline{\Pi}_{H' \to H}(\psi_i)$

b) for the approximation of functions of d variables, defined on the cube $I^d = (0,1)^d$ we can consider the dictionary $\mathcal{D}_n$ that consists of all functions of the type

$$u(x_1, \dots x_k) \cdot V(x_{k+1}, \dots x_d); \quad 1 \le k \le d,$$

c) for the approximation of functions of d variables by "free knots splines" we can consider dictionaries that consist of functions of the type

$$P \cdot \chi_\Delta$$

where P is a polynomial of the degree $\le r$ of d variables and χ_Δ is the characteristic function of the segment $\Delta \subset \mathbb{R}^d$

d) $\mathcal{D}$ is the set of ridge functions, i.e functions in $L^p(\Omega)$, $\Omega \subset \mathbb{R}^d$ of the type

$$u(x) = f(\langle x, \theta \rangle)$$

where f is a function of one variable, $x \in \Omega$, $\theta \in \mathbb{R}^d$, $|\theta| = 1$

e) $\mathcal{D}$ is the set of Gabor functions

$$\{ g_{a,b}(x-c); \; g_{a,b}(x) = e^{iax} \cdot e^{-bx^2}; \; a, c \in \mathbb{R} \; b > 0 \}$$

For each of those families of dictionaries there are results in A.T. that are valuable from the theoretical as well as practical point of view.

Following examples support the concept of n-term approximation.

1) Earlier we mentionned that

$$E_n^T(|x|, C[-\pi,\pi]) \asymp \frac{1}{n}$$

if $n \to \infty$

Theorem (Ismagilov, Maiorov... 1970th)

$$e_n(|x|, T, C[-\pi,\pi]) \asymp \frac{1}{n^{3/2}}$$

if $n \to \infty$

In this case the construction of the approximating polynomial is explicit.

So classical approximation is not the best in this case.

2) If Φ – complete O.N.S. in H and $f \in H$ then

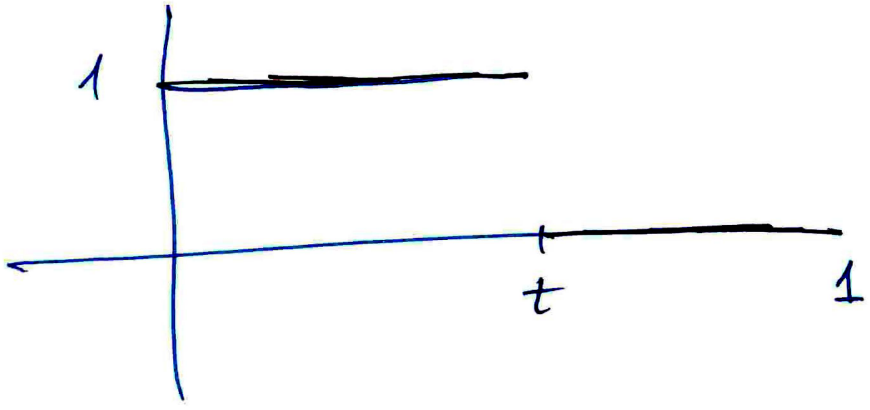
$$e_n(f, \Phi, H) = \left(\sum_{k \geq n+1} |c_k^*(f)|^2 \right)^{1/2},$$

where $\{c_k^*(f)\}_{k=1}^{\infty}$ is a non-increasing rearrangement of the sequence of absolute values of the Fourier coefficients of the function f with respect to complete O.N.S. Φ.

In many cases n-term approximation works much better than classical one.

Example. Let $\chi = \{\chi_t\}_{0<t<1} \subset L^2(0,1)$ where

$$\chi_t(x) = \begin{cases} 1, \text{ if } 0 \leq x \leq t \\ 0, \text{ if } x > t \end{cases}$$



Following result is well known

Theorem (Rudin, 1950th) For any O.N.S. $\Psi = \{\psi_k\} \subset L^2(0,1)$

$$E_n(X, \Psi, L^2(0,1)) \geq c\, n^{-1/2},$$

$c>0$, $n = 1, 2, \ldots$

But if $\Psi = \mathcal{H} = \{h_k\}$ is Haar system, then

$$e_n(X, \mathcal{H}, L^2(0,1)) \leq C\, 2^{-n/2},$$

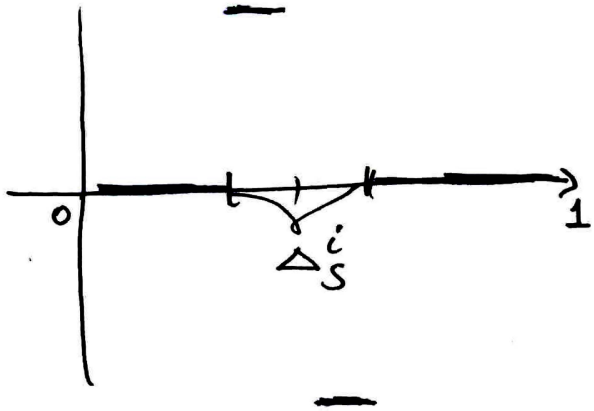
$n = 1, 2, \ldots$

Haar system is historically first and simpliest wavelet system.

Haar functions: $h_1 \equiv 1$ and if $k = 2^s + i$, $i = 1, \ldots 2^s$, $s = 0, 1, \ldots$ then

$$h_k(x) = \begin{cases} 0, & \text{if } x \notin \Delta_s^i \\ 2^{s/2}, & \text{if } x \in (\Delta_s^i)^+ \\ -2^{s/2}, & \text{if } x \in (\Delta_s^i)^- \end{cases}$$

where $\Delta_s^i = \left(\frac{i-1}{2^s}, \frac{i}{2^s}\right)$ and $(\Delta_s^i)^+, (\Delta_s^i)^-$ – left and right haeves of Δ_s^i correspondingly



3) Let us remark that in the previous example the decomposition of any $f \in X$ into Haar series is very sparse.

In general Sparsity is a very important concept in A.T. and applications. If any f from functional class K has sparse representation with respect to O.N.S. Φ then n-term approximation works very well and we can get very good estimates of

$$e_n(K, \Phi, L^2(\Omega)).$$

Current International Standard of Images Compression JPEG 2000 is based on n-term approximation with respect to some 2-d wavelet system Ψ because engineers agree that usual picture has sparse representation in this basis

(usual model of 2-d digital picture is a discrete function of two variables)

Let us consider the problem of approximation of the set

$$K_2 = \{\chi_A,\ A \subset [0,1]^2 - \text{convex}\} \subset$$
$$\subset L^2(0,1)^2$$

This problem is very natural from practical point of view.

For <u>classical</u> O. N. S. Φ

$$e_n(K_2, \Phi, L^2(0,1)^2) \geq C n^{-1/2}, \quad n = 1,2,\dots$$

<u>Theorem</u> (Candes, Donoho, 2002)

There exists tight frame Φ such that

$$e_n(K_2, \Phi, L^2(0,1)^2) \leq C \frac{\ln n}{n}^{3/2}, \quad n = 1,2,\dots$$

(Here for any $f \in K_2$ we consider expansion

$$f = \sum_{k=1}^{\infty} \langle f, \varphi_k \rangle \varphi_k$$

and take as n-term approximation the polynomial of the type

$$\sum_{k \in \Lambda_n} \langle f, \varphi_k \rangle \varphi_k, \quad \#\Lambda_n = n$$

with n biggest (in absolute value) coefficients.

Last result is almost optimal.
To prove it we need a method to get lower estimates for n-term approximation. In the case of orthonormal dictionary (or frame) the following theorem is widely used

Theorem (noncompressibility of hypercubes
B. Kashin, 1985)

Let $\Phi \subset H$ – O.N.S. Suppose that for some $n = 1, 2, \dots$ the set K contains the set of vertices of hypercube:

$$K \supset Q = \left\{ \sum_{i=1}^{2n} \varepsilon_i \psi_i,\ \varepsilon_i = \pm 1,\ \{\psi_i\}_{i=1}^{2n} \text{ – O.N.S} \right\}$$

then

$$e_n(K, \Phi, H) \geq c \cdot n^{1/2},$$

$c > 0$ – absolute constant.

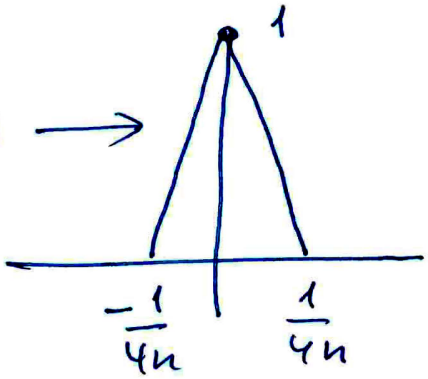
How it works? Let us show using Theorem above that for any O.N.S. Φ

$$e_n(\mathrm{Lip}\,\alpha, \Phi, L^2(0,1)) \geq c \cdot n^{-\alpha},$$
$$0 < \alpha \leq 1.$$

This result shows that in this case n-term approximation is not better than usual approximation by trigonometric polynomials of degree n.

For a given n consider the set of $2n$ functions

$$\psi_i(x) = \psi\left(x - \frac{i}{2n}\right) \qquad i = 1, 2, \dots 2n$$



It is easy to check that for $i = 1, 2, \dots 2n$

$$\|\psi_i\|_{L^2(0,1)} = (6n)^{-1/2}; \quad \psi_i (4n)^{-\alpha} \in \mathrm{Lip}\,\alpha$$

and $\Rightarrow$ for any signs $\varepsilon_i = \pm 1$

$$(8n)^{-\alpha} \cdot \sum_{i=1}^{2n} \varepsilon_i \psi_i \in \mathrm{Lip}\,\alpha$$

So

$$(8n)^{\alpha} \cdot \mathrm{Lip}\,\alpha \supset (6n)^{-1/2} \underbrace{\sum_{i=1}^{2n} \varepsilon_i \psi_i (6n)^{1/2}}_{\text{hypercube}}$$

Th. $\Rightarrow$

$$\sup_{f \in \mathrm{Lip}\,\alpha} e_n(f, \Phi, L^2(0,1)) \geq C n^{-\alpha}$$

□

GREEDY APPROXIMATION

In Applications of A.T. the results on existence of good approximation are not enough. We need a method of constructing such approximation.

That is why following topic became very popular.

We consider only the basic results concerning Greedy Approximation.

(See also V. Temlyakov, "Greedy Approximation", Cambridge Univ. Press, 2011)

Below H is a Hilbert space, $D \subset H$ is a dictionary. Suppose that

$$\|g\|_H = 1 \quad \text{for any } g \in D$$

For $f \in H$ try to find best 1-term approximation with respect to dictionary D

This best 1-term approximation is

$$\langle f, g_0 \rangle g_0$$

where g_0 is a solution of extremal problem

$$|\langle f\, g \rangle| \to \max\,,\ g \in D \qquad (*)$$

(we maximize the angle between f and elements of dictionary and suppose that this solution exists)

The idea of Greedy Approximation is to construct inductively approximants to f such way that at each step we take the best possible "1-term correction" to the already constructed. So on n-th step of Greedy Algorithm we have special n-term approximation and we expect this sequence of approximants to converge to f.

To be precise define:

$$G(f) := G(f, D) = \langle f, g_0 \rangle g_0$$

where g_0 is a solution of extremal problem (X);

$$R(f) := R(f, D) = f - G(f);$$

$$G_0(f) = 0, \quad f_0 \equiv R_0(f) := f$$

Then for $m = 1, 2, \ldots$ we inductively define

$$G_m(f) = G_{m-1}(f) + G(R_{m-1}(f));$$

$$f_m = R_m(f) = f - G_m(f) = R(R_{m-1}(f)).$$

Obviously

$$G_m(f) \in \sum_m = \left\{ \sum_{j=1}^{m} a_j g_j, \; a_j \in \mathbb{R}, \; g_j \in D \right\}.$$

Theorem (Jones, 1987) For any D and $f \in H$

$$G_m(f) \xrightarrow{H} f \quad \text{if} \quad m \to \infty.$$

Another version of G. A.
(weak greedy algorithm):

fix t: $0<t<1$ and instead of $G(f)$ defined by (*) take any element $\tilde{g} \in D$ such that

$$|\langle f, \tilde{g}\rangle| \geq t \sup_{g\in D} |\langle f, g\rangle|$$

If $\tilde{G}(f) = \langle f, \tilde{g}\rangle \tilde{g}$, $\tilde{R}(f) = f - \tilde{G}(f)$

then inductive process completely similar to G. A. converges to f for any $f \in H$ and any dictionary D.

If $D = \Phi = \{\varphi_k\}$ – O. N. S.

and

$$f = \sum a_k \varphi_k \qquad (**)$$

then

$$G_m(f, \Phi) = \sum_{k\in \Lambda_m} a_k \varphi_k$$

where

Λ_m – the support of m "biggest" coefficients of the series (**):

$$\forall k \in \Lambda_m, \forall k' \notin \Lambda_m \quad |a_k| \geq |a_{k'}|.$$

Rate of convergence of G. A.

To precise the setting of the problem we need the following notations:

$$A_1^0(D,M) =$$

$$= \left\{ f \in H : f = \sum_{k \in \Lambda} c_k g_k,\ g_k \in D,\ \sum_{k \in \Lambda} |c_k| \le M,\ \#\Lambda < \infty \right\}$$

$$A_1(D,M) = \text{closure of } A_1^0(D,M)$$

If $D = \{\varphi_k\}$ is O.N.S, then

$$\bigcup_M A_1(D,M) = \left\{ f \in H : \sum a_k \varphi_k = f,\ \sum |a_k| < \infty \right\}$$

In this case for any $f \in A_1(D,M)$

$$\| f - G_m(f) \|_H \le C \cdot M \cdot m^{-1/2}, \quad m = 1,2,\ldots$$

What happenes in the case of general dictionary? It turns out that the rate we can garantee is about $m^{-1/6}$ only.

The first result was obtained by De Vore, Temlyakov:

for any D and $f \in A_1(D, M)$

$$\| f - G_m(f) \|_H \leq C \cdot M \cdot m^{-1/6}, \quad m = 1, 2 \ldots$$

Best known result (Silnichenko)

$$1/6 \to 0.182$$

Lower estimate (Livshitz, Temlyakov):

$\exists D, f$ such that

$$\overline{\lim_{m \to \infty}} \| f - G_m(f) \|_H \cdot m^{0.1898} > 0.$$

A.T. in Compressed Sensing

Compressed Sensing is new and very dynamic branch of signal processing with wide spectrum of Applications. Crucial role in this area is played by Kolmogorov widths estimates.

Definition 3 If E is a normed space, $F \subset E$, then for $n = 0, 1, \ldots$ the Kolmogorov width is

$$d_n(F, E) = \inf_{\substack{L \subset E \\ \dim L \le n}} \sup_{f \in F} \operatorname{dist}_E(f, L),$$

where $\operatorname{dist}_E(f, L) = \inf_{y \in L} \|f - y\|_E$.

This definition was proposed by A.N. Kolmogorov in 1936.

Starting from 1970s a lot of widths estimates were obtained in A.T. Recently it became clear that this results could be succesfully applied in signal processing.

Compressed Sensing refers to the problem of economical recovery of an unknown vector $u \in \mathbb{R}^N$ from the information provided by linear measurements

$$\langle u, \varphi_j \rangle, \quad \varphi_j \in \mathbb{R}^N, \quad j = 1, 2, \dots n$$

in the case when

N is much bigger than n

The natural variant of such a setting uses the concept of sparsity.

Notatious:

$$\ell_p^N = (\mathbb{R}^N, \| \|_{\ell_p^N})$$

where for $x = \{x_j\}_{j=1}^N \in \mathbb{R}^N$

$$\|x\|_{\ell_p^N} = \begin{cases} \left(\sum_1^N |x_j|^p\right)^{1/p}, & \text{if } 1 \le p < \infty \\ \max\limits_{1 \le j \le N} |x_j|, & \text{if } p = \infty \end{cases}$$

$$B_p^N = \{x \in \mathbb{R}^N : \|x\|_{\ell_p^N} = 1\}.$$

We call a vector $u \in \mathbb{R}^N$ k-sparse if it has at most k nonzero coordinates.

We start with a problem: to find for a given pair (N, n) the biggest sparsity $k(N, n)$ such that there exists a set of vectors

$$\varphi_j \in \mathbb{R}^N, \quad j = 1, \dots, n$$

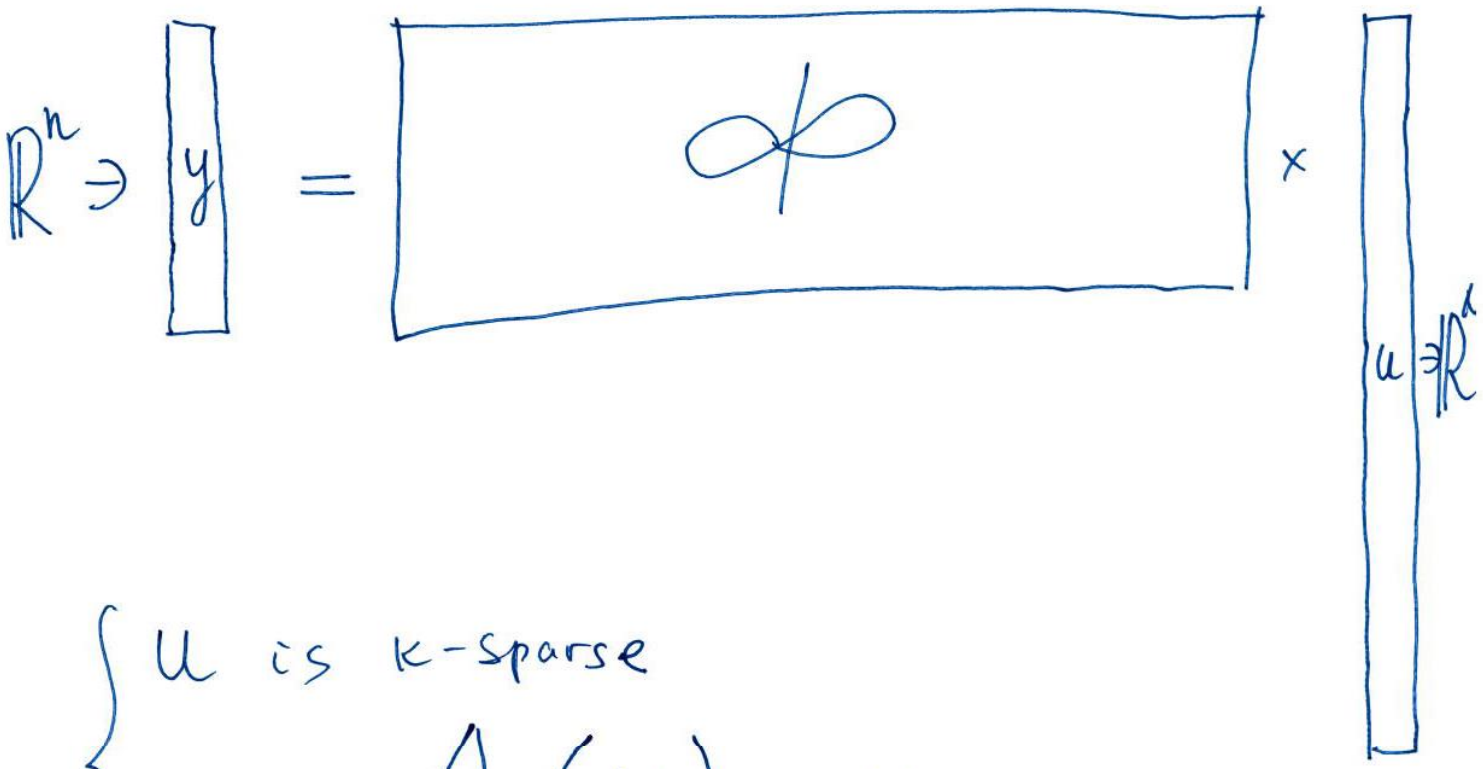
and an economical algorithm $A = A_\alpha$ mapping vector

$$y = (\langle u, \varphi_1 \rangle, \dots, \langle u, \varphi_n \rangle) \in \mathbb{R}^n$$

into $\mathbb{R}^N$ in such a way that for any u of sparsity $k(N, n)$ one would have an exact recovery

$$A(y(u)) = u.$$

In other words we want do describe matrices Φ with rows $\varphi_j \in \mathbb{R}^N$, $j = 1, \dots, n$ such that there exists an economical algorithm of solving the above formulated sparse recovery problem.



$$\begin{cases} u \text{ is } k\text{-sparse} \\ A_\Phi(y) = u \end{cases} \qquad (***)$$

If $2k \leq n$ it is easy to constuct Φ and <u>noneconomical</u> algorithm A such that $(***)$ holds true

But such a solution is useless for Applications

(Such Algorithm is connected with the solution of extremal problem

$$\|v\|_0 \to \min \text{ subject to } \Phi v = y,$$

where $\|v\|_0 := |\mathrm{supp}(v)|$, which is NP hard) P_0

In 2005 - 2006 Donoho and Candes, Romberg, Tao proved that if

$$K \leq C \frac{n}{\ln \frac{2N}{n}} \quad \text{then} \qquad (***)$$

there exists matrix Φ such that P_0 is equivalent to the following extremal problem

$$\|V\|_{\ell_1^N} \to \min \text{ subject to } \Phi V = y \qquad P_1$$

P_1 can be solved by linear programming techniques in a polynomial time

The estimate $(***)$ is best possible.

These results can be obtained using some widths estimates.

Notation For a positive a and $v \in \mathbb{R}^N$.

$$\sigma_a(v)_1 := \min_{\substack{w \in \mathbb{R}^N \\ |\mathrm{supp}(w)| \le a}} \|v - w\|_{\ell_1^N}$$

We say that a measurement matrix Φ has Strong Compressed Sensing Property (SCSP) if for any $u \in \mathbb{R}^N$ we have

$$\|u - A_\Phi(\Phi u)\|_{\ell_2^N} \le C k^{-1/2} \cdot \sigma_k(u)_1$$

for $k \asymp n / \log \frac{2N}{n}$

$$SCSP \Longrightarrow (***)$$

Theorem (Kashin, Temlyakov 2007)

Φ has SCSP if and only if

$$\mathrm{dist}_{\ell_\infty^N}(B_2^N, L_\Phi) \le \frac{C'}{n^{1/2}} \log^{1/2} \frac{2N}{n}$$

where L_Φ is the subspace in $\mathbb{R}^N$ generated by rows of Φ.

The existence of Φ such as in the theorem above follows from widths estimates

Theorem (Kashin, 1977 improved by Garnaev, Gluskin 1984)

$$d_n(B_2^N, \ell_\infty^N) \leq \frac{C}{n^{1/2}} \log^{1/2} \frac{2N}{n}$$

(and random n-dimensional subspace in $\mathbb{R}^N$ provides this estimate)