Approximation Rates for the Hierarchical Tensor Format in Periodic Sobolev Spaces

André Uschmajew (EPF Lausanne)

joint: Reinhold Schneider (TU Berlin)

Pro*Doc, Disentis, August 15, 2013

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

cadmos
center for advanced modeling science

Outline

1. Introduction: Linear approximation theory
2. Bilinear approximation
3. Tree based tensor networks

Outline

Motivation

- **Curse of dimension:**

  High-dimensional problems, e.g. eigenvalue problems for functions of many variables, become intractable when using standard discretization techniques due to the **exponential scaling** of the discretized systems.

- **Example:** Electronic Schrödinger equation $H\Psi = E\Psi$,

$$H = \frac{1}{2}\sum_{i=1}^{N}\Delta_i - \sum_{i=1}^{N}\sum_{\nu=1}^{K}\frac{Z_\nu}{|x_i - a_\nu|} + \frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\frac{1}{|x_i - x_j|},$$

  operates on functions $\Psi \in H^1(\mathbb{R}^{3N})$.

- **Approaches:**
  - **Sparse grids:** Based on *regularity*
  - **Low-rank tensor techniques:** Does regularity also help?

Isotropic Sobolev class:

Let $L_2(\pi_d)$ be the 2π periodic L_2 functions. Consider the following subclass:

$$B^s = \{f \in L_2(\pi_d) \colon \|f\|_s \leq 1\}, \quad \|f\|_s^2 = \max_{\mu=1,2,\ldots,d} \|f\|_{s,\mu}^2,$$

$$\|f\|_{s,\mu}^2 = \sum_{\mathbf{k}\in\mathbb{Z}^d} \bar{k}_\mu^{2s} |\widehat{f}(\mathbf{k})|^2, \quad \text{and} \quad \bar{k}_\mu = \begin{cases} |k_\mu|, & \text{for } k_\mu \neq 0, \\ 1 & \text{for } k_\mu = 0. \end{cases}$$

Regularity and linear approximation

- **Approximation by trigonometric polynomials:**

  Obviously, the best approximation of $f \in L_2(\pi_d)$ in the norm $\|\cdot\|_0$ by a trigonometric polynomial of degree at most n is

$$f_n = \sum_{|\mathbf{k}|_1 \le n} \widehat{f}(\mathbf{k}) \mathrm{e}^{i\mathbf{k}\cdot\mathbf{x}}.$$

- **Approximation error:**

  If $f \in B^s$, then

$$\|f - f_n\|_0^2 = \sum_{|\mathbf{k}|_1 > n} |\widehat{f}(\mathbf{k})|^2 \le n^{-2s} \sum_{|\mathbf{k}|_1 > n} |\mathbf{k}|_1^{2s} |\widehat{f}(\mathbf{k})|^2 \lesssim n^{-2s} \|f\|_s^2 \lesssim n^{-2s}.$$

Regularity and linear approximation

dof complexity:

The number of trigonometric polynomials of degree at most n grows like $\sim n^d$. Thus, to approximate $f \in B^s$ to an accuracy ε, we need an

$$N(\varepsilon) \lesssim \varepsilon^{-d/s} \quad (\varepsilon \to 0)$$

dimensional **linear subspace** in general.

Regularity and linear approximation

- **Kolmogorov N-width:**

  It is well known (Kolmogorov, 1936) that

$$d_N(B^s, L_2(\pi_d)) = \inf_{\substack{V_N \subset L_2(\pi_d) \\ \dim V_N = N}} \sup_{f \in B^s} \inf_{g \in V_N} \|f - g\|_0 \sim N^{-d/s} \quad (N \to \infty).$$

→ Approximation by trigonomeric polynomials is **asymptotically optimal**.

- **Curse of dimension:**

  To keep $N(\varepsilon) \sim \varepsilon^{-d/s}$ tolerable for $\varepsilon \to 0$, the **regularity needs to grow with dimension**:

$$s \sim d.$$

Mixed regularity and linear approximation

A partial way out ...

- **Mixed Sobolev class:**

  Consider functions from

$$B^{s,\text{mix}} = \{f \in L_2(\pi_d) \colon \|f\|_{s,\text{mix}} \le 1\},$$

$$\|f\|_{s,\text{mix}}^2 = \sum_{\mathbf{k} \in \mathbb{Z}^d} \Big(\prod_{\mu=1}^{d} \bar{k}_\mu \Big)^{2s} |\widehat{f}(\mathbf{k})|^2.$$

$\rightarrow$ **Mixed derivatives** up to order ds!

Mixed regularity and linear approximation

- **Hyperbolic cross approximation:**

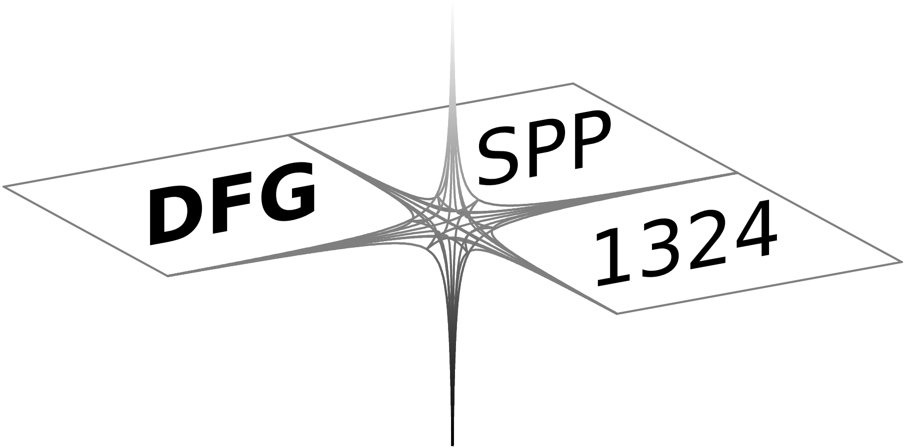
$$f_{\Gamma(n)} = \sum_{\mathbf{k}\in\Gamma(n)} \widehat{f}(\mathbf{k})\mathrm{e}^{i\mathbf{k}\cdot\mathbf{x}},$$

$$\Gamma(n) = \left\{\mathbf{k} \in \mathbb{Z}^d \colon \prod_{\mu=1}^{d} \bar{k}_\mu \leq n\right\}.$$

- **Approximation error:**

  By the same reasoning as before: If $f \in B^{s,\mathrm{mix}}$, then

$$\|f - f_{\Gamma(n)}\|_0 \lesssim n^{-s}.$$


DFG
SPP
1324

Mixed regularity and linear approximation

But this time...

dof complexity:

The space of polynomials with coefficients from the hyperbolic cross has dimension

$$|\Gamma(n)| \sim n^{-s} |\log n|^{s(d-1)} \quad (n \to \infty).$$

It can be shown that it follows

$$N(\varepsilon) \lesssim \varepsilon^{-1/s} |\log \varepsilon|^{d-1} \quad (\varepsilon \to 0).$$

Mixed regularity and linear approximation

- **Kolmogorov N-width:**

  It is known (Babenko, 1960) that

$$d_N(B^{s,\mathrm{mix}}, L_2(\pi_d)) \sim N^{-s} |\log N|^{s(d-1)} \quad (N \to \infty).$$

$\rightarrow$ Approximation by trigonomeric polynomials from the hyperbolic cross is **asymptotically optimal**.

- **Softened curse of dimension:**

  Leads to tolerable complexity at least up to $d = 10$ or so ...

- Yserentant's results: Regularity and approximability of electronic wave functions. Springer-Verlag, Berlin 2010.

Sums of separable functions

- **Tensor product structure:**

  The approximation by trigonometric polynomials yield approximations by a **sum of separable functions**:

$$\sum c_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}} = \sum c_{\mathbf{k}} \prod_{\mu=1}^{d} e^{ik_\mu x_\mu} = \sum_{\mathbf{k}} u^1_{k_1} \otimes \cdots \otimes u^1_{k_d}$$

  with **fixed** choice (dictionary) for the $u^\mu_{k_\mu}$.

- **Question:** Is there a (general) gain in complexity by not restricting the factors $u^\mu_{k_\mu}$ a priori, and if yes, in which function classes?

→ An appropriate, **non-tautological** answer is currently unknown.

- **In this talk:**

  The answer is probably: **asymptotically No** in the classes B^s and $B^{s,\mathrm{mix}}$.
  It is believed that the classical notion of smoothness is not appropriate.

Outline

Bilinear approximation

- **Best bilinear approximation error:**

  For $d \geq 2$, $1 \leq a < d$, $\mathbf{y} = (x_1, \ldots, x_a)$, $\mathbf{z} = (x_{a+1}, \ldots, x_d)$ let

$$\tau_R(f,a) = \inf_{\substack{u_1,\ldots,u_R \in L_2(\pi_a) \\ v_1,\ldots,v_R \in L_2(\pi_{d-a})}} \left\| f(\mathbf{x}) - \sum_{k=1}^{R} u_k(\mathbf{y}) v_k(\mathbf{z}) \right\|_0$$

- **How large one has to choose the rank R?**

  Study the quantities

$$\sup_{f \in B^s} \tau_R(f,a), \qquad \sup_{f \in B^{s,\mathrm{mix}}} \tau_R(f,a).$$

An equivalent formulation:

Let

$$A_f : L_2(\pi_{d-a}) \to L_2(\pi_a), \quad (A_f v)(\mathbf{y}) = \int f(\mathbf{y},\mathbf{z})\overline{v(\mathbf{z})}\, \mathrm{d}\mathbf{z}$$

denote the assosiated **Hilbert-Schmidt integral operator**. Then the bilinear approximation problem is equivalent to

$$\inf_{\operatorname{rank} A \le r} \|A_f - A\|_{HS}.$$

A link to operator ideals

- **"Solution": Schmidt expansion (SVD):**

  Let

$$A_f = \sum_{k=1}^{\infty} \sigma_k u_k \otimes v_k, \quad \{u_k\}, \{v_k\} \text{ ONS}, \ \sigma_k \geq 0,$$

  then

$$A = \sum_{k=1}^{R} \sigma_k u_k \otimes v_k$$

  satisfies

$$\tau_R(f,a) = \|A_f - A\|_{HS} = \sqrt{\sum_{k=R+1}^{\infty} \sigma_k^2}.$$

→ Need **singular value estimates** of integral operators with kernels from Sobolev classes

→ Close link to the theory of **operator ideals**.

Temlyakov's results

- In a series of papers (1986-1993) Temlyakov proved (amongst much more general results on L_p):

$$\sup_{f \in B^s} \tau_R(f,a) \sim R^{-s\max(1/a,1/(d-a))} \quad (R \to \infty),$$

$$R^{-2s}(\log R)^{2s(\min(a,d-a)-1)} \lesssim \sup_{f \in B^{s,\mathrm{mix}}} \tau_R(f,a) \lesssim R^{-2s}(\log R)^{2s(\max(a,d-a)-1)}$$

→ **Required rank $R(\varepsilon)$:**

Let $R(\varepsilon)$ denote the smallest r needed for accuracy ε, then

$$R(\varepsilon) \begin{cases} \sim \varepsilon^{-\min(a,d-a)/s} \quad (\varepsilon \to 0) & \text{for } f \in B^s, \\ \lesssim \varepsilon^{-1/(2s)} |\log \varepsilon|^{\max(a,d-a)-1} \quad (\varepsilon \to 0) & \text{for } f \in B^{s,\mathrm{mix}}. \end{cases}$$

A gain?

- **Number of required separable functions:** Example d even, $a = d/2$:

	$N(\varepsilon)$	$R(\varepsilon)$
$f \in B^s$	$\sim \varepsilon^{-d/s}$	$\sim \varepsilon^{-d/(2s)}$
$f \in B^{s,\text{mix}}$	$\sim \varepsilon^{-1/s}\lvert\log\varepsilon\rvert^{d-1}$	$\sim \varepsilon^{-1/2s}\lvert\log\varepsilon\rvert^{d/2-1}$

- **BUT:**

  While $N(\varepsilon)$ measures computational complexity (number of basis functions from a **fixed** basis), $R(\varepsilon)$ does not yet:

  Since **the singular vectors u_k, v_k are not known in advance**, we need to make sure we can approximate and store them efficiently.

  Griebel, Harbrecht: Approximation of bi-variate functions: singular value decomposition versus sparse grids, IMA J. Numer. Anal. 2013.

Approximation of singular vectors

- If we approximate u_k, v_k by $\tilde{u}_k, \tilde{v}_k$ to an accuracy to **accuracy** $\varepsilon R(\varepsilon)^{-1/2}/\sigma_k$ and put

$$\tilde{f} = \sum_{k=1}^{R(\varepsilon)} \sigma_k \tilde{u}_k \otimes \tilde{v}_k,$$

then

$$\|f - \tilde{f}\| \lesssim \varepsilon.$$

(Griebel and Harbrecht, 2013)

- **How many degrees of freedom do we have to spend to achieve this accuracy?**

Regularity of singular vectors

- **Mapping properties of integral operators:**

  The left singular vectors satisfy

$$\sigma_k u_k(y) = (A_f v_k)(y) = \sigma_k \int f(y,z) v_k(z)\,\mathrm{d}z$$

$\rightarrow \quad \|u_k\|_s \leq \|f\|_{s,1}/\sigma_k$ (similar for v_k)

- **Linear approximation:**

  Approximate u_k to accuracy $\varepsilon R(\varepsilon)^{-1/2}/\sigma_k$ requires (in general) $\sim (\varepsilon R(\varepsilon)^{-1/2})^{-1/s}$ dofs. This we have to do $2R(\varepsilon)$ times

$$\rightarrow \quad \mathbf{dof}(\varepsilon) \lesssim \varepsilon^{-1/s} R(\varepsilon)^{1+1/(2s)}.$$

Comparison

- **Required degrees of freedom:** Example $d = 2$, $a = 1$:

	$N(\varepsilon)$	$R(\varepsilon)$	$\mathrm{dof}(\varepsilon)$
$f \in B^s$	$\sim \varepsilon^{-2/s}$	$\sim \varepsilon^{-1/s}$	$\sim \varepsilon^{-2/s}\varepsilon^{-1/(2s^2)}$
$f \in B^{s,\mathrm{mix}}$	$\sim \varepsilon^{-1/s}\lvert\log \varepsilon\rvert$	$\sim \varepsilon^{-1/2s}$	$\sim \varepsilon^{-1/s}\varepsilon^{-1/(2s)-1/(4s^2)}$

Asymptotically, we lose!

- Does not even include the cost to compute the approximations.

Outline

Minimal subspaces

Recursively split variables...

- **Integral operators:**

  Let $f \in L_2(\pi_d)$, $t = \mu_1, \ldots, \mu_{|t|} \subsetneq 1, 2, \ldots, d$. Set $\mathbf{x}^t = (x_{\mu_1}, \ldots, x_{\mu_{|t|}})$ and $t^c = \{1, \ldots, d\} \setminus t$. Define the integral operator

$$(A_f^t v)(\mathbf{x}^t) = \int f(\mathbf{x}^t, \mathbf{x}^{t^c}) \overline{v(\mathbf{x}^{t^c})} \, \mathrm{d}\mathbf{x}^{t^c}$$

- **Minimal t-subspace:** **t-rank:**

  $U_f^t := \operatorname{ran}(A_f^t)$ $\operatorname{rank}_t(f) = \operatorname{rank}(A_f^t) = \dim(U_f^t)$

  Both definitions are due to Hitchcock (1927).

Nestedness

Main observation:

Let $t = t_1 \dot\cup t_2 \dot\cup \ldots \dot\cup t_N$, then

$$U_f^t \subseteq U_f^{t_1} \otimes U_f^{t_2} \otimes \cdots \otimes U_f^{t_N}.$$

In particular, if $t = \{1, 2, \ldots, d\}$ then

$$f \in U_f^{t_1} \otimes U_f^{t_2} \otimes \cdots \otimes U_f^{t_N}.$$

This **nestedness** is the starting point for the hierarchical Tucker representations.

Hierarchical tensor format

Hackbusch & Kühn (2009), Grasedyck (2010), Oseledets & Tyrtyshnikov (2009), Quantum chemistry ...

Dimension tree:

$T \subseteq 2^{\{1,2,\ldots,d\}}$ is called a **dimenson tree**, if

(i) the root is $t_r = \{1,2,\ldots,d\} \in T$,

(ii) every node $t \in T$ that is not a leaf has at least two *nonempty* sons $t_1, t_2, \ldots, t_{n_t} \in T$ such that $t = t_1 \cup t_2 \cup \cdots \cup t_{n_t}$ **is a disjoint union**,

(iii) the leaves are $\{\mu\}$, $\mu = 1,2,\ldots,d$.

- **HT format:**

  Let T be a dimension tree and $\mathbf{r} = (r_t)_{t \in T \setminus \{t_r\}}$ a set of **ranks** $r_t \in \mathbb{N} \cup \{+\infty\}$. Let $r_{t_r} = 1$ for the root. $f \in L_2(\pi_d)$ is $(T, \mathbf{r})$**-decomposable** if it can be decomposed in the following form.

  (i) To every node $t \in T \setminus \{t_r\}$ an r_t**-dimensional subspace** $U^t \subset L_2(\pi_{|t|})$ is associated in form of a **basis** $u_1^t, u_2^t, \ldots, u_{r_t}^t$. For the root let $u_1^{t_r} = f$.

  (ii) For every node $t \in T$ having sons $t_1, t_2, \ldots, t_{n_t}$ there exists a **transfer tensor** $\beta^t \in \mathbb{R}^{r_t \times r_{t_1} \times r_{t_2} \times \cdots \times r_{n_t}}$ such that it holds

$$u_k^t(\mathbf{x}^{t_1}, \mathbf{x}^{t_2}, \ldots, \mathbf{x}^{t_{n_t}}) = \sum_{k_1=1}^{r_{t_1}} \sum_{k_2=1}^{r_{t_2}} \cdots \sum_{k_{n_t}=1}^{r_{t_{n_t}}} \beta^t_{k,k_1,k_2,\ldots,k_{n_t}} u_{k_1}^{t_1}(\mathbf{x}^{t_1}) u_{k_2}^{t_2}(\mathbf{x}^{t_2}) \cdots u_{k_{n_t}}^{t_{n_t}}(\mathbf{x}^{t_{n_t}}).$$

- The set of $(T, \mathbf{r})$-decomposable functions will be denoted by $\mathscr{H}_{\leq \mathbf{r}, T}$.

Hierarchical tensor format

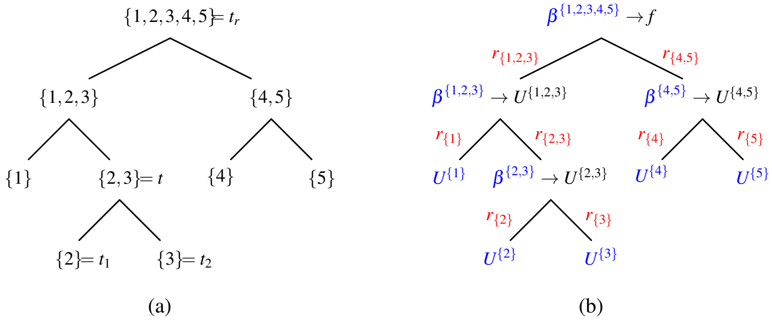
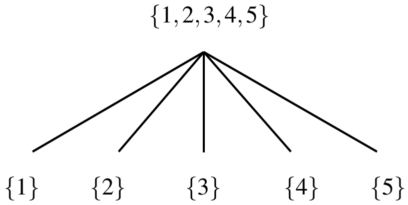


Figure : (a) A binary dimension tree for $\{1,2,3,4,5\}$, (b) parameters of the $(T,\mathbf{r})$-decomposition.

Tucker format



$f \in \mathscr{H}_{\leq \mathbf{r}, T}$ can be written as

$$f(x_1, \dots, x_d) = \sum_{k_1=1}^{r_1} \cdots \sum_{k_d=1}^{r_d} \beta_{k_1, \dots, k_d} u^1_{k_1}(x_1) \cdots u^d_{k_d}(x_d).$$

How to obtain approximations from $\mathscr{H}_{\leq \mathbf{r},T}$?

- **Single SVD projection:**

  Let $f^t(\mathbf{x}^t, \mathbf{x}^{t^c}) = \sum_{k_t=1}^{\infty} \sigma_{k_t}^t u_{k_t}^t(\mathbf{x}^t) v_{k_t}^t(\mathbf{x}^{t^c})$ be an **SVD at node** t. Let P_f^{t,r_t} be the **orthogonal projection onto** $\mathrm{span}\{u_1^t, \ldots, u_{r_t}^t\} \otimes \mathrm{span}\{v_1^t, \ldots, v_{r_t}^t\}$, that is,

$$P_f^{t,r_t} f = \sum_{k_t=1}^{r_t} \sigma_{k_t}^t u_{k_t}^t \otimes v_{k_t}^t.$$

- **HOSVD projection:**

  From leaves to root ...

$$P_f^{\mathbf{r}} = P_{f,L}^{\mathbf{r}} P_{f,L-1}^{\mathbf{r}} \cdots P_{f,1}^{\mathbf{r}}, \quad \text{with} \quad P_{f,l}^{\mathbf{r}} = \prod_{\mathrm{level}(t)=l} P_f^{t,r_t}.$$

$\rightarrow \quad P_f^{\mathbf{r}} f \in \mathscr{H}_{\leq \mathbf{r},T}$

Approximation error

- **Quasi-optimality of HOSVD:**

$$\|f - P_f^{\mathbf{r}} f\|_0^2 \leq \sum_{t \in T \setminus \{t_r\}} \|f - P_f^{t,r_t} f\|_0^2$$
$$= \sum_{t \in T \setminus \{t_r\}} \sum_{k_t \geq r_t + 1} (\sigma_{k_t}^t)^2 \leq (|T| - 1) \inf_{g \in \mathscr{H}_{\leq \mathbf{r},T}} \|f - g\|_0^2$$

De Lathauwer et al. (2000), Grasedyck (2010)

- It follows:

$$\tau_{\mathbf{r}}(f, T) = \inf_{\mathscr{H}_{\leq \mathbf{r},T}} \|f - g\| \sim \sum_{t \in T \setminus \{t_r\}} \tau_{r_t}(f, |t|)$$

Required degrees of freedom

Play the same game as before...

- **Required ranks:**

  Use Temlyakov's results on **bilinear approximation** to esimate the required ranks $r_t(\varepsilon)$ to achieve error ε in every term of $\sum_{t\in T\setminus\{t_r\}} \tau_{r_t}(f,|t|)$.

- **Overall cost of the hierarchical format:**

$$\mathrm{dof}(\varepsilon) \leq \sum_{t\,\in\,T\text{ not leaf}} r_t(\varepsilon)\prod_{i=1}^{n_t} r_{t_i}(\varepsilon) \qquad \text{(size of transfer tensors } \beta^t)$$

$$+\sum_{\mu=1}^{d} \text{dof to approximate } u_1^{\{\mu\}},\dots,u_{r_{\{\mu\}}(\varepsilon)}^{\{\mu\}} \text{ in the leaves}$$

- For the basis functions in the leaves, exploit again their **regularity**.

Results

- **Required degrees of freedom:**

  Let $\deg(T)$ denote the maximum degree of a node in T (number of sons + 1).

	$N(\varepsilon)$	$\mathrm{dof}(\varepsilon)$
$f \in B^s$	$\sim \varepsilon^{-d/s}$	$\begin{cases} \lesssim \varepsilon^{-d/s}, & \text{if } d > 2+1/(2s) \\ \sim \varepsilon^{-(2+1/(2s))/s}, & \text{else} \end{cases}$
$f \in B^{s,\mathrm{mix}}$	$\sim \varepsilon^{-1/s} \lvert \log \varepsilon \rvert^{d-1}$	$\begin{cases} \lesssim \varepsilon^{-\deg(T)/(2s)} \lvert \log \varepsilon \rvert^{N(T)}, & \text{if } \deg(T) \geq 3+1/(2s) \\ \lesssim \varepsilon^{-3/(2s)} \varepsilon^{-1/(4s^2)} \lvert \log \varepsilon \rvert^{(1+1/(2s))(d-2)}, & \text{else} \end{cases}$

Asymptotically, we lose!

- Does not even include the cost to compute the approximations.

Schneider & U. Preprint 2013

Discussion

- Only upper bounds for asymptotic rates...
- **Sparse transfer tensors?**

  The estimates are **upper bounds** and not necessarily sharp: For example $f_{\Gamma(n)}$ is a Tucker approximation with a sparse core tensor (hyperbolic cross).

  For binary trees it seems it would not help in the worst case!

- **Unfair comparison:**

  The mixed Sobolev spaces are by definition tailored to hyperbolic cross approximation.

- **Black-box character / universality of HOSVD:**

  For specific, **irregular** functions it might be much better (characteristic function on square). Given that, it could be worse :-)

  Open problem: What are the right function classes for tensor approximation?

Some remarks on the canonical format

- **Canonical low-rank approximation:**

  Isn't the following more natural to consider?

$$\inf \left\| f - \sum_{k=1}^{R} u_k^1 \otimes \cdots \otimes u_k^d \right\|_0$$

- **Again Temlyakov:**

$$\sup_{f \in B^{s,\text{mix}}} \inf \left\| f - \sum_{k=1}^{R} u_k^1 \otimes \cdots \otimes u_k^d \right\|_0 \lesssim R^{-sd/(d-1)}$$

No curse of dimension in the number of terms!

- **Even U. (2011):**

  If $f \in B^{s,\text{mix}}$ and $\|f - \sum_{k=1}^{R} u_k^1 \otimes \cdots \otimes u_k^d\|_0 =$ min, then all $u_k^\mu \in H^s$.

  $\rightarrow$ **Approximability!?**

The problem of instability

- **But...**

  When $d \geq 3$, then for given $r \geq 2$ a best approximation,

$$\left\| f - \sum_{k=1}^{R} u_k^1 \otimes \cdots \otimes u_k^d \right\|_0 = \min,$$

  might not exist!

  cf. De Silva & Lim 2008

- **Ill conditioning:**

  It is in line with this fact that

  - **No stable method** to calculate a solution close to the infimum is known.
  - **No reasonable bound on Sobolev norms** for the factors could be given in my paper, even if existence of a minimum is assumed.