Axis-alignment in low-rank and other structures

Nick Trefethen, Oxford and ENS Lyon

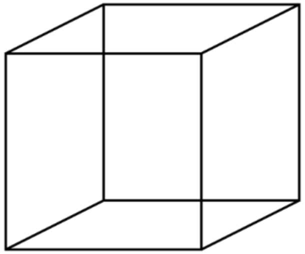
1. Low-rank approximation
2. Quasi-Monte Carlo
3. Sparse grids
4. Multivariate polynomials

Paper: "Cubature, approximation, and isotropy in the hypercube," *SIAM Review,* 2017

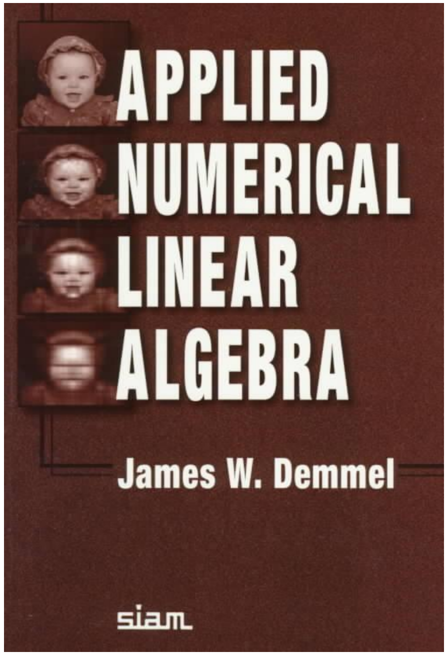
Figures in this talk: many stolen from the internet without attribution

1. Low-rank approximation

dimension s



$$f_r(\mathbf{x}) \approx \sum_{i=1}^{r} g_1^{(i)}(x_1)\, g_2^{(i)}(x_2) \cdots g_s^{(i)}(x_s)$$



APPLIED NUMERICAL LINEAR ALGEBRA

James W. Demmel

siam

1997

NETFLIX Challenge

Which one are your favourites?

STRANGER THINGS
LOVE
SENSE8
THE CROWN
CRAZY EX-GIRLFRIEND
THE GILMORE GIRLS
UNBREAKABLE KIMMY SCHMIDT
NEW GIRL

APARTMENT 23
GRACE AND FRANKIE
ORANGE IS THE NEW BLACK
13 REASONS WHY
NARCOS
GLOW
THE END OF THE F*** WORLD
WET HOT AMERICAN SUMMER

2007

2017

A few names:

Absil, Bader, Bebendorf, Beckermann, Beylkin, Carvajal, Chapman, Dahmen, de Lathauwer, DeVore, Dolgov, Drineas, Geddes, Goreinov, Gorodetsky, Grasedyck, Hackbusch, Halko, Hashemi, Kannan, Karaman, Khoromskij, Kolda, Kressner, Kroonenberg, Kühnemund, Kuske, Mahoney, Martinsson, Marzouk, Mohlenkamp, Oseledets, Rjasanow, Savostyanov, Süli, Tobler, Townsend, Tropp, Tyrtyshnikov, Uschmajew, Vandereycken, Vandewalle, Wilber, Zamarashkin,…

Aside

In 2D, low-rank approximation with greedy selection of pivots amounts to *Gaussian elimination as an iterative rather than direct algorithm.* So just like conjugate gradients, GE was born direct but turns out also to be iterative.

I got into this through chebfun: numerical computing with functions.

In 1D, Chebfun represents functions to 16 digits by Chebyshev series.
What to do in 2D? Bivariate Chebyshev series (= tensor products) ?

I suggested to Alex Townsend that he try low-rank approximations instead. He built Chebfun2 (2013).

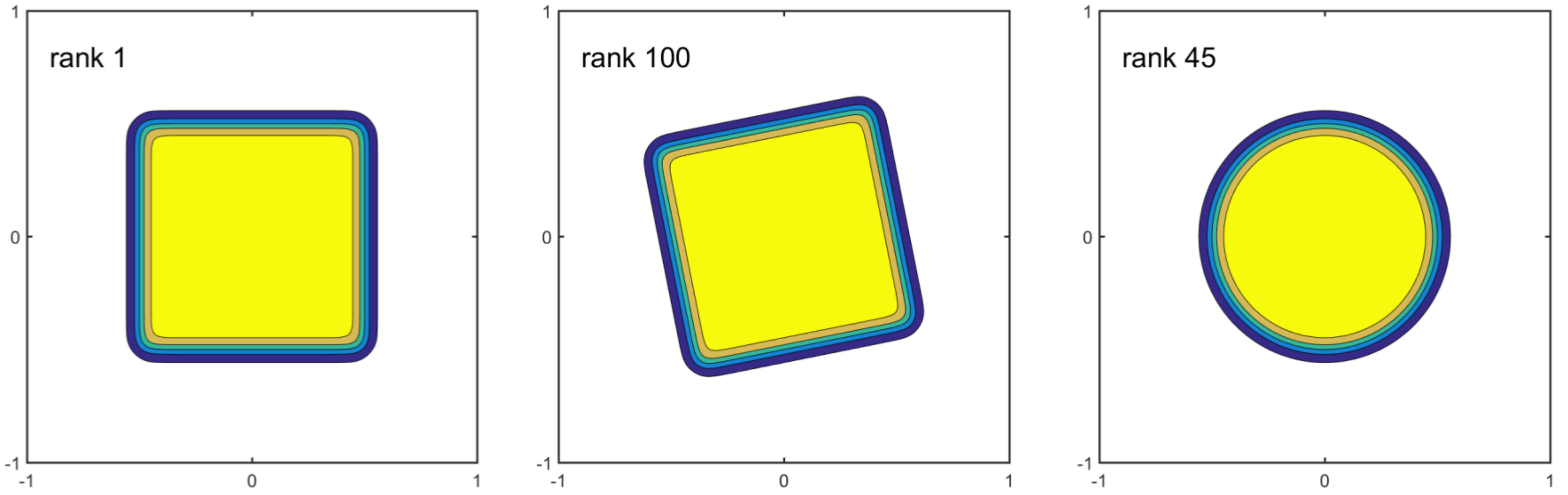
Alex Townsend
Cornell U.

Later Behnam Hashemi took us to 3D, using low-rank tensors. He built Chebfun3 (2017).

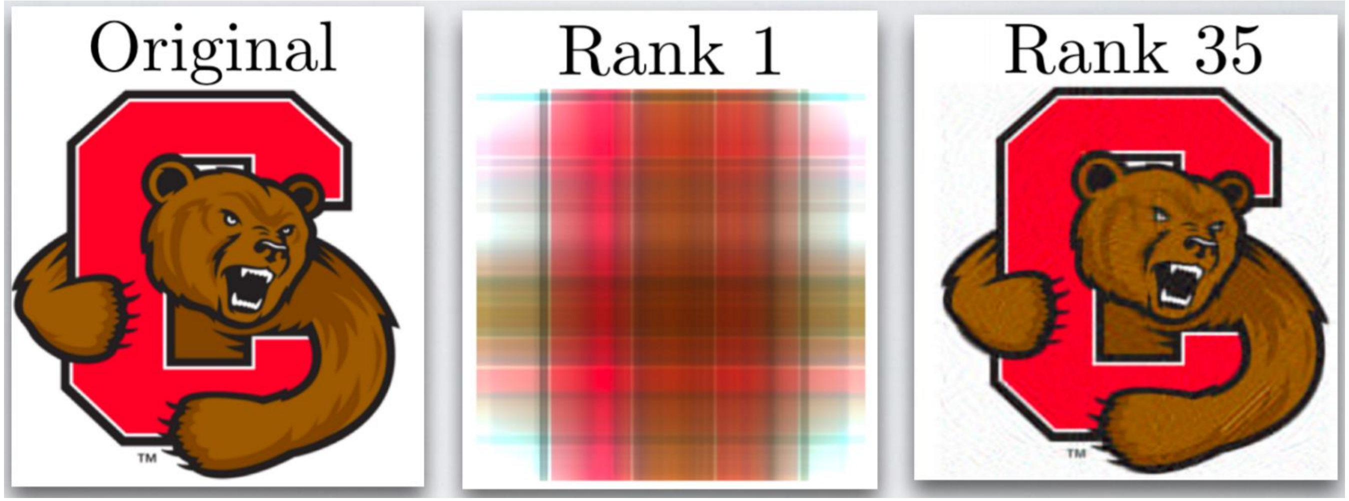
Behnam Hashemi
Shiraz U.

So, are the low-rank representations worthwhile?
Are they more efficient than tensor products?

The answer is: often yes, often no.
I believe the main reason for the “yes” cases is that interesting functions $f(x,y)$ or $f(x,y,z)$ often have some alignment with the axes.



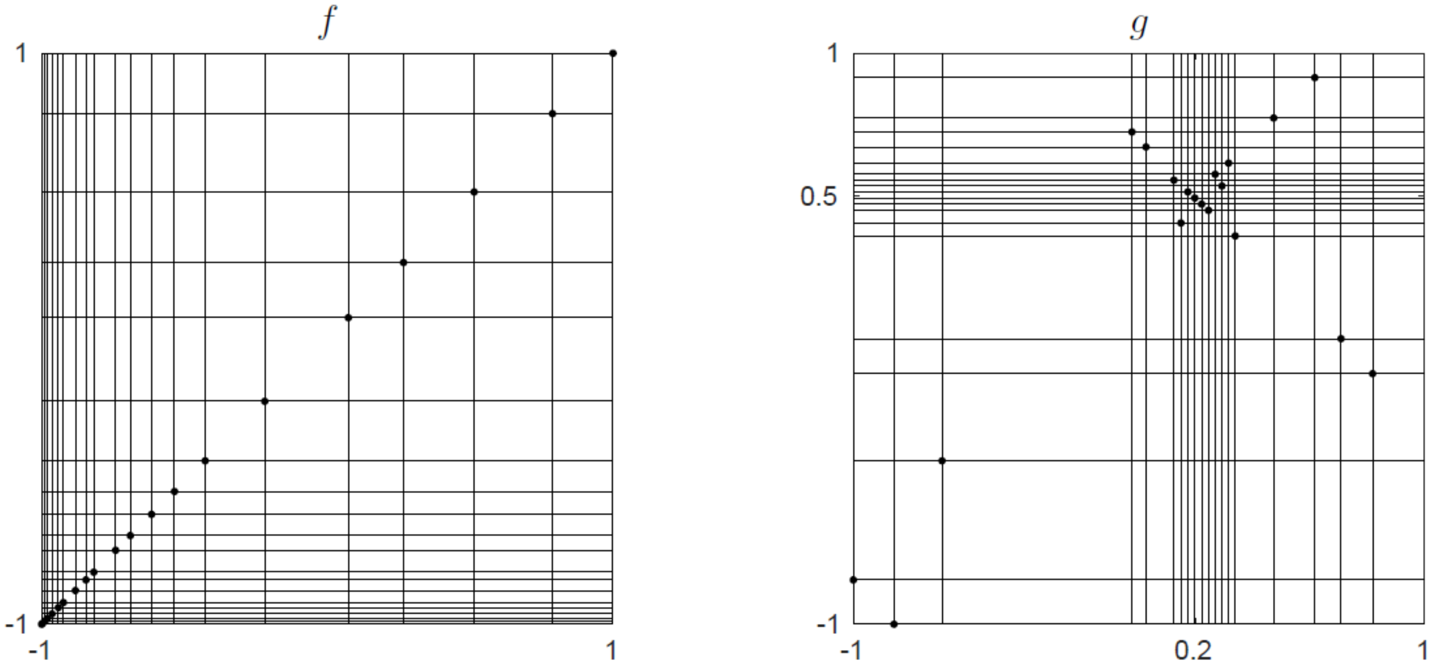
Low-rank compression methods are nonisotropic.



Alex Townsend’s compression of the Cornell bear

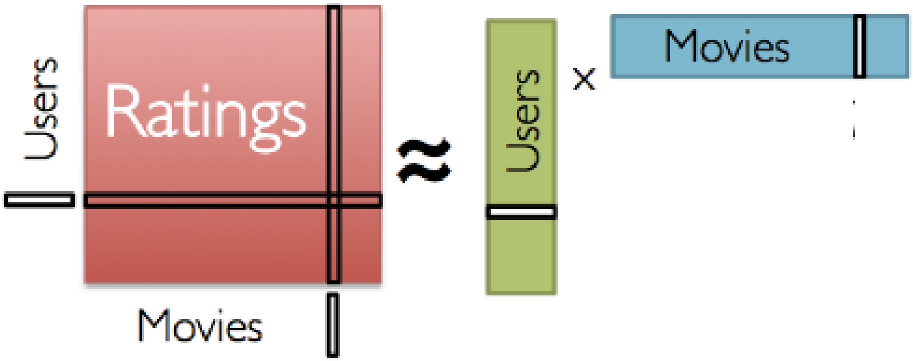
Function	Chebfun2 rank	Bivariate series size*	Comment
$\cos(100x)$	1	149 × 1	aligned
$\cos(100(x+y)/\sqrt{2}\,)$	2	115 × 115	lucky algebra
$J_0(100x)$	1	141 × 1	aligned
$J_0(100(x+y)/\sqrt{2}\,)$	61	115 × 115	no lucky algebra
$\tan(5x)$	1	116 × 1	aligned
$\tan(5(x+y)/\sqrt{2}\,)$	55	88 × 88	not aligned
$(1+25x^2)^{-1}$	1	185 × 1	aligned
$(1+25(x^2+y^2))^{-1}$	19	185 × 185	localized

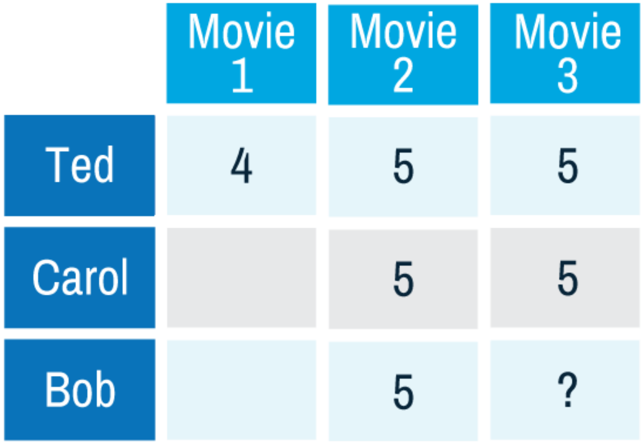
*These numbers should be divided by $\pi/2$ for a fair comparison, because polynomial series always waste a factor of $\pi/2$.


f
1
-1
-1
1
g
1
0.5
-1
-1
0.2
1

Functions vs. data

Sometimes alignment with axes is baked-in:





	Movie 1	Movie 2	Movie 3
Ted	4	5	5
Carol		5	5
Bob		5	?

One would hardly think of “rotating coordinates” to the new variables

$$\frac{\text{users} + \text{movies}}{\sqrt{2}}, \quad \frac{\text{users} - \text{movies}}{\sqrt{2}} \qquad !$$

But often the situation is not so clear. For example, what about —

- A 2D image? A (3D?) movie?
- A stochastic PDE with 20 parameters?
- A tensor-train electronic structure calculation with $s = 100$?
- A financial instrument represented in dimension $s = 360$?

The literature

Many papers prove theorems to show their low-rank methods are effective. Here are three ways in which this is done:

Asymptotic smoothness, for f dominated by singularities near the boundary.
Bebendorf, Brandt, Hackbusch, Tyrtyshnikov,….

PDE with low-rank right-hand side, where the solution inherits this structure.
Penzl, Dahmen-DeVore-Grasedyck-Süli,….

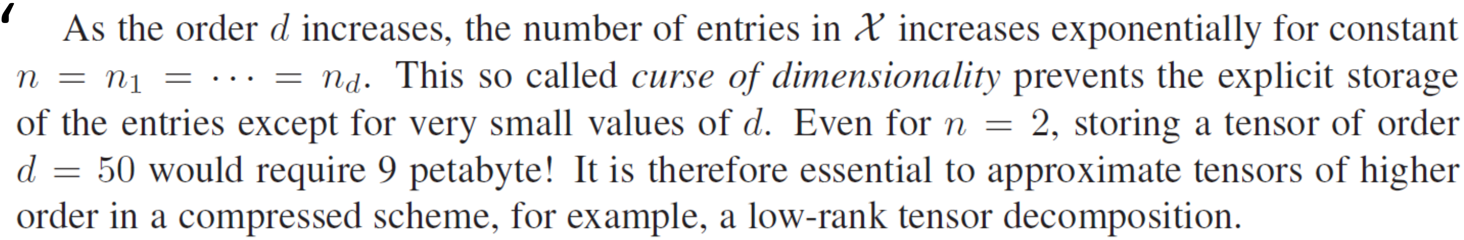
Hierarchical representations, where f is smooth away from the diagonal and one exploits this with an approximation built recursively of low-rank pieces.
Calderon-Zygmund, Beylkin-Coifman-Rokhlin, Greengard-Rokhlin, Hackbusch,….

No theorems, however, show that arbitrary functions can be compressed. Analogously, random matrices have approximately full numerical rank
(see e.g. Edelman 1988).

An analogy and a plea

The impossibility of compressing arbitrary data is not often highlighted. Perhaps typical is this statement from the excellent 2013 survey by Grasedyck-Kressner-Tobler :

“



As the order d increases, the number of entries in $\mathcal{X}$ increases exponentially for constant $n = n_1 = \cdots = n_d$. This so called *curse of dimensionality* prevents the explicit storage of the entries except for very small values of d. Even for $n = 2$, storing a tensor of order $d = 50$ would require 9 petabyte! It is therefore essential to approximate tensors of higher order in a compressed scheme, for example, a low-rank tensor decomposition.

”

But just because something is “essential” doesn’t make it possible!

Our task is to understand what structures are compressible and how problems can be reformulated to enhance this structure.

Analogy. Conjugate gradients and other Krylov iterations are useless for random matrices. Everybody knows this, and the discussion is always about enhancing structure via preconditioners.

I wish the low-rank compression literature were equally forthright.

2. Quasi-Monte Carlo (QMC)

A few names:

Caflisch, Dick, Glasserman, Halton, Hickernell, Kuo, L'Ecuyer, Lemieux, Moskowitz, Niederreiter, Novak, Nuyens, Pilichshammer, Schwab, Sloan, Sobol', Woźniakowski, Zaremba,…

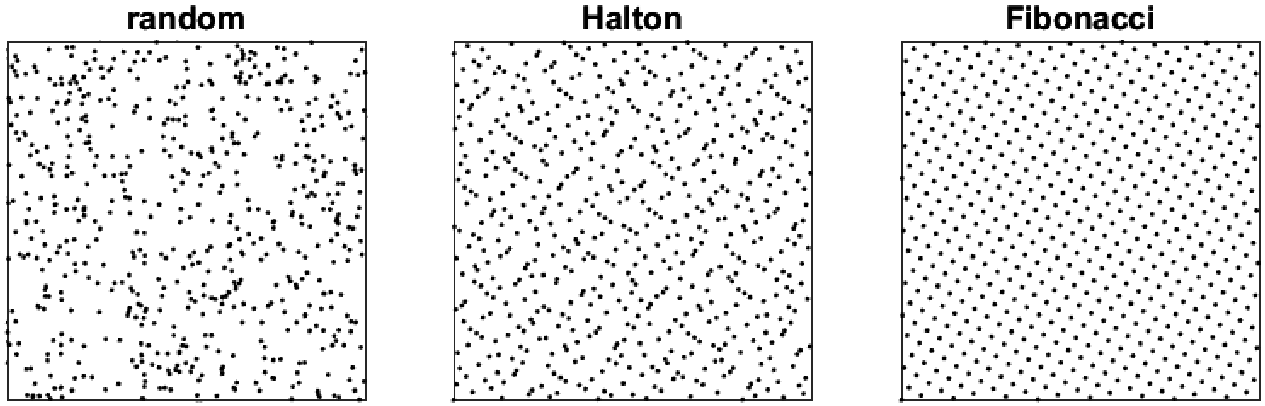
Equal-weight quadrature in hypercube of volume 1: $I_N = N^{-1}\sum_{i=1}^{N} f(s_i)$

Monte Carlo: nodes are random. Accuracy $O(N^{-1/2})$.

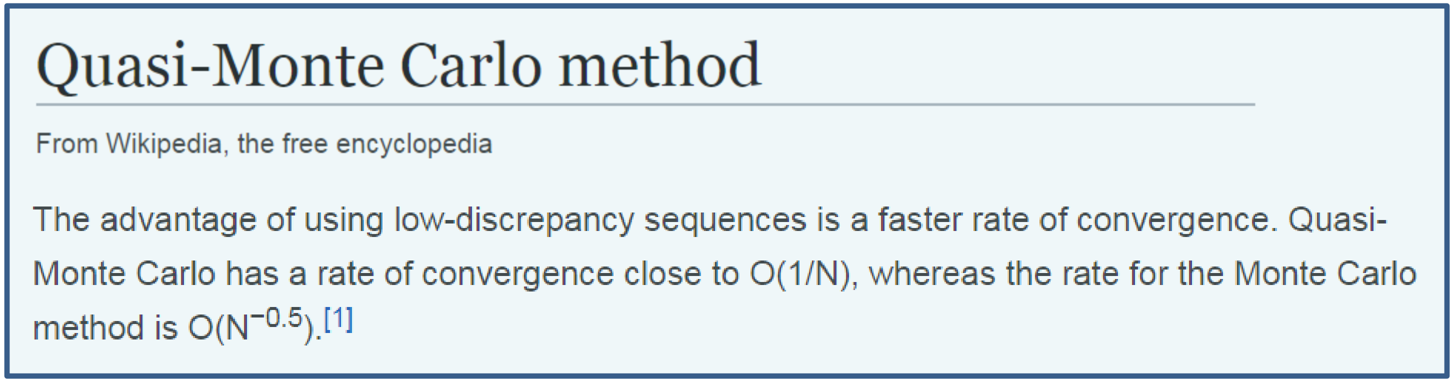
Quasi-Monte Carlo: nodes are more uniform than random.
The aim is to get accuracy closer to $O(N^{-1})$.

Very successful in practice. The method hit the headlines in 1995 with Papageorgiou-Paskov-Traub, valuation of financial derivatives in dim 360.

QMC methods are axis-dependent:



One can guess their success must depend on some kind of axis-alignment in applications. Experts know this, but don't always emphasize it.



Quasi-Monte Carlo method

From Wikipedia, the free encyclopedia

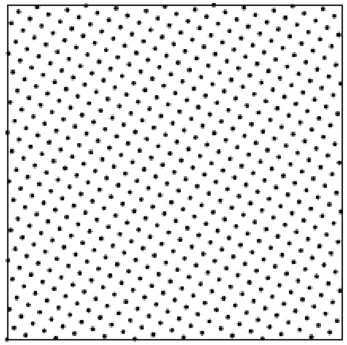
The advantage of using low-discrepancy sequences is a faster rate of convergence. Quasi-Monte Carlo has a rate of convergence close to O(1/N), whereas the rate for the Monte Carlo method is $O(N^{-0.5})$.[1]

Standard theory justifies QMC methods by the Koksma-Hlawka inequality:

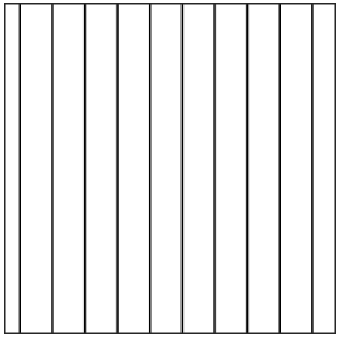
error ≤ star-discrepancy × Hardy-Krause variation

Star-discrepancy: a measure of irregularity of distribution of the sample pts with respect to rectangles aligned with the axes

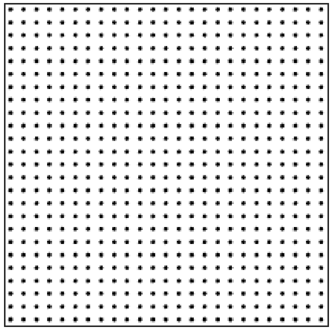
Hardy-Krause variation: a measure of nonsmoothness of the integrand with respect to derivatives aligned with the axes



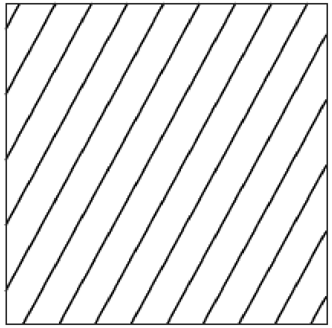
low-discrepancy grid



low-variation function



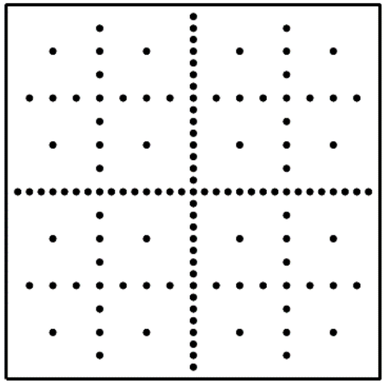
high-discrepancy grid

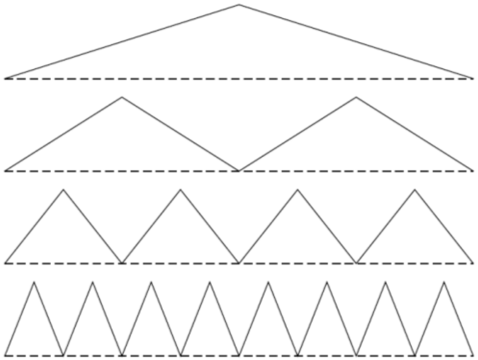


high-variation function

3. Sparse grids

Standard PDE grids are cheap for dimension $s = 2$, expensive for $s = 3$, and next to impossible for $s \geq 6$. Sparse grids are based on a recursive idea and have only logarithmically as many degrees of freedom.



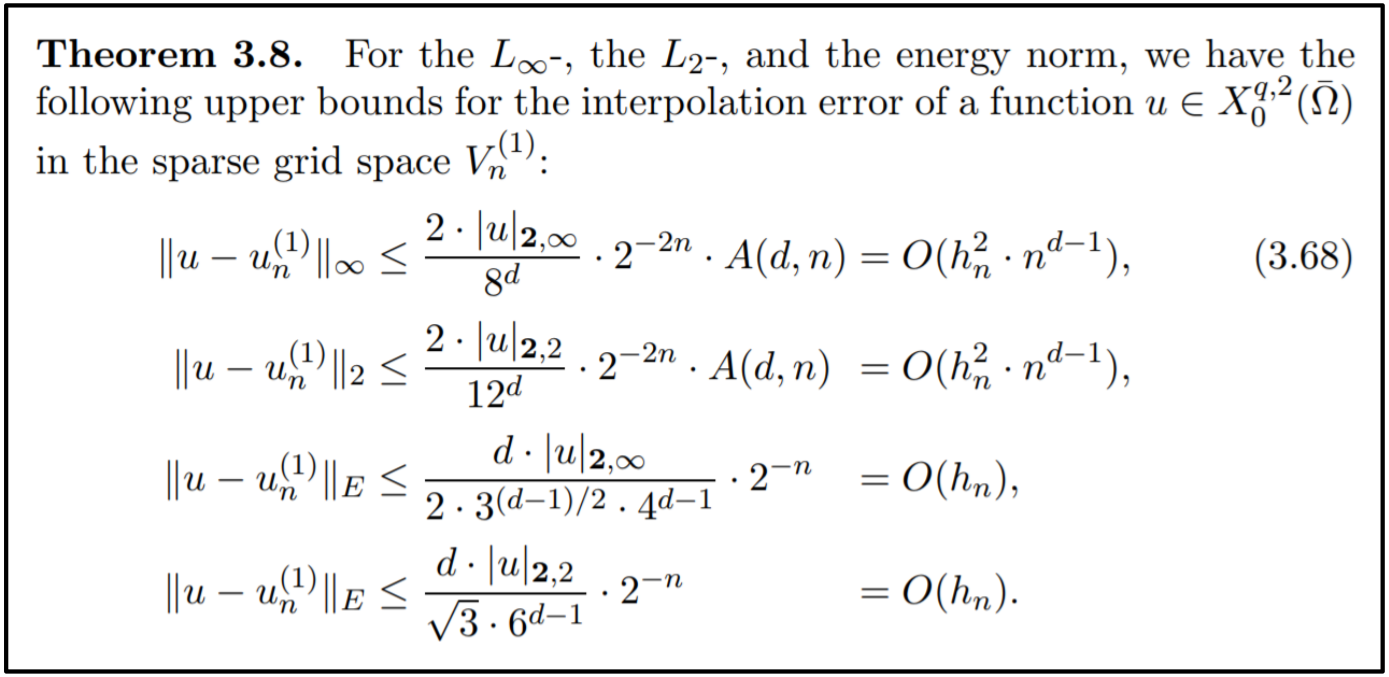


A few names:

Babenko, Bungartz, Garcke, Griebel, Gunzburger, Hegland, Holtz, Ma, Nobile, Rüde, Schwab, Smolyak, Temlyakov, Tempone, Webster, Wohlmuth, Yserentant, Zabaras, Zenger,….

Sparse grids have been very successful in some applications.
As the pictures suggest, they depend on grid alignment.

A typical theorem showing smooth functions are well approximated



Theorem 3.8. For the L_∞-, the L_2-, and the energy norm, we have the following upper bounds for the interpolation error of a function $u \in X_0^{q,2}(\bar{\Omega})$ in the sparse grid space $V_n^{(1)}$:

$$\|u - u_n^{(1)}\|_\infty \leq \frac{2 \cdot |u|_{\mathbf{2},\infty}}{8^d} \cdot 2^{-2n} \cdot A(d,n) = O(h_n^2 \cdot n^{d-1}), \tag{3.68}$$

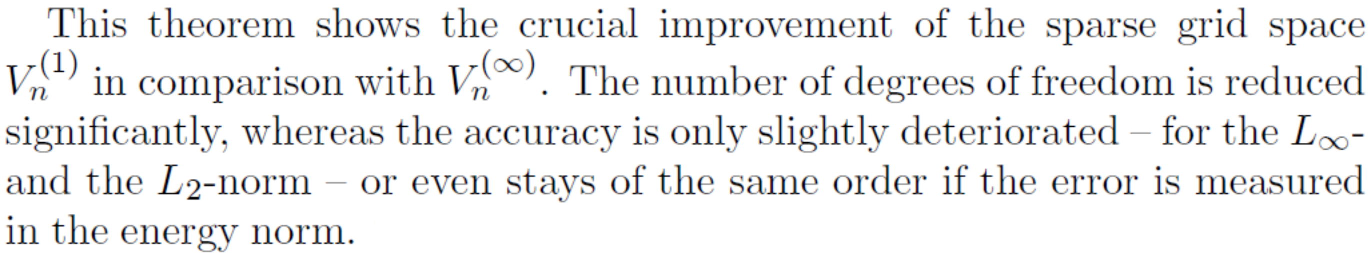
$$\|u - u_n^{(1)}\|_2 \leq \frac{2 \cdot |u|_{\mathbf{2},2}}{12^d} \cdot 2^{-2n} \cdot A(d,n) = O(h_n^2 \cdot n^{d-1}),$$

$$\|u - u_n^{(1)}\|_E \leq \frac{d \cdot |u|_{\mathbf{2},\infty}}{2 \cdot 3^{(d-1)/2} \cdot 4^{d-1}} \cdot 2^{-n} = O(h_n),$$

$$\|u - u_n^{(1)}\|_E \leq \frac{d \cdot |u|_{\mathbf{2},2}}{\sqrt{3} \cdot 6^{d-1}} \cdot 2^{-n} = O(h_n).$$

Bungartz-Griebel, *Acta Numerica*, 2004

“



This theorem shows the crucial improvement of the sparse grid space $V_n^{(1)}$ in comparison with $V_n^{(\infty)}$. The number of degrees of freedom is reduced significantly, whereas the accuracy is only slightly deteriorated – for the L_∞- and the L_2-norm – or even stays of the same order if the error is measured in the energy norm.

”

But what are $|u|_{\mathbf{2},\infty}$ and $|u|_{\mathbf{2},2}$? They are based on *mixed derivatives….*

Mixed derivatives

$$|u|_{\boldsymbol{\alpha},\infty} := \left\|D^{\boldsymbol{\alpha}} u\right\|_{\infty}, \qquad D^{\boldsymbol{\alpha}} u := \frac{\partial^{|\boldsymbol{\alpha}|_1} u}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \ ,$$

where the order of a mixed derivative is the ∞-norm of its multiindex $\boldsymbol{\alpha}$.

E.G., $\frac{\partial^4 f}{\partial x^2\, \partial y^2}$ is a mixed derivative of order 2, but $\frac{\partial^4 f}{\partial x^4}$ is of order 4.

So $\cos(100x)$, for example, is reasonably smooth, but becomes exponentially less smooth if you rotate the axes in $s \gg 1$ dimensions.

(Mathematically this is close to the Hardy-Krause variation.)

4. Multivariate polynomials

This is the central topic of algebraic geometry. (Interesting question: why?)

$$p(\mathbf{x}) = \sum_{i_1,\ldots,i_s} a_{i_1,\ldots,i_s} x_1^{i_1} \cdots x_s^{i_s}$$ (finite sum)

A few names:

Bates, de Boor, Bos, Cools, Hauenstein, Levenberg, Lyons, Novak, Ritter, Ron, Salzer, Sauer, Sommariva, Sommese, Stetter, Stillman, Sturmfels, Wampler, Xu,....

Degree (= total degree) of $x_1^{i_1} \ldots x_s^{i_s}$: $d = i_1 + \cdots + i_s$.

Reason #1: this definition is isotropic = invariant wrt rotations.

Reason #2: degree d terms are $O(h^d)$ as $h \to 0$.

Strong contrast with maximal degree: $d = \max\ \{ i_1 , \ldots , i_s \}$.

The degree of p is the maximum of the degrees of its monomials.

Applications of multivariate polynomials

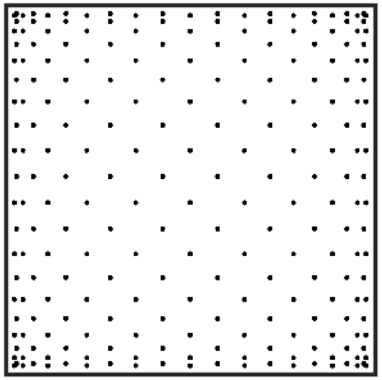
Polynomials of *maximal* degree d are easy to work with — tensor products. But the number of parameters is huge: for degree d, $\approx d^s$.

Total degree d is far more compact: $\approx d^s/s!$. Even in 3D this is 6 times less.

It is an old idea to seek algorithms based on polynomials of total degree d, though there are pervasive challenges of existence/uniqueness (unisolvency).

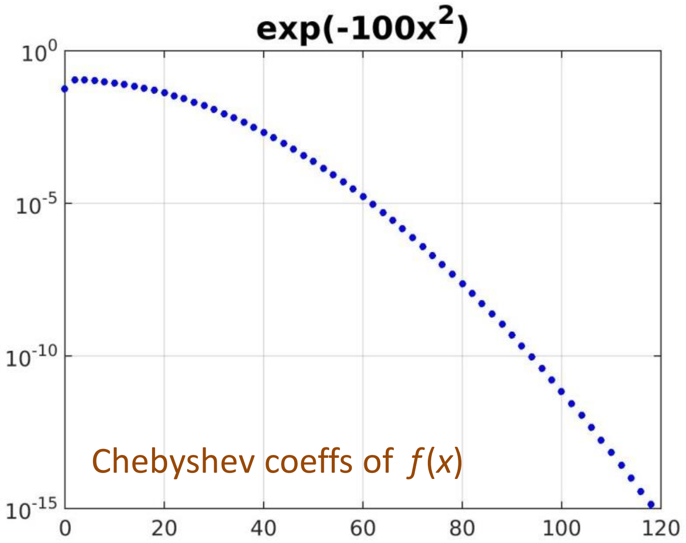
For cubature, i.e. multivariate integration, this begins with Maxwell 1877.

Recent development: Padua points. Unisolvency with small Lebesgue const.

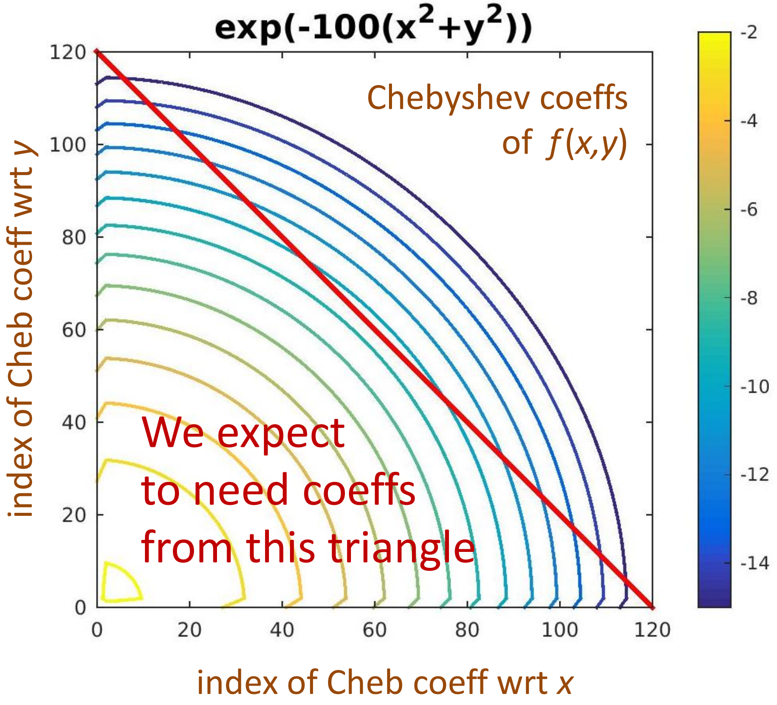


The anisotropy effect

$f(x) = \exp(-100x^2)$ can be resolved to 15 digits on $[-1,1]$ by $p(x)$ of degree 120.



Chebyshev coeffs of $f(x)$



Chebyshev coeffs of $f(x,y)$

We expect to need coeffs from this triangle

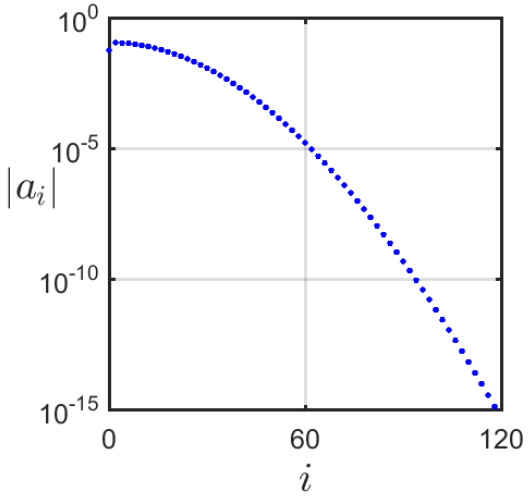
What degree $p(x,y)$ is needed for $f(x,y) = \exp(-100(x^2+y^2))$ on $[-1,1]^2$?

NB: f is isotropic, and multivariate polynomials are isotropic.

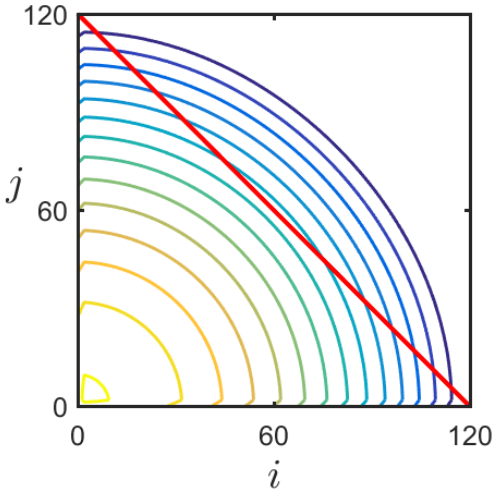
Wrong! Need degree $120\sqrt{2}$, not 120, to get 15 digits in the unit square.

Explanation: the square is $\sqrt{2}$ times longer along the diagonal.

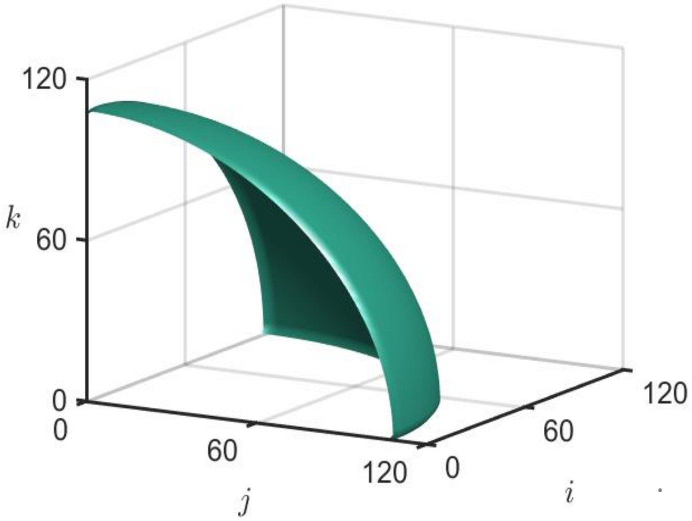
1D



2D



3D



```
f = @(x) exp(-100*x.^2);
plotcoeffs(chebfun(f),'.');
```

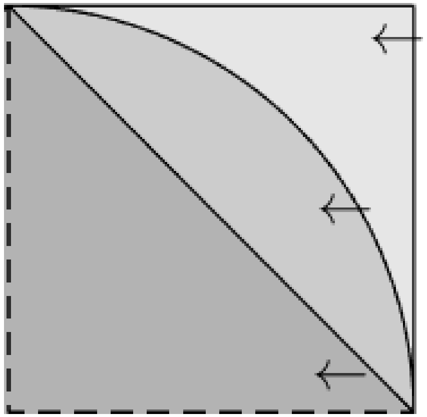
```
f = @(x,y) exp(-100*(x.^2+y.^2));
c = abs(chebcoeffs2(chebfun2(f)));
c = c(1:2:end,1:2:end); [m,n] = size(c);
contour(2*(0:n-1),2*(0:m-1),log10(c),-15:-2);
colorbar, axis equal
```

```
f = @(x,y,z) exp(-100*(x.^2+y.^2+z.^2));
c = abs(chebcoeffs3(chebfun3(f)));
c = c(1:2:end,1:2:end,1:2:end);
[m,n,p] = size(c);
isosurface(2*(0:n-1),2*(0:m-1),...
          2*(0:p-1),log10(c),-15);
axis equal, view(30,15)
```

Chebfun codes

Euclidean degree

We are thus led to conjecture that what matters for approximation in the s-hypercube is not total degree or maximal degree but *Euclidean degree*.



Maximal degree - ∞-norm of exponent vector

Euclidean degree - 2-norm of exponent vector

Total degree - 1-norm of exponent vector

Maximal degree (tensor products): $\sqrt{s}$ times finer resolution along diagonal

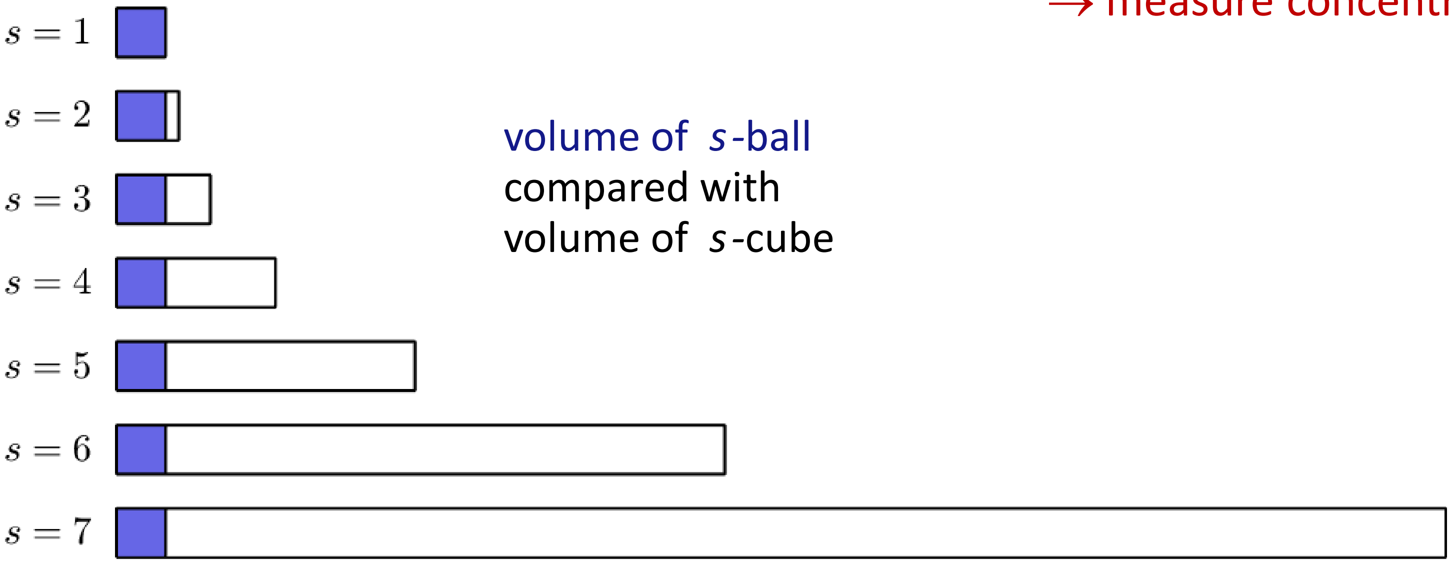
Euclidean degree: uniform resolution in all directions

Total degree: $\sqrt{s}$ times coarser resolution along diagonal;
isotropic in the s-ball, not the s-cube

Exponential dependence on dimension *s*

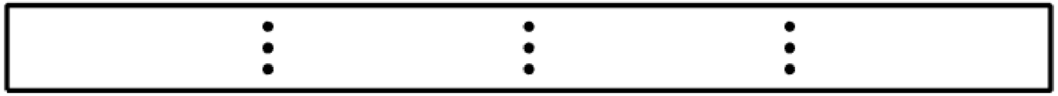
For $s \gg 1$, most of the volume of the hypercube is outside the inscribed hyperball (ratio $\approx s^{s/2}$).

→ measure concentration



volume of *s*-ball
compared with
volume of *s*-cube

Polynomials of given total degree in the 6-hypercube, one might say, have a nonuniformity analogous to this 3×3 grid in a rectangle:



Exponential dependence on dimension *s*, continued

Inefficiency of using maximal degree in the *s*-hypercube

$$\frac{\text{no. of parameters for approx. based on } d_{\max}}{\text{no. of parameters for approx. based on } d_E} \approx \frac{V_\infty}{V_2} \sim \sqrt{\pi s}\left(\frac{2s}{\pi e}\right)^{s/2}$$

Inefficiency of using total degree in the *s*-hypercube

$$\frac{\text{no. of parameters for approx. based on } d_T}{\text{no. of parameters for approx. based on } d_E} \approx \frac{s^{s/2} V_1}{V_2} \sim 2^{-1/2}\left(\frac{2e}{\pi}\right)^{s/2}$$

Bernstein (=Hooke) and Newton ellipses

Standard theory for approximation of $f(x)$ on $[-1,1]$ (Bernstein 1912): define Bernstein ρ-ellipse E_ρ: foci ± 1, semi-axis lengths sum to ρ.

Thm. $f(x)$ analytic for $x \in E_\rho \;\Rightarrow\; \| f - p_n^* \| = O_\varepsilon(\rho^{-n})$.

How to generalize this to approximation of $f(x_1, \dots, x_s)$ on $[-1,1]^s$?

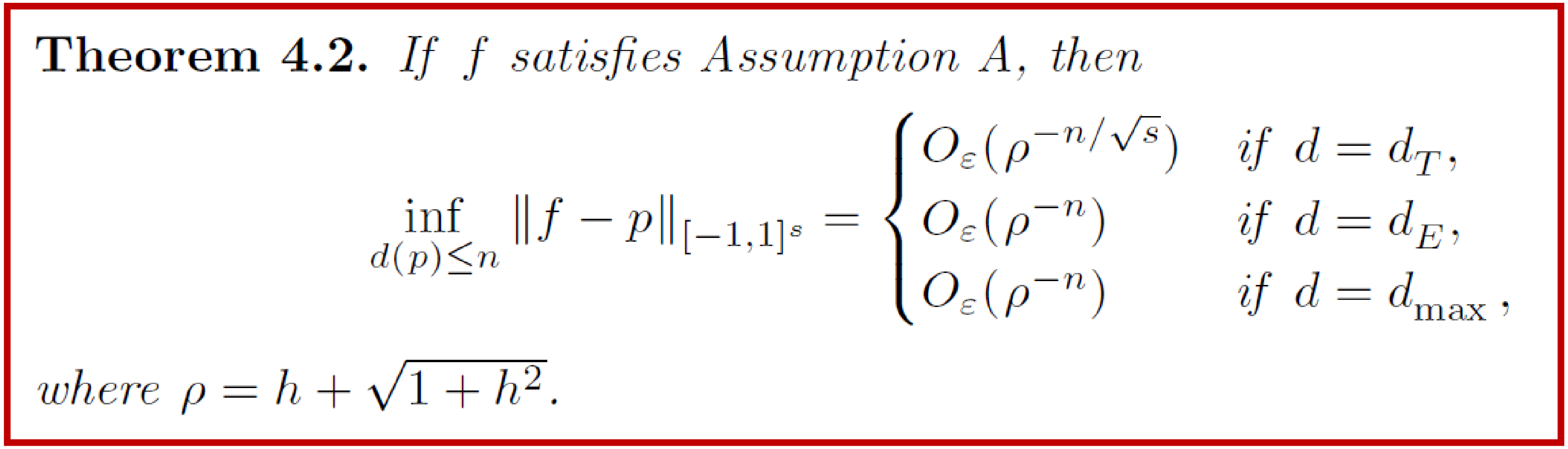
Note that $x \in E_\rho \Leftrightarrow x^2 \in N_{1,h^2}$ if we define $\rho = h + (1+h^2)^{1/2}$ and Newton ellipse N_{s,h^2}: foci $0, s$, leftmost point $-h^2$.

Thm. $f(x)$ analytic for $x^2 \in N_{1,h^2} \;\Rightarrow\; \| f - p_n^* \| = O_\varepsilon(\rho^{-n})$.

In s dimensions our region of analyticity will be $\boxed{x_1^2 + \dots + x_s^2 \in N_{s,h^2}}$.

Assumption A. For some $h > 0$, f is analytic for all $\boldsymbol{x}$ with $x_1^2 + \ldots + x_s^2 \in N_{s,h^2}$.

After some pretty algebra and appeal to Bochner and Martin 1948, we get:



Theorem 4.2. *If f satisfies Assumption A, then*

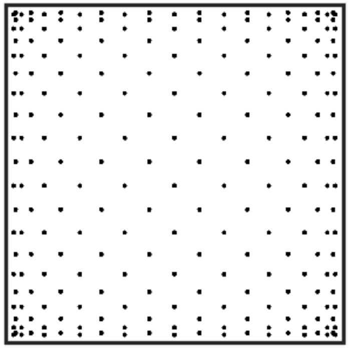
$$\inf_{d(p)\le n} \|f - p\|_{[-1,1]^s} = \begin{cases} O_\varepsilon(\rho^{-n/\sqrt{s}}) & \textit{if } d = d_T, \\ O_\varepsilon(\rho^{-n}) & \textit{if } d = d_E, \\ O_\varepsilon(\rho^{-n}) & \textit{if } d = d_{\max}, \end{cases}$$

where $\rho = h + \sqrt{1+h^2}$.

“Multivariate polynomial approximation in the hypercube,” *Proc. AMS,* 2017

Also related work by Bos & Levenberg, arXiv, 2017.

Cubes, balls, and a return to Padua points



Padua points live in the square, yet are made for interpolation by polynomials of given total degree – which are adapted to the disk. Mismatch.

Two research questions

1. Staying with the square: can we devise good points for Euclidean degree?
2. Staying with total degree: can we devise good points for the disk? [Bos]

Summary of polynomial approximation in the hypercube

Polynomials of total degree *d* in the hypercube make sense

- as $d \longrightarrow \infty$ in a ball (p-convergence)
- as $h \longrightarrow 0$ in any shape (h-convergence)

They don't make much sense as $d \longrightarrow \infty$ in a cube.

Axis-alignment
in low-rank and other structures

Nick Trefethen, Oxford and ENS Lyon

Thank You!

1. Low-rank approximation
2. Quasi-Monte Carlo
3. Sparse grids
4. Multivariate polynomials