# Non-polynomial discretizations: coping with redundancy and ill-conditioning

Daan Huybrechs

University of Leuven

9 October 2014, Woudschoten

Joint w. Sam Groth, Nele Lejon, Roel Matthysen, Peter Opsomer

# Outline

# Outline

1. Non-polynomial discretizations

2. Fourier extensions

# Introduction

**Many schemes are being developed for numerical wave simulations that are not based on (piecewise) polynomial approximations.** A few are:

- UWVF: the Ultraweak Variational Formulation
  - Cessenat and Deprés, 1994, 1998
  - Monk, Huttunen, Hiptmair
- PUFEM: partition of unity finite element method
  - Babuška and Melenk, 1997
- Plane wave basis in integral equations
  - de La Bourdonnaye 1994, Abboud, Perrey-Debain, Trevelyan
- Method of fundamental solutions
  - Barnett and Betcke, 2008, 2010
- WBM: the Wave Based Method (Desmet, 1998)

# Some observations

**These schemes have several things in common:**

- **oscillatory basis functions**: plane waves, Bessel functions, fundamental solutions
- Trefftz-type methods: approximate PDE solution using basic solutions of the same PDE[1]
- high-order convergence
- small number of degrees of freedom
- they often exhibit (extreme) **ill-conditioning**, yet **high accuracy**
- (they often involve having to evaluate highly oscillatory integrals)

[1] Trefftz, 1926: Ein Gegenstück zum Ritzschen Verfahren. Internat. congress on Applied Mechanics, Zürich.

# Why are oscillatory problems hard?

**Three things make life difficult when oscillations increase**:

1. Many degrees of freedom (dof) are required just to be able to represent the solution
   - 'resolving the oscillations'
2. Quite often, even more dof's are needed to solve a problem
   - due to pollution or dispersion errors[2]
3. Fast solvers for low-frequency problems typically fail (or need significant adjustments) for high-frequency problems
   - e.g. multigrid

---

[2] Babuska and Sauter, SIAM Review, 2000: Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?

# These methods exhibit ill-conditioning

**Ill-conditioning is usually problematic.**

What does it mean for $Ax = B$?

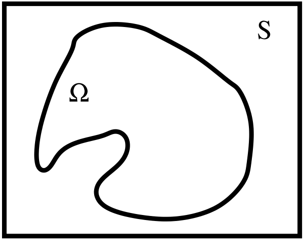
- no iterative solvers
- $A$ is singular. Hopefully $B$ lies in the range of $A$!
- If so, there is no uniqueness: many solution vectors $x$.
- For each possible solution vector, the residual $Ax - B$ is small.

**Can we exploit the redundancy and cope with the ill-conditioning?**

# The Wave Based Method (WBM)

Consider the Helmholtz equation

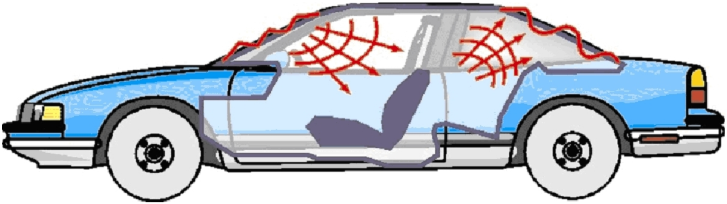
$$\Delta u + k^2 u = 0$$

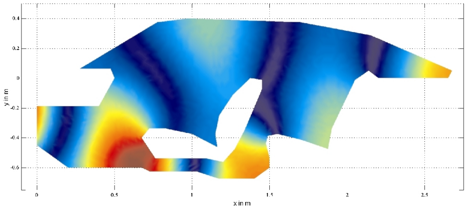



- Discretize on a **convex** and bounded domain $\Omega$
- using a set of solutions on a bigger **bounding box** $S$
- If the desired domain is not convex: subdivide into (few) convex subdomains

# A WBM simulation of a vibro-acoustic problem

**The full method is more capable:**





# WBM: basis functions

Using a bounding box with lengths $L_x$ and $L_y$, we write the pressure as

$$p(x,y) = \sum_{l=0}^{\infty} a_l \cos(k_{xl1}x)e^{-ik_{yl1}y} + \sum_{l=0}^{\infty} b_l e^{-ik_{xl2}} \cos(k_{yl2}y)$$

with

$$\begin{aligned}
(k_{xl1}, k_{yl1}) &= \left(\frac{l\pi}{L_x}, \pm\sqrt{k^2 - \left(\frac{l\pi}{L_x}\right)^2}\right) \\
(k_{xl2}, k_{yl2}) &= \left(\pm\sqrt{k^2 - \left(\frac{l\pi}{L_y}\right)^2}, \frac{l\pi}{L_y}\right)
\end{aligned}$$

Note that $k_x^2 + k_y^2 = k^2$.

# WBM: discretization

A weighted residual formulation (Galerkin) leads to:

$$Ax = B$$

with a highly ill-conditioned matrix $A$ and where $B$ corresponds to the boundary condition.

- entries of $A$ are computed accurately (quadrature)
- a direct solver is used. . .
- . . . and the solution satisfies Helmholtz and very accurately matches the boundary condition.

Why?

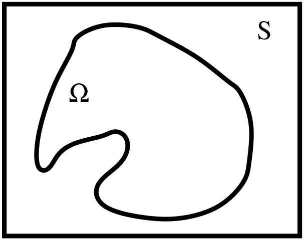
# Preliminary analysis

**Can one approximate solutions of $Lu = 0$ on $\Omega$ by solutions of $Lu = 0$ on $S \supset \Omega$?**

- For a second-order elliptic operators $L$: yes.
  P. Lax, 1956, *A stability theorem for solutions of abstract differential equations, and its application to the study of the local behavior of solutions of elliptic equations*.
- Yes in more general settings too:
  F. Browder, 1962, *Approximation by solutions of partial differential equations*.
- No convexity requirement ...

# Convexity: why

What is the extension of $u$ on $\Omega$ to $\tilde{u}$ on $S$?




**The continuation of $u$ outside $\Omega$ (for analytic $\Omega$) may develop singularities for two reasons:**[3]

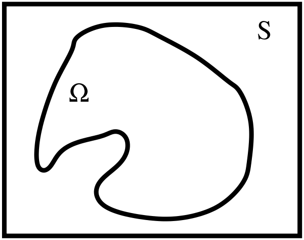
1. singularities due to the analytic continuation of the boundary data $f$
2. singularities due to the shape of $\partial\Omega$

Avoid 2 by using a convex domain. If not: slow convergence.

[3] R. F. Millar, 1980, The analytic continuation of solutions to elliptic boundary value problems in two independent variables

# More questions

To which extent can solutions be concentrated in $\Omega$ or in $S \setminus \Omega$?




**Are there solutions to $Lu = 0$ on $S$ that are small(ish) on $\Omega$ but large on $S \setminus \Omega$?**

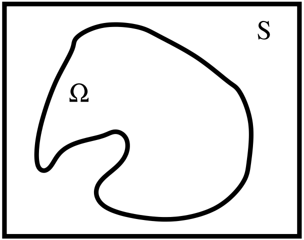
- How small can they possibly be?
- What is their discrete norm in the representation in our basis on $S$?
- compactly supported solutions are impossible
- Questions relate to singular values and vectors of the discretization matrix $A$.

# Outline

1. Non-polynomial discretizations
2. Fourier extensions

# Approximation by Fourier series

**Example:**




**For $\Omega \subset S := [0,1]^n$:**

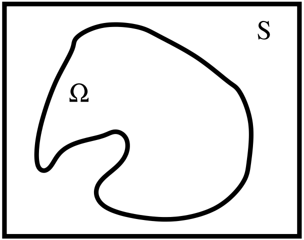
- approximate $f$ on $\Omega$ by Fourier series $f_S$ on $S$
- find best approximation in $L^2(\Omega)$ norm:

$$\min \|f - f_S\|_\Omega$$

- Does extension of $f$ from $\Omega$ to $S$ always exist? Yes. Whitney extension problem.
- Is it unique? No. Restriction of Fourier series on $S$ to $\Omega$ constitutes a frame for $L^2(\Omega)$.

# Approximation by Fourier series

**Example:**




**For $\Omega \subset S := [0,1]^n$:**

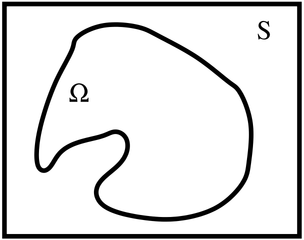
- approximate $f$ on $\Omega$ by Fourier series $f_S$ on $S$
- find best approximation in $L^2(\Omega)$ norm:

$$\min \|f - f_S\|_\Omega$$

- Does extension of $f$ from $\Omega$ to $S$ always exist? Yes. Whitney extension problem.
- Is it unique? No. Restriction of Fourier series on $S$ to $\Omega$ constitutes a frame for $L^2(\Omega)$.

# Approximation by Fourier series

**Example:**




**For $\Omega \subset S := [0,1]^n$:**

- approximate $f$ on $\Omega$ by Fourier series $f_S$ on $S$
- find best approximation in $L^2(\Omega)$ norm:

$$\min \|f - f_S\|_\Omega$$

- Does extension of $f$ from $\Omega$ to $S$ always exist? Yes. Whitney extension problem.
- Is it unique? No. Restriction of Fourier series on $S$ to $\Omega$ constitutes a frame for $L^2(\Omega)$.

# Recent results

**Fourier extension/Fourier continuation**

- origin in Fictitious Domain methods or Embedded Domain techniques for PDEs
- Boyd 2002, Bruno 2004: the Fourier extension problem
- H. 2009: analysis of exact solution in 1D
- Adcock, H. 2014: Fourier extensions are optimal for representing oscillatory functions
- Adcock, H., Martin-Vacquero, FoCM, 2014: proof of numerical stability
- Matthysen, H.: Fast construction of Fourier extension in 1D
- Lyon and Bruno, Lyon: time-steppers for PDEs, fast routine for special case

# On the representation of domains

**Approximating functions on a general domain $\Omega$ is hard:**

- Is the domain connected? Punctured? Open, closed or both?
- The boundary $\partial\Omega$ may have corners, cusps, . . .
- What is the dimension of $\Omega$?
- **Efficient spectral** approximation schemes known only for tensor-product domains
  - squares and rectangles, cubes, torus, . . .

**Representing a domain by approximating its boundary is very restrictive.**

# On the representation of domains

**Approximating functions on a general domain $\Omega$ is hard:**

- Is the domain connected? Punctured? Open, closed or both?
- The boundary $\partial\Omega$ may have corners, cusps, ...
- What is the dimension of $\Omega$?
- **Efficient spectral** approximation schemes known only for tensor-product domains
  - squares and rectangles, cubes, torus, ...

**Representing a domain by approximating its boundary is very restrictive.**

# The characteristic function

**The characteristic function $C(\Omega)$ turns out to be useful:**

$$C(x,y) = \begin{cases} 1, & \text{if } (x,y) \in \Omega, \\ 0, & \text{otherwise.} \end{cases}$$

**Examples**

- open circle: $C(x,y) \equiv x^2 + y^2 - R^2 < 0$
- Mandelbrot set:
  $C(x,y) \equiv$ iteration $z_{n+1} = z_n^2 + (x+iy)$ remains bounded

**We represent $\Omega$ by implementing $C(\Omega)$.**

# The characteristic function

**The characteristic function $C(\Omega)$ turns out to be useful:**

$$C(x,y) = \begin{cases} 1, & \text{if } (x,y) \in \Omega, \\ 0, & \text{otherwise.} \end{cases}$$

**Examples**

- open circle: $C(x,y) \equiv x^2 + y^2 - R^2 < 0$
- Mandelbrot set:
  $C(x,y) \equiv$ iteration $z_{n+1} = z_n^2 + (x+iy)$ remains bounded

**We represent $\Omega$ by implementing $C(\Omega)$.**

# The characteristic function

**The characteristic function $C(\Omega)$ turns out to be useful:**

$$C(x,y) = \begin{cases} 1, & \text{if } (x,y) \in \Omega, \\ 0, & \text{otherwise.} \end{cases}$$

**Examples**

- open circle: $C(x,y) \equiv x^2 + y^2 - R^2 < 0$
- Mandelbrot set:
  $C(x,y) \equiv$ iteration $z_{n+1} = z_n^2 + (x+iy)$ remains bounded

**We represent $\Omega$ by implementing $C(\Omega)$.**

# Advantages

Implementing the characteristic function has many advantages:

1. Very flexible, very **general**. No domain is a priori excluded.
2. Boundary can be anything, no need to represent it.
3. Simple **arithmetic**, e.g:

$$\Omega = A \cap B \quad \Rightarrow \quad C(x,y) = C_A(x,y) \textbf{ and } C_B(x,y)$$

4. **Implicit** definitions of domains can be used, e.g.

$$C(x,y) \equiv f(x,y) \geq c.$$

5. It is easy to **generate points** that belong to $\Omega$.

# Advantages

Implementing the characteristic function has many advantages:

1. Very flexible, very **general**. No domain is a priori excluded.
2. Boundary can be anything, no need to represent it.
3. Simple **arithmetic**, e.g:

$$\Omega = A \cap B \quad \Rightarrow \quad C(x,y) = C_A(x,y) \textbf{ and } C_B(x,y)$$

4. **Implicit** definitions of domains can be used, e.g.

$$C(x,y) \equiv f(x,y) \geq c.$$

5. It is easy to **generate points** that belong to $\Omega$.

# Advantages

Implementing the characteristic function has many advantages:

1. Very flexible, very **general**. No domain is a priori excluded.
2. Boundary can be anything, no need to represent it.
3. Simple **arithmetic**, e.g:

$$\Omega = A \cap B \quad \Rightarrow \quad C(x,y) = C_A(x,y) \textbf{ and } C_B(x,y)$$

4. **Implicit** definitions of domains can be used, e.g.

$$C(x,y) \equiv f(x,y) \geq c.$$

5. It is easy to **generate points** that belong to $\Omega$.

# Least squares approximation

**What does least squares approximation require?**

Find

$$f(x,y) \approx \sum_{i=1}^{N} c_i \phi_i(x,y)$$

which minimizes

$$\sum_{j=1}^{M} \left( \sum_{i=1}^{N} c_i \phi_i(x_j,y_j) - f(x_j,y_j) \right)^2$$

1. a set of **points**: sample the characteristic function $C(\Omega)$
2. a set of **functions**

# Fourier extension

**The Fourier extension scheme in 1D**

- represent a function on $[-1,1]$ by a Fourier series on $[-2,2]$

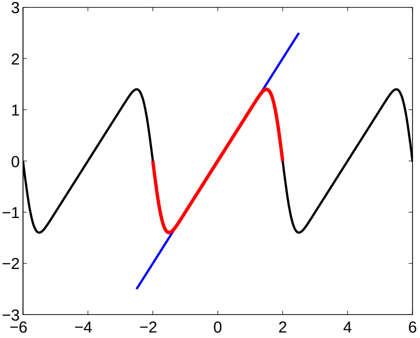
$$f(x) \approx \sum_{k=-N}^{N} c_k e^{\frac{\pi}{2} ikx}, \qquad x \in [-1,1],$$

rather than

$$f(x) \approx \sum_{k=-N}^{N} c_k e^{\pi ikx}, \qquad x \in [-1,1].$$

- no periodicity on $[-1,1]$ is required: no Gibbs phenomenon
- proposed (independently) by **Oscar Bruno** and **John Boyd**

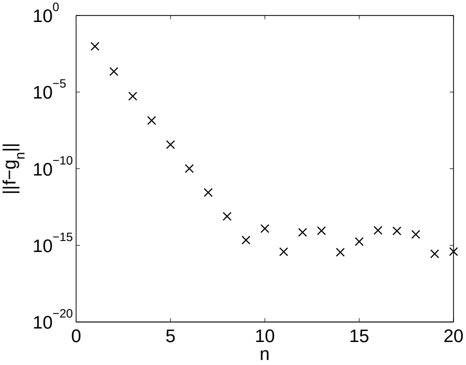
# An example: $f(x) = x$




The function $f(x) = x$ is not periodic on $[-1, 1]$, but smooth extensions periodic on $[-2, 2]$ exist.

# Convergence behaviour

Least squares approximation on $[-1, 1]$ using functions $e^{\frac{\pi}{2}ikx}$




**Stable**, spectral convergence to machine precision.

# Why is ill-conditioning natural?

**The least squares problem leads to a linear system**

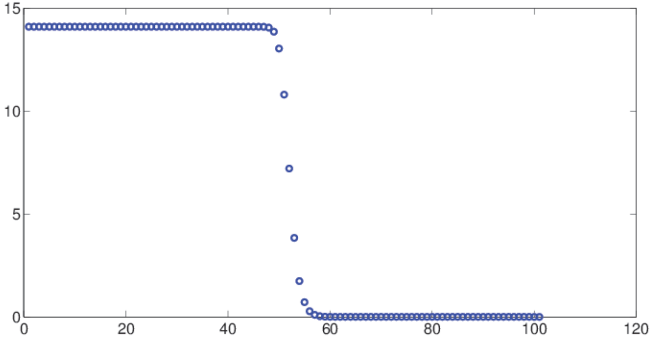
$$Ax = B$$

- $A \in \mathbb{C}^{M \times N}$ is rectangular: **overdetermined** least squares
  - elements are evaluations $\phi_i(x_j, y_j)$ (collocation)
  - alternative: $A_{m,n} = \langle \phi_m, \phi_n \rangle$ (projection)
- if the set $\{\phi_i\}$ is **complete and redundant**, columns are nearly **linearly dependent**
- hence $A$ is **extremely ill-conditioned**

# More on the matrix $A$

**The singular values of the rectangular matrix $A$:**

$$A_{m,n} = e^{\frac{\pi}{2} i (n - \frac{N}{2} - 1) x_m}$$



# What is the matrix $A$

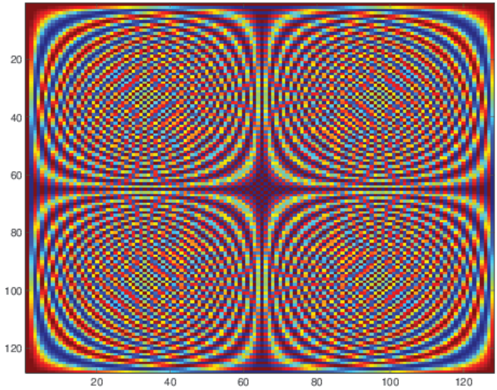
**$A$ is a rectangular subblock of the DFT matrix $D$**

$$X_k = \sum_{n=0}^{N-1} x_k e^{-i\pi kn/N}$$

$$A_{m,n} = e^{\frac{\pi}{2} i(n-\frac{N}{2}-1)x_m}$$

- subblocks of the DFT matrix have approximately low rank
- there is a fast matrix-vector product for $Ax$
- (at least with proper choices of parameters)

# The DFT matrix




**Subblocks have low-rank**

(Edelman et al, *The future Fast Fourier Transform?*, SISC, 1999)

# Singular values of $A$

**A lot can be said about the singular value decomposition**

$$A = U\Sigma V^*$$

**Columns of $U$ and $V$ are Periodic Discrete Prolate Spheroidal Sequences (P-DPSS)**

- related to discrete prolate spheroidal sequences (DPSS)
- related to prolate spheroidal wave functions
- popularized by Slepian in a series of papers I-V in the 60's and 70's

# Discrete prolate spheroidal sequences (1)

(Slepian, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty - V: the discrete case*, 1978)

**Question: which compactly supported sequence (in time) has maximally concentrated frequency spectrum?**

$$\left\{u_n^{(k)}(N,W)\right\}_{n=0}^{N-1} \quad \leftrightarrow \quad U_k(f;N,W) = \sum_{n=0}^{N-1} u_n^{(k)}(N,W)e^{-i\pi(N-1-2n)f}$$

Find the sequence $u_n^{(1)}$ that maximizes

$$\frac{\int_{-W}^{W} |U_1(f;N,W)|^2 \mathrm{d}f}{\int_{-\frac{1}{2}}^{\frac{1}{2}} |U_1(f;N,W)|^2 \mathrm{d}f}$$

# Discrete prolate spheroidal sequences (2)

We have

$$\int_{-W}^{W} |U_1(f;N,W)|^2 \mathrm{d}f = \lambda_1 \int_{-\frac{1}{2}}^{\frac{1}{2}} |U_1(f;N,W)|^2 \mathrm{d}f$$

with $\lambda_1$ close to 1.

Then, find the sequence $u_n^{(2)}$ that **maximizes concentration** of $U_2$ and that is **orthogonal to** $u_n^{(1)}$:

$$\int_{-W}^{W} |U_2(f;N,W)|^2 \mathrm{d}f = \lambda_2 \int_{-\frac{1}{2}}^{\frac{1}{2}} |U_2(f;N,W)|^2 \mathrm{d}f$$

And so on.

# Some interesting properties

The values $\lambda_k$ and sequences $u^{(k)}$ are eigenvalues and eigenvectors of the **prolate matrix** $\rho(N, W)$:

$$\rho(N, W)_{mn} = \frac{\sin 2\pi W(m-n)}{\pi(m-n)}.$$

This matrix **commutes with a tridiagonal matrix**.
The discrete prolate spheroidal wave functions satisfy an ODE and are eigenfunctions of an integral operator

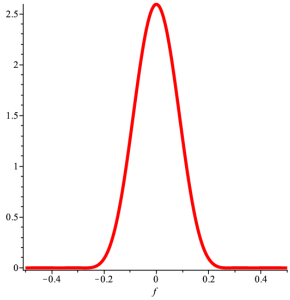
$$\int_{-W}^{W} \frac{\sin N\pi(f-f')}{\sin \pi(f-f')} U(f')\mathrm{d}f' = \lambda U(f)$$

They are **doubly orthogonal**: on $[-1/2, 1/2]$ and on $[-W, W]$.

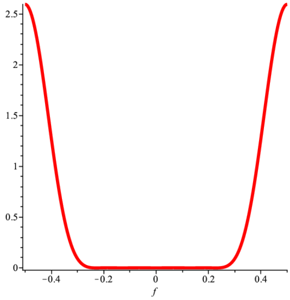
- If $A_{mn} = \langle \phi_m, \phi_n \rangle$ (projection) we have $A = \rho(N, \frac{1}{2T})$!

# An example

Set $W = 1/2$, $N = 11$.




(a) $\lambda_1 \approx 0.9999999$




(b) $\lambda_{11} \approx 0.00000001$

# How to use all this (1)

Let us precondition our system using an (appropriately sized) DFT matrix:

$$DAx = DB.$$

Then

$$DAx = \begin{bmatrix} D_1 A \\ D_2 A \end{bmatrix} x = \begin{bmatrix} D_1 B \\ D_2 B \end{bmatrix}$$

- $D_1 A$ is well-conditioned, $D_2 A$ is ill-conditioned
- because large eigenvalues have (nearly) bandlimited eigenvectors
- and small eigenvalues have high-frequency eigenvectors
- rank of $D_2 A$ is approximately $\log(N)$

# How to use all this (2)

$$\begin{bmatrix} D_1 A \\ D_2 A \end{bmatrix} x = \begin{bmatrix} D_1 B \\ D_2 B \end{bmatrix}$$

**We have a fast matrix-vector product, so we:**

- Solve $D_1 A x_1 = D_1 B$ with an iterative solver
- Construct $\log N$ random vectors in the null-space of $D_1 A$
- Using randomized linear algebra, use these to solve $D_2 A x_2 = D_2 B - D_2 A x_1$
- And add the two results together: $x = x_1 + x_2$

This is an $O(N \log N)$ algorithm. (With a fairly big constant).