Numerical Simulation and Scientific Computing I

Lecture 4:
Finite Difference Method,
Finite Precision Floating Point Arithmetics

Xaver Klemenschits, Paul Manstetten, and
Josef Weinbub

Institute for Microelectronics
TU Wien

nssc@iue.tuwien.ac.at

WS 2022

Quiz

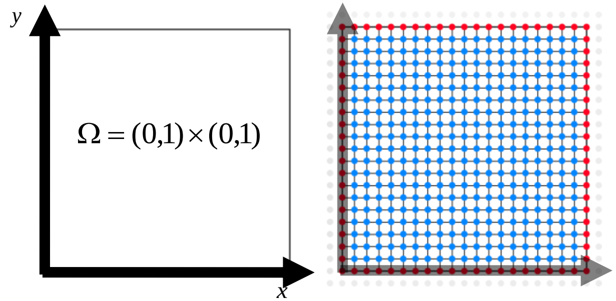
- Q1: Does C++ abstraction (classes, operator overloads, …) lead to run time overhead?
- Q2: How many choices are there to approximate a first derivative on a FDM grid?
- Q3: How big is the memory footprint of a 3D finite difference grid with uniform resolution of 1024 along each dimension?
- Q4: Assume someone provides you a solution uh to a specific linear equation system arising from applying the FDM to a problem, what could you do to check if it is indeed a solution?
- Q5: Assume additionally know the analytical solution u to the problem from Q4, how could you quantify the difference between the solution u and uh?

Outline

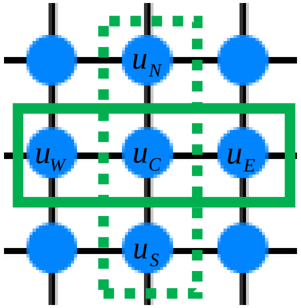
- Errors
- Vector and matrix norms
- Truncation error FD
- Boundary conditions FD
- Finite precision FP

Finite Difference Summary

- Starting point $\quad -(u_{xx}+u_{yy})=0 \quad in \quad \Omega=(0,1)\times(0,1) \quad +Dirichlet\ \ BCs$
- Domain discretization
  - N points per dimension
  - ~N² unknowns



- Approximation of (second) derivatives using FD



$$-\left(u_{xx}+u_{yy}\right)\approx-\frac{1}{h^2}\left(u_N+u_S+u_E+u_W-4u_C\right)=0$$

- Construction of linear equation system
  - A_h has size ~ (N^2 x N^2)

$$A_h\cdot u_h=b_h$$

$$\underbrace{\frac{1}{h^2}\begin{bmatrix}+4&-1&0&-1&0&0&0&0&0\\-1&+4&-1&0&-1&0&0&0&0\\0&-1&+4&0&0&-1&0&0&0\\-1&0&0&+4&-1&0&-1&0&0\\0&-1&0&-1&+4&-1&0&-1&0\\0&0&-1&0&-1&+4&0&0&-1\\0&0&0&-1&0&0&+4&-1&0\\0&0&0&0&-1&0&-1&+4&-1\\0&0&0&0&0&-1&0&-1&+4\end{bmatrix}}_{A_h}\cdot\underbrace{\begin{bmatrix}u_6\\u_7\\u_8\\u_{11}\\u_{12}\\u_{13}\\u_{16}\\u_{17}\\u_{18}\end{bmatrix}}_{x_h}=\underbrace{\frac{1}{h^2}\begin{bmatrix}u_1+u_5\\u_2\\u_3+u_9\\u_{10}\\0\\u_{14}\\u_{15}+u_{21}\\u_{22}\\u_{19}+u_{23}\end{bmatrix}}_{b_h}$$

Errors

- Discretized problem $$-\frac{1}{h^2}\left(u_N + u_S + u_E + u_W - 4u_C\right) = 0 \qquad A_h u_h - b_h = \underline{0}$$
- Total error $$u - \hat{u}_h = e_{\text{total}}$$
    - Difference between known solution and approximate solution of discretized problem
- Discretization/truncation error $$u - u_h = e_{\text{disc.}}$$
    - Difference between known solution and exact solution to discretized problem
    - To improve: numerical scheme, resolution
- Algebraic error $$u_h - \hat{u}_h = e_{\text{algeb.}}$$
    - Difference between exact and approximate solution of discretized problem
    - To improve: solution method, precision/order of finite arithmetic
- Residual $$A_h \hat{u}_h - b_h = \mathrm{r}_h$$
    - Deviation of RHS for approximate solution of discretized problem
    - Always available

- Relation to algebraic error
  - Interpretation: Residual is the perturbation of the RHS to make the approximate solution fit the discretized problem exactly

$$A_h\hat{u}_h = \hat{b}_h$$

$$A_h(u_h + \Delta u_h) = (b_h + \Delta b_h)$$

$$\underbrace{A_h u_h}_{b_h} + A_h \underbrace{\Delta u_h}_{e_{\text{algeb.}}} = b_h + \underbrace{\Delta b_h}_{r_h}$$

$$A_h e_{\text{algeb.}} = r_h$$

$$e_{\text{algeb.}} = A_h^{-1} r_h$$

- Condition number of A

$$\|A_h\| \cdot \|e_{\text{algeb.}}\| \geq \|A_h e_{\text{algeb.}}\| = \|r_h\|$$

$$\|e_{\text{algeb.}}\| = \|A_h^{-1} r_h\| \leq \|A_h^{-1}\| \cdot \|r_h\|$$

$$\frac{\|e_{\text{algeb.}}\|}{\|u_h\|} \leq \kappa(A) \frac{\|r_h\|}{\|b_h\|} = \|A_h^{-1}\|\|A_h\| \frac{\|r_h\|}{\|b_h\|}$$

Vector Norms

- Vector norms
  - positivity
  - homogeneity
  - subadditivity

$$\|x\| > 0 \qquad \forall x \neq 0$$

$$\|\alpha x\| = |\alpha| \cdot \|x\|$$

$$\|x + y\| \leq \|x\| + \|y\|$$

- Common norms
  - Manhattan norm
  - Euclidean norm
  - Maximum Norm

$$\|x\|_1 := \sum_{i=1}^{n} |x_i|$$

$$\|x\|_2 := \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$$

$$\|x\|_\infty := \max_{1 \leq i \leq n} |x_i|$$

- “Norm Sphere”

$$\|x\| = 1$$

Matrix Norms

- Induced matrix norm

$$\|A\| := \max_x \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

- Matrix norms
  - positivity
  - homogeneity
  - subadditivity
  - sub-multiplicativity
  - consistency

$$\|A\| > 0 \qquad \forall x \neq 0$$

$$\|\alpha A\| = |\alpha| \cdot \|A\|$$

$$\|A+B\| \leq \|A\| + \|B\|$$

$$\|AB\| \leq \|A\| \cdot \|B\|$$

$$\|Ax\| \leq \|A\| \cdot \|x\|$$

- Condition number
  - Deformation of how much “norm sphere” is deformed by the matrix

$$\kappa(A) := \frac{\max_{\|x\|=1} \|Ax\|}{\min_{\|x\|=1} \|Ax\|} = \|A^{-1}\| \|A\|$$

Norms Overview

$$\|x\| \qquad\qquad \|A\| := \max_{\|x\|=1} \|Ax\|$$

	Vector Norm	Induced Matrix Norm
Manhattan	$\|x\|_1 := \sum_{i=1}^{n} \lvert x_i \rvert$	$\|A\|_1 := \max_{1\le j\le n} \sum_{i=1}^{n} \lvert a_{ij} \rvert$
Euclidean	$\|x\|_2 := \left(\sum_{i=1}^{n} \lvert x_i \rvert^2 \right)^{1/2}$	$\|A\|_2 := \sigma_{\max}(A)$
Maximum	$\|x\|_\infty := \max_{1\le i\le n} \lvert x_i \rvert$	$\|A\|_\infty := \max_{1\le i\le m} \sum_{j=1}^{n} \lvert a_{ij} \rvert$

Normalizing Norms

- Comparing residual for different resolutions
  - Resolution N, grid spacing h

$$\left\| A_h \hat{u}_h - b_h \right\| = \left\| \mathrm{r}_h \right\|$$

$$\left\| x \right\| := \sqrt{\sum_{i=1}^{n} \left| x_i \right|^2}$$

  - Increase resolution and compare residual norms

$$\frac{\left\| \mathrm{r}_{h/2} \right\|}{\left\| \mathrm{r}_h \right\|} = ?$$

- Normalized sequence of norms

$$\left\| \underline{1} \right\|_{h_i} = 1 \quad \text{on} \quad \text{domain} \quad \Omega_\mathrm{i}$$

  - Norm for constant function on domain is 1
  - Examples

$$\left\| x \right\| := \max_{z \in \Omega_i} \left| x_z \right|$$

$$\left\| x \right\| := \sqrt{\frac{1}{\left| \Omega_i \right|} \sum_{z \in \Omega_i} \left| x_z \right|^2}$$

Overview FDM

Expectations when doubling the resolution?

Continuous Model

Finite differences
Resolution

$$\|u - u_h\| = \|e_{\text{disc.}}\|$$

Discrete Model

Solver
Finite precision/arithmetics

$$\|u_h - \hat{u}_h\| = \|e_{\text{algeb.}}\|$$

Approximate Solution

Norms

Residual

$$\|A_h \hat{u}_h - b_h\| = \|\mathrm{r}_h\|$$

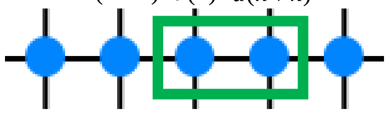
Taylor Series Expansion

- Approximate “nice” function around a point by its derivatives

$$u(x) = u(x_0) + u'(x_0)(x - x_0) + \frac{u''(x_0)}{2!}(x - x_0)^2 + \frac{u'''(x_0)}{3!}(x - x_0)^3 + \ldots + \frac{u^{(n)}(x_0)}{n!}(x - x_0)^n$$

- Forward finite difference for first derivative

$$\frac{\partial u}{\partial x} = u_{+x} \approx \frac{u_{+1} - u_0}{h}$$



$$\frac{u(x) - u(x_0)}{(x - x_0)} = u'(x_0) + \underbrace{\left(\sum_{n=2}^{\infty} \frac{u^{(n)}(x_0)}{n!}(x - x_0)^n \right) \cdot \frac{1}{(x - x_0)}}_{R(\ldots)}$$

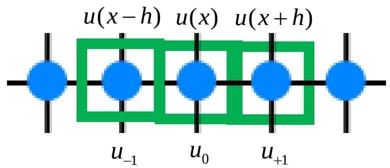
setting $\quad x = x_0 + h \quad$ and $\quad \xi$ between x_0 and $x_0 + h$

$$\frac{u(x_0 + h) - u(x_0)}{(h)} = u'(x_0) + \underbrace{\underbrace{\frac{u^{(2)}(\xi)}{2!}(h)^2}_{R(h^2)} \cdot \frac{1}{h}}_{O(h)}$$

- Remainder term is O(h)
  - Same holds for backward difference

Taylor Series Expansion

- Central difference for second derivative $\dfrac{\partial u^2}{\partial x^2} = u_{xx} \approx \dfrac{u_{+1} - 2u_0 + u_{-1}}{h^2}$



setting $\quad x = x_0 + h$

$$u(x_0+h) = u(x_0) + u'(x_0)(h) + \frac{u''(x_0)}{2!}(h)^2 + \frac{u'''(x_0)}{3!}(h)^3 + \ldots + \frac{u^{(n)}(x_0)}{n!}(h)^n$$

setting $\quad x = x_0 - h$

$$u(x_0-h) = u(x_0) - u'(x_0)(h) + \frac{u''(x_0)}{2!}(h)^2 - \frac{u'''(x_0)}{3!}(h)^3 + \ldots + \frac{u^{(2n)}(x_0)}{2n!}(h)^{2n} - \frac{u^{(2n+1)}(x_0)}{(2n+1)!}(h)^{2n+1}$$

$$u(x_0+h) + u(x_0-h) = 2u(x_0) + 2\frac{u''(x_0)}{2!}(h)^2 + 2\underbrace{\left(\sum_{n=2}^{\infty}\frac{u^{(2n)}(x_0)}{(2n)!}(h)^{2n}\right)}_{R(\ldots)}$$

$$\frac{u(x_0+h) + u(x_0-h) - 2u(x_0)}{(h)^2} = u''(x_0) + \underbrace{2\underbrace{\left(\frac{u^{(4)}(x_0)}{(4)!}(h)^4\right)}_{R(h^4)} \cdot \frac{1}{h^2}}_{O(h^2)}$$

- Remainder term is $O(h^2)$

Consistency of FDM

- Example $$-\Delta u = -(u_{xx} + u_{yy}) = 0 \quad in \quad \Omega = (0,1)\times(0,1)$$

- Using second order central differences

$$\frac{u(x_0+h)+u(x_0-h)-2u(x_0)}{(h)^2} = u''(x_0) + O(h^2)$$

- Test case
  - Find smooth analytic solution
  - Solve discretized problem for h, h/2, h/4, …
  - Assuming an algeb. error << disc. Error
  - Expected behavior of total error

$$\left\| u - \hat{u}_h \right\|_\infty \le Ch^2$$

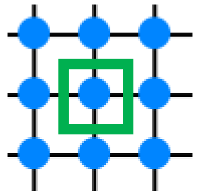
$$\left\| u - \hat{u}_{h/2} \right\|_\infty \le C\left(\frac{h}{2}\right)^2 = Ch^2\frac{1}{4}$$

  - Same holds for residual norm if condition of A_h is “ok”

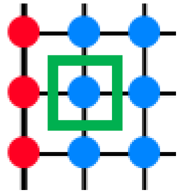
Boundary Conditions

- Interior point

$$-\frac{1}{h^2}\left(u_N + u_S + u_E + u_W - 4u_C\right) = 0$$



- Dirichlet BC

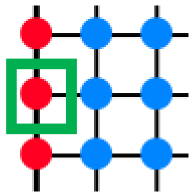


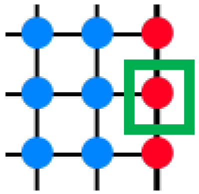
$$u_w = g_1$$

$$-\frac{1}{h^2}\left(u_N + u_S + u_E \qquad - 4u_C\right) = \frac{1}{h^2} g_1$$

- Neumann BC
  - Approximation of first derivative normal to the boundary

$$\frac{\partial u}{\partial n} = \frac{\partial u}{\partial x} = g_2$$





- Robin BC

$$\frac{\partial u}{\partial n} + \alpha u = g_3$$

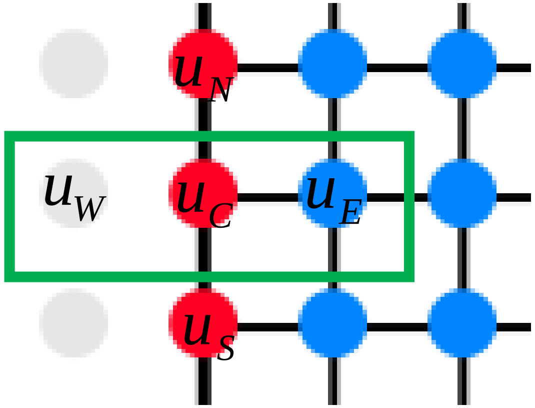
Boundary Conditions

- Neumann BC
  - Central difference (ghost point layer)

$$\frac{\partial u}{\partial n} = \frac{\partial u}{\partial x} = g_2 \approx \frac{u_E - u_W}{2h} \rightarrow u_W = -2hg_2 + u_E$$

$$-\frac{1}{h^2}\left(u_N + u_S + (-2hg_2 + u_E) + u_E - 4u_C\right) = 0$$

$$-\frac{1}{h^2}\left(u_N + u_S + 2u_E - 4u_C\right) = -\frac{2}{h} g_2$$



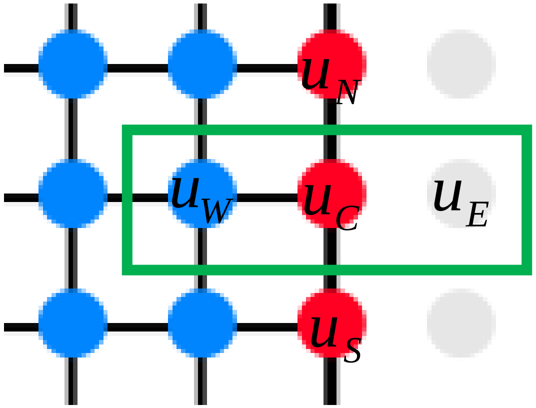
Boundary Conditions

- Robin BC
  - Central difference (ghost point layer)

$$\frac{\partial u}{\partial n}+\alpha u=g_3\approx\frac{u_E-u_W}{2h}+\alpha u_C\rightarrow u_E=2hg_3+u_W-2h\alpha u_C$$

$$-\frac{1}{h^2}\left(u_N+u_S+(2hg_3+u_W-2h\alpha u_C)+u_W-4u_C\right)=0$$

$$-\frac{1}{h^2}\left(u_N+u_S+2u_W-(4+2h\alpha)u_C\right)=\frac{2}{h}g_3$$



Consistency of FDM

- Example $$-\Delta u = -(u_{xx} + u_{yy}) = 0 \quad in \quad \Omega = (0,1)\times(0,1)$$
- Using second order central differences
$$\frac{u(x_0+h)+u(x_0-h)-2u(x_0)}{(h)^2} = u''(x_0) + O(h^2)$$
- Using forward/backward difference for Neumann or Robin BC
$$\frac{u(x_0)+u(x_0-h)}{(h)^2} = u'(x_0) + O(h) \qquad \frac{u(x_0+h)+u(x_0)}{(h)^2} = u'(x_0) + O(h)$$
- Test case
  - Find smooth analytic solution
  - Solve discretized problem for h, h/2, h/4, …
  - Assuming an algeb. error << disc. Error
  - Expected behavior of total error
$$\left\| u - \hat{u}_h \right\|_\infty \le Ch$$
$$\left\| u - \hat{u}_{h/2} \right\|_\infty \le C\frac{h}{2}$$

Finite Precision

$$u_h - \hat{u}_h = e_{\text{algeb.}} \qquad u - u_h = e_{\text{disc.}}$$

- We assumed algebraic error << discretization error
  - Condition number of the problem
  - Solver (e.g., number of iterations)
  - Finite representations and arithmetics used during calculation

- Example
  - Ill-conditioned system
  - Double precision FP
    - ~16 digits decimal precision
  - Single precision FP
    - ~7digits decimal precision

$$\underbrace{\begin{bmatrix} 0 & -1 \\ 0+\beta & -1 \end{bmatrix}}_{A} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 1+\beta \end{bmatrix}$$

$$y = -1$$

$$x = \frac{1+\beta-1}{\beta} = 1$$

$$\beta = 1\mathrm{e}^{-7} \rightarrow \kappa(A) = 2\mathrm{e}^{7}$$

$$\beta = 1\mathrm{e}^{-15} \rightarrow \kappa(A) = 2\mathrm{e}^{15}$$

Base 2 Floating Point Representation

- Reasons
  - Hardware implementation
  - Error analysis has tight bounds
  - Extra bit of precision (through normalization)

$$\pm s \cdot 1.dddddddddd \cdot 2^{eeeee}$$

- Number of significant decimal digits

$$2^{(binary\,precision)} = 10^{(decimal\,precision)}$$

$$\log_{10}(2^{(binary\,precision)}) = (decimal\ \ precision)$$

$$\log_{10}(2^{52+1}) \approx 16$$

$$\log_{10}(2^{23+1}) \approx 7.2$$

$$\log_{10}(2^{10+1}) \approx 3.3$$

Floating Point Representation

s	e	e	e	e	e	d	d	d	d	d	d	d	d	d	d

- IEEE 754 16bit floating point representation
  - Exponent with 5 digits “e”
  - Significant with 10 digits “d” (precision=10)
  - Sign encoded with 1 digits “s”
  - Base is 2, so digits are bits with state 0 or 1
  - Exponent
    - 00000 = subnormal numbers (for significant>0), otherwise zero
    - 00001 = min, 01111 = bias (=0), 11110 = max
    - 11111 = NaN (for significant>0), otherwise +-infinity

$$\pm s \cdot d.dddddddddd \cdot 2^{eeeee}$$

$$\pm 1 \cdot 2^{00000} \cdot 0.0000000000 = \pm 1 \cdot 2^{0} \cdot \left(0 + \frac{0}{2^{10}}\right) = \pm 0$$

$$\pm 1 \cdot 2^{01111} \cdot 1.0000000000 = \pm 1 \cdot 2^{15-15} \cdot \left(1 + \frac{0}{2^{10}}\right) = \pm 1$$

$$\pm 1 \cdot 2^{11110} \cdot 1.1111111111 = \pm 1 \cdot 2^{30-15} \cdot \left(1 + \frac{2^{10}-1}{2^{10}}\right) = \pm 65504$$

Quiz

Q1: What are the consequences/differences when using the Maximum norm or Euclidean norm to quantify the residual?

Q2: Which of the discussed matrix norms is the ‘cheapest’ in terms of computational effort?

Q3: What is the binary representation of “1000” in the IEEE 754 16bit/32bit/64bit FP format?

Q4: What is the difference between BLAS routine ‘dgemm’ and ‘sgemm’ / What does the LAPACK routine ‘dsysv’ do?

Q5: When would you advise to perform a LU decomposition of a matrix instead of a QR decomposition?