PROGRAM ROZWOJOWY
POLITECHNIKI WARSZAWSKIEJ

Advanced Computational Fluid Dynamics

Jacek Rokicki

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Table of content

1. Introduction

The content of this book covers lectures in Advanced Computational Fluid Dynamics held at the Faculty of Power and Aeronautical Engineering since 2006. The lectures evolved over the last 8 years broadening in scope and involving newer topics.

The original content was very much inspired by the works of Randall J. LeVeque and in particular by his book *Numerical Methods for Conservation Laws*[1].

The present book is organized by starting with the general equation of Fluid Mechanics and then by analysis of various model problems, which help to understand the complexity of the multidimensional, nonlinear Navier-Stokes, Euler equations and their discretisations. Various topics in numerical analysis and algebra (notably the algebraic eigenproblem) are also introduced to make the exposition complete for the reader. Certain topics are recalled from the more elementary course in Computational Fluid Dynamics held for the undergraduate students.

Warsaw, 2014-2016

[1] Randall J. LeVeque, Numerical Methods for Conservation Laws, 1990 Birkhäuser, ISBN 978-3-7643-2464-3.

2. Navier-Stokes equations

The Navier-Stokes equation for compressible medium are best presented in the unified manner which underlines their conservative structure.

$$\frac{\partial U}{\partial t} + \mathrm{Div}\, \mathbb{F}_c(U) = \mathrm{Div}\, \mathbb{F}_v(U, \nabla U) \tag{1}$$

where:

$$U = \begin{bmatrix} \rho \\ \rho V \\ \rho E \end{bmatrix} \in \mathbb{R}^5, \quad V = \begin{bmatrix} u \\ v \\ w \end{bmatrix} \in \mathbb{R}^3, \quad \mathbb{F}_c(U) = \begin{bmatrix} \rho V^T \\ \rho V V^T + p I_{3\times3} \\ (\rho E + p) V^T \end{bmatrix}_{5\times3} \qquad (2) \qquad (2)$$

In the above U stands for the composite conservative variable, ρ is the density, $V = [u, v, w]^T$ denotes a velocity vector, E is a total energy per unit mass, p stands for pressure, while $H = E + \frac{p}{\rho}$ is the total enthalpy per unit mass. The $\mathbb{F}_c(U)$ and $\mathbb{F}_v(U, \nabla U)$ stand for the convective and viscous fluxes respectively. The viscous flux can be expressed as:

$$\mathbb{F}_v(U, \nabla U) = \begin{bmatrix} 0 \\ \tau_{3\times3} \\ V^T \tau_{3\times3} - q^T \end{bmatrix}_{5\times3} \tag{3}$$

where q stands for a heat flux, while τ denotes the stress tensor. Both these quantities in Fluid Mechanics (and especially the stress tensor) can be defined via very different formulas, in particular when modelling of turbulence is attempted, nevertheless for simple Newtonian, linear fluid they are defined/calculated as:

$$\begin{gathered} \tau_{3\times3} = \mu \left[-\frac{2}{3} (\nabla \cdot V) I_{3\times3} + (\nabla V)^T + \nabla V \right] \\ q = -\lambda \, \nabla T \end{gathered} \tag{4}$$

where μ and λ stand for coefficients of dynamic viscosity and thermal conductivity respectively. It is good to remember that for air and water $\frac{\mu}{\rho}$ is very small and equals $\sim 10^{-5} \frac{m^2}{s}$ and $\sim 10^{-6} \frac{m^2}{s}$ respectively. This is the reason why in the Navier-Stokes equation the convective flux plays, in a sense, more important role than the viscous flux (at least for higher Reynolds numbers).

In addition, for perfect gas, the equation of state is assumed in the usual form: $\frac{p}{\rho} = RT$ (where T stands for temperature, while R is an ideal gas constant).

The tensor divergence operator Div present in (1) can be expressed for clarity in an extended form as:

$$\mathrm{Div}\,\mathbb{F}_c(U) = \begin{bmatrix} \mathrm{div}(\rho V) \\ \mathrm{div}(\rho u V) + \dfrac{\partial p}{\partial x} \\ \mathrm{div}(\rho u V) + \dfrac{\partial p}{\partial y} \\ \mathrm{div}(\rho w V) + \dfrac{\partial p}{\partial z} \\ \mathrm{div}(\rho H V) \end{bmatrix}_{5\times 1} \tag{5}$$

Where div denotes a usual scalar divergence operator acting on a vector functions.

3. Euler equations

The Euler equations are obtained from Navier-Stokes equation (1) by assuming that $\mu \equiv 0$ and $\lambda \equiv 0$ - the fluid is inviscid and does not conduct heat. In this case the equations are significantly simplified:

$$\frac{\partial U}{\partial t} + \mathrm{Div}\,\mathbb{F}_c(U) = 0 \tag{6}$$

And in slightly different notation they assume the form:

$$U = \begin{bmatrix} \rho \\ \rho V \\ \rho E \end{bmatrix} = \begin{bmatrix} \rho \\ m \\ \epsilon \end{bmatrix} \in \mathbb{R}^5, \qquad \mathbb{F}_c(U) = \begin{bmatrix} m^T \\ \rho^{-1}\, m m^T + p I_{3\times 3} \\ H m^T \end{bmatrix}_{5\times 3} \tag{7}$$

Where for perfect gas:

$$\begin{gathered} H = E + \frac{p}{\rho} = \frac{1}{\rho}(\epsilon + p) \\ p = (\gamma - 1)\left(\epsilon - \frac{m^T m}{2\rho}\right), \qquad \gamma = c_p/c_v \\ H = \frac{\gamma\epsilon}{\rho} - (\gamma - 1)\frac{m^T m}{2\rho^2} \end{gathered} \tag{8}$$

4. Euler equations in 1D

For 1D cases Euler equations are further simplified, to:

$$\frac{\partial U}{\partial t} + \frac{\partial}{\partial x}\mathbb{F}_c(U) = 0 \tag{9}$$

with:

$$U = \begin{bmatrix} \rho \\ \rho u \\ \rho E \end{bmatrix} = \begin{bmatrix} \rho \\ m \\ \epsilon \end{bmatrix} \in \mathbb{R}^3, \qquad \mathbb{F}_c(U) = \begin{bmatrix} m \\ \dfrac{m^2}{\rho} + p \\ H m \end{bmatrix}_{3\times 1} \tag{10}$$

and:

$$p = (\gamma - 1)\left(\epsilon - \frac{m^2}{2\rho}\right), \qquad \gamma = c_p/c_v \tag{11}$$

$$H = \frac{\gamma\epsilon}{\rho} - (\gamma - 1)\frac{m^2}{2\rho^2}$$

5. Model problems

In order to better understand the principles of discretisation, various model problems will be considered here, including:

1. 1D elliptic problem
2. 2D Poisson equation
3. Advection equation
4. Advection-diffusion equation
5. 1D parabolic problem
6. Telegraph equation
7. Multidimensional hyperbolic problem
8. Nonlinear advection problem
9. Burgers equation

The analysis of these discretisations will be made possible, by applying theoretical tools mainly related to the algebraic eigenvalue problem.

6. Scalar products and norms

Scalar product

Let's consider a vector space Vover $\mathbb{R}$ or $\mathbb{C}$, the elements of this space will be denoted by $u, v, w \in V$, while scalars are denoted as $\alpha, \beta, \gamma, \in \mathbb{R}$ or $\mathbb{C}$:

Definition

The scalar product is defined as a two argument function

$$(\cdot\,,\cdot): V \times V \to \mathbb{R} \text{ or } \mathbb{C} \qquad (12)$$

with the following axiomatic conditions:

$$\begin{aligned}
&\text{(i)} \quad (u\,,u) \geq 0, \ \ (u,u) = 0 \ \Leftrightarrow \ \ u \equiv 0 \\
&\text{(ii)} \ \ (\alpha u\,,v) = \alpha(u\,,v) \\
&\text{(iii)} \ (u\,,v) = \overline{(v\,,u)} \\
&\text{(iv)} \ (\alpha u + \beta v\,,w) = \alpha(u\,,w) + \beta(v\,,w)
\end{aligned} \qquad (13)$$

Examples:

The most popular formula for the finite dimension space $\mathbb{C}^n$ is:

$$(u\,,v) \overset{\text{def}}{=} \sum_{j=1}^{n} \overline{u_j} v_j \equiv u^H v \qquad (14)$$

But other forms are also possible:

$$(u\,,v)_* \overset{\text{def}}{=} \sum_{j=1}^{n} \beta_j^2 \cdot \overline{u_j} v_j \qquad (15)$$

$$\boxed{(u\,,v)_{**} \overset{\text{def}}{=} u^H A v \equiv \sum_{i=1}^{n} \sum_{j=1}^{n} \overline{u_j} a_{ij} v_j} \qquad (16)$$

where A Is positive definite ($\forall u \neq 0, u^H A u > 0$) Hermitian matrix ($A = A^H \equiv \overline{A^T}$)

Remark:

If the matrix is Hermitian, then $A = A^H \Leftrightarrow (Au, v)_{\mathbb{C}} = (u, Av)_{\mathbb{C}}$

Proof:

$$\forall u, v, \qquad (Au, v)_{\mathbb{C}} = (Au)^H v = u^H A v = u^H (Av) = (u, Av)_{\mathbb{C}} \qquad (17)$$

or:

$$\forall u, v, \qquad (Au, v)_{\mathbb{R}} = (Au)^T v = u^T A v = u^T (Av) = (u, Av)_{\mathbb{R}} \qquad (18)$$

This is the rationale for defining the Hermitian/symmetric matrices and subsequently operators via (17), (18) rather than by the definition using the matrix elements. The Hermitian operators are further on called self-adjoint.

The examples of scalar products for operators are given below:

$$(u\,,v)_{L^2(\Omega)} \overset{\text{def}}{=} \int_\Omega u(x)v(x)\,d\Omega,$$
$$(u\,,v)_{H^1(\Omega)} \overset{\text{def}}{=} \int_\Omega u(x)v(x) + \frac{du}{dx}\frac{dv}{dx}\,d\Omega, \tag{19}$$

The important property of each scalar product is now recalled, named a Cauchy-Schwarz inequality:

$$\forall u,v \quad |\,(u,v)| \le (u,u)(v,v), \tag{20}$$

As a result we are always able to define the angle between vectors u and v for every real valued scalar product, as:

$$u,v \in L^2(\Omega), \quad \cos\big(\theta(u,v)\big) \overset{\text{def}}{=} \frac{(u,v)}{\sqrt{(u,u)}\sqrt{(v,v)}} \tag{21}$$

This formula allows in turn to infer the value of the angle of θ.

Example:

Suppose now that $u(x) = x$, $v(x) = x^2$, while $\Omega = \langle 0,1\rangle$. We will calculate the angle θ between u and v:

$$\cos\theta = \frac{\int_0^1 x\,dx}{\sqrt{\int_0^1 x^2\,dx}\sqrt{\int_0^1 x^4\,dx}} = \frac{\frac{1}{4}}{\sqrt{\frac{1}{3}}\sqrt{\frac{1}{5}}} = \frac{\sqrt{15}}{4} \Longrightarrow \theta \approx 14.48° \tag{22}$$

Exercise: What is the angle between these functions in the H^1scalar product ?

Definition:

Vectors u and v are called *orthogonal* $\Leftrightarrow$ $(u,v) = 0$

The notion of *orthogonality* is very important and will be used extensively in the further exposition.

The space with the scalar product is called an inner-product space or unitary space.

Vector norms

The norm of a vector, is defined in a following axiomatic way as a non-negative function fulfilling three conditions ($u,v \in V,\ \ \alpha \in \mathbb{R}$):

$$\|\cdot\| : V \overset{\text{def}}{=} \mathbb{R}_+ \cup \{0\}$$

(i) $\|u\| \ge 0,\ \ \|u\| = 0 \Leftrightarrow\ \ u \equiv 0$

(ii) $\|\alpha u\| = |\alpha|\|u\|$ (23)

(iii) $\|u + v\| \le \|u\| + \|v\|$

Examples (Hölder norms):

$$u \in \mathbb{C}^n \text{ lub } \mathbb{R}^n, \qquad \|u\|_p = \sqrt[p]{\sum_{j=1}^{n} |u_j|^p} \tag{24}$$

In particular:

$$\|u\|_1 = \sum_{j=1}^{n} |u_j|, \quad \|u\|_2 = \sqrt{\sum_{j=1}^{n} |u_j|^2}, \quad \|u\|_\infty = \max_j |u_j| \qquad (25)$$

It is important to notice that each scalar product generates a norm:

$$\|u\| \stackrel{\text{def}}{=} \sqrt{(u,u)} \qquad (26)$$

thus each unitary space is also a normed space. This is not true the other way round, usually the norm is not generated by any scalar product. Out of all Hölder norms only $p = 2$ corresponds to a scalar product. For functions (infinite dimensional functional spaces) the norms are defined in a manner analogous to Hölder vector norms:

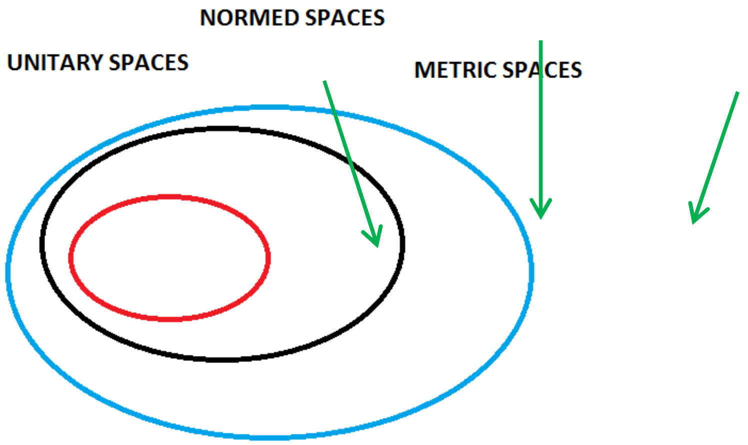
- For integrable functions: – $\|u\|_{L^1(\Omega)} \stackrel{\text{def}}{=} \int_\Omega |u(x)|\, dx$
- For square integrable functions: – $\|u\|_{L^2(\Omega)} \stackrel{\text{def}}{=} \int_\Omega |u(x)|^2\, dx$
- For continuous functions: – $\|u\|_{C(\Omega)} = \max_{x\in\Omega} |u(x)|$ (27)
- For functions with square integrable first derivative: – $\|u\|_{H^1(\Omega)} \stackrel{\text{def}}{=} \int_\Omega |u(x)|^2 + |u'(x)|^2\, dx$

Remark:

Each norm generates a metric (distance function) via the formula:

$$\rho(u,v) \stackrel{\text{def}}{=} \|u - v\| \qquad (28)$$

Thus every normed space is also a metric space (but again not vice versa). This is illustrated in the graph below:



7. Algebraic eigenproblem

For the purpose of further analysis we recall now the most important features of the algebraic (and also operator) eigenproblems.

Consider real or complex values matrices $A \in \mathbb{R}^{n\times n}$ or $A \in \mathbb{C}^{n\times n}$. The algebraic eigenvalue problem consists in finding nonzero $u \in \mathbb{R}^n$ or $\mathbb{C}^n$, such that:

$$Au = \lambda u \tag{29}$$

where $\lambda \in \mathbb{C}$ denotes eigenvalue corresponding to the eigenvector u.

Interpretation: We seek the "direction" u, which remains unchanged after u is multiplied by A, (only the length is modified).

Properties:

$$\begin{gathered}(A-\lambda I)u = 0 \Leftrightarrow \det(A-\lambda I) = 0 \\ \Updownarrow \\ c_n\lambda^n + c_{n-1}\lambda^{n-1} + \cdots + c_1\lambda^1 + c_0 = 0\end{gathered} \tag{30}$$

The last formula forms characteristic polynomial, which is obtained by calculation of the determinant above, indeed $c_0 = \det(A)$.

Therefore the singular matrix has at least one zero eigenvalue, while all eigenvalues of non-singular matrix are non-zero. From the properties of polynomials we see that the matrix has always n eigenvalues (not necessarily distinct and not always real valued, even for the real matrices).

The matrices A and A^T have the same eigenvalues The real matrix $A \in \mathbb{R}^{n\times n}$ may have complex eigenvalues which are then always pairwise conjugated.

There exist no-finite algorithm to find the eigenvalues of A, as there exist no finite algorithm to find the roots of the polynomial of sufficiently high degree (characteristic polynomial in this case).

Out of numerical reasons the coefficients of the characteristic polynomial should never be directly evaluated (as they accumulate all round-off errors).

Examples

1. The case when the matrix is the identity:

$$\begin{gathered}A = I \Leftrightarrow Iu = \lambda u \\ \lambda_1 = \lambda_2 = \cdots = \lambda_{n-1} = \lambda_n = 1\end{gathered} \tag{31}$$

The eigenvector can be quite arbitrary, but in particular, it can be a versor of one of the axes:

$$u_{(p)} = e_{(p)} = \begin{bmatrix}0\\ \vdots \\ 0\\ 1\\ 0\\ \vdots\\ 0\end{bmatrix} \Leftarrow \text{position } p, \quad p = 1,2,\dots,n \tag{32}$$

2. The case of the diagonal matrix:

$$A = D = \operatorname{diag}(d_j) = \begin{bmatrix} d_1 & 0 & \cdots & 0 & 0 \\ 0 & d_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \ddots & & \\ 0 & & & d_{n-1} & 0 \\ 0 & 0 & & 0 & d_n \end{bmatrix} \Leftrightarrow Du = \lambda u \tag{33}$$

$$\lambda_1 = d_1, \qquad \lambda_2 = d_2, \dots \quad , \lambda_n = d_n$$

The eigenvectors in this case are the same as previously (without however the possibility to choose the eigenvectors in a different way)

$$u_{(p)} = e_{(p)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Leftarrow \text{position } p \tag{34}$$

3. The case of the Jordan block

$$A = J = \begin{bmatrix} c & 1 & \cdots & 0 & 0 \\ 0 & c & \cdots & 0 & 0 \\ \cdots & \cdots & \ddots & 1 & \\ 0 & & & c & 1 \\ 0 & 0 & & 0 & c \end{bmatrix} \Leftrightarrow Ju = \lambda u$$

$$det(J - \lambda I) = (c - \lambda)^n \tag{35}$$

$$\Updownarrow$$

$$\lambda_1 = \lambda_2 = \cdots = \lambda_{n-1} = \lambda_n = c$$

We have again the case of multiple eigenvalue. However in this case the matrix has only one eigenvector. To verify this we consider the sequence of equations $Ju = cu$:

$$\begin{aligned} cu_1 + u_2 &= cu_1 && \Longrightarrow\ u_2 = 0 \\ cu_2 + u_3 &= cu_2 && \Longrightarrow\ u_3 = 0 \\ &\dots && \dots \\ cu_{n-1} + u_n &= cu_{n-1} && \Longrightarrow\ u_{n-1} = 0 \\ cu_n &= cu_n && \Longrightarrow\ u_n = 0 \end{aligned} \tag{36}$$

Therefore the only eigenvector has the form:

$$u_{(1)} = e_{(1)} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{37}$$

Summary: We have shown that matrices with multiple eigenvalues can have different number of eigenvectors.

Main properties of the eigenvalues and eigenvectors

1. If u is an eigenvector, then also $\alpha \cdot u$ is an eigenvector.

2. If $u_{(1)},\ldots,u_{(m)}$ are eigenvectors corresponding to the eigenvalue λ, then also $v=\sum_{j=1}^{m}\alpha_j\cdot u_{(j)}$ is an eigenvector corresponding to the same eigenvalue. To verify this we check: $\forall j,\ Au_{(j)}=\lambda u_{(j)}\ \Longrightarrow Av=\sum_{j=1}^{m}\alpha_j\cdot Au_{(j)}=\sum_{j=1}^{m}\alpha_j\cdot\lambda u_{(j)}=\lambda\sum_{j=1}^{m}\alpha_j\cdot u_{(j)}=\lambda v$, which shows that v is indeed an eigenvalue.
3. Eigenvectors corresponding to different eigenvalues are linearly independent.

$$\begin{aligned} &Au_{(1)}=\lambda_1 u_{(1)} && \text{and } \lambda_1\neq\lambda_2\ \Longrightarrow u_{(1)} \text{ and } u_{(2)}\\ &Au_{(2)}=\lambda_1 u_{(2)} && \text{are linearly independent}\end{aligned}\tag{38}$$

To verify this we present partial proof. Assume that $u_{(1)}$ and $u_{(2)}$ are linearly dependent, i.e., $u_{(1)}=\beta\, u_{(2)}$:

$$\begin{aligned} &Au_{(1)}=\lambda_1 u_{(1)} && \Longrightarrow(\lambda_1-\lambda_2)u_{(1)}=0\ \Longrightarrow\\ &A\beta u_{(1)}=\lambda_2\beta u_{(1)} && \lambda_1=\lambda_2 \text{ contradiction}\end{aligned}\tag{39}$$

4. If A has n distinct eigenvalues $|\lambda_1|<|\lambda_2|<\cdots<|\lambda_{n-1}|<\lambda_n$, then the corresponding eigenvectors $u_{(1)},\ldots,u_{(n)}$ form a basis in $\mathbb{R}^n$ (eigenvectors are linearly independent).
5. All eigenvalues of Hermitian matrix ($A=A^H\equiv\overline{A^T}$) are real. To prove this lets take an arbitrary eigenvector u and calculate the suitable scalar product. The fact that matrix is Hermitian implies that $(Au,u)=(u,A^Hu)=(u,Au)$:

$$\begin{aligned}(Au,u)&=(\lambda u,u)=\lambda(u,u)\\ (Au,u)&=(u,Au)=(u,\lambda u)=\bar\lambda(u,u)\end{aligned}\qquad\Rightarrow\lambda=\bar\lambda\tag{40}$$

As $(u,u)\neq 0$ this implies that the eigenvalue is always real.

6. Eigenvectors corresponding to distinct eigenvalues of Hermitian matrix are orthogonal. To prove this lets consider two eigenvectors u and v corresponding to the different eigenvalues λ and μ respectively.

$$\begin{gathered} A=A^H,\quad Au=\lambda u,\quad Av=\lambda v,\quad \lambda\neq\mu\\ (Au,v)=(\lambda u,v)=\lambda(u,v)\\ (Au,v)=(u,Av)=(u,\mu v)=\mu(u,v)\\ \Downarrow\\ (\lambda-\mu)(u,v)=0\ \Rightarrow\ (u,v)=0\end{gathered}\tag{41}$$

We have shown that the eigenvectors are indeed orthogonal.

7. The previous theorem can be further extended, as it appears that actually all eigenvectors of Hermitian matrix are orthogonal (be the eigenvalues distinct or not). In fact the eigenvectors of Hermitian matrix form a basis in $\mathbb{C}^n$.
8. The eigenvalues of symmetric (Hermitian) positive-definite matrix are positive. A symmetric (Hermitian) positive-definite real matrix is defined as such that for arbitrary non-zero vector u we have always $(Au,u)=u^TAu>0\ or\ (Au,u)=u^HAu>0$. To show that the eigenvalues are positive consider now the eigenvector u corresponding to the eigenvalue λ.

$$0<(Au,u)=\lambda(u,u)\ \Rightarrow\ \lambda>0\quad\text{as always }(u,u)>0\tag{42}$$

It is interesting to note that for arbitrary matrix A the matrix $A^H A$ is symmetric (Hermitian) and positive-definite, as

$$(A^H Au, u) = u^H A^H Au = (Au, Au) = \|Au\|_2^2 > 0 \tag{43}$$

thus the eigenvalues of $A^H A$ are always real and positive.

Similar matrices

Def. Matrices B and A are similar if for some invertible matrix Q:

$$B = Q^{-1}AQ \tag{44}$$

Properties:

1. Similar matrices have the same eigenvalues:

$$\det(B - \lambda I) = \det(Q^{-1}AQ - \lambda Q^{-1}IQ) = \det(Q^{-1}(A - \lambda I)Q) = \det(A - \lambda I) \tag{45}$$

2. The eigenvectors of the similar matrices are however different:

$$\begin{gathered} Au = \lambda u \;\Rightarrow\; Q^{-1}Au = \lambda Q^{-1}u \;\Rightarrow \\ (Q^{-1}AQ)(Q^{-1}u) = \lambda(Q^{-1}u) = Bv = \lambda v\,, \qquad v = (Q^{-1}u) \end{gathered} \tag{46}$$

Jordan matrix (Jordan normal form of a matrix A)

Theorem:

Any $n \times n$ square matrix A is similar to a Jordan matrix J (unique up to a permutation of blocks):

$$\begin{gathered} \forall A\ \exists Q, \qquad det(Q) \neq 0, \quad J = Q^{-1}AQ \\ J = \begin{bmatrix} J_1 & 0 & \cdots & 0 & 0 \\ 0 & J_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \ddots & & \\ 0 & & & J_{m-1} & 0 \\ 0 & 0 & & 0 & J_m \end{bmatrix}, \quad m \leq n \\ J_p = \begin{bmatrix} \lambda_p & 1 & \cdots & 0 & 0 \\ 0 & \lambda_p & \cdots & 0 & 0 \\ \cdots & \cdots & \ddots & 1 & \\ 0 & & & \lambda_p & 1 \\ 0 & 0 & & 0 & \lambda_p \end{bmatrix}_{n_p \times n_p} \qquad n_1 + n_2 + \cdots + n_m = n \end{gathered} \tag{47}$$

Properties:

1. The eigenvalues λ_p and λ_q from two different blocks are not necessarily distinct.
2. Each block corresponds to linearly independent eigenvector, thus matrix A has m linearly independent eigenvectors (the eigenvectors of J are $\left[e_{(1)},\ e_{(n_1+1)}, e_{(n_2+1)}, e_{(n_{m-1}+1)}\right]$)
3. If $m = n$ we call matrix A diagonalizable as J is strictly diagonal, and all blocs are (1×1):

$$J = \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \ddots & & \\ 0 & & & \lambda_{n-1} & 0 \\ 0 & 0 & & 0 & \lambda_n \end{bmatrix} = \operatorname{diag}(\lambda_p) \tag{48}$$

4. Hermitian matrices $A = A^H$ are diagonalisable
5. Normal matrices $AA^H = A^H A$ are diagonalizable
6. Many other matrices are diagonalizable
7. If A is diagonalizable, then:

$$\Lambda = Q^{-1}AQ \quad \Rightarrow AQ = \Lambda Q \tag{49}$$

Which means that eigenvectors of A are the columns of Q.

The Jordan theorem does not provide an aid in computations (it is not constructive), however it characterises all possible configurations of eigenvalues and eigenvectors the matrix can have. It also characterises an important class of diagonalizable matrices.

Power method to calculate eigenvalues

We will present a simplest method to calculate the largest eigenvalue of real and diagonalisable matrix $A, \left(Au_{(1)} = \lambda_1 u_{(1)}, \dots \ , Au_{(p)} = \lambda_p u_{(p)}, \dots \ Au_{(n)} = \lambda_n u_{(n)}, \right)$ with eigenvalues sorted in decreasing order $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots |\lambda_n|$, and the first eigenvalue being separated, i.e., larger than all others. We assume also that the eigenvectors $u_{(p)}, p = 1, \dots, n$ form the basis in $\mathbb{R}^n$.

The algorithm:

w_0 arbitrary vector (of random elements)

$w_k = Aw_{k-1}, \ \ k = 1,2, \dots$

$$\lim_{k\to\infty} w_k = u_{(1)} \quad \text{and} \quad \lambda_1 = \frac{\left(Au_{(1)}, u_{(1)}\right)}{\left(u_{(1)}, u_{(1)}\right)} \tag{50}$$

Proof of convergence:

$$w_0 = \sum_{p=1}^{n} \alpha_p u_{(p)} \quad \text{representation in the basis of eigenvectors}$$

$$w_1 = Aw_0 = \sum_{p=1}^{n} \alpha_p A u_{(p)} = \sum_{p=1}^{n} \alpha_p \lambda_p u_{(p)}$$

$$\Downarrow$$

$$w_k = Aw_{k-1} = \sum_{p=1}^{n} \alpha_p \lambda_p^{k-1} A u_{(p)} = \sum_{p=1}^{n} \alpha_p \lambda_p^{k} u_{(p)} = \tag{51}$$

$$= \alpha_1 \lambda_1^k u_{(1)} + \sum_{p=2}^{n} \alpha_p \lambda_p^k u_{(p)} = \lambda_1^k \left[\alpha_1 u_{(1)} + \sum_{p=2}^{n} \alpha_p \left(\frac{\lambda_p}{\lambda_1}\right)^k u_{(p)} \right] \rightarrow$$

$$\xrightarrow[k\to\infty]{} \lambda_1^k \alpha_1 u_{(1)}$$

With the sum in square bracket vanishing as $\left(\frac{\lambda_p}{\lambda_1}\right)^k \xrightarrow[k\to\infty]{} 0$

In practical computations w_k has to be normalised at each iteration in order to avoid its exponential growth (which is not dangerous in theoretical considerations).

Solving the linear equation having the eigenvalues and eigenvectors of the matrix

Suppose that we have Hermitian matrix $A = A^H$ with known eigenvalues and eigenvectors $Au_{(p)} = \lambda_p u_{(p)},\ p = 1, 2, \ldots, n$, and with the eigenvectors which are orthonormal $\left(u_{(p)}, u_{(q)}\right) = \delta_{pq}$

Suppose now that we want to solve the linear equation $Au = f$. The eigenvectors form an orthonormal basis in $\mathbb{R}^n$ and as a result the solution can be expressed as $u = \sum_{p=1}^{n} \alpha_p u_{(p)}$, where coefficients α_p are initially unknown.

We have $Au = \sum_{p=1}^{n} \alpha_p \lambda_p u_{(p)}$ and $\left(Au, u_{(q)}\right) = \sum_{p=1}^{n} \alpha_p \lambda_p \left(u_{(p),} u_{(q)}\right) = \alpha_q \lambda_q$, and finally:

$$\alpha_q = \frac{\left(f, u_{(q)}\right)}{\lambda_q} \quad \text{and the solution:} \quad u = \sum_{p=1}^{n} \frac{\left(f, u_{(p)}\right)}{\lambda_p} u_{(p)} \tag{52}$$

This is a very simple algorithm, nevertheless not very practical as eigenvectors and eigenvalues are much more difficult to obtain in contrast to the solution of the linear equations system by some standard method. However in rare cases, when eigenvectors and eigenvalues are indeed known *for free* (as is the case for discrete Poisson problem) this forms the basis of extremely efficient numerical procedure.

8. Eigenvectors (eigenfunctions) and eigenvalues of selected matrices (operators)

Eigenfunctions and eigenvalues of the 1D BVP operator

Suppose that we have finite dimensional operator $A_{n\times n}$ connected to the linear equation $A_{n\times n}\, u = f$ and the infinite dimensional operator L connected to 1D Boundary Value Problem (BVP):

$$Lu = \frac{d^2u}{dx^2} \qquad u \in V = \{u \in C^2\langle 0,1\rangle : u(0) = u(1) = 0\} \qquad (53)$$

Operator L is connected to the simple BVP below:

$$\begin{cases} \dfrac{d^2u}{dx^2} = f \\ u(x=0) = u(x=1) = 0 \end{cases} \qquad \text{with scalar product: } (u,v) \overset{\text{def}}{=} \int_0^1 u\cdot v\, dx \qquad (54)$$

We seek non-zero $u_{(p)}$ such that:

$$Lu_{(p)} = \lambda_p u_{(p)} \qquad u_{(p)} \in V \qquad (55)$$

We distinguish now two separate cases $\lambda = \mu^2$ positive, and $\lambda = -\sigma^2$ negative:

Positive $\lambda = \mu^2$ Negative $\lambda = -\sigma^2$

$$\begin{cases} \dfrac{d^2u}{dx^2} = \mu^2 u \\ u(0) = u(1) = 0 \end{cases} \qquad \begin{cases} \dfrac{d^2u}{dx^2} = -\sigma^2 u \\ u(0) = u(1) = 0 \end{cases} \qquad (56)$$

The general exact solution contains two constants C_1 and C_2, which have to be determined such that the boundary condition is fulfilled:

$$u(x) = C_1 e^{\mu x} + C_2 e^{-\mu x} \qquad u(x) = C_1 \cos\sigma x + C_2 \sin\sigma x$$

$$u(0) = C_1 + C_2 = 0 \Rightarrow C_1 = -C_2 \qquad u(0) = C_1 = 0 \Rightarrow C_1 = 0$$

$$u(1) = C_1[e^{\mu} - e^{-\mu}] = 0 \Rightarrow C_1 = C_2 = 0 \qquad u(1) = C_2 \sin\sigma = 0 \Rightarrow \sigma = k\pi \qquad (57)$$

as $e^{\mu} \neq e^{-\mu}$ for $\mu > 0$ $k = 1, 2, \ldots$

No solution for positive $\lambda = \mu^2$ The constant C_2 is arbitrary.

Therefore the eigenvalues λ_k and the eigenfunctions $u_{(k)}$ of the original operator, are:

$$\lambda_k = -\mathrm{k}^2\pi^2 \qquad (58)$$

$$u_{(k)} = \sin k\pi x\,, \quad k = 1, 2, \ldots \qquad (59)$$

We have obtained infinite sequence of eigenvalues and eigenvectors. This means that the operator has infinite dimension, which is typical for continuous operators and functional spaces.

All eigenvalues are real, which is connected to the fact that the operator L is “selfadjoint” (symmetric in the previous nomenclature).

$$(Lu, v) = (u, Lv\,), \qquad u, v \in V \qquad (60)$$

This is easy to show considering the definition of the scalar product (54) and taking advantage of the Green theorem:

$$(Lu,v) = \int_0^1 \frac{d^2u}{dx^2} \cdot v\ dx = \left[\frac{du}{dx} v|_0^1\right] - \int_0^1 \frac{du}{dx} \cdot \frac{dv}{dx} dx = \\ = \left[\frac{dv}{dx} u|_0^1\right] - \int_0^1 \frac{d^2v}{dx^2} \cdot u\, dx = (u, Lv) \tag{61}$$

Both terms in square brackets vanish as both functions $u, v \in V$ (53) vanish at the endpoints of the interval $\langle 0, 1\rangle$.

Eigenvalues and eigenvectors of the discrete 1D BVP operator

$$\begin{cases} \dfrac{d^2u}{dx^2} = f \\ u(x=0) = u(x=1) = 0 \end{cases} \Longrightarrow \qquad L_h = \begin{cases} u_0 = 0 \\ \dfrac{u_{j-1} - 2u_j + u_{j+1}}{h^2} = f_j, \\ u_{n+1} \end{cases} \tag{1}$$

$$x_j = 0 + jh, \qquad j = 1, \dots, n, \qquad h = \frac{1}{n+1} \tag{2}$$

The corresponding matrix $A_{n\times n}$ limited to $j = 1, \dots, n$ can be expressed as:

$$A_h = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}, \quad u_h = \begin{bmatrix} u_1 \\ \vdots \\ u_j \\ \vdots \\ u_n \end{bmatrix}, \quad f_h = \begin{bmatrix} f_1 \\ \vdots \\ f_j \\ \vdots \\ f_n \end{bmatrix} \tag{62}$$

The equation system is then expressed as:

$$A_h u_h = f\_h \tag{63}$$

We seek now the eigenvalues and eigenvectors of A_h. Through analogy with the continuous case we choose the eigenvectors in the form of complex exponents consisting of both cosine and sine functions $e^{i\pi kx} = \cos k\pi x + i \sin k\pi x$, $i = \sqrt{-1}$. Below we drop the lower index h to shorten the notation. The k -th eigenvector $u_{(k)}$ can be then expressed as:

$$u_{(k)} = \begin{bmatrix} e^{ik\pi x_1} \\ \vdots \\ e^{ik\pi x_j} \\ \vdots \\ e^{ik\pi x_n} \end{bmatrix} = \begin{bmatrix} e^{ik\pi h} \\ \vdots \\ e^{ik\pi jh} \\ \vdots \\ e^{ik\pi nh} \end{bmatrix} \tag{64}$$

This was our guess, and we have to prove now, that this are indeed eigenvectors of A

$$Au_{(k)} = \frac{u_{(k)j-1} - 2u_{(k)j} + u_{(k)j+1}}{h^2} = \frac{e^{ik\pi(j-1)h} - 2e^{ik\pi jh} + e^{ik\pi(j+1)h}}{h^2} = \tag{65}$$

$$= e^{ik\pi jh} \frac{e^{-ik\pi h} - 2 + e^{ik\pi h}}{h^2} = u_{(k)} \left[\frac{e^{-\frac{ik\pi h}{2}} - e^{\frac{ik\pi h}{2}}}{h}\right]^2 = u_{(k)} \left[-2i \sin\left(\frac{kh\pi}{2}\right)\right]^2 = \\ = u_{(k)}\lambda_k \tag{66}$$

We have obtained the proof, and the eigenvalues of A are listed below:

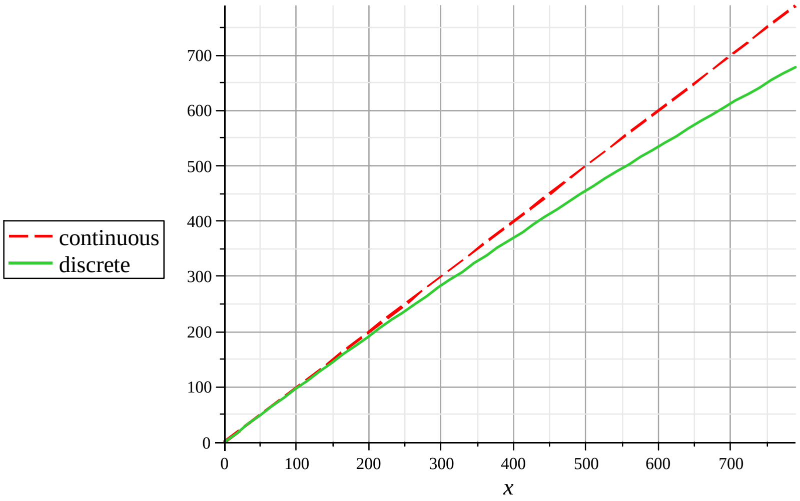
$$\lambda_k = -\frac{4\sin^2\left(\frac{kh\pi}{2}\right)}{h^2} \tag{67}$$

For small values of $k \ll n$ and large n

$$\lambda_k = -\frac{4\left(\frac{kh\pi}{2}\right)^2}{h^2} = -k^2\pi^2 \tag{68}$$

Which completely agrees with the first eigenvalues (58) of the continuous case. The ability to mimic the spectral properties of the continuous operator, by the discrete one, is an important property in numerical analysis.

The graph below shows both $-\lambda_{k(continuous)}$and $-\lambda_{k(discrete)}$ for different values of $x = k^2\pi^2$. The number of intervals in the discrete formulation is $n = 20$.



(69)

The value of $x = 400$ corresponds to roughly $k = 6$. Therefore 6 first eigenvalues are almost identical, which makes 30% of all eigenvalues of the discrete operator. From this we may infer that the finite dimensional operator will correctly resolve the longer waves, but will introduce significant error for much shorter waves with the wavelength close to the step-size h.

Eigenfunctions and eigenvalues of the 2D Poisson operator

$$\begin{cases} \mathrm{L}u \equiv \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} & (x,y) \in \Omega = \langle 0,1\rangle \times \langle 0,1\rangle \\ u|_{\partial\Omega} = 0 \end{cases} \tag{70}$$

$$\mathrm{L}u \equiv \mathrm{L}_x u + L_y u$$

Through analogy with the 1D problem the eigenfunctions of L are assumed as:

$$\mathrm{L}u_{(p,q)} = -\pi^2(p^2+q^2)u_{(p,q)} \tag{71}$$

and the eigenvalues are:

$$\lambda_{(p,q)} = -\pi^2(p^2+q^2) \tag{72}$$

Eigenfunctions and eigenvalues of the discrete 2D Poisson operator

$$\begin{cases} \mathrm{L}_h u\big|_{(x_i,y_j)} \equiv \dfrac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + \dfrac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h^2} \\ u_{i,j}\big|_{(x_i,y_j)\in\partial\Omega} = 0 \\ x_i = ih, \quad y_j = jh, \qquad h = \dfrac{1}{n+1}, \quad i,j = 1,\dots,n, \quad N = n^2 \end{cases} \tag{73}$$

$$(x_i, y_j) \in \Omega = \langle 0,1\rangle \times \langle 0,1\rangle$$

The corresponding matrix has the form:

$$\mathbb{A}_{N\times N} = \begin{bmatrix} \mathbb{T} & \mathbb{D} & & & \\ \mathbb{D} & \mathbb{T} & \mathbb{D} & & \\ & \ddots & \ddots & \ddots & \\ & & \mathbb{D} & \mathbb{T} & \mathbb{D} \\ & & & \mathbb{D} & \mathbb{T} \end{bmatrix} \tag{74}$$

where:

$$\mathbb{T}_{n\times n} = \frac{1}{h^2}\begin{bmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 1 \\ & & & 1 & -4 \end{bmatrix}, \quad \mathbb{D}_{n\times n} = \frac{1}{h^2}\begin{bmatrix} 1 & 0 & & & \\ 0 & 1 & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & 0 & 1 & 0 \\ & & & 0 & 1 \end{bmatrix} = \frac{1}{h^2}\mathbb{I}_{n\times n} \tag{75}$$

In analogy to 1D problem the eigenvectors are expressed as:

$$u_{(p,q)} = \begin{bmatrix} \sin(\pi hp)\sin(\pi hq) \\ \cdots \\ \sin(\pi hpi)\sin(\pi hqj) \\ \cdots \\ \sin(\pi hpn)\sin(\pi hqn) \end{bmatrix} \in \mathbb{R}^{N=n\cdot n} \tag{76}$$

The operator L_h can be expressed as $\mathrm{L}_h = \mathrm{L}_{hx} + \mathrm{L}_{hy}$, where both partial operators act independently on $u_{(p,q)}$:

$$\begin{gathered} L_{hx}u_{(p,q)} = -\frac{4\sin^2\left(\frac{ph\pi}{2}\right)}{h^2}u_{(p,q)} \\ L_{hy}u_{(p,q)} = -\frac{4\sin^2\left(\frac{qh\pi}{2}\right)}{h^2}u_{(p,q)} \end{gathered} \tag{77}$$

and therefore:

$$\lambda_{(p,q)} = -\frac{4}{h^2}\left[\sin^2\left(\frac{ph\pi}{2}\right) + \sin^2\left(\frac{qh\pi}{2}\right)\right] \tag{78}$$

9. Vector and matrix norms revisited

Vector norms

We recall now the properties of vector norms, which have the following properties ($u, v \in V,\ \ \alpha \in \mathbb{R}$):

$$\|\cdot\| : V \overset{\text{def}}{=} \mathbb{R}_+ \cup \{0\}$$

(iv) $\|u\| \geq 0, \quad \|u\| = 0 \Leftrightarrow \ u \equiv 0$

(v) $\|\alpha u\| = |\alpha|\|u\|$ (79)

(vi) $\|u + v\| \leq \|u\| + \|v\|$

Examples (Hölder norms):

$$u \in \mathbb{C}^n \text{ lub } \mathbb{R}^n, \qquad \|u\|_p = \sqrt[p]{\sum\nolimits_{j=1}^{n} |u_j|^p} \tag{80}$$

In particular:

$$\|u\|_1 = \sum\nolimits_{j=1}^{n} |u_j|, \ \ \|u\|_2 = \sqrt{\sum\nolimits_{j=1}^{n} |u_j|^2}, \ \ \|u\|_\infty = \max_j |u_j| \tag{81}$$

Matrix norms

Matrix (operator) norms have the following properties:

(i) $\|A\| \geq 0, \quad \|A\| = 0 \Leftrightarrow \ A \equiv 0$

(ii) $\|\alpha A\| = |\alpha|\|A\|$

(iii) $\|A + B\| \leq \|A\| + \|B\|$ (82)

(iv) $\|A \cdot B\| \leq \|A\| \cdot \|B\|$ (additional property)

Important class of matrix norms is generated (induced) by the vector norms.

Matrix norm $\|\cdot\|_M$ is induced by the vector norm $\|\cdot\|_V$, when:

$$\| A \|_M \overset{\text{def}}{=} \sup_{u \in V} \frac{\|Au\|_V}{\|u\|_V} = \sup_{\|u\|_V = 1} \|Au\|_V \tag{83}$$

Remark: The induced norm measures how much matrix A deforms a unit sphere $\|u\|_V = 1$.

Remark: The Euclidean matrix norm:

$$\|A\|_E = \sqrt{\sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{n} |a_{ij}|^2}$$

is not induced by any vector norm, and therefore it is of limited use (nevertheless it is consistent with the second vector norm, see below)

Definition: The matrix norm $\|\cdot\|_M$ is consistent with the vector norm $\|\cdot\|_V$ if:

$$\|Au\|_V \leq \| A \|_M \cdot \|u\|_V \tag{84}$$

Remark: Every induced norm is consistent as, from the definition of the induced norm:

$$\| A \|_M \geq \frac{\|Au\|_V}{\|u\|_V}$$

The following matrix norms are induced by the vector Hölder norms $\|u\|_{p:}$

Example:

$$\| A \|_\infty = \max_{i=1,..,n} \sum_{j=1}^{n} |a_{ij}| \tag{85}$$

Proof: see Annex 1

Example:

$$\| A \|_1 = \max_{j=1,..,n} \sum_{i=1}^{n} |a_{ij}| \tag{86}$$

Proof: see Annex 2

Example:

$$\| A \|_2 = \max_{\lambda \in \mathrm{spect}(A^H A)} \sqrt{\lambda} \tag{87}$$

and if $A = A^H$

$$\| A \|_2 = \max_{\lambda \in \mathrm{spect}(A)} \lambda = \lambda_{\max}(A) \tag{88}$$

Proof: see Annex 3

Remark: Second norm is always the smallest among all induced norms:

$$\| A \|_2 \leq \| A \|_p \tag{89}$$

Remark:

$$\| A \|_2 \leq \| A \|_E \leq \sqrt{n} \| A \|_2 \tag{90}$$

10. Iterative methods to solve large linear systems

The Jacobi iterative method

Suppose we have a large linear system, which is impractical or impossible to solve by the finite Gauss elimination method:

$$Au = b \qquad \text{with} \qquad A = [L + D + U] \tag{91}$$

where:

$$L = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ * & 0 & \dots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ * & * & * & 0 \end{bmatrix}, \qquad D = \begin{bmatrix} * & 0 & \cdots & 0 \\ 0 & * & \dots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & * \end{bmatrix}, \qquad U = \begin{bmatrix} 0 & * & \cdots & * \\ 0 & 0 & \cdots & * \\ \cdots & \cdots & \cdots & * \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{92}$$

The original equation can be thus rewritten as:

$$Du = b - [L + U]u \tag{93}$$

or alternatively as:

$$u = D^{-1}b - D^{-1}[L + U]u \tag{94}$$

The Jacobi iterative method generated by (94) starts with an arbitrary guess $u^{(0)}$ (e.g., zero) and generates the consecutive iterates according to the formula:

$$u^{(m+1)} = D^{-1}b - D^{-1}[L + U]u^{(m)}, \quad m = 0, 1, 2, \dots. \tag{95}$$

Or in the scalar form, for a fixed value of m:

$$u_j^{(m+1)} = \frac{1}{d_{jj}}\left[b_j - \sum_{i=1}^{j-1} a_{ij}u_j^{(m)} - \sum_{i=j+1}^{n} a_{ij}u_j^{(m)}\right], \quad j = 1, \dots, n \tag{96}$$

We will investigate the sufficient conditions for the convergence of this iterative procedure. Suppose now that u_* denotes the exact solution, we have:

$$u_* = D^{-1}b - D^{-1}[L + U]u_* \tag{97}$$

We can define now the error of each iteration as:

$$\varepsilon^{(m)} \overset{\text{def}}{=} u^{(m)} - u_* \quad \text{and} \quad e^{(m)} = \left\|\varepsilon^{(m)}\right\| \tag{98}$$

Subtracting the formulas (95) and (97) one obtains:

$$\varepsilon^{(m+1)} = -D^{-1}[L + U]\varepsilon^{(m)} \tag{99}$$

and a following estimation:

$$\begin{gathered} \left\|\varepsilon^{(m+1)}\right\|_V = \left\|-D^{-1}[L + U]\varepsilon^{(m)}\right\|_V \le \|D^{-1}[L + U]\|_M \left\|\varepsilon^{(m)}\right\|_V \\ e^{(m+1)} \le \|D^{-1}[L + U]\|_M e^{(m)} \end{gathered} \tag{100}$$

The iterations converge if for some matrix norm $\|\cdot\|_M$:

$$\|D^{-1}[L+U]\|_M < 1 \tag{101}$$

This is the necessary condition for convergence.

Remark: It is sufficient to prove the above for any consistent matrix norm.

Example:

Suppose that A Is strongly diagonally dominant:

$$|a_{ii}| > \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|$$

$$\Downarrow$$

$$\frac{1}{|a_{ii}|}\left[\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|\right] < 1 \tag{102}$$

$$B \overset{\text{def}}{\equiv} D^{-1}(L+U)$$

$$\|B\|_\infty = \max_i \frac{1}{|a_{ii}|}\left[\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|\right] < 1$$

Iterative method of Jacobi is therefore always convergent for the strongly diagonally dominant matrices

Remark:

If A is weakly diagonally dominant then $\|B\|_\infty \le 1$ and the necessary condition is not fulfilled. Possibly $\|B\|_2 < 1$ (as the second norm is smallest of all), but this is difficult or impossible to prove in a general case

Example:

Let's consider now a special case of weakly diagonally dominant matrix A, namely the matrix corresponding to the discrete Poisson operator:

$$\begin{aligned} \mathrm{L}_h u|_{(x_i,y_j)} &\equiv \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h^2} = \\ &= \frac{1}{h^2}\left[u_{i,j-1} + u_{i-1,j} - 4u_{i,j} + u_{i+1,j} + u_{i,j+1}\right] \end{aligned} \tag{103}$$

With weakly diagonally dominant matrix A

$$A = \frac{1}{h^2}[\ldots 1 \ldots 1 - 4\ 1 \ldots 1 \ldots] \tag{104}$$

$$B = D^{-1}[L+U] = \left[\ldots -\tfrac{1}{4} \ldots -\tfrac{1}{4}\ \ 0\ \ -\tfrac{1}{4} \ldots -\tfrac{1}{4} \ldots\right] = -\tfrac{1}{4}[Ah^2 + 4I] \tag{105}$$

Note that eigenvectors of A and B coincide, therefore the eigenvalues can be easily calculated”

On the other hand $\|B\|_\infty = 1$ which shows that sharper estimation of the matrix norm is essential to demonstrate convergence of the Jacobi iterative procedure.

$$\lambda_{pq}(B) = -\frac{1}{4}\left(h^2\lambda_{pq}(A) + 4\right) = -\left[\frac{h^2\lambda_{pq}(A)}{4} + 1\right] =$$

$$= -\left[1 - \sin^2\left(\frac{ph\pi}{2}\right) - \sin^2\left(\frac{qh\pi}{2}\right)\right] \qquad (106)$$

$$p, q = 1,2, \dots, n$$

$$\max_{p,q}\left|\lambda_{p,q}\right| = \left|\lambda_{1,1}\right| < 1$$

$$\Downarrow \qquad (107)$$

$$\|B\|_2 < 1$$

One can estimate that (for large values of n):

$$\|B\|_2 = \left|\lambda_{1,1}\right| \approx 1 - \frac{\pi^2h^2}{2} \qquad (108)$$

Therefore the Jacobi iterative method will remain convergent for the particular matrix (73) **Błąd! Nie można odnaleźć źródła odwołania.** corresponding to the discrete Poison problem.

Remark:

One can estimate number of iterations m necessary to lower the solution error by factor of $e \approx 2.71$

$$\|B\|_2^m = \frac{1}{e}$$

$$m\ \ln\|B\|_2 = -1 \quad \text{and with} \quad \ln(1-\alpha) \approx -\alpha$$

one obtains:

$$m\frac{\pi^2h^2}{2} \approx 1 \qquad (109)$$

$$m \approx \frac{2}{\pi^2h^2} = \frac{2}{\pi^2}(n+1)^2$$

The number of iterations grows as the square of the number of segments (steps) in one direction.

The total cost of finding the solution is (as matrix-vector multiplication requires n^2operations for dense matrices) therefore proportional to:

$$C(n+1)^2n^2 \sim Cn^4$$

This is a very expensive method and not at all practical. Krylov methods or multigrid algorithm are much faster. Nevertheless the method of Jacobi allows for very simple analysis and is the basis of multigrid approach. The method of Jacobi is easily transferable also to the nonlinear problems which are of our main interest.

The Gauss-Seidel iterative method

The Jacobi iterative method can be slightly modified (and simplified) by observing that the inner iterations are performed in a ordered sequence and some vector elements are available already from the current iteration, this modifies the original Jacobi algorithm (96) to get scalar version:

$$u_j^{(m+1)} = \frac{1}{d_{jj}}\left[b_j - \sum_{i=1}^{j-1} a_{ij}u_j^{(m+1)} - \sum_{i=j+1}^{n} a_{ij}u_j^{(m)}\right], \quad j = 1, \dots, n \tag{110}$$

Or in the matrix-vector representation:

$$u^{(m+1)} = [D + L]^{-1}\left[b - Uu^{(m)}\right], \quad m = 0, 1, 2, \dots. \tag{111}$$

The analysis of Gauss-Seidel algorithm is more involved than the analysis of Jacobi algorithm, therefore we will consider only the latter one.

Error analysis of consecutive iterations of the Jacobi iterative algorithm

Suppose we have a matrix corresponding to the 1D Poisson like problem:

$$A_{n\times n} = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}, \tag{112}$$

Convergence of the Jacobi iterative procedure depends on the properties of:

$$B = D^{-1}[L + U] = \begin{bmatrix} 0 & -\frac{1}{2} & & & \\ -\frac{1}{2} & 0 & -\frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{2} & 0 & -\frac{1}{2} \\ & & & -\frac{1}{2} & 0 \end{bmatrix}, \quad D = -\frac{2}{h^2}\mathbb{I} \tag{113}$$

Note that eigenevectors of A, D and B are the same, and:

$$B = D^{-1}A - \mathbb{I} \tag{114}$$

Since the eigenvalues and eigenvectors of A are known:

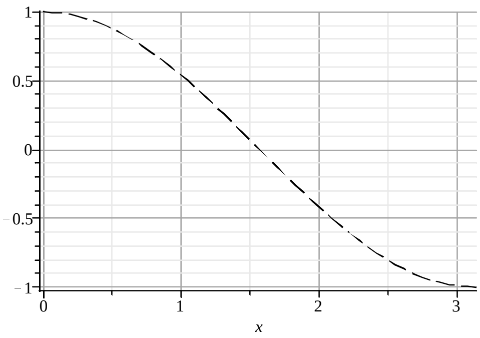
$$\lambda_p(A) = -\frac{4}{h^2}\sin^2\left(\frac{ph\pi}{2}\right), \quad p = 1,2, \dots, n \tag{115}$$

$$u_{(p)} = \begin{bmatrix} \sin(\pi hp1) \\ \cdots \\ \sin(\pi hpi) \\ \cdots \\ \sin(\pi hpn) \end{bmatrix} \in \mathbb{R}^N \tag{116}$$

The eigenvalues of B can be obtained by simple subtraction:

$$\lambda_p(B) = -\frac{h^2}{2}\lambda_p(A) - 1 \ = 2\sin^2\left(\frac{ph\pi}{2}\right) - 1 = -\cos(k\pi h) \tag{117}$$

The eigenvalues can be visualised graphically:



(118)

$x = k\pi h$

to demonstrate that low and high frequencies in the error (low and high components of the eigenvector- k small and large) are very weakly damped ($|\lambda| \approx 1$), while middle frequencies are strongly damped ($|\lambda| \ll 1$),.

Error analysis of Jacobi iterations with underrelaxation

We will repeat the previous analysis for the Jacobi method with underrelaxation (for the same matrix).

The simple Jacobi method (94)(95) can be written as:

$$u^{(m+1)} = D^{-1}b - D^{-1}[L+U]u^{(m)}$$

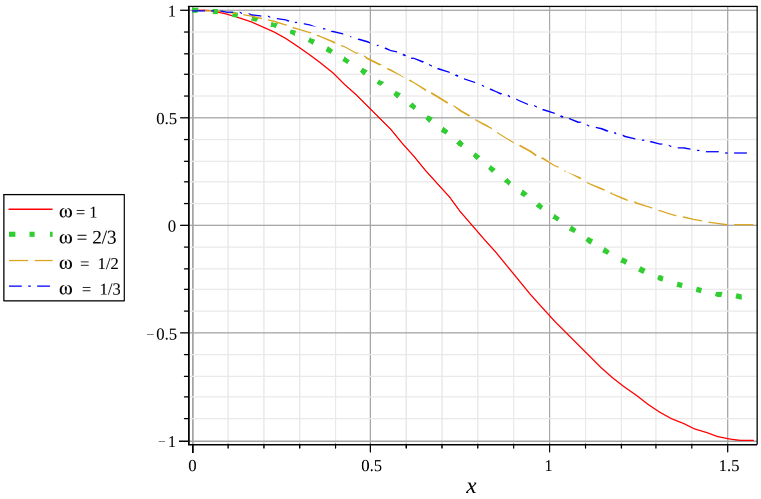
And with the aid of underrelaxation as:

$$\begin{aligned} u^{(m+1)} &= (1-\omega)u^{(m)} + \omega\left[D^{-1}b - D^{-1}[L+U]\right]\ u^{(m)} = \\ &= \omega D^{-1}b + \left[\mathbb{I} - \omega\mathbb{I} + \omega D^{-1}[L+U]\right]u^{(m)} = \\ &= \omega D^{-1}b + [\mathbb{I} - \omega D^{-1}A]u^{(m)} \end{aligned} \tag{119}$$

The convergence of the iterations depends on the eigenvalues of

$$\begin{aligned} B_\omega &= \mathbb{I} - \omega D^{-1}A \\ \lambda_p(B_\omega) &= 1 - \omega\frac{h^2}{2}\lambda_p(A) = 1 - 2\omega\sin^2\left(\frac{ph\pi}{2}\right) \\ \lambda_1(B_\omega) &= 1 - \omega\frac{h^2}{2}\lambda_1(A) = 1 - 2\omega\sin^2\left(\frac{h\pi}{2}\right) \approx 1 - \omega\frac{h^2\pi^2}{2} \end{aligned} \tag{120}$$

The eigenvalues are visualised for different values of ω in the Figure below with $x = ph\pi/2$, the smaller values of the modulus of the eigenvalue corresponding to the faster damping of the contribution of the corresponding eigenvector.



(121)

$x = \frac{ph\pi}{2}$

$x = \frac{ph\pi}{2}$	0	$\frac{\pi}{4}$	$\frac{\pi}{2}$	Comments
$\omega = 0$	1	1	1	Never convergent
$\omega = \frac{1}{3}$	1	$\frac{2}{3}$	$\frac{1}{3}$	Broadest spectrum of the damped frequencies
$\omega = \frac{1}{2}$	1	$\frac{1}{2}$	0	Very fast damping of high frequencies
$\omega = \frac{2}{3}$	1	$\frac{1}{3}$	$\frac{2}{3}$	Broad spectrum of the damped frequencies
$\omega = 1$	1	0	-1	Very good damping of middle frequencies

(122)

It is clear, that for $\omega = 1/2$ the high frequencies of the error are strongly damped while low frequencies are kept almost unchanged. Still slightly better properties can be observed for $\omega = 1/3$ for which the range of strongly damped frequencies is broader.

The Jacobi iterative procedure with underrelaxation is even slower in convergence than for $\omega = 0$ but the essential property is the smoothing (damping) of high frequencies which forms the basis of the multigrid method. Similar properties has the Gauss-Seidel iterative method.

11. Multigrid method

Suppose we have a sequence of meshes covering Ω with stepsizes $h, 2h, 4h, \dots$,etc. and a sequence of linear systems:

$$\begin{aligned} &A_h u_h = f_h \quad \text{(expensive to solve)} \\ &A_{2h} u_{2h} = f_{2h} \\ &A_{4h} u_{4h} = f_{4h} \quad \text{(much cheaper to solve)} \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \end{aligned} \tag{123}$$

How to use this sequence to find the solution of the first, expensive to solve equation, much faster ?

The naïve approach

The naïve approach consists in iterating first on the sparse mesh and then to interpolate the solution on the denser mesh and iterate in the finer mesh having interpolated solution as an initial guess.

The algorithm may look like below:

```
$H = h \cdot 2^L$,  L denotes number of Levels
$u_H^{(0)} \equiv 0$
for $l = L - 1$ step $-1$ until 0 do
        • $H = h \cdot 2^l$
        • Perform few Jacobi/Gauss-Seidel iterations to smooth out high frequency error in
          $A_H u_H = f_H$ starting from the available $u_H^{(0)}$
        • Interpolate $u_H$ on the finer mesh $u_{H/2}$to obtain the first guess for $u_{H/2}^{(0)}$
end do
```

This algorithm will work but will not be (if at all) faster, than the Jacobi/Gauss-Seidel original algorithm.

Full Multigrid Algorithm

Suppose now that, the single level of the Multigrid Algorithm MV_h performs the following recursive action:

$$v_h \leftarrow MV_h(v_h, f_h)$$

1. Perform few Jacobi/Gauss-Seidel iterations to smooth out high frequency error in $A_h u_h = f_h$ starting from the arbitrary v_h
2. If $\Omega_h$is the coarsest gird, go to 4, else do residual correction: (124)

$$f_{2h} \leftarrow I_h^{2h}(f_h - A_h v_h)$$

$$v_{2h} \leftarrow 0$$

$$v_{2h} \leftarrow MV_{2h}(v_{2h}, f_{2h})$$

3. Correct $v_h \leftarrow v_h + I_{2h}^{h} v_{2h}$
4. Perform (optionally) few Jacobi/Gauss-Seidel iterations to smooth out high frequency error in $A_h u_h = f_h$ starting from the current v_h

In the above I_h^{2h} denotes the restriction operator, which transfers the grid function v_h from Ω_h onto Ω_{2h}. The 1D example of such operator is given by:

$$\begin{gathered} v_{2h} \leftarrow I_h^{2h} v_h \\ v_{2h,i} = \frac{1}{4}\left(v_{h,2i-1} + 2v_{h,2i} + v_{h,2i+1}\right) \end{gathered} \tag{125}$$

As high frequency component of v_h is already smoothed out, this restriction will transfer almost all information to the coarse grid.

On the other hand I_{2h}^{h} stands for the prolongation operator, which transfers the grid function v_{2h} from Ω_{2h} onto Ω_h. The 1D example of such operator is given by:

$$\begin{gathered} v_h \leftarrow I_{2h}^{h} v_{2h} \\ v_{h,2i} = v_{2h,i} \\ v_{h,2i+1} = \frac{1}{2}\left(v_{2h,i} + v_{2h,i+2}\right) \end{gathered} \tag{126}$$

The restriction and prolongation operators (matrices) re as a rule related through the following requirement:

$$I_{2h}^{h} = c \cdot \left(I_h^{2h}\right)^T \tag{127}$$

where $c \in \mathbb{R}$ denotes a constant number.

In the example above, these transfer matrices have the following simple form:

$$I_{2h}^{h} = \frac{1}{2}\begin{bmatrix} 1 & & \\ 2 & & \\ 1 & 1 & \\ & 2 & \\ & 1 & 1 \\ & & 2 \\ & & 1 \end{bmatrix} \qquad I_h^{2h} = \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 & & & & \\ & & 1 & 2 & 1 & & \\ & & & & 1 & 2 & 1 \end{bmatrix} \tag{128}$$

The multigrid algorithm (124) offers significant acceleration of convergence, in comparison with the classical iterative schemes.

12. Matrix functions

Suppose we have an arbitrary scalar function $f(x)$, be it: x^2, $\sin x$, e^x, $\sqrt{x}$, or $|x|$. As we are interested in multidimensional problems we would like to define the meaning of such functions acting on matrices (and perhaps also on other operators).

It is obvious that the naïve element-wise definition:

$$f(A) \overset{\text{def}}{=} \begin{bmatrix} f(a_{11}) & \dots & f(a_{1n}) \\ \dots & \dots & \dots \\ f(a_{n1}) & \dots & f(a_{nn}) \end{bmatrix} \tag{129}$$

cannot be considered, as with this definition $A^2 \neq A \cdot A$.

We will therefore try to propose a more reasonable approach.

1. Suppose now that we start with polynomial function $f(x) = c_m x^m + \cdots + c_1 x^1 + c_0$. In this case the definition of $f(A)$ is quite straightforward:

$$f(A) \overset{\text{def}}{=} c_m A^m + \cdots + c_1 A^1 + c_0 I \tag{130}$$

2. Equally straightforward is the case when the scalar function is defined by the infinite power/Taylor series:

$$f(x) = \sum_{j=0}^{\infty} c_j x^j \tag{131}$$

As for example is the case for $e^x = \sum_{j=0}^{\infty} \frac{x^j}{j!}$. Then we can extend this formula to matrices:

$$f(A) \overset{\text{def}}{=} \sum_{j=0}^{\infty} c_j A^j \tag{132}$$

This series will remain convergent provided, that the corresponding scalar series is convergent:

$$f(\|A\|_M) = \sum_{j=0}^{\infty} c_j \|A\|_M^j \tag{133}$$

where $||A||_M$ denotes an arbitrary matrix norm.

In particular we can define

$$\sin(A) \overset{\text{def}}{=} A - \frac{A^3}{3!} + \frac{A^5}{5!} - \frac{A^7}{7!} + \cdots \tag{134}$$

and this definition will be consistent with the expected properties of the $\sin(A)$ function, in particular we will preserve the usual trigonometric identities, e.g., $\sin(2A) = \sin(A)\cos(A)$.

3. Suppose now that $f(x)$ is quite arbitrary, and the matrix A is diagonalisable, i.e., $A = Q^{-1}\Lambda Q$, where $\Lambda = \text{diag}(\lambda_j)$. Then we can define the matrix function as a scalar function acting on each eigenvalue separately:

$$f(A) \overset{\text{def}}{=} Q^{-1}f(\Lambda)Q = Q^{-1}\begin{bmatrix} f(\lambda_1) & 0 & \dots & 0 \\ 0 & f(\lambda_2) & \dots & 0 \\ \dots & \dots & \ddots & \dots \\ 0 & 0 & \dots & f(\lambda_n) \end{bmatrix} Q = Q^{-1}\mathrm{diag}\left[f(\lambda_p)\right]Q \qquad (135)$$

We will show that this definition is fully consistent with the previous one (via power series). For this purpose we observe that:

$$A^j = (Q^{-1}\Lambda Q)\cdot(Q^{-1}\Lambda Q)\cdot\ldots\cdot(Q^{-1}\Lambda Q) = Q^{-1}\Lambda^{\mathrm{j}}Q \qquad (136)$$

Therefore, one obtains the identity:

$$\begin{gathered} f(A) \overset{\text{def}}{=} \sum_{j=0}^{\infty} c_j A^j = \sum_{j=0}^{\infty} c_j\, Q^{-1}\Lambda^{\mathrm{j}}Q = Q^{-1}\left[\sum_{j=0}^{\infty} c_j\,\Lambda^{\mathrm{j}}\right]Q = Q^{-1}\mathrm{diag}\left[\sum_{j=0}^{\infty} c_j\,\lambda_p^j\right]Q = \\ = Q^{-1}\mathrm{diag}\left[f(\lambda_p)\right]Q \end{gathered} \qquad (137)$$

proving that both formulations are fully equivalent.

This approach allows to define nonanalytic functions like $\sqrt{A}$, or $|A|$ and expect them to behave as scalar counterparts do. However not all relations known from the scalar algebra are transferred to the matrix functions (mainly because matrices do not commute $AB \neq BA$). As a consequence of which, the well-known relation does not hold:

$$e^{A+B} \neq e^A \cdot e^B \qquad (138)$$

unless A and B have the same eigenvectors. However the diagonalisable matrices having the same eigenvectors do commute as $AB = Q^{-1}\Lambda_{\mathrm{A}}QQ^{-1}\Lambda_{\mathrm{B}}Q = Q^{-1}[\Lambda_{\mathrm{A}}\Lambda_{\mathrm{B}}]Q = Q^{-1}[\Lambda_{\mathrm{B}}\Lambda_{\mathrm{A}}]Q = BA$. Similarly do commute all matrices $f(A)$ and $g(A)$.

Matrix linear ODE

Knowing this we can attempt to solve the multidimensional linear Ordinary Differential Equation (ODE) in which matrix $A = Q^{-1}\Lambda Q$ is diagonalisable:

$$\begin{cases} \dfrac{du}{dt} = Au \\ u(t=0) = u_0 \end{cases}, \qquad A \in \mathbb{R}^{n\times n}, u \in \mathbb{R}^n \qquad (139)$$

By analogy to the scalar case we will assume that $u(t) = u_0 e^{At}$. To show that this is indeed a correct solution, it is enough to evaluate the derivative $\frac{d}{dt}e^{At}$, directly from the definition:

$$\begin{gathered} \frac{d}{dt}e^{At} \overset{\text{def}}{=} \lim_{\epsilon\to 0}\frac{e^{A(t+\epsilon)} - e^{At}}{\epsilon} = e^{At}\cdot\lim_{\epsilon\to 0}\frac{e^{A\epsilon} - I}{\epsilon} = \\ = e^{At}\left[\cdot\lim_{\epsilon\to 0}\frac{I + \frac{A\epsilon}{1!} + \frac{A^2\epsilon^2}{2!} + \cdots - I}{\epsilon}\right] = e^{At}\mathrm{A}\cdot\lim_{\epsilon\to 0}\left[I + \frac{A\epsilon}{2!} + \frac{A^2\epsilon^2}{3!}\ldots\right] = Ae^{At} \end{gathered} \qquad (140)$$

This confirms that the proposed solution indeed fulfils the equation (139).

13. Nonlinear equations

We will investigate now the possibility to solve the general nonlinear equation

$$\mathbb{F}(u) = 0 \tag{141}$$

in which u might be a scalar, a vector or a function (or even a vector function), while $\mathbb{F}$ is either a scalar function or a vector function or a functional operator respectively.

Problem A

To illustrate the first possibility we consider the simple problem to solve

$$e^{-x} = x \quad \text{or} \quad e^{-x} - x = 0 \tag{142}$$

where x is a scalar real variable and $\mathbb{F}(x) \equiv e^{-x} - x$.

Problem B

The second option is usually illustrated by a list of algebraic nonlinear equations, here however we propose the system of n equations which can be written using matrix-vector operators

$$\left(A + I \cdot e^{-x^T x}\right)x = b, \qquad x \in \mathbb{R}^n, b \in \mathbb{R}^n, A \in \mathbb{R}^{n\times n} \tag{143}$$

where I denotes the identity matrix, while $x^T x \equiv \|x\|^2$ is always a positive real number. Here $\mathbb{F}(x) \equiv \left(A + I \cdot e^{-x^T x}\right)x - b \equiv Ax + xe^{-x^T x} - b$. The matrix A and the RHS (Right Hand Side) vector b are given.

Problem C

The third option can be illustrated by any nonlinear ODE (Ordinary Differential Equation) or PDE (Partial Differential Equation), e.g., Euler or Navier-Stokes equations. For simplicity we propose the 1D nonlinear convection-diffusion BVP

$$\begin{cases} \dfrac{d^2u}{dx^2} + u\dfrac{du}{dx} = f(x) \\ \quad u(a) = u_a \\ \quad u(b) = u_b \end{cases} \tag{144}$$

where $u(x)$ is an unknown, sufficiently smooth function, $f(x)$ denotes the known function of the RHS, while $a < b, u_a, u_b$ represent the known real numbers. The nonlinearity $u\frac{du}{dx}$ present in this equation is similar to the convective nonlinearity of the Euler or Navier-Stokes equations. The function $\mathbb{F}(u)$ in this case is

$$\mathbb{F}(u) \equiv \begin{bmatrix} \dfrac{d^2u}{dx^2} + u\dfrac{du}{dx} - f(x) \\ u(a) - u_a \\ u(b) - u_b \end{bmatrix} \tag{145}$$

In the above, the first equation is nonlinear, while the 2nd and 3rd is linear. The RHS zero consist of one zero-function and two real zeros.

As all these problems are nonlinear we cannot be sure that the solution always exists and that in such cases is always unique. Nevertheless we will present few methods that allow to solve such systems in a systematic manner (if the solution exists and if the corresponding iterative procedure converges, which usually is not known *a priori*). However in some rare cases it is possible to decide about the existence and the uniqueness of the solution.

There exist many iterative algorithms to solve nonlinear problems, which are applicable under different conditions. The basic algorithms are:

- Fixed point iterations
- Bisection, which is generally applicable to scalar equations (omitted here)
- Secant method, which can be regarded as quasi-Newton method in which the derivative is replaced by the finite difference (again omitted)
- Method of frozen coefficients (applicable for some ODE's and PDE's)
- Newton method, in which derivative needs to be calculated
- Embedding in a pseudo-time problem (applicable for some ODE's and PDE's)

The Method of Frozen Coefficients

This method is purely heuristic and cannot be called "systematic". It bases on observation that higher derivatives are somehow "more important" in the equation than the function values itself. Therefore if the nonlinearity consists of the two, we may take the function values from the previous iteration, while the derivative from the present one. This is illustrated (for the Problem C) in the flowchart below:

1. At zero iteration take: $k = 0$, and $u_{(0)}(x) \equiv 0$,
2. Solve the linear BVP
$$\begin{cases} \dfrac{d^2u_{(k+1)}}{dx^2} + u_{(k)}\dfrac{du_k}{dx} = f(x) \\ u_{(k+1)}(a) = u_a \\ u_{(k+1)}(b) = u_b \end{cases}$$
3. Calculate the difference between $u_{(k)}(x)$ and $u_{(k+1)}(x)$
$$\delta \coloneqq \max_{a \le x \le b} \left| u_{(k)}(x) - u_{(k+1)}(x) \right|$$
4. If $\delta < \epsilon$ then STOP (the equation is solved with ϵ accuracy)
5. Substitute $k \coloneqq k + 1$ and return to step 2.

From numerical experience we know, that this iterative procedure will converge if the solution is "small" and will diverge if the solution is "large", as in this latter case the nonlinear term becomes more important than the second derivative. In Fluid Mechanics this technique will be successful for low Re and will fail for higher Re (say for $Re > 1000$).

The method of frozen coefficients is very simple and follows the engineering intuition which allows to solve the linear problem first and include nonlinearities at the later stage (e.g., when we solve the fluid flow or heat transfer problems we neglect in the first approximation the dependence of viscosity and thermal conductivity on temperature). Although simple, the procedure is quite arbitrary with respect to the choice of the element of the equation, that should be taken from the previous iterations. As a result the method will fail for more complex nonlinearities, such as

$$\frac{d^2u}{dx^2} + \left(\frac{du}{dx}\right)^3 = f(x) \tag{146}$$

In this case we have no indication how to proceed with the 3rd power of the first derivative.

Newton Method (Quasi-Linearisation)

The Newton method is best presented (in scalar case, i.e., for Problem A) by considering the Taylor expansion of $f(x_*)$ around certain point x

$$f(x_*) = f(x) + f'(x)(x_* - x) + \cdots \tag{147}$$

By rejection of higher-order terms and by assuming that x_* denotes the sought solution (i.e., $f(x_*) \equiv 0$), we obtain the linear equation for x_*

$$f'(x)(x_* - x) = -f(x) \tag{148}$$

Since this only a rough estimation, we replace x by the current approximation of the solution x_k, while x_* is replaced by the next approximation x_{k+1}

$$f'(x_k)(x_{k+1} - x_k) = -f(x_k) \tag{149}$$

And finally the Newton's formula is obtained:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \tag{150}$$

If we apply it to the Problem A $(f(x) \equiv \exp(-x) - x)$ we get

$$x_{k+1} = x_k + \frac{e^{-x_k} - x_k}{e^{-x_k} + 1} \tag{151}$$

The consecutive iterations (starting with $x_0 = 1$) will deliver:

k	x_k	Error = $f(x_k)$
0	1	-0.63
1	0.53	0.046
2	0.5669	0.00024494
3	0.567143285	6.92x10^{-9}
4	0.567143290409783869	5.54x10^{-18}
5	0.567143290409783872999968662210355549	3.54x10^{-36}

(note that 35 digits are accurate in the last iteration; this was possible because these computations were carried out using 120 digits of accuracy)

The Newton's method is very fast, if started from the close neighbourhood of the root. It usually fails if the first approximation is far from the root. It may also fail if the first derivative vanishes around the root.

The same approach can be used in the vector case again by considering the Taylor expansion

$$f(x_*) = f(x) + f'(x)(x_* - x) + \cdots$$

$$x \equiv \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, x_* \equiv \begin{bmatrix} x_{*1} \\ \vdots \\ x_{*n} \end{bmatrix}, f(x) \equiv \begin{bmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_n(x_1, \dots, x_n) \end{bmatrix}, \qquad f'(x) \equiv \frac{\partial f}{\partial x} \equiv \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \cdots & \ddots & \cdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix} \tag{152}$$

Replacing x_* by $x_{(k+1)}$ and x by $x_{(k)}$ we obtain the linear equation system for the current correction $\Delta \equiv x_{(k+1)} - x_{(k)}$

$$f'\left(x_{(k)}\right) \Delta = -f\left(x_{(k)}\right) \tag{153}$$

After this equation is solved, we obtain the next approximation by the formula $x_{(k+1)} = x_{(k)} + \Delta$.

In the case of the functional equation (Problem C) we first must understand what is the derivative of a functional operator $\mathbb{F}(u)$ with respect to u (in this case it will be an analogue of the directional derivative). Such analogue is called *Gatteaux derivative* and is defined as

$$\langle D\mathbb{F}(u), v\rangle \overset{\text{def}}{=} \lim_{\epsilon\to 0}\frac{\mathbb{F}(u+\epsilon v)-\mathbb{F}(u)}{\epsilon} \tag{154}$$

where the left hand side should be read as derivative of $\mathbb{F}$ at u in the direction of v. It should be additionally noted that the above formula is linear with respect v.

The *Gatteaux derivative* differs from the usual derivative (including the directional derivative) by incorporating the directional vector into the formula. This is illustrated in the table below

Case	Gatteaux derivative $\langle D\mathbb{F}(u), v\rangle$	Linear function	Gatteaux derivative $\langle D\mathbb{F}(u), v\rangle$
$\mathbb{F}: \mathbb{R} \to \mathbb{R}$ $x, u, v \in \mathbb{R}$	$\left.\frac{d\mathbb{F}}{dx}\right\vert_{x=u} \cdot v$	$\mathbb{F}(x) = c \cdot x$	$c \cdot v$
$\mathbb{F}: \mathbb{R}^n \to \mathbb{R}^n$ $x, u, v \in \mathbb{R}^n$	$\left[\left.\frac{\partial\mathbb{F}}{\partial x}\right\vert_{x=u}\right] \cdot v$	$\mathbb{F}(x) = A \cdot x$	$A \cdot v$

where c and A denote constant scalar and matrix respectively ($\partial\mathbb{F}/\partial x$ stands for the Jacobian matrix).

The corresponding Taylor formula for a functional case has a form

$$\mathbb{F}(u_*) = \mathbb{F}(u) + \langle D\mathbb{F}(u), u_* - u\rangle + \cdots \tag{155}$$

out of which, after rejection of higher order terms and by assuming that u_* is a solution, the linear problem is obtained

$$\langle D\mathbb{F}(u), u_* - u\rangle = -\mathbb{F}(u) \tag{156}$$

Replacing u_* by $u_{(k+1)}$ and u by $u_{(k)}$ we obtain the linear problem for the current correction $\Delta \equiv u_{(k+1)} - u_{(k)}$

$$\langle D\mathbb{F}\big(u_{(k)}\big), \Delta\rangle = -\mathbb{F}\big(u_{(k)}\big) \tag{157}$$

By solving this problem we obtain Δ and subsequently $u_{(k+1)} \coloneqq u_{(k)} + \Delta$

The algorithm will be presented for Problem 3, for which

$$\mathbb{F}(u) \equiv \begin{bmatrix} u'' + uu' - f \\ u(a) - u_a \\ u(b) - u_b \end{bmatrix} \tag{158}$$

where for simplicity derivatives were replaced by apostrophes.

The Gatteaux derivative is calculated from its definition

$$
\begin{aligned}
\langle D\mathbb{F}(u), v\rangle &\overset{\text{def}}{=} \lim_{\epsilon\to 0}\frac{1}{\epsilon}\begin{bmatrix} (u+\epsilon v)'' + (u+\epsilon v)(u+\epsilon v)' - f - (u''+uu'-f) \\ (u+\epsilon v)(a) - u_a - (u(a)-u_a) \\ (u+\epsilon v)(b) - u_b - (u(b)-u_b) \end{bmatrix} = \\
&= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\begin{bmatrix} \epsilon[v''+uv'+vu'] + \epsilon^2 vv' \\ \epsilon\, v(a) \\ \epsilon\, v(b) \end{bmatrix} = \begin{bmatrix} v''+uv'+vu' \\ v(a) \\ v(b) \end{bmatrix}
\end{aligned}
\qquad (159)
$$

The final algorithm to solve the nonlinear Problem C by the Newton's (Quasi-linearization) method is therefore

1. At zero iteration take: $k = 0$, and $u^{(0)}(x) \equiv 0$,
2. Solve the linear BVP
$$\begin{cases} \dfrac{d^2\Delta}{dx^2} + u_{(k)}\dfrac{d\Delta}{dx} + \dfrac{du_{(k)}}{dx}\Delta = -\left(\dfrac{d^2u_{(k)}}{dx^2} + u_{(k)}\dfrac{du_{(k)}}{dx} - f\right) \\ \Delta(a) = -(u_{(k)}(a) - u_a) \\ \Delta(b) = -(u_{(k)}(b) - u_b) \end{cases}$$
3. Calculate the norm of the increment Δ
$$\delta \coloneqq \max_{a\le x\le b} |\Delta(x)|$$
4. Calculate the next iteration $u_{(k+1)}(x) = u_{(k)}(x) + \Delta(x)$
5. If $\delta < \epsilon$ then STOP (the equation is solved with ϵ accuracy)
6. Substitute $k \coloneqq k + 1$ and return to step 2.

As noted earlier the Gatteaux derivative of a linear operator $\mathbb{F}(u) \coloneqq \mathbb{L}(u) \equiv \mathbb{L}u$ has a very simple form

$$\langle D\mathbb{L}(u), v\rangle \overset{\text{def}}{=} \lim_{\epsilon\to 0}\frac{\mathbb{L}(u+\epsilon v) - \mathbb{L}(u)}{\epsilon} = \lim_{\epsilon\to 0}\frac{\mathbb{L}u + \epsilon\mathbb{L}v - \mathbb{L}u}{\epsilon} = \mathbb{L}v \qquad (160)$$

Exercise:

Present the Newton algorithm for the Problem B and for the Problem C with the differential equation (146) and with the boundary condition $u(a) + 2\left(\frac{du}{dx}\right)^2(a) = u_{a1}$ and $u(b) = u_b$.

Exercise

Present the Newton Algorithm for the following BVP

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + (1+u^2)\dfrac{\partial^2 u}{\partial y^2} = 1 \qquad (x,y) \in \Omega = \langle 0,1\rangle \times \langle 0,1\rangle \\ u|_{\partial\Omega} = 0 \end{cases} \qquad (161)$$

14. Model scalar equations in 1D

The model equations described below are important to understand the behaviour of more complex PDE's (also nonlinear), but they also play a role in understanding certain behaviour of the discretisation schemes.

Advection equation (model hyperbolic equation)

The linear advection equation describes the simplest process of transport of the passive scalar. This equation involves the evolution in time and space and can be expressed as:

$$\begin{cases} \dfrac{\partial u}{\partial t} + c\dfrac{\partial u}{\partial x} = 0 \\ u(x, t = 0) = f(x) \end{cases}$$

or with an alternative notation as

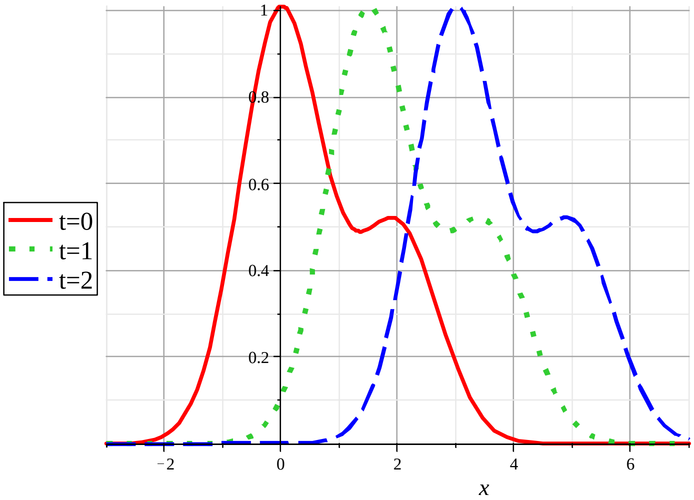
$$\begin{cases} u_t + cu_x = 0 \\ u(x, 0) = f(x) \end{cases} \tag{162}$$

where c is a constant value. The general solution to this Initial Value Problem (IVP) is given by:

$$u(x, t) = f(x - ct) \tag{163}$$

This is easily verified as $u_t = -cf'$, while $u_x = f'$ (therefore $u_t + cu_x = 0$).

Now it is clear that the solution of the advection equation is a function which moves to the right (if $c > 0$) with a constant speed c, while the shape of the function does not change. This is illustrated in the Figure below, which presents the solution at $t = 0, 1, 2$ $(c = 1.5)$.



It is also beneficial to perform complex Fourier analysis of this equation, and analyse the method which will be useful in the next cases. For this purpose we assume that we have complex valued function and the complex valued initial condition (in the form of complex Fourier mode $e^{ikx} = \cos kx + i \sin kx$, where $k > 0$ denotes the wave number):

$$\begin{cases} u_t + cu_x = 0 \\ u(x, 0) = e^{ikx} \end{cases} \tag{164}$$

The solution we seek in the form of:

$$u(x, t) = e^{i(kx - \omega t)} \tag{165}$$

where ω is an unknown coefficient (angular frequency), which has to be determined.

We have:

$$u_x = ike^{i(kx-\omega t)} = -iku \qquad u_t = -i\omega e^{i(kx-\omega t)} = -i\omega u \tag{166}$$

As a result the Partial Differential Equation (PDE) (164) is replaced by a simple algebraic equation allowing to determine ω:

$$-i\omega + ikc = 0 \Longrightarrow \ \omega = kc \tag{167}$$

The final solution is expressed as:

$$u(x,t) = e^{ik(x-ct)} \tag{168}$$

which perfectly agrees with the previous expression for the solution.

The most important properties of this exact solutions are:

- The amplitude of the solution does not change
- All waves move with the same speed c (irrespectively of the value of k).

Diffusion equation (model parabolic equation)

The diffusion PDE is a second order equation and describes phenomena related to viscous molecular or thermal diffusion:

$$\begin{cases} \dfrac{\partial u}{\partial t} = \nu \dfrac{\partial^2 u}{\partial x^2} \\ u(x,t=0) = e^{ikx} \end{cases} \quad \text{or with alternative notation as} \quad \begin{cases} u_t = \nu u_{xx} \\ u(x,0) = e^{ikx} \end{cases} \tag{169}$$

In this case we are unable to provide a simple solution for the initial condition $f(x)$, and only the Fourier mode initial condition is considered. In this case, assuming that the solution is $u(x,t) = e^{i(kx-\omega t)}$, we have again:

$$u_{xx} = -k^2 e^{i(kx-\omega t)} = -k^2 u \qquad u_t = -i\omega e^{i(kx-\omega t)} = -i\omega u \tag{170}$$

which allows to determine ω and the solution as:

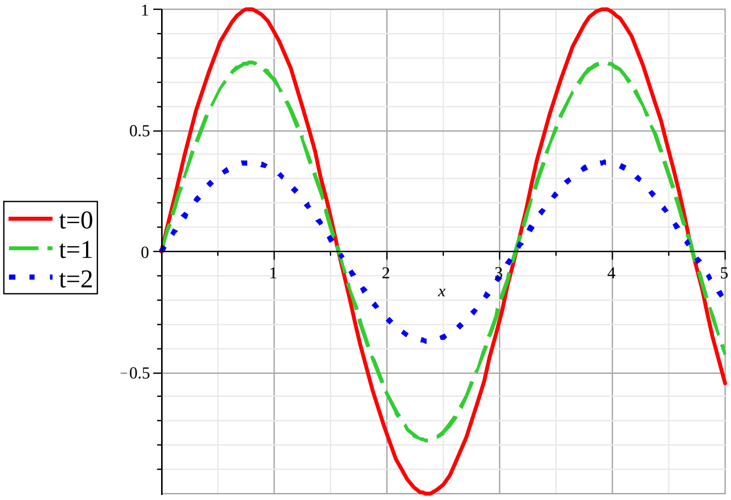
$$\begin{gathered} -i\omega = -\nu k^2 \Longrightarrow \omega = -i\nu k^2 \\ u(x,t) = e^{-\nu k^2 t}\, e^{ikx} \end{gathered} \tag{171}$$

The first exponential term plays a role of the decreasing (in time) amplitude of the initial Fourier mode.

The most important properties of this exact solutions are:

- The amplitude of the solution decreases (ν is always positive), decreases the faster the higher wave number is taken.
- The form of the initial Fourier mode is preserved.

These features of the solution are presented in the Figure below for the three consecutive moments of time:



Advection-Diffusion equation

The advection-diffusion PDE is a second order equation and describes phenomena related to the transport in conjunction with viscous, molecular or thermal diffusion:

$$\begin{cases} \dfrac{\partial u}{\partial t} + c\dfrac{\partial u}{\partial x} = \nu\dfrac{\partial^2 u}{\partial x^2} \\ u(x, t = 0) = e^{ikx} \end{cases}$$ or with alternative notation as $$\begin{cases} u_t + cu_x = \nu u_{xx} \\ u(x, 0) = e^{ikx} \end{cases} \quad (172)$$

In this case we are again unable to provide a simple solution for the initial condition $f(x)$, and only the Fourier mode initial condition is considered. In this case, assuming that the solution is $u(x,t) = e^{i(kx-\omega t)}$, we have again:

$$u_{xx} = -k^2 e^{i(kx-\omega t)} = -k^2 u \qquad u_t = -i\omega e^{i(kx-\omega t)} = -i\omega u$$
$$u_x = ike^{i(kx-\omega t)} = iku \quad (173)$$

which allows to determine ω and the solution as:

$$-i\omega + ikc = -\nu k^2 \Longrightarrow \omega = kc - i\nu k^2$$
$$u(x,t) = e^{-\nu k^2 t}\, e^{ik(x-ct)} \quad (174)$$

The most important properties of this exact solutions are:

- The amplitude of the solution decreases (ν is always positive), decreases the faster the higher wave number is taken.
- The initial Fourier mode travels with the speed c, but otherwise its form does not change.

Telegraph equation

The telegraph PDE is a third-order equation describing the wave propagation in the long telegraph lines:

$$\begin{cases} \dfrac{\partial u}{\partial t} + c\dfrac{\partial u}{\partial x} = -\lambda\dfrac{\partial^3 u}{\partial x^3} \\ u(x, t = 0) = e^{ikx} \end{cases}$$ or with alternative notation as $$\begin{cases} u_t + cu_x = -\lambda u_{xxx} \\ u(x, 0) = e^{ikx} \end{cases} \quad (175)$$

In this case we are again unable to provide a simple solution for the initial condition $f(x)$, and only the Fourier mode initial condition is considered (indeed this is of interest for telegraph lines in which propagation of waves are of main interest). In this case, assuming that the solution is $u(x,t) = e^{i(kx-\omega t)}$, we have again:

$$u_{xxx} = -ik^3 e^{i(kx-\omega t)} = -ik^3 u \qquad u_t = -i\omega e^{i(kx-\omega t)} = -i\omega u$$
$$u_x = ike^{i(kx-\omega t)} = iku \tag{176}$$

which allows to determine ω and the solution as:

$$-i\omega + ikc = i\lambda k^3 \Longrightarrow \omega = kc-i\lambda k^3 = k(c-\lambda k^2)$$
$$u(x,t) = e^{ik[x-(c-\lambda k^2)t]} \tag{177}$$

In the above $s(k) = c - \lambda k^2$ plays a role of the speed of wave propagation.

The most important properties of this exact solutions are:

- The amplitude of the solution does not change.
- The long waves $\lambda k^2 \ll c$ travel with the speed c, the shorter waves (k larger) depending on the sign of λ propagate faster (if $\lambda < 0$) or slower (if $\lambda > 0$).

The fact that the speed of propagation depends on the wave number is called a wave dispersion and this phenomenon constitutes the main reason of distortion of the signals in the telegraph (telephone) lines.

15. Multidimensional first order PDE

In analogy to the scalar equation (164) we will consider now the multidimensional case:

with

$$\begin{cases} \dfrac{\partial u}{\partial t} + A\dfrac{\partial u}{\partial x} = 0 \\ u(x,t=0) = f(x) \end{cases} \qquad u = \begin{bmatrix} u_1 \\ \dots \\ u_p \\ \dots \\ u_n \end{bmatrix}, f = \begin{bmatrix} f_1 \\ \dots \\ f_p \\ \dots \\ f_n \end{bmatrix} \qquad A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & \dots \\ \dots & \dots & \ddots & \dots \\ a_{n1} & \dots & \dots & a_{nn} \end{bmatrix} \tag{178}$$

However, as it will become evident through inspection of selected examples, this equation goes far beyond simple advection (or indeed a hyperbolic problem). In fact a very broad class of linear PDE's in (t,x) can be expressed in the form (178).

Example 1

Consider now the second-order wave equation:

$$\frac{\partial^2 v}{\partial t^2} - c^2\frac{\partial^2 v}{\partial x^2} = 0 \tag{179}$$

and a new vector variable $u(x,t) = \begin{bmatrix} u_1(x,t) \\ u_2(x,t) \end{bmatrix}$, where:

$$\begin{matrix} u_1(x,t) = \dfrac{\partial v}{\partial x} \\ u_2(x,t) = \dfrac{\partial v}{\partial t} \end{matrix} \quad \Rightarrow \quad \begin{matrix} \dfrac{\partial u_1}{\partial t} - \dfrac{\partial u_2}{\partial x} = 0 \\ \dfrac{\partial u_2}{\partial t} - c^2\dfrac{\partial u_1}{\partial x} = 0 \end{matrix} \tag{180}$$

The resulting system has the form:

$$\frac{\partial u}{\partial t} + \begin{bmatrix} 0 & -1 \\ -c^2 & 0 \end{bmatrix}\frac{\partial u}{\partial x} = 0, \quad A = \begin{bmatrix} 0 & -1 \\ -c^2 & 0 \end{bmatrix} \tag{181}$$

The characteristic polynomial of the matrix A is $w(\lambda) = \lambda^2 - c^2$, thus the two eigenvalues are real and equal $\lambda = \pm c$

Example 2

Consider now the second-order Laplace equation (written for variables t and x):

$$\frac{\partial^2 v}{\partial t^2} + \frac{\partial^2 v}{\partial x^2} = 0 \tag{182}$$

and a new vector variable $u(x,t) = \begin{bmatrix} u_1(x,t) \\ u_2(x,t) \end{bmatrix}$, where:

$$\begin{matrix} u_1(x,t) = \dfrac{\partial v}{\partial x} \\ u_2(x,t) = \dfrac{\partial v}{\partial t} \end{matrix} \quad \Rightarrow \quad \begin{matrix} \dfrac{\partial u_1}{\partial t} - \dfrac{\partial u_2}{\partial x} = 0 \\ \dfrac{\partial u_2}{\partial t} + \dfrac{\partial u_1}{\partial x} = 0 \end{matrix} \tag{183}$$

The resulting system has the form:

$$\frac{\partial u}{\partial t} + \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\frac{\partial u}{\partial x} = 0, \quad A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{184}$$

The characteristic polynomial of the matrix A is $w(\lambda) = \lambda^2 + 1$, thus the two eigenvalues are imaginary and equal $\lambda = \pm i$

Thus not all systems (178) have the same properties, and therefore are only partially of our interest.

We are here interested almost exlusively in the so called hyperbolic systems:

Definition:

We will call the system hyperbolic (178) if and only if the matrix $A_{n\times n}$ is diagonalisable and all the eigenvalues are real. In such case the right eigenvectors $r_{(1)}, r_{(2)}, \ldots, r_{(n)}$ form the basis in $\mathbb{R}^n$, and:

$$A = R\Lambda R^{-1}$$

where

$$R = \left[r_{(1)}, r_{(2)}, \ldots, r_{(n)}\right], \qquad Ar_p = \lambda_p r_p \tag{185}$$

$$\Lambda = RAR^{-1} \quad \text{and} \quad AR = R\Lambda$$

Now we are ready to present exact solution of the hyperbolic IVP problem (178), (185). For this purpose we left multiply (178) by a constant matrix R^{-1}, to obtain the equation:

$$\begin{cases} \dfrac{\partial(R^{-1}u)}{\partial t} + (R^{-1}AR)\dfrac{\partial(R^{-1}u)}{\partial x} = 0 \\ \quad R^{-1}u(x, t = 0) = R^{-1}f(x) \end{cases} \tag{186}$$

in which a new variable $v = R^{-1}u$ is introduced, to deliver a system of uncoupled equations:

$$\begin{cases} \dfrac{\partial v}{\partial t} + \Lambda\dfrac{\partial v}{\partial x} = 0, \qquad u = Rv \\ v(x, t = 0) = R^{-1}f(x) = g(x) \end{cases} \tag{187}$$

as a result we obtain a system of decoupled scalar advection equations:

$$\begin{cases} \dfrac{\partial v_1}{\partial t} + \lambda_1\dfrac{\partial v_1}{\partial x} = 0 \\ v_1(x, t = 0) = g_1(x) \end{cases}$$

$$\ldots \tag{188}$$

$$\begin{cases} \dfrac{\partial v_n}{\partial t} + \lambda_n\dfrac{\partial v_n}{\partial x} = 0 \\ v_n(x, t = 0) = g_n(x) \end{cases}$$

which we can solve in the finite form:

$$\begin{gathered} v_1(x, t) = g_1(x - \lambda_1 t) \\ v_2(x, t) = g_2(x - \lambda_2 t) \\ \ldots \\ v_n(x, t) = g_n(x - \lambda_n t) \end{gathered} \tag{189}$$

to finally get the solution in original variables:

$$u(x, t) = R\begin{bmatrix} g_1(x - \lambda_1 t) \\ g_2(x - \lambda_2 t) \\ \ldots \\ g_n(x - \lambda_n t) \end{bmatrix} \tag{190}$$

Example 3

Consider now again the example of the IVP for second-order wave equation (the initial value is prescribed for the function itself as well as for the time derivative, as the equation is of second order with respect to time t):

$$\begin{cases} \dfrac{\partial^2 w}{\partial t^2} - c^2 \dfrac{\partial^2 w}{\partial x^2} = 0 \\ w(x, t = 0) = a(x) \\ \dfrac{\partial w}{\partial t} = w_t(x, t = 0) = 0 \end{cases} \tag{191}$$

where $a(x)$ denotes a prescribed known function. The new vector variable $u(x,t) = \begin{bmatrix} u_1(x,t) \\ u_2(x,t) \end{bmatrix}$, is defined as:

$$\begin{aligned} u_1(x,t) &= \frac{\partial w}{\partial t} \equiv w_t \\ u_2(x,t) &= \frac{\partial w}{\partial x} \equiv w_x \end{aligned} \quad \Rightarrow \quad \begin{aligned} \frac{\partial u_1}{\partial t} - c^2 \frac{\partial u_2}{\partial x} = 0 \\ -\frac{\partial u_2}{\partial t} + \frac{\partial u_1}{\partial x} = 0 \end{aligned} \tag{192}$$

$$u(x, t = 0) = \begin{bmatrix} w_{t0} \\ w_{x0} \end{bmatrix} = \begin{bmatrix} 0 \\ a_x \end{bmatrix} \tag{193}$$

The resulting system has the form:

$$\frac{\partial u}{\partial t} + \begin{bmatrix} 0 & -1 \\ -c^2 & 0 \end{bmatrix} \frac{\partial u}{\partial x} = 0, \quad A = \begin{bmatrix} 0 & -1 \\ -c^2 & 0 \end{bmatrix} \tag{194}$$

The characteristic polynomial of the matrix A is $w(\lambda) = \lambda^2 - c^2$, thus the two eigenvalues are real and equal $\lambda_{1,2} = \pm c$. The eigenvectors $r_{(1)}, r_{(2)}$ and the R and R^{-1} matrices are

$$r_{(1)} = \begin{bmatrix} 1 \\ -c \end{bmatrix}, \quad r_{(2)} = \begin{bmatrix} 1 \\ c \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 1 \\ -c & c \end{bmatrix}, \quad R^{-1} = \frac{1}{2c} \begin{bmatrix} c & -1 \\ -c & 1 \end{bmatrix} \tag{195}$$

The transformed variables are:

$$\begin{gathered} v(x,t) = R^{-1} u(x,t), \quad u(x,t) = R v(x,t) \\ g(x) \equiv v(x, t = 0) = R^{-1} f(x) = \frac{1}{2c} \begin{bmatrix} c & -1 \\ -c & 1 \end{bmatrix} \begin{bmatrix} 0 \\ a_x \end{bmatrix} = \frac{1}{2c} \begin{bmatrix} -a_x \\ a_x \end{bmatrix} \end{gathered} \tag{196}$$

The solution of (194) can be therefore expressed as:

$$v(x,t) = \frac{1}{2c} \begin{bmatrix} -a_x(x - ct) \\ a_x(x + ct) \end{bmatrix} \tag{197}$$

Now the original solution $u(x,t) = Rv(x,t)$ is:

$$u(x,t) = \frac{1}{2c} \begin{bmatrix} 1 & 1 \\ -c & c \end{bmatrix} \begin{bmatrix} -a_x(x - ct) \\ a_x(x + ct) \end{bmatrix} = \frac{1}{2c} \begin{bmatrix} -a_x(x - ct) + a_x(x + ct) \\ c[a_x(x - ct) + a_x(x + ct)] \end{bmatrix} \tag{198}$$

The solution of the wave-equation (191) $w(x,t)$ is therefore:

$$\begin{aligned} w(x,t) = \int w_x \, dx = \int u_2(x,t) \, dx = \frac{1}{2} \int [a_x(x - ct) + a_x(x + ct)] dx = \\ = \frac{1}{2} [a(x - ct) + a(x + ct)] \end{aligned} \tag{199}$$

The exact solution presented in this is Section is of some theoretical interest, helping to understand the structure of the solution of the multidimensional linear hyperbolic problems. However it is not very useful for solving numerically the nonlinear Euler equations. Before we tackle nonlinear problems, we have to recall the discretisation schemes for the linear scalar advection type problems as well for the linear multidimensional hyperbolic problems.

Example 4

Suppose now, that we have the following initial value problem

$$\begin{cases} \dfrac{\partial u}{\partial t} + \begin{bmatrix} 1 & -3 \\ -2 & 2 \end{bmatrix} \dfrac{\partial u}{\partial x} = 0 \\ u(x, t = 0) = \begin{bmatrix} 0 \\ \sin(x) \end{bmatrix} \end{cases} \tag{200}$$

Analysing the matrix for eigenvalues we obtain

$$\begin{gathered} A = \begin{bmatrix} 1 & -3 \\ -2 & 2 \end{bmatrix}, \quad \lambda_1 = -1, \; \lambda_2 = 4, \; r_{(1)} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \quad r_{(1)} = \begin{bmatrix} 2 \\ -2 \end{bmatrix} \\ Q = \begin{bmatrix} 3 & 2 \\ 2 & -2 \end{bmatrix}, \quad Q^{-1} = \frac{1}{10} \begin{bmatrix} 2 & 2 \\ 2 & -3 \end{bmatrix} \end{gathered} \tag{201}$$

Therefore the initial condition for $v = Q^{-1}u$ is

$$v(x, t = 0) = Q^{-1}u(x, t = 0) = \frac{1}{10} \begin{bmatrix} 2\sin(x) \\ -3\sin(x) \end{bmatrix} \tag{202}$$

and the solution $v(x, t)$ can be expressed as:

$$v(x, t) = \frac{1}{10} \begin{bmatrix} 2\sin(x + 1t) \\ -3\sin(x - 4t) \end{bmatrix} \tag{203}$$

Returning to the original unknown function $u(x, t)$ we finally get the exact analytical solution

$$u(x, t) = Q \cdot v(x, t) = \frac{1}{10} \begin{bmatrix} 6\sin(x + t) - 6\sin(x - 4t) \\ 4\sin(x + t) + 6\sin(x - 4t) \end{bmatrix} \tag{204}$$

It can be easily verified that both the equation and the initial condition are fulfilled.

16. Discretisation of the scalar advection equation

We will recall here basic facts concerning the discretisation of the scalar advection equation, known from the earlier basic Computational Fluid Dynamics course:

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0 \qquad\qquad u_t + cu_x = 0 \qquad (205)$$

will be discretised on the space-time domain (see the Figure below).

$$u_j^p = u(x_j, t_p) = u(jh, p\Delta) \qquad j = 0, \pm 1, \pm 2, \dots. \quad p = 1,2,3, \dots \qquad (206)$$

The analysis of the discretisation formulas is based on the Lax theorem which states the that the convergence of finite difference discretisation to the exact solution is subject to two conditions:

- Consistency (the finite difference formula should properly approximate the differential formula)
- Stability (the numerical solution should not blow up in time)

Finite difference formulas

The basic finite difference formulas will be now recalled (see the online book Computational Fluid Dynamics by the present author), as they are used in further considerations. These formula were derived by the technique based on the Taylor expansion.

Derivative (A)	Finite Difference formula (B)	Error term (B-A)	Name
u_{xj}	$\frac{u_{j+1} - u_j}{h}$	$\frac{h}{2}u_{xxj} + \frac{h^3}{6}u_{xxxj} + \cdots$	One sided formula
u_{xj}	$\frac{u_{j+1} - u_{j-1}}{2h}$	$\frac{h^2}{6}u_{xxxj} + \frac{h^4}{120}u_{xxxxxj} + \cdots$	Central difference
u_{xxj}	$\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2}$	$\frac{h^2}{12}u_{xxxxj} + \frac{h^4}{360}u_{xxxxxxj} + \cdots$	---

Table 1, Finite difference formulas

These formulas are presented for the space derivatives u_x and u_{xx}, but the same formulas hold also for the time derivatives.

Explicit Euler Formula

The simplest possible discretisation of the advection equation with a forward time finite difference and with the central spatial finite difference formula:

$\frac{u_j^{p+1} - u_j^p}{\Delta} + c\frac{u_{j+1}^p - u_{j-1}^p}{2h} = 0$	The discretisation error is proportional to $O(\Delta) + O(h^2)$	(207)

allows to evaluate new value of the solution at the next time level u_j^{p+1}. This formula although straightforward and simple is numerically unusable being unconditionally unstable (the numerical solutions blows up in time – despite the fact the exact solution is fully bounded).

Explicit one sided formula

Interestingly less accurate discretisation with forward time finite difference and with the one sided spatial finite difference formula has much better properties:

$\frac{u_j^{p+1} - u_j^p}{\Delta} + c\frac{u_j^p - u_{j-1}^p}{h} = 0$	The discretisation error is proportional to $O(\Delta) + O(h)$	(208)

Despite being less accurate this discretisation formula is conditionally stable provided:

$$c > 0 \;\; and \; \Delta \leq \frac{h}{c} \tag{209}$$

The $\Delta \leq \frac{h}{c}$ condition is a typical Courant-Friedrichs-Levy condition (CFL) expressing the physical requirement that in one time step Δ the grid information cannot travel more than one computational cell h.

The symmetric formula:

$\frac{u_j^{p+1} - u_j^p}{\Delta} + c\frac{u_{j+1}^p - u_j^p}{h} = 0$	The discretisation error is proportional to $O(\Delta) + O(h)$	(210)

Is in turn stable for:

$$c < 0 \;\; and \; \Delta \leq \frac{h}{|c|} \tag{211}$$

This two formulas give rise to the slightly artificial in this context upwind formulation, characterised by the same discretisation error, but valid for all values of c:

$$\frac{u_j^{p+1} - u_j^p}{\Delta} + c_+\frac{u_j^p - u_{j-1}^p}{h} c_- \frac{u_{j+1}^p - u_j^p}{h} = 0$$
$$c_+ = \max(c, 0)\,, \qquad c_- = \min(c, 0) \tag{212}$$

This upwind formula has many far reaching generalisations for nonlinear and multidimensional problems. This formula is no longer linear as both c_+ and c_- are in principle nonlinear functions (this will become fully clear for multidimensional as well nonlinear problems).

The explicit Lax-Friedrichs formula

The explicit Euler formula can be improved, and made stable by replacing the u_j^p in the time derivative, by the spatial average of the solution at the previous time step:

$\frac{u_j^{p+1} - \frac{u_{j+1}^p + u_{j-1}^p}{2}}{\Delta} + c\frac{u_{j+1}^p - u_{j-1}^p}{2h} = 0$	The discretisation error is proportional to $O(\Delta) + O(h)$	(213)

with the same CFL condition for stability:

$$\Delta \leq \frac{h}{|c|} \tag{214}$$

It is also interesting to notice that the Lax-Friedrichs formula can be rewritten as explicit Euler formula with additional term on the right hand side:

$$\frac{u_j^{p+1}-u_j^p}{\Delta}+c\frac{u_{j+1}^p-u_{j-1}^p}{2h}=\frac{h^2}{2\Delta}\left(\frac{u_{j-1}^p-2u_j^p+u_{j+1}^p}{h^2}\right) \tag{215}$$

Which forms a valid discretisation of the advection equation, but at the same time looks as a second order spatial discretisation of the advection-diffusion equation:

$$u_t+cu_x=\epsilon u_{xx} \quad \text{where} \quad \epsilon=\frac{h^2}{2\Delta} \tag{216}$$

This modification can thus be understood as supplementing the original advection equation by a term of artificial viscosity (artificially added to stabilise the numerical system).

Implicit formulas

All formulas above can be made implicit in time, which generally makes them unconditionally stable, but at the very high numerical cost as implicit formulations require solving the large linear systems to obtain the numerical solution.

The examples for the Euler formula and for the upwind formulas are presented below:

$\frac{u_j^{p+1}-u_j^p}{\Delta}+c\frac{u_{j+1}^{p+1}-u_{j-1}^{p+1}}{2h}=0$	The discretisation error is proportional to $O(\Delta)+O(h^2)$

(217)

$$\frac{u_j^{p+1}-u_j^p}{\Delta}+c_+\frac{u_j^{p+1}-u_{j-1}^{p+1}}{h}+c_-\frac{u_{j+1}^{p+1}-u_j^{p+1}}{h}=0$$
$$c_+=\max(c,0)\,, \qquad c_-=\min(c,0) \tag{218}$$

Lax-Wendroff formula

The earlier formulations can be improved in time-discretisation accuracy to deliver $O(\Delta^2)+O(h^2)$, by extending the concept of artificial viscosity presented earlier. The Lax-Wendroff formulation can be expressed in the following form:

$$\frac{u_j^{p+1}-u_j^p}{\Delta}+c\frac{u_{j+1}^p-u_{j-1}^p}{2h}=\frac{c^2\Delta}{2}\left(\frac{u_{j-1}^p-2u_j^p+u_{j+1}^p}{h^2}\right) \tag{219}$$

which is table for $\Delta \leq \frac{h}{|c|}$

Beam-Warming formula

The analogous one side second order formulation $O(\Delta^2)+O(h^2)$ can be expressed by:

$$\frac{u_j^{p+1}-u_j^p}{\Delta}+c\frac{3u_j^p-4u_{j-1}^p+u_{j-2}^p}{2h}=\frac{c^2\Delta}{2}\left(\frac{u_j^p-2u_{j-1}^p+u_{j-2}^p}{h^2}\right) \tag{220}$$

which is stable for $c>0$ and $\Delta \leq \frac{2h}{|c|}$

The analogous symmetric formula can be shown to be stable for $c<0$.

Despite improved accuracy the higher-order formulas of Lax-Wendroff and Beam-Warming type are only of limited further interest in the context of simulation of compressible flows. In such flows discontinuities appear as rule, and in such cases other properties (monotonicity) are much more important than the formal accuracy of the scheme.

17. Discretisation of the multidimensional hyperbolic equation

We will attempt now to extend the scalar formulas of previous Section to the multidimensional hyperbolic case:

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} = 0 \qquad u = \begin{bmatrix} u_1 \\ \dots \\ u_p \\ \dots \\ u_n \end{bmatrix}, f = \begin{bmatrix} f_1 \\ \dots \\ f_p \\ \dots \\ f_n \end{bmatrix} \qquad A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & \dots \\ \dots & \dots & \ddots & \dots \\ a_{n1} & \dots & \dots & a_{nn} \end{bmatrix} \tag{221}$$

or

$$u_t + Au_x = 0$$

i.e., in which the matrix A is diagonalisable (the eigenvectors form the basis in $\mathbb{R}^n$), and the corresponding eigenvalues are real:

$$A = R\Lambda R^{-1}$$

where

$$R = \left[r_{(1)}, r_{(2)}, \dots, r_{(n)}\right], \qquad Ar_{(p)} = \lambda_p r_{(p)} \tag{222}$$

$$\Lambda = RAR^{-1} \quad \text{and} \quad AR = R\Lambda$$

As demonstrated by (187) the system can be expressed in the decoupled form ($v = R^{-1}u$), in which a system of decoupled scalar advection equations is obtained:

$$\begin{gathered} \frac{\partial v_1}{\partial t} + \lambda_1 \frac{\partial v_1}{\partial x} = 0 \\ \dots \\ \frac{\partial v_n}{\partial t} + \lambda_n \frac{\partial v_n}{\partial x} = 0 \end{gathered} \tag{223}$$

This scalar system can be discretised, e.g., by the Lax-Friedrichs scheme:

$$\begin{gathered} \frac{v_{1j}^{p+1} - \dfrac{v_{1j+1}^p + v_{1j-1}^p}{2}}{\Delta} + \lambda_1 \frac{v_{1j+1}^p - v_{1j-1}^p}{2h} = 0 \\ \dots \\ \frac{v_{nj}^{p+1} - \dfrac{v_{nj+1}^p + v_{nj-1}^p}{2}}{\Delta} + \lambda_n \frac{v_{nj+1}^p - v_{nj-1}^p}{2h} = 0 \end{gathered} \tag{224}$$

and expressed in the following vector form:

$$\frac{v_j^{p+1} - \dfrac{v_{j+1}^p + v_{j-1}^p}{2}}{\Delta} + \Lambda \frac{v_{j+1}^p - v_{j-1}^p}{2h} = 0 \tag{225}$$

Multiplying by R from the left side we obtain the scheme formulated for the original variables:

$$\frac{u_j^{p+1} - \dfrac{u_{j+1}^p + u_{j-1}^p}{2}}{\Delta} + \mathrm{A} \frac{u_{j+1}^p - u_{j-1}^p}{2h} = 0 \tag{226}$$

It is clear from the above that discretisation formulas do not change for vector systems of first order-equations (as both the equations and the formulas are linear). However for the vector upwind case, the formulas become more interesting:

$$\frac{v_{1j}^{p+1} - v_{1j}^{p}}{\Delta} + \lambda_{1+}\frac{v_{1j}^{p} - v_{1j-1}^{p}}{h} + \lambda_{1-}\frac{v_{1j+1}^{p} - v_{1j}^{p}}{h} = 0$$

$$\dots \tag{227}$$

$$\frac{v_{nj}^{p+1} - v_{nj}^{p}}{\Delta} + \lambda_{n+}\frac{v_{nj}^{p} - v_{nj-1}^{p}}{h} + \lambda_{n-}\frac{v_{nj+1}^{p} - v_{nj}^{p}}{h} = 0$$

Again multiplying by R from the left side we obtain the scheme formulated for the original variables:

$$\frac{u_{j}^{p+1} - u_{j}^{p}}{\Delta} + A_{+}\frac{u_{j}^{p} - u_{j-1}^{p}}{2h} + A_{-}\frac{u_{j+1}^{p} - u_{j}^{p}}{2h} = 0 \tag{228}$$

where:

$$A_{+} = R\Lambda_{+}R^{-1}, A_{-} = R\Lambda_{-}R^{-1}, \quad A = A_{+} + A_{-} \tag{229}$$

The matrices A_{+} and A_{-} filter the positive and negative eigenvalues of the matrix A, allowing for the stable discretisation. The CFL condition in this case has the following form:

$$\Delta \leq \frac{h}{\max_{j}|\lambda_{j}|} \tag{230}$$

This upwind method is no longer linear as it contain switching function between the sign of the eigenvalues. Basing on this nonlinear switching it is possible to improve the accuracy of the formula, circumventing the Godunov barrier.

18. Nonlinear hyperbolic equations

As presented in Chapter 4 the time dependent Euler equations have a form

$$\frac{\partial U}{\partial t}+\frac{\partial}{\partial x}\mathbb{F}_c(U)=0, \qquad U=\begin{bmatrix}\rho\\ \rho u\\ \rho E\end{bmatrix}=\begin{bmatrix}\rho\\ m\\ \epsilon\end{bmatrix}\in\mathbb{R}^3 \tag{231}$$

in which ρ, u, E stand for density, velocity and total energy per unit mass respectively. The scalar model equation (nonlinear advection equation) that will be considered and analysed is now:

$$\frac{\partial u}{\partial t}+\frac{\partial}{\partial x}f(u)=0 \qquad \text{or} \qquad u_t+[f(u)]_x=0 \tag{232}$$

where $u(x,t)$ is a scalar solution, while $f(u)$ is a known nonlinear function.

The simplest equation of this kind is the Burgers equation, for which $f(u)=u^2/2$:

$$\frac{\partial u}{\partial t}+\frac{\partial}{\partial x}\left(\frac{u^2}{2}\right)=0 \tag{233}$$

With the initial condition:

$$u(x,t=0)=u_0(x) \tag{234}$$

Equation (232) is presented in the so called “conservative form” as it describes the conservation of quantity represented by u, function $f(u)$ represents flux of this quantity. This equation can be transformed into the more familiar quasilinear form, by executing the differentiation with respect to x

$$\frac{\partial u}{\partial t}+\frac{\partial f}{\partial u}\frac{\partial u}{\partial x}=0$$

or (235)

$$\frac{\partial u}{\partial t}+a(u)\frac{\partial u}{\partial x}=0, \quad a(u)\stackrel{\text{def}}{=}\frac{\partial f}{\partial u}$$

This equation can be solved by the method of characteristics, which will be presented here in general and for three examples of increasing complexity.

The method of characteristics

The method of characteristics will be presented here for slightly more general equation than (235) but with the usual initial condition

$$\begin{cases}\dfrac{\partial u}{\partial t}+a(u,x)\dfrac{\partial u}{\partial x}=0\\ u(x,t=0)=u_0(x)\end{cases} \tag{236}$$

Let now $x=x_*(t)$ to denote an arbitrary curve in a (t,x) plane. On this line the solution $u(x,t)$ is equal to

$$u_*(t)\stackrel{\text{def}}{=}u(x_*(t),t) \tag{237}$$

We are looking for the (family) of lines on which the solution $u_*(t)$ is constant, i.e.,

$$\frac{du_*}{dt}=0 \Leftrightarrow \frac{\partial u}{\partial t}+\frac{dx_*}{dt}\cdot\frac{\partial u}{\partial x}=0 \tag{238}$$

This equation is identical with the nonlinear advection equation (236) provided the following ordinary differential equation is fulfilled

$$\begin{cases} \dfrac{dx_*}{dt} = a(u_*, x_*) \\ x_*(t=0) = x_0 \end{cases} \tag{239}$$

We do not know the solution $u(x,t)$ a priori, however we know that on the characteristic line $x_*(t)$ $u_*(t) = const \equiv u(x_0, t=0)$. The latter value is known and therefore

$$\begin{cases} \dfrac{dx_*}{dt} = a(u(x_0), x_*) \\ x_*(t=0) = x_0 \end{cases} \tag{240}$$

If we know the analytic solution to this equation, we obtain the family of curves parametrised with the value of x_0. In the following we will use the method of characteristics to solve the equation (236) graphically.

Linear equation with constant coefficient - a(u, x)=c

We know already the solution of this linear advection equation

$$u(x,t) = u_0(x - ct) \tag{241}$$

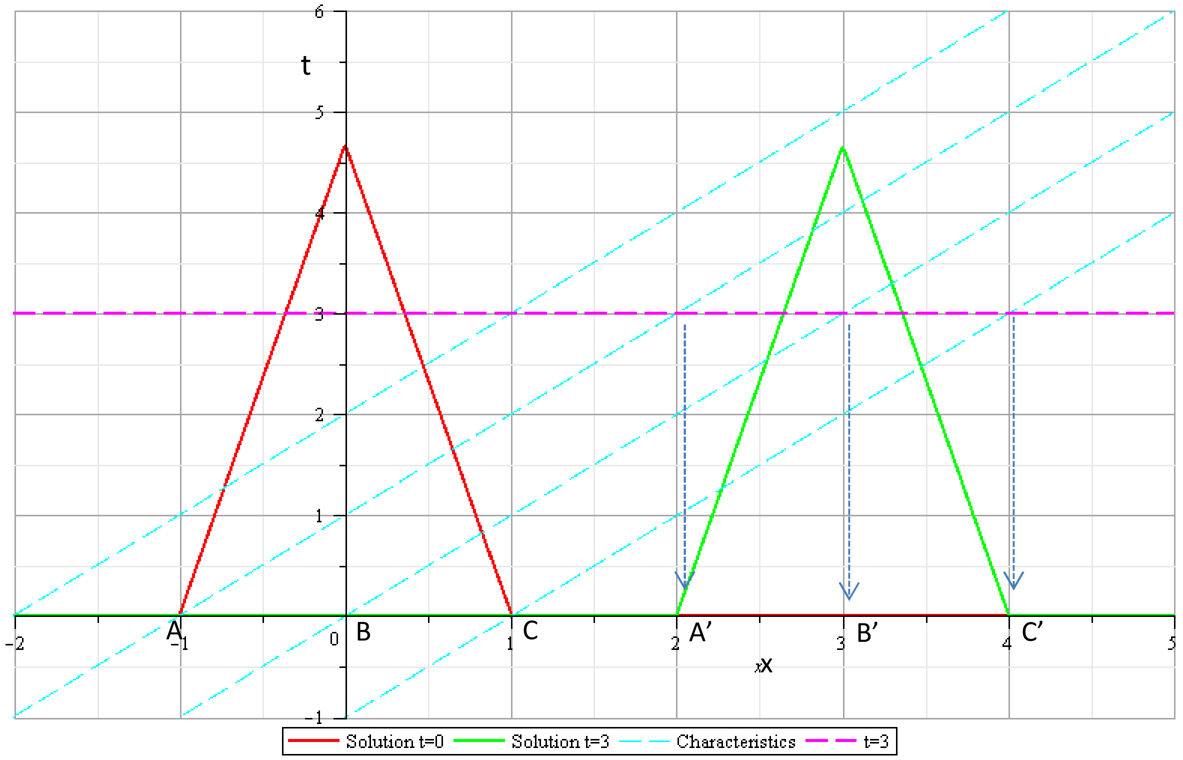
This means that on the straight lines $x = x_0 + ct$ the (x,t) plane) the solution is always constant (these lines are the characteristics). The characteristics can be alternatively found as a solution to the ODE (240)

$$\begin{cases} \dfrac{dx_*}{dt} = c \\ x_*(t=0) = x_0 \end{cases} \quad \Longrightarrow \quad x_*(t) = x_0 + ct \tag{242}$$

Let assume now that $u_0(x)$ in the initial condition has the form

$$u_0(x) = \begin{cases} 0 & \text{for } \; |x| > 1 \\ x+1 & \text{for } \; -1 < x < 0 \\ 1-x & \text{for } \; 0 < x < 1 \end{cases} \tag{243}$$

We seek now the form of the solution at time $t = t_1$. The geometric construction is based on the analysis of time evolution of the selected characteristic points of the function $u_0(x)$ (on the x-axis) A, B, C, which move to different locations A', B', C' (see next Fig.).



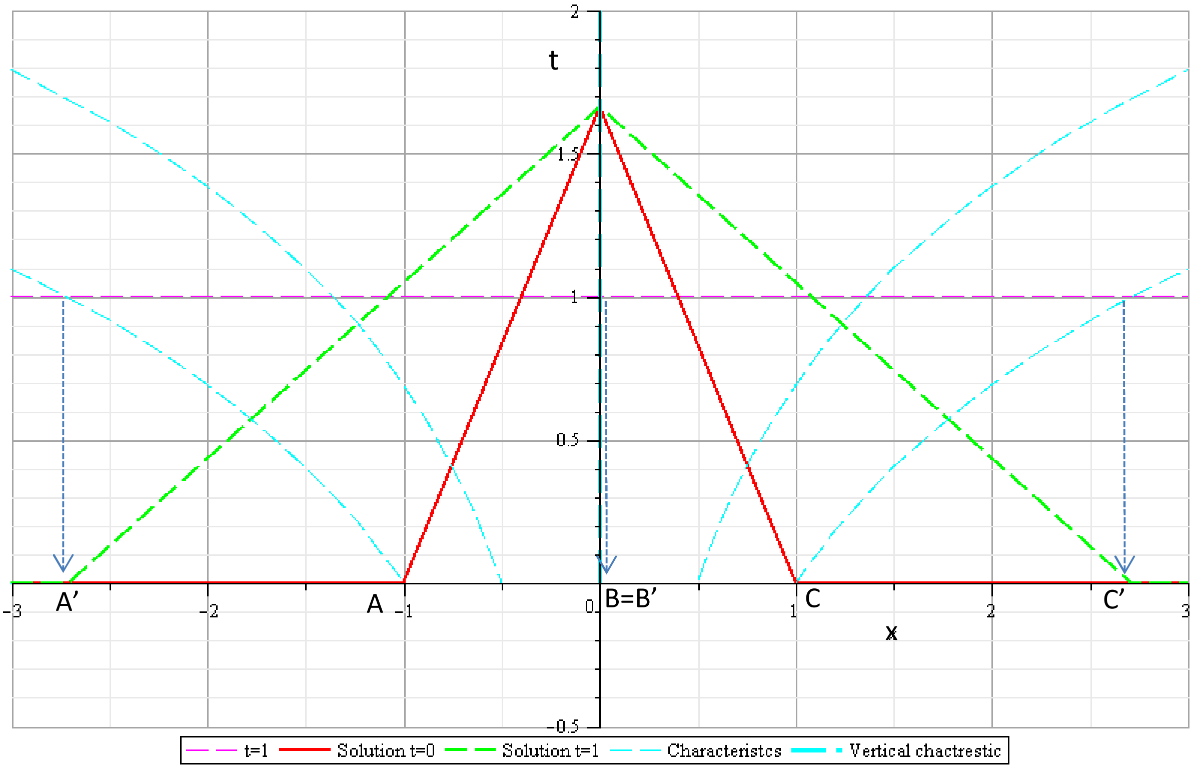
Linear equation with variable coefficient - a(u, x)=x

We will consider the case when $a(u,x) = x$ with a previous initial condition $u_0(x)$. We will seek the solution at time $t = t_1$. In this case the IVP has the following form

$$\begin{cases} \dfrac{\partial u}{\partial t} + x\dfrac{\partial u}{\partial x} = 0 \\ u(x, t = 0) = u_0(x) \end{cases} \tag{1}$$

and therefore the equation for characteristics and the family of curves are

$$\begin{cases} \dfrac{dx_*}{dt} = x_* \\ x_*(t = 0) = x_0 \end{cases} \quad \Longrightarrow \quad x_*(t) = x_0 e^t \tag{2}$$



The curvilinear characteristics are presented in the figure above as broken blue lines. The initial condition is denoted by the red solid line, while the solution at time $t = 1$ by the dashed green line.

Nonlinear equation - a(u, x)=u

This case corresponds to the Burgers equation

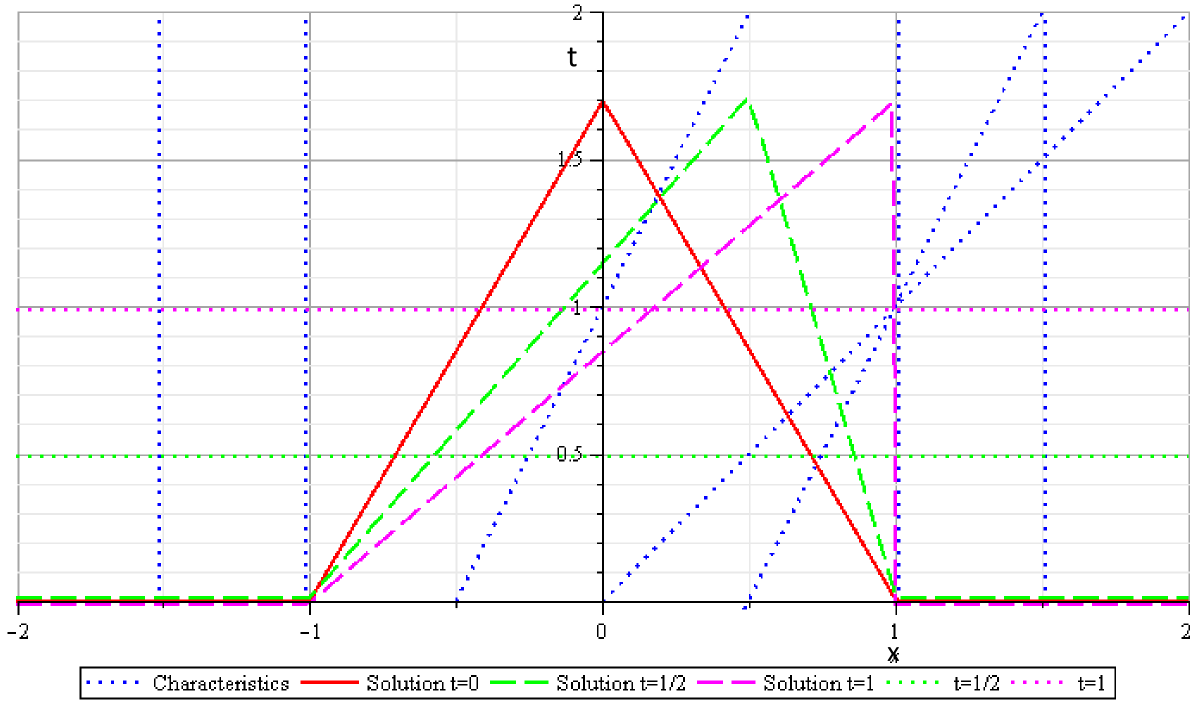
$$\begin{cases} \dfrac{\partial u}{\partial t} + u\dfrac{\partial u}{\partial x} = 0 \\ u(x, t = 0) = u_0(x) \end{cases} \tag{244}$$

This equation contains significant nonlinearity, which also can be found in the Euler and Navier-Stokes equations. The solution is sought for times t_1 and t_2.

The equation for the family of characteristics is:

$$\begin{cases} \dfrac{dx_*}{dt} = u(x_0) \\ x_*(t = 0) = x_0 \end{cases} \quad \Longrightarrow \quad x_*(t) = x_0 + u(x_0)t \tag{245}$$

Again the characteristics are the straight lines, with inclination depending on the initial value of the solution.



The analysis of the results shows (in this particular case) that above $t = t_* \equiv 1$ the characteristics start to overlap, which would mean that for a fixed argument the solution has two different values (as the characteristics carry different values of solution). Thus we have to conclude that above t_* the method of characteristics no longer can be used to predict the solution. Nevertheless we can expect that for $t = t_*$ the discontinuity appears in the solution.

This is a typical feature of the nonlinear hyperbolic equations. Even if the initial condition is continuous and regular (smooth) the solution remains smooth only for a finite time. Therefore each numerical scheme used to solve such equations must be able to cope with discontinuities.

It should also be investigated how these discontinuities evolve in time (how fast they propagate, are all discontinuities stable/permanent).

The discontinuities of Burgers equation correspond to similar features in Fluid Mechanics, i.e., shockwaves and contact discontinuities.

Weak solutions of the nonlinear hyperbolic equations

If as explained earlier the solution to the equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0 \qquad \text{or} \qquad u_t + [f(u)]_x = 0 \tag{246}$$

has discontinuities, we must understand what does this mean for the original problem (around discontinuity both derivatives u_t and u_x do not exist).

For this purpose we assume, that the so-called *test function* $\Phi(x,t) \in C_0^1(\mathbb{R} \times \mathbb{R}_+)$, i.e., that this function is continuously differentiable and vanishes at infinity (both in space and time).

We multiply now (246) by $\Phi(x,t)$ and integrate it over half-space $\langle -\infty, \infty\rangle \times \langle 0, \infty\rangle$

$$0 = \int_{-\infty}^{\infty}\int_{0}^{\infty} (u_t + [f(u)]_x) \cdot \Phi \; dt\, dx \tag{247}$$

After integration by parts we obtain:

$$\begin{gathered}\int_{0}^{\infty} u_t \cdot \Phi \; dt = -\int_{0}^{\infty} u \cdot \Phi_{\mathrm{t}} \; dt - u(x,0) \cdot \Phi(x,0) \\ \int_{-\infty}^{\infty} [f(u)]_x \cdot \Phi \, dx = -\int_{-\infty}^{\infty} f(u) \cdot \Phi_{\mathrm{x}} \, dx\end{gathered} \tag{248}$$

And therefore the whole equation is:

$$0 = -\int_{-\infty}^{\infty}\int_{0}^{\infty} \Phi_t u + \Phi_x \, f(u) \; dt\, dx + \int_{-\infty}^{\infty} \mathrm{u(x,0)} \cdot \Phi(x,0) \; dx \tag{249}$$

This equation is equivalent to (246) for differentiable $u(x,t)$,but admits also discontinuous solutions as it contains no derivatives of $u(x,t)$. This equation is called a weak version of (246).

A function $u(x,t)$ is called a weak solution of (246) if it fulfils the equation (249) for every function $\Phi(x,t) \in C_0^1(\mathbb{R} \times \mathbb{R}_+)$. It has to be stressed that weak solutions may not be unique, and additional conditions are needed to select the single correct solution. In fluid mechanics these conditions are based on the entropy condition (2nd law of thermodynamics).

Now we will try to investigate the evolution of discontinuities and their stability.

Propagation of discontinuities (scalar case)

To investigate how the discontinuities propagate, we shall consider the following Riemann initial value problem:

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x} f(u) = 0 \\ u(x,t=0) = \begin{cases} u_L & \text{for } \; x < 0 \\ u_R & \text{for } \; x \geq 0 \end{cases} \end{cases} \tag{250}$$

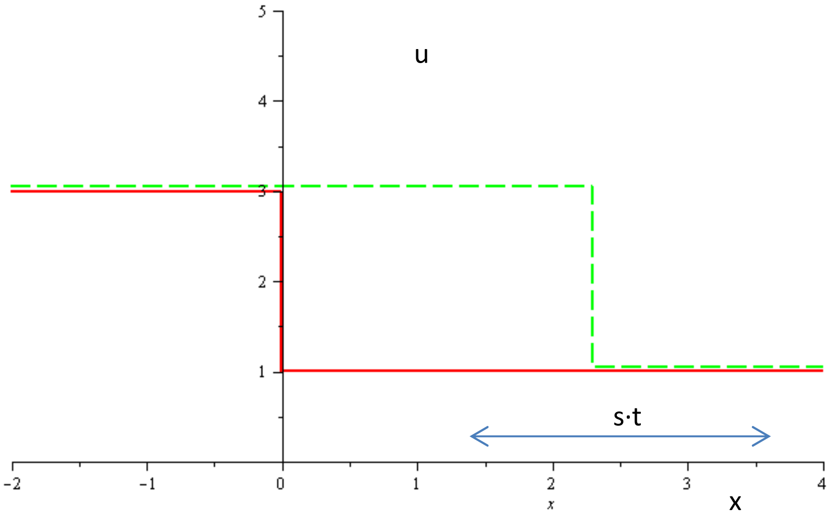
Here discontinuity is present already in the initial condition and therefore the solution is understood in the sense of the weak formulation (249).

We will seek the solution $u(x,t)$ in the form:

$$u(x,t) = \begin{cases} u_L & \text{for } \; x < st \\ u_R & \text{for } \; x \geq st \end{cases} \tag{251}$$

which means that the discontinuity moves without changing its intensity to the right with the speed s (if $s > 0$). This speed s is unknown and has to be determined.

One should note that $u(x,t) = const$ is an evident solution of(246) and therefore the assumption that the discontinuity moves without changing shape and with a constant speed seems natural. We will see further on, that this is not always the case.



Let now (251) be a solution (250) and let M be a large number. To determine the speed s we calculate now the definite integral of u_t

$$\int_{-M}^{M} u_t(x,t)dx = -\int_{-M}^{M} [f(u)]_x\, dx = f\big(u(-M,t)\big) - f\big(u(M,t)\big) = f(u_L) - f(u_R) \quad (252)$$

On the other hand we can calculate this integral directly (see Figure above)

$$\begin{gathered}\int_{-M}^{M} u(x,t)dx = (M+st)u_L + (M-st)u_R \\ \frac{d}{dt}\int_{-M}^{M} u(x,t)dx = s(u_L - u_R)\end{gathered} \quad (253)$$

Therefore

$$\begin{gathered}s(u_L - u_R) = f(u_L) - f(u_R) \\ s = \frac{f(u_L) - f(u_R)}{u_L - u_R}\end{gathered} \quad (254)$$

The last relation allows to calculate the speed of propagation of the discontinuity and is known as the Rankine-Hugoniot formula. In Fluid Mechanics analogous formula allows to determine the shockwave speed.

It must be stressed again, that the assumption of the particular form of the solution (251) may not be correct and in such cases also (254) is no longer valid.

Three special cases will now be considered, for which Rankine-Hugoniot formula will be evaluated:

a. Linear advection equation - $f(u) = cu$

$$s \equiv c \quad (255)$$

(For linear equations the discontinuities travel along the characteristics)

b. For Burgers equation - $f(u) = u^2/2$

$$\boxed{s = \frac{\frac{u_L^2}{2} - \frac{u_R^2}{2}}{u_L - u_R} \equiv \frac{u_L + u_R}{2}}$$

c. For “more” nonlinear equation - $f(u) = u^3$

$$\boxed{s = \frac{u_L^3 - u_R^3}{u_L - u_R} \equiv u_L^2 + u_L u_R + u_R^2}$$

Which discontinuities are permanent?

We will consider now the Burgers equation (244) with an initial condition (IC) consisting of two discontinuities (red solid line in the Figure below)

$$u_0(x) = \begin{cases} 0 & \text{for} \qquad x \le -1 \\ 1 & \text{for} \quad -1 < x < 1 \\ 0 & \text{for} \qquad x \ge 1 \end{cases} \tag{256}$$

In order to analyse whether the discontinuity is permanent, we shall consider a regularised $\tilde{u}_0(x)$ for which discontinuities are replaced by a very steep linear functions (in the ε neighbourhood) – see next Figure. The new initial condition (green broken line) is continuous and therefore a method of characteristics can be used. It can be easily noticed that the “right” discontinuity quickly reappears after time $t^* = 2\varepsilon$ (see magenta broken line), while the left one is further smoothed out - the inclination of the linear function drops down.

Further evolution of the former “left” discontinuity one can predict with the method of characteristics. The “right” discontinuity is permanent and its movement can be described by Rankine-Hugoniot relation (254).

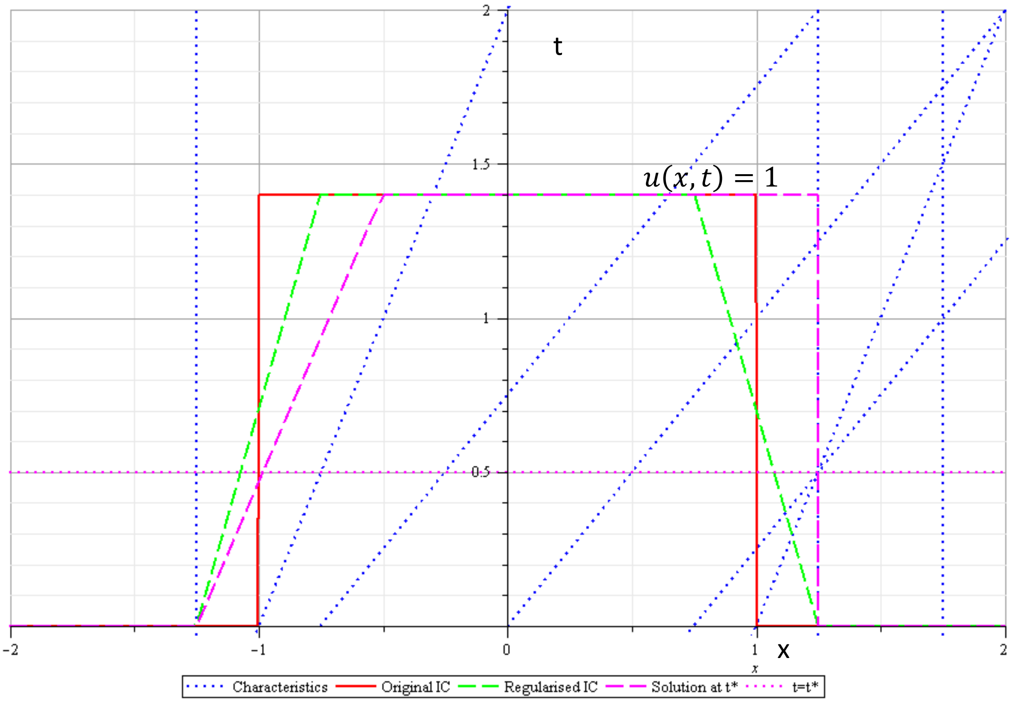


Figure 1

Therefore one may conclude that not all discontinuities are permanent, and as a consequence the week formulation (249) may contain solutions that need to be eliminated (by some additional argument). In Fluid Mechanics (for Euler equations) the identical phenomenon appears but there the physical argument is used instead of the presented geometric/kinematic reasoning (only compression shockwaves exist, the hypothetical rarefaction shockwaves violate the second law of thermodynamics).

For Navier-Stokes equation this additional conditions are not necessary as physical dissipation prevents from creation of unphysical solutions (the Navier-Stokes equations are not reversible in time - in contrast to the inviscid Euler equations). Nevertheless for small viscosity (large Reynolds number) the discretised Navier-Stokes equations (being underresolved on the insufficiently dense mesh) as a rule require additional artificial dissipation to prevent creation of false shocks.

It should be noted in addition, that for the linear hyperbolic equations (with constant coefficient c) all discontinuities are admissible and permanent (as in such case characteristics have same inclination and do not cross).

Conservative vs. quasilinear formulation

We have shown (250)-(254) how to calculate the shock speed basing it on the conservative version of the nonlinear advection equation

$$\frac{\partial u}{\partial t}+\frac{\partial}{\partial x}f(u)=0 \qquad (257)$$

This version is called conservative because it is the original form, when the equation is obtained from (some) conservation law. In contrast, if the derivative with respect to x is further evaluated, the nonconservative or quasilinear form is obtained

$$\frac{\partial u}{\partial t}+a(u)\frac{\partial u}{\partial x}=0 \quad \text{where} \quad a(u)\equiv\frac{\partial f}{\partial u} \qquad (258)$$

From numerical point of view it is important to know, (i) can this quasilinear version be discretised and solved to give correct solution and in particular (ii) will the obtained shock/discontinuity speed be the same as for the conservative version (257).

To investigate this latter question (in some indirect manner) we shall consider the Riemann problem for the Burgers equation

$$\frac{\partial u}{\partial t}+\frac{\partial}{\partial x}\left(\frac{u^2}{2}\right)=0 \qquad (259)$$

The speed of propagation of discontinuities is.

$$s=\frac{u_L+u_R}{2} \qquad (260)$$

The Burgers equation in the quasilinear form is

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}=0 \qquad (261)$$

the question remains, which value of shock speed can be associated with this equation (e.g., if solved numerically) – will it be in particular (260), or will it be perhaps some different value?. To indirectly answer this question we will multiply the quasilinear equation by u

$$u\frac{\partial u}{\partial t}+u^2\frac{\partial u}{\partial x}=0 \quad\Leftrightarrow\quad \frac{\partial}{\partial t}\left(\frac{u^2}{2}\right)+\frac{\partial}{\partial x}\left(\frac{u^3}{3}\right)=0 \tag{262}$$

and obtain different conservative equation

$$\frac{\partial v}{\partial t}+\frac{\partial}{\partial x}\left(\frac{2}{3}v^{3/2}\right)=0 \qquad \text{where} \qquad v=u^2 \tag{263}$$

We can calculate the speed of propagation of discontinuities for this equation:

$$s_*=\frac{2}{3}\frac{v_L^{3/2}-v_R^{3/2}}{v_L-v_R}\equiv\frac{2}{3}\frac{u_L^3-u_R^3}{u_L^2-u_R^2} \tag{264}$$

This formula is obviously different than (260) and thus, e.g., for $u_L=2$ and $u_R=0$ one obtains $s=1$ and $s_*=4/3$. This proves that the same quasilinear equation is equivalent to two different conservative equations with two different shock speeds. Therefore it must be concluded that the quasilinear equation, when discretised, will produce a solution with a false shock speed (and indeed this is the case when we try to solve the nonlinear discretisation).

For more advanced problems when the stationary solution is sought, the nonconservative equation will always produce the shock with a wrong intensity and location (this is a common observation for the transonic solutions of the Euler equations). This effect is not large yet it adversely affects the accuracy of computations, especially where drag estimation is concerned (shockwaves generate drag dependent directly on their intensity). Therefore all present numerical codes for compressible flows base on the conservative version of the Fluid Dynamic equations (be it Euler or Navier-Stokes).

Propagation of discontinuities (vector case)

We consider now the Riemann problem for the multidimensional nonlinear hyperbolic equation

$$\begin{cases} \dfrac{\partial u}{\partial t}+\dfrac{\partial}{\partial x}\mathbb{F}(u)=0 \\ u(x,t=0)=\begin{cases} u_L & \text{for} \quad x<0 \\ u_R & \text{for} \quad x\geq 0\end{cases}, \end{cases} \qquad u=\begin{bmatrix}u_1\\ \vdots\\ u_n\end{bmatrix}, \qquad \mathbb{F}(u)=\begin{bmatrix}F_1(u_1,\dots,u_n)\\ \vdots\\ F_n(u_1,\dots,u_n)\end{bmatrix}\in\mathbb{R}^n \tag{265}$$

and its quasilinear form

$$\begin{array}{c}\dfrac{\partial u}{\partial t}+A(u)\dfrac{\partial u}{\partial x}=0\\ u(x,t=0)=\begin{cases} u_L & \text{for} \quad x<0' \\ u_R & \text{for} \quad x\geq 0\end{cases}\end{array} \qquad A(u)\overset{\text{def}}{=}\frac{\partial\mathbb{F}(u)}{\partial u}\equiv\begin{bmatrix}\dfrac{\partial F_1}{\partial u_1} & \cdots & \dfrac{\partial F_1}{\partial u_n}\\ \vdots & \ddots & \vdots\\ \dfrac{\partial F_n}{\partial u_1} & \cdots & \dfrac{\partial F_n}{\partial u_n}\end{bmatrix} \tag{266}$$

The matrix A depends on the solution u, however in order to understand the nonlinear case we shall consider now the Riemann problem for $A=const$,i.e., for the linear vector hyperbolic equation (221)

$$\begin{cases} \dfrac{\partial u}{\partial t} + A\dfrac{\partial u}{\partial x} = 0 \\ u(x, t = 0) = \begin{cases} u_L & \text{for } x < 0 \\ u_R & \text{for } x \geq 0 \end{cases} \end{cases} \qquad u = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},\ u_L = \begin{bmatrix} u_{L1} \\ \vdots \\ u_{Ln} \end{bmatrix},\ u_R = \begin{bmatrix} u_{R1} \\ \vdots \\ u_{Rn} \end{bmatrix} \tag{267}$$

where the matrix A is diagonalisable

$$A = R\Lambda R^{-1} \qquad \begin{matrix} R = \left[r_{(1)}, r_{(2)}, \dots, r_{(n)}\right] \\ \Lambda = diag(\lambda_1, \dots, \lambda_n) \end{matrix} \qquad Ar_{(p)} = \lambda_p r_{(p)} \tag{268}$$

If the analysis carried out earlier for scalar equation is repeated, we may obtain the analogue of the Rankine-Hugoniot formula:

$$s\,(u_L - u_R) = A(u_L - u_R) \tag{269}$$

Which indicates that discontinuities (shocks) may travel unchanged with a speed s only if the jump vector $(u_L - u_R)$ is an eigenvector r_p of matrix A (and in such case $s \equiv \lambda_p$).

To further analyse this case we multiply now (as previously) the equation (267) by R^{-1} and cary out the following substitutions

$$v = R^{-1}u \qquad v_L = R^{-1}U_L \equiv \begin{bmatrix} v_{L1} \\ \vdots \\ v_{Ln} \end{bmatrix} \qquad v_R = R^{-1}U_R \equiv \begin{bmatrix} v_{R1} \\ \vdots \\ v_{Rn} \end{bmatrix} \tag{270}$$

to obtain

$$\begin{matrix} \dfrac{\partial v}{\partial t} + \Lambda\dfrac{\partial v}{\partial x} = 0 \\ v(x, t = 0) = \begin{cases} v_L & \text{for } x < 0 \\ v_R & \text{for } x \geq 0 \end{cases} \end{matrix} \quad \Leftrightarrow \quad \begin{matrix} \dfrac{\partial v_p}{\partial t} + \lambda_\mathrm{p}\dfrac{\partial v_p}{\partial x} = 0 \\ v_p(x, t = 0) = \begin{cases} v_{Lp} & \text{for } x < 0 \\ v_{Rp} & \text{for } x \geq 0 \end{cases} \end{matrix} \tag{271}$$

According to (255) the solution is:

$$v_p(x,t) = H(x, v_{Lp}, v_{Rp}, \lambda_p t) \stackrel{\text{def}}{=} \begin{cases} v_{Lp} & \text{for } x < \lambda_p t \\ v_{Rp} & \text{for } x \geq \lambda_p t \end{cases} \tag{272}$$

In the above $H(x, a, b, x_*)$ stands for the jump function (Heaviside like function), for which the discontinuity appears at the point x_*.

Finally the solution to the original problem can be expressed as

$$u(x,t) = R \cdot v(x,t) = R \cdot \begin{bmatrix} H(x, v_{L1}, v_{R1}, \lambda_1 t) \\ \vdots \\ H(x, v_{Ln}, v_{Rn}, \lambda_n t) \end{bmatrix} \tag{273}$$

Extension of this procedure to the fully nonlinear hyperbolic equation is difficult and requires further analysis.

Example 5

We shall present now the explicit solution to the following Riemann initial value problem

$$\begin{cases} \dfrac{\partial u}{\partial t} + \begin{bmatrix} 1 & -3 \\ -2 & 2 \end{bmatrix} \dfrac{\partial u}{\partial x} = 0 \\ u(x, t = 0) = \begin{bmatrix} u_1(x, t = 0) = \begin{cases} -1 & \text{for } x < 0 \\ 2 & \text{for } x \geq 0 \end{cases} \\ u_2(x, t = 0) = \begin{cases} 1 & \text{for } x < 2 \\ 3 & \text{for } x \geq 2 \end{cases} \end{bmatrix} \equiv \begin{bmatrix} H(x, -1, 2, 0) \\ H(x, 1, 3, 2) \end{bmatrix} \end{cases} \tag{274}$$

The following properties of A are now recalled

$$A = \begin{bmatrix} 1 & -3 \\ -2 & 2 \end{bmatrix}, \quad \lambda_1 = -1, \ \lambda_2 = 4, \ r_{(1)} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \quad r_{(1)} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$$
$$Q = \begin{bmatrix} 3 & 2 \\ 2 & -2 \end{bmatrix}, \quad Q^{-1} = \frac{1}{10} \begin{bmatrix} 2 & 2 \\ 2 & -3 \end{bmatrix} \tag{275}$$

Therefore the initial condition for $v = Q^{-1}u$ is

$$v(x, t = 0) = Q^{-1}u(x, t = 0) = \frac{1}{10} \begin{bmatrix} 2H(x, -1, 2, 0) + 2H(x, 1, 3, 2) \\ 2H(x, -1, 2, 0) - 3H(x, 1, 3, 2) \end{bmatrix} \tag{276}$$

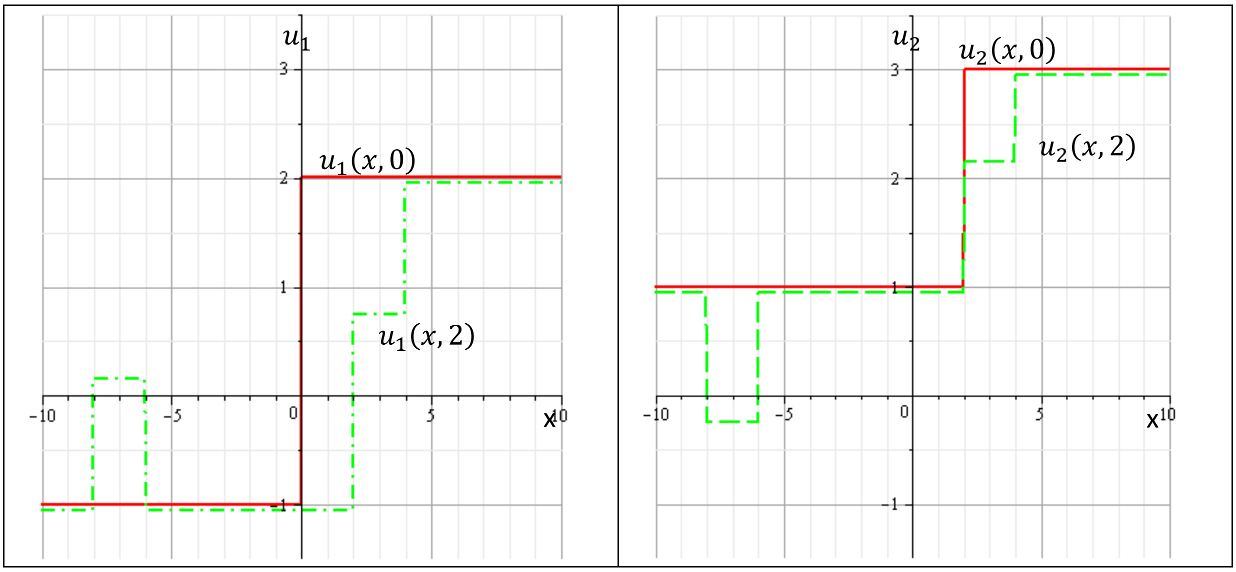
and the solution $v(x, t)$ can be expressed as:

$$v(x, t) = \frac{1}{10} \begin{bmatrix} 2H(x, -1, 2, 1t) + 2H(x, 1, 3, 2 + 1t) \\ 2H(x, -1, 2, -4t) - 3H(x, 1, 3, 2 - 4t) \end{bmatrix} \tag{277}$$

Returning to the original unknown function $u(x, t)$ we finally get the exact analytical solution

$$u(x, t) = Q \cdot v(x, t)$$
$$= \frac{1}{10} \begin{bmatrix} 6H(x, -1, 2, 1t) + 6H(x, 1, 3, 2 + 1t) + 4H(x, -1, 2, -4t) - 6H(x, 1, 3, 2 - 4t) \\ 4H(x, -1, 2, 1t) + 4H(x, 1, 3, 2 + 1t) - 4H(x, -1, 2, -4t) + 6H(x, 1, 3, 2 - 4t) \end{bmatrix} \tag{278}$$

It can be verified that both the equation and the initial condition are fulfilled. Both components of $u(x, t)$ are presented in Figure below ($t = 0$ red solid line, $t = 2$ green dashed line)



19. Godunov's order barrier theorem

Theorem

The monotonic, linear discretisation scheme for $\mathrm{u_t} + \mathrm{cu_x} = 0$ can be at most first order accurate.

This negative result adds additional difficulty in development of useful discretisation formulas for nonlinear hyperbolic systems (as obviously first order schemes are insufficiently accurate for hydrodynamic simulation purposes).

The additional requirement of monotonicity brought up by this theorem is motivated by two factors:

- The nonlinearity of equation of interest (and spontaneous generation of discontinuities)
- The oscillations appearing on discontinuities for higher order formulas

The practical consequence of Godunov theorem is that all discretisation formulas for the hyperbolic systems (e.g., Euler/Navier-Stokes equations) have to remain nonlinear, basing, e.g., either on the nonlinear upwind schemes or on the nonlinear artificial viscosity schemes.

20. Annex 1

How to calculate $\|A\|_\infty$ using the definition of the induced norm (83) ?

First we have the unit sphere and the operator acting on the sphere:

$$\|u\|_\infty \equiv \max_j |u_j| = 1$$

$$\|Au\|_\infty \equiv \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij} u_j| \leq \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|$$

$$\Downarrow \tag{1}$$

$$\|Au\|_\infty \leq \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|$$

We have shown that the norm is always smaller than some value. It is sufficient to show now that there exist unit vector for which this lower limit is actually achieved (this will be the value of the norm).

To show this we observe that, there must exist i_0 such that

$$\max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{i_0 j}| \tag{2}$$

We take now $u_* = \left[\text{sign}\, a_{i_0 1}, \text{sign}\, a_{i_0 2}, \dots, \text{sign}\, a_{i_0 n}\right]^T$, for which we have $\|u_*\|_\infty \equiv \max_j |u_{*j}| = 1$. We are now able to calculate:

$$\|Au_*\|_\infty \equiv \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij} u_{*j}| = \|u_*\| \sum_{j=1}^{n} |a_{i_0 j}| = \sum_{j=1}^{n} |a_{i_0 j}| \tag{3}$$

where the inequality was substituted by equality as all elements in the first sum are actually positive (u_* was selected in such a way to achieve this effect). Therefore:

$$\| A \|_\infty \overset{\text{def}}{=}= \sup_{\|u\|_V = 1} \|Au\|_\infty = \sum_{j=1}^{n} |a_{i_0 j}| = \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}| \tag{4}$$

The above concludes the proof.

21. Annex 2

How to calculate $\|A\|_\infty$ using the definition of the induced norm (83) ?

First we have the unit sphere and the operator acting on the sphere:

$$\|u\|_1 \equiv \sum_{j=1}^{n} |u_j| = 1$$

$$\|Au\|_1 \equiv \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} u_j \right| \le \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}| \qquad (5)$$

$$\Downarrow$$

$$\|Au\|_1 \le \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|$$

We have shown that the norm is always smaller than some value. It is sufficient to show now that there exist unit vector for which this lower limit is actually achieved (this will be the value of the norm).

To show this we observe, that there must exist j_0 such that

$$\max_{j=1,\dots,n} \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{ij_0}| \qquad (6)$$

We take now $u_* = e_{j_0} = [0, \dots, 0, 1, 0, \dots 0\,]^T$, with the only nonzero entry in the j_0 row. Therefore we have $\|u_*\|_1 \equiv \sum_{i=1}^{n} |u_{*j}| = 1$.

We are now able to calculate:

$$\|u_*\|_1 \equiv \sum_{j=1}^{n} |u_{*j}| = 1$$

$$\|Au_*\|_1 \equiv \sum_{i=1}^{n} |a_{ij_0}| = \qquad (7)$$

$$\|Au\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij_0}| = \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}| \cdot \|u_*\|_1$$

where the inequality was substituted by equality as only one column of the matrix remains (u_* was selected in such a way to achieve this effect). Therefore

$$\|A\|_1 = \sup_{\|u\|_1 = 1} \|Au\|_1 = \max_{j=1,\dots,n} \sum_{j=1}^{n} |a_{ij}| \qquad (8)$$

which concludes the proof.

22. Annex 3

How to calculate $\|A\|_2$ using the definition of the induced norm (83) ?

$$\| A \|_2 \overset{\text{def}}{=} \sup_{\|u\|_2=1} \|Au\|_2 = \sup_{\|u\|_2=1} \sqrt{u_H A^H A u} \tag{9}$$

Matrix $B \equiv A^H A = B^H$ is, as shown earlier, Hermitian and non-negative, therefore its eigenvalues λ_p are real and non-negative, while eigenvectors $v_{(p)}$ are orthogonal (in our case even orthonormal) and form the basis in $\mathbb{R}^n$ or $\mathbb{C}^n$:

$$B \cdot v_{(p)} = \lambda_p v_{(p)}, \quad p = 1,2, \dots, n \tag{10}$$

We will assume that:

$$\begin{gathered} 0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_p \leq \cdots \leq \lambda_n = \lambda_{\max}(B) \\ \left(v_{(p)}, v_{(q)}\right) = v_{(p)}{}^H \, v_{(q)} = \delta_{pq} \end{gathered} \tag{11}$$

Arbitrary $u \in \mathbb{R}^n$ or $\mathbb{C}^n$, $\|u\|_2 = 1$ can be expressed using basis functions:

$$\begin{gathered} u = \sum_{p=1}^{n} \alpha_p v_{(p)}, \quad \|u\|_2 = \sum_{p=1}^{n} \left|\alpha_p\right|^2 = 1 \\ Bu = \sum_{p=1}^{n} \alpha_p B v_{(p)} = \sum_{p=1}^{n} \alpha_p \lambda_p v_{(p)} \\ u^H B u = \sum_{p=1}^{n} \alpha_p \lambda_p \cdot \left(u, v_{(p)}\right) = \sum_{p=1}^{n} \alpha_p^2 \, \lambda_p \leq \lambda_{\max}(B) \end{gathered} \tag{12}$$

as a consequence:

$$\| A \|_2 \leq \sqrt{\lambda_{\max}(B)} \tag{13}$$

We have shown that the norm is always smaller than some value. It is sufficient to show now that there exist unit vector for which this lower limit is actually achieved (this lower limit will be the value of the norm).

Suppose we take $u = v_{(n)}$ the eigenvector corresponding to the maximum eigenvalue $\lambda_{\max}(B)$.

$$v_{(n)}^H B v_{(n)} = \lambda_{\max}(B) \tag{14}$$

Instead of inequality we obtain now the required lower limit, and as a consequence:

$$\| A \|_2 = \sqrt{\lambda_{\max}(B)} = \max_{\lambda \in \text{spect}(A^H A)} \sqrt{\lambda}$$

For Hermitian (symmetric real) matrices we have $A = A^H$ and $\lambda_p(B) = \lambda_p^2(A)$, and therefore:

$$\| A \|_2 = \lambda_{\max}(A) = \max_{\lambda \in \text{spect}(A)} |\lambda| \tag{15}$$

Remark:

The second matrix norm $\| A \|_2$ is difficult to calculate in general case, as it requires finding the maximum eigenvalue of the matrix A. Nevertheless in special cases (like for the discrete Poisson operator for the square/perpendicular domain) this value is readily available.