Lecture Notes for the Course on
NUMERICAL MODELS FOR NUCLEAR REACTORS

Prof. WALTER AMBROSINI
University of Pisa, Italy

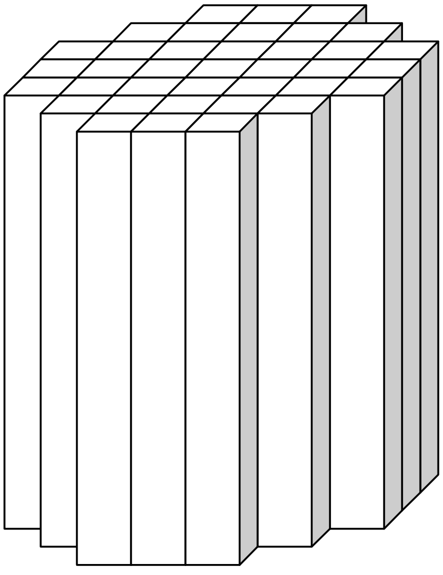
Unit 1 – Eigenvalue Problems with Neutron Diffusion and Solution Strategies: Inner Iterations

NOTICE: These notes were prepared by Prof. Ambrosini mainly **on the basis of the material adopted by Prof. Bruno Montagnini for the lectures he held up to years ago**, when he left to Prof. Ambrosini the charge of the course held for the Degree in Nuclear Engineering at the University of Pisa. This material is freely distributed to Course attendees or to anyone else requesting it. It has not the worth of a textbook and it is not intended to be an official publication. It was conceived as the notes that the teacher himself would take of his own lectures in the paradoxical case he could be both teacher and student at the same time (sometimes space and time stretch and fold in strange ways). It is also used as slides to be projected during lectures to assure a minimum of uniform, constant quality lecturing, regardless of the teacher's good and bad days. As such, the material contains reference to classical textbooks and material whose direct reading is warmly recommended to students for a more accurate understanding. In the attempt to make these notes as original as feasible and reasonable, considering their purely educational purpose, most of the material has been completely re-interpreted in the teacher's own view and personal preferences about notation. In this effort, errors in details may have been introduced which will be promptly corrected in further versions after discovery. Requests of clarification, suggestions, complaints or even sharp judgements in relation to this material can be directly addressed to Prof. Ambrosini at the e-mail address: walter.ambrosini@ing.unipi.it

EIGENVALUE (OR CRITICALITY) PROBLEMS

- **The static neutronic design of nuclear reactor core involves different aspects related to:**
  - ♦ **the calculation of the k_{eff} and of the flux spatial distribution in different configurations**
  - ♦ **the prediction of the evolution of reactor isotopical composition with the increase of burn-up: the so-called "life" or "depletion" calculations**

- **Eigenvalue problems, which can be expressed also in forms other than the simple calculation of k_{eff} (e.g., identifying the critical absorption cross section, liquid boron concentration or the critical reactor size) are quite relevant in the definition of reactor core parameters as the lattice pitch, the level of enrichment and control rod characteristics**
- **In heterogeneous reactors eigenvalue problems require the numerical solution of neutron diffusion (or more seldom transport) equations dealing with detailed geometry and isotopic composition data**
- **Both space and energy discretization is required in this step**

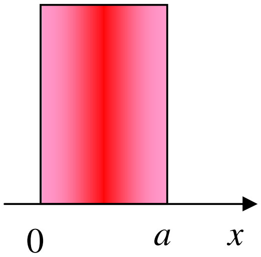


CRITICALITY CONDITIONS
FOR A PLANE HOMOGENEOUS SLAB

Reminder of the analytic solution

One-group differential problem

$$\begin{cases} D\dfrac{d^2\phi}{dx^2} - \Sigma_a\phi + \dfrac{\nu\Sigma_f}{k}\phi = 0 \\ \phi(0) = \phi(a) = 0 \end{cases}$$



The following positions can be adopted:

$$\frac{d^2\phi}{dx^2} = \frac{\Sigma_a - \nu\Sigma_f/k}{D}\phi = \mu\,\phi \qquad\qquad \mu = \mu(k) = \frac{\Sigma_a - \nu\Sigma_f/k}{D}$$

The general solution of the above equation is

$$\phi(x) = C_1 e^{\sqrt{\mu}x} + C_2 e^{-\sqrt{\mu}x}$$

Boundary conditions are then imposed

$$\phi(0) = C_1 + C_2 = 0$$

$$\phi(a) = C_1 e^{\sqrt{\mu}a} + C_2 e^{-\sqrt{\mu}a} = 0$$

This homogeneous algebraic linear system of equations has a non-trivial solution only if

$$\det\begin{bmatrix} 1 & 1 \\ e^{\sqrt{\mu}a} & e^{-\sqrt{\mu}a} \end{bmatrix} = 0$$

i.e., when

$$e^{-\sqrt{\mu}\,a} - e^{\sqrt{\mu}\,a} = 0 \;\Rightarrow\; e^{2\sqrt{\mu}\,a} = 1$$

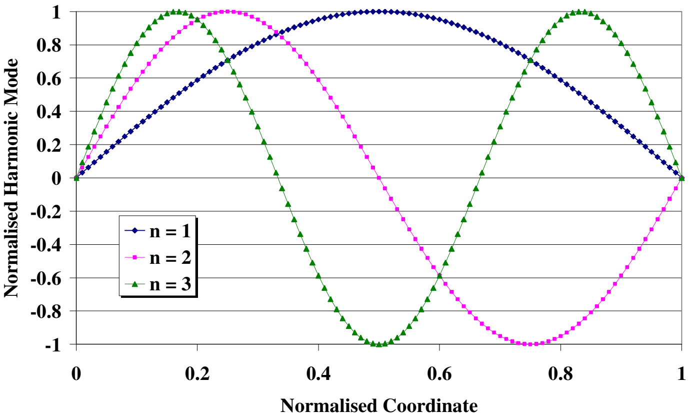
$$\Rightarrow\quad 2\sqrt{\mu}a = \ln(1) + i2n\pi = i2n\pi \quad (n = 0, \pm1, \pm2, \ldots)$$

In fact, it can be checked that it is

$$e^{\ln(1)+i2n\pi} = e^0[\cos(2n\pi) + i\,\mathrm{sen}(2n\pi)] = 1 \qquad (n = 0, \pm1, \pm2, \ldots)$$

Therefore, excluding $n = 0$ *leading to the trivial solution*, **it must be:**

$$\mu_n = -\frac{n^2\pi^2}{a^2} \qquad (n = 1, 2, \ldots)$$



Since in the definition of μ only k is variable, it is:

$$\mu_n = \frac{\Sigma_a - \nu\Sigma_f / k_n}{D} = -\frac{n^2\pi^2}{a^2} \quad \Rightarrow \quad k_n = \frac{\nu\Sigma_f}{\Sigma_a + D\dfrac{n^2\pi^2}{a^2}} \qquad (n = 1,2,\ldots)$$

As well known, it is $\phi_n(x) = C\,\mathrm{sen}\left(\dfrac{n\pi}{a}x\right) \qquad (n = 1,2,\ldots)$

Considering the different *harmonic modes* it can be observed that

- **the eigenvalues k are such as it is $k_1 > k_2 > \ldots$**
- ***the fundamental mode* is the only one being characterised by $\phi_1(x) > 0$ for $0 \le x \le a$;**
- **on the other hand, we know from Reactor Physics that it is the one on which the neutron flux settles in steady-state conditions and it is**

$$B^2 = B_1^2 = \frac{\pi^2}{a^2} \qquad \phi(x) = \phi_1(x) \qquad \frac{d^2\phi}{dx^2} = -B^2\phi$$

- **therefore, for steady-state problems, considering that the layer of multiplying material is homogeneous it can be shown that**

$$k_1 = \frac{\nu\Sigma_f}{\Sigma_a + DB^2} = \frac{\nu\Sigma_f\int_0^a \phi dx}{\left(\Sigma_a + DB^2\right)\int_0^a \phi dx} = \frac{\int_0^a \nu\Sigma_f\phi\, dx}{\int_0^a \Sigma_a\phi\, dx + \int_0^a DB^2\phi\, dx} = \frac{\int_0^a \nu\Sigma_f\phi\, dx}{\int_0^a \Sigma_a\phi\, dx - \int_0^a D\dfrac{d^2\phi}{dx^2}dx}$$

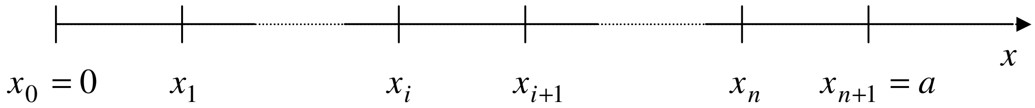
or

$$\boxed{k_1 = \frac{neutrons\ \ produced\ \ by\ \ fissions}{total\ \ number\ \ of\ \ neutrons\ \ in\ the\ \ previous\ \ generation\,(absorption + leakage)} = k_{eff}}$$

i.e., k_1 is found to be the *effective multiplication factor*

CRITICALITY CONDITIONS FOR A PLANE HOMOGENEOUS SLAB

Numerical Solution



The spatial dicretisation of the problem can be obtained by subdividing the range $(0, a)$ into segments having uniform length, h, and putting

$$\phi_i = \phi(x_i)$$

Different discretization approaches can be adopted for turning the differential equation into an algebraic one; here we choose

$$\left.\frac{d\phi}{dx}\right|_{x=x_i} \approx \frac{\phi_{i+1} - \phi_i}{h} \quad \textbf{\textit{(forward finite difference form)}}$$

However, we could also assume

$$\left.\frac{d\phi}{dx}\right|_{x=x_i} \approx \frac{\phi_i - \phi_{i-1}}{h} \quad \textbf{\textit{(backward finite difference form)}}$$

A closer examination, as well as the rigorous application of Taylor expansion, reveals that the two approximations are actually more accurate for expressing the derivatives at the locations $x_{i+1/2} = (x_i + x_{i+1})/2$ and $x_{i-1/2} = (x_{i-1} + x_i)/2$ respectively; that is:

$$\left.\frac{d\phi}{dx}\right|_{x=x_{i+1/2}} \approx \frac{\phi_{i+1} - \phi_i}{h} \qquad \left.\frac{d\phi}{dx}\right|_{x=x_{i-1/2}} \approx \frac{\phi_i - \phi_{i-1}}{h}$$

In order to obtain an approximation of the second order derivative, a symmetric expression ("centred" in x_i) can be then assumed

$$\left.\frac{d^2\phi}{dx^2}\right|_{x=x_i} \approx \frac{1}{h}\left[\left.\frac{d\phi}{dx}\right|_{x=x_{i+1/2}} - \left.\frac{d\phi}{dx}\right|_{x=x_{i-1/2}}\right] \approx \frac{1}{h}\left(\frac{\phi_{i+1} - \phi_i}{h} - \frac{\phi_i - \phi_{i-1}}{h}\right)$$

$$= \frac{\phi_{i+1} - 2\phi_i + \phi_{i-1}}{h^2}$$

Making use of these approximations, the differential equation expressing 1D diffusion in the layer can be reverted to this *finite difference* approximation form

$$D\frac{\phi_{i+1}-2\phi_i+\phi_{i-1}}{h^2}+\left(\frac{\nu\Sigma_f}{k}-\Sigma_a\right)\phi_i=0 \qquad (i=1,...,n)$$

while the boundary conditions take the form

$$\phi_0=0 \qquad \phi_{n+1}=0$$

A system of n equations in the n unknowns of the nodal fluxes $\{\phi_1,\phi_2,...,\phi_n\}$ is then obtained.

Putting

$$a_{i,i-1}=a_{i,i+1}=\frac{D}{h^2} \qquad a_{i,i}=-\frac{2D}{h^2}+\frac{\nu\Sigma_f}{k}-\Sigma_a$$

a *homogeneous linear algebraic system* is obtained, having the classical form of a *three-point equation*

$$a_{i,i-1}\phi_{i-1}+a_{i,i}\phi_i+a_{i,i+1}\phi_{i+1}=0 \qquad (i=1,...,n)$$

Writing the system in the form

$$\mathbf{A}\boldsymbol{\phi}=0$$

it can be noted that:

$$\mathbf{A}=\begin{bmatrix} \times & \times & 0 & 0 & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 & 0 & 0 & 0 \\ 0 & \times & \times & \times & 0 & 0 & 0 & 0 \\ 0 & 0 & \times & \times & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & \times & \times & \times & 0 & 0 \\ 0 & 0 & 0 & 0 & \times & \times & \times & 0 \\ 0 & 0 & 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & \times & \times \end{bmatrix}$$

- **in the system matrix the only coefficients being different from zero can be found in the main diagonal and the two adjoining ones: the matrix is *three-diagonal***
- **the neutron flux in each node is therefore directly related only to the one in the neighbouring nodes, a numerical effect clearly deriving by the discretization of the leakage term (the second order derivative).**

SOLUTION (at least from a purely theoretical standpoint):

- **The *"characteristic equation"***

  $$\det \mathbf{A}(k) = 0$$

  can be solved by any suitable means, identifying the n single or multiple *eigenvalues* $k_1, k_2, ..., k_n$.

- **The eigenvectors corresponding to each k_i are then identified:**
  - ♦ **an appropriate non-zero minor of $\mathrm{A}(k_i)$ having $n-1$ order can be selected;**
  - ♦ **the corresponding algebraic system of order $n-1$ obtained by assigning and arbitrary value to one of the unknowns ϕ_i can be solved by any suitable technique.**

Therefore, with respect to the analytic solution, the finite difference discretised equation provides <u>only</u> n <u>eigenvalues,</u> representing approximations of the eigenfunctions of the problem in the continuum space.

$$\boldsymbol{\phi}^{(1)} \equiv \begin{bmatrix} \phi_1^{(1)} \\ \circ \\ \circ \\ \phi_n^{(1)} \end{bmatrix} \qquad \boldsymbol{\phi}^{(n)} \equiv \begin{bmatrix} \phi_1^{(n)} \\ \circ \\ \circ \\ \phi_n^{(n)} \end{bmatrix}$$

For instance, in this particular case it can be demonstrated that (we skip the demonstration):

$$\phi_j^{(1)} = C \operatorname{sen} \frac{\pi\, hj}{a} \qquad\qquad (j = 1, ..., n)$$

and

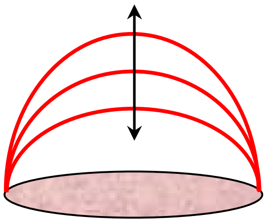
$$k = k_1 = \frac{\nu \Sigma_f}{\Sigma_a + \dfrac{2D}{h^2}\left[1 - \cos\left(\dfrac{\pi h}{a}\right)\right]} \underset{h \to 0}{\longrightarrow} \frac{\nu \Sigma_f}{\Sigma_a + D\dfrac{\pi^2}{a^2}}$$

<u>NOTES</u>:

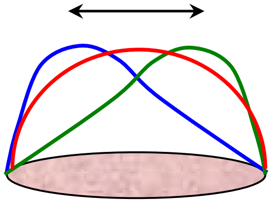
- **In most neutronic problems, it is just necessary to evaluate the *fundamental eigenvalue,* k_1, and the *corresponding fundamental eigenvector,* $\underline{\phi}^{(1)}$.**

- **Obviously in real life reactors, an accurate multidimensional space and energy description is necessary, aiming at determining reliable estimates of the k_{eff} and of the steady-state neutron flux distribution.**

- **Only in some selected cases, it is necessary to get information about the *higher order harmonic modes*, as it is the case for Boiling Water Reactors (BWRs) to evaluate the possible occurrence of "out-of-phase" or "regional" oscillations due to the thermal-hydraulic feedback.**



Core – wide
Power Oscillations



Regional
Power Oscillations

- **In fact, the thermal-hydraulic feedback may make more probable the occurrence of normally "subcritical" modes, which may mostly govern the transient evolution.**

MULTIGROUP EIGENVALUE PROBLEMS

A multigroup eigenvalue problem can be expressed making use of the diffusion theory in the form

$$div\, D_g(\vec{r}) grad_{\vec{r}} \phi_g(\vec{r}) - \Sigma_{r,g}(\vec{r})\phi_g(\vec{r}) + \sum_{g'<g} \Sigma_{s,g'\to g}(\vec{r})\phi_{g'}(\vec{r})$$

$$+\frac{\chi_g}{k}\sum_{g'=1}^{G}\nu\Sigma_{f,g'}(\vec{r})\phi_{g'}(\vec{r}) = 0 \qquad (\vec{r}\in V,\ g=1,...,G)$$

$$\phi_g(\vec{r}) + d_g\,\frac{d\phi_g(\vec{r})}{dn} = 0 \qquad (\vec{r}\in \partial V,\ \ g=1,...,G)$$

in which k is the parameter giving rise to eigenvalues; it is also

$$\chi_g = \int_{\Delta E_g}\chi(E)dE$$

i.e., χ_g is the fraction of neutrons produced by fission in g-th energy group and the *"removal cross section"* is defined as

$$\Sigma_{r,g}(\vec{r}) = \Sigma_{a,g}(\vec{r}) + \sum_{g'>g}\Sigma_{s,g\to g'}(\vec{r})$$

In order to solve these equations, it can be noted that, defining the *total fission neutron source* as

$$\psi(\vec{r}) = \sum_{g'=1}^{G}\nu\Sigma_{f,g'}(\vec{r})\phi_{g'}(\vec{r})$$

and putting

$$S_g(\vec{r}) = \chi_g\psi(\vec{r})$$

the steady-state diffusion equations with G energy groups become

$$div\, D_1(\vec{r}) grad_{\vec{r}}\phi_1(\vec{r}) - \Sigma_{r,1}(\vec{r})\phi_1(\vec{r}) + \frac{1}{k}S_1(\vec{r}) = 0$$

$$div\, D_2(\vec{r}) grad_{\vec{r}}\phi_2(\vec{r}) - \Sigma_{r,2}(\vec{r})\phi_2(\vec{r}) + \Sigma_{s,1\to 2}(\vec{r})\phi_1(\vec{r}) + \frac{1}{k}S_2(\vec{r}) = 0$$

$$div\, D_3(\vec{r}) grad_{\vec{r}}\phi_3(\vec{r}) - \Sigma_{r,3}(\vec{r})\phi_3(\vec{r}) + \sum_{g'=1}^{2}\Sigma_{s,g'\to 3}(\vec{r})\phi_{g'}(\vec{r}) + \frac{1}{k}S_3(\vec{r}) = 0$$

etc..

It can be noted that:

- **the first equation involves only the fission source and the neutron flux in the first energy group;**
- **the subsequent equations involve the fission source, the neutron flux in the group and the *slowing down* terms from previous (higher energy) groups;**
- **the initial *homogeneous problem* has been formally turned into a *non-homogeneous* one.**

It is therefore possible to envisage a solution procedure based on the following steps:

1. **a tentative distribution of the fission source, $\psi(\vec{r})$, and a tentative value of the multiplication factor, k, are assigned as $\psi^{(0)}(\vec{r})$ and $k^{(0)}$ respectively; this allows obtaining the group sources $S_g^{(0)}(\vec{r})$**
2. **after an appropriate *spatial discretization* (see later on), the group equations are solved in sequence:**
   - **the first equation is solved, obtaining an updated estimate of the neutron flux in the first group, $\phi_1^{(1)}(\vec{r})$;**
   - **the slowing down source from the first group appearing in the second equation can be therefore evaluated, then solving the equation for $\phi_2^{(1)}(\vec{r})$;**
   - **a similar procedure is adopted for all the higher index group equations, evaluating the appropriate slowing down sources from higher energy groups**
3. **the fission source can be then updated on the basis of the new fluxes, obtaining $\psi^{(1)}(\vec{r})$, and the estimate of the eigenvalues is also updated according to a "generational" formulation**

$$k^{(1)} = \frac{\text{neutrons generated by fissions at the new iteration}}{\text{neutron source adopted at the previous iteration}} = \frac{\int_V \psi^{(1)}(\vec{r})dV}{\frac{1}{k^{(0)}}\int_V \psi^{(0)}(\vec{r})dV}$$

then, the sources $S_g^{(1)}(\vec{r})$ are calculated again and the processi s repeated from step 2 until convergence.

The scheme includes two different phases:

- **iterations on the fission source, named *<u>outer iterations</u>*;**
- **the solution of the equation of single group equations, performed as if they were representing mono-energetic decoupled equations having the form**

$$div\, D(\vec{r})grad_{\vec{r}}\phi(\vec{r}) - \Sigma_a(\vec{r})\phi(\vec{r}) + S(\vec{r}) = 0$$

since very often the solution of these equations is performed by using iterative methods, this step is named *<u>inner iterations.</u>*

The adopted iteration scheme is described in the figure reported in the next page.

NOTES:

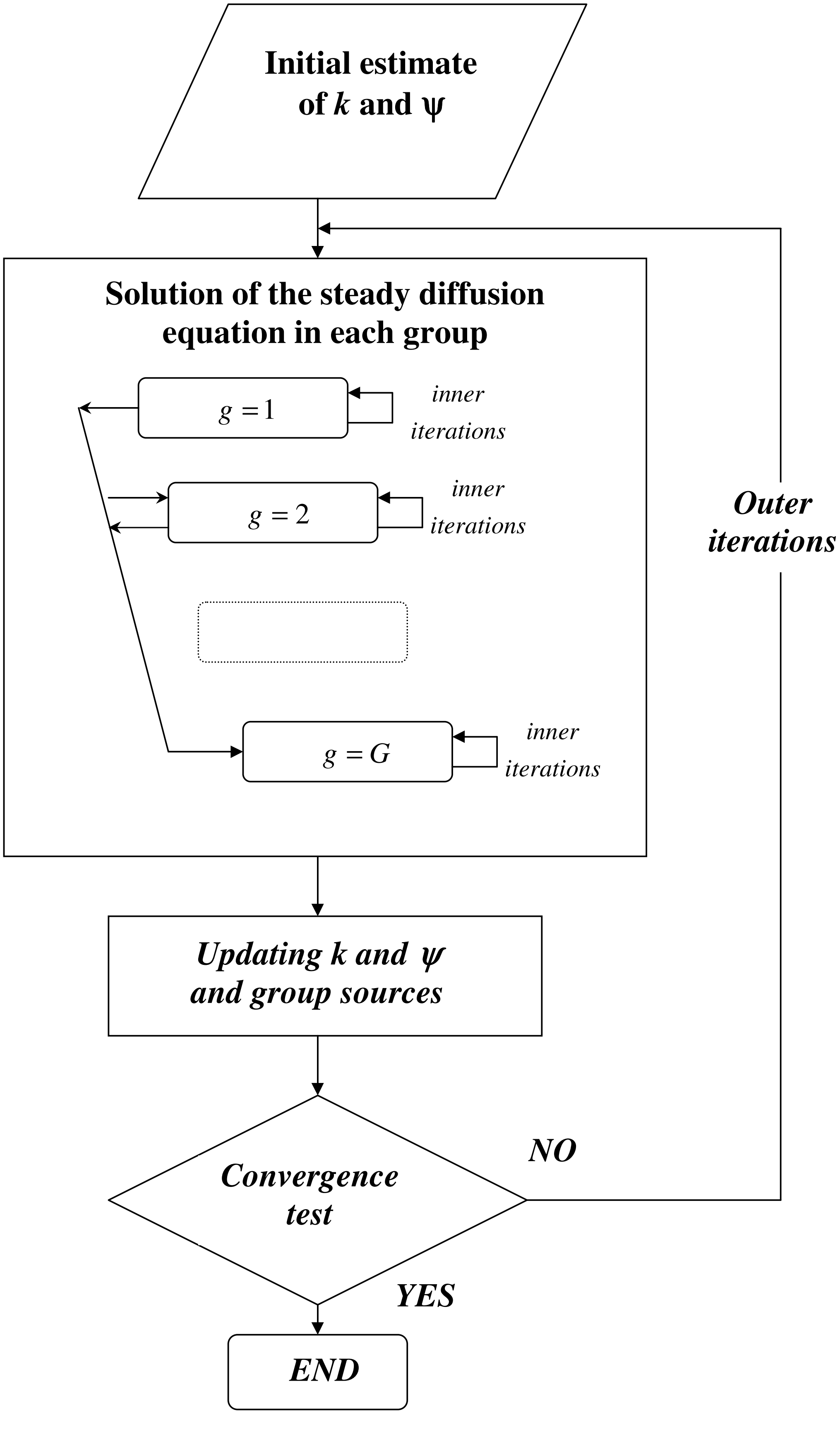
- **The scheme is general enough to be applied making use of any operator describing the neutronic behaviour, including the *transport equation*.**

  In fact, by changing the diffusion operator with the more accurate transport one, the same "lower triangular" structure of the slowing down matrix $\Sigma_{s,g'\to g}(\vec{r})$ appears, whenever a slowing down process is addressed. This allows to apply the same strategy, whenever a suitable spatial discretization of the transport operator is adopted.

- **The lower-triangular structure of the differential scattering matrix is obviously lost when *thermalisation problems* are dealt with.**

  In fact, when the energy of neutrons becomes comparable with that of target nuclei, "up-scattering" may occur in addition to "down-scattering".

  In such cases, the scattering source becomes as "coupled" as the fission source. A similar treatment of these terms can be therefore envisaged.


Initial estimate
of k and ψ
Solution of the steady diffusion
equation in each group
$g = 1$
inner
iterations
$g = 2$
inner
iterations
$g = G$
inner
iterations
Outer
iterations
Updating k and ψ
and group sources
Convergence
test
NO
YES
END

INNER ITERATIONS
FORM OF THE FINITE DIFFERENCE EQUATIONS

- **As explained above, a key point in solving eigenvalues problems is the numerical solution of group equations having the general monokinetic *non-homogeneous* form:**

$$-\,div\, D(\vec{r})grad_{\vec{r}}\phi(\vec{r})+\Sigma_a(\vec{r})\phi(\vec{r})=S(\vec{r})$$

- **In the light of numerical solution, it is important to highlight the main features of the algebraic linear systems obtained by the discretization of diffusion equations in one or more space dimensions**

One-dimensional Problems

- **In the case of the criticality problem related to a slab containing multiplicating material dealt with above, it was shown that the obtained linear system was homogeneous and had a tri-diagonal matrix; i.e., three-point equations were obtained:**

$$a_{i,i-1}\phi_{i-1}+a_{i,i}\phi_i+a_{i,i+1}\phi_{i+1}=0 \qquad (i=1,\ldots,n)$$

- **It is clearly understood that a similar discretization in the case of a fixed source (non-homogeneous) problem for a uniform property slab**

$$-D\frac{d^2\phi}{dx^2}+\Sigma_a\phi(x)=S \qquad \phi(0)=\phi(a)=0$$

provides three-point formulations

$$-D\frac{\phi_{i+1}-2\phi_i+\phi_{i-1}}{h^2}+\Sigma_a\phi_i=S_i \qquad \phi_0=\phi_{n+1}=0$$

In particolar, putting

$$a_{i,i-1}=a_{i,i+1}=-\frac{D}{h^2} \qquad a_{i,i}=\frac{2D}{h^2}+\Sigma_a$$

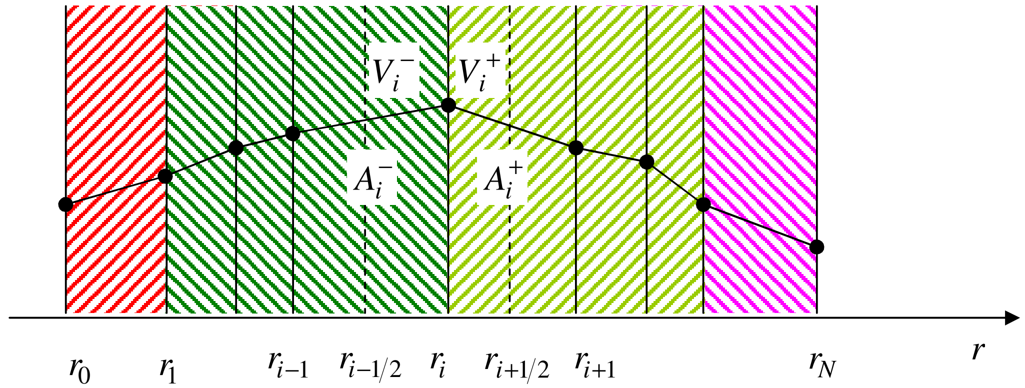
it is

$$a_{i,i-1}\phi_{i-1}+a_{i,i}\phi_i+a_{i,i+1}\phi_{i+1}=S_i \qquad (i=1,\ldots,n)$$

- **However, uniform property problems are scarcely relevant for practical applications; it is by far more interesting to deal with multi-layer problems**

EXERCISE
Let us consider *two different ways of providing a discretization of such problems by a "finite volume" technique*

1st METHOD: fluxes evaluated at the interfaces between nodes



- **We now assume the presence of material layers having different nuclear properties in 1D plane, cylindrical or spherical geometry**
- **Intervals ("nodes") are assigned such as each layer contains an integer number of them and each interval has uniform properties**
- ***Fluxes are evaluated at the interfaces between two intervals***
- **Each "flux point" is assigned a control volume $V_i = V_i^- \cup V_i^+$, made by half of the left and half of the right intervals; in general, these two halves have different nuclear properties**
- **Volumes V_i^- and V_i^+ the surfaces A_i^- e A_i^+ separating them are defined as follows:**
  - ♦ **plane geometry:** $V_i^- = r_i - r_{i-1/2}$**,** $V_i^+ = r_{i+1/2} - r_i$**,** $A_i^- = A_i^+ = 1$
  - ♦ **cylindrical geometry:** $V_i^- = \pi\left(r_i^2 - r_{i-1/2}^2\right)$**,** $V_i^+ = \pi\left(r_{i+1/2}^2 - r_i^2\right)$**,** $A_i^- = 2\pi r_{i-1/2}$**,** $A_i^+ = 2\pi r_{i+1/2}$

- ♦ **spherical geometry:** $V_i^- = \frac{4}{3}\pi\left(r_i^3 - r_{i-1/2}^3\right)$, $V_i^+ = \frac{4}{3}\pi\left(r_{i+1/2}^3 - r_i^3\right)$, $A_i^- = 4\pi\, r_{i-1/2}^2$, $A_i^+ = 4\pi\, r_{i+1/2}^2$

  in which it is $r_{i-1/2} = \left(r_{i-1} + r_i\right)/2$ **and** $r_{i+1/2} = \left(r_i + r_{i+1}\right)/2$

- **Diffusion equations are therefore integrated throughout** V_i**:**

$$-\int_{V_i} divD\; grad\,\phi\, dV + \int_{V_i} \Sigma_a \phi\, dV = \int_{V_i} S\, dV$$

  - ♦ **the Gauss (divergence) theorem is made use of**

$$\int_{V_i} divD\, grad\,\phi\, dV = \int_{A_i^+} D\frac{d\phi}{du}dA - \int_{A_i^-} D\frac{d\phi}{du}dA = D_i^+ A_i^+ \left.\frac{d\phi}{du}\right|_{r_{i+1/2}} - D_i^- A_i^- \left.\frac{d\phi}{du}\right|_{r_{i-1/2}}$$

  - ♦ **a linear distribution of the flux in each interval is assumed for purpose of approximating gradients**

$$\int_{V_i} divD\, grad\,\phi\, dV = D_i^+ A_i^+ \frac{\phi_{i+1} - \phi_i}{r_{i+1} - r_i} - D_i^- A_i^- \frac{\phi_i - \phi_{i-1}}{r_i - r_{i-1}}$$

  - ♦ **absorption and source terms are finally evaluated**

$$\int_{V_i} \Sigma_a \phi\, dV = \left(\Sigma_{a,i}^- V_i^- + \Sigma_{a,i}^+ V_i^+\right)\phi_i$$

$$\int_{V_i} S\, dV = S_i^- V_i^- + S_i^+ V_i^+$$

- **In the layers close to the boundaries 3rd kind boundary conditions are assumed**

$$\gamma_0 \phi(r_0) + \delta_0 \left.\frac{d\phi}{dr}\right|_{r_0} = 0 \qquad \gamma_N \phi(r_N) + \delta_N \left.\frac{d\phi}{dr}\right|_{r_N} = 0$$

  allowing to assume neutron flux zeroing at the physical or the extrapolated boundary by appropriate choices of the γ **and** δ **coefficients**

- **It is easy to check that also in this case three-point formulas are obtained**

$$a_{i,i-1}\phi_{i-1} + a_{i,i}\phi_i + a_{i,i+1}\phi_{i+1} = b_i \qquad (i = 0,\dots,N)$$

where the unknowns are $N+1$ since the boundary fluxes are generally non-zero.

It is, in fact:

$$-\int_{V_i} divD\; grad\,\phi\, dV + \int_{V_i} \Sigma_a \phi\, dV = \int_{V_i} S\, dV$$

$$-D_i^+ A_i^+ \frac{\phi_{i+1}-\phi_i}{r_{i+1}-r_i} + D_i^- A_i^- \frac{\phi_i - \phi_{i-1}}{r_i - r_{i-1}} + \left(\Sigma_{a,i}^- V_i^- + \Sigma_{a,i}^+ V_i^+\right)\phi_i = S_i^- V_i^- + S_i^+ V_i^+$$

$$-\frac{D_i^+ A_i^+}{r_{i+1}-r_i}\phi_{i+1} + \left[\frac{D_i^+ A_i^+}{r_{i+1}-r_i} + \frac{D_i^- A_i^-}{r_i - r_{i-1}}\right]\phi_i - \frac{D_i^- A_i^-}{r_i - r_{i-1}}\phi_{i-1}$$

$$+\left(\Sigma_{a,i}^- V_i^- + \Sigma_{a,i}^+ V_i^+\right)\phi_i = S_i^- V_i^- + S_i^+ V_i^+$$

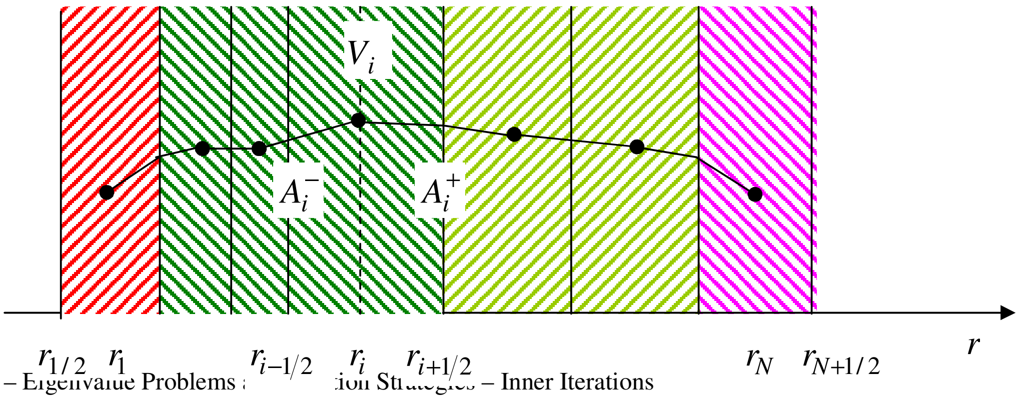
or

$$\underbrace{-\frac{D_i^- A_i^-}{r_i - r_{i-1}}}_{a_{i,i-1}}\phi_{i-1} + \underbrace{\left[\frac{D_i^+ A_i^+}{r_{i+1}-r_i} + \frac{D_i^- A_i^-}{r_i - r_{i-1}} + \Sigma_{a,i}^- V_i^- + \Sigma_{a,i}^+ V_i^+\right]}_{a_{i,i}}\phi_i \underbrace{-\frac{D_i^+ A_i^+}{r_{i+1}-r_i}}_{a_{i,i+1}}\phi_{i+1}$$

$$= \underbrace{S_i^- V_i^- + S_i^+ V_i^+}_{b_i}$$

The boundary conditions can be imposed by specializing the above equation, valid for internal nodes, to the case of "boundary" nodes.

2nd METHOD: fluxes evaluated at mid-node locations



- **Intervals ("nodes") are assigned such as each layer contains an integer number of them and each interval has uniform properties**
- **Fluxes are evaluated in the centres of each interval (nodes)**
- **Each node is assigned the volume V_i of the interval**
- **In this case, the volume V_i and the surfaces A_i^- e A_i^+ are defined by**
  - **plane geometry:** $V_i = r_{i+1/2} - r_{i-1/2}$, $A_i^- = A_i^+ = 1$
  - **cylindrical geometry:** $V_i = \pi\left(r_{i+1/2}^2 - r_{i-1/2}^2\right)$, $A_i^- = 2\pi r_{i-1/2}$, $A_i^+ = 2\pi r_{i+1/2}$
  - **spherical geometry:** $V_i = \frac{4}{3}\pi\left(r_{i+1/2}^3 - r_{i-1/2}^3\right)$, $A_i^- = 4\pi r_{i-1/2}^2$, $A_i^+ = 4\pi r_{i+1/2}^2$

  where $r_{i-1/2} = \left(r_{i-1} + r_i\right)/2$ **and** $r_{i+1/2} = \left(r_i + r_{i+1}\right)/2$
- **Diffusion equations are again integrated throughout each** V_i

$$-\int_{V_i} divD\ grad\ \phi\ dV + \int_{V_i} \Sigma_a \phi\ dV = \int_{V_i} S\ dV$$

  - **use is made of the divergence theorem**

$$\int_{V_i} divD\ grad\ \phi\ dV = \int_{A_i^+} D\frac{d\phi}{du}dA - \int_{A_i^-} D\frac{d\phi}{du}dA = D_i\ A_i^+ \left.\frac{d\phi}{du}\right|_{r_{i+1/2}} - D_i\ A_i^- \left.\frac{d\phi}{du}\right|_{r_{i-1/2}}$$

  - **<u>a linear distribution of the flux is assumed within each semi-interval connecting the node to each interface</u>**

$$\int_{V_i} divD\ grad\ \phi\ dV = D_i\ A_i^+ \frac{\phi_{i+1/2} - \phi_i}{r_{i+1/2} - r_i} - D_i\ A_i^- \frac{\phi_i - \phi_{i-1/2}}{r_i - r_{i-1/2}}$$

  - **<u>interfacial fluxes thus apperaing are eliminated by assuming the continuity of currents throuh the interfaces; e.g.</u>:**

$$-D_i \frac{\phi_{i+1/2} - \phi_i}{r_{i+1/2} - r_i} = -D_{i+1} \frac{\phi_{i+1} - \phi_{i+1/2}}{r_{i+1} - r_{i+1/2}}$$

$$\phi_{i+1/2}=\frac{\dfrac{D_i}{r_{i+1/2}-r_i}\phi_i+\dfrac{D_{i+1}}{r_{i+1}-r_{i+1/2}}\phi_{i+1}}{\dfrac{D_i}{r_{i+1/2}-r_i}+\dfrac{D_{i+1}}{r_{i+1}-r_{i+1/2}}}$$

- **absorption and source terms are then evaluated**

$$\int_{V_i}\Sigma_a\phi\,dV=\Sigma_{a,i}V_i\ \ \phi_i \qquad\qquad \int_{V_i}S\,dV=S_i\,V_i$$

- **It is then obtained**

$$-D_i\,A_i^+\frac{\phi_{i+1/2}-\phi_i}{r_{i+1/2}-r_i}+D_i\,A_i^-\frac{\phi_i-\phi_{i-1/2}}{r_i-r_{i-1/2}}+\Sigma_{a,i}V_i\ \ \phi_i=S_i\,V_i$$

that represents again a three-point formula in the nodal fluxes, if it is considered that

$$\phi_{i+1/2}=\frac{\dfrac{D_i}{r_{i+1/2}-r_i}\phi_i+\dfrac{D_{i+1}}{r_{i+1}-r_{i+1/2}}\phi_{i+1}}{\dfrac{D_i}{r_{i+1/2}-r_i}+\dfrac{D_{i+1}}{r_{i+1}-r_{i+1/2}}} \qquad \phi_{i-1/2}=\frac{\dfrac{D_{i-1}}{r_{i-1/2}-r_{i-1}}\phi_{i-1}+\dfrac{D_i}{r_i-r_{i-1/2}}\phi_i}{\dfrac{D_{i-1}}{r_{i-1/2}-r_{i-1}}+\dfrac{D_i}{r_i-r_{i-1/2}}}$$

In fact, let’s check that:

$$-D_i\,A_i^+\frac{\phi_{i+1/2}-\phi_i}{r_{i+1/2}-r_i}=-\frac{D_i\,A_i^+}{r_{i+1/2}-r_i}\left[\frac{\dfrac{D_i}{r_{i+1/2}-r_i}\phi_i+\dfrac{D_{i+1}}{r_{i+1}-r_{i+1/2}}\phi_{i+1}}{\dfrac{D_i}{r_{i+1/2}-r_i}+\dfrac{D_{i+1}}{r_{i+1}-r_{i+1/2}}}-\phi_i\right]$$

$$=-\frac{D_i\,A_i^+}{r_{i+1/2}-r_i}\left[\frac{\dfrac{D_i}{r_{i+1/2}-r_i}\phi_i+\dfrac{D_{i+1}}{r_{i+1}-r_{i+1/2}}\phi_{i+1}-\phi_i\left(\dfrac{D_i}{r_{i+1/2}-r_i}+\dfrac{D_{i+1}}{r_{i+1}-r_{i+1/2}}\right)}{\dfrac{D_i}{r_{i+1/2}-r_i}+\dfrac{D_{i+1}}{r_{i+1}-r_{i+1/2}}}\right]$$

$$=-\frac{\dfrac{D_i}{r_{i+1/2}-r_i}\,\dfrac{D_{i+1}}{r_{i+1}-r_{i+1/2}}}{\dfrac{D_i}{r_{i+1/2}-r_i}+\dfrac{D_{i+1}}{r_{i+1}-r_{i+1/2}}}A_i^+\left(\phi_{i+1}-\phi_i\right)$$

In order to understand the meaning of the above formulation, let us rewrite it as:

$$\textit{right volume current} \times \textit{surface area} = -D_i\, A_i^{+}\, \frac{\phi_{i+1/2} - \phi_i}{r_{i+1/2} - r_i}$$

$$= -\frac{\dfrac{D_i}{r_{i+1/2} - r_i}\;\dfrac{D_{i+1}}{r_{i+1} - r_{i+1/2}}}{\dfrac{D_i}{r_{i+1/2} - r_i} + \dfrac{D_{i+1}}{r_{i+1} - r_{i+1/2}}}\, A_i^{+}\left(\phi_{i+1} - \phi_i\right) = -A_i^{+}\; \underbrace{\overline{\left(\frac{D}{\Delta r}\right)}_{i,i+1}}_{\substack{\textit{weighetd averaged value of}\\ D/\Delta r \textit{ between nodes } i \textit{ and } i+1}} \left(\phi_{i+1} - \phi_i\right)$$

<u>**Suggested personal study activity**</u>
Prove that

$$D_i\, A_i^{-}\, \frac{\phi_i - \phi_{i-1/2}}{r_i - r_{i-1/2}} = \frac{\dfrac{D_{i-1}}{r_i - r_{i-1/2}}\;\dfrac{D_i}{r_{i+1/2} - r_i}}{\dfrac{D_{i-1}}{r_i - r_{i-1/2}} + \dfrac{D_i}{r_{i+1/2} - r_i}}\, A_i^{-}\left(\phi_i - \phi_{i-1}\right) = \overline{\left(\frac{D}{\Delta r}\right)}_{i-1,i}\, A_i^{-}\left(\phi_i - \phi_{i-1}\right)$$

as it can be obtained from the previous relation by an appropriate change of the indices.

Noting the above we recognise that the steady neutron balance equation in the volume

$$-D_i\, A_i^{+}\, \frac{\phi_{i+1/2} - \phi_i}{r_{i+1/2} - r_i} + D_i\, A_i^{-}\, \frac{\phi_i - \phi_{i-1/2}}{r_i - r_{i-1/2}} + \Sigma_{a,i} V_i\;\; \phi_i = S_i\, V_i$$

can be rewritten as:

$$\underbrace{-A_i^{+}\overline{\left(\frac{D}{\Delta r}\right)}_{i,i+1}\left(\phi_{i+1} - \phi_i\right)}_{\textit{leakage rate of neutrons at the right surface}} \;\; \underbrace{+A_i^{-}\overline{\left(\frac{D}{\Delta r}\right)}_{i-1,i}\left(\phi_i - \phi_{i-1}\right)}_{\textit{leakage rate of neutrons at the left surface}} \;+\; \underbrace{\Sigma_{a,i}\, V_i\, \phi_i}_{\textit{absorption rate}} \;=\; \underbrace{S_i\, V_i}_{\textit{source rate}}$$

holding for each internal point of the discretization. So, it is:

$$a_{i,i-1} = -A_i^{-}\overline{\left(\frac{D}{\Delta r}\right)}_{i-1,i} \qquad\qquad a_{i,i+1} = -A_i^{+}\overline{\left(\frac{D}{\Delta r}\right)}_{i,i+1}$$

$$a_{i,i} = A_i^{+}\overline{\left(\frac{D}{\Delta r}\right)}_{i,i+1} + A_i^{-}\overline{\left(\frac{D}{\Delta r}\right)}_{i-1,i} + \Sigma_{a,i}\, V_i \qquad\qquad b_i = S_i\, V_i$$

and

$$a_{i,i-1}\phi_{i-1} + a_{i,i}\phi_i + a_{i,i+1}\phi_{i+1} = b_i$$

In the layers close to the boundaries 3rd kind boundary conditions can be again assumed

$$\gamma_0 \phi(r_0) + \delta_0 \left.\frac{d\phi}{dr}\right|_{r_0} = 0 \qquad \gamma_N \phi(r_N) + \delta_N \left.\frac{d\phi}{dr}\right|_{r_N} = 0$$

thus closing the problem.

<u>Suggested personal study activity</u>

Try to express the formulations for different boundary conditions in the first and the last node.

Hint

Restart from the equation

$$-D_i\, A_i^+ \frac{\phi_{i+1/2} - \phi_i}{r_{i+1/2} - r_i} + D_i\, A_i^- \frac{\phi_i - \phi_{i-1/2}}{r_i - r_{i-1/2}} + \Sigma_{a,i} V_i\ \ \phi_i = S_i\, V_i$$

and assume:

a) $i = 1\ \ and\ \ \phi_{1/2} = 0$**; what does this condition mean? how does the equation for the first node look like?**

b) $i = 1\ \ and\ \ \frac{\phi_1 - \phi_{1/2}}{r_1 - r_{1/2}} = 0$**; what does this condition mean? how does the equation for the first node look like?**

c) $i = 1\ \ and\ \ \delta_{1/2} \frac{\phi_1 - \phi_{1/2}}{r_1 - r_{1/2}} + \gamma \phi_{1/2} = 0$**; what does this condition mean? how does the equation for the first node look like?**

Further Hint

For the first node the equation is:

$$-D_1\, A_1^+ \frac{\phi_{3/2} - \phi_1}{r_{3/2} - r_1} + D_1\, A_1^- \frac{\phi_1 - \phi_{1/2}}{r_1 - r_{1/2}} + \Sigma_{a,1} V_1\ \ \phi_1 = S_1\, V_1$$

or also (remember the above developments)

$$-A_i^+ \overline{\left(\frac{D}{\Delta r}\right)}_{1,2} (\phi_2 - \phi_1) + D_1\, A_1^- \frac{\phi_1 - \phi_{1/2}}{r_1 - r_{1/2}} + \Sigma_{a,1} V_1\ \ \phi_1 = S_1\, V_1$$

So, work on this form. You must end up with a formulation that has the form:

$$a_{1,1}\phi_1 + a_{1,2}\phi_2 = b_1$$

Repeat the above for the last node $i = N$ **with *N* the number of intervals.**

Suggested personal study activity

Try showing that in case of Cartesian coordinates, constant diffusion coefficient and uniform spatial discretization the equation

$$\underbrace{-A_i^{+}\overline{\left(\frac{D}{\Delta r}\right)}_{i,i+1}\left(\phi_{i+1}-\phi_i\right)}_{\textit{leakage rate of neutrons at the right surface}}\ \underbrace{+A_i^{-}\overline{\left(\frac{D}{\Delta r}\right)}_{i-1,i}\left(\phi_i-\phi_{i-1}\right)}_{\textit{leakage rate of neutrons at the left surface}}\ +\ \underbrace{\Sigma_{a,i}\,V_i\,\phi_i}_{\textit{absorption rate}}\ =\ \underbrace{S_i\,V_i}_{\textit{source rate}}$$

is equivalent to the "finite difference" approximation

$$-D\frac{\phi_{i+1}-2\phi_i+\phi_{i-1}}{h^2}+\Sigma_{a,i}\phi_i=S_i$$

<u>NB:</u> These discretization techniques can be used also for other physical problems (e.g., for heat conduction, electrostatic potential) having an elliptic character

$$-div\ k(\vec{r})\ grad_{\vec{r}}T(\vec{r})=q'''(\vec{r})$$

$$div\ \varepsilon(\vec{r})\ grad_{\vec{r}}V(\vec{r})=0$$

As a general conclusion:

The finite volume discretization of steady-state diffusion equations in 1D leads to linear algebraic systems having a three-diagonal matrix

This conclusion can be extended to some discretization schemes other than finite volumes, e.g., to "coarse-mesh" methods to be described later on

SOLUTION OF THREE-DIAGONAL MATRIX SYSTEMS

- **The solution of TDM systems is particularly simple by an algorithm being a special case of the Gauss elimination scheme.**
- **The matrix is represented in the form (e.g., for $N = 8$):**

$$\begin{bmatrix} v_1 & w_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ u_2 & v_2 & w_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & u_3 & v_3 & w_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & u_4 & v_4 & w_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & u_5 & v_5 & w_5 & 0 & 0 \\ 0 & 0 & 0 & 0 & u_6 & v_6 & w_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & u_7 & v_7 & w_7 \\ 0 & 0 & 0 & 0 & 0 & 0 & u_8 & v_8 \end{bmatrix} \cdot \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \\ \phi_7 \\ \phi_8 \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \\ S_7 \\ S_8 \end{bmatrix}$$

putting

$$u_i = a_{i,i-1} \qquad v_i = a_{i,i} \qquad w_i = a_{i,i+1}$$

The following *two-sweeps* technique is adopted known as *Thomas' algorithm (or TDMA=Three-Diagonal Matrix Algorithm)*

1. with increasing index (first sweep):

$$\alpha_1 = \frac{w_1}{v_1} \qquad \alpha_i = \frac{w_i}{v_i - u_i \alpha_{i-1}} \qquad (i = 2, ..., N-1)$$

$$\beta_1 = \frac{S_1}{v_1} \qquad \beta_i = \frac{S_i - u_i \beta_{i-1}}{v_i - u_i \alpha_{i-1}} \qquad (i = 2, ..., N)$$

2. with decreasing index (second sweep):

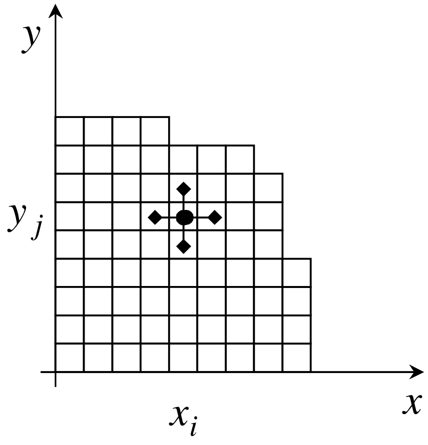
$$\phi_N = \beta_N \qquad \phi_i = \beta_i - \alpha_i \phi_{i+1}$$

In particular:

- **the first sweep eliminates an unknown in the equations;**
- **the second is needed to solve the lower triangular system thus obtained.**

Multidimensional Problems

- Let us consider the quarter reactor lattice represented in the figure



- It is discretised through a 2D grid with rectangular (or square) meshes

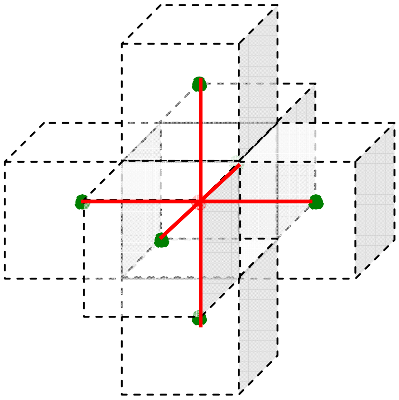
- If the reactor is homogeneous, the steady diffusion equation has the form

$$-D\left(\frac{d^2\phi}{dx^2}+\frac{d^2\phi}{dy^2}\right)+\Sigma_a\phi=S$$

- By a simple finite-difference equation, it is

$$-D\left(\frac{\phi_{i+1,j}-2\phi_{i,j}+\phi_{i-1,j}}{h_x^2}+\frac{\phi_{i,j+1}-2\phi_{i,j}+\phi_{i,j-1}}{h_y^2}\right)+\Sigma_a\phi_{i,j}=S_{i,j}$$

  clearly representing *a 5 point formulation*



- Similarly, for a three-dimensional problem

$$-D\left(\frac{d^2\phi}{dx^2}+\frac{d^2\phi}{dy^2}+\frac{d^2\phi}{dz^2}\right)+\Sigma_a\phi=S$$

  a 7 point formula is obtained

$$-D\left(\frac{\phi_{i+1,j,k}-2\phi_{i,j,k}+\phi_{i-1,j,k}}{h_x^2}\right.$$

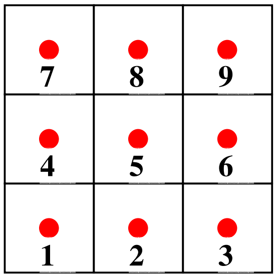
$$\left.+\frac{\phi_{i,j+1,k}-2\phi_{i,j,k}+\phi_{i,j-1,k}}{h_y^2}+\frac{\phi_{i,j,k+1}-2\phi_{i,j,k}+\phi_{i,j,k-1}}{h_z^2}\right)+\Sigma_a\phi_{i,j,k}=S_{i,j,k}$$

- For non-uniform properties, similar techniques as the ones shown for the 1D case can be applied

- In fact, though the above formulations are related to the case of uniform properties, it is quite clear that non-uniformity does not change the structure of the involved matrices

CHARACTERISTICS OF THE OBTAINED LINEAR SYSTEM MATRICES

- We just noted that in 1D cases three-diagonal matrices are obtained, to be efficiently dealt with by the *Thomas algorithm*



- In multi-dimensional cases, 5 or seven point formulas are instead obtained; the single-index numbering adopted for the grid nodes determines the structure of the matrix
- For instance, in the case of a homogeneous grid with 9 nodes numbered as in the figure, a following *banded matrix* system is obtained

$$\begin{bmatrix} a_{11} & a_{12} & 0 & a_{14} & 0 & 0 & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & a_{25} & 0 & 0 & 0 & 0 \\ 0 & a_{32} & a_{33} & 0 & 0 & a_{36} & 0 & 0 & 0 \\ a_{41} & 0 & 0 & a_{44} & a_{45} & 0 & a_{47} & 0 & 0 \\ 0 & a_{52} & 0 & a_{54} & a_{55} & a_{56} & 0 & a_{58} & 0 \\ 0 & 0 & a_{63} & 0 & a_{65} & a_{66} & 0 & 0 & a_{69} \\ 0 & 0 & 0 & a_{74} & 0 & 0 & a_{77} & a_{78} & 0 \\ 0 & 0 & 0 & 0 & a_{85} & 0 & a_{87} & a_{88} & a_{89} \\ 0 & 0 & 0 & 0 & 0 & a_{96} & 0 & a_{98} & a_{99} \end{bmatrix} \cdot \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \\ \phi_7 \\ \phi_8 \\ \phi_9 \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \\ S_7 \\ S_8 \\ S_9 \end{bmatrix}$$

where

$$a_{i,i} = \Sigma_a + \frac{4D}{h^2} \qquad\qquad a_{i,j} = -\frac{D}{h^2} \qquad\qquad (i \neq j)$$

In fact, from the notation with two indices

$$-D\left(\frac{\phi_{i+1,j} - 2\phi_{i,j} + \phi_{i-1,j}}{h_x^2} + \frac{\phi_{i,j+1} - 2\phi_{i,j} + \phi_{i,j-1}}{h_y^2} \right) + \Sigma_a \phi_{i,j} = S_{i,j}$$

reordering with a single index and assuming $h_x = h_y = h$ as in the figure, e.g. the equation for the node number 5 becomes

$$-D\left(\frac{\phi_6 - 2\phi_5 + \phi_4}{h^2} + \frac{\phi_8 - 2\phi_5 + \phi_2}{h^2} \right) + \Sigma_a \phi_5 = S_5$$

NB: We must easily understand shift from multiple index to single index numbering of the unknown variable.

It can be therefore noted that matrices obtained in 2D and 3D cases are *band matrices*. Moreover, the following general properties are obtained by most of the discretization procedures:

- ♦ $a_{i,i} > 0 \qquad (i = 1,...,N) \quad \Rightarrow$ ***the main diagonal is non-zero*** **(*1*)**
- ♦ $a_{i,j} = a_{j,i} \leq 0 \qquad i \neq j \quad \Rightarrow$ ***the matrix is symmetric and off-diagonal terms have opposite sign with respect to diagonal ones***
- ♦ $a_{i,i} > \sum_{i \neq j} {}_j \left| a_{i,j} \right| \qquad (i = 1,...,N) \quad \Rightarrow$ ***the matrix is diagonally dominant*** **(note that this occurs if** $\Sigma_a \neq 0$**!!!)**

[1] Obviously, a particular sign of the equations is assumed to write the inequalities so we obtained positive diagonal terms. Reversing the sign does not change the meaning of the above conclusions.

REMINDER ABOUT THE METHODS FOR SOLVING SYSTEMS OF LINEAR ALGEBRAIC EQUATIONS

- **The Cramer rule represents an inefficient method for solving linear algebraic equation systems, involving too many operations**
- **As known, other methods are therefore adopted, subdivided into two main categories**
  - ♦ ***Direct Methods:*** **they provide the virtually exact solution (only round-off error comes into play) with a finite number of operations**
  - ♦ ***Iterative Methods*****: they provide the solution as a limit a successive approximations**
- **In neutronic problems, as well as in many other engineering problems, the latter are the most frequently adopted ones: in our case, this is where *"internal iterations"* get their name from**
- **Let us consider an algebraic system specified in the form:** $A\phi = s$

Direct Methods

- ***Gauss or successive elimination method***
  - ♦ **By this method, equations are linearly combined in order to obtain an equivalent linear system having an *upper triangular matrix*, whose solution is straightforward**

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix} \cdot \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \end{bmatrix} \rightarrow \begin{bmatrix} \hat{a}_{11} & \hat{a}_{12} & \hat{a}_{13} & \hat{a}_{14} & \hat{a}_{15} \\ 0 & \hat{a}_{22} & \hat{a}_{23} & \hat{a}_{24} & \hat{a}_{25} \\ 0 & 0 & \hat{a}_{33} & \hat{a}_{34} & \hat{a}_{35} \\ 0 & 0 & 0 & \hat{a}_{44} & \hat{a}_{45} \\ 0 & 0 & 0 & 0 & \hat{a}_{55} \end{bmatrix} \cdot \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{bmatrix} = \begin{bmatrix} \hat{S}_1 \\ \hat{S}_2 \\ \hat{S}_3 \\ \hat{S}_4 \\ \hat{S}_5 \end{bmatrix}$$

  - ♦ **The solution of the obtained system is immediate, since it is just needed to solve the last (trivial) equation and to back-substitute in the previous ones**
  - ♦ **The steps to be followed in this respect are:**

1. **the first equation is added side by side to each subsequent one, after multiplying it by appropriate factors; in particular, for the *r*-th equation ($r \geq 2$), the factor is $-\dfrac{a_{r1}}{a_{11}}$; it is:**

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix} \cdot \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \end{bmatrix} \rightarrow \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a'_{22} & a'_{23} & a'_{24} & a'_{25} \\ 0 & a'_{32} & a'_{33} & a'_{34} & a'_{35} \\ 0 & a'_{42} & a'_{43} & a'_{44} & a'_{45} \\ 0 & a'_{52} & a'_{53} & a'_{54} & a'_{55} \end{bmatrix} \cdot \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{bmatrix} = \begin{bmatrix} S_1 \\ S'_2 \\ S'_3 \\ S'_4 \\ S'_5 \end{bmatrix}$$

2. **then, the second equation is added side by side to all the subsequent ones ($r \geq 3$) after multiplying it by $-\dfrac{a'_{r2}}{a'_{22}}$:**

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a'_{22} & a'_{23} & a'_{24} & a'_{25} \\ 0 & a'_{32} & a'_{33} & a'_{34} & a'_{35} \\ 0 & a'_{42} & a'_{43} & a'_{44} & a'_{45} \\ 0 & a'_{52} & a'_{53} & a'_{54} & a'_{55} \end{bmatrix} \cdot \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{bmatrix} = \begin{bmatrix} S_1 \\ S'_2 \\ S'_3 \\ S'_4 \\ S'_5 \end{bmatrix} \rightarrow \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a'_{22} & a'_{23} & a'_{24} & a'_{25} \\ 0 & 0 & a''_{33} & a''_{34} & a''_{35} \\ 0 & 0 & a''_{43} & a''_{44} & a''_{45} \\ 0 & 0 & a''_{53} & a''_{54} & a''_{55} \end{bmatrix} \cdot \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{bmatrix} = \begin{bmatrix} S_1 \\ S'_2 \\ S''_3 \\ S''_4 \\ S''_5 \end{bmatrix}$$

3. **a similar procedure is adopted up to the $(N-1)$-th equation, thus obtaining an upper triangular system matrix.**

♦ **In general, in order to reduce round-off errors, it is possible to change the order of equations in such a way that:**

$$|a_{11}| \geq |a_{r1}| \quad (r = 2,3,..) \qquad |a'_{22}| \geq |a'_{r2}| \quad (r = 3,4..) \quad \textbf{etc.}$$

a technique referred to as *partial pivoting*

♦ **As previously mentioned, *the Thomas algorithm represents a the particular formulation of the Gauss method for the simpler cases of three-diagonal matrix systems***

- ***Gauss-Jordan Method***

  - **In this case, equations are linearly combined in order to obtain *a diagonal matrix system***

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix} \cdot \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \end{bmatrix} \to \begin{bmatrix} \hat{a}_{11} & 0 & 0 & 0 & 0 \\ 0 & \hat{a}_{22} & 0 & 0 & 0 \\ 0 & 0 & \hat{a}_{33} & 0 & 0 \\ 0 & 0 & 0 & \hat{a}_{44} & 0 \\ 0 & 0 & 0 & 0 & \hat{a}_{55} \end{bmatrix} \cdot \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{bmatrix} = \begin{bmatrix} \hat{S}_1 \\ \hat{S}_2 \\ \hat{S}_3 \\ \hat{S}_4 \\ \hat{S}_5 \end{bmatrix}$$

  - **The procedure is similar to the one of Gauss method, but each equation is added to "all" the others (both preceding and following ones) *after multiplications by appropriate factors***

  - **Also in this case "pivoting" techniques can be used**

- ***Factorisation Methods***

  - **The Gauss method can be considered as a particular case of triangular factorisation methods**

  - **In general, by factorisation methods**

    - **the matrix $\mathbf{A}$ is factored in the product of a lower triangular matrix, $\mathbf{L}$, by an upper triangular one, $\mathbf{U}$**

$$\mathbf{A} = \mathbf{L}\mathbf{U}$$

    - **the solution of the given system is then reduced to the solution of two system with triangular matrix (that is immediate)**

$$\mathbf{L}\mathbf{U}\boldsymbol{\phi} = \mathbf{s} \qquad \Rightarrow \qquad \mathbf{L}\boldsymbol{\psi} = \mathbf{s} \qquad \mathbf{U}\boldsymbol{\phi} = \boldsymbol{\psi}$$

  The difficulty of the application of the method is therefore reduced to the step of factorisation, for which different algorithms (e.g., Doolittle, Crout, Cholesky) are available in classical textbooks ([2])

[2] See, e.g., A. Quarteroni, R. Sacco, F. Saleri, Numerical Mathematics, Springer, 2007.

Iterative Methods

- ***General Considerations***

A general iterative method for solving linear algebraic systems has the form

$$\boldsymbol{\phi}^{(k)} = \mathbf{H}\,\boldsymbol{\phi}^{(k-1)} + \mathbf{c} \qquad\qquad (k = 1, 2, \ldots)$$

where $\mathbf{H}$ is termed the iteration matrix and $\mathbf{c}$ is an appropriate vector, whose definition characterise the specific numerical scheme. If $\boldsymbol{\phi}$ is the exact solution of the system $\mathbf{A}\boldsymbol{\phi} = \mathbf{s}$, a *convergence condition* of the numerical scheme will be

$$\boldsymbol{\phi} = \mathbf{H}\,\boldsymbol{\phi} + \mathbf{c}$$

By subtracting the two latter relationships it is

$$\boldsymbol{\phi}^{(k)} - \boldsymbol{\phi} = \mathbf{H}\left[\boldsymbol{\phi}^{(k-1)} - \boldsymbol{\phi}\right] \qquad\qquad (k = 1, 2, \ldots)$$

and, introducing the error vector $\boldsymbol{\varepsilon}^{(k)} = \boldsymbol{\phi}^{(k)} - \boldsymbol{\phi}$, it is

$$\boldsymbol{\varepsilon}^{(k)} = \mathbf{H}\,\boldsymbol{\varepsilon}^{(k-1)} \qquad\qquad (k = 1, 2, \ldots)$$

and then

$$\boldsymbol{\varepsilon}^{(k)} = \mathbf{H}^{k}\boldsymbol{\varepsilon}^{(0)} \qquad\qquad (k = 1, 2, \ldots)$$

Making use of vector and matrix norms (see below), it is

$$\left\|\boldsymbol{\varepsilon}^{(k)}\right\| \le \left\|\mathbf{H}^{k}\right\|\left\|\boldsymbol{\varepsilon}^{(0)}\right\| \qquad\qquad (k = 1, 2, \ldots)$$

Necessary and sufficient condition for the convergence of the iterative scheme, i.e. in order to be

$$\lim_{k\to\infty}\left\|\boldsymbol{\varepsilon}^{(k)}\right\| = 0$$

***for any* $\boldsymbol{\varepsilon}^{(0)}$ is that**

$$\lim_{k\to\infty}\left\|\mathbf{H}^{k}\right\| = 0$$

that is sometimes expressed by saying that the iteration matrix must be convergent.

Definition. ***A matrix $\mathbf{H} \in \mathbb{C}_{n\times n}$ is said to be convergent if***

$$\lim_{k\to\infty} \mathbf{H}^k = \mathbf{0}$$

Definition. The spectral radius $\rho(\mathbf{H})$ of a matrix is the maximum modulus of any of its eigenvalues.

In our purposes it is sufficient to remember that:

Theorem: A necessary and sufficient condition for a matrix $\mathbf{H} \in \mathbb{C}_{n\times n}$ to be convergent is that $\rho(\mathbf{H})<1$.

> ***Therefore, in order to apply with success an iterative method it is necessary and sufficient that the iteration matrix has spectral radius $\rho(\mathbf{H})<1$.***

- ***Reminder about norms and spectral radius***

Definition: A vector norm is an application $n:\mathbb{C}_n \to \mathbb{R}$ such that, given two vectors $\mathbf{v}$ e $\mathbf{w} \in \mathbb{C}_n$ and $\alpha \in \mathbb{C}$, it is

$$n(\mathbf{v}) \geq 0,\ e\ n(\mathbf{v})=0 \ \textit{if and only if}\ \mathbf{v}=\mathbf{0}$$
$$n(\alpha\mathbf{v}) = |\alpha| n(\mathbf{v})$$
$$n(\mathbf{v}+\mathbf{w}) \leq n(\mathbf{v})+n(\mathbf{w})$$

The most common vector norms are:

♦ $\|\mathbf{v}\|_\infty = \max_i |v_i|$ **(L^∞-norm or also *maximum norm*)**

♦ $\|\mathbf{v}\|_1 = \sum_i |v_i|$ **(L^1-norm or also *absolute norm*)**

♦ $\|\mathbf{v}\|_2 = |\mathbf{v}|$ **(L^2-norm or also *Euclidean norm*)**

These norms can be formally obtained by putting $p=\infty, 1, 2$ in the general formulation

$$\|\mathbf{v}\|_p = \left(\sum_i |v_i|^p\right)^{1/p}$$

Given a vector norm, $\|\circ\|_p$, an application from $\mathbb{C}_{n\times n}$ to $\mathbb{R}$ is said to be a *compatible* (or *induced* or *subordinate*) *matrix norm* if it satisfies

$$\|\mathbf{H}\|_p = \sup_{\mathbf{v}\neq 0} \frac{\|\mathbf{H}\,\mathbf{v}\|_p}{\|\mathbf{v}\|_p}$$

It can be shown that the three norms that are compatible with (or are induced from or are subordinate to) the three above considered vector norms are

- ♦ $\|\mathbf{H}\|_\infty = \max_i \sum_k |h_{i,k}|$ ***maximum absolute row sum norm***
- ♦ $\|\mathbf{H}\|_1 = \max_k \sum_i |h_{i,k}|$ ***maximum absolute column sum norm***
- ♦ $\|\mathbf{H}\|_2 = \sqrt{\rho\left(\overline{\mathbf{H}}^T \mathbf{H}\right)}$ ***spectral norm***

These norms are also said *natural norms*. Some properties of the matrix norms are shortly recalled:

- ♦ $\|\mathbf{H}\| \geq 0 \quad \left(\|\mathbf{H}\| = 0 \Leftrightarrow \mathbf{H} = \mathbf{0}\right)$
- ♦ $\|\alpha\,\mathbf{H}\| = |\alpha|\,\|\mathbf{H}\|$
- ♦ $\|\mathbf{H}+\mathbf{J}\| \leq \|\mathbf{H}\| + \|\mathbf{J}\|$ $\qquad \|\mathbf{H}\cdot\mathbf{J}\| \leq \|\mathbf{H}\|\,\|\mathbf{J}\|$ $\qquad \|\mathbf{H}\mathbf{v}\| \leq \|\mathbf{H}\|\,\|\mathbf{v}\|$

In our purposes it is interesting to remember the following

Theorem (by Hirsch):

For any of the three natural norms it is

$$\rho(\mathbf{H}) \leq \|\mathbf{H}\|.$$

In fact, as a consequence of this theorem, it is:

A sufficient condition for a matrix $\mathbf{H}\in\mathbb{C}_{n\times n}$ to be convergent is that any of its natural norms be less than unity.

The Jacobi Method

To obtain the iterative scheme, the system matrix is decomposed in a diagonal part and an off-diagonal one:

$$\mathbf{A} \qquad = \qquad \mathbf{D} \qquad - \qquad \mathbf{E}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \circ & a_{1n} \\ a_{21} & a_{22} & a_{23} & \circ & a_{2n} \\ a_{31} & a_{32} & a_{33} & \circ & a_{3n} \\ \circ & \circ & \circ & \circ & \circ \\ a_{n1} & a_{n2} & a_{n3} & \circ & a_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & 0 & 0 & \circ & 0 \\ 0 & a_{22} & 0 & \circ & 0 \\ 0 & 0 & a_{33} & \circ & 0 \\ \circ & \circ & \circ & \circ & \circ \\ 0 & 0 & 0 & \circ & a_{nn} \end{bmatrix} - \begin{bmatrix} 0 & -a_{12} & -a_{13} & \circ & -a_{1n} \\ -a_{21} & 0 & -a_{23} & \circ & -a_{2n} \\ -a_{31} & -a_{32} & 0 & \circ & -a_{3n} \\ \circ & \circ & \circ & \circ & \circ \\ -a_{n1} & -a_{n2} & -a_{n3} & \circ & 0 \end{bmatrix}$$

The linear system is then rewritten as:

$$\mathbf{D}\boldsymbol{\phi} = \mathbf{E}\boldsymbol{\phi} + \mathbf{s}$$

It must be noted that, since in our cases it is always $a_{ii} \neq 0$, it is possible to calculate $\mathbf{D}^{-1}$, being the diagonal matrix having non-zero elements equal to the reciprocal of the non-zero elements of $\mathbf{D}$. As a consequence, it is:

$$\boldsymbol{\phi} = \mathbf{D}^{-1}\mathbf{E}\boldsymbol{\phi} + \mathbf{D}^{-1}\,\mathbf{s}$$

and, putting

$$\mathbf{H_J} = \mathbf{D}^{-1}\mathbf{E} \qquad\qquad \mathbf{q} = \mathbf{D}^{-1}\,\mathbf{s}$$

it is

$$\boldsymbol{\phi} = \mathbf{H_J}\,\boldsymbol{\phi} + \mathbf{q}$$

It can be recognised that it is $b_{ii} = 0$ and

$$\mathbf{H_J} = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \circ & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \circ & -\frac{a_{2n}}{a_{22}} \\ -\frac{a_{31}}{a_{33}} & -\frac{a_{32}}{a_{33}} & 0 & \circ & -\frac{a_{3n}}{a_{33}} \\ \circ & \circ & \circ & \circ & \circ \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & -\frac{a_{n3}}{a_{nn}} & \circ & 0 \end{bmatrix}$$

When A is diagonally dominant, it is therefore:

$$\sum_{j=1}^{n} \left| H_{J,ij} \right| < 1 \quad (i = 1,\ldots,n) \qquad \Rightarrow \qquad \left\| \mathbf{H_J} \right\|_{\infty} < 1 \qquad \Rightarrow \qquad \rho(\mathbf{H_J}) < 1$$

It is then straightforward defining the iterative process:

$$\phi^{(1)} = \mathbf{q} \qquad\qquad \phi^{(m+1)} = \mathbf{H_J}\,\phi^{(m)} + \mathbf{q}$$

By these definitions, it is:

$$\begin{aligned}
&\phi^{(1)} = \mathbf{q}\\
&\phi^{(2)} = \mathbf{H_J}\,\mathbf{q} + \mathbf{q}\\
&\phi^{(3)} = \mathbf{H_J}^{2}\,\mathbf{q} + \mathbf{H_J}\mathbf{q} + \mathbf{q}\\
&\circ\\
&\phi^{(m+1)} = \mathbf{H_J}^{m}\,\mathbf{q} + \mathbf{H_J}^{m-1}\,\mathbf{q} + \circ\circ\circ + \mathbf{H_J}\,\mathbf{q} + \mathbf{q}
\end{aligned}$$

For the above *von Neumann series* it can be recognised that

$$\lim_{m\to\infty} \phi^{(m+1)} = \left(\mathbf{I} - \mathbf{H_J}\right)^{-1} \mathbf{q} = \phi \quad \textit{where} \;\; \phi = \mathbf{H_J}\,\phi + \mathbf{q}$$

A simple proof of the above can be found in similarity with the treatment of scalar geometrical series considering that

$$\left(\mathbf{I} - \mathbf{H_J}\right)\phi^{(m+1)} = \left(\mathbf{I} - \mathbf{H_J}\right)\left[\mathbf{H_J}^{m}\,\mathbf{q} + \mathbf{H_J}^{m-1}\,\mathbf{q} + \circ\circ\circ + \mathbf{H_J}\,\mathbf{q} + \mathbf{q}\right]$$

Manipulating the RHS it can be easily found that

$$\begin{aligned}
&\left(\mathbf{I} - \mathbf{H_J}\right)\left[\mathbf{H_J}^{m}\,\mathbf{q} + \mathbf{H_J}^{m-1}\,\mathbf{q} + \circ\circ\circ + \mathbf{H_J}\,\mathbf{q} + \mathbf{q}\right]\\
&= \left[\mathbf{H_J}^{m}\,\mathbf{q} + \mathbf{H_J}^{m-1}\,\mathbf{q} + \circ\circ\circ + \mathbf{H_J}\,\mathbf{q} + \mathbf{q}\right]\\
&- \left[\mathbf{H_J}^{m+1}\,\mathbf{q} + \mathbf{H_J}^{m}\,\mathbf{q} + \circ\circ\circ + \mathbf{H_J}^{2}\,\mathbf{q} + \mathbf{q}\mathbf{H_J}\right]\\
&= \mathbf{q} - \mathbf{H_J}^{m+1}\,\mathbf{q} = \left(\mathbf{I} - \mathbf{H_J}^{m+1}\right)
\end{aligned}$$

and then

$$\left(\mathbf{I} - \mathbf{H_J}\right)\phi^{(m+1)} = \left(\mathbf{I} - \mathbf{H_J}^{m+1}\right)\mathbf{q}$$

or

$$\phi^{(m+1)} = \left(\mathbf{I} - \mathbf{H_J}\right)^{-1}\left(\mathbf{I} - \mathbf{H_J}^{m+1}\right)\mathbf{q}$$

Since $\mathbf{H_J}$ is a convergent matrix, it follows

$$\lim_{m\to\infty} \phi^{(m+1)} = \left(\mathbf{I} - \mathbf{H_J}\right)^{-1} \mathbf{q}$$

implying that

$$\lim_{m\to\infty} \phi^{(m+1)} = \left(\mathbf{I} - \mathbf{H_J}\right)^{-1} \mathbf{q} = \phi = \textit{exact solution}$$

since only for the exact solution it may be:

$$\phi = \mathbf{H_J}\,\phi + \mathbf{q}$$

NOTE:

The above arguments, together with the previously mentioned results on the convergence of general iterative schemes for the solution of linear algebraic systems, are enough for stating that the Jacobi method converges for diagonally dominant system matrices.

This is also enough to state that the Jacobi method converges for the linear systems obtained by the common space discretisations of the diffusion equation.

However, since we accepted most of the results without a rigorous proof, in order to better understand the underlying principles, the treatment of a special case will be easy and instructive.

<u>Case in which $\mathbf{H_J}$ is symmetric</u>

Let us consider the case in which $\mathbf{H_J}$ is a real and symmetric matrix. This is e.g. the case of the uniform property reactor with a uniform spatial discretisation.

This case is particularly simple, since it is recalled that *for any real symmetric matrix* (lets' call it A) an *orthogonal matrix* U always exists (i.e., a matrix whose transpose is equal to the inverse matrix, $\mathbf{U}^T = \mathbf{U}^{-1}$) such that

$$\mathbf{AU} = \mathbf{U}\ diag\{\lambda_1, \lambda_2, ..., \lambda_n\}$$

As a consequence, it can be inferred that:

- **the columns of U are eigenvectors of A, corresponding to the eigenvalues λ_i;**
- **these eigenvectors represent an *orthonormal basis* of $\mathbb{R}^n$.**

Such an orthonormal basis will be identified in the following as

$$\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, ..., \boldsymbol{\varphi}_n$$

With the above assumptions, the following can be shown.

<u>Theorem 1.</u> *Necessary and sufficient condition for convergence of the Jacobi method is that* $\rho(\mathbf{H_J}) < 1$

<u>Dem.:</u> Defining the error vector as

$$\boldsymbol{\varepsilon}^{(m)} = \boldsymbol{\phi}^{(m)} - \boldsymbol{\phi}$$

it is (see above)

$$\boldsymbol{\varepsilon}^{(m+1)} = \mathbf{H_J}^{m}\,\boldsymbol{\varepsilon}^{(1)}$$

If $\mathbf{H_J}$ is symmetric, its eigenvalues generate all $\mathbb{R}^n$. Therefore

$$\boldsymbol{\varepsilon}^{(1)} = \sum_{h=1}^{n} c_h^{(1)} \boldsymbol{\varphi}_h$$

It follows that

$$\boldsymbol{\varepsilon}^{(m+1)} = \mathbf{H_J}^{m}\left(\sum_{h=1}^{n} c_h^{(1)} \boldsymbol{\varphi}_h\right) = \sum_{h=1}^{n} c_h^{(1)} \mathbf{H_J}^{m} \boldsymbol{\varphi}_h = \sum_{h=1}^{n} c_h^{(1)} \lambda_h^{m} \boldsymbol{\varphi}_h$$

For any arbitrary set of $c_h^{(1)}$ it is therefore $\lim_{m\to\infty} \boldsymbol{\varepsilon}^{(m+1)} = 0$ *if and only if* $|\lambda_h| < 1$ for any *h*, i.e., *if and only if* $\rho(\mathbf{H_J}) < 1$.

NOTE:

Speaking in gross intuitive terms, the above means that the requirement $\rho(\mathbf{H_J}) < 1$ fulfils the request that $\mathbf{H_J}$ is an operator that *contracts* any vector to which it is applied; in particular it contracts the error vector.

<u>Summarising, it is:</u>

The strong (or strict) diagonal dominance of A
is a sufficient condition for the convergence of the Jacobi method.

It can be also shown that:

The weak diagonal dominance
and the irreducibility of the matrix A
are sufficient conditions for the Jacobi method convergence

Before commenting this result, it is recalled that:

<u>Def</u>.: ***A matrix is said weakly diagonally dominant if*** $\sum_{j\neq i}{}_{j} \left|a_{i,j}\right| \le \left|a_{i,i}\right|$***, but at least for a single value of*** r ***it is*** $\sum_{j\neq r}{}_{j} \left|a_{r,j}\right| < \left|a_{r,r}\right|$**.**

<u>Def</u>.: ***A $n \times n$ matrix is said to be reducible if and only if a permutation matrix P exists (i.e., a matrix that can be obtained by the identity matrix***

by permutation of the columns) such that $\mathbf{P}^T\mathbf{AP}$ ***is block triangular (or block diagonal as a special case).***

Def.: ***An irreducible matrix is not reducible.***

NOTE:

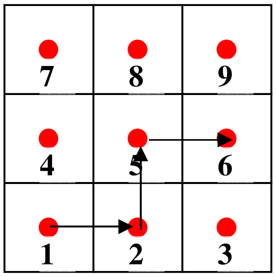
Let us comment this results on an intuitive physico-numerical basis. A reducible matrix is such that:

$$\mathbf{P}^T\mathbf{AP} = \begin{bmatrix} \hat{\mathbf{A}}_{11} & 0 & 0 & \circ & 0 \\ \hat{\mathbf{A}}_{21} & \hat{\mathbf{A}}_{22} & 0 & \circ & 0 \\ \hat{\mathbf{A}}_{31} & \hat{\mathbf{A}}_{32} & \hat{\mathbf{A}}_{33} & \circ & 0 \\ \circ & \circ & \circ & \circ & \circ \\ \hat{\mathbf{A}}_{k1} & \hat{\mathbf{A}}_{k2} & \hat{\mathbf{A}}_{k3} & \circ & \hat{\mathbf{A}}_{kk} \end{bmatrix} \qquad \mathbf{P}^T\mathbf{AP} = \begin{bmatrix} \hat{\mathbf{A}}_{11} & 0 & 0 & \circ & 0 \\ 0 & \hat{\mathbf{A}}_{22} & 0 & \circ & 0 \\ 0 & 0 & \hat{\mathbf{A}}_{33} & \circ & 0 \\ \circ & \circ & \circ & \circ & \circ \\ 0 & 0 & 0 & \circ & \hat{\mathbf{A}}_{kk} \end{bmatrix}$$

Block triangular **Block diagonal**

The condition of irreducibility is also expressed in terms of the *directed graph* of the matrix is *strictly connected*.

The graph is a geometrical figure (that in our purposes can be well represented by the discretisation lattice), where *edges* or *arcs* join the point P_r to point P_s if $a_{rs} \neq 0$. A graph is strictly connected if for any couple P_i and P_j of points there exists an oriented path leading from P_i to P_j.



For instance, in the figure it is possible to join P_1 to P_6 through the arcs $P_1 \rightarrow P_2$, $P_2 \rightarrow P_5$ and $P_5 \rightarrow P_6$ since it is $a_{12} \neq 0$, $a_{25} \neq 0$ and $a_{56} \neq 0$. Considering the system matrices obtained by the discretization of diffusion problems it can be understood that this is true for any other couple of points, since the discretised leakage terms always join a point to 2, 4 or 6 other points making a strictly connected chain of paths. Therefore, the graph is strictly coupled and the matrix irreducible.

Unfolding the mathematical jargon for our purposes

Let us now unfold the numerical jargon, to draw meaningful conclusions in our specific case:

- **the operation $\mathbf{P}^T\mathbf{AP}$ has the result of a *renumbering of the discretization nodes*;**
- ***whenever this renumbering would lead to a block-triangular or block-diagonal* (for a symmetric case) *system matrix*, *regions of the reactor would be "decoupled" from each other*, because the neutron flux in them would be independently determined;**
- **this situation does not normally occur, except in cases in which infinitely absorbing curtains (or empty spaces with zero flux) purposely separate reactor regions;**
- **therefore, *except for these extreme cases, the system matrices involved in the discretization of the diffusion equation are irreducible*;**
- **the strict diagonal dominance of the system matrix is assured only in cases in which absorption cross sections are non-zero; e.g., it was shown that for a simple 2D problem with uniform properties it is**

$$a_{i,i} = \Sigma_a + \frac{4D}{h^2} \qquad\qquad a_{i,j} = -\frac{D}{h^2} \qquad\qquad (i \neq j)$$

- **whenever Σ_a vanishes somewhere in the reactor, the strict diagonal dominance of the system matrix is lost;**
- **from a physical standpoint, nodes in which absorption is zero would have an unbounded flux in the presence of a source, unless leakage towards neighbouring nodes will take place;**
- **in this light, the role of irreducibility is to restore the possibility to calculate a limited neutron flux in poorly absorbing regions.**

It is interesting to consider that the concept of "nuclear reactor coupling" is frequently used to mean that reactor regions behave as a single system, with a strict interplay.

When a reactor is said to be "weakly coupled" it is often meant that the "migration length" is relatively small with respect to reactor size. In such a case, though different reactor regions do anyway interact among each other, this interaction may be weak enough to make them behave somehow separately: in fact, the probability of neutrons generated in a region to have influence on the behaviour of other regions may be small, <u>though never zero</u>.

The above demonstrates how much physical and numerical aspects interact in determining the conditions of convergence for the Jacobi method.

The simple idea behind the Jacobi method

Beyond the formalism that is necessary to convince about convergence properties, it should be noted that the simple idea behind the Jacobi scheme is to solve a given linear system of equations

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \ldots + a_{1,n}x_n = b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \ldots + a_{2,n}x_n = b_2 \\ \ldots \\ a_{n,1}x_1 + a_{n,2}x_2 + \ldots + a_{n,n}x_n = b_n \end{cases}$$

by the "strange" solution procedure

$$\begin{cases} x_1^{(m+1)} = \left(b_1 - a_{1,2}x_2^{(m)} \ldots - a_{1,n}x_n^{(m)}\right)/a_{1,1} \\ x_2^{(m+1)} = \left(b_2 - a_{2,1}x_1^{(m)} \ldots - a_{1,n}x_n^{(m)}\right)/a_{2,2} \\ \ldots \\ x_n^{(m+1)} = \left(b_n - a_{n,1}x_1^{(m)} \ldots - a_{n,n-1}x_{n-1}^{(m)}\right)/a_{n,n} \end{cases}$$

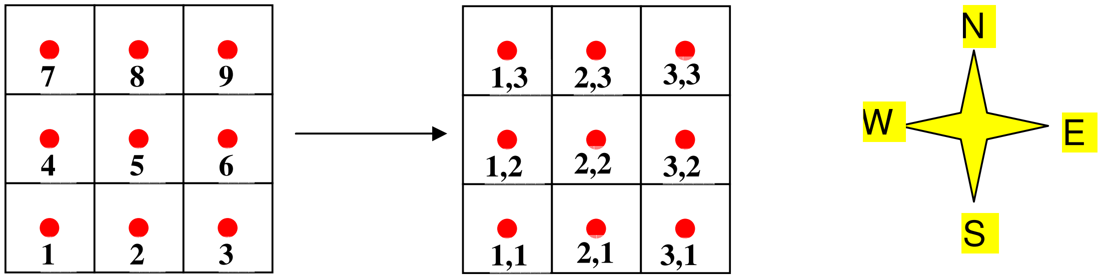
in which only the i-th unknown is retained at the left hand side and all the others are assigned a tentative value (or a previous iteration one) at the right hand side.

The fact that such a rough procedure may lead to convergence for a vast class of system matrices must be regarded as a consequence of the fact that the iteration matrix, having a spectral radius smaller than unity, contacts the error made with the initial guess in further iterations.

Practical implementation of the Jacobi method

Advantage is taken from the limited number of nodes involved in discretised diffusion equations (5 in 2D and 7 in 3D) and the nodes are numbered with multiple subscripts (one for each coordinate).

It is thus possible to define the coefficients of the equation related to the a given node, with reference to the neighbouring points



E.g., in two dimensions it is

$$-\frac{D}{h_x^2}\left(\phi_{i+1,j}-2\phi_{i,j}+\phi_{i-1,j}\right)-\frac{D}{h_y^2}\left(\phi_{i,j+1}-2\phi_{i,j}+\phi_{i,j-1}\right)+\Sigma_a\phi_{i,j}=S_{i,j}$$

that becomes

$$\phi_{i,j}=\mathcal{W}_{i,j}\phi_{i-1,j}+\mathcal{E}_{i,j}\phi_{i+1,j}+\mathcal{S}_{i,j}\phi_{i,j-1}+\mathcal{N}_{i,j}\phi_{i,j+1}+q_{i,j}$$

where

$$\mathcal{W}_{i,j}=\mathcal{E}_{i,j}=\frac{\dfrac{D}{h_x^2}}{\Sigma_a+\dfrac{2D}{h_x^2}+\dfrac{2D}{h_y^2}} \qquad \mathcal{S}_{i,j}=\mathcal{N}_{i,j}=\frac{\dfrac{D}{h_y^2}}{\Sigma_a+\dfrac{2D}{h_x^2}+\dfrac{2D}{h_y^2}}$$

$$q_{i,j}=\frac{S_{i,j}}{\Sigma_a+\dfrac{2D}{h_x^2}+\dfrac{2D}{h_y^2}}$$

The iterative scheme has therefore the form

$$\phi_{i,j}^{(m+1)}=\mathcal{W}_{i,j}\phi_{i-1,j}^{(m)}+\mathcal{E}_{i,j}\phi_{i+1,j}^{(m)}+\mathcal{S}_{i,j}\phi_{i,j-1}^{(m)}+\mathcal{N}_{i,j}\phi_{i,j+1}^{(m)}+q_{i,j}$$

In the 3D case other two additional coefficients would appear (e.g., $\mathcal{U}_{i,j}$ e $\mathcal{D}_{i,j}$).

<u>Convergence rate</u>

a) If the matrix $\mathbf{H_J}$ has a single eigenvalue having maximum modulus, λ_1, and its eigenvectors form a complete basis, it is:

$$\boldsymbol{\varepsilon}^{(m+1)}=\mathbf{H_J}^m\boldsymbol{\varepsilon}^{(1)}=\sum_{h=1}^{n}c_h^{(1)}\lambda_h^m\boldsymbol{\varphi}_h=\lambda_1^m\left\{c_1^{(1)}\boldsymbol{\varphi}_1+\sum_{h=2}^{n}c_h^{(1)}\frac{\lambda_h^m}{\lambda_1^m}\boldsymbol{\varphi}_h\right\}$$

therefore

$$\left|\varepsilon_j^{(m+1)}\right|\rightarrow C\left|\lambda_1^m\right|=C\rho^m\left(\mathbf{H_J}\right)$$

b) If B is symmetric (uniform reactor with uniform discrettzation) it can be shown that the eigenvalues are patterned in pairs $\pm\lambda_i$ and there are two eigenvalues having maximum modulus. Even, in this case it is anyway

$$\left\|\boldsymbol{\varepsilon}^{(m+1)}\right\|^2 = \left(\boldsymbol{\varepsilon}^{(m+1)}, \boldsymbol{\varepsilon}^{(m+1)}\right) = \left(\sum_{h=1}^{n} c_h^{(1)} \lambda_h^m \boldsymbol{\varphi}_h, \sum_{h=1}^{n} c_h^{(1)} \lambda_h^m \boldsymbol{\varphi}_h\right) = \sum_{h=1}^{n} \left|c_h^{(1)} \lambda_h^m\right|^2$$

therefore (we use the shorthand notation $\rho(\mathbf{H_J}) = \rho_{\mathbf{H_J}}$)

$$\left\|\boldsymbol{\varepsilon}^{(m+1)}\right\|^2 = \rho_{\mathbf{H_J}}^{2m} \left\{ \sum_{|\lambda_h| = \rho_{\mathbf{H_J}}} \left|c_h^{(1)}\right|^2 + \sum_{|\lambda_h| < \rho_{\mathbf{H_J}}} \left|c_h^{(1)}\right|^2 \frac{\left|\lambda_h^m\right|^2}{\rho_{\mathbf{H_J}}^{2m}} \right\} \to \rho_{\mathbf{H_J}}^{2m} \sum_{|\lambda_h| = \rho_{\mathbf{H_J}}} \left|c_h^{(1)}\right|^2$$

or

$$\left\|\boldsymbol{\varepsilon}^{(m+1)}\right\| \to C\,\rho_{\mathbf{H_J}}^m$$

Since after a convenient number of iterations the error decrease at each iteration is measured by the spectral radius, the *asymptotic rate of convergence* can be therefore defined as:

$$R_\infty = -\ln \rho\left(\mathbf{H_J}\right)$$

The number of iterations to be performed for reducing the error by a factor $\sigma < 1$ can be estimated assuming that at each iteration the error is reduced by a factor $\rho(\mathbf{B})$

$$\rho^m\left(\mathbf{H_J}\right) < \sigma \qquad \Rightarrow \qquad m \ln \rho\left(\mathbf{H_J}\right) < \ln \sigma \qquad \Rightarrow$$

$$m > \frac{\ln \sigma}{\ln \rho\left(\mathbf{H_J}\right)} = -\frac{\ln \sigma}{R_\infty}$$

In the purpose of a rapid convergence, it is therefore desirable that $\rho\left(\mathbf{H_J}\right) << 1$. It must be noted that:

- **if $\Sigma_a \to 0$ the degree of diagonal dominance of the system matrix is reduced; therefore it can be expected that $\rho\left(\mathbf{H_J}\right) \to 1$;**
- **similarly, if $h_x \to 0$ and $h_y \to 0$ the diagonal dominance tends to vanish and $\rho\left(\mathbf{H_J}\right) \to 1$; therefore, by refining the spatial grid a greater detail is obtained but the convergence rate decreases**

- ***Gauss-Seidel Method***

The iteration matrix of the Jacobi method, $\mathbf{H_J}$, is further split as follows:

$$\mathbf{H_J} \qquad\qquad = \qquad\qquad \mathbf{L} \qquad\qquad + \qquad\qquad \mathbf{U}$$

$$\begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \circ & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \circ & -\frac{a_{2n}}{a_{22}} \\ -\frac{a_{31}}{a_{33}} & -\frac{a_{32}}{a_{33}} & 0 & \circ & -\frac{a_{3n}}{a_{33}} \\ \circ & \circ & \circ & \circ & \circ \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & -\frac{a_{n3}}{a_{nn}} & \circ & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \circ & 0 \\ -\frac{a_{21}}{a_{22}} & 0 & 0 & \circ & 0 \\ -\frac{a_{31}}{a_{33}} & -\frac{a_{32}}{a_{33}} & 0 & \circ & 0 \\ \circ & \circ & \circ & \circ & \circ \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & -\frac{a_{n3}}{a_{nn}} & \circ & 0 \end{bmatrix} + \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \circ & -\frac{a_{1n}}{a_{11}} \\ 0 & 0 & -\frac{a_{23}}{a_{22}} & \circ & -\frac{a_{2n}}{a_{22}} \\ 0 & 0 & 0 & \circ & -\frac{a_{3n}}{a_{33}} \\ \circ & \circ & \circ & \circ & \circ \\ 0 & 0 & 0 & \circ & 0 \end{bmatrix}$$

it is therefore

$$\boldsymbol{\phi} = \mathbf{H_J}\,\boldsymbol{\phi} + \mathbf{q} \qquad \Rightarrow \qquad \boldsymbol{\phi} = \mathbf{L}\,\boldsymbol{\phi} + \mathbf{U}\,\boldsymbol{\phi} + \mathbf{q}$$

The following iteration process is then defined

$$\boldsymbol{\phi}^{(m+1)} = \mathbf{L}\,\boldsymbol{\phi}^{(m+1)} + \mathbf{U}\,\boldsymbol{\phi}^{(m)} + \mathbf{q}$$

Written in terms of components, this relationship is

$$\underbrace{\phi_i^{(m+1)}}_{\substack{value\ of\ \phi_i \\ at\ step\ m+1}} = \underbrace{\sum_{j=1}^{i-1} l_{i,j}\phi_j^{(m+1)}}_{\substack{summation\ over\ the\ components\ of\ \phi \\ already\ updated\ in\ step\ m+1}} + \underbrace{\sum_{j=i+1}^{n} u_{i,j}\phi_j^{(m)}}_{\substack{summation\ over\ the\ components\ of\ \phi \\ not\ yet\ updated\ in\ step\ m+1}} + q_i$$

NB: The basic idea is therefore to use immediately in the calculation the flux values updated in the current iteration step

The iteration matrix can be therefore identified as follows:

$$\boldsymbol{\phi}^{(m+1)} - \mathbf{L}\,\boldsymbol{\phi}^{(m+1)} = \mathbf{U}\,\boldsymbol{\phi}^{(m)} + \mathbf{q} \Rightarrow \boldsymbol{\phi}^{(m+1)} = \left(\mathbf{I} - \mathbf{L}\right)^{-1}\mathbf{U}\,\boldsymbol{\phi}^{(m)} + \left(\mathbf{I} - \mathbf{L}\right)^{-1}\mathbf{q}$$

or

$$\mathbf{H}_{GS} = \left(\mathbf{I} - \mathbf{L}\right)^{-1}\mathbf{U}$$

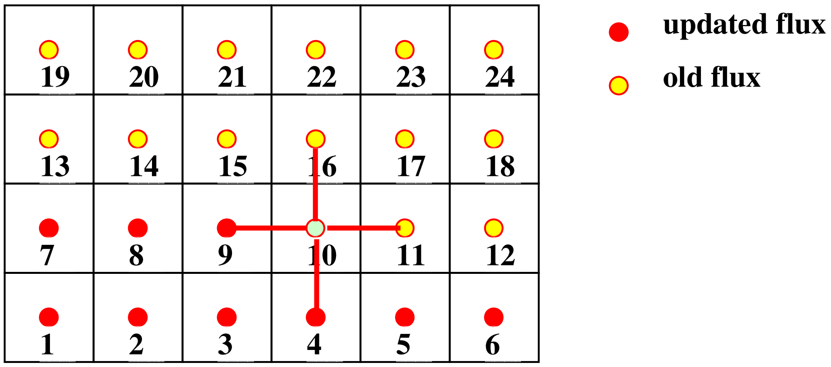
Under not too demanding assumptions, it can be shown that

$$\rho\left(\mathbf{H}_{GS}\right) = \rho^2\left(\mathbf{H_J}\right)$$

Therefore, the Gauss-Seidel method is expected to converge more rapidly than the Jacobi one, with a nearly halved number of iterations

Practical implementation of the Gauss-Seidel method

In the calculation of a new flux value, both updated and old iteration values are adopted



Using a two subscript notation for this 2D example, it is:

$$\phi_{i,j}^{(m+1)} = \mathcal{W}_{i,j}\phi_{i-1,j}^{(m+1)} + \mathcal{E}_{i,j}\phi_{i+1,j}^{(m)} + \mathcal{S}_{i,j}\phi_{i,j-1}^{(m+1)} + \mathcal{N}_{i,j}\phi_{i,j+1}^{(m)} + q_{i,j}$$

Successive OverRelaxation Method (SOR)

The basic idea is to extrapolate the prediction obtained by the Gauss-Seidel method, through an *overrelaxation factor* $\omega > 1$

$$\boldsymbol{\phi}^{(m+1)} = \boldsymbol{\phi}^{(m)} + \omega\left[\boldsymbol{\phi}^{(m+1)} - \boldsymbol{\phi}^{(m)}\right]_{GS}$$

where

$$\left[\boldsymbol{\phi}^{(m+1)} - \boldsymbol{\phi}^{(m)}\right]_{GS} = \mathbf{L}\,\boldsymbol{\phi}^{(m+1)} + \left(\mathbf{U} - \mathbf{I}\right)\boldsymbol{\phi}^{(m)} + \mathbf{q}$$

Therefore

$$\boldsymbol{\phi}^{(m+1)} = \boldsymbol{\phi}^{(m)} + \omega\left[\mathbf{L}\,\boldsymbol{\phi}^{(m+1)} + \left(\mathbf{U} - \mathbf{I}\right)\boldsymbol{\phi}^{(m)} + \mathbf{q}\right]$$

or

$$\boldsymbol{\phi}^{(m+1)} = \omega\mathbf{L}\,\boldsymbol{\phi}^{(m+1)} + \left[\left(1-\omega\right)\mathbf{I} + \omega\mathbf{U}\right]\boldsymbol{\phi}^{(m)} + \omega\mathbf{q}$$

Writing this vector relationship by components, it is

$$\phi_i^{(m+1)} = \omega\sum_{j=1}^{i-1} l_{i,j}\phi_j^{(m+1)} + (1-\omega)\phi_i^{(m)} + \omega\sum_{j=i+1}^{n} u_{i,j}\phi_j^{(m)} + \omega q_i$$

from which it can be checked that

$$\phi_i^{(m+1)} = \phi_i^{(m)} + \omega\left[\sum_{j=1}^{i-1} l_{i,j}\phi_j^{(m+1)} + \sum_{j=i+1}^{n} u_{i,j}\phi_j^{(m)} + q_i - \phi_i^{(m)}\right] = \phi_i^{(m)} + \omega\left[\phi_i^{(m+1)} - \phi_i^{(m)}\right]_{GS}$$

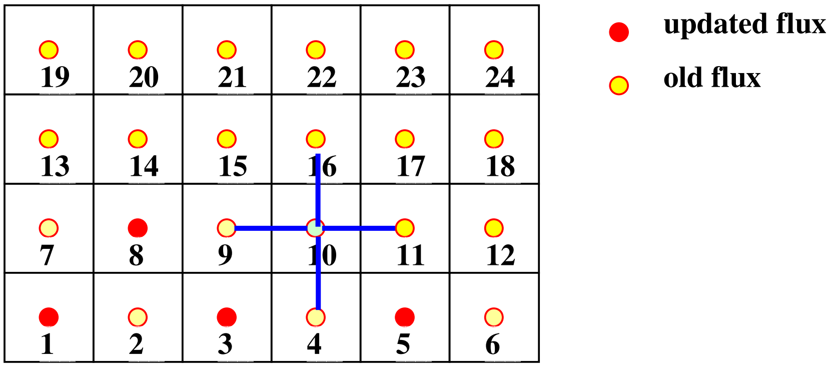
Using the two subscripts notation for a 2D case, it is

$$\begin{aligned}\phi_{i,j}^{(m+1)} &= \left(1-\omega\right)\phi_{i,j}^{(m)} + \omega\mathcal{W}_{i,j}\phi_{i-1,j}^{(m+1)} + \omega\mathcal{S}_{i,j}\phi_{i,j-1}^{(m+1)} + \omega\mathcal{E}_{i,j}\phi_{i+1,j}^{(m)} + \omega\mathcal{N}_{i,j}\phi_{i,j+1}^{(m)} + \omega q_{i,j} \\ &= \phi_{i,j}^{(m)} + \omega\left[\phi_{i,j}^{(m+1)} - \phi_{i,j}^{(m)}\right]_{GS}\end{aligned}$$

The numbering of nodes, which is irrelevant in the Jacobi method, is anyway important for the Gauss-Seidel and the SOR ones

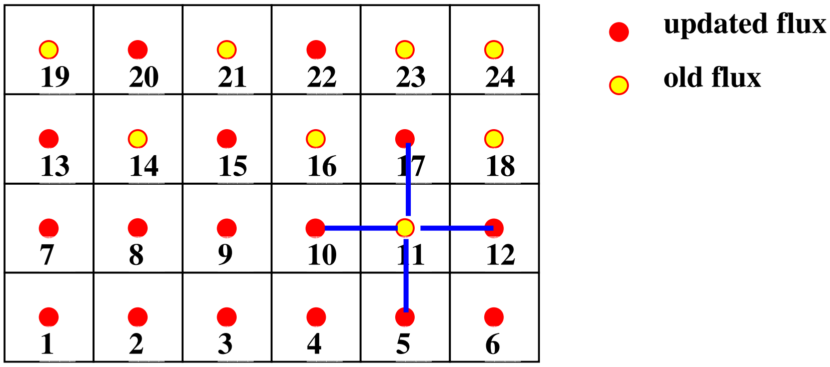
For both the Gauss-Seidel and the SOR method it is possible to adopt a “checkerboard” variant; two steps are performed at each iteration:

1) fluxes are updated in staggered nodes on the basis of old fluxes only



$$\phi_{i,j}^{(m+1)} = \mathcal{W}_{i,j}\phi_{i-1,j}^{(m)} + \mathcal{E}_{i,j}\phi_{i+1,j}^{(m)} + \mathcal{S}_{i,j}\phi_{i,j-1}^{(m)} + \mathcal{N}_{i,j}\phi_{i,j+1}^{(m)} + q_{i,j}$$

2) then, fluxes are updated in the remaining nodes on the basis of updated fluxes only



$$\phi_{i,j}^{(m+1)} = \mathcal{W}_{i,j}\phi_{i-1,j}^{(m+1)} + \mathcal{E}_{i,j}\phi_{i+1,j}^{(m+1)} + \mathcal{S}_{i,j}\phi_{i,j-1}^{(m+1)} + \mathcal{N}_{i,j}\phi_{i,j+1}^{(m+1)} + q_{i,j}$$

A greater symmetry is adopted avoiding the generation of spurious asymmetric components in the error

The iteration matrix in SOR is found considering that

$$\boldsymbol{\phi}^{(m+1)} = \left(\mathbf{I} - \omega\mathbf{L}\right)^{-1}\left[\left(1-\omega\right)\mathbf{I} + \omega\mathbf{U}\right]\boldsymbol{\phi}^{(m)} + \omega\left(\mathbf{I} - \omega\mathbf{L}\right)^{-1}\mathbf{q}$$

It can be found that the minimum of the spectral radius of the iteration matrix in SOR is found when:

$$\omega = \omega_{opt} = \frac{2}{1+\sqrt{1-\rho^2\left(\mathbf{H_J}\right)}} \qquad \textbf{(where } 1 \le \omega_{opt} \le 2\textbf{)}$$

In order to calculate ω_{opt} it is necessary to know $\rho\left(\mathbf{H_J}\right)$; therefore, it is advisable to perform a number of iterations with the Jacobi method to estimate its asymptotical convergence rate. For a symmetric $\mathbf{H_J}$, it is

$$\mathbf{q} = \sum_{h=1}^{n} q_h \boldsymbol{\varphi}_h \quad with \quad \boldsymbol{\varphi}_h \cdot \boldsymbol{\varphi}_j = \delta_{h,j}$$

then

$$\left\|\mathbf{H_J}^m\mathbf{q}\right\|^2 = \left(\mathbf{H_J}^m\mathbf{q}, \mathbf{H_J}^m\mathbf{q}\right) = \left(\sum_{h=1}^{n} q_h\lambda_h^m\boldsymbol{\varphi}_h, \sum_{h=1}^{n} q_h\lambda_h^m\boldsymbol{\varphi}_h\right) = \sum_{h=1}^{n}\left|q_h\right|^2\lambda_h^{2m}$$

$$= \sum_{|\lambda_h|=\rho_B}^{n}\left|q_h\right|^2\rho_{H_J}^{2m} + \sum_{|\lambda_h|<\rho_B}^{n}\left|q_h\right|^2\lambda_h^{2m} = \rho_{H_J}^{2m}\left\{\sum_{|\lambda_h|=\rho_B}^{n}\left|q_h\right|^2 + \sum_{|\lambda_h|<\rho_B}^{n}\left|q_h\right|^2\frac{\lambda_h^{2m}}{\rho_{H_J}^{2m}}\right\}$$

As a consequence:

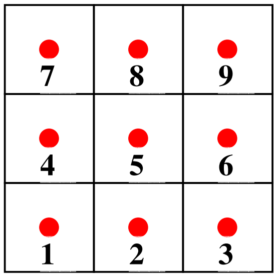
$$\lim_{m\to\infty}\frac{\left\|\mathbf{H_J}^{m+1}\mathbf{q}\right\|^2}{\left\|\mathbf{H_J}^{m}\mathbf{q}\right\|^2} = \lim_{m\to\infty}\rho_{\mathbf{B}}^2\left\{\frac{\displaystyle\sum_{|\lambda_h|=\rho_B}^{n}\left|q_h\right|^2 + \sum_{|\lambda_h|<\rho_B}^{n}\left|q_h\right|^2\frac{\lambda_h^{2m+2}}{\rho_{\mathbf{H_J}}^{2m+2}}}{\displaystyle\sum_{|\lambda_h|=\rho_B}^{n}\left|q_h\right|^2 + \sum_{|\lambda_h|<\rho_B}^{n}\left|q_h\right|^2\frac{\lambda_h^{2m}}{\rho_{\mathbf{H_J}}^{2m}}}\right\} = \rho_{\mathbf{H_J}}^2$$

It is:

$$\left(\rho_{SOR}\right)_{\omega=\omega_{opt}} = \frac{2}{1+\sqrt{1-\rho_{H_J}^2}} - 1$$

Line OverRelaxation Method (LOR)

- **Since the three-diagonal matrices are efficiently dealt with by the Thomas' algorithm, it is tried to make use of *the three-point structure existing in each direction in 2D and 3D problems***



- **For instance, in the case of the 9 node grid in the figure, the obtained banded matrix can be interpreted as *a block matrix whose diagonal blocks are tridiagonal***

$$\left[\begin{array}{ccc|ccc|ccc} a_{11} & a_{12} & 0 & a_{14} & 0 & 0 & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & a_{25} & 0 & 0 & 0 & 0 \\ 0 & a_{32} & a_{33} & 0 & 0 & a_{36} & 0 & 0 & 0 \\ \hline a_{41} & 0 & 0 & a_{44} & a_{45} & 0 & a_{47} & 0 & 0 \\ 0 & a_{52} & 0 & a_{54} & a_{55} & a_{56} & 0 & a_{58} & 0 \\ 0 & 0 & a_{63} & 0 & a_{65} & a_{66} & 0 & 0 & a_{69} \\ \hline 0 & 0 & 0 & a_{74} & 0 & 0 & a_{77} & a_{78} & 0 \\ 0 & 0 & 0 & 0 & a_{85} & 0 & a_{87} & a_{88} & a_{89} \\ 0 & 0 & 0 & 0 & 0 & a_{96} & 0 & a_{98} & a_{99} \end{array}\right] \cdot \left[\begin{array}{c} \phi_1 \\ \phi_2 \\ \phi_3 \\ \hline \phi_4 \\ \phi_5 \\ \phi_6 \\ \hline \phi_7 \\ \phi_8 \\ \phi_9 \end{array}\right] = \left[\begin{array}{c} S_1 \\ S_2 \\ S_3 \\ \hline S_4 \\ S_5 \\ S_6 \\ \hline S_7 \\ S_8 \\ S_9 \end{array}\right] \rightarrow \begin{bmatrix} \hat{\mathbf{A}}_{11} & \hat{\mathbf{A}}_{12} & 0 \\ \hat{\mathbf{A}}_{21} & \hat{\mathbf{A}}_{22} & \hat{\mathbf{A}}_{23} \\ 0 & \hat{\mathbf{A}}_{32} & \hat{\mathbf{A}}_{33} \end{bmatrix} \cdot \begin{bmatrix} \hat{\boldsymbol{\phi}}_1 \\ \hat{\boldsymbol{\phi}}_2 \\ \hat{\boldsymbol{\phi}}_3 \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{s}}_1 \\ \hat{\mathbf{s}}_2 \\ \hat{\mathbf{s}}_3 \end{bmatrix}$$

It should be noted that blocks $\hat{\mathbf{A}}_{i,i}$ <u>contain the coefficients of fluxes on a same line</u> and, therefore, they are three-diagonal as in 1D cases

On the other hand, blocks $\hat{\mathbf{A}}_{i,i-1}$ e $\hat{\mathbf{A}}_{i,i+1}$ contain the coefficients of the fluxes in the lines below and above the one in consideration

The idea is then pursued to write the system in the form

$$\hat{\mathbf{A}}_{i,i-1}\hat{\boldsymbol{\phi}}_{i-1} + \underbrace{\hat{\mathbf{A}}_{i,i}}_{threediagonal}\hat{\boldsymbol{\phi}}_i + \hat{\mathbf{A}}_{i,i+1}\hat{\boldsymbol{\phi}}_{i+1} = \hat{\mathbf{s}}_i$$

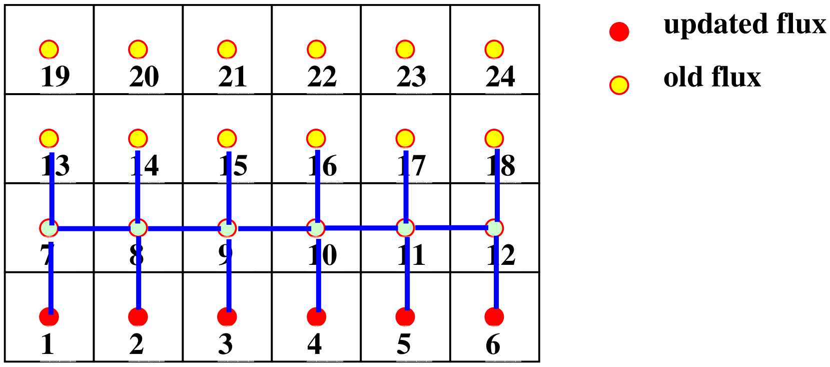
applying a Gauss-Seidel-like strategy

$$\overbrace{\underbrace{\hat{\mathbf{A}}_{i,i}}_{tridiagonal}\hat{\boldsymbol{\phi}}_i^{(m+1)}}^{line\ to\ be\ calculated} = \hat{\mathbf{s}}_i - \overbrace{\hat{\mathbf{A}}_{i,i-1}\hat{\boldsymbol{\phi}}_{i-1}^{(m+1)}}^{line\ already\ calculated} - \overbrace{\hat{\mathbf{A}}_{i,i+1}\hat{\boldsymbol{\phi}}_{i+1}^{(m)}}^{line\ not\ yet\ calculated}$$

Introducing also an overrelaxation, the LOR scheme is obtained

$$\underbrace{\hat{\mathbf{A}}_{i,i}}_{tri\,diagonal}\hat{\boldsymbol{\phi}}_i^{(m+1)} = \omega\hat{\mathbf{s}}_i - \omega\hat{\mathbf{A}}_{i,i-1}\hat{\boldsymbol{\phi}}_{i-1}^{(m+1)} + (1-\omega)\hat{\mathbf{A}}_{i,i}\hat{\boldsymbol{\phi}}_i^{(m)} - \omega\hat{\mathbf{A}}_{i,i+1}\hat{\boldsymbol{\phi}}_{i+1}^{(m)}$$

It is therefore possible to apply the direct Thomas algorithm on each line covering *"line-by-line"* the whole domain



- **With reference to the more relevant example in the above figure, adopting the two subscript notation, it is:**

$$\underbrace{\mathcal{W}_{i,j}\phi_{i-1,j}^{(m+1)}+O_{i,j}\phi_{i,j}^{(m+1)}+\mathcal{E}_{i,j}\phi_{i+1,j}^{(m+1)}}_{three\ point\ structure}=-\omega\mathcal{S}_{i,j}\phi_{i,j-1}^{(m+1)}-\omega\mathcal{N}_{i,j}\phi_{i,j+1}^{(m)}+\omega q_{i,j}$$

$$+(1-\omega)\left\{\mathcal{W}_{i,j}\phi_{i-1,j}^{(m)}+O_{i,j}\phi_{i,j}^{(m)}+\mathcal{E}_{i,j}\phi_{i+1,j}^{(m)}\right\}$$

where the right hand side exhibit a clear *three-point structure*

- **Though the technique can be applied preferentially in a single direction, it is more convenient to alternate the directions, pointing out the three-point structure alternatively in each one of them**

$$\underbrace{\mathcal{S}_{i,j}\phi_{i,j-1}^{(m+1)}+O_{i,j}\phi_{i,j}^{(m+1)}+\mathcal{N}_{i,j}\phi_{i,j+1}^{(m+1)}}_{three\ point\ structure}=-\omega\mathcal{W}_{i,j}\phi_{i-1,j}^{(m+1)}-\omega\mathcal{E}_{i,j}\phi_{i+1,j}^{(m)}+\omega q_{i,j}$$

$$+(1-\omega)\left\{\mathcal{S}_{i,j}\phi_{i-1,j}^{(m)}+O_{i,j}\phi_{i,j}^{(m)}+\mathcal{N}_{i,j}\phi_{i+1,j}^{(m)}\right\}$$

- **The following ADI method makes use of a similar technique**

<u>Alternating Direction Implicit Method (ADI)</u>

- **The method is due to Peaceman and Rachford (1955) and is based on the decomposition of matrix A in the form**

$$\mathbf{A} = \underbrace{\mathbf{H}}_{\substack{\text{"}horizontal\text{" } coefficients \\ of\ the\ discretized\ Laplacian}} + \underbrace{\mathbf{V}}_{\substack{\text{"}vertical\text{" } coefficients \\ of\ the\ discretized\ Laplacian}} + \underbrace{\mathbf{\Sigma}}_{\substack{coefficients \\ related\ to\ absorption}}$$

For a uniform property case, it is

$$\underbrace{-\frac{D}{h_x^2}\phi_{i-1,j}+\frac{2D}{h_x^2}\phi_{i,j}-\frac{D}{h_x^2}\phi_{i+1,j}}_{\substack{\downarrow \\ \mathbf{H}\phi}}\ \underbrace{-\frac{D}{h_y^2}\phi_{i,j-1}+\frac{2D}{h_y^2}\phi_{i,j}-\frac{D}{h_y^2}\phi_{i,j+1}}_{\substack{\downarrow \\ \mathbf{V}\phi}}+\Sigma_{a,ij}\phi_{i,j}=S_{i,j}$$

- **For purpose of illustration, it can be argued that the above splitting of matrix a is used to firstly obtain equivalent forms of the linear system**
  - **a first form is the following:**

$$\mathbf{A}\boldsymbol{\phi}=\mathbf{s} \Rightarrow \qquad [\mathbf{H}+\mathbf{V}+\mathbf{\Sigma}]\boldsymbol{\phi}=\mathbf{s} \qquad \Rightarrow \qquad \left[\left(\mathbf{H}+\frac{1}{2}\mathbf{\Sigma}\right)+\left(\mathbf{V}+\frac{1}{2}\mathbf{\Sigma}\right)\right]\boldsymbol{\phi}=\mathbf{s}$$

$$\left(\mathbf{H}+\frac{1}{2}\mathbf{\Sigma}\right)\boldsymbol{\phi}=\mathbf{s}-\left(\mathbf{V}+\frac{1}{2}\mathbf{\Sigma}\right)\boldsymbol{\phi} \qquad \Rightarrow \qquad \left[\mathbf{H}+\frac{1}{2}\mathbf{\Sigma}+\omega_m\mathbf{I}\right]\boldsymbol{\phi}=\mathbf{s}-\left[\mathbf{V}+\frac{1}{2}\mathbf{\Sigma}-\omega_m\mathbf{I}\right]\boldsymbol{\phi}$$

or

$$\left[\mathbf{H}+\frac{1}{2}\mathbf{\Sigma}+\omega_m\mathbf{I}\right]\boldsymbol{\phi}=\left[\omega_m\mathbf{I}-\mathbf{V}-\frac{1}{2}\mathbf{\Sigma}\right]\boldsymbol{\phi}+\mathbf{s}$$

  - **a second form is trivially obtained by exchanging the roles of H and V in the former one:**

$$\left[\mathbf{V}+\frac{1}{2}\mathbf{\Sigma}+\omega_m\mathbf{I}\right]\boldsymbol{\phi}=\left[\omega_m\mathbf{I}-\mathbf{H}-\frac{1}{2}\mathbf{\Sigma}\right]\boldsymbol{\phi}+\mathbf{s}$$

- **In the above, ω_m is a forcing parameter that could be thought of as the introduction of some degree of inertia in the iterations; this aspect will appear more clearly by treating the transient version of the algorithm; here this will become a key aspect for achieving fast convergence in some cases**

- **Taking profit of the three-point structure in the two directions, semi-iterations are performed adopting both forms:**

$$\underbrace{\left[\mathbf{H}+\frac{1}{2}\boldsymbol{\Sigma}+\omega_m\,\mathbf{I}\right]}_{\textit{three-point structure}}\;\boldsymbol{\phi}^{(m+1/2)}=\left[\omega_m\,\mathbf{I}-\mathbf{V}-\frac{1}{2}\boldsymbol{\Sigma}\right]\;\boldsymbol{\phi}^{(m)}+\mathbf{s}$$

$$\underbrace{\left[\mathbf{V}+\frac{1}{2}\boldsymbol{\Sigma}+\omega_m\,\mathbf{I}\right]}_{\textit{three-point structure}}\;\boldsymbol{\phi}^{(m+1)}=\left[\omega_m\,\mathbf{I}-\mathbf{H}-\frac{1}{2}\boldsymbol{\Sigma}\right]\;\boldsymbol{\phi}^{(m+1/2)}+\mathbf{s}$$

- ♦ **Since the matrices at the left hand side have a three-point structure (they are not both three-diagonal anyway), profit can be taken of the Thomas' algorithm to evaluate "implicitly" each row or column by "alternating" the sweeping directions: hence the name of the method**
- ♦ **It is interesting to rewrite the two semi-iterations in single step form**
  - ➢ **from the first semi-iteration it is**

$$\boldsymbol{\phi}^{(m+1/2)}=\left[\mathbf{H}+\frac{1}{2}\boldsymbol{\Sigma}+\omega_m\,\mathbf{I}\right]^{-1}\left[\omega_m\,\mathbf{I}-\mathbf{V}-\frac{1}{2}\boldsymbol{\Sigma}\right]\;\boldsymbol{\phi}^{(m)}+\left[\mathbf{H}+\frac{1}{2}\boldsymbol{\Sigma}+\omega_m\,\mathbf{I}\right]^{-1}\;\mathbf{s}$$

  - ➢ **substituting into the second and solving for $\boldsymbol{\phi}^{(m+1)}$ it is**

$$\boldsymbol{\phi}^{(m+1)}=\left[\mathbf{V}+\frac{1}{2}\boldsymbol{\Sigma}+\omega_m\,\mathbf{I}\right]^{-1}\left[\omega_m\,\mathbf{I}-\mathbf{H}-\frac{1}{2}\boldsymbol{\Sigma}\right]\left[\mathbf{H}+\frac{1}{2}\boldsymbol{\Sigma}+\omega_m\,\mathbf{I}\right]^{-1}\left[\omega_m\,\mathbf{I}-\mathbf{V}-\frac{1}{2}\boldsymbol{\Sigma}\right]\boldsymbol{\phi}^{(m)}$$
$$+\textit{source terms}$$

- ♦ **The resulting iteration matrix is, therefore**

$$\mathbf{J}_{\omega_m}=\left[\mathbf{V}+\frac{1}{2}\boldsymbol{\Sigma}+\omega_m\,\mathbf{I}\right]^{-1}\left[\omega_m\,\mathbf{I}-\mathbf{H}-\frac{1}{2}\boldsymbol{\Sigma}\right]\left[\mathbf{H}+\frac{1}{2}\boldsymbol{\Sigma}+\omega_m\,\mathbf{I}\right]^{-1}\left[\omega_m\,\mathbf{I}-\mathbf{V}-\frac{1}{2}\boldsymbol{\Sigma}\right]$$

It can be shown that

$$\rho\left(\mathbf{J}_{\omega_m}\right)<1\qquad \textit{if}\qquad \omega_m>0$$

- **The demonstration of this result is based on the following steps:**
  - **it is shown that $\mathbf{H}+\frac{1}{2}\mathbf{\Sigma}$ and $\mathbf{V}+\frac{1}{2}\mathbf{\Sigma}$ are *symmetric and positive definite (SPD)*, i.e., such that $\boldsymbol{\phi}^T\mathbf{A}\boldsymbol{\phi}>0,\ \forall\boldsymbol{\phi}\neq\mathbf{0}$; as such, all their eigenvalues are *real* and *positive*; ([3])**
  - **it is then shown that if $\omega_m>0$, any eigenvalue of $\mathbf{J}_{\omega_m}$, has a magnitude less than unity.**
- **Let us skip the demonstration and focus to understand why the ADI method is particularly successful for separable and nearly separable problems:**
  - **it is first recalled that for separable problems defined by the partial differential equation**

    $$\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}+B_{l,k}^2\phi=0$$

    the general solution has the form

    $$\phi=\varphi_{l,k}(x,y)\sim X_l(x)Y_k(y)$$

  - **the eigenfunctions $\varphi_{l,k}(x,y)$ are therefore such that**

    $$\frac{1}{X_l}\frac{d^2X_l}{dx^2}+\frac{1}{Y_k}\frac{d^2Y_k}{dy^2}+B_{l,k}^2=0$$

    and thanks to the separability it is

    $$\frac{1}{X_l}\frac{d^2X_l}{dx^2}+B_l^2=0 \qquad\qquad \frac{1}{Y_k}\frac{d^2Y_k}{dy^2}+B_k^2=0$$

  - **a direct consequence of the above is that $\varphi_{l,k}(x,y)$ satisfy simultaneously the equations**

    $$\frac{1}{\varphi_{l,k}}\frac{\partial^2\varphi_{l,k}}{\partial x^2}+B_l^2=0 \qquad\qquad \frac{1}{\varphi_{l,k}}\frac{\partial^2\varphi_{l,k}}{\partial y^2}+B_k^2=0$$

    i.e., $\varphi_{l,k}(x,y)$ is at the same time an eigenfunction of both the above eigenvalue equations, but with different eigenvalues;

[3] From the relationship $\boldsymbol{\phi}^T\mathbf{A}\boldsymbol{\phi}>0,\ \forall\boldsymbol{\phi}\neq\mathbf{0}$, assuming $\boldsymbol{\varphi}$ as an eigenvalue of $\mathbf{A}$, it follows $\boldsymbol{\varphi}^T\mathbf{A}\boldsymbol{\varphi}>0$ $\Rightarrow\boldsymbol{\varphi}^T\lambda\boldsymbol{\varphi}>0\Rightarrow\lambda\boldsymbol{\varphi}^T\boldsymbol{\varphi}>0\Rightarrow\lambda>0$. Moreover, it is recalled that the eigenvalues of a symmetric matrix are real as it is for any hermitian matrix.

- **these properties are retained also by the discretised form of the equations, so that the matrices H and V have the same eigenvectors, as it is of $\mathbf{H}+\frac{1}{2}\mathbf{\Sigma}$ and $\mathbf{V}+\frac{1}{2}\mathbf{\Sigma}$, though they have different eigenvalues**

$$\left(\mathbf{H}+\frac{1}{2}\mathbf{\Sigma}\right)\boldsymbol{\varphi}^{(k,l)}=\lambda_k\boldsymbol{\varphi}^{(k,l)} \qquad (k=1,\dots,n_x)$$

$$\left(\mathbf{V}+\frac{1}{2}\mathbf{\Sigma}\right)\boldsymbol{\varphi}^{(k,l)}=\mu_l\boldsymbol{\varphi}^{(k,l)} \qquad (l=1,\dots,n_y)$$

- **Considering the well known properties of functions of matrices, it can be understood that the spectral radius of matrix $\mathbf{J}_{\omega_m}$ is given by** [4] [5]

$$\rho\left(\mathbf{J}_{\omega_m}\right)=\max_{k,l}\left|\frac{(\omega_m-\lambda_k)(\omega_m-\mu_l)}{(\omega_m+\lambda_k)(\omega_m+\mu_l)}\right|<1 \qquad se\ \omega_m>0$$

where $\lambda_k>0$ e $\mu_l>0$ as they are eigenvector of the symmetric positive defined real matrices $\mathbf{H}+\frac{1}{2}\mathbf{\Sigma}$ and $\mathbf{V}+\frac{1}{2}\mathbf{\Sigma}$.

- **In fact, the matrices $\omega_m\mathbf{I}\pm\left(\mathbf{H}+\frac{1}{2}\mathbf{\Sigma}\right)$ and $\omega_m\mathbf{I}\pm\left(\mathbf{V}+\frac{1}{2}\mathbf{\Sigma}\right)$ have still $\boldsymbol{\varphi}^{(k,l)}$ as eigenvectors and the eigenvalues are $\omega_m\pm\lambda_k$ and $\omega_m\pm\mu_l$**
- **As a consequence, also $\mathbf{J}_{\omega_m}$ has $\boldsymbol{\varphi}^{(k,l)}$ as eigenvectors with the mentioned spectral radius.**
- **In theory, for separable cases it is possible to define a very efficient iterative procedure:**

[4] It is recalled that, given a function $f(z)$ of a complex variable z having a MacLaurin series expansion $f(z)=\sum_{n=0}^{\infty}a_nz^n$ that converges for $|z|<R$, the series $f(\mathbf{A})=\sum_{n=0}^{\infty}a_n\mathbf{A}^n$ converges if the eigenvalues of the matrix $\mathbf{A}$ have modulus less than *R*. The function is then called "well defined" (see e.g., R. Bronson, Matrix Operations, McGraw-Hill, 1989). As a consequence of the above, it can be easily verified that the matrix $f(\mathbf{A})$ has the same eigenvectors of $\mathbf{A}$ and its eigenvalues are $f(\lambda_i)$ where λ_i are the eigenvalues of $\mathbf{A}$.

[5] Moreover, if two matrices $\mathbf{A}$ and $\mathbf{B}$ have the same eigenvectors with eigenvalues λ_i and μ_i respectively, also $\mathbf{AB}$ has the same eigenvectors $\boldsymbol{\varphi}_i$ with eigenvalues $\lambda_i\mu_i$. In fact, it is easily seen that $\mathbf{AB}\boldsymbol{\varphi}_i=\mathbf{A}\mu_i\boldsymbol{\varphi}_i=\lambda_i\mu_i\boldsymbol{\varphi}_i$.

- **if p is the number of different eigenvalues λ_k e μ_l, p different values of ω_m, can be used in order to minimize the spectral radius of the overall iteration matrix**

  $$\left(\underline{\underline{\mathbf{J}}}_{\omega_m}\right)_{tot} = \underline{\underline{\mathbf{J}}}_{\omega_p}\,\underline{\underline{\mathbf{J}}}_{\omega_{p-1}}\ldots\underline{\underline{\mathbf{J}}}_{\omega_2}\,\underline{\underline{\mathbf{J}}}_{\omega_1}$$

  that has as a spectral radius the quantity

  $$\rho\left(\mathbf{J}_{\omega_m}\right)_{tot} = \max_{k,l}\left|\frac{(\omega_1-\lambda_k)(\omega_1-\mu_l)}{(\omega_1+\lambda_k)(\omega_1+\mu_l)}\cdots\frac{(\omega_p-\lambda_k)(\omega_p-\mu_l)}{(\omega_p+\lambda_k)(\omega_p+\mu_l)}\right|$$

- **putting at each one of the p steps ω_m equal to a given λ_k or μ_l, it would be obviously $\rho\left(\mathbf{J}_{\omega_m}\right)_{tot} = 0$**
- **since determining the eigenvalues λ_k e μ_l would require a considerable amount of computational work, it is instead accepted to choose the ω_m's with the criterion**

  $$\min_{\omega_1,\omega_2,\ldots,\omega_p}\ \max_{x\in(\alpha,\beta)}\left|\frac{(\omega_1-x)(\omega_2-x)\ldots(\omega_p-x)}{(\omega_1+x)(\omega_2+x)\ldots(\omega_p+x)}\right|$$

  where (α, β) contains the eigenvalues.

♦ **For non-separable problems, there is no rigorous theory for selecting the optimum forcing parameters; therefore:**

- **ADI has optimum performances for separable cases ([6]) but it might be not very efficient on some non-separable cases**
- **for nearly-separable cases (most of the addressed conditions) it is possible to evaluate the optimum values of ω_m referring to the separable case that can be considered closest to the addressed one, obtaining generally good performances;**
- **formulations defining the ω_m's are reported in relevant textbooks on this basis.**

([6]) It is remarked that in this context, by "separable" and "non-separable" problems it is simply meant that the reference partial differential equation problems are amenable or not to the solution by the technique of variable separation.

Gradient Methods

- **Considering the system**

$$\mathbf{A\,x} = \mathbf{s}$$

with $\mathbf{A}$ *real and symmetric*, we will identify with $\boldsymbol{\phi}$ its exact solution.

- **It is interesting to consider different quantitative definitions of the error:**
  - $E_1(\mathbf{x}) = (\boldsymbol{\phi}-\mathbf{x})^T \cdot (\boldsymbol{\phi}-\mathbf{x}) = \boldsymbol{\varepsilon}^T \cdot \boldsymbol{\varepsilon}$
  - $E_2(\mathbf{x}) = (\mathbf{s}-\mathbf{A\,x})^T \cdot (\mathbf{s}-\mathbf{A\,x}) = \mathbf{r}^T \cdot \mathbf{r}$
  - $E_3(\mathbf{x}) = (\boldsymbol{\phi}-\mathbf{x})^T (\mathbf{s}-\mathbf{A\,x}) = \boldsymbol{\varepsilon}^T \mathbf{A} \boldsymbol{\varepsilon}$

While the first and second definitions provide positive quantities (being scalar products of a vector by itself), the third one represents a quadratic form, having $\mathbf{A}$ as associated matrix, and can be greater or less than zero.

- **Recalling that, if A is *symmetric and positive definite* (SPD) the quadratic form is by definition positive (in fact, it is $\boldsymbol{\varepsilon}^T \mathbf{A} \boldsymbol{\varepsilon} > 0$ for any non-zero $\boldsymbol{\varepsilon}$), we will assume that this is the case, taking $E_3(\mathbf{x})$ as a measure of the error.**

- **It is therefore**

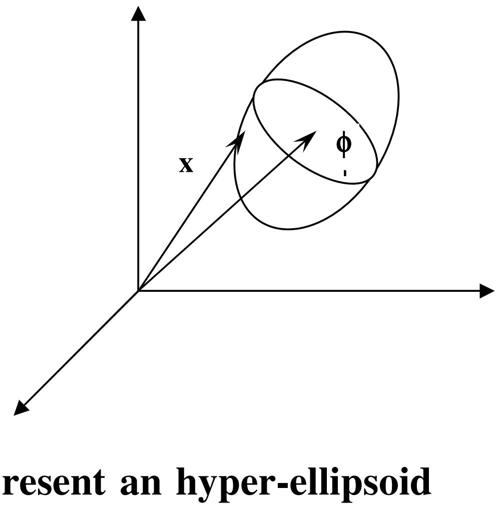
$$E_3(\mathbf{x}) \geq 0 \qquad \text{and} \qquad E_3(\mathbf{x}) = 0 \Rightarrow \mathbf{x} = \boldsymbol{\phi}$$

meaning that to find the system solution we should minimise $E_3(\mathbf{x})$

- **In this purpose, geometrical arguments can be set up by considering that the locus of points at constant error**

$$E_3(\mathbf{x}) = cost.$$

represents a *hyper-ellipsoid* in $\mathbb{R}^n$.



- **In fact, the canonical form of the quadratic polynomial can be found to represent an hyper-ellipsoid by making use of the transform**

$$-\boldsymbol{\varepsilon} = \mathbf{x} - \boldsymbol{\phi} = \mathbf{U}\,\mathbf{z}$$

where $\mathbf{U}$ is the orthogonal matrix ($\mathbf{U}^T\mathbf{U} = \mathbf{I}$ o $\mathbf{U}^T = \mathbf{U}^{-1}$) such that

$$\mathbf{U}^T\mathbf{A}\mathbf{U} = diag\{\lambda_1, \lambda_2, ..., \lambda_n\} = \boldsymbol{\Lambda}$$

(it always exists for symmetric matrices). The direction of the axes in the new reference system obtained by the transform are the eigenvectors of A, being an othonormal basis of $\mathbb{R}^n$.

It is:

$$E_3(\mathbf{z}) = \boldsymbol{\varepsilon}^T\mathbf{A}\,\boldsymbol{\varepsilon} = (\mathbf{U}\,\mathbf{z})^T\,\mathbf{A}\,\mathbf{U}\,\mathbf{z} = \mathbf{z}^T\mathbf{U}^T\mathbf{A}\mathbf{U}\,\mathbf{z} = \mathbf{z}^T\boldsymbol{\Lambda}\,\mathbf{z}$$

or

$$E_3(\mathbf{z}) = \lambda_1 z_1^2 + \lambda_2 z_2^2 + ... + \lambda_n z_n^2 = cost.$$

Since the eigenvalues of any SPD matrix are positive the above clearly represents a hyper-ellipsoid having axes $a_i \propto \sqrt{cost./\lambda_i}$, which are measures the error.

The basic idea in the gradient methods is therefore to search for better approximations of the solutions moving to hyper-ellipsoid with smaller and smaller axes.

Method of the “steepest descent”

The above idea is translated into mathematical formalism by the relationship

$$\boxed{\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \alpha^{(m)}\mathbf{p}^{(m)}}$$

where $\mathbf{x}^{(m)}$ represents the old approximation (on the old hyper-ellipsoid), $\mathbf{p}^{(m)}$ represents a vector identifying the direction of movement and $\alpha^{(m)}$ measures the distance along $\mathbf{p}^{(m)}$ necessary to reach $\mathbf{x}^{(m+1)}$

It must be understood that

- **the choice of $\mathbf{p}^{(m)}$ must be effective enough in order to move inwards the hyper-ellipsoid surface towards its centre**

- **computing $\mathbf{p}^{(m)}$ as a diametral direction would be too costly from the computational point of view: simpler recipes must be therefore identified.**

In the *"steepest descent"* method, the adopted recipe is to move inwards from the hyper-ellipsoid along the vector normal to the tangent plane, i.e. along the *"gradient"* of *E*$_3$ (hence, the name)

Therefore, in this case it is:

$\mathbf{p}^{(m)}$ **= vector normal to the surface in** $\mathbf{x}^{(m)}$

$\alpha^{(m)}$ **= scalar chosen as to minimise** $E_3\left(\mathbf{x}^{(m+1)}\right)$

In the canonical reference frame, it is conveniently assumed

$$\mathbf{p}'^{(m)} = -\mathbf{\Lambda}\,\mathbf{z}^{(m)}$$

since the gradient is immediately found to be

$$\frac{\partial E_3}{\partial z_i} = 2\lambda_i z_i \qquad \Rightarrow \qquad grad_z E_3\left(\mathbf{z}^{(m)}\right) \propto \mathbf{\Lambda}\,\mathbf{z}^{(m)}$$

However, since it is necessary to operate in the original reference frame it is needed to back-transform this result

$$\mathbf{p}^{(m)} = \mathbf{U}\,\mathbf{p}'^{(m)} = \mathbf{U}\mathbf{\Lambda}\,\mathbf{z}^{(m)} = \mathbf{U}\mathbf{U}^T\mathbf{A}\mathbf{U}\,\mathbf{z}^{(m)} = -\mathbf{A}\,\boldsymbol{\varepsilon}^{(m)}$$
$$= \mathbf{A}\left(\mathbf{x}^{(m)} - \boldsymbol{\phi}\right) = \mathbf{A}\,\mathbf{x}^{(m)} - \mathbf{s} = -\mathbf{r}^{(m)}$$

- $\alpha^{(m)}$ **is then chosen on the basis of the mentioned minimum criterion, thus obtaining**

$$\frac{\partial E_3}{\partial \alpha^{(m)}} = 0 \qquad\qquad \Rightarrow \qquad\qquad \alpha^{(m)} = -\frac{\mathbf{r}^{(m)T}\mathbf{r}^{(m)}}{\mathbf{r}^{(m)T}\mathbf{A}\mathbf{r}^{(m)}}$$

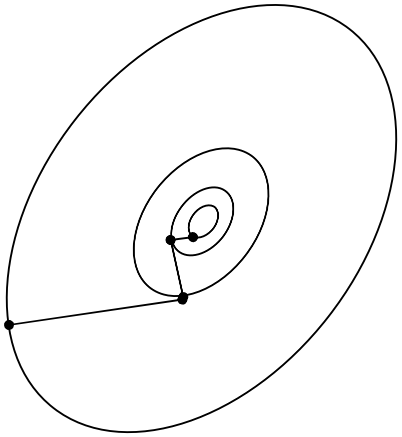
- **The final iterative scheme is therefore**

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \frac{\mathbf{r}^{(m)T}\mathbf{r}^{(m)}}{\mathbf{r}^{(m)T}\mathbf{A}\mathbf{r}^{(m)}}\mathbf{r}^{(m)} \qquad\qquad \mathbf{x}^{(0)} = \mathbf{x}_0$$

- **It should be noted that:**
  - **if** $z_i^{(m)} = 0 \Rightarrow p_i'^{(m)} = 0$**, i.e., whenever a component of the solution is exact, this advantage is preserved since the subsequent approximations will keep on the** $z_i = 0$ **plane;**

- **if the hyper-ellispoid is a solid of revolution along some axis (i.e., it is round, $\lambda_i = \lambda_j, i \neq j$), convergence is quicker; as a limiting case, if the hyper-ellipsoid is an hyper-sphere, convergence is achieved in a single iteration.**

- **An example related to *n=2* is shown in the figure below:**



Coniugate Gradient Method

- ♦ **It should be considered a direct method since, in principle, it allows to obtain the solution in *n* iterations (*n* is the dimension of the system matrix)**
- ♦ **Actually, as for any other direct method, round-off errors may accumulate, reducing the method to a *semi-iterative* one.**
- ♦ **The adopted formulations are, in this case**

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \alpha^{(m)}\mathbf{p}^{(m)}$$
$$\mathbf{p}^{(m+1)} = \mathbf{r}^{(m+1)} + \beta^{(m)}\mathbf{p}^{(m)}$$
$$\mathbf{r}^{(m+1)} = \mathbf{r}^{(m)} - \alpha^{(m)}\mathbf{A}\,\mathbf{p}^{(m)}$$

where $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}$ and $\alpha^{(m)}$ and $\beta^{(m)}$ are given by

$$\alpha^{(m)} = \frac{\mathbf{p}^{(m)T}\mathbf{r}^{(m)}}{\mathbf{p}^{(m)T}\mathbf{A}\,\mathbf{p}^{(m)}} \qquad \beta^{(m)} = -\frac{\mathbf{r}^{(m+1)T}\mathbf{A}\,\mathbf{p}^{(m)}}{\mathbf{p}^{(m)T}\mathbf{A}\,\mathbf{p}^{(m)}}$$

- ♦ **As it can be noted, these definitions, whose demonstration we omit, though different from the ones of the steepest descent method, are anyway simple; their meaning is the following:**
  - ▪ **vectors $\mathbf{p}^{(m)}$ constitute a set of *A-orthogonal* vectors (or *A-conjugate,* hence the name of the method*)*:**

$$\mathbf{p}^{(m+1)T}\mathbf{A}\,\mathbf{p}^{(k)} = 0 \qquad (k = 0,1,\ldots,m)$$

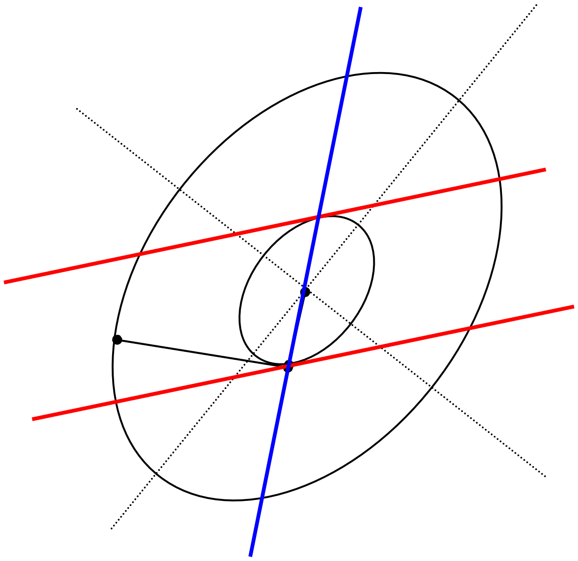
  - ▪ **with such definitions, the residuals $\mathbf{r}^{(m)}$, instead, turn out to be *orthogonal* among each other:**

$$\mathbf{r}^{(m+1)T}\mathbf{r}^{(k)} = 0 \qquad (k = 0,1,\ldots,m)$$

Since the n-th residual should be orthogonal to the others, $\mathbf{r}^{(0)}$, $\mathbf{r}^{(1)}$, …, $\mathbf{r}^{(n-1)}$, it should be zero.

$$\mathbf{r}^{(n)} = 0 \Rightarrow \qquad \mathbf{x}^{(n)} = \phi = exact\ \ solution$$

In the case *n*=2 the relation of A-conjugation between two directions defines the relation existing between the direction of two parallel tangent lines to an ellipsis and its diameter passing through the points of tangency: this illustrates the effectiveness of the method in finding the exact solution (i.e., the centre).



Modern versions of the conjugate gradient method can be applied to matrices the do not require to be SPD; the *bi-conjugate gradient method* is one such algorithm.

Sources and Suggested Readings

In English language

R. S. Varga, *Matrix Iterative Analysis*, Prentice Hall, 1962.

A. Quarteroni, R. Sacco, F. Saleri, *Numerical Mathematics*, Springer, 2007.

R. Bronson, *Matrix Operations*, McGraw-Hill, 1989.

In Italian language

G. Ghelardoni, P. Marzulli, *Argomenti di Analisi Numerica*, ETS Università, Pisa, 1980.

G. Gambolati, *Elementi di Calcolo Numerico*, Edizioni Libreria Cortina, Padova.

Web references

The Matrix Reference Manual, Mike Brookes, Imperial College, London, UK

http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html#Intro