Simulationsmethoden I WS 09 10

Lecture Notes

Finite element methods applied to solve PDE

Joan J. Cerdà [*]

December 14, 2009
ICP, Stuttgart

Contents

[*] `jcerda put here an at symbol icp.uni-stuttgart.de`

1 In this lecture we will talk about

1. FDM vs FEM.
2. Perspective: different ways of solving approximately a PDE.
3. Basic steps of any FEM intended to solve PDEs.
4. FEM in 1-D: heat equation for a cylindrical rod.
5. FEM in 2-D: the Poisson equation.
6. Sparse Matrixes (band matrixes) and FEM.
7. Other tricks for FEM and beyond.
8. Bibliography.

2 FDM vs FEM

1. With FEM you can handle better boundary problems with odd geometries.
2. With FEM it is easier to make more finer subdivisions of the space in regions where you need more accuracy. This is not so easy to do with FDM.

A word of caution: FEM as FDM are suitable for linear PDE's. If you have non-linear PDEs. You will have first to linearize it.

3 Perspective: different ways of solving approximately a PDE.

I have a PDE with certain bc (boundary conditions) to be solved, which options do I have:

1. Analytical solution: the best, but not always available.
2. Approximate solutions. There is always an error. The difference between the exact value and the approximate value is called the residual, we will denote it by R. For instance, let's suppose I have to solve the following ODE:

$$\frac{d\hat{\phi}}{dx} + \hat{\phi}(x) = 0 \tag{1}$$

if my approximate solution is $\phi(x) = 1 + a_1\ x + a_2\ x^2$, where a_1 and a_2 must be determined to get the ϕ as close to the exact solution $\hat{\phi}$ as I can. In that example the residual is

$$R(x; a_1, a_2) \equiv \frac{d\phi}{dx} + \phi(x) = 1\ +\ (1+x)\ a_1\ +\ (2x+x^2)\ a_2 \tag{2}$$

There are several ways of solving approximately a PDE, the most usual are:

1. **Ritz method (aka Goodman's method):** we get a_1 and a_2 via asking

$$\int_{Domain} R(x; a_1, a_2) dx = 0 \tag{3}$$

2. **Variational method (Rayleight-Ritz method):** from the PDE we get its variational integral. The function $\hat{\phi}$ that minimizes that variational integral is the solution to the problem. [FALTA]

3. **Weighted residual methods:** in this methods we assume the parameters a_1, a_2 (, ..., a_n if we have n in a general case) to be determined by asking the residual to obey a set of equations:

$$\int_{Domain} R(x; a_1, .., a_n)\ \omega_i(x)\ dx\ =\ 0 \tag{4}$$

   where $\omega_i(x)$ with $i = 1, ..n$ are n arbitrary functions called **weighting functions**. Some usual choices of $\omega_i(x)$ are

   a) **Collocation method:** we ask the residual R to be zero at n points $\{x_i\}_{i=1}^{i=n}$inside the domain. which is equivalent to say $\omega_i(x) = \delta(x - x_i)$.

   b) **Sub-domain method:** we split the domain in n regions and we ask the averaged R, $< R >$, to be zero inside each region.

   c) **Least-Squares method:** we set $\omega_i(x) = \frac{\partial R}{\partial a_i}$

   d) **Galerkin method:** if we use approximate solutions of the type $\phi(x) = 1 + a_1\ x + a_2\ x^2$, this means that $N_i(x) = x$ and $N_2(x) = x^2$ are a set of basis vectors for all our possible solutions. We set $\omega_i(x) = N_i(x)$,i.e., we have to solve n equations (in our example $n = 2$) like

$$\int_{Domain} R(x; a_1, .., a_n)\ N_i(x)\ dx\ =\ 0. \tag{5}$$

4. FDM: Finite Difference Methods.

5. FEM: Finite Element Methods.

6. FVM: Finite Volume Methods. A refined FDM popular in Computational fluid dynamics.

7. MOM:method of moments. You convert your differential equation into an integral equation. Used specially in electromagnetics.

4 Basic steps of any FEM intended to solve PDEs.

In all FEM variants there are always the same sequence of steps to be taken

1. **Discretize the continuum:** divide the solution into smaller regions that we call **elements**. The elements contain inside a certain number of points we call **nodes**. There are lots of shapes the elements can have. From segments of lines, triangles, squares, etc, to curved elements. The one/ones you will use depends on the problem you want to solve. For instance, for a 1D problem, like a cylindrical rod with radial symmetry, the most simple is to take the elements as linear segments with two nodes per segment (see Fig.1) and discretize it is as shown in figure 2. For a 2D problem, the most simple can be using triangular elements (see figure 3).

2. **Select the type of trial function to use, and in turn the shape functions:** We select what kind of functions we will take to describe the variation of the function ϕ inside each element (the trial function). This is equivalent to say, that we select the basis set of functions that will describe our solution. One of the usual choices is to take a polynomial like for instance $\phi(x) = a_0 + a_1 x$ (known as **linear element**) or $\phi(x) = a_0 + a_1 x + a_2 x^2$ (known as *quadratic element*). If we have n unknown coefficients $a_0, a_1, ..., a_{n-1}$ we will need the element to have n nodes to be able to determine them.

   Is it easy to show that the value of our trial function ϕ for a given position x inside an element can be written as a function of the values of ϕ at the N nodes the of element ($[\phi_1, \phi_2, ..., \phi_N]$) , i.e.

   $$\phi(x) = [\phi] \cdot [\mathbf{N}] = [\phi_\mathbf{1}, \phi_\mathbf{2}, ..., \phi_\mathbf{N}] \cdot [\mathbf{N_1(x)}, \mathbf{N_2(x)}, ..., \mathbf{N_N(x)}] \tag{6}$$

   where the $\{N_i(x)\}_{i=1}^{i=N}$ are known as the **Shape Functions**.

3. **The formulation:** Given the PDE you want to solve, now you must find a system of algebraic equations for each element "e" such that by solving it you got the values of of ϕ at the position of nodes of the element "e" ($[\phi_1, \phi_2, ..., \phi_N] \equiv [\phi]_e$), i.e., you must find for each element "e" the matrix $[\mathbf{K}]_e$ and the vector $[\mathbf{f}]_e$ such that,

   $$[\mathbf{K}]_e \cdot [\phi]_e = [\mathbf{f}]_\mathbf{e}. \tag{7}$$

   This is one of the more tricky parts of the FEM. There are different ways of getting the matrix $[\mathbf{K}]_e$, we will see it later.

4. **Assembling the equations for different elements**: one has to assemble the equations for all elements. The "problem" is that usually contiguous elements have nodes in common (see for instance figure 2, and therefore two algebraic equations may refer to the same node, and that has to be taken into account. At the end, if we have a total of M effective nodes[1] in the system, then we must build up a global matrix **[K]** of size $M \times M$ and a

[1] For instance in figure 1 we have a total of 5 elements that have each one two nodes, nonetheless due to the nodes in common, the number of effective nodes we have (the number of unknowns) is just $M = 6$ and not 10.

global vector $[\mathbf{f}]$ of size M such that the FEM problems "reduces" to solve the following matrix equation:

$$[\mathbf{K}] \cdot [\phi] = [\mathbf{f}]. \tag{8}$$

where $[\phi] \equiv [\phi_1, \phi_2, ..., \phi_M]$ is the value of the approximate solution ϕ at the position where the effective nodes are.

Notice that once we now $[\phi_1, \phi_2, ..., \phi_M]$ one is able to calculate the value of ϕ at whatever point x of the system, just make use of equation 6.

5. **Solve the system of equations:** In principle you can use whatever method you want, but the more the number of nodes the use, the better is the quality of the solution, the matrices we will build up can be very large, i.e. M becomes very large. As we will see, most of elements of matrix $[\mathbf{K}]$ are zero, *sparse matrix*, so you can save a lot of time and memory if you use special methods for sparse matrixes, we will see it later.

6. **Compute secondary quantities:** Once you know the ϕ values, you can compute other magnitudes using the values of ϕ.

Let's see now two examples about how this works in practice

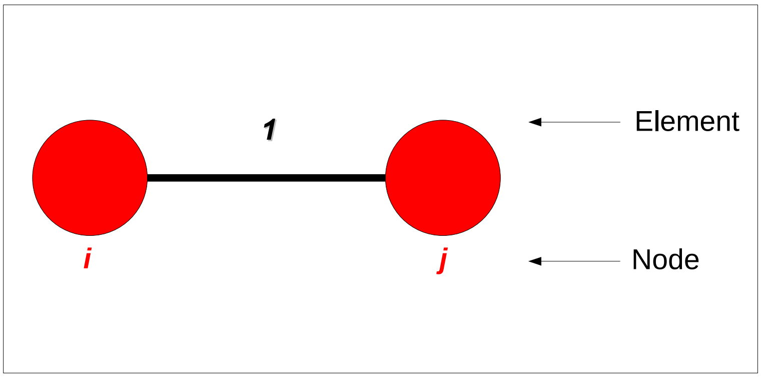

Figure 1: A single one dimensional element

5 FEM in 1-D: heat equation for a cylindrical rod.

In this section we plan to build up a very simple and basic one-dimensional FEM method.

Let's consider a cylindrical rod of radius R and length L with one end insulated an the other held to constant temperature T_{tip}, while the surrounding environment has a temperature T_{env} (see figure 2). We assume the cylinder lies along the x-axis, and the insulated tip is located at $x = 0$. The lost of heat occurs via the lateral surface of the cylinder at a rate characterized by the heat transfer coefficient h which units are $W/(m^2\ K)$, we assume that the thermal conductivity

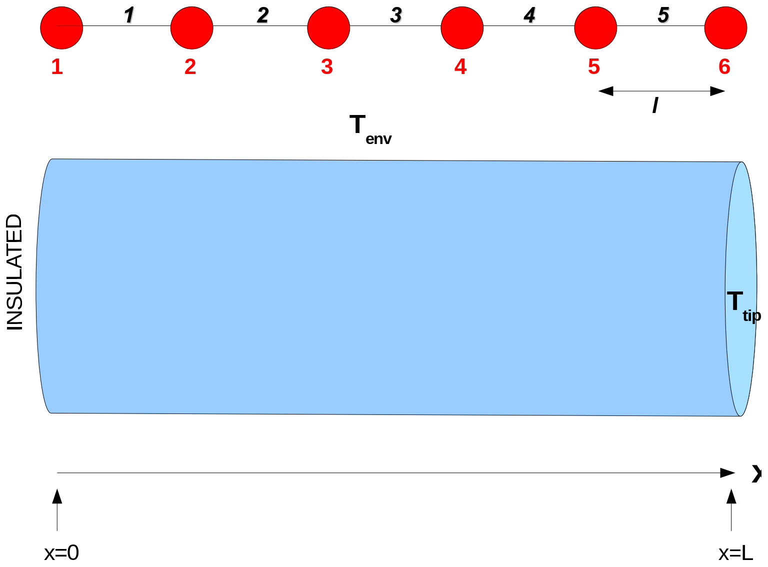

Figure 2: The 1D system we want to study using Galerkin FEM. The plot corresponds to using 5 one-dimensional linear elements (each element has 2 nodes), i.e. 6 effective nodes in total.

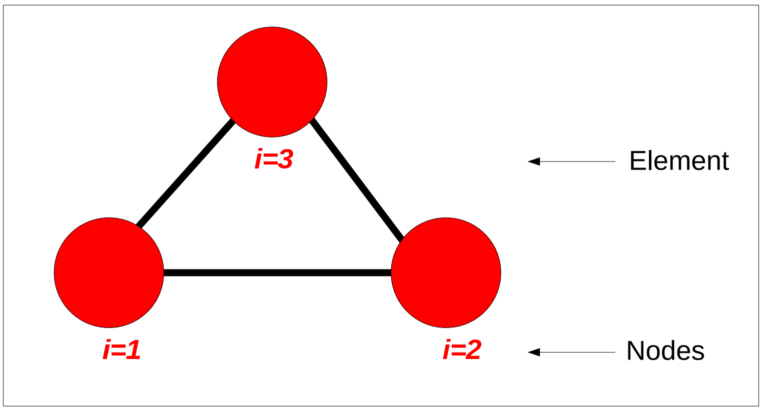

Figure 3: The most simple 2D element: a triangle.

for the bar is k^2. For such system, we can use the following coordinates

$$\begin{aligned} \hat{\theta} &= (T - T_{env}) && (10) \\ \eta &= \frac{x}{L} && (11) \\ \mu^2 &= L^2 \, \frac{2h}{kR} && (12) \end{aligned}$$

and it can be proved that the governing differential equation of the system is:

$$\frac{d^2\hat{\theta}}{d\eta^2} - \mu^2\hat{\theta} = 0 \qquad (13)$$

with the following boundary conditions

$$\begin{aligned} \frac{d\hat{\theta}}{d\eta} &= 0 \qquad at \;\; \eta = 0 && (14) \\ \hat{\theta} &= \hat{\theta}_{tip} = (T_{tip} - T_{env}) \qquad at \;\; \eta = 1. && (15) \end{aligned}$$

Proof: the heat balance on a differential volume of length dx gives, ($P = 2\pi R$,$A = \pi R^2$):

$$\begin{aligned} -kA \left.\frac{dT}{dx}\right|_x &= h \, P \, dx \, (T - T_{env}) - k \, A \, \left.\frac{dT}{dx}\right|_{x+dx} \\ &= h \, P \, dx \, (T - T_{env}) - k \, A \, \left.\frac{dT}{dx}\right|_x - k \, A \, \frac{d^2T}{dx^2} \, dx \qquad (16) \end{aligned}$$

We will use a FEM method known as *the Galerkin finite element method.* As we will see, the name comes from the request we do: we want the approximate solution to obey for each element the Galerkin approach in order to minimize the value of the residual. For our very simple one dimensional problem we choose the *elements* as a straight segments with two *nodes* located at the two ends of each segment, see figure 1. Therefore, we will split the rod bar in N elements, where two adjacent elements have always one node in common. Thus, for instance, in the case $N = 5$ elements, we will have a total number of $M = 6$ effective nodes (see figure 2). Let's call l to the length of each one of the elements. What we have done until now is the first step in any FEM problem, the **discretization of the continuum**: we divide the solution region into non-overlapping sub-regions that we call *elements*.

The second step in any FEM method is to choose which type of function will represent the variation of the field (in our case θ) along each element. This second stage is know as **Selection of the trial functions** which will lead us to determine which are the *shape functions* for our case

[2] The thermal conductivity k is the property of a material that indicates its ability to conduct heat. It arises in the Fourier´s law (aka law of heat conduction): $\mathbf{q} = -k \, \nabla T$, where $\mathbf{q}$ is the heat flux, i.e. the flow of energy per unit of area (units: W/m^2). The units of k are $W/(m \; K)$. Recall that the Integral form of the Fourier´s law is:

$$\frac{\partial Q}{\partial t} = -k \oint_S \nabla T \cdot \mathbf{dS}. \qquad (9)$$

(recall eq. 6) . The Shape functions are also known as *interpolation functions* or *basis functions*. Here, we will do the most basic assumption, i.e., for a segment with two nodes i and j which have values for θ function that are θ_i and θ_j, respectively, the change from one value to the other will be given by a linear function. Thus,

$$\theta(x) = a_1 + a_2\, x \tag{17}$$

where the constants a_1 and a_2 can be obtained via the two conditions

$$\begin{align} \theta_i &= a_1 + a_2\, x_i \tag{18}\\ \theta_j &= a_1 + a_2\, x_j \tag{19}\end{align}$$

it is simple to show that one can rewrite equation 17 as

$$\theta(x) = N_i(x)\,\theta_i + N_j(x)\,\theta_j \tag{20}$$

where

$$\begin{align} N_i(x) &= \frac{x_j - x}{l}, \tag{21}\\ N_j(x) &= \frac{x - x_i}{l}, \tag{22}\end{align}$$

are the so-called **Shape functions** for the 1D linear finite element. You should take into account that this shape functions, are defined to be strictly zero when we are out of the space assigned to the element associated to them. One should remark that using a first order polynomial like eq. 17 (*lineal element*) to account for the variation of the temperature θ inside an element is a quite coarse and rough approach. Thus for instance, one can assume a second degree polynomial like,

$$\theta(x) = a_1 + a_2\, x + a_3\, x^2 \tag{23}$$

but then, in order to calculate the three constants we need a third node to exist in the element (usually placed at the mid-point of the segment). Such type of element is know as a *quadratic element*. The quadratic element (or higher order elements), allow for a better description of what happens inside the element, but of course, the FEM method becomes more complex. At this moment we will keep things a simple as possible and we will assume we use a linear element such that equations 17 and 20 hold.

The third stage in any FEM method is know as the **Formulation** or the **obtaining of the element characteristics**. Basically in this step we determine the matrix equations that will govern the behavior of one single element. To do that, of course, we need to know which governing differential equation does the system (and therefore the element) obey, as well as what are the boundary conditions. In our simple case, the differential equation to obey and the conditions are given in eq. 13 and expressions 14 and 15. Before starting to derive the matrix equations for our element, notice that if equation 20 holds for our linear element, then

$$\frac{d\theta}{dx} = \frac{-1}{l}\theta_i + \frac{1}{l}\theta_j. \tag{24}$$

Because we assume θ within the element to be approached by a expression like eq. 20, we will always have an error respect what would be the exact solution. What we want is to minimize as far as possible the error when we compare the exact solution with the solution obtained via FEM. A nice method to get an approximate solution of the differential equation such that minimizes errors is known as the *Galerkin method* (and from there derives the name of the whole FEM method we are using). The Galerkin method applied to our case says that if our differential equation is eq. 13, our approximate solution θ will have the minimum possible error respect the exact solution if it obeys a set of equations:

$$\int_{rod} N_k \left(\frac{d^2\theta}{d\eta^2} - \mu^2 \theta \right) d\eta = 0 \tag{25}$$

where the number of equations is equal to the number of shape functions N_k we have (in our case, two per each element we use). Notice the shape functions N_k act as weighting functions. Given N_k is valid only for the element at which the node k belongs (notice that intermediate nodes will for one element act as the i-node, and for the neighbor node they will act as a j-node), we can reduce the integration over the whole rod to the integration over the element associated to the shape function N_k we consider.

Let's define $\eta_e \equiv l/L$. Because we have previously defined $\eta \equiv x/L$ (where L is the length of the rod), the following equalities hold for a linear element:

$$\begin{aligned} \frac{d\theta}{d\eta} &= L\,\frac{d\theta}{dx} = \frac{-1}{\eta_e}\theta_i + \frac{1}{\eta_e}\theta_j && (26)\\ N_i(\eta) &= 1 - \eta/\eta_e && (27)\\ N_j(\eta) &= \eta/\eta_e . && (28) \end{aligned}$$

Then, let's compute eq. 25 for the first element using the weighting functions $N_i(\eta)$ and $N_j(\eta)$ which are associated to such element. After a little bit of algebra involving the integration by parts, we get the following equations must be obeyed by the values of θ at nodes i and j of such element (i.e, θ_i and θ_j):

$$\begin{aligned} 0 &= \frac{1}{\eta_e}(\theta_i - \theta_j) + \frac{d\theta}{d\eta} + \frac{\mu^2\,\eta_e}{6}(2\theta_i + \theta_j) && (29)\\ 0 &= \frac{1}{\eta_e}(-\theta_i + \theta_j) - \frac{d\theta}{d\eta} + \frac{\mu^2\,\eta_e}{6}(\theta_i + 2\theta_j). && (30) \end{aligned}$$

The first equation corresponds to use N_i as a weighting function and the second equation is obtained when N_j is used as a weighting function in eq 25 (applied to the first element). Previous equations can be rewritten together in a matrix-like notation as:

$$\frac{1}{\eta_e}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\left\{ \begin{array}{c} \theta_i \\ \theta_j \end{array} \right\} + \frac{\mu^2\,\eta_e}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\left\{ \begin{array}{c} \theta_i \\ \theta_j \end{array} \right\} + \left\{ \begin{array}{c} \frac{d\theta}{d\eta} \\ -\frac{d\theta}{d\eta} \end{array} \right\} = \left\{ \begin{array}{c} 0 \\ 0 \end{array} \right\} \tag{31}$$

which is the **element characteristics** of the first element. Thus for instance, when $\mu^2 = 3$ and $\eta_e = 0.2$ (i.e., the system is split in 5 elements) the element characteristics becomes:

$$\begin{pmatrix} 5.2 & -4.9 \\ -4.9 & 5.2 \end{pmatrix}\left\{ \begin{array}{c} \theta_i \\ \theta_j \end{array} \right\} + \left\{ \begin{array}{c} \frac{d\theta}{d\eta} \\ -\frac{d\theta}{d\eta} \end{array} \right\} = \left\{ \begin{array}{c} 0 \\ 0 \end{array} \right\} \tag{32}$$

It is possible to show that for the other elements we can get a similar element characteristics.

Now in a fourth stage, we must **ensemble all the equations for all the elements**. This is very easily done in our 1D case as follows: if in a element characteristics matrix we have a row that refers to the same θ_x than a row in the characteristics matrix of another element, then we just add the two rows together. At the end we must get a matrix expression with as many rows as effective nodes we have. For instance, if we consider the first two elements (i.e. 3 nodes in total), the equations we get for the system will be:

$$\begin{pmatrix} 5.2 & -4.9 & 0 \\ -4.9 & 10.4 & -4.9 \\ 0 & -4.9 & 5.2 \end{pmatrix} \left\{ \begin{array}{c} \theta_1 \\ \theta_2 \\ \theta_3 \end{array} \right\} = \left\{ \begin{array}{c} 0 \\ 0 \\ \frac{d\theta}{d\eta} \end{array} \right\} \tag{33}$$

and if we consider the firsts three elements (4 nodes) the expression would be

$$\begin{pmatrix} 5.2 & -4.9 & 0 & 0 \\ -4.9 & 10.4 & -4.9 & 0 \\ 0 & -4.9 & 10.4 & -4.9 \\ 0 & 0 & -4.9 & 5.2 \end{pmatrix} \left\{ \begin{array}{c} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{array} \right\} = \left\{ \begin{array}{c} 0 \\ 0 \\ 0 \\ \frac{d\theta}{d\eta} \end{array} \right\} \tag{34}$$

and so on so forth. Notice that for the row that corresponds to θ_1 the $d\theta/d\eta$ is set to zero due to the boundary condition that we have for $\eta = 0$ (see eq.14).

Because the other boundary condition we have is that the last node should have $\theta = \theta_{tip}$, we can impose this directly in the system of equations to solve. Thus for instance, if we consider the 5 elements (6 effective nodes):

$$\begin{pmatrix} 5.2 & -4.9 & 0 & 0 & 0 & 0 \\ -4.9 & 10.4 & -4.9 & 0 & 0 & 0 \\ 0 & -4.9 & 10.4 & -4.9 & 0 & 0 \\ 0 & 0 & -4.9 & 10.4 & -4.9 & 0 \\ 0 & 0 & 0 & -4.9 & 10.4 & -4.9 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \left\{ \begin{array}{c} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \end{array} \right\} = \left\{ \begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \theta_{tip} \end{array} \right\}. \tag{35}$$

This last expression is the system of equations we have to solve to get the solution via the Galerkin FEM. The result can be generalized to whatever number of elements we want to have (but mind to change η_e). The solution of the system of algebraic equations is the fifth and final stage of the FEM (of course, later, once we know the θ values at each node, we can compute secondary quantities like for instance heat fluxes, etc.). The resolution of the system of equations, as in the case of the FDM method is done via one of the several techniques available. Notice that the matrix is sparse, in fact, band diagonal. The larger is the number of nodes we use, the more sparser becomes the matrix.

6 FEM in 2-D: the Poisson equation.

Let's now for a more complex case, a two dimensional problem. In this case, we want to solve the Poisson equation, namely,

$$\frac{\partial^2 \hat{V}}{\partial x^2} + \frac{\partial^2 \hat{V}}{\partial y^2} = -\rho(x, y). \tag{36}$$

In terms of the shape functions (still to be determined for this case) the trial solution $V(x, y)$ can be written as:

$$V(x,y) = \sum_e V_e(x,y) = \sum_e \sum_i v_{i,e} \, N_{i,e}(x,y) \tag{37}$$

where e is the index to denote the different *elements*, and i is the index to denote the different nodes inside the element e. $V_e(x, y)$ represents the trial solution associated to the element e. $N_{(i,e)}(x, y)$ is the shape function associated to the node i of the element e. The $v_{(i,e)}$ is the value of V (the potential) at the node i of the element e, .i.e., the stuff we want to know at the end of the day.

Let's take the most simple type of 2D element, i.e., a triangle element with 3 nodes located at the vertices of the triangular elements (see figure 3).

Let's take again linear shape functions, i.e., let's assume

$$N_{i,e}(x,y) = a_{i,e} + b_{i,e}\, x + c_{i,e}\, y \tag{38}$$

where $i = 1, 2, 3$ (we have 3 nodes). As in the 1D problem, we want each shape function to be 1 at the position of one of the nodes and zero on the position of the other nodes of the element. One can show that this can be obtained defining the Shape functions as the following determinants:

$$N_{1,e}(x,y) = \frac{1}{D_e} \begin{vmatrix} 1 & x & y \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}, \tag{39}$$

$$N_{2,e}(x,y) = \frac{1}{D_e} \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x & y \\ 1 & x_3 & y_3 \end{vmatrix}, \tag{40}$$

$$N_{3,e}(x,y) = \frac{1}{D_e} \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x & y \end{vmatrix}, \tag{41}$$

$$\tag{42}$$

where

$$D_e = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}. \tag{43}$$

and $[(x_1, y_1), (x_2, y_2), (x_3, y_3)]$ are the coordinates of the nodes (located at the vertices of the triangular element we use). In 3D the determinants will be 4×4 instead of 3×3 (we add a z column). Notice that you obtain the 1D linear elements also using the previous systematic way. Notice that by definition, the shape functions $N_{i,e}$ are zero out of the element e. Notice also that using the determinants notation, is very easy to check that they will be one at the node we want and zero at the other nodes.

Thus, once we know the position of the vertices of our triangular elements, using eqs. 39, 40 and 41 we can easily compute the coefficients $a_{i,e},\ b_{i,e},\ c_{i,e}$ that appear in equation 38. In this way, **given the positions of the vertices of our elements, we know the shape functions**.

A thing to take into account always is that **if our PDE to solve is of order m then the Shape functions we choose must be at least $m-1$ times derivable (i.e. different from zero)**, because as we will see later, we need up to the $m-1$ derivative of our shape functions to build-up the algebraic equations. If they were zero, then we would get an undetermined system of equations.

Now it comes the step of the **formulation**. In our 2D particular case, the residual is

$$R(x,y;[\mathbf{v}]) = \frac{\partial^2 V}{dx^2} + \frac{\partial^2 V}{dy^2} + \rho(x,y). \tag{44}$$

where V is given by equation 37, and $[\mathbf{v}]$ is the set of all the unknowns $v_{i,e}$ to be determined (the values of the potential in all the nodes we have). As in the 1D case, we choose the Galerkin method to minimize the value of the residual, i.e., we have to solve the set of equations:

$$\int\limits_{domain} N_{i,e}(x,y)\; R(x,y;[\mathbf{v}])\; dx\; dy = 0. \tag{45}$$

where we have an equation for each shape function $N_{i,e}(x,y)$ associated to an effective node. These equations can be rewritten as:

$$\int\limits_{e\ element} N_{i,e}(x,y)\; \left(\frac{\partial^2 V}{dx^2} + \frac{\partial^2 V}{dy^2}\right)\; dx\; dy \;\; + \int\limits_{e\ element} N_{i,e}(x,y)\; \rho(x,y)\; dx\; dy \;\; = 0. \tag{46}$$

Let's go for the first integral,

$$\int\limits_{e\ element} N_{i,e}(x,y)\; \frac{\partial^2 V}{dx^2}\; dx\; dy \tag{47}$$

it can be done by integrating by parts and we reduce the order of the partial derivatives from two to one because

$$\frac{\partial}{\partial x}\left\{N_{i,e}(x,y)\;\frac{\partial V}{\partial x}\right\} = \frac{\partial N_{i,e}(x,y)}{\partial x}\;\frac{\partial V}{\partial x} + N_{i,e}(x,y)\frac{\partial^2 V}{dx^2} \tag{48}$$

allows us to rewrite the integral with $N_{i,e}(x,y)\frac{\partial^2 V}{dx^2}$ as

$$\int\limits_{e\ element} \frac{\partial}{\partial x}\left\{N_{i,e}(x,y)\;\frac{\partial V}{\partial x}\right\}\; dx\; dy \;\; - \int\limits_{e\ element} \frac{\partial N_{i,e}(x,y)}{\partial x}\;\frac{\partial V}{\partial x}\; dx\; dy. \tag{49}$$

Now, using the divergence theorem in the plane[3], I can rewrite the first integral over the e-element as an integral over a closed path along the boundary of the element:

$$\int\limits_{e\ element} \frac{\partial}{\partial x}\left\{N_{i,e}(x,y)\;\frac{\partial V}{\partial x}\right\}\; dx\; dy = \oint\limits_{boundary\ of\ e} N_{i,e}(x,y)\;\frac{\partial V}{\partial x}\;\hat{n}_x\; dx\; dy \tag{51}$$

[3] See for instance the book Marsden-Tromba, *Vector calculus* . 3rd Edition, chap 8.1, ca. page 499.

$$\int\limits_{\partial Surface} \mathbf{F}\cdot\mathbf{n}\,dl \;=\; \int\limits_{Surface} \nabla\cdot\mathbf{F}\,dA \tag{50}$$

where $\hat{n}_x$ is the x component of the vector normal to the boundary of the element at each point. [FALTA i en els llocs on tenim el vertex com es defineix la normal ????]. A similar thing can be done for the y part of the integral and instead of $\hat{n}_x$ we will get a $\hat{n}_y$.

Despite this dependence with $\hat{n}_x$ and $\hat{n}_y$ looks at first sight like it will complicate our life, in fact it simplifies it a lot because those terms will cancel out among contiguous elements. The only kind of integrals of the style of eq. 51 that survive are those of those elements located at the boundary of the system (and therefore without a counter part), see figure 4.

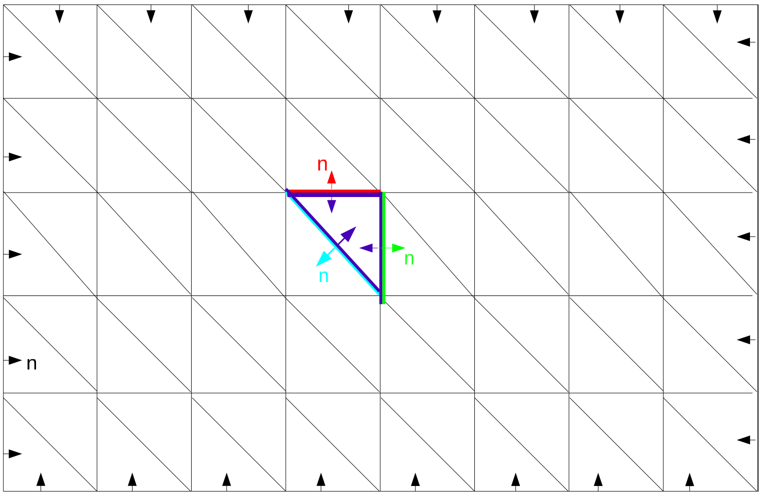

Figure 4: Suppose the following disposition of the triangular elements. The normals cancel out except at the boundaries , and so will do the contributions to the line integrals of the style of eq. 51.

For elements which are not in the boundary of the system (i.e., they are completely surrounded by other elements) it is easy to show that *the Galerkin condition applied to our element* e leads to three ($i = 1, 2, 3$) algebraic equations,

$$\frac{D_e}{2} \sum_j v_{j,e} \; (b_{i,e} \, b_{j,e} + c_{i,e} \, c_{j,e}) - \int_{e \; element} N_{i,e}(x, y) \; \rho(x, y) \; dx \; dy \; = \; 0 \tag{52}$$

where $j = 1, 2, 3$ represents also an index over the three different nodes we have in the element e, and $\frac{D_e}{2}$ is the area of the triangular element "e". If we write it in a matrix way:

$$\begin{aligned} [\mathbf{K}]_e[\mathbf{v}]_e &= [\mathbf{f}]_e, & (53)\\ K_{ij} &= \frac{D_e}{2} \; (b_{i,e} \, b_{j,e} + c_{i,e} \, c_{j,e}) \, , & (54)\\ f_i &= \int_{e \; element} N_{i,e}(x, y) \; \rho(x, y) \; dx \; dy. & (55)\\ & & (56) \end{aligned}$$

The different coefficients $b_{i,e}$, $b_{j,e}$ and $c_{i,e}$, $c_{j,e}$ are known quantities once we have defined the position of our nodes (use eqs. 39, 40 and 41).

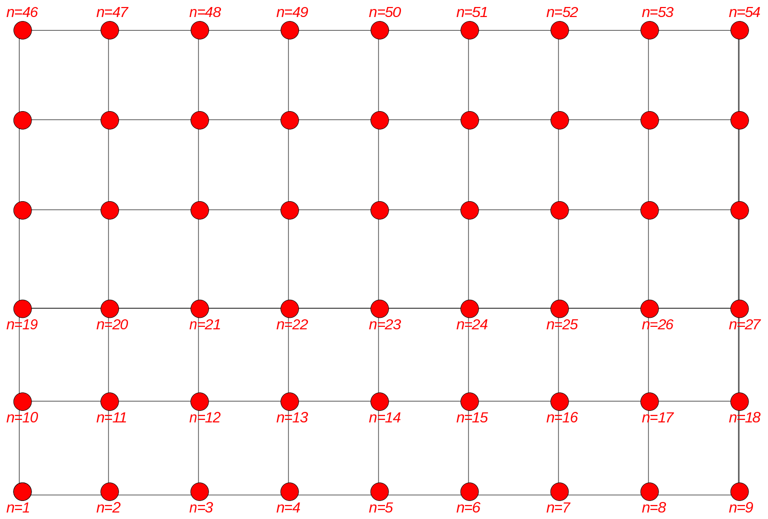

Figure 5: First step in the discretization of the 2D rectangular domain. We do a rectangular grid, and we assign a number n to the effective nodes as depicted in the plot.

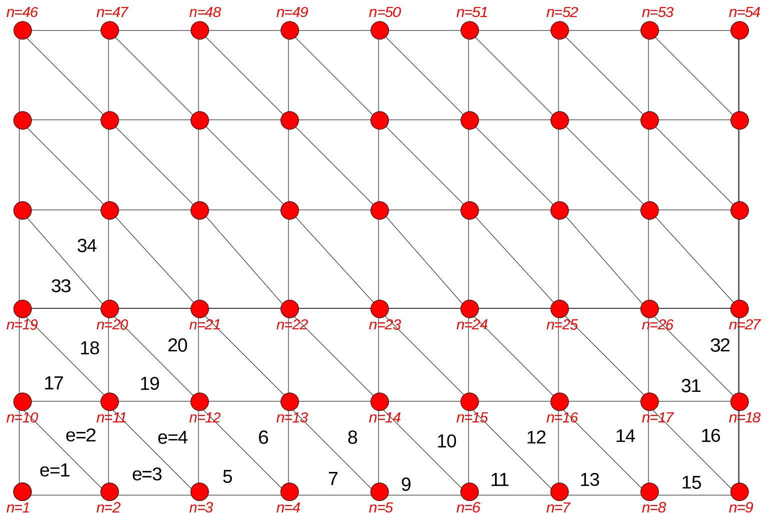

Figure 6: Second step in the discretization of the 2D rectangular domain. We create the triangle elements and we label them as depicted in the plot.

For elements in the boundary one has to add the terms of the style of eq. 51.

Now we know the equations for our elements (three for element), **we have to add the equations of all elements all together** taking into account that contiguous elements have nodes in common.

The most clean way of doing this is as follows: To give a particular example, let's suppose for simplicity we have to solve our Poisson equation onto a square domain. The boundary nodes those located in the perimeter of our square.

To obtain the global matrix [**K**] and the global vector [**f**], we do as follows:

- Discretize the square domain into a rectangular grid of $n = N_x \times N_y$ points (that will be our n *effective nodes*). See figure 5 .
- Now split the rectangular grid into triangular elements as shown in figure 6.
- The vertices in each triangle we will label them as $1, 2, 3$ in a clock-wise way, starting in the bottom left for odd elements, and up left for even elements.
- Notice the disposition of the odd and even labeled triangles. The mapping from the vertices of the triangle elements to the effective nodes involved is explained in tables 1 and 2.
- Now that we know the coordinates of the vertices for each element, we can compute the three shape functions $N_{i,e}$ for each element e. Make use of the determinant formulas we have seen before.
- Initialize the global matrix [**K**] and the global vector [**f**] with zeros.
- Now act as the following pseudocode says to build up the global matrix [**K**] and the global vector [**f**]:

```
For element=1  till element= the last element {
 for i=1 till i=3 {
  n = map_to_effective_node_the_vertice(i)
  Calc f_i  of the vector [f]_e see eq.53, store it as f
  for j=1 till j=3 {
    m = map_to_effective_node_the_vertice(j)
    Calc K_ij of the matrix [K]_e see eq.52, store it as K
    /* K_nm is the (n,m) element of the global matrix [K] */
    /* f_n  is the (n) element of the global vector [f] */
    if(n and m are not in the boundary) {
       K_nm += K
       f_n  += f
    } else {
      if( n is in the boundary) {
         set K_nn = 1
         f_n = v_n  ; /* if n is in the boundary, we know v_n */
```

```
          } else if (m is in the boundary) {
             f_n = f_n - K*v_m ; /* if m is in the boundary, we know v_m */
          }
        }
      }
    }
  }
```

- Now that you have the global matrix [**K**] and the global vector [**f**], just what remains to be done is to solve the system:

  $$[\mathbf{K}][\mathbf{v}] = [\mathbf{f}], \tag{57}$$

  in order to obtain the value of the potential V at position of the effective nodes [**v**].

Notice that in the pseudocode we use for the boundary nodes

```
if( n is in the boundary) {
   set K_nn = 1
   f_n = value for such point in the boundary
} else if (m in the boundary) {
   g_n = g_n - K*v_m
}
```

The first body of the if is to make sure that the values at the boundary nodes are fixed and are not changed. The second body (the "else if") is to incorporate the contributions from the line integrals of the nodes located at the boundary. The terms containing non zero line integral contributions can be easily added to the [**f**] vector (mind the minus sign due to the change of those terms from the left to the right hand side of the equations]. Thus, given a node n that has neighboring nodes m which are in the boundary, we have

$$f_n = f_n(arising\ from\ non\ bn)\ -\ \sum_{m}^{bn} K_{n,m}\, v_m \tag{58}$$

where bn stands for "boundary nodes near node n". Notice that because m is a node on the boundary, we now the value of v_m, and $K_{n,m}$ will contain also the result of the integral lines which can be easily evaluated.

Table 1: Odd labeled triangular elements. Let's suppose the triangle is in the rectangle of the α-th column and β-th row ($\alpha = 1$ and $\beta = 1$ is the rectangle at the bottom-left).

vertex num.	It has	Corresponds to
i	coordinates	the effective node
1	$(x_{\alpha-1}, y_{\beta-1})$	$n_1 \equiv (\beta - 1)N_x + \alpha$
2	$(x_{\alpha-1}, y_{\beta})$	$n_2 \equiv n_1 + N_x$
3	$(x_{\alpha}, y_{\beta-1})$	$n_3 \equiv n_1 + 1$

Table 2: Even labeled triangular elements. Let's suppose the triangle is in the rectangle of the α-th column and β-th row ($\alpha = 1$ and $\beta = 1$ is the rectangle at the bottom-left).

vertex num	It has	Corresponds to
i	coordinates	the effective node
1	$(x_{\alpha-1}, y_{\beta})$	n_2
2	(x_{α}, y_{β})	$n_2 + 1$
3	$(x_{\alpha}, y_{\beta-1})$	n_3

7 Sparse Matrixes (band matrixes) and FEM.

- A sparse matrix is a matrix with many elements that are zero.
- As we have seen in the 1D case, and is easy to see in 2D and 3D. FEM method leads to huge sparse matrix. This is because our node has direct connection only with its first neighbor nodes.
- The more the number of nodes, the larger the matrix to solve is. In 3D, this trouble is even more pronounced and the matrix is even more sparse.
- Storing zeros is a waste of memory. There are ways of storing matrices which have many zeros with a minimum amount of storage (See for instance Numerical Recipes, ca. Chapter 2.4). (Note: ca = circa (Latin) = around, near).
- If you solve the system
  $$[\mathbf{K}][\mathbf{v}] = [\mathbf{f}]. \tag{59}$$
  with methods like Gauss-Jordan. In the intermediate stage you will have huge matrices with non zeros. That's not nice, because you need to make many operations to solve the system, and you need huge amounts of memory.

- In fact, FEM leads usually to Band Diagonal Matrixes, i.e., the non-zero elements localize a few elements before, and after the diagonal. There are very efficient techniques to solve such band diagonal matrixes using minimum storage and minimum number of operations (fast calculation). (See for instance Numerical Recipes, ca. Chapter 2.4).

8 Other tricks for FEM and beyond.

- How to avoid the systematic addition of truncation errors by choosing the grid of nodes in a clever way. In this way, more accuracy without need of increasing the number of effective nodes in the system. This can be done by alternating the direction of the diagonal cuts. See plots in figure 7.

- A proper node labeling scheme can help to have a more narrow bandwidth in the matrixes. Which is important in order to have a faster solution of the matrix. The troubles with optimizing the labeling of the nodes is complex when we have complex geometries. There are programs and algorithms for doing such in an efficient way an reduce as maximum as possible the bandwidth of the matrixes.

- Notice that if you need to refine adding more elements in a certain zone because is the zone where you need more accuracy with FEM programs is very easy. An the the changes to do to the code of the program are minimum.

- The preparation of the FEM, selection of elements, mapping, etc. is very tedious. Nowadays there are excellent programs and algorithms intended to generate the meshes even for complex geometries.

- The use of higher orders than linear elements, is usually nice because with few high order terms you can save usually quite a lot of elements. The use of higher order elements is specially rewarding when the gradient of the field variable (θ in the 1D case, v in the 2D case) is expected to vary very rapidly. The use of the Pascal triangle can help to determine which should be the terms to include in higher orders in the 2D case.

- Notice that we have just worked out the FEM for bounded problems. For problems where no boundaries exist (e.g. scattering, radiation problems), there are special methods intended to solve such situations: use of infinite element methods, boundary element methods, or the adsorbing boundary condition.

- Notice also that we have not dealt with time dependent problems, but static ones. There are also FEM methods intended to solve problems to deal with time-dependent phenomena.

- The FEM method we have reviewed applies t finite linear equations. If you have a non linear PDE, you will have first to linearize it.

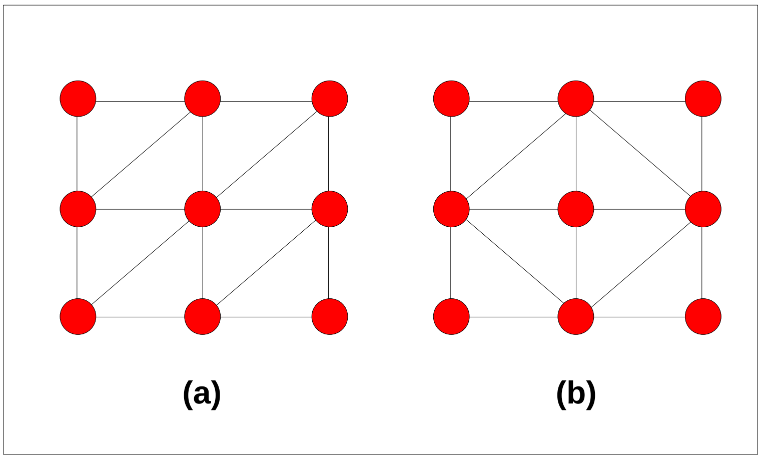

Figure 7: (a) All cuts in the same direction. (b) Alternance of the diagonal cuts to make easier truncation errors to cancel instead of adding up.

9 Bibliography

Suitable books to become more learned about FDM and FEM methods are:

- *Fundamentals of the Finite Element Method for Heat and Fluid Flow*, R.W. Lewis, P. Nithiarasu and K. N. Seetharamu. Wiley Ltd. (2004). ISBN: 0-470-84788-3.

- *Computational methods in physics and Engineering*, 2nd edition, Samuel S.M. Wong. World Scientific, (1997). ISBN:9810230176.

- *Numerical Techniques in Electromagnetics*, Matthew N. O. Sadiku. CRC Press (2001). ISBN: 0-8493-1395-3.

- *Numerical Recipes in C: The art of Scientific Computing*. Cambridge Univ. Press. (1992). ISBN: 0-521-43108-5.

There are also some nice web pages devoted to the finite element methods:

- The Internet finite element resources page: *http://homepage.usask.ca /ĩjm451/finite/ fe_resources/ fe_resources.html*