University of Bologna

Dept. of Electrical, Electronic and Information Engineering

Modelling and Computation of Electric and Magnetic Fields

Lecture notes

Prof. Andrea Cristofolini

ver. 2.10
September 30, 2019

Contents

Chapter 1

Vector algebra - A brief revision

1.1 Domains

A given metric space D is said connected if, for any two points A and B belonging to D, at least a path in D exists connecting A and B. A domain is an open connected subset of a finite-dimensional space. A domain D is said simply linearly-connected if, for any closed curve $\partial\Omega$ belonging to D, at least a surface S_Γ exists bounded by Γ and completely belonging to D. Alternatively, a domain D is said simply linearly-connected if any closed curve Γ can be reduced to a point by means of a continuous contraction.

A domain which is not simply connected is said multiple connected.

A domain D is said simply superficially-connected if, for any closed surface S belonging to D, the volume V_S bounded by S completely belongs to D. Alternatively, a domain D is said simply superficially-connected if any closed surface S can be reduced to a point by means of a continuous contraction.

A domain which is not simply superficially-connected is said multiply superficially-connected.

1.2 Vector operators

1.2.1 Gradient

The gradient is a first order vector differential operator describing the partial derivatives of a scalar function *f*. The gradient applied on a scalar differentiable function *f* yields a vector

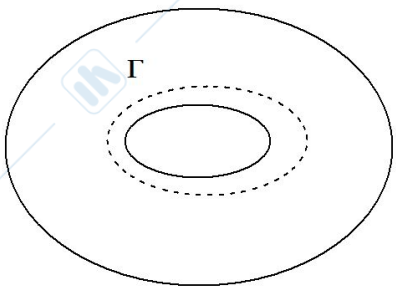


Figure 1.1: *The torus is a multiply linearly connected domain. Indeed, a surface S_Γ bounded by the closed curve Γ and entirely belonging to the torus can not be found. The torus is however simply superficially-connected.*

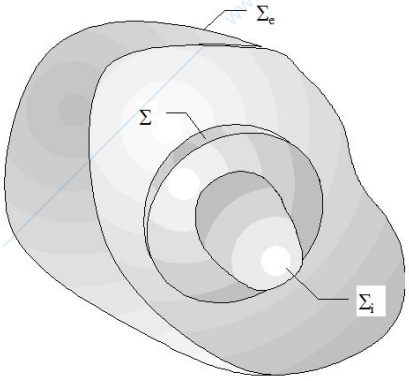


Figure 1.2: *The domain D, bounded by the surface Σ_e (Σ_e has been sectioned for sake of clarity), contains a cavity enclosed in the surface Σ_e, which does not belong to D. Therefore, the domain D is not a simply superficially-connected, as any closed surface Σ enclosing the cavity bounds a volume V which does not completely belongs to D.*

whose dot product with a $\hat{\boldsymbol{n}}$ unit vector is the directional derivative of f along the direction of $\hat{\boldsymbol{n}}$:

$$\nabla f \cdot \hat{\boldsymbol{n}} = \frac{\partial f}{\partial n} \tag{1.1}$$

In a cartesian three dimensional space, the gradient takes the following form:

$$\nabla f = \frac{\partial f}{\partial x}\hat{\boldsymbol{i}} + \frac{\partial f}{\partial y}\hat{\boldsymbol{k}} + \frac{\partial f}{\partial z}\hat{\boldsymbol{k}}. \tag{1.2}$$

The gradient direction is always perpendicular to the f function isolines. Indeed, said $\hat{\boldsymbol{t}}$ the unit vector tangent to the isoline of f at a point $\boldsymbol{p}$, the directional derivative of f along the direction of $\hat{\boldsymbol{t}}$ is equal to zero. Thus, for the definition of gradient (1.1), the dot product of ∇f with $\hat{\boldsymbol{t}}$ is equal to zero, that is ∇f is perpendicular to $\hat{\boldsymbol{t}}$:

$$\frac{\partial f}{\partial n} = 0 \implies \nabla f \cdot \hat{\boldsymbol{t}} = 0 \implies \nabla f \perp \hat{\boldsymbol{t}}$$

The line integral of a vector function $\boldsymbol{U}$ defined as the gradient of a scalar function f

$$\boldsymbol{U} = \nabla f \tag{1.3}$$

along a path Γ_{AB} between two points $\boldsymbol{A}$ and $\boldsymbol{B}$ is path independent and is defined only by the value of the function f at the endpoints of the integration path:

$$\int_{\Gamma_{AB}} \boldsymbol{U} \cdot \mathrm{d}\boldsymbol{l} = \int_{\Gamma_{AB}} \nabla f \cdot \mathrm{d}\boldsymbol{l} = \int_{\Gamma_{AB}} \frac{\partial f}{\partial l}\,\mathrm{d}l = f(\boldsymbol{B}) - f(\boldsymbol{A}).$$

As the circulation along a closed curve Γ can be regarded as line integral the two endpoints coincide, one can easily obtain:

$$\oint_{\Gamma} \boldsymbol{U} \cdot \mathrm{d}\boldsymbol{l} = \oint_{\Gamma} \nabla f \cdot \mathrm{d}\boldsymbol{l} = 0, \tag{1.4}$$

i.e. the gradient of a scalar function f is a circulation free vector function $\boldsymbol{U}$. A circulation free vector function $\boldsymbol{U}$ is said conservative. A vector field $\boldsymbol{U}$ derived as the gradient of a scalar function f is always conservative. Conversely, any conservative vector field $\boldsymbol{U}$ can be derived as the gradient of a scalar function f. The scalar function f is called scalar potential.

1.2.2 Divergence

The divergence is a first order vector differential operator describing the distribution in space of the sources or sink of a given vector field $\boldsymbol{U}$. In a region where a vector field $\boldsymbol{U}$ is defined, let us consider a volume V_S, containing a point $\boldsymbol{p}$ and bounded by a closed surface S. The divergence of the vector $\boldsymbol{U}$ at the point $\boldsymbol{p}$ is defined as the the ratio between the flux of $\boldsymbol{U}$ through the surface S and the volume V_S when V_S shrinks to the point $\boldsymbol{p}$:

$$\nabla \cdot \boldsymbol{U} = \lim_{V_S \to 0} \frac{\oint_{S_V} \boldsymbol{U} \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S}{V_S} \tag{1.5}$$

In a cartesian three dimensional space, the divergence takes the following form:

$$\nabla \cdot \boldsymbol{U} = \frac{\partial U_x}{\partial x} + \frac{\partial U_y}{\partial y} + \frac{\partial U_z}{\partial z}. \tag{1.6}$$

The divergence theorem (also known as Gauss theorem) provides an "integral" point of view of the definition (1.5). In a region where a vector field $\boldsymbol{U}$ is defined, considering a simply superficially-connected volume V, bounded by a closed surface S_V, the divergence theorem states that the net flux of the vector $\boldsymbol{U}$ through the surface S_V is equal to the integral of the divergence of $\boldsymbol{U}$ on the volume V:

$$\oint_{S_V} \boldsymbol{U} \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S = \int_V \nabla \cdot \boldsymbol{U} \, \mathrm{d}V \tag{1.7}$$

The divergence theorem can be regarded as a sort of "conservation law" for a vector field, stating that the total quantity of the vector $\boldsymbol{U}$ leaving the volume V (the flux passing through S_V in the the left hand side term in (1.7)) is equal to the sum of the source densities found in the volume V (the integral of the divergence of $\boldsymbol{U}$ on V in the right hand side term in (1.7)).

Let D be a domain in which a vector field $\boldsymbol{U}$ is defined. If the the condition:

$$\oint_S \boldsymbol{U} \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S = 0 \tag{1.8}$$

for any closed surface S, then the vector field $\boldsymbol{U}$ is solenoidal in D. The divergence theorem (1.7) allows to conclude that a solenoidal vector field in D is also a divergence free vector field:

$$\oint_S \boldsymbol{U} \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S = 0 \implies \nabla \cdot \boldsymbol{U} = 0, \quad \forall \boldsymbol{p} \in D$$

The opposite implication is necessarily true only if D is a simply superficially connected-domain.

1.2.3 Curl

The curl is a first order vector differential operator describing the tendency of a given vector field $\boldsymbol{U}$ to circulate a around a point $\boldsymbol{p}$. In a region where a vector field $\boldsymbol{U}$ is defined, let S be a surface, containing a point $\boldsymbol{p}$ and bounded by a closed curve Γ_S. The component of the curl of the vector $\boldsymbol{U}$ at the point $\boldsymbol{p}$ along the unit vector $\hat{\boldsymbol{n}}$ normal to the surface S is defined as the the ratio between the circulation of $\boldsymbol{U}$ along the curve Γ_S and the surface S when S shrinks to the point $\boldsymbol{p}$:

$$\nabla \times \boldsymbol{U} \cdot \hat{\boldsymbol{n}} = \lim_{S \to 0} \frac{\oint_{\Gamma_S} \boldsymbol{U} \cdot \mathrm{d}\boldsymbol{l}}{S} \tag{1.9}$$

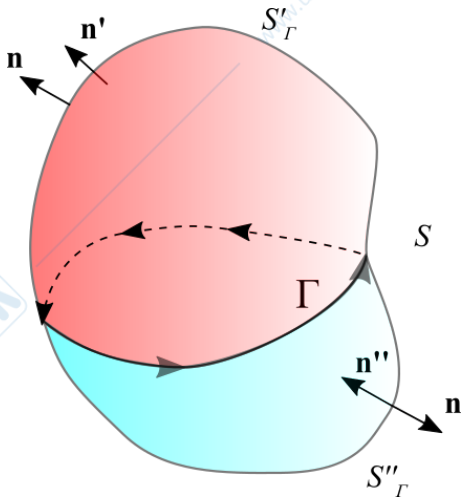


Figure 1.3: *Two surfaces S'_Γ and S_Γ" having the same boundary* Γ

Note that the sense of the unit vector $\hat{\boldsymbol{n}}$ is chosen on the basis of the orientation of the curve Γ_S according to the right hand rule. The Stokes' theorem provides an "integral" point of view of the definition (1.9). Let D be a simply connected domain where a vector field $\boldsymbol{U}$ is defined. Let S be a surface in D, bounded by a closed curve Γ_S. The Stokes' theorem states that the circulation of vector $\boldsymbol{U}$ along the curve Γ_S is equal to the flux of the curl of $\boldsymbol{U}$ through the surface S:

$$\oint_{\Gamma_S} \boldsymbol{U} \cdot \mathrm{d}\boldsymbol{l} = \int_S \nabla \times \boldsymbol{U} \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S \tag{1.10}$$

The Stokes' theorem has an important implication on vector field $\boldsymbol{F}$ derived as the curl of a vector $\boldsymbol{U}$:

$$\boldsymbol{F} = \nabla \times \boldsymbol{U}. \tag{1.11}$$

Indeed, if (1.11) holds on a domain D, the Stokes' theorem (1.10) ensures that the fluxes of $\boldsymbol{F}$ through surfaces S sharing the same boundary Γ are the same. In other words, the total flux of the vector $\boldsymbol{F}$ enclosed by the closed curve Γ does not depend on the surface S (bounded by Γ) on which the flux is actually evaluated, but is determined only by on the closed curve Γ. If this is the case, i.e. if the flux of $\boldsymbol{F}$ enclosed by Γ does not depend on the surface S, one can demonstrate that $\boldsymbol{F}$ is a solenoidal field.

Now, let S'_Γ and S_Γ" be two surfaces sharing the same boundary Γ, and let $\hat{\boldsymbol{n}}'$ and $\hat{\boldsymbol{n}}$" be the normal unit vectors of S'_Γ and S_Γ" respectively (see Fig. 1.3). The Stoke's theorem allows to state that the fluxes of the vector $\boldsymbol{F}$ defined in (1.11) through S'_Γ and S_Γ" are equal:

$$\begin{aligned}
\int_{S'_\Gamma} \boldsymbol{F} \cdot \hat{\boldsymbol{n}}' \, \mathrm{d}S &= \int_{S'_\Gamma} \nabla \times \boldsymbol{U} \cdot \hat{\boldsymbol{n}}' \, \mathrm{d}S = \oint_\Gamma \boldsymbol{U} \cdot \mathrm{d}\boldsymbol{l} \\
\int_{S_\Gamma"} \boldsymbol{F} \cdot \hat{\boldsymbol{n}}" \, \mathrm{d}S &= \int_{S_\Gamma"} \nabla \times \boldsymbol{U} \cdot \hat{\boldsymbol{n}}" \, \mathrm{d}S = \oint_\Gamma \boldsymbol{U} \cdot \mathrm{d}\boldsymbol{l} \\
\Longrightarrow \int_{S'_\Gamma} \boldsymbol{F} \cdot \hat{\boldsymbol{n}}' \, \mathrm{d}S &= \int_{S_\Gamma"} \boldsymbol{F} \cdot \hat{\boldsymbol{n}}" \, \mathrm{d}S
\end{aligned} \tag{1.12}$$

Now, let S be the closed surfaces obtained by joining the surfaces S' and S" along their common boundary Γ. While the normal unit vector $\hat{\boldsymbol{n}}'$ of the closed surface S points outward, $\hat{\boldsymbol{n}}'$ and $\hat{\boldsymbol{n}}$" must both comply with the orientation of the curve Γ. As a result:

$$\begin{aligned}
\hat{\boldsymbol{n}} &= \hat{\boldsymbol{n}}' \quad \text{on } S'_\Gamma \\
\hat{\boldsymbol{n}} &= -\hat{\boldsymbol{n}}" \quad \text{on } S_\Gamma"
\end{aligned}$$

By splitting the flux of $\boldsymbol{F}$ through the closed surface S in the two contributions through S' and S'', one obtains:

$$\oint_S \boldsymbol{F}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = \int_{S'_\Gamma} \boldsymbol{F}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S + \int_{S_\Gamma{}''} \boldsymbol{F}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = \int_{S'_\Gamma} \boldsymbol{F}\cdot\hat{\boldsymbol{n}'}\,\mathrm{d}S - \int_{S_\Gamma{}''} \boldsymbol{F}\cdot\hat{\boldsymbol{n}''}\,\mathrm{d}S,$$

and, for (1.12):

$$\oint_S \boldsymbol{F}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = 0. \tag{1.13}$$

Given the arbitrariness of the curve Γ and of the surfaces S' and S'', the result (1.13) can be extended to any closed surface S. According to the definition (1.8), a vector field $\boldsymbol{F}$ derived as the curl of a vector $\boldsymbol{U}$ is therefore a solenoidal field. Conversely, a solenoidal vector field can always be obtained as the curl of a vector function $\boldsymbol{U}$. The vector function $\boldsymbol{U}$ is called vector potential.

Another fundamental implication of the Stoke's theorem concerns the so called curl free vector field, that is function whose curl is equal to zero on a given domain. Let $\boldsymbol{U}$ be a curl free vector field defined on a simply linearly connected domain D:

$$\nabla\times\boldsymbol{U} = 0 \tag{1.14}$$

The Stoke's theorem ensures then that, for any closed curve Γ in D:

$$\oint_\Gamma \boldsymbol{U}\cdot\mathrm{d}\boldsymbol{l} = \int_S \nabla\times\boldsymbol{U}\,\mathrm{d}S = 0, \tag{1.15}$$

where S is a surface bounded by S and belonging to D. According to (1.4), the result (1.15) points out that the curl free vector field $\boldsymbol{U}$ is also a conservative field, and can be derived as the gradient of a scalar potential field f.

1.3 Solenoidal and conservative fields

As we have seen, a vector field $\boldsymbol{U}$ is called conservative on a given domain D when:

$$\oint_\Gamma \boldsymbol{U}\cdot\mathrm{d}\boldsymbol{l} = 0,$$

for any closed path Γ in D. A conservative vector field $\boldsymbol{U}$ is also a curl free field, that is:

$$\nabla\times\boldsymbol{U} = 0,$$

must hold on the entire domain. A curl free vector field is necessarily a conservative field only if the domain D is simply linearly connected. A conservative function can be derived as the gradient of a scalar potential:

$$\boldsymbol{U} = \nabla\cdot f \tag{1.16}$$

The scalar potential f is defined up to a constant, as the vector field $\boldsymbol{U}$ in (1.16) can be expressed as the gradient of

$$f^* = f + cost$$

where *cost* is an arbitrary constant uniformly defined on the entire domain. Let $\boldsymbol{U}$ be a conservative vector field, defined on a simply linearly connected domain D. Generally speaking, the function $\boldsymbol{U}$ will be completely determined by the following differential problem:

$$\begin{cases} \nabla\cdot\boldsymbol{U}=t(\boldsymbol{p}), \\ \nabla\times\boldsymbol{U}=0, \end{cases}$$

along with some boundary conditions. By expressing the conservative field $\boldsymbol{U}$ as the gradient of a scalar potential (1.16), one obtain the scalar Poisson's equation:

$$\nabla^2 f = t(\boldsymbol{p}), \tag{1.17}$$

where the ∇^2 is the second order differential operator called "Laplace operator" or laplacian" or "nabla":

$$\nabla^2 f = \nabla \cdot \nabla f. \tag{1.18}$$

In a cartesian three dimensional space, the nabla operator takes the following form:

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}. \tag{1.19}$$

A vector field $\boldsymbol{U}$ is called solenoidal on a given domain D when:

$$\oint_\Gamma \boldsymbol{U} \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S = 0,$$

for any closed surface § in D. A solenoidal vector field $\boldsymbol{U}$ is also a divergence free field, that is

$$\nabla \cdot \boldsymbol{U} = 0,$$

must hold on the entire domain. If the domain D is also simply superficially connected, a divergence free field is necessarily a solenoidal field. A solenoidal vector field can be expressed as the curl of a vector potential:

$$\boldsymbol{U} = \nabla \times \boldsymbol{F} \tag{1.20}$$

The vector potential f is defined up to a curl free vector function, which can be derived as the gradient of a scalar function. Thus the vector field $\boldsymbol{U}$ in (1.20) can also be expressed as the curl of

$$\boldsymbol{F}^* = \boldsymbol{F} + \nabla f.$$

where f is an arbitrary scalar function defined on the entire domain. Let $\boldsymbol{U}$ be a solenoidal vector field, defined on a simply superficially connected domain D. The function $\boldsymbol{U}$ is then completely determined by the following differential problem:

$$\begin{cases} \nabla \cdot \boldsymbol{U} = 0, \\ \nabla \times \boldsymbol{U} = T(\boldsymbol{p}), \end{cases}$$

along with some boundary conditions. By expressing the solenoidal field $\boldsymbol{U}$ as the curl of a vector potential (1.20), one obtains:

$$\nabla \times \nabla \times \boldsymbol{F} = \boldsymbol{T}(\boldsymbol{p}) \tag{1.21}$$

By introducing the following vector equation:

$$\nabla \times \nabla \times \boldsymbol{F} = \nabla(\nabla \cdot \boldsymbol{F}) - \nabla^2 \boldsymbol{F} \tag{1.22}$$

which define the "vector nabla" or vector laplacian" operator, and arbitrarily assuming a divergence free vector potential $\boldsymbol{F}$, one finally obtains the vector Poisson's equation:

$$\nabla^2 \boldsymbol{F} = -\boldsymbol{T}(\boldsymbol{p}), \tag{1.23}$$

In a cartesian three dimensional space, the second order differential operator ∇^2 applied to the vector field $\boldsymbol{F}$ here takes the following form:

$$\nabla^2 \boldsymbol{F} = \nabla^2 F_x \hat{\boldsymbol{i}} + \nabla^2 F_y \hat{\boldsymbol{j}} + \nabla^2 F_z \hat{\boldsymbol{k}}. \tag{1.24}$$

A conservative and solenoidal vector field will satisfy, on a simply (linearly and superficially) connected domain, the following differential problem:

$$\begin{cases} \nabla \cdot \boldsymbol{U} = 0, \\ \nabla \times \boldsymbol{U} = 0, \end{cases}$$

The scalar and the vector potential will be then solution of the scalar and the vector Laplace equation:

$$\nabla^2 f = 0 \tag{1.25}$$
$$\nabla^2 \boldsymbol{F} = 0 \tag{1.26}$$

A function satisfying the Laplace equation on a given domain is called Harmonic.

1.4 The Clebsch's theorem

Let $\boldsymbol{U}$ be a vector field defined on a simply (linearly and superficially) connected domain D:

$$\begin{cases} \nabla \cdot \boldsymbol{U} = t(\boldsymbol{p}), \\ \nabla \times \boldsymbol{U} = \boldsymbol{T}(\boldsymbol{p}). \end{cases} \tag{1.27}$$

Then, the field $\boldsymbol{U}$ can be expressed as the sum of a solenoidal component $\boldsymbol{U}_s$ and of a conservative component $\boldsymbol{U}_c$. Indeed, let $\boldsymbol{U}_c$ be a vector function such as:

$$\begin{cases} \nabla \cdot \boldsymbol{U}_c = t(\boldsymbol{p}), \\ \nabla \times \boldsymbol{U}_c = 0. \end{cases} \tag{1.28}$$

and:

$$\boldsymbol{U} = \boldsymbol{U}_c + \boldsymbol{U}_s. \tag{1.29}$$

Then, due to the linearity of the curl and divergence operators, we can write:

$$\begin{aligned} \boldsymbol{U}_s &= \boldsymbol{U} - \boldsymbol{U}_c. \\ \nabla \cdot \boldsymbol{U}_s &= \nabla \cdot \boldsymbol{U} - \nabla \cdot \boldsymbol{U}_c = t(\boldsymbol{p}) - t(\boldsymbol{p}) = 0 \\ \nabla \times \boldsymbol{U}_s &= \nabla \times \boldsymbol{U} - \nabla \times \boldsymbol{U}_c = \boldsymbol{T}(\boldsymbol{p}) - 0 = \boldsymbol{T}(\boldsymbol{p}) \end{aligned}$$

Therefore, the vector function $\boldsymbol{U}_s$ is a solenoidal function defined by the following differential problem:

$$\begin{cases} \nabla \cdot \boldsymbol{U}_s = 0, \\ \nabla \times \boldsymbol{U}_s = \boldsymbol{T}(\boldsymbol{p}). \end{cases} \tag{1.30}$$

and the original field $\boldsymbol{U}$ defined by (1.27) can be decomposed in the partial fields $\boldsymbol{U}_c$, conservative and satisfying (1.28) and $\boldsymbol{U}_s$, solenoidal and satisfying (1.30). The conservative field $\boldsymbol{U}_c$ can be derived as the gradient of a scalar potential f, while the solenoidal field $\boldsymbol{U}_s$ can be expressed as the curl of a vector potential $\boldsymbol{F}$. Thus, according to (1.29) the function $\boldsymbol{U}$ can be rewritten as:

$$\boldsymbol{U} = \nabla f + \nabla \times \boldsymbol{F}, \tag{1.31}$$

and the scalar and vector potential will be a solution of the scalar and vector Poisson's equations:

$$\begin{aligned} \nabla^2 f &= t(\boldsymbol{p}). \\ \nabla^2 \boldsymbol{F} &= \boldsymbol{T}(\boldsymbol{p}). \end{aligned}$$

Chapter 2

Fundamentals of Electromagnetics

2.1 Fundamental quantities in electromagnetism

2.1.1 Electric field and magnetic induction field

In a inertial reference system, the force $\boldsymbol{F}$ acting on an electric charge q and moving with a velocity $\boldsymbol{v}$ is given by the well known Lorentz equation:

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}). \tag{2.1}$$

Equation (2.1) can be considered as a definition of the electric field $\boldsymbol{E}$ and the magnetic flux density (also known as magnetic induction field) $\boldsymbol{B}$. So, We will say that an electric field $\boldsymbol{E}$ exists in a given region if any electrically charged particle at rest is subjected to a force $\boldsymbol{F}$ which is proportional to the electric charge. The electric field is then a vector equal to the force acting on the unit charge at rest, or:

$$\boldsymbol{E} = \frac{\boldsymbol{F}}{q}. \tag{2.2}$$

In the International System (SI), the electric charge is measured in coulombs (C) and the electric field is measured in volts per meter (V/m).

The presence of a magnetic flux density on a given region becomes evident on a charge q moving with a velocity $\boldsymbol{v}$. In this case, we will say that a magnetic induction field $\boldsymbol{B}$ exists if the electrically charged particle experiences a force $\boldsymbol{F}$, which is proportional both to the electric charge and to the velocity and directed in a perpendicular direction both to the velocity and to the magnetic induction field. Naming $v_\perp$ the velocity component perpendicular to the magnetic induction field, one can write:

$$B = \frac{F}{qv_\perp}. \tag{2.3}$$

that is, the magnetic induction magnitude is equal to magnitude of the force acting on a unitary charge which is moving with a unitary velocity $v_\perp$. The force $\boldsymbol{F}$, as said, is perpendicular both to the to the velocity $\boldsymbol{v}$ and to the magnetic induction field $\boldsymbol{B}$, and its direction and sense are determined by the right hand rule according to (2.1). In SI units, the magnetic induction field is measured in teslas (T).

2.1.2 Electric polarisation and magnetization in matter

Let ΔV be a volume in a given material, containing the point $\boldsymbol{x}$. Let $\Delta \boldsymbol{p}$ be the total electric dipole moment in the volume ΔV, that is the resultant of all the electric dipole moments $\boldsymbol{p}_i$ observed in ΔV due to the separation of bounded electric charges (see Fig. 2.1)

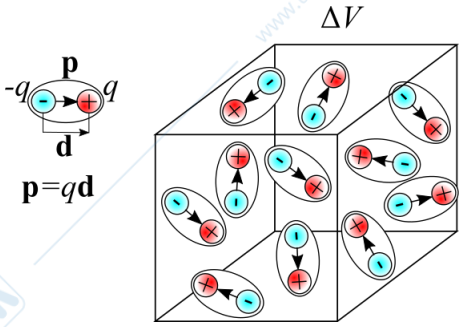


Figure 2.1: *Electric polarisation in a material. An electric dipole occurs when two charges q of opposite sign are separated by a distance* $\boldsymbol{d}$ *(oriented toward the positive charge). An electric dipole is associated to a moment* $boldsymbolp = q\boldsymbol{d}$*. Polarisation provides a macroscopic description of the electric dipole moment distribution in a medium*

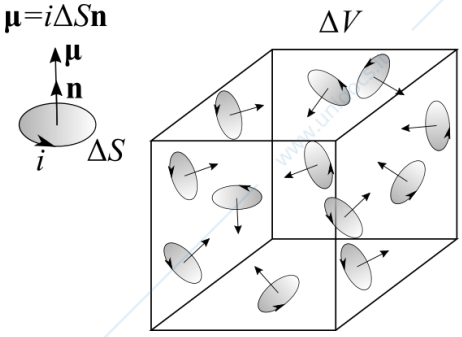


Figure 2.2: *Magnetisation in a material. An electric current i flowing around a surface* ΔS *produce a magnetic dipole moment* $\mu = i\Delta S\hat{\boldsymbol{n}}$*, where* $\hat{\boldsymbol{n}}$ *is the unit vector normal to* ΔS*. Magnetisation provides a macroscopic description of the magnetic dipole moment distribution in a medium*

$$\Delta \boldsymbol{p} = \sum_{\Delta V} \boldsymbol{p}_i. \tag{2.4}$$

Then, the electric polarisation at the point $\boldsymbol{x}$ is defined as the density of the total electric dipole moment when the volume ΔV shrinks to the point $\boldsymbol{x}$:

$$\boldsymbol{P} = \lim_{\Delta V \to 0} \frac{\Delta \boldsymbol{p}}{\Delta V}. \tag{2.5}$$

It should be noted that the limit in (2.5), while being perfectly admissible by the mathematical point of view, should be taken with some precautions. Indeed, as the distribution of the electric dipole moment in matter is not a continuous function in space (electric dipoles are localized where a separation of bounded electrical charges occurs, i.e. in molecules and atoms which constitute the matter), one can easily recognize that, shrinking ΔV to $\boldsymbol{x}$, one ends up obtaining a volume that does not contain any molecules at all, an thus has no inherent total electric dipole moment. Thus, the definition (2.5) should be taken as a way to represent on a "macroscopic" scale the average effects of a number of physical phenomena occurring on a "microscopic" scale. The volume ΔV should then shrink to the point $\boldsymbol{x}$ in "macroscopic" sense, so that ΔV is "very small" in the macroscopic scale. However ΔV should be still large enough in the microscopic scale to contain a sufficient number of dipole moments and to produce an average indication of the matter behaviour around the point $\boldsymbol{x}$.

The electric dipole moment and the polarization measure units in SI are coulomb meter (Cm) and coulomb per square meter (C/m^2) respectively. Magnetization in a material can be defined in a analogous way. Let ΔV be a volume in the material, containing the point $\boldsymbol{x}$

(see Fig. 2.2). Let $\Delta \boldsymbol{m}$ be the total magnetic dipole moment in the volume ΔV, that is the resultant of all the microscopic magnetic dipole moments $\boldsymbol{m}_i$ observed in ΔV.

$$\Delta \boldsymbol{m} = \sum_{\Delta V} \boldsymbol{m}_i. \tag{2.6}$$

The magnetization (or magnetic polarization) $\boldsymbol{M}$ at the point $\boldsymbol{x}$ is then defined as the limit of the magnetic dipole density when ΔV shrinks to $\boldsymbol{x}$

$$\boldsymbol{M} = \lim_{\Delta V \to 0} \frac{\Delta \boldsymbol{m}}{\Delta V}. \tag{2.7}$$

The magnetization $\boldsymbol{M}$ is a vector representing in the macroscopic scale the average effect of phenomena which occur on microscopic scale (electrons orbital motion around nuclei and electron spins). The limit appearing in (2.7) should be taken as a "macroscopic" one, in the sense that ΔV has to shrink to the point $\boldsymbol{x}$ in macroscopic sense, but remains large enough in the microscopic scale to contain a sufficient number of magnetic dipole moments and to produce an average indication of the matter behaviour around the point $\boldsymbol{x}$.

In SI, the magnetic dipole moment and the magnetization are measured in amperes square meter (Am^2) and amperes per meter (A/m) respectively.

2.1.3 Electric displacement and magnetic field

The electric displacement field $\boldsymbol{D}$ can be defined by the following equation:

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}. \tag{2.8}$$

The constant ϵ_0 is the free space permittivity (or vacuum permittivity, or electric constant). In SI units, $\epsilon_0 = 8.85418810^{-12}$ farads per meter (F/m). The electric displacement field is measured in the same units (C/m^2) as the electric polarisation.

In linear and isotropic materials, the electric polarization $\boldsymbol{P}$ has the same direction of the electric field $\boldsymbol{E}$ (which polarizes the material); polarization magnitude is also proportional to the magnitude of the electric field:

$$\boldsymbol{P} = \epsilon_0 \chi_e \boldsymbol{E}, \tag{2.9}$$

where the electric susceptibility χ_e is an adimensional scalar quantity which provides a measure of the extent to which a material electrically polarizes as a reaction to a given applied electric field.

Introducing (2.9) in (2.8), one obtains:

$$\begin{aligned} \boldsymbol{D} &= \epsilon_0 \boldsymbol{E} + \epsilon_0 \chi_e \boldsymbol{E}, \\ \Longrightarrow \boldsymbol{D} &= \epsilon_0 (1 + \chi_e) \boldsymbol{E}. \end{aligned} \tag{2.10}$$

Defining the material permittivity (or electric constant) ϵ as:

$$\epsilon = \epsilon_0 (1 + \chi_e), \tag{2.11}$$

and the material relative permittivity (often referred to as dielectric constant) $\epsilon_r = \epsilon / \epsilon_0$ as:

$$\epsilon_r = (1 + \chi_e), \tag{2.12}$$

one may rewrite (2.10) as follows:

$$\boldsymbol{D} = \epsilon_0 \epsilon_r \boldsymbol{E} \tag{2.13}$$

or

$$\boldsymbol{D} = \epsilon \boldsymbol{E}. \tag{2.14}$$

Most of dielectrics can be treated as homogeneous linear isotropic materials. The relative permittivity χ_e of solid dielectrics typically ranges between 2 and 10. Materials whose molecules exhibit a strong electric dipole moment are characterized by higher permittivity (the relative permittivity of water is about 80). The equations (2.9) and (2.10) can be modified to take into account of more complex behaviours. In anisotropic materials, the susceptibility, and thus the permittivity, are second rank tensors. Non-linear materials require an appropriate constitutive relations for defining the dependence of between $\boldsymbol{D}$ on $\boldsymbol{E}$:

$$\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{E}). \tag{2.15}$$

Analogously, the magnetic field $\boldsymbol{H}$ is defined as follows:

$$\boldsymbol{H} = \frac{\boldsymbol{B}}{\mu_0} - \boldsymbol{M}, \tag{2.16}$$

that is

$$\boldsymbol{B} = \mu_0(\boldsymbol{H} + \boldsymbol{M}). \tag{2.17}$$

The quantity μ_0 is the vacuum permeability (or free space permeability, or magnetic constant). In SI units, $\mu_0 = 4\pi 10^{-7} = 1.25663710^{-6}$ henries per meter (H/m). The magnetic field is measured in the same units (A/m) as the magnetisation.

In linear and isotropic medium, the magnetization $\boldsymbol{M}$ is related to the magnetic field $\boldsymbol{H}$ by a linear relation:

$$\boldsymbol{M} = \chi_m \boldsymbol{H}. \tag{2.18}$$

The magnetic susceptibility χ_m is an adimensional scalar quantity which gives a measure of how easily a material magnetizes when a given magnetic field is applied. Substituting (2.18) in (2.17), one gets:

$$\boldsymbol{B} = \mu_0(1 + \chi_m)\boldsymbol{H}. \tag{2.19}$$

Then, introducing the material permeability (or magnetic constant) μ:

$$\mu = \mu_0(1 + \chi_m), \tag{2.20}$$

and the material relative permeability $\mu_r = \mu/\mu_0$:

$$\mu_r = (1 + \chi_m), \tag{2.21}$$

equation (2.19) ca be rewritten as:

$$\boldsymbol{B} = \mu_0 \mu_r \boldsymbol{H} \tag{2.22}$$

or

$$\boldsymbol{B} = \mu \boldsymbol{H}. \tag{2.23}$$

A linear relation between $\boldsymbol{B}$ and $\boldsymbol{H}$ is in many cases sufficient to describe the magnetic properties of materials. Linear anisotropic materials are described by equations similar to (2.19) (2.22), and (2.23), where the and the material permeability are second rank tensors. The magnetic behaviour of non-linear materials is of course described by a non linear equation relating $\boldsymbol{B}$ and $\boldsymbol{H}$:

$$\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{H}). \tag{2.24}$$

The magnetic properties of materials will be discussed in detail in a future chapter.

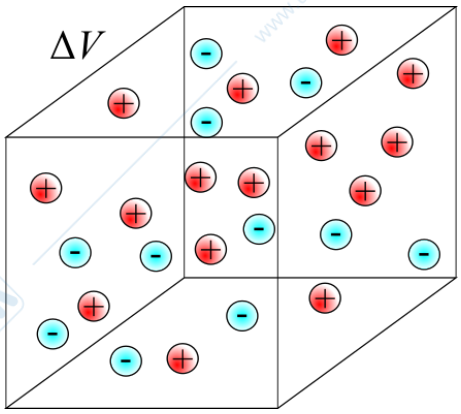


Figure 2.3: *Charge density in a medium. The total electric charge in the volume ΔV is obtained by an algebraic sum of the electric charges contained in ΔV.*

2.1.4 Charge density and electric current density

Let ΔV be a volume in a given material, containing the point $\boldsymbol{x}$. Let $N_{\Delta V,s}$ be the number of particles of a given species s in the volume ΔV. The number density of the species s in $\boldsymbol{p}$ is defined as:

$$n_s = \lim_{\Delta V \to 0} \frac{N_{\Delta V,s}}{\Delta V}. \tag{2.25}$$

Let ΔQ be the total electric charge in the volume ΔV, that is the algebraic sum of all the electric charges q_i observed in ΔV (see Fig. 2.3):

$$\Delta \boldsymbol{Q} = \sum_{\Delta V} q_i. \tag{2.26}$$

The volume charge density ρ at the point $\boldsymbol{x}$ is then defined as the limit of the ratio between ΔQ and ΔV when ΔV shrinks to the point $\boldsymbol{x}$

$$\rho = \lim_{\Delta V \to 0} \frac{\Delta Q}{\Delta V}. \tag{2.27}$$

Surface charge density ρ_Σ can be defined in a similar way. Given a surface ΔS, containing the point $\boldsymbol{x}$:

$$\rho_\Sigma = \lim_{\Delta S \to 0} \frac{\Delta Q_\Sigma}{\Delta S}. \tag{2.28}$$

where ΔQ_Σ is the net electric charge observed on the surface ΔS. Again, the limits appearing in (2.25), (2.27) and in (2.28) should be taken in "macroscopic" sense, being the volume ΔV and the surface ΔS small enough to provide a punctual informations in the macroscopic scale, but large enough in the microscopic scale to contain a sufficient number of electric charges and to produce an average indication of the net electric charge around the point $\boldsymbol{x}$. Now, let ΔS be a surface containing the point $\boldsymbol{x}$ and let $\hat{\boldsymbol{n}}$ be the unit vector normal to ΔS. Let ΔQ be the net electric charge crossing the surface ΔS during the time interval Δt. Then, the electric current density is defined as follows:

$$\boldsymbol{J} \cdot \hat{\boldsymbol{n}} = \lim_{\substack{\Delta S \to 0 \\ \Delta t \to 0}} \frac{\Delta Q}{\Delta S \Delta t}. \tag{2.29}$$

Also in this case, the limit in (2.29) is a "macroscopic" limit, as the surface ΔS and the time Δt have to be small enough in order to provide a punctual and instantaneous indication in

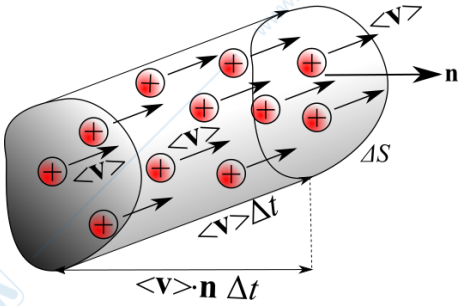


Figure 2.4: *Current density due to a collective motion of electric charges*

the macroscopic scale, but also large enough in the microscopic scale to allow the passage of a statistically relevant number of electric charges and to produce an average indication of the net electric charge flow around the point $\boldsymbol{x}$. The current density describe the collective motion of electric free charges.

In order do give an adequate expression of the electric current density, let us suppose that in a given material there is a collective motion of charges. Let us consider a given set of N particles, which move with a velocity $\boldsymbol{v}_i$ each, and are contained in a volume ΔV. The average velocity can be defined as:

$$< \boldsymbol{v} >= \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{v}_i. \tag{2.30}$$

Now, let ΔS be a surface and let $\hat{\boldsymbol{n}}$ be its normal unit vector; let Δt be a generic time interval. We construct a cylinder having ΔS as a base and the vector $< \boldsymbol{v} > \Delta t$ as a generatrix. Since the height of the cylinder is $< \boldsymbol{v} > \cdot \hat{\boldsymbol{n}} \Delta t$, the volume ΔV of the cylinder is $\Delta V = \Delta S < \boldsymbol{v} > \cdot \hat{\boldsymbol{n}} \Delta t$. We observe that, since during the time interval Δt each charged particle travels a distance $< \boldsymbol{v} > \Delta t$, any particle contained in the cylinder at a given time t will have passed through the surface ΔS at the time $t + \Delta t$. Being n_q the number density of the charged particles in the cylinder volume and q their electric charge, we can conclude that the total charge ΔQ that passes through the surface ΔS during Δt is:

$$\Delta Q = q n_q \Delta V = q n_q < \boldsymbol{v} > \Delta S \, \Delta t.$$

Comparing with the current density definition (2.29), one finally obtains:

$$\boldsymbol{J} = q n_q < \boldsymbol{v} > . \tag{2.31}$$

In metallic conductors, the current density is due to the drift motion of electrons, and then:

$$\boldsymbol{J} = -e n_e < \boldsymbol{v}_e > .$$

where e, n_e and $< \boldsymbol{v}_e >$ are the electron charge, the electron number density and the electrons average velocity respectively. There are materials in which the charge transport is due to two or more different species of charge carriers. For example, in electrolytic solutions and plasmas, two charge carriers are present (ions and electrons). Defining n_i and $< \boldsymbol{v}_i >$ as the ions number density and the ions average velocity respectively, the current density can be expressed as:

$$\boldsymbol{J} = e n_i < \boldsymbol{v}_i > - e n_e < \boldsymbol{v}_e > .$$

In linear and isotropic materials, the electric current density is related to the electric field by the following consitutive relation, often referred to as local Ohm's law:

$$\boldsymbol{J} = \sigma (\boldsymbol{E} + \boldsymbol{E}_i). \tag{2.32}$$

The term $\boldsymbol{E}_i$ represents the electromotive fields, i.e. fields that are not electric, but are still capable of causing a drift of the free charges (for example, electrochemical fields, temperature or electron density gradient...). The quantity σ is called conductivity, and describe macroscopically how easily a material let free electric charges drift. In SI units, the conductivity is measured in mhos per meter ($mhos\, m^{-1}$ or $(\Omega m)^{-1}$)

2.2 Fundamental laws of electromagnetics

2.2.1 The Ampere-Maxwell's law

The Ampere-Maxwell's law describes how a magnetic field $\boldsymbol{H}$ is generated. Let Γ be a closed curve; then:

$$\oint_\Gamma \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{l} = i_T \tag{2.33}$$

The right hand side in (2.33) is the total current enclosed by Γ, that is the flux of the total current density $\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}$ through any surface S_Γ bounded by Γ:

$$i_T = \int_{S_\Gamma} \left(\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t} \right) \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S. \tag{2.34}$$

It should be noted that the flux which defines i_T in the right hand side of (2.34) does not depend on the surface S_Γ on which the integral is evaluated, but only on the closed curve Γ bounding S_Γ. As explained in 1.2.3, this implies that the total current density is a solenoidal vector:

$$\oint_S \left(\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t} \right) \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S = 0. \tag{2.35}$$

where S is any closed surface. The total current density can be decomposed in two terms:

$$i_c = \int_{S_\Gamma} \boldsymbol{J} \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S, \tag{2.36}$$

$$i_d = \int_{S_\Gamma} \frac{\partial \boldsymbol{D}}{\partial t} \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S. \tag{2.37}$$

$$i_T = i_c + i_d. \tag{2.38}$$

The conduction current i_c, given by the flux of the current density $\boldsymbol{J}$, represents the drift motion of the free electric charges in a conductive medium. The i_d term is called displacement current, and is produced by a time varying displacement field. The displacement current is the mechanism which allows the total current to pass through the dielectric separating two plates of a capacitor.

It should be noted that, while the total current density is a solenoidal vector, the two terms $\boldsymbol{J}$ and $\frac{\partial \boldsymbol{D}}{\partial t}$ which constitute it are not necessarily solenoidal. In SI units, the electric currents i_t, i_c and i_d are measured in amperes (A).

As said, (2.33) states that the quantity i_T produces a magnetic field $\boldsymbol{H}$, which "wraps" its lines around the total current density. The Stokes theorem provides a way to formulate the Ampere-Maxwell's law in a local (or differential) form:

$$\nabla \times \boldsymbol{H} = J + \frac{\partial \boldsymbol{D}}{\partial t} \tag{2.39}$$

Applying the divergence theorems, the differential form of (2.35) can be found:

$$\nabla \cdot \left(\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t} \right) = 0. \tag{2.40}$$

2.2.2 The Faraday Neumann Lenz's law

The Faraday Newmann Lenz's law is a physical law that describes the phenomenon of electromagnetic induction, and relates the electric field with the rate of change of the magnetic induction field. Let Γ be a closed curve; then:

$$\oint_{\Gamma} \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{l} = -\frac{\mathrm{d}\Phi_{enc}}{\mathrm{d}t} \tag{2.41}$$

where the magnetic induction flux Φ_{enc} enclosed by Γ is defined as the flux of the magnetic induction field $\boldsymbol{B}$ through any surface S_Γ bounded by Γ:

$$\Phi_{enc} = \int_{S_\Gamma} \boldsymbol{B} \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S. \tag{2.42}$$

The flux of the magnetic induction field does not depend on the actual surface S_Γ on which the integral appearing in (2.42) is evaluated, but only on the closed curve Γ bounding S_Γ. Then, for the properties discussed in 1.2.3, the magnetic induction field $\boldsymbol{B}$ is a solenoidal vector:

$$\oint_{S} \boldsymbol{B} \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S = 0. \tag{2.43}$$

The Stokes theorem allows to obtain a differential form of (2.41):

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \tag{2.44}$$

The differential form of (2.43) can be found applying the divergence theorems to (2.43):

$$\nabla \cdot \boldsymbol{B} = 0. \tag{2.45}$$

2.2.3 Electrical charge conservation

The electrical charge conservation principle is formally expressed by the continuity equation. Let S be a closed surface, then the integral formulation of the continuity equation for the electric charge is:

$$\oint_{S} \boldsymbol{J} \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S = -\frac{\mathrm{d}Q_{enc}}{\mathrm{d}t}. \tag{2.46}$$

The right hand side term Q_{enc} is the total electric charge enclosed in the surface S. Assuming that only a volume charge density is present in the volume V_S bounded by S (and surface charge distributions are absent), Q_{enc} can be expressed as:

$$Q_{enc} = \int_{V_S} \rho \, \mathrm{d}V. \tag{2.47}$$

The continuity equation states that the electric charge is a conserved quantity, and can neither be created nor destroyed. Then the flux on the left hand side of (2.46), which represents the net charge flowing through S outward V_S in the time unit, is associated with the decrease in time of the total charge enclosed in S appearing in the right hand side of (2.46). Applying the divergence theorem to (2.46), one obtain the differential formulation of the continuity equation:

$$\nabla \cdot \boldsymbol{J} = -\frac{\partial \rho}{\partial t}. \tag{2.48}$$

2.2.4 Gauss' law

Gauss' law states that the electric charge produces a displacement field $\boldsymbol{D}$. For a closed surface S, the integral formulation of the Gauss' law is:

$$\oint_S \boldsymbol{D} \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S = Q_{enc}. \tag{2.49}$$

where Q_{enc} is again the total electric charge enclosed in the surface S. Assuming that only a volume charge density is present in the volume V_S, Q_{enc} is expressed by (2.47). The differential form of the Gauss' law is obtained applying the divergence theorem to (2.49):

$$\nabla \cdot \boldsymbol{D} = \rho. \tag{2.50}$$

2.3 Poynting theorem

Poynting's theorem allows us to define an important property of electromagnetic fields. Let V_S be a volume bounded by the closed surface S. For our convenience, let's rewrite the two of the fundamental law of electromagnetism we have discussed in the previous section:

$$\begin{aligned} \nabla \times \boldsymbol{E} &= -\frac{\partial \boldsymbol{B}}{\partial t} \\ \nabla \times \boldsymbol{H} &= \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t} \end{aligned}$$

that are the Faraday Neumann Lenz equation and the Ampere maxwell equation, both written in differential form. Dot-multiplying both sides of the first equation by $\boldsymbol{H}$, and both sides of the second one by $\boldsymbol{E}$, one obtains:

$$\boldsymbol{H} \cdot \nabla \times \boldsymbol{E} = -\boldsymbol{H} \cdot \frac{\partial \boldsymbol{B}}{\partial t} \tag{2.51}$$

$$\boldsymbol{E} \cdot \nabla \times \boldsymbol{H} = \boldsymbol{E} \cdot \boldsymbol{J} + \boldsymbol{E} \cdot \frac{\partial \boldsymbol{D}}{\partial t} \tag{2.52}$$

Subtracting (2.52) from (2.51) and recalling the vector identity:

$$\nabla \cdot (\boldsymbol{F} \times \boldsymbol{G}) = \boldsymbol{G} \cdot \nabla \times \boldsymbol{F} - \boldsymbol{F} \cdot \nabla \times \boldsymbol{G}.$$

one obtains:

$$\nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}) = -\boldsymbol{H} \cdot \frac{\partial \boldsymbol{B}}{\partial t} - \boldsymbol{E} \cdot \frac{\partial \boldsymbol{D}}{\partial t} - \boldsymbol{E} \cdot \boldsymbol{J} \tag{2.53}$$

Assuming that the materials in the region V_S are isotropic and linear, one can use the linear constitutive relation (2.14), (2.23) and (2.32) for manipulating the terms appearing on the right hand side of (2.53):

$$\boldsymbol{H} \cdot \frac{\partial \boldsymbol{B}}{\partial t} = \frac{1}{2\mu} \frac{\partial B^2}{\partial t}; \tag{2.54}$$

$$\boldsymbol{E} \cdot \frac{\partial \boldsymbol{D}}{\partial t} = \frac{1}{2} \epsilon \frac{\partial E^2}{\partial t}; \tag{2.55}$$

$$\boldsymbol{E} \cdot \boldsymbol{J} = \frac{1}{\sigma} J^2 - \boldsymbol{E}_i \cdot \boldsymbol{J}. \tag{2.56}$$

Then, inserting (2.54), (2.55) and (2.56) in (2.53), we can write:

$$\nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}) = -\frac{1}{2\mu}\frac{\partial B^2}{\partial t} - \frac{1}{2}\epsilon\frac{\partial E^2}{\partial t} - \frac{1}{\sigma}J^2 + \boldsymbol{E}_i \cdot \boldsymbol{J} \tag{2.57}$$

Equation (2.57) can be integrated on the domain V_S:

$$\int_{V_S} \nabla \cdot (\boldsymbol{E} \times \boldsymbol{H}) = -\frac{\mathrm{d}}{\mathrm{d}t}\int_{V_S}\frac{B^2}{2\mu}\,\mathrm{d}V - \frac{\mathrm{d}}{\mathrm{d}t}\int_{V_S}\frac{1}{2}\epsilon E^2\,\mathrm{d}V - \int_{V_S}\frac{1}{\sigma}J^2\,\mathrm{d}V + \int_{V_S}\boldsymbol{E}_i \cdot \boldsymbol{J}\,\mathrm{d}V \tag{2.58}$$

Applying now the divergence theorem (1.7) to the left hand side of (2.58) and reordering, one finally obtains:

$$\int_{V_S}\boldsymbol{E}_i \cdot \boldsymbol{J}\,\mathrm{d}V = \int_{V_S}\frac{1}{\sigma}J^2\,\mathrm{d}V + \frac{\mathrm{d}}{\mathrm{d}t}\int_{V_S}\frac{B^2}{2\mu}\,\mathrm{d}V + \frac{\mathrm{d}}{\mathrm{d}t}\int_{V_S}\frac{1}{2}\epsilon E^2\,\mathrm{d}V + \oint_S (\boldsymbol{E} \times \boldsymbol{H}) \cdot \hat{\boldsymbol{n}}\,\mathrm{d}S \tag{2.59}$$

Equation (2.59) is the formulation of the Poynting theorem for isotropic and homogeneous materials. In order to understand the physical meaning of the Poynting theorem, we briefly analyse each term appearing in (2.59). The term on the left hand side of (2.59) represents the total work per unit time done by the electromotive fields, that is the power produced by non-electrical forces acting on the free charges:

$$p_g = \int_{V_S}\boldsymbol{E}_i \cdot \boldsymbol{J}\,\mathrm{d}V \tag{2.60}$$

The first term in the right hand side of (2.59) represents the total energy per unit time dissipated by Joule effect, that is the power lost by the free charges due to collisions:

$$p_J = \int_{V_S}\frac{1}{\sigma}J^2\,\mathrm{d}V \tag{2.61}$$

The kernel $B^2/2\mu$ of the integral appearing in the second term in the right hand side of (2.59) is dimensionally an energy density, distributed in space due to the presence of a magnetic field. Integrating this kernel, one obtain the total magnetic energy in the volume V_S:

$$E_m = \int_{V_S}\frac{B^2}{2\mu}\,\mathrm{d}V \tag{2.62}$$

Analogously, the kernel $\epsilon E^2/2$ which appear in the integral in the third term at the right hand side of (2.59) is an energy density, distributed in space due to the presence of an electric field. The integration of this kernel represents then the total electrostatic energy in the volume V_S:

$$E_e = \int_{V_S}\frac{1}{2}\epsilon E^2\,\mathrm{d}V \tag{2.63}$$

The kernel $\boldsymbol{E}\times\boldsymbol{H}$ of the last integral in the right hand side of (2.59) is dimensionally a power per surface unit, and is called Poynting vector. The flux of the Poynting vector through the closed surface S represents the electromagnetic energy leaving the volume V_S through its boundary S:

$$p_P = \oint_S (\boldsymbol{E} \times \boldsymbol{H}) \cdot \hat{\boldsymbol{n}}\,\mathrm{d}S \tag{2.64}$$

Equation (2.59) is then a balance of the electromagnetic energy in the volume V_S. Indeed, (2.59) states that the electromagnetic power produced in V_S by the electromotive fields is partly dissipated in V_S by the Joule effect, partly produces a time variation of the magnetic and electrostatic energies in V_S, and partly leaves V_S through S:

$$p_g = p_J + \frac{\mathrm{d}}{\mathrm{d}t}(E_m + E_e) + p_P \tag{2.65}$$

2.4 Uniqueness theorem for the electromagnetic field

We are now interested in finding out what conditions should be met in order to formulate an electromagnetic problem that admits a unique solution. The Poynting theorem can be utilised for this purpose. Let V_S be a region bounded by the closed surface S. Let all the materials in V_S be isotropic and linear, and let the material properties ϵ, μ and σ be assigned (to strictly positive value) on the entire domain V_S. Then, the electromagnetic field (expressed by the fields $\boldsymbol{E}$, $\boldsymbol{D}$, $\boldsymbol{H}$, $\boldsymbol{B}$ and $\boldsymbol{J}$) is uniquely determined at any time t after the initial time t_0 if:

1. (Source terms): the distribution of the impressed electromotive field $\boldsymbol{E}_i$ over V_S is known for any time $t > t_0$.

2. (Boundary conditions): the tangential electric field $\boldsymbol{E}_t$ over S is known, or the tangential magnetic field $\boldsymbol{H}_t$ over S is known, or the tangential electric field $\boldsymbol{E}_t$ over part of S and the tangential magnetic field $\boldsymbol{H}_t$ over the rest of S are known, for any time $t > t_0$.

3. (Initial conditions): the electromagnetic field distribution is assigned at the initial time t_0.

To prove the theorem, we assume that there are two distinct solutions that obey the Maxwell equations (2.39), (2.44) and that meet the conditions set out in points 1), 2) and 3). Let us denote these solutions by $\boldsymbol{E}_1$, $\boldsymbol{D}_1$, $\boldsymbol{H}_1$, $\boldsymbol{B}_1$, $\boldsymbol{J}_1$ and by $\boldsymbol{E}_2$, $\boldsymbol{D}_2$, $\boldsymbol{H}_2$, $\boldsymbol{B}_2$, $\boldsymbol{J}_2$ respectively. Then, the difference between the two solutions can be defined as:

$$\begin{aligned}
\boldsymbol{E}_3 &= \boldsymbol{E}_1 - \boldsymbol{E}_2; \\
\boldsymbol{D}_3 &= \boldsymbol{D}_1 - \boldsymbol{D}_2; \\
\boldsymbol{H}_3 &= \boldsymbol{H}_1 - \boldsymbol{H}_2; \\
\boldsymbol{B}_3 &= \boldsymbol{B}_1 - \boldsymbol{B}_2; \\
\boldsymbol{J}_3 &= \boldsymbol{J}_1 - \boldsymbol{J}_2.
\end{aligned}$$

As the fields ($\boldsymbol{E}_1$, $\boldsymbol{D}_1$, $\boldsymbol{H}_1$, $\boldsymbol{B}_1$, $\boldsymbol{J}_1$) and ($\boldsymbol{E}_2$, $\boldsymbol{D}_2$, $\boldsymbol{H}_2$, $\boldsymbol{B}_2$, $\boldsymbol{J}_2$) both satisfy equations (2.39), (2.44), due to the linearity of the operators involved, we can write:

$$\nabla \times \boldsymbol{E}_3 = -\frac{\partial \boldsymbol{B}_3}{\partial t}; \tag{2.66}$$

$$\nabla \times \boldsymbol{H}_3 = \boldsymbol{J}_3 + \frac{\partial \boldsymbol{D}_3}{\partial t}. \tag{2.67}$$

The constitutive equations for the difference fields can be written as follows:

$$\boldsymbol{E}_1 = \epsilon \boldsymbol{D}_1;\ \boldsymbol{E}_2 = \epsilon \boldsymbol{D}_2; \implies \boldsymbol{E}_3 = \epsilon \boldsymbol{D}_3; \tag{2.68}$$

$$\boldsymbol{B}_1 = \mu \boldsymbol{H}_1;\ \boldsymbol{B}_2 = \mu \boldsymbol{H}_2; \implies \boldsymbol{B}_3 = \mu \boldsymbol{H}_3; \tag{2.69}$$

$$\boldsymbol{J}_1 = \sigma(\boldsymbol{E}_1 + \boldsymbol{E}_i);\ \boldsymbol{J}_2 = \sigma(\boldsymbol{E}_2 + \boldsymbol{E}_i); \implies \boldsymbol{J}_3 = \sigma \boldsymbol{E}_3. \tag{2.70}$$

It should be noted form (2.70) that no impressed electromotive field $\boldsymbol{E}_i$) is acting to produce the difference fields ($\boldsymbol{E}_3$, $\boldsymbol{D}_3$, $\boldsymbol{H}_3$, $\boldsymbol{B}_3$, $\boldsymbol{J}_3$). Applying the Poynting theorem to the difference fields, one then obtains:

$$0 = \int_{V_S} \frac{1}{\sigma} J_3^2 \, \mathrm{d}V + \frac{\mathrm{d}}{\mathrm{d}t} \int_{V_S} \frac{B_3^2}{2\mu} \, \mathrm{d}V + \frac{\mathrm{d}}{\mathrm{d}t} \int_{V_S} \frac{1}{2} \epsilon E_3^2 \, \mathrm{d}V + \oint_S (\boldsymbol{E}_3 \times \boldsymbol{H}_3) \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S \tag{2.71}$$

Moreover, as the the fields ($\boldsymbol{E}_1$, $\boldsymbol{H}_1$) and ($\boldsymbol{E}_2$, $\boldsymbol{H}_2$,) both have to satisfy the same boundary conditions, we can say that the the tangential electric field $\boldsymbol{E}_{t,3}$ over S is equal to zero, or the tangential magnetic field $\boldsymbol{H}_{t,3}$ over S is equal to zero, or the tangential electric field

$\boldsymbol{E}_{t,3}$ is equal to zero over part of S and the tangential magnetic field $\boldsymbol{H}_{t,3}$ is equal to zero over the rest of S, for any time $t > t_0$:

$$\begin{aligned}
\boldsymbol{E}_{t,1} = \boldsymbol{E}_{t,2} \text{ over } S \quad &\Longrightarrow \quad \boldsymbol{E}_{t,3} = 0 \text{ over } S, \text{ OR} \\
\boldsymbol{H}_{t,1} = \boldsymbol{H}_{t,2} \text{ over } S \quad &\Longrightarrow \quad \boldsymbol{H}_{t,3} = 0 \text{ over } S, \text{ OR} \\
\boldsymbol{E}_{t,1} = \boldsymbol{E}_{t,2} \text{ over part of } S, \text{ AND} & \\
\boldsymbol{H}_{t,1} = \boldsymbol{H}_{t,2} \text{ over the rest of } S \quad &\Longrightarrow \quad \boldsymbol{E}_{t,3} = 0 \text{ over part of } S, \text{ AND} \\
& \quad\;\; \boldsymbol{H}_{t,3} = 0 \text{ over the rest of } S; \qquad (2.72)
\end{aligned}$$

The conditions in (2.72) imply that the flux of the Poynting vector $\boldsymbol{E}_3 \times \boldsymbol{H}_3$ through S has to be zero. Indeed, the electric field $\boldsymbol{E}_3$ and the magnetic field $\boldsymbol{H}_3$ can be decomposed on the surface S in their tangential and normal components:

$$\begin{aligned}
\boldsymbol{E}_3 &= \boldsymbol{E}_{t,3} + E_{n,3}\,\hat{\boldsymbol{n}} \\
\boldsymbol{H}_3 &= \boldsymbol{H}_{t,3} + H_{n,3}\,\hat{\boldsymbol{n}}
\end{aligned}$$

Then, evaluating the term $(\boldsymbol{E}_3 \times \boldsymbol{H}_3) \cdot \hat{\boldsymbol{n}}$ one can write:

$$\begin{aligned}
\boldsymbol{E}_3 \times \boldsymbol{H}_3 \cdot \hat{\boldsymbol{n}} &= [(\boldsymbol{E}_{t,3} + E_{n,3}\,\hat{\boldsymbol{n}}) \times (\boldsymbol{H}_{t,3} + H_{n,3}\,\hat{\boldsymbol{n}})] \cdot \hat{\boldsymbol{n}} = \\
= (\boldsymbol{E}_{t,3} \times \boldsymbol{H}_{t,3}) \cdot \hat{\boldsymbol{n}} &+ \underbrace{H_{n,3}\,(\boldsymbol{E}_{t,3} \times \hat{\boldsymbol{n}}) \cdot \hat{\boldsymbol{n}}}_{=0 \text{ as } (\boldsymbol{E}_{t,3} \times \hat{\boldsymbol{n}}) \perp \hat{\boldsymbol{n}}} + \underbrace{E_{n,3}\,(\hat{\boldsymbol{n}} \times \boldsymbol{H}_{t,3}) \cdot \hat{\boldsymbol{n}}}_{=0 \text{ as } (\hat{\boldsymbol{n}} \times \boldsymbol{H}_{t,3}) \perp \hat{\boldsymbol{n}}} + \underbrace{E_{n,3}\,H_{n,3}\,(\hat{\boldsymbol{n}} \times \hat{\boldsymbol{n}})}_{=0 \text{ as } \hat{\boldsymbol{n}} \times \hat{\boldsymbol{n}} = 0} \Longrightarrow \\
&\Longrightarrow \boldsymbol{E}_3 \times \boldsymbol{H}_3 \cdot \hat{\boldsymbol{n}} = (\boldsymbol{E}_{t,3} \times \boldsymbol{H}_{t,3}) \cdot \hat{\boldsymbol{n}} \qquad (2.73)
\end{aligned}$$

As (2.72) prescribes that one of the two terms $\boldsymbol{E}_{t,3}$ or $\boldsymbol{E}_{t,3}$ appearing in the product on the right hand side of (2.73) has to be zero over S, one finally obtain that:

$$\oint_S (\boldsymbol{E}_3 \times \boldsymbol{H}_3) \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S = 0 \qquad (2.74)$$

Utilising (2.74) and reordering, we can rewrite (2.71) as:

$$-\int_{V_S} \frac{1}{\sigma} J_3^2\, \mathrm{d}V = \frac{\mathrm{d}}{\mathrm{d}\,t}\left(\int_{V_S} \frac{B_3^2}{2\mu}\, \mathrm{d}V + \int_{V_S} \frac{1}{2}\,\epsilon E_3^2\, \mathrm{d}V \right) \qquad (2.75)$$

Finally, as the fields $(\boldsymbol{E}_1, \boldsymbol{D}_1, \boldsymbol{H}_1, \boldsymbol{B}_1, \boldsymbol{J}_1)$ and $(\boldsymbol{E}_2, \boldsymbol{D}_2, \boldsymbol{H}_2, \boldsymbol{B}_2, \boldsymbol{J}_2)$ should comply with the same initial conditions, the difference fields have to be zero over V_S at $t = t_0$:

$$\begin{aligned}
\boldsymbol{E}_1 = \boldsymbol{E}_2 \text{ over } V_S \text{ for } t = t_0 \quad &\Longrightarrow \quad \boldsymbol{E}_3 = 0 \text{ over } V_S \text{ for } t = t_0; \\
\boldsymbol{B}_1 = \boldsymbol{B}_2 \text{ over } V_S \text{ for } t = t_0 \quad &\Longrightarrow \quad \boldsymbol{B}_3 = 0 \text{ over } V_S \text{ for } t = t_0. \qquad (2.76)
\end{aligned}$$

The left hand side of (2.75) is a non-positive quantity. Indeed, the integral kernel J_3^2/σ is a non-negative quantity, and thus:

$$-\int_{V_S} \frac{1}{\sigma} J_3^2\, \mathrm{d}V \leq 0 \quad \text{for any } t > t_0$$

Conversely, the right end side term of (2.75) represents the time derivative of the total energy associated with the fields $\boldsymbol{E}_3$ and $\boldsymbol{B}_3$. Being the magnetic and electrostatic energies obtained as the integral of non-negative quantities ($B_3^2/(2\mu) \geq 0$ and $\epsilon E_3^2/2) \geq 0$) over V_S, the

total energy is also a non-negative quantity. Furthermore, as a result of the initial conditions (2.76), the total energy is equal to zero at the time $t = t_0$:

$$\left(\int_{V_S} \frac{B_3^2}{2\mu} \mathrm{d}V + \int_{V_S} \frac{1}{2} \epsilon E_3^2 \mathrm{d}V \right) \Bigg|_{t=t_0} = 0.$$

The total electromagnetic energy associated with $\boldsymbol{E}_3$ and $\boldsymbol{B}_3$ is then at its minimum value at the time $t = t_0$. Any deviation after t_0 from the initial condition would lead to an increase of the energy and, thus, to a positive time derivative. Therefore, the term on the right hand side of (2.75) must remain constantly equal to zero ($\boldsymbol{E}_3$ and $\boldsymbol{B}_3$ remain equal to zero) or assume at some instant a positive value ($\boldsymbol{E}_3$ and $\boldsymbol{B}_3$ fields undergo some variation).

Therefore, (2.75) is verified if and only if both sides are equal to zero for any time $t > t_0$. As a consequence:

$$\boldsymbol{E}_3 = 0, \quad \boldsymbol{B}_3 = 0, \quad \boldsymbol{J}_3, = 0$$

and also, for the constitutive equations (2.68) and (2.68):

$$\boldsymbol{D}_3 = 0, \quad \boldsymbol{H}_3 = 0.$$

As the difference fields ($\boldsymbol{E}_3$, $\boldsymbol{D}_3$, $\boldsymbol{H}_3$, $\boldsymbol{B}_3$, $\boldsymbol{J}_3$) are equal to zero, then the two hypothetical solutions ($\boldsymbol{E}_1$, $\boldsymbol{D}_1$, $\boldsymbol{H}_1$, $\boldsymbol{B}_1$, $\boldsymbol{J}_1$) and ($\boldsymbol{E}_2$, $\boldsymbol{D}_2$, $\boldsymbol{H}_2$, $\boldsymbol{B}_2$, $\boldsymbol{J}_2$) are equal to each other.

Chapter 3

Mathematical tools

3.1 Green's identities

3.1.1 First Green's identity

Let V_S be a volume bounded by the closed surface S. Let φ and ψ be two scalar functions defined on V_S and twice continuously differentiable. Applying the divergence theorem (1.7) to the vector function $\psi\nabla\varphi$, one obtains:

$$\oint_S \psi\nabla\varphi\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = \int_{V_S}\nabla\cdot(\psi\nabla\varphi)\,\mathrm{d}V$$

end then, recalling that $\nabla\cdot(\psi\nabla\varphi) = \nabla\psi\cdot\nabla\varphi + \psi\nabla^2\varphi$:

$$\oint_S \psi\nabla\varphi\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = \int_{V_S}\nabla\psi\cdot\nabla\varphi\,\mathrm{d}V + \int_{V_S}\psi\nabla^2\varphi\,\mathrm{d}V. \tag{3.1}$$

Equation (3.1) is called first Green's identity.

3.1.2 Second Green's identity

Applying the same procedure to the vector function $\varphi\nabla\psi$, one can write as well:

$$\oint_S \varphi\nabla\psi\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = \int_{V_S}\nabla\psi\cdot\nabla\varphi\,\mathrm{d}V + \int_{V_S}\varphi\nabla^2\psi\,\mathrm{d}V. \tag{3.2}$$

Subtracting (3.2) from (3.1), one obtains the second Greens's identity:

$$\oint_S (\psi\nabla\varphi - \varphi\nabla\psi)\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = \int_{V_S}(\psi\nabla^2\varphi - \varphi\nabla^2\psi)\,\mathrm{d}V. \tag{3.3}$$

3.1.3 Third Green's identity

Let $\boldsymbol{p}_0$ and $\boldsymbol{p}$ be two points in the domain V_S, and let r be the distance between $\boldsymbol{p}_0$ and $\boldsymbol{p}$. One can easily recognise that, for any fixed point $\boldsymbol{p}_0$, the function:

$$\psi(\boldsymbol{p}) = \frac{1}{r} \tag{3.4}$$

is continuous and infinitely differentiable except when $\boldsymbol{p} = \boldsymbol{p}_0$. For $\boldsymbol{p} = \boldsymbol{p}_0$, the function ψ has instead a singularity. We now demonstrate that the function $\nabla\psi$ is a divergence free

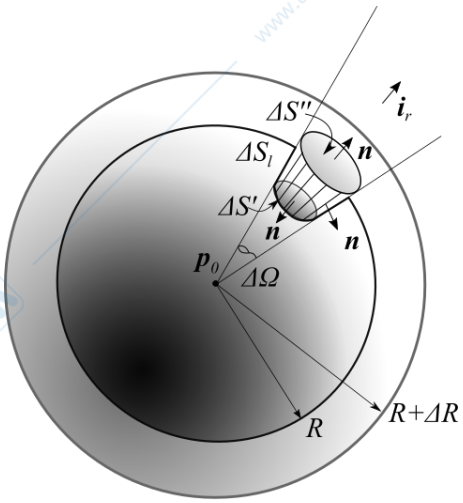


Figure 3.1:

vector in any point (except when $\boldsymbol{p} = \boldsymbol{p}_0$). Recalling the definition of the divergence operator given in (1.5):

$$\nabla \cdot \nabla\psi = \lim_{V_S \to 0} \frac{\oint_S \nabla\psi \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S}{V_S}, \tag{3.5}$$

we will prove that the flux of $\nabla\psi$ through any closed surface S bounding an infinitesimal volume V_S not enclosing $\boldsymbol{p}_0$ is equal to zero. Let V_S be the volume between the two spherical surfaces or radius R and $R + \Delta R$ respectively, centred at $\boldsymbol{p}_0$ and enclosed in the solid angle $\Delta\Omega$ (see Fig. 3.1). Assuming a spherical coordinate system centred at $\boldsymbol{p}_0$, the gradient of the ψ function can be expressed as:

$$\nabla\psi = -\frac{1}{r^2}\hat{\boldsymbol{i}}_r,$$

where $\hat{\boldsymbol{i}}_r$ in the radial unit vector. The surface S bounding V_S is constituted by the two spherical caps $\Delta S'$ and $\Delta S''$ encompassed by the solid angle $\Delta\Omega$ on the spherical surafaces above mentioned, and by the surface ΔS_l. Then, the flux of Ψ through S can be decomposed as:

$$\oint_S \nabla\psi \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S = -\oint_{\Delta S'} \frac{1}{r^2}\hat{\boldsymbol{i}}_r \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S - \oint_{\Delta S''} \frac{1}{r^2}\hat{\boldsymbol{i}}_r \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S - \oint_{\Delta S_l} \frac{1}{r^2}\hat{\boldsymbol{i}}_r \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S$$

Observing that $\hat{\boldsymbol{i}}_r \cdot \hat{\boldsymbol{n}} = -1$ on S', $\hat{\boldsymbol{i}}_r \cdot \hat{\boldsymbol{n}} = 1$ on S'' and $\hat{\boldsymbol{i}}_r \cdot \hat{\boldsymbol{n}} = 0$ on S_l ($since \hat{\boldsymbol{i}}_r \perp \hat{\boldsymbol{n}}$), and recalling that on a spherical surface $\mathrm{d}S = r^2 \mathrm{d}\Omega$ one obtains:

$$\oint_S \nabla\psi \cdot \hat{\boldsymbol{n}}\, \mathrm{d}S = \Delta\Omega - \Delta\Omega - 0 = 0$$

As the flux through S is identically equal to zero for any surface not enclosing $\boldsymbol{p}_0$, from (3.5) we obtain that the divergence of the vector function $\nabla\psi$, that is the laplacian of ψ, is equal to zero, in any point $\boldsymbol{p} \neq \boldsymbol{p}_0$:

$$\nabla \cdot \nabla\psi = \nabla^2\psi = \nabla^2 \frac{1}{r} = 0 \quad \forall \boldsymbol{p} \neq \boldsymbol{p}_0, \tag{3.6}$$

We want to apply now the second Green's identity to a volume V_S bounded by the closed surface S and containing the point $\boldsymbol{p}_0$, using the function ψ defined in (3.4). As noted before,

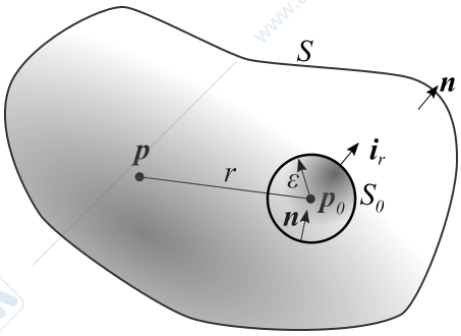


Figure 3.2:

$\psi(\boldsymbol{p})$ has a singularity in $\boldsymbol{p}_0$, so (3.3) can not be directly applied to the domain V_S. We will then utilise the domain V_S', obtained carving out from V_S a sphere V_0 of radius ϵ and centred in $\boldsymbol{p}_0$ (see Fig. 3.2). It should be noted that the boundary of the hollow domain obtained in this way is constituted by the external surface S (which also bounds V_S) and by the internal surface S_0, that is a spherical surface of radius ϵ and centred in $\boldsymbol{p}_0$. Then, inserting (3.4) and (3.6) in (3.3) , we obtain:

$$\oint_S (\frac{1}{r}\nabla\varphi - \varphi\nabla\frac{1}{r})\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S + \oint_{S_0}\frac{1}{r}\nabla\varphi\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S - \oint_{S_0}\varphi\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = \int_{V_S'}\frac{1}{r}\nabla^2\varphi\,\mathrm{d}V. \tag{3.7}$$

where $\varphi = \varphi(\boldsymbol{p})$ is a scalar continuously differentiable function. To bring us back to the initial domain V_S, we have to let the surface S_0 shrink to the point $\boldsymbol{p}_0$. To do this, we need to analyse the behaviour of the two integrals on the surface S_0 as ϵ tends to zero. Noting that $r = \epsilon$ on S_0, and that $\mathrm{d}S = \epsilon^2\,\mathrm{d}\Omega$, we have, for the second term of the left hand side of (3.7):

$$\lim_{\epsilon\to 0}\oint_{S_0}\frac{1}{r}\nabla\varphi\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = \lim_{\epsilon\to 0}\epsilon\oint_{4\pi}\nabla\varphi\cdot\hat{\boldsymbol{n}}\,\mathrm{d}\Omega = 0; \tag{3.8}$$

A similar procedure can be applied to the third term on the left hand side of (3.7):

$$\lim_{\epsilon\to 0}\left(-\oint_{S_0}\varphi\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S\right) = \lim_{\epsilon\to 0}\oint_{S_0}\frac{1}{r^2}\varphi\,\hat{\boldsymbol{i}}_r\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = \lim_{\epsilon\to 0}\oint_{4\pi}\varphi\,\hat{\boldsymbol{i}}_r\cdot\hat{\boldsymbol{n}}\,\mathrm{d}\Omega.$$

We note that $\hat{\boldsymbol{i}}_r\cdot\hat{\boldsymbol{n}} = -1$ on S_0: indeed, $\hat{\boldsymbol{i}}_r$ is a unit vector pointing radially outward from the origin $\boldsymbol{p}_0$, while $\hat{\boldsymbol{n}}$, being defined as the unit vector normal to the boundary of the domain V_S, conventionally points outward V_S, and thus toward $\boldsymbol{p}_0$. Noting also that, according to the hypothesis, the function $\varphi(\boldsymbol{p})$ tends to its value in $\boldsymbol{p}_0$ when ϵ tends to zero, one finally obtains:

$$\lim_{\epsilon\to 0}\left(-\oint_{S_0}\varphi\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S\right) = \lim_{\epsilon\to 0}\oint_{4\pi}\varphi\,\hat{\boldsymbol{i}}_r\cdot\hat{\boldsymbol{n}}\,\mathrm{d}\Omega = -4\pi\varphi(\boldsymbol{p}_0). \tag{3.9}$$

Utilising (3.8) and (3.9) and rearranging, we find the expression to which equation (3.7) reduces when S_0 shrinks to the point $\boldsymbol{p}_0$:

$$\varphi(\boldsymbol{p}_0) = -\frac{1}{4\pi}\int_{V_S}\frac{1}{r}\nabla^2\varphi\,\mathrm{d}V + \frac{1}{4\pi}\oint_S\frac{1}{r}\nabla\varphi\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S - \frac{1}{4\pi}\oint_S\varphi\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S \tag{3.10}$$

Recalling the definition of the gradient operator, one can write:

$$\nabla\varphi\cdot\hat{\boldsymbol{n}} = \frac{\partial\varphi}{\partial n}, \quad \nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}} = \frac{\partial}{\partial n}\left(\frac{1}{r}\right),$$

and then also:

$$\varphi(\boldsymbol{p}_0) = -\frac{1}{4\pi}\int_{V_S}\frac{1}{r}\nabla^2\varphi\,\mathrm{d}V + \frac{1}{4\pi}\oint_S \frac{1}{r}\frac{\partial\varphi}{\partial n}\mathrm{d}S - \frac{1}{4\pi}\oint_S \varphi\frac{\partial}{\partial n}\left(\frac{1}{r}\right)\mathrm{d}S \tag{3.11}$$

Equations (3.10), and its equavalent formulation given in(3.11), are known as third Green's identity. According to (3.10), a scalar function φ can be evaluated on any point $\boldsymbol{p}_0$ in V_S, knowing its laplacian over V_S, and the value of φ and of its normal derivative over the bounding surface S. However, we will see in the next sections that (3.10) can not be treated as a general solution of the Poisson's equation on a bounded domain. Indeed, the boundary conditions φ and of its normal derivative can not be assigned independently.

3.2 Regular boundary conditions at Infinity

Let us consider how the third Green' identity apply to the entire $\mathbb{R}^3$ space. To do this, we will apply (3.11) to a spherical domain V_R bounded by a spherical surface S_R of radius R centred in $\boldsymbol{p}_0$:

$$\begin{aligned}\varphi(\boldsymbol{p}_0) &= -\frac{1}{4\pi}\int_{V_R}\frac{1}{r}\nabla^2\varphi\,\mathrm{d}V + \frac{1}{4\pi}\oint_{S_R} \frac{1}{r}\frac{\partial\varphi}{\partial n}\mathrm{d}S - \frac{1}{4\pi}\oint_{S_R} \varphi\frac{\partial}{\partial n}\left(\frac{1}{r}\right)\mathrm{d}S = \\ &= -\frac{1}{4\pi}\int_{V_R}\frac{1}{r}\nabla^2\varphi\,\mathrm{d}V + \frac{1}{4\pi}\oint_{4\pi} R\frac{\partial\varphi}{\partial r}\mathrm{d}\Omega + \frac{1}{4\pi}\oint_{4\pi}\varphi\,\mathrm{d}\Omega.\end{aligned}$$

and let R tend to infinity. One can easily recognize that if the following conditions are satisfied:

$$\begin{aligned}\lim_{R\to\infty} R\varphi &= \alpha \\ \lim_{R\to\infty} R^2\frac{\partial\varphi}{\partial r} &= \beta\end{aligned} \tag{3.12}$$

the above expression, for R tending to infinity reduces to:

$$\varphi(\boldsymbol{p}_0) = -\frac{1}{4\pi}\int_{\mathbb{R}^3}\frac{1}{r}\nabla^2\varphi\,\mathrm{d}V \tag{3.13}$$

which is the solution of the scalar Poisson's equation in an unbounded $\mathbb{R}^3$ space. The conditions (3.12) are said regular boundary conditions at infinity, and specify that the function φ and its derivative $\frac{\partial\varphi}{\partial r}$ along the radial direction tends to zero at most as the functions $\frac{\alpha}{R}$ and $\frac{\beta}{R^2}$ do when R tends to infinity.

3.3 Harmonic functions

A twice continuously differentiable scalar function φ is said harmonic on a domain V_S if it satisfies the Laplace equation (1.25) over the entire domain:

$$\nabla^2\varphi = 0.$$

From (3.11), one recognise that a harmonic function is determined by its behaviour at the boundary:

$$\varphi(\boldsymbol{p}_0) = \frac{1}{4\pi}\oint_S \frac{1}{r}\frac{\partial\varphi}{\partial n}\mathrm{d}S - \frac{1}{4\pi}\oint_S \varphi\frac{\partial}{\partial n}\left(\frac{1}{r}\right)\mathrm{d}S.$$

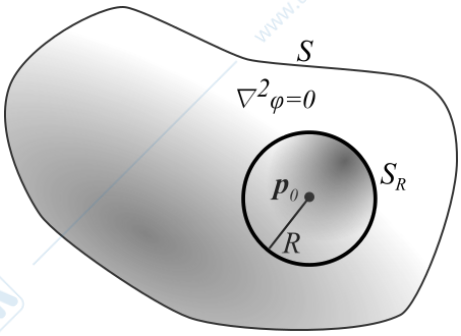


Figure 3.3:

3.3.1 Mean value theorem for harmonic functions

Let φ be a harmonic function, defined on a volume V_S bounded by the closed surface S. Let $\boldsymbol{p}_0$ be a point in V_S, and let V_R be a sphere centred in $\boldsymbol{p}_0$, bounded by the surface S_R of radius R, and entirely contained in V_S. Then, the value of the harmonic function φ at the point $\boldsymbol{p}_0$ is equal to the average value of the function φ on S_R. Indeed, applying (3.10) to the volume V_R, one can write:

$$\varphi(\boldsymbol{p}_0) = \frac{1}{4\pi}\oint_{S_R}\frac{1}{r}\nabla\varphi\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S - \frac{1}{4\pi}\oint_{S_R}\varphi\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S. \tag{3.14}$$

Noting that $r = R$ on the surface S_R, one can manipulate the first integral at the right hand side of (3.14) utilising the divergence theorem (1.7):

$$\oint_{S_R}\frac{1}{r}\nabla\varphi\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = \frac{1}{R}\oint_{S_R}\nabla\varphi\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = \frac{1}{R}\int_{V_R}\nabla\cdot\nabla\varphi\,\mathrm{d}V = \frac{1}{R}\int_{V_R}\nabla^2\varphi\,\mathrm{d}V = 0$$

The second integral at the right hand side of (3.14) can instead be rewritten as:

$$\oint_{S_R}\varphi\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = -\oint_{S_R}\frac{1}{r^2}\varphi\hat{\boldsymbol{i}}_r\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S = -\frac{1}{R^2}\oint_{S_R}\varphi\,\mathrm{d}S.$$

Inserting the last two results in (3.14), one finally obtains:

$$\varphi(\boldsymbol{p}_0) = \frac{1}{4\pi R^2}\oint_{S_R}\varphi\,\mathrm{d}S, \tag{3.15}$$

which proves the theorem. The theorem also implies that the value of the harmonic function φ at the point $\boldsymbol{p}_0$ is equal to the average value of the function φ on the sphere V_R. Indeed, integrating φ over the sphere V_R and applying (3.15):

$$\int_{V_R}\varphi\,\mathrm{d}V = \int_0^R\int_{S(r)}\varphi\,\mathrm{d}S\,\mathrm{d}r = 4\pi\varphi(\boldsymbol{p}_0)\int_0^R r^2\,\mathrm{d}r = \frac{4}{3}\pi R^3\varphi(\boldsymbol{p}_0),$$

and then:

$$\varphi(\boldsymbol{p}_0) = \frac{1}{\frac{4}{3}\pi R^3}\int_{V_R}\varphi\,\mathrm{d}V, \tag{3.16}$$

The mean value theorem for harmonic functions has some useful corollary

First corollary

A function $\varphi(\boldsymbol{p})$ harmonic on a connected domain V_S can not have local extrema in any point $\boldsymbol{p}_0$ internal to V_S. Indeed, reasoning ad absurdum, if the point $\boldsymbol{p}_0$ internal to the domain V_S was a local maximum, then there exists a sphere V_ϵ of radius ϵ in which $\varphi(\boldsymbol{p}_0) > \varphi(\boldsymbol{p}) \quad \forall \boldsymbol{p} \in V_\epsilon$. Then, we can define a scalar function $\varphi'(\boldsymbol{p}) = \varphi(\boldsymbol{p}) - \varphi(\boldsymbol{p}_0)$. The function $\varphi'(\boldsymbol{p})$ is still harmonic on V_S, as $\nabla^2\varphi(\boldsymbol{p}_0) = 0$, and:

$$\begin{aligned} \varphi'(\boldsymbol{p}_0) &= 0 \\ \varphi'(\boldsymbol{p}) &< 0 \quad \forall \boldsymbol{p} \in V_\epsilon \end{aligned}$$

Thus, applying the mean value theorem to the function function $\varphi'(\boldsymbol{p})$ on the volume V_ϵ, we get:

$$\varphi'(\boldsymbol{p}_0) = \frac{1}{\frac{4}{3}\pi\epsilon^3} \int_{V_R} \varphi' \, \mathrm{d}V, \tag{3.17}$$

The left hand side term in (3.17) is equal to zero, while the right hand side term, being an integral of a negative function (except in $\boldsymbol{p}_0$), must be a strictly negative quantity. Thus, the expression (3.17) can not hold, and since we derived it assuming a local maximum in $\boldsymbol{p}_0$ (assuming a local minimum, we would obtain an analogous result), we conclude that no extrema can exist in the points internal to V_S.

Second corollary

Any extrema of an harmonic function $\varphi(\boldsymbol{p})$ on a connected domain V_S is located on the boundary S of V_S.

Third corollary

If an harmonic function $\varphi(\boldsymbol{p})$ on a connected domain V_S assumes a constant value φ_0 on the entire boundary S of V_S, it remains constant on the entire domain V_S ($\varphi(\boldsymbol{p}) = \varphi_0 \quad \forall \boldsymbol{p} \in V_S$).

3.3.2 Reciprocity of harmonic functions

Let φ and ψ be two harmonic functions, defined on the domain V_S, which is bounded by S. Since $\nabla^2\varphi = 0$ and $\nabla^2\psi = 0$, we obtain from the second Green's identity:

$$\oint_S \psi \nabla\varphi \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S = \oint_S \varphi \nabla\psi \cdot \hat{\boldsymbol{n}} \, \mathrm{d}S, \tag{3.18}$$

that is:

$$\oint_S \psi \frac{\partial\varphi}{\partial n} \, \mathrm{d}S = \oint_S \varphi \frac{\partial\psi}{\partial n} \, \mathrm{d}S, \tag{3.19}$$

The equations (3.18) and (3.19) express the reciprocity property of the harmonic functions.

3.4 Uniqueness theorems for the Poisson's equation

Poisson's equation:

$$\nabla^2\varphi = t(\boldsymbol{p}),$$

which we introduced in section 1.3, stems from the formulation of a wide variety of physical problems (diffusion, thermal conduction...). The description of steady state electromagnetism in homogeneous media leads to the equation of Poisson (in scalar or vector form). This section gives some criteria that must be met in order for the solution of the Poisson's equation to be unique.

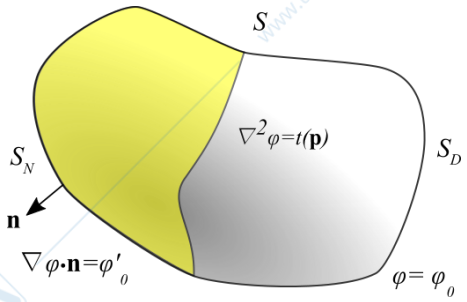


Figure 3.4: *Mixed boundary conditions: the normal derivative is assigned on S_N, while the value of the function is assigned on S_D.*

3.4.1 Dirichlet boundary conditions.

We analyse in the first place the case in which the unknown function is assigned on the entire boundary (Dirichlet boundary conditions). Let V_S be a connected domain, bounded by the closed surface S. Let $\varphi(\boldsymbol{p})$ be a scalar function, defined by the problem:

$$\begin{aligned} \nabla^2\varphi &= t(\boldsymbol{p}) \quad \forall \boldsymbol{p} \in V_S \\ \varphi(\boldsymbol{p}) &= \varphi_0(\boldsymbol{p}) \quad \forall \boldsymbol{p} \in S, \end{aligned} \tag{3.20}$$

where $\varphi_0(\boldsymbol{p})$ is a known function assigning the value of φ on the entire boundary S. We demonstrate now that the solution to the problem (3.20) is unique. Indeed, arguing ad absurdum, let us suppose that two different solutions exist, $\varphi_1(\boldsymbol{p})$ and $\varphi_2(\boldsymbol{p})$, both satisfying (3.20). Then, defined the scalar function $\varphi_3(\boldsymbol{p}) = \varphi_1(\boldsymbol{p}) - \varphi_2(\boldsymbol{p})$, for the linear property of the laplacian we have:

$$\begin{aligned} \nabla^2\varphi_3 = \nabla^2(\varphi_1 - \varphi_2) = \nabla^2\varphi_1 - \nabla^2\varphi_2 = t(\boldsymbol{p}) - t(\boldsymbol{p}) &= 0 \quad \forall \boldsymbol{p} \in V_S \\ \varphi_3(\boldsymbol{p}) = \varphi_1(\boldsymbol{p}) - \varphi_2(\boldsymbol{p}) = \varphi_0(\boldsymbol{p}) - \varphi_0(\boldsymbol{p}), &= 0 \quad \forall \boldsymbol{p} \in S. \end{aligned}$$

Then the function $\varphi_3(\boldsymbol{p})$ is a harmonic function over V_S, and assumes on the entire boundary S a constant value equal to zero. Thus, for the third corollary of the mean value for harmonic functions, $\varphi_3(\boldsymbol{p})$ has to be zero over the entire domain V_S, and $\varphi_1(\boldsymbol{p}) = \varphi_2(\boldsymbol{p})$, which proves the theorem.

3.4.2 Mixed boundary conditions

Now consider the case where the unknown function is assigned on part of the boundary (Dirichlet boundary conditions), while on the remaining part of the boundary the normal derivative is known (Neumann boundary conditions). Let V_S be a connected domain, bounded by the closed surface S. Let S_D be a part of the surface S on which the scalar function $\varphi(\boldsymbol{p})$ is assigned to a known value $\varphi_0(\boldsymbol{p})$; let S_N be the rest of S, on which the normal derivative of the scalar function $\varphi(\boldsymbol{p})$ is assigned to a known value $\varphi'_0(\boldsymbol{p})$. Let then $\varphi(\boldsymbol{p})$ be the scalar function defined by the problem:

$$\begin{aligned} \nabla^2\varphi &= t(\boldsymbol{p}) \quad \forall \boldsymbol{p} \in V_S \\ \varphi(\boldsymbol{p}) &= \varphi_0(\boldsymbol{p}) \quad \forall \boldsymbol{p} \in S_D, \\ \frac{\partial\varphi}{\partial n} &= \varphi'_0(\boldsymbol{p}) \quad \forall \boldsymbol{p} \in S_R, \end{aligned} \tag{3.21}$$

where $S_D \cup S_N = S$. We demonstrate now that the solution to the problem (3.20) is unique. Indeed, reasoning by reduction ad absurdum, let us suppose that two different solutions

exist, $\varphi_1(\boldsymbol{p})$ and $\varphi_2(\boldsymbol{p})$, both satisfying (3.21). Then, defined the scalar function $\varphi_3(\boldsymbol{p}) = \varphi_1(\boldsymbol{p}) - \varphi_2(\boldsymbol{p})$, for the linear property of the laplacian and of the normal derivative, we have:

$$\begin{aligned} \nabla^2\varphi_3 = \nabla^2(\varphi_1 - \varphi_2) = \nabla^2\varphi_1 - \nabla^2\varphi_2 = t(\boldsymbol{p}) - t(\boldsymbol{p}) &= 0 \quad \forall \boldsymbol{p} \in V_S \\ \varphi_3(\boldsymbol{p}) = \varphi_1(\boldsymbol{p}) - \varphi_2(\boldsymbol{p}) = \varphi_0(\boldsymbol{p}) - \varphi_0(\boldsymbol{p}), &= 0 \quad \forall \boldsymbol{p} \in S_D. \\ \frac{\partial\varphi_3}{\partial n} = \frac{\partial}{\partial n}(\varphi_1 - \varphi_2) = \frac{\partial\varphi_1}{\partial n} - \frac{\partial\varphi_2}{\partial n} = \varphi_0'(\boldsymbol{p}) - \varphi_0'(\boldsymbol{p}) &= 0 \quad \forall \boldsymbol{p} \in S_R, \end{aligned}$$

Applying the first Green's identity (3.1) with $\varphi = \psi = \varphi_3$, one obtains:

$$\oint_S \varphi_3 \frac{\partial\varphi_3}{\partial n}\,\mathrm{d}S = \int_{V_S} (\nabla\varphi_3)^2\,\mathrm{d}V + \int_{V_S} \varphi_3\nabla^2\varphi_3\,\mathrm{d}V.$$

One of the two factors in the product that appears in the kernel of the integral at the left hand side is equal to zero over the surface S. Furthermore, φ_3 is harmonic over V_S, and then:

$$\int_{V_S} (\nabla\varphi_3)^2\,\mathrm{d}V = 0. \tag{3.22}$$

Since the quantity $(\nabla\varphi_3)^2$ is necessarily non-negative, we can conclude that, in order for condition (3.22) to be met, we must have $\nabla\varphi_3 = 0$, that is the function φ_3 must assume a constant value over V_S. But, since $\varphi_3 = 0$ on S_D, we conclude that $\varphi_3 = 0$ over the entire domain V_S. Thus, $\varphi_1(\boldsymbol{p}) = \varphi_2(\boldsymbol{p})$, which proves the theorem.

3.4.3 Neumann boundary conditions

The last case we study occurs when the normal derivative of the unknown function $\varphi(\boldsymbol{p})$ is assigned to a given value $\varphi_0'(\boldsymbol{p})$ over the entire boundary (Neumann boundary conditions). Let V_S be a connected domain, bounded by the closed surface S. Let $\varphi(\boldsymbol{p})$ be a scalar function, defined by the problem:

$$\begin{aligned} \nabla^2\varphi &= t(\boldsymbol{p}) \quad \forall \boldsymbol{p} \in V_S \\ \varphi(\boldsymbol{p}) &= \varphi_0(\boldsymbol{p}) \quad \forall \boldsymbol{p} \in S, \end{aligned} \tag{3.23}$$

We see that in this case the solution to the problem (3.23) is not unique. Indeed, assuming the existence of two solutions $\varphi_1(\boldsymbol{p})$ and $\varphi_2(\boldsymbol{p})$, and defining their difference $\varphi_3(\boldsymbol{p}) = \varphi_1(\boldsymbol{p}) - \varphi_2(\boldsymbol{p})$, by a procedure similar to the one we used in the previous subsection, we obtain again:

$$\int_{V_S} (\nabla\varphi_3)^2\,\mathrm{d}V = 0.$$

We deduce than that the function φ_3 is constant over the domain V_S. However, in this case the value of this constant can not be determined. Thus we conclude that two different solutions $\varphi_1(\boldsymbol{p})$ and $\varphi_2(\boldsymbol{p})$ may exist, as long as their difference is equal to a constant. In other words, the solution to the problem (3.23) is defined up to a constant.

3.5 Green's function.

We want now to provide an explicit solution to Poisson's equation, defined on a connected and bounded domain V_S, when the boundary conditions are imposed in order to meet the criteria of uniqueness discussed in the previous section. For sake of generality, let us refer

to problem formulated in (3.21), where Dirichlet boundary condition are applied to a part of the bonding surface S, and Neumann conditions are imposed to the rest of S:

$$\begin{aligned} \nabla^2\varphi &= t(\boldsymbol{p}) \quad \forall \boldsymbol{p} \in V_S \\ \varphi(\boldsymbol{p}) &= \varphi_0(\boldsymbol{p}) \quad \forall \boldsymbol{p} \in S_D, \\ \frac{\partial\varphi}{\partial n} &= \varphi_0'(\boldsymbol{p}) \quad \forall \boldsymbol{p} \in S_R, \end{aligned}$$

The third Green's identity (3.11) can be conveniently used for the intended purpose.

$$\varphi(\boldsymbol{p}_0) = -\frac{1}{4\pi}\int_{V_S}\frac{1}{r}\nabla^2\varphi\,\mathrm{d}V + \frac{1}{4\pi}\oint_S\frac{1}{r}\frac{\partial\varphi}{\partial n}\,\mathrm{d}S - \frac{1}{4\pi}\oint_S\varphi\frac{\partial}{\partial n}\left(\frac{1}{r}\right)\mathrm{d}S$$

In order to find the value of $\varphi(\boldsymbol{p}_0)$ utilizing the third Green's identity (3.11), one has to know both $\varphi(\boldsymbol{p}_0)$ and $\frac{\partial\varphi}{\partial n}$ over the entire boundary. However, it should be noted that the boundary conditions do not do not provide explicitly the value of the function and its normal derivative on the boundary. More precisely, on S_D the value of φ is assigned, and the normal derivative is a result. The opposite happens on the surface, on which the normal derivative is set and the value of the function is the result. It is then necessary to find a way to eliminate this difficulty.

In order to do this, we look for a scalar function χ, harmonic in V_S:

$$\nabla^2\chi = 0$$

Then, applying the second Green's identity (3.3) with $\psi = \chi$:

$$\oint_S\chi\frac{\partial\varphi}{\partial n}\,\mathrm{d}S - \oint_S\varphi\frac{\partial\chi}{\partial n}\,\mathrm{d}S = \int_{V_S}\chi\nabla^2\varphi\,\mathrm{d}V,$$

and, multiplying by $\frac{1}{4\pi}$ and rearranging:

$$0 = -\frac{1}{4\pi}\int_{V_S}\chi\nabla^2\varphi\,\mathrm{d}V + \frac{1}{4\pi}\oint_S\chi\frac{\partial\varphi}{\partial n}\,\mathrm{d}S - \frac{1}{4\pi}\oint_S\varphi\frac{\partial\chi}{\partial n}\,\mathrm{d}S \tag{3.24}$$

Subracting (3.24) from the equation expressing the third Green's identity (3.11), one obtains:

$$\varphi(\boldsymbol{p}_0) = -\frac{1}{4\pi}\int_{V_S}\left(\chi-\frac{1}{r}\right)\nabla^2\varphi\,\mathrm{d}V + \frac{1}{4\pi}\oint_S\left(\chi-\frac{1}{r}\right)\frac{\partial\varphi}{\partial n}\,\mathrm{d}S - \frac{1}{4\pi}\oint_S\varphi\frac{\partial}{\partial n}\left(\chi-\frac{1}{r}\right)\mathrm{d}S \tag{3.25}$$

The function:

$$G(\boldsymbol{p}) = \chi - \frac{1}{r} \tag{3.26}$$

is called Green function. Inserting (3.26) in (3.25), we obtain:

$$\varphi(\boldsymbol{p}_0) = -\frac{1}{4\pi}\int_{V_S}G\nabla^2\varphi\,\mathrm{d}V + \frac{1}{4\pi}\oint_S G\frac{\partial\varphi}{\partial n}\,\mathrm{d}S - \frac{1}{4\pi}\oint_S\varphi\frac{\partial G}{\partial n}\,\mathrm{d}S \tag{3.27}$$

As the normal derivative of the function φ is assigned only on S_N, and is unknown on S_D, we want that $G(\boldsymbol{p}) = 0, \quad \forall \boldsymbol{p} \in S_D$, so that:

$$\oint_S G\frac{\partial\varphi}{\partial n}\,\mathrm{d}S = \int_{S_N}G\frac{\partial\varphi}{\partial n}\,\mathrm{d}S$$

In a similar way, as the value of the function φ is known only on S_D and not on S_N, we demand also that $\frac{\partial G}{\partial n} = 0, \quad \forall \boldsymbol{p} \in S_N$, in order to have:

$$\oint_S \varphi \frac{\partial G}{\partial n} \mathrm{d}S = \int_{S_D} \varphi \frac{\partial G}{\partial n} \mathrm{d}S$$

What is written above specifies that the Green's function is subject to homogeneous boundary conditions on S. Moreover, for (3.26), $G(\boldsymbol{p})$ is harmonic everywhere in V_S except in $\boldsymbol{p}_0$, where it presents a singularity. We can then formulate the Green's problem associated to the given boundary problem (3.21):

$$\begin{aligned} \nabla^2 G &= \delta(\boldsymbol{p} - \boldsymbol{p}_0) \quad \forall \boldsymbol{p} \in V_S \\ G(\boldsymbol{p}) &= 0 \quad \forall \boldsymbol{p} \in S_D, \\ \frac{\partial G}{\partial n} &= 0 \quad \forall \boldsymbol{p} \in S_R, \end{aligned} \tag{3.28}$$

where $\delta(\boldsymbol{p}-\boldsymbol{p}_0)$ is the Dirac delta function. Utilizing the Green's function G uniquely defined by the problem (3.28), the solution to the initial problem (3.21) takes the form:

$$\varphi(\boldsymbol{p}_0) = -\frac{1}{4\pi} \int_{V_S} G \nabla^2 \varphi \, \mathrm{d}V + \frac{1}{4\pi} \int_{S_N} G \frac{\partial \varphi}{\partial n} \mathrm{d}S - \frac{1}{4\pi} \int_{S_D} \varphi \frac{\partial G}{\partial n} \mathrm{d}S. \tag{3.29}$$

3.6 Volume potential, single and double layer potential

We now want to extend the theory developed up to now to the case where the function φ presents a discontinuity of some kind. Indeed, we recall that, in order to formulate the Green's identities, we assumed to operate on double continuously derivable functions. We then will depart from this assumption by considering a function $\varphi(\boldsymbol{p})$ that exhibits step discontinuities in its value and in its derivative. For sake of simplicity, let us consider a function $\varphi(\boldsymbol{p})$ defined on the entire $\mathbb{R}^3$, satisfying the regular conditions at infinity (3.12), and let us assume that we know the distribution of $\nabla^2\varphi$ over the domain. Let S_{sl} and S_{dl} be two surfaces in $\mathbb{R}^3$. We also assume that the normal derivative of φ to the surface S_{sl} has a step discontinuity over the entire S_{sl}, and that φ has a step discontinuity over the entire S_{dl}. The unit vector $\hat{\boldsymbol{n}}$ normal to the discontinuity surfaces S_{sl} and S_{dl} defines their positive and negative sides. Let $\left.\frac{\partial \varphi}{\partial n}\right|_+$ and $\left.\frac{\partial \varphi}{\partial n}\right|_-$ be the normal derivatives on the positive and negative sides of the discontinuity surfaces. Analogously, let φ_+ and φ_- be the values of the function φ at the positive and negative sides of the discontinuity surfaces.

Since the third Green's identity has been obtained under the assumption that the function is twice continuously derivable, we apply now (3.11)to the domain obtained by carving out from $\mathbb{R}^3$ two volumes, bounded by the closed surfaces Σ_1 and Σ_2, respectively, and containing the two discontinuity surfaces S_{sl} and S_{dl}, as shown in Fig. 3.5. Applying (3.10), we can write:

$$\begin{aligned} \varphi(\boldsymbol{p}_0) = -\frac{1}{4\pi} \int_{V_S} \frac{1}{r} \nabla^2 \varphi \, \mathrm{d}V + \\ + \frac{1}{4\pi} \oint_{\Sigma_1} \frac{1}{r} \nabla\varphi \cdot \hat{\boldsymbol{n}}_1 \, \mathrm{d}\Sigma_1 - \frac{1}{4\pi} \oint_{\Sigma_1} \varphi \nabla \frac{1}{r} \cdot \hat{\boldsymbol{n}}_1 \, \mathrm{d}\Sigma_1 + \\ + \frac{1}{4\pi} \oint_{\Sigma_2} \frac{1}{r} \nabla\varphi \cdot \hat{\boldsymbol{n}}_2 \, \mathrm{d}\Sigma_2 - \frac{1}{4\pi} \oint_{\Sigma_2} \varphi \nabla \frac{1}{r} \cdot \hat{\boldsymbol{n}}_2 \, \mathrm{d}\Sigma_2 \end{aligned} \tag{3.30}$$

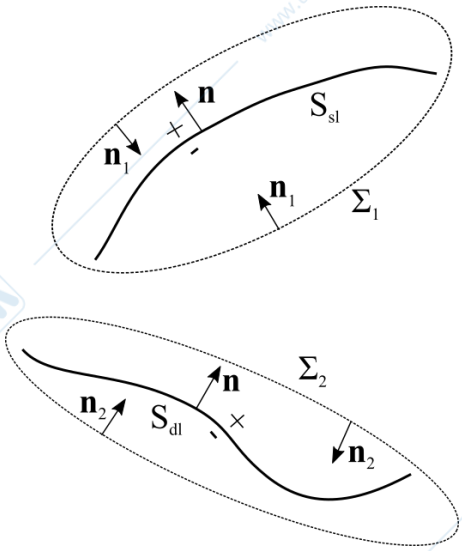


Figure 3.5: *Discontinuity surfaces: the normal derivative of φ has a step discontinuity over S_{sl}, and φ has a step discontinuity over the entire S_{dl}*

where $\hat{\boldsymbol{n}}_1$ and $\hat{\boldsymbol{n}}_2$ are the two normal unit vector, pointing outward the considered domain. To bring us back to the original domain $\mathbb{R}^3$, then we let the surfaces Σ_1 and Σ_2 shrink to S_{sl} and S_{dl}. Let us consider the surface Σ_1 tending to S_{sl}. In this process, one can observe that each infinitesimal surface element of S_{sl} will be approached by two elements of Σ_1 (see Fig. 3.6; one of them -we call it $\mathrm{d}\Sigma_{1,+}$- will tend to the positive side of S_{sl}, the other one, $\mathrm{d}\Sigma_{1,-}$, will tend to the negative one. Then, as Σ_1 tends to S_{sl}, we have that:

$$\lim_{\mathrm{d}\Sigma_{1,+}\to \mathrm{d}S} \begin{cases} \hat{\boldsymbol{n}}_1 \to -\hat{\boldsymbol{n}} \\ \nabla\varphi\cdot\hat{\boldsymbol{n}}_1 \to -\nabla\varphi\cdot\hat{\boldsymbol{n}} = -\frac{\partial\varphi}{\partial n}\Big|_+ \\ \varphi \to \varphi_+ \end{cases} \qquad \lim_{\mathrm{d}\Sigma_{1,-}\to \mathrm{d}S} \begin{cases} \hat{\boldsymbol{n}}_1 \to \hat{\boldsymbol{n}}, \\ \nabla\varphi\cdot\hat{\boldsymbol{n}}_1 \to \nabla\varphi\cdot\hat{\boldsymbol{n}} = \frac{\partial\varphi}{\partial n}\Big|_- , \\ \varphi \to \varphi_- \end{cases} \tag{3.31}$$

As a result, considering also that the function φ is continuous on Σ_1 and $\varphi_+ = \varphi_-$, the integrals over the surface Σ_1 in (3.30) reduces to:

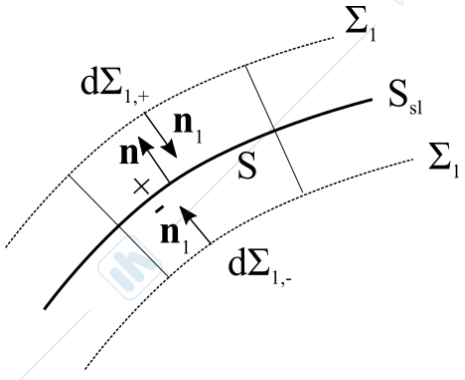


Figure 3.6: Σ_1 *tending to* S_{sl}

$$\frac{1}{4\pi}\oint_{\Sigma_1}\frac{1}{r}\nabla\varphi\cdot\hat{\boldsymbol{n}}_1\,\mathrm{d}\Sigma_1-\frac{1}{4\pi}\oint_{\Sigma_1}\varphi\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}_1\,\mathrm{d}\Sigma_1=$$
$$=-\frac{1}{4\pi}\int_{\S_{sl}}\frac{1}{r}\left(\left.\frac{\partial\varphi}{\partial n}\right|_+-\left.\frac{\partial\varphi}{\partial n}\right|_-\right)\mathrm{d}S+\frac{1}{4\pi}\int_{\S_{sl}}(\varphi_+-\varphi_-)\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S=$$
$$=-\frac{1}{4\pi}\int_{\S_{sl}}\frac{1}{r}\left(\left.\frac{\partial\varphi}{\partial n}\right|_+-\left.\frac{\partial\varphi}{\partial n}\right|_-\right)\mathrm{d}S$$

as Σ_1 tends to S_{sl}. In a similar fashion, recalling that the normal derivative is continuous on S_{dl}, we observe that the integrals over Σ_2 appearing in (3.30) reduces to:

$$\frac{1}{4\pi}\oint_{\Sigma_2}\frac{1}{r}\nabla\varphi\cdot\hat{\boldsymbol{n}}_2\,\mathrm{d}\Sigma_2-\frac{1}{4\pi}\oint_{\Sigma_2}\varphi\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}_2\,\mathrm{d}\Sigma_2=$$
$$=-\frac{1}{4\pi}\int_{\S_{dl}}\frac{1}{r}\left(\left.\frac{\partial\varphi}{\partial n}\right|_+-\left.\frac{\partial\varphi}{\partial n}\right|_-\right)\mathrm{d}S+\frac{1}{4\pi}\int_{\S_{dl}}(\varphi_+-\varphi_-)\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S=$$
$$=\frac{1}{4\pi}\int_{\S_{dl}}(\varphi_+-\varphi_-)\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S$$

as as Σ_2 tends to S_{dl}. We can conclude that, as Σ_1 and Σ_2 shrink to S_{sl} and S_{dl} respectively, (3.30) can be rewritten as:

$$\varphi(\boldsymbol{p}_0)=-\frac{1}{4\pi}\int_{V_S}\frac{1}{r}\nabla^2\varphi\,\mathrm{d}V-\frac{1}{4\pi}\int_{\S_{sl}}\frac{1}{r}\left(\left.\frac{\partial\varphi}{\partial n}\right|_+-\left.\frac{\partial\varphi}{\partial n}\right|_-\right)\mathrm{d}S+\frac{1}{4\pi}\int_{\S_{dl}}(\varphi_+-\varphi_-)\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S\quad(3.32)$$

The value of the unknown function φ at a given point $\boldsymbol{p}_0$ is then determined by the sum of the three terms at the right hand side of (3.32). We then refer to the first term as volume potential $\varphi_V(\boldsymbol{p}_0)$:

$$\varphi_V(\boldsymbol{p}_0)=-\frac{1}{4\pi}\int_{V_S}\frac{1}{r}\nabla^2\varphi\,\mathrm{d}V;\qquad(3.33)$$

the second term is called single layer potential $\varphi_s l(\boldsymbol{p}_0)$:

$$\varphi_s l(\boldsymbol{p}_0)=-\frac{1}{4\pi}\int_{\S_{sl}}\frac{1}{r}\left(\left.\frac{\partial\varphi}{\partial n}\right|_+-\left.\frac{\partial\varphi}{\partial n}\right|_-\right)\mathrm{d}S;\qquad(3.34)$$

the third term is the double layer potential $\varphi_d l(\boldsymbol{p}_0)$:

$$\varphi_d l(\boldsymbol{p}_0)=\frac{1}{4\pi}\int_{\S_{dl}}(\varphi_+-\varphi_-)\nabla\frac{1}{r}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}S\qquad(3.35)$$

In the following chapters we will be see the physical meaning of the expressions (3.33), (3.34) and (3.35).

Chapter 4

Steady state electromagnetics

4.1 Electrostatics

The electrostatic formulation is derived by assuming that fields are not varying in time, and that no macroscopic charge motion is observed:

$$\frac{\partial}{\partial t} = 0 \tag{4.1}$$

$$\boldsymbol{J} = 0 \tag{4.2}$$

The governing equations in differential form are then obtained by introducing the assumptions (4.1) and (4.2) in (2.44) and in (2.50), along with the relevant constitutive relation:

$$\nabla \times \boldsymbol{E} = 0 \tag{4.3}$$

$$\nabla \cdot \boldsymbol{D} = \rho \tag{4.4}$$

$$\boldsymbol{D} = \epsilon \boldsymbol{E} \tag{4.5}$$

Since, due to (4.3), the electric field is everywhere a curl free field, it is also a conservative field and can be then derived as the gradient of an electric scalar potential φ:

$$\boldsymbol{E} = -\nabla \varphi. \tag{4.6}$$

Introducing (4.6) and (4.5) in (4.4), we obtain the governing differential equation of electrostatics:

$$\nabla \cdot (\epsilon \nabla \varphi) = -\rho. \tag{4.7}$$

Assuming an infinite domain with a uniform permittivity ϵ, one obtains the Poisson equation for the electric scalar potential:

$$\nabla^2 \varphi = -\frac{\rho}{\epsilon}, \tag{4.8}$$

which, under the assumption of regular conditions at infinity, according to (3.13) yields the solution:

$$\varphi(\boldsymbol{p}_0) = \frac{1}{4\pi\epsilon} \int_{\mathbb{R}^3} \frac{\rho}{r} \, \mathrm{d}V. \tag{4.9}$$

4.2 Magnetostatics

The formulation of the magnetostatics is derived by assuming that fields are not varying in time:

$$\frac{\partial}{\partial t} = 0 \tag{4.10}$$

The governing equations in differential form are then obtained by introducing the assumptions (4.10) in (2.39), along with (2.45) and the relevant constitutive relation:

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} \tag{4.11}$$

$$\nabla \cdot \boldsymbol{B} = 0 \tag{4.12}$$

$$\boldsymbol{B} = \mu \boldsymbol{H} \tag{4.13}$$

Since, due to (4.12), the magnetic induction field is a solenoidal field, it can be then derived as the curl of a magnetic vector potential $\boldsymbol{A}$:

$$\boldsymbol{B} = \nabla \times \boldsymbol{A} \tag{4.14}$$

Introducing (4.14) and (4.13) in (4.11), we obtain the governing differential equation of magnetostatics:

$$\nabla \times (\frac{1}{\mu} \nabla \times \boldsymbol{A}) = \boldsymbol{J}. \tag{4.15}$$

Assuming an infinite domain with a uniform permeability μ, one obtains the Poisson vector equation for the magnetic vector potential:

$$\nabla^2 \boldsymbol{A} = -\mu \boldsymbol{J}. \tag{4.16}$$

The gauge $\nabla \cdot \boldsymbol{A}$ has been used to derive (4.16). Under the assumption of regular conditions at infinity, according to (3.13), we can express the solution of (4.16) as:

$$\boldsymbol{A}(\boldsymbol{p}_0) = \frac{\mu}{4\pi} \int_{\mathbb{R}^3} \frac{\boldsymbol{J}}{r} \, \mathrm{d}V. \tag{4.17}$$

4.3 Steady state electrodynamics

The formulation of the steady state electrodynamics is derived by assuming that fields are not varying in time:

$$\frac{\partial}{\partial t} = 0 \tag{4.18}$$

The governing equations in differential form are obtained by introducing the assumptions (4.18) in in (2.44) and in (2.40), along with the relevant constitutive relation:

$$\nabla \times \boldsymbol{E} = 0 \tag{4.19}$$

$$\nabla \cdot \boldsymbol{J} = 0 \tag{4.20}$$

$$\boldsymbol{J} = \sigma (\boldsymbol{E} + \boldsymbol{E}_i) \tag{4.21}$$

As in the electrostatic case, due to (4.19), the electric field is a conservative field, and can be derived as the gradient of an electric scalar potential φ (see (4.6)).

Introducing (4.6) and (4.21) in (4.19), we obtain the governing differential equation of steady state elecrodynamics:

$$\nabla \cdot (\sigma \nabla \varphi) = \nabla \cdot (\sigma \boldsymbol{E}_i). \tag{4.22}$$

Chapter 5

Fundamentals of numerical analysis

5.1 Introduction

Numerical analysis is a discipline whose aim is to provide convenient methods for solving mathematical problems with the aid of a computing machine. The process of problem solving is constituted by a sequence of possible various distinct phases. The first phase is the mathematical formulation of a model representing the physical phenomenon of interest. The purposes of a model can be different: the detailed analysis of a given system, or the prediction of how such a system will evolve over time starting from a certain initial condition, or even the design of a new device or system that relies to a known set of physical laws. Analytical solution, when available, may often be of unacceptable form and does not provide an easy to interpret numerical result. More often, there is no known method of obtaining the solution in analytical form.

The second phase of the problem solving process is then to formulate a numerical model approximating the original one, which can be solved with the aid of a computer. A numerical method which can be used to solve a problem will be called an algorithm. An algorithm is a complete and unambiguous set of procedures leading to the solution of a mathematical problem.

The third phase of problem solving is implementing the algorithm on a computer, and to make it run to produce a numerical solution.

The selection or construction of appropriate algorithms properly falls within the scope of numerical analysis. Having decided on a specific algorithm or set of algorithms for solving the problem, all the sources of error that may affect the results have to be considered.

In formulating a physical-mathematical model, one usually introduces some errors, constituted by the approximations made in deriving the model.

In deriving the numerical model from the physical-mathematical one, some approximation are made to convert analytical operators (derivatives, integrals,...) in an equivalent form which can be treated by a computer. We will refer to this kind of errors as truncation errors.

The next class of errors is due to the number representation system adopted by a computer. Computers necessarily represent each real number with a finite length word. In this way a round-off error is introduced.

The last source of errors is the uncertainty which affects on the input data.

Despite the fact that, among the types of errors listed, only the second and the third one are originated by the numerical procedure adopted, one should be always aware of how errors of any kind propagate within the algorithm and affect the final result.

5.2 Number representation

5.2.1 Fixed point representation

The representation system we are familiar with is a positional system, i.e. a system in which the position of the digit within the number indicates the power of the base which multiplies this figure. So, in order to represent a positive integer number q in a given base β, we generally use the notation $(a_n\, a_{n-1} \cdots a_1\, a_0)_\beta$ to mean:

$$q = (a_n\, a_{n-1} \cdots a_1\, a_0)_\beta = a_n\beta^n + a_{n-1}\beta^{n-1} + \cdots + a_1\beta^1 + a_0\beta^0. \tag{5.1}$$

Here, the base β is a positive integer with $\beta \geq 2$, and the digits a_k with $k = 0, 1, \cdot, n$ are positive integer $0 \leq a_k \leq \beta - 1$ and $a_n \neq 0$. For example, the notation in base 10 $(2805)_{10}$ is equivalent to:

$$(2805)_{10} = 2 \cdot 10^3 + 8 \cdot 10^2 + 0 \cdot 10^1 + 5 \cdot 10^1 = 2 \cdot 1000 + 8 \cdot 100 + 0 \cdot 10 + 5 \cdot 1.$$

The representation of a real number x involves the use of the radix point (the decimal point in base 10), which separates the integer part of x from the fractional one. For a positive real number x, we define the integer part $\lfloor x \rfloor$ as the maximum integer less than or equal to x. The fractional part is then:

$$\text{frac}(x) = x - \lfloor x \rfloor$$

The fractional part can be written, for a given base β, as:

$$(0.b_1\, b_2 \cdots b_n \cdots)_\beta = b_1\beta^{-1} + b_2\beta^{-2} + \cdots + b_n\beta^{-n} + \cdots. \tag{5.2}$$

where the digits b_k with $k = 0, 1, \cdots \infty$ are positive integer $0 \leq b_k \leq \beta - 1$. Then, in base 10, the notation $(0.407)_{10}$ means:

$$(0.407)_{10} = 4 \cdot 10^{-1} + 0 \cdot 10^{-2} + 7 \cdot 10^{-3}.$$

A positive real number x can then be represented by joining the integer part encoded utilising (5.1) with the representation of the fractional part (5.2):

$$x = (a_n\, a_{n-1} \cdots a_1\, a_0 . b_1\, b_2 \cdots b_n \cdots)_\beta, \tag{5.3}$$

implying that:

$$x = a_n\beta^n + a_{n-1}\beta^{n-1} + \cdots + a_1\beta^1 + a_0\beta^0 + b_1\beta^{-1} + b_2\beta^{-2} + \cdots + b_n\beta^{-n} + \cdots.$$

The fixed point representation system is an encoding method utilised by computers. Fixed point representation relies on the notation (5.3), using a fixed number t of digits, and keeping fixed the position of the radix point. In other words, in this representation system a fixed number of digits is assigned to represent both the integer part and the fractional one. A fixed point numerical system is obtained defining the fixed point number set $X(\beta, t, q)$:

$$\mathbf{X}(\beta, t, q) = \left\{ x \in \mathbb{R} = \text{sign}(x) \left[\sum_{k=0}^{t-(q+1)} a_k\beta^k + \sum_{k=1}^{q} b_k\beta^{-k} \right] \beta^p \right\} \tag{5.4}$$

where the base β is a positive integer with $\beta \geq 2$, the number of significant digits t is a positive integer $t \neq 0$ and q is a positive integer with $0 \leq q \leq t$ specifying the number of digits assigned to represent the fractional part. It should be noted that the fixed point numbers set $\mathbf{X}(\beta, t, q)$ is a discrete and countable set. One can easily recognise that the number of

positive elements in $\mathbf{X}(\beta, t, q)$ is $\beta^t - 1$, the maximum element in $\mathbf{X}$ assumes the value $x_M = (\beta^t - 1)\beta^{-q}$, while the minimum positive element is $x_m = \beta^{-q}$. So, a fixed point representation based on the set $\mathbf{X}(10, 4, 1)$ is numeral system in base 10 which allows to represent a positive real number utilising a total of 4 digits, one of which constitutes the fractional part (the integer part is then formed by 3 digits). It is clear that the maximum number which can be represented by this system is $x_M = (999.9)_{10}$, while the minimum strictly positive number is $x_m = (000.1)_{10}$. Furthermore, the discretisation (that is the distance between two consecutive number) of this system is constant and equal to $(000.1)_{10}$. Since the discretisation can be regarded as the maximum error introduced representing a real number by means of a fixed point scheme, we can conclude that the absolute maximum error is constant over the entire range, while the relative error varies and tends to its higher values when x approaches to x_m (for example, in the fixed point numeral system $\mathbf{X}(10, 4, 1)$, the absolute error in representing the two number $x_1 = 10^3/3 = 333.\overline{3}$ and $x_2 = 1/3 = 0.\overline{3}$ is the same, that is $E_1 = E_2 = 0.0\overline{3} = 10^{-1}/3$. However the relative errors in representing x_1 and x_2 are $e_1 = E_1/x_1 = 10^{-4}$ and $e_2 = E_2/x_2 = 10^{-1}$.

int	.	fract
999	.	9
999	.	8
999	.	7
⋯	.	⋯
⋯	.	⋯
001	.	1
001	.	0
000	.	9
⋯	.	⋯
000	.	3
000	.	2
000	.	1
000	.	0

Table 5.1: *Fixed point numbers set* $\mathbf{X}(10, 4, 1)$

The advantage of fixed point arithmetic the advantage lies in the efficiency and the speed with which the arithmetic operations are performed by computers. The drawback is the loss of range and precision when compared with other number representations.

For this reason, fixed point representation is used when the speed of calculation is preferable to the accuracy of the result. Graphic applications, computer games and DSP programming often resort to this type of numeral representation. Conversely, fixed point numbers are rarely used used in scientific computation, and the floating point representation is preferred instead.

5.2.2 Floating point representation of real numbers

An alternative way of representing a real number is offered by the normalised representation. It can be shown that any real number $x \neq 0$ can be represented in a given base $\beta \in \mathbb{Z}, \beta \geq 2$ as:

$$x = \text{sign}(x)\Big[\sum_{k=0}^{\infty} d_k \beta^{-k}\Big]\beta^p \tag{5.5}$$

where the digits d_k are positive integers satisfying the conditions:

$$\begin{aligned} 1 \leq d_0 \leq \beta - 1; \\ 0 \leq d_k \leq \beta - 1, \quad \forall k \geq 1; \\ \nexists j \ : d_k = \beta - 1, \quad \forall k > j. \end{aligned}$$

The term:

$$m = \sum_{k=0}^{\infty} d_k \beta^{-k} \tag{5.6}$$

is called mantissa or significand. Due to the properties of the d_k coefficients, one can observe that:

$$1 \leq m < \beta. \tag{5.7}$$

The exponent p is an integer and is called characteristic of x. The term β^p is called exponential part.

Floating point representation is based on the the floating point numbers set $\mathbf{F}(\beta, t, L, U) \subset \mathbb{R}$, defined as:

$$\mathbf{F}(\beta, t, L, U) = \{0\} \cup \left\{ x \in \mathbb{R} = \text{sign}(x) \left[\sum_{k=0}^{t-1} d_k \beta^{-k} \right] \beta^p \right\} \tag{5.8}$$

The base β is again a positive integer with $\beta \geq 2$; the number of significant digits t in the mantissa is a positive integer; the integers L (usually with $L < 0$) and U (usually with $U > 0$) respectively specify the lower and upper limits of the range within which the exponent p can vary; the digits d_k are positive integer, with:

$$\begin{aligned} 1 \leq d_0 \leq \beta - 1; \\ 0 \leq d_k \leq \beta - 1, \quad \forall k \geq 1; \end{aligned}$$

To figure out the difference between the fixed point and the floating point representations, let us define a floating point set $\mathbf{F}(10, 3, 0, 9)$ and let us compare it with the numeral system $\mathbf{X}(10, 4, 1)$ defined in the previous section. It can be noted that both representation uses four figures (four digits in the fixed point notation, tree digits for the mantissa and one for the exponential in the floating point representation). The maximum number in $\mathbf{F}(10, 3, 0, 9)$ is $x_M = (9.99)_{10} 10^9 \approx 10^{10}$. The minimum non negative number in $\mathbf{F}(10, 3, 0, 9)$ is $x_M = (1.00)_{10} 10^0 = 1$. The floating point set $\mathbf{F}(10, 3, 0, 9)$ then spans a range much wider than the fixed point counterpart (ten orders of magnitude for the first set, four orders of magnitude for the latter). The discretisation of floating point is not constant over the entire set, since the constant spacing of the mantissa is amplified or reduced according to the exponential part. As we will see, this peculiarity results in a relative error which remain constant over the entire represented range.

The process which translates a given real number x into its floating-point representation $fl(x) \in \mathbf{F}$ is called rounding. There are different possible ways of rounding a number x:

rounding to the nearest : x is translated into the nearest $fl(x) \in \mathbf{F}$;

rounding toward zero (often referred to as chopping): x is translated into the nearest $fl(x) \in \mathbf{F}$ between x and 0;

rounding toward ∞ : x is translated into the nearest $fl(x) \in \mathbf{F}$ between x and ∞;

rounding toward $-\infty$: x is translated into the nearest $fl(x) \in \mathbf{F}$ between x and $-\infty$;

m	β^p
9.99	10^9
9.98	10^9
9.97	10^9
...	...
1.01	10^9
1.00	10^9
9.99	10^8
...	...
...	...
1.02	10^0
1.01	10^0
1.00	10^0
0.000	

Table 5.2: *Floating point numbers set* $\mathbf{L}(10,3,0,9)$

If, for example, we use the floating point set $\mathbf{F}(10,3,0,9)$, the two real numbers 2/3 and -2/3 will be converted:

$$fl\left[\frac{(2)_{10}}{(3)_{10}}\right] = fl\left[(0.\overline{6})_{10}\right] = \begin{cases} 6.67\,10^{-1} & \text{rounding to the nearest} \\ 6.66\,10^{-1} & \text{rounding toward } 0 \\ 6.67\,10^{-1} & \text{rounding toward } \infty \\ 6.66\,10^{-1} & \text{rounding toward } -\infty \end{cases}$$

and

$$fl\left[-\frac{(2)_{10}}{(3)_{10}}\right] = fl\left[(-0.\overline{6})_{10}\right] = \begin{cases} -6.67\,10^{-1} & \text{rounding to the nearest} \\ -6.66\,10^{-1} & \text{rounding toward } 0 \\ -6.66\,10^{-1} & \text{rounding toward } \infty \\ -6.67\,10^{-1} & \text{rounding toward } -\infty \end{cases}$$

All the rounding introduce a round off error $x - fl(x)$. The relative round off error is expressed as:

$$\delta = \frac{|\delta(x)|}{|x|} = \frac{|x - fp(x)|}{|x|} \tag{5.9}$$

Assuming that the number x is within the operative range of the floating point set utilised for its representation and that a rounding toward 0 is utilised, we can then write:

$$\delta = \frac{\left|\text{sign}(x)\left[\sum_{k=0}^{\infty} d_k\beta^{-k}\right]\beta^p - \text{sign}(x)\left[\sum_{k=0}^{t-1} d_k\beta^{-k}\right]\beta^p\right|}{\left|\text{sign}(x)\left[\sum_{k=0}^{\infty} d_k\beta^{-k}\right]\beta^p\right|} = \frac{\sum_{k=t}^{\infty} d_k\beta^{-k}}{\sum_{k=0}^{\infty} d_k\beta^{-k}}$$

We want to find the maximum relative round off error δ. Recalling (5.7) we see that:

$$\delta = \frac{\sum_{k=t}^{\infty} d_k\beta^{-k}}{\sum_{k=0}^{\infty} d_k\beta^{-k}} = \frac{\beta^{-t}\sum_{k=0}^{\infty} d_{t+k}\beta^{-k}}{\sum_{k=0}^{\infty} d_k\beta^{-k}} \leq \beta^{-t}\sum_{k=0}^{\infty} d_{t+k}\beta^{-k} < \beta^{-t}\beta = \beta^{1-t}, \tag{5.10}$$

that is, the relative round off error is always less than the quantity β^{1-t}. The maximum relative round off error ϵ is called machine precision (or machine epsilon, or unit round off). In a floating point system, the machine precision depends on the floating point number

set utilised and on the rounding method utilised. The most useful rounding mode, and the one which is almost always used, is round to nearest, since this always produces the floating point number which is closest to x, thus reducing the round off error. One can easily recognise that the maximum relative error rounding to the nearest is half that which would occur whit other rounding procedure. Thus, the result in (5.10) can be generalised as:

$$\epsilon = k\,\beta^{1-t}, \quad \text{with } k = \begin{cases} \frac{1}{2} & \text{rounding to the nearest} \\ \\ 1 & \text{other runding methods} \end{cases} \tag{5.11}$$

The floating point representation of a given real number x can be expressed as:

$$fl(x) = x(1+\delta), \quad |\delta| \leq \epsilon \tag{5.12}$$

where the relative round off error in representing x is a function of $\delta = \delta(x)$ whose absolute value is bounded by the machine precision given in (5.11).

5.2.3 The binary numeral system

Although the decimal system is used and accepted worldwide, There is no inherent reason to use 10 as a base. Indeed, there are several applications where other numeral systems are preferable. In almost all computers and computer-based devices, whose internal logic is based on electrical pulses that can assume two values (ON or OFF), a binary numeral system is used. A base-2 or binary system is a positional notation with a base of 2. Each digit, often referred to as a bit, can assume two different values, 0 or 1. It is also evident that, since in a normalised binary floating point system the leading digit d_1 has to be a non-zero positive integer, we necessarily have that $d_1 = 1$. For this reason, the leading digit is often implicitly assumed to be equal to 1 en it is not stored (hidden bit technique).

In a binary system, the arithmetic operations are performed with the same rules and properties of the decimal system. In 5.3 are reported the elementary arithmetic operations.

+	0	1
0	0	1
1	1	10

×	0	1
0	0	0
1	0	1

Table 5.3: *Elementay arithmetic operations in a binary numeral system*

To get an insight of a floating point binary system, let us consider a set number set much smaller than the ones actually utilised in computers: suppose that the mantissa has only three digits and the exponent p can vary from -1 to 2. The "toy" floating number set $\mathbf{F}(2,3,-1,2)$ is shown in Figure 5.1. The mantissa of the toy system can assume four different values, that are:

$$\begin{aligned}
(1.00)_2 = 1\cdot 2^0 + 0\cdot 2^{-1} + 0\cdot 2^{-2} &= (1.0)_{10} = (\frac{(4)_{10}}{(4)_{10}} \\
(1.01)_2 = 1\cdot 2^0 + 0\cdot 2^{-1} + 1\cdot 2^{-2} &= (1.25)_{10} = \frac{(5)_{10}}{(4)_{10}} \\
(1.10)_2 = 1\cdot 2^0 + 1\cdot 2^{-1} + 0\cdot 2^{-2} &= (1.5)_{10} = \frac{(6)_{10}}{(4)_{10}} \\
(1.11)_2 = 1\cdot 2^0 + 1\cdot 2^{-1} + 1\cdot 2^{-2} &= (1.75)_{10} = \frac{(7)_{10}}{(8)_{10}}
\end{aligned}$$

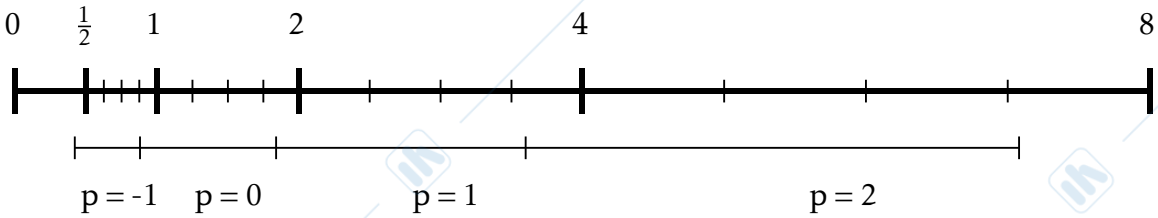


Figure 5.1: *Toy floating number set* $\mathbf{F}(2,3,-1,2)$

One can see that the spacing of the mantissa is 0.125 and that its value is determined by the less significant digit. The floating point set is obtained by multiplying the mantissa values by the exponential part, which can assume four possible value: $2^{-1} = (.5)_{10}$, $2^0 = (1)_{10}$, $2^1 = (2)_{10}$ and $2^2 = (4)_{10}$. One then obtain the set shown in table 5.4 in decimal notation: The positive elements of the toy floating point system are 32, and can divided in four groups,

$p = -1$	$p = 0$	$p = 1$	$p = 2$
$(0.5)_{10}$	$(1.0)_{10}$	$(2.0)_{10}$	$(4.0)_{10}$
$(0.625)_{10}$	$(1.25)_{10}$	$(2.5)_{10}$	$(5.0)_{10}$
$(0.75)_{10}$	$(1.5)_{10}$	$(3.0)_{10}$	$(6.0)_{10}$
$(0.875)_{10}$	$(1.75)_{10}$	$(3.5)_{10}$	$(7.0)_{10}$

Table 5.4: *Toy floating number set* $\mathbf{F}(2,3,-1,2)$

one for each exponent value. The spacing between the numbers is constant in each group, and increases with the exponential part. The region between 0 and the minimum floating point is called underflow region. We can see that the underflow range is much greater than the spacing of the first group (p=-1). The region of real numbers greater then the maximum in $\mathbf{F}$ is called overflow region.

The machine precision of an hypothetical computer representing the real numbers set by the toy floating number set $\mathbf{F}(2,3,-1,2)$ is $\epsilon = k\,2^{3-1} = k\,0.25$. This means that the computer will introduce a rounding error whenever a number x is converted in its representation $fp(x) \in \mathbf{F}(2,3,-1,2)$, and that the relative error δ will be at most 0.125 (12.5%) if a rounding to the nearest ($k = 0.5$) procedure is used, 0.25 (25.0%) if another rounding method is adopted.

Arithmetic operations between two floating point numbers are usually constructed so that the result is the floating point representation of the exact result. For example, the number z resulting from the addition between two real numbers x and y is represented by:

$$fl(z) = fl(x+y) = z(1+\delta(z))$$

The floating point result $fl(z)$ is generally not equal to z. It is worth noting that this implies that the properties of arithmetic operations in the real numbers set are no longer valid when operating on floating point numbers. For example, the sum of three real numbers x, y and z depends on the order by which the addition is executed, that is :

$$fl(fl(x+y)+z) \neq fl(fl(x+z)+y).$$

Indeed, computing the sum of $x = -1.75$, $y = -0.875$ and $z = 6.0$ in the toy floating point system with a rounding to the nearest mode:

$fl(fl(x+y)+z)$	$fl(fl(x+z)+y)$
$fl(fl((-1.75)+(-0.875))+(6.0)) =$	$fl(fl((-1.75)+(6.0))+(-0.875)) =$
$= fl(fl(-2.625)+(6.0)) =$	$= fl(fl(4.25)+(-0.875)) =$
$= fl((-2.5)+(6.0)) =$	$= fl((4.0)+(-0.875)) =$
$= fl(3.5) =$	$= fl(3.125) =$
$= 3.5$	$= 3.0$

The floating point systems used by computers does not differ substantially from the toy system $\mathbf{F}(2,3,-1,2)$, except of course for the length of the fields. In modern computers, the format of the floating point numbers binary representation and the floating point arithmetic are defined by the IEEE 754 standard. The IEEE 754 standard includes the technical specification of the rounding procedure, as well as the treatment of special cases (subnormal numbers, NaN (not a number), Inf (infinite)). The 32 bit (usually called single precision) and 64 bit (usually called double precision) representations are the main binary format in the IEEE 754 standard. A 32 bit number is divided into fields as follows: 1 bit for the sign, 8 bits for the exponent and 23 bits for the significand. Since the exponent field is 8 bits, it is used to represent exponents between -126 and 127. The format uses the hidden bit technique and does not store the leading digit, which is assumed to be equal to 1 for the binary normalized representation. As a result, the number of significant digits is a actually 24, that is the number of bits representing the significand (23) plus one. The machine precision is then $eps = k\,2^{24-1} = k\,1.192\,10^{-7}$. The 64 bit number representation is very similar to the 32 bit one, with wider fields: 1 bit is reserved for the sign, 11 bits for the exponent and 52 bits for the mantissa. The exponent ranges between -1022 and 1023, and the machine precision is $eps = k\,2^{52} = k\,2.22\,10^{-16}$.

5.3 Condition and stability of an algorithm

The condition of a procedure (not necessarily a numerical procedures) indicates the sensitivity of the procedure results to variations in input arguments. Considering a function f of the input data x, we can express the change of the function result Δf due to a (small) change δ in the input as:

$$\Delta f(x) = |f(x+\delta) - f(x)| \simeq |f(x) + f'(x)\delta) - f(x)| = |f'(x)\delta)|$$

where $f'(x)$ is the function derivative with respect to x. The relative variation in the function f is then:

$$\frac{\Delta f(x)}{|f(x)|} \simeq \left|\frac{f'(x)}{f(x)}\delta\right| = \left|\frac{f'(x)}{f(x)}x\right|\left|\frac{\delta}{x}\right|$$

The (local) condition can be defined as the relative change in the function output caused by a unit relative change in the argument:

$$A(x) = \left|\frac{f'(x)}{f(x)}x\right| \tag{5.13}$$

Definition (5.13) allows us to express the relative variation of the function $f(x)$ as:

$$\frac{\Delta f(x)}{|f(x)|} \simeq A(x)\left|\frac{\delta}{x}\right|$$

In a (locally) well conditioned problem, a small change in the function argument produces a relative change in the output of the same order of magnitude (i.e. $A \sim 1$). Conversely, in ill conditioned problems, small perturbation in the input argument are amplified and produces large relative change in the result ($A \gg 1$). We can also conveniently define a global condition (or condition number) as the maximum value of the local condition defined in (5.13):

$$A_g = \max\{A(x)\} \tag{5.14}$$

In a (globally) well conditioned problem (i.e. $A_g \sim 1$), small variations in the function argument will always produce small relative variations in the output. A (globally) ill conditioned problem (i.e. $A_g \sim 1$), instead, presents some (at least one) critical regions in its domain of definition where small changes in the input cause large variation of the result.

The related notion of stability describes how a numerical process is sensitive to the inevitable rounding errors committed during its execution in finite precision floating point arithmetic. The precise effect of these errors on the accuracy of the computed value for $f(x)$ is hard to determine except by actually carrying out the computations for particular finite precision arithmetic and comparing the computed answer with the exact analytical.

There is a subtle, but still important difference between the two concepts of condition and stability of an algorithm. While, for a well conditioned problem one can usually find a stable numerical procedure, an ill conditioned problem will generally require to be reformulated in a well conditioned equivalent form.

5.4 Numerical derivatives

Let $f(x)$ be a twice continuously derivable function defined on a given interval $[a,b]$ of the variable x. Let $x_1, x_2, \ldots, x_k, \ldots, x_n$ be a set of n evenly spaced values of the variable x, and let $\Delta = x_k - x_{k-1}$, with $k \neq 1$, be the constant spacing between these values. Then, the function $f(x)$ can be expressed as Taylor series centred in the generic point x_k:

$$f(x) = f(x_k) + \left.\frac{df}{dx}\right|_{x_k}(x-x_k) + \frac{1}{2}\left.\frac{d^2f}{dx^2}\right|_{x_k}(x-x_k)^2 + \frac{1}{6}\left.\frac{d^3f}{dx^3}\right|_{x_k}(x-x_k)^3 + \cdots$$

Then, indicating for sake of brevity $f_{k-1} = f(x_{k-1})$, $f_k = f(x_k)$ and $f_{k+1} = f(x_{k+1})$ we can write:

$$\begin{aligned} f_{k+1} &= f_k + \left.\frac{df}{dx}\right|_{x_k}\Delta + \frac{1}{2}\left.\frac{d^2f}{dx^2}\right|_{x_k}\Delta^2 + O(\Delta^3) \\ f_{k-1} &= f_k - \left.\frac{df}{dx}\right|_{x_k}\Delta + \frac{1}{2}\left.\frac{d^2f}{dx^2}\right|_{x_k}\Delta^2 + O(\Delta^3) \end{aligned} \tag{5.15}$$

From (5.15), the central finite difference formula for the first derivative can be obtained:

$$\left.\frac{df}{dx}\right|_{x_k} = \frac{f_{k+1} - f_{k-1}}{2\Delta} + O(\Delta^2) \tag{5.16}$$

In a a similar fashion, considering a Taylor series to the third derivative:

$$\begin{aligned} f_{k+1} &= f_k + \left.\frac{df}{dx}\right|_{x_k}\Delta + \frac{1}{2}\left.\frac{d^2f}{dx^2}\right|_{x_k}\Delta^2 + \frac{1}{6}\left.\frac{d^3f}{dx^3}\right|_{x_k}(\Delta)^3 + O(\Delta^4) \\ f_{k-1} &= f_k - \left.\frac{df}{dx}\right|_{x_k}\Delta + \frac{1}{2}\left.\frac{d^2f}{dx^2}\right|_{x_k}\Delta^2 - \frac{1}{6}\left.\frac{d^3f}{dx^3}\right|_{x_k}(\Delta)^3 + O(\Delta^4) \end{aligned} \tag{5.17}$$

we can derive a central finite difference expression for the second derivative:

$$\left.\frac{d^2 f}{d x^2}\right|_{x_k} = \frac{f_{k+1} - 2 f_k + f_{k-1}}{\Delta^2} + O(\Delta^2) \tag{5.18}$$

The term $O(\Delta^2)$ indicates that, calling E the truncation error affecting the formulas, then $\exists M \in \mathbb{R}, M > 0 : |E| < M \Delta^2$. That is, the truncation error tends to 0 as Δ^2 does. A similar approach allows us to find the forward and backward derivation formulas. Forward and backward finite difference formulas can be derived by a first degree Taylor series:

$$f_{k+1} = f_k + \left.\frac{d f}{d x}\right|_{x_k} \Delta + \cdots$$
$$f_{k-1} = f_k - \left.\frac{d f}{d x}\right|_{x_k} \Delta + \cdots,$$

from which one can obtain:

$$\left.\frac{d f}{d x}\right|_{x_k} = \frac{f_{k+1} - f_k}{\Delta} + O(\Delta) \tag{5.19}$$

$$\left.\frac{d f}{d x}\right|_{x_k} = \frac{f_k - f_{k-1}}{\Delta} + O(\Delta) \tag{5.20}$$

The forward (5.19) and backward (5.20) finite formulas introduce a greater truncation error with respect to the central finite difference (5.16) in evaluating the derivative.

5.5 Interpolation

In the most general sense, the interpolation meets the need to evaluate, given a set of data points contained within a certain range D of the variable $\boldsymbol{x}$, new data points for any value of the variable within the range D. Considering a function $f(x)$, whose values $f_1 = f(x_1)$, $f_2 = f(x_2)$, …, $f_n = f(x_n)$ are known for a certain set of n points x_1, x_2, …, x_n, hereafter referred to as nodes, interpolation consists in finding a function $\tilde{f}(x)$ such that $\tilde{f}(x_1) = f_1$, $\tilde{f}(x_2) = f_2$, …, $\tilde{f}(x_n) = f_n$. The data points $x_1; f_1, x_2; f_2, \ldots, x_n; f_n$ on the basis of which interpolation is constructed are often obtained by sampling or experimentation, and represent then the values of an unknown function $f(x)$ for a limited number of values of the independent variable. On other occasions, the data points are derived by sampling the values of a function that is known, but is too complex to evaluate efficiently. In these cases, interpolation provides a simpler function, with some loss in precision.

5.5.1 Polynomial interpolation

Polynomial interpolation consists in finding a polynomial of degree $\leq n - 1$ which interpolates a given set of n data points. The most common way to express a polynomial $p(x)$ of degree d is the power form:

$$p(x) = a_0 + a_1 x^1 + a_2 x^2 + \ldots + a_d x^d \tag{5.21}$$

There are indeed other formulations in which a polynomial can be expressed. Probably, the use of the Lagrange polynomials allows the most straightforward definition of a polynomial interpolating a given set of n data points $x_1; f_1, x_2; f_2, \ldots, x_n; f_n$. For any interpolation node x_k, the Lagrange polynomial $l_k(x)$ can be defined as a $n-1$ degree polynomial having roots at each nodes x_j, $j = 1, \ldots, n$, except for x_k, where it takes unit value:

$$l_k(x) = \frac{\prod_{\substack{j=1,\\ j\neq k}}^{n}(x - x_j)}{\prod_{\substack{j=1,\\ j\neq k}}^{n}(x_k - x_j)} \tag{5.22}$$

Once defined the Lagrange polynomials, the interpolating polynomial function $\tilde{f}(x)$ can be expressed as:

$$\tilde{f}(x) = f_1\, l_1(x) + f_2\, l_2(x) + \ldots + f_n\, l_n(x) = \sum_{k=1}^{n} f_k\, l_k(x) \tag{5.23}$$

Equation (5.23) shows that there is at least one polynomial of degree $\leq n-1$ which interpolates $f(x)$ at the n distinct points $x_1, x_2, \ldots, x_n$. However, the Lagrange formula (5.23) is not the most computationally efficient option. For practical purposes, the polynomial in its Newton form is preferred instead:

$$\tilde{f}(x) = c_1 + c_2(x - x_1) + c_3(x - x_2)(x - x_1) + \ldots + c_n(x - x_{n-1})(x - x_{n-2})\ldots(x - x_2)(x - x_1), \tag{5.24}$$

or its equivalent:

$$\begin{aligned}\tilde{f}(x) = c_1 + (x - x_1)\{c_2 + (x - x_2)[c_3 + (x - x_3)\langle c_4 + \ldots \\ \ldots + (x - x_{n-2})(c_{n-1} + (x - x_{n-1})c_n)\ldots\rangle]\},\end{aligned} \tag{5.25}$$

The nested Newton form (5.25) is particularly convenient for computational purposes, since it requires $2(n-1)$ additions and $n-1$ multiplications for evaluating $\tilde{f}(x)$ at any x. To compare with the Lagrange formulation, one should consider that, once the the data points are defined and the denominators of the Lagrange polynomials (5.22) are computed once for all, evaluating the interpolating function according to (5.23) requires $n(n-1)$ additions and $n(n-1)$ multiplications.

The coefficients $c_1, c_2, \ldots, c_n$, are obtained by imposing that the function $\tilde{f}(x)$ interpolates the data points: $\tilde{f}(x_1) = f_1$, $\tilde{f}(x_2) = f_2, \ldots, \tilde{f}(x_n) = f_n$. The n conditions form the linear system:

$$\begin{cases} c_1 & = f_1 \\ c_1 + c_2(x_2 - x_1) & = f_2 \\ \ldots & \\ c_1 + c_2(x_n - x_1) + c_3(x_n - x_2)(x_n - x_1) + \ldots + c_n(x_n - x_{n-1})(x_n - x_{n-2})\ldots(x_n - x_2)(x_n - x_1) & = f_n \end{cases}$$

Introducing the notation:

$$\begin{aligned} & f[x_1] = f_1; \\ & f[x_1, x_2] = \frac{f[x_2] - f[x_1]}{x_2 - x_1}; \\ & f[x_1, x_2, x_3] = \frac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1}; \\ & \ldots \\ & f[x_1, x_2, x_3, \ldots, x_k] = \frac{f[x_2, x_3, \ldots, x_k] - f[x_1, x_2, \ldots, x_{k-1}]}{x_k - x_1}; \\ & \ldots \\ & f[x_1, x_2, x_3, \ldots, x_n] = \frac{f[x_2, x_3, \ldots, x_n] - f[x_1, x_2, , \ldots, x_{n-1}]}{x_n - x_1}, \end{aligned} \tag{5.26}$$

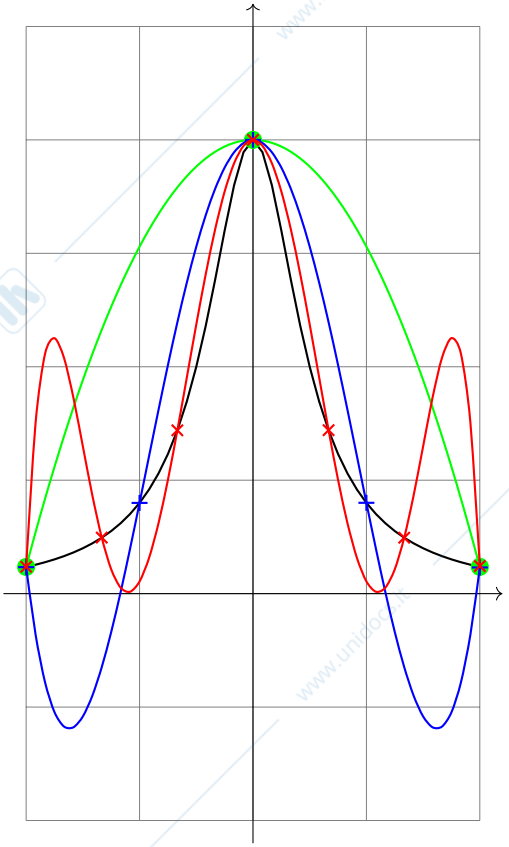


Figure 5.2: *Polynomial Interpolation of the Runge function (black line) in the [a,-a] interval: 2^{nd} degree polynomial (green line), 4^{th} degree polynomial (blue line), 6^{th} degree polynomial (red line). Sampling points used for evaluating the three interpolating polynomials are represented by dots, + and x respectively.*

the above system produces the solution:

$$
\begin{aligned}
c_1 &= f[x_1] \\
c_2 &= f[x_1, x_2] \\
c_3 &= f[x_1, x_2, x_3] \\
&\dots \\
c_k &= f[x_1, x_2, x_3, \dots, x_k] \\
&\dots \\
c_n &= f[x_1, x_2, x_3, \dots, x_n]
\end{aligned}
$$

The term $f[x_1, x_2, x_3, \dots, x_k]$ recursively defined in (5.26) is called the k^{th} divided difference of $f(x)$ at the points x_1, x_2, x_3, …, x_k.

Although the Weierstrass theorem states that for any continuous function $f(x)$ defined on a closed interval $[a, b]$ can be uniformly approximated by a polynomial function, there is no guarantee that the polynomial interpolating this function uniformly converges to the function itself increasing the number of the sampling points. Particularly, for evenly spaced data points, the interpolating polynomial can be affected by oscillations, which are more evident in proximity of the edges of the interval, and increase in amplitude and frequency for increasing number of data points (and then, for increasing degree of the interpolating

polynomial). This behaviour is called Runge phenomenon. A typical example occurs when attempting to interpolate the so called Runge function $r(x)$:

$$r(x) = \frac{1}{(1 + a^2 x^2)} \tag{5.27}$$

where a is any real constant. As can be observed in Fig. 5.2, increasing the number of data points for 3 to 5 and then to 7 results in increasingly oscillating polynomials. One can show that, in this case, the upper bound to the error between the generating function $r(x)$ and the interpolating polynomial increases with the number of sampling points n.

The oscillation problem of higher order interpolating polynomials can be avoided with a more adequate choice of interpolation nodes. By placing the nodes more densely towards the edges of the considered interval, one can generally avoid oscillation problems. In doing so, the frequency of the nodes become a variable of the problem, and has to be determined case by case. Additionally, on many occasions one does not have a-priori knowledge of the function to be interpolated, and does not completely control the nodes positioning; this is the case where an experimental data set has to be interpolated. A more flexible alternative is offered by piecewise polynomial interpolation.

5.5.2 Piecewise polynomial interpolation

In piecewise polynomial interpolation, one seek to construct the interpolating function on a given interval $[a, b]$ using a set of low order polynomials, each of them extending on small portion of $[a, b]$. Some conditions will also be enforced to satisfy the desired continuity properties of the interpolating function. In this approach the desired accuracy is achieved by increasing the number of pieces in which the interval $[a, b]$ is divided, thus increasing the number of polynomials, without increasing their degree.

Piecewise linear interpolation

Piecewise linear interpolation is ideally carried out by joining the data points by means of straight lines. In order to interpolate a given set of n data points x_1; f_1, x_2; f_2, …, x_n; f_n, (with $x_1 < x_2 < x_3 \cdots < n_{n-1} < x_n$) using this approach, the interval $[x_1, x_n]$ is divided in $n-1$ sub-intervals $[x_1, x_2]$, $[x_2, x_3]$, …, $[x_{n-1}, x_n]$. As shown in fig. 5.3, on each of these sub-intervals, the interpolation assumes a linear trend, while the first derivative is constant on the sub-intervals and shows a discontinuity at the nodes position x_1, x_2, …, x_n. Consequently, the second derivative is not defined at the nodes , and it is zero elsewhere. The piecewise linear interpolating function then belongs the to the C^0 differentiability class of functions.

A piecewise linear interpolation can be formally expressed defining a unit piecewise linear function $L_k(x)$ for each node $x_1, x_2, \ldots, x_n$. For a generic node x_k internal to the interval $[x_1, x_n]$, defining $\Delta_- = x_k - x_{k-1}$ and $\Delta_+ = x_{k+1} - x_k$, the function $L_k(x)$ can be expressed as:

$$L_k(x) = \begin{cases} 1 + \frac{x - x_k}{\Delta_-} & x \in [x_{k-1}, x_k] \\ 1 - \frac{x - x_k}{\Delta_+} & x \in [x_k, x_{k+1}] \\ 0 & x \notin [x_{k-1}, x_{k+1}] \end{cases} \tag{5.28}$$

The function $L_k(x)$ is then unitary at the node x_k, and it is zero at all other nodes. It has non-zero values only in the interval $[x_{k-1}, x_{k+1}]$ where it has a piecewise linear behaviour, with constant slopes $1/\Delta_-$ and $-1/\Delta_-$ in $]x_{k-1}, x_k[$ and $]x_k, x_{k+1}[$, respectively. We call support domain Ω_k of the node x_k the interval $[x_{k-1}, x_{k+1}]$ on which $L_k \neq 0$. The support domain is constituted by the sub-intervals having the node x_k as an edge, and constitutes the domain of

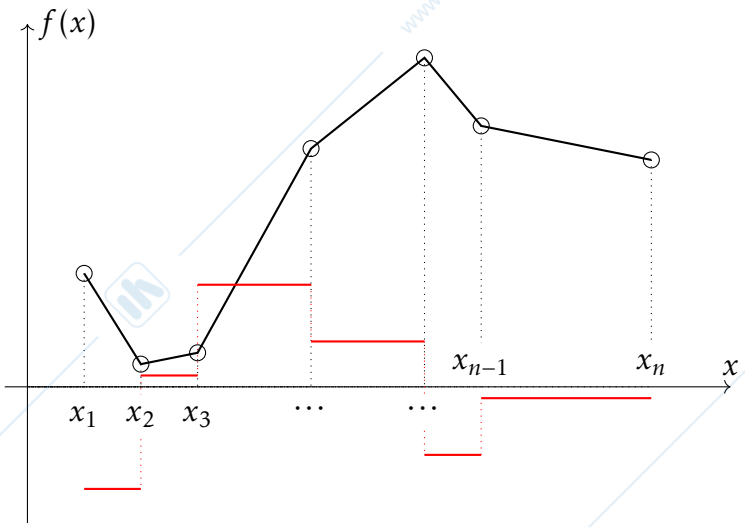


Figure 5.3: *Linear piecewise interpolation (black line) of a given set of data points (black circle). The red line represents the first derivative of the interpolating function.*

influence of k^{th} data point. The support domains of the edges of the interpolation domain x_1 and x_n extend of course only on one sub-interval, that are $[x_1, x_2]$ and $[x_{n-1}, x_n]$, respectively:

$$L_1(x) = \begin{cases} 1 - \frac{x - x_k}{\Delta_+} & x \in [x_1, x_2] \\ 0 & x \notin [x_1, x_2] \end{cases} \qquad L_n(x) = \begin{cases} 1 + \frac{x - x_n}{\Delta_-} & x \in [x_{n-1}, x_n] \\ 0 & x \notin [x_{n-1}, x_n] \end{cases} \tag{5.29}$$

As can be observed in Fig. 5.4, on any sub-interval $[x_{k-1}, x_k]$, only the two functions $[L_{k-1}(x)$ and L_k assumes non-zero values. The functions $L_k(x)$ constitute the basic building blocks for constructing the piecewise linear interpolation, which assumes the the form:

$$\tilde{f}(x) = f_1 L_1(x) + f_2 L_2(x) + \ldots + f_n L_n(x) = \sum_{k=1}^{n} f_k L_k(x) \tag{5.30}$$

Since the piecewise linear interpolation is linear on any sub-interval $[x_{k-1}, x_k]$, it can be shown that the error in interpolating any $f(x) \in C^2$ is bounded. Indeed, for any sub-interval $[x_{k-1}, x_k]$, calling $\Delta = x_k - x_{k-1}$ the width of the sub-interval:

$$|f(x) - \tilde{f}(x)| \leq \left(\frac{\Delta}{2}\right)^2 \max_{\xi \in [x_{k-1}, x_k]} \left\{ \frac{f''(\xi)}{2} \right\}$$

Piecewise linear interpolation provide then a second order accuracy in interpolating the function $f(x)$.

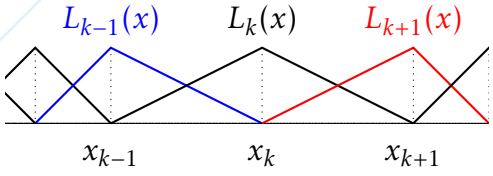


Figure 5.4: *Unit piecewise linear function* $L_k(x)$.

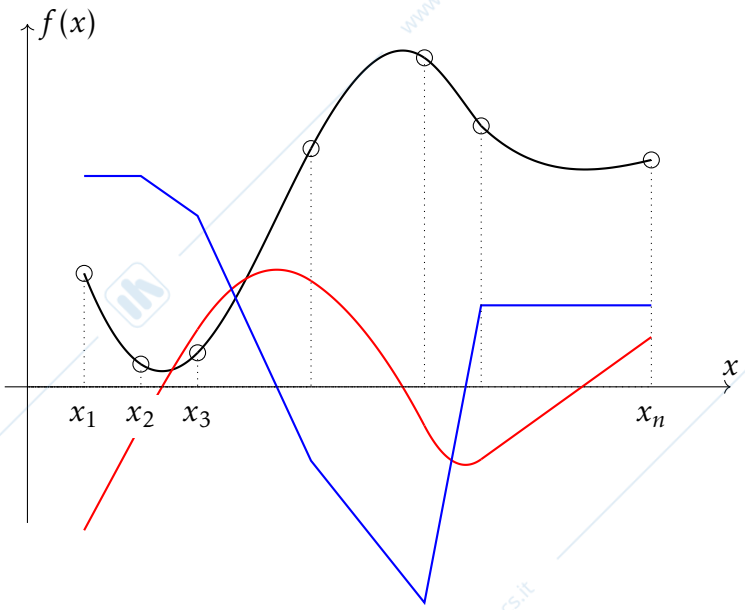


Figure 5.5: *Cubic spline interpolation (black line) of a given set of data points (black circle). The red and blue lines represent the first and the second derivatives of the interpolating function. Compare with the piecewise linear interpolation in Fig. 5.3*

Cubic spline

The degree of the piecewise polynomial can be increased to enhance the smoothness of the interpolating function. A cubic spline provides an interpolating piecewise polynomial $\tilde{f}(x) \in C^2$, that is a piecewise polynomial continuous up to the second derivative. A cubic spline interpolating a given set of n data points $x_1; f_1, x_2; f_2, \ldots, x_n; f_n$ may be constructed correcting the linear piecewise expression (5.30) by adding on each sub-interval $[x_i, x_i]$ two functions cubic polynomials $C_i(x)$ and $C_j(x)$, so that:

$$\tilde{f}(x) = f_i L_i(x) + f_j L_j(x) + C_i + C_j \tag{5.31}$$

Since $\tilde{f}(x)$ has to interpolate the given set of data points, we must have $\tilde{f}(x_i) = f_i$ and $\tilde{f}(x_j) = f_j$, and also:

$$\begin{cases} C_i(x_i) = 0 \\ C_i(x_j) = 0 \end{cases} \quad a) \qquad \begin{cases} C_j(x_i) = 0 \\ C_j(x_j) = 0 \end{cases} \quad b) \tag{5.32}$$

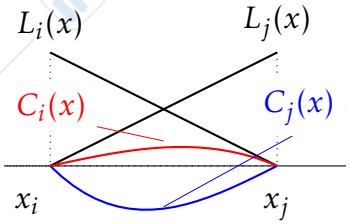


Figure 5.6: *Unit piecewise linear function $L_i(x)$ and $L_j(x)$, along with the cubic corrections $C_i(x)$ (red) and $C_j(x)$ (blue).*

Calling $\Delta = x_i - x_j$ the width of the considered sub-interval, we observe that the expression of the two functions $L_i(x)$ and $L_j(x)$ on $[x_i, x_j]$ are:

$$L_i(x) = 1 - \frac{x - x_i}{\Delta} \quad L_j(x) = 1 + \frac{x - x_j}{\Delta}$$

We can then express the first and second derivative of the interpolating function $\tilde{f}(x)$ as:

$$\frac{\mathrm{d}\tilde{f}}{\mathrm{d}x} = \frac{f_j - f_i}{\Delta} + \frac{\mathrm{d}C_i}{\mathrm{d}x} + \frac{\mathrm{d}C_j}{\mathrm{d}x} \tag{5.33}$$

$$\frac{\mathrm{d}^2\tilde{f}}{\mathrm{d}x^2} = \frac{\mathrm{d}^2 C_i}{\mathrm{d}x^2} + \frac{\mathrm{d}^2 C_j}{\mathrm{d}x^2}. \tag{5.34}$$

Since we require that the second derivative of the interpolating function is a continuous function in the interval $[x_1, x_n$, we express it as a piecewise linear interpolation of the n data points $x_1; f''_1, x_2; f''_2, \ldots, x_n; f''_n$:

$$\frac{\mathrm{d}^2\tilde{f}}{\mathrm{d}x^2} = f''_1 L_1(x) + f''_2 L_2(x) + \ldots + f''_n L_n(x) = \sum_{k=1}^{n} f''_k L_k(x)$$

On the generic sub-interval $[x_i, x_j]$ we have that:

$$\frac{\mathrm{d}^2\tilde{f}}{\mathrm{d}x^2} = f''_i L_i(x) + f''_j L_j(x),$$

whence, comparing with (5.34), we obtain:

$$\frac{\mathrm{d}^2 C_i}{\mathrm{d}x^2} = f''_i L_i(x), \tag{5.35a}$$

$$\frac{\mathrm{d}^2 C_i}{\mathrm{d}x^2} = f''_j L_j(x) \tag{5.35b}$$

The equations (5.35), along with the boundary conditions (5.32), constitute two differential problems which may be solved giving:

$$C_i(x) = \frac{\Delta^2}{6} f''_i (L_i(x)^2 - 1) L_i(x), \tag{5.36a}$$

$$C_j(x) = \frac{\Delta^2}{6} f''_j (L_j(x)^2 - 1) L_j(x), \tag{5.36b}$$

In order to determine the unknown parameters $f''_1, f''_2, \ldots, f''_n$ we have to enforce the continuity condition for the first derivative. Introducing (5.36) in (5.33), one can express the first derivative of $\tilde{f}$ on x_i, x_j as:

$$\frac{\mathrm{d}\tilde{f}}{\mathrm{d}x} = \frac{f_j - f_i}{\Delta} + \frac{\Delta}{6}[f''_j (3L_j^2 - 1) - f''_i (3L_i^2 - 1)] \tag{5.37}$$

Consider now two contiguous sub-intervals x_{k-1}, x_k and x_k, x_{k+1}, joining at the node x_k. Using (5.37) for expressing the derivative at x_k for the left ($i = k-1$ and $j = k$) sub-interval ad for the right one ($i = k$ and $j = k+1$), we can enforce the continuity conditions for any $k = 2, \cdot, n-1$, obtaining:

$$\left.\frac{\mathrm{d}\tilde{f}}{\mathrm{d}x}\right|_{x_{k-1,k}} = \frac{f_k - f_{k-1}}{\Delta_-} + \frac{\Delta_-}{6}(2f''_k + f''_{k-1}) = \frac{f_{k+1} - f_k}{\Delta_+} - \frac{\Delta_+}{6}(2f''_k + f''_{k+1}) = \left.\frac{\mathrm{d}\tilde{f}}{\mathrm{d}x}\right|_{x_{k,k+1}},$$

whence:

$$\Delta_- f''_{k-1} + 2(\Delta_- + \Delta_+) f''_k + \Delta_+ f''_{k+1}) = 6\left(\frac{f_{k-1} - f_k}{\Delta_-} + \frac{f_{k+1} - f_k}{\Delta_+}\right), \tag{5.38}$$

where $\Delta_- = x_k - x_{k-1}$ and $\Delta_+ = x_{k+1} - x_k$. The $n-2$ equations obtained by condition (5.38) $\forall k = 2, 3, \cdots, n-1$ can be completed by two conditions, wich specify the desired behaviour of the interpolating function at the nodes x_1 and x_n. The so called natural conditions assign a zero value to the parameters f''_1 and f''_n:

$$f''_1 = 0 \tag{5.39a}$$
$$f''_n = 0 \tag{5.39b}$$

Another commonly adopted conditions are the parabolic conditions, that are:

$$f''_1 - f''_2 = 0 \tag{5.40a}$$
$$f''_n - f''_{n-1} = 0 \tag{5.40b}$$

The system obtained from (5.38) along with (5.39) (or (5.40)) is then solved to give the values of the unknown parameters $f''_1, f''_2, \ldots, f''_n$ which define the cubic spline.

Piecewise polynomial interpolation in two dimensions

The concept of piecewise polynomial interpolation can be extended to represent functions depending on two or more variables. In this section, we will mainly discuss the linear piecewise polynomial interpolation of a function, whose values $f_1 = f(x_1, y_1)$, $f_2 = f(x_2, y_2)$, $\cdots$, $f_n = f(x_n, y_n)$ are known at a given set of n nodes $\boldsymbol{p}_1 = (x_1; y_1)$, $\boldsymbol{p}_2 = (x_2; y_2)$, $\cdots$, $\boldsymbol{p}_n = (x_n; y_n)$, although higher degree interpolations are possible once an appropriate set of basis functions are defined.

A polynomial piecewise interpolation $\tilde{f}(x, y)$ of the given set of data points can be obtained dividing the interpolation domain in a set of subdomains (or elements), on which $\tilde{f}(x, y)$ assumes a pre-defined polynomial behaviour. A linear piecewise interpolation of a function defined on a plane domain requires then a triangulation of the domain itself using the given set of nodes. The triangulation consists in dividing the domain into triangles, each of which having vertices at the nodes. Any two triangular elements in the triangulation may intersect at one common edge, or does not intersect at all. A very commonly used triangulation is the Delaunay triangulation. A Delaunay triangulation of a given set of nodes maximize the minimum angle of all the angles of the triangles in the triangulation. In other worlds, a Delaunay triangulation tends to avoid sharp triangles with small acute angles.

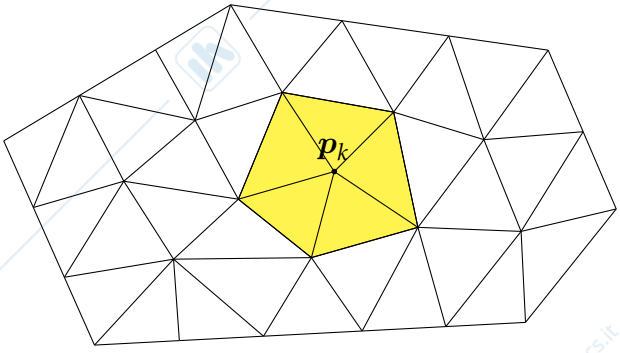


Figure 5.7: *triangulation of a plane domain. The support domain of the node $\boldsymbol{p}_k$ is highlighted.*

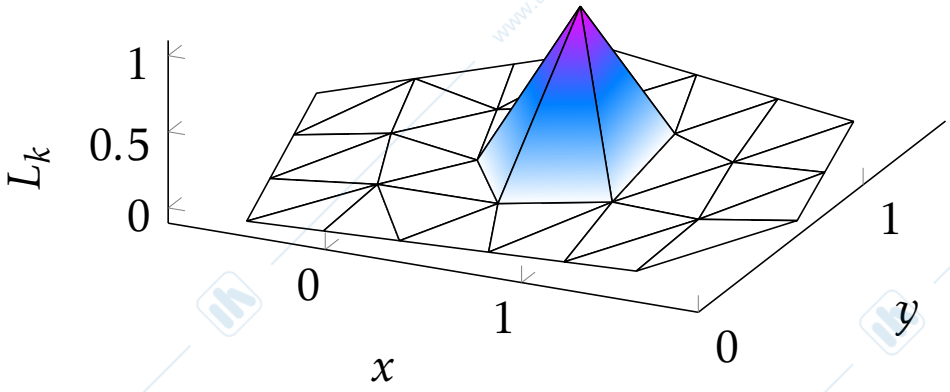


Figure 5.8: *unitary piecewise linear function* $S_k(x,y)$.

Once a triangulation of the given nodes set has been defined we have to define a set of base functions in order to formulate the interpolating function $\tilde{f}(x,y)$. Proceeding in a similar manner to what was done in the case of interpolation in one dimension, we define the piecewise linear unit function $L_k(x,y)$ at the generic node $\boldsymbol{p}_k = (x_k;y_k)$ requiring that $L_k(x_k,y_k) = 1$, $L_k(x_j,y_j) = 0, \forall j \neq k$. Then, defining the support domain Ω_k of the node $\boldsymbol{p}_k = (x_k;y_k)$ as the set of triangles having $\boldsymbol{p}_k$ as a vertex, the function $L_k(x,y) \neq 0$ only if $\boldsymbol{p} = (x;y) \in \Omega_k$, and $L_k(x,y)$ is linear on each triangle of Ω_k. We consider now a triangle $T_{ijk} \in \Omega_k$, having vertices at the node $\boldsymbol{p}_k = (x_k;y_k)$, $\boldsymbol{p}_i = (x_i;y_i)$ and $\boldsymbol{p}_j = (x_j;y_j)$. The area S of the triangle can be expressed as:

$$S = \frac{1}{2}\begin{vmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix}$$

A generic point $\boldsymbol{p} = (x;y)$ internal to T_{ijk} introduce in T_{ijk} a subdivision in three triangles, having vertices in $\boldsymbol{p}$ and in two of the three nodes $\boldsymbol{p}_k$, $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$, as shown in fig 5.9. The area of these triangles will be expressed as:

$$S_i = \frac{1}{2}\begin{vmatrix} 1 & x & y \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix} \quad S_j = \frac{1}{2}\begin{vmatrix} 1 & x_i & y_i \\ 1 & x & y \\ 1 & x_k & y_k \end{vmatrix} \quad S_k = \frac{1}{2}\begin{vmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x & y \end{vmatrix}$$

One can easily verify that the above expressions $S_i(x,y)$, $S_j(x,y)$ and $S_k(x,y)$ are linear functions of the point $\boldsymbol{p}$ coordinates x and y. Thus, the ratio:

$$\frac{S_k(x,y)}{S}$$

is a linear function assuming unitary value at $\boldsymbol{p}_k$, and a zero value at the other two vertices $\boldsymbol{p}_i$ =, and $\boldsymbol{p}_j$, and can be taken as the definition of the unit piecewise linear function on the element T_{kij}. Since T_{kij} also belongs to the support domains Ω_i and Ω_j, we can express in a similar way the other two unit piecewise linear functions $L_i(x,y)$ and $L_j(x,y)$ assuming non zero values on T_{kij}. We have then:

$$L_i(x,y) = \frac{S_k(x,y)}{S} \quad L_j(x,y) = \frac{S_k(x,y)}{S} \quad L_k(x,y) = \frac{S_k(x,y)}{S} \tag{5.41}$$

The three functions $L_i(x,y)$, $L_j(x,y)$ and $L_k(x,y)$ are also called area coordinates or barycentric coordinates, as any position $\boldsymbol{p}$ within the triangle T_{kij} can be expressed as:

$$\boldsymbol{p} = L_k\,\boldsymbol{p}_j + L_i\,\boldsymbol{p}_i + L_j\,\boldsymbol{p}_j.$$

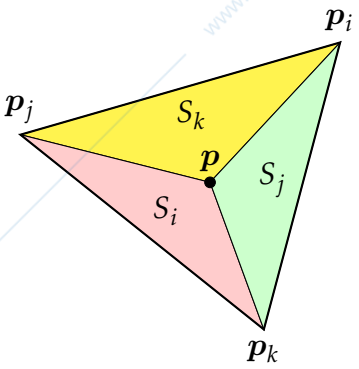


Figure 5.9: *Definition of the areas $S_i(x,y)$, $S_j(x,y)$ and $S_k(x,y)$.*

The area coordinates are not mutually independent as they are related by the identity:

$$L_i(x,y) + L_j(x,y) + L_k(x,y) = 1 \tag{5.42}$$

The set of functions $L_k(x,y), k = 1,\cdot,n$ constitutes the required basis for the linear piecewise interpolation of a function, whose values $f_1 = f(x_1,y_1)$, $f_2 = f(x_2,y_2)$, $\cdots$, $f_n = f(x_n,y_n)$ are known at a given set of n nodes $\boldsymbol{p}_1 = (x_1;y_1)$, $\boldsymbol{p}_2 = (x_2;y_2)$, $\cdots$, $\boldsymbol{p}_n = (x_n;y_n)$

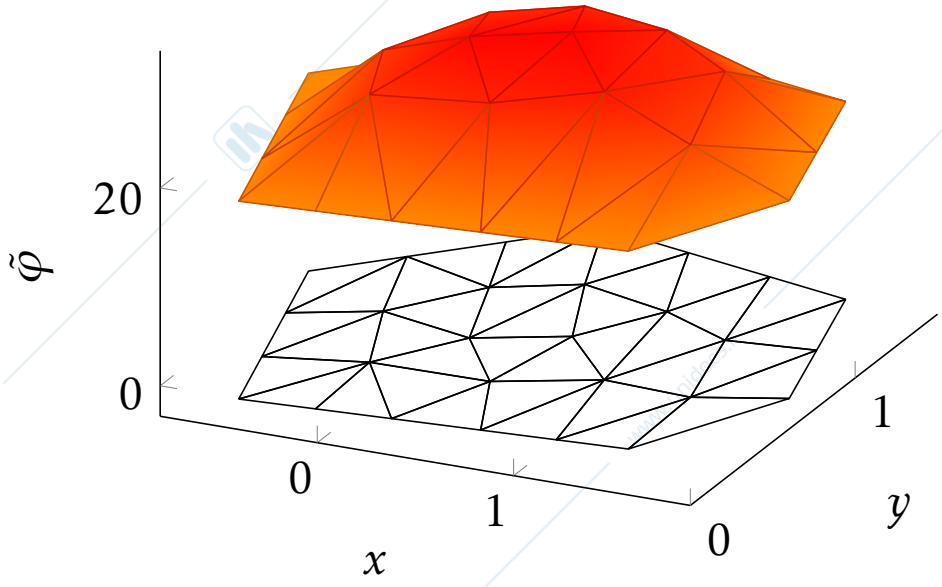


Figure 5.10: *two dimensional piecewise interpolating function $\tilde{\varphi}(x,y)$.*

Chapter 6

Finite difference method

The numerical solution of a differential equation can be performed by means of a finite difference method (FDM). In FDMs the derivatives of the unknown function are expressed by means of the finite difference formulas (5.16), (5.18), (5.19) and (5.20) as algebraic expressions involving the values of the unknown function on a a pre-defined grid of points. To give an example, we apply the method to the solution of a one-dimensional Poisson problem.

6.1 1D Poisson's equation

The scalar Poisson equation (1.17) in a one-dimensional formulation (i.e. $\frac{\partial}{\partial y} = 0$ and $\frac{\partial}{\partial z} = 0$) reduces to the ordinary differential equation:

$$\frac{\mathrm{d}^2 \varphi}{\mathrm{d}\, x^2} = t(x)$$

which can be solved when appropriate boundary conditions are specified. Let us consider the following one-dimensional differential problem, defined on a domain $x \in [a, b]$:

$$\begin{cases} \frac{\mathrm{d}^2 \varphi}{\mathrm{d}\, x^2} = t(x), \\ \varphi(a) = \varphi_a, \\ \left.\frac{\mathrm{d}\varphi}{\mathrm{d}\, x}\right|_{x=b} = \varphi'_b. \end{cases} \tag{6.1}$$

Defining a set of n evenly spaced points $x_1 = a$, x_2, …, x_k, …, $x_n = b$, with $\Delta = \frac{b-a}{n-1}$ the constant spacing between two adjacent points, we see that the central finite difference formula (5.18) can be used to express the left hand side the differential equation in (6.1) in a discrete form. For any internal point we can write:

$$\left.\frac{\mathrm{d}^2 \varphi}{\mathrm{d}\, x^2}\right|_{x=x_k} = t(x_k) \implies \frac{\varphi_{k-1} - 2\,\varphi_k + \varphi_{k+1}}{\Delta^2} + O(\Delta^2) = t_k,$$

that is, allowing a $O(\Delta^2)$ error:

$$\varphi_{k-1} - 2\,\varphi_k + \varphi_{k+1} = \Delta^2\, t_k, \quad k = 2, 3, \ldots, n-1. \tag{6.2}$$

We have now to treat the boundary conditions at the point $x_1 = a$ and $x_n = b$. The first one is a Dirichlet boundary condition assigning the value of the function at the point x_1. We can then write:

$$\varphi_1 = \varphi_a. \tag{6.3}$$

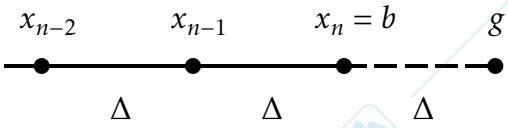


Figure 6.1: *Fictitious point g is inserted to preserve the $O(\Delta^2)$ order of accuracy*

To impose the condition on the derivative at the point x_n we can resort to the backward finite difference formula (5.20):

$$\left.\frac{d\varphi}{dx}\right|_{x=x_n} = \frac{\varphi_n - \varphi_{n-1}}{\Delta} + O(\Delta) = \varphi'_n$$

In doing so, we introduce an error ($O(\Delta)$), which results in a lower order of accuracy with respect to (6.2). To preserve the order of accuracy ($O(\Delta^2)$) of (6.2), we use the ghost point technique: as shown in Fig. 6.1, a fictitious point is inserted at the position $g = x_n + \Delta = b * \Delta$, allowing then to use the central difference formulas. From (5.16) and (5.18), we can write two conditions at the point x_n:

$$\left.\frac{\mathrm{d}^2\varphi}{\mathrm{d}x^2}\right|_{x=x_n} = t(x_n) \implies \frac{\varphi_{n-1} - 2\varphi_n + \varphi_g}{\Delta^2} + O(\Delta^2) = t_n,$$

$$\left.\frac{d\varphi}{dx}\right|_{x=x_n} = \frac{\varphi_g - \varphi_{n-1}}{2\Delta} + O(\Delta^2) = \varphi'_b$$

and, by eliminating φ_g:

$$2\varphi_{n-1} - 2\varphi_n = \Delta^2 t_n - 2\Delta\varphi'_b. \tag{6.4}$$

Equations (6.2), (6.3) and (6.4) can be put together to form a system of n linear algebraic equations in n unknowns, that are the values of the function φ at the points $x_1, x_2, \ldots, x_n$:

$$[K]\{\varphi\} = \{s\} \tag{6.5}$$

where

$$K = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & \dots & \dots & 0 \\ 1 & -2 & 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & \dots & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & 1 & -2 & 1 & & \vdots \\ \vdots & & & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -1 \end{bmatrix}, \quad \{\varphi\} = \begin{Bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \vdots \\ \vdots \\ \vdots \\ \varphi_{n-1} \\ \varphi_n \end{Bmatrix}, \quad \{s\} = \begin{Bmatrix} \varphi_a \\ \Delta^2 t_2 \\ \Delta^2 t_3 \\ \vdots \\ \vdots \\ \vdots \\ \Delta^2 t_{n-1} \\ \Delta^2 t_n - \Delta\varphi'_b \end{Bmatrix}$$

We observe that the nonzero coefficient of the matrix $[K]$ are located only on the main diagonal, and on the two diagonals immediately above and below the main diagonal. A matrix that has this property is called tridiagonal. Tridiagonal matrices are a particular kind of sparse matrices. A matrix is said to be sparse when the most of its elements are zero.

Diagonally dominant matrices

A square matrix $[A]$ with rank n is called strictly (row) diagonally dominant when:

$$|a_{i,i}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|, \quad \forall i = 1, \dots, n; \tag{6.6}$$

is called weakly (row) diagonally dominant when:

$$|a_{i,i}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|, \quad \forall i = 1, \dots, n. \tag{6.7}$$

If a strict inequality holds in (6.7) at least for one row i, then the matrix is said to be irreducibly (row) diagonally dominant.

The Gershgorin's theorem and its corollaries allow us to state that strictly and irreducibly diagonally dominant matrices are non singular, and can then be inverted.

We can observe that the coefficient matrix $[K]$ of the linear system (6.5) is an irreducibly diagonally dominant matrix, and is therefore invertible. Then (6.5) admits a solution. It is also worth noting that the irreducibility character is conferred to $[K]$ by its first row, which derives by the Dirichlet condition at $x = a$.

6.2 2D Poisson equation

Let us consider the Poisson equation in a cartesian coordinates system, assuming that $\frac{\partial}{\partial z} = 0$. Under this assumption, the Poissons equation is:

$$\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = t(x, y).$$

Consider a rectangular domain, of dimensions L_x and L_y along the x and y directions, respectively. We know from the uniqueness theorems that some kind of conditions has to be specified on the entire boundary. Without loss of generality, we consider a mixed type boundary conditions, with Dirichlet B.C. (i.e. φ assigned) on the south and west edge of the domain, and Neumann conditions (i.e. normal derivative assigned) on the rest of the boundary, that is constituted by the north and east edges. Placing the origin of the coordinates system in the south west corner, the problem is formulated as follows:

$$\begin{cases} \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = t(x, y), \\ \varphi(0, y) = \varphi_W(y), \\ \varphi(x, 0) = \varphi_S(x), \\ \left.\frac{\partial \varphi}{\partial x}\right|_{x=L_x} = \varphi'_E(y), \\ \left.\frac{\partial \varphi}{\partial y}\right|_{x=L_y} = \varphi'_N(x), \end{cases} \tag{6.8}$$

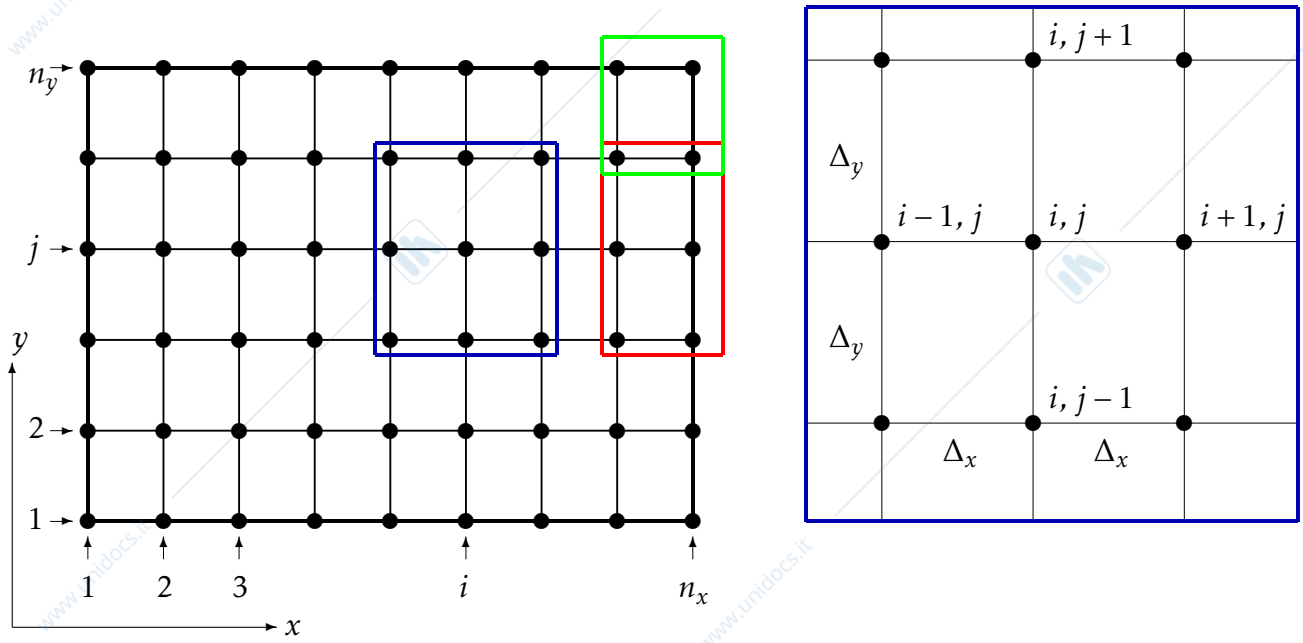


Figure 6.2: *two dimensional grid for FDM: 5 points stencil*

To solve problem (6.8), we define a grid of evenly spaced points along the x and y direction. Calling n_x and n_y the number of points along the two directions x and y, the constant spacing along x and y are $\Delta_x = \frac{L_x}{n_x - 1}$ and $\Delta_y = \frac{L_y}{n_y - 1}$. Any points is then identified by a pair of indexes i and j, which define its position in the grid. Let us consider an internal point (i,j). The central finite difference formula (5.18) can be used to express the partial derivatives in the Poisson's equation appearing in (6.8):

$$\left.\frac{\partial^2 \varphi}{\partial x^2}\right|_{x=x_i, y=y_j} = \frac{\varphi_{i-1,j} - 2\,\varphi_{i,j} + \varphi_{i+1,j}}{\Delta_x^2} + O(\Delta_x^2),$$

$$\left.\frac{\partial^2 \varphi}{\partial y^2}\right|_{x=x_i, y=y_j} = \frac{\varphi_{i,j-1} - 2\,\varphi_{i,j} + \varphi_{i,j+1}}{\Delta_y^2} + O(\Delta_y^2).$$

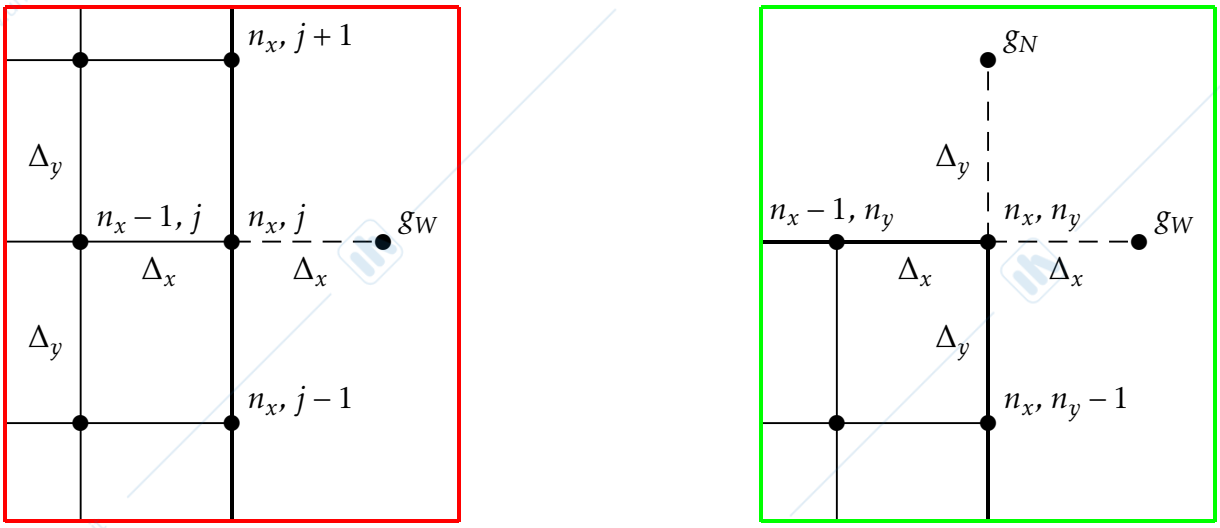
For any internal point the Poisson's equation can then be discretised with a second order accuracy $O(\Delta_x^2, \Delta_y^2)$ in the following algebraic equation:

$$\frac{1}{\Delta_y^2}\varphi_{i,j-1} + \frac{1}{\Delta_x^2}\varphi_{i-1,j} - 2\left(\frac{1}{\Delta_x^2} + \frac{1}{\Delta_y^2}\right)\varphi_{i,j} + \frac{1}{\Delta_x^2}\varphi_{i+1,j} + \frac{1}{\Delta_y^2}\varphi_{i,j+1} = t(x_i, y_j). \tag{6.9}$$

The discretised equation (6.9) involves the unknown values of the function φ at the points $(x_i,\ y_{j-1})$, $(x_{i-1},\ y_j)$, $(x_i,\ y_j)$, $(x_{i+1},\ y_j)$, and $(x_i,\ y_{j+1})$. The set of the points involved in a discretised equation is usually called stencil. It is worth noting that, when φ is a harmonic function (i.e. $t(x,y) = 0$) and the spacings in the x and y directions are equal (i.e. $\Delta_x = \Delta_y = \Delta$), (6.9) reduces to:

$$\varphi_{i,j} = \frac{1}{4}(\varphi_{i,j-1} + \varphi_{i-1,j} + \varphi_{i+1,j} + \varphi_{i,j+1}). \tag{6.10}$$

As the mean value theorem for harmonic functions suggests, the value of the value of φ in the central point of the stencil is equal to the arithmetic average of the function values at the neighbouring nodes.



a) b)

Figure 6.3: *4 points stencil for treating Neumann B.C: on east edge (a, and three points stencil for treating the NE corner (b*

The Dirichlet Boundary condition can be expressed as an assignment for the points on the south and west edges:

$$\begin{aligned} \varphi_{i,1} &= \varphi_S(x_j), \quad i = 1,2,\ldots,n_x \\ \varphi_{1,j} &= \varphi_W(y_j), \quad i = 1,2,\ldots,n_y \end{aligned} \tag{6.11}$$

For treating the Neumann boundary condition on the east and north edges of the rectangle, The fictitious point technique can be conveniently used. Considering a boundary point on the east edge, and referring to Fig. 6.3(a), one can write, with a second order accuracy $O(\Delta_x^2, \Delta_y^2)$:

$$\frac{1}{\Delta_y^2}\varphi_{n_x,j-1} + \frac{1}{\Delta_x^2}\varphi_{n_x-1,j} - 2\left(\frac{1}{\Delta_x^2}+\frac{1}{\Delta_y^2}\right)\varphi_{n_x,j} + \frac{1}{\Delta_x^2}\varphi_{g,E} + \frac{1}{\Delta_x^2}\varphi_{n_x,j+1} = t(x_{n_x}, y_j)$$

$$\frac{\varphi_{g,E} - \varphi_{n_x-1,j}}{2\,\Delta_x} = \varphi'_E(y_j), \quad j = 2,3,\ldots,n_y-1$$

from which, eliminating the fictitious variable $\varphi_{g,E}$, we get:

$$\frac{1}{\Delta_y^2}\varphi_{n_x,j-1} + \frac{2}{\Delta_x^2}\varphi_{n_x-1,j} - 2\left(\frac{1}{\Delta_x^2}+\frac{1}{\Delta_y^2}\right)\varphi_{n_x,j} + \frac{1}{\Delta_x^2}\varphi_{n_x,j+1} = t(x_{n_x}, y_j) - 2\,\frac{\varphi'_E(y_j)}{\Delta_x},$$

$$j = 2,3,\ldots,n_y-1 \tag{6.12}$$

The B.C. on the north edge can the treated in an analogous fashion, yielding:

$$\frac{2}{\Delta_y^2}\varphi_{i,n_y-1} + \frac{1}{\Delta_x^2}\varphi_{i,n_y} - 2\left(\frac{1}{\Delta_x^2}+\frac{1}{\Delta_y^2}\right)\varphi_{i,n_y} + \frac{1}{\Delta_x^2}\varphi_{i,n_y+1} = t(x_i, y_{n_y}) - 2\,\frac{\varphi'_N(x_i)}{\Delta_y},$$

$$i = 2,3,\ldots,n_x-1. \tag{6.13}$$

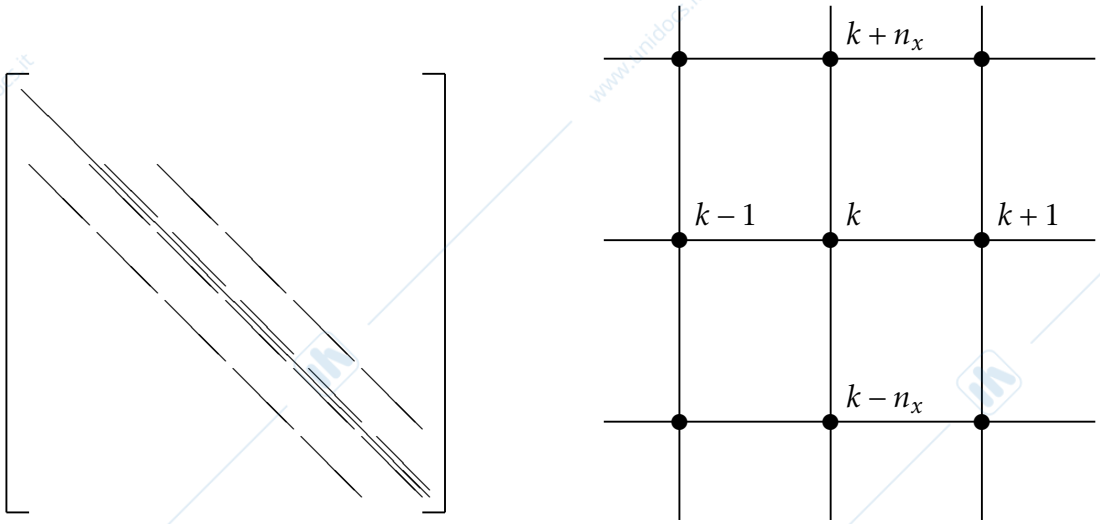


Figure 6.4: *Coefficient matrix structure (left) and numbering of the stencil for an internal point (right). The black lines in the matrix sketch show the position of the non-zero elements resulting from the node numbering*

At the NE corner, two conditions on the normal derivative have to be simultaneously satisfied. Using in this case two fictitious points as shown in Fig. 6.3(b), we can write:

$$\frac{1}{\Delta_y^2}\varphi_{n_x,n_y-1}+\frac{1}{\Delta_x^2}\varphi_{n_x-1,n_y}-2\left(\frac{1}{\Delta_x^2}+\frac{1}{\Delta_y^2}\right)\varphi_{n_x,n_y}+\frac{1}{\Delta_x^2}\varphi_{g,E}+\frac{1}{\Delta_x^2}\varphi_{g,N}=t(x_{n_x},y_{n_y})$$

$$\frac{\varphi_{g,N}-\varphi_{n_x,n_y-1}}{2\,\Delta_y}=\varphi'_N(x_{n_x}),$$

$$\frac{\varphi_{g,E}-\varphi_{n_x-1,n_y}}{2\,\Delta_x}=\varphi'_E(y_{n_y}),$$

whence:

$$\frac{2}{\Delta_y^2}\varphi_{n_x,n_y-1}+\frac{2}{\Delta_x^2}\varphi_{n_x-1,n_y}-2\left(\frac{1}{\Delta_x^2}+\frac{1}{\Delta_y^2}\right)\varphi_{n_x,n_y}=t(x_{n_x},y_{n_y})-2\,\frac{\varphi'_N(x_{n_x})}{\Delta_y}-2\,\frac{\varphi'_E(y_{n_y})}{\Delta_x}. \tag{6.14}$$

We can adopt a row-by-row numbering scheme, where each node (i,j) in the grid is assigned an index $k=(j-1)n_x+i$. Equations (6.8), (6.11), (6.12), (6.13) and (6.14) constitute a linear system:

$$[K]\{\varphi\}=\{s\},$$

with rank $n_p = n_x \times n_y$. In the chosen node numbering, the structure of $[K]$ can be sketched as shown in Fig. 6.4: the non-zero elements are found only in the main diagonal, in the two diagonals immediately above and below the main diagonal, and in the two diagonals placed n_x positions to the left and to the right of the main diagonal. The coefficients matrix $[K]$ is a sparse and irreducibly (due to the Dirichlet boundary condition on the west and south edges) diagonally dominant matrix. Then, the matxix $[K]$ is invertible and the linear system derived from the discretisation admits a solution $\{\varphi\}=[K]^{-1}\{s\}$.

Chapter 7

Finite element method

Finite elements methods (FEM) are a powerful class of numerical techniques widely utilised in any fields of science and engineering for evaluating the approximate solution of partial differential equations $D[f(x))] = 0$.

In a finite element method, the problem domain is divided in a certain number of subdomains, called elements, on which the unknown function is approximated with a piecewise polynomial interpolation $\tilde{f}(x)$ of some kind. This approach results in a great flexibility in treating problems defined on complex domains. The discretised formulation of the problem is derived enforcing some conditions on the residual $r(x) = D[\tilde{f}(x)]$, which represent a measure of the extent to which $\tilde{f}(x)$ fails to satisfy the given differential equation.

7.1 1D Poisson problem

Let us consider a one dimensional domain $x \in [a, b]$, where the following equation:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[p(x)\frac{\mathrm{d}\varphi}{\mathrm{d}x}\right] = t(x)$$

is defined. The above equation reduces to the one-dimensional formulation of the scalar Poisson equation (1.17) when the real parameter $p(x)$ is uniform on the domain $[a, b]$, and can be solved when appropriate boundary conditions are given:

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}x}\left[p(x)\frac{\mathrm{d}\varphi}{\mathrm{d}x}\right] - t(x) = 0, \\ \varphi(a) = \varphi_a, \\ \left.\frac{\mathrm{d}\varphi}{\mathrm{d}x}\right|_{x=b} = \varphi'_b. \end{cases} \tag{7.1}$$

The first step for solving the problem (7.1) using the finite element method is to introduce in the domain $[a, b]$ n points (or nodes) $x_1 = a$, x_2, …, $x_n = b$, dividing the the domain $[a, b]$ into $n-1$ sub-intervals (or elements) $[x_1, x_2]$, $[x_1, x_2]$, …, $[x_{n-1}, x_n]$. We then approximate the unknown function $\varphi(x)$ using a piecewise linear test function $\tilde{\varphi}(x)$:

$$\tilde{\varphi}(x) = \varphi_1 L_1(x) + \varphi_2 L_2(x) + \ldots + \varphi_n L_n(x) = \sum_{k=1}^{n} \varphi_k L_k(x) \tag{7.2}$$

The functions $L_1(x)$, $L_2(x)$, …, $L_n(x)$, which constitute the basis of the interpolating function, are the unit piecewise linear functions defined in (5.28) and (5.29).

We then define the residual $r(x)$ as:

$$r(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[p(x)\frac{\mathrm{d}\tilde{\varphi}}{\mathrm{d}x}\right] - t(x) \neq 0 \tag{7.3}$$

Assuming that $\tilde{\varphi}(x) \in C^2$, we could require that the weighted residual is zero on the domain $[a,b]$:

$$\int_a^b w(x)\, r(x) \mathrm{d}x = 0$$

that is:

$$\int_a^b w(x) \frac{\mathrm{d}}{\mathrm{d}x}\left[p(x)\frac{\mathrm{d}\tilde{\varphi}}{\mathrm{d}x}\right]\mathrm{d}x - \int_a^b w(x)\, t(x) \mathrm{d}x = 0 \tag{7.4}$$

where $w(x)$ is a weighting function defined on $[a,b]$. The choice of the weighting function $w(x)$ is not unique. However, as the function representation (7.2) reduces the problem in finding the n values φ_1, φ_2, …, φ_n at the nodes x_1, x_2, …, x_n, we should be able to derive from (7.5) a set of n equations constituting a system that can be solved. Some commonly adopted choice are the Dirac delta function $\delta(x - x_k)$, centred on a nodal point x_k, or a constant unitary step function equal to one on the interval $[(x_{k-1} + x_k)/2, (x_k + x_{k+1})/2]$, or the shape function $L_k(x)$ used for the piecewise representation of the unknown.

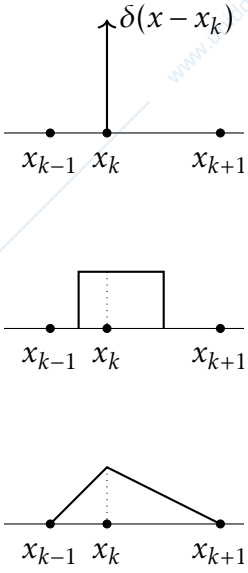


Figure 7.1: *weighing functions at the point x_k: the Dirac delta (up), the unitary step (centre) and the unit piecewise linear function (down)*

Being a piecewise linear interpolation, the function $\tilde{\varphi}(x) \in C^0$, and its second derivative is not defined, along with the residual (7.3) and the weighted residual appering in (7.4). In order to overcome this limitation, condition (7.4) can be relaxed integrating by part the first integral on the left hand side:

$$\int_a^b p(x) \frac{\mathrm{d}w}{\mathrm{d}x} \frac{\mathrm{d}\tilde{\varphi}}{\mathrm{d}x} \mathrm{d}x = \left[w(x)\, p(x) \frac{\mathrm{d}\tilde{\varphi}}{\mathrm{d}x}\right]_a^b - \int_a^b w(x)\, t(x) \mathrm{d}x \tag{7.5}$$

Condition (7.5) is called weak formulation of the given problem, and can be applied to the function $\tilde{\varphi}(x) \in C^0$.

According to the Galerkin approach, we use the function $L_k(x)$ as a weighting function $w(x)$ in (7.5). Since $L_k(x)$ in not equal to zero only on its support domain Ω_k, we obtain:

$$\int_{\Omega_k} p(x) \frac{\mathrm{d}L_k}{\mathrm{d}x} \frac{\mathrm{d}\tilde{\varphi}}{\mathrm{d}x} \mathrm{d}x = \left[L_k(x)\, p(x) \frac{\mathrm{d}\tilde{\varphi}}{\mathrm{d}x}\right]_a^b - \int_{\Omega_k} L_k(x)\, t(x) \mathrm{d}x \tag{7.6}$$

For any internal points x_k, $k = 2,\dots,x_{n-1}$, the support domain is constituted by two elements, the left hand element $\omega_{k-1,k} \equiv [x_{k-1}, x_k]$ and the right hand one $\omega_{k,k+1} \equiv [x_k, x_{k+1}]$. Furthermore, the weighting function $L_k(x)$ is zero at the points $x_1 = a$ and $x_n = b$. As a result, (7.6) can be rewritten for any internal point ($k = 2,\dots,n-1$) as:

$$\int_{\omega_{k-1,k}} p(x)\frac{\mathrm{d}L_k}{\mathrm{d}x}\frac{\mathrm{d}\tilde{\varphi}}{\mathrm{d}x}\mathrm{d}x + \int_{\omega_{k,k+1}} p(x)\frac{\mathrm{d}L_k}{\mathrm{d}x}\frac{\mathrm{d}\tilde{\varphi}}{\mathrm{d}x}\mathrm{d}x = -\int_{\omega_{k-1,k}} L_k(x)\,t(x)\mathrm{d}x - \int_{\omega_{k,k+1}} L_k(x)\,t(x)\mathrm{d}x \tag{7.7}$$

On the support domain, the piecewise linear test function $\tilde{\varphi}(x)$ can be expressed in the form:

$$\tilde{\varphi}(x) = \begin{cases} \varphi_{k-1}\,L_{k-1}(x) + \varphi_k\,L_k(x), & x \in \omega_{k-1,k} \\ \varphi_k\,L_k(x) + \varphi_{k+1}\,L_{k+1}(x), & x \in \omega_{k,k+1} \end{cases}$$

that, introduced in (7.7), finally yields:

$$\begin{aligned} &\left[\int_{\omega_{k-1,k}} p(x)\frac{\mathrm{d}L_k}{\mathrm{d}x}\frac{\mathrm{d}L_{k-1}}{\mathrm{d}x}\mathrm{d}x\right]\varphi_{k-1} + \left[\int_{\omega_{k-1,k}} p(x)\frac{\mathrm{d}L_k}{\mathrm{d}x}\frac{\mathrm{d}L_k}{\mathrm{d}x}\mathrm{d}x\right]\varphi_k + \\ &+\left[\int_{\omega_{k,k+1}} p(x)\frac{\mathrm{d}L_k}{\mathrm{d}x}\frac{\mathrm{d}L_k}{\mathrm{d}x}\mathrm{d}x\right]\varphi_k + \left[\int_{\omega_{k,k+1}} p(x)\frac{\mathrm{d}L_k}{\mathrm{d}x}\frac{\mathrm{d}L_{k+1}}{\mathrm{d}x}\mathrm{d}x\right]\varphi_{k+1} = \\ &= -\int_{\omega_{k-1,k}} L_k(x)\,t(x)\mathrm{d}x - \int_{\omega_{k,k+1}} L_k(x)\,t(x)\mathrm{d}x \end{aligned} \tag{7.8}$$

Equation (7.8) represents an algebraic equation in the unknown φ_{k-1}, φ_k, φ_{k+1} of the test function at the nodal points x_{k-1}, x_k, x_{k+1}. The values of the coefficient in (7.8) can be found recalling the expression of the functions $L_{k-1}(x)$, $L_k(x)$ and $L_{k+1}(x)$ on ω_+, whence the derivatives can be calculated:

$$L_{k-1}(x) = \begin{cases} 1 - \frac{x-x_{k-1}}{\Delta_{k-1,k}}, & x \in \omega_{k-1,k} \\ 0, & x \in \omega_{k,k+1} \end{cases} \qquad \frac{\mathrm{d}L_{k-1}}{\mathrm{d}x} = \begin{cases} -\frac{1}{\Delta_{k-1,k}}, & x \in \omega_{k-1,k} \\ 0, & x \in \omega_{k,k+1} \end{cases}$$

$$L_k(x) = \begin{cases} 1 + \frac{x-x_k}{\Delta_{k-1,k}}, & x \in \omega_{k-1,k} \\ 1 - \frac{x-x_k}{\Delta_{k,k+1}}, & x \in \omega_{k,k+1} \end{cases} \qquad \frac{\mathrm{d}L_k}{\mathrm{d}x} = \begin{cases} \frac{1}{\Delta_{k-1,k}}, & x \in \omega_{k-1,k} \\ -\frac{1}{\Delta_{k,k+1}}, & x \in \omega_{k,k+1} \end{cases}$$

$$L_{k+1}(x) = \begin{cases} 0, & x \in \omega_{k-1,k} \\ 1 + \frac{x-x_{k+1}}{\Delta_{k,k+1}}, & x \in \omega_{k,k+1} \end{cases} \qquad \frac{\mathrm{d}L_{k+1}}{\mathrm{d}x} \begin{cases} 0, & x \in \omega_{k-1,k} \\ \frac{1}{\Delta_{k,k+1}}, & x \in \omega_{k,k+1} \end{cases}$$

where $\Delta_{k-1,k} = x_k - x_{k-1}$ and $\Delta_{k,k+1} = x_{k+1} - x_k$. The above result can be introduced in (7.8), yielding:

$$-\frac{p_{k-1,k}}{\Delta_{k-1,k}}\varphi_{k-1} + \left(\frac{p_{k-1,k}}{\Delta_{k-1,k}} + \frac{p_{k,k+1}}{\Delta_{k,k+1}}\right)\varphi_k - \frac{p_{k,k+1}}{\Delta_{k,k+1}}\varphi_{k+1} = s_{k,-} + s_{k,+}, \tag{7.9}$$

where:

$$p_{k-1,k}- = \frac{1}{\Delta_{k-1,k}}\int_{\omega_{k-1,k}} p(x)\,\mathrm{d}x, \quad p_{k,k+1} = \frac{1}{\Delta_{k,k+1}}\int_{\omega_{k,k+1}} p(x)\,\mathrm{d}x$$

and:

$$s_{k,-}- = -\int_{\omega_{k-1,k}} L_k(x)\,t(x)\,\mathrm{d}x, \quad s_{k,+} = -\int_{\omega_{k,k+1}} L_k(x)\,t(x)\,\mathrm{d}x$$

In order to enforce the Dirichlet B.C. at the point $x_1 = a$, we can just write:

$$\varphi_1 = \varphi_a \tag{7.10}$$

For treating the Neumann condition at the node $x_n = n$, we rewrite (7.6), noting that the support domain Ω_n for this node is constituted by the element $[x_{n-1}, x_n]$ only ($\Omega_n = [x_{n-1}, x_n] \equiv \omega_{n,-}$), and that $L_n(a) = 0$ and $L_n(b) = 1$:

$$\int_{\omega_{n-1,n}} p(x) \frac{\mathrm{d}L_n}{\mathrm{d}x} \frac{\mathrm{d}\tilde{\varphi}}{\mathrm{d}x} \mathrm{d}x = p(x_n) \frac{\mathrm{d}\tilde{\varphi}}{\mathrm{d}x}\bigg|_{x=x_n} - \int_{\omega_{n-1,n}} L_k(x)\, t(x) \mathrm{d}x \tag{7.11}$$

Since $\tilde{\varphi}(x) = \varphi_{n-1} L_{n-1}(x) + \varphi_n L_n(x)$ on $[x_{n-1}, x_n]$, we also have:

$$\begin{aligned} \left[\int_{\omega_{n-1,n}} p(x) \frac{\mathrm{d}L_n}{\mathrm{d}x} \frac{\mathrm{d}L_{n-1}}{\mathrm{d}x} \mathrm{d}x\right] \varphi_{n-1} + \left[\int_{\omega_{n-1,n}} p(x) \frac{\mathrm{d}L_n}{\mathrm{d}x} \frac{\mathrm{d}L_n}{\mathrm{d}x} \mathrm{d}x\right] \varphi_n = \\ = p(x_n) \frac{\mathrm{d}\tilde{\varphi}}{\mathrm{d}x}\bigg|_{x=x_n} - \int_{\omega_{n-1,n}} L_n(x)\, t(x) \mathrm{d}x \end{aligned} \tag{7.12}$$

and, introducing the expression of the derivatives of the functions L_{n-1} and L_n:

$$-\frac{p_{n-1,n}}{\Delta_{n-1,n}} \varphi_{n-1} + \frac{p_{n-1,n}}{\Delta_{n-1,n}} \varphi_n = p(x_n) \frac{\mathrm{d}\tilde{\varphi}}{\mathrm{d}x}\bigg|_{x=x_n} + s_{n,-}, \tag{7.13}$$

where $p_{k-1,k}$ and s_- have the same meaning as in (7.9).

Equations (7.9), (7.10), and (7.13) constitute an algebraic linear system:

$$[K]\{\varphi\} = \{s\}$$

where:

$$[K] = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & \dots & \dots & 0 \\ -\frac{p_{2,-}}{\Delta_{2,-}} & \left(\frac{p_{2,-}}{\Delta_{2,-}} + \frac{p_{2,+}}{\Delta_{2,+}}\right) & -\frac{p_{2,+}}{\Delta_{2,+}} & 0 & \dots & \dots & \dots & 0 \\ 0 & -\frac{p_{3,-}}{\Delta_{3,-}} & \left(\frac{p_{3,-}}{\Delta_{3,-}} + \frac{p_{3,+}}{\Delta_{3,+}}\right) & -\frac{p_{3,+}}{\Delta_{3,+}} & \dots & \dots & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & -\frac{p_{k-1,k}}{\Delta_{k-1,k}} & \left(\frac{p_{k-1,k}}{\Delta_{k-1,k}} + \frac{p_{k,k+1}}{\Delta_{k,k+1}}\right) & -\frac{p_{k,k+1}}{\Delta_{k,k+1}} & & \vdots \\ \vdots & & & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -\frac{p_{n-2,n-1}}{\Delta_{n-2,n-1}} & \left(\frac{p_{n-2,n-1}}{\Delta_{n-2,n-1}} + \frac{p_{n-1,n}}{\Delta_{n-1,n}}\right) & -\frac{p_{n-1,n}}{\Delta_{n-1,n}} \\ 0 & 0 & 0 & 0 & \dots & 0 & -\frac{p_{n-1,n}}{\Delta_{n-1,n}} & \frac{p_{n-1,n}}{\Delta_{n-1,n}} \end{bmatrix}$$

and:

$$\{\varphi\} = \begin{Bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \vdots \\ \vdots \\ \vdots \\ \varphi_{n-1} \\ \varphi_n \end{Bmatrix}, \quad \{s\} = \begin{Bmatrix} \varphi_a \\ s_{2,-} + s_{2,+} \\ s_{3,-} + s_{3,+} \\ \vdots \\ s_{k,-} + s_{k,+} \\ \vdots \\ s_{n-1,-} + s_{n-1,+} \\ s_{n,-} + p(x_n)\, \varphi'_b \end{Bmatrix}$$

As in finite difference method, the non-zero coefficient in the matrix $[K]$ is a symmetric tridiagonal matrix. The presence of the Dirichlet B.C. at $x = a$ makes $[K]$ an irreducibly (row) diagonally dominant matrix. This implies that $[K]$ can be inverted, and the linear system solved to yield the solution $\{\varphi\}$. It is worth noting that, for a uniform $p(x) = 1$, $t(x) = T$, and a uniform grid spacing, $\Delta_{k-1,k} = \Delta_{k,k+1} = \Delta$, $p_{k-1,k} = p_{k,k+1} = 1$, and $s_{k-1,k} = s_{k,k+1} = -T\,\Delta/2$. Equations (7.9) and (7.13) reduce then to:

$$\varphi_{k-1} - 2\,\varphi_k + \varphi_{k+1} = T\,\Delta^2 \varphi_{n-1} - \varphi_n = T\,\Delta^2 -$$

and

$$\varphi_{n-1} - \varphi_n = -T\,\Delta^2 + \varphi_b',$$

that is the same result that would be obtained by applying the finite difference method. In can be shown that the discretisation scheme shown has a second order accuracy (O^2):

The following two equations shows the FEM discretization (7.9) obtained for two contiguous points x_k and x_{k+1}.

$$\left[\int_{\omega_{k-1,k}} p(x)\frac{\mathrm{d}L_k}{\mathrm{d}x}\frac{\mathrm{d}L_{k-1}}{\mathrm{d}x}\mathrm{d}x\right]\varphi_{k-1} + \left[\int_{\omega_{k-1,k}} p(x)\frac{\mathrm{d}L_k}{\mathrm{d}x}\frac{\mathrm{d}L_k}{\mathrm{d}x}\mathrm{d}x\right]\varphi_k +$$
$$+\left[\int_{\omega_{k,k+1}} p(x)\frac{\mathrm{d}L_k}{\mathrm{d}x}\frac{\mathrm{d}L_k}{\mathrm{d}x}\mathrm{d}x\right]\varphi_k + \left[\int_{\omega_{k,k+1}} p(x)\frac{\mathrm{d}L_k}{\mathrm{d}x}\frac{\mathrm{d}L_{k+1}}{\mathrm{d}x}\mathrm{d}x\right]\varphi_{k+1} =$$
$$= -\int_{\omega_{k-1,k}} L_k(x)\,t(x)\mathrm{d}x - \int_{\omega_{k,k+1}} L_k(x)\,t(x)\mathrm{d}x$$

$$\left[\int_{\omega_{k,k+1}} p(x)\frac{\mathrm{d}L_{k+1}}{\mathrm{d}x}\frac{\mathrm{d}L_k}{\mathrm{d}x}\mathrm{d}x\right]\varphi_k + \left[\int_{\omega_{k,k+1}} p(x)\frac{\mathrm{d}L_{k+1}}{\mathrm{d}x}\frac{\mathrm{d}L_{k+1}}{\mathrm{d}x}\mathrm{d}x\right]\varphi_{k+1} +$$
$$+\left[\int_{\omega_{k+1,k+2}} p(x)\frac{\mathrm{d}L_{k+1}}{\mathrm{d}x}\frac{\mathrm{d}L_{k+1}}{\mathrm{d}x}\mathrm{d}x\right]\varphi_{k+1} + \left[\int_{\omega_{k+1,k+2}} p(x)\frac{\mathrm{d}L_{k+1}}{\mathrm{d}x}\frac{\mathrm{d}L_{k+2}}{\mathrm{d}x}\mathrm{d}x\right]\varphi_{k+2} =$$
$$= -\int_{\omega_{k,k+1}} L_{k+1}(x)\,t(x)\mathrm{d}x - \int_{\omega_{k+1,k+2}} L_{k+1}(x)\,t(x)\mathrm{d}x$$

The element $\omega_{k,k+1}$ is the intersection between the two support domains Ω_k and Ω_{k+1}, and contributes to the equations for the nodes x_k and x_{k+1}, and for these only. As can be observed, the highlighted contributes of the element $\omega_{k,k+1}$ to the ones coming from other elements in forming the coefficient matrix (yellow highlight) ant the right hand side (green highlight) of the linear systems. This suggests that, defining for each element $\omega_{k,k+1}$ an element matrix $[K]_{k,k+1}$ and an element RHS array $\{s\}_{k,k+1}$:

$$K_{k,k+1} = \begin{bmatrix} \int_{\omega_{k,k+1}} p(x)\frac{\mathrm{d}L_k}{\mathrm{d}x}\frac{\mathrm{d}L_k}{\mathrm{d}x}\mathrm{d}x & \int_{\omega_{k,k+1}} p(x)\frac{\mathrm{d}L_k}{\mathrm{d}x}\frac{\mathrm{d}L_{k+1}}{\mathrm{d}x}\mathrm{d}x \\ \int_{\omega_{k,k+1}} p(x)\frac{\mathrm{d}L_{k+1}}{\mathrm{d}x}\frac{\mathrm{d}L_k}{\mathrm{d}x}\mathrm{d}x & \int_{\omega_{k,k+1}} p(x)\frac{\mathrm{d}L_{k+1}}{\mathrm{d}x}\frac{\mathrm{d}L_{k+1}}{\mathrm{d}x}\mathrm{d}x \end{bmatrix},$$

$$\{s\}_{k,k+1} = -\begin{Bmatrix} \int_{\omega_{k,k+1}} L_k(x)\,t(x)\mathrm{d}x \\ \int_{\omega_{k,k+1}} L_{k+1}(x)\,t(x)\mathrm{d}x \end{Bmatrix},$$

the linear system can be constructed by assembling, for each element, the entries of $[K]_{k,k+1}$ in the appropriate position of the coefficient matrix, and the entries of $\{s\}_{k,k+1}$ in the appropriate position of system RHS.

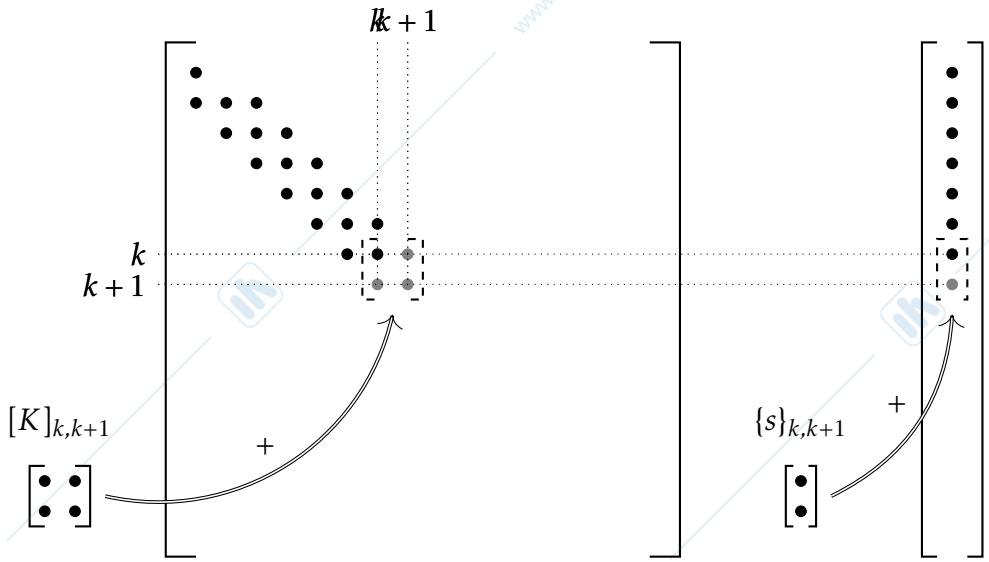


Figure 7.2: *assembly of the coefficient matrix and of the RHS of the algebraic system obtained by the FEM discretisation.*

The accuracy of the method described so far can be increased by increasing the degree of the shape functions constituting the basis of the piecewise polynomial interpolation. A second degree test function $\tilde{f}(x) \in C^0$ can be obtained adding a node $x_m = \frac{x_k + x_{k+1}}{2}$ at the midpoint of each element $\omega_{k,k+1}$, and defining for each element three second degree shape functions:

$$\begin{aligned} N_k &= L_k(2L_k - 1) \\ N_{k+1} &= L_{k+1}(2L_{k+1} - 1) \\ N_m &= 4L_k L_{k+1} \end{aligned}$$

The test function $\tilde{f}(x)$ is then:

$$\tilde{f}(x) = N_k(x)\varphi_k + N_m(x)\varphi_m + N_{k+1}(x)\varphi_{k+1}$$

on each element. The element matrix $[K]_{k,k+1}$ and an element RHS array $\{s\}_{k,k+1}$ become:

$$K_{k,k+1} = \begin{bmatrix} \int_{\omega_{k,k+1}} p(x) \frac{\mathrm{d}N_k}{\mathrm{d}x} \frac{\mathrm{d}N_k}{\mathrm{d}x} \mathrm{d}x & \int_{\omega_{k,k+1}} p(x) \frac{\mathrm{d}N_k}{\mathrm{d}x} \frac{\mathrm{d}N_m}{\mathrm{d}x} \mathrm{d}x & \int_{\omega_{k,k+1}} p(x) \frac{\mathrm{d}N_k}{\mathrm{d}x} \frac{\mathrm{d}N_{k+1}}{\mathrm{d}x} \mathrm{d}x \\ \int_{\omega_{k,k+1}} p(x) \frac{\mathrm{d}N_m}{\mathrm{d}x} \frac{\mathrm{d}N_k}{\mathrm{d}x} \mathrm{d}x & \int_{\omega_{k,k+1}} p(x) \frac{\mathrm{d}N_m}{\mathrm{d}x} \frac{\mathrm{d}N_m}{\mathrm{d}x} \mathrm{d}x & \int_{\omega_{k,k+1}} p(x) \frac{\mathrm{d}N_m}{\mathrm{d}x} \frac{\mathrm{d}N_{k+1}}{\mathrm{d}x} \mathrm{d}x \\ \int_{\omega_{k,k+1}} p(x) \frac{\mathrm{d}N_{k+1}}{\mathrm{d}x} \frac{\mathrm{d}N_k}{\mathrm{d}x} \mathrm{d}x & \int_{\omega_{k,k+1}} p(x) \frac{\mathrm{d}N_{k+1}}{\mathrm{d}x} \frac{\mathrm{d}N_m}{\mathrm{d}x} \mathrm{d}x & \int_{\omega_{k,k+1}} p(x) \frac{\mathrm{d}N_{k+1}}{\mathrm{d}x} \frac{\mathrm{d}N_{k+1}}{\mathrm{d}x} \mathrm{d}x \end{bmatrix},$$

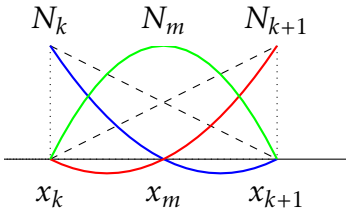


Figure 7.3: *second degree shape functions $N_k(x)$, $N_m(x)$ and $N_{k+1}(x)$*

$$\{s\}_{k,k+1} = -\begin{Bmatrix} \int_{\omega_{k,k+1}} N_k(x)\, t(x) \mathrm{d}x \\ \int_{\omega_{k,k+1}} N_m(x)\, t(x) \mathrm{d}x \\ \int_{\omega_{k,k+1}} N_{k+1}(x)\, t(x) \mathrm{d}x \end{Bmatrix},$$

Also in this case, the algebraic system resulting from the discretisation can be obtained by assembling the element matrix $[K]_{k,k+1}$ and an element RHS array $\{s\}_{k,k+1}$ into the coefficient matrix $[K]$ and the RHS array $\{s\}$.

7.2 2D Poisson problem

In this section, we will discuss the application of the finite element method to a problem governed by the equation:

$$\begin{cases} \nabla \cdot [p(\boldsymbol{x}) \nabla \varphi] = s(\boldsymbol{x}) \\ \varphi(\boldsymbol{x}) = \varphi_0(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in S_D, \\ \frac{\partial \varphi}{\partial n} = \varphi_0'(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in S_N, \end{cases} \tag{7.14}$$

on a three dimensional domain V, bounded by a closed surface S. According to (7.14), we assume that Dirichlet boundary conditions are applied to a part of S (S_D), and Neumann boundary conditions are applied to the rest of S (S_N). One can easily recognise that, for a uniform $p(\boldsymbol{x})$ on V, equation (7.14) reduces to Poisson equation (1.17).

In order to apply the weighted residual approach described in the previous section, we define a test function $\tilde{\varphi}(\boldsymbol{x})$ as a piecewise polynomial interpolation using a given set of unit piecewise polynomial functions $N_1(\boldsymbol{x}), N_2(\boldsymbol{x}), \cdots, N_n(\boldsymbol{x})$, where n is the number of nodes in the domain discretisation:

$$\tilde{\varphi}(\boldsymbol{x}) = \varphi_1 N_1(\boldsymbol{x}) + \varphi_2 N_2(\boldsymbol{x}) + \cdots + \varphi_n N_n(\boldsymbol{x}) \tag{7.15}$$

Assuming a twice continuously derivable test function, the residual $r(x,y)$ can the be defined as a measure of how much the test function fails to satisfy the given governing equation:

$$r(\boldsymbol{x}) = \nabla \cdot [p(\boldsymbol{x}) \nabla \tilde{\varphi}] - s(\boldsymbol{x}) \tag{7.16}$$

We can then require that the weighted residual is zero on the domain V:

$$\int_V w(\boldsymbol{x})\, r(\boldsymbol{x})\, \mathrm{d}V = \int_V w(\boldsymbol{x}) \nabla \cdot [p(\boldsymbol{x}) \nabla \tilde{\varphi}] \mathrm{d}V - \int_V w(\boldsymbol{x})\, s(\boldsymbol{x})\, \mathrm{d}V = 0 \tag{7.17}$$

where $w(\boldsymbol{x})$ is a weighting function defined on V. In order to obtain a weak formulation of (7.17), we consider the vector function $\boldsymbol{F} = p(\boldsymbol{x}) \nabla \tilde{\varphi}$ and the scalar function $\psi = \nabla \cdot [w(\boldsymbol{x}) \boldsymbol{F}] = \nabla \cdot [w(\boldsymbol{x})\, p(\boldsymbol{x}) \nabla \tilde{\varphi}]$. Integrating ψ over V and applying the divergence theorem, we get:

$$\int_V \nabla \cdot [w(\boldsymbol{x})\, p(\boldsymbol{x}) \nabla \tilde{\varphi}]\, \mathrm{d}V = \oint_S w(\boldsymbol{x})\, p(\boldsymbol{x}) \nabla \tilde{\varphi}\, \mathrm{d}S.$$

On the other hand, knowing that $\nabla \cdot [w(\boldsymbol{x}) \boldsymbol{F}] = \nabla w \cdot \boldsymbol{F} + w(\boldsymbol{x}) \nabla \cdot \boldsymbol{F}$, we can also write:

$$\int_V \nabla \cdot [w(\boldsymbol{x})\, p(\boldsymbol{x}) \nabla \tilde{\varphi}]\, \mathrm{d}V = \int_V \nabla w \cdot [p(\boldsymbol{x}) \nabla \tilde{\varphi}]\, \mathrm{d}V + \int_V w(\boldsymbol{x}) \nabla \cdot [p(\boldsymbol{x}) \nabla \tilde{\varphi}]\, \mathrm{d}V$$

Comparing the two above equations, we can derive that:

$$\int_V w(\boldsymbol{x}) \nabla \cdot [p(\boldsymbol{x}) \nabla \tilde{\varphi}]\, \mathrm{d}V = -\int_V \nabla w \cdot [p(\boldsymbol{x}) \nabla \tilde{\varphi}]\, \mathrm{d}V + \oint_S w(\boldsymbol{x})\, p(\boldsymbol{x}) \nabla \tilde{\varphi} \cdot \hat{\boldsymbol{n}} \mathrm{d}S.$$

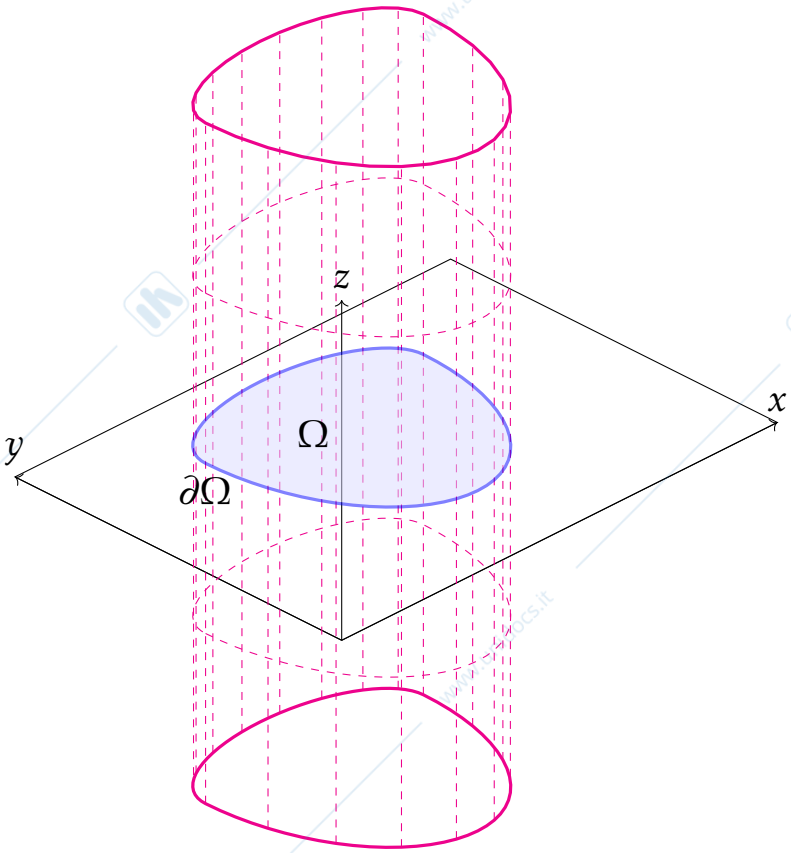


Figure 7.4: *Three dimensional representation of the domain in a two dimensional problem.*

that introduced in (7.17) yields:

$$\int_V \nabla w \cdot [p(\boldsymbol{x}) \nabla \tilde{\varphi}] \, \mathrm{d}V = -\int_V w(\boldsymbol{x}) s(\boldsymbol{x}) \, \mathrm{d}V + \oint_S w(\boldsymbol{x}) p(\boldsymbol{x}) \nabla \tilde{\varphi} \cdot \hat{\boldsymbol{n}} \mathrm{d}S. \tag{7.18}$$

Equation (7.18) is the weak formulation of the considered problem on a three dimensional domain, and can be applied to a piecewise polynomial test functions $\tilde{\varphi}(\boldsymbol{x}) \in C^2$.

Assuming now an orthogonal reference system xyz, (7.14) reduces to a two dimensional problem if one can assume that $\frac{\partial}{\partial z} = 0$. In this case, the three dimensional domain can be considered as the volume bounded by the cylindrical surface having a closed curve $\partial\Omega$ as a directrix, and generatrices along the z axis, as shown in fig. 7.4. The property $p = p(x, y)$ and the source term $s(x, y)$ are both functions of the independent variables x and y. The domain discretisation is then carried out on a two dimensional domain Ω lying on a xy plane and bounded by the closed curve $\partial\Omega$. Domain triangulation (i.e. subdivision of the domain in a set of triangular elements) is a commonly used approach in discretising a plane domain.

Considering a piecewise linear interpolation based on a triangulation of the domain, the shape functions are the unit piecewise linear functions discussed in section 5.5.2 and defined for each triangle by (5.41). In a triangular mesh with n nodes, the test function can then be expressed as:

$$\tilde{\varphi}(x, y) = \varphi_1 L_1(x, y) + \varphi_2 L_2(x, y) + \cdots + \varphi_n L_n(x, y) \tag{7.19}$$

The two dimensional weak formulation of the problem can be obtained applying (7.18) to a section of the cylindrical domain previously described:

$$\int_\Omega \nabla w \cdot [p(x,y)\nabla\tilde{\varphi}]\,\mathrm{d}S = -\int_\Omega w(x,y)\,s(x,y)\,\mathrm{d}S + \oint_{\partial\Omega} w(x,y)\,p(x,y)\nabla\tilde{\varphi}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}l. \tag{7.20}$$

Following the Galerkin approach, the discrete algebraic equation for the generic point k is obtained by introducing in (7.20) the shape function L_k as a weighting function (i.e., $w(x,y) = L_k(x,y)$):

$$\int_\Omega \nabla L_k \cdot [p(x,y)\nabla\tilde{\varphi}]\,\mathrm{d}S = -\int_\Omega L_k(x,y)\,s(x,y)\,\mathrm{d}S + \oint_{\partial\Omega} L_k(x,y)\,p(x,y)\nabla\tilde{\varphi}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}l. \tag{7.21}$$

As we will see, (7.21) yields an algebraic equation involving the unknown values of the test function $\tilde{\varphi}$ at the mesh nodes (more precisely, the value of the function $\tilde{\varphi}$ at the node k and at all the vertices in the support domain ω_k). We can then expect that we can construct an algebraic system by writing n equations for each node, using (7.21) or some special treatment for the boundary nodes.

The integral at the left hand side of (7.21) is formally executed on the entire plane domain Ω. However, since L_k, its gradient, and, as a consequence, the kernel of the integral, assume non zero values only in its support domain Ω_k, we can write:

$$\int_\Omega \nabla L_k \cdot [p(x,y)\nabla\tilde{\varphi}]\,\mathrm{d}S = \int_{\Omega_k} \nabla L_k \cdot [p(x,y)\nabla\tilde{\varphi}]\,\mathrm{d}S.$$

Furthermore, the same integral can be expressed as the sum of the partial integrals over each triangular element $\omega_{i_{el}}$ constituting the support domain Ω_k:

$$\int_{\Omega_k} \nabla L_k \cdot [p(x,y)\nabla\tilde{\varphi}]\,\mathrm{d}S = \sum_{\omega_{i_{el}}\in\Omega_k} \int_{\omega_{i_{el}}} \nabla L_k \cdot [p(x,y)\nabla\tilde{\varphi}]\,\mathrm{d}S.$$

We can now evaluate the integral on a generic triangular element $\omega_{i_{el}}$, having vertices at the point k, i and j. On this element, the test function behaviour is only governed by the shape functions $L_k(x,y)$, $L_i(x,y)$ and $L_j(x,Y)$:

$$\tilde{\varphi}(x,y) = \varphi_k\,L_k(x,y) + \varphi_i\,L_i(x,y) + \varphi_j\,L_j(x,y).$$

and then:

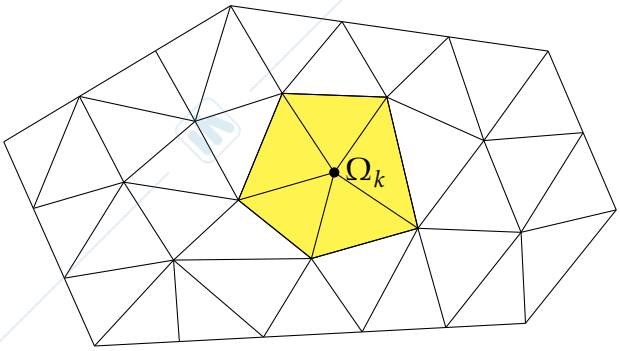


Figure 7.5: *triangulation of a plane domain.*

$$\int_{\omega_{i_{el}}} \nabla L_k \cdot [p(x,y)\nabla\tilde{\varphi}]]\,\mathrm{d}S =$$

$$= \left\{\int_{\omega_{i_{el}}} \nabla L_k \cdot [p\nabla L_k]\,\mathrm{d}S\right\}\varphi_k + \left\{\int_{\omega_{i_{el}}} \nabla L_k \cdot [p\nabla L_i]\,\mathrm{d}S\right\}\varphi_i + \left\{\int_{\omega_{i_{el}}} \nabla L_k \cdot [p\nabla L_j]\,\mathrm{d}S\right\}\varphi_j \quad (7.22)$$

Thus, the integral over a single triangular element in (7.22) results in three constants multiplying the value of the unknown function at the vertices of the element. The left hand side of (7.21) will be obtained by summing the integration results (7.22) carried over each element in the support domain Ω_k. We can collect the coefficients appearing at the right hand side in (7.22) in a matrix, called element matrix $[K]_{el}$:

$$[K]_{el} = \int_{\omega_{i_{el}}} \begin{bmatrix} \nabla L_i \cdot (p\nabla L_i) & \nabla L_i \cdot (p\nabla L_j) & \nabla L_i \cdot (p\nabla L_k) \\ \nabla L_j \cdot (p\nabla L_i) & \nabla L_j \cdot (p\nabla L_j) & \nabla L_j \cdot (p\nabla L_k) \\ \nabla L_k \cdot (p\nabla L_i) & \nabla L_k \cdot (p\nabla L_j) & \nabla L_k \cdot (p\nabla L_k) \end{bmatrix} \mathrm{d}S$$

The coefficients in (7.22) can be seen at the third row of the matrix $[K]_{el}$. The coefficients at the first and the second rows have been obtained using the shape functions L_i and L_k as a weighting functions, respectively. The coefficients in the first, second and third rows are to be multiplied by value of the unknown function at the vertices of the element φ_i , φ_j , and φ_k , respectively.

Introducing the matrix $[\nabla L]_{el}$:

$$[\nabla L]_{el} = \begin{bmatrix} \frac{\partial L_i}{\partial x} & \frac{\partial L_j}{\partial x} & \frac{\partial L_k}{\partial x} \\ \frac{\partial L_i}{\partial y} & \frac{\partial L_j}{\partial y} & \frac{\partial L_k}{\partial y} \end{bmatrix}, \quad (7.23)$$

the element matrix $[K]_{el}$ can be expressed in a more compact form:

$$[K]_{el} = \int_{\omega_{i_{el}}} [\nabla L]^T [\nabla L]\, p(x,y)\mathrm{d}S. \quad (7.24)$$

The first integral at the right hand side term of (7.21) may be treated in a similar fashion. Noting that the integral kernel assumes non zero values only on the support domain Ω_k, and that can then decomposed in the sum of the contributions given by the integration over each single triangular element $\omega_{i_{el}}$ forming Ω_k, we can write:

$$-\int_{\Omega} L_k(x,y)\, s(x,y)\,\mathrm{d}S = -\int_{\omega_{i_{el}}} L_k(x,y)\, s(x,y)\,\mathrm{d}S = -\sum_{\omega_{i_{el}} \in \Omega_k} \int_{\omega_{i_{el}}} L_k(x,y)\, s(x,y)\,\mathrm{d}S.$$

Since the element $\omega_{i_{el}}$ belongs also to the support domains of the nodes i and j, it will contribute also to right hand side of the equation (7.21) when using the shape functions L_i and L_j as weighting functions. We can store these contributions in the right hand side array $\{rhs\}_{el}$ of the element $\omega_{i_{el}}$:

$$\{rhs\}_{el} = -\int_{\omega_{i_{el}}} \begin{Bmatrix} L_i(x,y)\, s(x,y) \\ L_j(x,y)\, s(x,y) \\ L_k(x,y)\, s(x,y) \end{Bmatrix} \mathrm{d}S$$

Introducing the element shape functions array $\{L\}$:

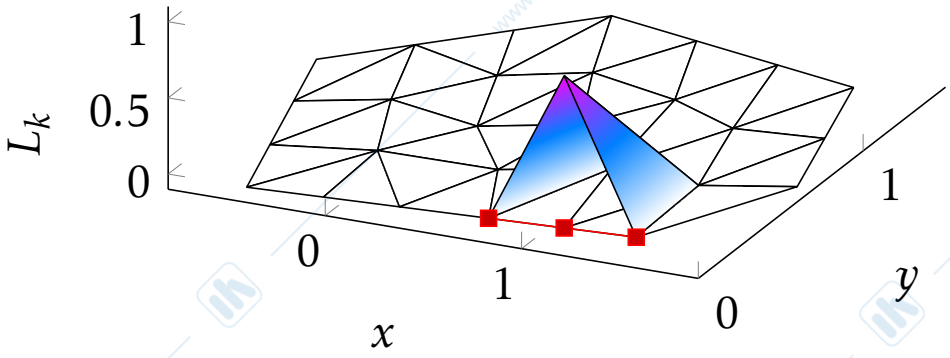


Figure 7.6: *linear shape function $L_k(x,y)$ for a boundary node.*

$$\{L\} = \begin{Bmatrix} L_i(x,y) \\ L_j(x,y) \\ L_k(x,y) \end{Bmatrix}$$

we can write the right hand side array $\{rhs\}_{el}$ in a compact form:

$$\{rhs\}_{el} = -\int_{\omega_{i_{el}}} \{L\}\, s(x,y)\, \mathrm{d}S. \tag{7.25}$$

Consider now the last term at the right hand side of (7.21). This integral is calculated over the closed curve $\partial\Omega$ which is the boundary of the calculation domain. One can easily recognise that, for any internal point k, the weighting function L_k is zero at the boundary. Thus, the kernel of the integral (and therefore the integral itself) is identically zero. The integral then produces a non zero result only if the point k is on the boundary of the calculation domain, where the assigned boundary conditions have to be applied.

If the point k belongs to a portion of boundary on which a Dirichlet condition is given, then the weighted residual approach procedure described so far is not applied. Instead, the value specified by the boundary condition is assigned to nodal value of the test function:

$$\varphi_k = \varphi_0(x_k, y_k) \tag{7.26}$$

If the point k belongs to a portion of boundary on which Neumann condition is given, we must refer to (7.21), taking into account that in this case the last integral at the right hand side gives some contribution. Since L_k is not zero only in the support domain Ω_k, the integral can be restricted to the two element edges e_1 and e_2 in Ω_k lying on the boundary $\partial\Omega$ (the two edges are marked in red in fig. 7.6). Knowing that, at a given point k placed on the Neumann boundary we have:

$$\nabla\tilde{\varphi}\cdot\hat{\boldsymbol{n}} = \frac{\partial\tilde{\varphi}}{\partial n} = \varphi_0'(x_k, y_k)$$

we can then write

$$\oint_{\partial\Omega} L_k(x,y)\, p(x,y)\, \nabla\tilde{\varphi}\, \hat{\boldsymbol{n}}\, \mathrm{d}l = \int_{e_1} L_k(x,y)\, p(x,y)\, \varphi_0'(x_k,y_k)\, \mathrm{d}l + \int_{e_2} L(x,y)\, p(x,y)\, \varphi_0'(x_k,y_k)\, \mathrm{d}l. \tag{7.27}$$

Thus, (7.27) allows to evaluate the correction to be applied to right hand side array of the solving algebraic system to take into account of the Neumann boundary conditions.

The solving algebraic system

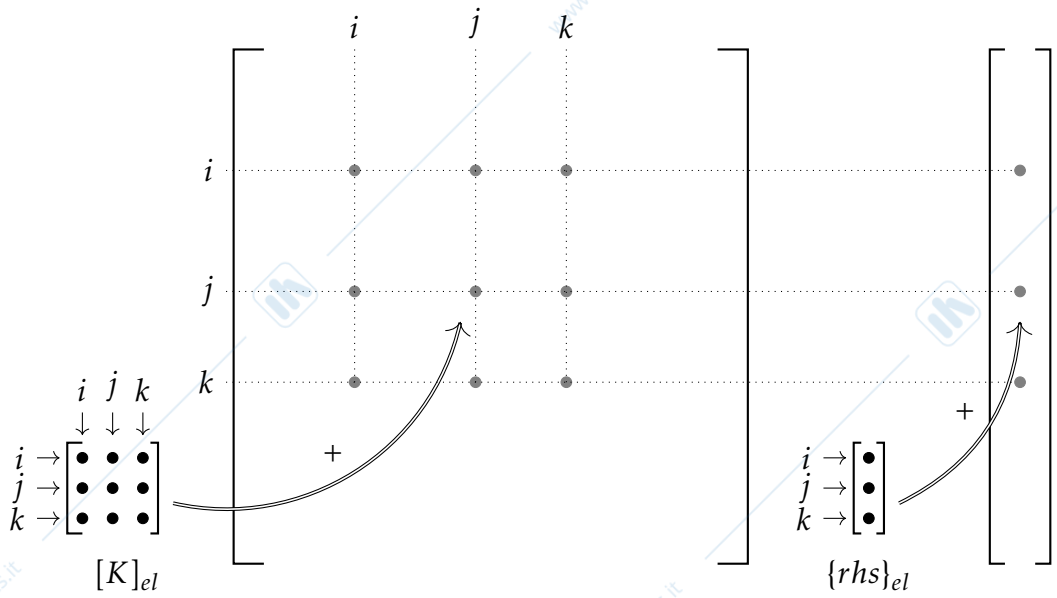


Figure 7.7: *assembly of the coefficient matrix and of the RHS of the algebraic system obtained by the FEM discretisation for a plane problem with piecewise linear shape functions.*

$$[K]\{\varphi\} = \{rhs\} \tag{7.28}$$

can now be constructed. The coefficient matrix $[K]$ is obtained by assembling the element matrices $[K]_{el}$ expressed in (7.24) evaluated for each elements. In order to describe the assembly process, we will assume that the first, the second and third vertex of a given element are the nodes i, j, and k respectively. Then. as shown in fig. 7.7, the coefficients in the first row of $[K]_{el}$ will be added to the coefficients in the i^{th} row of $[K]$; the coefficients in the second row of $[K]_{el}$ will be added to the coefficients in the j^{th} row of $[K]$; the coefficients in the third row of $[K]_{el}$ will be added to the coefficients in the k^{th} row of $[K]$. The coefficients in the first column of $[K]_{el}$ will be added to the coefficients in the i^{th} column of $[K]$; the coefficients in the second column of $[K]_{el}$ will be added to the coefficients in the j^{th} column of $[K]$; the coefficients in the third column of $[K]_{el}$ will be added to the coefficients in the k^{th} column of $[K]$.

A similar procedure is adopted to obtain the right hand side array, which is constructed by assembling the element right hand side arrays $\{rhs\}_{el}$ calculated as in (7.25). As sketched in fig. 7.7, the first element in $\{rhs\}_{el}$ is added to the system right hand side array $\{rhs\}$ at the i^{th} position; the second element of $\{rhs\}_{el}$ is added to $\{rhs\}$ at the j^{th} position; the third element of $\{rhs\}_{el}$ is added to $\{rhs\}$ at the k^{th} position. The system right hand side $\{rhs\}$ is then modified to take into account of the Neumann boundary conditions, according to (7.27).

The coefficient matrix $[K]$ is a sparse symmetric matrix. When Dirichlet boundary conditions are applied to some portion of the boundary, $[K]$ is also an irreducibly (row) diagonally dominant matrix. This implies that $[K]$ can be inverted, and that the linear system can be solved to yield the solution $\{\varphi\}$.

As we have seen for one dimensional FEM formulation, the accuracy of the method can be increased by increasing the degree of the shape functions. In order to define a second degree test function $\tilde{f}(x) \in C^0$, a node has to be added at the midpoint of all the edges of the triangular elements. On each element, the test function behaviour is then described by six

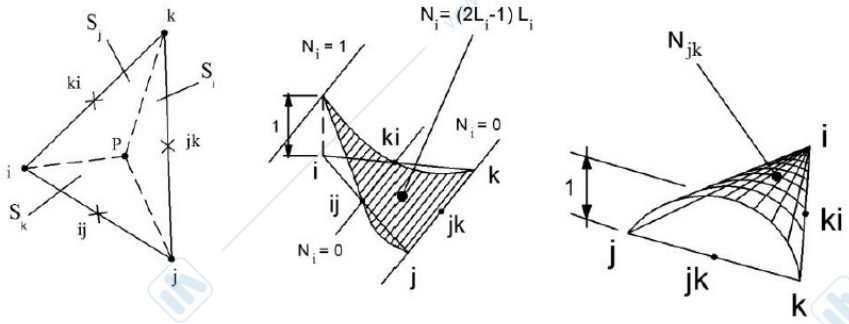


Figure 7.8: *Quadratic shape function $N_k(x,y)$ on a triangular element.*

second degree polynomial shape functions:

$$\begin{aligned} N_i &= L_i(2L_i-1) \\ N_j &= L_j(2L_j-1) \\ N_k &= L_k(2L_k-1) \\ N_{ij} &= 4L_iL_j \\ N_{jk} &= 4L_jL_k \\ N_{ki} &= 4L_kL_i \end{aligned}$$

The test function $\tilde{f}(x)$ is then:

$$\tilde{f}(x) = N_i(x,y)\varphi_i + N_j(x,y)\varphi_j + N_k(x,y)\varphi_k + N_{ij}(x,y)\varphi_{ij} + N_{jk}(x,y)\varphi_{jk} + N_{ki}(x,y)\varphi_{ki}$$

on each element.

Using a higher degree piecewise polynomial representation of the test function does not substantially change the Galerkin approach previously described. In this case, the discrete algebraic equation for the generic point k (or ki, since also the edge midpoints have to be taken into account) is obtained by introducing in (7.20) the shape function N_k (or N_{ki}) as a weighting function:

$$\int_\Omega \nabla N_k \cdot [p(x,y)\nabla\tilde{\varphi}]\,\mathrm{d}S = -\int_\Omega N_k(x,y)s(x,y)\,\mathrm{d}S + \oint_{\partial\Omega} N_k(x,y)p(x,y)\nabla\tilde{\varphi}\cdot\hat{\boldsymbol{n}}\,\mathrm{d}l. \qquad (7.29)$$

The algebraic system is obtained by assembling in the coefficient matrix $[K]$ the element matrices:

$$[K]_{el} = \int_{\omega_{i_{el}}} [\nabla N]^T [\nabla N]\, p(x,y)\mathrm{d}S, \qquad (7.30)$$

where:

$$[\nabla N]_{el} = \begin{bmatrix} \frac{\partial N_i}{\partial x} & \frac{\partial N_j}{\partial x} & \frac{\partial N_k}{\partial x} & \frac{\partial N_{ij}}{\partial x} & \frac{\partial N_{jk}}{\partial x} & \frac{\partial N_{ki}}{\partial x} \\ \frac{\partial L_i}{\partial y} & \frac{\partial N_j}{\partial y} & \frac{\partial N_k}{\partial y} & \frac{\partial N_{ij}}{\partial y} & \frac{\partial N_{jk}}{\partial y} & \frac{\partial N_{ki}}{\partial y} \end{bmatrix}, \qquad (7.31)$$

and by assembling the element right hand side array:

$$\{rhs\}_{el} = -\int_{\omega_{i_{el}}} \{N\}\, s(x,y)\,\mathrm{d}S. \qquad (7.32)$$

in the system right hand side array $\{rhs\}$, with:

$$\{N\} = \begin{Bmatrix} N_i(x,y) \\ N_j(x,y) \\ N_k(x,y) \\ N_{ij}(x,y) \\ N_{jk}(x,y) \\ N_{ki}(x,y) \end{Bmatrix}.$$

7.3 Solving steady state problems by FEM

The finite element method described in the previous sections can be utilised to solve a number of different physical problems. Steady state electromagnetic problems are ideally suited for being treated by means of FEM.

7.3.1 Electrostatics

Electrostatic problems are governed by (4.7), which has the form appearing in (7.14). A two dimensional problem can then be solved using the Galerkin formulation (7.29):

$$\int_\Omega \nabla N_k \cdot [\epsilon(x,y)\nabla\tilde{\varphi}]\,\mathrm{d}S = \int_\Omega N_k(x,y)\rho(x,y)\,\mathrm{d}S + \oint_{\partial\Omega} N_k(x,y)\epsilon(x,y)\nabla\tilde{\varphi}\cdot\hat{\boldsymbol{n}}\mathrm{d}l. \tag{7.33}$$

The algebraic system is obtained by assembling the element matrices:

$$[K]_{el} = \int_{\omega_{i_{el}}} [\nabla N]^T [\nabla N]\epsilon(x,y)\mathrm{d}S, \tag{7.34}$$

in the coefficient matrix $[K]$ and by assembling the element right hand side array:

$$\{rhs\}_{el} = \int_{\omega_{i_{el}}} \{N\}\rho(x,y)\,\mathrm{d}S. \tag{7.35}$$

in the system right hand side array $\{rhs\}$.

7.3.2 Magnetostatics

The governing equation of magnetostatics is given in (4.15) where the unknown function is the vector potential $\boldsymbol{A}$. Unlike the governing equation for electrostatics, (4.15) is a vector equation, which generally involves the solution of three scalar equations in the three directions of the space. As a result, the residual is a vector quantity defined as:

$$\boldsymbol{r}(\boldsymbol{x}) = \nabla\times\left[\frac{1}{\mu(\boldsymbol{x})}\nabla\times\tilde{\boldsymbol{A}}(\boldsymbol{x})\right] - \boldsymbol{J}(\boldsymbol{x}), \tag{7.36}$$

and the weighted residual approach yields:

$$\int_V \boldsymbol{w}(\boldsymbol{x})\cdot\boldsymbol{r}(\boldsymbol{x}) = \int_V \boldsymbol{w}(\boldsymbol{x})\cdot\nabla\times\left[\frac{1}{\mu(\boldsymbol{x})}\nabla\times\tilde{\boldsymbol{A}}(\boldsymbol{x})\right]\mathrm{d}V - \int_V \boldsymbol{w}(\boldsymbol{x})\cdot\boldsymbol{J}(\boldsymbol{x})\,\mathrm{d}V = 0. \tag{7.37}$$

It should be noted that the weighting function $\boldsymbol{w}(\boldsymbol{x})$ appearing in (7.37) is an arbitrarily oriented vector function. Thus, the condition described by (7.37) ensures that the weighted

residual projection along the direction of $\boldsymbol{w}(\boldsymbol{x})$ is equal to zero. In a truly three dimensional problem, (7.37) can be used to weight the residual along three independent directions, thus producing three scalar equations at any given point in the domain.

A two dimensional plane formulation is obtained when all quantities does not vary along the z axis, and when the current density $\boldsymbol{J}$ is directed along the z axis ($\boldsymbol{J} = J_z \hat{\boldsymbol{k}}$). Under these assumptions, the magnetic induction field lies on the xy plane (i.e. $\boldsymbol{B} = B_x \hat{\boldsymbol{i}} + B_y \hat{\boldsymbol{j}}$) and the only relevant component of the magnetic vector potential is along the z axis ($\boldsymbol{A} = A_z \hat{\boldsymbol{k}}$). The two dimensional formulation can then be derived noting that:

$$\nabla \times \left[\frac{1}{\mu} \nabla \times A_z \hat{\boldsymbol{k}} \right] = -\nabla \cdot \left[\frac{1}{\mu} \nabla A_z \right] \hat{\boldsymbol{k}} \tag{7.38}$$

which allows to rewrite (4.15) as:

$$\nabla \cdot \left[\frac{1}{\mu} \nabla A_z \right] = -J_z. \tag{7.39}$$

The two dimensional weighted residual approach is obtained by applying (7.37) along the only relevant direction, that is by using $\boldsymbol{w} = w_z(x,y)\hat{\boldsymbol{k}}$ as weighting function:

$$\int_V w_z(\boldsymbol{x}) \nabla \cdot \left[\frac{1}{\mu(\boldsymbol{x})} \nabla \tilde{A}_z(\boldsymbol{x}) \right] \mathrm{d}V + \int_V w_z(\boldsymbol{x}) J_z(\boldsymbol{x}) \, \mathrm{d}V = 0. \tag{7.40}$$

Equation (7.40) is analogous to the general formulation given in (7.18), and can be manipulated to obtain the weak Galerkin formulation on the two dimensional $\Omega domain$:

$$\int_\Omega \nabla N_k \cdot \left[\frac{1}{\mu(x,y)} \nabla \tilde{A}_z \right] \mathrm{d}S = \int_\Omega N_k(x,y) J_z(x,y) \, \mathrm{d}S + \oint_{\partial\Omega} N_k(x,y) \frac{1}{\mu(x,y)} \nabla \tilde{A}_z \cdot \hat{\boldsymbol{n}} \mathrm{d}l. \tag{7.41}$$

The element matrices:

$$[K]_{el} = \int_{\omega_{i_{el}}} [\nabla N]^T [\nabla N] \frac{1}{\mu(x,y)} \mathrm{d}S, \tag{7.42}$$

and the element right hand side array:

$$\{rhs\}_{el} = \int_{\omega_{i_{el}}} \{N\} J_z(x,y) \, \mathrm{d}S. \tag{7.43}$$

can be assembled to form the the coefficient matrix $[K]$ and right hand side array of the solving algebraic system.

7.3.3 Steady state electrodynamics

Also steady state electrodynamics can be treated by means of The FEM approach. Assuming, for sake of simplicity, that the impressed electric field is zero on the calculation domain, the governing equation (4.22) becomes:

$$\nabla \cdot [\sigma \nabla \varphi] = 0 \tag{7.44}$$

The Galerkin approach applied to the two dimensional formulation of (7.44) yields:

$$\int_\Omega \nabla N_k \cdot [\sigma \nabla \tilde{\varphi}] \, \mathrm{d}S = \oint_{\partial\Omega} N_k(x,y) \, \sigma \nabla \tilde{\varphi} \cdot \hat{\boldsymbol{n}} \mathrm{d}l. \tag{7.45}$$

The element matrices:

$$[K]_{el} = \int_{\omega_{i_{el}}} [\nabla N]^T [\nabla N] \sigma \mathrm{d}S, \tag{7.46}$$

can be assembled to form the the coefficient matrix $[K]$ of the solving algebraic system. In this case, the only non-zero entries in the system right hand side array are only due to the boundary conditions, being identically equal to zero all the element right hand side arrays.

Chapter 8

Quasi-Stationary Electromagnetics

8.1 From Maxwell equations to wave equations

The aim of this section is to provide a derivation of the wave equations for the scalar (φ) and vector ($\boldsymbol{A}$) electromagnetic potentials. Therefore, here are Maxwell's equations, written in the local form:

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0} \tag{8.1}$$

$$\nabla \cdot \boldsymbol{B} = 0 \tag{8.2}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \tag{8.3}$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}, \tag{8.4}$$

the constitutive relations:

$$\boldsymbol{D} = \epsilon \boldsymbol{E} \tag{8.5}$$

$$\boldsymbol{B} = \mu \boldsymbol{H}, \tag{8.6}$$

and Ohm's law (in local form):

$$\boldsymbol{J} = \sigma(\boldsymbol{E} + \boldsymbol{E}_i), \tag{8.7}$$

where $\boldsymbol{E}_i$ $[\frac{V}{m}]$ represents any externally imposed electric field.
Thanks to Eq. (8.2), any magnetic field can be expressed as the curl of a vector potential $\boldsymbol{A}$ with:

$$\boldsymbol{B} = \nabla \times \boldsymbol{A}. \tag{8.8}$$

$\boldsymbol{A}$ takes the name of *magnetic vector potential*, MVP from here onwards. The substitution of Eq. (8.8) in Eq. (8.3) yields:

$$\nabla \times \boldsymbol{E} = -\frac{\partial}{\partial t}(\nabla \times \boldsymbol{A}). \tag{8.9}$$

If the physical domain under exam is static, the time derivative can be brought inside the curl operation:

$$\nabla \times \boldsymbol{E} = -\nabla \times \frac{\partial \boldsymbol{A}}{\partial t} \implies \nabla \times \left(\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t} \right) = 0. \tag{8.10}$$

Now the quantity $\left(\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t}\right)$ is irrotational everywhere, and therefore can be expressed as the gradient of a scalar function φ:

$$\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t} = -\nabla \varphi \implies \boldsymbol{E} = -\nabla \varphi - \frac{\partial \boldsymbol{A}}{\partial t}. \tag{8.11}$$

The substitution of this new expression for $\boldsymbol{E}$ in Eq. (8.4), applying Eq. (8.5), (8.6) and (8.8), together with the assumption of a uniform magnetic permeability μ, yields:

$$\nabla \times \frac{\nabla \times \boldsymbol{A}}{\mu} = \boldsymbol{J} + \epsilon \frac{\partial}{\partial t}\left(-\nabla\varphi - \frac{\partial \boldsymbol{A}}{\partial t}\right) \implies \nabla \times (\nabla \times \boldsymbol{A}) = \mu \boldsymbol{J} - \mu\epsilon\left(\frac{\partial}{\partial t}\nabla\varphi + \frac{\partial^2 \boldsymbol{A}}{\partial t^2}\right), \tag{8.12}$$

from which, through vector analysis:

$$\nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A} = \mu \boldsymbol{J} - \mu\epsilon\left(\frac{\partial}{\partial t}\nabla\varphi + \frac{\partial^2 \boldsymbol{A}}{\partial t^2}\right) \tag{8.13}$$

As the choice of the divergence of the MVP, $\nabla \cdot \boldsymbol{A}$, is arbitrary, the so-called *Lorenz gauge* can be imposed:

$$\nabla(\nabla \cdot \boldsymbol{A}) = -\mu\epsilon\nabla\frac{\partial\varphi}{\partial t} \implies \nabla \cdot \boldsymbol{A} = -\mu\epsilon\frac{\partial\varphi}{\partial t} \tag{8.14}$$

The substitution of Eq. (8.14) in Eq. (8.13) yields the wave equation[1] for the vector potential:

$$-\nabla^2 \boldsymbol{A} = \mu \boldsymbol{J} - \mu\epsilon\frac{\partial^2 \boldsymbol{A}}{\partial t^2}. \tag{8.15}$$

For what concerns the scalar potential φ, by substituting Eq. (8.11) in Eq. (8.1) through Eq. (8.5) assuming **uniform permittivity** ϵ one gets:

$$\nabla \cdot \left(-\nabla\varphi - \frac{\partial \boldsymbol{A}}{\partial t}\right) = \frac{\rho}{\epsilon} \implies \nabla^2\varphi + \frac{\partial}{\partial t}(\nabla \cdot \boldsymbol{A}) = -\frac{\rho}{\epsilon}. \tag{8.16}$$

Choosing again to use the *Lorenz gauge*, the wave equation for the scalar potential is obtained:

$$-\nabla^2\varphi = \frac{\rho}{\epsilon} - \mu\epsilon\frac{\partial^2\varphi}{\partial t^2}. \tag{8.17}$$

Notably, the two equations are decoupled due to the Lorenz gauge imposition. They are non-homogeneous too, as both their right-hand sides (*rhs* from here onwards) are non-null. One of the fundamental properties of any wave is its *propagation speed*. The electromagnetic (EM) field makes indeed no exception, and it's propagation speed equals the speed of light c. Therefore, in the vacuum, the following relation holds true:

$$\mu_0\epsilon_0 = \frac{1}{c^2}. \tag{8.18}$$

Another interesting property of the EM field is that its components, i.e. the electric and magnetic fields, are always orthogonal to each other (see Fig. 8.1). Moreover, their propagation occurs as described by:

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \qquad \nabla \times \boldsymbol{H}\Big|_{(\boldsymbol{J}=0)} = \frac{\partial \boldsymbol{D}}{\partial t}, \tag{8.19}$$

These two equations outline that some sort of feedback relation takes place between the electric and magnetic components of the electromagnetic field. In particular, any variation of the magnetic field generates a variable electric field, whose variation onsets a displacement current $\boldsymbol{D}$, which in turn generates a magnetic field $\boldsymbol{B}$ and so on. Therefore, the propagation of electromagnetic waves is a self-sustaining mechanism.

[1] The term *wave equation* refers to a second-order linear **hyperbolic** partial differential equation, describing the temporal and spatial propagation of a function

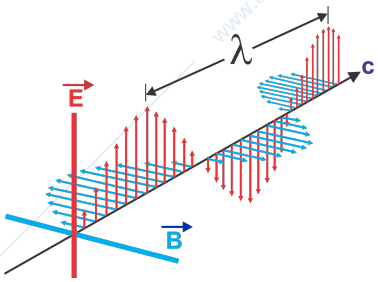


Figure 8.1: *electromagnetic field*

The solution of the wave equation for an infinite, linear and homogeneous media are the so-called *retarded potentials*, expressed by:

$$\boldsymbol{A}^*(\boldsymbol{p}_0, t) = \frac{\mu}{4\pi} \int_V \frac{\boldsymbol{J}\left(\boldsymbol{p}, t - \frac{r_{\boldsymbol{p},\boldsymbol{p}_0}}{v}\right)}{r_{\boldsymbol{p},\boldsymbol{p}_0}} d\tau_{\boldsymbol{p}} \tag{8.20}$$

$$\varphi^*(\boldsymbol{p}_0, t) = \frac{1}{4\pi\epsilon_0} \int_V \frac{\rho\left(\boldsymbol{p}, t - \frac{r_{\boldsymbol{p},\boldsymbol{p}_0}}{v}\right)}{r_{\boldsymbol{p},\boldsymbol{p}_0}} d\tau_{\boldsymbol{p}}, \tag{8.21}$$

where $v = \frac{c}{\sqrt{\epsilon_r \mu_r}}$ represents the speed of light in the considered media ($c \simeq 3 \cdot 10^8 m/s$). As shown in Fig. 8.1, $\boldsymbol{p}$ is the point where the field sources are located, whereas $\boldsymbol{p}_0$ is the point where the field is evaluated. The adjective *retarded* refers to the fact that, as outlined in Eq. (8.20) and (8.21), the fields are evaluated in point $\boldsymbol{p}_0$ at instant t using the sources $\boldsymbol{J}$ and ρ "as they were" at a previous time, $t - \frac{r_{\boldsymbol{p},\boldsymbol{p}_0}}{v}$. The time delay is caused by the fact that the electromagnetic interaction propagates with a finite velocity, v. As it will be shown in the next section, there are physical situations where the *retarded* nature of the potentials can be neglected. In particular, the condition to be met is the following:

$$\frac{r_{\boldsymbol{p},\boldsymbol{p}_0}|_{max}}{v} \ll \frac{1}{2\pi f}, \tag{8.22}$$

where $r_{\boldsymbol{p},\boldsymbol{p}_0}|_{max}$ represents the maximum linear extension of the considered domain. This means that the *finite* propagation speed of the EM interaction can be neglected, if (for the given frequency f) the domain is small enough.

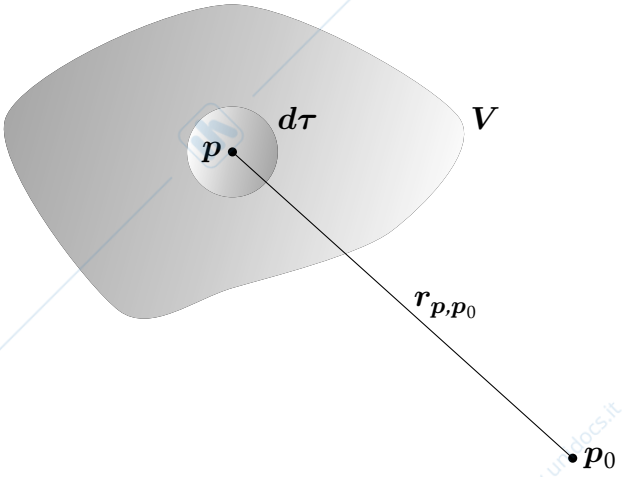


Figure 8.2: *retarded potentials example*

8.2 Quasi-stationary approximation

As already emphasized, the wave propagation originates from the interaction between the time-varying electric and magnetic components of the EM field. However, considering once again Eq. (8.19), if some requirements (described in the following lines) are met, one of the two following approximations can, in turn, be adopted:

- **electric** quasi-stationary problem $\Longrightarrow \frac{\partial B}{\partial t} = 0$
- **magnetic** quasi-stationary problem $\Longrightarrow \frac{\partial D}{\partial t} = 0$

The **electric quasi-stationary approximation** holds if the solenoidal electric field produced by the time variation of the magnetic field is negligible with respect to the electrostatic irrotational electric field. In this case, the electrostatic energy outweighs the magnetic energy and $\boldsymbol{E}$ can be considered irrotational, thus leading to:

$$\nabla \times \boldsymbol{E} = 0, \tag{8.23}$$

which is the same expression adopted for the stationary conditions. However, it is worth highlighting that both the continuity equation and the Ampère-Maxwell equations still hold their time-varying terms:

$$\nabla \cdot \boldsymbol{J} = -\frac{\partial \rho}{\partial t} \qquad \nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}. \tag{8.24}$$

In a dual way, the **magnetic quasi-stationary approximation** is applicable if the contribution given to the magnetic field generation by the time-varying charge density ρ (hence the displacement current) is negligible with respect to the one given by the conduction current $\boldsymbol{J}$, thus making the magnetic energy prevail over the electrostatic one. If this is the case, the *continuity equation* can be modified, neglecting the contribution given by the charge density variation:

$$\nabla \cdot \boldsymbol{J} = -\frac{\partial \rho}{\partial t} \Longrightarrow \nabla \cdot \boldsymbol{J} = 0, \tag{8.25}$$

and allowing the Ampère law to be used for the calculation of the magnetic fields.
In other words, adopting the electric quasi-static approximation means assuming that the electric field at the given temporal instant t is generated only by the electrostatic distribution of charges (neglecting the contributions of the time-varying magnetic field), whereas when the magnetic quasi-static approximation is employed the magnetic field at instant t is assumed to be generated only by the solenoidal current distribution (neglecting the effects of any time-varying electric field). For each of the two quasi-stationary cases, the self-sustaining mechanism that has been described in the last section no longer takes place.
Let's focus our attention to the **magnetic quasi-stationary approximation**, being the most common in practical cases. Neglecting the displacement current ($\frac{\partial D}{\partial t}$) contribution to the total current implies assuming an infinite propagation speed for the electromagnetic interaction instead of the "usual" speed of light c.
Maxwell's equation for the magnetic quasi-stationary electromagnetism take the following form:

$$\nabla \cdot \boldsymbol{E} = 0 \tag{8.26}$$

$$\nabla \cdot \boldsymbol{B} = 0 \tag{8.27}$$

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \tag{8.28}$$

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \cancel{\frac{\partial \boldsymbol{D}}{\partial t}} \Longrightarrow \nabla \times \boldsymbol{H} = \boldsymbol{J}. \tag{8.29}$$

In the last section expressions Eq. (8.15) and (8.16) for the propagation of the vector and scalar potential respectively have been derived; it is worth repeating the same procedure for the quasi-magnetostatic case.
Recalling the non-stationary expression for the electric field:

$$\boldsymbol{E} = -\nabla\varphi - \frac{\partial \boldsymbol{A}}{\partial t}, \tag{8.30}$$

and substituting it in the local form of Ohm's law Eq. (8.1) yields:

$$\boldsymbol{J} = \sigma\left(-\nabla\varphi - \frac{\partial \boldsymbol{A}}{\partial t}\right) + \sigma\boldsymbol{E}_i = \sigma\left(\boldsymbol{E}_i - \nabla\varphi\right) - \sigma\frac{\partial \boldsymbol{A}}{\partial t} = \boldsymbol{J}_0 - \sigma\frac{\partial \boldsymbol{A}}{\partial t}, \tag{8.31}$$

where $\boldsymbol{J}_0 = \sigma\left(\boldsymbol{E}_i - \nabla\varphi\right)$.
The new quantity $\boldsymbol{J}_0$ represents the currents that one would find *if* the problem was stationary, whereas the $-\sigma\frac{\partial \boldsymbol{A}}{\partial t}$ term accounts for the currents produced by the magnetic induction, through Faraday's law. Assuming now constant electrical properties for the considered physical media ($\sigma, \mu = const$), as well as the absence of externally imposed electric fields ($\boldsymbol{E}_i = 0$), Eq. (8.31) can be rewritten:

$$\nabla \times (\nabla \times \boldsymbol{A}) = -\mu\sigma\nabla\varphi - \mu\sigma\frac{\partial \boldsymbol{A}}{\partial t} \Longrightarrow \nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2\boldsymbol{A} = -\mu\sigma\nabla\varphi - \mu\sigma\frac{\partial \boldsymbol{A}}{\partial t}. \tag{8.32}$$

Once again, as the divergence of the magnetic vector potential is up to choose, and by setting:

$$\nabla \cdot \boldsymbol{A} = -\mu\sigma\varphi. \tag{8.33}$$

The *diffusion equation*[2] for the magnetic vector potential $\boldsymbol{A}$ is then finally obtained:

$$\nabla^2\boldsymbol{A} = \mu\sigma\frac{\partial \boldsymbol{A}}{\partial t}. \tag{8.34}$$

As it can be noticed, under the aforementioned assumptions the spatial and temporal evolution of the magnetic vector potential is no longer described by a wave equation, as it used to be in Eq. (8.15). Instead, when the magneto quasi-static approximation is adopted $\boldsymbol{A}$, *diffuses* through the considered material. Comparing the diffusion equation (8.34) to the wave equation (8.15), two things should be pointed out:

- the coefficient before the derivative, $\sigma\mu$, is no longer a speed. Indeed, its unit measure is $\left[\frac{s}{m^2}\right]$. Still, the product $\sigma\mu$ determines the speed of the diffusion mechanism (as one can see, the higher $\sigma\mu$, the lower the MVP diffusion);
- the second-order derivative which can be found in Eq. (8.15) has become a first-order derivative, and represents the speed of propagation of the phenomena.

A classic physical phenomena that applies to this mathematical description is heat diffusion, expressed through the so-called *heat equation*:

$$\nabla^2 T = \frac{\rho C_p}{K}\frac{\partial T}{\partial t}, \tag{8.35}$$

[2]Differently from wave equations, these are **parabolic** partial differential equations

where T $[K]$ is the temperature, C_p $[J/K]$ the specific heat at constant pressure, K $[W/(m \cdot K)]$ the thermal conductivity and ρ $[kg/m^3]$ the density.
Fig. 8.3 shows an example of evolution over time of the temperature of a physical system with uniform thermal properties, starting from τ_0. As one can see, as the time advances ($\tau_0 \rightarrow \tau_1 \rightarrow \tau_2$) the temperature distribution tends to an equilibrium, i.e. the gradient is subjected to a damping. Now this is substantially different from a wave propagation case, as $\forall t \neq \tau_0$ the distribution function $T(t)$ is $\neq 0$ $\forall x$. However, as it can be seen $T(\tau_0) \neq 0$ only in $x = 0$. This means that the physical interaction takes place with an *infinite* speed, i.e. each point of the geometry is "aware" of what happens in every other one.

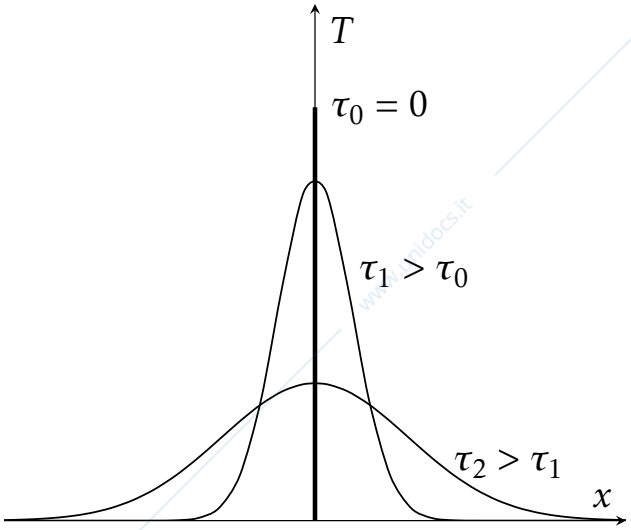


Figure 8.3: *heat gradient diffusion*

8.3 Skin Effect

The diffusion equation for the MVP allows drawing some important conclusions on how the current density $\boldsymbol{J}$ distributes itself over a conductor. Let's start once again from the diffusion equation:

$$\nabla^2 \boldsymbol{A} = \mu\sigma \frac{\partial \boldsymbol{A}}{\partial t}. \tag{8.36}$$

For the sake of simplicity, it will be assumed that the MVP and the current density are directed along the z axis, and that they both depend only on the y variable:

$$\boldsymbol{A} = A_z(y,t)\hat{\boldsymbol{k}} \qquad \boldsymbol{J} = J_z(y,t)\hat{\boldsymbol{k}} \tag{8.37}$$

This assumption grants that the magnetic field produced by the current $\boldsymbol{J}$ will have only the x component: $\boldsymbol{B} = B_x(y,t)\hat{\boldsymbol{i}}$. Thanks to the aforementioned assumptions, Eq.8.36 can be rewritten as:

$$\frac{\partial^2 A_z}{\partial y^2} - \mu\sigma \frac{\partial A_z}{\partial t} = 0 \tag{8.38}$$

If we now assume that the system under analyis operates under sinusoidal regime conditions, with frequency f, the Steinmetz transform can be applied to the MVP thus yielding:

$$\underline{A_z} = A_z e^{j\alpha} = A_z(y) e^{j\alpha(y)}. \tag{8.39}$$

The unknown $\boldsymbol{A}_z$ has been transformed into a complex phasor, eliminating its time-dependant nature. It's module is expressed by $A_z(y)$, and its phase angle by $\alpha(y)$. Notice that, as indicated, both the module and the phase of $\underline{A_z}$ depend on the spatial variable y. Having now applied the Steinmetz transform, the diffusion equation can be transformed too:

$$\frac{d^2\underline{A_z}}{dy^2} - j\omega\mu\sigma\underline{A_z} = 0, \tag{8.40}$$

where $\omega = 2\pi f$, $\mu = \mu_0 \cdot \mu_r$. As it can be seen, thanks to the elimination of time the equation depends now on the only y variable, thus allowing the use of *ordinary derivatives*. Eq. (8.40) is a second-order complex homogeneous **ordinary differential equation** (ODE), which allows a simple analytical solution that will be reported here. Let's write down the characteristic polynomial associated to the equation:

$$\underline{\lambda}^2 + j\omega\mu\sigma \Longrightarrow \lambda = \pm\sqrt{-j\omega\mu\sigma}. \tag{8.41}$$

The polynomial has two distinct solutions, that can be rewritten recalling that $j = e^{j\pi/2}$,

$$\underline{\lambda} = \pm\sqrt{\omega\mu\sigma\ e^{-j\pi/2}} = \pm\sqrt{\omega\mu\sigma}\ e^{-j\pi/4} = \pm\sqrt{\frac{\omega\mu\sigma}{2}}\ (1+j). \tag{8.42}$$

The quantity δ is called *Skin depth*, and it provides an indication of "how much" the MVP is able to penetrate the considered physical media:

$$\delta = \sqrt{\frac{2}{\omega\mu\sigma}} = \sqrt{\frac{1}{\pi f\mu\sigma}}. \tag{8.43}$$

Now, the solution of the differential equation, i.e. $\underline{A_z}$ will be give by the combination of two exponential functions:

$$\underline{A_z} = \underline{C_1}e^{+(1+j)y/\delta} + \underline{C_2}e^{-(1+j)y/\delta} \tag{8.44}$$

Right away, we can discard the positive exponential term, $\underline{C_1}e^{+(1+j)y/\delta}$, as it represents an un-physical solution (we do want a decreasing MVP for y approaching infinity). Therefore, having set $\underline{C_1} = 0$, our problem is restricted to finding the constant $\underline{C_2}$ of this exponential equation:

$$\underline{A_z} = \underline{C_2}e^{-(1+j)y/\delta}. \tag{8.45}$$

In order to accomplish this task, some kind of boundary condition must be employed. Let's say that the magnetic field's value it's known for $y = 0$, i.e. $B\big|_{y=0} = B_0$. In order to use this information, we must use the relation between the magnetic field and the MVP:

$$B = \nabla \times A, \tag{8.46}$$

which expressed in Cartesian coordinates yelds:

$$\boldsymbol{B} = det\begin{bmatrix} \hat{\boldsymbol{i}} & \hat{\boldsymbol{j}} & \hat{\boldsymbol{k}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ \cancel{A_x} & \cancel{A_y} & A_z \end{bmatrix} = \frac{\partial A_z}{\partial y}\hat{\boldsymbol{i}} = B_x\hat{\boldsymbol{i}}. \tag{8.47}$$

Eq. (8.47) allows writing a relation analogous to Eq. (8.45) for the magnetic field:

$$B_x = -\frac{1+j}{\delta}\underline{C_2}\,e^{-(1+j)y/\delta} = B_0\,e^{-y/\delta}e^{-jy/\delta}. \tag{8.48}$$

This equation's result must be B_0 for $y = 0$, hence yielding the value of C_2:

$$B_0 = -\frac{1+j}{\delta}\underline{C_2} \Longrightarrow \underline{C_2} = -\frac{\delta}{1+j}B_0 = -\frac{1-j}{2}\delta B_0 = \frac{\sqrt{2}}{2}e^{j\pi/4}\delta B_0. \tag{8.49}$$

Now we can substitute the expression of C_2 in Eq. (8.45), obtaining:

$$\underline{A_z} = \frac{\sqrt{2}}{2}\delta e^{j3/4\pi}\underline{B_0}e^{-(1+j)y/\delta} = \frac{\sqrt{2}}{2}\delta\underline{B_0}e^{-y/\delta}e^{-y/\delta-3/4\pi}. \tag{8.50}$$

If we want to derive the behaviour of the current density J_z, we can use the Ampère law (which is still valid, due to the quasi-stationary magnetic approximation):

$$\nabla \times H = J. \tag{8.51}$$

Expressing the rotor in Cartesian coordinates yields:

$$\boldsymbol{J} = det\begin{bmatrix} \hat{\boldsymbol{i}} & \hat{\boldsymbol{j}} & \hat{\boldsymbol{k}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ H_x & \cancel{H_y} & \cancel{H_z} \end{bmatrix} = \frac{\partial H_x}{\partial y}\hat{\boldsymbol{k}} = J_z\hat{\boldsymbol{k}}, \tag{8.52}$$

and along with $B = \mu H$ allows writing:

$$J_z = -\frac{1}{\mu}\frac{\partial B_x}{\partial y} = \frac{1}{\mu}\underline{B_0}\frac{1+j}{\delta}e^{-(1+j)y/\delta} = \frac{1}{\delta\mu}\underline{B_0}\sqrt{2}\; e^{j\pi/4}e^{-(1+j)y/\delta}. \tag{8.53}$$

Looking at the highlighted part of Eq. (8.53), we can define:

$$J_0 = \frac{\sqrt{2}}{\delta\mu}\underline{B_0} = \frac{\sqrt{2}}{\delta}\underline{H_0}. \tag{8.54}$$

Hence, Eq. (8.53) can be further developed:

$$J_z = J_0\; e^{j\pi/4}e^{-(1+j)y/\delta} = J_0 e^{-y/\delta}e^{-j(y/\delta-\pi/4)} = J_0 e^{-y/\delta}e^{-jy/\delta}e^{-j\pi/4} \tag{8.55}$$

At this point we can proceed anti transforming B_x and J_z, restoring the time-dependency. As one can see, the equations are travelling waves, dumped by the exponential term:

$$B_x(y,t) = B_0\; e^{-y/\delta}cos\left(\omega t - \frac{y}{\delta}\right) \tag{8.56}$$

$$J_z(y,t) = J_0\; e^{-y/\delta}cos\left(\omega t - \frac{y}{\delta} + \frac{\pi}{4}\right) \tag{8.57}$$

Interestingly, the two quantities have a phase offset of 45°. Fig. 8.4 shows an example of (8.57), computed for two different time instants τ_1 and τ_2, with $\tau_2 > \tau_1$. As one can see δ, the *skin depth*, provides an indication of how deep the current density can *penetrate* the given material. It is worth noticing that δ depends on the frequency f and on the electrical properties of the material σ and μ. Hence, the higher the frequency, the lower will be the area of the conductor where the current flows.
As an example, let's take the most common conductor, copper, and calculate the skin depth at $f = 50\ Hz$. Assuming a resistivity $\rho_{Cu} = 1.68 \cdot 10^-8$, thanks to $\rho = 1/\sigma$ the copper conductivity will be $\sigma_{Cu} = 5.95 \cdot 10^{-7}$. As the relative permeability can be assumed being μ_r, Eq. (8.43) yields:

$$\delta_{Cu} = \sqrt{\frac{1}{\pi f \mu_0 \sigma_{Cu}}} \approx 0.0095\ m. \tag{8.58}$$

Hence, no matter how thick the used (copper) conductor can be, the majority of the current density will be distributed over its first 9.5 *mm* of depth. More precisely, recalling Eq. (8.57) the amplitude of the current density J_z at a distance $y = \delta$ will be $J_0\; e^{-1} = 0.368$. This means that once the skin depth is reached, the current density is just the 36.8% of the value assumed at the material surface.

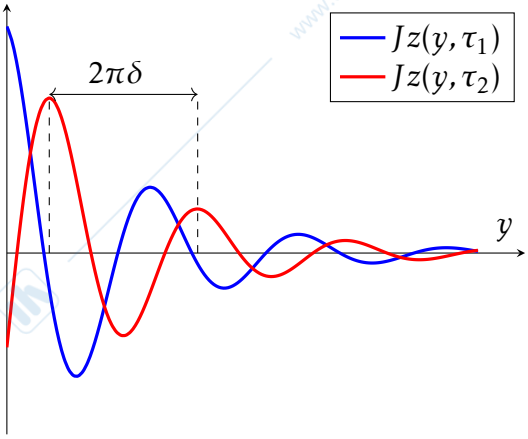


Figure 8.4: *diffusion of J_z for two different time-instants*

A numerical approach for diffusion equations

Explicit Euler method

Let's take the diffusion equation for $\boldsymbol{A}$, the magnetic vector potential (MVP), as derived in sec. 8.2:

$$\nabla^2 \boldsymbol{A} = \mu\sigma \frac{\partial \boldsymbol{A}}{\partial t}, \tag{8.59}$$

where $\boldsymbol{A} = \boldsymbol{A}(x,y,z,t)$. As it can be seen, the equation has both spatial and temporal dependencies, in the left and right-hand sides respectively. For the sake of simplicity, let's say that the MVP has only the z component, and that it depends only on the variable y. This allows writing:

$$\boldsymbol{A} = \boldsymbol{A}(y,t) = A_z \hat{\boldsymbol{k}}, \tag{8.60}$$

which implies that the current density has this very same property:

$$\boldsymbol{J} = \boldsymbol{J}(y,t) = J_z \hat{\boldsymbol{k}}. \tag{8.61}$$

Therefore, the magnetic vector $\boldsymbol{B}$ will have only the x component: $\boldsymbol{B}(y,t) = B_x \hat{\boldsymbol{i}}$. Fig. 8.5 represents the described situation.

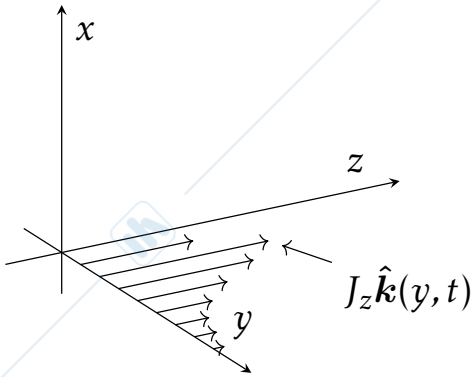


Figure 8.5: *current density distribution*

Thanks to the aforementioned assumptions, the problem is now mono-dimensional, and thus Eq. (8.34) can be re-written as:

$$\frac{\partial^2 A_z}{\partial y^2} = \mu\sigma \frac{\partial A_z}{\partial t} \tag{8.62}$$

In order to solve numerically the given equation, two different discretization processes have to be performed: a first one regarding the spatial variable y, introducing a certain number of points at a distance of Δy one from the other, and a second one regarding the time t, defining a time-interval Δt. Looking at Fig. 8.6, each "column" represents the evolution over time of the given point marked on the y axis. As an example, let's take the ith point, and express the derivatives through the finite difference formulas. The forward Euler method employs a central difference approximation for the spatial derivative term, and a forward difference approximation for the time-derivative; the solution A_z is assumed to be known at instant (k):

$$\left.\frac{\partial^2 A_z}{\partial y^2}\right|_{y=y_i}^{t=t_k} = \frac{A_{i+1}^{(k)} - 2A_i^{(k)} + A_{i-1}^{(k)}}{\Delta y^2} + O(\Delta y)^2 \qquad \left.\frac{\partial A_z}{\partial t}\right|_{y=y_i}^{t=t_k} = \frac{A_i^{(k+1)} - A_i^{(k)}}{\Delta t} + O(\Delta t), \tag{8.63}$$

As one can notice, the use of a central difference formula leads to a second-order truncation error; however, the use of a forward difference for the time-derivative discretization results in a first-order error. The performed discretization allows rewriting Eq. (8.62) as:

$$\frac{A_{i+1}^{(k)} - 2A_i^{(k)} + A_{i-1}^{(k)}}{\Delta y^2} = \mu\sigma \frac{A_i^{(k+1)} - A_i^{(k)}}{\Delta t}. \tag{8.64}$$

Thanks to the assumption made on $A_z^{(t)}$, the only unknown term of Eq. (8.64) is $A_i^{(k+1)}$, and the expression can be re-ordered as:

$$A_i^{(k+1)} = A_i^{(k)} + \Delta t \frac{1}{\mu\sigma} \frac{A_{i+1}^{(k)} - 2A_i^{(k)} + A_{i-1}^{(k)}}{\Delta y^2}. \tag{8.65}$$

As it can be observed, the *rhs* of this latter equation contains both the MVP of the central point i and the *diffusion* term, which contains the values of the MVP for three different points, all belonging to the same temporal instant (k). This kind of approach is called **forward Euler method**, and it yields an **explicit** equation for each point, as the MVP at point i at instant $(k+1)$ is expressed only through the values of MVP at the preceding instant (k). Thanks to the discretization process, a continuous differential equation has been transformed into an ensemble of discrete difference equations (the "price to be paid" for this is, of course, a certain degree of approximation). For the considered diffusion equation, the stability condition of the method is given by:

$$\Delta t < \frac{1}{2}\mu\sigma(\Delta y)^2. \tag{8.66}$$

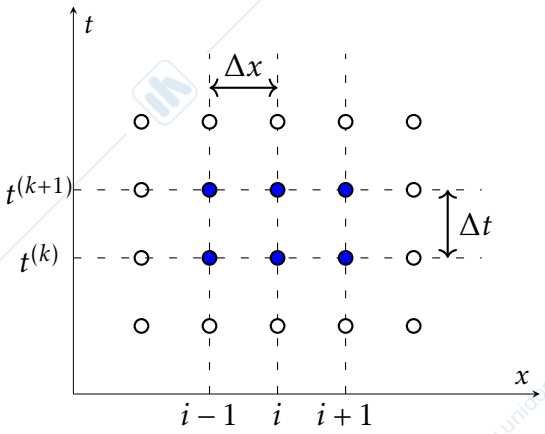


Figure 8.6: *discretization grid*

In the next section an example of the problems arising if the condition expressed in Eq. (8.66) is not met will be examined.

Implicit Euler method

Differently from the explicit method, the **implicit Euler method** uses a backward difference formula in order to estimate the time-derivative value at instant $(k+1)$:

$$\left.\frac{\partial^2 A_z}{\partial y^2}\right|_{y=y_i}^{t=t_{k+1}} = \frac{A_{i+1}^{(k+1)} - 2A_i^{(k+1)} + A_{i-1}^{(k+1)}}{\Delta y^2} + O(\Delta y)^2 \qquad \left.\frac{\partial A_z}{\partial t}\right|_{y=y_i}^{t=t_{k+1}} = \frac{A_i^{(k+1)} - A_i^{(k)}}{\Delta t} + O(\Delta t), \quad (8.67)$$

thus yielding:

$$\frac{A_{i+1}^{(k+1)} - 2A_i^{(k+1)} + A_{i-1}^{(k+1)}}{\Delta y^2} = \sigma\mu\frac{A_i^{(k+1)} - A_i^{(k)}}{\Delta t}. \quad (8.68)$$

This time Eq. (8.68) has only one known term, $A_i^{(k)}$, and four unknowns. Thus, it cannot be solved without adding some further information. In order to advance from the *(k)*th to the *(k+1)*th time instant, this **implicit** method requires writing (n) equations, where (n) stands for the number of points employed in the domain discretization. The question that might arise now is "why should we choose to employ an implicit method, when an explicit one would save us the trouble of writing and solving a linear system for each time-step?". The answer is that the explicit Euler method can be flawed by serious instability problems, due to the assumption that the time derivative (the diffusion term) of Eq. (8.64) remains constant through each time step of length Δt. As illustrated in Fig. 8.7, if Δt is too large compared to the time-scale of the physical phenomenon under analysis, one can get a solution $\widetilde{f_i}(t)$ which might dramatically differ from the analytical (and thus correct) one, marked as $f_i(t)$. In particular, what might be expected is the onsetting of oscillations, whose amplitude will increase over time.

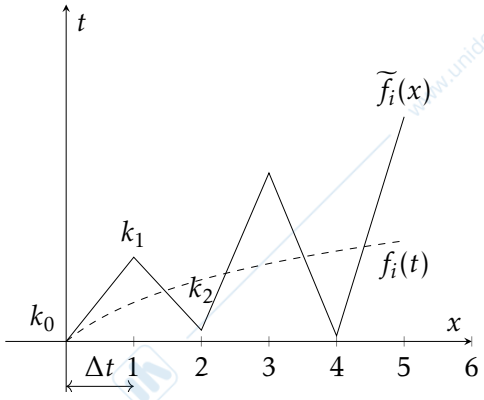


Figure 8.7: *explicit Euler method (instability issues)*

To explain this, let's look once again at Fig. 8.7: at instant (k_0) the numerical and analytical solutions $\widetilde{f_i}(t)$ and $f_i(t)$ do coincide. The explicit method implies that the value of $\mu\sigma\frac{\partial \boldsymbol{A}}{\partial t}$, computed in (k_0), remains constant throughout the time that separates (k_0) and (k_1), i.e. Δt. That equals to assuming that the gradient causing the flux of MVP through point (i) doesn't change over Δt. This assumption is of course false, as each flux will inevitably reduce the gradient by which is caused. Practically, once the approximated solution is obtained at instant (k_1), the overestimation of the (ingoing or outgoing) flux will determine in turn an

overestimated gradient in (k_1). Hence the sharp behaviour of $\widetilde{f}_i(x)$ from (k_1) to (k_2), which will once again determine an even higher gradient once reached (k_2), and so on, leading to the aforementioned oscillations of increasing amplitude. This problematic can be resolved in two main ways, either by (a) decreasing Δt so that the oscillations remain "bounded" or (b) switching to a more stable computational method. The implicit Euler method doesn't have any stability issue, and thus doesn't place any constraint in the choice of Δt. However, some attention should still be paid, as both the forward and backward Euler methods are *first order* accurate in time. Therefore, the two Euler methods will be affected by first-order truncation errors. Some kind of *remedy* to flaw issue exist though, as discussed in the next subsection.

Crank - Nicolson method

The so-called **Crank - Nicolson** aims at overcoming the main issue of the Euler methods, that is their temporal first-order nature. In this spirit, the given equation is discretized at an intermediate time instant, that is $(k+1/2)$ (see Fig. 8.8). Therefore, the problem can be

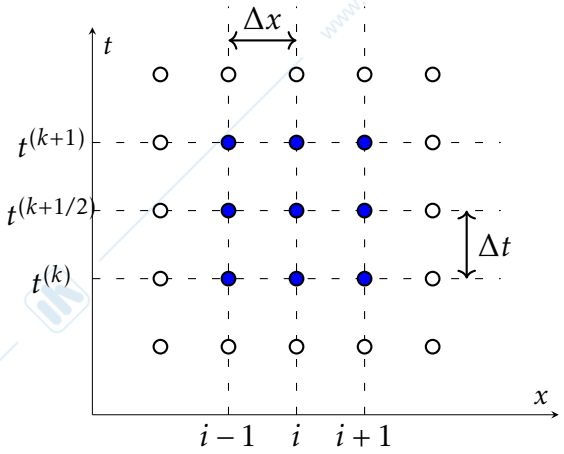


Figure 8.8: *discretization grid (Crank - Nicolson method)*

rewritten in the following way:

$$\frac{A_{i+1}^{(k+1/2)} - 2A_i^{(k+1/2)} + A_{i-1}^{(k+1/2)}}{\Delta y^2} + O(\Delta y)^2 = \sigma\mu \frac{A_i^{(k+1)} - A_i^{(k)}}{\Delta t} + O(\Delta t^2). \tag{8.69}$$

The expression for the time derivative has not changed with respect to Eq. (8.64) and Eq. (8.68), but it now represents the derivative computed at instant $(k+1/2)$:

$$\left.\frac{\partial A_z}{\partial t}\right|_{y=y_i}^{t=t_{k+1/2}} = \frac{A_i^{(k+1)} - A_i^{(k)}}{\Delta t} \tag{8.70}$$

However, most importantly, this derivative is now computed through a **central difference**, granting this technique a *second-order* degree of truncation. Of course, the values of $(A_{i+1},\, A_i,\, A_{i-1})$ appearing in the left-hand side of Eq. (8.69) have to be considered as an average of the values of the same points ($A_{i+1},\, A_i,\, A_{i-1}$) between instants (k) and $(k+1)$. Here, as an example:

$$A_i^{k+1/2} = \frac{A_i^{(k+1)} + A_i^{(k)}}{2}. \tag{8.71}$$

This method assumes that the unknown function evolves linearly between each couple of adjacent time instants. It should be also noticed that the only known-term here is $A_i^{(k)}$, thus making this one an *implicit* method, therefore requiring a linear system to be solved in order to *time-march* to the next time instant.

8.4 Magnetic quasi-stationary planar problems

Let's see how the Finite Element Method (FEM) can be applied to the solution of the diffusion equation for the magnetic vector potential. Recalling the formulation of the problem:

$$\nabla \times \left(\frac{1}{\mu} \nabla \times \boldsymbol{A} \right) = \boldsymbol{J}_0 - \sigma \frac{\partial \boldsymbol{A}}{\partial t}, \tag{8.72}$$

this equation describes a general three-dimensional case. For the sake of simplicity, we would like to reduce this one to a planar problem. Hence, let's assume that $\boldsymbol{A}$ and $\boldsymbol{J}$ have non-null components only along the z direction. This allows writing:

$$\boldsymbol{A} = A_z(x,y,t)\hat{\boldsymbol{k}} \qquad \boldsymbol{J} = J_z(x,y,t)\hat{\boldsymbol{k}}. \tag{8.73}$$

Weak formulation derivation

Let's consider a linear operator L, such that:

$$L[\boldsymbol{A}] = \nabla \times \left(\frac{1}{\mu} \nabla \times \boldsymbol{A} \right) - J_0 + \sigma \frac{\partial \boldsymbol{A}}{\partial t} = 0. \tag{8.74}$$

Our goal is to determine the values of A_z that verify $L[A] = 0$. In order to solve this problem on a generic domain, let's employ a triangular mesh. This allows us adopting a piecewise linear interpolation for the unknown:

$$A_z(x,y,t) \approx \tilde{A} = N_1(x,y)A_1(t) + N_2(x,y)A_2(t) + \ldots + N_n(x,y)A_n(t) = \sum_{i=1}^{n} N_i(x,y)A_i(t), \tag{8.75}$$

where A_i is the value of A_z sampled on the ith node of the mesh, whereas N_i is the ith shape function employed for the interpolation. It should be noticed that the shape functions are only spatial-dependant, whereas the values of the MVP are only time-dependant. As one can see, the so-called *weighted residual approach* is being applied, the same way described in the context of the FEM formulation of the Poisson problem. Once the interpolation has been performed, a local error function is defined:

$$\boldsymbol{e}(x,y,z,t) = L[\tilde{\boldsymbol{A}}] = \nabla \times \left(\frac{1}{\mu} \nabla \times \tilde{\boldsymbol{A}} \right) - \boldsymbol{J}_0 + \sigma \frac{\partial \tilde{\boldsymbol{A}}}{\partial t} \neq 0. \tag{8.76}$$

As the elimination of the local error in each point would be impossible, a average condition of null error can be enforced:

$$\int_V \boldsymbol{W} \cdot \boldsymbol{e}\, dV = \int_V \boldsymbol{W} \cdot \left[\nabla \times \left(\frac{1}{\mu} \nabla \times \tilde{\boldsymbol{A}} \right) - \boldsymbol{J}_0 + \sigma \frac{\partial \tilde{\boldsymbol{A}}}{\partial t} \right] dV = 0 \tag{8.77}$$

However, thanks to the assumption of a planar problem, the above condition has to be enforced only along the k axis (corresponding to the z component of the MVP), leading to:

$$-\int_V W_k \nabla \cdot \left(\frac{1}{\mu} \nabla \tilde{A}_z \right) dV = \int_V W_z J_{0,z} dV - \int_V W_z \sigma \frac{\partial \tilde{A}_z}{\partial t} dV. \tag{8.78}$$

The situation should look now very similar to the 2D Poisson formulation, with the left hand side of Eq. (8.78) presenting the issue a second-order derivative of $\tilde{A}_z$, which is a C^0 function. The problem can now be re-arranged in an equivalent *weak formulation*:

$$\int_\Omega \nabla W_z \cdot \left(\frac{1}{\mu}\nabla \tilde{A}_z\right) dS - \oint_{\partial\Omega} W_z \frac{1}{\mu}\frac{\partial \tilde{A}_z}{\partial n} dl = \int_\Omega W_z J_{0,z} dS - \int_\Omega W_z \sigma \frac{\partial \tilde{A}_z}{\partial t} dS. \tag{8.79}$$

It should also be noticed that thanks to the assumptions on the nature of $\boldsymbol{A}$ and $\boldsymbol{J}$, the discretization can be now performed on the two-dimensional domain Ω.

Galerkin approach and discretization

Adopting the Galerkin approach means choosing the shape functions as weighting functions:

$$W_z(x,y) \Rightarrow N_i(x,y) \tag{8.80}$$

Finally, a suitable expression for a computer implementation can finally be obtained:

$$\int_\Omega \nabla N_i \cdot \left(\frac{1}{\mu}\nabla \tilde{A}_z\right) dS - \oint_{\partial\Omega} N_i \frac{1}{\mu}\frac{\partial \tilde{A}_z}{\partial n} dl = \int_\Omega N_i J_{0,z} dS - \int_\Omega N_i \sigma \frac{\partial \tilde{A}_z}{\partial t} dS. \tag{8.81}$$

The first three terms of Eq. (8.81) have already been analysed in the context of the 2-D Poisson problem, and the only new contribution requiring further discussion is $\int_\Omega N_i \sigma \frac{\partial \tilde{A}_z}{\partial t} dS$. The first three terms represent the *static contribution* to the problem, and would therefore be found even in the formulation of a magnetostatic problem. This last term, instead, takes into account the contributions given to $\boldsymbol{A}$ by the phenomenon of the diffusion. As the ith shape function is non-null only on the ith support domain Ω_i:

$$\int_\Omega N_i \sigma \frac{\partial \tilde{A}_z}{\partial t} dS = \int_{\Omega_i} N_i \sigma \frac{\partial \tilde{A}_z}{\partial t} dS = \sum_{\omega_i \in \Omega_i} \int_{\omega_i} N_i \sigma \frac{\partial \tilde{A}_z}{\partial t} dS. \tag{8.82}$$

For the ith element, Eq. (8.75) takes now the following form:

$$\frac{\partial \tilde{A}_z}{\partial t} = N_i \frac{\partial \tilde{A}_{z,i}}{\partial t} + N_j \frac{\partial \tilde{A}_{z,j}}{\partial t} + N_k \frac{\partial \tilde{A}_{z,k}}{\partial t}. \tag{8.83}$$

The substitution of this result in Eq. (8.82) yields:

$$\int_\Omega N_i \sigma \frac{\partial \tilde{A}_z}{\partial t} dS = \left[\int_{\omega_i} N_i \sigma N_i \, dS\right] \frac{\partial \tilde{A}_{z,i}}{\partial t} + \left[\int_{\omega_i} N_i \sigma N_j \, dS\right] \frac{\partial \tilde{A}_{z,j}}{\partial t} + \left[\int_{\omega_i} N_i \sigma N_k \, dS\right] \frac{\partial \tilde{A}_{z,k}}{\partial t} \tag{8.84}$$

If the element ω_i is has the nodes i, j, k as its vertices this is the so-called element matrix:

$$S_{el}(\omega_i) = \begin{bmatrix} \int_{\omega_i} N_i \sigma N_i \, dS & \int_{\omega_i} N_i \sigma N_j \, dS & \int_{\omega_i} N_i \sigma N_k \, dS \\ \int_{\omega_i} N_j \sigma N_i \, dS & \int_{\omega_i} N_j \sigma N_j \, dS & \int_{\omega_i} N_j \sigma N_k \, dS \\ \int_{\omega_i} N_k \sigma N_i \, dS & \int_{\omega_i} N_k \sigma N_i \, dS & \int_{\omega_i} N_k \sigma N_k \, dS \end{bmatrix} \tag{8.85}$$

which can be expressed in a more compact form as:

$$S_{el}(\omega_i) = \int_{\omega_i} [N]^t \, \sigma \, [N] \, dS. \tag{8.86}$$

Assembling and solving the linear system

The computed element matrices $[K_{el}], [S_{el}]$ and $\{rhs_{el}\}$ are assembled in order to form the global matrices, here expressed using the same notation of the element matrices, but without the subscripts:

$$[K]\left\{\tilde{A}_z\right\} + [S]\left\{\frac{d\tilde{A}_z}{dt}\right\} = \{rhs\}. \tag{8.87}$$

Here the matrices $\{A_z\}$ and $\left\{\frac{\partial A_z}{\partial t}\right\}$ gather the nodal values of the MVP and its time-derivative respectively:

$$\left\{\tilde{A}_z\right\} = \begin{Bmatrix} A_{z,1} \\ \vdots \\ A_{z,i} \\ \vdots \\ A_{z,n} \end{Bmatrix} \qquad \left\{\frac{d\tilde{A}_z}{dt}\right\} = \begin{Bmatrix} \frac{dA_{z,1}}{dt} \\ \vdots \\ \frac{dA_{z,i}}{dt} \\ \vdots \\ \frac{dA_{z,n}}{dt} \end{Bmatrix}. \tag{8.88}$$

Summarizing, a problem described by a partial differential equation expressed in a strong formulation with $\nabla \times \left(\frac{1}{\mu}\nabla \times A_z\right) = J_{0,z} - \sigma\frac{\partial A_z}{\partial t}$, has been reformulated by means of a weak formulation and discretized. Its approximated solution, $\tilde{A}_z$, can be found by solving the linear system described by Eq. (8.87). Intuitively, the reliability of the approximated solution, i.e. how close it will be to the analytical (exact) one, is somehow related to the refinement of the employed mesh.

Time discretization - general case

Until now, only the space variables have been discretized, and the issue constituted by the time has not been took into account. An idea for performing this task could be the use of the aforementioned **Crank - Nicolson** method, which features a second-order accuracy for the time-derivative process. Our goal is finding a procedure that, given the spatial distribution of $\tilde{A}_z$ at a generic time instant (k), allows computing it the subsequent $(k+1)$ and so on. That process is called **time marching**, and it is performed as described in 8.3 through the discretization the time derivative as:

$$\left.\frac{dA_{z,i}}{dt}\right|_{x=x_i}^{t=t_{k+1/2}} = \frac{A_{z,i}^{(k+1)} - A_{z,i}^{(k)}}{\Delta t}. \tag{8.89}$$

Also, the MVP at instant $(k+1/2)$ using the centered difference formula as an average betweens its values at instants (k) and $(k+1)$ respectively.

$$A_i^{k+1/2} = \frac{A_{z,i}^{(k+1)} + A_{z,i}^{(k)}}{2}. \tag{8.90}$$

Following this approach, Eq. (8.87) can be re-written as:

$$\frac{1}{2}[K]\left(\{A_z\}^{(k+1)} + \{A_z\}^{(k)}\right) + \frac{1}{\Delta t}[S]\left(\{A_z\}^{(k+1)} + \{A_z\}^{(k)}\right) = \{rhs\}^{k+1/2}, \tag{8.91}$$

From which, isolating the unknown matrix $[A_z]^{(k+1)}$:

$$\left(\frac{1}{2}[K] + \frac{1}{\Delta t}[S]\right)\{A_z\}^{(k+1)} = \{rhs\}^{k+1/2} + \left(\frac{1}{\Delta t}[S] - \frac{1}{2}[K]\right)\{A_z\}^{(k)}. \tag{8.92}$$

If we now define two auxiliar matrices $[M_1]$ and $[M_2]$, as:

$$[M_1] = \frac{1}{2}[K] + \frac{1}{\Delta t}[S] \qquad [M_2] = \frac{1}{\Delta t}[S] - \frac{1}{2}[K], \tag{8.93}$$

we get the final expression:

$$[M_1]\{A_z\}^{(k+1)} = \{rhs\}^{k+1/2}[M_2]\{A_z\}^{(k)}, \tag{8.94}$$

which is once again a linear system. Therefore, if the MVP is know on the domain at the instant (k), thanks to 8.94 the distribution of $\tilde{A}_z$ at instant $(k+1)$ can be implicitly found. The presented method can be used for the assessment of any time-variant diffusion problem. However, as for the majority of the application in the power-engineering field the assumption of sinusoidal steady-state is acceptable, it's worth reviewing a dedicated method.

Time discretization - sinusoidal steady-state

If we want to make the assumption of a system governed by a sinusoidal law, the following conditions must be met:

1. **Sinusoidal and iso-frequential eccitation** $(\boldsymbol{J}_0)$ of the system;
2. **Linearity of the employed materials**. As an example, let's take a look at the case of a geometry involving some kind of iron among the meshed materials. Iron's magnetic relative permeability μ_r is non-linear, and thus (upon reaching saturation) upper harmonics would inevitably appear in the currents. The superposition of any upper harmonic frequency to the fundamental would inevitably distort the waveform (which would nevertheless be periodic), and determine a *non-sinusoidal* regime. Some kind of analogy with this case will be valid for any non-linearity of the materials, that must be therefore neglected in this particular context.

If the excitation of the system a frequency f $[Hz]$, $\omega = 2\pi f$ $[rad/s]$ can be derived and used to perform the so-called **Steinmetz transform**. This allows eliminating the time-dependency of the variables:

$$A_{z,i}(t) \rightarrow \underline{A_{z,i}} = A_{z,i}e^{j\alpha_i} \qquad \frac{\partial A_{z,i}(t)}{\partial t} \rightarrow j\omega\underline{A_{z,i}} = j\omega A_{z,i}e^{j\alpha_i}, \tag{8.95}$$

where α_i is the angular phase of the complex number $\underline{A_{z,i}}$.. Thanks to this transformation, the linear system can now be re-written and solved:

$$[K]\left\{\underline{\tilde{A}_z}\right\} + j\omega[S]\left\{\underline{\tilde{A}_z}\right\} = \{\underline{rhs}\} \tag{8.96}$$

$$([K] + j\omega[S])\left\{\underline{\tilde{A}_z}\right\} = \{\underline{rhs}\} \tag{8.97}$$

$$[M]\left\{\underline{\tilde{A}_z}\right\} = \{\underline{rhs}\}. \tag{8.98}$$

Finally, a linear non-stationary problem, described by a partial differential equation $\in \mathbb{R}$ has been brought back to a stationary complex problem, whose unknowns are now phasors.