CHAPTER 3

Solving for the Scalar Magnetic Potential

3.1 THE "VARIATIONAL" FORMULATION

We now treat the problem we arrived at for what it is, an *equation*, to be studied and solved as such: Given a bounded domain D, a number I (the mmf), and a function μ (the permeability), subject to the conditions $0 < \mu_0 \le \mu \le \mu_1$ of Eq. (18) in Chapter 2,

$$(1) \qquad \left| \begin{array}{l} \textit{find } \varphi \in \Phi^I = \{\varphi \in \Phi : \ \varphi = 0 \text{ on } S^h_0, \ \varphi = I \text{ on } S^h_1\} \textit{ such that} \\ \\ \quad \int_D \mu \, \text{grad}\, \varphi \cdot \text{grad}\, \varphi' = 0 \quad \forall \ \varphi' \in \Phi^0. \end{array} \right.$$

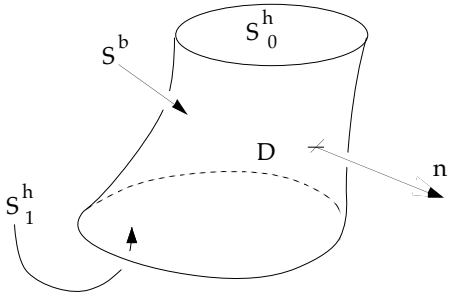


FIGURE 3.1. The situtation, reduced to its meaningful geometrical elements.

All potentials φ and test functions φ' belong to the encompassing linear space Φ of piecewise smooth functions on D (cf. 2.4.2), and the geometrical elements of this formulation, surface $S = S^h \cup S^b$, partition $S^h = S^h_0 \cup S^h_1$ of the "magnetic wall" S^h (Fig. 3.1), are all that we abstract from the concrete situation we had at the beginning of Chapter 2. We note that the magnetic energy (or rather, coenergy, cf. Remark 2.6) of

$h = \text{grad}\,\varphi$, that is,

$$F(\varphi) = \tfrac{1}{2} \int_D \mu\, |\text{grad}\,\varphi|^2,$$

is finite for all elements of Φ. The function F, the type of which is *FIELD* → *REAL*, and more precisely, $\Phi \to \mathbb{R}$, is called the *(co)energy functional*.

Remark 3.1. The use of the quaint term "functional" (due to Hadamard), not as an adjective here but as a somewhat redundant synonym for "function", serves as a reminder that the argument of F is not a simple real- or vector-valued variable, but a point in a space of infinite dimension, representative of a field. This is part of the "functional" point of view advocated here: One *may* treat complex objects like fields as mere "points" in a properly defined functional space. ◊

Function F is quadratic with respect to φ, so this is an analogue, in infinite dimension, of what is called a *quadratic form* in linear algebra. Quadratic forms have associated polar forms. Here, by analogy, we define the *polar form* of F as $\mathcal{F}(\varphi, \psi) = \int_D \mu\, \text{grad}\,\varphi \cdot \text{grad}\,\psi$, a bilinear function of two arguments, that reduces to F, up to a factor 2, when both arguments take the same value.

The left-hand side of (1) is thus $\mathcal{F}(\varphi, \varphi')$. This cannot be devoid of significance, and will show us the way: In spite of the dimension being infinite, let us try to apply to the problem at hand the body of knowledge about quadratic forms. There is in particular the following trick, in which only the linearity properties are used, not the particular way F was defined: For any real λ,

$$0 \le F(\varphi + \lambda\psi) = F(\varphi) + \lambda\, \mathcal{F}(\varphi, \psi) + \lambda^2\, F(\psi) \quad \forall\, \psi \in \Phi. \tag{2}$$

One may derive from this, for instance, the Cauchy–Schwarz inequality, by noticing that the discriminant of this binomial function of λ must be nonnegative, and hence

$$\mathcal{F}(\varphi, \psi) \le 2\, [F(\varphi)]^{1/2}\, [F(\psi)]^{1/2},$$

with equality only if $\psi = a\varphi + b$, with a and b real, $a \ge 0$. Here we shall use (2) for a slightly different purpose:

Proposition 3.1. *Problem* (1) *is equivalent to*

$$\textit{Find } \varphi \in \Phi^I \textit{ such that } \quad F(\varphi) \le F(\psi) \quad \forall\, \psi \in \Phi^I, \tag{3}$$

the coenergy minimization *problem*.

Proof. Look again at Fig. 2.8, and at Fig. 3.2 below. If φ solves (3), then $F(\varphi) \le F(\varphi + \lambda\varphi')$ for all φ' in Φ^0, hence $\lambda\, \mathcal{F}(\varphi, \varphi') + \lambda^2 F(\varphi') \ge 0$ for all $\lambda \in \mathbb{R}$, which implies (the discriminant, again[1]) that $\mathcal{F}(\varphi, \varphi') = 0$ for all φ' in Φ^0, which is (1). Conversely, if φ solves (1), and ψ belongs to Φ^I, then (cf. (2)) $F(\psi) = F(\varphi) + \mathcal{F}(\varphi, \psi - \varphi) + F(\psi - \varphi) = F(\varphi) + F(\psi - \varphi) \ge F(\varphi)$, since $\psi - \varphi \in \Phi^0$ and $F(\psi - \varphi) \ge 0$. $\Diamond$

This confirms our intuitive expectation that the physical potential should be the one, among all eligible potentials, that minimizes the coenergy. Problem (3) is called the *variational form* of the problem. In the tradition of mathematical physics, a problem has been cast in variational form when it has been reduced to the minimization of some function subject to some definite conditions, called "constraints". The constraint, here, is that φ must belong to the affine subspace Φ^I (an *affine constraint,* therefore). Such problems in the past were the concern of the calculus of variations, which explains the terminology. Nowadays, Problem (1) is often described as being "in variational form", but this is an abuse of language, for such a weak formulation does not necessarily correspond to a minimization problem: In harmonic-regime high-frequency problems, for instance, a complex-valued functional is stationarized, not minimized. For the sake of definiteness, I'll refer to (1) as "the weak form" and to (3) as "the variational form".

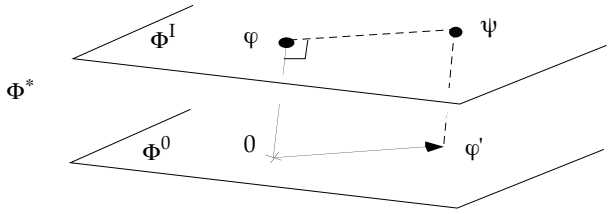


FIGURE 3.2. Geometry of the variational method ($\psi = \varphi + \varphi'$).

Conversely, however, variational problems with affine constraints have as a rule a weak form, which can be derived by consideration of the *directional derivative*[2] of F at point φ. By definition, the latter is the linear map

$$(4) \qquad \psi \to \lim_{\lambda \to 0} [F(\varphi + \lambda\psi) - F(\varphi)]/\lambda.$$

If φ yields the minimum, the directional derivative of F should vanish

[1]Alternatively, first divide by λ, then let λ go to 0.

[2]Known as the *Gâteaux derivative*.

at φ, for all directions that satisfy the constraint. The condition obtained that way is called the *Euler equation* of the variational problem.

Here, (4) is the map $\psi \to \mathcal{F}(\varphi, \psi)$, after (2). Therefore, Problem (1), which expresses the cancellation of this derivative in all directions parallel to Φ^0, is the Euler equation of the coenergy minimization problem (3).

Exercise 3.1. Find variational forms for Problems (2.34) and (2.36).

In the space Φ^* of the last chapter (Exer. 2.9), which is visualized as ordinary space in Fig. 3.2, we may define a *norm*, $\|\varphi\|_\mu = (2F(\varphi))^{1/2} \equiv [\int_D \mu\, |\text{grad}\,\varphi|^2]^{1/2}$, hence a notion of distance: The *distance in energy* of two potentials is $d_\mu(\varphi,\psi) = \|\varphi-\psi\|_\mu \equiv [\int_D \mu\, |\text{grad}(\varphi-\psi)|^2]^{1/2}$. The variational problem can then be described as the search for this potential in Φ^I that is closest to the origin, in energy: in other words, the *projection* of the origin on Φ^I.

Moreover, this norm stems from a scalar product, which is here, by definition, $(\varphi, \psi)_\mu = \int_D \mu\, \text{grad}\,\varphi \cdot \text{grad}\,\psi$ ($\equiv \mathcal{F}(\varphi, \psi)$, the polar form), with $\|\varphi\|_\mu = [(\varphi, \varphi)_\mu]^{1/2}$. The weak form also then takes on a geometrical interpretation: It says that vector φ is orthogonal to Φ^0, which amounts to saying (Fig. 3.2) that point φ is the *orthogonal projection* of the origin on Φ^I. The relation we have found while proving Prop. 3.1,

$$F(\psi) = F(\varphi) + F(\psi - \varphi) \quad \forall\, \psi \in \Phi^I, \tag{5}$$

if φ is the solution, then appears as nothing but the Pythagoras theorem, in a functional space of infinite dimension.

Exercise 3.2. Why the reference to Φ^*, and not to Φ ?

This is our first encounter with a *functional space:* an affine space, the elements of which can usually be interpreted as functions or vector fields, equipped with a notion of distance. When, as here, this distance comes from a scalar product on the associated vector space, we have a *pre-Hilbertian* space. (Why "pre" will soon be explained.) The existence of this metric structure (scalar product, distance) then allows one to speak with validity of the "closeness" of two fields, of their orthogonality, of converging sequences, of the continuity of various mappings, and so forth. For instance (and just for familiarization, for this is a trivial result), if we call $\varphi(I)$ the solution of (1) or (3), considered as a function of the mmf I, we have

Proposition 3.2. *The mapping* $I \to \varphi(I)$ *is continuous in the energy metric.*

Proof. By (1), $\varphi(I) = I\,\varphi(1)$, hence $\|\varphi(I)\|_\mu = |I|\,\|\varphi(1)\|_\mu$, that is, $\|\varphi(I)\|_\mu \leq$

$C\,|I|$ for all I, where C is a constant, thus satisfying the criterion for continuity of linear operators. ◊

Exercise 3.3. Show that $\mathcal{J}$ (notation of Exer. 2.9) is continuous on Φ^*.

As for the functional point of view, also heralded before, we now have a good illustration of it: Having built a functional space of eligible potentials, we search for a distinguished one, here the orthogonal projection of the origin on Φ^I.

Remark 3.2. Once we have this solution φ, then, by the integration by parts formula, $\int_D \mu\,|\mathrm{grad}\,\varphi|^2 = \int_S \mathbf{n}\cdot\mathbf{b}\;\varphi = I\int_{S^h_1} \mathbf{n}\cdot\mathbf{b} = I^2/R$, by definition of the reluctance R (cf. 2.4.1). So, finding the magnetic coenergy will give access to R. We'll return to this in Chapter 4 (Subsection 4.1.3). ◊

3.2 EXISTENCE OF A SOLUTION

After this promising commencement, the bad news: Problems (1) or (3) *may fail to have a solution*.

3.2.1 Trying to find one

Call d the distance of the origin to Φ^I, that is, $d = \inf\{\|\psi\|_\mu : \psi \in \Phi^I\}$. For each integer n, there certainly exists some φ_n in Φ^I such that $\|\varphi_n\|_\mu \le d + 1/n$. (Otherwise, d would be lower than the infimum.) Moreover, $d = \lim_{n\to\infty} \|\varphi_n\|_\mu$. One says that the φ_ns form a *minimizing sequence*, which we may expect to converge towards a limit φ. If so, this limit will be the solution.

Indeed, by developing $\|\varphi_n \pm \varphi_m\|_\mu^2 = \int_D \mu\,|\mathrm{grad}(\varphi_n \pm \varphi_m)|^2$, we have

$$\|\varphi_n - \varphi_m\|_\mu^2 + \|\varphi_n + \varphi_m\|_\mu^2 = 2(\|\varphi_n\|_\mu^2 + \|\varphi_m\|_\mu^2).$$

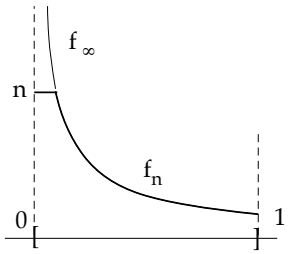
The point $(\varphi_n + \varphi_m)/2$ belongs to Φ^I, so its distance to 0 is no smaller than d; therefore $\|\varphi_n + \varphi_m\|_\mu^2 \ge 4d^2$, and hence,

$$\|\varphi_n - \varphi_m\|_\mu^2 \le 2(\|\varphi_n\|_\mu^2 + \|\varphi_m\|_\mu^2) - 4d^2. \tag{6}$$

Now, let n and m tend to infinity; the right-hand side tends to 0, so $\|\varphi_n - \varphi_m\|_\mu$ tends to 0; this qualifies $\{\varphi_n : n \in \mathbb{N}\}$ as a *Cauchy sequence*, which a converging sequence must be (cf. Appendix A, Subsection A.4.1).

But this necessary condition is not sufficient. Just as the set of rational numbers does not contain all the limits of its Cauchy sequences, functional

spaces which do not contain the limits of their own abound. For instance (inset), in the space of piecewise smooth functions over [0, 1], equipped with the norm $\| f \| = \int_0^1 |f(x)|\, dx$, the sequence $f_n = x \to \inf(n, 1/\sqrt{x})$ is Cauchy (**Exercise 3.4**: prove it), but the would-be limit $x \to 1/\sqrt{x}$ is *not* piecewise smooth. Hence the necessity of the following definition:



Definition 3.1. *A metric space* X *is* complete *if all Cauchy sequences in* X *converge towards an element of* X.

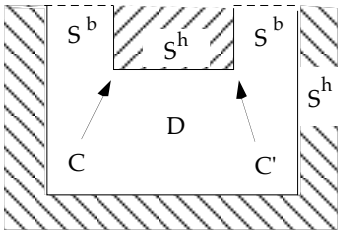
In particular, a complete normed space is called a *Banach space*, and a complete pre-Hilbertian space is called a *Hilbert space*.

If our underlying space Φ^* was complete (and then each slice Φ^I, being closed by Exer. 3.3, would be complete), the above reasoning would thus establish the existence of a solution to (1) or (3), which we already know is unique. But Φ^* is not complete, as counterexamples built on the same principle as in Exer. 3.4 will show. What this points to, however, is a failure of our *method*. Conceivably, a piecewise smooth solution could exist in spite of our inability to prove its existence a priori by this minimizing sequence approach. One might even be tempted to say, "Never mind, we know this solution exists on physical grounds, and we shall be content with an approximation (that is, an element of high enough rank of some minimizing sequence). Moreover, didn't we *prove*, with this Weyl lemma, that φ would be smooth in homogeneous regions? If so the present space Φ, though not complete, is rich enough. So let's proceed and focus on finding a usable approximation."

Such a stand would not be tenable. First, there is a logical fallacy: smoothness was proved, in Chapter 1, but in case the solution *exists*, which is what we want to assess. Besides, you don't prove something "on physical grounds". Rather, modelling sets up a correspondence between a segment of reality and a mathematical framework, by which some empirical *facts* have mathematical *predicates* as counterparts. The truth of such predicates must be proved *within the model*, and failure to achieve that just invalidates *the modelling*. So the responsibility of asserting the existence of a solution to (1) or (3), within this mathematical framework, is ours.

Alas, not only can't we prove piecewise smoothness a priori, but we can build counterexamples, that correspond to quite realistic situations. We shall display one.

3.2.2 Φ^* is too small



Refer to Fig. 2.6 (recalled in inset), and imagine the system as so long in the z–direction that all field lines are in the x–y plane, which makes a 2D modelling feasible. It is well known that the field will be infinite at the tips of the "re-entrant corners" C and C' with such geometry. (This is the same phenomenon as the "spike effect" in electrostatics.) By doing Exercises 5 and 6, you should be able to see why: An analogue of Problems (1) and (3), in an appropriately simplified two-dimensional setting, can be solved in closed form, and its solution exhibits a mild singularity at the origin (which corresponds to corner C). The potential is well-behaved (cf. Fig. 3.6, p. 89), but its gradient becomes infinite at the origin, in spite of the magnetic coenergy[3] being bounded.

This is an idealization, but it points to an unacceptable weakness of our modelling: The restriction to piecewise smooth potentials, which seemed quite warranted, bars the existence of such mild singularities,[4] whereas physics requires they be accounted for, as something that can happen. Our space is too small: The frame is too narrow.

Of course, we could blame this failure on too strict a definition of smoothness, and revise the latter in the light of new data, contriving to accept mildly singular fields as "smooth" according to some new, looser definition. But first, this kind of "monster-barring" [La] would lead to even more technical concepts and (likely) to something more esoteric than the radical solution we shall eventually adopt. Moreover, it might be only the beginning of an endless process: One may easily imagine how fractal-like boundaries, for instance, could later be invoked to invalidate our attempts to deal with corners.

The radical (and right) solution is *completion:* Having a non-complete functional space, immerse it into a larger, complete space. Then the above method works: The solution exists, in the completed space, as the limit of a minimizing sequence. (All of them will yield the same limit.)

[3]In such a 2D modelling, the μ-norm corresponds to the (co)energy contributed by the region of space lying between two horizontal planes, one unit of length apart.

[4]Precisely: the singularity at 0 makes it impossible to extend φ to a domain that would contain the origin and where its gradient would be finite, which is required by 1-smoothness "over" D, as we defined it.

3.2.3 Completing Φ^*

Completion logically belongs to the mathematical Appendix of this book, but the idea is so important, and so germane to what physicists do spontaneously when they define "generalized solutions" to problems which have no "classical" ones, that it may be worthwhile to discuss it here.

First, note this is not the same thing as *closure*. Indeed, if A is a part of a metric space $\{X, d\}$, sequences which fail to converge in A may converge to an element of its closure, so if X is complete, the completion of A will be its closure $\overline{A}$. But there, A is already immersed in a pre-existing metric space. If such an encompassing complete space does not yet exist, we can't proceed that way.

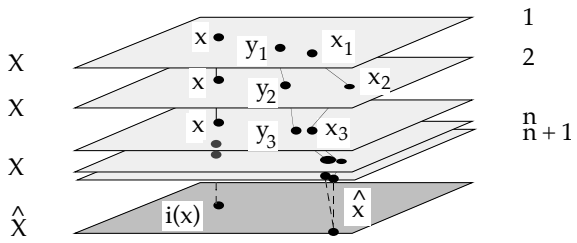


FIGURE 3.3. The idea of completion. $\hat{X}$ is an abstract space: its elements are of a different type that those of X. But there is a natural injection of X into $\hat{X}$, so the process can be seen, very informally, as "plugging the holes" in X.

The idea (Fig. 3.3) conforms with the usual method for building new mathematical objects from old ones: define equivalence relations (cf. A.1.6), and take equivalence *classes*. This is how, one will remember, rational numbers are built from pairs of integers, and real numbers from sequences of rationals. Completion is quite analogous to the construction of $\mathbb{R}$ in this respect. Suppose $\{X, d\}$ is a metric space, that is, a set X equipped with a distance d. Let X° be the set of all Cauchy sequences in X. Two elements of X°, say $x^\circ = \{x_1, x_2, \ldots, x_n, \ldots\}$ and $y^\circ = \{\ldots, y_n, \ldots\}$, will be deemed equivalent if $\lim_{n\to\infty} d(x_n, y_n) = 0$ (Fig. 3.3). That is easily seen to be an equivalence relation, under the hypothesis that we are dealing with Cauchy sequences. We thus consider the quotient $\hat{X}$, we give it a distance, $\hat{d}(\hat{x}, \hat{y}) = \lim_{n\to\infty} d(x_n, y_n)$, where $\{x_n\}$ and $\{y_n\}$ are representatives of the classes $\hat{x}$ and $\hat{y}$, and we define the newly found metric space $\{\hat{X}, \hat{d}\}$ as *the completion of* X. This is a bold move, for the elements of $\hat{X}$, being sets of sequences of elements of X, seem of a completely different

nature than those of X. But there is a natural way to inject X into $\hat{X}$: to $x \in X$, associate the class $\hat{x} = i(x)$ of the constant sequence $x^\circ = \{x, x, \ldots, x, \ldots\}$. This way, $\hat{X}$ appears as an extension of X (note that $\hat{d}$ restricts to d on the image $i(X)$ of X under the injection i). Moreover, as proved in A.4.1, $\{\hat{X}, \hat{d}\}$ is complete, and X, or rather its image $i(X)$, is dense in $\hat{X}$.

This mechanism does not guarantee that the completion of a functional space will be a functional space: its elements being equivalence classes of sequences of functions, some of these classes might not be identifiable[5] with any classically defined function. As a rule, one must invoke other mathematical theories to establish the functional nature of the elements of the completion—when such is the case.

The classical example is $L^2(D)$, the prototypal Hilbert space: $L^2(D)$ is defined as the completion of[6] $C_0^\infty(D)$ with respect to the norm $\|f\| = (\int_D |f|^2)^{1/2}$. A central result of Lebesgue integration theory, then, is that $L^2(D)$ coincides with the space of square-integrable functions over D, or rather, of equivalence *classes* of such functions, with respect to the "a.e. =" relation (equality except on a negligible set). If this sounds complex, it's because it really is . . . (see Appendix A, Subsection A.4.2). Fortunately, this complexity can be circumscribed, and once in possession of $L^2(D)$, and of its analogue $\mathbb{L}^2(D)$ (square integrable vector fields), completion is an easy task, as we shall see later.

Completion corresponds to a very natural idea in physics. Many problems are idealizations. For instance, there is no such thing in nature as a sharp corner, but the sharp corner idealization helps understand what happens near a surface with high curvature. In this respect, the whole *family* of solutions, parameterized by curvature, contains information that one solution for a finite curvature would fail to give. This information is summarized by the singular solution, which belongs to the completion, because an element of the completion *is*, in the sense we have seen, a sequence of smooth solutions.

[5]This is no hair-splitting: For instance, the completion of $C_0^\infty(E_2)$ with respect to the norm $\varphi \to (\int_{E_2} |\text{grad } \varphi|^2)^{1/2}$ is *not* a functional space, not even a space of distributions [DL]. Still, this *Beppo Levi space* is home to electric or magnetic potentials in 2D problems. This reflects the intrinsic difficulty of dimension 2, for in 3D, Beppo Levi's space is functional, being continuously injectable in the space $L^6(E_3)$ of functions with integrable sixth power. We'll see that in Chapter 7.

[6]Note that, since a space is dense in its completion, spaces in which $C_0^\infty(D)$ is dense, with respect to the same quadratic norm, have the same completion. So we would obtain the same result, $L^2(D)$, by starting from $C^1(\overline{D})$, or $C^0(\overline{D})$, or for that matter, from the space of piecewise smooth functions over D.

A little more abstractly, suppose the problem has been cast in the form $Ax = b$, where b symbolizes the data, x the solution, and A some mapping of type $SOLUTION \rightarrow DATA$. Solving the problem means, at a high enough level of abstraction, finding[7] the inverse A^{-1}, which may not be defined for some values of b (those corresponding to sharp corners, let's say, for illustration). But if there is a solution x_n for each element b_n of some sequence that converges toward b, it's legitimate to define the limit $x = \lim_{n \rightarrow \infty} x_n$ as the solution, if there is such a limit, and *if there isn't, to invent one.* That's the essence of completion. Moreover, attributing to Ax the value b, whereas A did not make sense, a priori, for the *generalized solution* x, constitutes a prolongation of A beyond its initial domain, a thing which goes along with completion (cf. A.2.3). Physicists made much mileage out of this idea of a generalized solution, as the eventual limit of a parameterized family, before the concepts of modern functional analysis (complete spaces, distributions, etc.) were elaborated in order to give it status.

Summing up: We now attribute the symbol Φ^* to the completion of the space of piecewise smooth functions in D, null on S^h_0 and equal to some constant on S^h_1, with respect to the norm $\|\varphi\|_\mu = [\int_D \mu \, |\text{grad}\, \varphi|^2]^{1/2}$. Same renaming for Φ^I (which is now the closure of the previous one in Φ^*). Equation (1), or Problem (3), has now a (unique) solution. The next item in order[8] is to *solve* for it.

3.3 DISCRETIZATION

But what do we mean by that? Solving an equation means being able to answer specific questions about its solution with controllable accuracy, whichever way. A century ago, or even more recently in the pre-computer era, the only way was to represent the solution "in closed form", or as the sum of a series, thus allowing a numerical evaluation with help of formulas and tables. Computers changed this: They forced us to work from the outset with *finite* representations. Eligible fields and solutions must

[7] An unpleasantly imprecise word. What is required, actually, is some *representation* of the inverse, by a formula, a series, an algorithm . . . anything that can give *effective* access to the solution.

[8] Whether Problem (3) is well posed (cf. Note 1.16) raises other issues, which we temporarily bypass, as to the continuous dependence of φ on data: on I (Prop. 3.2 gave the answer), on μ (cf. Exers. 3.17 and 3.19), on the dimensions and shape of the domain (Exers. 3.18 and 3.20).

therefore be parameterized, with a perhaps very large, but finite, number of parameters.

3.3.1 The Ritz–Galerkin method

Suppose we have a finite catalog of elements of Φ, $\{\lambda^i : i \in \mathcal{I}\}$, often called *trial functions*, where $\mathcal{I}$ is a (finite) set of indices. Each λ^i must be a simple function, one which can be handled in closed form. If we can find a family of real parameters $\{\boldsymbol{\varphi}_i : i \in \mathcal{I}\}$ such that $\sum_i \boldsymbol{\varphi}_i \lambda^i$ is an approximation of the solution, this will be enough to answer questions the modelling was meant to address, provided the approximation is good enough, because all the data-processing will be done via the λ^is. The parameters $\boldsymbol{\varphi}_i$ (set in bold face) are called the *degrees of freedom* (abbreviated as DoFs or DoF, as the case may be) of the field they generate. We shall denote by $\boldsymbol{\varphi}$, bold also, the family $\boldsymbol{\varphi} = \{\boldsymbol{\varphi}_i : i \in \mathcal{I}\}$.

The Ritz–Galerkin idea consists in restricting the search for a field of least energy to those of the form $\sum_{i \in \mathcal{I}} \boldsymbol{\varphi}_i \lambda^i$ that belong to Φ^I. The catalog of trial functions is then known as a *Galerkin basis*. (We shall say that it defines an *approximation method*, and use the subscript m to denote all things connected with it, when necessary; most often, the m will be understood.) This is well in the line of the above contructive method for proving existence, for successive enlargements of the Galerkin basis will generate a sequence with *decreasing* energy, and if, moreover, this is a minimizing sequence, the day is won.

To implement this, let us introduce some notation: Φ_m is the finite dimensional space of linear combinations of functions of the catalog, that is to say, the space spanned by the λ^is, and we define

$$\Phi^I{}_m = \Phi_m \cap \Phi^I, \qquad \Phi^0{}_m = \Phi_m \cap \Phi^0. \tag{7}$$

The approximate problem is thus:

$$\textit{Find } \varphi_m \in \Phi^I{}_m \textit{ such that } \quad F(\varphi_m) \le F(\psi) \quad \forall\, \psi \in \Phi^I{}_m . \tag{8}$$

This problem has a solution (by the compactness argument of A.2.3), since we are considering here a positive definite quadratic form on a *finite*-dimensional space. If in addition we assume that $\Phi^I{}_m$ and $\Phi^0{}_m$ are parallel, just as Φ^I and Φ^0 were in Fig. 3.2 (this is not automatic, and depends on a sensible choice of trial functions), then (8) is equivalent, by exactly the same reasoning we made earlier, to the following Euler equation:

(9) $\quad Find\ \varphi_m \in \Phi^{\mathrm{I}}{}_m\ such\ that\quad \int_D \mu\ \mathrm{grad}\ \varphi_m \cdot \mathrm{grad}\ \varphi' = 0\quad \forall\ \varphi' \in \Phi^0{}_m.$

This parallelism condition is usually easy to achieve: It is enough that some specific combination $\varphi^{\mathrm{I}}{}_m = \sum_i \boldsymbol{\varphi}_i^{\mathrm{I}} \lambda^i$ satisfy $\varphi^{\mathrm{I}}{}_m = 0$ on $S^h{}_0$ and $\varphi^{\mathrm{I}}{}_m = \mathrm{I}$ on $S^h{}_1$. If necessary, such a function will be built on purpose and added to the list of trial functions. Now all functions of $\Phi^{\mathrm{I}}{}_m$ are of the form $\varphi^{\mathrm{I}}{}_m + \varphi$ with $\varphi \in \Phi^0{}_m$, which we can write in compact form like this:

(10) $\quad \Phi^{\mathrm{I}}{}_m = \varphi^{\mathrm{I}}{}_m + \Phi^0{}_m.$

In words: $\Phi^{\mathrm{I}}{}_m$ is the *translate* of $\Phi^0{}_m$ by vector $\varphi^{\mathrm{I}}{}_m$ (Fig. 3.4).

It would now be easy to show that (9) is a *regular linear system*. The argument relies on uniqueness and on the equality between the number of unknowns (which are the degrees of freedom) and the number of equations in (9), which is the dimension of $\Phi^0{}_m$ (cf. Exer. 2.7). We defer this, however, as well as close examination of the properties of this linear system, till we have made a specific choice of trial functions.

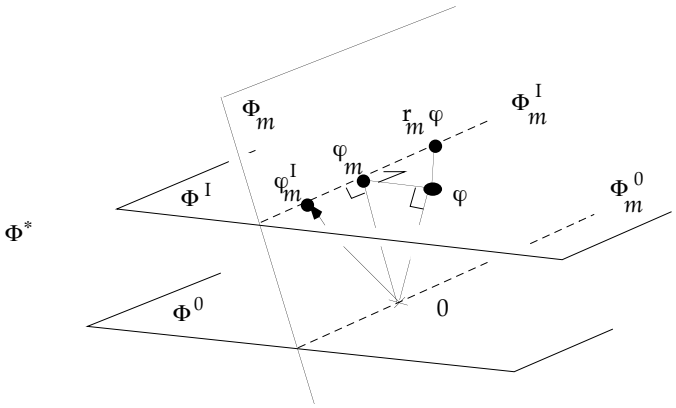


FIGURE 3.4. Geometry of the Ritz–Galerkin method.

Problem (9) is called the "discrete formulation", as opposed to the "continuous formulation" (1). Both are in weak form, but (9) is obviously "weaker", since there are fewer test functions. In particular, the weak solenoidality of $b = \mu\ \mathrm{grad}\ \varphi$ has been destroyed by restricting to a *finite* set of test functions. The span of such a set cannot be dense in Φ^0, so the proof of Prop. 2.3 is not available, and we can't expect Eqs. (2.23) and (2.24) in Chapter 2 (about $\mathrm{div}\ b = 0$ and $n \cdot b = 0$) to hold for $b_m = \mu\ \mathrm{grad}\ \varphi_m$. Still, something must be preserved, which we shall call, for lack of a better term, "*m*-weak solenoidality[9] of b" and "*m*-weak

enforcement of the $n \cdot b$ boundary condition" on S^b. This also will wait (till the next chapter).

Meanwhile, it's interesting to examine the geometry of the situation (Fig. 3.4). The figure suggests that φ_m, which is the projection of 0 on Φ^I_m, is also the projection of φ (the exact solution) on Φ^I_m. This is correct: To see it, just restrict the test functions in (1) to elements of Φ^0_m, which we have assumed (cf. (7)) are contained in Φ^0, which gives

$$\int_D \mu \text{ grad } \varphi \,.\, \text{grad } \varphi' = 0 \quad \forall \varphi' \in \Phi^0_m \,.$$

But by (9) we also have

$$\int_D \mu \text{ grad } \varphi_m \,.\, \text{grad } \varphi' = 0 \quad \forall \varphi' \in \Phi^0_m \,,$$

therefore, by difference,

$$\int_D \mu \text{ grad}(\varphi_m - \varphi) \cdot \text{grad } \varphi' = 0 \quad \forall \varphi' \in \Phi^0_m \,, \tag{11}$$

which expresses the observed orthogonality.

The figure also suggests a general method for error estimation. Let $r_m\varphi$ be an element of Φ^I_m that we would be able to associate with φ, as an approximation, *if* we knew it. Then we have, as read off the figure, and proved by setting $\varphi' = \varphi_m - r_m\varphi$ in (11),

$$\|\varphi - \varphi_m\|_\mu \le \|\varphi - r_m\varphi\|_\mu$$

($r_m\varphi$ is farther from φ, in energy, than φ_m is). So if we are able somehow to bound $\|\varphi - r_m\varphi\|_\mu$, an error bound on $\varphi_m - \varphi$ will ensue. The potential of the idea for error control is obvious, and we whall return to it in Chapter 4, with a specific Galerkin basis and a specific r_m.

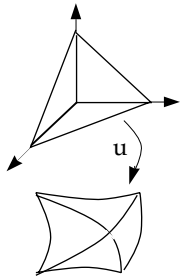
The Ritz–Galerkin method is of surprising efficiency. If trial functions are well designed, by someone who has good feeling for the real solution, a handful of them may be enough for good accuracy in estimating the functional. But it's difficult to give guidelines of general value in this respect, especially for three-dimensional problems. Besides, the computer changed the situation. We can afford many degrees of freedom nowadays (some modern codes use millions [We]) and can lavish machine time on the systematic design of Galerkin bases in a *problem-independent* way: This is what *finite elements* are about.

[9]The terminology is hesitant: Some say these equations are approximately satisfied "in the sense of weighted residuals", or "in the weak sense of finite elements", or even simply "in the weak sense", which may induce confusion. "Discrete" solenoidality might be used as a more palatable alternative to "m-weak" solenoidality.

3.3.2 Finite elements

So let be given a bounded domain $D \subset E_3$ with a piecewise smooth boundary S, and also inner boundaries, corresponding to material interfaces (discontinuity surfaces of μ, in our model problem).

A *finite element mesh* is a tessellation of D by volumes of various shapes, but arranged in such a way that two of them intersect, if they do, along a common face, edge, or node,[10] and never otherwise. We shall restrict here to tetrahedral meshes, where all volumes have six edges and four faces, but this is only for clarity. (In practice, hexahedral meshes are more popular.[11]) Note that a volume is not necessarily a straight tetrahedron, but may be the image of some "reference tetrahedron" by a smooth mapping u (inset).[12] This may be necessary to fit curved boundaries, or to cover infinite regions. Usually, one also arranges for material interfaces to be paved by faces of the mesh.



Exercise 3.7. Find all possible ways to mesh a cube by tetrahedra, under the condition that no new vertex is added.

Drafting a mesh for a given problem is a straightforward, if tedious, affair. But designing *mesh generators* is much more difficult, a scientific specialty [Ge] and an industry. We shall not touch either subject, and our only concern will be for the output of a mesh-generation process. The mesh is a complex data structure, which can be organized in many different ways, but the following elements are always present, more or less directly: (1) a list of nodes of the mesh, pointing to their locations; (2) a list of edges, faces, and volumes, with indirections allowing one to know which nodes are at the ends of this and that edge, etc.; (3) parameters describing the mapping of each volume to the reference one; (4) for each volume, parameters describing the material properties (for instance, the average value of μ, in our case).

For maximum simplicity in what follows, we assume that all volumes are straight tetrahedra. This can always be enforced, by distorting D to a polyhedron with plane faces, which is then chopped into tetrahedra.

[10]Or vertex. For some, "vertex" and "node" specialize in distinct meanings, vertices being the tips of the elementary volumes, and nodes the points that will support degrees of freedom. This distinction will not be made here.

[11]Most software systems offer various shapes, including tetrahedra and prisms, to be used in conjunction. This is required in practice for irregular regions.

[12]A more precise definition will be given in Chapter 7.

(This changes the model a little, of course, and adds some error to the approximation error inherent in the finite element method.)

We shall use the following simple description of the mesh: (1) four *sets*, denoted $\mathcal{N}$, $\mathcal{E}$, $\mathcal{F}$, $\mathcal{T}$, for nodes, edges, faces, and tetrahedra; (2) *incidence relations*, on which more below; (3) the *placement* of the mesh: this is a function $n \to x_n$, from $\mathcal{N}$ to $\overline{D}$, giving for each node n its position x_n in D or on S. In the case of straight tetrahedra, this is enough to determine the location of all *simplices* (the generic name for node, edge, face, etc.), and no other placement parameters are needed.

Thanks to this placement, one can confuse under a single expression, for example, "tetrahedron T", two conceptually different things: here the element T of $\mathcal{T}$, which is a mere label, and the tetrahedron T, a part of D, which is its image under the placement. It's a convenient and not too dangerous abuse,[13] which I'll commit freely, for all simplices. Symbols $\mathcal{F}(e)$, $\mathcal{N}(T)$, and other similar ones, will stand for, respectively, the subset of all faces that contain edge e, the subset of all nodes that are contained in tetrahedron T, and other similar subsets for various simplices. The purpose of the incidence relations, which we shall wait until Chapter 5 to describe in full detail, is to point to the faces of a given tetrahedron, the edges of a given face, etc., and thus to give full knowledge of subsets like $\mathcal{F}(e)$ or $\mathcal{N}(T)$. Finally, we shall denote by D_n the subdomain of D obtained by putting together all tetrahedra of the subset $\mathcal{T}(n)$, and use similar notation for D_e and D_f, calling D_s the *cluster* of tetrahedra *around* simplex s (Fig. 3.5). (No attempt is made to distinguish between open and closed clusters, as that will be clear from context.)

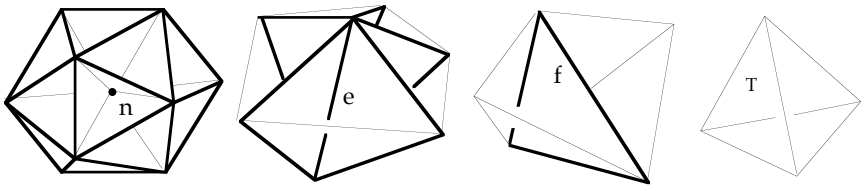


FIGURE 3.5. Clusters of tetrahedra around simplex s (s being, from left to right, node n, edge e, face f, and tetrahedron T). For better view, faces containing the simplex are supposed to be opaque, and others transparent.

Let's now recall the notion of barycentric coordinates. Four points x_1, x_2, x_3, x_4 in three-dimensional space are in *generic position* if the determinant $\det(x_2 - x_1, x_3 - x_1, x_4 - x_1)$ does not vanish. In that case, they

[13] Mathematicians use s for the simplex as an algebraic object and $|s|$ for its image.

form a tetrahedron. Four real numbers $\lambda^1, \lambda^2, \lambda^3, \lambda^4$ such that $\sum_i \lambda^i = 1$ determine a point x, the *barycenter* of the x_is for these *weights*, uniquely defined by

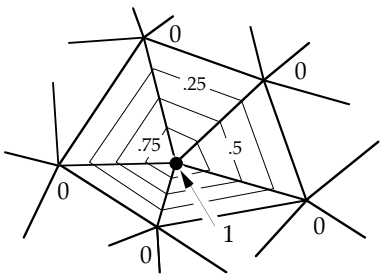
$$(12) \qquad x - x_0 = \sum_{i = 1, 4} \lambda^i (x_i - x_0),$$

where x_0 is any origin (for instance, one of the x_is). Conversely, any point x has a unique representation of the form (12), and the weights λ^i, considered as four functions of x, are the *barycentric coordinates* of x in the *affine basis* provided by the four points. Note that x belongs to the tetrahedron if $\lambda^i(x) \geq 0$ for all i. The λ^is are affine functions of x.

Remark 3.3. Consequently, a function p which is polynomial with respect to the three Cartesian coordinates can be expressed as a polynomial expression $x \to P(\lambda^1(x), \ldots, \lambda^4(x))$ of the barycentric coordinates, where P is another polynomial, *of the same maximum degree* as p, with four variables. This possibility is often used, usually without warning. ◊

Now, consider our paving of $\overline{D}$ by tetrahedra. To each node n of the mesh, let us attribute a function, defined as follows: Its value at point x is 0 if the cluster D_n does not contain x, and if it does, it is the barycentric coordinate of x with respect to n, in the affine basis provided by the tetrahedron to which x belongs. (There is no ambiguity in that, because if x belongs to a simplex s, and thereby, to all tetrahedra of the cluster of s, its barycentric coordinates with respect to vertices of s are all the same, whatever the tetrahedron one considers to reckon them.) We shall reattribute to this *nodal function* the symbol λ^n. Note that, by construction, $\lambda^n(x) \geq 0$, its support is $\overline{D}_n$, its domain is $\overline{D}$ (but doesn't go beyond), and

$$(13) \qquad \sum_{n \in \mathcal{N}} \lambda^n(x) = 1 \text{ for all } x \in \overline{D}.$$



The λ^ns themselves are often called "barycentric coordinates", though they coincide with the previous λ^is only for the nodes around x. This abuse is harmless, but I'll stick to "nodal functions", notwithstanding.

A shorter way to describe them is to say: λ^n is the only *piecewise affine* function[14] that takes the value 1 at node n and 0 at all other nodes. The inset shows the pattern of level lines of λ^n in the 2D case (triangulation

[14] Meaning: affine by restriction to each tetrahedron. I will use "*mesh-wise*" in such cases: mesh-wise affine, mesh-wise quadratic, etc. (this is not standard terminology).

of a plane domain D). It is easy from this to imagine the graph of the corresponding function, and to understand why the λ^ns are often called "hat functions".

Exercise 3.8. Prove that the hat functions are linearly independent.

Exercise 3.9. Compute the average of λ^n over (1) an edge e, (2) a face f, (3) a tetrahedron T, all containing n.

Remark 3.4. Two things are essential in this construction: (1) each λ^n is supported on the cluster of n, (2) they form a *partition of unity* over D, i.e., $\sum_{n\in\mathcal{N}} \lambda^n = 1$, relation (13). The affine character is secondary, and is lost in case of curved tetrahedra.[15] But it considerably simplifies the programming, in conjunction with Remark 3.3, as we'll see. ◊

Well, that's all: *The finite element method is the Ritz–Galerkin method, the basis functions being a partition of unity associated with a mesh,* as above.

There are many ways to devise such a partition of unity, and the use of barycentric functions is only the simplest. When one refers to "a" finite element, it's this whole procedure one has in mind, not only the analytical expression of the basis functions. However, the latter suffices in many cases. Here, for instance, the restrictions of the λ^ns to individual tetrahedra are affine functions, that is, polynomials of maximum degree 1 of the Cartesian coordinates (one calls them "P^1 elements" for this reason), and this is enough characterization.[16]

Let us give another example, which demonstrates the power of this notation. What are "P^2 elements"? This means functions with small support, like the above λ^ns, which restrict to each tetrahedron as a second-degree polynomial, and therefore (Remark 3.3) are in the span of the products $\lambda^n\lambda^m$. This is enough to point to the partition of unity, for the set $\{\lambda^n\lambda^m:\ n\in\mathcal{N},\ m\in\mathcal{N}\}$ is perfect in this respect: we do have

$$\sum_{n,\,m\in\mathcal{N}} \lambda^n\lambda^m = \sum_{n\in\mathcal{N}}[\lambda^n(\sum_{m\in\mathcal{N}}\lambda^m)] = \sum_{n\in\mathcal{N}}\lambda^n = 1$$

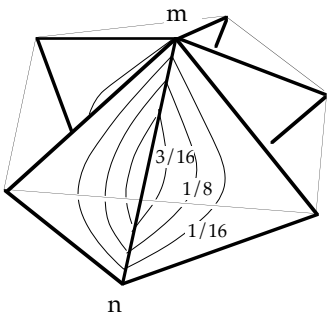
after (13), and the support of $\lambda^n\lambda^m$ is either the cluster of n, if n = m, or the cluster of edge n to m, if n and m are neighbors (the inset, next page,

[15]What is affine, then, is the "pull-back" of λ^n onto the reference tetrahedron. For this notion, push a little forward (Note 7.9).

[16]There is in finite element theory a traditional distinction between "basis functions", like the λ^n, and "shape functions", which are their restrictions to mesh volumes. As one sees here, shape functions are more simply characterized. Theory, on the other hand, is easier in terms of basis functions.

shows the level lines of $\lambda^n \lambda^m$). Note how the coefficients $\boldsymbol{\varphi}_{n\,m}$ in the expansion $\varphi = \sum_{n,\,m} \boldsymbol{\varphi}_{n\,m} \lambda^n \lambda^m$ are determined by the values of φ at the nodes and the mid-edges (Exer. 3.11).

Exercise 3.10. Compute the averages of $\lambda^n \lambda^m$ and $\lambda^n \lambda^m \lambda^\ell$ on a tetrahedron, in all cases, $n \neq m$, $n = m$, etc.



Exercise 3.11. Devise a set of P^2 functions w_{mn} such that $w_{mn} = 1$ at the middle of edge {m, n}, or at node n if $n = m$, and 0 at all other nodes and mid-edges.

Exercise 3.12 (Gaussian quadrature formulas). The average of an affine function over a tetrahedron is the average of its nodal values. The average of a *quadratic* function is a *weighted* average of its nodal and mid-edge values. Which weights? What about triangles?

Finite elements with degrees of freedom attached to specific points (cf. Note 10), like the P^1 and P^2 elements, are called *Lagrangian* [CR]. There are other varieties, built on hexahedra or other shapes, or with derivatives as DoFs (those are *Hermitian* elements), and so forth. Refer to specialized books such as [Ci]. There are also vector-valued finite elements, to which we shall return in Chapters 5 and 6.

3.3.3 The linear system

Generated by these basis elements, the finite dimensional subspace Φ_m contains all functions of the form

$$\varphi = \sum_{n \in \mathcal{N}} \boldsymbol{\varphi}_n \lambda^n. \tag{14}$$

There is one degree of freedom $\boldsymbol{\varphi}_n$ for each node n, equal to the value of φ at node n. The family $\boldsymbol{\varphi} = \{\boldsymbol{\varphi}_n : n \in \mathcal{N}\}$ can be construed as a vector of an N-dimensional space, where $N = \#\mathcal{N}$ is the number of nodes in the mesh. We shall denote this vector space by $\boldsymbol{\Phi}_m$ (and drop the m, which can be done without any risk of confusion while we are dealing with *one* mesh at a time). Of course Φ_m and $\boldsymbol{\Phi}$ are isomorphic, but they are objects of different kinds, and we shall keep the difference in mind. To stress it, let us call p_m the injective map from $\boldsymbol{\Phi}$ into Φ defined by (14), which sends $\boldsymbol{\varphi}$ to $\varphi_m = p_m(\boldsymbol{\varphi})$. Then, $\Phi_m = p_m(\boldsymbol{\Phi})$. Similar notation will be used throughout, with capitals for spaces, and boldface connoting degrees of freedom and

the vector spaces they span. In particular, we shall denote with bold parentheses the Euclidean scalar product of two elements of $\mathbf{\Phi}$, like this:

$$(15) \qquad (\boldsymbol{\varphi}, \boldsymbol{\varphi}') = \sum_{n \in \mathcal{N}} \varphi_n \varphi'_n.$$

To introduce Φ^I_m, first call $\mathcal{N}(S^h)$ the set of all boundary nodes that belong to S^h, including those on the frontier between S^h and S^b. Formally, $\mathcal{N}(S^h) = \{n \in \mathcal{N} : x_n \in \mathrm{cl}(S^h)\}$, where cl stands for the closure relative to S. Let $\mathcal{N}(S^h_0)$ and $\mathcal{N}(S^h_1)$ similarly be defined. Then, define

$$(16) \qquad \mathbf{\Phi}^I = \{\boldsymbol{\varphi} \in \mathbf{\Phi} : \varphi_n = 0 \text{ if } n \in \mathcal{N}(S^h_0),\ \varphi_n = I \text{ if } n \in \mathcal{N}(S^h_1)\}$$

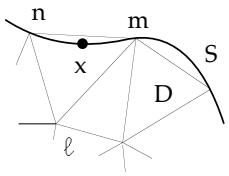
and, similarly, $\mathbf{\Phi}^0$, two parallel subspaces of $\mathbf{\Phi}$. Finally, let us set

$$(17) \qquad \Phi^0_m = p_m(\mathbf{\Phi}^0), \quad \Phi^I_m = p_m(\mathbf{\Phi}^I).$$

Relation (10), $\Phi^I_m = \varphi^I_m + \Phi^0_m$, has a counterpart here. Let us construct $\boldsymbol{\varphi}^1$, a special vector, with all components $\varphi^1_n = 0$ except for $n \in \mathcal{N}(S^h_1)$, where they are set to 1. Then, with $\boldsymbol{\varphi}^I$ defined as $\boldsymbol{\varphi}^I = I\,\boldsymbol{\varphi}^1$,

$$(18) \qquad \mathbf{\Phi}^I = \boldsymbol{\varphi}^I + \mathbf{\Phi}^0.$$

Remark 3.5. If you try to check (7) at this stage, you will see that it fails if the faces at the boundary do not fit it exactly. Cf. the inset: a piecewise affine function that vanishes at n and m, but not at ℓ, cannot be zero at x. Because of this tiny difference, Φ^I_m is not contained in Φ^I, and applying the geometrical reasonings suggested by Fig. 3.4 would be a "variational crime", in the sense of Strang and Fix [SF]. This (jocular) charge should not deter anyone from using a mesh similar to the one in inset in case of a curved boundary. This is perfectly right! What is not, and would constitute the crime, would be to apply the simple convergence proof that will follow to such a situation, which calls for more cumbersome treatment. Thanks to our decision to deform D into a polyhedron before meshing, we do have $\Phi^I_m = \Phi_m \cap \Phi^I$ and $\Phi^0_m = \Phi_m \cap \Phi^0$, as announced in (7). But this will not be effectively used before we address convergence and error analysis, and what immediately follows does not depend on the truth of these assertions. ◊



We want now to interpret problem (9), that is,

$$(9') \qquad find\ \varphi_m \in \Phi^I_m\ such\ that \quad \int_D \mu\ \mathrm{grad}\,\varphi_m \cdot \mathrm{grad}\,\varphi' = 0 \quad \forall\, \varphi' \in \Phi^0_m,$$

in algebraic terms. Since $\Phi^I_m = p_m(\mathbf{\Phi}^I)$, this is a linear system with respect

to the "free" degrees of freedom, that is, those not constrained by (16), which are the nodes of the subset $\mathcal{N}_0 = \mathcal{N} - \mathcal{N}(S^h)$. On the other hand, since $\Phi^0{}_m$ is parallel to $\Phi^I{}_m$, there are as many equations as unknowns in (9'). Our aim is to rewrite (9') in terms of the degrees of freedom. For this, let us set, for any two nodes n and m,

(19) $$\mathbf{M}_{n\,m} = \int_D \mu \, \mathrm{grad}\, \lambda^n \cdot \mathrm{grad}\, \lambda^m,$$

and form the symmetric matrix $\mathbf{M}$, indexed on $\mathcal{N} \times \mathcal{N}$, of which this is the entry at row n and column m. Then (just write φ_m and φ' as in (14), and expand), (9') is equivalent to

(9'') $$\textit{find } \boldsymbol{\varphi} \in \boldsymbol{\Phi}^I{}_m \textit{ such that } \quad (\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}') = 0 \quad \forall\, \boldsymbol{\varphi}' \in \boldsymbol{\Phi}^0,$$

via the correspondence $\varphi_m = p_m(\boldsymbol{\varphi})$. As a matter of course (we did that twice already), this is equivalent to the variational problem

(9''') $$\textit{find } \boldsymbol{\varphi} \in \boldsymbol{\Phi}^I \textit{ such that } \quad \mathbf{F}(\boldsymbol{\varphi}) \le \mathbf{F}(\boldsymbol{\psi}) \quad \forall\, \boldsymbol{\psi} \in \boldsymbol{\Phi}^I,$$

where $\mathbf{F}(\boldsymbol{\varphi}) = \frac{1}{2}(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi})$.

$\mathbf{M}$ is traditionally dubbed the *stiffness matrix* of the problem, because of the origins of the finite elements method: In mechanics, the analogue of our $\boldsymbol{\varphi}$ is most often a displacement vector, and $\mathbf{F}(\boldsymbol{\varphi})$ is deformation energy, so $\mathbf{M}\boldsymbol{\varphi}$ is a force vector, and a force-to-displacement ratio is a stiffness. (One could make a case for *admittance* matrix, in our context.)

As a last step, let us write $\mathbf{M}$ in block form, by partitioning the indexing set $\mathcal{N}$ as[17] $\mathcal{N} = \mathcal{N}_0 + \mathcal{N}(S^h)$. With ad-hoc but obvious notation,

$$\mathbf{M} = \begin{vmatrix} {}^{00}\mathbf{M} & {}^{01}\mathbf{M} \\ {}^{10}\mathbf{M} & {}^{11}\mathbf{M} \end{vmatrix},$$

where the submatrix ${}^{00}\mathbf{M}$ is indexed over $\mathcal{N}_0$ and thus operates in the subspace $\boldsymbol{\Phi}^0$ of genuine unknowns (those not constrained by essential boundary conditions). We also write vectors in block form, $\boldsymbol{\varphi} = \{{}^0\boldsymbol{\varphi}, {}^1\boldsymbol{\varphi}\}$, and $\boldsymbol{\varphi}^I = \{\mathbf{0}, {}^1\boldsymbol{\varphi}^I\}$. Thanks to this and to (18), we see that (9'') is equivalent to *find* ${}^0\boldsymbol{\varphi} \in \boldsymbol{\Phi}^0$ *such that*

(20) $${}^{00}\mathbf{M}\,{}^0\boldsymbol{\varphi} = -\,{}^{01}\mathbf{M}\,{}^1\boldsymbol{\varphi}^I,$$

at last a standard linear system, since the right-hand side $-\,{}^{01}\mathbf{M}\,{}^1\boldsymbol{\varphi}^I$ is known.

[17] The union sign $\cup$ is replaced by $+$ when sets are disjoint, as here.

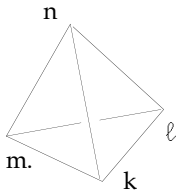
3.3.4 "Assembly", matrix properties

Methods to effectively *solve* this linear system are beyond our scope. One may refer to many excellent handbooks for this, among which are [Cr, GL, Gv, Va]. The choice of methods, however, strongly depends on the structure and properties of ${}^{00}\mathbf{M}$, so we need a few indications on this. And of course, we must address the practical problem of computing the entries (19).

Let us say from the outset that it's not a good idea to concentrate on the "useful" matrix ${}^{00}\mathbf{M}$ of (20), thus forgetting about $\mathbf{M}$, for two reasons. First, the properties of $\mathbf{M}$ are simpler to discover, and those of its *principal* submatrices (i.e., diagonal sub-blocks), like ${}^{00}\mathbf{M}$, easily follow. Next, the boundary conditions one wishes to consider may change during the study of a given problem, thus changing the set $\mathcal{N}_0$. Finally, as we shall see in the next chapter, some data one wishes to access require the knowledge of all $\mathbf{M}$.

The first concern is for the computation of the entries of $\mathbf{M}$. With ∇ standing for grad for shortness, let us define (cf. (19))

$$\mathbf{M}^{\mathbf{T}}_{nm} = \int_T \mu \, \nabla\lambda^n \cdot \nabla\lambda^m,$$



so that $\mathbf{M}_{nm} = \sum_{T \in \mathcal{T}} \mathbf{M}^{\mathbf{T}}_{nm}$. If one replaces μ by its average $\mu(T)$ over the tetrahedron, this is easy to compute, as we now show.

Call $\{k, \ell, m, n\}$ the vertices of T, and suppose for definiteness they are placed as shown in inset, vectors[18] $k\ell$, km, kn forming a positively oriented frame. Notice that $|k\ell \times km|/2$ is the area of face $\{k, \ell, m\}$ and $1/|\nabla\lambda^n|$ the height of the tetrahedron above that face, which results in $\nabla\lambda^n = (k\ell \times km)/(6\ \mathrm{vol}(T))$. Now,

$$(21) \qquad \int_T \nabla\lambda^n \cdot \nabla\lambda^m = \frac{1}{36\ \mathrm{vol}(T)}\ (k\ell \times km) \cdot (\ell n \times \ell k)$$

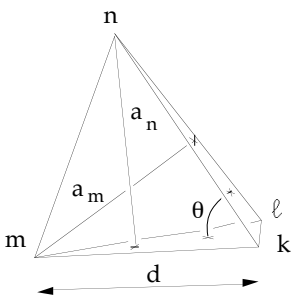
$$= \frac{1}{36\ \mathrm{vol}(T)} [\ (k\ell \cdot \ell n)\ (\ell k \cdot km) + |k\ell|^2\ km \cdot \ell n]$$

by a well-known formula[19] (which shows, incidentally, that the result is insensitive to the orientation of T). There is another expression for (21), known as the *cotangent formula*, which gives useful insight. The dot product of $\nabla\lambda^n$ and $\nabla\lambda^m$ is $-(1/a_m)(1/a_n)\cos\theta$, with the notation given

[18] A symbol like km denotes the vector from point x_k (the location of node k) to point x_m. This is a gross abuse of notation, but an innocuous one.

[19] $(a \times b) \cdot (c \times d) = (a \cdot c)\ (b \cdot d) - (a \cdot d)\ (b \cdot c)$.

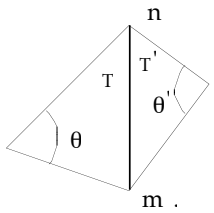
in inset (a for "altitude"), and $\mathrm{vol}(T) = a_n d\,|k\ell|/6$. But $a_m = d \sin\theta$, where d is the distance from m to the line supporting $k\ell$, and thus,



$$\begin{aligned}\int_T \nabla\lambda^n \cdot \nabla\lambda^m &= -\,a_n\,|km|\,|k\ell|\cos\theta/(6\,a_m a_n)\\ &= -\,\tfrac{1}{6}\,|k\ell|\cot\theta.\end{aligned}$$

Adding all contributions, we thus have, with ad-hoc and obvious notation,

$$\mathbf{M}_{nm} = \int_D \mu\,\nabla\lambda^n \cdot \nabla\lambda^m = -\,\tfrac{1}{6}\sum_{T\in\mathcal{T}(\{n,m\})} \mu(T)\,|k\ell|_T \cot\theta_T\,. \tag{22}$$



In dimension 2 (notation in inset), this reduces to

$$\mathbf{M}_{nm} = -\,\tfrac{1}{2}\,(\mu(T)\cot\theta + \mu(T')\cot\theta'). \tag{23}$$

Remark 3.6. Diagonal entries $\mathbf{M}_{nn}$ cannot be obtained by this formula, but since $\nabla\lambda^n = -\sum_{m\neq n}\nabla\lambda^m$, one has $\sum_m \mathbf{M}_{nm} = \sum_m \int_D \mu\ \mathrm{grad}\,\lambda^n \cdot \mathrm{grad}\,\lambda^m = \int_D \mu\ \mathrm{grad}\,\lambda^n \cdot \mathrm{grad}(\sum_m \lambda^m) = 0$, hence $\mathbf{M}_{nn} = -\sum_{m\neq n}\mathbf{M}_{nm}$, and also $\mathbf{M}^{\mathbf{T}}{}_{nn} = -\sum_{m\neq n}\mathbf{M}^{\mathbf{T}}{}_{nm}$ by summing over T instead of D. So (22) is enough. ◊

Since the data structure gives access to the node locations, and hence to the components of the edge vectors, the simplest programming is via (21), which requires no more than coding a handful of determinants (the volume itself is $|\det(k\ell, km, kn)|/6$). This will be the basic subroutine for the assembly program. Running it for all pairs of nodes gives the "elementary matrix" $\mathbf{M}^{\mathbf{T}}$ and then $\mathbf{M} = \sum_{T\in\mathcal{T}} \mathbf{M}^{\mathbf{T}}$ by looping over the tetrahedra. This is the *assembly* process, by which the matrix is constructed from the mesh data structure.

This way, only terms which do contribute to the matrix are evaluated. A priori, of course, $\mathbf{M}_{nm} = 0$ for pairs of nodes which are not linked by a common edge, that is, most of them: $\mathbf{M}$ is *sparse*, that is to say, has a small percentage of nonzero entries. This has consequences, also, on the way these entries are stored (the precise coding of the assembly depends on options taken at this level) and on the algorithms for solving the linear system [BR, GL].

This sparsity is perhaps the most important property of finite-element matrices. (The Galerkin method generates full matrices, unless the supports

of the basis functions are small, which is precisely what finite elements achieve.) Other properties we now list are not specific to finite elements, but depend on the "partition of unity" feature (the equality $\sum_{n \in \mathcal{N}} \lambda^n = 1$). For shortness, $\mathbf{1}$ will denote the vector of $\mathbf{\Phi}$ all components of which are equal to 1, and $\vee\{\boldsymbol{\varphi}, \boldsymbol{\psi}, \ldots\}$ the *span* of a family of vectors $\boldsymbol{\varphi}$, $\boldsymbol{\psi}$, ... (cf. A.2.2).

Proposition 3.3. $\mathbf{M}$ *is symmetric, and nonnegative definite, that is* (cf. Section B.1),

$$(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}) \geq 0 \quad \forall\, \boldsymbol{\varphi}.$$

Proof. $(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}) = \|p(\boldsymbol{\varphi})\|_\mu^2 \equiv \int_D \mu \, |\operatorname{grad} p(\boldsymbol{\varphi})|^2 \geq 0$. ◊

Proposition 3.4. $\ker(\mathbf{M}) = \vee\{\mathbf{1}\}$.

Proof. We already know that $\mathbf{M1} = 0$ (Remark 3.6). Conversely, if $\mathbf{M}\boldsymbol{\varphi} = 0$, then $(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}) = 0$, hence grad $p(\boldsymbol{\varphi}) = 0$, hence $p(\boldsymbol{\varphi}) = c$, a constant, and $\sum_n (\boldsymbol{\varphi}_n - c)\, \lambda^n = 0$, hence $\boldsymbol{\varphi} = c\,\mathbf{1}$, if the λ^ns are independent, which we know is the case for hat functions, by Exer. 3.8. ◊

Exercise 3.13. If $\mathbf{M}$ is symmetric and nonnegative definite, show that $\ker(\mathbf{M})$, which is defined as $\{\boldsymbol{\varphi} : \mathbf{M}\boldsymbol{\varphi} = 0\}$, is equal to $\{\boldsymbol{\varphi} : (\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}) = 0\}$.

Proposition 3.5. *Apart from* $\mathbf{M}$ *itself, all principal submatrices of* $\mathbf{M}$ *are positive definite* (cf. Section B.1)*, and hence regular.*

Proof. Let $\mathcal{N}_0$ be a part of $\mathcal{N}$, and consider a block partitioning of $\mathbf{M}$ on the basis of the $\mathcal{N} = \mathcal{N}_0 + (\mathcal{N} - \mathcal{N}_0)$ partitioning of the node set. To avoid vertical displays, let us write this $\mathbf{M} = \{\{{}^{00}\mathbf{M}, {}^{01}\mathbf{M}\}, \{{}^{10}\mathbf{M}, {}^{11}\mathbf{M}\}\}$, by rows of blocks, according to the standard convention. Then ${}^{00}\mathbf{M}$ is (by definition) a *principal* submatrix of $\mathbf{M}$. Suppose there is a vector ${}^0\boldsymbol{\varphi}$, supported on $\mathcal{N}_0$, such that $({}^{00}\mathbf{M}\, {}^0\boldsymbol{\varphi}, {}^0\boldsymbol{\varphi}) = 0$, and build from it a vector $\boldsymbol{\varphi}$ supported on all $\mathcal{N}$ by attributing the value 0 to all DoFs in $\mathcal{N} - \mathcal{N}_0$. Then $(\mathbf{M}\boldsymbol{\varphi}, \boldsymbol{\varphi}) = 0$, which we know implies $\boldsymbol{\varphi} = c\,\mathbf{1}$. But if $\mathcal{N}_0 \neq \mathcal{N}$, then some components of $\boldsymbol{\varphi}$ vanish, hence $c = 0$, and ${}^0\boldsymbol{\varphi} = \mathbf{0}$. Then, by the result of Exercise 3.13, ${}^{00}\mathbf{M}$ is regular. ◊

A particular case is when $\mathcal{N}_0$ reduces to *one* node n, showing that $\mathbf{M}_{nn} > 0$. So all diagonal coefficients of $\mathbf{M}$ are positive, and the sum of entries of a same row, or column, is zero, by Prop. 3.4.

Now, a property which is more closely linked with the use of barycentric functions. Formula (22) shows that in case of acute dihedral angles, all off-diagonal entries are nonpositive. Symmetric positive definite matrices with ≤ 0 off-diagonal coefficients are called *Stieltjes matrices* [Va] and are important because of the following property:

Proposition 3.6. *If* $\mathbf{A}$ *is a Stieltjes matrix, all entries of its inverse are nonnegative.*[20]

Proof. Let's agree to write $\mathbf{v} \geq 0$ if no component of vector $\mathbf{v}$ is negative. Let $\mathbf{u}$ be the solution of the linear system $\mathbf{A}\,\mathbf{u} = \mathbf{b}$, and suppose $\mathbf{b} \geq 0$. Write $\mathbf{u}$ in the form $\mathbf{u} = \mathbf{u}^+ - \mathbf{u}^-$, with $\mathbf{u}^+ \geq 0$ and $\mathbf{u}^- \geq 0$, and remark that $\mathbf{u}^+_n \mathbf{u}^-_n = 0$ for all $n \in \mathcal{N}_0$ (if we continue to call $\mathcal{N}_0$ the indexing set of $\mathbf{S}$, $\mathbf{u}$, and $\mathbf{b}$). Now,

$$0 \leq (\mathbf{b}, \mathbf{u}^-) = (\mathbf{A}(\mathbf{u}^+ - \mathbf{u}^-), \mathbf{u}^-) = (\mathbf{A}\mathbf{u}^+, \mathbf{u}^-) - (\mathbf{A}\mathbf{u}^-, \mathbf{u}^-).$$

But $(\mathbf{A}\mathbf{u}^+, \mathbf{u}^-) \leq 0$, because only off-diagonal entries of $\mathbf{A}$ contribute to this scalar product, so $(\mathbf{A}\mathbf{u}^-, \mathbf{u}^-) = 0$, hence $\mathbf{u}^- = 0$. Thus $\mathbf{A}^{-1}\mathbf{b}$ has no negative components if $\mathbf{b}$ has none, hence the result. ◊

This applies to our system matrix ${}^{00}\mathbf{M}$ in the case where no dihedral angle is obtuse, with interesting consequences that we shall discover in the next chapter.

EXERCISES

Texts for Exercises 3.1 to 3.4 are on pp. 64 and 66.

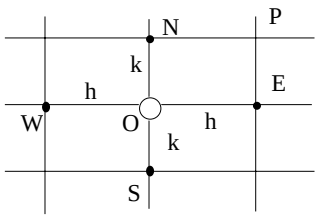
Exercise 3.5. Reinforce your knowledge of the following facts: The real and imaginary parts of a function which is holomorphic in some domain of the complex plane are harmonic. Conformal transformations preserve harmonicity.

Exercise 3.6. In the plane $\{x, y\}$, find a function φ which is harmonic in the domain $\{\{x, y\} : x < 0 \text{ or } y < 0\}$ and null on the axes $y = 0$ and $x = 0$. Take its restriction to the domain D obtained by clipping the regions $y \leq -1$ and $x \leq -1$. Examine the singularity of φ at 0. Is $\nabla\varphi$ square-integrable in D ? Show that this is the idealization of a situation which can happen physically.

Exercise 3.7 is on p. 74. Exers. 3.8 to 3.12 are on pp. 77 to 79, and Exer. 3.13 on p. 83.

[20]Matrices with this property are called "monotone". (They are akin to "M-matrices" [BP, Jo, Na]. Beware the terminological confusion around this concept.) Notice that at least one term on each row of the inverse must be positive.

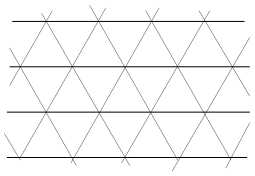
Exercise 3.14. In the *finite differences* method, potential values at the nodes of a so-called "orthogonal grid" are the unknowns, and equations are obtained via local Taylor expansions of the unknown potential [Va]. For instance (inset), if φ must satisfy $-\Delta\varphi = 0$, the values of φ at a node O and at neighboring nodes E, N, W, S, will approximately satisfy



$$\varphi_O = [k^2(\varphi_E + \varphi_W) + h^2(\varphi_N + \varphi_S)] / 2(k^2 + h^2). \tag{24}$$

There is one equation of this kind for each node like O, and altogether they form a linear system similar to (20), when one takes into account boundary nodal values. Describe this method as a special case of the finite element method.

Exercise 3.15. The method of finite differences does not adapt easily to domains with complicated boundaries, and the finite element method has a decisive advantage in this respect. However, it's intuitive that inside physically homogeneous regions (constant coefficients), one should use meshes as "regular" (that is, uniform and repetitive, crystal-like) as one can devise. For instance, in 2D, equilateral triangles (inset) are a good idea. As far as tetrahedral elements are concerned, do we have the equivalent of this in 3D? Can one pave space with regular tetrahedra?



Exercise 3.16. The next best thing to a regular tetrahedron is an *isosceles* tetrahedron, one for which opposite edges are equal, two by two [Co]. Find an isosceles tetrahedron that will pave.

Exercise 3.17. One expects the reluctance of a circuit to decrease when the permeability increases anywhere inside. Show that, indeed, if $\mu_2 \geq \mu_1$ a.e. in D in our model problem, the corresponding reluctances satisfy $R_2 \leq R_1$.

Exercise 3.18. One expects the reluctance to decrease when all the dimensions of the device increase proportionally. Prove it.

Exercise 3.19. Study the continuity of φ with respect to μ in the model problem.

Exercise 3.20 (research project). Study the continuity of φ with respect to the shape of the domain.

HINTS

3.1. The question amounts to finding a quadratic functional whose directional derivative at point φ would be

$$\varphi' \to \int_D \mu_0 \, \mathrm{grad}\, \varphi \cdot \mathrm{grad}\, \varphi' + \mu_0 \int_D \mathrm{m} \cdot \mathrm{grad}\, \varphi',$$

and we know the answer when $\mathrm{m} = 0$. What remains to be found is a function of φ (obviously, a linear function, since m does not depend on φ) having $\varphi' \to \mu_0 \int_D \mathrm{m} \cdot \mathrm{grad}\, \varphi'$ as its directional derivative.

3.2. On Φ, $\| \ \|_\mu$ is not a norm: Properties $\|\lambda \varphi\|_\mu = |\lambda| \, \|\varphi\|_\mu$, $\|\varphi + \psi\|_\mu \le \|\varphi\|_\mu + \|\psi\|_\mu$ and $\|\varphi\|_\mu \ge 0$ do hold, but $\|\varphi\|_\mu = 0$ does not entail $\varphi = 0$. We have only a *semi-norm* there. How can that be cured?

3.3. The goal is to find a constant C such that $|\mathcal{J}(\psi)| \le C \, \|\psi\|_\mu$, where ψ is any member of Φ^*, not one that satisfies (1) necessarily. But on the other hand, the solution of (1) which has the same mmf as ψ is a good reference, after Fig. 3.2 and the proof of Prop. 3.1.

3.6. This is a simple exercise in conformal transformations. First find a harmonic function in a half-plane that vanishes on the boundary, then map the half-plane onto the desired region.

3.7. Call "small diagonals" and "large diagonals" the segments joining two vertices, depending on whether they belong to the cube's surface or not. Show that at most one large diagonal can exist in the mesh. If there is one, show that at least three inner faces must have it as an edge. (Beware, it's a challenging exercise.)

3.8. Look at the nodal values of $\sum_n \boldsymbol{\alpha}_n \lambda^n$.

3.9. In particular, $(\int_T \lambda^n)/\mathrm{vol}(T) = 1/4$, where vol denotes the volume, and the general case is, obviously, $(\int_s \lambda^n)/\mathrm{meas}(s) = 1/(p+1)$ for a simplex s of dimension p, where meas for "measure" stands for length, area, etc. To say "all sums $\int_s \lambda^n$ for $n \in \mathcal{N}(s)$ are equal, and they add to meas(s), by (13)" is a fine symmetry argument, but why this equality? It stems from general results on change of variables in integration—but rather try a pedestrian and straightforward "calculus proof".

3.10. Probably the simplest way is to use the calculus proof to compute $\int_s (\lambda^n)^2$, then the symmetry argument for $m \ne n$.

3.11. Up to the factor 4, $\lambda^n \lambda^m$ is right. As for $\lambda^n \lambda^n$, look at its behavior along a typical edge {n, m}, and rectify at mid-edge.

3.12. Combine Exers. 3.10 and 3.11.

3.13. The same trick as in Prop. 3.1.

3.14. Grid cells must be cut in two, so if point P for instance is linked with O by an edge, making them "neighbors" on the finite element mesh, one expects a nonzero entry in the stiffness matrix at row O and column P, which is *not* the case of the finite-difference scheme (24). Explaining why this term vanishes is the key. It has to do with the right angle, obviously.

3.15. If paving was possible, tetrahedra around a given edge would join without leaving any gap, so the dihedral angle would have to be $2\pi/n$ for some integer n. Is that so?

3.17. Suppose $I = 1$ for simplicity. Then $R_1^{-1} = \inf(\int_D \mu_1 \mid \nabla\varphi \mid^2 : \varphi \in \Phi^1\}$. Replace φ by φ_2, the solution for $\mu = \mu_2$.

3.18. Map the problem concerning the enlarged region onto the reference one, and see how this affects μ.

3.19. Consider two problems corresponding to permeabilities μ_1 and μ_2, all other things being equal. Denote the respective solutions by φ_1 and φ_2. Let $\|\varphi\|_1$ or $\|\varphi\|_2$ and $(\varphi, \varphi')_1$ or $(\varphi, \varphi')_2$ stand for $\|\varphi\|_\mu$ and $(\varphi, \varphi')_\mu$, depending on the value of μ. One has

$$\int_D \mu_i \operatorname{grad} \varphi_i \cdot \operatorname{grad} \varphi' = 0 \quad \forall\ \varphi' \in \Phi^0, \text{ for } i = 1, 2. \tag{25}$$

Choose appropriate test functions, combine both equations, and apply the Cauchy–Schwarz inequality.

3.20. If the deformation is a homeomorphism, the same mapping trick as in Exer. 3.18 reduces the problem to analyzing the dependence with respect to μ, with a new twist, however, for μ will become a tensor. You will have to work out a theory to cover this case first.

SOLUTIONS

3.1. Let $\mathcal{M}(\varphi) = \int_D m \cdot \operatorname{grad} \varphi$. Since $\mathcal{M}(\varphi + \lambda\varphi') = \mathcal{M}(\varphi) + \lambda\mathcal{M}(\varphi')$, the directional derivative of $\mathcal{M}$ is $\varphi' \to \lim_{\lambda \to 0}(\mathcal{M}(\varphi + \lambda\varphi') - \mathcal{M}(\varphi))/\lambda$, that is, $\varphi' \to \int_D m \cdot \operatorname{grad} \varphi'$, the same formally[21] as $\mathcal{M}$ itself. This holds for

all linear functionals, so we shall not have to do it again. The variational forms of (2.34) and (2.36) thus consist in minimizing the functionals

$$\varphi \to \tfrac{1}{2}\int_D \mu_0 \,|\,\mathrm{grad}\,\varphi\,|^2 + \mu_0 \int_D \mathrm{m}\cdot\mathrm{grad}\,\varphi \quad \text{on } \Phi^{\mathrm{I}},$$

$$\varphi \to \tfrac{1}{2}\int_D \mu_0 \,|\,\mathrm{grad}\,\varphi\,|^2 - \mathrm{F}\,\mathcal{J}(\varphi) \quad \text{on } \Phi^{*},$$

respectively.

3.2. On Φ, $\|\varphi\|_\mu = 0$ implies a constant value of φ, but not $\varphi = 0$, so $\|\ \|_\mu$ is not a norm, whereas its restriction to Φ^* is one. This hardly matters, anyway, since two potentials which differ by an additive constant have the same physical meaning. So another possibility would be for us to define the quotient $\Phi/\mathbb{R}$ of Φ by the constants, call that $\dot{\Phi}$, and give it the norm $\|\dot{\varphi}\|_\mu = \inf\{c \in \mathbb{R} :\ \|\varphi + c\|_\mu\}$, where φ is a member of the class $\dot{\varphi} \in \dot{\Phi}$. Much trouble, I'd say, for little advantage, at least for the time being. Later, we'll see that what happens here is a general fact, which has to do with *gauging*: It's *equivalence classes* of potentials, not potentials themselves, that are physically meaningful, so this passage to the quotient I have been dodging here will have to be confronted.

3.3. Take $\psi \in \Phi^*$, and let $\mathrm{I} = \mathcal{J}(\psi)$. Then (Fig. 3.2) $\|\varphi(\mathrm{I})\|_\mu \le \|\psi\|_\mu$. Using Prop. 3.2, we thus have

$$|\mathcal{J}(\psi)| = |\mathrm{I}| = [\|\varphi(1)\|_\mu]^{-1}\,\|\varphi(\mathrm{I})\|_\mu \le [\|\varphi(1)\|_\mu]^{-1}\,\|\psi\|_\mu.$$

3.4. If $\mathrm{m} > \mathrm{n}$, $\int |\mathrm{f_n} - \mathrm{f_m}| = \int_{[1/\mathrm{n},\,1/\mathrm{m}]} \mathrm{dx}/\sqrt{\mathrm{x}} = 2/\sqrt{\mathrm{n}} - 2/\sqrt{\mathrm{m}} < 2/\sqrt{\mathrm{n}}$ tends to zero.

3.5. Let $\mathrm{f}(\mathrm{z}) = \mathrm{P}(\mathrm{x}, \mathrm{y}) + \mathrm{i}\,\mathrm{Q}(\mathrm{x}, \mathrm{y})$. *Holomorphy* of f inside D means differentiability in the *complex* field $\mathbb{C}$, that is, for all $\mathrm{z} \in \mathrm{D}$, existence of a complex number $\partial\mathrm{f}(\mathrm{z})$ such that $\mathrm{f}(\mathrm{z} + \mathrm{dz}) = \mathrm{f}(\mathrm{z}) + \partial\mathrm{f}(\mathrm{z})\,\mathrm{dz} + \mathrm{o}(\mathrm{dz})$ for all dz in $\mathbb{C}$. Cauchy conditions for holomorphy are $\partial_\mathrm{x}\mathrm{P} = \partial_\mathrm{y}\mathrm{Q}$ and $\partial_\mathrm{y}\mathrm{P} = -\partial_\mathrm{x}\mathrm{Q}$, so $\partial_{\mathrm{xx}}\mathrm{P} = \partial_{\mathrm{xy}}\mathrm{Q} = \partial_{\mathrm{yx}}\mathrm{Q} = -\partial_{\mathrm{yy}}\mathrm{P}$, hence $\Delta\mathrm{P} = 0$, and the same for Q. In dimension 2, *conformal mappings* (those which preserve angles, but not distances) are realized by holomorphic maps from $\mathbb{C}$ to $\mathbb{C}$, and holomorphy is preserved by composition.

3.6. A harmonic function in the upper half-plane $\mathrm{y} > 0$ which vanishes for $\mathrm{y} = 0$ is $\{\mathrm{x}, \mathrm{y}\} \to \mathrm{y}$, the function denoted Im (for imaginary part). The *fan map* $\mathrm{g} = \mathrm{z} \to \mathrm{i}\,\mathrm{z}^{3/2}$ sends the upper half-plane to the domain

[21]But not conceptually. The argument of $\mathcal{M}$ is a point in an *affine* space, whereas φ', in the expression of the directional derivative, is an element of the associated *vector* space.

$\{\{x, y\} : x < 0$ **or** $y < 0\}$. Composition of Im and g^{-1} yields the desired function (cf. Fig. 3.6), better expressed in polar coordinates:

$$\varphi(r, \theta) = r^{2/3} \sin((2\theta - \pi)/3).$$

(Note that φ cannot be extended to the whole plane. Note also that it is not piecewise k-smooth for $k > 0$, in the sense we adopted in Chapter 2.) Its gradient is infinite at the origin, where its modulus behaves like $r^{-1/3}$. Since $\int_0^R (r^{-1/3})^2 \, r \, dr = \int_0^R r^{1/3} \, dr$ converges, this is a potential with finite associated magnetic (co)energy.

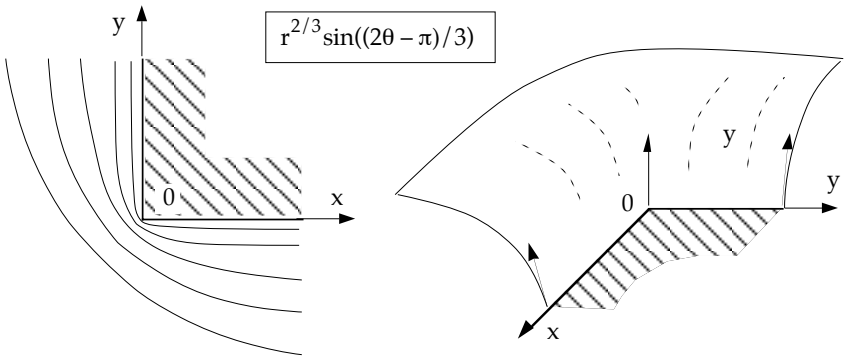


FIGURE 3.6. The function $f(r, \theta) = r^{2/3} \sin((2\theta - \pi)/3)$ of plane polar coordinates, for $\pi/2 \le \theta \le 2\pi$. Left: level lines. Right: perspective view of the graph.

Now imagine one of the level surfaces of φ (a cylinder along 0z) is lined up by some perfectly permeable material. This potential is then the solution of a two-dimensional analogue[22] of our model problem, in which the system would be infinite in the z–direction.

3.7. Two large diagonals would cut at the center, so there can't be more than one. At least three faces must hinge on it, since dihedral angles are less than π, and at most six, corresponding to the six possible vertices. So, see how to leave out one, two, or three of these. Fig. 3.7 gives the result. Alternatively, one may consider whether opposite faces are cut by parallel

[22]A genuinely three-dimensional example would be more demonstrative. See [Gr] for the (more difficult) techniques by which such examples can be constructed.

or anti-parallel small diagonals. (This is meaningful when one thinks of stacking cubes in order to make a tetrahedral mesh.) Whichever way, it's pretty difficult to prove the enumeration complete!

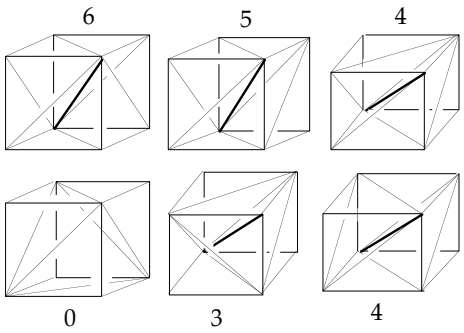


FIGURE 3.7. All ways to mesh a cube, depending on the number of inner faces that have a large diagonal as one of their edges.

3.8. $\sum_n \boldsymbol{\alpha}_n \lambda^n(x_m) = \boldsymbol{\alpha}_m$, by construction of the nodal functions, so if $\sum_n \boldsymbol{\alpha}_n \lambda^n = 0$, then all $\boldsymbol{\alpha}_n$s vanish.

3.9. One may invoke the general result about "change of variables" in integration, $\int_{u(D)} f J_u = \int_D u^* f$, where $u^* f$ is the pull-back $x \to f(u(x))$ and J_u the Jacobian of the mapping u, for there is an affine map from T to itself that swaps n and m, λ^n and λ^m, which is volume preserving ($J_u = 1$). A much more elementary but safer alternative is, in Cartesian coordinates: Place the basis of tetrahedron T in plane x–y, and let n be the off-plane node, at height h. Then $\lambda^n(x, y, z) = z/h$. If A is the area of the basis, then $\int_T \lambda^n = A \int_0^h dz\, (1 - z/h)^2\, z/h = hA/12$, hence $\int_T \lambda^n = \mathrm{vol}(T)/4$. Same thing for a triangle: basis on x axis, height h, etc. You may prefer a proof by recurrence on the dimension. Anyway, once in possession of these basic symmetry results, further computations (cf. Exer. 3.10) simplify considerably.

3.10. First compute $I_{nn} = \int_T (\lambda^n)^2 = A/h^2 \int_0^h dz\, (h - z)^2\, z^2/h^2 = \mathrm{vol}(T)/10$, and similarly, obtain the equality $I_{nnn} = \int_T (\lambda^n)^3 = \mathrm{vol}(T)/20$. Then $I_{nn} = \int_T \lambda^n (1 - \sum_{m \neq n} \lambda^m) = \mathrm{vol}(T)/4 - 3\, I_{nn}$), hence $I_{nn} = \mathrm{vol}(T)/20$. And so on. The general formula,

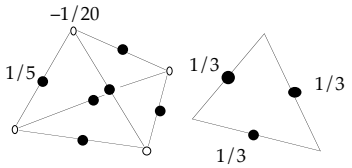
$$\int_T (\lambda^\ell)^i\, (\lambda^m)^j\, (\lambda^n)^k = 6\ \mathrm{vol}(T)\ i!\ j!\ k!\ /(i+j+k+3)!$$

(cf. [Sf]), may save you time someday. (Thanks to (13) and Remark 3.3,

this is enough to sum any polynomial over T.) The analogue on faces [SF] is $\int_f (\lambda^m)^i (\lambda^n)^j = 2\,\mathrm{area}(f)\, i!\, j!/(i+j+2)!$.

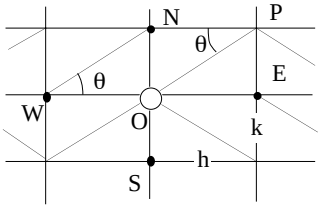
3.11. On $[0, 1]$, the function $w = x \to 2x^2 - x$ behaves as requested, that is, $w(0) = w(1/2) = 0$ and $w(1) = 1$. Therefore, $w_{nm} = 4\,\lambda^n\lambda^m$ and $w_{nn} = \lambda^n(2\lambda^n - 1)$.

3.12. Weights $1/5$ at mid-edges and $-1/20$ at nodes. For the triangle, amusingly, $1/3$ at mid-edges and 0 at nodes (inset).



3.13. For all $\boldsymbol{\psi}$, $0 \le \mathbf{(M(\varphi + \lambda\psi),\ \varphi + \lambda\psi)} = 2\lambda\,\mathbf{(M\varphi,\ \psi)} + \lambda^2\,\mathbf{(M\psi,\ \psi)}$, so $\mathbf{(M\varphi,\ \psi)} = 0$ for all $\boldsymbol{\psi}$, which implies $\mathbf{M\varphi} = 0$. ◊

3.14. Use the cotangent formula (23). There are two right angles in front of edge OP, hence the nullity of the coefficient A_{OP}. Formula (24) comes immediately in the case of node pairs like O–N, O–E, etc., by using (23) and the relation $\tan\theta = k/h$, where θ is the angle shown in the inset. Note how (23) gives the same value for matrix entries corresponding to such pairs whatever the diagonal along which one has cut the rectangular cell. (Further study: Consider orthogonal, but not uniformly spaced, grids. Generalize to dimension 3.)



3.15. No, the regular tetrahedron is not a "space-filling" solid. Its dihedral angle, easily computed, is about 70°32′, hence a mismatch. See [Ka] or [Si] on such issues.

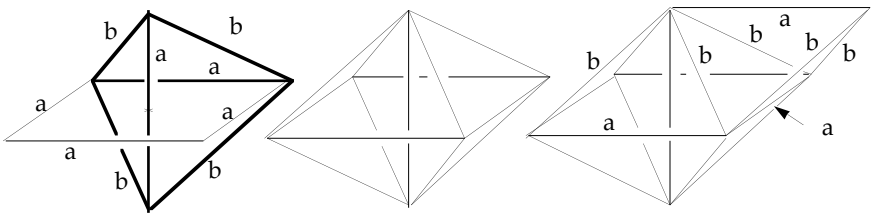


FIGURE 3.8. Left: The tetrahedron (in thick lines). Middle: Assembly of four copies of it into an octahedron, by rotation around the middle vertical pole. Right: Sticking a fifth copy to the upper right flank. A sixth copy will be attached to the lower left flank in the same way, hence a paving parallellepiped.

3.16. Cf. Fig. 3.8. Length b should equal $a\sqrt{2}/3$. There is numerical evidence [MP] that such tetrahedra yield better accuracy in some computations than the standard "cubic" grid, subdivided as Fig. 3.7 suggests. (A suitable combination of regular octahedra and tetrahedra, by which one *can* pave [Ka], may also be interesting in this respect.)

3.17. Call φ_1 and φ_2 the solutions corresponding to μ_1 and μ_2. Then $R_1^{-1} = \inf\{\int_D \mu_1 |\nabla\varphi|^2 : \varphi \in \Phi^1\} \le \int_D \mu_1 |\nabla\varphi_2|^2 \le \int_D \mu_2 |\nabla\varphi_2|^2 = R_2^{-1}$.

3.18. With respect to some origin, map D to D_λ by $x \to \lambda x$, with $\lambda > 0$, and assign to D_λ the permeability μ_λ defined by $\mu_\lambda(\lambda x) = \mu(x)$. If φ is an admissible potential for the problem on D, then φ_λ, similarly defined by $\varphi_\lambda(\lambda x) = \varphi(x)$, is one for the problem on D_λ. Changing variables, one sees that $\int_{D_\lambda} \mu_\lambda |\nabla\varphi_\lambda|^2 = \int_D \lambda\mu |\nabla\varphi|^2$, so it all goes as if μ had been multiplied by λ (in vacuum, too!). Hence the result by Exer. 3.17.

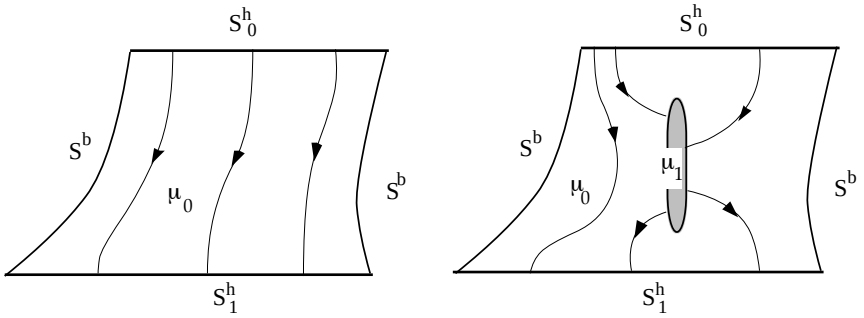


FIGURE 3.9. Exercise 3.19. How the presence in the domain under study of a highly permeable part ($\mu_1 >> \mu_0$), even of very small relative volume, is enough to distort the field. (Two-dimensional drawing, for clarity. In the case of Fig. 3.1, a similar effect would be achieved by putting a high-μ thin sheet inside D.)

3.19. Since both φ_1 and φ_2 belong to Φ^I, one can set $\varphi' = \varphi_1 - \varphi_2$ in both equations (25), and subtract, which yields

$$(\varphi_1, \varphi_1 - \varphi_2)_1 + (\varphi_2, \varphi_2 - \varphi_1)_2 = 0.$$

Therefore,

$$\|\varphi_1 - \varphi_2\|_1^2 = \int_D (\mu_1 - \mu_2)\, \nabla\varphi_2 \cdot \nabla(\varphi_2 - \varphi_1) \equiv ((1 - \mu_2/\mu_1)\varphi_2, \varphi_2 - \varphi_1)_1,$$

hence $\|\varphi_1 - \varphi_2\|_1 \le C(\mu)\, \|\varphi_2\|_1$ by Cauchy–Schwarz, where $C(\mu)$ is an upper bound for $|1 - \mu_2/\mu_1|$ over D. Hence the continuity with respect to μ (a small uniform variation of μ entails a small change of the solution), but

only, as mathematicians say, "in the L^∞ norm". The result cannot be improved in this respect: A large variation of μ, even concentrated on a small part of the domain, can change the solution completely, as Fig. 3.9 suggests.

REFERENCES

[BP] A. Berman, R.J. Plemmons: **Nonnegative Matrices in the Mathematical Sciences,** Academic Press (New York), 1979.

[BR] J.R. Bunch, D.J. Rose: **Sparse Matrix Computations,** Academic Press (New York), 1976.

[Ci] P.G. Ciarlet: **The Finite Element Method for Elliptic Problems**, North-Holland (Amsterdam), 1978.

[CR] P.G. Ciarlet, P.A. Raviart: "General Lagrange and Hermite interpolation in $\mathbb{R}^n$ with applications to finite element methods", **Arch. Rat. Mech. Anal., 46** (1972), pp. 177–199.

[Cr] P.G. Ciarlet: **Introduction à l'analyse numérique matricielle et à la programmation**, Masson (Paris), 1982.

[Co] N.A. Court: **Modern pure solid geometry,** Chelsea (Bronx, NY), 1964. (First edition, Macmillan, 1935.)

[DL] J. Deny, J.L. Lions: "Les espaces du type de Beppo Levi", **Ann. Inst. Fourier**, **5** (1953–1954), pp. 305–370.

[Ge] P.L. George: **Génération automatique de maillages. Applications aux méthodes d'éléments finis,** Masson (Paris), 1990.

[GL] A. George, J.W. Liu: **Computer Solution of Large Sparse Positive Definite Systems**, Prentice-Hall (Englewood Cliffs, NJ), 1981.

[Gv] G.H. Golub, C.F. Van Loan: **Matrix Computations**, North Oxford Academic (Oxford) & Johns Hopkins U.P. (Baltimore), 1983.

[Gr] P. Grisvard: **Elliptic Problems in Nonsmooth Domains,** Pitman (Boston), 1985.

[Jo] C.R. Johnson: "Inverse M-matrices", **Lin. Alg. & Appl., 47** (1982), pp. 195–216.

[Ka] J. Kappraff: **Connections**: The Geometric Bridge Between Art and Science, McGraw-Hill (New York), 1991.

[La] I. Lakatos: **Proofs and Refutations,** Cambridge U.P. (Cambridge), 1976.

[MP] P. Monk, K. Parrott, A. Le Hyaric: "Analysis of Finite Element Time Domain Methods in Electromagnetic Scattering", Int. report 96/25, Oxford University Computing Laboratory (Oxford, U.K.), 1996.

[Na] R. Nabben: "Z-Matrices and Inverse Z-Matrices", **Lin. Alg. & Appl., 256** (1997), pp. 31–48.

[Sf] P.P. Silvester, R. Ferrari: **Finite Elements for Electrical Engineers,** Cambridge University Press (Cambridge), 1991.

[Si] J. Sivardière: **La symétrie en Mathématiques, Physique et Chimie**, Presses Universitaires de Grenoble (Grenoble), 1995.

[SF] G. Strang, G.J. Fix: **An Analysis of the Finite Element Method,** Prentice-Hall (Englewood Cliffs, NJ), 1973.

[Va] R.S. Varga: **Matrix Iterative Analysis**, Prentice-Hall (Englewood Cliffs, NJ), 1962.

[We] T. Weiland: "Time Domain Electromagnetic Field Computation with Finite Difference Methods", **Int. Journal of Numerical Modelling, 9** (1996), pp. 295–319.