APPENDIX A

Mathematical Background

This is not a tutorial, just background material. It contains two interleaved texts. One is rather formal, with definitions, mainly, and occasionally, proofs, arranged in logical order: Things are defined in terms of primitive concepts and of previously defined things. The other part, where examples will be provided, is a commentary on why and how these notions and properties can be useful in computational electromagnetism. There, of course, one feels free to invoke not yet formally defined entities.

The treatment is neither exhaustive nor balanced. The space devoted to each notion does not necessarily reflect its intrinsic importance. Actually, most important notions will be familiar to the reader already and will cursorily be treated, just enough to provide a context for the ones I have chosen to emphasize: those that are (in my opinion) both important and generally underrated.

Most definitions in the formal part are implicit: When a new concept or object is introduced, its name is set in *italics*,[1] and the context provides a definition. The index should help locate such definitions, when needed.

A.1 BASIC NOTIONS

Notions we choose to consider as primitive, and that we shall not define, are those of set theory: sets, elements, subsets, equality, inclusion (symbols $=$, $\in$, and $\subseteq$), finite and infinite sets, and of logic: propositions, or "predicates", true or false. Basic notions that follow are defined in terms of primitive notions.

[1] Italics also serve to put emphasis on some words, according to standard practice. This should cause no confusion.

A.1.1 Sets, properties

If X is a set, $\mathcal{P}(X)$ will denote the set of all its parts (or *power set)*, and $\varnothing$ the empty set. Don't confuse elements, such as x, with one-element subsets, denoted $\{x\}$. The *Cartesian product* of sets X and Y is denoted by $X \times Y$. It is made of all *pairs* $\{x, y\}$, with $x \in X$ and $y \in Y$. The product $A \times B$ of two parts $A \subseteq X$ and $B \subseteq Y$ is the set of pairs $\{x, y\}$, with $x \in A$ and $y \in B$.

When speaking of pairs, order counts: x first, then y. If X and Y are different sets, no problem. But some confusion may occur when $X = Y$. If $x \neq y$, are $\{x, y\}$ and $\{y, x\}$ different elements of $X \times X$? Yes, of course, so $\{x, y\} \neq \{y, x\}$. But this same notation, $\{x, y\}$, is often used also for something else, namely, the subsets of X composed of two elements, and now, $\{x, y\}$ and $\{y, x\}$ point to the same object, an element of $\mathcal{P}(X)$. So we are dealing with a different concept here, that of *unordered* pair.

Some have tried to promote the use of a different word for unordered pairs ("couple", for instance) to stress the difference. But then it's difficult to remember which is which. So if you see $\{x, y\}$ at any place in this book, be it called couple or pair, assume the order counts, unless the context warns you otherwise. (Fortunately, the confusion is most often harmless.) The natural extension[2] of the pair concept is the *n-tuple,* $\{x_1, x_2, \dots, x_n\}$. Unordered n-tuples are subsets containing n elements, all different.

Propositions and predicates are statements which can assume the value **true** or **false** (with a special face, because **true** and **false** are labels for the two elements of a special set, "Boolean algebra", to which we shall return). Some examples of predicates: $x \in X$, or $x \neq y$, or $A \subseteq B$, or else $x \in A$ **and** $y \notin B$, etc. (Again, **and** is a logical operation in Boolean algebra, about which we shall have more to say.) The difference between "proposition" and "predicate" is semantical: Predicates can contain variables, whose values may affect the truth value of the predicate. For instance, speaking of real numbers, $x > 0$ is a predicate, the truth value of which depends on the value of the free variable[3] x, whereas $2 < 1$ is just a proposition[4] (its value is **false**).

[2] It's recursively defined: a triple $\{x, y, z\}$ is the pair $\{\{x, y\}, z\}$, where the first element is itself a pair, and so forth. Of course, only *finite* strings can be formed this way, but we'll soon do better.

[3] "Free variables" are those whose value matters to the expression containing them, like x in $x^2 + 2x + 1$ (whose value depends on x), as opposed to "bound" or "dummy variables", like y in $\int f(y)\, dy$. I shall not attempt to be more rigorous (see [Ha] and the article "Symbolic logic" in [It]). Instead, I hope to convey some feeling for this by accumulating examples.

Properties—for example, positivity of real numbers, positive-definiteness of matrices, solenoidality of vector-fields, etc.—are predicates involving such objects. If p is a property, one denotes by $\{x \in X : p(x)\}$ the subset of X made of all elements for which this property holds true. For some immediate examples, consider a subset R of the Cartesian product $X \times Y$. Its *section* by x is $R_x = \{y \in Y : \{x, y\} \in R\}$. (Beware it's a part of Y, not of X! Cf. Fig. A.1.) Its *projection* on X is $p_X(R) = \{x \in X : R_x \neq \emptyset\}$. Any property thus defines a subset, and a subset A defines a property, which is $x \in A$. Since subsets and properties are thus identified, operations on sets translate into operations on properties: thus, for example, $\{x \in X : p(x)$ **and** $q(x)\} = \{x \in X : p(x)\} \cap \{x \in X : q(x)\}$, and the same with **or** and $\cup$.

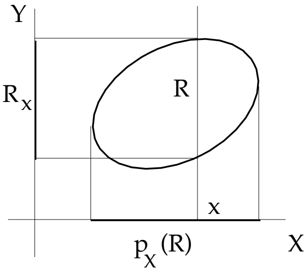


FIGURE A.1. Notions of section and projection.

A.1.2 Relations and functions

A *relation* is a triple $r = \{X, Y, R\}$, where X and Y are two sets and R a subset of $X \times Y$, called the *graph* of the relation. Objects x and y such that $\{x, y\} \in R$ are said to be related (or linked) by r. There are various shorthands for the predicate $\{x, y\} \in R$, such as $r(x, y)$ or (more often) the so-called "infix" notation $x\ r\ y$. (Familiar examples of the latter are $x \leq y$, $u \perp v$, etc.) The *domain* and *codomain* of r are the projections $p_X R$ and $p_Y R$, thus denoted:

[4] Of course, when a sentence such as "$x > 0$" or "div $b = 0$" appears in a text the aim of which is not primarily mathematical, we assume this predicate has the value **true**. In fact, the author is usually telling us just that, but won't risk the ridicule of saying "I have just proven that the predicate 'div $b = 0$' is true". However, the occurrence in such texts of bits of formal reasoning, as for instance when discussing the truth of the statement, "if div $b = 0$, there exists some a such that $b = \text{rot } a$", clearly shows that maintaining the distinction between a predicate and the assertion that this predicate is true (which is of course, another predicate) is not only a formal game. Sometimes, it's the only way to settle an argument.

$$\mathrm{dom}(r) = \{x \in X : R_x \neq \varnothing\}, \quad \mathrm{cod}(r) = \{y \in Y : R_y \neq \varnothing\}.$$

So, informally, the domain dom(r) contains all those x of X that relate to some y in Y, and the codomain cod(r) is the symmetrical concept: all y's related to some x. (The codomain of r is also called its *range*.) The *inverse* of r is the relation $r^{-1} = \{Y, X, R\}$, and the domain of one is the codomain of the other.

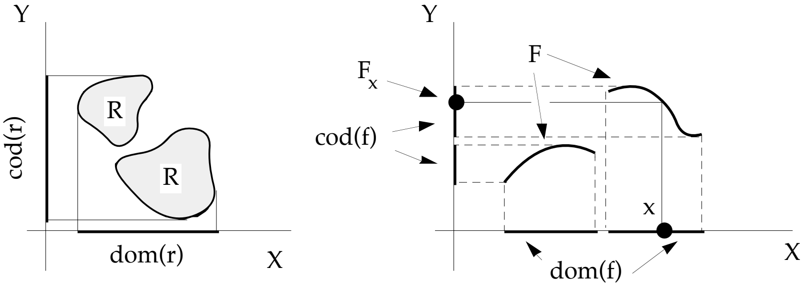


FIGURE A.2. Left: Domain and codomain. Right: A functional graph F.

The graph F of a relation $f = \{X, Y, F\}$ is *functional* if each section F_x contains at most one element of Y (Fig. A.2, right). The relation f is then called a *function* "from X to Y", or a "Y-valued function on X". The function f is *surjective* if $\mathrm{cod}(f) = Y$, *one-to-one* if its inverse f^{-1} (which is always defined, but only as a relation, a priori) is also a function, then called the *inverse function,* or *reciprocal* of f. The set of all functions from X into Y will be denoted by $X \to Y$, and if f is such a function, we'll say that "the *type* of f is $X \to Y$". The construct "$f \in X \to Y$" thus makes sense, under the convention that "$\to$" has precedence over "$\in$", and I occasionally use it, but "$f : X \to Y$" is the standard way to introduce a function of type $X \to Y$.

Thus, all functions are a priori *partial,* that is, dom(f) may be strictly smaller than X. *Total* functions are those for which $\mathrm{dom}(f) = X$. A total function is *injective* if it is one-to-one, surjective, as we just said, if $\mathrm{cod}(f) = Y$, and *bijective* if both properties hold. Total functions are called *mappings* or *maps,* but again, excessive emphasis on such fine semantic distinctions is not very productive. Better take "function" and "mapping" as synonyms, and call attention on whether $\mathrm{dom}(f) = X$ or not, when necessary. Most functions in this book are partial.

Some relations (in the common sense of the word) between physical entities are better conceived as general relations than as functional ones. A good example is provided by "Bean's law", an idealization of what happens in a type-II superconductor when all currents flow parallel to some given direction. The scalar components j and e of the current density and the electric field along this direction are then supposed to be related as follows: if $e \neq 0$ at some point, then j at this point is equal to some characteristic value j_c, called the *critical current*, and the sign of j is that of e ; if $e = 0$, then any value of j between $-j_c$ and j_c is possible, and which one actually occurs at any instant depends on the past evolution of e. This relatively complex prescription is elegantly summarized by a (non-functional) relation: The pair {e, j} must belong to the graph of Fig. A.3, left.

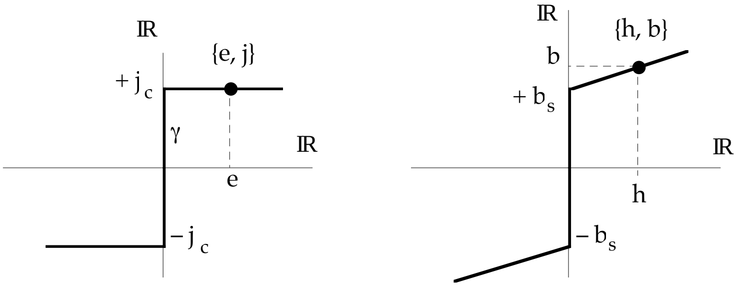


FIGURE A.3. Left: Bean's law, for type-II superconductors, expressed as a non-functional relation {IR, IR, γ}. Right: A similar idealization of the b–h characteristic of a "soft" ferromagnetic material.

The same trick is useful to express b–h constitutive laws in similar circumstances (horizontal currents, vertical magnetic field). If, as it happens for instance in induction heating simulations, one works over a large range of values of h (some 10^5 A/m, say), the hysteretic cycle is so narrow, relatively speaking, that one may as well ignore hysteresis. Hence the behavior depicted in Fig. A.3, right. Again, this b–h relationship is conveniently expressed by a non-functional relation, i.e., a graph.

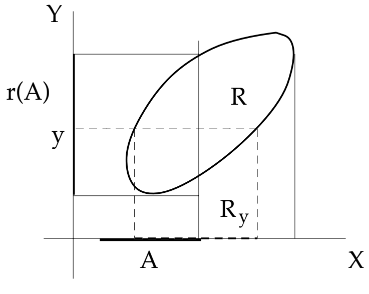


FIGURE A.4. Image of A under r = {X, Y, R}.

Let's proceed with concepts that are common to relations and functions. If A is a part of X, its *image* under r, denoted by $r(A)$, is

$$r(A) = \{y \in Y : R_y \cap A \neq \varnothing\}$$

(cf. Fig. A.4). Note that $r(X) = \mathrm{cod}(r)$. If $A = \{a\}$, a single element set, we write $r(a)$ instead of $r(\{a\})$, and call this set the image of a under r. Note that $r(x) = R_x$, hence the syntax $y \in r(x)$ as another way to say that x and y are r-related. For $B \subseteq Y$, the set $\{x \in X : R_x \cap B \neq \varnothing\}$, denoted $r^{-1}(B)$, is called the *inverse image* or *pre-image* of B, and $r^{-1}(Y) = \mathrm{dom}(r)$. Remark that $\mathrm{cod}(r) = \{y \in Y : r^{-1}(y) \neq \varnothing\}$. Relation $s = \{X, Y, S\}$ is *stronger* than relation r if $S \subseteq R$. This is obviously another relation,[5] between relations, which can logically be denoted by $s \subseteq r$.

We need mechanisms to build new relations from old ones.[6] Since relations are graphs, and hence sets, operations on sets apply to relations: given $r = \{X, Y, R\}$ and $s = \{X, Y, S\}$, one can form the new relation $\{X, Y, R \cap S\}$ (stronger than both r and s), which we denote by r **and** s. Similarly, r **or** $s = \{X, Y, R \cup S\}$ (weaker than both r and s). In the special case $S = A \times Y$, where A is a part of X, r **and** s is called the *restriction* of r to A (Fig. A.5). Its domain is $\mathrm{dom}(r) \cap A$. It's usually denoted by $r_{|A}$. If $r = s_{|A}$ for some $A \subseteq X$, one says that s is an *extension* of r.

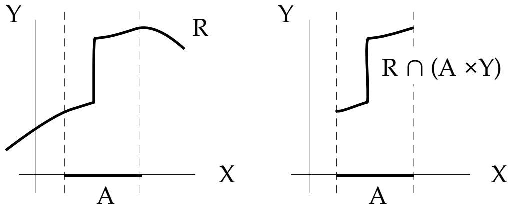


FIGURE A.5. Restricting relation $r = \{X, Y, R\}$ to A. An alternative definition of the restriction is $r_{|A} = r \circ \mathrm{id}(A)$, where $\mathrm{id}(A) = \{X, X, \Delta \cap (A \times X)\}$ (see the definition of Δ p. 270).

[5] Can you describe its graph? Is it an order (as defined below)? A total one or only a partial one?

[6] *Computer programming* is basically just that. A program is a function p defined on the set S of possible states of the machine; entering data selects some $s \in S$, and $p(s)$ is the final state, including output display; the game consists in building p from a set of basic "instructions", which are, as one may show, functions of type $S \to S$. So programming consists, indeed, in building new functions from old ones. (For serious developments on this, see a treatise on "functional programming", for instance [Hn].) Here, we need not go so far as formally presenting a programming language (the rare bits of programs that appear in this book should be self-explanatory), but some awareness of the underlying mechanisms may be useful.

Relations can also be composed, when their sets match properly: Given $r = \{X, Y, R\}$ and $s = \{Y, Z, S\}$, the *composition* of r and s, denoted by $s \circ r$, is

$$s \circ r = \{X, Z, \cup \{r^{-1}(y) \times s(y) : y \in Y\}\}.$$

This amounts to saying that $z \in (s \circ r)(x)$ if and only if there is at least one y such that $y \in r(x)$ and $z \in s(y)$ (Fig. A.6). The composition $g \circ f$ of two functions of respective types $X \rightarrow Y$ and $Y \rightarrow Z$ is a function of type $X \rightarrow Z$.

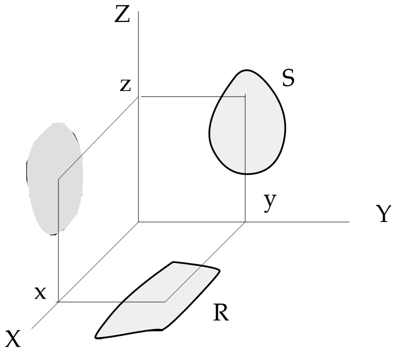


FIGURE A.6. Composing two relations. When {x, y} and {y, z} span R and S respectively, {x, z} spans the graph of the relation s o r.

A.1.3 Families

Family is just another name, more appropriate in some contexts, for "total function". If $\mathcal{J}$ is a set (finite or not), and X is a set, a *family of objects of type* X, *indexed by* $\mathcal{J}$, is a mapping from $\mathcal{J}$ into X, denoted by $\{x_i : i \in \mathcal{J}\}$. To each label i, taken in $\mathcal{J}$, thus corresponds an object x_i, of type X. It is convenient—although a bit confusing, perhaps—to denote the set of all such families by $X^{\mathcal{J}}$. The set $\mathcal{J}$ can be finite, in which case it may seem we have redefined n-tuples. Not so: for $\mathcal{J}$, finite or not, is *not* supposed to be ordered. So we really have a new concept[7] here.

The distinction may sometimes be useful. Think of the nodes of a finite element mesh. They form a set $\mathcal{N}$, usually finite. Let us call N the

[7] It's not "unordered n-tuple", either, which we chose earlier to interpret as an n-element subset of X. In the case of families, repetitions are allowed, and two labels i and j can point to the same object, $x_i = x_j$. A part Y of X can always be considered as a family, however, by indexing it over itself: $Y = \{x_y : y \in Y\}$, where $x_y = y$. So one should not worry too much about such fine distinctions.

number of nodes (that is, $N = \#\mathcal{N}$, if one wishes to use this convenient shorthand for the number of elements in a set). Suppose a real-valued "degree of freedom" (DoF) is assigned to each node. We thus have a family $\{\mathbf{u}_n : n \in \mathcal{N}\}$ of N real numbers, indexed over $\mathcal{N}$, that will be denoted by **u**, in boldface. Here, X is the set $\mathbb{R}$ of real numbers, the index set $\mathcal{J}$ is $\mathcal{N}$, and **u** is thus a member of $\mathbb{R}^{\mathcal{N}}$, with the above notation. If you think, "This **u** is an N-dimensional real vector", you are right, for indeed, $\mathbb{R}^{\mathcal{N}}$ is a real vector space of dimension N. But you should resist the natural compulsion to say, "So, this is an element of $\mathbb{R}^N$ (the Cartesian product of $\mathbb{R}$ by itself, N times) and hence, an N-tuple". An N-element family is not an N-tuple, because no order among its members is implied. Nodes are labelled, not numbered. A family is less structured than an N-tuple, in this respect.

Remark A.1. Nodes are not *yet* numbered, that is. True, at some stage in the process of finite element modelling, a numbering scheme is introduced: When solving for the DoF, by using the Gauss–Seidel method for instance, there will be a first DoF, a second DoF, etc. But sound programming methodology demands that this numbering be deferred to the very stage where it becomes relevant and useful. Moreover, one may have to deal with several different numbering schemes for the same set of nodes—if only to test numbering schemes [KB] for efficiency (they affect the bandwidth of the matrices and the speed of iterative methods). Such a *numbering* will be a (bijective) mapping from $\mathcal{N}$ onto $[1, N]$, the segment of the first N nonzero integers, and it will assign to each N-member family an N-tuple, in a one-to-one way. So once a numbering is given, there is an identification (an isomorphism—see Note 11, p. 276) between $\mathbb{R}^{\mathcal{N}}$ and $\mathbb{R}^N$. But this is not a "canonical" identification, since it depends on the numbering. $\mathbb{R}^{\mathcal{N}}$ and $\mathbb{R}^N$ are definitely not the same object. ◊

A.1.4 Binary relations

We now look at the case where $Y = X$. A relation $r = \{X, X, R\}$, then called a *binary* relation in X, confers on X some structure, that X alone did not possess. Thus, the compound "X as equipped with the relation r" (that is, the pair $\{X, r\}$), is a new object, for which the notation $\{X, r\}$ is appropriate.

Two standard examples of binary relations, equivalence and order, will come to mind. Let us call the part $\Delta = \{\{x, y\} \in X \times X : x = y\}$ of $X \times X$ the *diagonal*, and the relation $\mathrm{id} = \{X, X, \Delta\}$ the *identity*. A relation r is *reflexive* if its graph contains Δ, that is, if $\mathrm{id} \subseteq r$, *symmetric* if $r^{-1} = r$, *antisymmetric* if $(r \textbf{ and } r^{-1}) \subseteq \mathrm{id}$, *transitive* if $(r \circ r) \subseteq r$. A reflexive, transitive, and symmetric (resp. antisymmetric) relation is an *equivalence* (resp. an *order,* or *ordering*). Cf. Fig. A.7.

Generic notation for equivalences and orders is $\equiv$ and $\leq$ (or $\subseteq$), and one uses expressions such as *lesser than, greater than,* etc., instead of symbols, occasionally. If $r = \{X, X, R\}$ is an order (e.g., the relation $\leq$ in $\mathbb{R}$), one calls $\{X, X, R - \Delta\}$ the *strict* associated relation (example: $<$ in $\mathbb{R}$ is thus associated with $\leq$), enunciated as *strictly lesser than*, etc. Such relations, for which the generic notation is $<$, are *not* orders (beware!), so one tends to avoid them; hence the use of contrived expressions, such as *nonnegative* for $\geq$, to avoid the ambiguous "positive" (is it $>$ or $\geq$?).

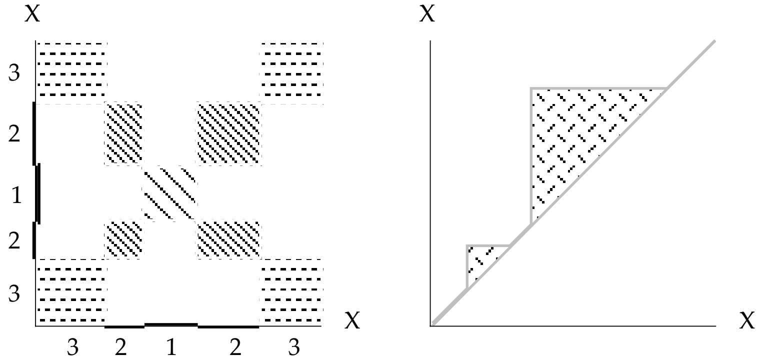


FIGURE A.7. Structure of the graph R (the shaded set) for an equivalence (left) and a partial order (right). Notice how R, on the left, splits into separate parts of the form $X_i \times X_i$ (i = 1, 2, 3 here, corresponding to three different shading textures), where each X_i is an equivalence class (see below, A.1.6). The shaded set includes Δ in both cases. See also Fig. A.8.

A.1.5 Orders

An order r is *total* if $r \cup r^{-1} = \{X, X, X \times X\}$, that is, for all pairs $\{x, y\}$, either $x \in r(y)$ or $y \in r(x)$. Total orders, like $\leq$ in $\mathbb{R}$ or $\mathbb{N}$, are also called *linear* orders, which makes intuitive sense: They line up things.

A nice standard example of partial order is divisibility in $\mathbb{N}$. (See also Note 5.) A more topical one, for us, is on the set $\mathcal{M}$ of all possible finite element meshes of a given region: A mesh *m'* is a refinement of a mesh *m* (one may say *m'* is *finer than m*) if each edge, face or volume of *m* is properly meshed by a suitable restriction of *m'*. Two different meshes may have a common refinement without any of them being finer than the other, so the order is only partial.

The supremum should not be confused with what is called a *maximal element* in A, that is, some x in A such that $x \leq y$ never holds, whatever y in A. In Fig. A.8, f and g are maximal in A, d and e are *minimal* (and a is minimal in X). Maximal elements are not necessarily unique, and may not exist at all (the open interval]0, 1[has none).

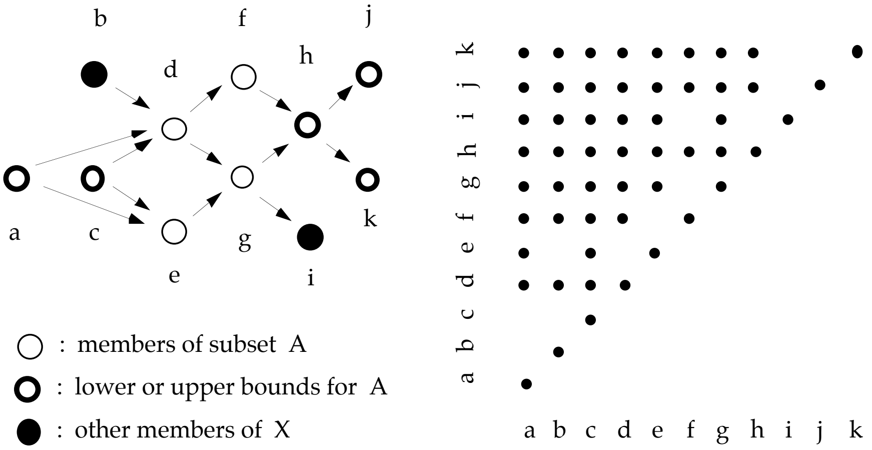
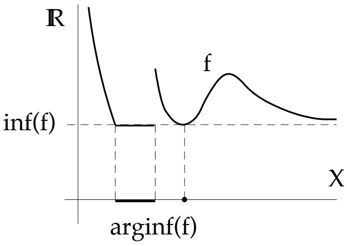


FIGURE A.8. A partial order on an 11-element set, $X = \{a, b, \ldots, k\}$. Right: According to the graphic convention of Fig. A.7. Left: As a *sagittal* graph (i.e., with arrows), which is much more convenient in the case of transitive relations (because most arrows can be omitted, as for instance the one from a to f).



Notions of inf and sup pass to functions $f \in X \to Y$, when Y is ordered ($Y = \mathbb{R}$, most often). The infimum inf(f) of a function is the infimum inf(f(X)) of its image. More explicit notation, such as $\inf\{f(x) : x \in X\}$, is generally used. The pre-image of inf(f(X)), that is to say, the subset of elements of X that realize the minimum of f, is denoted arginf(f), or $\text{arginf}(\{f(x) : x \in X\})$ (but one will not bother with the double system of parentheses, usually). Note that arginf(f) is not an element of X but of $\mathcal{P}(X)$, which can be the empty set. One says that f "reaches its minimum" on arginf(f).

The index set $\mathcal{J}$ of a family may be ordered. If the order is total, a family $\{x_i : i \in \mathcal{J}\}$ is called a *sequence*. (The use of this word, in general, rather implies that $\mathcal{J} = \mathbb{N}$, or some subset of $\mathbb{N}$. If $\mathcal{J} = \mathbb{R}$, or some interval of $\mathbb{R}$, one will rather say something like *trajectory*.) This is the generalization of n-tuples, when $\mathcal{J}$ is infinite. If the order is only partial, we have a *generalized sequence*. The family of all finite element approximate solutions to a field problem is one.

Finally, a *minimizing sequence* for a function $f : X \to Y$ is a family $\{x_n : n \in \mathbb{N}\}$ of elements of X such that $\inf\{f(x_n) : n \in \mathbb{N}\} = \inf(f)$. The standard way to prove that $\operatorname{arginf}(f)$ is not empty is to look for the limit of a minimizing sequence.

A.1.6 Equivalence classes, "gauging"

If r is an equivalence in X, the set $r(x)$ is called the *equivalence class* of x. Equivalence classes are disjoint, and their union is all of X (Fig. A.7, left). In other words, an equivalence relation generates a *partition* of X into equivalence classes (and the other way around: A partition induces an equivalence relation).

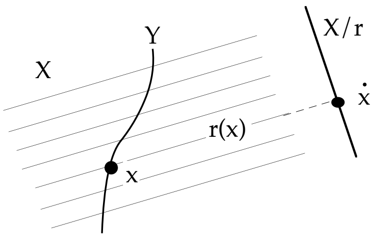


FIGURE A.9. Equivalence classes, quotient, representative section.

This provides one of the most powerful mechanisms for creating new objects in mathematics (and this is why the previous notions deserved emphasis). When objects of some kind are equivalent in some respect, it's often worthwhile to deal with them wholesale, by dumping all of them into an equivalence class, and treating the latter as a new, single object. If X is the initial set and r the equivalence relation, the set of classes is then called the *quotient* of X by r, with various denotations, such as X/r for instance. Don't confuse the quotient, elements of which are not of type X, with what one may call a *representative section* (Fig. A.9), which is a subset of X "transverse to the classes", so to speak, obtained by picking one element x (the *representative element*) in each equivalence class. The quotient is in one-to-one correspondence with each representative section. Among such sections, some may be more remarkable than others, depending on which structure X possesses.

Examples abound, and many appear in this book. Let us just mention the following, of special interest in electromagnetism. If some field b is divergence-free in some region of space, there may exist, under conditions which are not our present concern, a field a such that $b = \text{rot } a$, called a "vector potential" for b. Such a field is not unique (one may always add a gradient to it). Among vector potentials, the relation $\text{rot } a_1 = \text{rot } a_2$ is an equivalence, the classes of which are obviously in one-to-one relation with the b's.

The various representations of the electric field e provide a more involved example. From Faraday's law ($\partial_t b + \text{rot } e = 0$), and by using the above representation $b = \text{rot } a$, we have $e = -\partial_t a - \text{grad } \psi$, where ψ is called the "scalar (electric) potential". Calling A and Ψ the sets of suitable potentials a and ψ (they should satisfy some qualifying assumptions, which we need not give here explicitly), we have an equivalence relation in $A \times \Psi$: Two pairs $\{a_1, \psi_1\}$ and $\{a_2, \psi_2\}$ are equivalent if they correspond to the same electric field, i.e., if $\partial_t a_1 + \text{grad } \psi_1 = \partial_t a_2 + \text{grad } \psi_2$, over some specified span of time.

One may conceive all pairs $\{a, \psi\}$ in a given class as mere representations (all equivalent) of some electric field, but the mathematical point of view is bolder: The equivalence *class,* taken as a whole, *is* the same object as the electric field. Dealing with e (in numerical simulations, for instance), or dealing with the whole class of $\{a, \psi\}$s, is the same thing. But of course, a vector field and a class of pairs of fields are objects of very different nature, and doing the mental identification may not be easy. Hence the more conservative approach that consists in selecting among the members of an equivalence class some distinguished one, as representative of the class.

For instance, we may privilege among the pairs $\{a, \psi\}$ of a given class (i.e., a given e) the one for which $\text{div } a = 0$. (There is only one, if we work in the whole space. Otherwise, additional boundary conditions are needed to select a unique a.) Such a specification for selecting one member in each equivalence class is called a *gauging* procedure. (The previous one is "Coulomb gauge". Imposing $c^2 \, \partial_t \psi + \text{div } a = 0$ is "Lorenz[8] gauge".) Now, one may feel more assertive in dealing with *the* pair $\{a, \psi\}$ as a representation of e. Things go sour, however, when one entertains the illusion that this pair would be more deserving, more "physical", than its siblings of the same class, that it would be the "right" one, in some way. Such futile concerns about gauging have delayed the implementation of 3D eddy-current codes for years in some institutes.

Remark A.2. The same delusion seems to be at the root of persistent misunderstanding about the Aharonov–Bohm effect. (Cf. [AB]: Interference experiments on electrons detect the existence of an induction flux inside an

[8] According to [NC], it's L. Lorenz, not H.A. Lorentz.

extremely thin tightly wound solenoid, in spite of the field being null outside it. This is paradoxical only when one insists on thinking of electrons as *localized* classical objects which, according to such a naive view, "have to pass" in the region where $b = 0$ and thus "cannot feel the influence" of b, whereas they could "feel" that of a.) The effect, some argue, points to the vector potential a rather than the induction b as the "most fundamental" descriptor[9] of the field. If the issue really is, "Of the two *mathematical* objects, b on the one hand, and the whole class of a's such that $b = \text{rot } a$ on the other hand, which one must be considered as 'the primitive concept' ?", one may conceivably take sides. This is a choice between two different *formalisms* for the *same* theory, since the two objects are in one-to-one correspondence and describe exactly the same physics (which is why no experiment can resolve such an issue). Undeniably, when it comes to quantum field–particle interactions, having a in the equations is more convenient. But since the equations are gauge-invariant, none of the representatives of the class is thus privileged, so there is nothing in the AB effect that would give arguments to consider the "Coulomb gauged", or the "Lorenz gauged" vector potential as the *physical* one. The frequent claim (cf., e.g., [Kn]) that AB would allow one to *measure* the Coulomb gauged vector potential is totally misleading. What can be measured, by determining the electron's phase shift, is the induction flux, from which of course the Coulomb gauged a is readily derived in the axisymmetric situation usually (and needlessly) assumed. ◊

A.1.7 Operations, structured sets

Operations are functions of type $X \to X$ (*unary* operations), $X \times X \to X$ (*binary* operations) etc., i.e., functional relations that map n-tuples of X to elements of X. The reader may wish to translate the standard concepts of *commutativity* and *associativity* of operations in terms of graphs of such relations.

Sets under consideration in specific questions are not naked, but structured, by relations and—mostly—operations. We saw how an equiva-

[9] In some cases, which verge on the tendentious, the observed interferences are said to be "due to the vector potential", and hence (the reader is subtly led to conclude, although it's never explicitly said) "not due" to the (magnetic induction) field. This is silly. The involved phase factor can be computed indifferently in terms of a (as $\exp[-iq/\hbar \int_\gamma \tau \cdot a]$, where γ is a loop around the solenoid), or b (as $\exp[-iq/\hbar \int_\Sigma n \cdot b]$, where Σ is some surface bounded by γ, and therefore, pierced by the solenoid). Both expressions yield the same value, by the Stokes theorem. The latter may be less *convenient*, in the thin-solenoid case usually discussed, for b is then a distribution, but this is a side issue. Anyhow, there is no need to postulate a "thin" solenoid to discuss the AB effect (cf. Exers. 8.5 and 8.6).

lence or an order could, already by itself, structure a set. But usually there is much more: On $\mathbb{R}$, for instance, there is order, addition, multiplication, division, and all these relations interact to give the set the structure we are used to. The same goes for all standard sets, such as $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Z}$, $\mathbb{C}$, and so forth. So what is called $\mathbb{R}$ is in fact $\{\mathbb{R}, \leq, +, *, \ldots\}$, that is, the set in full gear, equipped with all its structuring relations, and this is where the concept of *type* is useful: A type is a structured set.[10]

Moreover, objects of different types may interact via other relations, hence an encompassing structure, informally called an *algebra*. (Appendix B gives a detailed example), which further stretches the notion of type. So when I say that x is "of type X", as has happened several times already, I mean not only that x is a member of the set X, but that x can enter in relation with other objects, belonging to X or to other sets, to all the extent allowed by the rules of the algebra. For instance, when X is $\mathbb{N}$, the integer n can be added to another integer m, can be multiplied by it, etc., but can also be added to a real number, serve as exponentiation factor, etc. The type of an object, in short, encompasses all one can do to it and with it.[11]

The practice of denoting the type and the underlying set the same way has its dark side: If one is fond of very compact symbols, as mathematicians are, some overloading is unavoidable, for the same symbol will have to represent different types. For example, $\mathbb{R}^3$ normally stands for the set of triples of real numbers. It is quite tempting to use this symbol also to denote structures which are isomorphic to $\mathbb{R}^3$, like the three-dimensional real vector space, or ordinary 3D space. This is highly questionable, for the operations allowed on triples, on 3-vectors, and on points of space are not the same. (We'll elaborate on that later. But already it is clear that points, as geometric objects, cannot be added or multiplied by scalars, the way vectors can.) Hence the occasional appearance in this book of long symbols like *POINT* or *VECTOR* to denote different types built upon the same underlying set (namely, $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$).

We'll work out the simple example of *Boolean algebra*, denoted B, to see how a few operations can give rich structure to even the less promising set, one with two elements, labeled **true** and **false**. Two relations will structure it. (The corresponding type is the one called *LOGICAL* in most programming languages.) The first operation is **not** = {B, B, NOT}, the

[10] An *isomorphism* between $\{X, r_1, r_2, \ldots\}$ and $\{X', r'_1, r'_2, \ldots\}$ is then a one-to-one map f such that $x \; r_i \; y \Leftrightarrow f(x) \; r'_i \; f(y)$ for all x, y, i, that is, a structure-preserving bijective map. If there is only one obvious sensible choice for f, one says the isomorphism is *canonical*.

[11] "Object", here, has the same sense as in "object-oriented programming" [Me].

graph of which, NOT, is depicted in Fig. A.10. It's the subset of B × B consisting of the two pairs {**true**, **false**} and {**false**, **true**}, out of four in all. This graph is functional, and the unary operation **not** it defines is indeed what was expected, turning **true** into **false**, and vice versa. The second function is **and** = {B × B, B, AND}, where the graph AND, functional again, now contains four elements out of the possible eight, as shown in Fig. A.10 (where the limits of the graphical representation we used till now become obvious; hence the preferred use of *tabular* representations for binary operations, as in Fig. A.11). We may now define the new function **or** by x **or** y = **not**(**not**(x), **not**(y)), and combine them in various ways.

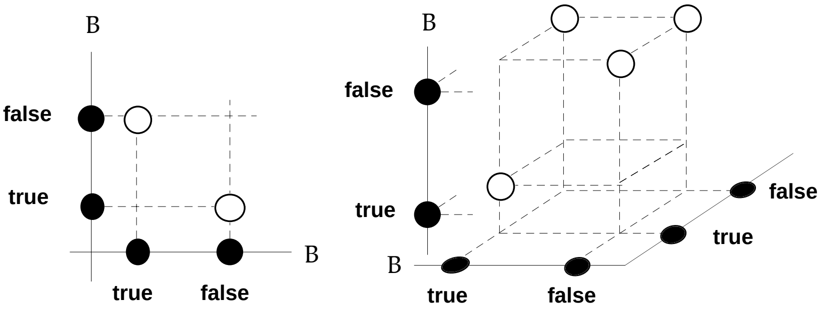


FIGURE A.10. Left: Graph of the "unary" function **not**. Right: Graph of the binary function **and**.

	T	T
T	T	F
F	F	F

and

	T	F
T	T	T
F	T	F

or

	T	F
T	F	T
F	T	F

xor

	T	F
T	T	F
F	T	T

⇒

FIGURE A.11. Some operations in Boolean algebra (**T** and **F** stand for **true** and **false**), in tabular representation. (Variable x spans rows of the table, y spans columns, and the entry at x—y is the truth value of x r y.) Note how the symmetry of relations, or lack thereof, is rendered.

A.1.8 Logic

Which leads us into propositional calculus, and logic. Propositions, and predicates more generally, can be seen as B-valued functions, whose domain is the set of all possible "well-formed" expressions (that is, all strings of symbols that conform to some specific grammar, which only logicians and programming-language designers take the trouble to make explicit). Since predicates are thus relations, the above building mechanisms apply to them. For instance, if p and q are predicates, p **and** q is one, too: Its value, according to the above definition of **and**, is **true** when both p and q assume the value **true**, and **false** otherwise. The algebraic structure of B allows a lot of similar constructions: **not** p, p **or** q, p **and** (q **or** r), etc.

One among these constructs, q **or not** p, whose value is **false** when p = **true** and q = **false**, and **true** in all other cases (Fig. A.11), is so frequently used that it deserves a special notation: $p \Rightarrow q$.[12] The abuse, and perhaps even the use of this, should be discouraged. Better use "if p, then q". The difference is only a matter of concrete syntax,[13] but it seems to matter much, and such a plain sentence is less prone to confusion than "$p \Rightarrow q$".

Two other important shorthands should be known, $\forall$ and $\exists$. They too allow new predicates to be built from old ones. Suppose p(x) is some predicate containing the variable x in which x is free. Then $p(x)\ \forall\, x$ is a new predicate, the value of which is **true** if and only if p(x) = **true** for all possible values of x, and now the variable x is *bound*.[14] Symmetrically, the truth value of $p(x)\ \exists\, x$ (or, with a more readable syntax, $\exists\, x : p(x)$) is **false** if and only if p(x) = **false** for all possible values of x. These symbols are dreaded by many engineers, and perhaps not without some reason, for their abuse in mathematical training during the 1970s has done much harm, worldwide. They should be used very sparingly, especially the latter, and there is alternative concrete syntax, such as p(x) **for all** x, or even "p(x) holds for all x" and "p(x) holds for some x", much closer to natural language, while still being unambiguous.

[12] So, be wary of the informal use of $p \Rightarrow q$, voiced as "p implies q". (See [Hr] for a nice discussion of this and similar issues.) The risk is high that "q is true" will be understood, which may be wrong. The safe use of this in reasoning demands that *two* different statements be proved: that $p \Rightarrow q$ is true (whatever the values of the free parameters in both p and q, which may affect their truth values) *and* that p is true. One can then conclude that q is true. Misuses of this basic and celebrated logical mechanism are too often seen. The most common mistake consists in carefully proving that $p \Rightarrow q$ holds, while overlooking that p can be false (for some values of the free variables that appear in it), and to go on believing that q has been proved.

[13] Abstract syntax deals with the deep structure of formal expressions (including programming languages). Concrete syntax is concerned with the choice of symbols, their position, how they are set, etc. See [Mr].

[14] One may fear some logical loophole here, but no worry: the definition of "free" and "bound" (they are antonyms) is recursive. If a variable appears at only one place, it's free, and the only way to bind two occurrences of the same variable is to invoke one among a limited list of binding mechanisms, including the use of the so-called "quantifiers" $\forall$ and $\exists$, as described above.

There is no reason to deny oneself the convenience of a shorthand, however—for instance, in constructs of the form "*Find* h *such that*

(p) $\quad i\omega \int \mu\, h \cdot h' + \int \sigma \operatorname{rot} h \cdot \operatorname{rot} h' = 0 \quad \forall\, h' \in \mathbb{H}$."

This one means that the equality $i\omega \int \mu\, h \cdot h' + \int \sigma \operatorname{rot} h \cdot \operatorname{rot} h' = 0$ should hold whatever the test field h', provided the latter is selected within the allowed class of such fields, which class is denoted by $\mathbb{H}$. The predicate (p), in which variable h' is bound whereas h is free, thus expresses a property of h, the unknown field, the property that characterizes h as the solution to the problem at hand. To say "*find* h *such that* (. . .) = 0 *for all* h' *which belong to* $\mathbb{H}$" is the same prescription, only a little more verbose. But to omit the clause "∀ h' . . . " or "for all h' . . . ", whatever the concrete syntax, would be a capital sin, turning the precise statement of a problem into gibberish. The issue is not tidiness, or lack thereof, but much more importantly, *meaning:* Without the clause "$\forall\, h' \in \mathbb{H}$", the problem is not posed at all.

A.1.9 A notation for functions

Before leaving this Section, I wish to explain an idiosyncrasy that you also may find convenient at times.

Many functions are defined via algebraic expressions. Take for instance the expression[15] $x^2 + 2x + 1$. The set $\{\{x, y\} \in \mathbb{R} \times \mathbb{R} : y = x^2 + 2x + 1\}$ defines a functional relation, f say. I find convenient, time and again, to write this

(e) $\quad f = x \rightarrow x^2 + 2x + 1,$

which should be parsed as suggested by Fig. A.12 and understood as follows: "Let's name f the function that maps the real number x to the real number that results from evaluating the expression $x^2 + 2x + 1$." (Of course, the arrow should not be read as "tends to", according to the more standard convention, which I avoid, except in unambiguous constructs such as $\lim_{\varepsilon \to 0} \ldots$, etc.)

[15] An *expression* is just a combination of symbols that conforms to some definite syntax. In *algebraic* expressions, two kinds of symbols are allowed: variables or constants, of definite types (here, the real x and the integer 2), and relational symbols (here, + and the exponentiation) that belong to a specific algebra (here, standard arithmetic). To *evaluate* the expression consists in assigning to the variables definite values, and doing the computation according to the rules of the algebra. Note that the same expression could make sense in other algebras: x can be a matrix, for instance.

$$\boxed{f = \boxed{x \to \boxed{\boxed{x^2} + 2x + 1}}}$$

FIGURE A.12. Parsing (e), that is, finding its logical structure, here indicated by the hierarchy of nested boxes, as a syntactic analyzer would do, if instructed of precedence rules: Multiplication and exponentation take precedence over addition, the arrow is weaker than all operational symbols, and = is the weakest link of all.

On both sides of the equal sign in (e) we have the same mathematical object, a function, only differently tagged: by its name f on the left, and by the whole expression $x \to x^2 + 2x + 1$ on the right. (Variable x is bound in this expression: another example of the binding mechanism.) So the equal sign is quite legitimate at this place. On the other hand, the arithmetic expression $x^2 + 2x + 1$ (where x is a free variable) is *not* the function, and to write $f = x^2 + 2x + 1$ would be highly incorrect.[16]

Why not simply $f(x) = x^2 + 2x + 1$? This is a bit ambiguous, because it can also stand for the statement of an equality, unless you declare explicitly your intention to use it as a function definition. Hence the frequent use of a special symbol, ≙ or :=, for "is defined as", like this: $f(x) \triangleq x^2 + 2x + 1$. But if special symbol there must be, better choose the arrow, which puts emphasis on the right object, the defined one, which is f, not[17] $f(x)$. Also, the arrowed notation can be nested without limits, as the following example will show.

Given a function q on 3D-space, which may represent for instance an electric charge density, one may define its Newtonian *potential* (the electric potential, in that case, up to the factor ε_0) as follows:

$$\psi = x \to \frac{1}{4\pi} \int \frac{q(y}{|x - y|}\, dy,$$

where dy is the volume element and $|x - y|$ the distance between points x and y. (Observe that y is bound in the integral—yet another binding

[16] If you think I insist too much on such trivia, pay attention to the practice of physics journals: most often, $f(x)$ refers to a *function*, and f to its *value*. Mathematicians do exactly the opposite: f is the function, $f(x)$ its value at x. This schism is all the more detrimental to science in that it goes generally unnoticed.

[17] One might argue that $f(x) \triangleq x^2 + 2x + 1$ needs some quantifier, such as perhaps $\forall$, to really define f. Actually, a quantifier has been designed for just that purpose: the λ of "lambda-calculus". Cf. [Kr].

mechanism—and x free, and how the arrow binds x.) Now, one may define a new function, of higher level: $G = q \to \psi$, that is, the operator (named after Green) that maps q to ψ. Instead of this two-step definition, we may, thanks to the arrowed notation, write

$$G = q \to (x \to \frac{1}{4\pi} \int \frac{q(y)}{|x - y|}\, dy),$$

in one stroke. (Parentheses force the correct parsing.) This is a precious shortcut at times, to be used sparingly, of course.

This is the first example we encounter of a function defined on a set whose elements are themselves functions, and which maps them to other functions. For clarity, such functions are called *operators* (especially when, as in the present case, the correspondence is linear). When the set they map to is $\mathbb{R}$, the word *functional* is used (cf. p. 62).

The arrowed notation is especially useful when variables and parameters occur together. Examine this:

$$\operatorname{grad}(y \to \frac{1}{|x-y|}) = y \to \frac{x-y}{|x-y|^3}$$

(an equality between two vector fields, since $x - y$ is a vector). Here, x is the parameter, y the variable, and there is no ambiguity as to which gradient, with respect to x or to y, we mean.

Expressions other than arithmetic can be put on the right of the arrow: conditional expressions, and even whole programs. For example, we may have this

$$g = x \to \textbf{if } x \geq 0 \textbf{ then } x^2 + 2x + 1 \textbf{ else } 0,$$

$$h = x \to \textbf{if } x \geq 0 \textbf{ then } x^2 + 2x + 1.$$

The difference between g and the previously defined f (cf. (e)) is clear (they differ for $x \leq 0$), but what about h with respect to f? They surely differ, since $\operatorname{dom}(f) = \mathbb{R}$, whereas $\operatorname{dom}(h) = \{x \in \mathbb{R} : x \geq 0\}$. Yet their defining expressions are the same. But h is the restriction of f to the positive half-line. As this example shows, the domain of a function should always be described with precision, for the expression or fomula or recipe for evaluating the function may well make sense *beyond* this domain.[18] We'll see this phenomenon recur when we study the differential operators grad, rot, div. *Different* operators will similarly be called

[18] Consider $f = x \to (x^2 - 2x + 1)/(x - 1)$, with $\operatorname{dom}(f) = \mathbb{R} - \{1\}$.

"gradient", for instance, and will differ by the extent of their respective domains.

A.2 IMPORTANT STRUCTURES

A.2.1 Groups

Groups are important, not only because many mathematical structures like linear space, algebra, etc., are first and foremost groups, with added features, but as a key to *symmetry*.

A *group* is a set equipped with an associative binary operation, with a neutral element and for each element, an inverse. Examples: the group $\mathbb{Z}$ of relative integers, the regular matrices of some definite order, etc.

As these two examples show, the group operation may or may not be commutative, hence a notational schism. Commutative, or *Abelian,* groups, like $\mathbb{Z}$, are often denoted additively. But in the general case, the operation is called a product, denoted without any symbol, by simple juxtaposition, the neutral element is 1, and the inverse of g is g^{-1}.

A group G *acts* on a set X if for each $g \in G$ there is a map from X to X, that we shall denote by $\pi(g)$, such that $\pi(1)$ is the identity map, and $\pi(gh) = \pi(g) \circ \pi(h)$.[19] Observe, by taking $h = g^{-1}$, that $\pi(g)$ must be bijective, so $\pi(g)$ is a *permutation* of X. The set $\{\pi(g) : g \in G\}$ is thus a group of permutations,[20] the group law being composition of maps. Let's denote this set by $\pi(G)$.

The same abstract group can act in different ways on various related geometric objects: points, vectors, plane figures, functions, fields, tensors, etc. What counts with groups is their actions. Hence the importance of the related vocabulary, which we briefly sketch.

The action is *faithful,* or *effective,* if $\pi(g) = 1$ implies $g = 1$. (Informally, an action on X is effective if all group elements "do something" on X.) In that case, G and $\pi(G)$ are isomorphic, and $\pi(G)$ can be seen as a "concrete" realization of the "abstract" group G. This justifies writing gx, instead of $\pi(g)x$, for the image of x by $\pi(g)$. The *orbit* of x under the action of G is the set $\{gx : g \in G\}$ of transforms of x. Points x and y are in the same

[19] This is called an action *on the left,* or *left action*, as opposed to a *right action*, which would satisfy $\pi(gh) = \pi(h) \text{ o } \pi(g)$, the other possible convention. A non-Abelian group can act differently from the left and from the right, on the same set. All our group actions will be on the left.

[20] A subgroup of the "symmetric group" $S(X)$, which consists of *all* permutations on X, with composition as the group law.

orbit if there exists some group element g that transforms x into y. This is an equivalence relation, the classes of which are the orbits. If all points are thus equivalent, i.e., if there is a single orbit, one says the action is *transitive*. The *isotropy* group (or *stabilizer*, or *little group)* of x is the subgroup $G_x = \{g \in G : gx = x\}$ of elements of G that fix x. A transitive action is *regular* if there are no fixed points, that is, $G_x = 1$ for all x (where 1 denotes the trivial group, reduced to one element).

In the case of a regular action, X and G look very much alike, since they are in one-to-one correspondence. Can we go as far as saying they are identical? No, because the group has more structure than the set it acts upon. For a simple example, imagine a circle. No point is privileged on this circle, there is no mark to say "this is the starting point". On the other hand, the group of planar rotations about a point (where there *is* a distinguished element, the identity transform) acts regularly on this circle. Indeed, the circle and this group (traditionally denoted SO_2) *can* be identified. But in order to *do* this identification, we must select a point of the circle and decide that it will be paired with the identity transform. The identification is not canonical, and there is no group structure on the circle before we have made such an identification.

The concept of *homogeneous space* subsumes these observations. It's simply a set on which some group acts transitively and faithfully. If, moreover, the little group is trivial (regular action), the only difference between the homogeneous space X and the group G lies in the existence of a distinguished element in G, the identity. Selecting a point O in X (the origin) and then identifying gO with g —hence O in X with 1 in G —provides X with a group structure.

So when homogeneity is mentioned, ask what is supposed to be homogeneous (i.e., ask what the elements of X are) and ask about the group action. (As for isotropy and other words in tropy, it's just a special kind of homogeneity, where the group has to do with rotations in some way.)

A.2.2 Linear spaces: V_n, A_n

I don't want to be rude by recalling what a *vector space* (or *linear space)* is, just to stress that a vector space V is already a group (an Abelian one), with the notion of scalar[21] multiplication added, and appropriate axioms. The *span* $\vee\{v_i : i \in \mathcal{J}\}$ of a family of vectors of V is the set of all weighted sums $\sum_{i \in \mathcal{J}} \alpha^i v_i$, with scalar coefficients α^i only a *finite*

[21] Unless otherwise specified, the field of scalars is $\mathbb{R}$.

number of which are nonzero (otherwise there is nothing to give sense to the sum). This span, which is a vector space in its own right, is a *subspace* of V. A family is linearly *independent* if the equality $\sum_i \alpha^i v_i = 0$ forces all $\alpha^i = 0$. The highest number of vectors in a linearly independent family is the *dimension* of its span, if finite; otherwise we have an *infinite dimensional* subspace. The notion applies as a matter of course to the family of all vectors of V. If the dimension dim(V) of V is n, one may, by picking a basis (n independent vectors $e_1, \ldots, e_n$), write the generic vector v as $\mathbf{v}^1 e_1 + \ldots + \mathbf{v}^n e_n$, hence a one-to-one correspondence $v \leftrightarrow \{\mathbf{v}^1, \ldots, \mathbf{v}^n\}$ between v and the n-tuple of its *components*. So there is an isomorphism (non-canonical) between V and $\mathbb{R}^n$, which authorizes one to speak of *the* n-dimensional real vector space. That will be denoted V_n. Don't confuse V_n and $\mathbb{R}^n$, however, as already explained. In an attempt to maintain awareness of the difference between them, I use boldface for the components,[22] and call the n-tuple $\mathbf{v} = \{\mathbf{v}^1, \ldots, \mathbf{v}^n\}$ they form, not only a vector (which it is, as an element of the vector space $\mathbb{R}^n$), but a **vector**. Notation pertaining to $\mathbb{R}^n$ will as a rule be in boldface.

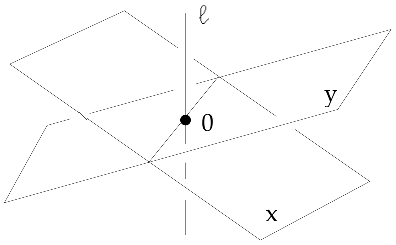
A relation r = {V, W, R}, where V and W are vector spaces is *linear* if the graph R is a vector space in its own right, that is, a subspace of the product V × W. If the graph is functional, we have a *linear map*. Linear maps $s : V \rightarrow W$ are thus characterized by $s(x + y) = s(x) + s(y)$ and $s(\lambda x) = \lambda s(x)$ for all factors. Note that dom(s) and cod(s) are subspaces of V and W.

Next, affine spaces. Intuitively, take V_n, forget about the origin, and what you have got is A_n, the n-dimensional affine space. But we are now equipped to say that more rigorously. A vector space V, considered as an additive group, acts on itself (now considered just as a set) by the mappings $\pi(v) = x \rightarrow x + v$, called *translations*. This action is transitive, because for any pair of points {x, y}, there is a vector v such that $y = x + v$, and regular, because $x + v \neq x$ if $v \neq 0$, whatever x. The structure formed by V as a set[23] equipped with this group action is called the *affine space* A *associated with* V. Each vector of V has thus become a point of A, but there is nothing special any longer with the vector 0, as a point in A.

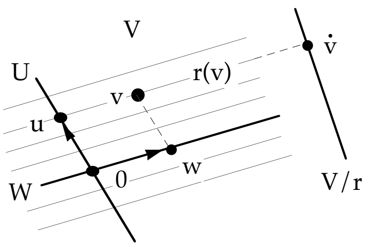
More generally, an *affine space* A is a homogeneous space with respect to the action of some vector space V, considered as an additive group. By

[22] At least, when such components can be interpreted as degrees of freedom, in the context of the finite element method. Our DoF-vectors are thus **vectors**. (Don't expect absolute consistency in the use of such conventions, however, as this can't be achieved.)

[23] Be well aware that V is first *stripped* of its operations, thus becoming a mere set, then *refurnished* with this group action, to eventually become something else, namely A.



selecting a point 0 in A to play origin, we can identify vector v of V with point $0 + v$ of A. But there may be no obvious choice for an origin. For example [Bu], having selected a point x and a line ℓ through x in 3D space, all planes passing through x and not containing ℓ form an affine space (inset). None of them is privileged, and the group action is not obvious.[24] For an easier example, consider a subspace W of some vector space V, and define an equivalence r by $u \; r \; v \Leftrightarrow v - u \in W$. Equivalence classes have an obvious affine structure (W acts on them regularly by $v \to v + w$) and are called *affine subspaces* of V, *parallel* to W. Of course, no point of an affine subspace qualifies more than any other as origin.



Remark A.3. The latter is not just any equivalence relation, but one which is compatible[25] with the linear structure: if $x \; r \; y$, then $\lambda x \; r \; \lambda y$, and $(x + z) \; r \; (y + z)$. This way, the quotient X/r is a vector space. Now if one wants to select a representative section, it makes sense to preserve this compatibility, by requesting this section to be a vector subspace U of V (inset, to be compared with Fig. A.9), which is said to be *complementary* with respect to W. Then each $v \in V$ can uniquely be written as $v = u + w$, with $u \in U$ and $w \in W$. Again, don't confuse the quotient V/r with the complement U, although they are isomorphic. ◊

Affine space is perhaps the most fundamental example of homogeneous space. From a philosophical standpoint, the fact that we chose to do almost all our applied physics in the framework provided by A_3 (plus, when needed, a time parameter) reflects the *observed* homogeneity of the space around us.

[24] Take a vector u parallel to ℓ, and two parallels ℓ' and ℓ'' to u, distinct from ℓ. They pierce plane x at x' and x''. The "translation" associated with $\{\lambda', \lambda''\} \in \mathrm{IR}^2$ is the mapping $x \to \{$the plane determined by $0, \; x' + \lambda' u, x'' + \lambda'' u\}$.

[25] Note the importance of this concept of compatibility between the various structures put on a same set.

What you can do on *VECTORS* may not be doable on *POINTS*. Indeed, the product λx is meaningless in an affine space: What makes sense is *barycenters*. The barycenter of points x and y with respective weights λ and $1 - \lambda$ is $x + \lambda(y - x)$. Generalizing to n points is easy. Affine independence, dimension of the affine space, and affine subspaces follow from the similar concepts as defined about the vector space. *Barycentric coordinates* could be discussed at this juncture, if this had not already been done in Chapter 3.[26]

Affine relations are characterized by affine graphs. If the graph is functional, we have an *affine map*. Affine maps on A_n are those that are linear with respect to the $(n + 1)$-**vector** of barycentric coordinates. *Affine subspaces* are the pre-images of affine maps. Affine subspaces of a vector space are of course defined as the affine subspaces of its associate. The sets of solutions of equations of the form $Lx = k$, where L is a linear map (from V_n to V_m, $m \leq n$) and k a vector, are affine subspaces, and those corresponding to the same L and different k's are parallel. The one corresponding to $k = 0$ (called *kernel* of L, denoted $\ker(L)$) is the vector subspace parallel to them all.

If x is a point in affine space A, vectors of the form $y - x$ are called *vectors at* x. They form of course a vector space isomorphic with the associate V, called *tangent space at* x, denoted T_x. (In physics, elements of V are often called *free vectors*, as opposed to *bound vectors*, which are vectors "at" some point.) The tangent space to a curve or a surface that contains x is the subspace of T_x formed by vectors at x which are tangent to this curve or surface. Note that vector fields are maps of type *POINT* → *BOUND_VECTOR*, actually, with the restriction that the value of v at x, denoted $v(x)$, is a vector at x. The distinction between this and a *POINT* → *FREE_VECTOR* map, which may seem pedantic when the point spans ordinary space, must obviously be maintained in the case of fields of tangent vectors to a surface.

A *convex set* in an affine space is a part C such that

$$(x \in C \text{ and } y \in C) \Leftrightarrow \lambda x + (1 - \lambda)y \in C \quad \forall \lambda \in [0, 1].$$

Affine subspaces are convex. The intersection of a family of convex sets is a convex set. The *convex hull* of a part K is the intersection of all convex sets containing K, and thus the smallest of such sets. It coincides with the union of all barycenters, with nonnegative weights, of pairs of points of K.

[26] One may—but it's a bit more awkward than the previous approach—define affine spaces ab initio, without first talking of vector spaces, by axiomatizing the properties of the *barycentric map*, which sends $\{x, y, \lambda\}$ to $\lambda x + (1 - \lambda)y$.

Let us finally discuss *orientation* of vector and affine spaces (cf. 5.2.3). This cannot be ignored, because of the prominent role played in electromagnetism by the cross product and the curl operator—both "sensitive to orientation", in a sense we shall discover later.

A *frame* in V_n is an (ordered) n-tuple of linearly *independent* vectors. Select a basis (which is thus a frame among others), and look at the determinant of the n vectors of a frame, hence a $FRAME \rightarrow REAL$ function. This function is basis-dependent, of course. But the equivalence relation defined by "$f \equiv f'$ if and only if frames f and f' have determinants of same sign" does not depend on the basis, and is thus intrinsic to the structure of V_n. There are two equivalence classes with respect to this relation.

Orienting V_n consists in designating one of them as the class of "positively oriented" frames. This amounts to defining a function, which assigns to each frame a label, like e.g., "direct" and "skew". There are two such functions, therefore two possible orientations. (Equivalently, one may define an oriented vector space as a pair {vector space, privileged basis}, provided it's well understood that this basis plays no other role than specifying the orientation.)

Subspaces of V_n also can be oriented, by the same procedure, and orientations on different subspaces are unrelated things. Affine subspaces are oriented by orienting the parallel vector subspace. For consistency, one agrees that the subspace {0} can be oriented, too, by giving it a sign, + 1 or – 1. Connected patches of affine subspaces, such as polygonal faces, or line segments (and also, after the previous sentence, points), can be oriented by orienting the supporting subspace. Lines and surfaces as a whole are oriented by conferring orientations to all their tangents or tangent planes in a consistent[27] way, if that can be done. (It cannot in the case of a Möbius band, for instance.)

There is another kind of orientation of subspaces (and hence, of lines, surfaces, etc.), called *outer* orientation. By definition, an outer orientation of a p-dimensional subspace W of V_n is an orientation of one of its complementary subspaces U, as previously defined (Remark A.3). As we saw in Chapter 5 in the case $n = 3$, this formalizes the concepts of "crossing direction" when $p = 2$, and of "way of turning around" a line when $p = 1$. *If* the ambient space is oriented, outer orientation of W determines an inner orientation: Given a frame in W, made of p vectors, one may add to them the $n - p$ vectors of a *positively* oriented frame in U, hence a

[27] I hope this makes intuitive sense. One cannot be more precise without introducing manifolds and charts, that is, starting a differential geometry course.

frame of V_n, which falls into one of the two orientation classes, hence the orientation of the original frame.

A.2.3 Metric spaces

A *metric space* $\{X, d\}$ is a set X equipped with a *distance*, that is, a function $d : X \times X \rightarrow \mathbb{R}$ such that $d(x, y) = d(y, x) \geq 0$ $\forall$ $x, y \in X$, with $d(x, y) > 0$ if $x \neq y$, and $d(x, z) + d(z, y) \geq d(x, y)$ $\forall$ x, y, z.

Metric-related notions we may have to use are *open ball* $B(x, r) = \{y \in \mathbb{R} : d(x, y) < r\}$, *open set* (a part A which if it contains x also contains an open ball centered at x), *closed* set (one the complement of which is open), *distance* $d(x, A)$ of a point x *to a part* A (which is $\inf\{d(x, y) : y \in A\}$), *adherence* or *closure* $\overline{A}$ of a part A (all points x of X such that $d(x, A) = 0$), *interior* of A (points such that $d(x, X - A) > 0$, set denoted $\text{int}(A)$ when we need it), *boundary* ∂A of A (points for which $d(x, A) = 0$ and $d(x, X - A) = 0$). A sequence $\{x_n \in X : n \in \mathbb{N}\}$ *converges* if $\lim_{n \rightarrow \infty} d(x_n, x) = 0$ for some $x \in X$, called the *limit*, which one immediately sees must then be unique. By taking all the limits of all sequences whose elements belong to a part A, one obtains its closure.[28] A part A is *dense* in B if $\overline{A}$ contains B, which means (this is what counts in practice) that for any $b \in B$ and any $\varepsilon > 0$, there is some $a \in A$ such that $d(a, b) < \varepsilon$. This is strictly the same as saying that one can form a sequence $\{a_n \in A : n \in \mathbb{N}\}$ that converges towards b. A metric space X is *separable* if it contains a denumerable dense part.

Weaker than the notion of limit is that of *accumulation point*: Let us say (this is not part of the received terminology) that a family "clusters" at x if one can extract from it a sequence that converges to x (then called an accumulation point for this family). Convergent sequences cluster at their limit (and at no other point). Some sequences may not cluster at all. *Compact* parts of a metric space are closed parts in which any sequence must cluster at some point.

These notions are useful in approximation theory (Chapter 4). For instance, the union of all approximation spaces, for all imaginable meshes of a given domain, form a dense set in the set of all eligible fields, for the energy distance. Moreover, the family of approximate solutions, indexed over these meshes, clusters at the right solution. But knowing that is not enough. What one wants, which is more difficult, is to devise "refinement rules" which, starting from any mesh, generate a sequence of finer meshes

[28] Beware: This equivalence, like some others this list suggests, may not hold in topological spaces whose topology cannot be described by a distance.

with the property of convergence of the corresponding approximate solutions. (Incidentally, the functional space must be separable for this to be possible, since elements of the form $\sum_{i \in \mathcal{J}} \varphi_i \lambda^i$, where the coefficients φ_i take rational values only, and $\mathcal{J}$ is the union of the sequence of Galerkin bases, form a dense denumerable set. Most usual functional spaces are separable, as a corollary of the Weierstrass theorem[29] on polynomial approximation of continuous functions.)

A function f from {X, d} to {X', d'} is *continuous* if it maps converging sequences of X to converging sequences of X'. This is equivalent to saying that the pre-image of a closed set is closed. The function f is *uniformly* continuous if for each $\varepsilon > 0$, there exists $\delta(\varepsilon)$ such that, for any pair {x, y} of points taken in dom(f), $d(x, y) < \delta(\varepsilon)$ implies $d'(f(x), f(y)) < \varepsilon$. Obviously (this is a standard exercise in manipulating quantifiers), uniform continuity is logically stronger than simple continuity, but the two notions coincide when f is affine. An *isometry* from {X, d} to {X', d'} is a function f such that $d'(f(x), f(y)) = d(x, y)$ for all {x, y} in dom(f). This implies one-to-oneness, and uniform continuity of f and of its reciprocal, and therefore, *homeomorphism* (existence of a one-to-one map continuous in both directions), but is stronger.[30]

Remark A.4. You may be excused for guessing that continuous functions are functional relations with a closed graph, for this seems so natural. But it's wrong . . . For instance, the "weak gradient" of Chapter 5 has a closed graph in $L^2(D) \times \mathbb{L}^2(D)$, but is not continuous. (Relations between continuity of functions and closedness of their graphs are governed by the "Banach theorem", a deep result which belongs to the hard core of functional analysis, but is not used here. See [Br].) ◊

If a sequence clusters, its image by a continuous function clusters, too, so the continuous image of a compact part is compact. In particular, a real-valued continuous function whose domain is compact reaches its minimum and its maximum, since its image is closed.

Remark A.5. This obvious result is important in proving existence of equilibrium configurations in many physical situations. Suppose the set of states S can be described as a normed space (see below), the norm of a state being precisely its energy. (This is what we do in Chapters 2, 3, and 4.) States of bounded energy that satisfy specific constraints (of the kind $f(x) = 0$, where f is a continuous function) then form a closed bounded set.

[29] Given a nonempty compact subset K of $\mathbb{R}^n$, a continuous function $f : K \rightarrow \mathbb{R}$, and $\varepsilon > 0$, there exists a polynomial $p : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $|f(x) - p(x)| < \varepsilon$ for all $x \in K$.

[30] So strong actually, that it implies much more. For instance, a theorem by Mazur and Ulam asserts that a surjective isometry between real Banach spaces is affine [MU].

Such a set is compact if S is of finite dimension, because then any bounded sequence must cluster somewhere (as shown by the well-known Bolzano–Weierstrass proof). But such a sequence need not cluster in infinite dimension, since there are an infinity of directions in which to go. This is why existence proofs for variational problems in infinite-dimensional functional spaces always require some structural element to supply, or replace, the missing compactness. Quite often (as will be the case with the projection theorem proved a little later, and in 3.2.1) *convexity,* associated with completeness (see A.4.1 below) is this element. ◊

The *support* of a real- or vector-valued function f on a metric space X is the closure of the set $\{x \in X : f(x) \neq 0\}$. Don't confuse support and domain.

A useful density result, often invoked is this book, can informally be expressed as follows: *Smooth* fields over E_3 form a dense set among fields, in the energy norm. To be more precise, let h be a vector field such that $\int_{E_3} |h(x)|^2 \, dx < \infty$, and denote by $\|h\|$ the square root of this energy-like integral. Take a *mollifier* ρ, that is, a real-valued C^∞ function on E_3, nonnegative, with bounded support, and such that $\int \rho = 1$. Define the sequence h_n by[31]

$$(*) \qquad h_n(x) = \int_{E_3} \rho(y) \, h(x - y/n) \, dy,$$

or in more compact notation, $h_n = \rho_n * h$, where $*$ denotes the *convolution product*, and $\rho_n = x \rightarrow n^{1/3} \rho(nx)$. The outcome of such a product is as smooth as the smoothest factor ("regularizing property" of convolution), so h_n is $\mathbb{C}^\infty$. (This result, if not its proof, is intuitive: If ρ is smooth, one can differentiate under the summation sign in $(*)$, indefinitely.) Now, evaluate the quadratic norm of $h - h_n$: A tricky computation, which makes use of Fubini's theorem (see, e.g., [Pr]), and thus roots deeply in Lebesgue integration theory, will show that this norm tends to 0. Hence the density. The result is easily extended to fields over D by restriction.

We made repeated use of this when invoking the following argument, either in this form or in a closely related one: Suppose $f \in L^2(D)$ and $\int f \varphi = 0$ for all $\varphi \in C^\infty(D)$. By density, there is a sequence φ_n in $C^\infty(D)$ which converges towards f. Each term of the sequence $\int f \varphi_n$ is zero, and its limit is $\int |f|^2$ by continuity of the scalar product, hence $f = 0$ a.e. See for instance the proof of Prop. 2.1, p. 43.

[31] This way, $h_n(x)$ is a weighted average of values of h at points close to x. One often assumes a nice shape for the graph of ρ (centered at the origin, invariant by rotation, etc.), but this is not required, as far as theory is concerned.

A *path* in a metric space is a continuous mapping $c : [0, 1] \to X$. A *circuit,*[32] or *loop,* is a path that closes on itself ($c(0) = c(1)$). Let's call *patch* a continuous mapping $C : [0, 1] \times [0, 1] \to X$. A part of a metric space is *connected* if two of its points, x and y, can always be joined by a path c, that is, with $c(0) = x$ and $c(1) = y$. A connected open set is called a *domain*. We know the word in a different sense already, but this dual use should not be too confusing, for the "domain of definition" of a function or a field is often a domain in the present topological sense, and the context always makes clear what one means.

Two circuits c_0 and c_1 are *homotopic* if there is a patch C such that $c_0 = s \to C(s, 0)$ and $c_1 = s \to C(s, 1)$, with $C(0, t) = C(1, t)$ for all t in $[0, 1]$. This means one can be continuously deformed into the other, all intermediate steps being loops.[33] A metric space X is *simply connected* if any circuit is continuously reducible to a point, that is, homotopic to a circuit of the form $c(t) = x \ \forall\, t \in [0, 1]$, where x is a point of X.

A.2.4 Normed spaces, Euclidean norms

Being metric and being a vector or affine space are two different things, but if a set bears both structures, they had better be compatible. Suppose X is an affine space, with associated vector space V, and a distance d. The two structures are *compatible* if $d(x + v, y + v) = d(x, y)$, for all points x, y and all translation vectors $v \in V$. Then, once selected an origin 0, the real-valued function on V defined by $\|v\| = d(0, 0 + v)$ has the following properties, which characterize, by definition, a *norm:* $\|v\| > 0$ unless $v = 0$, $\|\lambda v\| = |\lambda|\|v\|$ for all v in V and real λ, and $\|v + w\| \leq \|v\| + \|w\|$. If, conversely, a vector space has a norm $\|\ \|$, the distance this induces on the associated affine space, $d(x, y) = \|x - y\|$, is compatible with the affine structure. A *normed space* is, in principle, a vector space V equipped with a norm, but you will often realize, thanks to the context, that what is really implied is the associated affine metric space.

Norms often stem from a *scalar product,* that is, a real-valued mapping, denoted $(\ ,\ )$, from $V \times V$ to $\mathbb{R}$, linear with respect to both arguments, symmetrical (i.e., $(v, w) = (w, v)$ for all $\{v, w\}$), and overall, *positive*

[32] For some, "circuit" implies more smoothness than mere continuity of c.

[33] More generally, two *maps* g_0 and g_1 from Y into X are *homotopic* if there is a continuous map f from $Y \times [0, 1]$ into X such that $f(s, 0) = g_0(s)$ and $f(s, 1) = g_1(s)$. One of the rare merits of some recent science-fiction movies has been to popularize this notion, since it happens to be the formalization of the concept of "morphing". The case of loops is when Y is a circle.

definite, that is

$$(v, v) > 0 \Leftrightarrow v \neq 0.$$

The norm of a vector v is then defined as $\|v\| = [(v, v)]^{1/2}$. Thereby, notions such as *orthogonality* (v and w are orthogonal if $(v, w) = 0$, which one may denote $v \perp w$) and *angle* (the angle of two nonzero vectors v and w is $\arccos((v, w)/\|v\|\,\|w\|)$), make sense. A vector space with scalar product is a "pre-Hilbertian space", a structure we shall study in its own right later.

The most familiar example is when $V = V_n$. Then we rather denote the norm $\|v\|$ by $|v|$, the scalar product (v, w) by $v \cdot w$, call them the *modulus* and the *dot product* respectively, and say that they confer a *Euclidean structure* (or a *metric*) on V_n and its affine associate A_n, via the *Euclidean distance* $d(x, y) = |x - y|$. *Euclidean geometry* is the study of the affine metric space $\{A_n, d\}$, called n-dimensional *Euclidean space*. (Why the singular in "space" will be discussed below.)

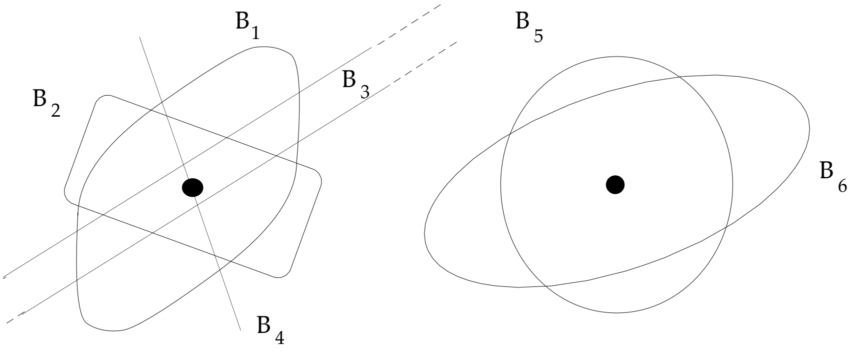


FIGURE A.13. Convex sets B_1, B_2, B_5, and B_6 are barrels, but B_3 (not bounded) and B_4 (not absorbing) are not. Observe the various degrees of symmetry of each barrel. Only those on the right generate *Euclidean* norms.

Other norms than Euclidean ones can be put on V_n, as suggested by Fig. A.13. All it takes is what is aptly called a *barrel,* that is a bounded, closed, absorbing, and balanced convex set B: *Balanced* means that $-v \in B$ if $v \in B$, *absorbing* that for every v, there exists $\lambda > 0$ such that λv belongs to B, *closed*, that the real interval $I(v) = \{\lambda : \lambda v \in B\}$ is closed for all v, and *bounded*, that I(v) is bounded for all $v \neq 0$. One then sets $|v|_B = 1/\sup\{\lambda : \lambda \in I(v)\}$, and this function (the inverse of which is called the *gauge* associated with B) is easily seen to be a norm. (Check that $|v|_B = \inf\{\lambda : v \in \lambda B\}$.) The closed unit ball for this norm is then the barrel B itself. (Notice how the two different notions of closedness we had up to now are thus reunited.)

Barrels that generate Euclidean norms (that is, ellipsoids) are obviously "more symmetrical" than others (Fig. A.13). A bit of group theory confirms this intuition. Let's denote by GL_n, and call *linear group*, the group of all bijective linear maps on V_n. (It's isomorphic, via a choice of basis, with the group of $n \times n$ regular matrices, but should not be confused with it.) Let us set $G(B) = \{g \in GL_n : \ v \in B \Leftrightarrow gv \in B\}$, the subgroup of linear transforms that leave B globally invariant, that is to say, the little group of B relative to the action of GL_n on subsets of V_n. We see immediately, on the examples of Fig. A.13, where $n = 2$, what "more symmetrical" means: $G(B_1)$ has only two elements (the identity and the reflection $v \rightarrow - v$ with respect to the origin), $G(B_2)$ has four, whereas $G(B_5)$ and $G(B_6)$ have an infinity (both groups are isomorphic with SO_2, actually). It can be shown (but this is beyond our reach here) that Euclidean barrels are those with *maximal*[34] isotropy groups, and thus indeed, the most symmetrical barrels that can exist.

This symmetry is what is so special about Euclidean norms. A bit earlier, we remarked that physical homogeneity of the space around us was reflected in the choice of affine space as framework for most our modellings. One may add to this that *isotropy* of physical space is reflected in the use of Euclidean norms, hence their prominent role. Indeed, a Euclidean norm *privileges no direction*:[35] If v and w both belong to the surface of a Euclidean barrel B, there is a linear transform $g \in G(B)$ such that $w = gv$. In other words, the action of G(B) on the surface of B, that is to say, its action on *directions*, is transitive. This is not true of other barrels.

Alternatively (as one easily sees, the two properties are equivalent), one may say that GL_n acts transitively on the set of all Euclidean barrels. In other words, given two scalar products " $\cdot$ " and " $\circ$ ", there is a linear

[34] More precisely: Define the *unimodular* group SL_n as the subgroup of linear maps that are represented, in some basis, by matrices of determinant 1. (This definition is in fact basis independent.) Then maximal proper subgroups of SL_n, all isomorphic with the group of orthogonal $n \times n$ matrices, are the isotropy groups of Euclidean barrels.

[35] *In spite* of what Fig. 13, right part, seems to suggest: One may feel like objecting "what about the principal directions of the ellipsoids B_5 or B_6 ?" But there is nothing special with these directions. They are spurious, due to the fact that we commit ourselves to a specific basis *just to do the drawing.* They will disappear if one selects the eigenvectors as basis (then B becomes a disk). Contrast this with the two "axes" of B_2, which no change of basis can erase. Denizens of a flat world with a metric governed by the barrel B_2 would be able to recognize the existence of privileged directions in their universe. Ask any New Yorker.

invertible map L such that $v \circ w = L v \cdot Lw$. Which is why one can speak of *the* Euclidean geometry, *the* Euclidean norm, in the face of apparent evidence about their multiplicity. For after all, each choice of basis in V_n generates a particular dot product, to wit $v \cdot w = \sum_i \mathbf{v}^i \mathbf{w}^i$, where $\mathbf{v}$ and $\mathbf{w}$ are the component **vector**s, so there seems to be as many Euclidean geometries as possible bases. Why then *the* Euclidean space? Because all Euclidean structures on V_n are equivalent, up to linear transformation. We shall see this in more concrete terms in the next Section.

A.3 OUR FRAMEWORK FOR ELECTROMAGNETISM: E_3

In this book, we work in 3D affine space, and it is assumed all along that a specific choice of dot product *and* orientation has been made once and for all. Thus, what is called E_3 in the main text is always *oriented Euclidean three-dimensional space.*

Note that the all-important notion of *cross product* would not make sense without orientation: By definition, $u \times v$ is orthogonal to both u and v and its modulus is $|u|\,|v| \sin\theta$, where $\theta = \arccos(u \cdot v/|u|\,|v|)$, but all this specifies $u \times v$ only up to sign, hence the rule that u, v, and $u \times v$, in this order, should make a "positively oriented" frame (cf. p. 287). This assumes that one of the two classes of frames has been designated as the "positive", or "direct" one.

A.3.1 Grad, rot, and div

This subsection discusses the classical differential operators in relation with these structures.

We pointed out the essential uniqueness of Euclidean space, all Euclidean structures being equivalent via linear transformations. This is so ingrained in us that we forget about the multiplicity of Euclidean metrics, and it may be appropriate to tip the scales the other way for a while.

Consider two different Euclidean structures on A_3, as provided by two different dot products " $\cdot$ " and " $\circ$ ", and denote them E_3 and $\mathbb{E}_3$ respectively. (No orientation yet.) Let $\varphi : A_3 \rightarrow \mathrm{IR}$, a smooth scalar field, be given. Its existence owes nothing to the Euclidean structure, obviously. But what of its gradient? If we define grad φ as the vector field such that $(\text{grad}\ \varphi)(x) \cdot v(x) = \lim_{\lambda \rightarrow 0} (\varphi(x + \lambda v) - \varphi(x))$ —and who would object to that?[36]—then grad φ *does* depend on the Euclidean structure, and we have two different gradients: one, grad, for E_3, and another one, $\mathbb{grad}$ say, for $\mathbb{E}_3$. Coming back to A_3, where the notions of scalar field and vector field do make intrinsic sense, without any need for a metric, we have obtained two differential operators of type *SCALAR_FIELD* → *VECTOR_FIELD*, respectively grad and $\mathbb{grad}$, which are *different.* Of course—and this is where the equivalence of Euclidean structures lets itself be felt—they are closely related: As a consequence of $(\mathbb{grad}\,\varphi)(x) \circ v = (\text{grad}\,\varphi)(x) \cdot v$, one has $L^tL(\mathbb{grad}\,\varphi)(x) = (\text{grad}\,\varphi)(x)$, at all points, hence $L^tL\,\mathbb{grad} = \text{grad}$, which corresponds to a change of basis.

Such equivalence suggests that all these different gradients are mere avatars of a *single,* intrinsically defined, operator that would make sense on A_3. Indeed, this operator exists: It's the "exterior derivative" of differential geometry, denoted d ; but to develop the point would lead us into this "radical way" alluded to in 2.2.3, but not taken.

The situation with the curl operator is even worse, because not only the Euclidean structure, but the orientation of space plays a role in its definition.

Given u, we may—by using the Stokes theorem backwards—define rot u as the vector field such that its flux through a surface equals the circulation of u along this surface's boundary. Both words "through" and "along" refer to orientation, but the former connotes outer orientation of the surface, and the latter, inner orientation of the rim. Since both orientations can be defined independently, defining rot requires they be related in some arbitrary but definite way. When the ambient space is oriented, it becomes possible to establish such a relation (by the corkscrew rule), as we saw in 5.2.1 and p. 287. So it's only in *oriented* three-dimensional Euclidean space that rot makes sense. Another way to express this is to say that for a given dot product, there are *two* curl operators in three-space, one for each possible orientation, which deliver opposite fields when fed with one.

So if we insist on imitating the previous development on E_3, $\mathbb{E}_3$, grad and $\mathbb{grad}$, we must distinguish ${}^+E_3$ and ${}^-E_3$, say, to account for the two possible orientations, as well as ${}^+\mathbb{E}_3$ and ${}^-\mathbb{E}_3$, hence four operators ${}^-$rot, ${}^+$rot, ${}^-\mathbb{rot}$, ${}^+\mathbb{rot}$, of type *VECTOR_FIELD* → *VECTOR_FIELD*, defined in A_3. Again, if the new metric $\circ$ is given by $u \circ v = Lu \cdot Lv$, and if L preserves orientation (which one can always assume), one has[37] ${}^+\mathbb{rot} =$

[36] Well . . . One often sees (grad f)(x) defined as the vector of coordinates $\{\partial_1 f, \partial_2 f, \partial_3 f\}$ at point x. This is a different notion: The entity thus defined is a *covector,* that is to say, an element of the dual of V_3, not a vector. The $\partial_i f$'s are what is called "covariant components" of grad f. Only in the case of an orthonormal frame do they coincide with its ordinary ("contravariant") components.

[37] The new cross-product $\mathbb{\times}$ is then given by $L(u \mathbb{\times} v) = Lu \times Lv$. The reader is challenged to prove the formulas of Fig. A.14 by proper application of the Stokes theorem.

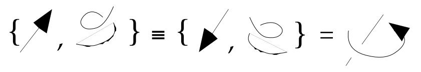
$\det(L^{-1})\ {}^+\mathrm{rot}\, L^tLu$, and this plus the obvious relations ${}^+\mathrm{rot}\, u = -{}^-\mathrm{rot}\, u$ and ${}^+\mathbb{rot}\, u = -{}^-\mathbb{rot}\, u$ makes such changes of metric and orientation manageable (cf. Fig. A.14), but the lesson is clear: *Classical differential operators are definitely impractical*[38] when it comes to such changes. Better stay with the same metric and the same orientation all over. Problems where this is too cumbersome, such as computations involving moving (and possibly, deformable) conductors, call for the more elaborate framework provided by differential geometry, as discussed in 2.2.3.

$$\begin{tikzcd}
\bullet \arrow[r, "\mathrm{grad}"] & \bullet \arrow[r, "\mathrm{rot}"] & \bullet \arrow[r, "\mathrm{div}"] & \bullet \\
\bullet \arrow[r, "\mathbb{grad}"] \arrow[u, "1"] & \bullet \arrow[r, "\mathbb{rot}"] \arrow[u, "L^tL"] & \bullet \arrow[r, "\mathbb{div}"] \arrow[u, "\det(L)"] & \bullet \arrow[u, "\det(L)"]
\end{tikzcd}$$

FIGURE A.14. Relations between the differential operators associated with two different dot products. This is what is called a *commutative diagram*: Each arrow is marked with an operator, and by composing operators along a string of arrows that joins two dots, one obtains something which depends only on the extreme points, not on the path followed. Note that $\mathbb{div}\ v = \mathrm{div}\ v$.

A.3.2 Digression: the so-called "axial vectors"

As if this was not complicated enough, someone invented the following devilish device. Let's start from V_3 with a metric, but no orientation. Using our freedom to create new geometric objects thanks to the mechanism of equivalence relations and classes, let's introduce pairs {v, Or}, where v is a vector, and Or one of the two classes of frames, decree that {v, Or} and {– v, – Or}, where – Or is of course the other class, are equivalent, and call the equivalence class an *axial vector*. To soothe the vexed ordinary vector, call it a *polar* vector. (As suggested in inset, the right icon for an axial vector is not the arrow, but a segment with a sense of rotation around it. Note how, just as a polar vector orients its supporting line, an axial vector "outer orients" this line. Note also that axial and polar vectors can be associated in one-to-one correspondence, but in two different ways, one for each orientation of



[38] Some solace can be found in the invariance of the divergence with respect to changes of metric: div v = $\mathbb{div}$ v. If v is interpreted as the velocity field of a fluid mass, its divergence is the rate of change of the volume occupied by this mass, and though the volume depends on the metric, volume *ratios* do not.

ambient space.) Now define a new operator $\overrightarrow{\mathrm{rot}}$ as follows. Start from a field $\overset{\curvearrowright}{\mathrm{v}}$ of axial vectors. Select a representative {v, Or}, and define $\overrightarrow{\mathrm{rot}}\,\overset{\curvearrowright}{\mathrm{v}}$ as $^{\mathrm{Or}}\mathrm{rot}\ \mathrm{v}$, where $^{\mathrm{Or}}\mathrm{rot}$ is the operator associated with *this* choice of orientation. This is a consistent definition, because $^{-\mathrm{Or}}\mathrm{rot}(-\mathrm{v}) = {}^{\mathrm{Or}}\mathrm{rot}(\mathrm{v})$, and thus $\overrightarrow{\mathrm{rot}}\,\overset{\curvearrowright}{\mathrm{v}}$ is a well-defined *polar* (yes, polar!) vector field. Now, lo and behold, the new operator $\overrightarrow{\mathrm{rot}}$ does not depend on orientation.[39]

To make use of it, one must then confer the axial status to some of the vector fields in Maxwell equations. The electric field, akin to a force, has nothing to do with orientation and is thus polar. Then b *must* be axial, and also h, because of $\mathrm{b} = \mu\mathrm{h}$, and of course d and j are polar. Excessive emphasis on such notions, sometimes combined with obscure considerations of the "axial character" of some physical entities, on "the way vectors behave under mirror reflection", and so on, generates much undue confusion. The tiny advantage of not depending on orientation ($\overrightarrow{\mathrm{rot}}$ continues to depend on the metric, anyway), is thus dearly paid for.

The key to clarity is to stay aware of the distinction between physical entities and their mathematical representations. A vector field is a vector field is a vector field . . . But it often happens to be just an element, the main one but not the only one, in the description of a physical entity, to which other elements, standing in background, also contribute.

For instance, the electric field, as a physical object, can be represented by three mathematical objects, acting in conjunction: affine space, a dot product, and (the main item) a vector field denoted e. The magnetic field, still as a physical object, demands a little more: space, dot product, *an orientation,* and (the main item, again) a vector field b. Among these four elements, the first three can be fixed once and for all, thus forming a background, or "mathematical framework", here symbolized by E_3, which can be used for all electromagnetic entities. Hence the expression of a physical law such as, for instance, Faraday's, as a differential relation between vector fields, namely $\partial_t \mathrm{b} + \mathrm{rot}\ \mathrm{e} = 0$.

However, there is some leeway in the choice of items that will be kept in background. As the concept of axial vector suggests, one may decide *not* to include orientation among them, and have the actors on the stage (now axial vectors and polar vectors, depending) carry this information with them all the time. Hence such orientation-free but also, terribly contrived, formulations as $-\,\partial_t \vec{\mathrm{d}} + \overrightarrow{\mathrm{rot}}\,\overset{\curvearrowright}{\mathrm{h}} = \vec{\mathrm{j}}$, and symmetrically, $\partial_t \overset{\curvearrowright}{\mathrm{b}} + \overset{\curvearrowright}{\mathrm{rot}}\,\vec{\mathrm{e}} = 0$, where $\overset{\curvearrowright}{\mathrm{rot}}$ is the operator of Note 39.

[39] A similar operator, $\overset{\curvearrowright}{\mathrm{rot}}$, also orientation-independent, will act on a polar vector to give an axial one: Just define $\overset{\curvearrowright}{\mathrm{rot}}\,\vec{\mathrm{v}}$ as the class of the pair $\{^{\mathrm{Or}}\mathrm{rot}\ \mathrm{v}, \mathrm{Or}\}$.

But then, why not also bring *metric,* which is at least as versatile as orientation, to the foreground? This is possible by treating b and e as differential forms. Then Faraday's law takes the form $\partial_t b + \mathrm{d}e = 0$, where d is the exterior derivative, which is metric- and orientation-independent. Axial vectors thus appear as an awkward device, which leaves us with a job less than half-done, at the price of considerable conceptual complexity.

A.3.3 The Poincaré lemma

Curl-free fields are gradients, locally, and divergence-free fields are curls. The Poincaré lemma is the precise statement of this well-known and important property.

A domain D of E_3 is *star-shaped* if it contains a privileged point x_0 such that if $x \in D$, then $x_0 + \lambda(x - x_0)$ belongs to D for all $\lambda \in [0, 1]$. One may always select x_0 as origin, which we do in what follows.

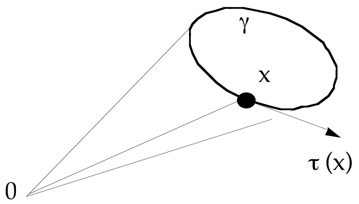
Poincaré's lemma. *Let* e, b, *and* q *be two vector fields and a function, smooth over a star-shaped domain* D, *such that* $\mathrm{rot}\, e = 0$ *and* $\mathrm{div}\, b = 0$ *in* D. *There exists a smooth function* ψ *and smooth fields* a *and* j *such that* $e = \mathrm{grad}\, \psi$, $b = \mathrm{rot}\, a$, *and* $q = \mathrm{div}\, j$, *in all* D.

There are explicit formulas for ψ, a, and j, as follows:

(p1) $$\psi(x) = \int_0^1 x \cdot e(\lambda x)\, d\lambda,$$

(p2) $$a(x) = -\int_0^1 x \times b(\lambda x)\, \lambda\, d\lambda,$$

(p3) $$j(x) = \int_0^1 x\, q(\lambda x)\, \lambda^2\, d\lambda,$$



where x is a fixed point of D and λ the integration variable, and the proof is a verification—not that straightforward. For (p2), for instance, take the circulation of a along a small loop γ (inset), compare the result with the flux of b across the surface of the cone centered at 0 generated by γ, and apply Stokes to the sole of the cone.

Note that an open ball is star-shaped, so the lemma is always valid *locally,* in the neighborhood of a point. What is at stake here is the *global* result ("in *all* D"). It holds in all dimensions, and studying the proof (as given in [BS], after pp. 94–95 of [Sp], or [Co], or [Sc], p. 140) reveals what is important in the hypothesis: not D being star-shaped in the strict sense, but the existence of a *deformation–retract,* that is, a family g_t of maps from D into itself, continuous with respect to t and x, which satisfies $g_1(x) = x$ and $g_0(x) = x_0$ for all $x \in D$. (In the language of Note

33, this is a homotopy between the identity map $x \to x$ and the constant map $x \to x_0$.) A metric space is *contractible* if it can be, so to speak, collapsed onto one of its points by such a deformation-retract (here, $g_t(x) = tx$). The Poincaré lemma is thus valid for contractible domains of E_3, actually, even if simple formulas like (p1)–(p3) may not be available.

All simply connected sets of E_2 are contractible. In E_3, this condition is implied by contractibility, but the latter is stronger. One can prove that bounded simply connected regular domains with a connected boundary are contractible. (This is the criterion we use in Chapter 5.) For simply connected regions with non-connected boundary, it is still true that curl-free fields are gradients, although solenoidal fields may not be curls.

Note that, contrary to what is often lightly asserted, domains where all irrotational fields are global gradients need *not* be simply connected. Figure 8.8, Chapter 8, offers a counter-example.

Formula (p2) is important in electromagnetism, where it is called "Poincaré gauge" [BS]. A "gauge", as we saw on p. 274, is a rule by which, given a solenoidal b, one can select a particular representative in the class of vector potentials a that satisfy $\text{rot } a = b$ —if there is one, which Poincaré lemma shows to be the case in contractible regions. As pointed out in [Sk], the gauge implied by (p2) is the obvious condition $x \cdot a(x) = 0$, which does not coincide with either the Coulomb or the Lorenz gauge. In particular, note that $\text{div } a \neq 0$ in (p2). Poincaré gauge might have useful applications in some modellings [Ma], and should be better known.

The central importance of Poincaré's lemma, however, lies elsewhere: the fact that, for a contractible domain of E_3, the sequence

$$\begin{tikzcd} C^\infty(\overline{D}) \arrow[r, "\text{grad}"] & \mathbb{C}^\infty(\overline{D}) \arrow[r, "\text{rot}"] & \mathbb{C}^\infty(\overline{D}) \arrow[r, "\text{div}"] & C^\infty(\overline{D}) \end{tikzcd}$$

is *exact*, in the sense of Chapter 5 (the codomain of each operator fills out the kernel of the next operator in the sequence). Moreover, when the sequence is *not* exact, i.e., when either of the quotients

$$\ker(\text{rot}\,;\ \mathbb{C}^\infty(\overline{D}))/\text{grad}(C^\infty(\overline{D})), \qquad \ker(\text{div}\,;\ \mathbb{C}^\infty(\overline{D}))/\text{rot}(\mathbb{C}^\infty(\overline{D}))$$

has nonzero dimension, some topological peculiarities of D (presence of "loops" and "holes", respectively, as explained in Chapter 5) can be inferred.

A.3.4 Symmetry in E_3

In most modellings, there is some geometrical symmetry of the domain of interest, that can be exploited to reduce the size of the computational domain, hence substantial economies. The idea is to perform the computation on a subdomain, called the *symmetry cell*, containing only one point per orbit under the action of the symmetry group. Thus, the union of images of the closure of the cell is identical with the closure of the original computational domain. But this supposes a proper setting of boundary conditions on "new boundaries" thus introduced (on symmetry planes, for instance), and for this, the formal notions about symmetry that follow may be helpful.

The *isometries* of a metric space X are the transformations (functions of type $X \to X$, defined over all X) that preserve distances. (This implies bijectivity.) Isometries of E_3 are the rotations, the translations, the mirror symmetries, and their compositions. We'll say an isometry is *skew* or *direct*, according to whether it changes the orientation of a reference frame or not. (Alternatively, one could say *odd* or *even*, but we reserve these words for a different use.)

Let D be a regular bounded domain in E_3.

Definition A.1. *An isometry*[40] i *of* E_3 *is a* symmetry *of domain* D *if it leaves* D *globally unchanged:* $i(D) = D$.

Symmetries of D form of course a group (denoted G_D or simply G in what follows).[41] This group has two elements in the case that first comes to mind when symmetry is mentioned, which is bilateral symmetry: the identity and the mirror symmetry h with respect to a plane Σ (group denoted C_{1h}). But there may be much more: for instance, all the $2\pi/n$ rotations around some straight line a (called a "repetition axis of order n"), group denoted C_n. Other frequently encountered symmetry groups are D_n, C_{nh}, C_{nv}, obtained by combining the rotations in C_n with, respectively, the half-turn around an axis orthogonal to a, the reflection h with respect to a plane orthogonal to a, and the reflection v with respect to a plane containing a, and D_{nh}, which is obtained by composing the rotations of D_n with h. For concrete examples, think of a three-blade propeller (group D_3 or C_3, depending on whether the propeller's action is reversible or not), a triumph arch (C_{2v}), the Eiffel tower (C_{4v}), a brick (D_{2h}).

[40] Some interesting symmetries are not isometries. One may conceive of objects with "fractal" structure, invariant with respect to some non-distance-preserving transformations, dilatations for instance. The exploitation of symmetries of this kind is an open problem.

[41] The little group of D under the action of isometries on parts of E_3.

A symmetry of D is direct or skew according to whether the isometry it comes from is itself direct or skew. Elements of G_D which are direct symmetries form a subgroup of G_D.

Let i be an isometry of E_3. If v is a vector at x which has its tip at y (which is another way to say that $y = x + v$), it is natural to define the transform of v under i as the vector at ix which has its tip at iy, that is, $iy - ix$. We shall denote this vector by $i_* v$.

By restriction to D, one may similarly define the effect of a symmetry s of D on a vector at x, for x in D or its boundary. If now $v = x \to v(x)$ is a vector *field* over D, we'll denote by Sv the transform of v under the action of s, thus defined:

$$(1) \qquad (Sv)(sx) = s_*(v(x)),$$

that is to say, $Sv = x \to s_*(v(s^{-1}x))$. Thus if s is, for instance, the mirror reflection in a plane, and if v is represented, according to a popular (and quite unfortunate) graphic convention, by a bundle of arrows, Sv is imaged by the set of reflections of these arrows. Functions transform under a symmetry the same way vector fields do: If φ is a function defined over D, we may set $(S\varphi)(sx) = \varphi(x)$, on the model of (1), that is $S\varphi = x \to \varphi(s^{-1}x)$ (the "push-forward" of φ by s, cf. Note 7.10). All this suggests the following definition:

Definition A.2. *A symmetry* s *of* D *is a symmetry* of the vector field v *[resp.* of the function φ] *if and only if* $Sv = v$ *[resp.* $S\varphi = \varphi$].

Note how this provides a concrete example of a family of group actions, all different, of the same group, here G_D, on different geometrical objects. General notions as given earlier apply. In particular, the symmetries of a vector field or a function form a subgroup of G_D, denoted G_v or G_φ if a name is needed, called as we know the isotropy group (or little group) of v or φ. By the Stokes theorem, the little group of a function φ [resp. of a field h, a field j] can be embedded[42] in the little group of grad φ [resp. of rot h, of div j].

When we refer to the symmetries of a *problem*, it means more than the symmetries of D. Symmetries of the material properties also should be considered. This is, in all generality, a difficult subject, if one wishes to take into account the deformability of materials, and possible anisotropies. For homogeneous materials and non-changing geometries, however, it's simple. All we have to do is consider the symmetry groups

[42] It's not simply an isomorphism, because rot v, for instance, may be much more symmetrical than v. (Think of some undistinguished v for which rot v = 0.)

of the functions σ, μ, and ε, and take their intersection with G_D. The subgroup of G_D thus obtained is the symmetry group of the problem.

Many symmetries are *involutions*, in the sense that $s^2 = 1$ (the identity): symmetries with respect to a point, a straight line or a plane, are involutions. For these, the following notion applies:

Definition A.3. *A function* φ *is said to be* even *[resp.* odd*] with respect to the involutive symmetry* s if $S\varphi = \varphi$ *[resp.* $S\varphi = -\varphi$]. *A vector field* v *is* even *[resp.* odd] *if* $Sv = v$, *that is to say* $s_* v = v$ *at all points [resp.* $s_* v = -v$].

It's easy to see that if a function is even or odd, its gradient has the same property, and that the divergence of a vector field has the same parity as the field. In contrast, the curl of an even or odd field has *opposite* parity in the case of a *skew* symmetry (reflection with respect to a point or a plane) and the same parity in the case of a direct symmetry (half-turn around some axis). This reflects the "sensitivity to orientation" of rot, as earlier remarked.

These properties rule the setting of boundary conditions, in a quite simple way, at least as far as mirror symmetries are concerned. Suppose (which is the general case) the source of the field is a given current density j^g. If j^g is even[43] [resp. odd], j has the same property, and hence e (at least in conductors) is even [resp. odd], provided σ is even with respect to this mirror. By Faraday's law, b then has the symmetry of rot e, which means odd [resp. even]. And so forth, for all fields. Once the parity of all fields has thus been determined, boundary conditions follow from simple rules: For fields which are, like b, associated with surfaces (fields d and j), the boundary condition is $n \cdot b = 0$ in case of even fields, no condition at all in case of odd fields. For fields like h which are associated with lines (fields e and a), it's the opposite: The boundary condition is $n \times h = 0$ in case of odd fields, no condition at all in case of even fields. Since h and b (or d and e, or j and e) have same parity, boundary conditions on symmetry planes are complementary: $n \cdot b = 0$ on some, $n \times h = 0$ on others. We had a concrete example of this with the Bath-cube problem.

For more on this subject, see [B1, B2].

Let's now give a few other practical examples, also borrowed from the TEAM workshop trove.

[43] It's always possible to express j^g as the sum of an even and an odd component, and to do this repeatedly for all mirror symmetries, thus forming kind of "Fourier components" of the source. One then solves one reduced problem (on the symmetry cell) for each of these components, and adds the results.

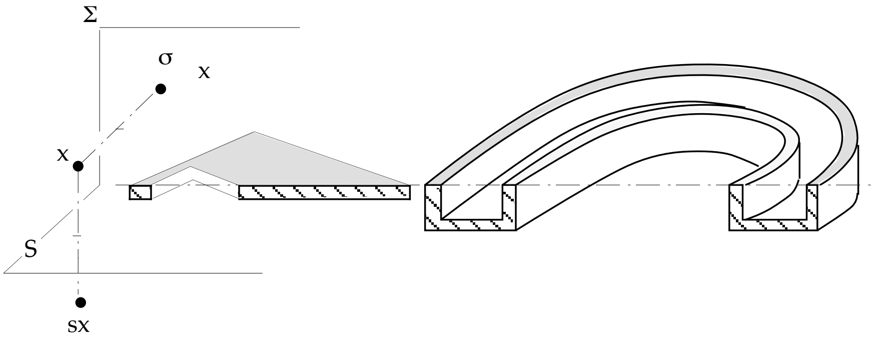


FIGURE A.15. Sketch of TEAM Pbs. 7, 14, with common symmetry D_2. (Only the "material symmetry cell", below S and behind Σ, is shown. This is a part of the passive conductor that generates all of it by letting the symmetry operations act.) The inductor, not represented here, may not share the symmetries of the passive conductor, but this does not impede the exploitation of symmetry [B1]. TEAM Problems 8, on the detection of a crack inside an iron piece, and 19, on a microwave cavity, have the same kind of symmetry.

Problems 7 (the misnamed "asymmetrical" plate with a hole), 8 (coil over a crack) and 14 (the "Euratom casing" [R&]) fall into a category described by the group C_{2v}, with four elements. It is generated by reflections s and σ with respect to two orthogonal planes S and Σ. Its elements are thus $\{1, s, \sigma, s\,\sigma\}$ (Fig. A.15).

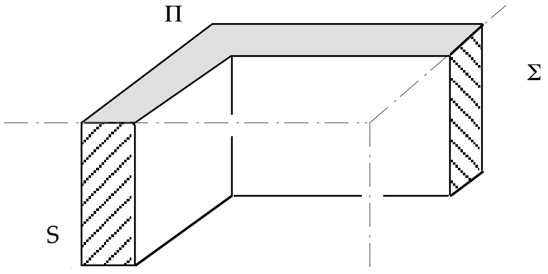


FIGURE A.16. Symmetry D_{2h}, common to Pbs. 3, 4, 1. (The example shown is the symmetry cell of the "Felix brick".)

When the group is generated by reflections s, σ, π with respect to *three* orthogonal planes, it is called D_{2h} (Fig. A.16). It has 8 elements, and is relevant to Pbs. 3 (the "Bath ladder") and 4 (the "Felix brick" [T&]). Problem 12 (the cantilevered flexible plate [C&]) should be included in this category, because all computations relative to it can be conducted in the so-called "reference configuration" of the conductive plate, with negligible error (because of small deformations), and the symmetry group of this reference configuration is D_{2h}.

Some problems are "much more symmetrical" than any of the above, having infinite symmetry groups. The frequent (and never formally explained . . .) references in this book to "2D modelling" have to do with geometrical symmetry: 2D modelling is relevant when the symmetry group of the problem contains all translations along some direction. Symmetry reduction by one spatial dimension also occurs in case of axisymmetry (group SO_2 or larger). For instance, TEAM Pbs. 1 and 2, the "Felix cylinders", and Pb. 9 on the far-field effect in a tube, have symmetry group O_2, composed of all rotations around a fixed axis, combined with reflections with respect to an axial plane.

Even more extended symmetry can happen. Problems 6 and 11 on eddy currents induced in a hollow sphere have symmetry group O_3: all rotations around a fixed point, combined with reflections with respect to the origin. Fourier series is the right (and well known) tool for such cases. It goes quite far, up to giving *exact* solutions, by formulas, in some cases (the hollow sphere, for instance).

A.4 GLIMPSES OF FUNCTIONAL ANALYSIS

A.4.1 Completion, principle of extension by continuity

Cauchy sequences in a metric space $\{X, d\}$ are sequences $\{x_n : n \in \mathbb{N}\}$ such that $d(x_n, x_m)$ tends to zero when both indices n and m tend to infinity. Convergent sequences are Cauchy. A space is *complete* if, conversely, all its Cauchy sequences converge. A normed vector space with this property is called a *Banach space.*

In applied mathematics, the only good spaces are complete spaces, as we experienced in Chapter 3. So let's give in full this construction of complete spaces that proved so important then:

Theorem A.1 (of completion). *Given a metric space* $\{X, d\}$, *there exists a space* $\{\hat{X}, \hat{d}\}$ *and an isometry* $i \in X \to \hat{X}$, *such that* $\hat{X}$ *be complete and* cod(i) *dense in* $\hat{X}$.

Proof. The key proof-ideas were given in 3.2.3, and we just fill in details. $X°$ being the set of all Cauchy sequences $x° = \{x_1, \ldots, x_n, \ldots\}$ of elements of X, set $x° \sim y°$ if $\lim_{n \to \infty} d(x_n, y_n) = 0$. This is an equivalence, because $x° \sim y°$ and $y° \sim z°$ imply that $d(x_n, z_n) \leq d(x_n, y_n) + d(y_n, z_n)$, tends to 0, by the triangular inequality. Now define $\hat{X}$ as the quotient $X°/\sim$, and set

$\hat{d}(\hat{x}, \hat{y}) = \lim_{n \to \infty} d(x_n, y_n)$. Then $\hat{d}(\hat{x}, \hat{z}) \le \hat{d}(\hat{x}, \hat{y}) + \hat{d}(\hat{y}, \hat{z})$, still by the triangular inequality, it's obvious that $\hat{d}(\hat{y}, \hat{x}) = \hat{d}(\hat{x}, \hat{y})$, and if $\hat{d}(\hat{x}, \hat{y}) = 0$, classes $\hat{x}$ and $\hat{y}$ coincide, since two representatives $x°$ and $y°$ will satisfy $\lim_{n \to \infty} d(x_n, y_n) = 0$ and thus be equivalent. Let $i(x)$ be the sequence $\{x, \dots, x, \dots\}$. Then $\hat{d}(i(x), i(y)) = d(x, y)$, so i is an isometry. The image $i(X)$ is dense in $\hat{X}$, because if $x° = \{x_1, \dots, x_n, \dots\}$ is a representative of $\hat{x}$, then $\hat{d}(\hat{x}, i(x_n)) = \lim_{m \to \infty} d(x_m, x_n)$, which tends to 0 by definition of a Cauchy sequence. Finally, if $\{\hat{x}^n : n \in \mathbb{N}\}$ is a Cauchy sequence of $\hat{X}$, select a sequence $\{\varepsilon_n : n \in \mathbb{N}\}$ of reals which tend to 0, and for each n, choose $x_n \in X$ such that $\hat{d}(i(x_n), \hat{x}^n) \le \varepsilon_n$, which the density of $i(X)$ makes possible. As $d(x_n, x_m) = \hat{d}(i(x_n), i(x_m)) \le \hat{d}(i(x_n), \hat{x}^n) + \hat{d}(\hat{x}^n, \hat{x}^m) + \hat{d}(\hat{x}^m, i(x_m)) \le \varepsilon_n + \hat{d}(\hat{x}^n, \hat{x}^m) + \varepsilon_m$, which tends to 0, the x_ns form a Cauchy sequence. Let $\hat{x}$ be its class. Then $\hat{d}((\hat{x}, \hat{x}^n) \le \hat{d}(\hat{x}, i(x_n)) + \hat{d}(i(x_n), \hat{x}^n) \le \lim_m \hat{d}(x_m, x_n) + \varepsilon_n$, which goes to 0 as n increases, showing that $\hat{x}$ is the limit of $\{\hat{x}^n\}$. ◊

Note that one can legitimately refer to *the* completion, because if one can find, by some other method, another dense injection j of X into some complete space $X^{\wedge}$, then elements of $\hat{X}$ and $X^{\wedge}$ are in isometric correspondence, so the completion is unique up to isometry. The proof is *constructive*, giving us one of these isometric complete spaces in explicit form. One can argue that $\hat{X}$ is not necessarily "the right one", however. Indeed, our intuitive notion of completion seems to require embedding X into a space made of objects of the *same* type as those of X. Hence the search, in most cases, for such a "concrete" complete space. For instance, if X is a space of functions defined on a domain of E_3, one will try[44] to identify its completion with a similar functional space. An important example will be given below, where $L^2(D)$, the completion of $C(D)$, is embedded in a space of functions defined on D, thanks to Lebesgue integration theory.

There is a companion result to the completion theorem:

Theorem A.2 (of extension by continuity). *Let* X *and* Y *be metric spaces, both complete,* U *a* dense *part of* X, *and* $f_U \in X \to Y$, *with* $\mathrm{dom}(f_U) = U$, *a* uniformly *continuous function. There is an extension* f *of* f_U *to all* X *that is continuous, and it's the only one.*

Proof. Take $x \in X$ and let $\{x_n \in U\}$ be a sequence that converges to x. Because of *uniform* continuity, the $f_U(x_n)$ form a Cauchy sequence, which converges, since Y is complete, towards a point that one can denote $f(x)$, because it does not depend on the chosen sequence. As $f(x) = f_U(x)$ if $x \in U$,

[44] And when this fails, never mind: A cunning extension of the very notion of function will often save the day.

one thus obtains an extension of f_U the domain of which is all X, and one easily checks that f is (uniformly) continuous. If g is another continuous extension of f_U, then $\lim_{n \to \infty} g(x_n) = g(x)$ by continuity, so $g(x) = f(x)$. ◊

Obviously, continuity of f_U might not be enough (Fig. A.17).

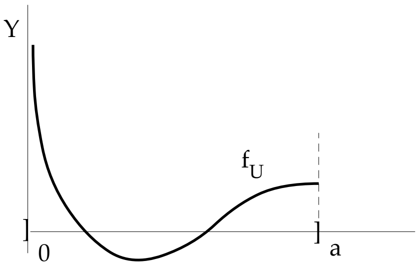


FIGURE A.17. Here, $X = [0, a]$, part of IR, $U =]0, a]$ and $Y =$ IR. In spite of its continuity, f_U has no continuous extension to X.

This result is almost often applied to the extension of *linear* (or affine) maps, between normed spaces, and *then* continuity is enough, because affine continuous maps are uniformly continuous. It works as follows: When a linear map (or as one prefers to say then, an "operator") $L : X \to Y$ of domain U is continuous, one can extend it into an operator from the completion X of U to the completion of Y, since U is dense in X; just apply the previous result to the composition $i_Y \circ L$, where i_Y is the canonical injection of Y into its completion.

A.4.2 Integration

I assume you know about integration, though not necessarily about *Lebesgue* integration *theory*. It's an ample and difficult theory, which cannot even be sketched here. And yet, some of its results are absolutely essential when it comes to weak formulations, complete spaces, existence proofs, etc. Fortunately, one can live in blissful ignorance of the theory, provided one is aware of what it does better than the older and (only apparently) easier Riemann theory.[45]

What Riemann's theory does, and does fairly well, is to give sense to the concept of *average value* of a *continuous* function over a set where concepts such as "length", "area", or "volume" make sense. (The generic

[45] As stressed in [Bo], the standard comment, "in one case you divide up the x-axis and in the other you divide up the y-axis", is totally misleading in its emphasis on a tiny technical difference.

term is "measure". For instance, on the real line, the measure of an interval [a, b] is $|b - a|$, in both theories.) After some work on so-called "Riemann sums", one obtains a sensible definition of the integral $I([a, b], f) = \int_{[a, b]} f$ of f over [a, b], also denoted $\int_{[a, b]} f(x)\, dx$, or $\int_a^b f(x)\, dx$, which embodies the concept of "area under the graph", when $f \geq 0$. (The average is then $I / |b - a|$.) Extensions to sets other than intervals, and to several variables, then follow; hence a map, the type of which is

$$PART\_OF\_A\_MEASURED\_SPACE \times FUNCTION \rightarrow REAL,$$

with the right properties: additivity with respect to the set, linearity with respect to the function. The integral of f over A is denoted $\int_A f$.

Remark A.6. This reduced notation, recommended in Note 1.8 and largely used in this book, reflects the functional character of integration: All that is left is the operator symbol $\int$ and the two arguments, A and f. There is no ambiguity when A is a part of a set X on which exists a standard measure (which is the case of E_3), and if A is all X, one may even not mention it. Developed notation such as $\int_X f(x)\, dx$ or $\int_X f(x)\, d\mu(x)$ may be useful when one must be explicit about the underlying measure (because several of them can appear simultaneously, for instance, or to sort out multiple integrals: several examples appear in Chapter 1), but in such cases, x is a bound variable, that must appear (at least) twice in an expression, as argument of the function *and* of the measure element. Expressions like $\int_X f(x)\, dV$, for instance, are not well-formed in this respect, and should not be used. ◊

There is however an essential flaw in this theory. When a sequence of functions f_n converges pointwise[46] towards some function f, one cannot assert that $\int_X f = \lim_{n \to \infty} \int_X f_n$, if only because the limit f may be outside the domain of the above mapping, and thus not be "integrable in the Riemann sense". Because of this shortcoming, one cannot safely permute integration and passage to the limit, like this: $\int_X \lim_{n \to \infty} f = \lim_{n \to \infty} \int_X f_n$. The Lebesgue theory corrects that by enlarging the domain of the map: There are more functions integrable "in the Lebesgue sense", on more exotic sets. This advantage, by itself, is marginal, for it's not so often that one *must* compute the average of an everywhere discontinuous function on a Cantor set, or similar. The point is elsewhere: In Lebesgue theory, one *can* permute limit and integration, under the condition of *dominated convergence*, that is, when there exists a function g, integrable itself, such that $|f_n(x)| \leq g(x)$ whatever x and n. This commutativity between two such fundamental operations is the great triumph of Lebesgue's theory,

[46] That is, $f_n(x)$ tends to f(x), as a sequence of real numbers, for a fixed x.

because it legitimizes a series of basic manipulations in calculus: differentiation under a summation sign (hence the possibility to permute differentiation and convolution alluded to p. 290), change of order of summation in multiple integrals (Fubini's theorem), and so forth.

How is this achieved? Very roughly, take the space of continuous functions C(X) on some metric space X, and give it a norm, by setting $\| f \|_1 = \int_X | f(x) | \, dx$, where the integral is understood in the Riemann sense. This is not a *complete* space (as discussed in 3.2.1). So take its completion, call this enlarged space $L^1(X)$, and now that every Cauchy sequence does converge, define the integral of the limit f of $\{f_n\}$ (which has just been defined into existence by the process of completion) as the limit of the integrals $\int f_n$. It's so simple that one may wonder where the difficulty is that makes books so thick [Ha, Lo]: Again, it comes from the completion being an abstract space, not a priori a functional one (the above limit f is an abstract object, not yet a function), and the hard work consists in embedding this abstract completion $L^1(X)$ into some functional space.

One thus introduces (after a copious measure of measure theory) a concept of "measurable function", which is very encompassing,[47] and an equivalence relation, the "almost everywhere equal (a.e.)" relation alluded to at places in this book: $f \overset{a.e.}{=} g$ (or $f(x) = g(x)$ a.e.) if points x where $f(x) \neq g(x)$ form a "negligible" set, that is, one to which Lebesgue's measure theory attributes the measure 0. Once all this, which is an impressive piece of work, is said and done, one can identify the elements of $L^1(X)$ with *equivalence classes* of a.e.-equal measurable functions. Yet, one continues to call "functions" the elements of $L^1(X)$, and this abuse is natural enough: Two almost everywhere equal functions belong to the same class and have *the same integral*, so from the point of view of integration theory they are "the same" indeed. This is even more justified when one realizes that a *continuous* function is *alone* in its own class (because two a.e.-equal continuous functions must coincide).

Once in possession of $L^1(X)$, one can define $L^2(X)$ as the space of "functions" the square of which is in $L^1(X)$, or more precisely, as the completion of C(X) with respect to the *quadratic* norm, $\|f\|_2 = (\int |f|^2)^{1/2}$ instead of the L^1-norm $\| f \|_1 = \int | f |$. This L^2-norm is associated with a scalar product, namely $(f, g) = \int f(x) \, g(x) \, dx \equiv \int fg$, so L^2 is pre-Hilbertian, and being also complete, is a Hilbert space, *the* essential concrete realization of this abstract notion.

[47] It's such a large class that no constructive examples of *non*-measurable functions exist; one must invoke the axiom of choice to get them.

Most of the time, it's convenient to think of elements of $L^2(X)$ as functions, though they are actually *classes* of functions. But there is one case in which awareness of the real nature of L^2 is important: when one tries to define the restriction of a function of $L^2(D)$, where D is our usual "computational domain", to its boundary S. The boundary being a negligible set, values of f on S can be changed at will without changing the *class* f belongs to, which means that "restriction to S" is a meaningless expression as regards elements of $L^2(D)$. And yet we need such a concept to deal with boundary-value problems! Hence the introduction of the relatively sophisticated notion of *trace*: The trace γf of a continuous function f is just its restriction to S. Now if f is a generic element of $L^2(D)$, there is, by construction, a Cauchy sequence of continuous functions f_n which tends towards f in quadratic norm. So we define γf, the trace of f, as the limit in $L^2(S)$ of the sequence of restrictions γf_n, *provided this sequence converges*, which may not be the case: Some "functions" of $L^2(D)$ have traces, some have not. The question is discussed in Chapter 7, where it is shown in detail how functions the gradient of which (in the weak sense) is square summable in D do have traces, even though they need not be continuous. All this makes only the beginning of the (difficult) theory of Sobolev spaces, but what precedes is enough baggage for our needs.

Apart from this all-important extension of scope, Lebesgue theory does not bring anything new when it comes to the more mundane aspects of integration as used in calculus, such as integration by parts, change of variables, and the like. Let's just stress two points of special importance, the definition of *circulations* and *fluxes*.

Let c denote a bounded curved line in E_3. On c, the Euclidean distance existing in E_3 induces a notion of length of curved segments, which turns c into a measured space, on which integration makes sense: If f is a function whose domain contains c, the integral $\int_c f$ is the average of f on c, multiplied by the length of c.

Now, let's equip c with a field of tangent vectors. For this, take a parameterization of c, that is to say, a smooth map, still denoted c, from $[0, 1]$ into E_3, having the curved line as codomain. (The deliberate confusion between the path $c : [0, 1] \rightarrow E_3$ and the curve proper, which is only the codomain of this path, has obvious advantages, provided one stays aware of the distinction.) Assume the derivative $\partial_t c(t)$, which is a vector of V_3, does not vanish for $t \in [0, 1]$. Set $\tau(t) = \partial_t c(t)/|\partial_t c(t)|$: This is the *unit tangent vector* at point $c(t)$. (Obviously, whatever the parameterization, there are only two possible fields τ, each corresponding to one of the two possible orientations of c. Cf. p. 287 and 5.2.1.)

Finally, let u be a smooth vector field, the domain of which contains c. By taking the dot product $\tau(x) \cdot u(x)$ for each point x of c, one obtains a smooth real-valued function of domain c, naturally denoted by $\tau \cdot u$. This function can be integrated on c, hence a number $\int_c \tau \cdot u$. This is, by definition, the *circulation of* the vector field u *along* c, *as oriented by* τ. (Of course it reverses sign with orientation.)

The same things exactly can be said about a smooth patch C, mapping $[0, 1] \times [0, 1]$ into E_3, and such that vectors $\partial_s C$ and $\partial_t C$ at point $C(s, t)$ don't vanish. One then forms a normal field $n(s, t) = N(s, t)/ |N(s, t)|$, where $N(s, t) = \partial_s C \times \partial_t C$, again with only two possible outcomes, corresponding to orientations of C. Again, $n \cdot u$ is a scalar function on C, whose integral $\int_C n \cdot u$ is called the *flux* of u *through* C *as oriented by* n. By sewing patches together, and orienting them consistently, one can thus define fluxes relative to smooth orientable surfaces. This is the case, in particular, of the surface S of a computational domain D, and we have often had to deal with integrals like $\int_S n \cdot u$, especially when using the two basic integration by parts formulas, established in 2.3.1 and 2.3.2:

(2) $$\int_D \varphi \operatorname{div} b = -\int_D b \cdot \operatorname{grad} \varphi + \int_S n \cdot b \ \varphi,$$

(3) $$\int_D h \cdot \operatorname{rot} a = \int_D a \cdot \operatorname{rot} h - \int_S n \times h \cdot a.$$

These formulas concern smooth fields, but thanks to the good behavior of Lebesgue integrals with respect to passages to the limit, one can extend these formulas by continuity to $\varphi \in L^2_{grad}(D)$, $b \in \mathbb{L}^2_{div}(D)$, h and a in $\mathbb{L}^2_{rot}(D)$, as defined in Chapter 5, thus giving them enlarged validity. See Section 5.1 for this important development.

A.4.3 Hilbert spaces

A real[48] vector space X is *pre-Hilbertian* when equipped with a scalar product, as previously defined. The function $\| \ \| = x \rightarrow (x, x)^{1/2}$ is then a norm, which confers a metric on X. (The triangular inequality comes from

(4) $$|(x, y)| \leq \|x\| \, \|y\|,$$

which is the *Cauchy–Schwarz* inequality.) Note that $(\ , \ )$ is continuous with respect to both its arguments. A simple computation yields the following *parallelogram equality*:

[48] That is, built on $\mathbb{R}$ as scalar field. Complex spaces are not less important, but there is some gain in simplicity in treating them apart, as we do a little later.

$$(5) \qquad \|x - y\|^2 + \|x + y\|^2 = 2(\|x\|^2 + \|y\|^2) \qquad \forall\, x, y \in X.$$

The existence of a scalar product gives sense to the notions of *orthogonality* in X (x and y are orthogonal if $(x, y) = 0$, which one may denote $x \perp y$) and *angle* (the angle of two nonzero vectors x and y is $\arccos((x, y)/\|x\|\,\|y\|)$), so all the concepts of Euclidean geometry make sense: The Pythagoras theorem holds, and (5) is nothing else than a generalization of the metric relation between median and sides in elementary geometry of the triangle (Fig. A.18). Such things cannot be said of any normed vector space, only if (5) is valid for the given norm $\|\ \|$, for then one can prove that $\{x, y\} \to (\|x + y\|^2 - \|x - y\|^2)/4$ is a scalar product. Pre-Hilbertian spaces, and their affine associates, are therefore those spaces in which notions and concepts of ordinary Euclidean geometry hold, without any restriction on the dimension: *their theory extends intuitive geometry to infinite dimension.*

A *Hilbert space* is a *complete* pre-Hilbertian space, and we saw many examples, almost all of them related with the spaces L^2 or $\mathbb{L}^2$.

The basic result about Hilbert spaces is this:

Theorem A.3 (of projection). *Let* C *be a* closed convex *part of a* Hilbert *space* X. *The function "distance to* C*", i.e.,* $d_C = x \to \inf\{\|x - y\| : y \in C\}$, *reaches its lower bound at a unique point of* C, *called the* projection *of* x *on* C, *here denoted* $p_C(x)$.

Proof. Most of the proof appears in 3.2.1, the only difference being that there, C was not only convex but an affine subspace. In particular, the key concept of *minimizing sequence* was introduced there. So let's be terse: The lower bound $d = d_C(x) = \inf\{\|x - y\| : y \in C\}$ can't be reached, if it is reached at all, at more than one point, for if $\|x - y\| = \|x - z\| = d$ for $y \neq z$, then $u = (y + z)/2$ would belong to C by convexity, whereas $\|x - u\| < d$ after (5), hence a contradiction. As for existence, let $y_n \in C$ be a minimizing sequence, i.e., $\|x - y_n\|$ converges towards $d_C(x)$. It's a Cauchy sequence, because

$$\|y_n - y_m\|^2 + 4\,d^2 \leq \|y_n - y_m\|^2 + 4\,\|x - (y_n + y_m)/2\|^2$$
$$= 2(\|x - y_n\|^2 + \|x - y_m\|^2),$$

thanks to (5) and to the convexity of C, and the right-hand side tends to $4d^2$. Since X is complete, there is a limit, which belongs to C, since C is closed. ◊

Remark A.7. The inequality that characterizes the projection, that is

$$(6) \qquad \|p_C(x) - x\|^2 \leq \|y - x\|^2 \quad \forall\, y \in C,$$

can also be written as (develop the scalar product)

$$(7)\qquad (x - p_C(x), y - p_C(x)) \leq 0 \quad \forall y \in C.$$

This is called a "variational inequality", or *variational inequation*, if considered as the problem "given x, find $p_C(x)$". Observe how this is "read off" Fig. A.18, right, confirming the remark on Hilbertian geometry as the natural extension of Euclidean geometry to infinite dimensions. Equation (7) is called the *Euler equation* of the *variational problem* (6). ◊

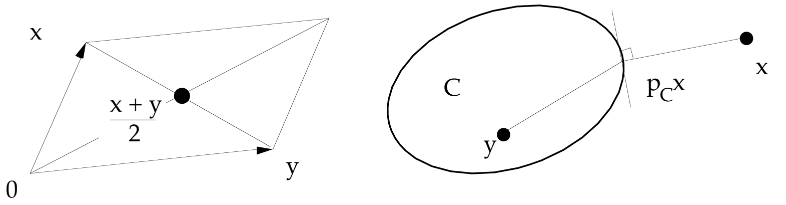


FIGURE A.18. The parallelogram equality (left) and inequality (7).

Remark A.8. The map p_C is a *contraction*, in the sense that

$$\|p_C(x) - p_C(y)\| \leq \|x - y\| \quad \forall x, y \in X.$$

To see it, replace y by $p_C(y)$ in (7), permute x and y, add, and apply Cauchy–Schwarz. ◊

In the particular case when C is a closed subspace Y of X, (7) becomes an equality, or variational equation:

$$(x - p_Y x, y) = 0 \qquad \forall y \in Y.$$

The vector subspace formed by all elements of X orthogonal to Y is called the *orthocomplement* of Y, or more simply, its *orthogonal*, denoted $Y^\perp$. It is closed, as easily checked (cf. Remark A.9). One can therefore apply Theorem A.3 to it, and the projection of x on $Y^\perp$ appears to be $x - p_Y x$. Thus, any x in X can be written as the sum of two orthogonal vectors, one in Y, one in its orthocomplement. Moreover, this decomposition is unique, for $y_1 + z_1 = y_2 + z_2$, with y_i in Y and z_i in $Y^\perp$, $i = 1, 2$, implies $y_1 - y_2 = z_2 - z_1$ at the same time as $y_1 - y_2 \perp z_2 - z_1$, and hence $y_1 = y_2$ and $z_1 = z_2$. One says that Y and $Y^\perp$ have X as *direct sum*, and this is denoted $X = Y \oplus Y^\perp$. Note that $Y^{\perp\perp} = Y$.

Remark A.9. If the subspace Y is *not* closed, one may still define its orthocomplement by $Y^\perp = \{z \in X : (y, z) = 0 \ \ \forall y \in Y\}$. It's closed, because if $z_n \in Y^\perp$ converges to z, then $(y, z) = \lim_{n \to \infty} (y, z_n) = 0$ for all y in Y, by continuity of the scalar product. By applying the projection theorem to $Y^{\perp}$, one sees that $Y^{\perp\perp}$ is not Y, but its *closure* in X. ◊

A second special case is when Y is the kernel of a linear continuous functional $f : X \to \mathbb{R}$. Then Y is closed indeed, and does not coincide with X if f is not trivial, so there exists in $Y^{\perp}$ some nonzero vector z. The equality $x = x - \theta z + \theta z$ then holds for all x and all real θ. But $x - \theta z$ belongs to Y if $\theta = f(x)/f(z)$, so $z \perp (x - \theta z)$ for this value θ, and hence $(x, z) = \theta\, \|z\|^2$, that is, finally,

$$f(x) = (x, z\, f(z)/\|z\|^2) \quad \forall\, x \in X.$$

So there exists a vector z_f that "represents f ", in the precise sense that its scalar product with x is f(x), and this vector is $z_f = z\, f(z)/\|z\|^2$. Moreover (apply (4)), $\|z_f\| = \sup\{|f(x)| / \|x\| : x \neq 0\}$, that is $\|z_f\| = \|f\|$. The correspondence between f and z_f thus achieved is therefore a linear isometry, and we may conclude:

Theorem A.4 (F. Riesz). *To each linear continuous real-valued function* f *on a real Hilbert space* X, *there corresponds a unique vector* z_f *such that* $f(x) = (x, z_f)\ \forall\, x \in X$, *and* $\|f\| = \|z_f\|$.

In this respect, a Hilbert space is "its own dual". But beware there can be other isomorphisms between a concrete Hilbert space and its dual than the Riesz one, which is both an asset (one can solve boundary-value problems that way) and an inexhaustible source of puzzlement. See for example the several isomorphisms between $H^{1/2}(S)$ and its dual in Section 7.4.

Third special case: when C is some *affine* closed subspace X^g, with X^0 as parallel vector subspace, and the point to be projected is the origin. Calling x the projection, we see that x solves the problem, *find* $x \in X^g$ *such that* $(x, x') = 0\ \ \forall\, x' \in X^0$. As the slight change in notation should help one to realize, this result is the paradigm of our existence proofs in Chapters 4 and 6: By adopting the energy-related scalar product, we were able to apply the projection theorem directly in this form. It's not always convenient, however, and the following generalization then comes handy:

Lax–Milgram's lemma. *Let* $a : X \times X \to \mathbb{R}$ *be a bilinear map, continuous with respect to both arguments, and such that*

$$(8) \qquad a(x, x) \geq \alpha\, \|x\|^2 \quad \forall\, x \in X,$$

where α *is a strictly positive real number* (coercivity *of* a). *Given a linear continuous functional* $f \in X \to \mathbb{R}$, *the problem* find $x \in X$ such that

$$(9) \qquad a(x, x') = f(x') \quad \forall\, x' \in X$$

has a unique solution x_f, *and the mapping* $f \rightarrow x_f$ *is* continuous.

Proof. Since $x' \rightarrow a(x, x')$ is continuous, there exists, by the Riesz theorem, some element of X, which can be denoted Ax (a single symbol, for the time being), such that $(Ax, x') = a(x, x')$ for all x'. This defines a linear continuous operator A from X into itself, injective by virtue of (8), and Eq. (9) can then be written as $Ax = z_f$, where z_f is the Riesz vector of f. This is equivalent to $x - \rho(Ax - z_f) = x$, where $\rho \neq 0$ is a parameter that can be chosen at leisure. Let $\{x_n\}$ be the sequence defined by $x_0 = 0$ and $x_{n+1} = (1 - \rho A)x_n + \rho z_f$. If it does converge, the limit is the solution x_f of $Ax = z_f$, and $\alpha \|x_f\| \leq \|z_f\| \equiv \|f\|$ after (9), hence the continuity of $f \rightarrow x_f$. The sequence will converge if $\|1 - \rho A\| < 1$, so let's compute:

$$\|x - \rho Ax\|^2 \leq \|x\|^2 - 2\rho\,(Ax, x) + \rho^2\,\|Ax\|^2 \leq \|x\|^2 - 2\rho\alpha\|x\|^2 + \rho^2\,\|Ax\|^2$$

after (8), so $\|1 - \rho A\| < 1$ if $0 < \rho < \|A\|^2/2\alpha$. (Note that no *symmetry* of a was assumed or used.) ◊

The standard application is then to the problem, *find* $x \in U^g$ *such that* $a(x, x') = 0 \ \forall x' \in U^0$, where a is a continuous bilinear map. By picking some x^g in U^g, this amounts to finding x in U^0 such that $a(x^0 + x^g, x') = 0 \ \forall x' \in U^0$. As seen by setting $f(x') = -a(x^g, x')$ and $X = U^0$, the lemma applies if the restriction of a to U^0 is coercive.

As mentioned in Note 48, the need arises to extend all these notions and results to complex spaces. This is most easily, if not most compactly, done by *complexification*. The *complexified* U^c of a vector space U is the set $U \times U$ with composition laws induced by the following prescription: An element $\mathrm{U} = \{u_R, u_I\}$ of U^c being written in the form $\mathrm{U} = u_R + iu_I$, one applies the usual rules of algebra, with $i^2 = -1$. Thus, $\mathrm{U} + \mathrm{U}' = u_R + iu_I + u'_R + iu'_I = u_R + u'_R + i(u_I + u'_I)$, and if $\Lambda = \lambda_R + i\,\lambda_I$, then

$$\Lambda\mathrm{U} = (\lambda_R + i\,\lambda_I)(u_R + iu_I) = \lambda_R u_R - \lambda_I u_I + i(\lambda_R u_I - \lambda_I u_R).$$

The *Hermitian* scalar product (U, V) of two complex vectors $\mathrm{U} = u_R + iu_I$ and $\mathrm{V} = v_R + iv_I$ is by convention the one obtained by developing the product $(u_R + iu_I, v_R - iv_I)$ after the same rules, so $(\mathrm{U}, \mathrm{V}) = (u_R, v_R) + (u_I, v_I) + i(u_I, v_R) - i\,(u_R, v_I)$. The norm of U is given by $|\mathrm{U}|^2 = (\mathrm{U}, \mathrm{U})$. (Be aware that a different convention is adopted in Chapters 8, where expressions such as $(\mathrm{rot}\,\mathrm{U})^2$ are understood as $\mathrm{rot}\,\mathrm{U} \cdot \mathrm{rot}\,\mathrm{U}$, not as $|\mathrm{rot}\,\mathrm{U}|^2$.)

Now, when X is complex, all things said up to now remain valid, if (x, y) is understood as the Hermitian scalar product, with obvious adjustments: f is *complex*-valued, and the Riesz vector is no longer linear, but *anti*-linear with respect to f (to multiply f by λ multiplies x_f by λ^*). The form a in the Lax–Milgram lemma becomes "sesqui"-linear (anti-linear with respect to the second argument), and the same computation as above yields the same result, provided

$$\mathrm{Re}[a(x, x)] \geq \alpha \,\|x\|^2 \quad \forall\, x \in X,$$

with $\alpha > 0$, which is what "coercive" means in the complex case.

Remark A.10. The lemma remains valid if λa is coercive, in this sense, for some complex number λ. We make use of this in 8.1.3, where the problem is of the form *find* $x \in U^g$ *such that* $a(x, x') = 0 \quad \forall\, x' \in U^0$, with $(1 - i)a$ coercive over U^0. ◊

The theory does not stop there. Next steps would be about orthonormal bases and Fourier coefficients, whose treatment here would be out of proportion with the requirements of the main text. Let's just mention (because it is used once in Chapter 9) the notion of weak convergence: A sequence $\{x_n : n \in \mathbb{N}\}$ *weakly converges* toward x if

$$\lim_{n \to \infty} (x_n, y) = (x, y) \quad \forall\, y \in X.$$

This is usually denoted by $x_n \rightharpoonup x$. By continuity of the scalar product, convergence in the standard sense (then named "strong convergence" for contrast) implies weak convergence, but not the other way around: for instance, the sequence of functions $x \to \sin nx$, defined on $[-1, 1]$, converges to 0 weakly, but not strongly. However, weak convergence *plus* convergence of the *norm* is equivalent to strong convergence.

Compact operators are those that map weakly convergent sequences to strongly convergent ones. It's not possible to do justice to their theory here. Let's just informally mention that, just as Hilbert space is what most closely resembles Euclidean space among infinite-dimensional functional spaces, compact operators are the closest kin to matrices in infinite dimension, with in particular similar spectral properties (existence of eigenvalues and associated eigenvectors). An important result in this theory, *Fredholm's alternative*, is used in Chapter 9. Cf. (for instance) [Yo] on this.

A.4.4 Closed linear relations, adjoints

The notion of *adjoint* is essential to a full understanding of the relations between grad and div, the peculiarities of rot, and integration by parts formulas involving these operators.

We know (p. 284) what a linear relation $A : X \to Y$ is: one the graph $\mathcal{A}$ of which is a subspace of the vector space $X \times Y$. If the relation is functional, i.e., if the section $\mathcal{A}_x$ contains no more than one element, we have a linear operator. By linearity, this amounts to saying that the only pair in X × Y of the form {0, y} that may belong to $\mathcal{A}$ is {0, 0}.

Suppose now X and Y are Hilbert spaces, with respective scalar products $(\ ,\ )_X$ and $(\ ,\ )_Y$. Whether $\mathcal{A}$ is closed, with respect to the metric induced on X × Y by the scalar product $(\{x, y\}, \{x', y'\}) = (x, x')_X + (y, y')_Y$, is a legitimate question. If A is continuous, its graph is certainly closed, for if a sequence of pairs $\{x_n, Ax_n\}$ belonging to $\mathcal{A}$ converges to some pair {x, y}, then y = Ax. The converse is not true (Remark A.4), so we are led to introduce the notion of *closed* operator, as one the graph of which is closed.

Now if the graph of a linear relation {X, Y, $\mathcal{A}$} is not closed, why not consider its *closure* {X, Y, $\overline{\mathcal{A}}$}? We get a new relation this way, which is an extension of the given one. But it may fail to be functional, because pairs of the form {0, y} with y ≠ 0 may happen to be adherent to $\mathcal{A}$. Hence the following definition: An operator is *closable* if the closure of this graph is functional. In Chapter 5, we work out in detail the case of div: $\mathbb{L}^2(D) \rightarrow L^2(D)$, with domain $\mathbb{C}^\infty(\overline{D})$, find it closable, and define the "weak" divergence as its closure. The new operator thus obtained has an enlarged domain (denoted $\mathbb{L}^2_{\text{div}}(D)$) and is, of course, closed, but not continuous on $\mathbb{L}^2(D)$.

There is a way to systematically obtain closed operators. Start from some operator A, and take the orthogonal $\mathcal{A}^\perp$ of its graph in X × Y. This is, as we know, a closed subspace of the Cartesian product. Now consider the relation {Y, X, $\mathcal{A}^\perp$ }, with X as target space. *If* this happens to be a functional relation, we denote by – A* the corresponding operator, which thus will satisfy the identity

$$(11) \qquad (x, A^*\, y)_X = (y, Ax)_Y \quad \forall\ \{x, y\} \in \mathcal{A},$$

and call A*—an operator of type Y → X—the *adjoint*[49] of A.

So when is $\mathcal{A}^\perp$ functional? The following statement gives the answer:

Proposition A.1. *Let* A = {X, Y, $\mathcal{A}$ } *be a given linear relation. The relation* {Y, X, $\mathcal{A}^\perp$} *is functional if and only if* dom(A) *is dense in* X.

Proof. If $\{x, 0\} \in \mathcal{A}^\perp$, then $(x, \xi)_X = (0, A\xi)_Y \equiv 0$ for all $\xi \in$ dom(A), after (11). So if dom(A) is dense, then x = 0, and $\mathcal{A}^\perp$ is functional. Conversely, if dom(A) is not dense, there is some x ≠ 0 in the

[49] Not to be confused with the *dual* of A, similarly defined, but going from the dual Y' of Y to the dual X' of X. The notion of adjoint is specifically Hilbertian.

orthocomplement of dom(A) with respect to X, and hence a nontrivial pair $\{x, 0\} \in \mathcal{A}^{\perp}$. ◊

Remark A.11. After (11), the domain of A^* is made of all y such that the linear partial function $x \to (y, Ax)_Y$ be continuous on dom(A), with respect to the metric of X. This can be used as an alternative definition of A^*: first define its domain this way, then define the image A^*y as the Riesz vector of the linear continuous mapping obtained by extending $x \to (y, Ax)_Y$ to the closure of dom(A), i.e., all X, by continuity. ◊

If dom(A) is not dense, we can always consider A as being of type $X^0 \to Y$, where X^0 is the closure of dom(A) (equipped with the same scalar product as X, by restriction), and still be able to define an adjoint, now of type $Y \to X^0$.

Note that $(\mathcal{A}^{\perp})^{\perp}$ is the closure of $\mathcal{A}$. Therefore, if an operator A has an adjoint, and if dom(A^*) is dense, the closure of A is A^{**}, the adjoint of its adjoint. Therefore,

Proposition A.2. *Let* $A : X \to Y$ *be a linear operator with dense domain. If* dom(A^*) *is dense in* Y, A *is closable.*

Its closure is then A^{**}. This is how we proved that div was closable, in Chapter 5: The domain of its adjoint is dense because it includes all functions $\varphi \in C_0^{\infty}(D)$. Indeed, the map $b \to \int_D \varphi \operatorname{div} b \equiv -\int_D b \cdot \operatorname{grad} \varphi$ is $\mathbb{L}^2$-continuous for such a φ, due to the absence of a boundary term. As we see here, the weak divergence is simply the adjoint of the operator grad : $C_0^{\infty}(D) \to \mathbb{C}_0^{\infty}(D)$, the closure of which in $L^2(D) \times \mathbb{L}^2(D)$, in turn, is a *strict* restriction (beware!) of the weak gradient.

The reader is invited to play with these notions, and to prove what follows: The boundary of D being partitioned $S = S^h \cup S^b$ as in the main chapters, start from grad and − div, acting on smooth fields, but restricted to functions which vanish on S^h and to fields which vanish on S^b, respectively. Show that their closures (that one may then denote grad_h and $-\text{div}_b$) are mutual adjoints. Same thing with rot_h and rot_b.

REFERENCES

[AB] Y. Aharonov, D. Bohm: "Significance of Electromagnetic Potentials in the Quantum Theory", **Phys. Rev., 115** (1959), pp. 485–491.

[Bo] R.P. Boas: "Can we make mathematics intelligible?", **Amer. Math. Monthly, 88** (1981), pp. 727–731.

[B1] A. Bossavit: "The Exploitation of Geometrical Symmetry in 3-D Eddy-currents Computations", **IEEE Trans., MAG-21**, 6 (1985), pp. 2307–2309.

[B2] A. Bossavit: "Boundary value problems with symmetry, and their approximation by finite elements", **SIAM J. Appl. Math., 53,** 5 (1993), pp. 1352–1380.

[BS] W.E. Brittin, W.R. Smythe, W. Wyss: "Poincaré gauge in electrodynamics", **Am. J. Phys., 50**, 8 (1982), pp. 693–696.

[Bu] W.L. Burke: **Applied Differential Geometry**, Cambridge University Press (Cambridge, UK), 1985.

[Co] F.H.J. Cornish: "The Poincaré and related gauges in electromagnetic theory", **Am. J. Phys., 52,** 5 (1984), pp. 460–462.

[C&] Y. Crutzen, N.J. Diserens, C.R.I. Emson, D. Rodger: **Proc. European TEAM Workshop on Electromagnetic Field Analysis** (Oxford, England, 23–25 April 1990), Commission of the European Communities (Luxembourg), 1990.

[Ha] P.R. Halmos: **Measure Theory**, Van Nostrand (Princeton), 1950.

[Hl] P.R. Halmos: **Naive Set Theory**, Van Nostrand (Princeton), 1960.

[Hn] P. Henderson: **Functional Programming**, Prentice-Hall (Englewood Cliffs, NJ), 1980.

[Hr] R. Hersh: "Math Lingo vs. Plain English: Double Entendre", **Amer. Math. Monthly, 104**, 1 (1997), pp. 48–51.

[It] K. Ito (ed.): **Encyclopedic Dictionary of Mathematics** (2nd ed.), The MIT Press (Cambridge, MA), 1987.

[KB] A. Kaveh, S.M.R. Behfar: "Finite element ordering algorithms", **Comm. Numer. Meth. Engng., 11,** 12 (1995), pp. 995–1003.

[Kn] E.J. Konopinski: "What the electromagnetic vector potential describes", **Am. J. Phys., 46**, 5 (1978), pp. 499–502.

[Kr] J.-L. Krivine: "Fonctions, programmes et démonstration", **Gazette des Mathématiciens,** 60 (1994), pp. 63–73.

[Lo] M. Loève: **Probability Theory**, Van Nostrand (Princeton), 1955.

[Ma] I. Mayergoyz: **Nonlinear Diffusion of Electromagnetic Fields,** Academic Press (Boston), 1998.

[MU] S. Mazur, S. Ulam: "Sur les transformations isométriques d'espaces vectoriels normés", **C.R. Acad. Sci. Paris, 194** (1932), pp. 946–948.

[Me] B. Meyer: **Object-oriented Software Construction**, Prentice Hall (New York), 1988.

[Mr] B. Meyer : **Introduction to the Theory of Programming Languages**, Prentice-Hall (New York), 1990.

[NC] R.D. Nevels, K.J. Crowell: "A Coulomb gauge analysis of a wave scatterer", **IEE Proc.-H, 137**, 6 (1990), pp. 384–388.

[Pr] J.D. Pryce: **Basic Methods of Linear Functional Analysis**, Hutchinson & Co, Ltd. (London), 1973.

[R&] K.R. Richter, W.M. Rucker, O. Biro (Eds.): **4th IGTE Symposium & European TEAM 9** (Graz, Austria, 10–12 Oct. 1990), Technische Universität Graz (Graz), 1990.

[Sc] B. Schutz: **Geometrical Methods of Mathematical Physics**, Cambridge University Press (Cambridge, U.K.), 1980.

[Sk] B.-S.K. Skagerstam: "A note on the Poincaré gauge", **Am. J. Phys., 51,** 12 (1983), pp. 1148–1149.

[Sp] M. Spivak: **Calculus on Manifolds**, Benjamin, (New York), 1965.

[T&] L. Turner, H. Sabbagh et al. (Eds.): **Proceedings of the Toronto TEAM/ACES Workshop at Ontario Hydro** (25–26 Oct. 1990), Report ANL/FPP/TM-254, the Fusion Power Program at Argonne Nat. Lab., Argonne, Ill., 60439–4814.

[Yo] K. Yosida: **Functional Analysis**, Springer-Verlag (Berlin), 1965.